STAT 3008: Applied Regression Analysis
2019/20 Term 2 Mid-Term Examination Quick Answers

Summary Statistics on mid-term scores: Q1= 70, Q2=84, Q3=87.25, Q4=98

- Most students lost at least 4 points from Prob 4(b)(c), and at least 7 points from Prob 5(a)
- In case of grading issues, please feel free to email Dr. Philip Lee at *pklee@sta.cuhk.edu.hk*

Problem 1:

(a) Let $g(\beta_1,\beta_2)=\sum_{i=1}^{n}(y_i-\beta_1 x_1-\beta_2 x_2)^2$. Differentiate g wrt β_1 *and* β_2,

$\frac{\partial g}{\partial \beta_1}=-2\sum_{i=1}^{n} x_{i1}(y_i-\beta_1 x_{i1})$ and $\frac{\partial g}{\partial \beta_2}=-2\sum_{i=1}^{n} x_{i2}(y_i-\beta_1 x_{i2})$.

Put $\left.\frac{\partial g}{\partial \beta_1}\right|_{\hat\beta_1=\hat\beta_2=0}=\left.\frac{\partial g}{\partial \beta_2}\right|_{\hat\beta_1=\hat\beta_2=0}=0 \Rightarrow \hat\beta_1=\frac{\sum_{i=1}^{n} x_{i1}y_i}{\sum_{i=1}^{n} x_{i1}^2}$ and $\hat\beta_2=\frac{\sum_{i=1}^{n} x_{i2}y_i}{\sum_{i=1}^{n} x_{i2}^2}$.

Since $g(\beta)$ is a convex function in β_1 and β_2, the above is an absolute minimum point.

(b) $l(\beta_1,\beta_2,\sigma^2)=-\frac{n}{2}\sum_{i=1}^{n}\ln(2\pi\sigma^2)-\sum_{i=1}^{n}\frac{1}{2\sigma^2}\left(y_i-\beta_1 x_{i1}-\beta_2 x_{i2}\right)^2$

(c) Yes, because $l(\beta_1,\beta_2,\sigma^2)=-\frac{n}{2}\sum_{i=1}^{n}\ln(2\pi\sigma^2)-\frac{1}{2\sigma^2}g(\beta_1,\beta_2)$ => Maximize $l(\beta_1,\beta_2,\sigma^2)$ based on β_1 and β_2 is equivalent to minimize $g(\beta_1,\beta_2)$.

(d) Yes, $E\left[\hat\beta_1\right]=\frac{1}{\sum_{i=1}^{n} x_{i1}^2}\sum_{i=1}^{n} x_{i1}E[y_i]=\frac{1}{\sum_{i=1}^{n} x_{i1}^2}\sum_{i=1}^{n} x_{i1}\left(\beta_1 x_{i1}+\beta_2 x_{i2}\right)=\beta_1$

(e) Yes, because $\hat\beta_1\overline{x_1^2}+\hat\beta_2\overline{x_2^2}=\frac{1}{n}\sum x_{i1}y_i+\frac{1}{n}\sum x_{i2}y_i$

Problem 2:(a) **ABA≠A Since AB = $\mathbf{0}_{n\times n}$**

(b) Since $\mathbf{B}^2=\mathbf{I}_n-2\mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}+\mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'X}(\mathbf{X'X})^{-1}\mathbf{X'}=\mathbf{I}_n-\mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}=\mathbf{B}$ and $\mathbf{A^2}=\mathbf{A}$,

⇨ $\mathbf{A}^5=\mathbf{A}=\mathbf{I_n}\text{-}\mathbf{B}=\mathbf{I_n}\text{-}\mathbf{B}^7$

(c)
$$E[\mathbf{e}'\mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X}'\mathbf{e}]=tr(E[\mathbf{e}'\mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X}'\mathbf{e}])=E(tr(\mathbf{e}'\mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X}'\mathbf{e}))=E(tr(\mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X}'\mathbf{ee}'))$$
$$=tr(\mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X}'E(\mathbf{ee}'))=tr(\mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}_n))=\sigma^2 tr((\mathbf{X'X})^{-1}\mathbf{X'X})=\sigma^2 tr(\mathbf{I_{p+1}})=(p+1)\sigma^2$$

Problem 3:

(a)

Coefficient Table

Variable	Coefficient	Std. Error	t-statistic	p-value
Constant	-9.9081	**5.3871**	-1.8392	**0.07234**
X	**0.6579**	0.2309	2.849	0.006535

ANOVA Table

Source	df	SS	MS	F	p-value
Regression	**1**	150.00	150.000	8.118	0.006535
Residuals	**46**	**850.00**	18.478		
Total	47	1000.00			

(b) (Step 1) H_o: β_o = -2.0 vs H_1: β_o > -2.0

(Step 2) t_o = (-9.9081-(-2))/5.3871 = -1.468

(Step 3) Since p-value = Pr(t_{46} > t_o) = 0.9255 > 0.05, we do not reject H_o at α = 0.05.

(Step 4) We do not have sufficient evidence that β_o is greater than -2.0.

Problem 4:

(a)

Model	EV	df	RSS
1	Null	53	1,145.7
2	1	52	194.5
3	2	52	921.3
4	3	52	10.15
5	12	51	13.10
6	13	51	5.186
7	23	51	7.314
8	123	50	3.812

(b) Yes, because R^2(Model 5) = 99.11% ≈ 1 based on x_1 and x_2, and R^2(Model 4) = 98.86% ≈ 1 based on x_3. [or simply based of the fact that $\hat{\rho}(y, x_1) = 91.118\%$ and $\hat{\rho}(y, x_3) = 99.556\%$ => x_1 and x_3 has to be highly correlated with each other]

(c) Yes. Because if $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ are orthogonal (i.e. 0 correlation), they should be orthogonal to the error variable **e** in Model 5: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$ => $SS_{total} = SS_{reg} + RSS$, where $SS_{reg} = SS_{reg(x1)} + SS_{reg(x2)}$. Now SS_{reg} = 1145.7-13.10 = 1132.6, $SS_{reg(x1)}$ = 1145.7 – 194.5 = 951.2 and $SS_{reg(x2)}$ = 1145.7 – 921.3 = 224.4 from Model 2. Since $SS_{reg(x1)} + SS_{reg(x2)}$ = 951.2 + 224.4 = 1175.6 ≈ 1132.6 = SS_{reg}, correlation between x_1 and x_2 should be close to 0.

Problem 5:

(a) If we express **Total = FirstFloor+SecondFloor+Basement**, we have

$$\begin{aligned}\hat{y} &= \hat{\gamma}_0 + \hat{\gamma}_{Year}\text{Year} + \hat{\gamma}_{FirstFloor}\text{FirstFloor} + \hat{\gamma}_{SecondFloor}\text{SecondFloor} + \hat{\gamma}_{Basement}\text{Basement} \\ &= 0.8947 + 0.0035231 \times \text{Year} + (0.00003378 + 0.0003954) \times \text{FirstFloor} \\ &\quad + 0.0003954 \times \text{SecondFloor} + (-0.0002274 + 0.0003954) \times \text{Basement} \qquad \text{- Equation (1)}\end{aligned}$$

(I) Note that $\hat{\beta}_{Basement}$ = -0.0002274 is the difference between $\hat{\gamma}_{\text{Basement}}$ and $\hat{\gamma}_{\text{SecondFloor}}$ from Equation (1),

$\hat{\beta}_{Basement} < 0 \Leftrightarrow \hat{\gamma}_{\text{Basement}} < \hat{\gamma}_{\text{SecondFloor}}$, which is intuitive because basement is typically used for storage and garage, which should be cheaper than 2/F (which is mainly utilized as bedrooms).

[Alternatively, you can view $\hat{\beta}_{Basement}$ in the original model as the change in log-price for 1 sq.ft increase in the Basement, while keeping (1) Total = FirstFloor+SecondFloor+Basement and (2) FirstFloor unchanged. (1) and (2) implies that 2/F has to be decreased by 1 sq. ft, and you are going to come up with the same conclusion.]

(II) $\hat{\beta}_{\text{FirstFloor}}$ = 0.00003378 is the difference between $\hat{\gamma}_{\text{FirstFloor}}$ and $\hat{\gamma}_{\text{SecondFloor}}$ from Equation (1),

$\hat{\beta}_{\text{FirstFloor}} > 0 \Leftrightarrow \hat{\gamma}_{\text{FirstFloor}} > \hat{\gamma}_{\text{SecondFloor}}$ is intuitive because 1/F is the place family stays most (e.g. living room, dining room, kitchen, …etc), which should be more expensive than the 2/F (mainly as bedrooms).

(II) $\hat{\beta}_{\text{Total}}$ = 0.0003954 > $\hat{\beta}_{\text{FirstFloor}}$ = 0.00003378 $\Leftrightarrow 2\hat{\gamma}_{\text{SecondFloor}} > \hat{\gamma}_{\text{FirstFloor}}$ from Equation (1), which should be intuitive since the 1 sq. ft of 1/F should not more expensive than 2 sq. ft of the 2/F.

(b) n = 1169+5 = 1174.