**37252 Regression and Linear Models**

**Lab 6: Non-linear Regression**

This lab is marked out of 26.

Please save your file in PDF format with name

**37252\_Lab6\_Surname\_FirstName**

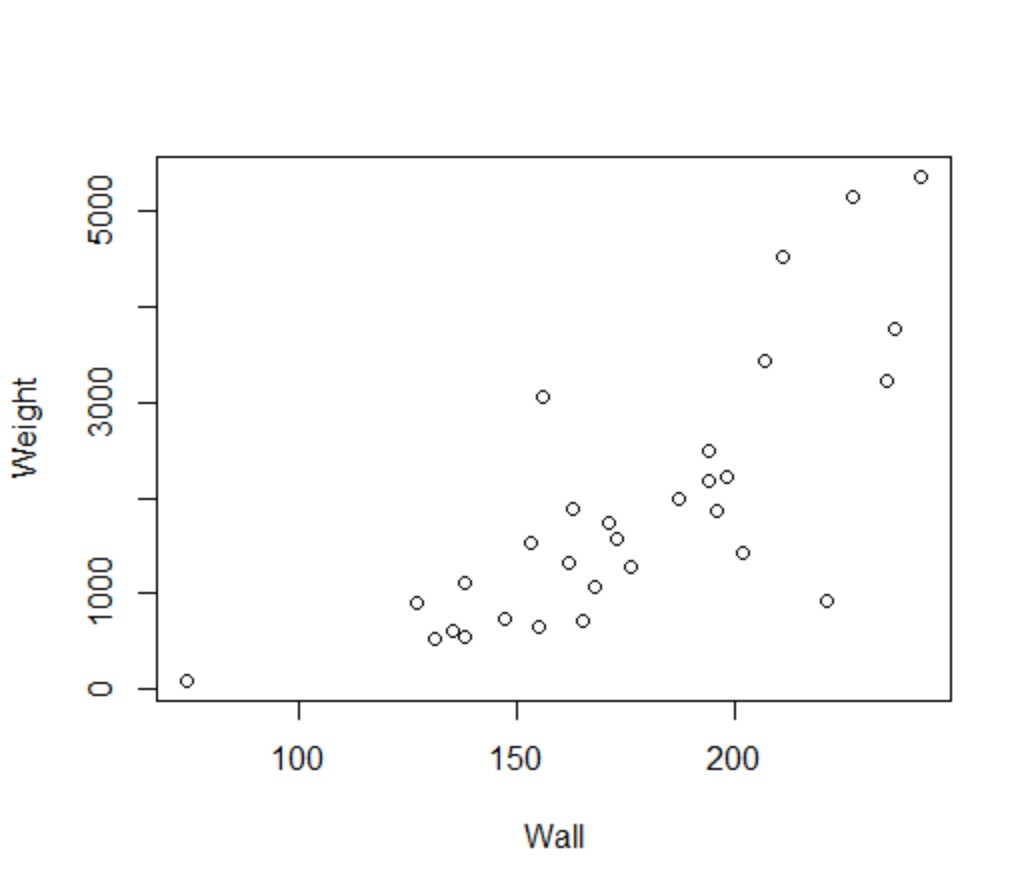
**Due: 12 noon Wednesday 24 April 2024**

In this week’s lab model octopus beak weight. The data are a sample from 30 octopi and available in **37252\_Lab6\_data.csv** which can be downloaded from Canvas.

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

|  |  |  |
| --- | --- | --- |
| **Name** | **Role** | **Description** |
|  | response | weight of octopus beak |
|  | predictor | length of lateral wall of octopus beak |

Below is a scatter plot of the data (see previous labs for graphing instructions).



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

Direction – positive **[1 mark]**.

Type – positively curved **[1 mark]**.

Strength – medium-strong **[1 mark]**.

1. If a straight-line model was fitted to the data, what two problems would we see with the residuals **[2 marks]**?

Curvature implying serial correlation and therefore breach of assumption of independence **[1 mark]**.

Increasing variance **[1 mark]**.

Let’s ignore the curvature in the data (hint!) and build a straight-line model, making sure to request a histogram and PP plot and to save the standardised residuals.

1. Write down the estimated regression model **[1 mark]** and provide interpretations of the estimated beta coefficients **[2 marks]**.

> weightdat <- read.csv("~/2024\_37252/Labs/Lab6/37252\_Lab6\_data.csv")

> mod1 <- lm(weight ~ wall, data = weightdat)

> summary(mod1)

Call:

lm(formula = weight ~ wall, data = weightdat)

Residuals:

Min 1Q Median 3Q Max

-2254.7 -381.3 -196.5 299.4 1808.8

Coefficients:

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

(Intercept) -2987.678 785.816 -3.802 0.000713 \*\*\*

wall 27.907 4.362 6.398 6.31e-07 \*\*\*

---

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

Residual standard error: 904.8 on 28 degrees of freedom

Multiple R-squared: 0.5938, Adjusted R-squared: 0.5793

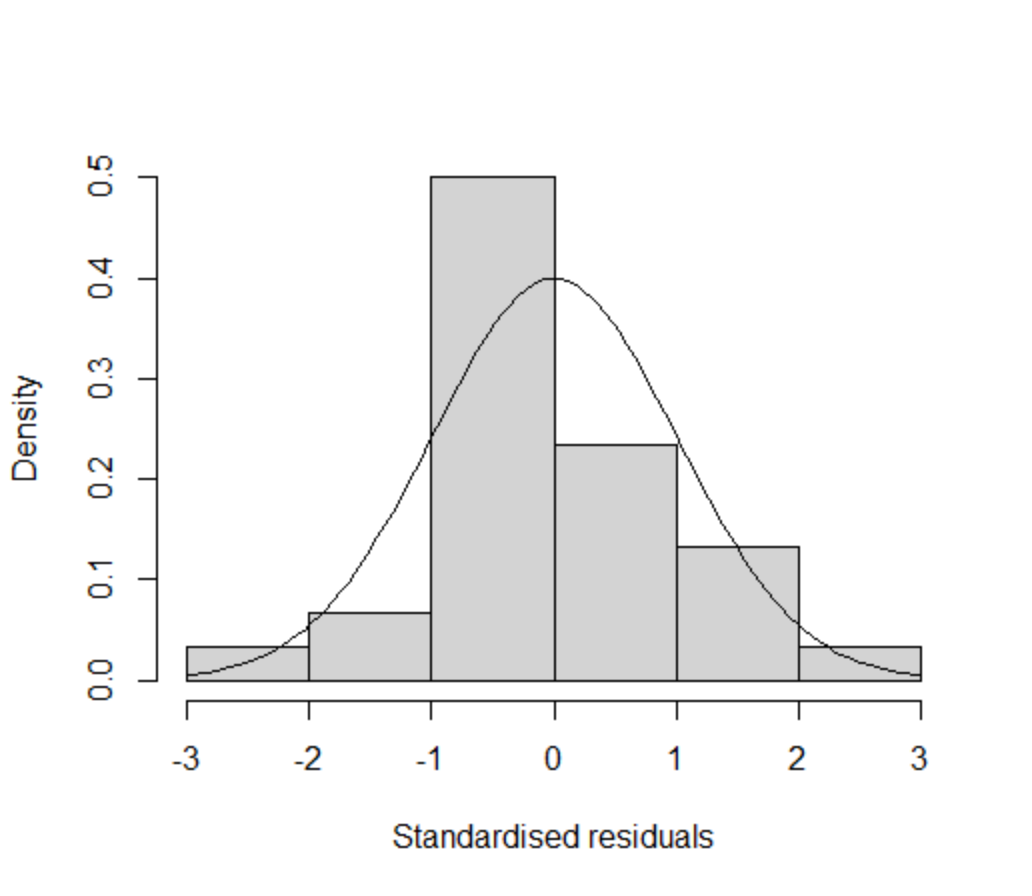
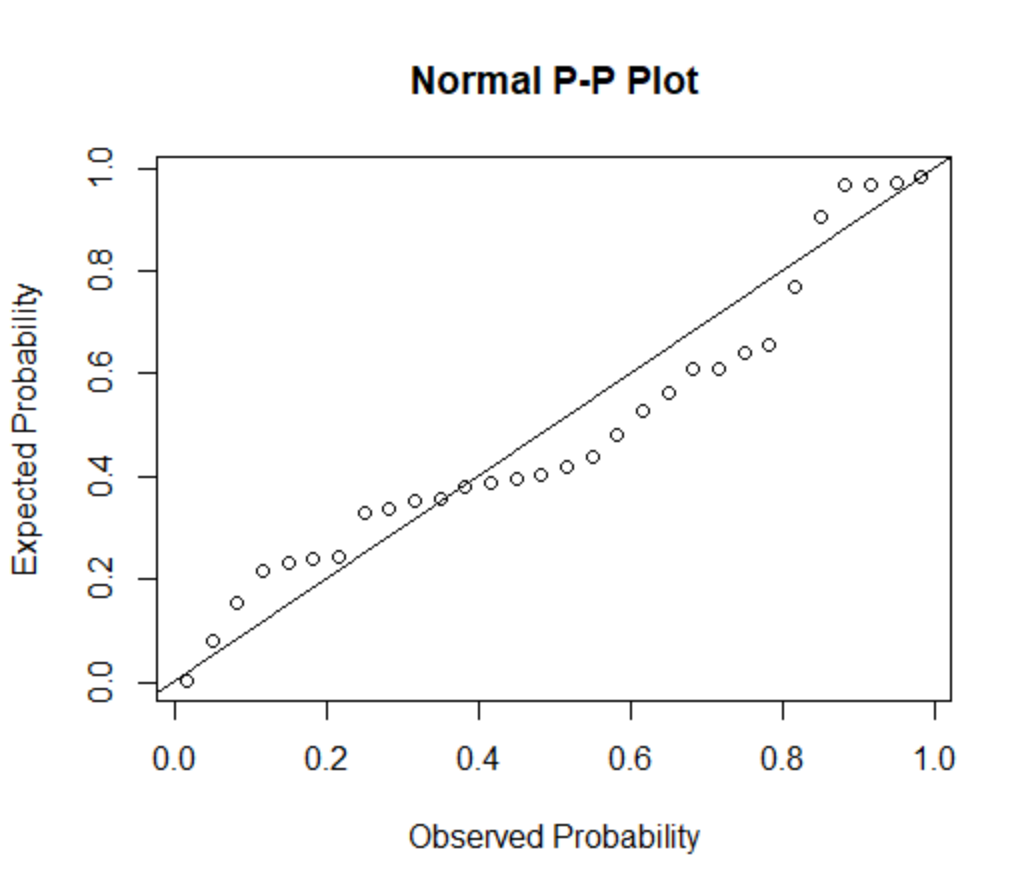
F-statistic: 40.93 on 1 and 28 DF, p-value: 6.315e-07

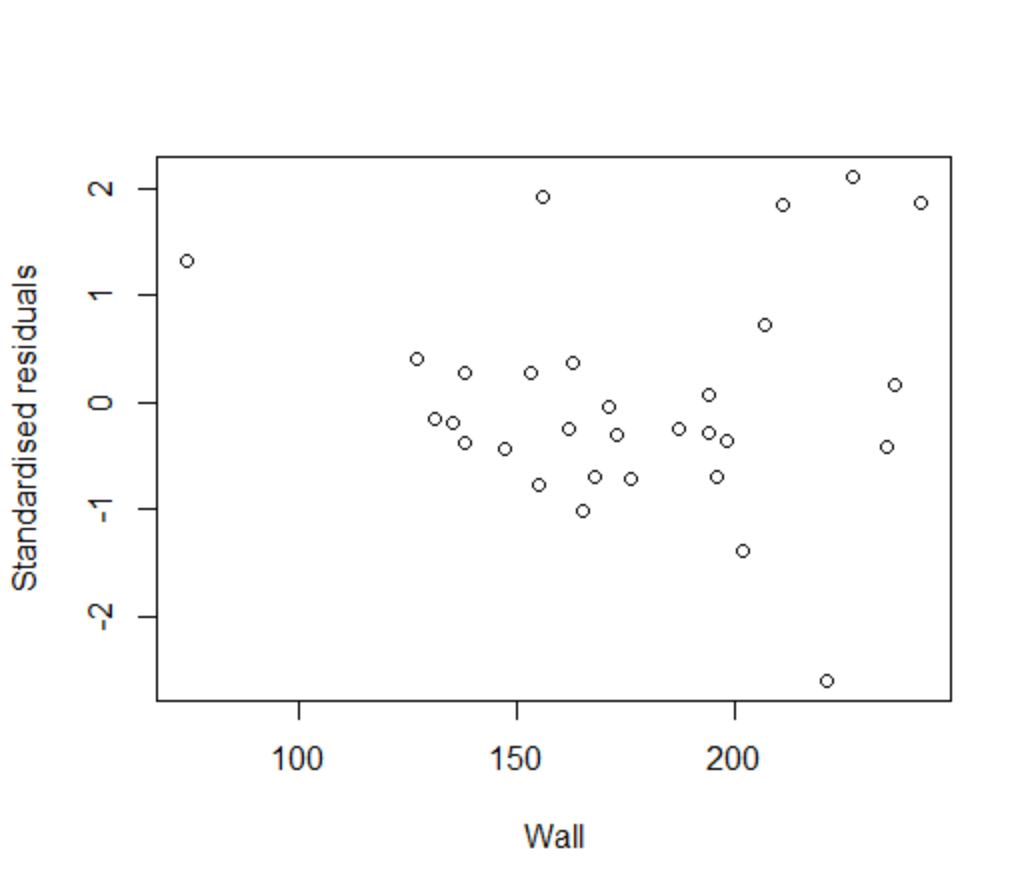
**[1 mark]**

The coefficient is the predicted when **[1 mark]**. Note that this makes no sense – in making this interpretation we have extrapolated to far outside of the range of sample on which the model was built.

The coefficient is the predicted change in when increases by 1 **[1 mark]**.

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



Normality assumption – PP plot in particular shows large departures from normality **[1 mark]**.

Independence assumption – curvature implying serial correlation in breach of assumption **[1 mark]**.

Constant variance assumption – increasing variance, so breach of this assumption **[1 mark]**.

Let’s now try to deal with the curvature with a polynomial regression by replacing with as the predictor (you will need to create the variable – see previous labs for instructions). Again, request a histogram and PP plot and save the standardised residuals.

Write down the estimated regression model **[1 mark]** and provide interpretations of the estimated beta coefficients **[2 marks]**.

> weightdat$wallSqrd <- weightdat$wall^2

> mod2 <- lm(weight ~ wallSqrd, data = weightdat)

> summary(mod2)

Call:

lm(formula = weight ~ wallSqrd, data = weightdat)

Residuals:

Min 1Q Median 3Q Max

-2367.4 -450.8 -137.0 347.2 1813.8

Coefficients:

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

(Intercept) -776.01542 419.38900 -1.850 0.0748 .

wallSqrd 0.08330 0.01198 6.952 1.47e-07 \*\*\*

---

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

Residual standard error: 859.8 on 28 degrees of freedom

Multiple R-squared: 0.6332, Adjusted R-squared: 0.6201

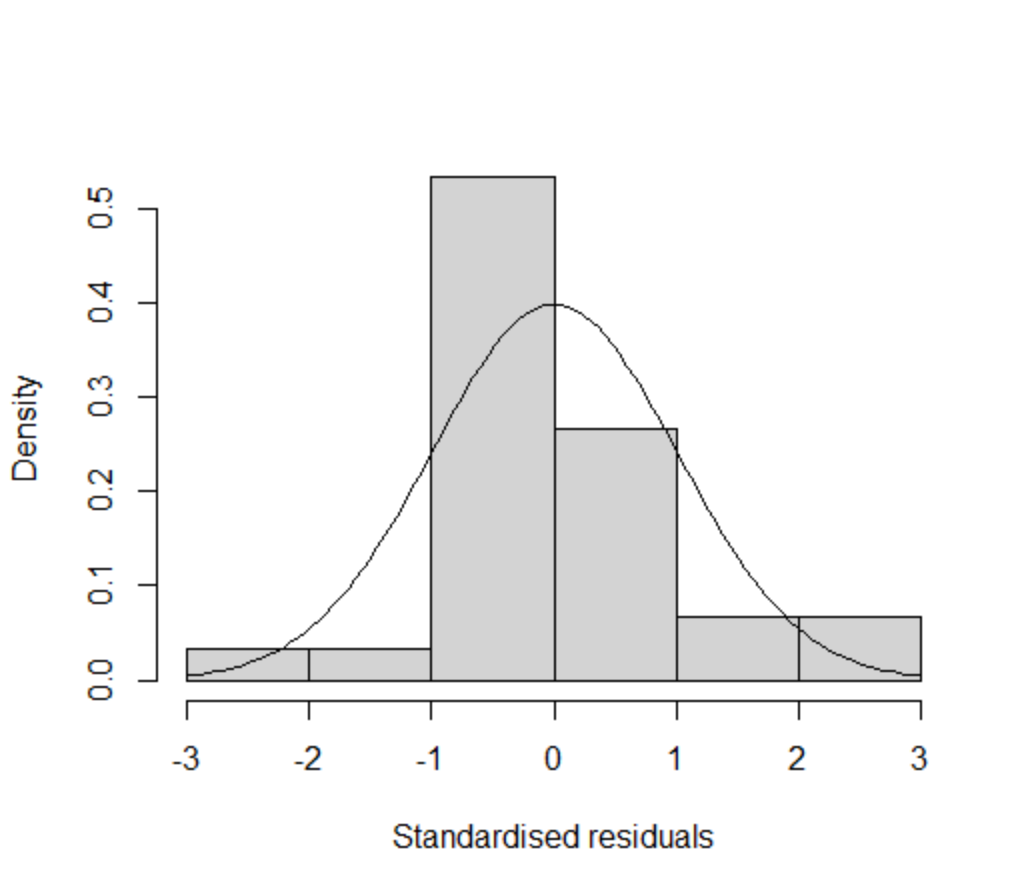
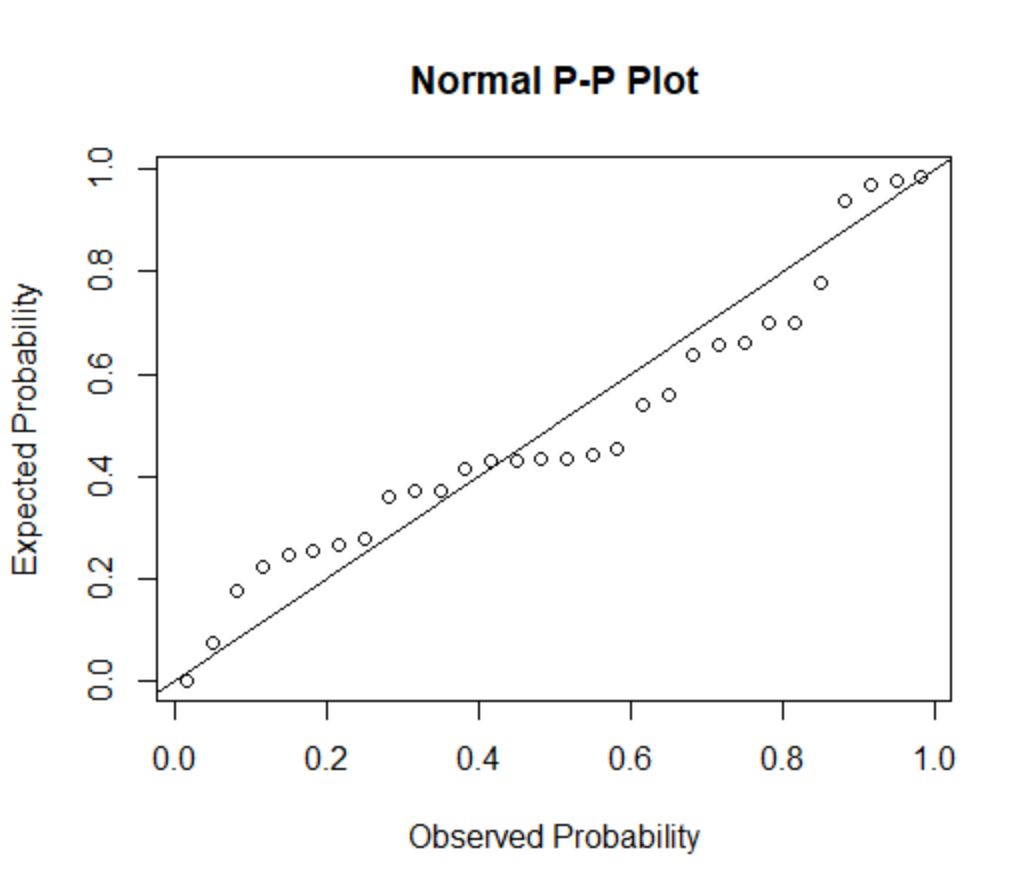
F-statistic: 48.33 on 1 and 28 DF, p-value: 1.473e-07

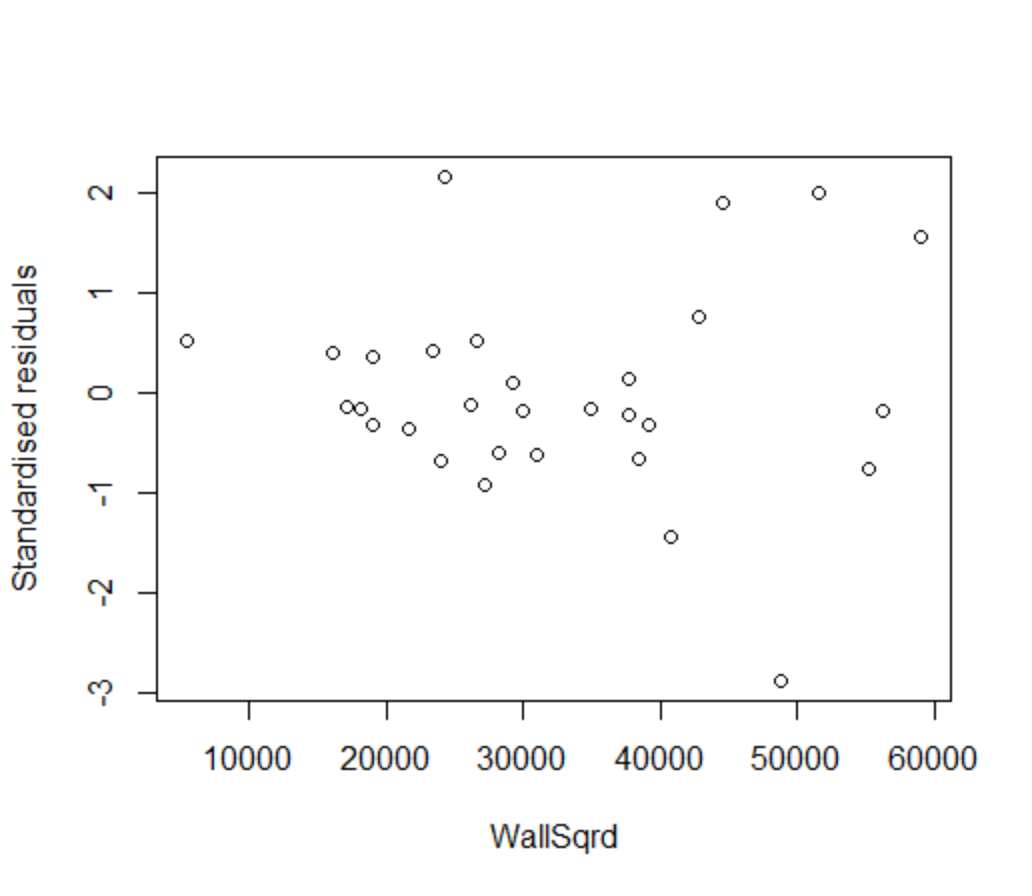
**[1 mark]**

The coefficient is the predicted when **[1 mark]**.

The coefficient is the predicted change in when increases by 1 **[1 mark]**.

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



Normality assumption – PP plot in particular shows large departures from normality **[1 mark]**.

Independence assumption – a small hint of curvature now, but assumption OK **[1 mark]**.

Constant variance assumption – increasing variance, so breach of this assumption **[1 mark]**.

The polynomial model has removed the curvature in the residuals but the increasing variance remains (hint!). There are a couple of ways we can approach this; next week we will look at weighted least squares (WLS) regression but here we will make a log transform of weight.

Construct a model with as response and as predictor, again requesting residuals plots and saving the standardised residuals and Cook’s distances.

1. Write down the estimated regression model in log-units **[1 mark]** and provide interpretations of the estimated beta coefficients **[2 marks]**. Re-write the regression equation in original units of **[1 mark]** and re-interpret the estimated exponential of the beta coefficients **[2 marks]**.

> weightdat$weightLn <- log(weightdat$weight)

> mod3 <- lm(weightLn ~ wallSqrd, data = weightdat)

> summary(mod3)

Call:

lm(formula = weightLn ~ wallSqrd, data = weightdat)

Residuals:

Min 1Q Median 3Q Max

-1.38767 -0.24868 0.01323 0.32098 1.16444

Coefficients:

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

(Intercept) 5.618e+00 2.538e-01 22.135 < 2e-16 \*\*\*

wallSqrd 5.116e-05 7.252e-06 7.055 1.13e-07 \*\*\*

---

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

Residual standard error: 0.5204 on 28 degrees of freedom

Multiple R-squared: 0.64, Adjusted R-squared: 0.6271

F-statistic: 49.77 on 1 and 28 DF, p-value: 1.129e-07

**Log-units**

**[1 mark]**

The coefficient is the predicted when **[1 mark]**.

The coefficient is the predicted change in when increases by 1 **[1 mark]**.

**Original units**

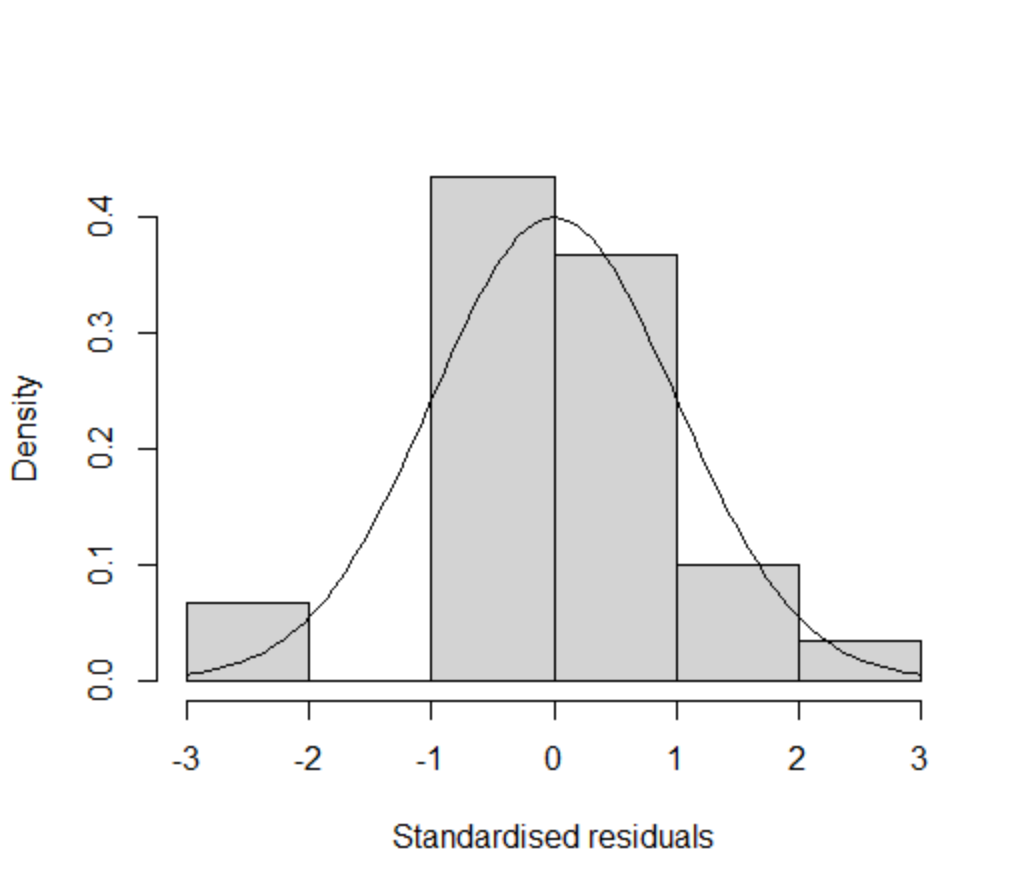
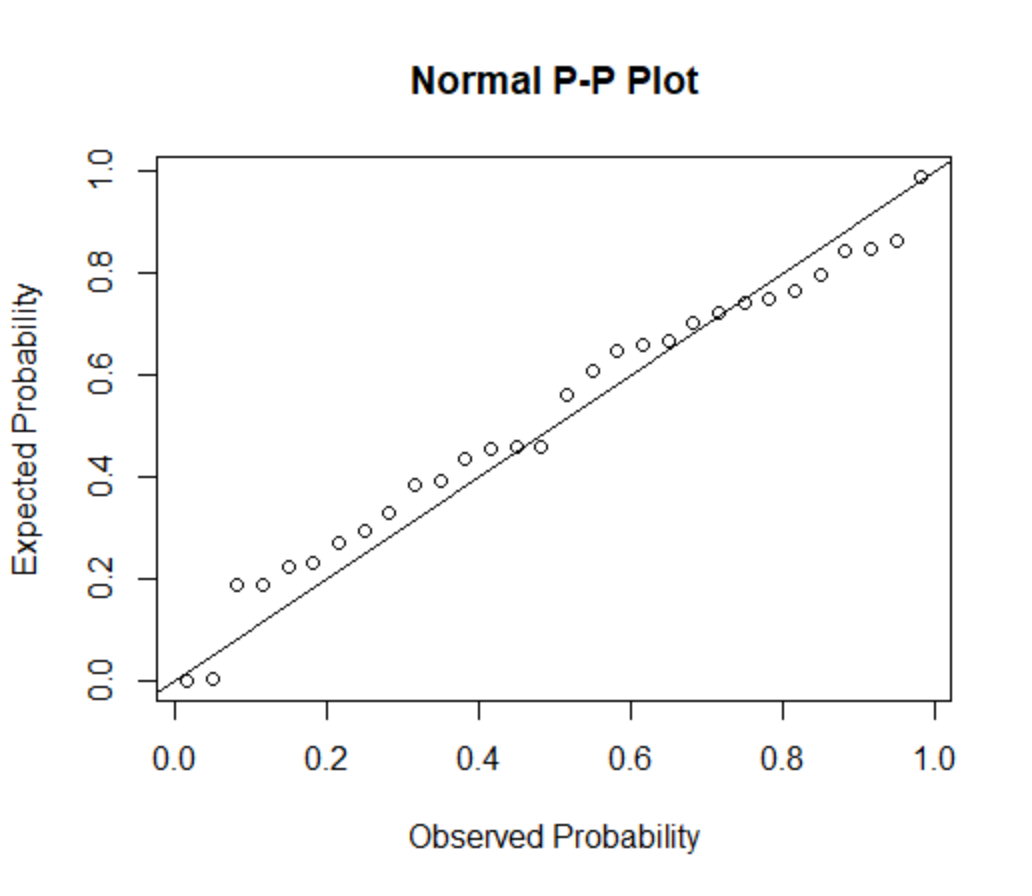
**[1 mark]**

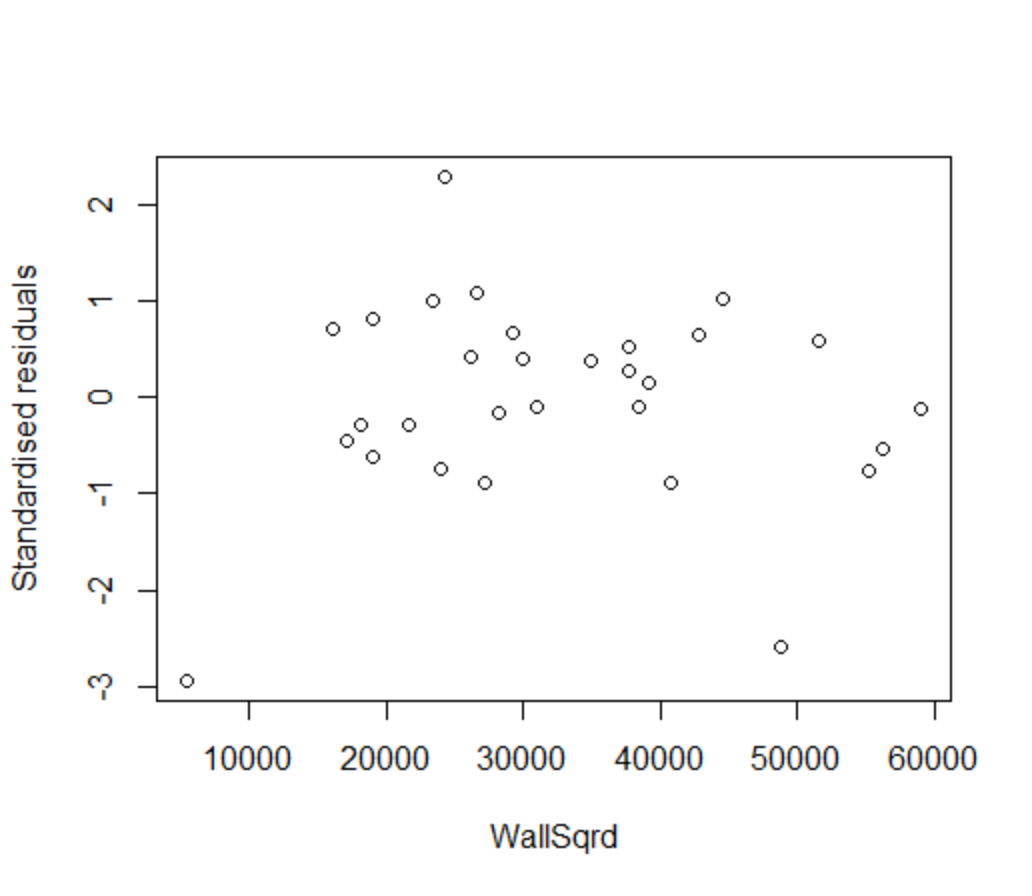
The exponential of the coefficient is the predicted when **[1 mark]**.

The exponential of the coefficient is the predicted multiple of when increases by 1, i.e.

**[1 mark]**

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



Normality assumption – PP plot in particular shows large departures from normality **[1 mark]**.

Independence assumption – no curvature now, so assumption OK **[1 mark]**.

Constant variance assumption – constant variance, so assumption OK **[1 mark]**.

**Extra work [0 marks].** Identify and filter out any influential points, run the model again and see if this improves the residuals with respect to the assumption of normality.