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
sl	response	length of flower sepal
pl	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

$$SL = \beta_0 + \beta_1 * pl + \epsilon$$

and fitted model

$$\widehat{sl} = 4.3066 + 0.40892 * pl.$$

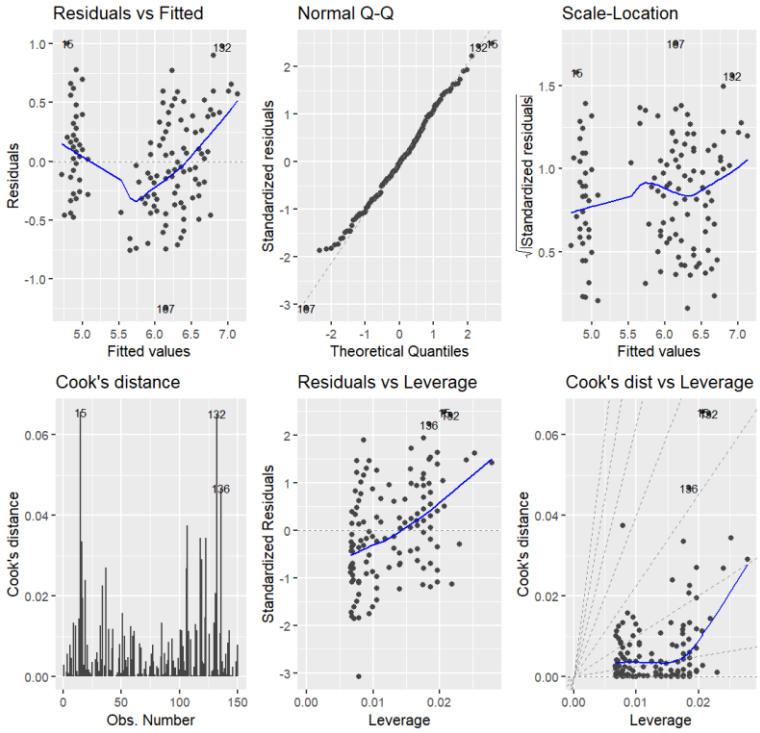
To assess compliance with the modelling assumptions we check the residuals

$$\hat{\epsilon}_i = sl_i - \widehat{sl}_i, \qquad i \in \{1,2, \dots ,150\}.$$

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

We begin with a visual analysis of the residuals.

(a) 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.

(b) Using significance level $\alpha = 0.05$, 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]**.

```
        Shapiro-Wilk normality test

data:  data1$resid
W = 0.99298, p-value = 0.6767
```

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.

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

```
        Durbin-Watson test

data:  model1
DW = 1.8673, p-value = 0.3705
alternative hypothesis: true autocorrelation is not 0
```

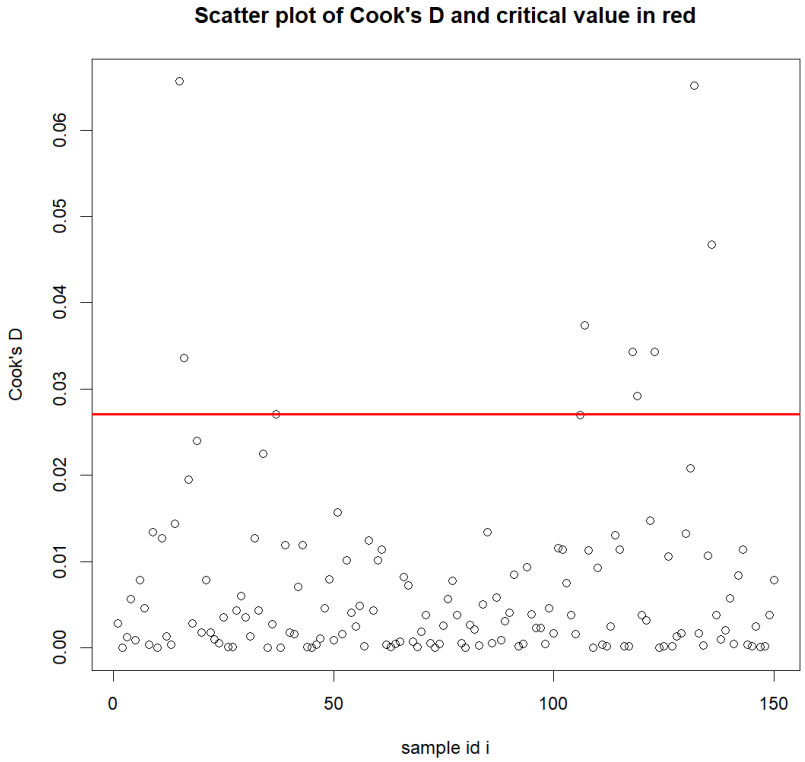
Null Hypothesis: The residuals are not autocorrelated.
Test statistic: 1.8673
p-value: 0.3705
Test decision: Since p-value is greater than α, we don't have evidence to reject the null hypothesis.
Conclusion: The residuals are not autocorrelated; the assumption is a match.

(d) The coefficient of determination for the model is $R^2 = 0.76$. 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.

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



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:

	sl	sw	pl	pw	species	cooksD
15	5.8	4.0	1.2	0.2	setosa	0.06567190
16	5.7	4.4	1.5	0.4	setosa	0.03357130
37	5.5	3.5	1.3	0.2	setosa	0.02706010
107	4.9	2.5	4.5	1.7	virginica	0.03741390
118	7.7	3.8	6.7	2.2	virginica	0.03433850
119	7.7	2.6	6.9	2.3	virginica	0.02916103
123	7.7	2.8	6.7	2.0	virginica	0.03433850
132	7.9	3.8	6.4	2.0	virginica	0.06520640
136	7.7	3.0	6.1	2.3	virginica	0.04677382

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.

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

New estimates:

```
Call:
lm(formula = sl ~ pl, data = data1.reduced)

Residuals:
     Min       1Q   Median       3Q      Max
-0.72463 -0.26505 -0.00502  0.23495  0.81568

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.30535    0.07230   59.55   <2e-16 ***
pl           0.39978    0.01774   22.53   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3575 on 139 degrees of freedom
Multiple R-squared:  0.7851,    Adjusted R-squared:  0.7835
F-statistic: 507.7 on 1 and 139 DF,  p-value: < 2.2e-16
```

Full dataset estimates:

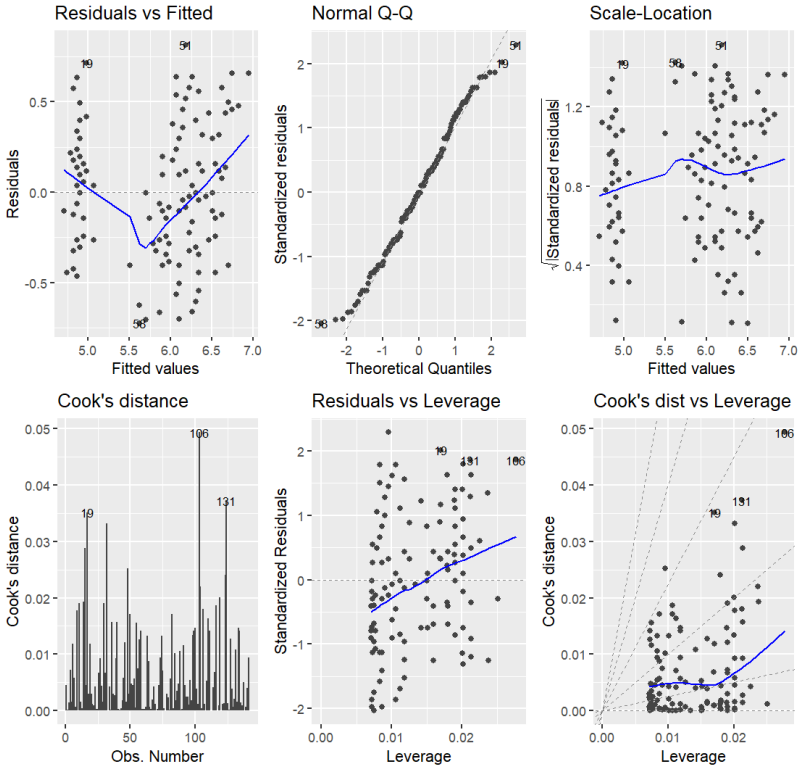
```
Call:
lm(formula = sl ~ pl, data = data1)

Residuals:
     Min       1Q   Median       3Q      Max
-1.24675 -0.29657 -0.01515  0.27676  1.00269

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.30660    0.07839   54.94   <2e-16 ***
pl           0.40892    0.01889   21.65   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4071 on 148 degrees of freedom
Multiple R-squared:   0.76,     Adjusted R-squared:  0.7583
F-statistic: 468.6 on 1 and 148 DF,  p-value: < 2.2e-16
```

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.

(g) 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

(h) Using significance level $\alpha = 0.05$, 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]**.

```
        Shapiro-Wilk normality test

data:  model1.reduced$resid
W = 0.98606, p-value = 0.1652
```

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.

(i) 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]**.

```
        Durbin-Watson test

data:  model1.reduced
DW = 1.7384, p-value = 0.09984
alternative hypothesis: true autocorrelation is not 0
```

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.