Linear Regression, Classification

Theoretical Exercise

Problem 1, U

1. We assume the linear regression model has the form $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$

$$\begin{aligned}
\text{SSE} &= \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \\
\frac{\partial}{\partial \beta_0} &= -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0 \\
0 &= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\beta_0 - \sum_{i=1}^{n}\beta_1 x_i \\
\sum_{i=1}^{n}\beta_0 &= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\beta_1 x_i \\
\hat{\beta}_0 &= \bar{y} - \beta_1 \bar{x}
\end{aligned}$$

We now substitute $\hat{\beta}_0$ below to solve for β_1

$$\begin{aligned}
\text{SSE} &= \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \\
\frac{\partial}{\partial \beta_1} &= -2\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) = 0 \\
0 &= \sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) \\
0 &= \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n}\beta_0 x_i - \sum_{i=1}^{n}\beta_1 x_i^2 \\
0 &= \sum_{i=1}^{n}(x_i y_i - \beta_0 x_i - \beta_1 x_i^2) \\
&\text{(substitute: } \beta_0 = \bar{y} - \beta_1 \bar{x}) \\
0 &= \sum_{i=1}^{n}(x_i y_i - (\bar{y} - \beta_1 \bar{x})x_i - \beta_1 x_i^2) \\
0 &= \sum_{i=1}^{n}(x_i y_i - \bar{y} x_i + \beta_1 \bar{x} x_i - \beta_1 x_i^2) \\
0 &= \sum_{i=1}^{n}(x_i y_i - \bar{y} x_i) - \beta_1 \sum_{i=1}^{n}(x_i^2 - \bar{x} x_i)
\end{aligned}$$

$$\beta_1 \sum_{i=1}^{n}(x_i^2 + \bar{x}x_i) = \sum_{i=1}^{n}(x_i y_i - \bar{y}x_i)$$
$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i y_i - \bar{y}x_i)}{\sum_{i=1}^{n}(x_i^2 + \bar{x}x_i)}$$
$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i y_i) - n\bar{x}\bar{y})}{\sum_{i=1}^{n}(x_i^2) + n\bar{x}^2}$$
$$(\text{note: } \sum_{i=1}^{n}(\bar{x}^2 - x_i\bar{x}) = 0, \sum_{i=1}^{n}(\bar{x}\bar{y} - y_i\bar{x}) = 0)$$
$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i y_i - \bar{y}x_i) + \sum_{i=1}^{n}(\bar{x}\bar{y} - y_i\bar{x})}{\sum_{i=1}^{n}(x_i^2 + \bar{x}x_i) + \sum_{i=1}^{n}(\bar{x}^2 - x_i\bar{x})}$$
$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

2. In part 1, we showed $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$. By this argument, we can get $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1\bar{x}$.
3. $\hat{\beta} \sim \mathcal{N}(\beta, (X^T X)^{-1}\sigma^2)$

   $E[\hat{\beta}_1] = E\left[\frac{\sum_{i=1} n(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] = \frac{\sum_{i=1}^{n}(x_i - \bar{x})E(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\beta_1 \sum_{i=1}^{n}(x_i = \bar{x})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \beta_1$

   $var(\hat{\beta}_1) = \frac{1}{(\sum_{i=1}^{n}(x_i - \bar{x})^2)^2}\sum_{i=1}^{n}(x_i - \bar{x})^2 var(y_i) = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}$

   $\hat{\beta}_1 \sim \mathcal{N}(\beta_1, \frac{\sigma^2}{\sum(x_i - \bar{x})^2})$

   $E[\beta_0] = E(\bar{y} - \hat{\beta}_1\bar{x}) = E[\beta_0 + \beta_1\bar{x} - \hat{\beta}_1\bar{x}] = \beta_0 + \beta_1\bar{x} - \beta_1\bar{x} = \beta_0$

   $var(\hat{\beta}_0) = \frac{\sigma^2}{n} + \frac{\sigma^2\bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}\left(\sum_{i=1}^{n}(x_i - \bar{x})^2 + n\bar{x}\right) = \frac{\sigma^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}$

   $\hat{\beta}_0 \sim \mathcal{N}(\beta_0, \frac{\sigma^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2})$

Problem 2 Exercise 3.7: conceptual question

(a) We would expect the cubic regression to perform as good as or better than linear regression. This is because cubic regression model can fit the data better, thus have more flexibility. We can set the higher order terms of degree 2 or larger to zero, and we would get a linear regression. Thus, cubic regression can achieve a lower training RSS.

(b) Linear regression would have a lower test RSS. This is because the underlying relationship between X and y is linear. Fitting a higher degree polynomial would overfit the data resulting in low bias but high variance, thus having a higher RSS for the test data.

(c) Cubic regression will have a lower RSS for training data, because cubic regression is more flexible, meaning it can fit a linear regression model by setting some coefficients to zero, as well as a cubic higher degree.

(d) There is not enough information. The cubic regression could have a higher RSS because it is overfitting the training data, so resulting in higher RSS for test data. Cubic regression could also have a lower RSS, because data is of cubic degree, thus resulting both a lower RSS for train and test data.

Problem 3

1. In total, if each null hypothesis has a Type I error, then there are m Type I errors.
2. For m tests that are performed independently, then we can use Bonferroni correction. The family-wise error rate for each hypothesis is α/m.

Problem 4

1. Show that $\frac{RSS}{\sigma^2} \sim \chi^2(N-p-1)$

   $\hat{y} = X\hat{\beta} = X(X^TX)^{-1}X^Ty = Hy$, by the derivation result of least squares from earlier, where H is the hat matrix that is symmetric and idempotent.

   $\epsilon = y - \hat{y} = (I-H)y = (I-y)(X\beta+\epsilon) = (I-H)\epsilon$

   $RSS = \epsilon^T\epsilon = \epsilon^T(I-H)^T(I-H)\epsilon = \epsilon^T(I-H)\epsilon$

   Then, $\frac{(N-p)s^2}{\sigma^2} = \frac{\epsilon^T\epsilon}{\sigma^2} = (\frac{\epsilon}{\sigma})^T(I-H)(\frac{\epsilon}{\sigma}) \sim \chi^2(N-p-1)$

   By the Fisher-Cochran theorem, the degrees of freedom is the rank of $I-H$. Since (I_H) is idempotent, the rank would be the its trace, $tr(H) = tr(X(X^TX)^{-1}X^T) = p$

2. Since $\frac{\sum_{i=1}^n(y_i-\beta^Tx_i)^2}{\sigma^2} \sim \chi^2(N-1)$ and $\frac{RSS}{\sigma^2} \sim \chi^2(N-p-1)$, by the additive property of χ^2-distribution. Their difference would be a χ^2-distribution, and with degrees of freedom $(N-1)-N(N-p-1) = p+1$.

3. Suppose $BA = 0$, we diagonalize $A = UDU^+$ where $U^+U = I$, D is diagonal. Then, $BA = 0$ implies $BUD = 0$. Let $BU = [c_1, c_2]$ where c_1 column has the same number of columns as that of the diagonal element of D. Then $c_1, D_1 = 0$ iff $c_{1,ij}, D_{i,j} = 0$ for all i, j. Let $y = U^+x$. y is then also a standard normal vector.

   $Q = x^+UDU^+x = y^+Dy = y_1^+D_1y_1$, and $T = Bx = BUy = cy$. Because there exists no j such that both $c_{1,ij}$ are nonzero. In other words, Q and T compose of different y_i's. Therefore, Q and T are independent of each other.

4. For normal distribution, uncorrelatednes implies independence. To show that RSS and $\hat{\beta}$ are uncorrelated, it is sufficient to show that $E[\epsilon\hat{\beta}^T] = 0$

   $E[\epsilon\hat{\beta}^T] = E[((I-H)\epsilon)((X^TX)^{-1}X^Ty)^T] = E[(I-H)\epsilon y^TX(X^TX)^{-1}] = (I-H)E[\epsilon y^T]X(X^TX)^{-1} = (I-H)XI\sigma^2(X^TX)^{-1} = 0$, the last equation is due to $(I-H)X = 0$.

5. Earlier we showed that $RSS/\sigma^2 \sim \chi^2(N-p-1)$, then divide both sides by $(N-p-1)$, we get $\frac{RSS}{\sigma^2(N-p-1)} \sim \frac{\chi^2(N-p-1)}{N-p-1}$. Then,

$$\frac{\hat{\beta}_j - \beta_j}{s_{\hat{\beta}_j}} \sim \frac{z}{\sqrt{\frac{\chi^2(N-p-1)}{N-p-1}}}$$

   since z and χ^2 are independent, we get,

$$\frac{\hat{\beta}_j - \beta_j}{s_{\hat{\beta}_j}} \sim t(N-p-1)$$

Problem 5

1. Solution in r attached

Problem 6: LDA vs Logistics Regression

1. Solution in r attached