**Understanding Data and Statistical Design (60117)**

**Lab 6: Simple Linear Regression II**

This lab is marked from 18.

Please submit via Canvas.

**Due by the conclusion of the lab class**

In this week’s lab we look at assessing model fit for simple linear regression.

**QUESTION 1. Iris flower petal characteristics**

In this week’s lab we revisit the model from last week and assess its fit. We again consider the variables in the table below.

|  |  |  |
| --- | --- | --- |
| **Name** | **Type** | **Description** |
|  | response | length of flower sepal |
|  | predictor | length of flower petal |

The data is from the “iris” data set built into R (see accompanying R code file).

Recall the hypothesised population model

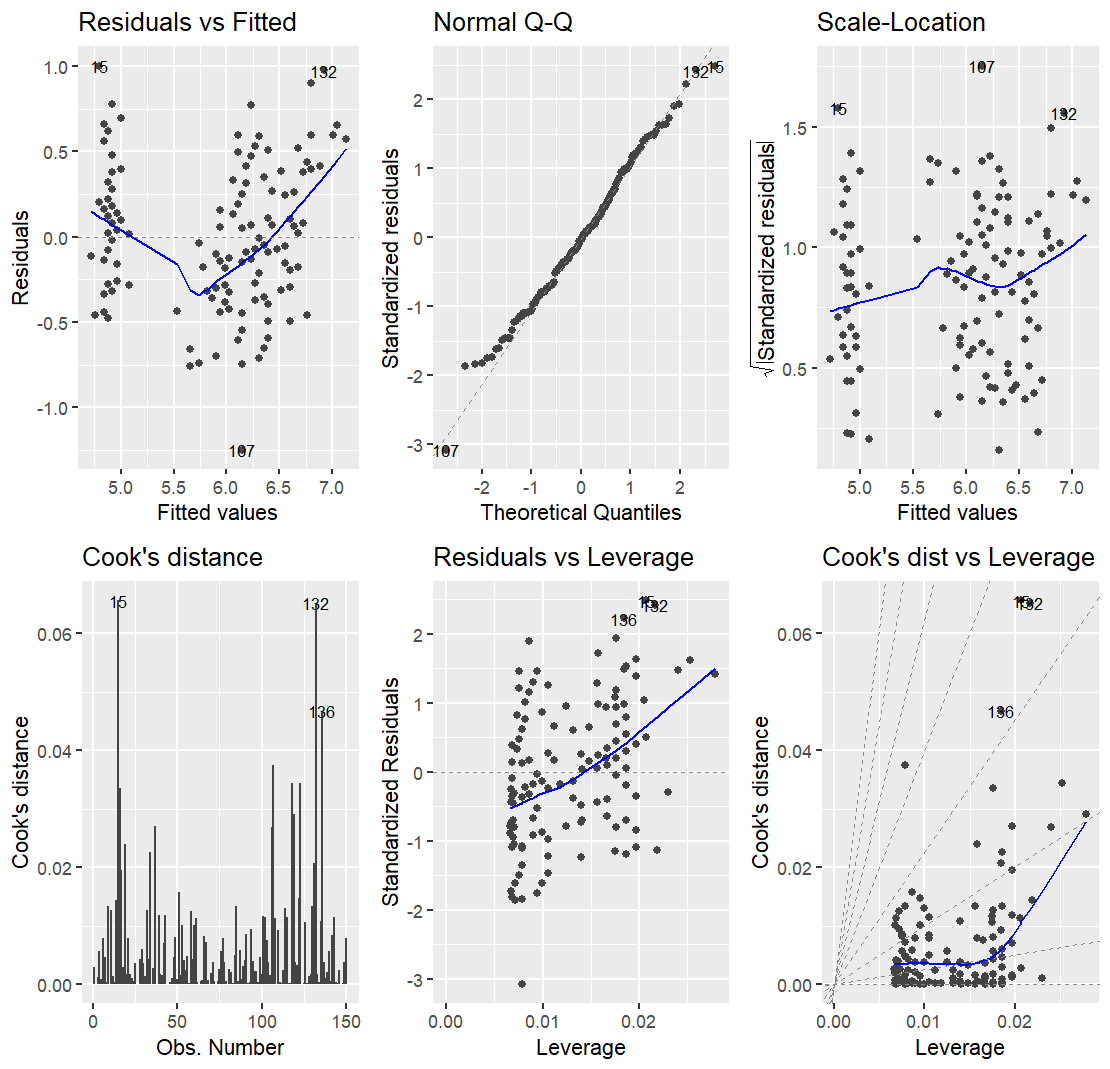
and fitted model

To assess compliance with the modelling assumptions we check the residuals

as a proxy for checking the noise terms (we don’t have these to check directly).

We begin with a visual analysis of the residuals.

1. Produce appropriate diagnostic plots and determine if the assumptions of normality, constant variance and independence appear to have been satisfied **[3 marks]**.



**Normality:** The Q – Q plot seems to have problems since there are residual data points at the tails that do not align with the line

**Constant Variance:** The residual scatter plot shows a constant variance.

**Independence:** The plot does not show any pattern on the residuals, suggesting independence.

We can also check normality of the residuals with a hypothesis test.

1. Using significance level , test if the residuals are normally distributed. Write down the hypotheses, the test statistic and p-value, the test decision (with reason) and a conclusion using a minimum of mathematical language **[3 marks]**.

A black background with white text

Description automatically generated

**Null Hypothesis:** The residuals are normally distributed. **Test statistic:** 0.99298

**p-value:** 0.6767

**Test decision:** Since p-value is greater than α,we don’t have evidence to reject the null hypothesis.

**Conclusion:** The residuals are normally distributed, assumption is match.

The independence assumption can also be investigated with the Durbin-Watson statistic, which assesses the degree of autocorrelation in a data sample. Note that significant autocorrelation would violate the independence assumption, but lack of autocorrelation does not necessarily imply independence.

1. Providing a reason for your answer, determine if there is any statistical evidence that the residuals are not independent **[3 marks]**.

A black screen with white text

Description automatically generated

**Null Hypothesis:** The residuals are not autocorrelated. **Test statistic:** 1.8673

**p-value:** 0.3705

**Test decision:** Since p-value is greater than α,we don’t have evidence to reject the null hypothesis.

**Conclusion:** The residuals are not autocorrelated; the assumption is a match.

1. The coefficient of determination for the model is . Other than via the fitted regression model, how else is this related to the variables in the model **[3 marks]**?

Having a positive slope of 0.3715907, it is possible to know that x and y have a positive relationship, and by calculating R, 0.8717798, we conclude that it is a strong relationship.

In conclusion, it has a strong positive relationship.

We are now going to identify data points with large influence on the estimated regression equation. We do this using Cook’s D statistic.

1. Calculating a relevant statistic, identify any potentially influential points **[3 marks]**.

A diagram with dots and numbers

Description automatically generated with medium confidence

**Critical cooks:** 0.02702

Any point above the red line, the critical point, is potentially influential. The values that are above the line are the following:

A screen shot of a computer screen

Description automatically generated

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

Create a new data excluding the 9 points identified above and re-run the regression on this reduced data set.

1. Calculate the proportional changes in the estimated beta coefficients between the reduced and full data set models **[3 marks]**.

New estimates:

A computer screen shot of a black screen

Description automatically generated

Full dataset estimates:

A computer screen with white text

Description automatically generated

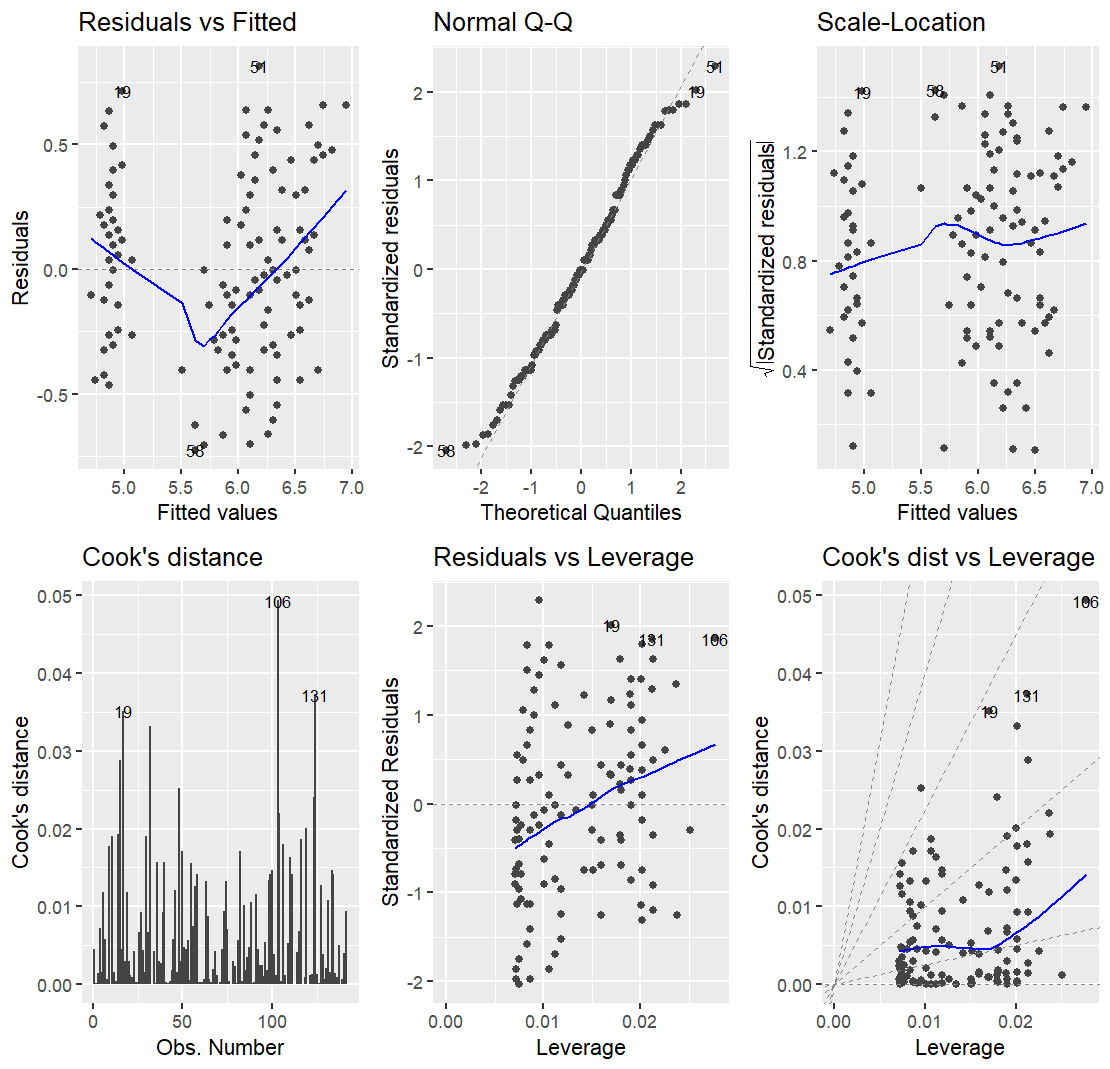
**Intercept change:** -0.02902522%

**Slope change:** -2.235156%

The proportional changes in the estimated beta coefficients are quite small, so the full data set model should be preferred on this basis.

To finish our analysis, we will check the assumptions for the model fitted to the reduced data set.

1. Produce appropriate diagnostic plots and determine if the assumptions of normality, constant variance and independence appear to have been satisfied for the model on the reduced data set **[not assessed]**.



**Normality:** The Q – Q plot seems to have problems since there are residual data points at the tails that do not align with the line.

**Constant Variance:** The residual scatter plot shows a constant variance.

**Independence:** The plot does not show any pattern on the residuals, suggesting independence.

Same results as the full dataset

1. Using significance level , test if the residuals of the reduced data set model are normally distributed. Write down the hypotheses, the test statistic and p-value, the test decision (with reason) and a conclusion using a minimum of mathematical language **[not assessed]**.

A black background with white text

Description automatically generated

**Null Hypothesis:** The residuals are normally distributed. **Test statistic:** 0.98606

**p-value:** 0.1652

**Test decision:** Since p-value is greater than α,we don’t have evidence to reject the null hypothesis.

**Conclusion:** The residuals are normally distributed, assumption is match.

1. Providing a reason for your answer, determine if there any statistical evidence that the residuals of the reduced data set model are not independent **[not assessed]**.

A black screen with white text

Description automatically generated

**Null Hypothesis:** The residuals are not autocorrelated. **Test statistic:** 1.7384

**p-value:** 0.09984

**Test decision:** Since p-value is greater than α,we don’t have evidence to reject the null hypothesis.

**Conclusion:** The residuals are not autocorrelated; the assumption is a match.