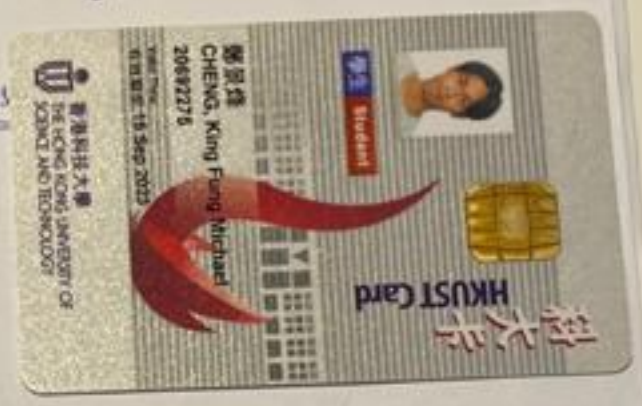
Cheng King Fung Michael SID: 2069 2275

1.a) log transformation. From scatter plot 1, we can see that the log-log scatter plot have a linear pattern while the orginal does not.



1(b) i

$e^{-13.363} = 1.572 \times 10^{-6}$

$e^{-7.861} = 3.839 \times 10^{-4}$

we are 95% confident that 'medv' is in between the range of 1.572×10^{-6} to 3.839×10^{-4}

1(b)ii The linearity, equal variance and the normality. From the residual plot, we can see that there is a funnel in the error (violated equal var). More the normal QQ plot have right skew (normality and linearity)

Cheng King Fung Michael SID: 2069 2275

1 c i)

I agree with him as the two models both have the same amount of predictors and response

I don't agree with him as the response of model 1 is 'crim' while model 2 is log('crim'), so we cannot compare directly

1 c) ii

$e^{-0.91993} = 0.3985$

$e^{3.765798} = 43.198$

The 95% prediction interval is (0.3985, 43.198)

---

2 a) There are 7 dummies in the full model

2 b) i Dallas and RH

b) ii No, as the coefficient of 'Type_of_location_urban' is only 0.0165 ≈ 0, so it makes no significant difference

P.2

Chang Hing Fung Michael SID: 2069 2275

2b) 771

$$y = 0.1935 + 0.0156x_1 - 0.0017x_2 + 0.0035x_3$$
$$- 0.0141x_4 + 0.0071x_5 + 3.328\times10^{-5}x_6 +$$
$$-0.0435 + 0.0724 + 0.1481 + 0.1481 \times (-0.0033)x_5$$

$$y = 0.3705 + 0.0151x_1 - 0.0017x_2 + 0.0035x_3 - 0.0141x_4$$
$$+ 6.61127\times10^{-3}x_5 + 3.326\times10^{-3}x_6$$

2) b) TV The β_0 changes from 0.1935 to 0.3705

While $\beta_{mx.nusorr}$ changes from 0.0071 to 6.61127×10^{-3}

R.3

Cheng King Fung Michael   SID : 2069 2275

3a) Administration, as it has the highest p value
(0.6 >> 0.05)
(p-value)

b) ~~y = 50120 + 0.8057~~
y = 50120 + 0.8057 × R&D − 0.0268 × administration + 0.0272 × marketing

c) $H_0 = F < 4$

$F = 296$ p value $= 4.53 \times 10^{-30}$

∴ Reject H_0

d) $0.602 > 0.05$

Reject H_0

e) $RMSE = \sqrt{(1 - R^2_{adj})\, S_y^2}$

$= \sqrt{(1 - 0.948)(40306)^2}$

$= 9191.22$

P.4

Cheng Hong Fung Michael SID: 2069 2275

3 f)

$$Y = 50120 + 7500(0.8057) + 150000(-0.0268) + 210100(0.0272)$$

$$Y = 111967.5$$

expected profit is $111967.5 the performance of the company is very good. It outperforms our model.

3g) No, as the VIF of all predictors << 10

P.T

Cheng King Fung Michael SID: 2069 2275

4 a) Best for 1 predictor = Age)

Best for 2 predictors = (Age, HP)

Best for 3 predictors = (Age, KM, HP)

b)

1		(Age)
2		(Age, HP)
3		(Age, KM, HP)

c) $1 + p\left(\frac{p+1}{2}\right)$

$= 1 + 3\frac{(3+1)}{2} = 7$

d)

1	4979.9561	(Age)
2	4965.8036	(Age, HP)
3	4967.1074	(Age, KM HP)

Cheng Hing Fung Michael SID: 2069 2275

5a) No, because classification method requires a response (Y), but customer segmentation in this case does not

5b) Label = 2

c) Yes, we can see that the plot decrease dramatically after k=3

d) Label = 0 (window shopper)
Label = 1 (loyal customer)
Label = 2 (normal customer)

It is because label 0 spent the less money and the most in time (it's window shopper) while label 1 spent a most money and a large period of time, so it is loyal. while the leftover is normal customer (label 2)

p.7

Cheng King Fung Michael [signature] SID: 2069 2275

5e) ~~[illegible]~~ Yes, as he can use hierarchical clustering to visualize the behaviours by a dendrogram

---

6.

3rd $92.8 + 0.507(x)$

4th $91.701 + 0.522(x)$

3rd

$\hat{y}_1 = 123.22$
$\hat{y}_2 = 111.052$
$\hat{y} = 110.038$
$\hat{y} = 120.685$

3rd MSE $= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

$= 1.227773$ //

4th

$\hat{y}_1 =$	123.027
$\hat{y}_2$	110.493
$\hat{y}_3$	109.449
$\hat{y}_4$	120.411

4th MSE $= 0.806012$ //

4-fold CV MSE $= 0.6241 + 0.690561 + 1.227773 + 0.806012$

$= 3.348$ //

Name: Chang Heng Fung Michael Student ID: 20692275

7. a) 0

b) $TPR = \frac{(\text{Real and Predict Y})}{(\text{Real Y})} = \frac{5}{9} = 0.556$

c) $FPR = \frac{0}{5} = 0$

d) $\frac{14}{28} = 35.7\%$

e) No, it is only achieved when

$TP + FN = FP + TN$

---

8 a) $\frac{1}{1 + e^{-(12.2871 + 3.5312(3.2) + 0.2149(9))}}$

≈ 1

b) $L = 12.2871 + 3.5312 \times CGA + 0.2149 \times Hours$

d) positive

Chang Long Ting Michael 20692275

e)

It's 90.625 % chance will give be correct [illegible] about the true positive (predict passed and really pass)

f)

KS measure is
Around [illegible] 0.67

[illegible]

1,10