The University of Sydney

STAT5002 Introduction to Statistics

Semester 1 **Tutorial 9** 2018

1. The effect of height (H) and weight (W) on catheter length (L) on $n = 12$ children with congenital heart disease was analyzed with a multiple regression model and some results are presented after the computer problems. Use this output to answer the following.

(a) Write down the fitted multiple regression model for this situation. Include an estimate for the error variance.

(b) Calculate a 90% confidence interval for the true coefficient of H. What can you conclude from your interval?

(c) What is the multiple correlation coefficient between L and (H, W)?

(d) Are there any high leverage points and/or any outlier in the data?

(e) Comment on the model diagnostic plots.

(f) Fit a simple linear regression of L on W and comment on the model diagnostic plots.

(g) Calculate the sample correlation coefficient between the L and W values.

(h) Compare the R^2 for the two regression models.

(i) What is the prediction of the expected catheter length when the weight is 90 under the simple linear regression model of L on W?

Output for Problem 1

```
H = c(42.8, 63.5, 37.5, 39.5, 45.5, 38.5, 43, 22.5, 37, 23.5,
    33, 58)
W = c(40, 93.5, 35.5, 30, 52, 17, 38.5, 8.5, 33, 9.5, 21, 79)
L = c(37, 49.5, 34.5, 36, 43, 28, 37, 20, 33.5, 30.5, 38.5, 47)
dat = data.frame(H, W, L)
dat

##       H    W    L
## 1  42.8 40.0 37.0
## 2  63.5 93.5 49.5
## 3  37.5 35.5 34.5
## 4  39.5 30.0 36.0
## 5  45.5 52.0 43.0
## 6  38.5 17.0 28.0
## 7  43.0 38.5 37.0
## 8  22.5  8.5 20.0
## 9  37.0 33.0 33.5
## 10 23.5  9.5 30.5
## 11 33.0 21.0 38.5
## 12 58.0 79.0 47.0

M1 = lm(L ~ H + W, data = dat)
summary(M1)

##
## Call:
## lm(formula = L ~ H + W, data = dat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.048 -1.258 -0.259  1.899  7.004
```

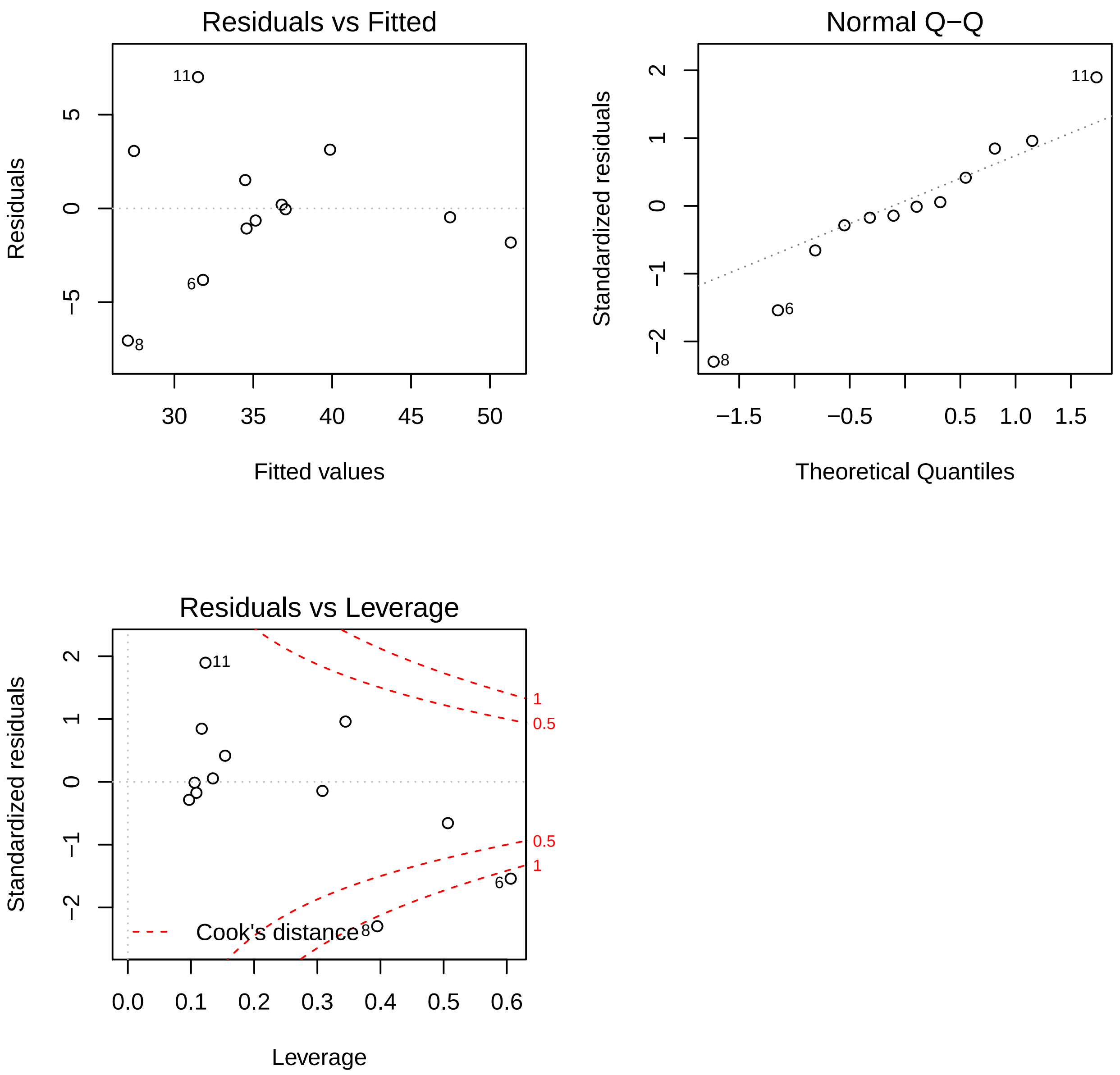
```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  21.0084     8.7512   2.401   0.0399 *
## H             0.1964     0.3606   0.545   0.5993
## W             0.1908     0.1652   1.155   0.2777
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.943 on 9 degrees of freedom
## Multiple R-squared:  0.8053,Adjusted R-squared:  0.7621
## F-statistic: 18.62 on 2 and 9 DF,  p-value: 0.0006336
```

```
par(mfrow = c(2, 2))
plot(M1, which = c(1, 2, 5), add.smooth = FALSE)
round(lm.influence(M1)$h, 3)
```

```
##     1     2     3     4     5     6     7     8     9    10
## 0.106 0.507 0.109 0.154 0.117 0.606 0.135 0.395 0.097 0.345
##    11    12
## 0.123 0.308
```

```
round(cooks.distance(M1), 3)
```

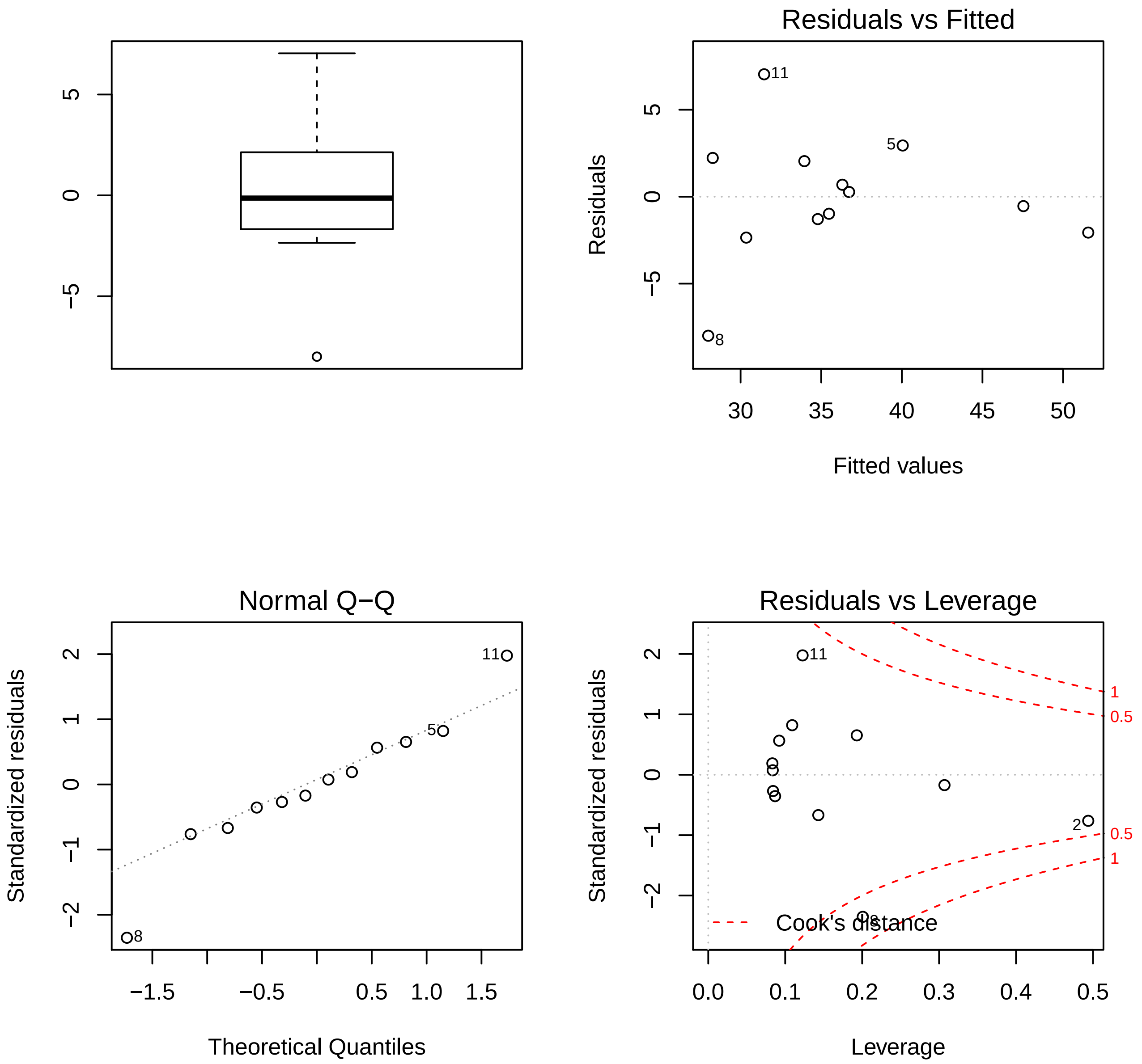
```
##     1     2     3     4     5     6     7     8     9    10
## 0.000 0.148 0.001 0.011 0.032 1.218 0.000 1.149 0.003 0.162
##    11    12
## 0.168 0.003
```



```
M2 = lm(L ~ 1 + W, data = dat)
summary(M2)

##
## Call:
## lm(formula = L ~ 1 + W, data = dat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.994 -1.481 -0.135  2.091  7.040
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.63746    2.00421  12.792 1.60e-07 ***
## W            0.27727    0.04399   6.303 8.87e-05 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.802 on 10 degrees of freedom
## Multiple R-squared:  0.7989,Adjusted R-squared:  0.7788
## F-statistic: 39.73 on 1 and 10 DF,  p-value: 8.871e-05

par(mfrow = c(2, 2))
boxplot(M2$residuals)
plot(M2, which = c(1, 2, 5), add.smooth = FALSE)
```



2. In an experiment to determine the source from which corn plants in various soils obtain their phosphorous, the concentration of inorganic phosphorous (x_1) and of two types of organic phosphorous (x_2, x_3) in the soil, and also the phosphorous content (y) of the plants, were measured. The data are

```
x1 = c(0.4, 0.4, 3.1, 0.6, 4.7, 1.7, 9.4, 10.1, 11.6, 12.6, 13.8,
    10.9, 23.1, 29.9)
x2 = c(53, 23, 19, 34, 24, 65, 44, 31, 29, 58, 55, 37, 46, 51)
x3 = c(158, 163, 37, 157, 59, 123, 46, 117, 173, 112, 117, 111,
    114, 124)
y = c(64, 60, 71, 61, 54, 77, 81, 93, 93, 51, 60, 76, 96, 99)
```

(a) Create a data-frame `dat`, produce the correlation matrix (`cor`), and have a look at the pairwise scatterplots (`pairs`). Comment on these in terms of obvious outliers, shape of plots and homoscedasticity.

(b) Use `lm()` to fit the regression model

$$Y_i = \beta_0 + \beta_1\ x_{i1} + \beta_2\ x_{i2} + \beta_3\ x_{i3} + \epsilon_i\,, \quad \epsilon_i \sim NID(0, \sigma^2),$$

show the summary of the `lm()` output and give the fitted value and residual for the first observation.

(c) Calculate a 95% confidence interval for β_3 and give all plausible values to 1dp.

(d) What is the square of the multiple correlation coefficient for this model?

(e) What observation has fourth largest leverage and are there any high leverage points?

(f) Using the `lm()` output it seems that x_3 can be dropped from the model. It also seems that x_2 can be dropped. Why?

(g) Determine a reasonable model with x_1 alone for predicting the expected phosphorous content of the plants. What is the R^2 value for your final model? How does this compare with your answer in (d)?

(h) For the regression of phosphorous content on inorganic phosphorous only and calculate a prediction for $Y|x_1 = 40$ based on this model.

(i) Show that for the full model the largest Cook's distance is 0.63 and for the simple linear regression model 0.19 (2dp). Why are the values different?

3. Given is the linear regression model $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $i = 1\ldots, n \geq 2$. Because of convenience the scale of the x values is changed (e.g. from inches to centimeters) and the transformed explanatory values $z = x/c$, where c is a constant, are used instead. Write the new model as

$$Y_i = \gamma_0 + \gamma_1 z_i + \varepsilon_i.$$

(a) Represent estimates of γ_0 and γ_1 in terms of $\hat{\beta}_0$ and $\hat{\beta}_1$.

(b) Show that r^2 is invariant.

4. In an experiment to investigate the amount of a drug retained in the liver of a rat, 19 rats were randomly selected, weighed, placed under light ether anesthesia and given an oral dose of the drug. The dose an animal received was determined as approximately 40mg of the drug per kilogram of body weight, since liver weight is known to be strongly related to body weight and it was felt that large livers would absorb more of a given dose than smaller livers. After a fixed length of time each rat was sacrificed, the liver weighed, and the percent of the dose in the liver determined.

The experimental hypothesis was that, for the method of determining the dose, there is no relationship between the percentage of the dose in the liver (Y) and the body weight (x_1), liver weight (x_2), and relative dose (x_3).

The data (Weisberg, S. *Applied Linear Regression* (1980)) are given in `ratliver.txt`.

(a) Consider the model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i, \quad \epsilon_i \sim NID(0, \sigma^2).$$

Give the fitted least squares multiple regression equation.

(b) Predict the expected Y value when $x_1 = 165$, $x_2 = 8.0$ and $x_3 = 0.85$.

(c) Can any of the explanatory variables be dropped from the model?

(d) Determine if there are any high leverage points. Check if the x-values of any high leverage points are extreme.

(e) Are there any outliers?

(f) Refit the model ignoring the outlier. Determine the square of the multiple correlation coefficient. Comment.

(g) Calculate 95% confidence intervals for the parameter estimates in the model. Do any of the intervals contain 0?

5. In order to compare four different brands of golf balls, five balls from each brand are placed in a machine that exerts the force produced by a 200 metre drive. The number of simulated drives needed to crack or chip each ball is recorded below.

A	B	C	D
281	270	218	364
220	334	244	302
274	307	225	325
242	290	273	337
251	331	249	355

Consider the model

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

where $\epsilon_{ij} \sim N(0, \sigma^2)$ and Y_{ij} denotes the jth observation on brand i. Which brand performs best under this model?

6. The data `tooth.txt` contains results from an (old) experiment into the effects of vitamin C on tooth growth and will be used in Lecture 17. Thirty guinea pigs were divided (at random) into three groups of ten and dosed with vitamin C (administered in orange juice). Group 1 dose was low (0.5mg vit. C), group 2 dose was medium (1mg vit. C) and group 3 dose was high (2mg vit. C). Length of odontoblasts (teeth) measured as response variable.

(a) Store the data `tooth.txt` in a data frame called `dat`. Explore the structure of the data by using both, `str(dat)` and `summary(dat)`.

(b) Produce a boxplot of `Length` by `Dose.fac` and comment on the homoscedasticity assumption of the errors.

(c) Consider fitting the dosage group as numerical value in one model and factor in another model. Which one is more appropriate?