THE UNIVERSITY OF AUCKLAND

SEMESTER ONE 2018
Campus: SOUTHWEST UNIVERSITY, CHONGQING, CHINA

STATISTICS

Data Analysis

(Time allowed: TWO Hours)

INSTRUCTIONS

- Attempt **all** questions.

1. This question refers to the **Auckland Rental Data** in **Appendix A**.
[Total: 23 marks]

   a. Using the pairs plot, comment on the most important features of the data. Limit your comments to three or four separate features.
   [3 marks]

   b. Three models have been fitted: `rent1.fit`, `rent2.fit`, and `rent3.fit`. Explain the differences between them, and why these changes were made.
   [4 marks]

c. The variable `beds.F` is not significant in model `rent2.fit`. However, it is significant in the model `rent4.fit`, at the end of **Appendix A**. Explain why.

[3 marks]

d. Provide an equation for the model `rent3.fit`, stating its assumptions.

[4 marks]

e. Interpret the effect of each of the variables remaining in the model `rent3.fit` on the rental price.

[9 marks]

2. This question refers to the **Lobster Survival Data** in **Appendix B**.
[Total: 14 marks]

   a. The model `lobster1.fit` was fitted by the biologists, and appeared in a paper they published. Why is this model **NOT** appropriate?
   [3 marks]

   b. What is a name given to the type of model fitted in `lobster2.fit`?
   [1 marks]

   c. Provide the equation of the model `lobster2.fit`, stating its assumptions.
   [3 marks]

d. Do you think the model `lobster2.fit` is appropriate? Explain why, or why not, with reference to its assumptions.

[4 marks]

e. Interpret the effect of lobster size on survival, using the model `lobster2.fit`.

[3 marks]

3. This question refers to the **Southwest University Test Data** in **Appendix C**.

[Total: 13 marks]

   a. Provide an equation for the model `test.fit`, stating its assumptions.

   [4 marks]

   b. Inspect the output from the `anova()` function used in **Appendix C**. What is the null hypothesis associated with the p-value in the final column? What do you conclude from this?

   [3 marks]

c. Compare the p-value in the `anova()` output to the p-values in the `summary()` output. They appear to contradict each other: the `anova()` p-value is significant (i.e., it is greater than 0.05), but one of the p-values in the `summary()` table is not. Explain this apparent contradiction.

[2 marks]

d. The confidence intervals in the output from `confint()` appear to be different to the confidence intervals in the output from `multipleComp()`. Explain why.

[2 marks]

e. From the `summary()` and `confint()` output, the lecturer concludes that there is evidence to suggest that, on average, Class 2 had better students than Class 1. On average, we expect Class 2 students to score between 0.07 and 5.77 marks higher than Class 1 students. Do you agree with this conclusion? Explain why, or why not.

[2 marks]

4. This section contains multiple-choice questions.

[Total: 10 marks]

- Answer **ALL** questions.
- For each question, select the **ONE** answer by circling the number it is next to.
- If you give more than one answer to any question, you will receive zero marks for that question.
- If you wish to change your answer, make it very obvious what your final answer is.
- Each question is worth two marks.
- Each question has a single correct answer.

a. Which of the following is **NOT** an assumption of a generalised linear model with a Poisson response distribution? Let μ_i be the expectation of the response variable for the ith observation.

(1) The variance of the response for the ith observation is μ_i.

(2) The observations are independent of one another.

(3) Each observation's response variable has a Poisson distribution.

(4) $\log[(\mu_i/(1-\mu_i)]$ is a linear combination of the explanatory variables.

b. Watson is a dog who likes to eat sausages. Ben asks Watson to do five tricks. For each trick he successfully completes, Ben gives Watson a sausage. Which of the following distributions is most appropriate for the number of sausages that Watson receives?

(1) Chi-squared.

(2) Binomial.

(3) Poisson.

(4) Normal.

c. A generalised linear model was fitted with the `glm()` function in R, using the argument `family = "quasibinomial"`. Which of the following statements **IS** an assumption of the model? Let Y_i be the number of succesful trials for the ith observation, out of a total of n_i trials. The expected number of succesful trials is $n_i p_i$.

(1) $\text{Var}(Y_i) = n_i p_i (1 - p_i)$.

(2) Constant variance of the response across all observations.

(3) $\log[p_i/(1-p_i)]$ is a linear combination of the explanatory variables.

(4) Y_i has a binomial distribution.

d. Which of the following statements about the analysis of a two-by-two contingency table is **TRUE**?

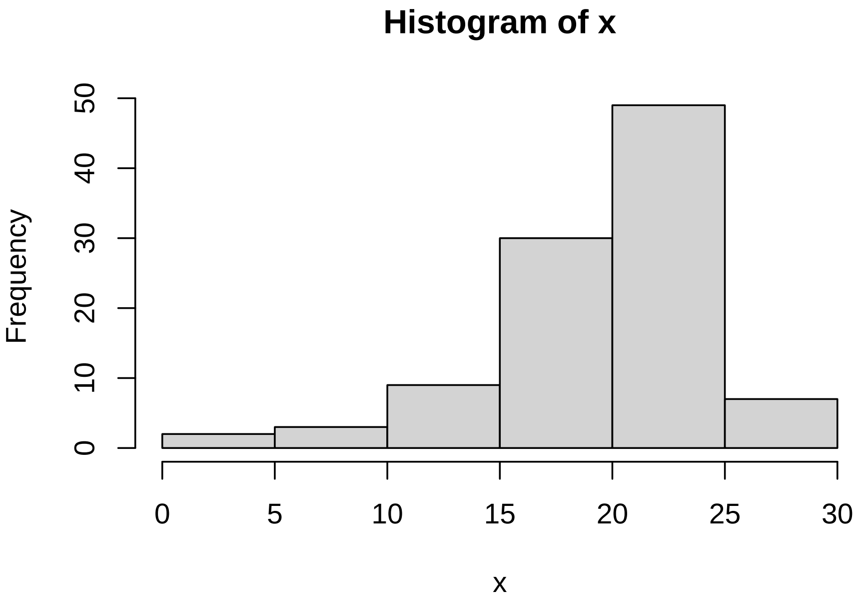
(1) The null hypothesis of the Chi-squared test is that there is an association between the two categorical variables.

(2) We compute the p-value from a Chi-squared test by comparing the Chi-squared test statistic to an F-distribution.

(3) If the expected value of at least one cell is less than five, then we should use Fisher's exact test instead of a Chi-squared test.

(4) If the observed counts are close to the expected counts, then we would expect the Chi-squared test statistic to be large.

e. In total, 100 observations of some variable x were recorded, and are plotted in the histogram below. Which statement most accurately describes the relationship between the mean, $\bar{x}$, and the median, $\tilde{x}$?



(1) $\bar{x} \approx \tilde{x}$

(2) $\bar{x} > \tilde{x}$

(3) $\bar{x}$ and $\tilde{x}$ are both undefined.

(4) $\bar{x} < \tilde{x}$

APPENDICES BOOKLET FOLLOWS

APPENDICES BOOKLET

CONTENTS

Auckland Rental Data

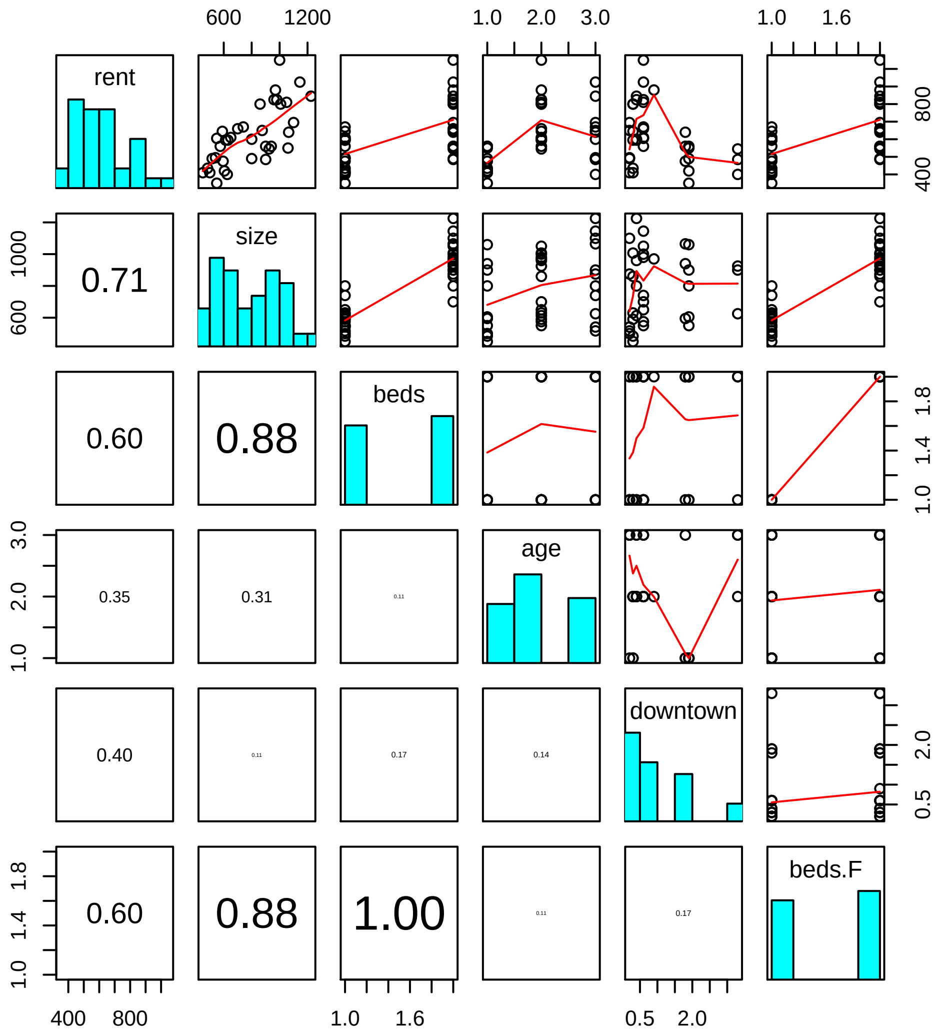
Data were collected on the monthly rental and other characteristics of 36 randomly selected apartments in Auckland. We wish to build a model to explain the monthly rental of an apartment. The variables measured were

`rent`	The monthly rental (in NZ$)
`size`	The apartment size (in square feet)
`beds`	The number of bedrooms (either 1 or 2)
`age`	The age of the apartment building (`new`, `recent`, or `old`)
`downtown`	The distance from the city centre (in miles)

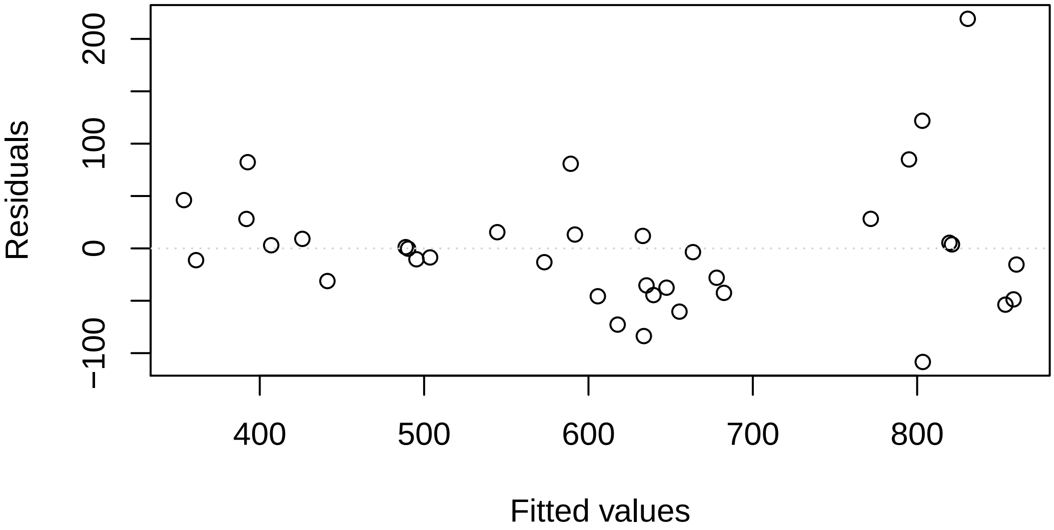
```
> ## Printing the first six observations.
> head(rent.df)
  rent size beds age downtown
1  810 1050    2 Old      0.6
2  560  575    1 Old      0.6
3  550 1060    2 New      1.9
4  610  650    1 Old      0.6
5  800 1007    2 Old      0.3
6  435  484    1 New      0.3
```

```
> rent.df = within(rent.df, {
+     beds.F = factor(beds)
+ })
> summary(rent.df$age)
   New    Old Recent 
    10     15     11 
> summary(rent.df$size)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  450.0   593.8   800.0   791.0   972.5  1225.0 
> summary(rent.df$downtown)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.2000  0.3000  0.6000  0.9861  1.8000  3.3000 
```

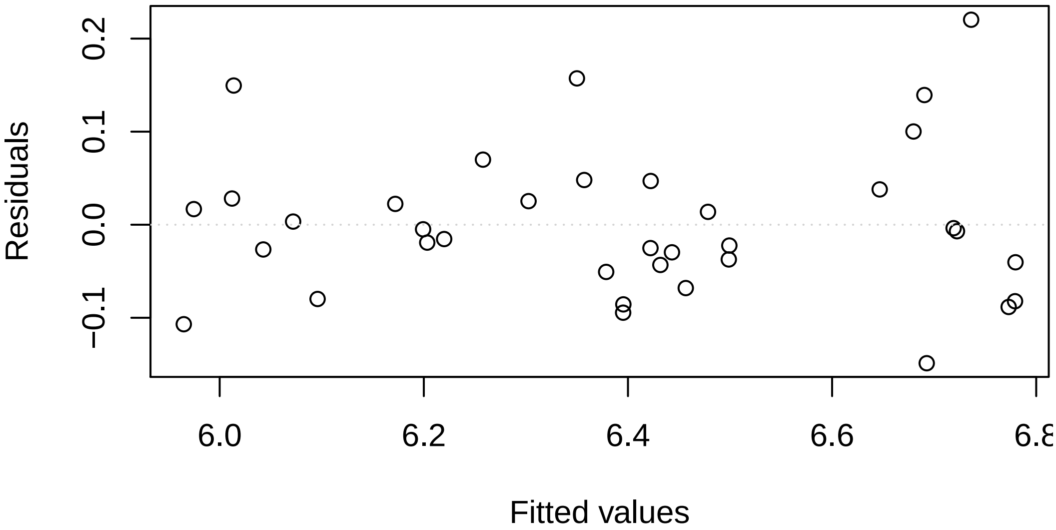
```
> pairs20x(rent.df)
```



```
> rent1.fit=lm(rent ~ size + downtown + beds.F + age,
+              data = rent.df)
> eovcheck(rent1.fit)
```

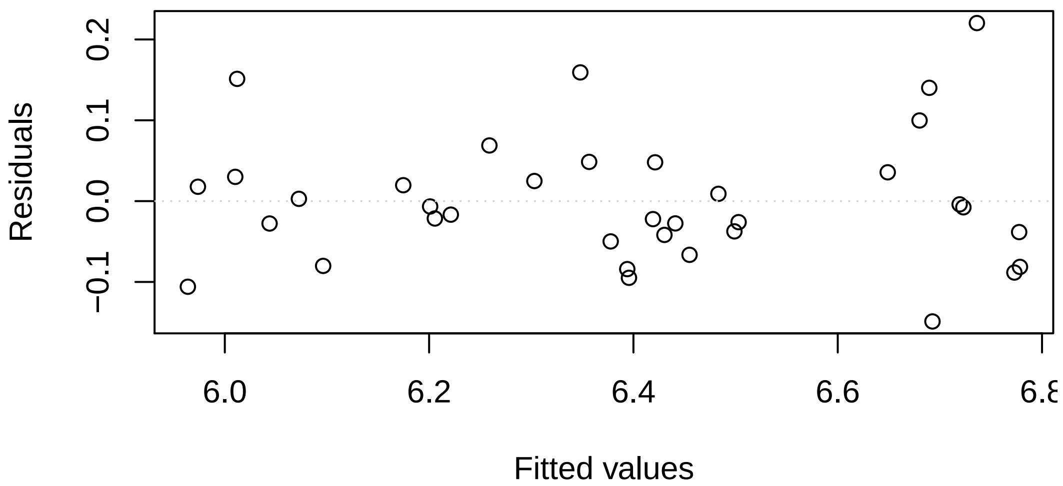



```
> rent2.fit=lm(log(rent) ~ size + downtown + beds.F + age,
+              data = rent.df)
> eovcheck(rent2.fit)
```

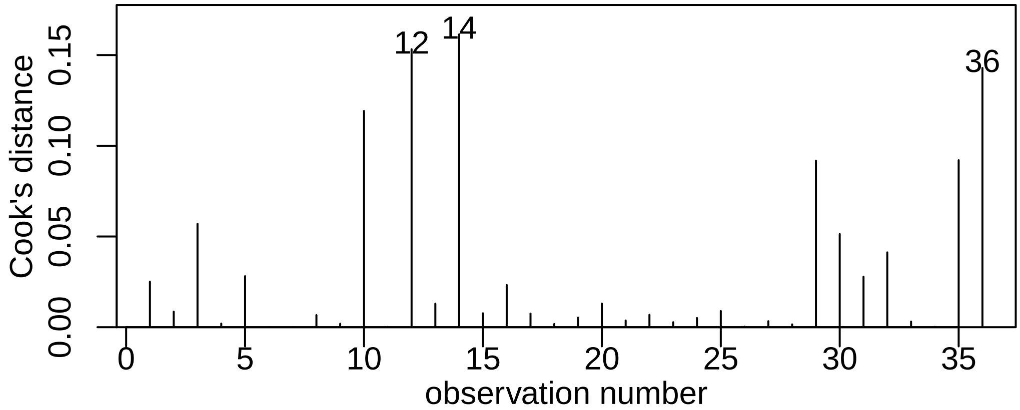


```
> summary(rent2.fit)

Call:
lm(formula = log(rent) ~ size + downtown + beds.F + age, data = rent.df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.14881 -0.04509 -0.01117  0.03061  0.22029

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.6866876  0.0919170  61.868  < 2e-16 ***
size         0.0008595  0.0001548   5.551 4.92e-06 ***
downtown    -0.1024085  0.0168748  -6.069 1.15e-06 ***
beds.F2     -0.0076631  0.0657676  -0.117   0.9080
ageOld       0.2591978  0.0385543   6.723 1.89e-07 ***
ageRecent    0.0887542  0.0430398   2.062   0.0479 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08653 on 30 degrees of freedom
Multiple R-squared:  0.9104,Adjusted R-squared:  0.8955
F-statistic: 60.98 on 5 and 30 DF,  p-value: 8.289e-15
```

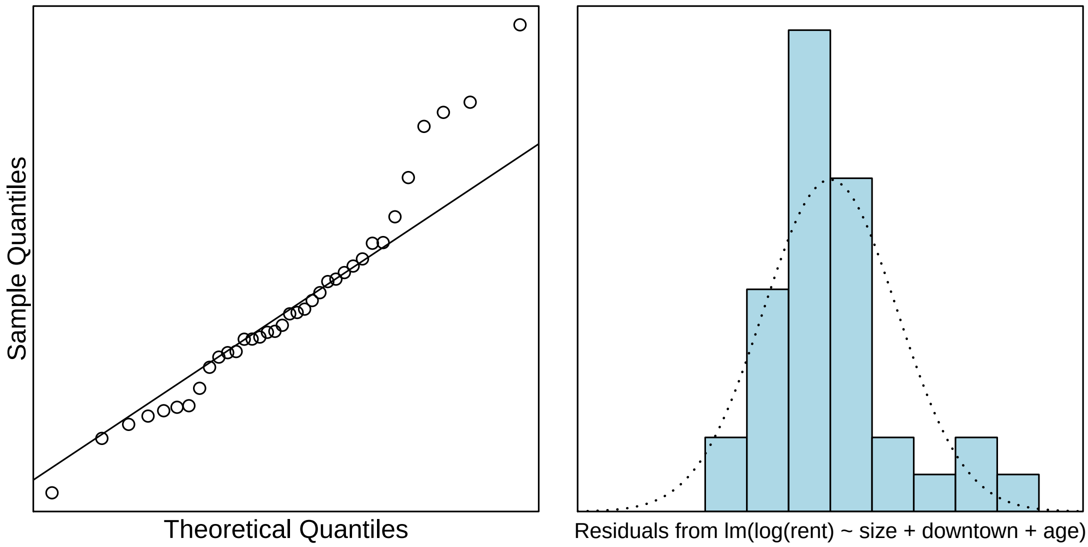
```
> rent3.fit = lm(log(rent) ~ size + downtown + age, data = rent.df)
> summary(rent3.fit)

Call:
lm(formula = log(rent) ~ size + downtown + age, data = rent.df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.14880 -0.04378 -0.01215  0.03145  0.22034

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.695e+00  5.609e-02 101.540  < 2e-16 ***
size         8.434e-04  6.983e-05  12.078 2.94e-13 ***
downtown    -1.027e-01  1.640e-02  -6.264 5.80e-07 ***
ageOld       2.593e-01  3.793e-02   6.837 1.16e-07 ***
ageRecent    9.038e-02  4.005e-02   2.256   0.0312 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08514 on 31 degrees of freedom
Multiple R-squared:  0.9104,Adjusted R-squared:  0.8988
F-statistic: 78.73 on 4 and 31 DF,  p-value: 8.741e-16
> eovcheck(rent3.fit)
```



```
> cooks20x(rent3.fit)
```




```
> normcheck(rent3.fit, main = "")
```



```
> exp(confint(rent3.fit))
                  2.5 %      97.5 %
(Intercept) 265.2570610 333.4455396
size          1.0007013   1.0009863
downtown      0.8727017   0.9330716
ageOld        1.1995592   1.4002531
ageRecent     1.0087282   1.1877668
```

```
> 100 * (exp(confint(rent3.fit)[2:5, ]) - 1)
                 2.5 %      97.5 %
size        0.07012659  0.09863448
downtown  -12.72982697 -6.69283734
ageOld     19.95592060 40.02531162
ageRecent   0.87282378 18.77668023
```

```
> ## For a 100-unit change in size.
> 100 * (exp(100 * confint(rent3.fit)[2, ]) - 1)
   2.5 %   97.5 %
 7.26176 10.36092
```

For Question (c) Only

```
> rent4.fit = lm(log(rent) ~ beds.F, data = rent.df)
> summary(rent4.fit)

Call:
lm(formula = log(rent) ~ beds.F, data = rent.df)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.36894 -0.21071 -0.01347  0.16764  0.40800 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6.22687    0.05227 119.131  < 2e-16 ***
beds.F2      0.32167    0.07195   4.471 8.24e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2155 on 34 degrees of freedom
Multiple R-squared:  0.3702,Adjusted R-squared:  0.3517 
F-statistic: 19.99 on 1 and 34 DF,  p-value: 8.243e-05
```

Lobster Survival Data

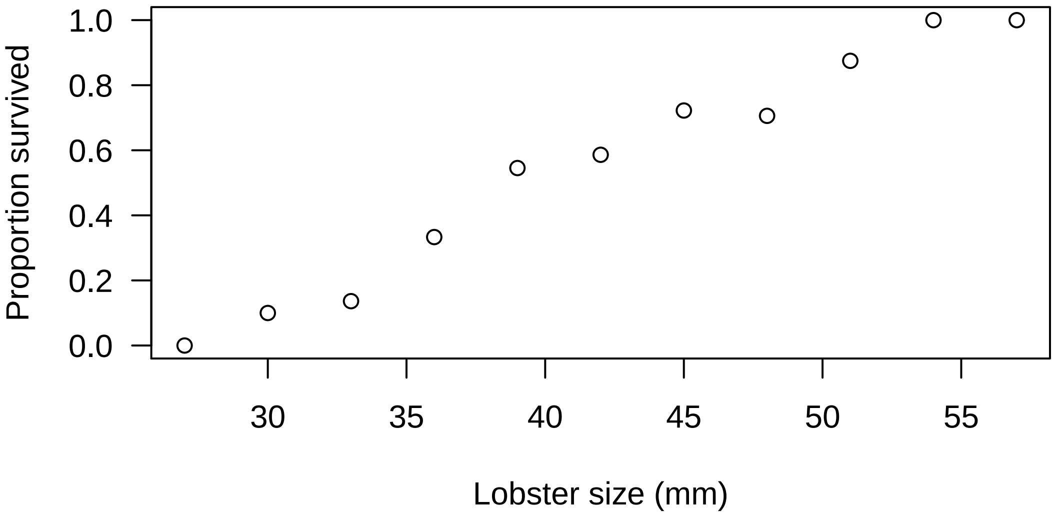
Biologists collected data to investigate how a lobster's size affects its survival. In total, they collected 159 juvenile lobsters from their natural habitat, and measured their size. They tethered the lobsters to the ocean floor for one night. Any lobsters that were missing were assumed to have been eaten by a predator. The surviving lobsters were released.

The variables in the data set are

`size`	Lobster length, measured to the nearest 3 mm
`n`	The number of lobsters of a particular length
`survived`	The number of lobsters of a particular length that survived

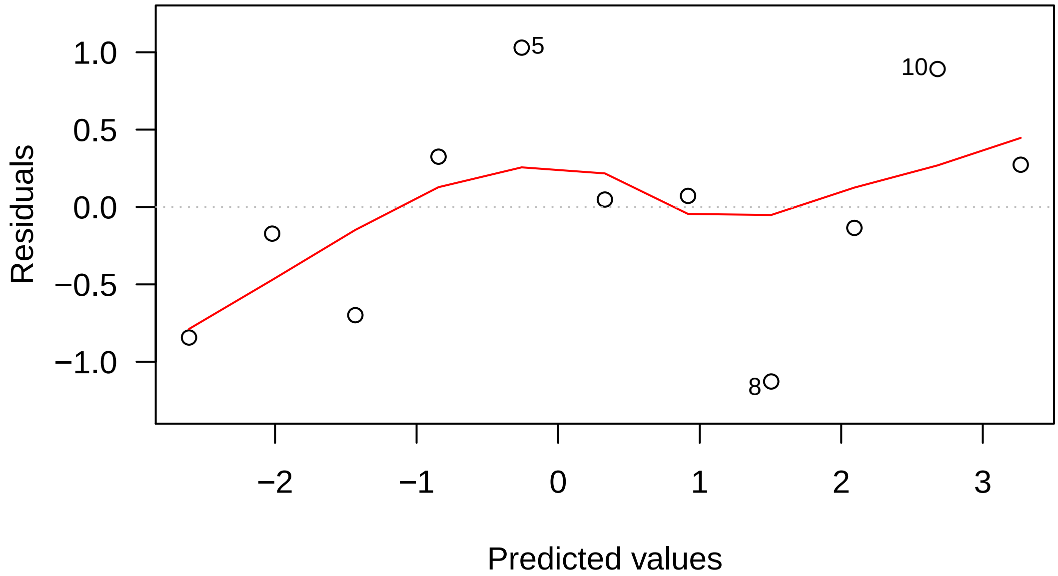
```
> lobster.df = within(lobster.df, {
+     p = survived/n
+ })
> lobster.df
   size  n survived         p
1    27  5        0 0.0000000
2    30 10        1 0.1000000
3    33 22        3 0.1363636
4    36 21        7 0.3333333
5    39 22       12 0.5454545
6    42 29       17 0.5862069
7    45 18       13 0.7222222
8    48 17       12 0.7058824
9    51  8        7 0.8750000
10   54  6        6 1.0000000
11   57  1        1 1.0000000
```

```
> plot(p ~ size, data = lobster.df,
+      xlab = "Lobster size (mm)",
+      ylab = "Proportion survived")
```




```
> lobster1.fit = lm(p ~ size, data = lobster.df)
> summary(lobster1.fit)

Call:
lm(formula = p ~ size, data = lobster.df)

Residuals:
      Min        1Q    Median        3Q       Max
-0.089376 -0.036212  0.000887  0.033829  0.106301

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.948038   0.086867  -10.91 1.72e-06 ***
size         0.035569   0.002017   17.63 2.75e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06348 on 9 degrees of freedom
Multiple R-squared:  0.9719,Adjusted R-squared:  0.9687
F-statistic: 310.8 on 1 and 9 DF,  p-value: 2.752e-08
```

```
> lobster2.fit = glm(p ~ size, family = "binomial", weights = n, data = lobs
> summary(lobster2.fit)

Call:
glm(formula = p ~ size, family = "binomial", data = lobster.df,
    weights = n)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.12729  -0.43534   0.04841   0.29938   1.02995

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.89597    1.38501  -5.701 1.19e-08 ***
size         0.19586    0.03415   5.735 9.77e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 52.1054  on 10  degrees of freedom
Residual deviance:  4.5623  on  9  degrees of freedom
AIC: 32.24

Number of Fisher Scoring iterations: 4


> plot(lobster2.fit, which = 1)
```




```
> lobster2.fit$deviance
[1] 4.562321
> lobster2.fit$df.residual
[1] 9
> 1 - pchisq(lobster2.fit$deviance, lobster2.fit$df.residual)
[1] 0.8706732
> confint(lobster2.fit)


Waiting for profiling to be done...


                  2.5 %     97.5 %
(Intercept) -10.8034921 -5.3449644
size          0.1329987  0.2675871
> exp(confint(lobster2.fit))


Waiting for profiling to be done...


                   2.5 %      97.5 %
(Intercept) 2.032839e-05 0.004772121
size        1.142249e+00 1.306807434
```

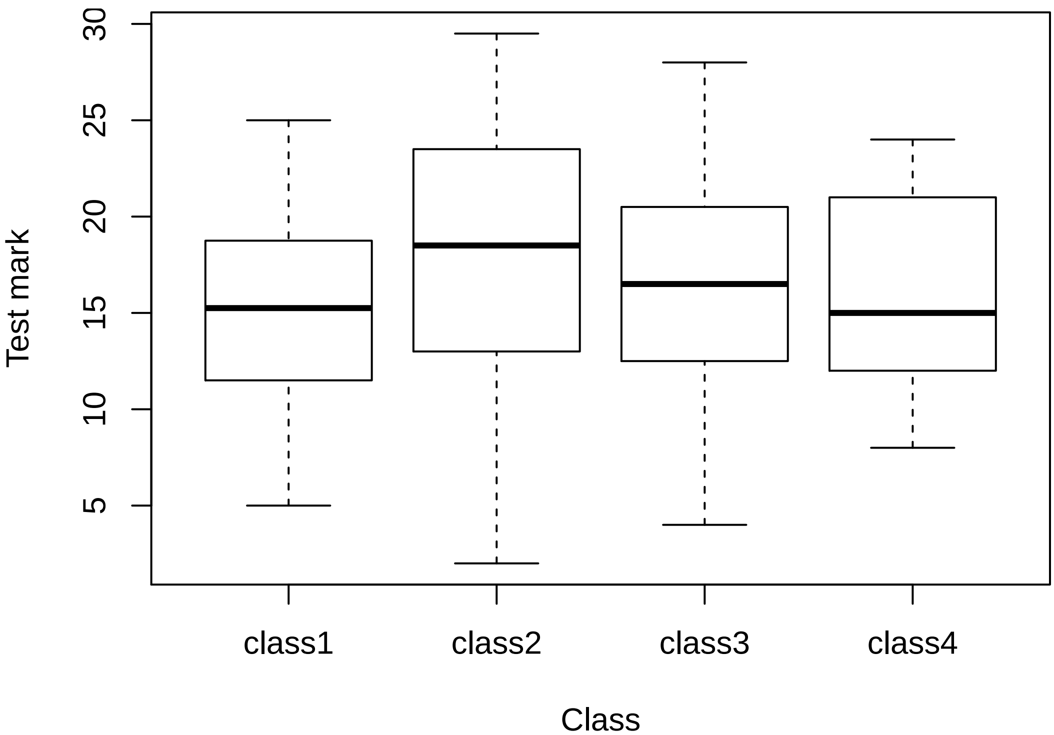
Southwest University Test Data

STATS 201 students at Southwest University completed a mid-semester test on 16 April 2018. Each student in the course belongs to one of four 'classes'. The lecturer was interested in whether or not some classes have better students than others, on average.

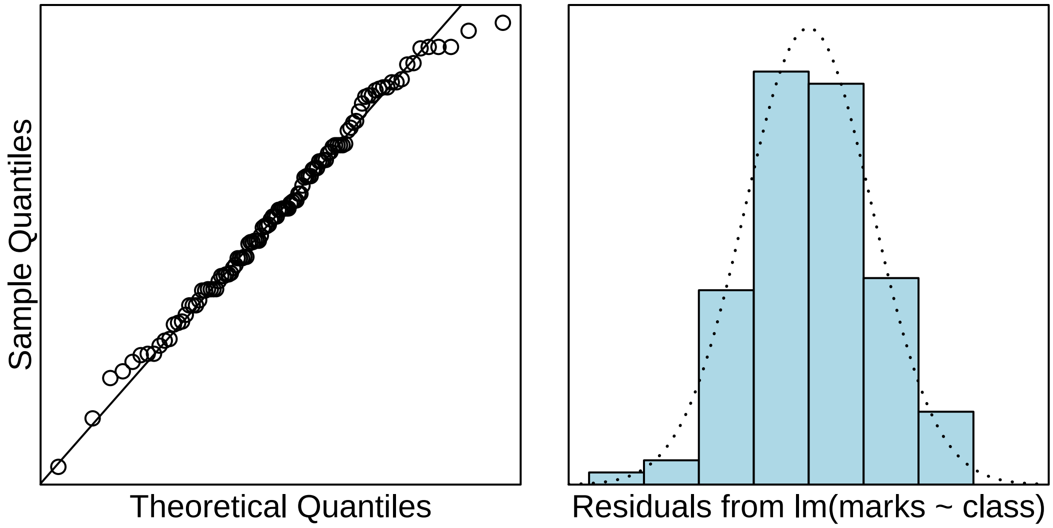
The variables in the data set are

`marks`	The student's score on the test
`class`	The student's class; either `class1`, `class2`, `class3`, or `class4`

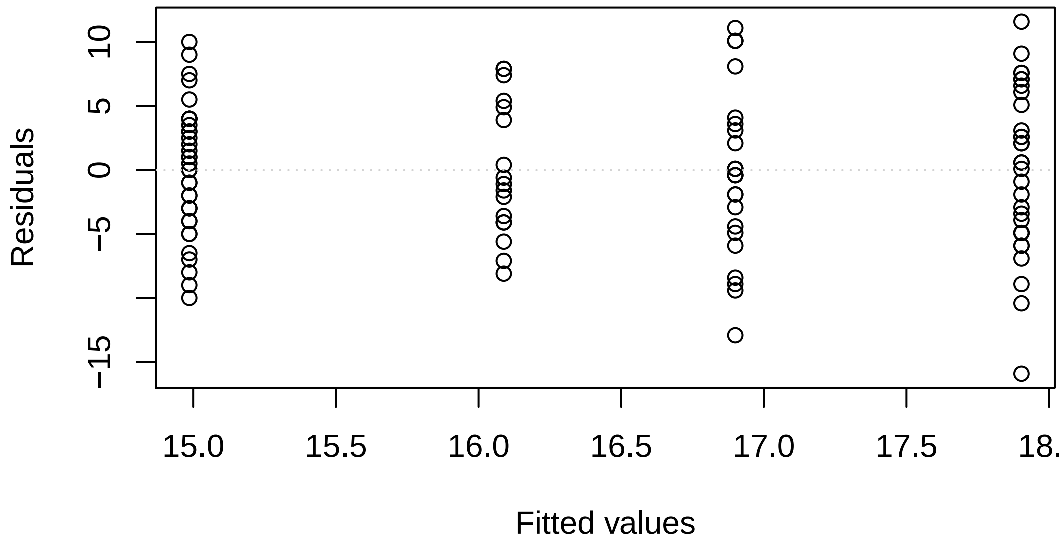
```
> boxplot(marks ~ class, data = test.df, xlab = "Class",
+         ylab = "Test mark")
```



```
> test.fit = lm(marks ~ class, data = test.df)
> normcheck(test.fit)
```




```
> eovcheck(test.fit)
```



```
> anova(test.fit)
Analysis of Variance Table

Response: marks
           Df Sum Sq Mean Sq F value Pr(>F)
class       3  149.8  49.947  1.4523 0.2319
Residuals 105 3611.1  34.391


> summary(test.fit)

Call:
lm(formula = marks ~ class, data = test.df)

Residuals:
     Min       1Q   Median       3Q      Max
-15.9032  -4.0882   0.0139   4.0139  11.5968

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.9861     0.9774  15.333   <2e-16 ***
classclass2   2.9171     1.4369   2.030   0.0449 *
classclass3   1.9139     1.5267   1.254   0.2128
classclass4   1.1021     1.7258   0.639   0.5245
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.864 on 105 degrees of freedom
Multiple R-squared:  0.03984,Adjusted R-squared:  0.01241
F-statistic: 1.452 on 3 and 105 DF,  p-value: 0.2319


> confint(test.fit)
                  2.5 %    97.5 %
(Intercept) 13.04810861 16.924114
classclass2  0.06799374  5.766236
classclass3 -1.11336779  4.941146
classclass4 -2.31977970  4.524028
```

```
> multipleComp(test.fit)
                     Estimate Tukey.L Tukey.U Tukey.p
class1  -  class2 -2.9171147 -6.6684  0.8342  0.1836
class1  -  class3 -1.9138889 -5.8997  2.0719  0.5944
class1  -  class4 -1.1021242 -5.6075  3.4033  0.9193
class2  -  class3  1.0032258 -3.1122  5.1187  0.9200
class2  -  class4  1.8149905 -2.8055  6.4355  0.7349
class3  -  class4  0.8117647 -4.0011  5.6246  0.9713
```