STAT 632, HW 2
Due: Friday, February 14

Reading: Chapter 3, pp.45–83, from *A Modern Approach to Regression.* (skip Section 3.3.3)
Optional Reading: Chapter 3, pp. 92–99, from *An Introduction to Statistical Learning.*

Directions: Please submit your completed assignment to Blackboard. For the concept questions, your solutions may be typed (using LaTeX or equation editor in Word), or handwritten and then scanned. For the data analysis questions, which require R, you must type your solutions. I suggest using R Markdown and knitting to PDF or HTML. Include all R code in your answers to each data analysis question.

Exercise 1.

(a) What are the assumptions for the simple linear regression model? Describe at least two diagnostics that are commonly used to check these assumptions.

(b) What does it mean for a point to be an outlier? For simple linear regression, what rule is commonly used to classify points as outliers?

(c) What does it mean for a point to have high leverage? For simple linear regression, what rule is commonly used to classify points of high leverage?

(d) For simple linear regression, what are the formulas for the error, e_i, residual, $\hat{e}_i$, and standardized residual, r_i? What is $Var(e_i)$ and $Var(\hat{e}_i)$ (just write down the formulas, no derivation necessary)? Describe two reasons why it is useful to look at a plot of the standardized residuals versus the fitted values.

Exercise 2. Mark the following as either True or False. Provide a brief explanation if you marked your answer False.

(a) A plot of the residuals versus fitted values is especially useful for checking the assumptions linearity and constant variance.

(b) The log transformation is most commonly used to stabilize the variance for count data.

(c) When considering transformations for a simple linear regression model, it is always necessary to transform both the predictor and response variable.

(d) When fitting a simple linear regression model, the most important piece of information is the R^2 (coefficient of determination). An R^2 close to 1 always indicates that a straight line is a good fit to the data.

(e) Transformations are useful for linearizing the relationships between the explanatory (X) and response (Y) variables, and for overcoming problems due to nonconstant variance.

Exercise 3.[1] This exercise uses a data set on national statistics obtained from the United Nations, and collected between 2009-2011. To load the data into R run the following command:

```
UN11 <- read.csv("https://ericwfox.github.io/data/UN11.csv")
```

The data set contains several variables, including `ppgdp`, the gross national product per person in US dollars, and `fertility`, the total fertility rate (number of children per woman).

(a) Make a scatterplot with `fertility` on the y-axis and `ppgdp` on the x-axis. Explain why we should consider log transformations for this data.

(b) Make a scatterplot of `log(fertility)` versus `log(ppgdp)` and add the least squares regression line. Does the association appear to be reasonably linear?

(c) Use the `lm()` function to fit a simple linear regression model with `log(fertility)` as the response variable, and `log(ppgdp)` as the explanatory variable. Use the `summary()` function to print the results.

(d) Write down the equation for the least squares line.

(e) Interpret the slope of the regression model.

(f) For a locality not in the data with `ppgdp` $= 1000$, obtain a point prediction and a 95% prediction interval for `log(fertility)`. If the interval (a, b) is a 95% prediction interval for `log(fertility)`, then a 95% prediction interval for fertility is given by $(\exp(a), \exp(b))$. Use this results to get a 95% prediction interval for `fertility`.

(g) Make a plot of the standardized residuals versus fitted values, and a QQ plot of the standardized residuals. Comment on whether or not the assumptions for simple linear regression appear to be satisfied.

(h) Which countries are flagged as outliers? That is, which countries have standardized residuals outside the interval from -2 to 2. In your view, does it seem necessary to remove any of these points, and then refit the model?

Bonus. [1 point] Consider the model $Y_i = \mu + e_i$, where μ is a parameter representing the population mean, and $e_i \sim N(0, \sigma^2)$ independently for $i = 1, \cdots, n$. The prediction for the response variable is given by $\hat{y} = \bar{y}$, where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is the sample mean. Derive a $1 - \alpha$ prediction interval for a single, new value of the response variable. Is the prediction interval wider or narrower than a $1 - \alpha$ confidence interval for the population mean μ? Assume that the population variance σ^2 is unknown and needs to be estimated.

[1] From Weisberg S.,*Applied linear regression*, fourth edition, Exercise 2.16, with slight modifications. The data set `UN11` was obtained from the `alr4` package.