**37252 Regression and Linear Models**

**Lab 2: Simple Linear Regression II**

This lab is marked out of 18.

Please save your file in PDF format with name

**37252\_Lab2\_Surname\_FirstName**

**Due: 12 noon Wednesday 20 March 2024**

In this week’s lab we analyse the fit of the simple linear regression model we built last week. The data are taken from 20 students and available in **37252\_Lab2\_data.csv** which can be downloaded from Canvas.

|  |  |  |
| --- | --- | --- |
| **Name** | **Role** | **Description** |
| $$score$$ | response | examination score |
| $$hours$$ | predictor | hours spent on revision |

We now rebuild the model from last week, this time requesting a few extras to aid our analysis.

> scoredat <- read.csv("~/2024\_37252/Labs/Lab2/37252\_Lab2\_data.csv")

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

> mod1.st.resid<-rstandard(mod1)

> hist(mod1.st.resid, xlab = "Standardised residuals", freq = F, main = "")

> curve(dnorm, add = T)

> probDist <- pnorm(mod1.st.resid)

> plot(ppoints(length(mod1.st.resid)), sort(probDist), main = "Normal P-P Plot", xlab = "Observed Probability", ylab = "Expected Probability")

> abline(0,1)

We begin by checking the residuals as proxy for checking the errors for compliance with the modelling assumptions (we don’t have the errors to check directly).

First the assumption of normality. R produced a histogram and PP Plot, which are copied below.

 

To check the normality assumption visually we check the empirical plot against the theoretical plot. If they match (approximately) we conclude the assumption is met, otherwise we conclude the assumption is not met.

1. Using the histogram **[2 marks]** and PP-plot **[1 mark]** of the “Standardized” residuals, comment on the assumption of normality.

Due to the small sample size, the histogram is quite “blocky”, but does appear symmetrical with no outliers, so nothing contradicting normality **[2 marks]**.

The PP plot shows the residuals tracking the line of theoretical behaviour, so nothing contradicting normality **[1 mark]**.

We can also check normality statistically via a hypothesis test.

> shapiro.test(mod1.st.resid)

 Shapiro-Wilk normality test

data: mod1$residuals

W = 0.97799, p-value = 0.9055

We are testing the following hypotheses

$H\_{0}:$ the residuals $\hat{ϵ}\_{i}$ are normally distributed

$H\_{A}:$ the residuals $\hat{ϵ}\_{i}$ are not normally distributed.

1. Are the residuals normally distributed at the 0.05 significance level? **[2 marks]**?

Yes – with a p-value reported as 0.906 (greater than 0.05) **[1 mark]** we can retain the null hypothesis that the residuals behave as normally-distributed RVs **[1 mark]**.

Next we check the residuals for the assumptions of independence and constant variance. We use a scatter plot for this.

> plot(scoredat$hours, mod1.st.resid, xlab = "hours", ylab = "Standarised residuals")



For independence we want to see no recognisable patterns (e.g. lines, curves etc.). For constant variance we want to see similar vertical variation as we move along horizontally.

1. Using the scatterplot, comment on the assumptions of independence **[1 mark]** and constant variance **[1 mark]**.

Independence – no signs of patterns indicating serial correlation, so no breach of assumption **[1 mark]**.

Constant variance – no evidence of increasing variance, so no breach of assumption **[1 mark]**.

We can also check the independence assumption with the Durbin-Watson statistic.

> library('car')

> durbinWatsonTest(mod1)

 lag Autocorrelation D-W Statistic p-value

 1 -0.2086104 2.393437 0.298

 Alternative hypothesis: rho != 0

1. Is there any statistical evidence that the residuals are not independent **[2 marks]**.

No – the DW statistic is between 1 and 3 indicating no problem with serial correlation **[2 marks]** and p value is > 0.05.

Now we check for influential points using Cook’s D.

> cooksD<-cooks.distance(mod1)

> boxplot(cooksD, main = "Boxplot of Cook's Distance")

> abline(h = 0.22)



> 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

$$D\_{critical}=\frac{4}{n-m-1}=\frac{4}{20-2}≈0.22.$$

**[1 mark]**

We see that record 13 has Cook’s D in excess of the critical value indicating this point is potentially influential **[1 mark]**.

When we identify potentially influential points we exclude them and rebuild the model. If the estimated beta-coefficients in the reduced dataset model have changed significantly from the full dataset model we retain the reduced dataset model, otherwise we return to the full dataset model. Additionally, if excluding the points improves the behaviour of the residuals with regard to the assumptions we retain the reduced dataset model, otherwise we return to the full dataset model.

> scoredat\_reduced <- scoredat[-13,]

> mod2 <- lm(score ~ hours, data = scoredat\_reduced)

> summary(mod2)

1. Calculate the proportional changes in the beta coefficients between the filtered and full dataset models **[2 marks]**.

Coefficients:

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

(Intercept) 13.9408 6.8678 2.030 0.0583 .

hours 1.2232 0.1773 6.898 2.58e-06 \*\*\*

**Full dataset model**

$$\hat{score}=22.164+0.992×hours$$

**Filtered dataset model**

$$\hat{score}=13.941+1.223×hours$$

**Proportional changes in beta-coefficients**

Proportional change in estimated $β\_{0}$: $\frac{13.941-22.164}{22.164}≈-37.101\%$

Proportional change in estimated $β\_{1}$: $\frac{1.223-0.992}{0.992}≈23.286\%$

**[2 marks]**

The proportional changes in the beta-coefficients are large enough to worry about, so the filtered dataset model should be preferred on this basis.

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

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| $$t\_{0.005}$$ | $$t\_{0.01}$$ | $$t\_{0.025}$$ | $$t\_{0.05}$$ | $$t\_{0.1}$$ | $$t\_{0.9}$$ | $$t\_{0.95}$$ | $$t\_{0.975}$$ | $$t\_{0.99}$$ | $$t\_{0.995}$$ |
| -2.88 | -2.55 | -2.10 | -1.73 | -1.33 | 1.33 | 1.73 | 2.10 | 2.55 | 2.88 |

1. Using 0.05 significance level, perform a hypothesis test to determine if the true value of the slope coefficient in the filtered dataset model is less than 1.537. Write down the hypotheses **[1 mark]**, the test statistic **[1 mark]**, the result of the test **[1 mark]** and a conclusion in non-mathematical language **[1 mark]**.

**Hypotheses**

$H\_{0}:$ $β\_{1}=1.537$

$H\_{A}:$ $β\_{1}<1.537$ **[1 mark]**

**Test statistic**

$$t=\frac{1.223-1.537}{0.177}=-1.77$$

**Test decision**

Reject the null hypothesis as $t<t\_{0.05}=-1.73$ **[1 mark]**.

**Conclusion**

There is evidence that for each extra hour of revision, examination score is predicted to increase by less than 1.537 **[1 mark]**.

1. The coefficient of determination for the filtered dataset model is $R^{2}=0.737$. How is this related to the variables in the model **[1 mark]**?

The square root of the coefficient of determination is the sample correlation between $hours$ and $score$, i.e. $R=corr(hours\_{i},score\_{i})$ **[1 mark]**.