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 **NZ Car Data** in **Appendix A**.

[Total: 17 marks]

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

   [3 marks]

   b. Provide the equation of model the model `nzcars.fit`, stating any assumptions that are being made.

   [3 marks]

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

[3 marks]

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]

e. According to Toyota, the power output for a Toyota Aurion is 205 kilowatts. Calculate the residual using your answer from the previous question. Remember the residual is the difference between the observed response and its expectation, where the observed response is assumed to have a normal distribution.

[2 marks]

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

[3 marks]

2. This question refers to the **Hull Damage Data** in **Appendix B**.

[Total: 19 marks]

a. What is a name given to the type of model fitted in `ship1.fit` and `ship2.fit`?

[1 marks]

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

[2 marks]

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

[3 marks]

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

[4 marks]

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

[4 marks]

f. 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 larger for hulls made by Company B than it is for hulls made by Company A. What model could be fitted to test this claim? You may either explain in words, or write a line of R code.

[2 marks]

g. 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 may be the case.

[3 marks]

3. This question refers to the **Beer Quality Data** in **Appendix D**.
[Total: 14 marks]

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

   b. Calculate a confidence interval for this estimate.
   [3 marks]

c. Provide an interpretation of your confidence interval.

[3 marks]

d. Using your confidence interval, is there evidence to suggest that the price of beer is related to its perceived quality? Provide an explanation.

[3 marks]

e. Inspect the output from the `chisq.test()` function in **Appendix D**. What is the null hypothesis associated with the p-value?

[1 marks]

f. Provide an interpretation of the output from the hypothesis test.

[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 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.

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.

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.

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

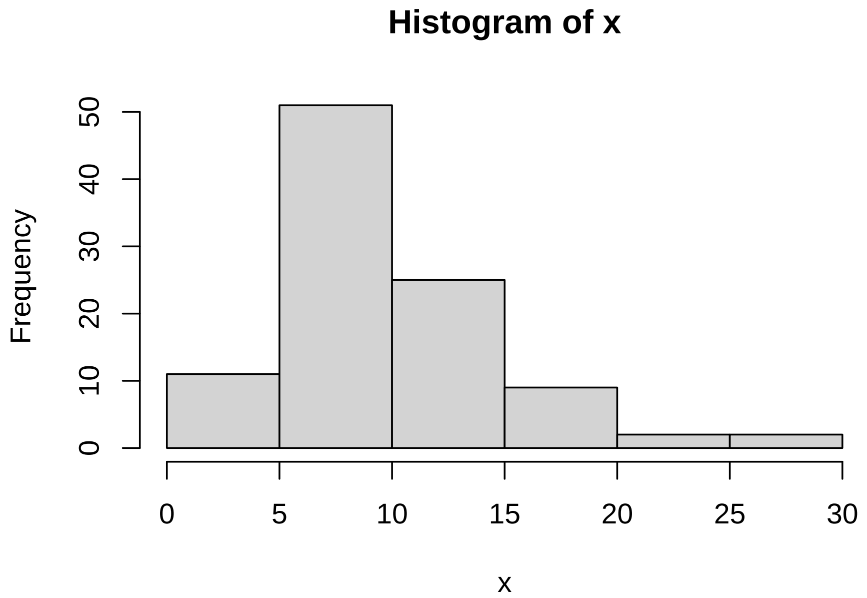
(1) Tukey-adjusted confidence intervals were invented by turkeys, but standard confidence intervals were invented by human beings.

(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.

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

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

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

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

APPENDICES BOOKLET FOLLOWS

APPENDICES BOOKLET

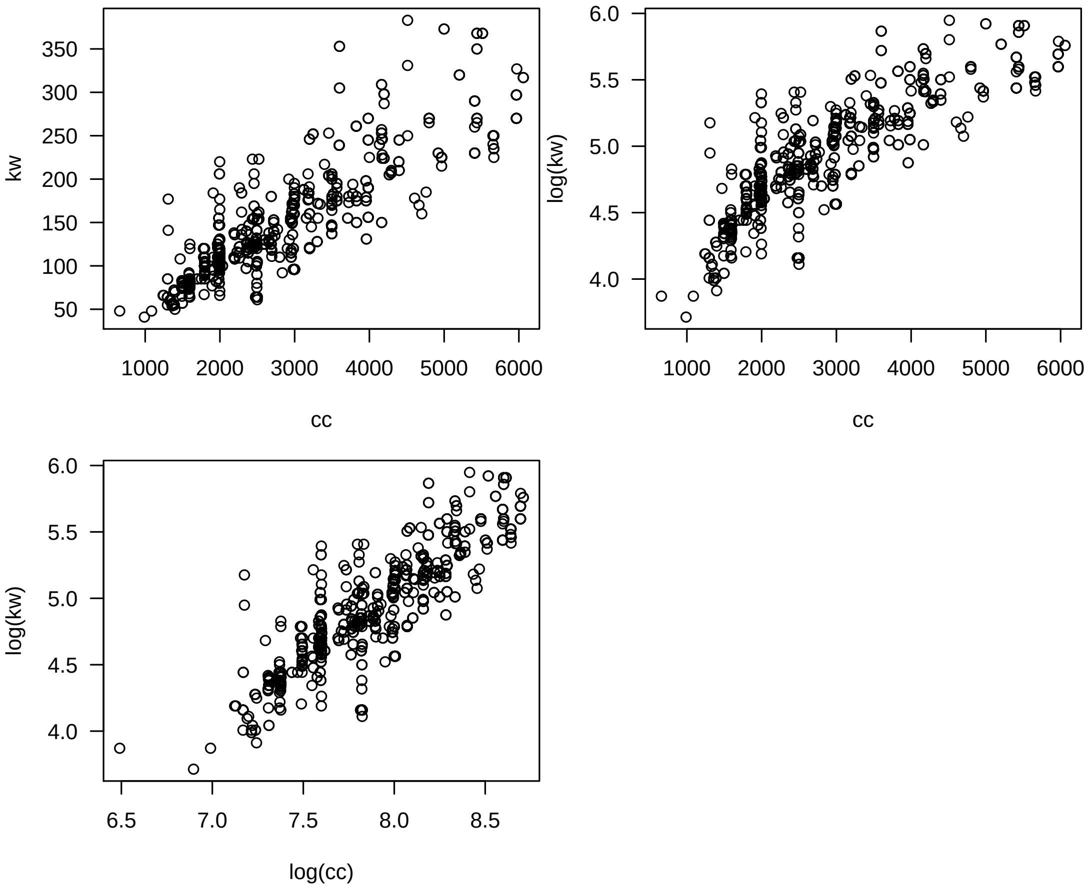
CONTENTS

NZ Car Data

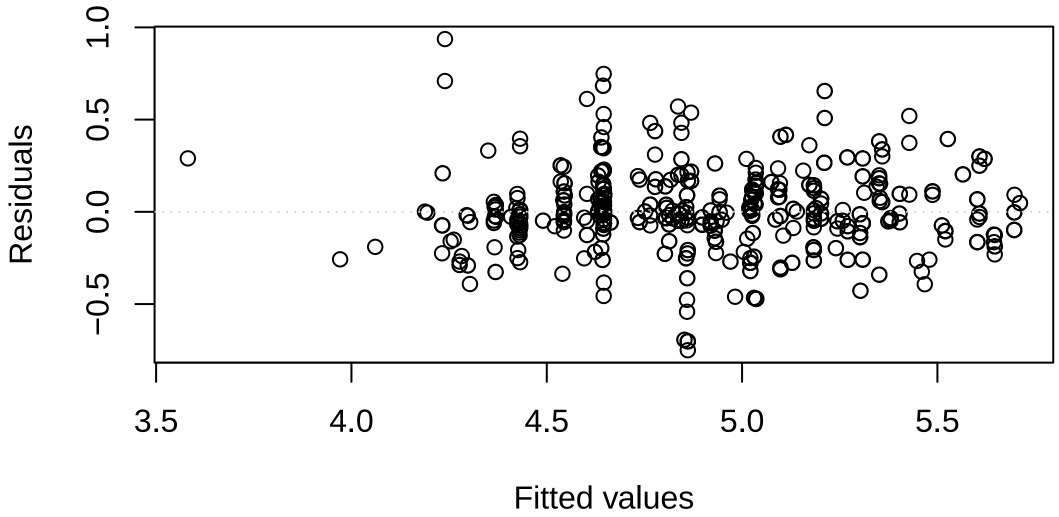
The New Zealand Automobile Association (AA) publish various data about models of car on their website. Although there are many variables available, here it is of interest to determine the relationship between the power output of the cars and the size of their engines.

The variables in the data set are

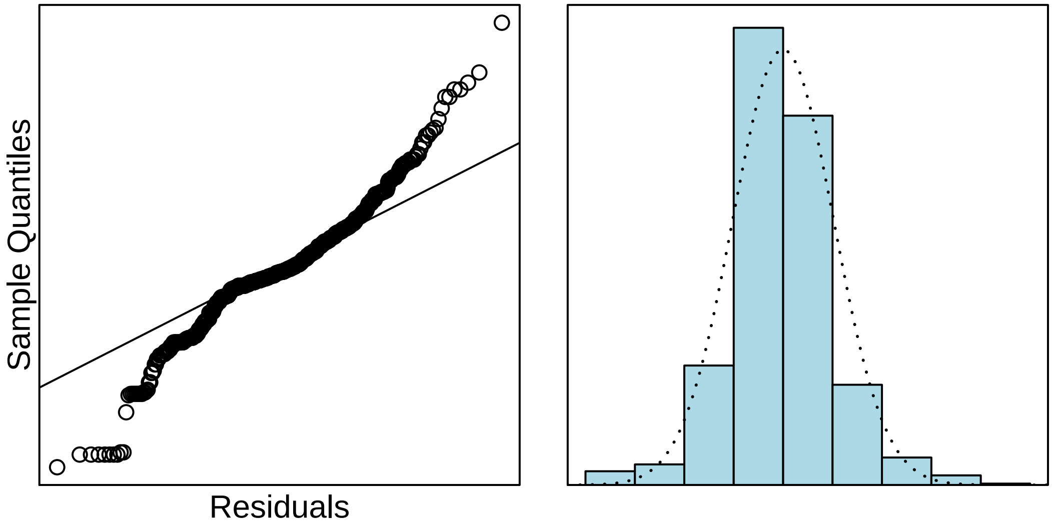
`kw`	The power output of the car measured in kilowatts (kW)
`cc`	The engine size measured in cubic centimetres (cm^3)



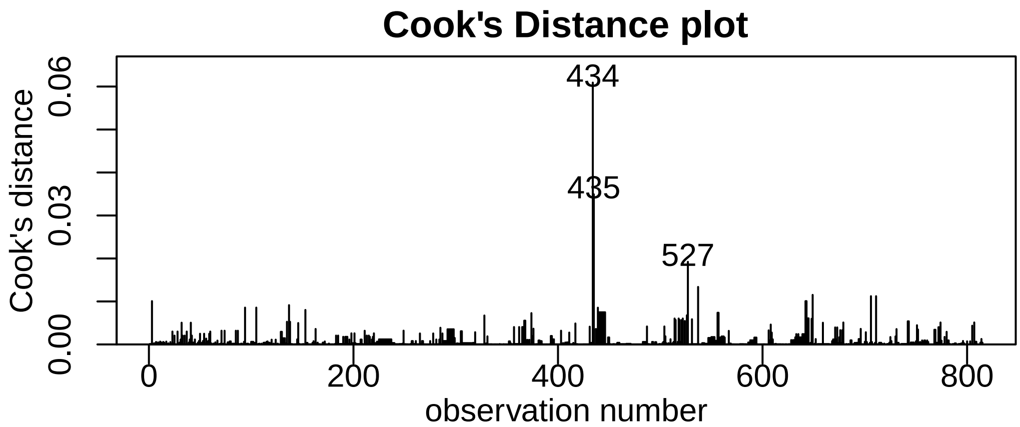
```
> nzcars.fit = lm(log(kw) ~ log(cc), data = nzcars.df)
> eovcheck(nzcars.fit)
```



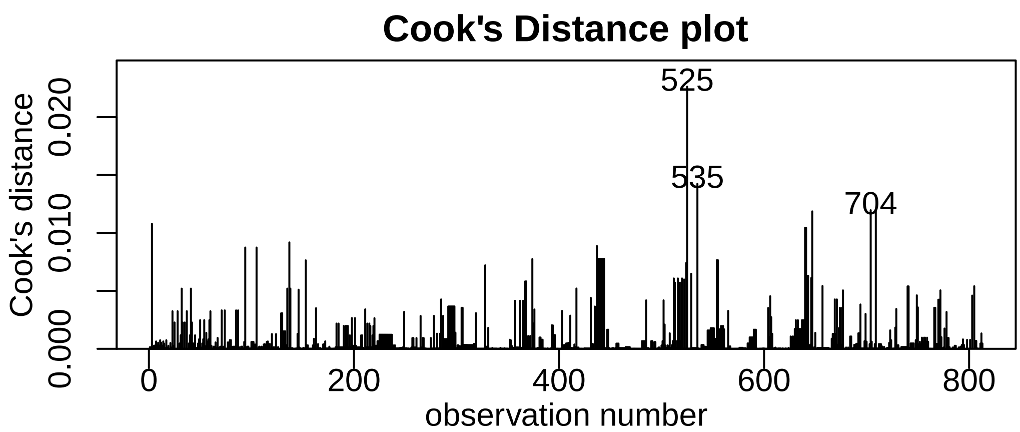

```
> normcheck(nzcars.fit, main = "", xlab = "Residuals")
```



```
> 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
```

Hull Damage Data

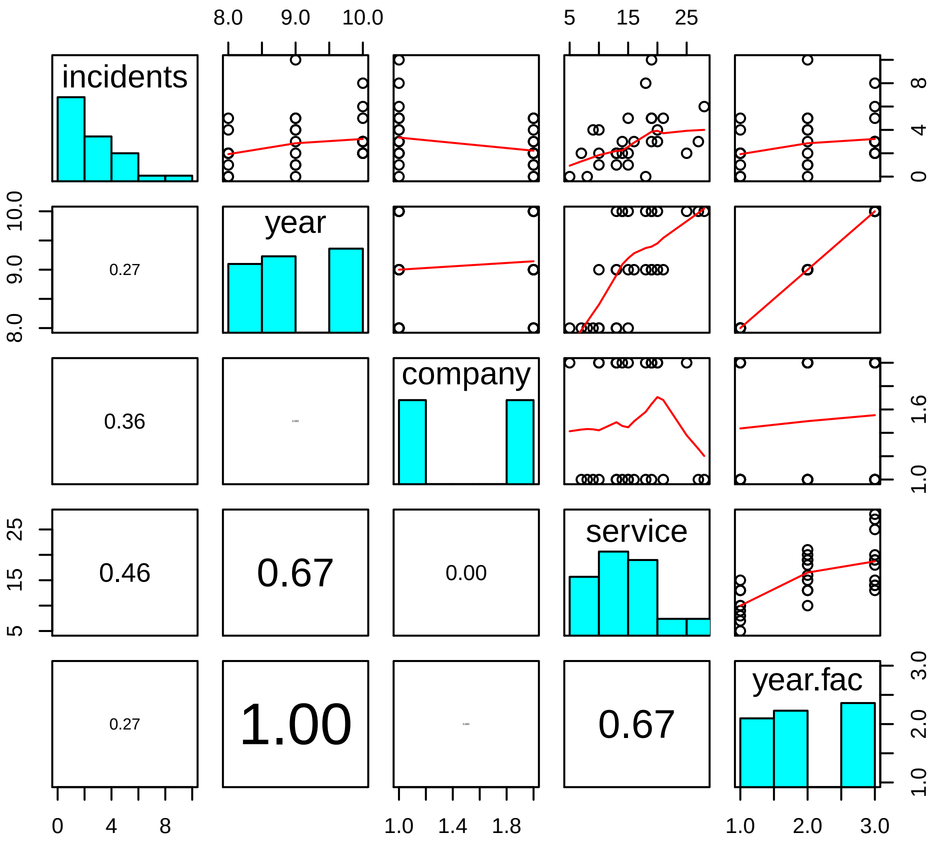
This question refers to data that come from a study investigating a particular type of minor damage caused by waves to the forward sections of ships' hulls. In total, 30 randomly selected ships were inspected for hull damage, and the number of damage incidents were recorded from each. Hull construction 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)
```

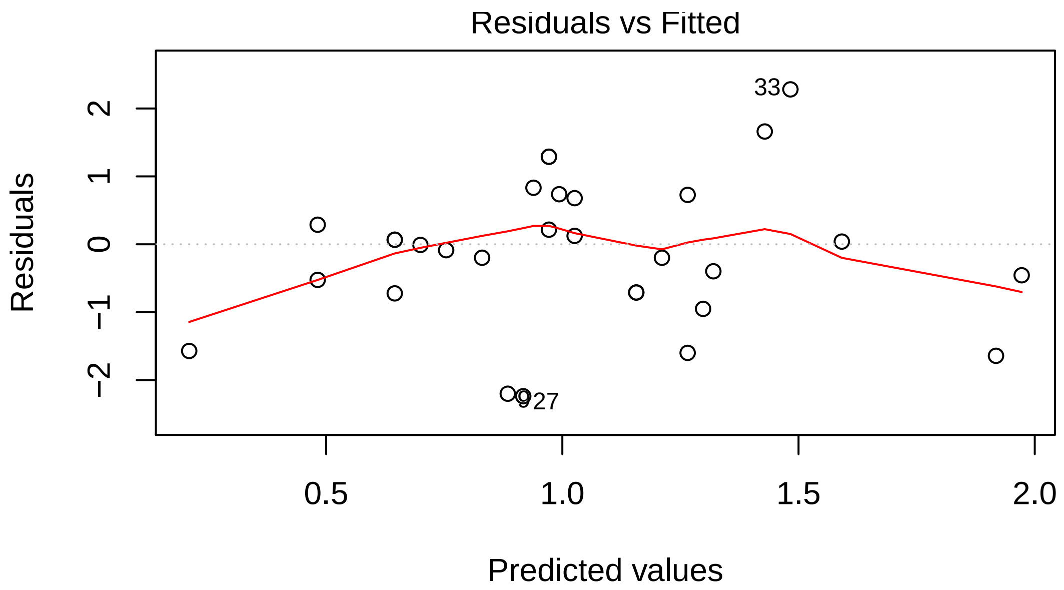



```
> ship1.fit <- glm(incidents ~ company + service + year.fac,
+                   family = "poisson", data = ship.df)
```

```
> 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
```

For Part (g) Only

```
> 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

```
> beer.table
           quality.poor quality.good
price.low            24           36
price.high           12           48
> chisq.test(beer.table)

Pearson's Chi-squared test with Yates' continuity correction

data:  beer.table
X-squared = 4.8016, df = 1, p-value = 0.02843
```