STAT525 HOMEWORK#3

1. KNNL Problem 2.57.
2. KNNL Problem 2.61.
3. KNNL Problem 3.19
4. (SAS Exercise) Use the **crime rate** data described in KNNL Problem 1.28.
   (a) Describe the distribution of the explanatory variable.
   (b) Run the linear regression to predict the county crime rate from the percentage of individuals having at least a high school diploma.
   (c) Plot the residuals versus the explanatory variable and briefly describe the plot noting any unusual patterns or points.
   (d) Examine the distribution of the residuals by getting a histogram and a normal probability plot of the residuals by using the `HISTOGRAM` and `QQPLOT` statements in `PROC UNIVARIATE`. What do you conclude?
5. (SAS Exercise) Use the **crime rate** data described in KNNL Problem 1.28. Change the data set by changing the value of the crime rate for the last observation from 7582 to 758 (e.g., a typo). You can do this in a data step. For example,

```
DATA a2; SET a1; IF _n_ EQ 84 THEN y=758;
```

   An alternative is to simply edit the data file.
   (a) Make a table comparing the results of this analysis with the results of the analysis of the original data. Include in the table the following: fitted equation, t-test for the slope, with standard error and p-value, R^2, and the estimate of σ^2. Briefly summarize the differences.
   (b) Repeat parts (c) and (d) from the previous problem for this altered data set analysis and summarize how these plots help you to detect the unusual observation.
6. (SAS Exercise) Use the **sales growth** data described in KNNL Problem 3.17.
   (a) Generate a scatterplot of the data and discuss the appropriateness of using a linear regression model.
   (b) Using `PROC TRANSREG`, which power transformation of Y (i.e., value of λ) is most appropriate to use here?
   (c) Apply this transformation of Y and generate a scatterplot. Again comment on the appropriateness of using a linear regression model.
   (d) Run the regression model using the transformed data and generate a residual plot (using X or $\hat{Y}$) and a normal probability plot. What do the plots show?
   (e) Express the estimated regression function in the original units.

2.57. The normal error regression model (2.1) is assumed to be applicable.

a. When testing H_0: $\beta_1 = 5$ versus H_a: $\beta_1 \neq 5$ by means of a general linear test, what is the reduced model? What are the degrees of freedom df_R?

b. When testing H_0: $\beta_0 = 2$, $\beta_1 = 5$ versus H_a: not both $\beta_0 = 2$ and $\beta_1 = 5$ by means of a general linear test, what is the reduced model? What are the degrees of freedom df_R?

a) : The reduced model is $Y_i - 5X_i = \beta_0 + \varepsilon_i$ with $df_R = n-1$ by (2.71).

b) : With $\beta_0 = 2$ and $\beta_1 = 5$, the reduced model is $Y_i - 5X_i - 2 = \varepsilon_i$ with $df_R = n$ because we have no parameters lost.

2.61. Show that the ratio $SSR/SSTO$ is the same whether Y_1 is regressed on Y_2 or Y_2 is regressed on Y_1. [*Hint:* Use (1.10a) and (2.51).]

By 1.10a, $b_1 = \dfrac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$ and by 2.51, $SSR = b_1^2 \sum(X_i - \bar{X})^2$

$$\frac{SSR}{SSTO} = \frac{\left(\dfrac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^{\cancel{2}}}\right)^2 \cancel{\sum(X_i - \bar{X})^2}}{\sum(Y_i - \bar{Y})^2} = \frac{\sum(X_i - \bar{X})^2(Y_i - \bar{Y})^2}{\sum(X_i - \bar{X})^2(Y_i - \bar{Y})^2}$$

$$= \left[\frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})(Y_i - \bar{Y})}\right]^2$$

With Y_1 regressing on Y_2, $X_i = Y_{i1}$, $Y_i = Y_{i2}$, $\bar{X} = \bar{Y}_1$, and $\bar{Y} = \bar{Y}_2$:

$$\left[\frac{\sum(Y_{i1} - \bar{Y}_1)(Y_{i2} - \bar{Y}_2)}{\sum(Y_{i1} - \bar{Y}_1)(Y_{i2} - \bar{Y}_2)}\right]^2$$

With Y_2 regressing on Y_1, $X_i = Y_{i2}$, $\bar{X} = \bar{Y}_2$, $Y_i = Y_{i1}$, and $\bar{Y} = \bar{Y}_1$:

$$\left[\frac{\sum(Y_{i2} - \bar{Y}_2)(Y_{i1} - \bar{Y}_1)}{\sum(Y_{i2} - \bar{Y}_2)(Y_{i1} - \bar{Y}_1)}\right]^2$$

and we see these are identical.

3.19. A student fitted a linear regression function for a class assignment. The student plotted the residuals e_i against Y_i and found a positive relation. When the residuals were plotted against the fitted values $\hat{Y}_i$, the student found no relation. How could this difference arise? Which is the more meaningful plot?

The positive relation for the e_i vs. Y_i plot suggests that the model may not be correctly capturing the true trend of the data; that is, the model becomes less accurate for larger values of y. The more meaningful plot is the e_i vs. $\hat{Y}_i$ plot, which should ideally show no relation/pattern.