STAT 540, Fall 2015

Due on **12:00 A.M., Friday, Nov 13th**

Homework 10

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) Show the LS estimates for β_1 or β_2 are not affected by X_2 or X_1 if the predictors are linearly independent for model $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$. That is, show the results of $\hat{\beta}_1$ and $\hat{\beta}_2$ on page 25 in lecture 10.

(2) Global warming is a contentious environmental issue. The data in `temperature.txt` on the web site are measurements of the worldwide average annual temperature for 108 years from 1880 to 1987. The response variable is expressed as a temperature anomaly. This is the deviation of that year's temperature from the average over all 108 years. The trend in the temperature anomaly is the same as the trend in the temperature. What do these

data tell us about the temperature trend? Assessment of the uncertainty in the trend is as important as assessment of the trend.

(a) Fit a usual linear regression and estimate the slope (temperature change per year). Estimate the standard error of the slope assuming that errors are independent.

(b) Is the assumption of a linear trend reasonable? Explain why or why not. You are free to choose your favorite lack of fit method, but I strongly suggest you consider more than just a residual plot. If want to use a polynomial, please consider whether that polynomial is appropriate for the apparent trend. All subsequent parts will use the linear model temp $= \beta_0 + \beta_1$year $+ \epsilon$ even though that isn't really appropriate.

(c) Estimate the lag-1 correlation in the residuals from the linear model, i.e. the correlation between e_t and e_{t-1}.

(d) The Durbin-Watson statistic is a test statistic in both Statistics and Econometrics used to detect the presence of autocorrelation in the residuals from a regression analysis. It is defined as $\sum_{t=2}^{T}(e_t - e_{t-1})^2 / \sum_{t=1}^{T} e_t^2$ whose distribution has been casted in numerical tables. For your convenience, it has been included in the `R` package `lmtest` as a function `dwtest` and you can load it directly. Test for a significant lag-1 correlation using the Durbin-Watson test.

(3) Simulation studies.

(a) Take the matrix $\mathbf{X}$ given in `hw.dat`, and set $\boldsymbol{\beta} = (1, 1)$, $\sigma^2 = 2$. Run a simulation experiment in which you are going to generate a draw for the vector $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ and use $\mathbf{y}, \mathbf{X}$ to calculate $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}^2$, repeating this for 1000 times. That is you will have 1000 simulated $\widehat{\boldsymbol{\beta}}$ vectors and 1000 simulated $\widehat{\sigma}^2$ values. Calculate the sample mean vector and sample joint variance matrix of $(\widehat{\boldsymbol{\beta}}', \widehat{\sigma}^2)$, and compare to their theoretical values.

(b) According to the theory we learned in lecture, $\widehat{\sigma}^2$ should be independent of $\widehat{\boldsymbol{\beta}}$. Is the variance matrix from your simulation experiment consistent with this?

(c) Now, instead generating y_i as

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \nu_i - 1$$

where ν_i are independent χ_1^2 random variables. In this setting, do $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}^2$ appear to be unbiased? Explain what should be the theoretical mean and variance of them.

(d) For $i = 1, \ldots, 100$, create your own $\mathbf{X}$ in the following way: let $x_{i1} = 1$ and x_{i2} be i.i.d $N(0, 1)$. Let

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \nu_i - 1$$

where ν_i are independent χ_1^2 random variables. Simulate 1000 draws for $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}^2$, and create histograms of their marginal distributions. Do these distributions appear to be approximately normal? For $\widehat{\beta}_1$ and $\widehat{\beta}_2$, theoretical results are known that

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to N(\mathbf{0}, \sigma^2\{\mathbb{E}(\mathbf{x}_i\mathbf{x}_i')\}^{-1})$$

in distribution so that approximate distribution of $\widehat{\boldsymbol{\beta}}$ is

$$\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2/n\{\mathbb{E}(\mathbf{x}_i\mathbf{x}_i')\}^{-1}).$$

Based on the model setting, calculate the theoretical means and variances for $\boldsymbol{\beta}$ (under the asymptotic approximation) and compare these to the simulated means and variances.

(e) Assume that $y_1 = 0$, and for $t = 2, \ldots, T$, that

$$y_t = \alpha y_{t-1} + \epsilon_t$$

where ϵ_t are i.i.d $N(0,1)$. Let $\widehat{\alpha}$ be the OLS coefficient in a regression of y_t on y_{t-1} (without the intercept).

Simulate the distribution of $\widehat{\alpha}$ for $T = 20$, for different possible values of α: $\alpha = 0.5, 0, 9$ and 1. For each value of α, generate 1000 draws and plot the histograms of the draws for $\widehat{\alpha}$, and calculate the bias. What do you observe?

(4) Analysis on the NFL football data. The course web site provides data from the National Football League (`nfl-05.csv`). The data set consists of statistics for each team on various offensive measures of performance (offense is when the listed team is in control of the ball). For example, the variable PASS is the total number of yards gained by throwing the ball and the variable RUSH is the total number of yards gained by running the ball up the field. These statistics are the totals over all games in the 2005 season. The variable YDS is equal to the sum of PASS and RUSH. YPG means average "yards per game". The variable PTS is the total number of points earned by the team during the season, and PTS/G is the average points per game. Wins is the number of games won. It is expected the chances of a team winning the game increase as the team gains more yards during the course of a game. There are two leagues ("conferences") within which teams play: the NFC and the AFC. The two conferences play other teams within the same conference during the regular season. The goal of this assignment is to investigate the relationship between the number of wins (the response) and several predictor variables.

(a) Obtain the estimated regression function when Wins is regressed on RUSH and PASS.

(b) Test whether PASS can be dropped from the model given that RUSH is retained.

(c) Add the variable YDS to the model given in part (a) and obtain the estimated regression function. Discuss your results briefly.

(d) Let C be the indicator of conference being AFC, and obtain the regression model for Wins against YDS, the conference type, and their interactions.

(e) Using the model from part (d), predict the number of wins for an AFC team with YDS$= 5000$. Be sure to include an appropriate measure of uncertainty like prediction intervals.

(f) For the model in part (d), test whether there is a difference between AFC and NFC.

(g) Add PTS/G (points per game) into the model from part (d). Include its pairwise and three-way interactions. Provide point estimates of wins for all combinations of the following variables: conference being AFC or NFC, PTS/G= 18 or 22, and YDS= 4500 or 5500.

(h) For the model in part (g), test whether there is a difference between AFC and NFC.

(5) Model selection (practice): The `modelsel1.txt` data set contains four predictor variables and $n = 50$ observations. Using the `R` package `leaps`, you can easily get Mallows' C_p, AIC, BIC, etc.

(a) Find the best model using the stepwise selection, using $\alpha_{\text{entry}} = \alpha_{\text{remove}} = 0.2$.

(b) Do you end up with the same model using using $\alpha_{\text{entry}} = \alpha_{\text{remove}} = 0.1$?

(c) Use `R` to calculate Mallow's C_p for all possible models. Identify the variables in the 3 models with the three smallest C_p values.

(d) If you use AIC or BIC, do you select the same best model? The same best three models?

(6) Textbook Problems:

(a) # 8.13: Portray response curves.

(b) # 8.14: Gender coding scheme.

(c) # 8.15: Copier maintenance.

(d) # 8.19: Copier maintenance continued.

(e) # 8.34: Interpret indicator variables.