Chapter 3 - Linear Regression

1. This is the result of regressing *sales* onto *TV*, *radio* and *newspaper*. From the table we can conclude the following:

   (a) For a 1000 unit increase in *TV* based ads, keeping *radio* and *newspaper* based ads unchanged, we will observe a 46 unit increase in *sales*.

   (b) For a 1000 unit increase in *radio* based ads, keeping *TV* and *newspaper* based ads unchanged, we will observe a 189 unit increase in *sales*.

   (c) We also see that the t-statistic of the coefficient for *newspaper* is very small. Consequently the p-value is close to 1, thus suggesting that *newspaper* is statistically insignificant.

2. For an input x the KNN-classifier will find the K closest points to x and assign the most frequently occurring class amongst the set of close points to x. The KNN-regressor on the other hand will assign the average of the K closest points as the predicted value.

3. X_1 is GPA, X_2 is IQ, X_3 is Gender (1 for Female and 0 for Male), X_4 is the interaction between GPA and IQ, and X_5 is the interaction between GPA and Gender. Thus salary (S) can be equated as follows:

$$S = 50 + 20X_1 + 0.07X_2 + 35X_3 + 0.01X_1X_2 - 10X_1X_3$$

   (a) $S_F = 50 + 20X_1 + 0.07X_2 + 35 + 0.01X_1X_2 - 10X_1$
   $S_M = 50 + 20X_1 + 0.07X_2 + 0.01X_1X_2$

$$S_F - S_M = 35 - 10X_1$$

   Now if $35 > 10X_1$ then $S_F > S_M$ else $S_F <= S_M$. Thus for a given IQ and GPA, males earn more than females provided GPA is high enough.

   (b) $S_F = 50 + 20(4) + 0.07(110) + 35 + 0.01(4)(110) - 10(4)$
   $S_F = 137.1$ units.

(c) The significance of a predictor cannot be accurately determined from its coefficient (β). The p-value for the t-statistic needs to be computed. The t-statistic is got by dividing β by the standard error of β ($SE(\beta)$). So if $SE(\beta) <<< \beta$ then β will be significant.

4. (a) When comparing based on training RSS, the cubic fit is expected to be slightly better or very similar to the linear fit.

   (b) When comparing based on test RSS, the cubic fit is expected to be worse than the linear fit due to overfitting.

   (c) When comparing based on training RSS, the cubic fit will be much better than the linear fit as the true relationship between the predictors and response is non-linear.

   (d) When comparing based on test RSS, we have two situations that might arise. In the first case if the true relationship is closer to linear than non-linear than the linear fit is expected to do better. If the true relationship is very far from linear, then the cubic fit is expected to have a better test RSS.

5. $\hat{y}_i = x_i\hat{\beta}$ where $\hat{\beta} = \frac{\sum\limits_{i=1}^{n} x_i y_i}{\sum\limits_{j=1}^{n} x_j^2}$

$$\hat{y}_i = x_i \frac{\sum\limits_{k=1}^{n} x_k y_k}{\sum\limits_{j=1}^{n} x_j^2}$$

$$\hat{y}_i = \frac{x_i x_1 y_1 + x_i x_2 y_2 + \ldots x_i x_n y_n}{\sum\limits_{j=1}^{n} x_j^2}$$

$$\hat{y}_i = \frac{x_i x_1 y_1}{\sum\limits_{j=1}^{n} x_j^2} + \frac{x_i x_2 y_2}{\sum\limits_{j=1}^{n} x_j^2} + \ldots \frac{x_i x_n y_n}{\sum\limits_{j=1}^{n} x_j^2}$$

$$\hat{y}_i = \sum_{k=1}^{n} \frac{x_i x_k}{\sum\limits_{j=1}^{n} x_j^2} y_k$$

Therefore, $a_k = \frac{x_i x_k}{\sum\limits_{j=1}^{n} x_j^2}$

6. $Y = \beta_0 + \beta_1 X$

$$\beta_0 = \bar{y} - \beta_1 \bar{x} \text{ and } \beta_1 = \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum\limits_{j=1}^{n}(x_j - \bar{x})^2}$$

$$Y = \bar{y} - \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum\limits_{j=1}^{n}(x_j - \bar{x})^2}\bar{x} + \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum\limits_{j=1}^{n}(x_j - \bar{x})^2}X$$

Now if a line passes through $(\bar{x}, \bar{y})$, then for $X = \bar{x}, Y$ will be equal to $\bar{y}$. Plugging $X = \bar{x}$ in the above equation we see that the second and third terms of the equation cancel each other and we are left with $Y = \bar{y}$.

7. $Cor(x, y) = \dfrac{\sum\limits_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum\limits_{j=1}^{n}(x_j - \bar{x})^2}\sqrt{\sum\limits_{j=1}^{n}(y_j - \bar{y})^2}}$

Given $\bar{x} = \bar{y} = 0$, $Cor(x, y) = \dfrac{\sum\limits_{i=1}^{n} x_i y_i}{\sqrt{\sum\limits_{j=1}^{n} x_j^2}\sqrt{\sum\limits_{j=1}^{n} y_j^2}}$

$$R^2 = \frac{TSS - RSS}{TSS} = \frac{\sum\limits_{i=1}^{n}(y_i - \bar{y})^2 - \sum\limits_{k=1}^{n}(\hat{y_k} - y_k)^2}{\sum\limits_{j=1}^{n}(y_j - \bar{y})^2}$$

$$R^2 = \frac{\sum\limits_{i=1}^{n} y_i^2 - \sum\limits_{i=k}^{n}(\hat{y_k} - y_k)^2}{\sum\limits_{j=1}^{n} y_j^2}$$

$$R^2 = 1 - \frac{\sum\limits_{i=k}^{n}(\hat{y_k} - y_k)^2}{\sum\limits_{j=1}^{n} y_j^2}$$

Now, $\hat{y_k} = \beta_0 + \beta_1 x_k$. Now since $\bar{x} = \bar{y} = 0$, $\beta_0 = 0$ and $\beta_1 = \dfrac{\sum\limits_{i=1}^{n} x_i y_i}{\sum\limits_{j=1}^{n} x_j^2}$.

$$\sum_{k=1}^{n}(\hat{y_k} - y_k)^2 = \sum_{k=1}^{n}(\beta_1 x_k - y_k)^2$$

$$\sum_{k=1}^{n}(\hat{y_k}-y_k)^2=\beta_1^2\sum_{k=1}^{n}x_k^2+\sum_{k=1}^{n}y_k^2-2\beta_1\sum_{k=1}^{n}x_ky_k$$

$$\frac{\sum_{k=1}^{n}(\hat{y_k}-y_k)^2}{\sum_{j=1}^{n}y_j^2}=\beta_1^2\frac{\sum_{k=1}^{n}x_k^2}{\sum_{j=1}^{n}y_j^2}+1-2\beta_1\frac{\sum_{k=1}^{n}x_ky_k}{\sum_{j=1}^{n}y_j^2}$$

$$R^2=1-\frac{\sum_{i=k}^{n}(\hat{y_k}-y_k)^2}{\sum_{j=1}^{n}y_j^2}=2\beta_1\frac{\sum_{k=1}^{n}x_ky_k}{\sum_{j=1}^{n}y_j^2}-\beta_1^2\frac{\sum_{k=1}^{n}x_k^2}{\sum_{j=1}^{n}y_j^2}$$

Substituting $\beta_1=\frac{\sum_{i=1}^{n}x_iy_i}{\sum_{j=1}^{n}x_j^2}$ in the above equation we get,

$$R^2=2\frac{\sum_{i=1}^{n}x_iy_i}{\sum_{j=1}^{n}x_j^2}\frac{\sum_{k=1}^{n}x_ky_k}{\sum_{j=1}^{n}y_j^2}-\frac{(\sum_{i=1}^{n}x_iy_i)^2}{(\sum_{j=1}^{n}x_j^2)^2}\frac{\sum_{k=1}^{n}x_k^2}{\sum_{j=1}^{n}y_j^2}$$

Thus, $R^2=\frac{(\sum_{i=1}^{n}x_iy_i)^2}{\sum_{j=1}^{n}x_j^2\sum_{j=1}^{n}y_j^2}=Cor(x,y)^2.$