STAT 331 Fall 2018 Midterm 2 Review – Solution

Lec # 1

1. (a)

(i)

Solution: The `(Intercept)` coefficient is the average market value in millions of USD for a company in the energy sector with 5 billion USD in assets and 30,000 employees.

(ii)

Solution: The `Assets5B` coefficient is the increase in average market value in millions of USD for every million USD in assets, if the sector and number of employees remains the same.

Or, it is the difference in average market value in millions of USD for two companies in the same sector with the same number of employees, but with 1 million USD difference in assets.

(iii)

Solution: The `SectorFinance` coefficient is the difference in average market value in millions of USD between a company in the finance sector and a company in the energy sector, where both have the same asset value and number of employees.

(b)

(i)

Solution:

```
M2 <- lm(MarketValue ~ Assets5B + Employees30K*Sector, data = forbes)
anova(M1, M2)
```

(ii)

Solution: There are 7 parameters in the reduced model (`M1` $= M_{\text{red}}$), 4 of which correspond to levels of `Sector` which are not subsumed by the intercept. Since the interation model adds one parameter for each of these 4 levels, there are $7 + 4 = 11$ in the interaction model (M_{full}). The F-distribution has two parameters, the first of which is the number of parameters set to 0 under H_0, which is 4, and the second is the degrees of freedom in the full model, which is $79 - 11 = 68$. Therefore, the null distribution of the F-statistic is $\mathcal{F}(4, 68)$.

(c)

Solution:

We are interested estimating

$$\begin{aligned}\gamma =& E[\texttt{MarketValue} \,|\, \texttt{Assets5B} = a, \texttt{Employees30K} = 5, \texttt{Sector} = \texttt{Finance}] \\ &- E[\texttt{MarketValue} \,|\, \texttt{Assets5B} = a, \texttt{Employees30K} = -1, \texttt{Sector} = \texttt{Retail}] \\ =& (\beta_0 + a\beta_1 + 5\beta_2 + \beta_3) - (\beta_0 + a\beta_1 - \beta_2 + \beta_6) \\ =& 6\beta_2 + \beta_3 - \beta_6,\end{aligned}$$

where $\beta = (\beta_0, \ldots, \beta_6)$ are the regression coefficients in the order in which they are displayed at the beginning of the question. To obtain a point estimate for γ, we substitute $\hat{\beta}$ for β, such that

$$\hat{\gamma} = 6\hat{\beta}_2 + \hat{\beta}_3 - \hat{\beta}_6 = 6 \cdot 130 - 2300 + 7600 = 6080.$$

2. (a)

Solution: Let $SS_{\text{err}}^{(0)}$ and $SS_{\text{err}}^{(1)}$ denote the residual sum-of-squares for models M_0 and M_1 respectively. Then the F-statistic is given by

$$F = \frac{(SS_{\text{err}}^{(0)} - SS_{\text{err}}^{(1)})/5}{SS_{\text{err}}^{(1)}/(133-7)}.$$

In this problem, we are given the unbiased estimators for each model, namely $\hat{\sigma}^2_{(0)} = SS_{\text{err}}^{(0)}/(133-2) = 76.84$ and $\hat{\sigma}^2_{(1)} = SS_{\text{err}}^{(1)}/(133-7) = 71.32$. Therefore, the F-statistic is given by

$$F = \frac{\left[(133-2)\cdot\hat{\sigma}^2_{(0)} - (133-7)\cdot\hat{\sigma}^2_{(1)}\right]/5}{\hat{\sigma}^2_{(1)}} = 3.03.$$

(b)

Solution: $F \mid H_0 \sim \mathcal{F}(5, 126)$ and $p = P(F > F_{\text{obs}})$.

(c)

Solution: Using the variable names from the dataset, model M_1 can be written as

$$\begin{aligned}\texttt{Test}_i = \beta_0 &+ \beta_1 \cdot \texttt{Poverty}_i + \beta_2 \cdot I[\texttt{City}_i = \texttt{Davenport}] \\ &+ \beta_3 \cdot I[\texttt{City}_i = \texttt{DesMoines}] + \beta_4 \cdot I[\texttt{City}_i = \texttt{IowaCity}] \\ &+ \beta_5 \cdot I[\texttt{City}_i = \texttt{SiouxCity}] + \beta_6 \cdot I[\texttt{City}_i = \texttt{Waterloo}] + \epsilon_i, \qquad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2).\end{aligned}$$

Therefore,

$$\begin{aligned} E[\texttt{Test} \mid \texttt{Poverty} = 0.25, \texttt{City} = \texttt{CedarRapids}] &= \beta_0 + 0.25\beta_1 \\ E[\texttt{Test} \mid \texttt{Poverty} = 0.5, \texttt{City} = \texttt{Davenport}] &= \beta_0 + 0.5\beta_1 + \beta_2, \end{aligned}$$

such that $\tau = -0.25\beta_1 - \beta_2$.

(d)

Solution: Let $\hat{\tau} = -0.25\hat{\beta}_1 - \hat{\beta}_2$. Using the `R` output, we know that

$$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}, \sigma^2 \begin{bmatrix} 0.17 & -0.027 \\ -0.027 & 0.097 \end{bmatrix} \right).$$

Therefore, $\hat{\tau}$ is a linear combination of normals and must therefore be normal, with

$$\begin{aligned} E[\hat{\tau}] &= -0.25E[\hat{\beta}_1] - E[\hat{\beta}_2] = -0.25\beta_1 - \beta_2 = \tau, \\ \text{var}(\hat{\tau}) &= \sigma^2 \begin{bmatrix} -0.25 \\ -1 \end{bmatrix} \begin{bmatrix} 0.17 & -0.027 \\ -0.027 & 0.097 \end{bmatrix} \begin{bmatrix} -0.25 & -1 \end{bmatrix} = \sigma^2 \cdot 0.094. \end{aligned}$$

Thus, we have $\text{se}(\hat{\tau}) = \hat{\sigma}_{(1)} \cdot \sqrt{0.094} = 2.59$, such that a 95% confidence interval for τ is

$$\hat{\tau} \pm 1.98 \cdot \text{se}(\hat{\tau}) = (19.89 - 5.13, 19.89 + 5.13) = (14.76, 25.01).$$

(e)

Solution:

```
plot(predict(M1), resid(M1),
     xlab = "Predicted Test Scores", ylab = "Residual Test Scores")
```

Lec # 2

1. (a)

Solution: Let $\tilde{y}_i = \log(y_i)$ and $\tilde{\varepsilon}_i = \log(\varepsilon_i)$. Then

$$\tilde{y}_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 2\log(|x_{i1}x_{i2}| + 1) + \tilde{\varepsilon}_i, \qquad \tilde{\varepsilon}_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2),$$

which is just an ordinary multiple regression model. With

$$\tilde{y} = \begin{bmatrix} \tilde{\varepsilon}_1 \\ \vdots \\ \tilde{\varepsilon}_n \end{bmatrix}, \qquad \tilde{X} = \begin{bmatrix} x_{11} & x_{12} & 2\log(|x_{11}x_{12}| + 1) \\ \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & 2\log(|x_{n1}x_{n2}| + 1) \end{bmatrix},$$

we have $\hat{\beta} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{y}$.

(b)

Solution: Since

$$\operatorname{var}\left(\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}\right) = \sigma^2 \begin{bmatrix} 3.4 & 0 \\ 0 & 1.3 \end{bmatrix},$$

$\hat{\beta}_1$ and $\hat{\beta}_1$ are uncorrelated and thus

$$\operatorname{var}(\hat{\tau}) = \operatorname{var}(\hat{\beta}_1) + 2^2 \operatorname{var}(\hat{\beta}_2) = \sigma^2(3.4 + 4 \times 1.3) = \sigma^2 \times 8.6,$$

such that $\operatorname{sd}(\hat{\tau})/\sigma = \sqrt{8.6} = 2.93$.

(c)

Solution: Since $\hat{\tau} = \hat{\beta}_1 - 2\hat{\beta}_2$ is a linear combination of a multivariate normal, it is also normal, with $E[\hat{\tau}] = E[\hat{\beta}_1] - 2E[\hat{\beta}_2] = \tau$ and variance from (b) calculated as $\text{var}(\hat{\tau}) = \sigma^2 \cdot 8.6$. Since

$$Z = \frac{\tau - \hat{\tau}}{\text{sd}(\hat{\tau})} \sim \mathcal{N}(0, 1)$$

is independent of $\hat{\sigma} \times (n-3)/\sigma^2 \sim t_{(n-3)}$, we have

$$\frac{\tau - \hat{\tau}}{\text{se}(\hat{\tau})} \sim t_{(n-3)}.$$

Therefore, a 95% confidence interval for τ is of the form $\hat{\tau} \pm q \cdot \text{se}(\hat{\tau})$, where

$$\hat{\tau} = \hat{\beta}_1 - 2\hat{\beta}_2 = -0.98, \qquad \text{se}(\hat{\tau}) = \hat{\sigma} \cdot 2.93 = 0.059,$$

and

$$P(|T_{(n-3)}| < q) = 0.95 \iff P(T_{(n-3)} > q) = 0.025 \implies q = 2.57.$$

Therefore, the confidence interval for τ is $(-1.13, -0.83)$. Moreover, if L and U are random variables such that $P(L < \tau < U) = 0.95$ for any value of β and σ, then

$$P(L < \tau < U) = P(1/U < 1/\tau < 1/L) = 095,$$

such that a 95% confidence interval for $\gamma = 1/\tau$ is $(-1.21, -0.88)$.

2. (a)

Solution: Taking logs, the model becomes

$$\log(y_i) = \log(\gamma_0) + \beta_1 \cdot \log(x_i) + \beta_2 \cdot 2\log(x_1 + 1) + \log(\varepsilon_i), \qquad \log(\varepsilon_i) \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2).$$

Therefore, if we let

$$z = \begin{bmatrix} \log(y_1) \\ \vdots \\ \log(y_n) \end{bmatrix} \qquad \text{and} \qquad W = \begin{bmatrix} 1 & \log(x_1) & 2\log(x_1+1) \\ \vdots & \vdots & \vdots \\ 1 & \log(x_n) & 2\log(x_n+1) \end{bmatrix},$$

the model becomes $z \sim \mathcal{N}(W\alpha, \sigma^2 I)$, where $\alpha = (\log(\gamma_0), \beta_1, \beta_2)$. The MLE of α is then calculated as $\hat{\alpha} = (W'W)^{-1}W'z$. By the plug-in principle, the MLEs of γ_0, β_1, and β_2 are

$$\hat{\gamma}_0 = \exp(\hat{\alpha}_0), \qquad \hat{\beta}_1 = \hat{\alpha}_1, \qquad \hat{\beta}_2 = \hat{\alpha}_2.$$

(b)

Solution: Note that y_i has a log-normal distribution:

$$\log(y_i) \sim \mathcal{N}\Big(\log(\gamma_0) + \beta_1 \cdot \log(x_i) + \beta_2 \cdot 2\log(x_1 + 1), \sigma^2 \Big).$$

Therefore,

$$\tau = E[y_i \,|\, x_i = 2.5] = \exp\Big(\log(\gamma_0) + \beta_1 \cdot \log(2.5) + \beta_2 \cdot 2\log(2.5 + 1) + \tfrac{1}{2}\sigma^2 \Big),$$

such that a point estimate for τ is

$$\hat{\tau} = \exp\Big(\log(5) + -3.2 \cdot 0.92 + 1.6 \cdot 2.51 + \tfrac{1}{2}0.2^2 \Big) = 14.97.$$

(c)

Solution: By the usual confidence interval procedure, we have

$$\begin{aligned} 95\% &= P\Big(\hat{\beta}_2 - 1.98 \cdot \mathrm{se}(\hat{\beta}_2) < \beta_2 < \hat{\beta}_2 + 1.98 \cdot \mathrm{se}(\hat{\beta}_2)\Big) \\ &= P\Big(\exp\{\hat{\beta}_2 - 1.98 \cdot \mathrm{se}(\hat{\beta}_2)\} - 1 < \underbrace{\exp(\beta_2) - 1}_{=\lambda} < \exp\{\hat{\beta}_2 + 1.98 \cdot \mathrm{se}(\hat{\beta}_2)\} - 1\Big), \end{aligned}$$

such that a 95% confidence interval for λ is

$$\Big(\exp(1.6 - 1.98 \cdot 0.3) - 1, \exp(1.6 + 1.98 \cdot 0.3) - 1\Big) = (1.73, 7.97).$$

3. (a)

Solution: Note that $y \sim \mathcal{N}(X\beta, \sigma^2 V)$ is multivariate normal. Therefore $y^\star = L^{-1}y$ must also be multivariate normal with

$$\begin{aligned} E[y^\star] &= L^{-1}E[y] = L^{-1}X\beta, \\ \mathrm{var}(y^\star) &= L^{-1}\mathrm{var}(y)[L^{-1}]' = \sigma^2 L^{-1}LL'[L']^{-1} = \sigma^2 I. \end{aligned}$$

(b)

Solution: Letting $X^\star = L^{-1}X$, we have $y^\star \sim \mathcal{N}(X^\star\beta, \sigma^2 I)$ which is the usual regression setting. Therefore,

$$\begin{aligned} \hat{\beta} = (X^{\star\prime}X^\star)^{-1}X^{\star\prime}y^\star &= (X'[L^{-1}]'L^{-1}X)^{-1}X'[L^{-1}]'L^{-1}y \\ &= (X'[L']^{-1}L^{-1}X)^{-1}X'[L']^{-1}L^{-1}y \\ &= (X'[LL']^{-1}X)^{-1}X'[LL']^{-1}y = (X'V^{-1}X)^{-1}X'V^{-1}y. \end{aligned}$$

4. (a)

Solution: Since $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X'X)^{-1})$, $\hat{\gamma} = A\hat{\beta}$ is a linear combination of normals, it is thus normal with mean $E[\hat{\gamma}] = AE[\hat{\beta}]$ and variance $\text{var}(\hat{\gamma}) = A\text{var}(\hat{\beta})A'$, such that

$$\hat{\gamma} \sim \mathcal{N}(\gamma, \sigma^2 A(X'X)^{-1}A').$$

(b)

Solution: Let $V = LL'$ be the Cholesky decomposition of V. Then $Z = L^{-1}(Y - \mu) \sim \mathcal{N}(0, I_q)$, such that $Z = (Z_1, \ldots, Z_q)$ are iid standard normals. Therefore,

$$Z'Z = \sum_{j=1}^{q} Z_j^2 \sim \chi^2_{(q)}.$$

On the other hand,

$$\begin{aligned} Z'Z &= (Y - \mu)'[L^{-1}]'L^{-1}(Y - \mu) \\ &= (Y - \mu)'[L']^{-1}L^{-1}(Y - \mu) \\ &= (Y - \mu)'[LL']^{-1}(Y - \mu) = (Y - \mu)'V^{-1}(Y - \mu), \end{aligned}$$

which gives the desired result.

(c)

Solution: Using the result of parts (a) and (b), we know that

$$\hat{\gamma} \mid H_0 \sim \mathcal{N}(\gamma_0, \sigma^2 A(X'X)^{-1}A'),$$

such that if $a = \gamma_0$ and $M = A(X'X)^{-1}A'$, under H_0 we have

$$W_1 = (\hat{\gamma} - a)'[\sigma^2 M]^{-1}(\hat{\gamma} - a) = \frac{(\hat{\gamma} - a)' M^{-1}(\hat{\gamma} - a)}{\sigma^2} \sim \chi^2_{(q)}.$$

Note that $W_1 = g(\hat{\beta})$, such that it is independent of

$$W_2 = e'e/\sigma^2 = h(e) \sim \chi^2_{(n-p)}.$$

By definition, an F-distribution is a ratio of independent χ^2 random variables scaled by their degrees of freedom, i.e.,

$$T = \frac{W_1/q}{W_2/(n-p)} = \frac{(\hat{\gamma} - a)' M^{-1}(\hat{\gamma} - a)/(q\sigma^2)}{e'e/((n-p)\sigma^2)} = \frac{(\hat{\gamma} - a)' M^{-1}(\hat{\gamma} - a)/q}{\hat{\sigma}^2} \sim \mathcal{F}(q, n-p).$$

Thus we have $c = q$, $a = \gamma_0$, and $M = A(X'X)^{-1}A'$.