E6690 Stats Learning HW1

Yihang Ding yd2459 yd2459@columbia.edu

P1.

(a). $\sum_{i=1}^{n}(X_i^2) = \sum_{i=1}^{n}[(X_i-\bar{X})+\bar{X}]^2$

$= \sum_{i=1}^{n}(X_i-\bar{X})^2 + 2\bar{X}\sum_{i=1}^{n}(X_i-\bar{X}) + \sum_{i=1}^{n}\bar{X}^2$

As $\bar{X} = \frac{1}{n}\sum_{i=1}^{n}X_i$, $\sum_{i=1}^{n}(X_i-\bar{X}) = \sum_{i=1}^{n}X_i - n\bar{X} = 0$, we can get

$\sum_{i=1}^{n}X_i^2 = \sum_{i=1}^{n}(X_i-\bar{X})^2 + n\bar{X}^2$

$(n-1)S^2+n\bar{X}^2 = (n-1)\cdot\frac{1}{n-1}\cdot\sum_{i=1}^{n}(X_i-\bar{X})^2+n\bar{X}^2 = \sum_{i=1}^{n}(X_i-\bar{X})^2+n\bar{X}^2.$

So the equation is correct.

(b) As $\frac{n-1}{\sigma^2}S^2 \sim \chi^2_{n-1}$, with n-1 degrees of freedom

$E(\frac{n-1}{\sigma^2}S^2) = E(\chi^2_{n-1}) = n-1$

Thus $E(S^2) = E(\frac{\sigma^2}{n-1}\cdot\frac{n-1}{\sigma^2}\cdot S^2)$

$= \frac{\sigma^2}{n-1}\cdot E(\frac{n-1}{\sigma^2}S^2)$

$= \frac{\sigma^2}{n-1}\cdot n-1 = \sigma^2.$

So $S^2$ is an unbiased estimator of $\sigma^2$

(C) To prove $\bar{X}$ is independent of $X_i-\bar{X}$, we can prove $Cov(\bar{X}, X_i-\bar{X})=0$.

$Cov(\bar{X}, X_i-\bar{X}) = Cov(\bar{X}, X_i) - Cov(\bar{X},\bar{X})$

$= Cov(\frac{1}{n}\sum_{i=1}^{n}X_i, X_i) - Var(\bar{X})$

$= \frac{1}{n}Cov(\sum_{i=1}^{n}X_i, X_i) - \frac{S^2}{n}$

$= \frac{1}{n}[Cov(X_0,X_i)+Cov(X_1,X_i)+\cdots+Cov(X_i,X_i)] - \frac{S^2}{n}.$

As $X_1, X_2, \dots X_n$ are i.i.d., $Cov(X_i,X_j)\ (i\neq j) = 0.$

So $Cov(\bar{X}, X_i-\bar{X}) = \frac{1}{n}Cov(X_i,X_i) - \frac{\sigma^2}{n}$

$= \frac{\sigma^2}{n} - \frac{\sigma^2}{n} = 0.$

So $\bar{X}$ is independent of $X_i-\bar{X}$.

(d). As $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$, as proved in (C), $\bar{X}$ is independent of $X_i-\bar{X}$, thus $\bar{X}$ is independent of $S^2$.

P2

As $\bar{x} = \bar{y} = 0.$

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

$\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} y_i x_i}{\sum x_i^2}$ , $\hat{\beta}_0 = 0$ , i.e. $\hat{y} = \hat{\beta}_1 x$

To proof $R^2 = r^2$:

$$\frac{\sum_{i=1}^{n} \hat{y}_i^2}{\sum_{i=1}^{n} y_i^2} = \frac{\left(\sum_{i=1}^{n} x_i y_i\right)^2}{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}$$

$$\sum_{i=1}^{n} \hat{y}_i^2 = \left(\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}\right) \cdot \left(\sum_{i=1}^{n} x_i y_i\right)$$

$$\sum_{i=1}^{n} \hat{y}_i^2 = \hat{\beta}_1 \cdot \sum_{i=1}^{n} x_i y_i$$

$$\sum_{i=1}^{n} \hat{\beta}_1^2 x_i^2 = \hat{\beta}_1 \sum_{i=1}^{n} x_i y_i$$

$$\sum_{i=1}^{n} \hat{\beta}_1 x_i \cdot x_i = \sum_{i=1}^{n} x_i y_i$$

$$\cancel{\sum_{i=1}^{n} x_i y_i}$$

$$\sum_{i=1}^{n} x_i (\hat{\beta}_1 x_i - y_i) = 0.$$

$$\sum_{i=1}^{n} x_i \left(\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} x_i - y_i\right) = 0.$$

$$\sum_{i=1}^{n} x_i y_i \left(\sum_{i=1}^{n} x_i \cdot \frac{x_i}{\sum_{i=1}^{n} x_i^2} - 1\right) = 0.$$

$$\sum_{i=1}^{n} x_i y_i \cdot 0 = 0.$$

As the equation is correct, it's proved that $R^2 = r^2$.

P6

$\bar{x} = \sum_{i=1}^{20} x_i/20 = 0.428$. $\bar{y} = \sum_{i=1}^{20} y_i/20 = 19.91$.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{20}(y_i-\bar{y})(x_i-\bar{x})}{\sum_{i=1}^{20}(x_i-\bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - \bar{x}\sum_{i=1}^{n} y_i - \bar{y}\sum_{i=1}^{n} x_i + n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2} = 30.101$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 7.039$$

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

$$= \frac{1}{n-2}\sum_{i=1}^{n}(y_i-(\hat{\beta}_0+\hat{\beta}_1 x_i))^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i^2 - 2(\hat{\beta}_0+\hat{\beta}_1 x_i)y_i + (\hat{\beta}_0+\hat{\beta}_1 x_i)^2\right) = 1.36$$

when $x=0.5$, $y = 22.090$

$$R^2 = \frac{ESS}{TSS} = \frac{\sum(\hat{y}_i-\bar{y})^2}{\sum(y_i-\bar{y})^2} = \frac{\sum(\hat{\beta}_0+\hat{\beta}_1 x_i)^2 - 2\bar{y}\sum(\hat{\beta}_0+\hat{\beta}_1 x_i)^2 + \bar{y}^2}{\sum y_i^2 - 2\bar{y}\sum y_i + \bar{y}^2} = 0.977$$

P7

$$MSR = \frac{RSS}{p-1} \sim \lambda^2$$

$$MSE = \frac{ESS}{n-p} \sim \lambda^2$$

ESS has $n-p$ degrees of feedom associated with it since $p$ parameters need to be estimated in the regression function.

RSS has $p-1$ degrees of feedom associated with it, representing the number of $X$ variables $X_1, \dots X_{p-1}$. here $p-1=6$. $p=7$.

To test whether there's a regression relation between the response $Y$ and the set of $X$ variables $X_1, \dots X_{p-1}$, i.e. to choose between alternatives

$H_0$. $\beta_1 = \beta_2 = \dots = \beta_{p-1} = 0$.

$H_1$. not all $\beta_k$ ($k=1, \dots p-1$) equal zero.

$$F^* = \frac{MSR}{MSE} = \frac{\frac{ESS}{6}}{\frac{RSS}{45-7}} = \frac{\frac{TSS-RSS}{6}}{\frac{RSS}{38}} = 1.89$$

The p-value can be calculated by F-test $\rightarrow$ p-value calculator.

p-value is 0.108 for $F(6,38)$ with $F^* = 1.89$