THE UNIVERSITY OF AUCKLAND

SEMESTER ONE 2018
Campus: SOUTHWEST UNIVERSITY, CHONGQING, CHINA

STATISTICS

Data Analysis

(Time allowed: ONE Hours)

INSTRUCTIONS

- Attempt **all** questions.

1. This question refers to the **NZ Car Data** in **Appendix A**.

[Total: 17 marks]

a. Inspect the three scatter plots on the first page of **NZ Car Data**. What model do they suggest should be fitted to the data? Explain why.

[3 marks]

A model using a log transformation for both the explanatory variable `cc` and the response variable `kw` should be fitted. This appears to be a linear relationship with constant scatter, but the others do not.

One mark for suggesting a model with log *transformations for both variables. Two marks for a sensible explanation.*

b. Write down the equation of model being fitted here. State any assumptions that are being made.

[3 marks]

$\log(\text{kw}_i) = \beta_0 + \beta_1 \text{cc}_i + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$.

We also assume the observations are independent.

Three marks for an appropriate answer. Take off half marks for any small errors, or a whole mark for a larger error.

c. Discuss the difference between models `nzcars.fit` and `nzcars1.fit`. Are there any practical consequences of these differences? Justify your answer.

[3 marks]

The model `nzcars1.fit` has removed two observations with the largest Cook's distances. However, this change has not affected the parameter estimates very much, so there is not really a practical consequence. We should probably leave these observations in the model.

One mark for saying that ***`nzcars1.fit`*** *drops two of the obseravations. Two marks for saying that there are no practical consequences, with a suitable explanation.*

d. A Toyota Aurion has a 3456 cm^3 engine. Using the model `nzcars.fit`, calculate a point estimate for its power output. You do **not** need to provide an interval.

[3 marks]

$-2.65 + 0.96 \log(3456) = -2.65 + 0.96 \times 8.15 = 5.18.$

$\exp(5.18) = 176$ kilowatts.

e. Interpret the relationship between power output and engine size from the analysis of **NZ Car Data** in **APPENDIX A**.

[3 marks]

For every 1% increase in engine size, the expected power output increases by between 0.92% and 1.00%.

Three marks for a suitable interpretation. Note that students should not *use the* ***`exp(confint(nzcars.fit))`*** *output.*

2. This question refers to the **Auckland Rental Data** in **Appendix B**.

[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]

Rental price seems to be positively related to apartment size and the number of bedrooms. The relationship with distance from downtown is less clear, but it seems those that are far away are cheaper. There are some relationships between explanatory variables; for example, the size of an apartment seems highly correlated with the number of bedrooms.

Award one mark for each sensible comment, up to a maximum of three marks. They do not need to go into as much detail as I have provided above. Just picking out three relationships is enough.

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]

Model `rent1.fit` is a model that uses all explanatory variables, with `rent` as the response variable. However, there appeared to be nonconstant variance in the residuals, and so model `rent2.fit` was fitted, using `log(rent)` as the response. The `beds` variable was not significant, and so it was dropped for the model `rent3.fit`.

One mark for saying `rent1.fit` has nonconstant variance. One mark for saying `rent2.fit` has a log-transformation on the response variable. One mark for saying that the `beds` variable was not significant in `rent2.fit`. One mark for saying that `rent3.fit` has dropped this variable.

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 B**. Explain why.

[3 marks]

From the pairs plot, it looks as though there is a relationship between `rent` and `beds`. However, there is a strong relationship between the two explanatory variables `size` and `beds`, so it is likely that the variable `size` is already explaining the relationship between `beds` and `rent`. We don't need both in the model. The large p-value in `rent2.fit` is probably due to multicollinearity.

Three marks for a suitable explanation. Using the word 'multicollinearity' would be great, but they can explain the idea just fine without it. Take off half marks for any small errors, or a whole mark for larger errors.

d. Provide an equation for the model `model3.fit`, stating its assumptions.

[4 marks]

$\log(p_i) = \beta_0 + \beta_1 s_i + \beta_2 d_i + \beta_3 o_i + \beta_4 r_i + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$,

OR

$mu_i = \beta_0 + \beta_1 s_i + \beta_2 d_i + \beta_3 o_i + \beta_4 r_i$, $\log(p_i) \sim N(\mu_i, \sigma^2)$,

Where, for the ith apartment, p_i is its rental price, s_i is its size, d_i is its distance from downtown, o_i is a dummy variable indicating whether or not its age is 'old', and r_i is a dummy variable indicating whether or not its age is 'recent'.

We also assume the observations are independent.

Four marks for a suitable answer. Take a mark off for errors like forgetting the log *transformation. Take a half mark off for small errors, and a whole mark for larger errors.. They don't need to define their terms in as much detail as I have.*

e. Interpret the effect of each of the variables remaining in your chosen model on the rental price.

[9 marks]

Holding all else constant:

For every 100 square-foot increase in apartment size, we estimate the median rental price increases by between 7.3% and 10.4%.

For every one-mile increase in distance from downtown, the median rental price decreases by between 6.7% and 12.7%.

The median rental price of apartments that are old is between 20.0% and 40.0% higher than the median price of new apartments.

The median rental price of apartments that are recently built is between 0.9% and 18.8% higher than the median price of new apartments.

Two marks for each correct interpretation. One mark for remembering to refer to the median. Take off half a mark for small errors, or one *mark for larger errors. Remember alternative interpretations can still be correct; for example, we could say that the median rental price of old houses is between* 1.2 *and* 1.4 *times the median rental price of new houses.*

3. This question refers to the **Lobster Survival Data** in **Appendix C**.

[Total: 13 marks]

a. The model `lobster1.fit` was fitted by the biologists, and appeared in a paper they published. Why is this model inappropriate?

[3 marks]

They have used the proportion of surviving lobsters as the response variable. However, this will not have a normal distribution, because a proportion must be between 0 and 1. They have also assumed a linear relationship between size and the proportion; the expected value for this proportion will therefore also be oustide the $[0, 1]$ interval for lobsters of particular sizes. Finally, they have assumed constant variance, but this is unlikely to be met; very large and very small size classes are likely to have smaller variance than those in between, due to the variance of the binomial distribution.

Award three marks for any single sensible correct answer. No need to provide three, as I have done above.

b. Provide the equation of the model `lobster2.fit`, stating its assumptions.

[3 marks]

$\log(p_i/(1-p_i)) = \beta_0 + \beta_1 s_i$, $Y_i \sim \text{Binomial}(n_i, p_i)$, where s_i is the size of the ith class of lobsters, p_i is the probability a lobster in this size class survives, Y_i is the number of lobsters in the size class that survived, and n_i is the total number of lobsters in the size class. We assume size classes are independent.

One mark for having the correct linear combination on the right hand side. One mark for using the logistic link function. One mark for stating the number of lobsters that survived is a binomial random variable. One mark for mentioning independence.

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

[4 marks]

Yes. The residual plot looks OK. The null hypothesis that the model is appropriate is not rejected by the deviance statistic. We should have independence, given the design of the survey.

One mark for saying yes. One mark for mentioning the residual plot. One mark for mentioning the deviance. One mark for mentining independence.

d. Interpret the effect of lobster size on survival, using the model `lobster.fit2`.

[3 marks]

For every 1 mm increase in lobster size, the odds of surviving are multiplied by between 1.14 and 1.31.

OR

For every 1 mm increase in lobster size, the odds of surviving are increased by between 14 and 31%.

Three marks for a correct interpretation. There are others that are correct, but the above two are the most likely to be used.

4. This question refers to the **Hull Damage Data** in **Appendix D**.

[Total: 18 marks]

a. What is the difference between `ship1.fit` and `ship2.fit`? Why was this change made?

[2 marks]

The model `ship1.fit` has the explanatory variable `year.fac`, but `ship1.fit` does not. The variable was dropped because it was not significant.

One mark for saying `ship2.fit` does not have `year.fac` as an explanatory variable. One mark for a sensible explanation.

b. Provide the equation of the model `ship2.fit`, stating its assumptions.

[3 marks]

$\log(\mu_i) = \beta_0+\beta_1 b_i+\beta_2 s_i$, $Y_i \sim \text{Poisson}(\mu_i)$, where $b_i = 1$ if the ith ship had a hull made by Company B, s_i is the number of months it has been in service, and Y_i is its number of damage incidents. Observations are assumed to be independent.

One mark for the right linear combination on the right-hand side. One mark for correctly including the log *link function. One mark for specifying the response has a Poisson distribution. One mark for all the terms being correctly defined. Don't remove any marks if they forget about independence.*

c. Do you think the model `ship2.fit` is appropriate? Explain why, or why not, with reference to its assumptions.

[4 marks]

Yes, it seems appropriate. The residual plot looks OK, and the deviance is suitably low so that we do not reject the null hypothesis that the model is appropriate. We have independence as the ships were randomly selected.

One mark for saying yes. One mark for mentioning the residual plot. One mark for saying the deviance is low enough. One mark for independence.

If they say the model is not appropriate for a sensible reason, then feel free to award full marks. For example, some students may say that the residual plot has curvature. I don't think this curvature is clear enough to obviously be a real effect, but I appreciate that this is a judgement call, so award the marks in this case.

d. Interpret the effect of the explanatory variables in the model `ship2.fit` on the number of hull damage incidents.

[4 marks]

We estimate that ships with hulls made by Company B have between 8% and 61% fewer damage incidents than ships made by Company A.

We estimate that, for every passing year of service, the number of damage incidents on a ship's hull increases by between 26% and 189%.

Two marks for correct interpretation of each interval. Note that there are many appropriate interpretations: students may provide a percentage change as above, but they also may make a multiplicative statement. They may also describe how Company A as more damage incidents than Company B, rather than how Company B has fewer than Company A.

e. The CEO of Company A claims that hulls made by Company A age better than hulls made by Company B. In other words, they claim that the effect of service is less pronounced for hulls made by Company A than it is for hulls made by Company B. What model could be fitted to test this claim? You may either explain in words, or write a line of R code.

[2 marks]

A model with an interaction between `company` and `service` should be fitted.

`glm(incidents  company*service, family = "poisson", data = ship.df)`

Two marks for stating an interaction should be fitted. If they provide R code, do not deduct marks for small mistakes—just lok for the interation.

f. Note that the variable `year.fac` is significant in the model `ship3.fit`, but it is not in the model `ship1.fit`. Explain why this is the case.

[3 marks]

Two possible answers:

The variables `year.fac` and `service` are correlated: the earlier a ship was made, the longer it has been in service. Including them both in the model results in `year.fac` not being significant, because the effect of the age of the boat on damage incidents is already being explained by its length in service. Once `service` is removed, `year.fac` therefore becomes significant.

OR

The model `ship3.fit` does not appear to be appropriate, because its residual deviance is much larger than the associated degrees of freedom. We shouldn't really trust the p-value for `year.fac` in this model. The Poisson assumption does not seem reasonable; once we change to a quasi-Poisson model, it is likely that `year.fac` will no longer be significant.

Three marks for a sensible interpretation.

5. This question refers to the **Beer Quality Data** in **Appendix E**.

[Total: 10 marks]

a. Calculate an estimate of an odds-ratio to explain the relationship between the price of beer and its perceived quality.

[2 marks]

$(24 \times 48)/(36 \times 12) = 2\frac{2}{3}$

Two marks for a correct odds-ratio. It is fine to calculate this the other way around to get $(36 \times 12)/(24 \times 48) = 0.375$, *in which case all answers below need to be inverted, too.*

b. Calculate a confidence interval for this estimate. Provide an interpretation.

[5 marks]

Standard error of $\log(\text{OR})$ is $\sqrt{1/24 + 1/36 + 1/48 + 1/12} \approx 0.417$.

CI for $\log(\text{OR})$ is $\log(2\frac{2}{3}) \pm 1.96 \times 0.417 = (0.164, 1.797)$.

CI for OR is $\exp(0.164, 1.797) = (1.18, 6.03)$.

The odds of a person perceiving the beer to be of 'good' quality are between 18% and 503% higher for people who think the beer's price is 'high' than they are for people who think the beer's price is 'low'.

One mark for calculating the standard error of the log of the odds-ratio. One mark for calculating the confidence interval for the log of the odds-ratio. One mark for exponentiating to give the confidence interval for the odds ratio. Two marks for a good interpretation.

c. Is there evidence to suggest that the price of beer is related to its perceived quality? Provide an explanation.

[3 marks]

Yes. The confidence interval for the odds ratio does not contain 1, so it is not plausible that the odds of perceiving the beer to be of 'good' quality is the same for both the people who though the beer had a low price, and the people who thought the beer had a high price.

One mark for saying yes. Two marks for a sensible explanation, either using the confidence interval above, or by calculating a Chi-square test statistic (although this is time consuming).

6. This question refers to the **Southwest University Test Data** in **Appendix F**.

[Total: 13 marks]

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

[4 marks]

$\text{marks}_i = \beta_0 + \beta_1 b_i + \beta_2 c_i + \beta_3 d_i + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$.

OR

$\mu_i = \beta_0 + \beta_1 b_i + \beta_2 c_i + \beta_3 d_i$, $\text{marks}_i \sim N(\mu_i, \sigma^2)$.

Here, b_i is a dummy variable for Class 2, c_i is a dummy variable for Class 3, and d_i is a dummy variable for Class 4. We also assume the observations are independent.

Four marks for a suitable answer. Take off half a mark for small errors, and a whole mark for larger errors.

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

[3 marks]

The null hypothesis is that all group means are the same. We have no evidence to reject this null hypothesis.

Two marks for the correct null hypothesis. One mark for stating that it is not rejected. They may write the null hypothesis in words (as above) or they may write something like "the null hypothesis is $\mu_1 = \mu_2 = \cdots = \mu_4$" or even "the null hypothesis is $\beta_1 = \beta_2 = ... = \beta_3 = 0$". All of these are correct.

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 contradiction.

[2 marks]

The small p-value in the `summary()` output is likely due to the multiple comparisons problem. We should put more trust in the p-value from the `anova()` output, because it tests the composite null hypothesis that all coefficients of the dummy variables are equal to 0 in one go.

Two marks for mentioning the multiple comparisons problem.

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]

The confidence intervals in the `multipleComp()` output have been adjusted for multiple comparisons.

Two marks for stating that `multipleComp()` has adjusted for multiple comparisons. If a student says that `multipleComp()` shows all possible comparisons between levels but `confint()` does not, and does not mention the multiple-comparison adjustnment, then only award one mark.

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. The average test mark for Class 2 students is somewhere between 0.07 and 5.77 marks higher than that of Class 1 students. Do you agree with this conclusion? Explain why, or why not.

[2 marks]

No, this is not a sensible conclusion. The lecturer is basing his interpretation on p-values and confidence intervals that have not been adjusted for multiple comparisons. After this adjustment has taken place, the p-value testing the null hypothesis of no difference between Class 1 and Class 2 is large (0.18), and the confidence interval is much wider (-0.83 to 6.67).

Two marks for saying no, because the lecturer has failed to adjust for multiple comparisons.

7. This section contains multiple-choice questions.

- 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 binomial response distribution? Let p_i be the probability of a 'successful' trial in the ith observation.

(1) The response variable for each observation has a binomial distribution.

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

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

(4) The observations are independent of one another.

(3)

b. Blackburn Rovers is a football club based in Lancashire, England. Which of the following distributions is most appropriate for the number of goals they score next season?

(1) Poisson.

(2) Binomial.

(3) Normal.

(4) This is not a random variable. We already know they won't score any goals.

(1)

c. A generalised linear model was fitted with the `glm()` function in R, using the argument `family = "quasipoisson"`. Which of the following **IS** an assumption of this model? Let Y_i be the response of the ith observation, with expectation μ_i.

(1) $\text{Var}(Y_i) = k\mu_i$, where k is estimated by the model.

(2) μ_i is a linear combination of the explanatory variables.

(3) The errors have a normal distribution.

(4) Y_i has a Poisson distribution with expectation μ_i.

(1)

d. How are Tukey-adjusted confidence intervals different from the standard confidence intervals that are provided by the `confint()` function?

(1) Tukey-adjusted confidence intervals were invented by turkeys, but standard confidence intervals were invented by human beings.

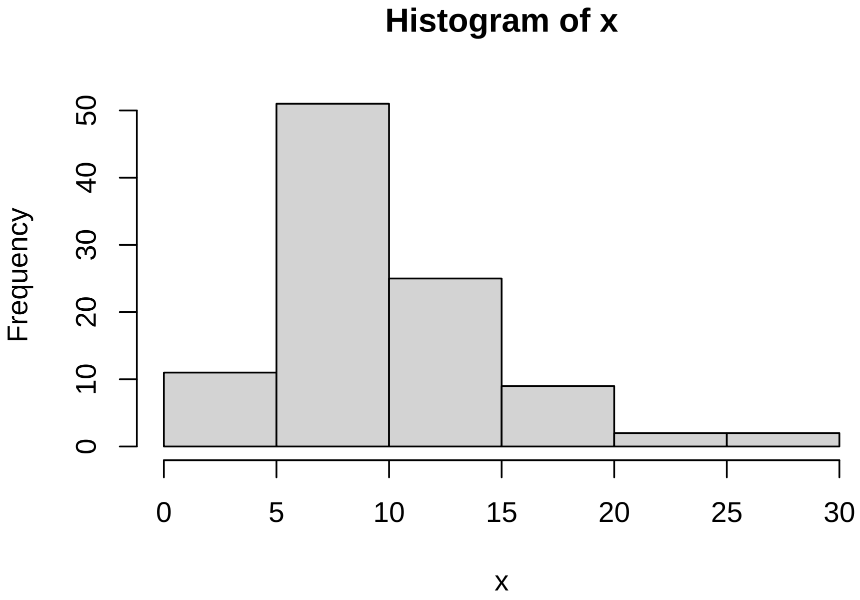
(2) Tukey-adjusted confidence intevals account for multicollinearity, but standard confidence intervals do not.

(3) Tukey-adjusted confidence intervals account for multiple comparisons, but standard confidence intervals do not.

(4) Standard confidence intervals are not appropriate when the response variable comes from a Tukey distribution. Instead, we should use Tukey-adjusted confidence intervals.

(3)

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



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

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

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

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

(3)

8. This section contains multiple-choice questions.

   - 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) The response variable for each observation has a Poisson distribution.

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

(4)

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.

(2)

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.

(3)

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.

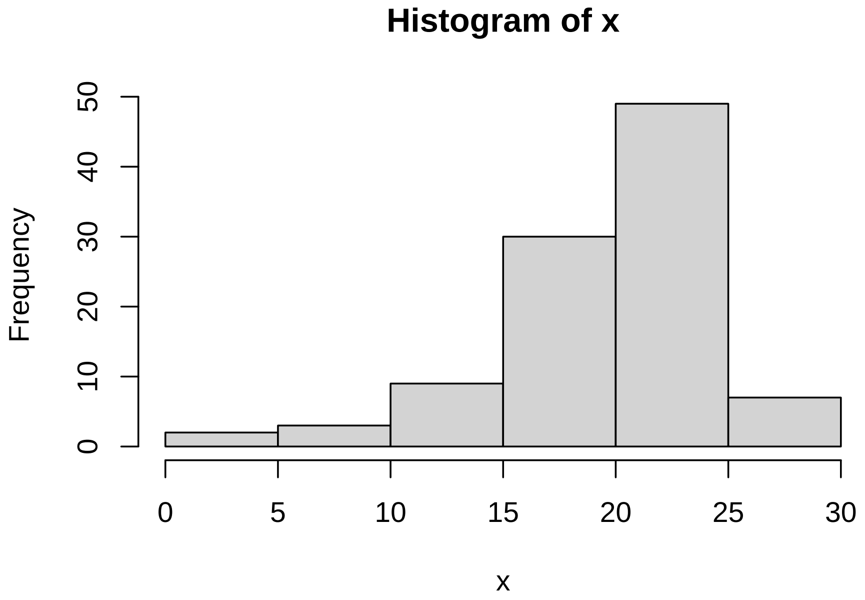
(2) We compute the p-value from a Chi-squared test by comparing the Chi-squared test statistic to an F-distribution.

(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.

(3)

e. In total, 100 observations of some variable x were recorded, and are plotted in the histogram below. Which statement 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) $\bar{x} < \tilde{x}$

(4)

APPENDICES BOOKLET FOLLOWS

APPENDICES BOOKLET

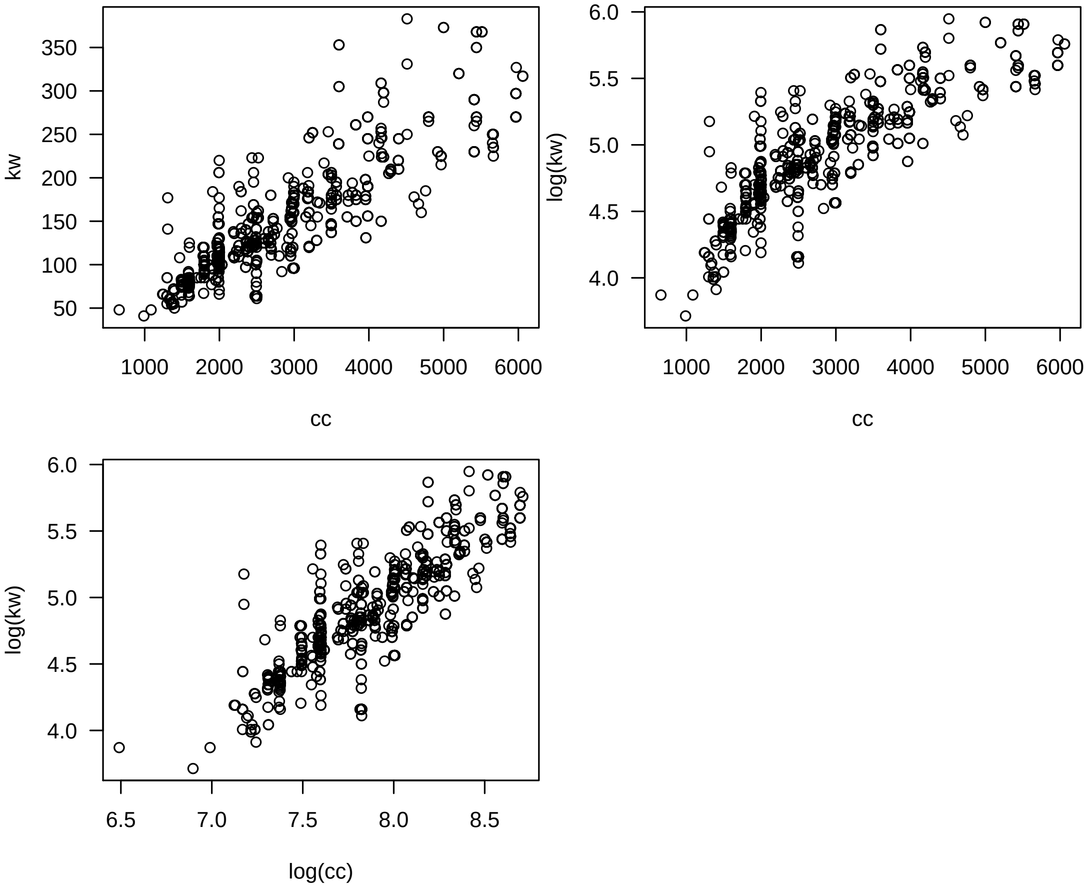
CONTENTS

NZ Car Data

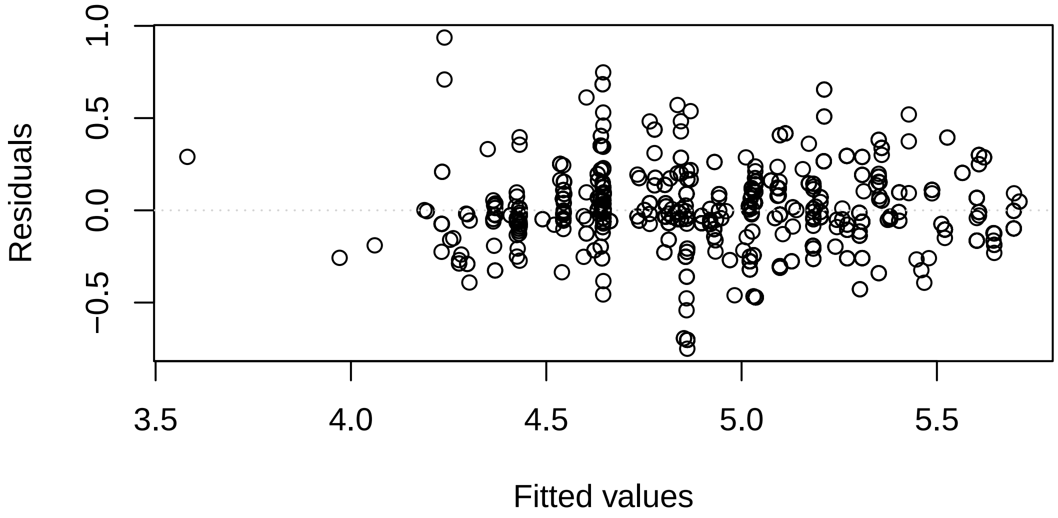
The New Zealand Automobile Association (AA) publish various data about models of car on their website. Although there are many variables available, here it is of interest to determine the relationship between the power output of the cars and the size of their engines.

The variables in the data set are

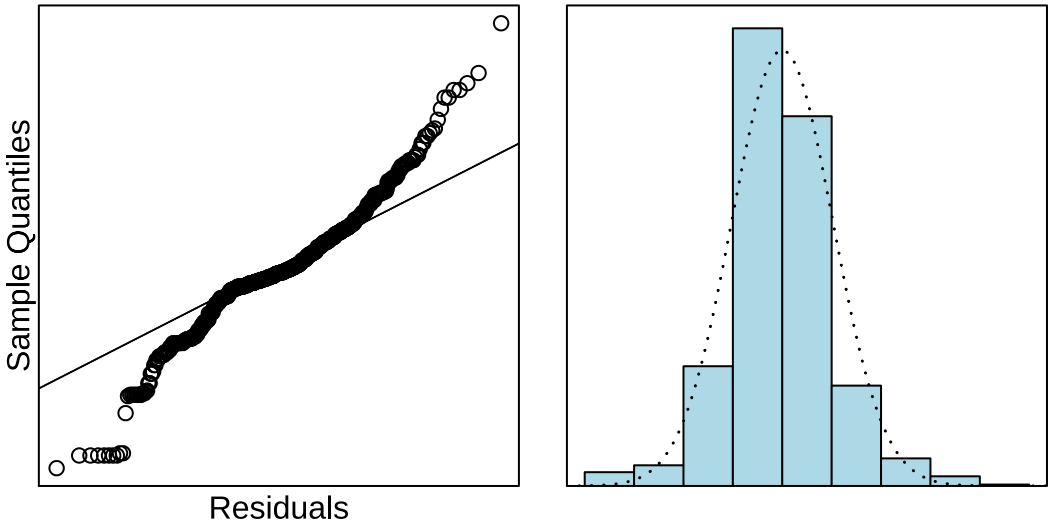
`kw`	The power output of the car measured in kilowatts (kW)
`cc`	The engine size measured in cubic centimetres (cm^3)



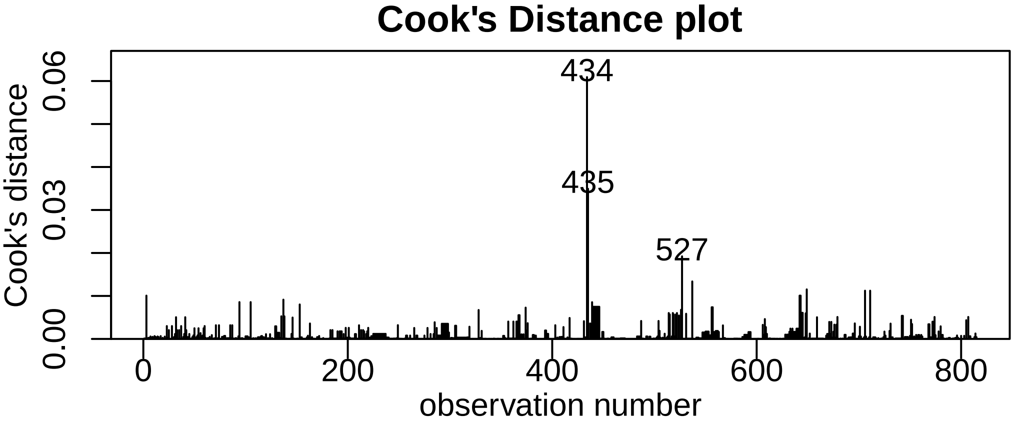
```
> nzcars.fit = lm(log(kw) ~ log(cc), data = nzcars.df)
> eovcheck(nzcars.fit)
```



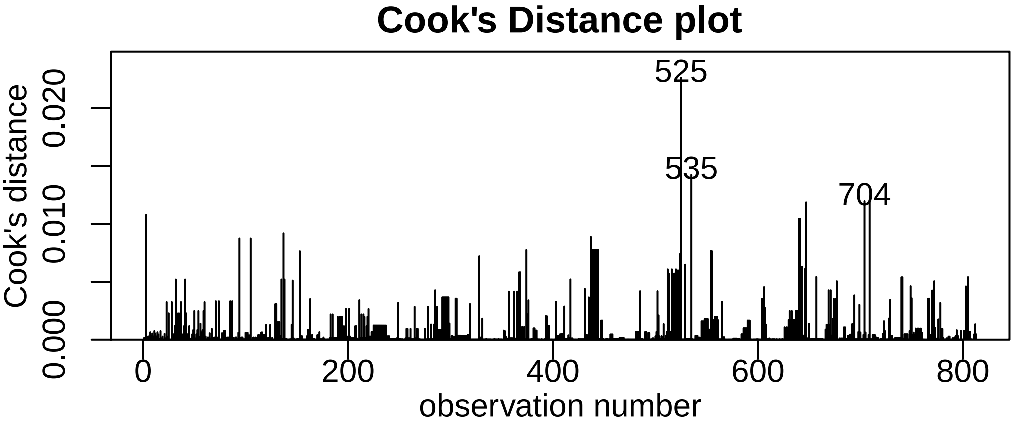
```
> normcheck(nzcars.fit, main = "", xlab = "Residuals")
```



```
> cooks20x(nzcars.fit)
```



```
> nzcars.df[434:435,]
     make model doors   cc  kw manual  auto
434 Mazda   RX8     2 1308 177  61995  <NA>
435 Mazda   RX8     2 1308 141     NA 61995
> nzcars1.fit = lm(log(kw)~log(cc),
+                   data = nzcars.df[-c(434,435), ])
> cooks20x(nzcars1.fit)
```



```
> summary(nzcars.fit)

Call:
lm(formula = log(kw) ~ log(cc), data = nzcars.df)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.75009 -0.07331 -0.00998  0.10614  0.93675 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.65055    0.15153  -17.49   <2e-16 ***
log(cc)      0.96010    0.01916   50.10   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 0.2051 on 813 degrees of freedom
Multiple R-squared:  0.7553,Adjusted R-squared:  0.755 
F-statistic:  2510 on 1 and 813 DF,  p-value: < 2.2e-16 
> summary(nzcars1.fit)

Call:
lm(formula = log(kw) ~ log(cc), data = nzcars.df[-c(434, 435), 
    ])

Residuals:
     Min       1Q   Median       3Q      Max 
-0.74728 -0.06717 -0.00918  0.10694  0.75313 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.73545    0.14929  -18.32   <2e-16 ***
log(cc)      0.97060    0.01888   51.41   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 0.2011 on 811 degrees of freedom
Multiple R-squared:  0.7652,Adjusted R-squared:  0.7649 
```

```
F-statistic:  2643 on 1 and 811 DF,  p-value: < 2.2e-16


> confint(nzcars.fit)
                 2.5 %     97.5 %
(Intercept) -2.9479738 -2.3531209
log(cc)      0.9224856  0.9977227
> exp(confint(nzcars.fit))
                 2.5 %     97.5 %
(Intercept) 0.05244587 0.09507199
log(cc)     2.51553523 2.71209843
> confint(nzcars1.fit)
                 2.5 %    97.5 %
(Intercept) -3.0284991 -2.442401
log(cc)      0.9335391  1.007653
> exp(confint(nzcars1.fit))
                 2.5 %     97.5 %
(Intercept) 0.04838821 0.08695183
log(cc)     2.54349505 2.73916420
```

Auckland Rental Data

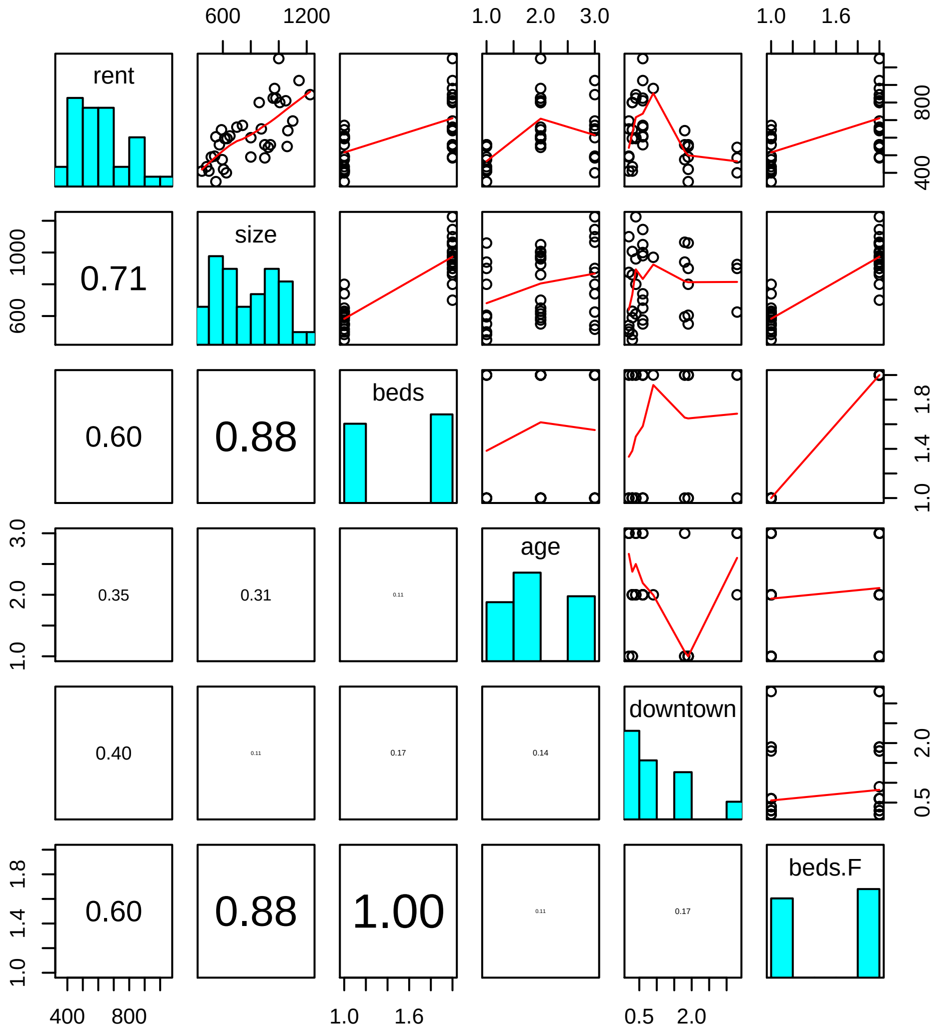
Data were collected on the monthly rental and other characteristics of 36 randomly selected apartments in Auckland. We wish to build a model to explain the monthly rental of an apartment. The variables measured were

`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)

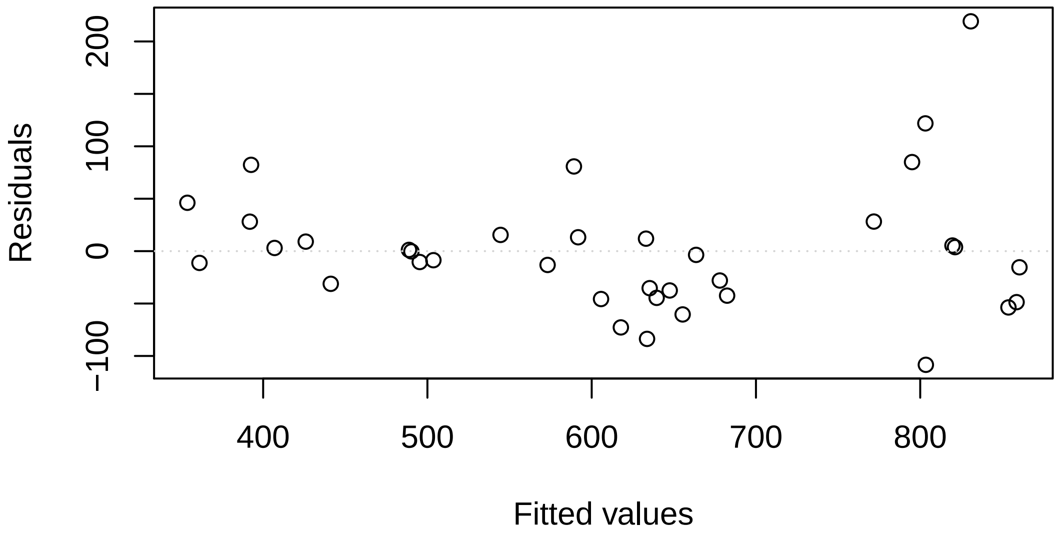
```
> ## Printing the first six observations.
> head(rent.df)
  rent size beds age downtown
1  810 1050    2 Old      0.6
2  560  575    1 Old      0.6
3  550 1060    2 New      1.9
4  610  650    1 Old      0.6
5  800 1007    2 Old      0.3
6  435  484    1 New      0.3
```

```
> 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 
```

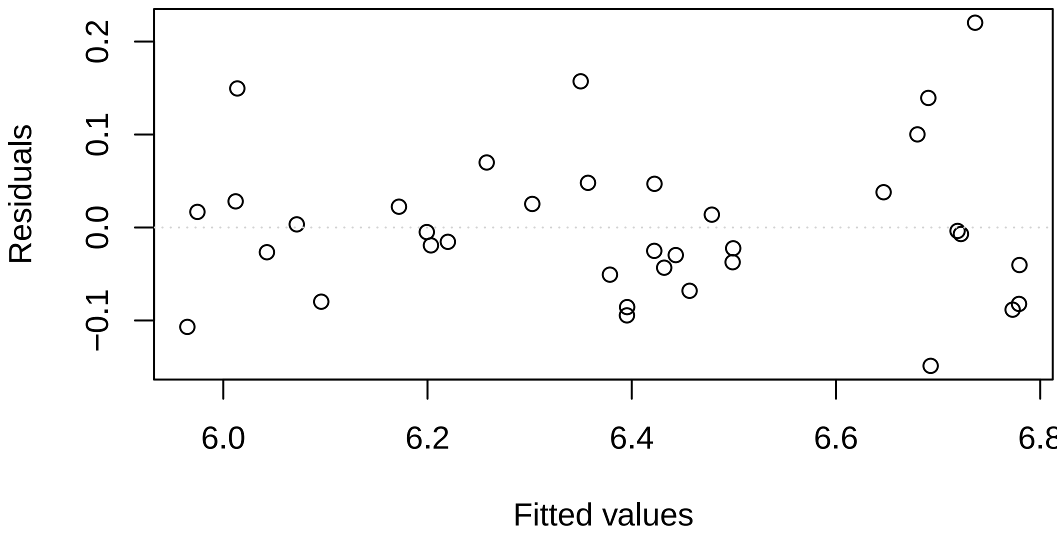
```
> pairs20x(rent.df)
```



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

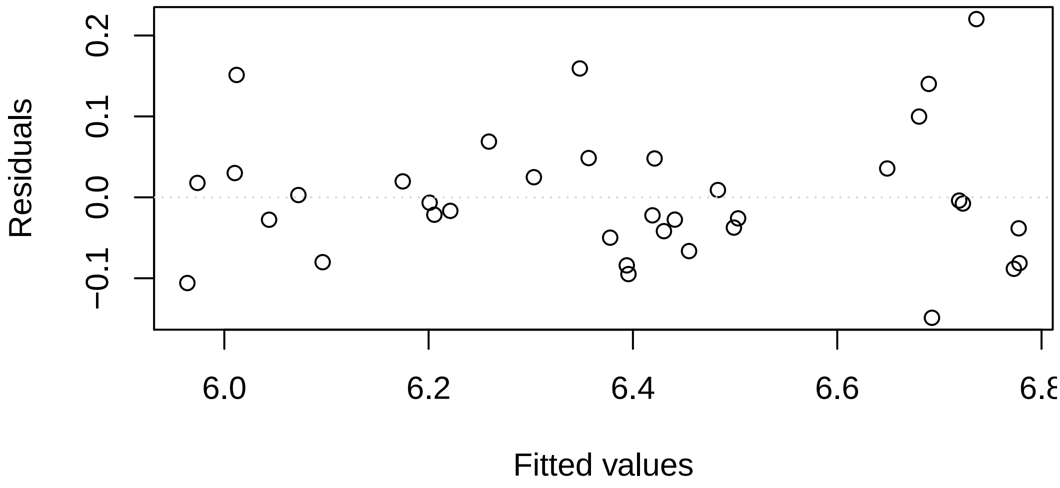


```
> rent2.fit=lm(log(rent) ~ size + downtown + beds.F + age,
+              data = rent.df)
> eovcheck(rent2.fit)
```

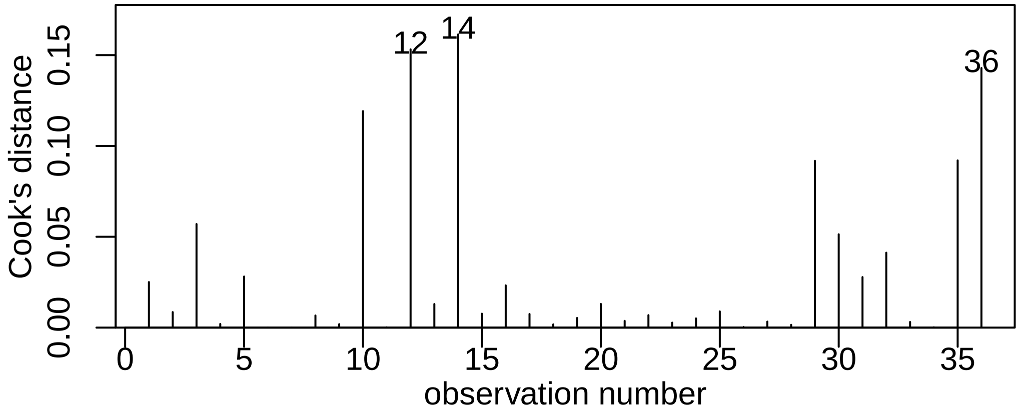


```
> summary(rent2.fit)

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
```

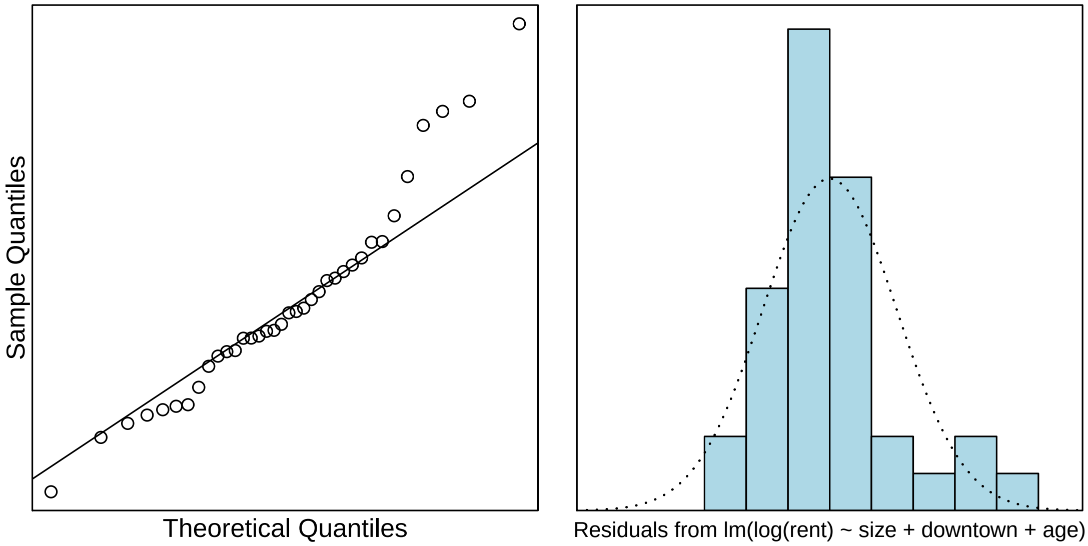
```
> rent3.fit = lm(log(rent) ~ size + downtown + age, data = rent.df)
> summary(rent3.fit)

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

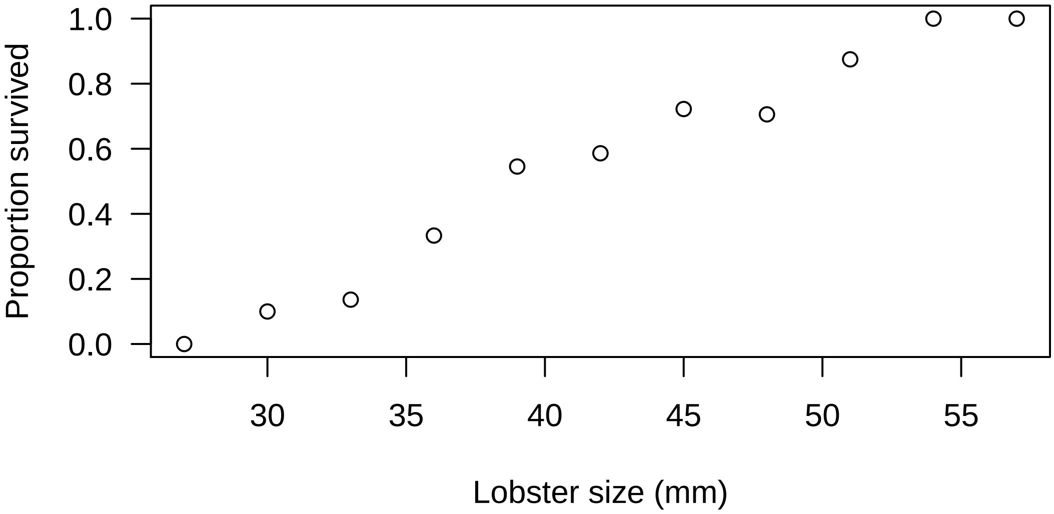
Biologists collected data to investigate how a lobster's size affects its survival. In total, they collected 159 juvenile lobsters from their natural habitat, and measured their size. They tethered the lobsters to the ocean floor for one night. Any lobsters that were missing were assumed to have been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

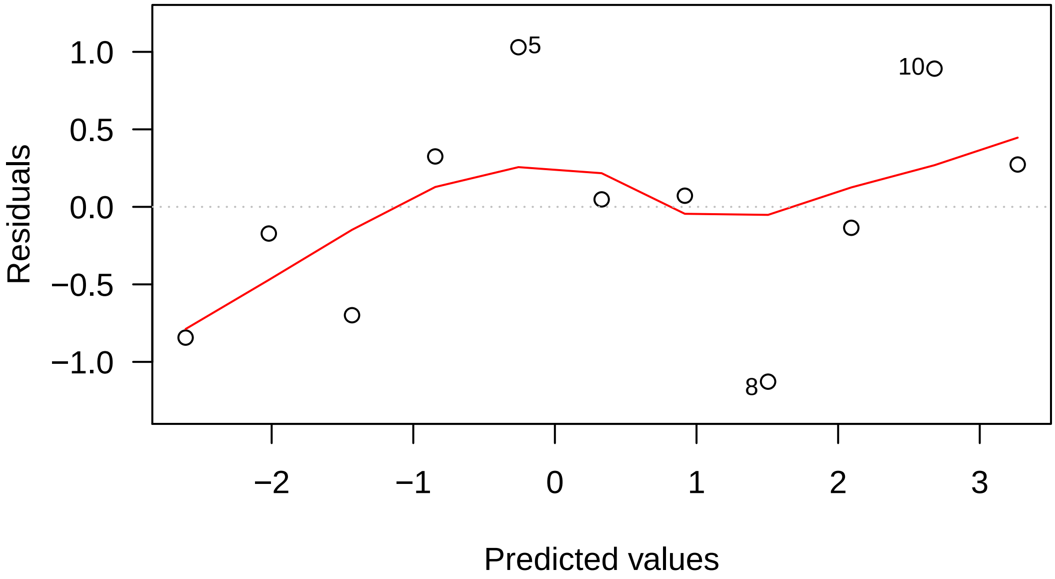
```
> lobster.df = within(lobster.df, {
+     p = survived/n
+ })
> lobster.df
   size  n survived         p
1    27  5        0 0.0000000
2    30 10        1 0.1000000
3    33 22        3 0.1363636
4    36 21        7 0.3333333
5    39 22       12 0.5454545
6    42 29       17 0.5862069
7    45 18       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
```

Hull Damage Data

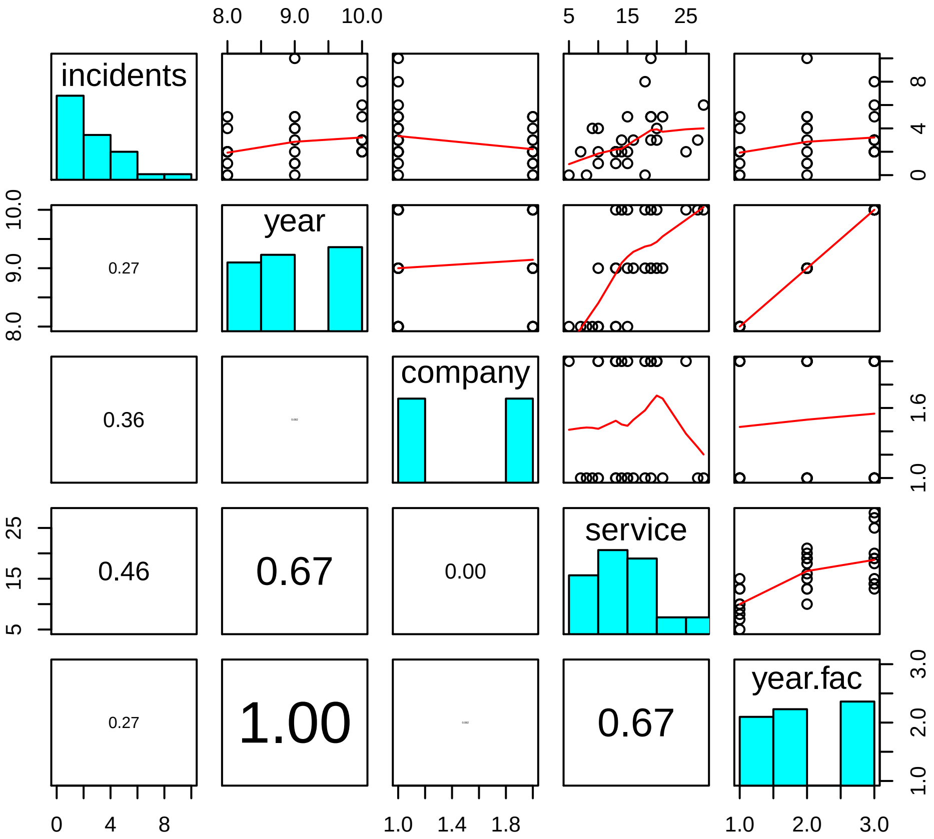
This question refers to data that come from a study investigating a particular type of minor damage caused by waves to the forward sections of ships' hulls. In total, 30 randomly selected ships were inspected for hull damage, and the number of damage incidents were recorded from each. Hull construction engineers are interested in determining if the design of the hull is related to the number of observed damage incidents. Hull designs vary across manufacturers, and potentially improve from year to year. Also, we might expect more damage incidents for ships that have been in service for longer periods of time.

The variables in the data set are

`incidents` The number of damage incidents detected on the ship's hull
`year` The year of construction: `8` for 2008, `9` for 2009, and `10` for 2010
`company` The company that constructed the ship; either `A` or `B`
`service` The number of months the ship was in service

```
> ship.df <- within(ship.df, {
+     year.fac = as.factor(year)
+ })
> summary(ship.df$incidents)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   2.000   2.500   3.067   4.000  10.000
> summary(ship.df$year.fac)
 8  9 10
 9 10 11
> summary(ship.df$company)
 A  B
15 15
> summary(ship.df$service)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   5.00   13.00   15.00   15.53   19.00   28.00
```

```
> pairs20x(ship.df)
```



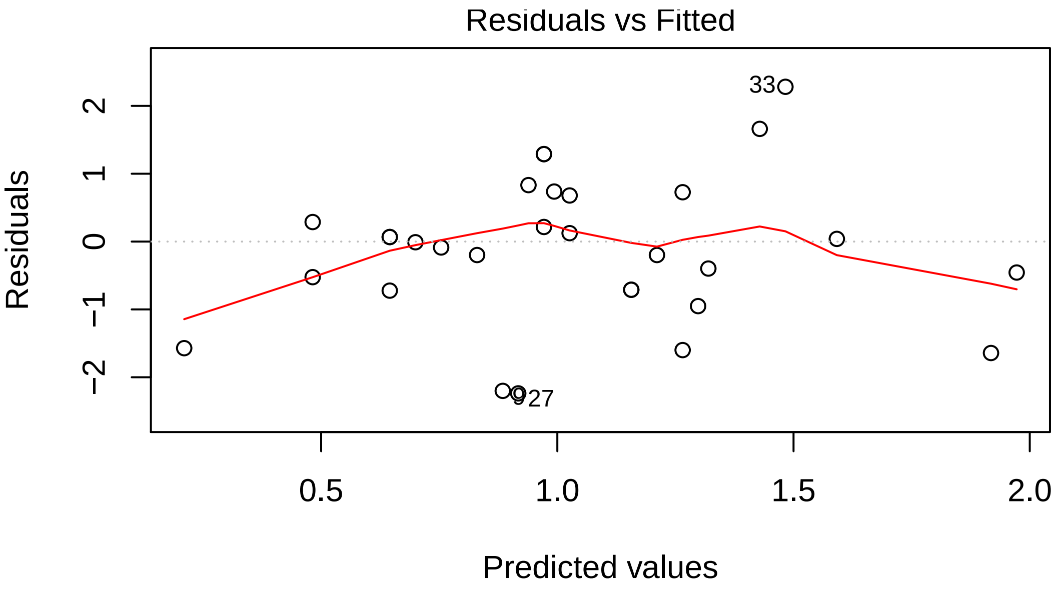
```
> ship1.fit <- glm(incidents ~ company + service + year.fac,
+                   family = "poisson", data = ship.df)
```

```
> anova(ship1.fit, test = "Chisq")
Analysis of Deviance Table

Model: poisson, link: log

Response: incidents

Terms added sequentially (first to last)


         Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                        29     50.195
company   1   6.3339        28     43.861  0.01185 *
service   1   9.4511        27     34.410  0.00211 **
year.fac  2   0.8783        25     33.532  0.64459
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> summary(ship1.fit)

Call:
glm(formula = incidents ~ company + service + year.fac, family = "poisson",
    data = ship.df)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.35927  -0.70150  -0.09147   0.45431   1.99681

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.38336    0.35939   1.067   0.2861
companyB    -0.52422    0.21857  -2.398   0.0165 *
service      0.04931    0.02396   2.058   0.0396 *
year.fac9    0.27684    0.32885   0.842   0.3999
year.fac10   0.13834    0.38027   0.364   0.7160
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 50.195  on 29  degrees of freedom
Residual deviance: 33.532  on 25  degrees of freedom
AIC: 123.49

Number of Fisher Scoring iterations: 5


> ship1.fit$deviance
[1] 33.5321
> ship1.fit$df.residual
[1] 25
> 1 - pchisq(ship1.fit$deviance, ship1.fit$df.residual)
[1] 0.1182934


> ship2.fit <- glm(incidents ~ company + service,
+                  family = "poisson", data = ship.df)


> summary(ship2.fit)

Call:
glm(formula = incidents ~ company + service, family = "poisson",
    data = ship.df)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.23706  -0.70999  -0.04805   0.58199   2.28167

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.44937    0.33636   1.336  0.18156
companyB    -0.51125    0.21658  -2.361  0.01825 *
service      0.05439    0.01747   3.114  0.00185 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 50.195  on 29  degrees of freedom
Residual deviance: 34.410  on 27  degrees of freedom
AIC: 120.37

Number of Fisher Scoring iterations: 5
> 1 - pchisq(ship2.fit$deviance, ship2.fit$df.residual)
[1] 0.1544394

> plot(ship2.fit, which = 1)
```



```
> confint(ship2.fit)


Waiting for profiling to be done...


                  2.5 %      97.5 %
(Intercept) -0.23128811  1.08940658
companyB    -0.94508880 -0.09286586
service      0.01990593  0.08849989
```

```
> exp(confint(ship2.fit))


Waiting for profiling to be done...


                2.5 %    97.5 %
(Intercept) 0.7935108 2.9725096
companyB    0.3886451 0.9113157
service     1.0201054 1.0925341


> ## For twelve-month change in service.
> exp(12 * confint(ship2.fit)[3, ])


Waiting for profiling to be done...


   2.5 %   97.5 %
1.269815 2.892146


> ship3.fit <- glm(incidents ~ company + year.fac,
+                   family = "poisson", data = ship.df)
> summary(ship3.fit)

Call:
glm(formula = incidents ~ company + year.fac, family = "poisson",
    data = ship.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2483  -0.7988  -0.3150   0.5687   2.2444

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.9080     0.2452   3.704 0.000213 ***
companyB     -0.5708     0.2166  -2.636 0.008392 **
year.fac9     0.5900     0.2902   2.033 0.042068 *
year.fac10    0.6285     0.2857   2.200 0.027789 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 50.195  on 29  degrees of freedom
Residual deviance: 37.809  on 26  degrees of freedom
AIC: 125.76

Number of Fisher Scoring iterations: 5
> anova(ship3.fit, test = "Chisq")
Analysis of Deviance Table

Model: poisson, link: log

Response: incidents

Terms added sequentially (first to last)


         Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                        29     50.195
company   1   6.3339        28     43.861  0.01185 *
year.fac  2   6.0522        26     37.809  0.04850 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Beer Quality Data

Researchers were interested in determining whether or not the price of beer affected its perceived quality. In total, 60 people were given identical bottles of beer to try. They were split into two groups of 30: one group was told that the beer was cheap, and the other group was told that the beer was expensive. Each person then rated the beer as either ‘good’ or ‘poor’.

The data are shown below.

		Quality	
		Poor	Good
Price	Low	24	36
	High	12	48

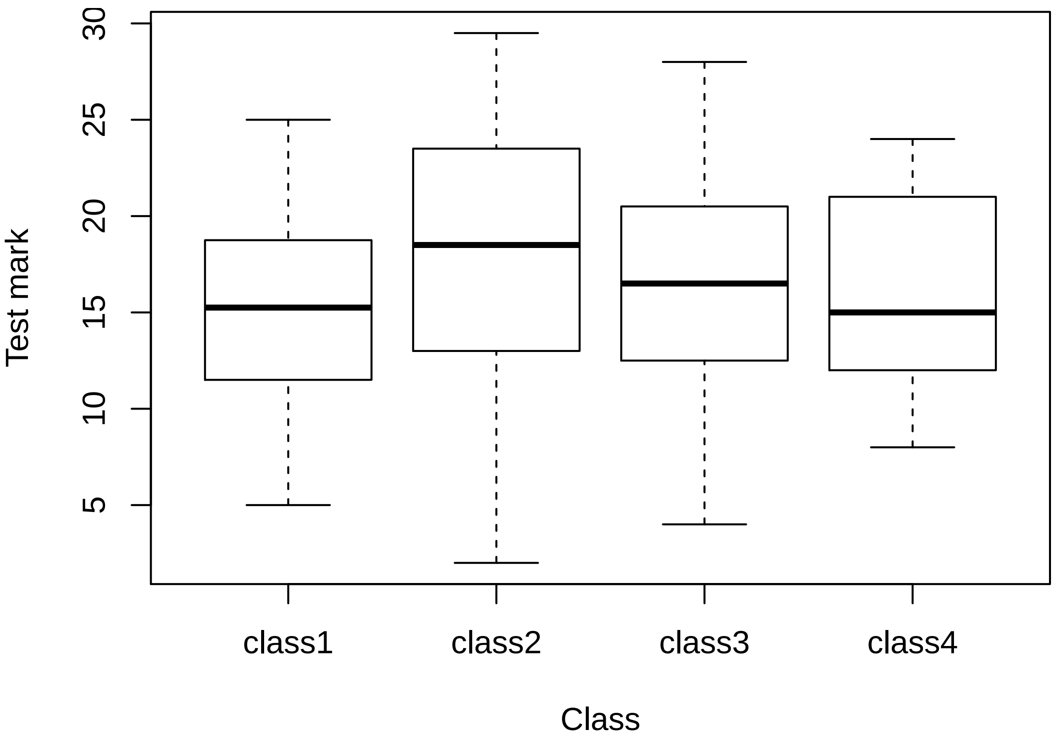
Southwest University Test Data

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.

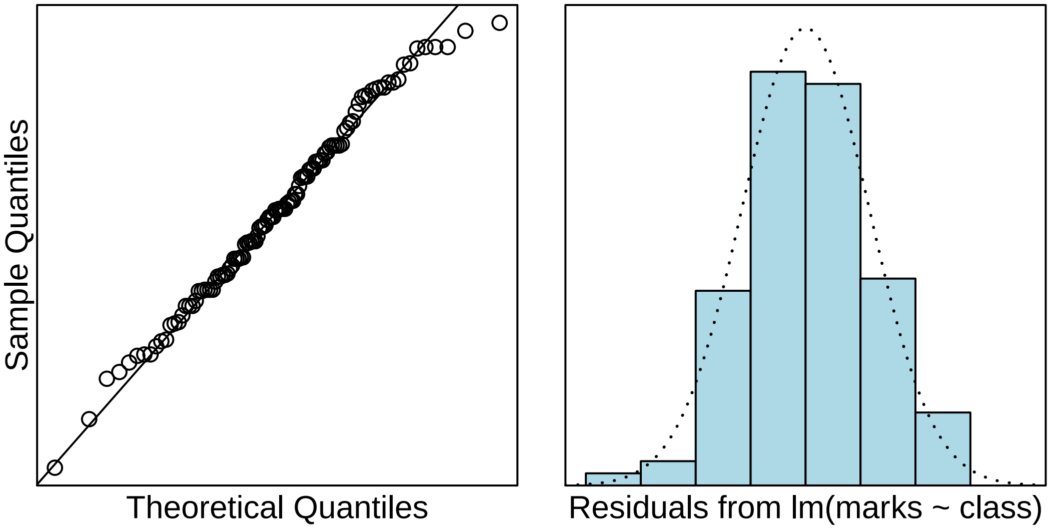
The variables in the data set are

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

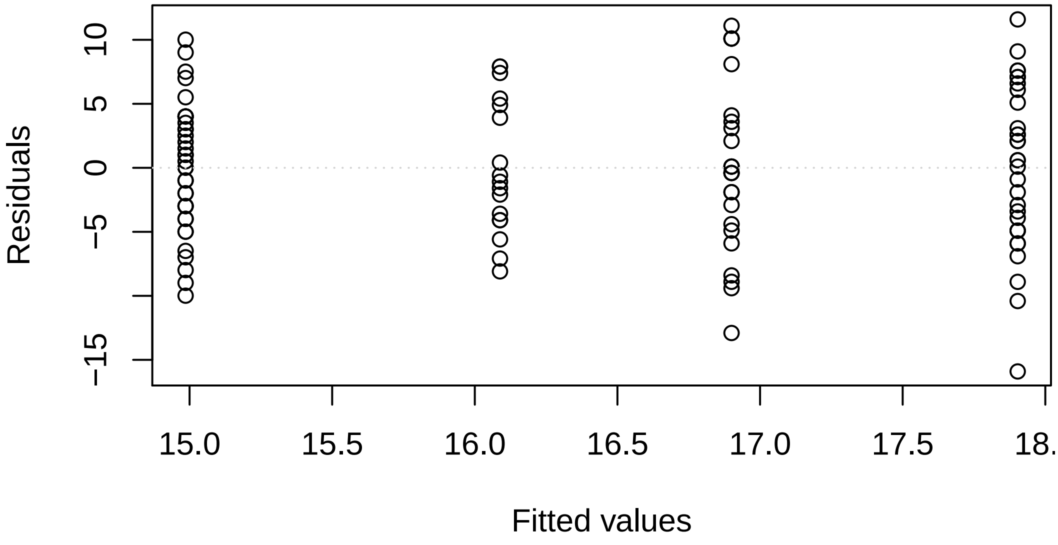
```
> boxplot(marks ~ class, data = test.df, xlab = "Class",
+         ylab = "Test mark")
```



```
> 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
```