University of California, Los Angeles
Department of Statistics

Statistics 100C **Instructor: Nicolas Christou**

Multiple regression

The multiple regression model
Let Y be the response variable and let $x_1, x_2, \ldots, x_k$ the predictor variables. The multiple regression model in coordinate form is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \epsilon_i, i = 1, 2, \ldots, n.$$

It is very helpful to present this model in a more compact notation using matrix and vector form as follows.
Note: Matrices and vectors will be denoted with uppercase or lowercase boldface letter.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

Where:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix}, \text{ and } \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

In multiple regression the matrix $\mathbf{X}$ and the vector $\boldsymbol{\beta}$ are:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & x_{23} & \cdots & x_{2k} \\ 1 & x_{31} & x_{32} & x_{33} & \cdots & x_{3k} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} & \cdots & x_{nk} \end{pmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}.$$

Therefore,

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & x_{23} & \cdots & x_{2k} \\ 1 & x_{31} & x_{32} & x_{33} & \cdots & x_{3k} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Verify that $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \epsilon_i$.

Let $\mathbf{x_i'}$ be the ith row of $\mathbf{X}$. Write an expression for y_i.

We can also express $\mathbf{X}$ as follows:

$$\mathbf{X} = [\mathbf{1}, \mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_k}].$$

Note:
The dimensions of the vectors and matrices above are:

$$\mathbf{y} : n \times 1, \ \mathbf{X} : n \times (k+1), \ \boldsymbol{\beta} : (k+1) \times 1, \ \boldsymbol{\epsilon} : n \times 1.$$

Where k is the number of predictor variables in the model. Why do we need the extra column $\mathbf{1}$?

Consider the multiple regression model without the intercept β_0. What changes would you make in the notes above?

In simple regression $k = 1$ and the dimensions of $\mathbf{X}$ is $n \times 2$, and $\boldsymbol{\beta}$ is 2×1.

In simple regression, $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ the matrix $\mathbf{X}$ has the following form:

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \text{ and the vector } \boldsymbol{\beta} \text{ is } \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}.$$

Gauss-Markov conditions in matrix/vector form:
$\epsilon_1, \ldots, \epsilon_n$ are i.i.d. $N(0, \sigma)$. This is the same as (use matrix approach):

Find $E[\mathbf{y}]$ and $\mathrm{var}[\mathbf{y}]$.

As with simple regression, using the method of least squares we minimize the sum of the squared residuals in order to get the estimates of $\boldsymbol{\beta}$. More specifically the following quantity is minimized: $\sum_{i=1}^{n} \epsilon_i^2$, or in matrix form

$$\begin{aligned}
\min S &= \boldsymbol{\epsilon}'\boldsymbol{\epsilon} \quad \text{Now replace } \boldsymbol{\epsilon} \text{ with} \ldots \\
\min S &= \\
\min S &= \\
\min S &=
\end{aligned}$$

Now, to find the least squares estimate of the vector $\boldsymbol{\beta}$ we will take the derivative of S above with respect to the vector $\boldsymbol{\beta}$. At this point let's review some elements of matrix and vector differentiation.
Matrix and vector differentiation
Let

$$\boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{pmatrix}$$

be a p-dimensional vector and let $f(\boldsymbol{\theta})$ be a function of $\boldsymbol{\theta}$. When the derivative of $f(\boldsymbol{\theta})$ is taken with respect to the vector $\boldsymbol{\theta}$ we mean that the partial derivative of $f(\boldsymbol{\theta})$ is taken with respect to each element of $\boldsymbol{\theta}$, i.e.

$$\frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \frac{\partial f(\boldsymbol{\theta})}{\partial \theta_1} \\ \frac{\partial f(\boldsymbol{\theta})}{\partial \theta_2} \\ \vdots \\ \frac{\partial f(\boldsymbol{\theta})}{\partial \theta_p} \end{pmatrix}$$

We will present now two important results of matrix/vector differentiation.

1. Let $\boldsymbol{\theta}$ as define above and $\mathbf{c}' = (c_1, c_2, \ldots, c_p)$. If $f(\boldsymbol{\theta}) = \mathbf{c}'\boldsymbol{\theta}$ it follows that

$$\frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{c}.$$

2. Let $\mathbf{A}$ be a $p \times p$ symmetric matrix and let $\boldsymbol{\theta}$ as define above. Define now the quadratic expression $f(\boldsymbol{\theta}) = \boldsymbol{\theta}'\mathbf{A}\boldsymbol{\theta}$. It follows that

$$\frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 2\mathbf{A}\boldsymbol{\theta}.$$

The proof of results (1) and (2) above are left as an exercise.

Least squares estimates of $\boldsymbol{\beta}$
Now that we are familiar with matrix differentiation we can find the least squares estimates of $\boldsymbol{\beta}$ as follows by applying results (1) and (2) from the previous section. We minimize

$$\min S = \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}.$$

Verify that $\mathbf{y}'\mathbf{X}$ is a row vector:

Verify that $\mathbf{X}'\mathbf{X}$ is symmetric matrix and also write few elements of this matrix.
$\mathbf{X}'\mathbf{X} =$

What are the dimensions of $\mathbf{X}'\mathbf{X}$?

Now apply the two results from vector and matrix differentiation to get:

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \qquad\qquad\qquad\qquad = \mathbf{0}$$

to obtain the least squares normal equations

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y}.$$

Therefore,

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{pmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

We now continue to find the expected value and variance of $\hat{\boldsymbol{\beta}}$. Before we do this, we will review the mean and variance of random vectors.

Mean and variance of a random vector and properties

Let $\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$ be a random vector with $E\mathbf{Y} = \begin{pmatrix} EY_1 \\ EY_2 \\ \vdots \\ EY_n \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} = \boldsymbol{\mu}$. The variance covariance matrix of $\mathbf{Y}$ denoted with $var(\mathbf{Y})$ is defined as follows:

$$\begin{aligned}
var(\mathbf{Y}) &= E(\mathbf{Y}-\boldsymbol{\mu})(\mathbf{Y}-\boldsymbol{\mu})' \\
&= E\begin{pmatrix} Y_1-\mu_1 \\ Y_2-\mu_2 \\ \vdots \\ Y_n-\mu_n \end{pmatrix}(Y_1-\mu_1, Y_2-\mu_2, \ldots, Y_n-\mu_n) \\
&= E\begin{pmatrix} (Y_1-\mu_1)^2 & (Y_1-\mu_1)(Y_2-\mu_2) & \ldots & (Y_1-\mu_1)(Y_n-\mu_n) \\ (Y_2-\mu_2)(Y_1-\mu_1) & (Y_2-\mu_2)^2 & \ldots & (Y_2-\mu_2)(Y_n-\mu_n) \\ \vdots & \vdots & \vdots & \vdots \\ (Y_n-\mu_n)(Y_1-\mu_1) & (Y_n-\mu_n)(Y_2-\mu_2) & \ldots & (Y_n-\mu_n)^2 \end{pmatrix} \\
&= \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \ldots & \sigma_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \ldots & \sigma_n^2 \end{pmatrix} = \boldsymbol{\Sigma}.
\end{aligned}$$

So $\boldsymbol{\Sigma}$ is the variance covariance matrix of the vector $\mathbf{Y}$. It is symmetric and positive definite. Two important results are given below that will help us find the expected value and variance of $\hat{\boldsymbol{\beta}}$.

1. Expected value and variance os a linear combination of $\mathbf{Y}$. Let $\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}$ be a vector of constants and let $q = \mathbf{a}'\mathbf{Y}$. Then $E(q) = E(\mathbf{a}'\mathbf{Y}) = \mathbf{a}'E(\mathbf{Y}) = \mathbf{a}'\boldsymbol{\mu}$. The variance of q can be found as follows:

$$\begin{aligned}
var(q) = E(q-\mu_q)^2 &= E(\mathbf{a}'\mathbf{Y} - \mathbf{a}'\boldsymbol{\mu})^2 \\
&= E(\mathbf{a}'\mathbf{Y} - \mathbf{a}'\boldsymbol{\mu})(\mathbf{a}'\mathbf{Y} - \mathbf{a}'\boldsymbol{\mu}) \\
&= \mathbf{a}'E(\mathbf{Y}-\boldsymbol{\mu})(\mathbf{Y}-\boldsymbol{\mu})'\mathbf{a} \\
&= \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}.
\end{aligned}$$

Note: q is a scalar and therefore its variance should be a scalar and not a matrix. We can verify that $var(q) = \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}$ is 1×1.

2. Let $\mathbf{A}$ be a $p \times n$ matrix of contents. We will examine now $\mathbf{Q} = \mathbf{AY}$ is a $p \times 1$ vector and therefore its variance should be a $p \times p$ matrix. Let's find the expected value of $\mathbf{Q}$ first. $E(\mathbf{Q}) = E(\mathbf{AY}) = \mathbf{A}E(\mathbf{Y}) = \mathbf{A}\boldsymbol{\mu}$.

$$\begin{aligned} var(\mathbf{Q}) = E(\mathbf{Q} - E(\mathbf{Q}))(\mathbf{Q} - E(\mathbf{Q}))' &= E(\mathbf{AY} - \mathbf{A}\boldsymbol{\mu})(\mathbf{AY} - \mathbf{A}\boldsymbol{\mu})' \\ &= \mathbf{A}E(\mathbf{Y} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu})'\mathbf{A}' \\ &= \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'. \end{aligned}$$

We will use these results to find the mean and variance of $\hat{\boldsymbol{\beta}}$.
Mean and variance of $\hat{\boldsymbol{\beta}}$
We found earlier that $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. Therefore, using results (1) and (2) we can find the mean and variance of $\hat{\boldsymbol{\beta}}$. Before we do this, let's revisit the multiple regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ to state the assumptions in matrix form. The assumption that $E(\epsilon_i) = 0$ in matrix form is expressed as $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and therefore $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$. The assumption that $var(\epsilon_i) = \sigma^2$ and $cov(\epsilon_i, \epsilon_j) = 0$, for $i \neq j$ in matrix form is expressed as

$$var(\boldsymbol{\epsilon}) = \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix} = \sigma^2\mathbf{I}.$$

and therefore, $var(\mathbf{Y}) = var(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$, where $\mathbf{I}$ is the $n \times n$ identity matrix. We are ready now to find the mean and variance of $\hat{\boldsymbol{\beta}}$.

$$E(\hat{\boldsymbol{\beta}}) = E\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\right) =$$

This shows that $\hat{\boldsymbol{\beta}}$ is unbiased estimator of $\boldsymbol{\beta}$, i.e. $E(\hat{\beta}_0) = \beta_0, E(\hat{\beta}_1) = \beta_1, \dots, E(\hat{\beta}_k) = \beta_k$. Now we find the variance of $\hat{\boldsymbol{\beta}}$.

$$var(\hat{\boldsymbol{\beta}}) \quad = \quad var\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\right) =$$

Note: The matrix $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ plays the role of matrix $\mathbf{A}$ of result (2) of the previous section. Therefore, $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and since $(\mathbf{X}'\mathbf{X})$ is symmetric (and therefore its inverse is also symmetric) it follows that $\mathbf{A}' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$. Therefore,

$$var(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2 \begin{pmatrix} v_{00} & v_{01} & \dots & v_{0k} \\ v_{10} & v_{11} & \dots & v_{1k} \\ \vdots & \vdots & \vdots & \vdots \\ v_{k0} & v_{k1} & \dots & v_{kk} \end{pmatrix} = \begin{pmatrix} var(\hat{\beta}_0) & cov(\hat{\beta}_0, \hat{\beta}_1) & \dots & cov(\hat{\beta}_0, \hat{\beta}_k) \\ cov(\hat{\beta}_1, \hat{\beta}_0) & var(\hat{\beta}_1) & \dots & cov(\hat{\beta}_1, \hat{\beta}_k) \\ \vdots & \vdots & \ddots & \vdots \\ cov(\hat{\beta}_k, \hat{\beta}_0) & cov(\hat{\beta}_k, \hat{\beta}_1) & \dots & var(\hat{\beta}_k) \end{pmatrix},$$

where v_{ij} are the elements of the inverse of $\mathbf{X}'\mathbf{X}$. In another notation, we can expressed the variances and covariances of $\hat{\boldsymbol{\beta}}$ using the v_{ij} are the elements.
Find the following in terms of v_{ij}:
var($\hat{\beta}_0$)=

var($\hat{\beta}_1$) =

var($\hat{\beta}_k$) =

cov($\hat{\beta}_i, \hat{\beta}_j$) =

cor($\hat{\beta}_i, \hat{\beta}_j$) =

A different representation of the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.
Partition $\mathbf{X}$ and $\boldsymbol{\beta}$ as follows:
$\mathbf{X} = \left[\begin{array}{cc} \mathbf{1} & \mathbf{X_{(0)}} \end{array}\right]$ and $\boldsymbol{\beta} = \left[\begin{array}{c} \beta_0 \\ \boldsymbol{\beta_{(0)}} \end{array}\right]$.

What is $\mathbf{X_{(0)}}$?

What is $\boldsymbol{\beta_{(0)}}$?

Now express the model based on the partition above:
$\mathbf{y} =$

Now express the least squares estimator $\hat{\boldsymbol{\beta}}$ using this partition:
$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} =$

Fitted values

Recall that for the simple regression model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ the fitted values are given by $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, i = 1, \ldots, n$.

For the multiple regression model $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \epsilon_i, i = 1, 2, \ldots, n$ the fitted values are given by

$\hat{y}_i =$. In vector form: $\hat{y}_i =$

Some details:

$$\begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \\ \vdots \\ \hat{y}_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & x_{23} & \cdots & x_{2k} \\ 1 & x_{31} & x_{32} & x_{33} & \cdots & x_{3k} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{pmatrix}$$

Or $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$.

Replace $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ to get:

$\hat{\mathbf{y}} =$

which can be expressed as $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$, where $\mathbf{H} =$.

Note: $\mathbf{H}$ is the so called “hat” matrix with the following properties:

1. $\mathbf{H}$ is symmetric. Why?

2. $\mathbf{H}$ is idempotent: $\mathbf{HH} = \mathbf{H}$. Why?

3. What is $\mathbf{HX}$?

4. It follows from (3) that

   $\mathbf{H1} =$

   $\mathbf{Hx_1} =$

   $\vdots$

   $\mathbf{Hx_k} =$

5. $\text{tr}(\mathbf{H}) =$

   Note: The trace of a square matrix is the sum of its diagonal elements. A very useful property of the trace is the cyclical property:

   $\text{tr}(\mathbf{ABC}) = \text{tr}(\mathbf{BCA}) = \text{tr}(\mathbf{CAB}) \neq \text{tr}(\mathbf{BAC})$.

6. $h_{ii} = \mathbf{x_i'}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_i}$, where $\mathbf{x_i'}$ is the *ith* row of $\mathbf{X}$.

7. $h_{ij} = \mathbf{x_i'}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_j}$.

Mean and variance of $\hat{\mathbf{y}}$.

$E[\hat{\mathbf{y}}] =$

$\text{var}[\hat{\mathbf{y}}] =$

Some details:

$$\text{var}(\hat{\mathbf{Y}}) = \begin{pmatrix} \text{var}(\hat{y}_1) & \text{cov}(\hat{y}_1, \hat{y}_2) & \text{cov}(\hat{y}_1, \hat{y}_3) & \cdots & \cdots & \text{cov}(\hat{y}_1, \hat{y}_n) \\ \text{cov}(\hat{y}_2, \hat{y}_1) & \text{var}(\hat{y}_2) & \text{cov}(\hat{y}_2, \hat{y}_3) & \cdots & \cdots & \text{cov}(\hat{y}_2, \hat{y}_n) \\ \cdots & \cdots & \ddots & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \ddots & \cdots & \cdots \\ \text{cov}(\hat{y}_n, \hat{y}_1) & \text{cov}(\hat{y}_n, \hat{y}_2) & \text{cov}(\hat{y}_n, \hat{y}_3) & \cdots & \cdots & \text{var}(\hat{y}_n) \end{pmatrix}$$

$$\text{var}(\hat{\mathbf{Y}}) = \sigma^2 \begin{pmatrix} h_{11} & h_{12} & h_{13} & \cdots & \cdots & h_{1n} \\ h_{21} & h_{22} & h_{23} & \cdots & \cdots & h_{2n} \\ \cdots & \cdots & \ddots & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \ddots & \cdots & \cdots \\ h_{n1} & h_{n2} & h_{n3} & \cdots & \cdots & h_{nn} \end{pmatrix}$$

Where, h_{ij} is the ij_{th} element of the hat matrix $\mathbf{H}$. Therefore the variance of the i_{th} fitted value is $\text{var}(\hat{y}_i) = \sigma^2 h_{ii}$. Therefore, $h_{ii} \geq 0$. Note: We will show later that $\frac{1}{n} \leq h_{ii} \leq 1$.

Note: For simple regression we have seen that

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

and

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Residuals

The residual values are given by $e_i = y_i - \hat{y}_i, i = 1, \ldots, n$.

Some details:

$$\begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_n \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} - \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \\ \vdots \\ \hat{y}_n \end{pmatrix}$$

Or $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$. Replace $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ to get: $\mathbf{e} =$

Verify that $\mathbf{I} - \mathbf{H}$ is symmetric and idempotent.

Another expression for $\mathbf{e}$ (useful for distribution theory later) is the following: In $\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{y}$ replace $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ to get:
$\mathbf{e} =$.

Note: We can compute the residuals using $\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{y}$. The expression $\mathbf{e} = (\mathbf{I} - \mathbf{H})\boldsymbol{\epsilon}$ is used in distribution theory, not to compute the residuals, because $\boldsymbol{\epsilon}$ is not observed.

Mean and variance of $\mathbf{e}$

$E[\mathbf{e}] =$

$\text{var}[\mathbf{e}] =$

Some details:

$$\text{var}(\mathbf{e}) = \begin{pmatrix} \text{var}(e_1) & \text{cov}(e_1, e_2) & \text{cov}(e_1, e_3) & \cdots & \cdots & \text{cov}(e_1, e_n) \\ \text{cov}(e_2, e_1) & \text{var}(e_2) & \text{cov}(e_2, e_3) & \cdots & \cdots & \text{cov}(e_2, e_n) \\ \cdots & \cdots & \ddots & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \ddots & \cdots & \cdots \\ \text{cov}(e_n, e_1) & \text{cov}(e_n, e_2) & \text{cov}(e_n, e_3) & \cdots & \cdots & \text{var}(e_n) \end{pmatrix}$$

$$\text{var}(\mathbf{e}) = \sigma^2 \begin{pmatrix} 1 - h_{11} & -h_{12} & -h_{13} & \cdots & \cdots & -h_{1n} \\ -h_{21} & 1 - h_{22} & -h_{23} & \cdots & \cdots & -h_{2n} \\ \cdots & \cdots & \ddots & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \ddots & \cdots & \cdots \\ -h_{n1} & -h_{n2} & -h_{n3} & \cdots & \cdots & 1 - h_{nn} \end{pmatrix}$$

Where, $1 - h_{ij}$ is the ij_{th} element of the matrix $\mathbf{I} - \mathbf{H}$. Therefore the variance of the i_{th} residual is $\text{var}(e_i) = \sigma^2(1 - h_{ii})$. Therefore $h_{ii} \leq 1$.

Gauss-Markov theorem in multiple regression
In simple regression we showed that among all the linear unbiased estimators of β_0 and β_1 the least squares estimates are BLUE (Best Linear Unbiased Estimators) in the sense that they have the least variance. We will prove the same theorem in multiple regression. We will show that if $\mathbf{b}$ is another unbiased estimator of $\boldsymbol{\beta}$ its variance covariance matrix will exceed the variance covariance matrix of $\hat{\boldsymbol{\beta}}$ by a positive semidefinite matrix, i.e. $var(\mathbf{b}) \geq var(\hat{\boldsymbol{\beta}})$.

Proof
The OLS estimates are given by $\hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-\mathbf{1}}\mathbf{X'Y}$. Let $\mathbf{b} = \mathbf{M}^*\mathbf{Y}$ be another unbiased estimator of $\boldsymbol{\beta}$. Let's define $\mathbf{M} = \mathbf{M}^* - (\mathbf{X'X})^{-\mathbf{1}}\mathbf{X'}$ and therefore $\mathbf{M}^* = \mathbf{M} + (\mathbf{X'X})^{-\mathbf{1}}\mathbf{X'}$. Since $\mathbf{b}$ is unbiased it follows that $E(\mathbf{b}) = E(\mathbf{M}^*\mathbf{Y}) = \boldsymbol{\beta}$.
Find the condition on $\mathbf{M}$ that must hold so that $\mathbf{b}$ is an unbiased estimator of $\boldsymbol{\beta}$.

Now let's examine the variance of $\mathbf{b}$.

$$\text{var}(\mathbf{b}) = \text{var}(\mathbf{M}^*\mathbf{Y}) = \text{var}\left[\mathbf{M} + (\mathbf{X'X})^{-\mathbf{1}}\mathbf{X'}\right]\mathbf{Y}$$

Note: This has the form $\mathbf{AY}$ and therefore $\text{var}(\mathbf{AY}) = \mathbf{A\Sigma A'}$. Here, $\mathbf{\Sigma} =$.
Continue to show that $\text{var}(\mathbf{b}) = \sigma^2\mathbf{MM'} + \sigma^2(\mathbf{X'X})^{-\mathbf{1}}$

What is $\sigma^2(\mathbf{X'X})^{-\mathbf{1}}$?
Therefore what else do wee need to show?

Show that $\mathbf{MM'}$ is a positive semidefinite matrix.
Let $\mathbf{a}$ be a non-zero vector: Then $\mathbf{a'MM'a} \geq 0$. Why?

And therefore $\mathbf{MM'}$ is a positive definite matrix.

The Gauss-Markov theorem can be extended to a linear combination of the vector $\hat{\boldsymbol{\beta}}$. Let $\mathbf{a'}\hat{\boldsymbol{\beta}}$ be a linear combination of $\hat{\boldsymbol{\beta}}$. Find the variance of $\mathbf{a'}\hat{\boldsymbol{\beta}}$.

Now let $\mathbf{a'b}$ be another unbiased estimator of $\mathbf{a'}\boldsymbol{\beta}$. Find the variance of $\mathbf{a'b}$. Note: We have an expression for $\text{var}(\mathbf{b})$ from earlier note.

Therefore, $var(\mathbf{a'b}) \geq var(\mathbf{a'}\hat{\boldsymbol{\beta}})$.

Consider the special case where $\mathbf{a} = (0, 0, \ldots, 1, 0, \ldots, 0)'$. What result do we get here?

Multivariate normal distribution and distribution theory in multiple regression

We say that a random vector $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)'$ with mean vector $\boldsymbol{\mu}$ and variance covariance matrix $\boldsymbol{\Sigma}$ follows the multivariate normal distribution if its probability density function is given by

$$f(\mathbf{Y}) = \frac{1}{(2\pi)^{\frac{n}{2}}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{Y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y}-\boldsymbol{\mu})}, \tag{1}$$

and we write, $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Moment generating function

If $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ using the transformation $\mathbf{Y} = \boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{Z} + \boldsymbol{\mu}$ we find that $M_{\mathbf{Y}}(\mathbf{t}) = e^{\mathbf{t}'\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}}$.

Theorem 1

Let $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and let $\mathbf{A}$ be an $m \times n$ matrix of rank m and $\mathbf{c}$ be an $m \times 1$ vector.
Then $\mathbf{AY} + \mathbf{c} \sim N_m(\mathbf{A}\boldsymbol{\mu} + \mathbf{c}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$.
and $\mathbf{AY} \sim N_m(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$

Apply theorem 1 in multiple regression:

Theorem 2

Let $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Sub-vectors of $\mathbf{Y}$ follow the multivariate normal distribution and linear combinations of $Y_1, Y_2, \ldots, Y_n$ follow the univariate normal distribution. For example, suppose $\mathbf{Y}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$ are partitioned as follows $\mathbf{Y} = \begin{pmatrix} \mathbf{Q_1} \\ \mathbf{Q_2} \end{pmatrix}, \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$, where $\mathbf{Q_1}$ is $p \times 1$. It follows that $\mathbf{Q_1} \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ and $\mathbf{Q_2} \sim N_{n-p}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$. For a linear combination of $Y_1, Y_2, \ldots, Y_n$, i.e. $a_1Y_1 + a_2Y_2 + \ldots + a_nY_n = \mathbf{a}'\mathbf{Y}$, it follows that, $\mathbf{a}'\mathbf{Y} \sim N(\mathbf{a}'\boldsymbol{\mu}, \sqrt{\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}})$.

Example

Let $\mathbf{Y} = \left(\begin{array}{c} Y_1 \\ Y_2 \\ \hline Y_3 \\ Y_4 \\ Y_5 \end{array}\right), \boldsymbol{\mu} = \left(\begin{array}{c} \mu_1 \\ \mu_2 \\ \hline \mu_3 \\ \mu_4 \\ \mu_5 \end{array}\right), \boldsymbol{\Sigma} = \left(\begin{array}{cc|ccc} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} & \sigma_{15} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \sigma_{24} & \sigma_{25} \\ \hline \sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{34} & \sigma_{35} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 & \sigma_{45} \\ \sigma_{51} & \sigma_{52} & \sigma_{53} & \sigma_{54} & \sigma_5^2 \end{array}\right)$, then if $\mathbf{Q_1} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$,

it follows that $\mathbf{Q_1} \sim N\left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}\right]$.

Apply theorem 2 in multiple regression

Theorem 3: Statistical independence
Suppose $\mathbf{Y}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$ are partitioned as in theorem 2. We say that $\mathbf{Q_1}, \mathbf{Q_2}$ are statistically independent if and only if $\boldsymbol{\Sigma}_{12} = \mathbf{0}$.

Application
Suppose $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and define the following two vectors $\mathbf{Q_1} = \mathbf{AY}$ and $\mathbf{Q_2} = \mathbf{BY}$. Then, $\mathbf{Q_1}$ and $\mathbf{Q_2}$ are independent if $cov(\mathbf{Q_1}, \mathbf{Q_2}) = \mathbf{A\Sigma B'} = \mathbf{0}$. We stack the two vectors as follows: $\mathbf{Q} = \begin{pmatrix} \mathbf{Q_1} \\ \mathbf{Q_2} \end{pmatrix} = \begin{pmatrix} \mathbf{A} \\ \mathbf{B} \end{pmatrix} \mathbf{Y} = \mathbf{LY}$. Therefore using theorem 1 we find that $\mathbf{Q} \sim N(\mathbf{L}\boldsymbol{\mu}, \mathbf{L\Sigma L'})$ or $\mathbf{Q} \sim N\left[\begin{pmatrix} \mathbf{A} \\ \mathbf{B} \end{pmatrix} \boldsymbol{\mu}, \begin{pmatrix} \mathbf{A\Sigma A'} & \mathbf{A\Sigma B'} \\ \mathbf{B\Sigma A'} & \mathbf{B\Sigma B'} \end{pmatrix}\right]$, and we conclude that $\mathbf{Q_1}$ and $\mathbf{Q_2}$ are independent if and only if $\mathbf{A\Sigma B'} = \mathbf{0}$. Here, we can just simply find the covariance between the vectors $\mathbf{Q_1}$ and $\mathbf{Q_2}$ and if it is $\mathbf{0}$ then we conclude that $\mathbf{Q_1}$ and $\mathbf{Q_2}$ are independent.

Or we can find $\text{cov}(\mathbf{AY}, \mathbf{BY}) = \mathbf{A\Sigma B'}$. Why? If $\text{cov}(\mathbf{AY}, \mathbf{BY}) = \mathbf{0}$ then $\mathbf{AY}, \mathbf{BY}$ are independent.

Apply theorem 3 in multiple regression

Estimation using the method of maximum likelihood
Consider the multiple regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $E(\boldsymbol{\epsilon}) = \mathbf{0}$, $\text{var}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$, and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$.

1. Write the likelihood function.

2. Write the log likelihood function.

3. Estimate $\boldsymbol{\beta}$ using MML and show that the MLE estimator is the same as with OLS estimator.

4. Find the MLE estimator of σ^2.

5. Find $E[\hat{\sigma^2}] = E[\frac{1}{n}\mathbf{e}'\mathbf{e}]$. We can use $E[\frac{1}{n}\text{tr}(\mathbf{e}'\mathbf{e})]$. Why?

Conditional distributions using multivariate normal
Consider the bivariate normal distribution (see page 1). From theorem 1 it follows that $Y_1 \sim N(\mu_1, \sigma_1)$. This is also called the marginal probability distribution of Y_1. We want to find the conditional distribution of Y_2 given Y_1.

From the conditional probability law, $f_{Y_2|Y_1}(y_2|y_1) = \frac{f_{Y_1Y_2}(y_1,y_2)}{f_{Y_1}(y_1)}$, and after substituting the bivariate density and the marginal density it can be shown that the conditional probability density function of Y_2 given Y_1 is given by

$$f_{Y_2|Y_1}(y_2|y_1) = \frac{1}{\sqrt{\sigma_2^2(1-\rho)^2}\sqrt{2\pi}} exp\left[-\frac{1}{2}\left(\frac{Y_2-\mu_2-\rho\frac{\sigma_2}{\sigma_1}(Y_1-\mu_1)}{\sigma_2^2(1-\rho^2)}\right)\right].$$

We recognize that this is a normal probability density function with mean $\mu_{Y_2|Y_1} = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(Y_1-\mu_1)$ and variance $\sigma^2_{Y_2|Y_1} = \sigma_2^2(1-\rho^2)$.

In general:

Suppose that $\mathbf{Y}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$ are partitioned as follows $\mathbf{Y} = \begin{pmatrix} \mathbf{Y_1} \\ \mathbf{Y_2} \end{pmatrix}, \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$, and $\mathbf{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. It can be shown that the conditional distribution of $\mathbf{Y_1}$ given $\mathbf{Y_2}$ is also multivariate normal, $\mathbf{Y_1}|\mathbf{Y_2} \sim N(\boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2})$, where $\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{Y_2} - \boldsymbol{\mu}_2)$, and $\boldsymbol{\Sigma}_{1|2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$.

Apply these results in multiple regression.

Expectation of a quadratic expression using properties of the trace (example)
Consider the multiple regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Let $\mathbf{C}$ be a $m \times k+1$ matrix of constants and $\boldsymbol{\gamma}$ be a $m \times 1$ vector of constants.

Find $E\left[(\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\gamma})'(\mathbf{C}(\mathbf{X'X})^{-\mathbf{1}}\mathbf{C}')^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\gamma})\right]$.

1. What are the dimensions of $\mathbf{C}(\mathbf{X'X})^{-\mathbf{1}}\mathbf{C}'$.

2. Verify that the entire expression is a scalar and therefore you can use properties of the trace to find the expected value.

3. You will need the following:
   $\mathrm{var}[(\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\gamma})] =$

4. $E[(\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\gamma})] =$

5. Now use $E[\mathrm{tr(scalar)}]$ result to find the expectation of the expression.

Partial regression
Introduction
Consider the model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, \ldots, n$. We can obtain the estimator of β_1 using the following two-stage procedure:

1. Regress $\mathbf{y}$ on $\mathbf{1}$. This means we are using the model $y_i = \beta_0 + \epsilon_i, i = 1, \ldots n$. What is the residual vector here? Denote it with $\mathbf{y}^*$.

2. Regress $\mathbf{x}$ on $\mathbf{1}$. This means we are using the model $x_i = \delta_0 + \eta_i, i = 1, \ldots n$. What is the residual vector here? Denote it with $\mathbf{x}^*$.

3. Finally regress $\mathbf{y}^*$ on $\mathbf{x}^*$. Therefore the model we are using here is $y_i^* = \alpha_0 + \beta_1 x_i^* + \epsilon_i, i = 1, \ldots, n$. Verify that the estimate of the slope of this model is the usual $\hat{\beta}_1$ that we have seen in simple regression.

Generalize the previous result in multiple regression
$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Partition $\mathbf{X} = [\mathbf{X_1}\mathbf{X_2}]$ and therefore $\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta_1} \\ \boldsymbol{\beta_2} \end{pmatrix}$.
Here is an example:
Suppose $k = 5$ and let $\mathbf{X_1} = \begin{bmatrix} \mathbf{1} & \mathbf{x_1} & \mathbf{x_2} & \mathbf{x_3} \end{bmatrix}$ and $\mathbf{X_2} = \begin{bmatrix} \mathbf{x_4} & \mathbf{x_5} \end{bmatrix}$.
Then $\boldsymbol{\beta_1} =$

Then $\boldsymbol{\beta_2} =$

Verify

$$
\begin{aligned}
\mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\
\mathbf{y} &= \begin{bmatrix} \mathbf{1} & \mathbf{x_1} & \mathbf{x_2} & \mathbf{x_3} & \mathbf{x_4} & \mathbf{x_5} \end{bmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix} + \boldsymbol{\epsilon} \\
\mathbf{y} &= \beta_0 \begin{bmatrix} \mathbf{1} \end{bmatrix} + \beta_1 \begin{bmatrix} \mathbf{x_1} \end{bmatrix} + \beta_2 \begin{bmatrix} \mathbf{x_2} \end{bmatrix} + \beta_3 \begin{bmatrix} \mathbf{x_3} \end{bmatrix} + \beta_4 \begin{bmatrix} \mathbf{x_4} \end{bmatrix} + \beta_5 \begin{bmatrix} \mathbf{x_5} \end{bmatrix} + \boldsymbol{\epsilon} \\
\mathbf{y} &= \begin{bmatrix} \mathbf{1} & \mathbf{x_1} & \mathbf{x_2} & \mathbf{x_3} \end{bmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{bmatrix} \mathbf{x_4} & \mathbf{x_5} \end{bmatrix} \begin{pmatrix} \beta_4 \\ \beta_5 \end{pmatrix} + \boldsymbol{\epsilon} \\
\mathbf{y} &= \mathbf{X_1}\boldsymbol{\beta_1} + \mathbf{X_2}\boldsymbol{\beta_2} + \boldsymbol{\epsilon}
\end{aligned}
$$

Consider the following three models:

1. $\mathbf{y} = \mathbf{X_1}\boldsymbol{\beta_1} + \boldsymbol{\epsilon}$. (Short regression.)
   Then $\hat{\boldsymbol{\beta}}_\mathbf{1} =$

2. $\mathbf{y} = \mathbf{X_2}\boldsymbol{\beta_2} + \boldsymbol{\epsilon}$. (Short regression.)
   Then $\hat{\boldsymbol{\beta}}_\mathbf{2} =$

3. $\mathbf{y} = \mathbf{X_1}\boldsymbol{\beta_1} + \mathbf{X_2}\boldsymbol{\beta_2} + \boldsymbol{\epsilon}$. (Long regression, same as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.)
   How about the estimators of $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$ in this model? Are they the same as above?

Obtain $\hat{\boldsymbol{\beta}}_{\mathbf{1.2}}$ and $\hat{\boldsymbol{\beta}}_{\mathbf{2.1}}$ using partial regression. We will prove the following theorem:
The estimator of $\boldsymbol{\beta}_{\mathbf{2}}$ in the long regression can be obtained as follows.

1. Regress $\mathbf{y}$ on $\mathbf{X_1}$ and compute the residuals $\mathbf{y}^*$. Is $\mathbf{y}^*$ a vector or a matrix? How would you express $\mathbf{y}^*$ using the "residual maker" matrix for this model?

2. Regress each column of $\mathbf{X_2}$ on $\mathbf{X_1}$ and compute the residuals $\mathbf{X_2^*}$. Is $\mathbf{X_2^*}$ a vector or a matrix? How would you express $\mathbf{X_2^*}$ using the "residual maker" matrix for this model?

3. Finally regress $\mathbf{y}^*$ on $\mathbf{X_2^*}$ to obtain $\hat{\boldsymbol{\beta}}_{\mathbf{2.1}}$.

Proof
Recall the least squares normal equations from earlier material:

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y}.$$

Replace $\mathbf{X} = [\mathbf{X_1}\mathbf{X_2}]$ and $\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_{\mathbf{1}} \\ \boldsymbol{\beta}_{\mathbf{2}} \end{pmatrix}$ and multiply to get two normal equations:

Solve equation (1) in terms of $\hat{\boldsymbol{\beta}}_{\mathbf{1.2}}$

As an aside comment, what happens if $\mathbf{X}_{\mathbf{1}}'\mathbf{X_2} = \mathbf{0}$. The question here is to compare $\hat{\boldsymbol{\beta}}_{\mathbf{1.2}}$ under this condition, with $\hat{\boldsymbol{\beta}}_{\mathbf{1}}$ from the short regression of $\mathbf{y}$ on $\mathbf{X_1}$.

Back to the proof:
Replace $\hat{\boldsymbol{\beta}}_{\mathbf{1.2}}$ in the second normal equation to get

$$\mathbf{X_2'X_1(X_1'X_1)^{-1}X_1'y - X_2'X_1(X_1'X_1)^{-1}X_1'X_2}\hat{\boldsymbol{\beta}}_{\mathbf{2.1}} + \mathbf{X_2'X_2}\hat{\boldsymbol{\beta}}_{\mathbf{2.1}} = \mathbf{X_2'y}$$

Rearrange and solve for $\hat{\boldsymbol{\beta}}_{\mathbf{2.1}}$. Note: Your goal, when you rearrange these expressions, is to get the residual maker matrix $\mathbf{I - H_1 = I - X_1(X_1'X_1)^{-1}X_1'}$. Show that

$$\mathbf{X_2'(I - H_1)X_2}\hat{\boldsymbol{\beta}}_{\mathbf{2.1}} = \mathbf{X_2'(I - H_1)y}$$

Since $\mathbf{I - H_1}$ is idempotent insert another $\mathbf{I - H_1}$ on both sides of the equation above. What do you observe? Do you get residuals on both sides of the equations?

Two special cases of partial regression
Case A
Partition $\mathbf{X}$ and $\boldsymbol{\beta}$ as follows:
$\mathbf{X} = \begin{bmatrix} \mathbf{1} & \mathbf{X_{(0)}} \end{bmatrix}$ and $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \boldsymbol{\beta_{(0)}} \end{bmatrix}$.

The goal here is to use partial regression to find an expression for $\hat{\boldsymbol{\beta}}_{\mathbf{(0)}}$.
Think about the following: What plays the role of $\mathbf{X_1}$ and $\mathbf{X_2}$ here?

What is the $\mathbf{H_1}$ hat matrix in this situation?
$\mathbf{H_1} =$

Write in words the two stage procedure that will give the vector of the residuals. (See the partial regression theorem.)

Express mathematically the estimator of $\boldsymbol{\beta_{(0)}}$.

Note: The same estimator $\hat{\boldsymbol{\beta}}_{(\mathbf{0})}$ can be obtained using directly the least squares estimator
$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\boldsymbol{\beta}}_{(\mathbf{0})} \end{pmatrix} = \begin{pmatrix} n & \mathbf{1}'\mathbf{X}_{(\mathbf{0})} \\ \mathbf{X}_{(\mathbf{0})}'\mathbf{1} & \mathbf{X}_{(\mathbf{0})}'\mathbf{X}_{(\mathbf{0})} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{1}'\mathbf{y} \\ \mathbf{X}_{(\mathbf{0})}'\mathbf{y} \end{pmatrix}$.

To show this we need to use the inverse of a partitioned matrix as given below.

If all inverses exist,

$$\begin{pmatrix} \mathbf{A_{11}} & \mathbf{A_{12}} \\ \mathbf{A_{21}} & \mathbf{A_{22}} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{A_{11}^{-1}} + \mathbf{B_{12}}\mathbf{B_{22}^{-1}}\mathbf{B_{21}} & -\mathbf{B_{12}}\mathbf{B_{22}^{-1}} \\ -\mathbf{B_{22}^{-1}}\mathbf{B_{21}} & \mathbf{B_{22}^{-1}} \end{pmatrix} = \begin{pmatrix} \mathbf{C_{11}^{-1}} & -\mathbf{C_{11}^{-1}}\mathbf{C_{12}} \\ -\mathbf{C_{21}}\mathbf{C_{11}^{-1}} & \mathbf{A_{22}^{-1}} + \mathbf{C_{21}}\mathbf{C_{11}^{-1}}\mathbf{C_{12}} \end{pmatrix}$$

where
$\mathbf{B_{22}} = \mathbf{A_{22}} - \mathbf{A_{21}}\mathbf{A_{11}^{-1}}\mathbf{A_{12}}$
$\mathbf{B_{12}} = \mathbf{A_{11}^{-1}}\mathbf{A_{12}}$
$\mathbf{B_{21}} = \mathbf{A_{21}}\mathbf{A_{11}^{-1}}$

and
$\mathbf{C_{11}} = \mathbf{A_{11}} - \mathbf{A_{12}}\mathbf{A_{22}^{-1}}\mathbf{A_{21}}$
$\mathbf{C_{12}} = \mathbf{A_{12}}\mathbf{A_{22}^{-1}}$
$\mathbf{C_{21}} = \mathbf{A_{22}^{-1}}\mathbf{A_{21}}$

Case B

Begin with the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ and then add one extra predictor $\mathbf{z}$. Let c be the slope of this new predictor. Write the new model:

$\mathbf{y} =$

Apply the partial regression theorem here in order to find $\hat{c}$.

1. Regress ________ on ________ to get the residuals ________.

2. Regress ________ on ________ to get the residuals ________.

3. Finally to get $\hat{c}$ regress ________ on ________

Express mathematically $\hat{c}$.

$\hat{c} =$

A note on the error sum of squares
For the the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, show that $\mathbf{y}'\mathbf{y} = \mathbf{e}'\mathbf{e} + \hat{\mathbf{y}}'\hat{\mathbf{y}}$.
Use $\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{y}$ and $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$.

In the equation above, subtract $n\bar{y}^2$ in both sides and simplify to show that, as with simple regression, $SST = SSE + SSR$.

Coefficient of determination R^2.
It is defined as $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$.

Change in the error sum of squares when an extra predictor is added in the model
Here, we are comparing SSE_X (regression of $\mathbf{y}$ on $\mathbf{X}$) with SSE_{Xz} (regression of $\mathbf{y}$ on $\mathbf{X}$ and $\mathbf{z}$). These are the two models:
$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$
Give an expression of the residuals:
$\mathbf{e} =$

$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + c\mathbf{z} + \boldsymbol{\epsilon}$
Give an expression of the residuals:
$\mathbf{u} =$

Write the two normal equations using the long regression. Note: Denote the estimator of $\boldsymbol{\beta}$ with $\hat{\boldsymbol{\delta}}$ in the long regression.

Find $\hat{\boldsymbol{\delta}}$. Verify that $\hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\beta}} -$ [something that involves $\hat{c}$].

Now back to the vector of the residuals $\mathbf{u}$ in the long regression to show that $\mathbf{u} = \mathbf{e} - (\mathbf{I} - \mathbf{H})\mathbf{z}\hat{c}$.

$\mathbf{u} =$

Finally compute $SSE_{Xz} = \mathbf{u}'\mathbf{u} =$

Change in R^2. Show that $R^2_{Xz} \geq R^2_X$.

Partial correlations

Consider the model $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, i = 1, \ldots, n$.
Find the correlation between y and x_2 with x_1 in the model.

1. Regress $\mathbf{y}$ on $\mathbf{x_1}$ and compute the residuals $\mathbf{y}^*$.
   Give an expression for $\mathbf{y}^*$.

2. Regress $\mathbf{x_2}$ on $\mathbf{x_1}$ and compute the residuals $\mathbf{x_2^*}$.
   Give an expression for $\mathbf{x_2^*}$.

3. The square of the partial correlation coefficient is computed as follows:
   $$r^2_{yx_2|x_1} = \frac{\text{cov}^2(y^*, x_2^*)}{\text{var}(x_2^*)\text{var}(y^*)}$$
   Noting that $\bar{y^*} = 0$ and $\bar{x_2^*} = 0$ simplify this expression.

Apply the same idea to the following model with 5 predictors.
$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \epsilon_i, i = 1, \ldots, n$.
Find the correlation between y and x_5 with x_1, x_2, x_3, x_4 in the model.

1. Regress ________ on ________ and compute the residuals $\mathbf{y}^*$.

2. Regress ________ on ________ and compute the residuals $\mathbf{x_5^*}$.

3. Compute the square of the correlation between y and x_5 with x_1, x_2, x_3, x_4 in the model.

Another method:
$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \epsilon_i, i = 1, \ldots, n$.

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i, i = 1, \ldots, n$.

$$r^2_{yx_5|x_1,x_2,x_3,x_4} = \frac{SSE[y \text{ on } x_1, x_2, x_3, x_4] - SSE[y \text{ on } x_1, x_2, x_3, x_4, x_5]}{SSE[y \text{ on } x_1, x_2, x_3, x_4]}$$

Constrained least squares

This topic is connected with hypothesis testing, because under H_0 we have a constrained least squares problem to solve. Suppose we want to estimate the vector $\boldsymbol{\beta}$ of the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, subject to a set of m linear constraints of the form $\mathbf{C}\boldsymbol{\beta} = \boldsymbol{\gamma}$. For example, if $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$, $\mathbf{C} = \begin{pmatrix} 2 & -1 & 1 \\ 1 & 2 & 3 \end{pmatrix}$, and $\boldsymbol{\gamma} = \begin{pmatrix} 10 \\ 20 \end{pmatrix}$, then we have two linear constraints, so $m = 2$. These are the two constraints:

$$\begin{aligned} 2\beta_0 - \beta_1 + \beta_2 &= 10 \\ \beta_0 + 2\beta_1 3\beta_2 &= 20. \end{aligned}$$

We still want to minimize $\sum_{i=1}^n \epsilon_i^2 = \boldsymbol{\epsilon}'\boldsymbol{\epsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ but now the minimization is subject to the linear constraints $\mathbf{C}\boldsymbol{\beta} = \boldsymbol{\gamma}$. This can be done using the method of Lagrange multipliers. We need one Lagrange multiplier for each constraint. Let $\boldsymbol{\lambda} = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_m \end{pmatrix}$. Where else did we use Lagrange multipliers in the course?

Here is the minimization.

$$\min Q = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + 2\boldsymbol{\lambda}'(\mathbf{C}\boldsymbol{\beta} - \boldsymbol{\gamma})$$

Use matrix and vector differentiation to find the estimator of $\boldsymbol{\beta}$. We will denote this estimator with $\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}$. Take the partial derivative with respect to $\boldsymbol{\beta}$ and set it equal to zero and solve for $\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}$.

$\frac{\partial Q}{\partial \boldsymbol{\beta}} = \mathbf{0}$

We need to find $\boldsymbol{\lambda}$. Multiply both sides by $\mathbf{C}$ and solve for $\boldsymbol{\lambda}$.

Finally, find $\hat{\boldsymbol{\beta}}_{\boldsymbol{c}}$

Show that $\hat{\boldsymbol{\beta}}_{\boldsymbol{c}} = \hat{\boldsymbol{\beta}} -$ something

$\hat{\boldsymbol{\beta}}_{\boldsymbol{c}} =$

Fitted values of the constrained least squares
Show that $\hat{\mathbf{y}}_{\mathbf{c}} = \hat{\mathbf{y}} -$ something:

$\hat{\mathbf{y}}_{\mathbf{c}} =$

Residual values of the constrained least squares
Show that $\mathbf{e}_{\mathbf{c}} = \mathbf{e} +$ something.

$\mathbf{e}_{\mathbf{c}} = \mathbf{y} - \hat{\mathbf{y}}_{\mathbf{c}} =$

Error sum of squares of the constrained least squares
Show that $SSE_c = SSE +$ something ≥ 0

$SSE_c = \mathbf{e}_{\mathbf{c}}'\mathbf{e}_{\mathbf{c}} =$

Find $E[\mathbf{e}_{\mathbf{c}}'\mathbf{e}_{\mathbf{c}}]$. Where do you think we need this expectation?

A different method to find $\hat{\boldsymbol{\beta}}_{\mathbf{c}}$
Use the canonical form of the model: Solve for the constraint and transform the model to the canonical form.

We are using the constraint $\mathbf{C}\boldsymbol{\beta} = \boldsymbol{\gamma}$, which can be expressed as $\mathbf{C_1}\boldsymbol{\beta_1} + \mathbf{C_2}\boldsymbol{\beta_2} = \boldsymbol{\gamma}$. So the idea here is that we partition $\mathbf{C} = \begin{pmatrix} \mathbf{C_1} & \mathbf{C_2} \end{pmatrix}$ and $\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta_1} \\ \boldsymbol{\beta_2} \end{pmatrix}$
For example, suppose $k = 4$ and consider the two constraints

$$\begin{aligned} 3\beta_0 + 5\beta_1 + 4\beta_2 - 2\beta_3 + 3\beta_4 &= 5 \\ 2\beta_0 - 2\beta_1 + 4\beta_2 + 3\beta_3 - 5\beta_4 &= 8 \end{aligned}$$

which can be expressed as $\begin{pmatrix} 3 & 5 & 4 & -2 & 3 \\ 2 & -2 & 4 & 3 & -5 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} 5 \\ 8 \end{pmatrix}$ or

$$\begin{pmatrix} 3 & 5 \\ 2 & -2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} 4 & -2 & 3 \\ 4 & 3 & -5 \end{pmatrix} \begin{pmatrix} \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} 5 \\ 8 \end{pmatrix}.$$

In this example $\mathbf{C_1} = \begin{pmatrix} 3 & 5 \\ 2 & -2 \end{pmatrix}$, $\mathbf{C_2} = \begin{pmatrix} 4 & -2 & 3 \\ 4 & 3 & -5 \end{pmatrix}$ and $\boldsymbol{\beta_1} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$, $\boldsymbol{\beta_2} = \begin{pmatrix} \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}$.

Important note: You can partition $\mathbf{C}$ arbitrarily but either $\mathbf{C_1}$ or $\mathbf{C_2}$ must be non-singular.

Assume $\mathbf{C_1}$ is non-singular and solve for $\boldsymbol{\beta_1}$.
$\mathbf{C_1}\boldsymbol{\beta_1} + \mathbf{C_2}\boldsymbol{\beta_2} = \boldsymbol{\gamma}$

$\boldsymbol{\beta_1} =$

Now back to the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ and write it as
$\mathbf{y} = \mathbf{X_1}\boldsymbol{\beta}_1 + \mathbf{X_2}\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$ (must be partitioned according to the partition of $\mathbf{C}$.)

Substitute $\boldsymbol{\beta_1}$ in the model, rearrange, and express the model as $\mathbf{y_r} = \mathbf{X_{2r}}\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$.

What is $\mathbf{y_r}$?
What is $\mathbf{X_{2r}}$?

The estimator of $\boldsymbol{\beta}_2$ is $\hat{\boldsymbol{\beta}}_{\mathbf{2c}}$ that subvector of the $\hat{\boldsymbol{\beta}}_{\mathbf{c}}$ from the Lagrange multipliers method.

$\hat{\boldsymbol{\beta}}_{\mathbf{2c}} =$

$\hat{\boldsymbol{\beta}}_{\mathbf{1c}} =$

Note:
$\hat{\boldsymbol{\beta}}_{\mathbf{c}} = \begin{pmatrix} \hat{\boldsymbol{\beta}}_{\mathbf{1c}} \\ \hat{\boldsymbol{\beta}}_{\mathbf{2c}} \end{pmatrix}$.

Distribution of quadratic forms of normally distributed random variables

a. Let $\mathbf{Z} \sim N_n(\mathbf{0}, \mathbf{I})$.

   What is the distribution of Z_i?

   What is the distribution of Z_i^2?

   What is the distribution of $\sum_{i=1}^n Z_i^2$?

   Express $\sum_{i=1}^n Z_i^2$ in vector form.

b. Let $\mathbf{Z} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$.

   Repeat the previous questions. The goal here to find the distribution of $\sum_{i=1}^n \frac{Z_i^2}{\sigma^2}$?

   Express $\sum_{i=1}^n \frac{Z_i^2}{\sigma^2}$ in vector form.

c. Let $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \sigma^2\mathbf{I})$.
   Assume first that $\boldsymbol{\mu} = \mu\mathbf{1}$.

   What is the distribution of Y_i?

   Standardize Y_i so that it follows a $N(0, 1)$ distribution:

   What is the distribution of $\sum_{i=1}^n \frac{(Y_i-\mu)^2}{\sigma^2}$?

   What if the means are not the same, i.e. $E[Y_i] = \mu_i, i = 1, \ldots, n$?

   Express $\sum_{i=1}^n \frac{(Y_i-\mu_i)^2}{\sigma^2}$ in vector form.

d. Use (c) in regression: Find χ^2 distributions using $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$.

e. Suppose $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
Begin with the following transformation:
$\mathbf{V} = \boldsymbol{\Sigma}^{-\frac{1}{2}}[\mathbf{Y} - \boldsymbol{\mu}]$.

Useful notes on linear algebra
Eigenvalues (characteristic values) and eigenvectors (characteristic vectors):
Let $\mathbf{A}$ be a $k \times k$ square matrix and $\mathbf{I}$ be the $k \times k$ identity matrix. Then the scalars $\lambda_1, \lambda_2, \ldots, \lambda_k$ of the solution of $|\mathbf{A} - \lambda \mathbf{I}| = 0$ are called the eigenvalues or characteristic values. The equation $|\mathbf{A} - \lambda \mathbf{I}| = 0$ is a polynomial function of λ.

For each eigenvalue there is a corresponding eigenvector:
Let $\mathbf{A}$ be a $k \times k$ matrix and let λ be an eigenvalue of $\mathbf{A}$. If $\mathbf{x}$ is a nonzero $k \times 1$ vector such that $\mathbf{Ax} = \lambda \mathbf{x}$ we say that $\mathbf{x}$ is an eigenvector associated with the eigenvalue λ.

Now suppose $\mathbf{A}$ is a symmetric matrix. It can be decomposed as $\mathbf{A} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}'$, where $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_k)$, and $\mathbf{P} = [\mathbf{e_1}, \mathbf{e_2}, \ldots, \mathbf{e_k}]$ is the matrix of the normalized eigenvectors.

Orthogonal matrix:
A matrix $\mathbf{P}$ is said to be orthogonal if $\mathbf{PP}' = \mathbf{P}'\mathbf{P} = \mathbf{I}$, or $\mathbf{P}' = \mathbf{P}^{-1}$.

Square root matrix and inverse square root matrix of a symmetric matrix:

1. $\mathbf{A}^{\frac{1}{2}} = \mathbf{P}\boldsymbol{\Lambda}^{\frac{1}{2}}\mathbf{P}'$.
2. $\mathbf{A}^{\frac{1}{2}}\mathbf{A}^{\frac{1}{2}} = \mathbf{A}$.
3. $\mathbf{A}^{-\frac{1}{2}} = \mathbf{P}\boldsymbol{\Lambda}^{-\frac{1}{2}}\mathbf{P}'$.
4. $\mathbf{A}^{-\frac{1}{2}}\mathbf{A}^{-\frac{1}{2}} = \mathbf{A}^{-1}$.

Now back to the transformation $\mathbf{V} = \boldsymbol{\Sigma}^{-\frac{1}{2}}[\mathbf{Y} - \boldsymbol{\mu}]$.
Note: $\mathbf{V}$ is of the form $\mathbf{AY}$. Therefore use properties of random vectors to find the following:

$E[\mathbf{V}] =$

$\text{var}[\mathbf{V}] =$

What is the distribution of $\mathbf{V}$ and why? (A theorem from multivariate normal distribution.)

What is the distribution of $\mathbf{V}'\mathbf{V}$? Note: $\mathbf{V}$ looks the same as in (a).

Finally replace $\mathbf{V} = \boldsymbol{\Sigma}^{-\frac{1}{2}}[\mathbf{Y} - \boldsymbol{\mu}]$ to get:

f. Use (e) to find a χ^2 distribution associated with $\hat{\boldsymbol{\beta}}$.

First we note that $\hat{\boldsymbol{\beta}} \sim N_{k+1}[\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}]$.

Since $\mathbf{X}'\mathbf{X}$ is symmetric, $(\mathbf{X}'\mathbf{X})^{\mathbf{-1}}$ is also symmetric. Why do we need this information?

What transformation would you choose here? Note: When $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ we use the inverse square root matrix: $\mathbf{V} = \boldsymbol{\Sigma}^{-\frac{1}{2}}[\mathbf{Y} - \boldsymbol{\mu}]$. (From $\boldsymbol{\Sigma}$ to $\boldsymbol{\Sigma}^{-\frac{1}{2}}$.)

Therefore $\mathbf{V} =$

Find $E[\mathbf{V}]$.

Find var$[\mathbf{V}]$.

Therefore $\frac{\mathbf{V}'\mathbf{V}}{\sigma^2} \sim$.

Replace now $\mathbf{V} = (\mathbf{X}'\mathbf{X})^{\frac{1}{2}}[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}]$ to get:

g. Use the results from (d) and (f) to show that $\frac{(n-k-1)s_e^2}{\sigma^2} \sim \chi^2_{n-k-1}$.

From (d) we have $\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{\sigma^2} \sim$

Continue by adding/subtracting $\mathbf{X}\hat{\boldsymbol{\beta}}$

h. A different method to show $\frac{(n-k-1)s_e^2}{\sigma^2} \sim \chi^2_{n-k-1}$.

1. Let $\mathbf{Y} \sim N_n(\mathbf{0}, \mathbf{I})$. If $\mathbf{P}$ is orthogonal matrix (i.e. $\mathbf{P}'\mathbf{P} = \mathbf{I}$) then $\mathbf{Z} = \mathbf{P}'\mathbf{Y} \sim N_n(\mathbf{0}, \mathbf{I})$. Why?

2. Let $\mathbf{Y} \sim N_n(\mathbf{0}, \mathbf{I})$, and let $\mathbf{A}$ be a symmetric and idempotent matrix. Then $\mathbf{Y}'\mathbf{A}\mathbf{Y} \sim \chi^2_r$, where r is the number of eigenvalues of $\mathbf{A}$ equal to 1. The other $n-r$ eigenvalues are equal to zero (see previous handout).

   First show that a symmetric idempotent matrix has eigenvalues 0 or 1.

   Now for the proof:

   $\mathbf{Y}'\mathbf{A}\mathbf{Y} =$ (use spectral decomposition on $\mathbf{A}$.)

i. Use the results from (h) to show that $\frac{(n-k-1)s_e^2}{\sigma^2} \sim \chi^2_{n-k-1}$.

Efficiency of the least squares estimator $\hat{\boldsymbol{\beta}}$
Multi parameter case

Let $\hat{\boldsymbol{\theta}}$ be the estimator of $\boldsymbol{\theta}$ ($p \times 1$ vector). For example, in the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ we have $\boldsymbol{\theta} = \left(\quad \right)$. We say that $\hat{\boldsymbol{\theta}}$ is an efficient estimator of $\boldsymbol{\theta}$ if

1. $E[\hat{\boldsymbol{\theta}}] = \boldsymbol{\theta}$.
2. $\mathrm{var}[\hat{\boldsymbol{\theta}}] = \mathbf{I}^{-1}(\boldsymbol{\theta})$, where $\mathbf{I}(\boldsymbol{\theta})$ is the information matrix.

The information matrix is computed as follows:

$$\mathbf{I}(\boldsymbol{\theta}) = -E\begin{bmatrix} \frac{\partial^2 \mathrm{lnL}}{\partial \beta_0^2} & \frac{\partial^2 \mathrm{lnL}}{\partial \beta_0 \partial \beta_1} & \cdots & \frac{\partial^2 \mathrm{lnL}}{\partial \beta_0 \partial \beta_k} & \frac{\partial^2 \mathrm{lnL}}{\partial \beta_0 \partial \sigma^2} \\ \frac{\partial^2 \mathrm{lnL}}{\partial \beta_1 \partial \beta_0} & \frac{\partial^2 \mathrm{lnL}}{\partial \beta_1^2} & \cdots & \frac{\partial^2 \mathrm{lnL}}{\partial \beta_1 \partial \beta_k} & \frac{\partial^2 \mathrm{lnL}}{\partial \beta_1 \partial \sigma^2} \\ \vdots & \vdots & \ddots & \vdots & \\ \frac{\partial^2 \mathrm{lnL}}{\partial \beta_k \partial \beta_0} & \frac{\partial^2 \mathrm{lnL}}{\partial \beta_k \partial \beta_1} & \cdots & \frac{\partial^2 \mathrm{lnL}}{\partial \beta_k^2} & \frac{\partial^2 \mathrm{lnL}}{\partial \beta_k \partial \sigma^2} \\ \frac{\partial^2 \mathrm{lnL}}{\partial \sigma^2 \partial \beta_0} & \frac{\partial^2 \mathrm{lnL}}{\partial \sigma^2 \partial \beta_1} & \cdots & \frac{\partial^2 \mathrm{lnL}}{\partial \sigma^2 \partial \beta_k} & \frac{\partial^2 \mathrm{lnL}}{\partial \sigma^{2(2)}} \end{bmatrix} = -E\begin{bmatrix} \frac{\partial^2 lnL}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta'}} & \frac{\partial^2 lnL}{\partial \boldsymbol{\beta} \partial \sigma^2} \\ \frac{\partial^2 lnL}{\partial \sigma^2 \partial \boldsymbol{\beta'}} & \frac{\partial^2 lnL}{\partial \sigma^{2(2)}} \end{bmatrix} = -E\left[\frac{\partial^2 lnL}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta'}}\right]$$

Write the log-likelihood function based on the normality assumption

$lnL =$

Find the following

$$\frac{\partial lnL}{\partial \boldsymbol{\beta}} =$$

$$\frac{\partial^2 lnL}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta'}} =$$

$$\frac{\partial^2 lnL}{\partial \boldsymbol{\beta} \partial \sigma^2} =$$

$$\frac{\partial^2 lnL}{\partial \sigma^{2(2)}} =$$

Find the information matrix: $-E\left[\frac{\partial^2 lnL}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta'}}\right]$.

Note that the information matrix is block diagonal. Therefore, the inverse of the upper left block matrix of the information matrix is $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ which is the same as the variance of $\hat{\boldsymbol{\beta}}$. We also know that $\hat{\boldsymbol{\beta}}$ is unbiased estimator of $\boldsymbol{\beta}$ and therefore we conclude that $\hat{\boldsymbol{\beta}}$ is an efficient estimator of $\boldsymbol{\beta}$.

Centered model

Consider the usual multiple regression model is $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. If we partition $\mathbf{X} = (\mathbf{1}, \mathbf{X_{(0)}})$ and $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \boldsymbol{\beta_{(0)}} \end{pmatrix}$ we can write the model as $\mathbf{y} = \beta_0\mathbf{1} + \mathbf{X_{(0)}}\boldsymbol{\beta_{(0)}} + \boldsymbol{\epsilon}$. Suppose now we add and subtract $\frac{1}{n}\mathbf{1}\mathbf{1}'\mathbf{X_{(0)}}\boldsymbol{\beta_{(0)}}$. Rearrange and complete the next model equation using this information. Note: In your answer you should include the mean sweeper matrix which centers the predictors:

$$\mathbf{y} = \mathbf{1}\left(\beta_0 + \frac{1}{n}\mathbf{1}'\mathbf{X_{(0)}}\boldsymbol{\beta_{(0)}}\right) + \underline{\hspace{6cm}} + \boldsymbol{\epsilon}$$

Is $\left(\beta_0 + \frac{1}{n}\mathbf{1}'\mathbf{X_{(0)}}\boldsymbol{\beta_{(0)}}\right)$ a scalar? Write it as a function of the sample means of the predictors

Replace this scalar with γ_0 and denote the centered predictors with $\mathbf{Z}$

Finally we get

$$\mathbf{y} =$$

This is called the centered model, where,

$$\mathbf{1} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \ \boldsymbol{\beta_{(0)}} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}, \ \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}, \ \mathbf{Z} = \begin{pmatrix} z_{11} & z_{12} & z_{13} & \cdots & z_{1k} \\ z_{21} & z_{22} & z_{23} & \cdots & z_{2k} \\ z_{31} & z_{32} & z_{33} & \cdots & z_{3k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & z_{n3} & \cdots & z_{nk} \end{pmatrix}.$$

Another way to get the centered model above is to look at the regression model equation for each y_i.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \epsilon_i.$$

What do we need to do? This is similar to the simple regression centered model.

Estimation of the centered model:
Before we start, find the following: $\mathbf{1}'\mathbf{Z}$ and $\mathbf{Z}'\mathbf{1}$. These results will be helpful in the estimation of the centered model.

Now write the normal equations using the centered model and estimate γ_0 and $\boldsymbol{\beta}_{(\mathbf{0})}$.

Does the estimator of $\boldsymbol{\beta}_{(\mathbf{0})}$ remind you any previous material we discussed?

We want to show next that the fitted values and residuals of the centered model are the same with those of the non centered model.

Express the fitted values of the centered model using $\hat{\gamma}_0$ and $\hat{\boldsymbol{\beta}}_{(\mathbf{0})}$.

How about the residuals? Are they the same?

Distribution theory
Assume that $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. What is the distribution of $\mathbf{y}$ using the centered model?

Find a χ^2 distribution originated from the distribution above.

We can use this χ^2 distribution to show that $\frac{(n-k-1)S_e^2}{\sigma^2} \sim \chi^2_{n-k-1}$. What do you suggest to do, beginning from the χ^2_n distribution above?

Hypothesis testing
F test for the general linear hypothesis
Consider the regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \epsilon_i, \quad i = 1, \ldots, n.$$

Also, $E(\epsilon_i) = 0$, $E(\epsilon_i \epsilon_j) = 0$ for $i \neq j$, and $\mathrm{var}(\epsilon_i) = \sigma^2$.

Suppose we want to test the following linear hypotheses:

a. $H_0 : \beta_2 = 0$
   $H_a : \beta_2 \neq 0$
b. $H_0 : \beta_2 = 3$
   $H_a : \beta_2 \neq 3$
c. $H_0 : \beta_1 = \beta_5$ or $\beta_1 - \beta_5 = 0$
   $H_a : \beta_1 \neq \beta_5$ or $\beta_1 - \beta_5 \neq 0$
d. $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$
   $H_a :$ At least one $\beta_i \neq 0$

   This hypothesis can be expressed as:
   $H_0 : \boldsymbol{\beta}_{(\mathbf{0})} = \mathbf{0}$
   $H_a : \boldsymbol{\beta}_{(\mathbf{0})} \neq \mathbf{0}$
   where, $\boldsymbol{\beta}_{(\mathbf{0})} = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5))'$
e. $H_0 : (\beta_2, \beta_5)' = \mathbf{0}$
   $H_a : (\beta_2, \beta_5)' \neq \mathbf{0}$

All these hypotheses above can be expressed through the general linear hypothesis:
$H_0 : \mathbf{C}\boldsymbol{\beta} - \boldsymbol{\gamma} = \mathbf{0}$
$H_a : \mathbf{C}\boldsymbol{\beta} - \boldsymbol{\gamma} \neq \mathbf{0}$

Let's find the matrix $\mathbf{C}$ and the vector $\boldsymbol{\gamma}$ for each one of the hypotheses (a)-(e) above:

a. $\mathbf{C} =$ $\qquad\qquad$ $\boldsymbol{\gamma} =$

b. $\mathbf{C} =$ $\qquad\qquad$ $\boldsymbol{\gamma} =$

c. $\mathbf{C} =$ $\qquad\qquad$ $\boldsymbol{\gamma} =$

d. This is also called the overall significance of the model.

   $\mathbf{C} =$ $\qquad\qquad$ $\boldsymbol{\gamma} =$

e. We are testing here whether the two parameters (β_2, β_5) are significant simultaneously.

   $\mathbf{C} =$ $\qquad\qquad$ $\boldsymbol{\gamma} =$

   Note: In general $\mathbf{C}$ is $m \times (k+1)$ matrix and $\boldsymbol{\gamma}$ is $m \times 1$ vector.

Test statistic
We will develop here the F statistic for the general linear hypothesis. $H_0 : \mathbf{C}\boldsymbol{\beta} = \boldsymbol{\gamma}$. We can find this F statistic using the following three methods:

A. Ratio of two independent χ^2 random variables.

B. Extra sum of squares principle.

C. Likelihood ratio test.

A. Ratio of two independent χ^2 random variables.
We are testing $\mathbf{C}\boldsymbol{\beta} = \boldsymbol{\gamma}$, or $\mathbf{C}\boldsymbol{\beta} - \boldsymbol{\gamma} = \mathbf{0}$.

Consider the estimator of $\mathbf{C}\boldsymbol{\beta} - \boldsymbol{\gamma} = \mathbf{0}$. What is it?

Find the distribution of $\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\gamma}$ under H_0:

$\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\gamma} \sim$

Find a χ^2 distribution originated from this multivariate normal distribution:
Use the result from quadratic forms:

Which other χ^2 distribution do we need?

Noting that $\hat{\boldsymbol{\beta}}$ and S_e^2 are independent we construct the F statistic using the definition of F distribution: Ratio of two independent χ^2 random variables, each one divided by its degrees of freedom:

Approximately what is the expected value of this F statistic?

Reject H_0 if $F > F_{1-\alpha;m,n-k-1}$.

B Extra sum of squares principle.
The test statistic above can also be computed using the full and reduced models.
Under H_0 we have a constrained least squares model (reduced model). We estimate the model under H_0 to obtain the constrained error sum of squares:

$SSE_R =$

We also compute the error sum of squares under no restrictions:

$SSE_F =$

The F statistic is then computed as follows: $\frac{\frac{SSE_R - SSE_F}{df_R - df_F}}{\frac{SSE_F}{df_F}}$

What is SSE_R?

What is SSE_F?

Therefore what is $SSE_R - SSE_F$?

What is df_R?

What is df_F?

Therefore $df_R - df_F$?

Finally show that the F statistic using this method is exactly the same as in A (ratio of two independent χ^2 random variables.)

Example:
Suppose, $k = 5$ and we are testing $H_0 : \beta_4 = 0, \beta_5 = 0$.
The reduced model is the regression of $\mathbf{y}$ on the predictors ________________
The full model is the regression of $\mathbf{y}$ on the predictors ________________
$m =$

Hypothesis testing using the t statistic

Consider the hypothesis $H_0 : \beta_1 = 0$ against the alternative $H_a : \beta_1 \neq 0$. This hypothesis, as we have seen above, can be expressed in the form $H_0 : \mathbf{C}\boldsymbol{\beta} = \boldsymbol{\gamma}$ and therefore it can be tested using $\frac{(\mathbf{C}\hat{\boldsymbol{\beta}}-\boldsymbol{\gamma})'\left[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}'\right]^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}}-\boldsymbol{\gamma})}{mS_e^2}$. This test statistic follows $F_{m,n-k-1}$, with $m = 1$. This suggests that there is an equivalent t statistic. (It should be $t^2_{n-k-1} = F_{1,n-k-1}$.) Let's explore this result.

Since $\hat{\boldsymbol{\beta}} \sim N_{k+1}\left[\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right]$ it follows that under $H_0 : \beta_1 = 0$

$\hat{\beta}_1 \sim$ (Note: The elements of $(\mathbf{X}'\mathbf{X})^{-1}$ are denoted with v_{ij}.).

And also, $\frac{(n-k-1)S_e^2}{\sigma^2} \sim \chi^2_{n-k-1}$.

Use the two distributions above to construct a t statistic:

Now, let's see if the square of this t statistic is the same as the F statistic given by $\frac{(\mathbf{C}\hat{\boldsymbol{\beta}}-\boldsymbol{\gamma})'\left[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}'\right]^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}}-\boldsymbol{\gamma})}{mS_e^2}$.

We are testing $H_0 : \beta_1 = 0$. Find the following:

$\mathbf{C} =$

$\boldsymbol{\gamma} =$

$\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\gamma} =$

How about $\left[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}'\right]^{-1}$?

$m =$

With the above, $\frac{(\mathbf{C}\hat{\boldsymbol{\beta}}-\boldsymbol{\gamma})'\left[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}'\right]^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}}-\boldsymbol{\gamma})}{mS_e^2} =$

This is exactly the same as t^2, because $\frac{\hat{\beta}_1}{S_e\sqrt{v_{11}}} \sim t_{n-k-1}$.

C. Likelihood ratio test.
We can obtain the same F statistic using the likelihood ratio test. We begin with the likelihood ratio test $\Lambda = \frac{L(\hat{\omega})}{L(\hat{\Omega})}$, where $L(\hat{\omega})$ and $L(\hat{\Omega})$ are the maximized likelihood functions under the restrictions imposed by the null hypothesis and under no restrictions respectively. We reject H_0 if $\Lambda < k$. We need to find the MLEs with and without restrictions. We assume $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$:

1. Under H_0 we have a constrained least squares problem and therefore the estimator is the constrained least squares estimator that we have seen in previous lectures, $\hat{\boldsymbol{\beta}}_{\mathbf{c}}$. The MLE of σ^2 under the null hypothesis is $\hat{\sigma_0^2} = \frac{\mathbf{e_c'e_c}}{n}$.
2. Under no restrictions the estimator is the usual least squares estimator $\hat{\boldsymbol{\beta}}$. The MLE of σ^2 under no restrictions is $\hat{\sigma_1^2} = \frac{\mathbf{e'e}}{n}$.

Now back to the likelihood ratio test. Reject H_0 if

$$\Lambda = \frac{L(\hat{\omega})}{L(\hat{\Omega})} < k$$

$$\Lambda = \frac{\left(2\pi\hat{\sigma_0^2}\right)^{-\frac{n}{2}} e^{-\frac{1}{2\hat{\sigma_0^2}}\mathbf{e_c'e_c}}}{\left(2\pi\hat{\sigma_1^2}\right)^{-\frac{n}{2}} e^{-\frac{1}{\hat{\sigma_1^2}}\mathbf{e'e}}} < k, \quad \text{but } n\hat{\sigma_0^2} = \mathbf{e_c'e_c} \text{ and } n\hat{\sigma_1^2} = \mathbf{e'e}$$

$$\Lambda = \left(\frac{\hat{\sigma_1^2}}{\hat{\sigma_0^2}}\right)^{\frac{n}{2}} \frac{e^{-\frac{n}{2}}}{e^{-\frac{n}{2}}} < k, \quad \text{substitute } \hat{\sigma_0^2} \text{ and } \hat{\sigma_1^2}$$

$$\Lambda = \frac{\mathbf{e'e}}{\mathbf{e_c'e_c}} < k^{\frac{2}{n}}, \quad \text{but } \mathbf{e_c'e_c} = \mathbf{e'e} + (\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\gamma})' \left[\mathbf{C}(\mathbf{X'X})^{-1}\mathbf{C}'\right]^{-1} (\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\gamma})$$

$$\Lambda = \frac{\mathbf{e'e}}{\mathbf{e'e} + (\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\gamma})' \left[\mathbf{C}(\mathbf{X'X})^{-1}\mathbf{C}'\right]^{-1} (\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\gamma})} < k^{\frac{2}{n}}, \quad \text{divide by } \mathbf{e'e}$$

$$\Lambda = \frac{1}{1 + \frac{(\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\gamma})' \left[\mathbf{C}(\mathbf{X'X})^{-1}\mathbf{C}'\right]^{-1} (\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\gamma})}{\mathbf{e'e}}} < k^{\frac{2}{n}}, \quad \text{if } H_0 \text{ is true, then } \Lambda \approx 1.$$

$$\Lambda = \frac{(\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\gamma})' \left[\mathbf{C}(\mathbf{X'X})^{-1}\mathbf{C}'\right]^{-1} (\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\gamma})}{\mathbf{e'e}} > k^{-\frac{2}{n}} - 1$$

$$\Lambda = \frac{(\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\gamma})' \left[\mathbf{C}(\mathbf{X'X})^{-1}\mathbf{C}'\right]^{-1} (\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\gamma})}{(n-k-1)S_e^2} > k^{-\frac{2}{n}} - 1$$

$$\Lambda = \frac{(\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\gamma})' \left[\mathbf{C}(\mathbf{X'X})^{-1}\mathbf{C}'\right]^{-1} (\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\gamma})}{mS_e^2} > \left(k^{-\frac{2}{n}} - 1\right) \frac{n-k-1}{m}$$

$$\Lambda = \frac{(\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\gamma})' \left[\mathbf{C}(\mathbf{X'X})^{-1}\mathbf{C}'\right]^{-1} (\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\gamma})}{mS_e^2} > k'$$

This is exactly the same F statistic we found above using the ratio of two independent χ^2 random variables or using the extra sum of squares principle. To find the rejection region k' we choose the significance level α and continue as follows: $P(F_{m,n-k-1} > k') = \alpha$. Therefore, $k' = F_{1-\alpha,m,n-k-1}$.

Power calculations in multiple regression

The previous test statistics are the central F statistics. They follow a central F distribution (or a central t distribution) under H_0. When H_0 is not true the test statistic follows a non central F distribution (or a non central t distribution). We need the non central distributions to compute the power of the test. The power of a test is the probability of rejecting the null hypothesis when the null is not true.

We will first need few results on the non central χ^2 distribution.

1. Definition: If $\mathbf{Y} \sim N(\boldsymbol{\mu}, \mathbf{I})$ we say that $\mathbf{Y'Y} \sim \chi^2_n(NCP = \boldsymbol{\mu'\mu})$.
2. If $\mathbf{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then $\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{Y} - \boldsymbol{\mu}) \sim N(\mathbf{0}, \mathbf{I})$ and therefore $(\mathbf{Y} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \boldsymbol{\mu}) \sim \chi^2_n$ (this is called the central χ^2_n).
3. Now consider this transformation: $\boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{Y}$. This follows $N(\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\mu}, \mathbf{I})$. This has the form of (1) and therefore, $\mathbf{Y}'\boldsymbol{\Sigma}^{-1}\mathbf{Y} \sim \chi^2_n(NCP = \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu})$ (this is called the non central χ^2_n with non centrality parameter $\theta = \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$).
4. Another example: Suppose $\mathbf{Y} \sim N(\boldsymbol{\mu}, \sigma^2\mathbf{I})$. Then this will look like (1) if we use the transformation $\frac{\mathbf{Y}}{\sigma}$. Now, $\frac{\mathbf{Y}}{\sigma} \sim N(\frac{\boldsymbol{\mu}}{\sigma}, \mathbf{I})$. Therefore, $\frac{\mathbf{Y'Y}}{\sigma^2} \sim \chi^2_n(NCP = \frac{\boldsymbol{\mu'\mu}}{\sigma^2})$. This can also be expressed as $\frac{\sum_{i=1}^n Y_i^2}{\sigma^2} \sim \chi^2_n(NCP = \frac{\sum_{i=1}^n \mu_i^2}{\sigma^2})$.
5. Moment generating function of non central χ^2 distribution. In general, a random variable Q that has m.g.f. of the form $M_Q(t) = (1-2t)^{-\frac{n}{2}} e^{\theta\frac{t}{1-2t}}$ follows the χ^2 distribution with non centrality parameter θ. We write $Q \sim \chi^2(n, \theta)$. For example, in part (4) $\theta = \frac{\boldsymbol{\mu'\mu}}{\sigma^2}$.
6. If U, V are independent with $U \sim \chi^2_{n_1}(NCP = \theta)$ and $V \sim \chi^2_{n_2}$ then
$\frac{\frac{U}{n_1}}{\frac{V}{n_2}} \sim F_{n_1,n_2}(NCP = \theta)$.

Back to hypothesis testing: $H_0 : \mathbf{C}\boldsymbol{\beta} = \boldsymbol{\gamma}$. What is the distribution of the F statistic under H_0 and when H_0 is not true?

$$\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\gamma} \sim N\left[\mathbf{0}, \sigma^2\mathbf{C}(\mathbf{X'X})^{-1}\mathbf{C}'\right] \quad \text{under } H_0 \text{ and therefore,}$$

$$\frac{(\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\gamma})'\left[\mathbf{C}(\mathbf{X'X})^{-1}\mathbf{C}'\right]^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\gamma})}{\sigma^2} \sim \chi^2_m.$$

And the following has the central F distribution:

$$\frac{(\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\gamma})'\left[\mathbf{C}(\mathbf{X'X})^{-1}\mathbf{C}'\right]^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\gamma})}{ms_e^2} \sim F_{m,n-k-1}.$$

If H_0 is not true then

$$\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\gamma} \sim N\left[\mathbf{C}\boldsymbol{\beta} - \boldsymbol{\gamma}, \sigma^2\mathbf{C}(\mathbf{X'X})^{-1}\mathbf{C}'\right]$$

Let's find a non central χ^2 distribution. We need to transform the previous distribution into $N\left[\text{something}, \mathbf{I}\right]$. Therefore, we need to multiply $\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\gamma}$ by what?

Finally we conclude that

$$\frac{(\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\gamma})'\left[\mathbf{C}(\mathbf{X'X})^{-1}\mathbf{C}'\right]^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\gamma})}{\sigma^2} \sim \chi^2_m.$$

with non centrality parameter $\theta =$

Therefore.

$$\frac{(\mathbf{C}\hat{\boldsymbol{\beta}}-\boldsymbol{\gamma})'\left[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{C}'\right]^{-\mathbf{1}}(\mathbf{C}\hat{\boldsymbol{\beta}}-\boldsymbol{\gamma})}{ms_e^2} \sim F_{m,n-k-1},$$

with non centrality parameter

$$\theta = \frac{(\mathbf{C}\boldsymbol{\beta}-\boldsymbol{\gamma})'\left[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{C}'\right]^{-\mathbf{1}}(\mathbf{C}\boldsymbol{\beta}-\boldsymbol{\gamma})}{\sigma^2}.$$

Note: To compute the non centrality parameter we need values for the $\boldsymbol{\beta}$ vector and for σ^2. .

Compute the power:

$$\begin{aligned} 1-\beta &= P\left[\text{reject} \quad H_0 \quad \text{when it is false}\right] \\ &= P\left[F_{m,n-k-1}(\theta) > F_{1-\alpha;m,n-k-1}\right]. \end{aligned}$$

Hypothesis testing
Example 1:
Consider the following data with 155 observations of soil concentrations on lead, cadmium, copper, and zinc.

```
a <- read.table("http://www.stat.ucla.edu/~nchristo/statistics100c/soil_complete.txt", header=TRUE)

#Response variable:
y <- a$lead

#Predictors:
x1 <- a$cadmium
x2 <- a$copper
x3 <- a$zinc
```

You will test the hypothesis
$H_0 : \beta_2 = 0$
$H_a : \beta_2 \neq 0$
using three different methods:

a. F test for the general linear hypothesis.

b. t test.

c. Extra sum of squares principle.

You will need some of the following information:

1. Vector $\hat{\boldsymbol{\beta}}$:

```
> beta_hat
            [,1]
ones   7.2010775
x1   -14.1775608
x2    -0.1865834
x3     0.4251507
```

2. Inverse of the matrix $\mathbf{X}'\mathbf{X}$:

```
> solve(t(X) %*% X)
              ones            x1            x2            x3
ones  4.494339e-02  8.964032e-03 -1.557152e-03 -1.023796e-05
x1    8.964032e-03  4.718014e-03 -3.741216e-04 -1.957493e-05
x2   -1.557152e-03 -3.741216e-04  9.581905e-05 -2.323903e-06
x3   -1.023796e-05 -1.957493e-05 -2.323903e-06  3.565241e-07
```

3. Variance of y:

```
> var(y)
[1] 12392.15
```

4 . Regression of $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} \epsilon_i$.

```
> qf <- lm(y ~ x1+x2+x3)
> summary(qf)
Call:
lm(formula = y ~ x1 + x2 + x3)

(OTHER INFORMATION FROM THE OUTPUT WAS REMOVED).

Residual standard error: 25.79063 on 151 degrees of freedom
```

5. Regression of $y_i = \beta_0 + \beta_1 x_{1i} + \beta_3 x_{3i} \epsilon_i$.

```
> qr <- lm(y ~ x1+x3)
> summary(qr)
Call:
lm(formula = y ~ x1 + x3)

(OTHER INFORMATION FROM THE OUTPUT WAS REMOVED).

Residual standard error: 25.75210 on 152 degrees of freedom
```

6. Some percentiles:

```
> qt(0.975, 151)
[1] 1.975799
>
> qt(0.975, 152)
[1] 1.975694
>
> qf(0.95, 1, 151)
[1] 3.903781
>
> qf(0.95, 1, 152)
[1] 3.903366
>
> qf(0.95, 3, 151)
[1] 2.664504
>
> qf(0.95, 2, 152)
[1] 3.055558
```

Example 2:
Access the data:

```
a <- read.table("http://www.stat.ucla.edu/~nchristo/statistics100C/body_fat.txt", header=TRUE)

#Response variable:
a$y

#Predictors:
x1 <- a$x6
x2 <- a$x7
x3 <- a$x8
x4 <- a$x9
x5 <- a$x10
```

a. $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$
   H_a : At least one $\beta_i \neq 0$

```
ones <- rep(1, nrow(a))

#Construct the design matrix X:
X <- as.matrix(cbind(ones,x1,x2,x3,x4,x5))

#Estimate the beta vector:
beta_hat <- solve(t(X) %*% X) %*% t(X) %*% a$y

#Define the matrix C:
q <- c(0,1,0,0,0,0,0,0,1,0,0,0,0,0,0,1,0,0,0,0,0,0,1,0,0,0,0,0,0,1)
C <- matrix(q,5,6,byrow=TRUE)

#Vector gamma:
g <- c(0,0,0,0,0)

#Compute se^2:
se2 <- (t(a$y) %*% a$y - t(beta_hat) %*% t(X) %*% a$y) / (nrow(a)-5-1)

#Compute the F statistics:
F <- (t(C%*%beta_hat-g)%*%solve(C%*%solve(t(X)%*%X)%*%t(C))%*%(C %*%beta_hat-g)) / (5*se2)
```

b. $H_0 : (\beta_2, \beta_5)' = \mathbf{0}$
   $H_a : (\beta_2, \beta_5)' \neq \mathbf{0}$

```
#Define matrix C and vector gamma:
q <- c(0,0,1,0,0,0,0,0,0,0,0,1)
C <- matrix(q, 2,6, byrow=TRUE)

g <- c(0,0)

F <- (t(C%*%beta_hat-g)%*%solve(C%*%solve(t(X)%*%X)%*%t(C))%*%(C%*%beta_hat-g)) / (2*se2)
```

c. $H_0 : \beta_2 = 0$
   $H_a : \beta_2 \neq 0$

```
#Define matrix C and vector gamma:

q <- c(0,0,1,0,0,0)
C <- matrix(q, 1,6, byrow=TRUE)

g <- c(0)

F <- (t(C%*%beta_hat-g)%*%solve(C%*%solve(t(X)%*%X)%*%t(C))%*%(C%*% beta_hat-g)) / (1*se2)
```

Extra notes on hypothesis testing

Consider the multiple regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Assume the Gauss-Markov conditions hold and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. Suppose there are 5 predictors. Answer the following questions:

a. Test the overall significance of the model using the F test for the general linear hypothesis:

$$F = \frac{(\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\gamma})' \left[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}'\right]^{-1} (\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\gamma})}{m s_e^2},$$

What is $\mathbf{C}$?

What is $\boldsymbol{\gamma}$?

b. Use the F test for the general linear hypothesis to test:

$$\begin{aligned} H_0 &: (\beta_1, \beta_3)' = \mathbf{0} \\ H_a &: (\beta_1, \beta_3)' \neq \mathbf{0} \end{aligned}$$

What is $\mathbf{C}$?

What is $\boldsymbol{\gamma}$?

c. Test the hypothesis in question (b) using the extra sum of squares principle.

d. Consider the full model. Test the hypothesis that $\beta_4 = 0$ using the t statistic.

e. Consider the following two linear constraints:

$$\begin{aligned} H_0 &: \beta_1 + \beta_2 - 3\beta_5 = 2, \beta_3 + \beta_4 + \beta_5 = 3 \\ H_a &: \text{Not true} \end{aligned}$$

Use the canonical form of the model to test the hypothesis with the method of extra sum of squares.

f. Repeat (e) using the F test for the general linear hypothesis.

Confidence intervals
Find a confidence interval for β_1

We have seen that $\hat{\boldsymbol{\beta}} \sim N_{k+1}\left[\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right]$.

Therefore, $\hat{\beta}_1 \sim$ (Note: The elements of $(\mathbf{X}'\mathbf{X})^{-1}$ are denoted with v_{ij}.).

And also, $\frac{(n-k-1)S_e^2}{\sigma^2} \sim \chi^2_{n-k-1}$.

Use the two distributions above to construct a ratio that follows a t distribution. This will be the pivot to help us construct a $1-\alpha$ confidence interval for β_1:

Now use $P[-t_{\frac{\alpha}{2};n-k-1} < t_{n-k-1} < t_{\frac{\alpha}{2};n-k-1}] = 1-\alpha$ to construct the interval.

In general: Find a $1-\alpha$ confidence interval for $\mathbf{a}'\boldsymbol{\beta}$

Find the distribution of $\mathbf{a}'\hat{\boldsymbol{\beta}}$

And use $\frac{(n-k-1)S_e^2}{\sigma^2} \sim \chi^2_{n-k-1}$ to construct a a ratio that follows a t distribution:

Finally a $1-\alpha$ confidence interval for $\mathbf{a}'\boldsymbol{\beta}$ is given by

Prediction intervals
Consider the multiple regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Assume the Gauss-Markov conditions hold and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. Therefore, $y_i = \mathbf{x_i'}\boldsymbol{\beta} + \epsilon_i$. Suppose we want to predict a new value y_0.

Based on the model above we should have $y_0 = \mathbf{x_0'}\boldsymbol{\beta} + \epsilon_0$. Note: y_0 is not one of $y_1, \ldots, y_n$.

What is $\mathbf{x_0'}$?

What would the predictor of y_0 based on the least squares estimator $\hat{\boldsymbol{\beta}}$?

$\hat{y}_0 =$

Now consider the error of prediction $y_0 - \hat{y}_0$.

Find $E[y_0 - \hat{y}_0]$

Find $\text{var}[y_0 - \hat{y}_0]$. Are y_0 and $\hat{y}_0$ independent? Why?

What is the distribution of $y_0 - \hat{y}_0$?

Finally use $\frac{(n-k-1)S_e^2}{\sigma^2} \sim \chi^2_{n-k-1}$ to construct a a ratio that follows a t distribution and a $1-\alpha$ prediction interval for y_0.

Confidence interval for $E[y_0] = \mathbf{x_0'}\boldsymbol{\beta}$. Here, begin with the distribution of $\hat{y}_0$.

Prediction problem - revisited

Consider the multiple regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. The Gauss-Markov conditions hold. A prediction of a new value of the response variable is given by $\hat{y}_0 = \mathbf{x_0}'\hat{\boldsymbol{\beta}}$. Here we will find the predicted value by minimizing the mean-squared error, $E(Y_0 - \hat{Y}_0)^2$. We will follow these steps:

1. Assume that the predictor assumption is a linear combination of the observed y_i's, i.e. $\hat{y}_0 = \mathbf{w}'\mathbf{y}$, where $\mathbf{w}$ are unknown constants. We want this predictor to be unbiased, $E[\hat{Y}_0] = \mathbf{x_0}'\boldsymbol{\beta}$. Find the constraint that must hold in order for $\hat{y}_0$ to be unbiased.

2. Now we will minimize the the mean-squared error, $E(y_0 - \hat{y}_0)^2$. It will be easier however to minimize $\mathrm{var}(y_0 - \hat{y}_0)$. The two minimizations are the same. Why?

3. Now minimize $\mathrm{var}(y_0 - \hat{y}_0)$ subject to the constraint you found in part (1). Your goal is to find the weights $\mathbf{w}$ and finally show that $\hat{y}_0 = \mathbf{w}'\mathbf{y}$ is equal to $\mathbf{x_0}'\hat{\boldsymbol{\beta}}$ (therefore it will be the same as the predictor based on least squares).

Centering and scaling

Multicollinearity is a problem in multiple regression when some predictors are highly correlated with other predictors. We will explain multicollinearity in the next pages, but first we will discuss the centered and scaled model.

We discussed the centered model earlier. The centered model can be expressed as

$$y_i = \gamma_0 + \beta_1(x_{i1} - \bar{x}_1) + \beta_2(x_{i2} - \bar{x}_2) + \ldots + \beta_k(x_{ik} - \bar{x}_k) + \epsilon_i.$$

The centered and scaled model can be obtained as follows. We multiply and divide each centered predictor in the previous equation by the quantity $\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}$ for $j = 1 \ldots k$ to get:

$$\begin{aligned}
y_i &= \gamma_0 + \beta_1 \frac{\sqrt{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2}}{\sqrt{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2}}(x_{i1} - \bar{x}_1) + \ldots + \beta_k \frac{\sqrt{\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2}}{\sqrt{\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2}}(x_{ik} - \bar{x}_k) + \epsilon_i \text{ or} \\
y_i &= \gamma_0 + \delta_1 \frac{z_{i1}}{\sqrt{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2}} + \ldots + \delta_k \frac{z_{ik}}{\sqrt{\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2}} + \epsilon_i \text{ or} \\
y_i &= \gamma_0 + \delta_1 Z_{si1} + \ldots + \delta_k Z_{sik} + \epsilon_i,
\end{aligned}$$

This is the centered and scaled model, where, $\delta_j = \beta_j \sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}$ and $Z_{sij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}}$.

Connection between β_j and δ_j.

We can also expressed the centered and scaled model in matrix/vector form. Define the matrix $\mathbf{D}$ as follows:

$$\mathbf{D} = \begin{pmatrix}
\sqrt{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2} & 0 & 0 & \ldots & 0 \\
0 & 0 & 0 & \ldots & 0 \\
\vdots & \ddots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & 0 & \sqrt{\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2}
\end{pmatrix}$$

The centered model in matrix/vector form was expressed as

$$\mathbf{y} = \gamma_0 \mathbf{1} + \mathbf{Z}\boldsymbol{\beta}_{(\mathbf{0})} + \boldsymbol{\epsilon}$$

Use the matrix $\mathbf{D}$ to transformed the centered model into its centered and scaled form:

$$\mathbf{y} = \gamma_0 \mathbf{1} + \mathbf{Z}\boldsymbol{\beta}_{(\mathbf{0})} + \boldsymbol{\epsilon} \text{ (transform } Z \text{ and } \boldsymbol{\beta}_{(\mathbf{0})})$$

What are the centered and scaled predictors?

What is the vector of the slopes of the centered and scaled predictors?

Estimation of the centered and scaled model:
First we see that
$\mathbf{1'Z_s} =$ and
$\mathbf{Z_s'1} =$

This will help next when we write the normal equations. Complete the normal equations:

$$\begin{pmatrix} \end{pmatrix} \begin{pmatrix} \hat{\gamma}_0 \\ \hat{\boldsymbol{\delta}}_{\mathbf{(0)}} \end{pmatrix} = \begin{pmatrix} \\ \end{pmatrix} \mathbf{y}$$

But, $\mathbf{1'Zs} = \mathbf{0}$ and $\mathbf{Zs'1} = \mathbf{0}$. Therefore,

$$\begin{pmatrix} \hat{\gamma}_0 \\ \hat{\boldsymbol{\delta}}_{\mathbf{(0)}} \end{pmatrix} =$$

It follows that, $\hat{\gamma}_0 = \bar{y}$ and $\hat{\boldsymbol{\delta}}_{\mathbf{(0)}} = (\mathbf{Zs'Zs})^{-1}\mathbf{Zs'y}$. But, $\mathbf{Zs'Zs} = \mathbf{R}$ (correlation matrix of the k predictors - see next page). Finally, $\hat{\boldsymbol{\delta}}_{\mathbf{(0)}} = \mathbf{R}^{-1}\mathbf{Zs'y}$.

Find $E[\hat{\boldsymbol{\delta}}_{\mathbf{(0)}}] =$

Find $\text{var}[\hat{\boldsymbol{\delta}}_{\mathbf{(0)}}] =$

Summary:

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \;\; \text{non-centered model} \\ \mathbf{y} &= \beta_0\mathbf{1} + \mathbf{X}_{(\mathbf{0})}\boldsymbol{\beta}_{\mathbf{(0)}} + \boldsymbol{\epsilon} \\ \\ \mathbf{y} &= \gamma_0\mathbf{1} + \mathbf{Z}\boldsymbol{\beta}_{\mathbf{(0)}} + \boldsymbol{\epsilon} \;\; \text{centered model} \\ \\ \mathbf{y} &= \gamma_0\mathbf{1} + \mathbf{Z_s}\boldsymbol{\delta}_{\mathbf{(0)}} + \boldsymbol{\epsilon} \;\; \text{centered and scaled model} \end{aligned}$$

These three models have the same
fitted values
residuals
SSR
SSE
R^2
F statistic for testing the overall significance of the model
t statistics for testing individual β_i coefficients.

Useful notes:
$\hat{\delta}_j = \hat{\beta}_j\sqrt{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}$, therefore $\hat{\beta}_j = \frac{\hat{\delta}_j}{\sqrt{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}}$.

$\hat{\boldsymbol{\delta}}_{\mathbf{(0)}} = \mathbf{D}\hat{\boldsymbol{\beta}}_{\mathbf{(0)}}$, therefore, $\hat{\boldsymbol{\beta}}_{\mathbf{(0)}} = \mathbf{D}^{-1}\hat{\boldsymbol{\delta}}_{\mathbf{(0)}}$.

$\hat{\beta}_0 = \hat{\gamma}_0 - \bar{\mathbf{x}}'\hat{\boldsymbol{\beta}}_{\mathbf{(0)}} = \bar{y} - \bar{x}_1\hat{\beta}_1 - \ldots - \bar{x}_k\hat{\beta}_k$.

We can verify that $\mathbf{Zs}'\mathbf{Zs} = \mathbf{R}$ from the following:

$$\mathbf{Zs}'\mathbf{Zs} = \begin{pmatrix} \frac{x_{11}-\bar{x}_1}{\sqrt{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2}} & \frac{x_{21}-\bar{x}_1}{\sqrt{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2}} & \cdots & \cdots & \frac{x_{n1}-\bar{x}_1}{\sqrt{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2}} \\ \frac{x_{12}-\bar{x}_2}{\sqrt{\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)^2}} & \frac{x_{22}-\bar{x}_2}{\sqrt{\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)^2}} & \cdots & \cdots & \frac{x_{n2}-\bar{x}_2}{\sqrt{\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)^2}} \\ \frac{x_{13}-\bar{x}_3}{\sqrt{\sum_{i=1}^{n}(x_{i3}-\bar{x}_3)^2}} & \frac{x_{23}-\bar{x}_3}{\sqrt{\sum_{i=1}^{n}(x_{i3}-\bar{x}_3)^2}} & \cdots & \cdots & \frac{x_{n3}-\bar{x}_3}{\sqrt{\sum_{i=1}^{n}(x_{i3}-\bar{x}_3)^2}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{x_{1k}-\bar{x}_k}{\sqrt{\sum_{i=1}^{n}(x_{ik}-\bar{x}_k)^2}} & \frac{x_{2k}-\bar{x}_k}{\sqrt{\sum_{i=1}^{n}(x_{ik}-\bar{x}_k)^2}} & \cdots & \cdots & \frac{x_{nk}-\bar{x}_k}{\sqrt{\sum_{i=1}^{n}(x_{ik}-\bar{x}_k)^2}} \end{pmatrix} \begin{pmatrix} \frac{x_{11}-\bar{x}_1}{\sqrt{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2}} & \frac{x_{12}-\bar{x}_2}{\sqrt{\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)^2}} & \cdots & \cdots & \frac{x_{1k}-\bar{x}_k}{\sqrt{\sum_{i=1}^{n}(x_{ik}-\bar{x}_k)^2}} \\ \frac{x_{21}-\bar{x}_1}{\sqrt{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2}} & \frac{x_{22}-\bar{x}_2}{\sqrt{\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)^2}} & \cdots & \cdots & \frac{x_{2k}-\bar{x}_k}{\sqrt{\sum_{i=1}^{n}(x_{ik}-\bar{x}_k)^2}} \\ \frac{x_{31}-\bar{x}_1}{\sqrt{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2}} & \frac{x_{32}-\bar{x}_2}{\sqrt{\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)^2}} & \cdots & \cdots & \frac{x_{3k}-\bar{x}_k}{\sqrt{\sum_{i=1}^{n}(x_{ik}-\bar{x}_k)^2}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{x_{n1}-\bar{x}_1}{\sqrt{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2}} & \frac{x_{n2}-\bar{x}_2}{\sqrt{\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)^2}} & \cdots & \cdots & \frac{x_{nk}-\bar{x}_k}{\sqrt{\sum_{i=1}^{n}(x_{ik}-\bar{x}_k)^2}} \end{pmatrix}$$

Therefore,

$$\mathbf{Zs}'\mathbf{Zs} = \begin{pmatrix} 1 & r_{12} & r_{13} & r_{14} & \cdots & \cdots & r_{1k} \\ r_{21} & 1 & r_{23} & r_{24} & \cdots & \cdots & r_{2k} \\ r_{31} & r_{32} & 1 & r_{34} & \cdots & \cdots & r_{3k} \\ \vdots & \vdots & \vdots & \vdots & \cdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \cdots & \cdots & \vdots \\ r_{k1} & r_{k2} & r_{k3} & r_{k4} & \cdots & \cdots & 1 \end{pmatrix} = \mathbf{R}.$$

Multicollinearity - theory

Using the centered and scaled model we showed that the variance covariance matrix of $\hat{\boldsymbol{\delta}}_{\mathbf{(0)}}$ is equal to $var(\hat{\boldsymbol{\delta}}_{\mathbf{(0)}}) = \sigma^2\mathbf{R}^{-1}$. We want to find an expression for $var(\hat{\delta}_1)$. This is equal to $\sigma^2 \times$(position (1,1) of $\mathbf{R^{-1}}$). First we will partition $\mathbf{R}$ as follows:

$$\mathbf{R} = \begin{pmatrix} 1 & r_{12} & r_{13} & r_{14} & \ldots & \ldots & r_{1k} \\ r_{21} & 1 & r_{23} & r_{24} & \ldots & \ldots & r_{2k} \\ r_{31} & r_{32} & 1 & r_{34} & \ldots & \ldots & r_{3k} \\ \vdots & \vdots & \vdots & \vdots & \ldots & \ldots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ldots & \ldots & \vdots \\ r_{k1} & r_{k2} & r_{k3} & r_{k4} & \ldots & \ldots & 1 \end{pmatrix} = \begin{pmatrix} 1 & \mathbf{r}' \\ \mathbf{r} & \mathbf{R_{22}} \end{pmatrix}.$$

To find the inverse of the partitioned matrix we will use the following result from linear algebra:

$$\begin{pmatrix} \mathbf{A_{11}} & \mathbf{A_{12}} \\ \mathbf{A_{21}} & \mathbf{A_{22}} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{C_{11}}^{-1} & -\mathbf{C_{11}}^{-1}\mathbf{C_{12}} \\ -\mathbf{C_{21}}\mathbf{C_{11}}^{-1} & \mathbf{A_{22}}^{-1} + \mathbf{C_{21}}\mathbf{C_{11}}^{-1}\mathbf{C_{12}} \end{pmatrix}.$$

where,

$$\begin{aligned} \mathbf{C_{11}} &= \mathbf{A_{11}} - \mathbf{A_{12}}\mathbf{A_{22}}^{-1}\mathbf{A_{21}} \\ \mathbf{C_{12}} &= \mathbf{A_{12}}\mathbf{A_{22}}^{-1} \\ \mathbf{C_{21}} &= \mathbf{A_{22}}^{-1}\mathbf{A_{21}} \end{aligned}$$

Using this result we can find the inverse of the partitioned $\mathbf{R}$ matrix. In particular, we are interested in finding the element at position (1,1) of $\mathbf{R}^{-1}$. It will correspond to $\mathbf{C_{11}}^{-1}$.

Therefore, $var(\hat{\delta}_1) =$

We will show next that $var(\hat{\delta}_1) = \frac{\sigma^2}{1-R_1^2}$, where R_1^2 is the R^2 of the regression of x_1 on $x_2, x_3, \ldots, x_k$.

Note: We have seen that the three models (non-centered, centered, centered/scaled) have the same R^2. Find R_1^2 using the centered and scaled model. This is the model equation:

$$Zs_{i1} = \alpha_0 + \alpha_2 Zs_{i2} + \alpha_3 Zs_{i3} + \ldots + \alpha_k Zs_{ik} + \epsilon_i$$

As always, $R_1^2 = \frac{SSR}{SST}$.

Noting that

$$\mathbf{Zs_1} = \begin{pmatrix} \frac{x_{11}-\bar{x}_1}{\sqrt{\sum_{i=1}^n (x_{i1}-\bar{x}_1)^2}} \\ \frac{x_{21}-\bar{x}_1}{\sqrt{\sum_{i=1}^n (x_{i1}-\bar{x}_1)^2}} \\ \vdots \\ \vdots \\ \frac{x_{n1}-\bar{x}_1}{\sqrt{\sum_{i=1}^n (x_{i1}-\bar{x}_1)^2}} \end{pmatrix}.$$

find SST:

$SST =$

Now let's find SSR. We know that $SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n}\hat{y}_i^2 - n\bar{y}^2 = \hat{\mathbf{y}}'\hat{\mathbf{y}} - n\bar{y}^2$. In the model we are using here, the response variable is Zs_1, and because $\bar{Z}s_1 = 0$ it follows that $SSR = \hat{\mathbf{Z}}\mathbf{s_1}'\hat{\mathbf{Z}}\mathbf{s_1} = \mathbf{Zs_1}'\mathbf{HZs_1}$.

Which $\mathbf{H}$ is this?

Let $\mathbf{Z}^* = [\mathbf{Z_{s_2}}, \mathbf{Z_{s_3}}, \ldots, \mathbf{Z_{s_k}}]$. Therefore $\mathbf{H} =$

It follows that $SSR =$

Finally, $R_1^2 = \mathbf{r}'\mathbf{R_{22}}^{-1}\mathbf{r}$.

We just showed that $var(\hat{\delta}_1) = \frac{\sigma^2}{1-\mathbf{r}'\mathbf{R_{22}}^{-1}\mathbf{r}} = \frac{\sigma^2}{1-R_1^2}$. Since $\delta_j = \beta_j\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}$, it follows that
$\hat{\beta}_1 =$

and

$var(\hat{\beta}_1) =$

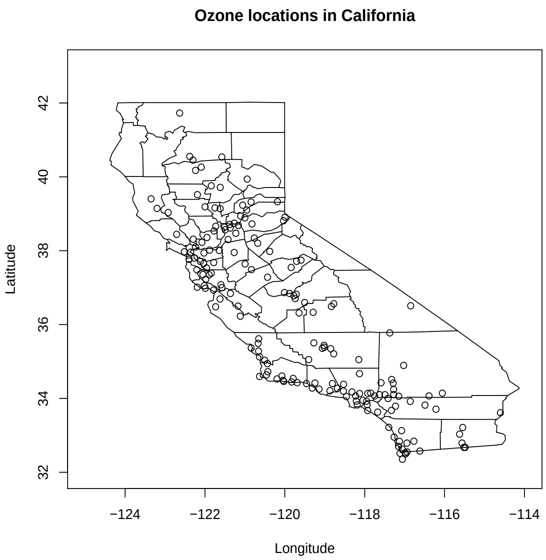
We see that if the R^2 of the regression of predictor j on the other $k-1$ predictors is large (close to 1) the variance of the predictor of $\hat{\beta}_j$ will be inflated, and therefore the corresponding t statistic will be small.

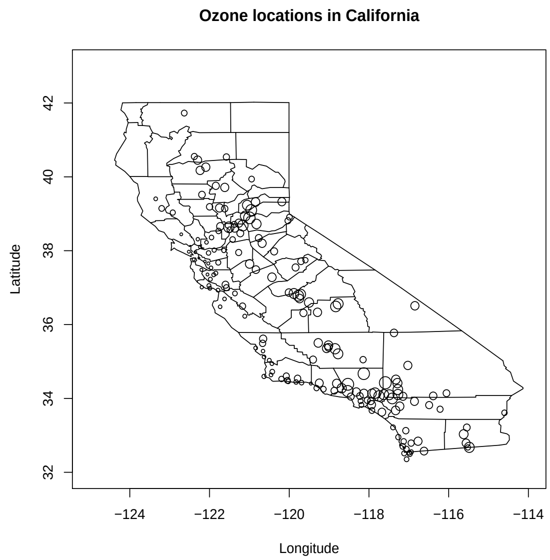
Detection of multicollinearity - variance inflation factor (VIF)
The variance inflation factor is given by $VIF_j = \frac{1}{1-R_j^2}$, and because $var(\hat{\beta}_j) = \frac{\sigma^2}{(1-R_1^2)\sum_{i=1}^{n}(x_{ij}-x_j)^2}$ it can be expressed as $VIF_j = \frac{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}{\sigma^2}var(\hat{\beta}_j) = \frac{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}{\sigma^2}\sigma^2 V_{jj} = (n-1)S_{xj}^2 V_{jj}$, where, V_{jj} is the (j,j)th element of $(\mathbf{X}'\mathbf{X})^{-1}$.

Generalized and weighted least squares

So far we assumed that the Gauss-Markov conditions hold. For the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ we have assumed that $E[\boldsymbol{\epsilon}] = 0$ and $\text{var}[\boldsymbol{\epsilon}] = \sigma^2\mathbf{I}$.

These assumptions may not hold in certain cases, therefore $\hat{\boldsymbol{\beta}}$ will not be the best estimator. For example, observations taken over time may have serial correlation and observations taken over space may exhibit spatial correlation. Consider the following two plots from a data set on ozone at 175 ozone monitoring stations in California.





How does this change the assumptions?

Assume that $E[\boldsymbol{\epsilon}] = 0$ and $\text{var}[\boldsymbol{\epsilon}] = \sigma^2\mathbf{V}$, where $\mathbf{V}$ is a full rank symmetric matrix of known constants.

For the ozone data above, $\mathbf{V}$ can be constructed using the exponential covariance function $c(h_{ij}) = c_1 e^{-\frac{h_{ij}}{\alpha}}$ where c_1 and α are certain parameters and h_{ij} is the distance between data points i and j. Write few elements of $\mathbf{V}$:

$$\mathbf{V} = \begin{pmatrix} \quad & \quad \\ \quad & \quad \end{pmatrix}.$$

Suppose we decided to use the usual ordinary least squares estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. Since $E[\boldsymbol{\epsilon}] = 0$ and $\text{var}[\boldsymbol{\epsilon}] = \sigma^2\mathbf{V}$ it follows that

$E[\mathbf{y}] =$

$\text{var}[\mathbf{y}] =$

Find $E[\hat{\boldsymbol{\beta}}] =$

Find $\text{var}[\hat{\boldsymbol{\beta}}] =$

What do you observe? Think in terms of the Gauss-Markov theorem (which uses the Gauss-Markov conditions).

In addition, consider the estimator $\mathbf{c}'\hat{\boldsymbol{\beta}}$.

Is $\mathbf{c}'\hat{\boldsymbol{\beta}}$ an unbiased estimator of $\mathbf{c}'\boldsymbol{\beta}$?

Show that $\text{var}[\mathbf{c}'\hat{\boldsymbol{\beta}}] = \sigma^{\mathbf{2}}\mathbf{q}'\mathbf{V}\mathbf{q}$. What is $\mathbf{q}$?

Conclusion: Ignoring correlation can cause problems in inference.

Can we transform the model so that the transformed vector of the error terms (and therefore the vector $\mathbf{y}$) satisfies the Gauss-Markov conditions? To answer this question, consider a similar situation, even though in a different context, when we discussed the distribution of quadratic forms. Aside note: When $\mathbf{y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ we transform $\mathbf{y}$ in such a way so that the new vector has variance equal to $\mathbf{I}$.

Also, $\mathbf{V}$ is symmetric: Use spectral decomposition and the inverse square root matrix of $\mathbf{V}$.

Model transformation:
Initial model:
$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.

Transformed model:

The model $\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\epsilon^*}$ is of the form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ and therefore using $\mathbf{X}^*$ and $\mathbf{y}^*$ we get:

$\hat{\boldsymbol{\beta}}_{\boldsymbol{GLS}} =$

Replace now $\mathbf{X}^* = \mathbf{V}^{-\frac{1}{2}}\mathbf{X}$ and $\mathbf{y}^* = \mathbf{V}^{-\frac{1}{2}}\mathbf{y}$:

$\hat{\boldsymbol{\beta}}_{\boldsymbol{GLS}} =$

Find the expected value and variance of $\hat{\boldsymbol{\beta}}_{\boldsymbol{GLS}}$.

$E[\hat{\boldsymbol{\beta}}_{\boldsymbol{GLS}}]=$

$\text{var}[\hat{\boldsymbol{\beta}}_{\boldsymbol{GLS}}] =$

Estimation by direct minimization of the error sum of squares of the model $\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\epsilon^*}$.
We have estimated $\boldsymbol{\beta}$ by recognizing its connection to $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where instead of $\mathbf{X}$ and $\mathbf{y}$ we have used $\mathbf{X}^*$ and $\mathbf{y}^*$.

Here we minimize $\boldsymbol{\epsilon^{*\prime}\epsilon^*}$. Using the model $\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\epsilon^*}$ replace $\boldsymbol{\epsilon^*} = \mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta}$ and continue with the minimization.

Estimation using the method of maximum likelihood
As we have seen in previous material, if we assume that $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ then $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ and the estimation of $\boldsymbol{\beta}$ and σ^2 can be obtained using the method of maximum likdelihood.

Similarly, if we assume that $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{V})$ then $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{V})$.

What is the likelihood function of $\mathbf{y}$?

$\mathbf{L} =$

What is the log likelihood function of $\mathbf{y}$?

$\ln\mathbf{L} =$

$\frac{\partial \ln \mathbf{L}}{\partial \boldsymbol{\beta}} =$

$\frac{\partial \ln \mathbf{L}}{\partial \sigma^2} =$

Use properties of the trace of a matrix to find $E[\hat{\sigma^2}] = \frac{1}{n}\mathbf{e'_{GLS}}\mathbf{e_{GLS}}$.

In the process we will need the following:

$E[\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\boldsymbol{GLS}}] =$

$\text{var}[\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\boldsymbol{GLS}}] =$

Back to the expected value of $\hat{\sigma^2}$:

$$\begin{aligned} E[\hat{\sigma^2}] &= \frac{1}{n}E\mathbf{e'_{GLS}}\mathbf{e_{GLS}} \\ &= \frac{1}{n}E[\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\boldsymbol{GLS}}]'\mathbf{V^{-1}}[\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\boldsymbol{GLS}}] \end{aligned}$$

Weighted least squares
For the previous results we assumed that $\mathbf{V}$ is a full rank symmetric matrix of constants without any other assumptions. If we assume that $\mathbf{V}$ is diagonal, then the error terms are independent but with unequal variances. In this case we obtained the so called weighted least squares, $\hat{\boldsymbol{\beta}}_{\boldsymbol{WLS}}$. The expressions though we obtained above are the same.

Example:
Consider the simple regression model without intercept $y_i = \beta_1 x_i + \epsilon_i$ with $E(\epsilon_i) = 0$, $\epsilon_1, \ldots, \epsilon_n$ are independent, and $\text{var}(\epsilon_i) = \sigma^2 x_i^2$. We see that the variances are unequal. Derive the weighted least squares estimate of β_1 and obtain its variance.

Distribution theory, hypothesis testing, confidence intervals
All the results we obtained earlier for distribution theory (quadratic forms), hypothesis testing, confidence intervals can be applied for the generalized/weighted least squares model.

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{GLS}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X'V^{-1}X})^{-1})$$

$$\frac{(\hat{\boldsymbol{\beta}}_{\boldsymbol{GLS}}-\boldsymbol{\beta})'\mathbf{X'V^{-1}X}(\hat{\boldsymbol{\beta}}_{\boldsymbol{GLS}}-\boldsymbol{\beta})}{\sigma^2} \sim \chi^2_{k+1}$$

$$\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{\sigma^2} \sim \chi^2_n$$

$$\frac{(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\boldsymbol{GLS}})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\boldsymbol{GLS}})}{\sigma^2} \sim \chi^2_{n-k-1}$$

Suppose we are testing the hypothesis
$H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$
$H_a : \mathbf{C}\boldsymbol{\beta} \neq \mathbf{0}$

What is the F statistic here? Note: $\mathbf{C} : m \times (k+1)$.

Comparing regression equations

Using two different data sets on the same variables we build the following two regression models

$$\begin{aligned}\mathbf{y_1} &= \mathbf{X_1}\boldsymbol{\beta_1} + \boldsymbol{\epsilon_1}\\ \mathbf{y_2} &= \mathbf{X_2}\boldsymbol{\beta_2} + \boldsymbol{\epsilon_2}\end{aligned}$$

where,
$\mathbf{y_1}$ is $n_1 \times 1$, $\mathbf{y_2}$ is $n_2 \times 1$,
$\mathbf{X_1}$ is $n_1 \times (k+1)$, $\mathbf{X_2}$ is $n_2 \times (k+1)$
$\boldsymbol{\beta_1}$ is $(k+1) \times 1$, $\boldsymbol{\beta_2}$ is $(k+1) \times 1$.

Suppose the first p elements of the vectors $\boldsymbol{\beta_1}$ and $\boldsymbol{\beta_2}$ are the same. Then, we can write the two vectors as

$$\boldsymbol{\beta_1} = \begin{pmatrix} \boldsymbol{\beta^1} \\ \boldsymbol{\beta_1^2} \end{pmatrix} \text{ and } \boldsymbol{\beta_2} = \begin{pmatrix} \boldsymbol{\beta^1} \\ \boldsymbol{\beta_2^2} \end{pmatrix}$$

where,
$\boldsymbol{\beta^1}$ is $p \times 1$, $\boldsymbol{\beta_1^2}$ is $(k+1-p) \times 1$, and $\boldsymbol{\beta_2^2}$ is $(k+1-p) \times 1$.

We wish to test the hypothesis
$H_0 : \boldsymbol{\beta_1^2} - \boldsymbol{\beta_2^2} = \mathbf{0}$
$H_a : \boldsymbol{\beta_1^2} - \boldsymbol{\beta_2^2} \neq \mathbf{0}$.

We can test this hypothesis using the F test for the general linear hypothesis that was discussed in previous lectures. But first, let's partition the matrices $\mathbf{X_1}$ and $\mathbf{X_2}$ as follows:

$$\mathbf{X_1} = \begin{pmatrix} \mathbf{X_1^1} & \mathbf{X_1^2} \end{pmatrix} \text{ and } \mathbf{X_2} = \begin{pmatrix} \mathbf{X_2^1} & \mathbf{X_2^2} \end{pmatrix}$$

Note: The columns of $\mathbf{X_1^1}$ correspond to $\boldsymbol{\beta^1}$ and the columns of $\mathbf{X_1^2}$ correspond to $\boldsymbol{\beta_1^2}$. Similarly, the columns of $\mathbf{X_2^1}$ correspond also to $\boldsymbol{\beta^1}$ and the columns of $\mathbf{X_2^2}$ correspond to $\boldsymbol{\beta_2^2}$.

The two models can be combined into one regression model as follows. Complete the regression matrix by expressing it as a block matrix using $\mathbf{X_1^1}$, $\mathbf{X_1^2}$, $\mathbf{X_2^1}$, and $\mathbf{X_2^2}$:

$$\begin{pmatrix} \mathbf{y_1} \\ \mathbf{y_2} \end{pmatrix} = \begin{pmatrix} \qquad\qquad\qquad\qquad \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta^1} \\ \boldsymbol{\beta_1^2} \\ \boldsymbol{\beta_2^2} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\epsilon_1} \\ \boldsymbol{\epsilon_2} \end{pmatrix}.$$

This model that combines all the information is of the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

and the hypothesis $\boldsymbol{\beta_1^2} - \boldsymbol{\beta_2^2} = \mathbf{0}$ can be tested using
$H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$
$H_a : \mathbf{C}\boldsymbol{\beta} \neq \mathbf{0}$
with the general F test

$$F_{m,n-k-1} = \frac{(\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\gamma})' \left[\mathbf{C}(\mathbf{X'X})^{-1}\mathbf{C'}\right]^{-1} (\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\gamma})}{ms_e^2},$$

where the matrix $\mathbf{C}$ is defines as follows. Remember: We are testing for the equality of two subvectors that have dimensions $(k+1-p) \times 1$. So the matrix $\mathbf{C}$ will consist of 0s, 1s, -1s. Complete the matrix below:

$$\mathbf{C} = \begin{pmatrix} \qquad\qquad\qquad\qquad \\ \\ \end{pmatrix}$$

Therefore the dimensions of $\mathbf{C}$ are:

Example:
Suppose $k = 5$ and $p = 3$ and we have two data sets on the same variables. We want to test
$H_0 : \beta_3^1 = \beta_3^2, \beta_4^1 = \beta_4^2, \beta_5^1 = \beta_5^2$
H_a : Not true

Things to do:

1. Write the regression matrix when we combine the two data sets. Note: Express the predictors as $\mathbf{x_1^1}, \mathbf{x_2^1}, \ldots$, where, $\mathbf{x_1^1}$ is predictor 1 in data set 1, $\mathbf{x_2^1}$ is predictor 2 in data set 1, etc.

2. Write the vector $\boldsymbol{\beta}$.

3. Write the matrix $\mathbf{C}$ for testing the hypothesis above.

Test the hypothesis using the extra sum of squares method.
As we have seen a hypothesis test can also be done using the extra sum of squares principle. The general idea is that we transform the model taking into account the null hypothesis. Using the transformed model (reduced model) we compute the error sum of squares of the reduced model and together with the error sum of squares of the full model we construct the F statistic.
We wish to test the hypothesis
$H_0 : \boldsymbol{\beta_1^2} - \boldsymbol{\beta_2^2} = \mathbf{0}$
$H_a : \boldsymbol{\beta_1^2} - \boldsymbol{\beta_2^2} \neq \mathbf{0}$.

Express the model under the null hypothesis.

$$\begin{pmatrix} \mathbf{y_1} \\ \mathbf{y_2} \end{pmatrix} = \begin{pmatrix} \quad & \quad \\ \quad & \quad \end{pmatrix} \begin{pmatrix} \quad \\ \quad \end{pmatrix} + \begin{pmatrix} \boldsymbol{\epsilon_1} \\ \boldsymbol{\epsilon_2} \end{pmatrix}.$$

What is SSE_R?

What is the full model?

What is SSE_F?

What are the degrees of freedom?
$df_F =$

$df_R =$

Write the F statistic based on the full and reduced model:

Example
Suppose $k = 5$ and we we want to test
$H_0 : \beta_4^1 = \beta_4^2, \beta_5^1 = \beta_5^2$
H_a : Not true

Note:
Subscript ij refers to observation i for variable j.
Superscript 1 or 2 refer to dataset 1 or 2.

This is the formulation:

$$\begin{pmatrix} y_{11} \\ y_{21} \\ y_{31} \\ \vdots \\ y_{n_11} \\ y_{12} \\ y_{22} \\ y_{32} \\ \vdots \\ y_{n_22} \end{pmatrix} = \begin{pmatrix} 1 & x_{11}^1 & x_{12}^1 & x_{13}^1 & x_{14}^1 & x_{15}^1 & 0 & 0 \\ 1 & x_{21}^1 & x_{22}^1 & x_{23}^1 & x_{24}^1 & x_{25}^1 & 0 & 0 \\ 1 & x_{31}^1 & x_{32}^1 & x_{33}^1 & x_{34}^1 & x_{35}^1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n_11}^1 & x_{n_12}^1 & x_{n_13}^1 & x_{n_14}^1 & x_{n_15}^1 & 0 & 0 \\ 1 & x_{11}^2 & x_{12}^2 & x_{13}^2 & 0 & 0 & x_{14}^2 & x_{15}^2 \\ 1 & x_{21}^2 & x_{22}^2 & x_{23}^2 & 0 & 0 & x_{24}^2 & x_{25}^2 \\ 1 & x_{31}^2 & x_{32}^2 & x_{33}^2 & 0 & 0 & x_{34}^2 & x_{35}^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n_21}^2 & x_{n_22}^2 & x_{n_23}^2 & 0 & 0 & x_{n_24}^2 & x_{n_25}^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4^1 \\ \beta_5^1 \\ \beta_4^2 \\ \beta_5^2 \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{21} \\ \epsilon_{31} \\ \vdots \\ \epsilon_{n_11} \\ \epsilon_{11} \\ \epsilon_{22} \\ \epsilon_{32} \\ \vdots \\ \epsilon_{n_21} \end{pmatrix}$$

In this example we have $k = 5, p = 4$ and $\mathbf{C}$ is given by $\mathbf{C} = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \end{pmatrix}$.

Diagnostics
Influential data points
Deleting a single point in regression
In this document we will explore the effect of deleting a single point in multiple regression. Let's partition the vector $\mathbf{y}$, the matrix $\mathbf{X}$, and the vector $\boldsymbol{\epsilon}$ as follows:

$$\mathbf{y} = \begin{pmatrix} \mathbf{y_{(i)}} \\ y_i \end{pmatrix} = \mathbf{X}\boldsymbol{\beta} + \begin{pmatrix} \boldsymbol{\epsilon_{(i)}} \\ \epsilon_i \end{pmatrix} = \begin{pmatrix} \mathbf{X_{(i)}} \\ \mathbf{x_i'} \end{pmatrix} \boldsymbol{\beta} + \begin{pmatrix} \boldsymbol{\epsilon_{(i)}} \\ \epsilon_i \end{pmatrix}.$$

Some notation: The subscript (i) means that the ith data point is removed, and $\mathbf{x_i'}$ is the ith row of the $\mathbf{X}$ matrix. We know already the solution of least squares when none of the points is removed. The usual OLS solution is: $\hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}\mathbf{X'y}$. The model we are using here is:

$$\mathbf{y_{(i)}} = \mathbf{X_{(i)}}\boldsymbol{\beta} + \boldsymbol{\epsilon_{(i)}}.$$

and therefore, $\hat{\boldsymbol{\beta}}_{\boldsymbol{(i)}} =$.

We need an expression for $(\mathbf{X_{(i)}}'\mathbf{X_{(i)}})^{-1}$ and for $\mathbf{X}'_{(i)}\mathbf{y_{(i)}}$.

Use the partition of $\mathbf{X} = \begin{pmatrix} \mathbf{X_{(i)}} \\ \mathbf{x_i'} \end{pmatrix}$ to find:

$$\mathbf{X'X} =$$

It follows that

$$\mathbf{X}'_{(i)}\mathbf{X_{(i)}} =$$

A useful result from linear algebra will be used here. Let $\mathbf{A}$ be a matrix and $\mathbf{b}$ be a vector. Then,

$$[\mathbf{A} - \mathbf{bb'}]^{-1} = \mathbf{A}^{-1} + \frac{\mathbf{A}^{-1}\mathbf{bb'A}^{-1}}{1 - \mathbf{b'A}^{-1}\mathbf{b}}, \text{ provided that } \mathbf{A} \text{ is invertible and } 1 - \mathbf{b'A}^{-1}\mathbf{b} \neq 0.$$

Similarly,

$$[\mathbf{A} + \mathbf{bb'}]^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{bb'A}^{-1}}{1 + \mathbf{b'A}^{-1}\mathbf{b}}, \text{ provided that } \mathbf{A} \text{ is invertible and } 1 + \mathbf{b'A}^{-1}\mathbf{b} \neq 0.$$

We can now use the first result to find the inverse of $(\mathbf{X_{(i)}}'\mathbf{X_{(i)}})^{-1}$.

$$\begin{aligned} [\mathbf{X_{(i)}}'\mathbf{X_{(i)}}]^{-1} &= [\mathbf{X'X} - \mathbf{x_{(i)}x'_{(i)}}]^{-1} \\ &= (\mathbf{X'X})^{-1} + \frac{(\mathbf{X'X})^{-1}\mathbf{x_i x_i'}(\mathbf{X'X})^{-1}}{1 - \mathbf{x_i'}(\mathbf{X'X})^{-1}\mathbf{x_i}}, \end{aligned}$$

Does the denominator of the last term of the previous expression remind anything? Therefore,

$$[\mathbf{X_{(i)}}'\mathbf{X_{(i)}}]^{-1} = (\mathbf{X'X})^{-1} + \frac{(\mathbf{X'X})^{-1}\mathbf{x_i x_i'}(\mathbf{X'X})^{-1}}{}.$$

Now find an expression of $\mathbf{X}'_{(i)}\mathbf{y_{(i)}}$. Begin with $\mathbf{X'y}$:

$$\mathbf{X'y} =$$

Now let's compute the estimate of the $\boldsymbol{\beta}$ vector, which after the deletion of data point i will be denoted with $\boldsymbol{\beta}_{(i)}$. The OLS vector will be denoted $\hat{\boldsymbol{\beta}}_{(i)}$.

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{(i)} &= [\mathbf{X_{(i)}}'\mathbf{X_{(i)}}]^{-1}\mathbf{X}'_{\mathbf{(i)}}\mathbf{y_{(i)}} \\
&= \left[(\mathbf{X}'\mathbf{X})^{-1} + \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}'_\mathbf{i}\mathbf{x_i}(\mathbf{X}'\mathbf{X})^{-1}}{1-h_{ii}}\right]\mathbf{X}'_{\mathbf{(i)}}\mathbf{y_{(i)}}
\end{aligned}$$

Replace now $\mathbf{X}'_{\mathbf{(i)}}\mathbf{y_{(i)}} = \mathbf{X}'\mathbf{y} - \mathbf{x_{(i)}}y_i$. Therefore,

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{(i)} &= \left[(\mathbf{X}'\mathbf{X})^{-1} + \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_i}\mathbf{x}'_\mathbf{i}(\mathbf{X}'\mathbf{X})^{-1}}{1-h_{ii}}\right][\mathbf{X}'\mathbf{y} - \mathbf{x_{(i)}}y_i] \\
&= \hat{\boldsymbol{\beta}} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_i}y_i \\
&+ \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_i}\mathbf{x}'_\mathbf{i}\hat{\boldsymbol{\beta}}}{1-h_{ii}} \\
&- \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_i}h_{ii}y_i}{1-h_{ii}}
\end{aligned}$$

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{(i)} &= \hat{\boldsymbol{\beta}} - \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_i}y_i(1-h_{ii})}{1-h_{ii}} + \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_i}\mathbf{x}'_\mathbf{i}\hat{\boldsymbol{\beta}}}{1-h_{ii}} - \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_i}h_{ii}y_i}{1-h_{ii}} \\
&= \hat{\boldsymbol{\beta}} - \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_i}y_i}{1-h_{ii}} + \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_i}\mathbf{x}'_\mathbf{i}\hat{\boldsymbol{\beta}}}{1-h_{ii}} \\
&= \hat{\boldsymbol{\beta}} - \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_i}}{1-h_{ii}}[y_i - \mathbf{x}'_\mathbf{i}\hat{\boldsymbol{\beta}}]
\end{aligned}$$

Note: We recognize that $y_i - \mathbf{x}'_\mathbf{i}\hat{\boldsymbol{\beta}} =$ Therefore,

$\hat{\boldsymbol{\beta}}_{(i)} = \hat{\boldsymbol{\beta}} - \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_i}}{1-h_{ii}}e_i$, and the influence of the ith data point on the vector $\hat{\boldsymbol{\beta}}$ is given by

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)} = \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_i}}{1-h_{ii}}e_i.$$

The vector $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}$ is often called $DFBETA_i$.

The quantity, $\mathbf{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})$ has the units of $\hat{\mathbf{y}}$ and its squared length is equal to:

$$\begin{aligned}
[\mathbf{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})]'[(\mathbf{X}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})] &= (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}) = \\
&= \frac{e_i^2}{(1-h_{ii})^2}\mathbf{x}'_\mathbf{i}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_i}] = \frac{h_{ii}}{(1-h_{ii})^2}e_i^2.
\end{aligned}$$

This squared length is the basis for Cook's distance. It is computed as follows:

$$D_i = \frac{h_{ii}}{(1-h_{ii})^2}\frac{e_i^2}{(k+1)s_e^2}.$$

We can also compute the effect of deleting a data point on the predicted value $\hat{y}_i$. The new predicted value is denoted with $\hat{y}_i(i)$, and the difference $\hat{y}_i - \hat{y}_i(i)$ is denoted with $DFFITS_i$ and it is computed as follows:

$$DFFITS_i = \hat{y}_i - \hat{y}_i(i) = \mathbf{x}'_\mathbf{i}\hat{\boldsymbol{\beta}} - \mathbf{x}'_\mathbf{i}\hat{\boldsymbol{\beta}}_{(i)} = \mathbf{x}'_\mathbf{i}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}) = \mathbf{x}'_\mathbf{i}\frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_i}}{1-h_{ii}}e_i = \frac{h_{ii}}{1-h_{ii}}e_i.$$

We would also like to develop an expression that connects s_e^2 and $s_e^2(i)$, where s_e^2 is the unbiased estimate of σ^2 using all the n data points and $s_e^2(i)$ is the unbiased estimate of σ^2 when the ith data point is deleted. It follows that $s_e^2(i)$ is the unbiased estimator of σ^2 (after deleting data point i.)

$$s_e^2(i) = \frac{1}{n-k-2} \sum_{l=1, l \neq i}^{n} (y_l - \mathbf{x}_l' \hat{\boldsymbol{\beta}}_{(i)})^2,$$

and it should have the properties of s_e^2, i.e., it is unbiased, and also $\frac{(n-k-2)s_e^2(i)}{\sigma^2} \sim \chi^2_{n-k-2}$. The expression of $s_e^2(i)$ can be expressed in terms of s_e^2, e_i, h_{ii} as follows:

The matrix $\mathbf{H}$ is idempotent, therefore $\mathbf{HH} = \mathbf{H}$, which implies that $\sum_{l=1}^n h_{il}^2 = hii$. Also, since $\mathbf{He} = \mathbf{H}(\mathbf{I}-\mathbf{H})\mathbf{y} = \mathbf{0}$ it follows that $\sum_{l=1}^n h_{il} e_l = 0$. Using these results and also the result from above, $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)} = \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_i}}{1-h_{ii}} e_i$ we get:

$$\begin{aligned}
\sum_{l=1, l \neq i}^{n} (y_l - \mathbf{x}_l' \hat{\boldsymbol{\beta}}_{(i)})^2 &= \sum_{l=1, l \neq i}^{n} (y_l - \mathbf{x}_l' \hat{\boldsymbol{\beta}} + \mathbf{x}_l' \hat{\boldsymbol{\beta}} - \mathbf{x}_l' \hat{\boldsymbol{\beta}}_{(i)})^2 \\
&= \sum_{l=1, l \neq i}^{n} (e_l + \mathbf{x}_l' (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}))^2 = \sum_{l=1, l \neq i}^{n} \left(e_l + \mathbf{x}_l' \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_i}}{1-h_{ii}} e_i \right)^2 \\
&= \sum_{l=1, l \neq i}^{n} (e_l + h_{il} \frac{e_i}{1-h_{ii}})^2 = \sum_{l=1}^{n} (e_l + h_{il} \frac{e_i}{1-h_{ii}})^2 - (e_i + h_{ii} \frac{e_i}{1-h_{ii}})^2 \\
&= \sum_{l=1}^{n} e_l^2 + \frac{e_i^2}{(1-h_{ii})^2} \sum_{l=1}^{n} h_{il}^2 + 2 \frac{e_i}{1-h_{ii}} \sum_{l=1}^{n} e_l h_{il} - \frac{e_i^2}{(1-h_{ii})^2} \\
&= \sum_{l=1}^{n} e_l^2 - \frac{e_i^2}{1-h_{ii}}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
s_e^2(i) &= \frac{1}{n-k-2} \sum_{l=1, l \neq i}^{n} (y_l - \mathbf{x}_l' \hat{\boldsymbol{\beta}}_{(i)})^2 = \frac{1}{n-k-2} \left(\sum_{l=1}^{n} e_l^2 - \frac{e_i^2}{1-h_{ii}} \right) \Rightarrow \\
(n-k-2) s_e(i)^2 &= (n-k-1) s_e^2 - \frac{e_i^2}{1-h_{ii}}.
\end{aligned}$$

Example:
Let's compute some of the expressions above.

```
a <- read.table("http://www.stat.ucla.edu/~nchristo/statistics100C/rain_wheat.txt", header=TRUE)
```

	rain	wheat
1	12	310
2	14	320
3	13	323
4	16	330
5	18	334
6	20	348
7	19	352
8	22	360
9	22	370
10	20	344
11	23	370
12	24	380
13	26	385
14	27	393
15	28	395
16	29	400
17	30	403
18	31	406
19	26	383
20	27	388
21	28	392
22	29	398
23	30	400
24	31	403
25	20	270
26	50	260

Let's compute the $DFBETA_i$ vector, $DFFITS_i$ vector, Cook's distance, and $s_e^2(i)$:

```
k <- 1
ones <- rep(1, nrow(a))
X <- as.matrix(cbind(ones, a$rain))
H <- X %*% solve(t(X) %*% X) %*% t(X)
betahat <- solve(t(X) %*% X) %*% t(X) %*% a$wheat
se2 <- (t(a$wheat) %*% a$wheat - t(betahat) %*% t(X) %*% a$wheat) / (nrow(a)-k-1)

e <- a$wheat - X %*% betahat
h <- diag(H)

#Compute DFBETAi vector:
dfbeta <- c(0,0)
for(i in 1:26){
dfb <- t( solve(t(X) %*% X) %*% X[i,] * e[i]/(1-h[i]) )
dfbeta <- rbind(dfbeta, dfb)
}

#Compute DFFITSi vector:
dffits <- rep(0,26)
for(i in 1:26){
dffits[i] <- h[i]*e[i]/(1-h[i])
}

#Compute Cook's distance:
D <- rep(0,26)
for(i in 1:26){
D[i] <- h[i]*e[i]^2/((1-h[i])^2*(k+1)*se2)
}
```

```
#Compute se^2(i):
se2i <- rep(0,26)
for(i in 1:26){
se2i[i] <- ( (nrow(a)-k-1)*se2-e[i]^2/(1-h[i]) )/(nrow(a)-k-2)
}

> head(dfbeta[-1,])
       ones
 -10.656807 0.36677131
  -7.090793 0.23679665
  -6.706132 0.22762529
  -4.333757 0.13865488
  -3.193471 0.09566092
  -1.063000 0.02839604

> dffits
 [1]   -6.2555515   -3.7756396   -3.7470035   -2.1152784   -1.4715742   -0.4950788
 [7]   -0.2445059    0.0261643    0.4687776   -0.7124230    0.3915380    0.7341148
[13]    0.8776548    1.2481878    1.4157437    1.8021571    2.1665339    2.6187348
[19]    0.7940537    1.0240767    1.2677504    1.6914258    1.9775995    2.4020277
[25]   -4.7332902 -119.3509570

> D
 [1] 8.323144e-02 3.865679e-02 3.364939e-02 1.569233e-02 9.877095e-03 1.432397e-03
 [7] 3.098183e-04 4.864369e-06 1.561500e-03 2.966132e-03 1.159646e-03 4.207036e-03
[13] 5.781940e-03 1.093786e-02 1.284058e-02 1.864538e-02 2.386102e-02 3.065808e-02
[19] 4.732884e-03 7.362709e-03 1.029634e-02 1.642449e-02 1.988084e-02 2.579396e-02
[25] 1.309305e-01 9.019692e+00

> se2i
 [1] 1659.489 1687.662 1698.529 1708.213 1712.159 1728.521 1731.602 1732.311
 [9] 1727.234 1724.446 1728.290 1717.193 1712.359 1697.099 1694.751 1683.711
[17] 1677.632 1671.058 1715.982 1708.614 1702.196 1689.502 1686.755 1680.779
[25] 1384.469  296.899
```

All the above can be obtained much easier using:

```
q <- lm(a$wheat ~ a$rain )
> influence(q)
```

Adding a single point in regression

In this document we will explore the effect of adding a single point in multiple regression. We will add a new y value, y_0 and a new row of the $\mathbf{X}$ matrix, $\mathbf{x_0'}$. The new model in matrix form is expressed as follows:

$$\begin{pmatrix} \mathbf{y} \\ y_0 \end{pmatrix} = \begin{pmatrix} \mathbf{X} \\ \mathbf{x_0'} \end{pmatrix} \boldsymbol{\beta} + \begin{pmatrix} \boldsymbol{\epsilon} \\ \epsilon_0 \end{pmatrix}.$$

Or

$$\mathbf{y}_{new} = \mathbf{X}_{new}\boldsymbol{\beta} + \boldsymbol{\epsilon}_{new}.$$

Therefore $\hat{\boldsymbol{\beta}}_{new} =$

We need an expression for $(\mathbf{X'}_{new}\mathbf{X}_{new})^{-1}$ and for $\mathbf{X'}_{new}\mathbf{y_{new}}$.

$\mathbf{X'}_{new}\mathbf{X}_{new} =$

$\mathbf{X'}_{new}\mathbf{y_{new}} =$

A useful result from linear algebra will be used here. Let $\mathbf{A}$ be a matrix and $\mathbf{b}$ be a vector. Then,

$$[\mathbf{A} + \mathbf{bb'}]^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{bb'}\mathbf{A}^{-1}}{1 + \mathbf{b'}\mathbf{A}^{-1}\mathbf{b}}, \text{ provided that } \mathbf{A} \text{ is invertible and } 1 + \mathbf{b'}\mathbf{A}^{-1}\mathbf{b} \neq 0.$$

We can now use this result to find the inverse of $(\mathbf{X'}_{new}\mathbf{X}_{new})^{-1}$. Therefore

$$\begin{aligned}
(\mathbf{X'}_{new}\mathbf{X}_{new})^{-1} &= (\mathbf{X'X} + \mathbf{x_0x_0'})^{-1} \\
&= (\mathbf{X'X})^{-1} - \frac{(\mathbf{X'X})^{-1}\mathbf{x_0x_0'}(\mathbf{X'X})^{-1}}{1 + \mathbf{x_0'}(\mathbf{X'X})^{-1}\mathbf{x_0}},
\end{aligned}$$

The denominator of the last term of the previous expression is denoted with $1 + h_{00}$. This is just a notation because h_{00} is not a leverage value. Why?

$$(\mathbf{X'}_{new}\mathbf{X}_{new})^{-1} = (\mathbf{X'X})^{-1} - \frac{(\mathbf{X'X})^{-1}\mathbf{x_0x_0'}(\mathbf{X'X})^{-1}}{1 + h_{00}}.$$

Now let's compute the estimator of the $\boldsymbol{\beta}$ vector. The OLS vector will be denoted $\hat{\boldsymbol{\beta}}_{new}$.

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{new} &= (\mathbf{X'}_{new}\mathbf{X}_{new})^{-1}\mathbf{X'}_{new}\mathbf{y}_{new} \\
&= \left[(\mathbf{X'X})^{-1} - \frac{(\mathbf{X'X})^{-1}\mathbf{x_0x_0'}(\mathbf{X'X})^{-1}}{1 + h_{00}}\right] \begin{pmatrix} \mathbf{X'} & \mathbf{x_0} \end{pmatrix} \begin{pmatrix} \mathbf{y} \\ y_0 \end{pmatrix} \\
&= \left[(\mathbf{X'X})^{-1} - \frac{(\mathbf{X'X})^{-1}\mathbf{x_0x_0'}(\mathbf{X'X})^{-1}}{1 + h_{00}}\right] [\mathbf{X'y} + \mathbf{x_0}y_0] \\
&= \hat{\boldsymbol{\beta}} + (\mathbf{X'X})^{-1}\mathbf{x_0}y_0 - \frac{(\mathbf{X'X})^{-1}\mathbf{x_0x_0'}\hat{\boldsymbol{\beta}}}{1 + h_{00}} - \frac{(\mathbf{X'X})^{-1}\mathbf{x_0}h_{00}y_0}{1 + h_{00}} \\
&= \hat{\boldsymbol{\beta}} + \frac{(\mathbf{X'X})^{-1}\mathbf{x_0}y_0(1 + h_{00})}{1 + h_{00}} - \frac{(\mathbf{X'X})^{-1}\mathbf{x_0x_0'}\hat{\boldsymbol{\beta}}}{1 + h_{00}} - \frac{(\mathbf{X'X})^{-1}\mathbf{x_0}h_{00}y_0}{1 + h_{00}} \\
&= \hat{\boldsymbol{\beta}} + \frac{(\mathbf{X'X})^{-1}\mathbf{x_0}y_0}{1 + h_{00}} - \frac{(\mathbf{X'X})^{-1}\mathbf{x_0x_0'}\hat{\boldsymbol{\beta}}}{1 + h_{00}} \\
&= \hat{\boldsymbol{\beta}} + \frac{(\mathbf{X'X})^{-1}\mathbf{x_0}}{1 + h_{00}}[y_0 - \mathbf{x_0'}\hat{\boldsymbol{\beta}}].
\end{aligned}$$

Now let $e_0 = y_0 - \mathbf{x_0'}\hat{\boldsymbol{\beta}}$. Note: e_0 is not a residual. This is only a notation. Therefore,

$$\hat{\boldsymbol{\beta}}_{new} = \hat{\boldsymbol{\beta}} + \frac{(\mathbf{X'X})^{-1}\mathbf{x_0}}{1 + h_{00}}e_0.$$

Influential analysis
Using the residuals and leverage values (the diagonal of the hat matrix) we can find interesting diagnostics for identifying unusual observations.

$\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{y} = (\mathbf{I} - \mathbf{H})\boldsymbol{\epsilon}$

$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and therefore $h_{ii} = \mathbf{x_i'}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_i}$

Internally studentized residuals
Find the distribution of e_i (it will involve h_{ii}).

We also know that $\frac{(n-k-1)S_e^2}{\sigma^2} \sim \chi^2_{n-k-1}$.

However the expression $r_i = \frac{e_i}{S_e\sqrt{1-h_{ii}}}$ (which is called the internally studentized residual) does not follow a t distribution because e_i is not independent of S_e^2.

Instead we will show that $\frac{r_i^2}{n-k-1} \sim \text{beta}(\frac{1}{2}, \frac{1}{2}(n-k-2))$.

Proof
We see that $\frac{r_i^2}{n-k-1} = \frac{e_i^2}{S_e^2(n-k-1)(1-h_{ii})}$
Since $S_e^2 = \frac{\mathbf{e}'\mathbf{e}}{n-k-1}$ we can replace $S_e^2(n-k-1) = \mathbf{e}'\mathbf{e}$ to get $\frac{r_i^2}{n-k-1} = \frac{e_i^2}{\mathbf{e}'\mathbf{e}(1-h_{ii})}$.

Now express e_i as $e_i = \mathbf{c_i'}\mathbf{e}$, where $\mathbf{c_i'} = (0, 0, 0, \ldots, 0, 1, 0, \ldots, 0)$ (a row vector of zeros with 1 at position i).

Replace now $\mathbf{e} = (\mathbf{I} - \mathbf{H})\boldsymbol{\epsilon}$ to get $e_i = \mathbf{c_i'}(\mathbf{I} - \mathbf{H})\boldsymbol{\epsilon}$.

Since e_i is a scalar we can also write e_i as the transpose of the previous expression:

$e_i =$

Also express $\mathbf{e}'\mathbf{e}$ as a function of $\boldsymbol{\epsilon}$
$\mathbf{e}'\mathbf{e} =$

Now back to the expression $\frac{r_i^2}{n-k-1}$. We can write it as
$\frac{r_i^2}{n-k-1} = \frac{e_i^2}{(n-k-1)S_e^2(1-h_{ii})} = \frac{\boldsymbol{\epsilon}'(\mathbf{I}-\mathbf{H})\mathbf{c_i}\mathbf{c_i'}(\mathbf{I}-\mathbf{H})\boldsymbol{\epsilon}}{\boldsymbol{\epsilon}'(\mathbf{I}-\mathbf{H})\boldsymbol{\epsilon}(1-h_{ii})}$ and divide both the numerator and denominator by σ^2 to get

$$\frac{\frac{\boldsymbol{\epsilon}'}{\sigma}\frac{(\mathbf{I}-\mathbf{H})\mathbf{c_i}\mathbf{c_i'}(\mathbf{I}-\mathbf{H})}{1-h_{ii}}\frac{\boldsymbol{\epsilon}}{\sigma}}{\frac{\boldsymbol{\epsilon}'}{\sigma}(\mathbf{I}-\mathbf{H})\frac{\boldsymbol{\epsilon}}{\sigma}}$$

Assuming that $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ it follows that $\frac{\boldsymbol{\epsilon}}{\sigma} \sim$

Now let $\mathbf{z} = \frac{\boldsymbol{\epsilon}}{\sigma}$ and $\mathbf{Q} = \frac{(\mathbf{I}-\mathbf{H})\mathbf{c_i}\mathbf{c_i'}(\mathbf{I}-\mathbf{H})}{1-h_{ii}}$.

What are dimensions of
$\mathbf{c_i}$

$\mathbf{I} - \mathbf{H}$

$\mathbf{Q}$

So far

$$\frac{r_i^2}{n-k-1} = \frac{\mathbf{z}'\mathbf{Q}\mathbf{z}}{\mathbf{z}'(\mathbf{I}-\mathbf{H})\mathbf{z}}$$

These are quadratic expressions and because
$\mathbf{Z} \sim N(\mathbf{0}, \mathbf{I})$ they must follow ...
One suggestion is to claim that this ratio follows an F distribution
However the two quadratic expressions $\mathbf{z}'\mathbf{Q}\mathbf{z}$ and $\mathbf{z}'(\mathbf{I}-\mathbf{H})\mathbf{z}$
are not independent. Why?

We can add/subtract $\mathbf{z}'\mathbf{Q}\mathbf{z}$ in the denominator to get

$$\frac{r_i^2}{n-k-1} = \frac{\mathbf{z}'\mathbf{Q}\mathbf{z}}{\mathbf{z}'(\mathbf{I}-\mathbf{H}-\mathbf{Q})\mathbf{z} + \mathbf{z}'\mathbf{Q}\mathbf{z}}$$

Now the quadratic expressions in the denominator are independent. Why?

Notes:

1. Find $\mathbf{c_i}'(\mathbf{I}-\mathbf{H})\mathbf{c_i}$

2. Show that $\mathbf{Q}$ is symmetric and idempotent.
   $\mathbf{Q}' =$

   $\mathbf{Q}\mathbf{Q} =$

3. Find the trace of $\mathbf{Q}$.

4. Find the distribution of $\mathbf{z}'\mathbf{Q}\mathbf{z}$.

5. Show that $\mathbf{I}-\mathbf{H}-\mathbf{Q}$ is symmetric and idempotent.
   $[\mathbf{I}-\mathbf{H}-\mathbf{Q}]' =$

   $[\mathbf{I}-\mathbf{H}-\mathbf{Q}][\mathbf{I}-\mathbf{H}-\mathbf{Q}] =$

6. Find the trace of $\mathbf{I}-\mathbf{H}-\mathbf{Q}$.

7. Find the distribution of $\mathbf{z}'(\mathbf{I}-\mathbf{H}-\mathbf{Q})\mathbf{z}$.

So far we showed that $\mathbf{z}'\mathbf{Q}\mathbf{z} \sim \chi^2_1$ or $\Gamma(\frac{1}{2}, 2)$ and $\mathbf{z}'(\mathbf{I} - \mathbf{H} - \mathbf{Q})\mathbf{z} \sim \chi^2_{n-k-2}$ or $\Gamma(\frac{n-k-2}{2}, 2)$.

The following result from mathematical statistics will help us identify the distribution of $\frac{r_i^2}{n-k-1} = \frac{\mathbf{z}'\mathbf{Q}\mathbf{z}}{\mathbf{z}'(\mathbf{I}-\mathbf{H}-\mathbf{Q})\mathbf{z}+\mathbf{z}'\mathbf{Q}\mathbf{z}}$.

Joint probability distribution of functions of random variables
The idea of the distribution of a function of a random variable can be extended to bivariate and multivariate random vectors as follows.

Let X_1, X_2 be jointly continuous random variables with pdf $f_{X1X2}(x_1, x_2)$. Suppose $Y_1 = g_1(X_1, X_2)$ and $Y_2 = g_2(X_1, X_2)$. We want to find the joint pdf of Y_1, Y_2. We follow this procedure:

1. Solve the equations $y_1 = g_1(x_1, x_2)$ and $y_2 = g_2(x_1, x_2)$ for x_1 and x_2 in terms of y_1 and y_2 to get $x_1 = h_1(y_1, y_2)$ and $x_2 = h_2(y_1, y_2)$.
2. Compute the Jacobian: $\mathbf{J} = \begin{vmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_2} \\ \frac{\partial g_2}{\partial x_1} & \frac{\partial g_2}{\partial x_2} \end{vmatrix}$. ($\mathbf{J}$ is the determinant of the matrix of partial derivatives.)

To find the joint pdf of Y_1, Y_2 use the following result: $f_{Y_1,Y_2}(y_1, y_2) = f_{X_1,X_2}(x_1, x_2)|\mathbf{J}|^{-1}$, where $|\mathbf{J}|$ is the absolute value of the Jacobian. Here, x_1, x_2 are the expressions obtained from step (1) above, $x_1 = h_1(y_1, y_2)$ and $x_2 = h_2(y_1, y_2)$.

Example
Suppose X and Y are independent random variables with $X \sim \Gamma(\alpha_1, \beta)$ and $Y \sim \Gamma(\alpha_2, \beta)$. Compute the joint pdf of $U = X + Y$ and $V = \frac{X}{X+Y}$ and find the distribution of U and the distribution of V. Also show that U, V are independent.

Solution:
A random variable X is said to have a gamma distribution with parameters α, β if its probability density function is given by

$$f(x) = \frac{x^{\alpha-1}e^{-\frac{x}{\beta}}}{\Gamma(\alpha)\beta^\alpha}, \quad \alpha, \beta > 0, x \geq 0.$$

Here $X \sim \Gamma(\alpha_1, \beta)$ and $Y \sim \Gamma(\alpha_2, \beta)$, therefore,

$$f_X(x) = \frac{x^{\alpha_1-1}e^{-\frac{x}{\beta}}}{\Gamma(\alpha_1)\beta^{\alpha_1}}, \text{ and } f_Y(y) = \frac{y^{\alpha_2-1}e^{-\frac{y}{\beta}}}{\Gamma(\alpha_2)\beta^{\alpha_2}}$$

Because X, Y are independent, the joint pdf of X and Y is the product of the two marginal pdfs:

$$f_{XY}(x, y) = f_X(x)f_Y(y) = \frac{x^{\alpha_1-1}e^{-\frac{x}{\beta}}}{\Gamma(\alpha_1)\beta^{\alpha_1}} \frac{y^{\alpha_2-1}e^{-\frac{y}{\beta}}}{\Gamma(\alpha_2)\beta^{\alpha_2}} = \frac{x^{\alpha_1-1}y^{\alpha_2-1}e^{-\frac{x+y}{\beta}}}{\Gamma(\alpha_1)\Gamma(\alpha_2)\beta^{\alpha_1+\alpha_2}}.$$

Now follow the two steps above:

1. Solve the equations $u = x + y$ and $v = \frac{x}{x+y}$ in terms of x and y. We get: $x = uv$ and $y = u(1 - v)$.
2. Compute the Jacobian: $\mathbf{J} = \begin{vmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{vmatrix} = \begin{vmatrix} 1 & 1 \\ \frac{y}{(x+y)^2} & -\frac{x}{(x+y)^2} \end{vmatrix} = -\frac{1}{x+y} = -\frac{1}{u}$.

Finally to find the joint pdf of U, V use $x = uv$ and $y = u(1 - v)$ in the joint pdf of X, Y: $\mathrm{f}_{UV}(u, v) = \frac{(uv)^{\alpha_1-1}[u(1-v)]^{\alpha_2-1}e^{-\frac{u}{\beta}}u}{\Gamma(\alpha_1)\Gamma(\alpha_2)\beta^{\alpha_1+\alpha_2}}$, multiply by $\frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1+\alpha_2)}$ and rearrange to get :

$$f_{UV}(u, v) = \frac{u^{\alpha_1+\alpha_2-1}e^{-\frac{u}{\beta}}}{\Gamma(\alpha_1 + \alpha_2)\beta^{\alpha_1+\alpha_2}} \times \frac{v^{\alpha_1-1}(1 - v)^{\alpha_2-1}\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}.$$

Therefore,

$$f_{UV}(u,v) = \frac{u^{\alpha_1+\alpha_2-1}e^{-\frac{u}{\beta}}}{\Gamma(\alpha_1+\alpha_2)\beta^{\alpha_1+\alpha_2}} \times \frac{v^{\alpha_1-1}(1-v)^{\alpha_2-1}}{B(\alpha_1,\alpha_2)},$$

where, $B(\alpha_1,\alpha_2) = \int_0^1 v^{\alpha_1-1}(1-v)^{\alpha_2-1}dv = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1+\alpha_2)}$ is the Beta function.

We observe that

a. U, V are independent.

b. $U \sim \Gamma(\alpha_1+\alpha_2, \beta)$.

c. $V \sim \text{Beta}(\alpha_1, \alpha_2)$.

Use the previous result to find the distribution of $\frac{r_i^2}{n-k-2}$. For our discussion here:

What is X?

What is Y?

What is $X+Y$?

Externally studentized residuals

Consider the ratio $t_i = \frac{e_i}{S_{e(i)}\sqrt{1-h_{ii}}}$, where $S^2_{e(i)}$ is the unbiased estimator of σ^2 after data point i is deleted from the data set. Multiply/divide $t_i^2 = \frac{e_i^2}{S^2_{e(i)}(1-h_{ii})}$ by $n-k-2$

$$\frac{e_i^2}{S^2_{e(i)}(1-h_{ii})} = \frac{e_i^2(n-k-2)}{(n-k-2)S^2_{e(i)}(1-h_{ii})}$$

We have seen that $(n-k-2)S^2_{e(i)} = (n-k-1)S_e^2 - \dfrac{e_i^2}{1-h_{ii}}$

Replace this in the denominator

$$= \frac{e_i^2(n-k-2)}{[(n-k-1)S_e^2 - \frac{e_i^2}{1-h_{ii}}](1-h_{ii})}$$

$$= \frac{e_i^2(n-k-2)}{(n-k-1)S_e^2(1-h_{ii}) - e_i^2}$$

Note: $r_i^2 = \dfrac{e_i^2}{S_e^2(1-h_{ii})}$

Replace $e_i^2 = r_i^2 S_e^2(1-h_{ii})$

Show that the ratio is equal to $\frac{B}{1-B}(n-k-2))$, where $B = \frac{r_i^2}{n-k-1}$.

Note:

$\frac{r_i^2}{n-k-1} \sim \text{beta}(\frac{1}{2}, \frac{1}{2}(n-k-2))$ (internally studentized residual).

Finally, it can be shown (see homework 10) that if $B \sim \text{beta}(\frac{1}{2}\alpha, \frac{1}{2}\beta)$ then $\frac{\beta B}{\alpha(1-B)} \sim F_{\alpha,\beta}$.

Use this result to show that $t_i^2 = \frac{B}{1-B}(n-k-2) \sim F_{1,n-k-2}$ and therefore, $t_i = \frac{e_i}{S_{e(i)}\sqrt{1-h_{ii}}} \sim t_{n-k-2}$.

Variable selection

Some general comments
Suppose we have several predictors available.

a. We would like to use only few of them. Why?

b. If multicollinearity is not present, how would you choose the predictors that stay in the model?

c. However, when multicollinearity is present, the decision about which predictors we should keep becomes more difficult. Why?

d. In general, the predictors selected to be removed should have the least effect on the response variable.

e. If we remove important predictors what happens to the least squares estimator? Think in terms of "short" and "long" regression. In addition, what happens to the estimator of σ^2?

f. Therefore, we should keep a small number of predictors to reduce multicollinearity, but at the same time make sure that the bias and estimate of σ^2 is low.

Effects on the regression when predictors are removed from the model.

A. Effect on $\boldsymbol{\beta}$.

Suppose the correct model is $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. After partitioning $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2})$ and $\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta_1} \\ \boldsymbol{\beta_2} \end{pmatrix}$ we can write the model as $\mathbf{y} = \mathbf{X_1}\boldsymbol{\beta_1} + \mathbf{X_2}\boldsymbol{\beta_2} + \boldsymbol{\epsilon}$, therefore $E\mathbf{y} = \mathbf{X_1}\boldsymbol{\beta_1} + \mathbf{X_2}\boldsymbol{\beta_2}$.

Suppose now we remove the term $\mathbf{X_2}\boldsymbol{\beta_2}$, so the model we decided to use is $\mathbf{y} = \mathbf{X_1}\boldsymbol{\beta_1} + \boldsymbol{\epsilon}$.

Therefore $\hat{\boldsymbol{\beta}}_{\mathbf{1}} =$ and

$E(\hat{\boldsymbol{\beta}}_{\mathbf{1}}) =$

What do you observe?

B. Effect on the estimator of σ^2.

Based on the full model we have $S_e^2 = \frac{\mathbf{y}'(\mathbf{I}-\mathbf{H})\mathbf{y}}{n-k-1}$. Suppose that the short regression has p parameters. Then $S_p^2 = \frac{\mathbf{y}'(\mathbf{I}-\mathbf{H_1})\mathbf{y}}{}$. (Complete the denominator). When would S_p^2 be unbiased?

We know that $E(S_e^2) = \sigma^2$, but what about S_p^2? Using properties of the trace we find the following:

$$\begin{aligned}
ES_p^2 &= \frac{1}{n-p} E\mathrm{tr}\mathbf{y}'(\mathbf{I}-\mathbf{H_1})\mathbf{y} \\
&= \frac{1}{n-p}\mathrm{tr}(\mathbf{I}-\mathbf{H_1})E(\mathbf{y}\mathbf{y}') \\
&= \frac{1}{n-p}\mathrm{tr}(\mathbf{I}-\mathbf{H_1})(\sigma^2\mathbf{I}+\mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}'\mathbf{X}') \\
&= \sigma^2 + \frac{\boldsymbol{\beta}'\mathbf{X}'(\mathbf{I}-\mathbf{H_1})\mathbf{X}\boldsymbol{\beta}}{n-p}
\end{aligned}$$

We conclude that $E(S_p^2 - S_e^2) =$

C. Effect on the variance covariance matrix of $\hat{\boldsymbol{\beta}}$.

If we use the short regression then the variance of the estimator of $\boldsymbol{\beta_1}$ is $\mathrm{var}(\hat{\boldsymbol{\beta}}_\mathbf{1}) =$

In the long regression the variance of the estimator of $\boldsymbol{\beta_1}$ using partial regression will be: $\mathrm{var}(\hat{\boldsymbol{\beta}}_\mathbf{1.2}) = \sigma^2 \left[\mathbf{X_1^*}'\mathbf{X_1^*}\right]^{-1}$. What is $\mathbf{X_1^*}$?

Simplify $\mathrm{var}(\hat{\boldsymbol{\beta}}_\mathbf{1.2})$ by replacing $\mathbf{X_1^*}$ and expanding.

$\mathrm{var}(\hat{\boldsymbol{\beta}}_\mathbf{1.2}) =$

We need to find a way to compare $\mathrm{var}(\hat{\boldsymbol{\beta}}_\mathbf{1})$ with $\mathrm{var}(\hat{\boldsymbol{\beta}}_\mathbf{1.2})$. A result from linear algebra will help us. In general, if $\mathbf{A}$ and $\mathbf{B}$ are matrices, and $\mathbf{A^{-1}} \geq \mathbf{B^{-1}}$ then $\mathbf{A} \leq \mathbf{B}$. Therefore if $\mathbf{A^{-1}} - \mathbf{B^{-1}} \geq \mathbf{0}$ then $\mathbf{A} - \mathbf{B} \leq \mathbf{0}$. So let's compare the inverse of $\mathrm{var}(\hat{\boldsymbol{\beta}}_\mathbf{1})$ and $\mathrm{var}(\hat{\boldsymbol{\beta}}_\mathbf{1.2})$.

$$\begin{aligned}
\left[\mathrm{var}(\hat{\boldsymbol{\beta}}_\mathbf{1})\right]^{-1} - \left[\mathrm{var}(\hat{\boldsymbol{\beta}}_\mathbf{1.2})\right]^{-1} &= \\
\left[\mathrm{var}(\hat{\boldsymbol{\beta}}_\mathbf{1})\right]^{-1} - \left[\mathrm{var}(\hat{\boldsymbol{\beta}}_\mathbf{1.2})\right]^{-1} &\geq
\end{aligned}$$

We conclude that

Therefore the variance covariance matrix decreases when we drop predictors.

As an example of the above consider the two models:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \epsilon_i \\ y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i \end{aligned}$$

Show that $var(\hat{\beta}_1) \leq var(\hat{\beta}_{1.2})$. Note: Consider $var(\hat{\beta}_1)$ from simple regression and then find $var(\hat{\beta}_{1.2})$ using the results from the discussion on multicollinearity.

D. Effect on the fitted values.
If we use the short regression, the fitted values are $\hat{\mathbf{y}}_{\mathbf{p}} =$
Therefore, $E(\hat{\mathbf{y}}_{\mathbf{p}}) = \mathbf{H_1 X}\boldsymbol{\beta} \neq \mathbf{X}\boldsymbol{\beta}$, therefore $\hat{\mathbf{y}}_{\mathbf{p}}$ is biased. The bias is the difference between the expected value of $\hat{\mathbf{y}}_{\mathbf{p}}$ and the expected value of $\hat{\mathbf{y}}$.

$$\begin{aligned} \text{Bias}(\hat{\mathbf{y}}_{\mathbf{p}}) = \mathbf{B} &= E(\hat{\mathbf{y}}_{\mathbf{p}}) - E(\hat{\mathbf{y}}) = \mathbf{H_1 X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta} = -(\mathbf{I} - \mathbf{H_1})\mathbf{X}\boldsymbol{\beta} \\ & \quad \text{Compute the sum of the squared bias values of the vector} \\ & \quad -(\mathbf{I} - \mathbf{H_1})\mathbf{X}\boldsymbol{\beta} \\ \mathbf{B'B} &= \\ & \quad \text{Standardize it by dividing by } \sigma^2 \\ \frac{\mathbf{B'B}}{\sigma^2} &= \\ & \quad \text{From part [B] we have} \\ ES_p^2 &= \sigma^2 + \frac{\boldsymbol{\beta}'\mathbf{X}'(\mathbf{I} - \mathbf{H_1})\mathbf{X}\boldsymbol{\beta}}{n-p} \\ \boldsymbol{\beta}'\mathbf{X}'(\mathbf{I} - \mathbf{H_1})\mathbf{X}\boldsymbol{\beta} &= \\ \frac{\mathbf{B'B}}{\sigma^2} &= \frac{E[SSE_p]}{\sigma^2} - (n-p) \end{aligned}$$

An estimate of this standardized bias is $\frac{SSE_p}{S_e^2} - (n-p)$, where S_e^2 is the estimator of σ^2 in the full model. If this is close to zero it means that the bias introduced in the model by dropping those particular predictors is small.

Example:
Suppose we have 8 predictors and we are considering the following two reduced models:
(a) $\mathbf{y}$ on $\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}, \mathbf{x_4}, \mathbf{x_5}$
(b) $\mathbf{y}$ on $\mathbf{x_1}, \mathbf{x_2}, \mathbf{x_3}$

Explain how we apply the previous result to choose between the two models.

Answer:

E. Mallows' C_p criterion.
The previous note takes into account only the bias of $\hat{\mathbf{y}}_\mathbf{p}$. What if we also consider the variance of $\hat{\mathbf{y}}_\mathbf{p}$. So let's examine the MSE of $\hat{\mathbf{y}}_\mathbf{p}$.

Aside note: If $\hat{\theta}$ is the estimate of θ then the MSE (mean square error) is defined as $MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = \mathrm{var}(\hat{\theta}) + B^2$, where $B = E(\hat{\theta}) - \theta$.

In the discussion here, we are dealing with the vector $\hat{\mathbf{y}}_\mathbf{p}$ and its MSE is defined as:
$MSE(\hat{\mathbf{y}}_\mathbf{p}) = \mathrm{var}(\hat{\mathbf{y}}_\mathbf{p}) + \mathbf{B}\mathbf{B}'$.

What is $\mathrm{var}(\hat{\mathbf{y}}_\mathbf{p})$?

What is $\mathbf{B}$?

Therefore,
$MSE(\hat{\mathbf{y}}_\mathbf{p}) =$

It is easier to compute the trace of the $MSE(\hat{\mathbf{y}}_\mathbf{p})$.

$\mathrm{tr}[MSE(\hat{\mathbf{y}}_\mathbf{p})] =$

Divide this by σ^2 to standardize and use the result from part (D) we get the Mallows' C_p criterion $C_p = \frac{SSE_p}{S_e^2} - (n - p) + p$.

We conclude that if the bias introduced from dropping predictors from the model is small then $C_p \approx p$.