University of California, Los Angeles
Department of Statistics

Statistics 100C **Instructor: Nicolas Christou**

Homework 5

Exercise 1
Please refer to homework 4, exercise 3.

a. Test the overall significance of the model. The easiest way to do this is to find first SSE and SST. Then you can compute SSR and then the F statistic.

b. Test the following hypothesis:
$H_0 : \beta_1 - 2\beta_2 = 0$
$H_0 : \beta_1 - 2\beta_2 \neq 0$
The test statistic will be:

$$t = \frac{\mathbf{a}'\hat{\boldsymbol{\beta}} - 0}{s_e\sqrt{\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}}}.$$

Before you compute test statistic above write the vector $\mathbf{a}$, that will help you to extract the elements needed from $(\mathbf{X}'\mathbf{X})^{-1}$ to find the $\text{var}(\hat{\beta}_1 - 2\hat{\beta}_2)$.

c. Find a confidence interval for $E(y_g)$ when $\mathbf{x}'_{\mathbf{g}} = (1 \quad 24 \quad 29)$. Use:

$$\hat{y}_g \pm t_{\frac{\alpha}{2};n-k-1}s_e\sqrt{\mathbf{x}'_{\mathbf{g}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_{\mathbf{g}}}.$$

d. Compute R^2 for these data.

Exercise 2
Show that the error sum of squares $SSE = \mathbf{e}'\mathbf{e}$, is equal to the following expressions:

$$SSE = \mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} = \mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\hat{\mathbf{Y}}.$$

Exercise 3
Suppose for a multiple regression problem the units of the the i_{th} independent variable are in millimeters. Explain what would happen to the estimate $\hat{\beta}_i$ of β_i and to its variance if we express the i_{th} independent variable in meters instead of millimeters. *Hint:* Multiply $\mathbf{X}$ by a diagonal matrix containing 0.001 in the $(i,i)_{th}$ position and 1's in the other diagonal positions.

Exercise 4
Most people like to eat out at restaurants that offer supposedly top quality food. But do we pay for the quality of food or for something else... One hundred restaurants were selected from the area of Westwood, Brentwood, and Santa Monica (these are 2000 data). The source of this data set is from http://www.zagat.com and can be accessed in `R` as follows:

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

In this data set there are four variables:

food: The food rating for each restaurant on a scale from 1-30 (30 being the best).

decor: The decor rating for each restaurant on a scale from 1-30.

ser: The service rating for each restaurant on a scale from 1-30.

cost: The cost ($) for dinner including one drink and the tip for each restaurant.

Answer the following questions:

a. Construct the scatterplots of the variable *cost* on each of the other three variables.

b. Run the following 3 regressions:

   i. *cost* on $food$

   ii. *cost* on *decor*

   iii. *cost* on *ser*

c. From your answer to question (b), with which of the three independent variables is *cost* most correlated?

d. Use the fitted line of the best of the three regressions to predict the cost of a dinner at a restaurant which has food rating 20, decor rating 16, and service rating 13.

e. Now, add the three ratings ($food$, *deco*, and *service*) to create a total rating variable. So, now you have a new variable called *total*. Plot *cost* against *total*.

f. Regress *cost* on *total*. Is there a stronger relationship (R^2) between *cost* and *total*, than any of the previous regressions from question (b)? Write down the fitted regression line.

g. Check the assumptions of the model of part (f). Plot and print the residuals against the fitted values and against *total*. Are there any violations of the assumptions?

h. Using the model of part (f) predict the cost of a dinner at a restaurant of your choice (perhaps your favorite restaurant if you have one). You will need of course the values of $food$, *ser*, and *decor*.

This regression model was first applied in data from restaurants in New York City by Professor Jeff Simonoff of the Statistics Department of the Stern School of Business at New York University.

Exercise 5
Use the data of exercise 4. Consider the multiple regression model

$$cost_i = \beta_o + \beta_1 food_i + \beta_2 decor_i + \beta_3 ser_i + \epsilon_i$$

a. Construct the $n \times 4$ design matrix $\mathbf{X}$, and compute the $\mathbf{X}'\mathbf{X}$.

b. Obtain the least squares estimates $\hat{\boldsymbol{\beta}}$ using matrix and vector operations.

c. Verify that your answers are correct by using the `lm` function in `R`.

d. Compute the hat matrix $\mathbf{H}$ and show only the elements of the first 5 rows and columns.