Stats 201B (W14) Regression Analysis: Model Building, Fitting, and Criticism

- Textbook: J. J. Faraway (2005). "Linear Models with R," Chapman & Hall.

Chapter 1. Introduction

Before you start

- Look before you leap!
- Understand the physical problem and objective
- Put the problem into stat. terms
- Understand how the data were collected.
  - experimental data
  - observational data (survey data)
  - nonresponses/ missing values?
  - how are the data coded (for qualitative variables)?
  - possible errors?

Initial data analysis

- numerical summaries (mean, sd, min, max, cor, etc.)
- graphical summaries (boxplots, histograms, scatter plots, etc.)
- look for outliers, data-entry errors, skewed or unusual distributions
- cleaning and preparing data for analysis is an important part in practice

Example data(pima)

- Any missing or unusual observations?
- What does 0 represent for each variable?
- missing values are coded in various ways in practice!

Regression Analysis models the relationship between

- response y (output or dependent variable)
- and predictors $x_1, \ldots, x_k$ (input, independent or explanatory variables)

Types of regressions

- simple regression: $k = 1$
- multiple regression: $k > 1$
- response y is often continuous
- predictors can be continuous or discrete (categorical)
- ANOVA (analysis of variance): all predictors are qualitative
- ANCOVA (analysis of covariance): a mix of continuous and discrete predictors
- When response y is discrete, **logistic regression** or **Poisson regression** could be used.
- multivariate multiple regression: multiple responses $y_1, \ldots, y_m$

Objectives of regression analysis

- prediction of future observations
- assessment of effects of, or relationship betweens x's and y

- description of data structure

History

- Gauss developed **least squares** methods in early 1800.
- Galton coined the term "regression to mediocrity" in 1825 when he studied the heights of sons and fathers as

$$\frac{y-\bar{y}}{SD_y} = r\frac{x-\bar{x}}{SD_x}$$

  where r is the correlation between x and y.
- phenomenon (**regression effect**): sons of tall fathers tend to be tall but not as tall as their fathers while sons of short fathers tend to be short but not as short as their fathers.

Example data(stat500)

Chapter 2. Linear Model/Estimation

1. Linear Model

Consider a general model

$$Y = f(x_1, \ldots, x_k) + \epsilon$$

where $f()$ is an unknown function and ϵ is random error. This model is too general and not useful at all. A more practical approach is to restrict $f()$ to some parametrical form, say f is a linear function of the predictors as follows.

Linear Model:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \epsilon,$$

where y is the response (dependent variable), x_i are predictors (covariates, independent variables), β_i are unknown regression coefficients, and ϵ is a random error. Assume that $E(\epsilon) = 0$ and $var(\epsilon) = \sigma^2$.

Remarks:

1. In a linear model, the parameters β_i enter linearly, the predictors do not have to be linear.

   $$y = \beta_0 + \beta_1 x_1 + \beta_2 \log(x_2) + \beta_3 x_1 x_2 + \epsilon$$

   is a linear model.

   $$y = \beta_0 + \beta_1 x_1^{\beta_2} + \epsilon$$

   is not a linear model.

2. linear models seem restrictive, but are actually very flexible, because predictors can be transformed and combined in many ways.

3. Truly nonlinear models are rarely absolutely necessary (unless supported by theory).

For data with n observations $(y_i, x_{i1}, \ldots, x_{ik})$, the model is:

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

- Assume that ϵ_i are independent with mean 0 and common variance σ^2.

In matrix form, the model is

$$y = X\beta + \epsilon, E(\epsilon) = 0, cov(\epsilon) = \sigma^2 I$$

where

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \ldots & x_{nk} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

- X, an $n \times p$ matrix, is called the **model matrix**, where $p = k + 1$ here.
- β is a vector of unknown regression coefficients.
- ϵ is a vector of random errors.

2. Estimation

The **least squares estimate** of β minimizes

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \epsilon^T \epsilon = (y - X\beta)^T (y - X\beta)$$

Let

$$\frac{\partial L}{\partial \beta} = -2X^T(y - X\beta) = 0$$

The **normal equations** are

$$X^T X\beta = X^T y$$

The **least squares estimate** (LSE) is

$$\hat{\beta} = (X^T X)^{-1} X^T y,$$

The **predicted or fitted values** are

$$\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y = Hy,$$

where $H = X(X^T X)^{-1} X^T$ is the **hat matrix**.

- The hat matrix plays an important role in residual analysis.
- Property: H is idempotent, i.e., $H^2 = H$.

The **residuals** are

$$\hat{\epsilon} = y - X\hat{\beta} = y - Hy = (I - H)y$$

The residual sum of squares (RSS) is

$$RSS = \hat{\epsilon}^T \hat{\epsilon} = y^T (I - H)(I - H)y = y^T (I - H)y$$

An unbiased estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{RSS}{n - p} = MSE,$$

where $n - p$ is the **degrees of freedom** of the model.

The least squares estimate has a nice **geometrical interpretation** (Fig. 2.1).

Properties of the least squares estimate $\hat{\beta}$

1. LSE $\hat{\beta}$ is an unbiased estimator of β, i.e., $E(\hat{\beta}) = \beta$.
2. The covariance depends on σ and the model matrix, $cov(\hat{\beta}) = \sigma^2(X^TX)^{-1}$.
3. LSE $\hat{\beta}$ is max. likelihood estimator (MLE) if ϵ has a normal distribution.
4. **Gauss-Markov Theorem**: If the model is correct and $E(\epsilon) = 0$ and $cov(\epsilon) = \sigma^2 I$, then $\hat{\beta}$ is the best linear unbiased estimator (BLUE).

Examples of calculating $\hat{\beta}$

1. mean only model: $y_i = \mu + \epsilon_i$

$$X = 1_n,\ \beta = \mu,$$

$$\hat{\beta} = (X^TX)^{-1}X^Ty = (1^T1)^{-1}1^Ty = \frac{1}{n}1^Ty = \bar{y}$$

2. simple linear regression (one predictor): $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

$$y = X\beta + \epsilon \text{ means } \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

A simpler approach:

$$y_i = \beta_0 + \beta_1\bar{x} + \beta_1(x_i - \bar{x}) + \epsilon_i = \beta_0' + \beta_1(x_i - \bar{x}) + \epsilon_i$$

Then

$$X = \begin{pmatrix} 1 & x_1 - \bar{x} \\ \vdots & \vdots \\ 1 & x_n - \bar{x} \end{pmatrix}, X^TX = \begin{pmatrix} n & 0 \\ 0 & \sum(x_i - \bar{x})^2 \end{pmatrix}$$

$$\begin{pmatrix} \hat{\beta}_0' \\ \hat{\beta}_1 \end{pmatrix} = (X^T X)^{-1} X^T y$$

$$= \begin{pmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{\sum(x_i - \bar{x})^2} \end{pmatrix} \begin{pmatrix} 1 & \cdots & 1 \\ x_1 - \bar{x} & \cdots & x_n - \bar{x} \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

$$\hat{\beta}_0' = \frac{1}{n} \sum y_i = \bar{y}, \quad \hat{\beta}_1 = \frac{\sum(x_i - \bar{x}) y_i}{\sum(x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Remarks:

- In general, it is quite complicated to compute the inverse of $X^T X$.
- However, the computation is easy if $X^T X$ is diagonal, which happens for designed experiments.
- When $X^T X$ is diagonal, estimates of $\hat{\beta}_1, \ldots, \hat{\beta}_p$ are uncorrelated. (why?)

3. Goodness of fit

The coefficient of multiple determination

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{\sum(\hat{y}_i - y_i)^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{RSS}{TotalSS}$$

is a measure of goodness of fit.

- It tells the percentage of variation in y explained by the regression model.
- $0 \leq R^2 \leq 1$.
- Large R^2 (close to 1) indicates a good fit.

- A large R^2 does NOT necessarily imply that the regression model is a good one.
- Caution: If the model does not include the intercept, R^2 defined here is not meaningful.

Example data(gala)

4. Identifiability

LSE normal equations:

$$X^T X\hat{\beta} = X^T y$$

where X is an $n \times p$ matrix, $X^T X$ is a $p \times p$ matrix.

- If $X^T X$ is singular ($rank(X^T X) < p$), the inverse does not exist and $\hat{\beta}$ is unidentifiable.
- It happens when some columns of X are linearly dependent.
- Common in experimental design, e.g., 2-sample comparison experiment or one-way layout.

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad i = 1, 2; j = 1, \ldots, m$$

- Caution: R fits the largest identifiable model automatically.
- In practice, be careful about near un-identifiability (near collinearity).

Example data(gala)

Gauss-Markov Theorem: LSE $\hat{\beta}$ is BLUE.

Proof. Let $A = (X^TX)^{-1}X^T$. Then $AX = (X^TX)^{-1}X^TX = I$,

$$AA^T = (X^TX)^{-1}X^TX(X^TX)^{-1} = (X^TX)^{-1}.$$

LSE $\hat{\beta} = Ay$ is a linear combination of the random response vector y.

Let $\hat{\beta}^* = By$ be any other linear transformation of y, where B is a $p \times n$ matrix of fixed coefficients. Denote $D = B - A$. Then

$$E(\hat{\beta}^*) = E(By) = BE(y) = BX\beta = (D + A)X\beta = DX\beta + \beta.$$

So $\hat{\beta}^*$ is unbiased iff $DX = 0$.

$$V(\hat{\beta}^*) = V(By) = BV(y)B^T = B(\sigma^2 I)B^T = \sigma^2 BB^T$$

$$= \sigma^2(A + D)(A + D)^T = \sigma^2(AA^T + DD^T)$$

Note $AD^T = (X^TX)^{-1}X^TD^T = 0$ and $D^TA = 0$ since $DT = 0$ for any unbiased estimator. So

$$V(\hat{\beta}^*) = \sigma^2((X^TX)^{-1} + DD^T)$$

$$V(\hat{\beta}^*) = V(\hat{\beta}) + \sigma^2 DD^T \geq V(\hat{\beta})$$

because $DD^T \geq 0$ is semi-positive definite (i.e., for any vector a, $a^TDD^Ta = \tilde{a}^T\tilde{a} = \sum \tilde{a}_i^2 \geq 0$).

Gauss-Markov Theorem implies that for any linear combination of β, say $a^T\beta$, LSE $a^T\hat{\beta}$ is BLUE.

$$V(a^T\hat{\beta}) = a^TV(\hat{\beta})a = \sigma^2 a^T(X^TX)^{-1}a$$

$$V(a^T\hat{\beta}^*) = a^TV(\hat{\beta}^*)a = a^TV(\hat{\beta})a + \sigma^2 a^TDD^Ta \geq \sigma^2 a^T(X^TX)^{-1}a.$$

Chapter 3. Linear Model/Inference

$$y = X\beta + \epsilon, E(\epsilon) = 0, cov(\epsilon) = \sigma^2 I$$

For statistical inference, it is further assumed that the errors follow a **normal distribution**, i.e., $\epsilon \sim N(0, \sigma^2 I)$.

3.1 Hypothesis Tests

to compare two **nested** linear models,

- Null H_0: smaller model ω, $dim(\omega) = q$
- Alternative H_1: larger model Ω, $dim(\Omega) = p$

Basic idea:

- Geometric representation (Fig. 3.1)

- reject H_0 if $RSS_\omega - RSS_\Omega$ is large.
- or reject H_0 if the ratio $(RSS_\omega - RSS_\Omega)/RSS_\Omega$ is large.

The F test statistic

$$F = \frac{(RSS_\omega - RSS_\Omega)/(df_\omega - df_\Omega)}{RSS_\Omega/df_\Omega} = \frac{(RSS_\omega - RSS_\Omega)/(p-q)}{RSS_\Omega/(n-p)}$$

- This is indeed the maximum likelihood ratio test statistic.
- When $\epsilon \sim N(0, \sigma^2 I)$, the F statistic has an F distribution with df $p-q$ and $n-p$, i.e., $F \sim F_{p-q,n-p}$, under the null model ω.
- Reject H_0 at level α if $F > F_{\alpha,p-q,n-p}$. (upper αth percentile, $F^{(\alpha)}_{p-q,n-p}$)
- P-value $= P(F_{p-q,n-p} > F)$.
- works when ω is a subset or subspace of Ω.

3.2 Testing Examples

a. Test of all the predictors. Are any of the predictors useful? Model: $y = X\beta + \epsilon$ or $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \epsilon$

$$\begin{aligned} H_0 &: \beta_1 = \ldots = \beta_k = 0 \\ H_1 &: \beta_j \neq 0 \text{ for at least one } j \geq 1 \end{aligned}$$

Here $p = k+1$, $q = 1$, ω is $y = \beta_0 + \epsilon$ and $\hat{\beta}_0 = \bar{y}$ under ω. So

$$\begin{aligned} RSS_\omega &= (y-\bar{y})^T(y-\bar{y}) = \sum (y_i - \bar{y})^2 = SS_T \text{ (or } TSS) \\ RSS_\Omega &= (y - X\hat{\beta})^T(y - X\hat{\beta}) = SS_E \text{ (or } RSS) \end{aligned}$$

ANOVA table for testing all predictors

Source	DF	SS	MS	F
Regression	$p-1$	SS_{reg}	$MS_{reg} = \frac{SS_{reg}}{p-1}$	$F = \frac{MS_{reg}}{MS_E}$
Error (or residual)	$n-p$	SS_E	$MS_E = \frac{SS_E}{n-p}$	
Total	$n-1$	SS_T		

P-value $= P(F_{p-1,n-p} > F)$

Remarks

- A failure of reject the null is not the end of the analysis!
- "Fail to reject" the null is not the same as "accept" the null.
- Rejecting the null does not imply the alternative model is the best model.

Example data(savings)

b. Testing just one predictor. Can one particular predictor be dropped?
Model: $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \epsilon$

$$H_0 : \beta_j = 0, \quad H_1 : \beta_j \neq 0$$

Here $p = k + 1$, $q = p - 1 = k$ and ω has one parameter less than Ω. So RSS_ω is obtained by fitting a regression without predictor x_j.

The F test statistic is

$$F = \frac{(RSS_\omega - RSS_\Omega)/(p-q)}{RSS_\Omega/(n-p)} = \frac{(RSS_\omega - RSS_\Omega)}{RSS_\Omega/(n-p)}$$

$$\sim F_{1,n-p} \text{ under } H_0$$

Alternatively, use a t-test

$$t = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{(X^T X)^{-1}_{jj}}} \sim t_{n-p} \text{ under } H_0$$

- Reject H_0 at level α if $|t| > t_{\alpha/2,n-p}$
- P-value $= P(|t_{n-p}| \geq |t|)$.

Remarks

- Fact: $F = t^2$ when testing one predictor only.
- The result (of testing $\beta_j = 0$) depends on other predictors x_i in the model. (Why?)

Example data(savings)

c. Testing a pair of predictors.
Suppose both x_i and x_j have p-values > 0.05. Can we eliminate both from the model?

d. Testing a subspace. For model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon$$

Can we combine two variables $(x_1 + x_2)$ to simplify the model as

$$y = \beta_0 + \beta_{12}(x_1 + x_2) + \beta_3 x_3 + \ldots + \beta_k x_k + \epsilon$$

The null hypothesis is a linear subspace of the original model.
What is the null hypothesis here?

Example data(savings)

3.3 Permutation Tests

- The F or t test assumes the errors are normally distributed.
- A permutation test does not require the normality assumption.
- Procedure:

- consider all $n!$ permutations of the response variable
- compute the F statistic for each permutation
- compute proportion of F statistics exceed the observed F statistic for the original response

- The proportion is estimated by the p-value calculated in the usual way (assuming normal errors).
- When n is large, do a random sample of the $n!$ permutations.
- To test one predictor (x_i): permute that predictor (x_i) rather than the response y.

Example data(savings)

3.4 Confidence Intervals (CIs) for β

The 100(1-α)% CI for β_j is

$$\hat{\beta}_j \pm t_{\alpha/2,n-p} \cdot se(\hat{\beta}_j) = \hat{\beta}_j \pm t_{\alpha/2,n-p}\hat{\sigma}\sqrt{(X^TX)^{-1}_{jj}}$$

- The $100(1-\alpha)\%$ CI for β_j does not contain 0 if and only if the t test rejects $H_0 : \beta_j = 0$ (vs $H_1 : \beta_j \neq 0$) at level α.

The 100(1-α)% confidence region for β is

$$(\hat{\beta}-\beta)^T(X^TX)(\hat{\beta}-\beta) \leq p\hat{\sigma}^2 F_{\alpha,p,n-p}$$

- The confidence region is ellipsoidally shaped.
- The $100(1-\alpha)\%$ confidence region for β does not contain the origin if and only if the F test rejects $H_0 : \beta = 0$ at level α.
- CIs and confidence region may give conflict conclusions. Why?

- The conclusion from confidence region is preferred to individual CIs.

Example data(savings)

- Distinguish the correlation between predictors, $cor(x_i, x_j)$, and correlation between estimators, $cor(\hat{\beta}_i, \hat{\beta}_j)$

3.5 CIs for Predictions

For a new set of predictors x_0, the predicted response is $\hat{y}_0 = x_0^T\hat{\beta}$.

$$var(\hat{y}_0) = var(x_0^T\hat{\beta}) = x_0^T cov(\hat{\beta})x_0 = \sigma^2 x_0^T(X^TX)^{-1}x_0$$

- The 100(1-α)% **CI** on the **mean response** is

$$\hat{y}_0 \pm t_{\alpha/2,n-p}\hat{\sigma}\sqrt{x_0^T(X^TX)^{-1}x_0}$$

- The 100(1-α)% **prediction interval** for a **future observation** is

$$\hat{y}_0 \pm t_{\alpha/2,n-p}\hat{\sigma}\sqrt{1 + x_0^T(X^TX)^{-1}x_0}$$

- The prediction intervals become wider as x_0 moves away from the data center.
- Caution on extrapolation (when the new x_0 is outside the range of the original data)

Example data(gala)

3.6 Designed Experiments

In a designed experiment, we have some control over X and can make it **orthogonal** or near orthogonal. Two important design features:

- orthogonality
- randomization

which allows us to make stronger conclusions from the analysis.

Suppose $X = (X_1 X_2)$ such that $X_1^T X_2 = 0$ and $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$.

$$y = X\beta + \epsilon = X_1\beta_1 + X_2\beta_2 + \epsilon$$

$$X^T X = \begin{pmatrix} X_1^T X_1 & X_1^T X_2 \\ X_2^T X_1 & X_2^T X_2 \end{pmatrix} = \begin{pmatrix} X_1^T X_1 & 0 \\ 0 & X_2^T X_2 \end{pmatrix}$$

$$\hat{\beta} = (X^T X)^{-1} X^T y = \begin{pmatrix} X_1^T X_1 & 0 \\ 0 & X_2^T X_2 \end{pmatrix}^{-1} \begin{pmatrix} X_1^T y \\ X_2^T y \end{pmatrix}$$

$$\hat{\beta}_1 = (X_1^T X_1)^{-1} X_1^T y, \quad \hat{\beta}_2 = (X_2^T X_2)^{-1} X_2^T y$$

- Estimate of β_1 does not depend on whether X_2 is in the model or not.
- Does the significance of β_1 depend on whether X_2 is in the model or not? Why?
- Randomization provides protection against the influence of unknown **lurking** variables

Example data(odor) vs data(savings)

3.7 Observational Data

- Interpreting models built on observational data is problematic.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

- What does $\hat{\beta}_1$ mean?

- "A unit change in x_1 will produce a change of $\hat{\beta}_1$ in the response"?
- "$\hat{\beta}_1$ is the effect of x_1 when all the other specified predictors are held constant"?
- prediction is more stable than parameter estimation.

Example data(savings)

3.8 Practical Difficulties

$$y = X\beta + \epsilon$$

We have gone over the linear model theory of estimation and inference. What are the difficulties?

Albert Einstein: "So far as theories of mathematics are about reality; they are not certain; so far as they are certain, they are not about reality."

- Nonrandom samples
- Choice and range of predictors
- Model misspecification. George Box:
  "All models are wrong but some are useful."
- Publication and experimenter bias
- Practical and statistical significance

Chapter 4. Linear Model/Diagnostics

$$Y = X\beta + \epsilon,\ \epsilon \sim N(0, \sigma^2 I)$$

Regression Diagnostics: Use residuals to check assumptions

- Errors: independent, equal variance, normally distributed?
- Model structure: $E(y) = X\beta$?
- Unusual observations: outliers, influential cases?

4.1 Checking Error Assumptions

Residuals. $\hat{\epsilon} = y - \hat{y} = (I - H)y$,

$\text{cov}(\hat{y}, \hat{\epsilon}) = \text{cov}(Hy, (I-H)y) = H\text{cov}(y)(I-H)^T = \sigma^2 H(I-H) = 0$

- $\hat{y}$ and $\hat{\epsilon}$ are uncorrelated if errors are independent.

  $\text{cov}(\hat{\epsilon}) = \text{cov}((I - H)y) = (I - H)\text{cov}(y)(I - H)^T = \sigma^2(I - H)$

- $\hat{\epsilon}_i$ may be correlated even if ϵ_i are independent.
- $\hat{\epsilon}_i$ may not have equal variance even if ϵ_i do.
- but we still use residuals as the impact is usually small.

Residual plots

- residuals vs. fitted: plot $\hat{\epsilon}_i$ (or r_i defined in section 4.2) vs. $\hat{y}_i$
- residuals vs. predictors: plot $\hat{\epsilon}_i$ (or r_i) vs. x_{ij} for each j
- check model structure, constant variance, nonlinearity, outliers (see Fig. 4.1)

Example data(savings)

How to deal with nonconstant variance?

- Use weighted least squares (Section 6.1) if we know the form of nonconstant variance.
- Transform the response

How to choose a function $h()$ so that $\mathrm{var}(h(y))$ is constant?

$$h(y) = h(Ey) + (y - Ey)h'(Ey) + \cdots$$

$$\mathrm{var}(h(y)) = h'(Ey)^2 \mathrm{var}(y) + \cdots$$

Ignoring higher order terms, for $\mathrm{var}(h(y))$ to be constant, make

$$h'(Ey) \propto (\mathrm{var}(y))^{-1/2}$$

$$h(y) = \int (\mathrm{var}(y))^{-1/2} dy = \int \frac{1}{sd(y)} dy$$

Two commonly used transformations:

- If $\mathrm{var}(y) = \mathrm{var}(\epsilon) \propto (Ey)^2$, then $h(y) = \log(y)$
- If $\mathrm{var}(y) = \mathrm{var}(\epsilon) \propto (Ey)$, then $h(y) = \sqrt{y}$

Example data(gala)

Normality? Normal Q-Q plot:

- plot sorted $\hat{\epsilon}_i$ (or r_i) against standard normal quantiles $\Phi^{-1}(\frac{i}{n+1})$ for $i = 1, \ldots, n$, where Φ is the cdf for $N(0, 1)$.
- points shall close to a straight line if the model is correct and errors are $N(0, \sigma^2 I)$.

What to do when non-normality is found?

- For short-tailed distribution, not a serious problem and can be ignored.
- For skewed errors, try to transform the response.
- For long-tailed errors, need to do something ...

Correlated errors?

- For temporal or spatial data, errors are often correlated.
- graphical check:
  - plot $\hat{\epsilon}$ against time (or run order)
  - plot $\hat{\epsilon}_i$ against $\hat{\epsilon}_{i-1}$
- Durbin-Watson test

$$DW = \frac{\sum_{i=2}^{n}(\hat{\epsilon}_i - \hat{\epsilon}_{i-1})^2}{\sum_{i=2}^{n}\hat{\epsilon}_i^2}$$

Example data(airquality)

4.2 Finding unusual observations

Leverage: $h_i = h_{ii}$, the ith diagonal element of $H = X(X^TX)^{-1}X^T$

$$\text{cov}(\hat{\epsilon}) = \sigma^2(I - H) \text{ so } \text{var}(\hat{\epsilon}_i) = \sigma^2(1 - h_{ii})$$

- A large leverage h_{ii} will make $\text{var}(\hat{\epsilon}_i)$ small.
- Points far from the center of the x space have large leverages.
- $\sum_i h_{ii} = p$.
- Rule of thumb: look more closely if $h_{ii} > 2p/n$.

Standardized Residuals (or internally studentized residuals):

$$r_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

- r_i has approximately variance 1 when the model is correct and assumptions hold.
- r_i may be preferred to $\hat{\epsilon}_i$ (e.g., in R)

Outliers? Which one is an outlier in Fig. 4.10, p. 66?
How to test whether case i is a potential outlier?

- Remove it from the data and recompute the estimates.
- Let $\hat{\beta}_{(i)}$ and $\hat{\sigma}_{(i)}$ be the estimate of β and σ without case i.
- The fitted value for case i is

$$\hat{y}_{(i)} = x_i^T \hat{\beta}_{(i)}$$

- If $y_i - \hat{y}_{(i)}$ is large, case i is an outlier.

$$\text{var}(y_i - \hat{y}_{(i)}) = \text{var}(y_i) + \text{var}(\hat{y}_{(i)}) = \sigma^2 + \sigma^2 x_i^T (X_{(i)}^T X_{(i)})^{-1} x_i$$

Studentized residuals (or jackknife residuals):

$$t_i = \frac{y_i - \hat{y}_{(i)}}{\hat{\sigma}_{(i)}\sqrt{1 + x_i^T (X_{(i)}^T X_{(i)})^{-1} x_i}}$$

- $t_i \sim t_{n-p-1}$ if case i is not an outlier. Why?
- $\hat{\sigma}^2_{(i)}$ is an unbiased estimator of σ^2 and has df=$n-p-1$.
- t_i is a function of r_i (so no need to fit n regressions)

$$t_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}} = r_i \sqrt{\frac{n-p-1}{n-p-r_i^2}}$$

- declare case i as an outlier (with Bonferroni correction) if $|t_i| > t_{\alpha/(2n),n-p-1}$.
- may not work if there are two or more outliers next to each other.
- An outlier in one model may not be an outlier in another model.

What should be done about outliers? Read page 68.

Example data(savings) and data(star)

Influential Cases and Cook's Distance
Let $\hat{\beta}_{(i)}$ be the estimate of β without case i.
Let $\hat{y}_{(i)} = X\hat{\beta}_{(i)}$ be the fitted values without case i.

Cook's distance is

$$D_i = \frac{(\hat{y}_{(i)} - \hat{y})^T(\hat{y}_{(i)} - \hat{y})}{p \cdot \hat{\sigma}^2} = \frac{(\hat{\beta}_{(i)} - \hat{\beta})^T(X^TX)(\hat{\beta}_{(i)} - \hat{\beta})}{p \cdot \hat{\sigma}^2}$$

- Large D_i implies case i is influential (the estimates and fitted values will be quite different with or without case i).

$$D_i = \frac{r_i^2}{p}\frac{h_{ii}}{1 - h_{ii}}$$

- A highly influential case must have a large leverage or a large standardized residual.
- An influential case may, but not necessarily, be an outlier; see Fig. 4.10, p. 66.

Example data(savings)

4.3 Checking the model structure

- residual plots ($\hat{\epsilon}$ vs. $\hat{y}$ and x_i) may reveal problems of model structure, but
- other predictors can impact the relationship between y and x_i

Added variable plot (or partial regression plot) can help isolate the effect of x_i on y

- regress y on all x except x_i, and get residuals $\hat{\delta}$.
- regress x_i on all x except x_i, and get residuals $\hat{\gamma}$.
- plot $\hat{\delta}$ against $\hat{\gamma}$.
- fit a regression line to the plot. The slope is $\hat{\beta}_i$.

Partial residual plot

$$y - \sum_{j \neq i} x_j \hat{\beta}_j = \hat{y} + \hat{\epsilon} - \sum_{j \neq i} x_j \hat{\beta}_j = \hat{\epsilon} + x_i \hat{\beta}_i$$

- plot $\hat{\epsilon} + \hat{\beta}_i x_i$ against x_i.
- The slope of the fitted line is also $\hat{\beta}_i$.
- better for nonlinearity detection than added variable plot.

Example data(savings)

Chapter 5. Problems with the Predictors

What will change if we scale the predictors or response?

- If $x_i \longrightarrow (x_i + a)/b$, then $\hat{\beta}_i \longrightarrow b\hat{\beta}_i$
- If $y \longrightarrow by$, then $\hat{\beta} \longrightarrow b\hat{\beta}$ and $\hat{\sigma} \longrightarrow b\hat{\sigma}$
- The t-tests, F-tests and R^2 are unchanged.
- Scaling or standardizing the predictors and response makes comparisons simpler.

Example data(savings)

Collinearity

- If X^TX is close to singular, we have collinearity (or multicollinearity).
- Collinearity causes serious problems with the estimation of β and interpretation.

Ways to detect collinearity

- Examine correlation matrix.
- Regress x_j on all other predictors, get R_j^2. A large R_j^2 (close to one) indicates a problem.
- Examine the eigenvalues of X^TX. Let $\lambda_1 \geq \ldots \geq \lambda_p$ be the eigenvalues.
- The **condition number**

$$\kappa = \sqrt{\frac{\lambda_1}{\lambda_p}}$$

where $\kappa \geq 30$ is considered large.

Let

$$S_{x_j x_j} = \sum_i (x_{ij} - \bar{x}_j)^2$$

Then

$$\text{var}(\hat{\beta}_j) = \sigma^2 \left(\frac{1}{1 - R_j^2} \right) \frac{1}{S_{x_j x_j}}$$

- $\text{var}(\hat{\beta}_j)$ will be large if $S_{x_j x_j}$ is small.
- $\text{var}(\hat{\beta}_j)$ will be large if R_j^2 is close to 1.
- $\text{var}(\hat{\beta}_j)$ will be minimized if $R_j^2 = 0$.
- **variance inflation factor**: $\text{vif} = \frac{1}{1-R_j^2}$
- design strategy:
  - spread X to make $S_{x_j x_j}$ large.
  - make predictors orthogonal to make $R_j^2 = 0$.
- one cure of collinearity is amputation, i.e., drop some (redundant) predictors.

Example data(seatpos)

Chapter 6. Problems with the Errors

$$Y = X\beta + \epsilon,\ \epsilon \sim N(0, \sigma^2 I)$$

- Assume errors are independent, equal variance, normally distributed
- What to do if errors are not?

6.1 Generalized Least Squares

If we know the error variance structure, say,

$$\text{var}(\epsilon) = \sigma^2\Sigma = \sigma^2 SS^T,$$

where $\Sigma = SS^T$ is the Choleski decomposition and S is a triangular matrix.

$$Y = X\beta + \epsilon$$
$$S^{-1}Y = S^{-1}X\beta + S^{-1}\epsilon$$

Consider the transformed variables and model,

$$Y' = S^{-1}Y, X' = S^{-1}X, \epsilon' = S^{-1}\epsilon$$
$$Y' = X'\beta + \epsilon'$$
$$\text{var}(\epsilon') = \text{var}(S^{-1}\epsilon) = S^{-1}\text{var}(\epsilon)S^{-T} = S^{-1}\sigma^2 SS^T S^{-T} = \sigma^2 I$$

So the Generalized LS (GLS) is an ordinary LS (OLS) for the transformed model,

$$\hat{\beta} = (X'^T X')^{-1} X'^T Y' = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y$$
$$\text{var}(\hat{\beta}) = \sigma^2(X'^T X')^{-1} = \sigma^2(X^T\Sigma^{-1}X)^{-1}$$
$$RSS = (Y' - X'\hat{\beta})^T(Y' - X'\hat{\beta}) = (Y - X\hat{\beta})^T\Sigma^{-1}(Y - X\hat{\beta})$$

Diagnostics should be applied to the residuals $\hat{\epsilon}' = S^{-1}\hat{\epsilon}$.

- Main problem in practice: Σ is unknown
- need to estimate Σ from data, not easy
- need to guess the form of Σ in many cases
- Two popular correlation structures:
  - completely symmetrical (CS): $\Sigma_{ij} = \rho$ for $i \neq j$
  - autoregressive AR(1) : $\Sigma_{ij} = \rho^{|i-j|}$

$$\epsilon_{i+1} = \rho\epsilon_i + \delta_i$$

Example data(longley)

6.2 Weighted Least Squares

WLS is a special case of GLS,

- errors are uncorrelated but have unequal variances

$$\Sigma = diag(1/w_1, \ldots, 1/w_n)$$

- so $S = diag(\sqrt{1/w_1}, \ldots, \sqrt{1/w_n})$ and
- regress $\sqrt{w_i}y_i$ on $\sqrt{w_i}x_i$ (and $\sqrt{w_i}1$)
- E.g., if $\text{var}(\epsilon_i) \propto x_i$ suggest $w_i = 1/x_i$.
- E.g., if y_i is the average of n_i observations, $\text{var}(\epsilon_i) = \sigma^2/n_i$ suggest $w_i = n_i$.
- use $\sqrt{w_i}\hat{\epsilon}_i$ for diagnostics

Example data(fps)

6.3 Testing for Lack of Fit

- If the model is correct, $\hat{\sigma}^2$ is an unbiased estimator of σ^2.
- but we do not know σ^2 and
- $\hat{\sigma}^2$ depends on the model.
- If we have **replicates** in the data, we can have an estimate of σ^2 that does not depend on any model.

Let y_{ij} be the ith observation in the group of replicates j.

- "pure error" estimate of σ^2 is SS_{pe}/df_{pe}, where

$$SS_{pe} = \sum_j \sum_i (y_{ij} - \bar{y}_j)^2$$

$$df_{pe} = \sum_j (\#replicates - 1) = n - \#groups$$

- this is indeed the $\hat{\sigma}^2$ from the one-way ANOVA model with group as a factor.

$$y_{ij} = \mu + \tau_j + \epsilon_{ij}$$

- Caution: whether the replicates are genuine?
- Caution: avoid overfitting with complex model.

Example data(corrosion)

6.4 Robust Regression

- LSE is clearly the best when the errors are normal, but
- other methods are preferred for long-tailed error distributions.

M-Estimation chooses β to minimize

$$\sum_{i=1}^{n} \rho(y_i - x_i^T \beta)$$

- $\rho(x) = x^2$ is just least squares (LS)
- $\rho(x) = |x|$ is called least absolute deviation (LAD) regression or L_1 regression
- Huber's method is a compromise between LS and LAD

$$\rho(x) = \begin{cases} x^2/2 & \text{if } |x| \le c \\ c\,x - c^2/2 & \text{otherwise} \end{cases}$$

  where c should be a robust estimate of σ, e.g., a value proportional to the median of $\hat{\epsilon}$.

Robust regression is similar to WLS with weight function $w(u) = \rho'(u)/u$; see page 99.

- LS: $w(u)$ is constant
- LAD: $w(u) = 1/|u|$
- Huber:

$$w(u) = \begin{cases} 1 & \text{if } |u| \le c \\ c/|u| & \text{otherwise} \end{cases}$$

- weights $w_i = w(u_i)$ depend on the residuals $u_i = y_i - \sum_{j=1}^{p} x_{ij}\hat{\beta}_j$

- so use an iteratively reweighted least squares (IRWLS) approach to fit.

$$\widehat{\text{var}}(\hat{\beta}) = \hat{\sigma}^2 (X^T W X)^{-1}$$

where $W = diag(w_1, \ldots, w_n)$.

Example data(gala)

Least Trimmed Squares (LTS) minimizes

$$\sum_{i=1}^{q} \hat{\epsilon}^2_{(1)}$$

where q is some number less than n and

- $\hat{\epsilon}_{(1)} \leq \ldots \leq \hat{\epsilon}_{(n)}$ are ordered residuals.
- LTS is a **resistant** regression method and can tolerate a large number of outliers.
- default choice of q in `ltsreg` is $\lfloor n/2 \rfloor + \lfloor (p+1)/2 \rfloor$ where $\lfloor x \rfloor$ is the largest integer $\leq x$.

Bootstrap is a general method to obtain standard errors or confidence intervals

- Generate ϵ^* by **sampling with replacement** from $\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n$
- Form $y^* = X\hat{\beta} + \epsilon^*$
- Compute $\hat{\beta}^*$ from (X, y^*)

Read summary on page 105-106.

Example data(gala) and data(star)

Chapter 7 Transformation

7.1 Transforming the Response

Box-Cox Method For $y > 0$, consider $y \longrightarrow g_\lambda(y)$

$$g_\lambda(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log y & \lambda = 0 \end{cases}$$

- choose λ to maximize the log-likelihood

$$L(\lambda) = -\frac{n}{2}\log(RSS_\lambda/n) + (\lambda - 1)\sum \log y_i$$

- A $100(1-\alpha)\%$ CI for λ is

$$\lambda : L(\lambda) > L(\hat{\lambda}) - \frac{1}{2}\chi_1^{2(1-\alpha)}$$

- Box-Cox method gets upset by outliers
- If $\max_i y_i / \min_i y_i$ is small, the Box-Cox method will not have much real effect.
- choose a convenient λ for easy of interpretation
- remember to transform back to the original scale. For example,

$$\log(y) = \beta_0 + \beta_1 x + \epsilon$$

$$y = \exp(\beta_0 + \beta_1 x)\exp(\epsilon)$$

the errors enter **multiplicatively** not **additively** as they usually do.

Example data(savings)

7.2 Transforming the Predictors

- replace x with more than one term, say $f(x) + g(x) + \cdots$
- add cross-product terms, say $x_i x_j$, etc.
- fit different models in different regions, e.g., subset regression

Broken Stick Regression is continuous but non-smooth

$$y = \beta_0 + \beta_1 B_l(x) + \beta_2 B_r(x) + \epsilon$$

where $B_l(x)$ and $B_r(x)$ are two hockey-stick functions:

$$B_l(x) = \begin{cases} c - x & \text{if } x < c \\ 0 & \text{otherwise} \end{cases} \quad B_r(x) = \begin{cases} x - c & \text{if } x > c \\ 0 & \text{otherwise} \end{cases}$$

Polynomials are flexible and smooth

$$y = \beta_0 + \beta_1 x + \ldots + \beta_d x^d + \epsilon$$

- Two ways to choose d: sequentially add or delete terms
- Do no eliminate lower order terms from the model
- Orthogonal polynomials are useful

  $$y = \beta_0 + \beta_1 \phi_1(x) + \ldots + \beta_d \phi_d(x) + \epsilon$$

  where $\phi_i(x)$ is a polynomials of order i, and orthogonal, i.e., $\sum_x \phi_i(x)\phi_j(x) = 0$ for $i \neq j$.
- Response surface models for more than one predictors

  $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon$$

Example data(savings)

Regression Splines are smooth and have local influence property

$$y = \beta_0 + \beta_1\phi_1(x) + \ldots + \beta_d\phi_d(x) + \epsilon$$

where $\phi_i(x)$ are B-spline basis functions on the interval $[a, b]$

- requires knotpoints $t_1, \ldots, t_k$
- A basis function is continuous and smooth (derivative $\phi_i'(x)$ is continuous).
- A cubic B-spline basis function is nonzero on an interval defined by four successive knots and zero elsewhere.
- A cubic B-spline basis function is a cubic polynomial for each subinterval between successive knots.
- A basis function integrates to one over its support.
- The broken stick regression is an example of linear splines.

Example simulated data

$$y = \sin^3(2\pi x^3) + \epsilon, \quad \epsilon \sim N(0, (0.1)^2)$$

Remarks

- many possible ways to transform the predictors
- many different models with various complexity
- Complex models may be good for prediction but difficult to interpret
- For small data sets or where the noise level is high, standard regression is most effective.

Chapter 8 Variable Selection

Intended to select the "best" subset of predictors. Why?

- To explain the data in the simplest way. "The simplest is best".
- To reduce the noise or variation of the estimation.
- To avoid collinearity

How to choose subset?

- **Backward elimination**: Start with all predictors, get rid of the least significant one, repeat until some criterion is reached (i.e., all predictors are significant at a chosen level)
- **Forward selection**: Start with no predictor, add the most significant predictor, repeat until some criterion is reached (i.e., all predictors are not significant at a chosen level)
- **Stepwise selection**: combination of backward and forward selection.
- **Criterion-based procedures**: search over various models (e.g., all subset regressions) and choose a model based on a criterion (e.g., AIC, BIC, Adjusted R^2, C_p, etc.)

Akaike Information Criterion (AIC)

$$AIC = -2 \max \text{log-likelihood} + 2p$$

For regression, $-2 \max \text{log-likelihood} = n \log(RSS/n) + constant$.

Bayes Information Criterion (BIC)

$$BIC = -2 \max \text{log-likelihood} + p \log n$$

Adjusted R^2

$$R_a^2 = 1 - \frac{RSS/(n-p)}{TSS/(n-1)} = 1 - \left(\frac{n-1}{n-p}\right)(1-R^2) = 1 - \frac{\hat{\sigma}^2_{model}}{\hat{\sigma}^2_{null}}$$

Mallow's C_p

$$C_p = \frac{RSS_p}{\hat{\sigma}^2} + 2p - n$$

- RSS_p is the RSS from the model with p predictors
- $\hat{\sigma}^2$ is from the model with all predictors
- for the model with all predictors, $\hat{\sigma}^2 = RSS_p/(n-p)$ so $C_p = (n-p) + 2p - n = p$
- If a model fits poorly, then RSS_p will be large and $C_p > p$.
- choose models with small p and C_p around or less than p

Example data(state)

Chapter 9 Shrinkage Methods

9.1 Principal Component Regression

9.2 Partial Least Squares

9.3 Ridge Regression $\hat{\beta} = (X^TX + \lambda I)^{-1}X^Ty$

Chapter 10 Statistical Strategy and Model Uncertainty

Diagnostics, Transformation, Variable selection, Diagnostics, ...

Chapter 11 A Complete Example

Chapter 12 Missing Data: Imputation

Chapter 13 Analysis of Covariance

- deal with a mixture of quantitative (numeric) and qualitative (categorical) predictors
- use **dummy variables** or **contrast coding** for qualitative predictors
- for a factor of k groups (or levels), define k dummy variables, $d_1, \ldots, d_k$

$$d_j = \begin{cases} 1 & \text{if a case belongs to group } j \\ 0 & \text{otherwise} \end{cases}$$

- Note: $d_1 + \cdots + d_k = 1$
- so only need $k - 1$ dummy variables
- or use any $k - 1$ linear combinations of the k dummy variables
- The default choice in R is the **treatment coding**, which use the last $k-1$ dummy variables, so the first group is the reference (or base) group.
- The choice of coding does not affect the R^2, $\hat{\sigma}^2$ and overall F-statistic.
- but it does affect the $\hat{\beta}$.
- The interpretation of the estimation depends on the coding.
- interaction between a categorical and numeric variable is often of interest.

Example data(sexab)
$y = ptsd$ (Post-traumatic stress disorder on standard scale)
$x = cpa$ (Childhood physical abuse on standard scale)
$d = csa$ (Childhood sexual abuse - abused or not abused)

$$d = csa = \begin{cases} 1 & \text{Abused} \\ 0 & \text{NotAbused} \end{cases}$$

Consider 3 possible models

$$\begin{aligned} model\ \ 1: &\ \ y = \beta_0 + \beta_1 x + \epsilon \\ model\ \ 2: &\ \ y = \beta_0 + \beta_1 x + \beta_2 d + \epsilon \\ model\ \ 3: &\ \ y = \beta_0 + \beta_1 x + \beta_2 d + \beta_3 x \cdot d + \epsilon \end{aligned}$$

- model 1: a single regression line (csa has no effect on ptsd)
- model 2: two parallel regression lines with same slope
  (the effect of csa does not depend on cpa)
- model 3: two separate regression lines
  (the effect of csa depends on cpa)

- Textbook: J. J. Faraway (2006). "Extending the Linear Model with R," Chapman & Hall.

Chapter 2 Binomial Data

Linear model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \epsilon_i,$$

where the error ϵ_i are independent and follow $N(0, \sigma^2)$, so

- the Y_i are independent and follow $N(\mu_i, \sigma^2)$ where

$$\mu_i = E(Y_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$$

2.2 Binomial Regression Model

Consider data

$$(y_i, n_i - y_i, x_{i1}, \ldots, x_{ik}), \text{ for } i = 1, \ldots, n$$

where

- y_i and $n_i - y_i$ are the number of "successes" and "failures", respectively, out of n_i independent trials
- so the response has a **binomial distribution**, $Y_i \sim B(n_i, p_i)$

$$P(Y_i = y_i) = \binom{n_i}{y_i} p_i^{y_i} (1 - p_i)^{n_i - y_i}$$

$$E(Y_i) = n_i p_i, \quad \text{var}(Y_i) = n_i p_i (1 - p_i)$$

- assume that the Y_i are independent
- $x_{i1}, \ldots, x_{ik}$ are covariates

- model the relationship between the p_i and predictors $x_{i1}, \ldots, x_{ik}$ as
$$\eta_i = g(p_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$$
- $\eta = g(p)$ is called the **link function**
- since $0 \leq p_i \leq 1$, it is not appropriate to let $\eta_i = p_i$
- link function g is monotone and satisfies $0 \leq p = g^{-1}(\eta) \leq 1$

Common choices of link functions $\eta = g(p)$

- Logit: $\eta = \log(p/(1-p))$ (**logistic regression**)
- Probit: $\eta = \Phi^{-1}(p)$ where $\Phi(x)$ is the cdf of $N(0,1)$
- Complementary log-log: $\eta = \log(-\log(1-p))$

Inverse of link functions $p = g^{-1}(\eta)$

- Inverse of Logit: $p = \frac{\exp(\eta)}{1+\exp(\eta)}$
- Inverse of Probit: $p = \Phi(\eta) = P(Z \leq \eta)$ where $Z \sim N(0,1)$
- Inverse of Complementary log-log: $p = 1 - \exp(-\exp(\eta))$

Max Likelihood Estimator (MLE) $\hat{\beta}$ max the log-likelihood

$$l(\beta) = \sum_{i=1}^{n} \left[y_i \log(p_i) + (n_i - y_i) \log(1-p_i) + \log \binom{n_i}{y_i} \right]$$

For the logit link function, this is equivalent to

$$l(\beta) = \sum_{i=1}^{n} \left[y_i \eta_i - n_i \log(1 + \exp(\eta_i)) + \log \binom{n_i}{y_i} \right]$$

Example data(orings)

2.3 Inference

To compare two models ω and Ω with s and l parameters, $s < l$, the likelihood ratio test (LRT) statistic is

$$\Delta = 2\log\frac{L(\Omega)}{L(\omega)} = 2\log L(\Omega) - 2\log L(\omega)$$

- is approximately χ^2 distributed with $df = l-s$ under the smaller model, if the n_i are relatively large.
- For a χ^2 r.v. with $df = d$,

$$E(\chi^2_d) = d,\ \text{var}(\chi^2_d) = 2d$$

When the larger model is saturated, $\hat{p}_i = y_i/n_i$, and the LRT statistic becomes

$$D = 2\sum_{i=1}^{n}\left[y_i\log\frac{y_i}{\hat{y}_i} + (n_i - y_i)\log\frac{n_i - y_i}{n_i - \hat{y}_i}\right]$$

where $\hat{y}_i$ are the fitted values from the smaller model.

- The D is called **deviance** and measures how good the smaller model fits (compared to the saturated model).
- The deviance is a measure of goodness of fit.
- The deviance D is approximately χ^2 distributed with $df = n-s$, if the n_i are relatively large (and the model is correct).
- For a good approximation, it is often requested that $n_i \geq 5$.
- **Residual deviance** in R is the deviance for the current model.
- **Null deviance** is the deviance for a model with intercept only.
- approximate CI for β_i is: $\hat{\beta}_i \pm z_{\alpha/2}se(\hat{\beta}_i)$ or use `confint()`

Example data(orings)

2.5 Interpreting Odds

- Odds are used to represent chance in bets.
- Let p be the probability of success and o be the odds,

$$\frac{p}{1-p} = o \quad p = \frac{o}{1+o}$$

- A 3-1 **on** bet has $o = 3$ and $p = 3/4$.
- A 3-1 on bet would pay only \$1 for every \$3 bet.

For logistic regression,

$$\log(odds) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

- β_i has a simple interpretation in terms of log-odds
- a unit increases in x_1 with all other predictors held fixed increases the log-odds of success by β_1 (or increases the odds of success by a factor of $\exp(\beta_1)$).

Example data(babyfood)

2.7 Choice of Link Function

- when p is moderate (not close to 0 or 1), the link functions are similar.
- Larger differences are apparent in the tails; see R plots.
- but for very small p, one needs a very large amount of data to obtain just a few successes

Example data(bliss)

2.10 Prediction and Effective Doses

- To predict the outcome for given covariates, say x_0,

$$\hat{\eta} = x_0^T \hat{\beta}$$

$$\hat{\text{var}}(\hat{\eta}) = x_0^T \hat{\text{cov}}(\hat{\beta}) x_0 = x_0^T (X^T W X)^{-1} x_0$$

  where $\hat{\text{cov}}(\hat{\beta}) = (X^T W X)^{-1}$ can be extracted using the `cov.unscaled` component of the model summary.

- To get an answer in probability, transform back using $p = g^{-1}(\eta)$

Effective Doses: to determine the x value for a given p

- For a simple logistic model

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

  the effective dose x_p for prob. of success p is

$$x_p = \frac{\text{logit}(p) - \beta_0}{\beta_1}$$

- ED50 stands for effective dose for which there will be a 50% chance of success; $ED50 = -\beta_0/\beta_1$.

- To determine the standard error, use the delta method

$$\text{var}(h(\hat{\theta})) \approx h'(\hat{\theta})^T \text{var}(\hat{\theta}) h'(\hat{\theta})$$

$$\theta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \; h(\theta) = \frac{\text{logit}(p) - \beta_0}{\beta_1}, \; h'(\theta) = \begin{pmatrix} -1/\beta_1 \\ \beta_0/\beta_1^2 \end{pmatrix}$$

Example data(bliss)

Chapter 3.1 Poisson Regression

If Y has a Poisson distribution with mean $\mu > 0$,

$$P(Y = y) = \frac{\mu^y}{y!} e^{-\mu}, \quad y = 0, 1, 2, \ldots$$

- $E(Y) = \text{var}(Y) = \mu$
- Poisson distributions arise naturally when we count the number of (independent) events in a given time period.
- Binomial $B(n, p)$ has a fixed total number of counts, which is n
- If n is large and p is small, Poisson(μ) is a good approximation of $B(n, p)$ with $\mu = np$.

Poisson regression model (with link function $\eta = \log(\mu)$)

$$\eta_i = \log(\mu_i) = x_i^T \beta$$

The log-likelihood is

$$l(\beta) = \sum_{i=1}^{n} (y_i \log(\mu_i) - \mu_i - \log(y_i!)) = \sum_{i=1}^{n} (y_i x_i^T \beta - \exp(x_i^T \beta) - \log(y_i!))$$

Differentiating wrt β_j gives the MLE $\hat{\beta}$ as the solution to

$$\sum_{i=1}^{n} (y_i - \exp(x_i^T \hat{\beta})) x_{ij} = 0 \text{ for any } j$$

which can be written as

$$X^T y = X^T \hat{\mu}, \quad \hat{\mu} = \exp(X \hat{\beta})$$

Deviance (also known as the G-statistic)

$$D = 2 \sum_{i=1}^{n} (y_i \log(y_i / \hat{\mu}_i) - (y_i - \hat{\mu}_i))$$

Example data(gala)

Overdispersion

- Poisson (or Binomial) distributions have only one parameter μ (or p) so it is not very flexible for empirical fitting purposes.
- Overdispersion or underdispersion can occur in Poisson (or Binomial) models; see text for possible reasons.
- When overdispersion happens, the standard errors of the estimates are inflated.
- One approach is to introduce a **dispersion parameter** ϕ, which can be estimated as

$$\hat{\phi} = \frac{X^2}{n-p}$$

where X^2 is the Pearson's X^2 statistic for goodness of fit.

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$$

- For overdispersion: $\phi > 1$; for underdispersion: $\phi < 1$.
- When overdispersion is considered, an F-test rather than a χ^2 test should be used.

Example data(gala)

Chapter 6 Generalized Linear Models

6.1 GLM Definition

The distribution of response Y is from the **exponential family**:

$$f(y|\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right]$$

- θ is called the **canonical parameter** and represents the location
- ϕ is called the **dispersion parameter** and represents the scale.

Examples of exponential family

- Normal or Gaussian:

$$f(y|\theta,\phi) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right]$$

$$= \exp\left[\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right)\right]$$

  so we can write $\theta = \mu, \phi = \sigma^2$,

$$a(\phi) = \phi, b(\theta) = \theta^2/2, c(y,\phi) = -(y^2/\phi + \log(2\pi\phi))/2$$

- Poisson:

$$f(y|\theta,\phi) = \frac{\mu^y}{y!}e^{-\mu} = \exp[y\log(\mu) - \mu - \log(y!)]$$

  so we can write $\theta = \log(\mu), \phi \equiv 1$,

$$a(\phi) = 1, b(\theta) = \exp(\theta), c(y,\phi) = -\log(y!)$$

- Binomial $B(m, p)$:

$$\begin{aligned} f(y|\theta,\phi) &= \binom{m}{y} p^y (1-p)^{m-y} \\ &= \exp\left[y\log(p) + (m-y)\log(1-p) + \log\binom{m}{y} \right] \\ &= \exp\left[y\log\frac{p}{1-p} + m\log(1-p) + \log\binom{m}{y} \right] \end{aligned}$$

so we can write $\theta = \log\frac{p}{1-p}$, $\phi \equiv 1$, $a(\phi) = 1$,

$$b(\theta) = -m\log(1-p) = m\log(1+e^{\theta})), c(y,\phi) = \log\binom{m}{y}$$

mean and variance of exponential family distributions

$$E(Y) = \mu = b'(\theta), \quad \text{var}(Y) = b''(\theta)a(\phi)$$

- mean is a function of θ only;
- variance is a product of functions of the location and the scale.
- $V(\mu) = b''(\theta)$ is called the **variance function** and describes how the variance relates to the mean.
- For Gaussian case, $b''(\theta) = 1$ and so the variance is independent of the mean.

Linear predictor

$$\eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k = x^T\beta$$

Link function

$$\eta = g(\mu)$$

- g links mean response, $\mu = E(Y)$, to covariates through η

- g should be monotone, continuous and differentiable
- **canonical link** function satisfies

$$\eta = g(\mu) = \theta \longrightarrow g(b'(\theta)) = \theta$$

Family	Canonical Link	Variance function $b''(\theta)$
Normal	$\eta = \mu$	1
Poisson	$\eta = \log \mu$	μ
Binomial	$\eta = \log \frac{p}{1-p}$	$mp(1-p)$

- If a canonical link is used, $X^T Y$ is **sufficient** for β.

6.2 Fitting a GLM

One-step Taylor expansion

$$g(y) \approx g(\mu) + (y - \mu)g'(\mu) \equiv z$$

so

$$\text{var}(z) = (g'(\mu))^2 \text{var}(y) = (g'(\mu))^2 V(\mu) a(\phi)$$

The following IRWLS procedure is used to fit a GLM:

1. Set initial estimates $\hat{\mu}_0$ and $\hat{\eta}_0 = g(\hat{\mu}_0)$.
2. Form the "adjusted dependent variable" $z_0 = \hat{\eta}_0 + (y - \hat{\mu}_0)g'(\hat{\mu}_0)$.
3. Form the weights $w_0 = [(g'(\hat{\mu}_0))^2 V(\hat{\mu}_0)]^{-1}$.
4. Re-estimate β using WLS with response z_0 and weights w_0.
5. Update $\hat{\eta}_1 = x^T \hat{\beta}$ and $\hat{\mu}_1 = g^{-1}(\hat{\eta}_1)$.
6. Iterate steps 2–5 until convergence.

Estimates of variance are obtained from:

$$\widehat{\text{var}}(\hat{\beta}) = (X^T W X)^{-1}\hat{\phi}$$

Example data(bliss)

For a binomial response, consider p (rather than $\mu = mp$),

$$\eta = g(p) = \log\frac{p}{1-p}, g'(p) = \frac{1}{p(1-p)}, V(p) = \frac{p(1-p)}{m}, w = mp(1-p)$$

6.3 Hypothesis Tests

- For a saturated (or full) model $\hat{\mu} = y$.
- Likelihood ratio test (LRT) statistic for comparing the current model with a saturated model is

$$2(l(y, \phi|y) - l(\hat{\mu}, \phi|y))$$

- when $a(\phi) = \phi/w_i$, this simplifies to

$$D(y, \hat{\mu})/\phi = \sum_i 2w_i(y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i))/\phi$$

  where $\tilde{\theta}$ and $\hat{\theta}$ are the estimates under the full and current model, respectively.

- $D(y, \hat{\mu})$ is called the **deviance**.
- $D(y, \hat{\mu})/\phi$ is called the **scaled deviance**.

GLM	Deviance $D(y, \hat{\mu})$
Normal	$\sum_i (y_i - \hat{\mu}_i)^2$
Poisson	$2\sum_i (y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i))$
Binomial	$2\sum_i \left[y_i \log\frac{y_i}{\hat{\mu}_i} + (n_i - y_i)\log\frac{n_i - y_i}{n_i - \hat{\mu}_i}\right]$

- An alternative to deviance is Pearson's X^2 statistic:

$$X^2 = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

  where $V(\hat{\mu}) = \text{var}(\hat{\mu})$.

- The scaled deviance $D(y, \hat{\mu})/\phi$ and Pearson's X^2 statistic are asymptotically χ^2 distributed.
- If we know ϕ, we can test the goodness of fit with a χ^2 test.
- For nested models ω and Ω, $D_\omega - D_\Omega$ is asymptotically χ^2.
- If we know ϕ, use χ^2 test.
- If we don't know ϕ, use an F-statistic (which is approximately F distributed)

$$\frac{(D_\omega - D_\Omega)/(df_\omega - df_\Omega)}{\hat{\phi}}$$

  where $\hat{\phi} = X^2/(n-p)$ is a good estimate of the dispersion.

- For the Gaussian model, $\hat{\phi} = RSS_\Omega/df_\Omega$, the F test is exact.

Example data(bliss)

6.4 GLM Diagnostics

Residuals

- The **response residual** $\hat{\epsilon} = y - \hat{\mu}$, which do not have constant variance.
- **Pearson residual**
$$r_P = \frac{y - \hat{\mu}}{\sqrt{V(\hat{\mu})}}$$
where $V(\mu) \equiv b''(\theta)$. Note $\sum r_P^2 = X^2$.
- Let deviance $D = \sum d_i = \sum r_D^2$. The **deviance residual** is
$$r_D = sign(y - \hat{\mu})\sqrt{d_i}$$
- The **working residual** is a by-product of the IRWLS fitting procedure.

Leverage and influence:

- The **hat matrix** is
$$H = W^{1/2}X(X^TWX)^{-1}X^TW^{1/2}$$
where $W = \text{diag}(w)$ and w are the weights used in fitting.

Standardized residuals :
$$r_{SD} = \frac{r_D}{\sqrt{\hat{\phi}(1 - h_{ii})}}$$

Cook's distance is
$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})^T(X^TWX)(\hat{\beta}_{(i)} - \hat{\beta})}{p \cdot \hat{\phi}}$$

Example data(bliss)

Model diagnostics

- Constant variance is not assumed, so be careful on the choice of residual plots.
- The response residuals do not have constant variance.
- The deviance residuals are expected to have constant variance.
- Plot the deviance residuals against the fitted linear predictor $\hat{\eta}$.
- however, residual plots are not helpful in some cases; see p.127 for reasons.
- Normality is not assumed except for the Gaussian.
- Use half-normal plot to check **unusual points** (such as outliers or influential cases)

Difficulty: How to judge the **curvilinear** relationship between the response and the predictors?

- Use the **linearized response**

$$z = \eta + (y - \mu)g'(\mu)$$

- plot z versus linear predictor $\hat{\eta}$ to check link function $\eta = g(\mu)$
- partial residual plot: $z - \eta + \hat{\beta}_j x_j$ versus x_j

Example data(gala)