**37252 Regression and Linear Models**

**Lab 3: Multiple Linear Regression I**

This lab is marked out of 26.

Please save your file in PDF format with name

**37252\_Lab3\_Surname\_FirstName**

**Due: 12 noon Wednesday 27 March 2024**

In this week’s lab we extend our model from the previous two labs into a multiple regression model. The data are available in **37252\_Lab3\_data.csv** which can be downloaded from Canvas.

The variables we now consider are summarised in the table below.

|  |  |  |
| --- | --- | --- |
| **Name** | **Role** | **Description** |
| $$score$$ | response | examination score |
| $$hours$$ | predictor | hours spent on revision |
| $$anxiety$$ | predictor | anxiety level |
| $$aPoints$$ | predictor | A-level entry points |

From Lab 1 we know the nature of the relationship between $score$ and $hours$, so we look now at the relationship between the response and other predictors.

> pairs(~ score + hours + anxiety + aPoints, data = scoredat)



1. Describe the direction, type and strength of the relationship between $score$ and $anxiety$ **[3 marks]** and between $score$ and $aPoints$ **[3 marks]**.

$score$ **v.** $anxiety$

Direction – negative **[1 mark]**.

Type – linear **[1 mark]**.

Strength – weak **[1 mark]**.

$score$ **v.** $aPoints$

Direction – positive **[1 mark]**.

Type – linear **[1 mark]**.

Strength – strong **[1 mark]**.

We now build our first multiple regression model.

> mod1 <- lm(score ~ ., data = scoredat)

> summary(mod1)

1. Write down the estimated regression equation **[1 mark]** and provide interpretations of the estimated beta coefficients **[4 marks]**.

Coefficients:

 Estimate Std. Error t value Pr(>|t|)

(Intercept) -11.82254 8.80575 -1.343 0.198143

hours 0.55114 0.17086 3.226 0.005284 \*\*

anxiety 0.10352 0.05762 1.796 0.091327 .

aPoints 1.98888 0.46918 4.239 0.000625 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.468 on 16 degrees of freedom

Multiple R-squared: 0.8602, Adjusted R-squared: 0.834

F-statistic: 32.81 on 3 and 16 DF, p-value: 4.563e-07

$$\hat{score}=-11.823+0.551×hours+0.104×anxiety+1.989×aPoints$$

**[1 mark]**

The coefficient $\hat{β}\_{0}=-11.823$ is the predicted $score$ when $hours=anxiety=aPoints=0$ **[1 mark]**.

The coefficient $\hat{β}\_{h}=0.551$ is the predicted change in $score$ when $hours$ increases by 1 and with $anxiety$ and $aPoints$ held constant **[1 mark]**.

The coefficient $\hat{β}\_{a}=0.104$ is the predicted change in $score$ when $anxiety$ increases by 1 and with $hours$ and $aPoints$ held constant **[1 mark]**.

The coefficient $\hat{β}\_{aP}=1.989$ is the predicted change in $score$ when $aPoints$ increases by 1 and with $hours$ and $anxiety$ held constant **[1 mark]**.

1. Test if the overall regression is significant at the 0.05 level (“overall regression is significant” is code for an F-test in multiple regression context). Write down the null and alternative hypotheses **[1 mark]**, the value of the test statistic and associated p–value **[1 mark]**, the result of the test **[1 mark]** and your conclusion in non-mathematical language **[1 mark]**.

**Hypotheses**

$$H\_{0}: β\_{h}=β\_{a}=β\_{aP}=0$$

$H\_{A}: β\_{j}\ne 0$ for at least one $j\in \left\{h,a,aP\right\}$ **[1 mark]**

**Test statistic and p-value**

Test statistic $f=32.81$ with p-value $p=0.000$ **[1 mark]**.

**Test result**

Reject $H\_{0}$ as $p<0.05$ **[1 mark]**.

**Conclusion**

The overall regression is significant **[1 mark]**.

Below is a table of some quantiles from the relevant Student’s T distribution.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| $$t\_{0.005}$$ | $$t\_{0.01}$$ | $$t\_{0.025}$$ | $$t\_{0.05}$$ | $$t\_{0.1}$$ | $$t\_{0.9}$$ | $$t\_{0.95}$$ | $$t\_{0.975}$$ | $$t\_{0.99}$$ | $$t\_{0.995}$$ |
| -2.92 | -2.58 | -2.12 | -1.75 | -1.34 | 1.34 | 1.75 | 2.12 | 2.58 | 2.92 |

1. Using 0.05 significance level, test if each extra point achieved in A-level is associated with more than an extra 6/5 points in examination score. Write down the null and alternative hypotheses **[1 mark]**, the value of the test statistic **[1 mark]**, the result of the test **[1 mark]** and your conclusion in non-mathematical language **[1 mark]**.

**Hypotheses**

$$H\_{0}: β\_{aP}=1.2$$

$H\_{A}: β\_{aP}>1.2$ **[1 mark]**

**Test statistic**

$$t=\frac{1.989-1.2}{0.469}≈1.68$$

**[1 mark]**

**Test result**

Retain $H\_{0}$ as $t<t\_{0.95}$ **[1 mark]**.

**Conclusion**

The evidence is not quite strong enough to conclude that each extra point achieved in A-level is associated with more than an extra 6/5 points in examination score **[1 mark]**.

1. Perform a visual analysis of the residuals for compliance with the normality, independence and constant variance assumptions **[3 marks]**.

 



Normality assumption – histogram looks OK (for small sample size) and PP plot shows compliance with normality, so no problem with this assumption **[1 mark]**.

Independence assumption – no signs of patterns that would indicate serial correlation, so no problem with this assumption **[1 mark]**.

Constant variance assumption – no evidence of changing variance, so no problem with this assumption **[1 mark]**.

1. Giving a reason for your answer, determine if multicollinearity is a problem with this model **[2 marks]**.

> vif(mod1)

 hours anxiety aPoints

2.288305 1.130329 2.152911

There is no problem with multicollinearity **[1 mark]** as all variance inflation figures are below the warning threshold of 5 **[1 mark]**.

Now we check for influential points.

> library('olsrr')

> ols\_plot\_cooksd\_bar(mod1)



1. Calculating a relevant statistic **[1 mark]**, identify any potentially influential points **[1 mark]**.

The critical Cook’s D for this model is

$$D\_{critical}=\frac{4}{n-m-1}=\frac{4}{20-4}=0.25 or \frac{4}{n}=0.2 $$

**[1 mark]**

We see that records 15 and 13 have Cook’s D in excess of the critical value indicating these points are potentially influential **[1 mark]**.