**37252 Regression and Linear Models**

**Lab 1: Simple Linear Regression I**

This lab is marked out of 20.

Please save your file in PDF format with name

**37252\_Lab1\_Surname\_FirstName**

**Due: 12 noon Wednesday 13 March 2024**

In this week’s lab we build a model to predict university examination scores based on hours spent on revision. The data are taken from 20 students and available in **37252\_Lab1\_data.csv** which can be downloaded from Canvas.

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

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

In order to know what type of model to build we need to examine the nature of the relationship between the variables. First we look at this visually with a scatter plot.

> scoredat <- read.csv("~/2024\_37252/Labs/Lab1/37252\_Lab1\_data.csv")

> plot(scoredat$hours, scoredat$score, xlab = "hours", ylab = "score")



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

We can also measure the strength of a relationship statistically. One such measure is Pearson’s correlation coefficient, which measures the strength of a linear relationship.

> cor.test(scoredat$hours, scoredat$score, method = "pearson")

 Pearson's product-moment correlation

data: scoredat$hours and scoredat$score

t = 6.1012, df = 18, p-value = 9.166e-06

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

 0.5944695 0.9268088

sample estimates:

 cor

0.8210109

The **Sig. (2-tailed)** field is the p-value from the test of the hypotheses

$$H\_{0}: ρ=0$$

$$H\_{A}: ρ\ne 0$$

where

$$ρ=corr\left(score,hours\right)$$

is the (unknown) population correlation coefficient.

Note that

$$\hat{ρ}=corr\left(score\_{i},hours\_{i}\right)=0.821$$

is the sample correlation coefficient.

1. Is the population correlation coefficient statistically different from zero at the 0.05 significance level? **[1 marks]**. Compare the sample correlation coefficient to your observations from part (a) **[2 marks]**.

We now build our first simple linear regression model.

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

> summary(mod1)

Call:

lm(formula = score ~ hours, data = scoredat)

Residuals:

 Min 1Q Median 3Q Max

-10.6747 -4.1513 0.1568 4.4249 12.2210

Coefficients:

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

(Intercept) 22.1640 6.5257 3.396 0.00322 \*\*

scoredat$hours 0.9920 0.1626 6.101 9.17e-06 \*\*\*

---

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

Residual standard error: 6.431 on 18 degrees of freedom

Multiple R-squared: 0.6741, Adjusted R-squared: 0.656

F-statistic: 37.22 on 1 and 18 DF, p-value: 9.166e-06

> confint(mod1)

 2.5 % 97.5 %

(Intercept) 8.4540373 35.874044

hours 0.6503953 1.333562

1. Write down the estimated regression equation **[1 mark]** and provide interpretations of the estimated beta coefficients **[2 marks]**.
2. What restrictions should be placed on the values that $hours$ can take **[1 mark]**?

For a model to be useful we need it to be “statistically significant”. For simple linear regression this means we have to show that $β\_{1}\ne 0$, one of which ways to do so is via a T-test.

1. Is the regression model significant at the 0.05 level? Write down the hypotheses **[1 mark]**, the test statistic and p-value **[1 mark]**, the result of the test **[1 mark]** and a conclusion in non-mathematical language **[1 mark]**?
2. Test whether or not $β\_{0}=36$ at the 0.05 level. Write down the hypotheses **[1 mark]**, the result of the test with reason **[2 marks]** and a conclusion in non-mathematical language **[1 mark]**.

We can now use the regression model to predict $score$ when $hours=33$ (or any other non-negative value) and to find a 95% confidence interval for the mean predicted $score$ at this level of $hours$. We can also ask for the 95% confidence interval for the “individual” predicted $score$ at this level of $hours$. This can be done by hand using the formulae in the notes (see pages 32 and 33 Lecture 2 Notes), but we can also get R to do it for us.

> newdata = data.frame(hours=33)

> predict(mod1, newdata,interval="confidence", level = 0.95)

 fit lwr upr

1 54.89933 51.21958 58.57909

> predict(mod1, newdata,interval="predict", level = 0.95)

 fit lwr upr

1 54.89933 40.89609 68.90258

1. With reference to equations (21) and (23) in Lecture 2, explain why the 95% confidence interval for an individual predicted value is wider than the 95% confidence interval for the mean predicted value **[1 mark]**.
2. Suppose that the observed value of the $score$ when $hours=33$ was 57. Calculate the residual at this point **[1 mark]**.