STAT 540, Fall 2015

Due on **Friday, October 2nd**

Homework 5

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) Consider the problem (7 (b)) in HW 1, i.e. the problem 1.27 in the textbook. Using data from that problem and based on your answer in HW 1, provide solution to the following questions.
   (a) Examine the normal probability plot of the residuals and state your conclusion.
   (b) Examine the plot of the residuals against the ages of the women in the study and state your conclusion.

(2) Textbook problems and relevant problems:
   (a) Problem 4.8.
   (b) Problem 4.22 (first consider Bonferroni inequality for $g = 3$, then consider general g)

(c) For n events, $A_1, \dots, A_n$, show the inclusion-exclusion formula that is

$$\begin{aligned}\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) &= \sum_{i=1}^{n} \mathbb{P}(A_i) - \sum_{1\le i<j\le n} \mathbb{P}(A_iA_j) + \dots \\ &+ (-1)^{k-1} \sum_{1\le i_1<\dots<i_k\le n} \mathbb{P}(A_{i_1}\cdots A_{i_k}) \\ &+ \dots + (-1)^{n-1}\mathbb{P}(A_1\cdots A_n)\end{aligned}$$

(d) Using the inclusion-exclusion formula, show that for n events $\{A_i\}$

$$\sum_{i=1}^{n} \mathbb{P}(A_i) - \sum_{1\le i<j\le n} \mathbb{P}(A_iA_j) \le \mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \le \sum_{i=1}^{n} \mathbb{P}(A_i)$$

(3) Observations on the yield of a chemical reaction taken at various temperatures were recorded as follows, where x_i and y_i are temperature and yield, respectively.

(x_i)	(y_i)
150	77.4
150	76.7
150	78.2
200	84.1
200	84.5
200	83.7
250	88.9
250	89.2
250	89.7
300	92.8
300	92.7
300	93.9

(a) Compute least square estimates of the intercept β_0 and slope β_1 of a simple linear regression of yield y on temperature x. Report the estimates and their standard errors.

(b) Construct a normal plot of the residuals from fitting the model in part (a). (Do not submit this plot.) What does this plot reveal?

(c) Construct a plot of y_{ij} versus x_i and a plot of the residuals $e_{ij} = y_{ij} - \hat{y}_{ij}$ versus x_i. What does this plot reveal?

(d) Partition the residual sum of squares into two parts, $\text{SS}_{\text{pure error}}$ and SS_{lof}.

(e) Use result from above to test the null hypothesis

$$H_0 : \mathbb{E}(y_i|x_i) = \beta_0 + \beta_1 x_i \text{ for } x_i = 150, 200, 250, 300$$

against the general alternative. Report the value of an F-statistic, its degrees of freedom, and the associated p-value. State your conclusion.

(f) Evaluate least squares estimates for the parameters in the quadratic polynomial model

$$y_i = \alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 + \epsilon_i$$

and report estimates, standard errors, t-statistics, and p-values for all three parameters.

(g) Partition the residual sum of squares for the model in part (f) into two parts $\text{SS}_{\text{pure error}}$ and SS_{lof}; and report the values of the F-statistic for the lack of fit test, its degrees of freedom, and the resulting p-value. State your conclusion.

(h) Give an interpretation of the values of α_0, α_1 and α_2 with respect to the effect of temperature on yield.

(4) Suppose, as in Problem (3) above, a few observations of the yield y of a chemical process were taken at each of four temperatures x, and you are only given information on the sample means and standard deviations for the observed yields at each temperature. The summary data are

Temperature	150	200	250	300
sample mean	75	85	89	91
sample variance	1.15	1.00	1.25	0.60
sample size	2	5	5	3

(a) Use this information to complete the least squares estimates of β_0 and β_1 for the simple linear regression model. (These are not the same data used in Problem (3)). Report the least squares estimates of the coefficients and their standard errors.

(b) Complete the following ANOVA table:

Source of variation	df	Sum of Squares	Mean Square
Regression on x			
Residuals			
Lack-of-fit			
Pure error			
Corrected total			

(c) State your conclusion for the lack of fit test.

(5) For each of the following models, indicate if it is a linear model, a nonlinear model, or an intrinsically linear model (a nonlinear model that can be transformed into a linear model).. For an intrinsically linear model, identify the transformation that produces a linear model. In each case, ϵ_i denotes a random error with variance σ^2. In parts (a), (c), (d) and (e), $\mathbb{E}(\epsilon_i) = 0$ and $\mathbb{E}(\epsilon_i) = 1$ in parts (b) and (f).

(a) $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 \log(x_{i2}) + \beta_3 \sin(x_{3i}) + \epsilon_i$

(b) $y_i = \epsilon_i \exp(\beta_0 + \beta_1 x_i)$

(c) $y_i = \beta_0 + \log(\beta_1 x_{1i}) + \beta_2 x_{2i} + \epsilon_i$

(d) $y_i = \beta_0 \exp(\beta_1 x_{1i}) + \epsilon_i$

(e) $y_i = (1 + \exp(\beta_0 + \beta_1 x_{1i} + \epsilon_i))^{-2}$

(f) $y_i = (\beta_0 + \beta_1 x_i)\epsilon_i$