STAT511 HW#11

Reading: CH11 in O&L
See Canvas Calendar for Due Date.

44 points total, **4** points per problem part unless otherwise noted.

1. How does length of service (LOS) related to wages? For this question, we will work with data on the LOS (in months) and wages of 60 women who work in Indiana banks. The banks were classified as Large or Small. Wages are yearly total income divided by the number of weeks worked. The wages have been multiplied by a constant for reasons of confidentiality. The data is available as BankSalaries.csv from Canvas.

   The data were originally reported in a publication "Is small beautiful? Work-family tension, work conditions and organization size" by MacDermid (Family Relations, 44 (1994), p159-167), but were taken from Problem 10.1 in Introduction to the Practice of Statistics (3rd Edition) by Moore and McCabe.

   NOTE: In STAT512, we will use ANCOVA to analyze data where we have both a continuous and a categorical predictor (in this case LOS and size). Here we will simply fit a separate model for Large and Small banks.

   Since we will be running the analysis separately for Large and Small banks, it helps to split the data into separate objects. To do this you will use code like:

```
Large <- subset(Salaries, Size=="Large")
Small <- subset(Salaries, Size=="Small")
```

   A. Create a scatterplot of Wages vs LOS with observations identified as coming from Large or Small Banks. Overlay regression lines for Large and Small Banks. Include this plot in your assignment. (**2 pts**)
   To do this you will use code like:

```
library(lattice)
xyplot(Wages ~ LOS, data=Salaries, groups=Size, type=c("p","r"),
auto.key=list(space="right"))
```

   B. Regress Wages (Y) against LOS (X) separately for Large and Small banks and report the regression equation for each. **(2 pts)**

   C. Give the estimate and corresponding 95% confidence interval for the intercepts for both Large and Small banks. Based on salary, would a brand new employee (LOS=0) be better off at a Large or a Small bank? Justify your response based on model estimates.

   D. Give the estimate and interpretation (in the context of this problem) of the slopes for both Large and Small banks.

   E. Using $\alpha=0.05$, test the null hypothesis that the slope of the true regression line is zero versus an alternative that it is not zero for both Large and Small banks. Give the p-value and conclusion. Do we have evidence that LOS is (linearly) related to wages?

F. Give the estimate and corresponding 95% confidence interval for mean Wages for an employee with 8 years of experience (LOS=96 months) for both Large and Small banks. Based on salary, would an employee with 8 years of experience (LOS=96) be better off at a Large or a Small bank? Justify your response.

G. Would the prediction intervals corresponding to **part F**, be wider or more narrow than the confidence intervals? (**2 pts**) You do not need to include the prediction intervals in your assignment, just state whether they would be wider or more narrow.

H. From the scatterplot in **part A**, there appears to be an outlier from the Large banks (with a particularly high salary). Provide the LOS and Wages for this observation as well as the Rstudent residual. Compute the bonferonni adjusted p-value for this observation. Note that the adjustment should be based on just Large banks (n=35).

I. Provide the estimated correlation (R) and p-value corresponding to the null hypothesis that population correlation is zero versus an alternative that it is not zero for both Large and Small banks. How do these p-values compare to the ones you reported in **part E**? (**2 pts**)

J. So far our comparison of Large and Small banks has been based on salary. Another measure related to employee satisfaction is stability (which we will loosely translate into longer LOS). Assuming the data are based on random samples from Large and Small banks, can we conclude that the mean LOS is different between Large and Small banks? What test can be used? What is the p-value?

2. Data on age in coating Thickness (X) and Strength (Y) from an experiment involving steel are available from Canvas as Steel.csv.

A. Regress Strength against Thick and look at (i) the plot of Strength versus Thick (ii) studentized residuals versus predicted values and (iii) qqplot of studentized residuals. Include these plots in your assignment. Do the regression assumptions appear to be met? **Note:** To get studentized residuals, use the stdres() function from the MASS library.

B. Perform an F-test for "lack of fit". Give your p-value and make a conclusion.

C. Now perform a quadratic regression and create a scatterplot with the fitted curve overlaid. Include the summary table and plot in your assignment. This can be done with code like the following. **Note** that b0,b1,b2 will be replaced with estimated coefficients from the quadratic regression model.

```
Fit<-lm(Strength ~ Thick + I(Thick^2), data=Steel)
plot(Strength ~ Thick, data=Steel)
curve(b0 + b1*x + b2*x^2, add=TRUE)
```