Linear Regression
HW 2
Due 9/8 at 11:59pm

Directions: Submit a .pdf file containing your responses for the homework. The .pdf can be converted from Latex file, pictures of your handwriting solutions, word files, markdown files, etc (anything that can be converted into .pdf). If there are coding problems, only include the theory responses question in the .pdf, and upload a separate notebook for Python code.

Textbook Questions:

1. p89: # 2.1
2. p90: # 2.3
3. p91: # 2.10
4. p92: # 2.22
5. p146: # 3.2
   Hint: for each case, draw an imaginary residual plot of residuals vs. fitted values by hand. (1) assumes $\hat{\beta}_1$ is positive.

Other questions:

1. Recall the following statement about in-sample predictions: assuming (x_i, y_i) belongs to the training sample we have used to fit the SLR, then the residual e_i has mean 0 and variance $\sigma^2[1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{SSX}]$. Now imagine after fitting a regression line and obtaining the parameter estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, we are predicting an observation that's out of sample, i.e. the observation has independent variable value $X = x_0$, where x_0 does not belong to the data used to fit the model. The true response variable

   $$y_0 = \beta_0 + \beta_1 x_0 + \epsilon_0$$

   is predicted by

   $$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

   What is the mean and variance of the out-of-sample prediction bias $e_0 = y_0 - \hat{y}_0$? Hint: since this is an out-of-sample point, ϵ_0 is uncorrelated with all the in sample random errors $\epsilon_i, \quad i = 1, ..., n$.

2. **DO NOT USE PYTHON OR R.** Calculate manually or just with a simple calculator, to fill out the ANOVA table for SLR using the following data, then perform the F-test by stating the hypothesis (H_0 and H_1) and providing the test statistic and decision of the test:

X	Y
3	10
3.5	11.5
5	12
6	14

Coding questions:

1. The data "hospital_infection.csv" recorded the patients infection rates data from 58 hospitals. Let's take `InfctRisk` as the response and `Stay` (average staying days) as the predictor and perform a regression analysis using Python. You are not restricted to the libraries and functions I have used in class. Upload your python file in the homework submission. In your .pdf, answer the following questions:
   (a) From scatter plot: is there any initial relationship you can observe? i.e. linear or not, positive or negative, etc.
   (b) From simple linear regression: write down the equation of the fitted line.
   (c) From the summary: what is the test statistic and p-value of the coefficient of `Stay`. How does that suggest the impact of `Stay` to `InfctRisk`?
   (d) Calculate the 95% confidence interval for β_1. Does it match with the result from the summary output?
   (e) Run ANOVA: what is the test statistic and p-value of the F test? What does that suggest about the significance of the predictor in your model?
   (f) Verify that squaring the t-statistic associated with testing the null hypothesis $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$ is the same value as F-statistic generated from the ANOVA. Also verify that the p-values are approximately the same for the two tests.
   (g) From the summary: what is the value of R^2? What does it suggest about the goodness-of-fit of your model?
   (h) Predict the infection risk for a hospital with an average stay length of 32 days.

2. Continue with the hospital infection example from above. Use python to draw the residual v.s. fitted value plot and QQ plot based on your model. Based on your plots, evaluate the following assumptions. A brief/roughly observed comment is acceptable (i.e., seems to be false because...; hard to judge becuase...; etc.).
   (a) Infection risk and stay length have a linear relationship.
   (b) The assumption of error terms have constant variance is valid.
   (c) The assumption of error terms are independent is valid.
   (d) The assumption of error terms are normally distributed is valid.

3. Continue with the hospital infection example from above. Use Python to give the 95% confidence interval for the mean infection risk when the average stay is 32 days, and the 95% prediction interval of the infection risk for a new patient who would stay 32 days.