Assignment 10

① ⓐ $y_0, y_1, \dots, y_n$ are independent normal, so y_0, $\hat{\beta}$ & $\hat{\varepsilon}$ are independent multivariate normal. y_0 & $x_0'\hat{\beta}$ are independent normal, $E(y_0) = E(x_0'\hat{\beta}) = x_0'\beta$, and $var(y_0 - x_0'\hat{\beta}) = var(y_0) + var(x_0'\hat{\beta})$

$$= \sigma^2 + x_0' \sigma^2 (X'X)^{-1} x_0' = \sigma^2(1 + x_0'(X'X)^{-1}x_0),$$

and $y_0 - x_0'\hat{\beta} \sim N(0, \sigma^2(1 + x_0'(X'X)^{-1}x_0))$. Then

$$z = \frac{y_0 - x_0'\hat{\beta}}{\sqrt{\sigma^2(1 + x_0'(X'X)^{-1}x_0}} \sim N(0,1), \text{ independent of}$$

$w = \frac{SSE}{\sigma^2}$ because SSE is a function of $\hat{\varepsilon}$, which is independent of both y_0 & $\hat{\beta}$. Now $w \sim \chi^2(n-k-1)$, so

$$t = \frac{z}{\sqrt{w/(n-k-1)}} = \frac{\dfrac{y_0 - x_0'\hat{\beta}}{\sqrt{\cancel{\sigma^2}(1 + x_0'(X'X)^{-1}x_0}}}{\sqrt{\dfrac{SSE}{\cancel{\sigma^2}}/(n-k-1)}}$$

$$= \frac{y_0 - x_0'\hat{\beta}}{\sqrt{MSE(1 + x_0'(X'X)^{-1}x_0}} \sim t(n-k-1)$$

(1b) $1-\alpha = P\left\{-t_{\alpha/2} < t < t_{\alpha/2}\right\}$

$$= P\left\{-t_{\alpha/2} < \frac{y_0 - x_0'\hat{\beta}}{\sqrt{MSE(1+x_0'(X'X)^{-1}x_0)}} < t_{\alpha/2}\right\}$$

$$= P\left\{-t_{\alpha/2}\sqrt{MSE(1+x_0'(X'X)^{-1}x_0)} < y_0 - x_0'\hat{\beta} < t_{\alpha/2}\sqrt{MSE(1+x_0'(X'X)^{-1}x_0)}\right\}$$

$$= P\left\{x_0'\hat{\beta} - t_{\alpha/2}\sqrt{MSE(1+x_0'(X'X)^{-1}x_0)} < y_0 < x_0'\hat{\beta} + t_{\alpha/2}\sqrt{MSE(1+x_0'(X'X)^{-1}x_0}\right\},$$

or $x_0'\hat{\beta} \pm t_{\alpha/2}\sqrt{MSE(1+x_0'(X'X)^{-1}x_0')}$

(2) Simple random sampling is regression with $k=0$ explanatory variables. The X matrix is a column of ones, and $(X'X)^{-1} = \frac{1}{n}$. $\hat{\beta} = \bar{y}$, $x_0 = 1$, $MSE = s^2$ (the usual sample variance) and from the formula sheet, the prediction interval is

$$x_0'\hat{\beta} \pm \sqrt{MSE(1 + x_0'(X'X)^{-1}x_0)}$$

$$= \bar{y} \pm \sqrt{s^2(1+\frac{1}{n})}$$

Deriving it, use $t = \frac{z}{\sqrt{w/\nu}}$, with

$$w = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$$

$$\bar{y} \sim N(\mu, \frac{\sigma^2}{n}), \quad y_0 \sim N(\mu, \sigma^2)$$

Independent, so $y_0 - \bar{y} \sim N(0, \sigma^2 + \sigma^2/n)$

$$z = \frac{y_0 - \bar{y}}{\sqrt{\sigma^2(1+\frac{1}{n})}} \sim N(0,1)$$

Independent of s^2 (because $\bar{y}$ & s^2 are independent and y_0 & s^2 are independent), so

$$t = \frac{\frac{y_0 - \bar{y}}{\sqrt{\cancel{\sigma^2}(1+\frac{1}{n})}}}{\sqrt{\frac{\cancel{(n-1)}s^2}{\cancel{\sigma^2}} / \cancel{(n-1)}}} = \frac{y_0 - \bar{y}}{\sqrt{s^2(1+\frac{1}{n})}} \sim t(n-1)$$

(2 continued.)

Then, $1-\alpha = P\left\{ -t_{\alpha/2} < \dfrac{y_0 - \bar{y}}{\sqrt{s^2(1+\frac{1}{n})}} < t_{\alpha/2} \right\}$

$$= P\left\{ \bar{y} - t_{\alpha/2}\sqrt{s^2(1+\tfrac{1}{n})} < y_0 < \bar{y} + t_{\alpha/2}\sqrt{s^2(1+\tfrac{1}{n})} \right\}$$

or $\bar{y} \pm \sqrt{s^2(1+\frac{1}{n})}$, which is what we got from the formula sheet.

For $\bar{y} = 7.5$, $s^2 = 3.82$ & $n = 14$, R gives a critical value qt(0.975, 13) = 2.16

The 95% margin of error is

$$t_{\alpha/2}\sqrt{s^2(1+\tfrac{1}{n})} = 2.16\sqrt{3.82\left(1+\tfrac{1}{14}\right)}$$

$= 4.37$, and the prediction interval is

$$(7.5 - 4.37,\ 7.5 + 4.37)$$

$$= (3.13,\ 11.87)$$

3 (a) $w \sim N\left(\sum_{j=n+1}^{n+m} x_j' \beta,\ m\sigma^2\right)$

(b) $\hat{w} = \sum_{j=n+1}^{n+m} x_j' \hat{\beta} = v'\hat{\beta}$,

where $v = \sum_{j=n+1}^{n+m} x_j$. This notation will save a lot of writing.

(c) $\hat{w} \sim N(v'\beta,\ \sigma^2 v'(X'X)^{-1}v)$

(d) $\hat{w}$ & w are independent, since w is based on a new set of observations. $E(w) = v'\beta$, so

$$w - \hat{w} \sim N(0,\ \sigma^2(m + v'(X'X)^{-1}v))$$

(e)

$$z = \frac{w - \hat{w}}{\sqrt{\sigma^2(m + v'(X'X)^{-1}v)}} \sim N(0,1)$$

(f)

$$t = \frac{\dfrac{w - \hat{w}}{\sqrt{\cancel{\sigma^2}(m + v'(X'X)^{-1}v)}}}{\sqrt{\dfrac{SSE}{\cancel{\sigma^2}} / (n-k-1)}}$$

$$= \frac{w - \hat{w}}{\sqrt{MSE(m + v'(X'X)^{-1}v)}} \sim t(n-k-1)$$

(3g) $w_{n+1}, \ldots w_{n+m}$ and $\hat{\beta}$ and $\hat{\varepsilon}$ are all independent. The numerator is a function of $w_{n+1}, \ldots w_{n+m}$ and $\hat{\beta}$, while the denominator is a function of $\hat{\varepsilon}$.

(h) $1-\alpha = P(-t_{\alpha/2} < t < t_{\alpha/2})$

$$= P\left(-t_{\alpha/2} < \frac{w - \hat{w}}{\sqrt{MSE(m + v'(X'X)^{-1}v}} < t_{\alpha/2}\right)$$

= Copying the steps of problem 1b,

$$= P\{\hat{w} - t_{\alpha/2}\sqrt{MSE(m + v'(X'X)^{-1}v} < w < \hat{w} + t_{\alpha/2}\sqrt{MSE(m + v'(X'X)^{-1}v}\}$$

Where again

$$w = \sum_{j=n+1}^{n+m} y_j$$

$$\hat{w} = \sum_{j=n+1}^{n+m} x_j' \hat{\beta} \quad \text{and}$$

$$v = \sum_{j=n+1}^{n+1} x_j$$

(4) $v = \begin{pmatrix} 3 \\ 36.3 \\ 450.29 \\ 220 \end{pmatrix}$, MSE = 6.889327

(a) $\hat{w} = 69.27362$

(b) $(\cancel{58.01}, \cancel{80.84})$ (59.38053, 79.16671)

(5) $tr(H) = tr(X(X'X)^{-1}X') = tr(X'X(X'X)^{-1})$
$= tr(I_{k+1}) = k+1 = \sum_{i=1}^{n} h_{ii}$, so mean hat value
is $\frac{1}{n}\sum_{i=1}^{n} h_{ii} = \frac{k+1}{n}$

(6) $H = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \frac{1}{n}(1 \ 1 \cdots 1) = \frac{1}{n}\begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}$

a) $h_{ii} = \frac{1}{n}$

b) $h_{ij} = \frac{1}{n}$

(7) Noting that $v_i' X = x_i'$,

$$h_{ii} = v_i' H v = v_i' X (X'X)^{-1} X' v_i$$

$$= x_i' (X'X)^{-1} x_i'' = x_i' (X'X)^{-1} x_i > 0$$

because $(X'X)^{-1}$ is positive definite.

(8) Because the model has an intercept

$$\sum_{i=1}^n \hat{\varepsilon}_i = \sum_{i=1}^n (y_i - \hat{y}_i) = \sum_{i=1}^n y_i - \sum_{i=1}^n \hat{y}_i = 0$$

and $\bar{\hat{y}} = \bar{y}$.

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

$$= \left(\frac{\sum_{i=1}^n (y_i - \bar{y}) \hat{y}_i - \sum_{i=1}^n (y_i - \bar{y}) \bar{y}}{\sqrt{SST * SSR}} \right)^2$$

(8 continued)

$$= \frac{\left(\sum_{i=1}^{n} y_i \hat{y}_i - \bar{y} \sum_{i=1}^{n} \hat{y}_i - \bar{y} \underbrace{\sum_{i=1}^{n} (y_i - \bar{y})}_{=0} \right)^2}{SST \quad SSR}$$

Vector notation

$$= \frac{\left(y' \hat{y} - \bar{y} n \bar{y} \right)^2}{SST \; SSR} = \frac{\left(y' H y - n \bar{y}^2 \right)^2}{SST \quad SSR}$$

$$= \frac{\left(y' H' H y - n \bar{y}^2 \right)^2}{SST \quad SSR} = \frac{\left((Hy)' H y - n \bar{y}^2 \right)^2}{SST \quad SSR}$$

$$= \frac{\left(\hat{y}' \hat{y} - n \bar{y}^2 \right)^2}{SST \quad SSR} = \frac{\left(\sum_{i=1}^{n} \hat{y}_i^2 - n \bar{y}^2 \right)^2}{SST \quad SSR}$$

$$= \frac{\left(\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \right)^2}{SST \; SSR} = \frac{SSR^2}{SST \; SSR}$$

$$= \frac{SSR}{SST} = R^2$$

(9) (a) Yes $\hat{y} = X\hat{\beta}$, and $\hat{\beta}$ is independent of $\hat{\varepsilon}$. Functions of independent random vectors are independent.

(b) There should be no visible pattern.

(10) $Cov(\hat{y}, y) = Cov(Hy, y) = H\,cov(y, y)$

$= H\sigma^2 I \neq 0$, so not independent.

(11) $cov(y, \hat{\varepsilon}) = cov(y, (I-H)y)$

$= cov(y, y)(I-H)' = \sigma^2 I (I-H)$

$= \sigma^2 (I-H) \neq 0$, so not independent

(12) A plot of x_i and $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ would be a straight line, regardless of model correctness.

13 (a) True

(b) True : $E(\hat{\varepsilon}) = E(y - \hat{y}) = E(y) - E(\hat{y})$

$$= X\beta - E(X\hat{\beta}) = X\beta - X E(\hat{\beta}) = X\beta - X\beta = \underset{\sim}{0}$$

True for $i = 1, \dots, n$.

(c) ~~True~~ False.

(d) False : $\mathrm{cov}(\hat{\varepsilon}) = \mathrm{cov}((I-H)y)$

$$= (I-H)\,\mathrm{cov}(y)\,(I-H)' = (I-H)\sigma^2 I (I-H)$$

$$= \sigma^2 (I-H)^2 = \sigma^2 (I-H), \text{ so}$$

$\mathrm{var}(\hat{\varepsilon}_i) = \sigma^2(1 - h_{ii}) \neq 0$

(e) True

(f) True : $y \sim N_n(X\beta, \sigma^2 I)$, so

$$\hat{\varepsilon} = (I-H)y \sim N_n(0, \sigma^2(I-H))$$

And the marginals of a multivariate normal are normal.

(g) True

(h) False : $\mathrm{Cov}(\hat{\varepsilon}) = \sigma^2(I-H)$, not diagonal.

(14) $\hat{\varepsilon} = y - \hat{y} = y - X\hat{\beta}$

$$\Longrightarrow \quad y = X\hat{\beta} + \hat{\varepsilon}$$

(15) Because the x variables are constants. They are independent of everything.

(16)

$$r = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(\hat{\varepsilon}_i - 0)}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2 \sum_{i=1}^{n}(\hat{\varepsilon}_i - 0)^2}}$$

Just look at the numerator

$$\text{Numerator} = \sum_{i=1}^{n} x_{ij}\,\hat{\varepsilon}_i - \bar{x}_j \sum_{i=1}^{n} \hat{\varepsilon}_i = \sum_{i=1}^{n} x_{ij}\,\hat{\varepsilon}_i$$

$= 0$, one of the zeros in $X'\hat{\varepsilon} = 0$.

To see this another way note that the $n \times 1$ vector

$$\begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix} = X p_j,$$

where p_j is a $(k+1) \times 1$ vector with a one in position j, and zero elsewhere.

$$\text{Then } \sum_{i=1}^{n} x_{ij}\,\hat{\varepsilon}_i = (X p_j)'\hat{\varepsilon} = p_j' X'\hat{\varepsilon} = p_j' \underset{(k+1)\times 1}{\mathbf{0}} = \underset{1\times 1}{0}$$

(17) $$R = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})(\hat{\varepsilon}_i - 0)}{\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{\varepsilon}_i - 0)^2}}$$

$$\text{Numerator} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})\hat{\varepsilon}_i = \sum_{i=1}^{n}\hat{y}_i\hat{\varepsilon}_i - \bar{y}\underbrace{\sum_{i=1}^{n}\hat{\varepsilon}_i}_{=0}$$

$$= \hat{y}'\hat{\varepsilon} = (X\hat{\beta})'\hat{\varepsilon} = \hat{\beta}'X'\hat{\varepsilon} = \hat{\beta}'0 = 0$$

(18) (a)

i) Yes, I suppose. Treatment increases the rate of the chemical reaction. $t = -2.133$, $p = 0.0455$.

ii) Yes. Higher concentration speeds up the reaction.

iii) Okay! I did.

(b) i) $t = -5.424$, $p = 0.0000311$

ii) $\hat{\beta}$ is negative.

(c) $t = 4.029$, $p = 0.000787$. Yes, cubic helps.

(d) Yes. The residual is less than -30.

(e) (i) Critical value $= 3.607$

ii) Yes, one outlier. Concentration $= 0.02$, Reaction rate $= 47$, treated.