STAT 540, Fall 2015

Due on **12:00 A.M., Friday,Nov 6th**

Homework 9

Homework format for all STAT 540 homework this term: Please label all problems clearly and turn in an organized homework assignment. You don't need to spend hours producing beautifully typeset homework, but you won't get credit if we can't find or read your answer. Unless noted otherwise, turn in the following (as appropriate for the problem).

- Theoretical derivation (when asked for).
- Numerical results **with an explanation of your solution**, written in complete sentences. If computer code is absolutely necessary to provide context here, then include it–nicely formatted–within the solution (otherwise, see below).
- Appropriate graphics. Use informative labels, including titles and axis labels. Try to put multiple plots on the page by using, for example, the R command `par(mfrow=c(2,2))`.
- **Only as necessary**: Final clean computer code used to answer the problem **attached to the end of your homework**. Only include the rare code excerpts without which we wouldn't be able to figure out what you did. Annotate your code. Number and order the code in order of the problems. When in doubt, leave it out; consider that we will probably never read it.
- Some problems will be relatively open-ended, such as "Here are some data. Analyze them and write a report." I will provide further instructions about reports later. They should be self-contained, with suitable EDA, graphs, numerical results, and **scientific interpretation**. No computer code should be included. The report should be concise: "no longer than necessary".

(1) Case study 5.1.1. from the book *The Statistical Sleuth* describes a dietary restriction study. Female mice were assigned to one of the following six treatment groups:

- NP: unlimited, nonpurified, standard feed
- N/N85: normal diet before weaning and normal diet (85 kcal/week) after weaning
- N/R50: normal diet before weaning and reduced calorie (50 kcal/week) after weaning
- R/R50: reduced calorie diet before and after weaning (50 kcal/week)
- N/R50 lopro: normal diet before weaning, reduced calorie (50 kcal/week) after weaning, and reduced protein

- N/R40: normal diet before weaning and severely reduced calorie (40 kcal/week) after weaning

The response of interest was mouse lifetime in months. Download the corresponding data file at `http://www.science.oregonstate.edu/ schafer/Sleuth/data-sets.html` or access it by installing and loading the `R` package `Sleuth3` and examining `case0501`. Complete the following parts under the assumption that a Gauss-Markov model with normal errors and a separate mean for each of the 6 treatment groups is appropriate for these data.

(a) Create side-by-side boxplots of the response for this dataset, with one boxplot for each treatment group. Be sure to clearly label the axes of your plot.
(b) Find the SSE (sum of squared errors) for the full model with 6 means.
(c) Compute $\hat{\sigma}^2$ for the full model with 6 means.
(d) Find the SSE for a reduced model that has 1 common mean for the N/R50 and R/R50 treatment groups and a distinct mean for each of the other 4 treatment groups.
(e) Use the answer from parts (b) through (d) to compute an F statistic for testing the null hypothesis that mean of the response vector is in the column space associated with the reduced model versus the alternative that the mean of the response vector is in the column space of the full model but not in the column space of the reduced model.
(f) Explain to the scientists conducting this study what the F statistic in part (e) can be used to test. Consider the context of the study and use terms non-statistician scientists will understand.

(2) Using the data from KNNL problem 6.15 for patient satisfication, do the following:
(a) Fit the model with X_1, X_2, X_3 (full model). Report the summary table for the model. Fit a model with X_2 only. Report the summary table for the model. Compare and explain the results.
(b) Examine the scatter plot matrix (pairs). Examine the sample correlation matrix (cor). What do these plots suggest about part (a)?
(c) Compute the VIF values for the predictors in the full model. Does this change your answer in part (a)?
(d) Does SSR(X_1) equal SSR(X_3)? Does SSR(X_2) equal to SSR($X_2|X_3$)? Interpret the results.

(3) The data in `avp.txt` were constructed to make some points about assessing the form of $f(X)$ in a multiple regression. The data set has 3 variables, X_1, X_2, and X_3 that are to be used to predict the response Y. All my questions based on the model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i.$$

The investigators who collected these data are especially concerned about lack of this multiple regression.

(a) Fit the regression and plot the residuals against each of the X variables. Is there any indication of lack of fit? Explain why or why not?

(b) Fit a full quadratic (i.e. including $X_1^2, X_2^2, X_3^2, X_1X_2, X_1X_3, X_2X_3$ in the model). Use the results from this model and the original regression to test for lack of fit. Is there any evidence of lack of fit?

(c) Fit a Lowess non-parametric regression using a smoothing parameter of 0.6. Use the results from this model and the original regression to test for lack of fit. Is there any evidence of lack of fit? (If you need any help with Lowess, please refer to the package `lowess` in `R` and discuss it with me).

(d) Examine added variable plots (a.k.a. partial regression residual plots) for each variable? Is there any indication of lack of fit? Explain your answer.

(4) Website development problem. The data in website.txt are from an observational study of productivity of developers at a website development company. The eventual goal of the analysis is to "determine which variables have the greatest impact on the number of websites delivered". Consider the following model

$$\text{Deliver}_i = \beta_0 + \beta_1 \text{BACKLOG}_i + \beta_2 \text{EXPERIENCE}_i + \beta_3 \text{PROCESS}_i + \beta_4 \text{YEAR}_i + \epsilon.$$

Fit the regression model in `R`, then consider the following question. Please consider to be a medium size sample.

(a) Consider each point. Any concerns about regression outliers? Is so, list the points (by id#) that are a concern and briefly explain why.

(b) Do any points raise concerns about usually large influence on the fitted values? If so, list the id #'s that are a concern and briefly explain why.

(c) Since the objective here is to examine regression coefficients, do any points have usually large influence on the estimated regression coefficients? Again, if so, list the points (by id#) that are a concern and briefly explain why.

(d) Are there any concerns with multicollinearity for any variables? Explain why or why not?

(e) Plot residuals versus predicted values. Any issues that concern you? Explain why or why not.

(f) The CEO of this company reminds you that she hired you to "determine which variables have the greatest impact on the number of websites delivered". Use standardized regression coefficients to answer her question.

(5) Study on the deleted studentized residuals t_i's. You may need to use the Sherman-Morris theorem (no need to show it, you can use it directly) that for matrices $\mathbf{A}, \mathbf{U}, \mathbf{B}$ and vectors $\mathbf{u}, \mathbf{v}$

$$(\mathbf{A} + \mathbf{u}\mathbf{v}')^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}'\mathbf{A}^{-1}}{1 + \mathbf{v}'\mathbf{A}^{-1}\mathbf{u}}$$
$$(\mathbf{A} + \mathbf{U}\mathbf{B}\mathbf{U}')^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}\mathbf{B}(\mathbf{I} + \mathbf{U}'\mathbf{A}^{-1}\mathbf{U}\mathbf{B})^{-1}\mathbf{U}'\mathbf{A}^{-1}$$

(a) Show that $\mathbf{X}'\mathbf{X} = \mathbf{X}'_{(-i)}\mathbf{X}_{(-i)} + \mathbf{x}_i'\mathbf{x}_i$ and $\mathbf{X}'\mathbf{y} = \mathbf{X}'_{(-i)}\mathbf{y}_{(-i)} + \mathbf{x}_i'y_i$.

(b) Show that

$$\widehat{\boldsymbol{\beta}}_{(-i)} = \widehat{\boldsymbol{\beta}} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i' y_i + \frac{1}{1-h_{ii}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i'\mathbf{x}_i[\widehat{\boldsymbol{\beta}} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i' y_i]$$

so that

$$\widehat{\boldsymbol{\beta}}_{(-i)} - \widehat{\boldsymbol{\beta}} = \frac{\widehat{y}_i - y_i}{1-h_{ii}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i'$$

(c) Based on (a) and (b), show that the deleted residuals satisfy

$$d_i = \frac{e_i}{1-h_{ii}}.$$

(d) Show that $1 + \mathbf{x}_i \left(\mathbf{X}_{(-i)}'\mathbf{X}_{(-i)}\right)^{-1} \mathbf{x}_i' = (1-h_{ii})^{-1}$ so that by lecture note $s^2(d_i) = \mathrm{MSE}_{(-i)}(1-h_{ii})^{-1}$.

(e) Recall that $\mathrm{SSE} = \mathbf{y}'(\mathbf{I}-\mathbf{H})\mathbf{y} = \mathbf{y}'\mathbf{y} - \mathbf{y}'\widehat{\mathbf{y}}$ and $\mathbf{y}'\mathbf{y} = \mathbf{y}_{(-i)}'\mathbf{y}_{(-i)} + y_i^2$, show that

$$(n-p)\mathrm{MSE} = (n-p-1)\mathrm{MSE}_{(-i)} + \frac{e_i^2}{1-h_{ii}}$$

(Hint: what is $\mathrm{SSE}_{(-i)}$?)

(f) Show the deleted studentized residual t_i's expression on page 11 in the lecture 10 based on above parts.

(6) Study on the diagnostic tools.

(a) Show that $h_{ii} \in [0,1]$ where $\mathbf{H} = (h_{ij})_{1\le i,j\le n}$ denotes the projection matrix.

(b) Show that DFFITS_i equals to $t_i\sqrt{h_{ii}(1-h_{ii})^{-1}}$.

(c) Show that the Cook's distance statistic satisfies

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

where r_i is the studentized residual.

(d) (Optional) Show that the VIF for the kth variable VIF_k equals to $(n-1)((\mathbf{X}_c'\mathbf{X})^{-1})_{kk}$ where $\mathbf{X}_c$ is the centered design matrix

(7) Textbook problems:

(a) Problem 7.22; 7.29