UMJI-SJTU

FALL 2020
Mock Exam

Applied Regression Using R

(**Time allowed:** 100 mins)

INSTRUCTIONS

- Answers should be written on the special **ANSWER BOOKLET** provided.
- Attempt **all** questions.

1. This question refers to the **Bikepath movement data** in **Appendix A**.
[Total 15 marks]

   a. Give a sensible reason why you would expect the usage of cycle paths to be lower on weekends/public holidays compared to regular weekdays. [2 marks]

   b. Write brief **Methods And Assumption Checks** for the analysis of the **Bikepath movement data** in **APPENDIX A**. [5 marks]

   c. Give a brief **Executive Summary** of the main conclusions of the analysis of the **Bikepath movement data** in **Appendix A**.
   Note: remember to address the questions asked. [7 marks]

   d. A log-normal model is an alternative model for this problem. When we fit Poisson regression we interpret in terms of expected (mean) counts. What do we interpret in terms when we fit a log-normal model? [1 mark]

2. This question refers to the **Coronary heart disease data** in **Appendix B**.
[Total 8 marks]

   a. Explain in one or two sentences why logistic regression is a sensible approach to use for this problem. [1 mark]

   b. Write down an equation of the final model fitted to the data, just as you would for a Method and Assumption Checks section. [2 marks]

   c. Give a brief **Executive Summary** of the main conclusions of the **Coronary heart disease data** analysis in **Appendix B**. [5 marks]

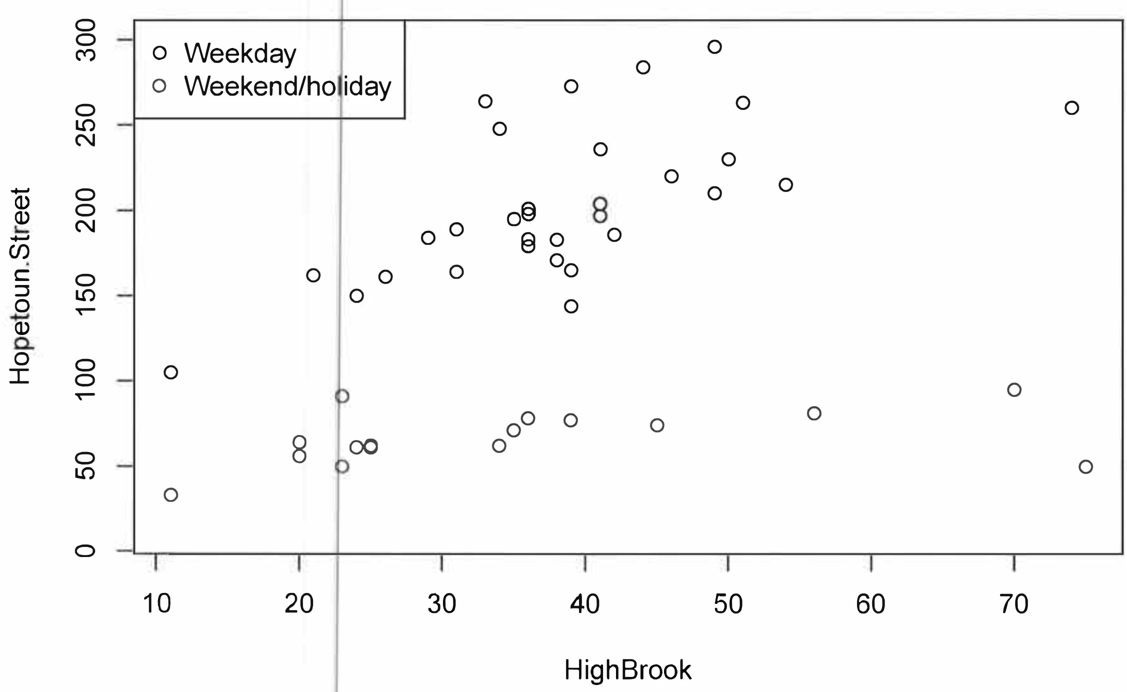
Appendix A Bikepath movement data

The number of cycle movements in Auckland is collected at sites across the region using permanent, automated cycle-monitoring equipment. This data is useful to monitor the usages of bikepaths and cycle traffic patterns. Summary of monthly/daily movements is available at `https://at.govt.nz/cycling-walking/research-monitoring/monthly-cycle-monitoring/`. 50 randomly selected days between 1 Jan 2019 to 31 August 2019 are analysed and data collected. The variables are:

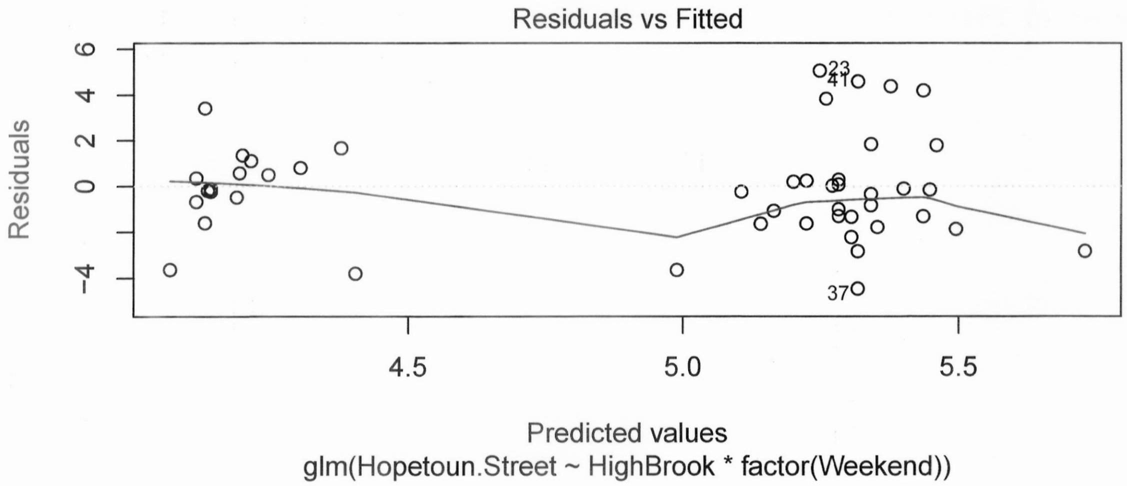
`Weekend`	indicator if day is weekend/public holiday (1 = weekends/public holidays and 0 = weekdays),
`HighBrook`	number of cycle movements on High Brook cycle path that day,
`Hopetoun.Street`	number of cycle movements on High Hopetoun Street cycle path that day.

A lecturer was interested how usage of cycle paths was related. Was the number of cyclists using the Hopetoun Street path related to the numbers using the High Brook cycle path? Also, did any relationship depend on the type of day: regular weekday verses weekends and public holiday days? How did usage differ between these types of days?

```
> plot(Hopetoun.Street~HighBrook,data=bike.df,col=1+as.numeric(Weekend)
+       ,ylim=c(10,300));
> legend("topleft",c("Weekday","Weekend/holiday"),pch=c(1,1),col=c(1:2))
```

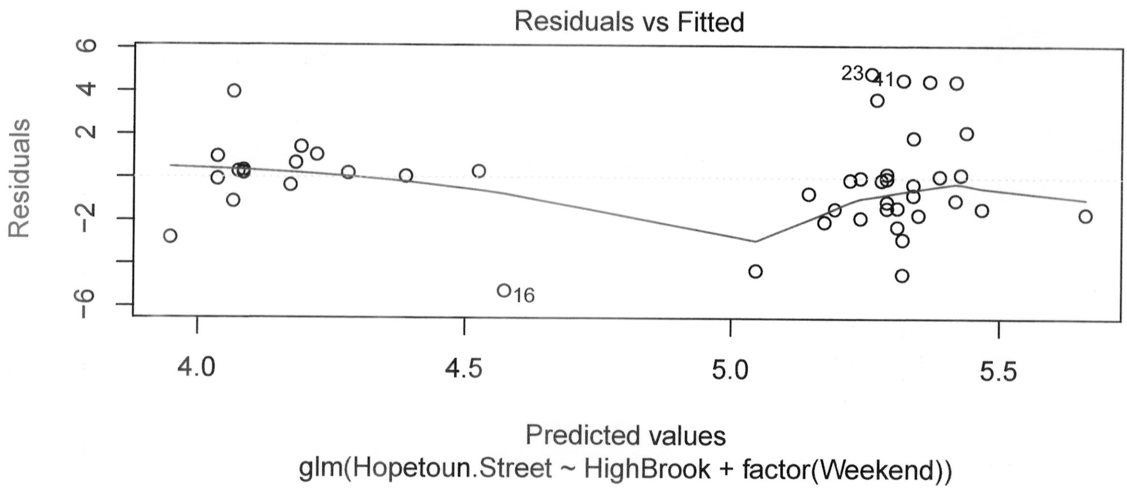


```
> bike.fit1 <- glm(Hopetoun.Street ~ HighBrook*factor(Weekend),
+                  family=poisson,data=bike.df)
> summary(bike.fit1)

Call:
glm(formula = Hopetoun.Street ~ HighBrook * factor(Weekend),
    family = poisson, data = bike.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.4450  -1.3334  -0.2217   0.7398   5.0512

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                 4.859711   0.044224 109.890  < 2e-16 ***
HighBrook                   0.011769   0.001074  10.961  < 2e-16 ***
factor(Weekend)1           -0.849313   0.081230 -10.456  < 2e-16 ***
HighBrook:factor(Weekend)1 -0.006515   0.001972  -3.303 0.000956 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1868.72  on 49  degrees of freedom
Residual deviance:  245.37  on 46  degrees of freedom
AIC: 592.06

Number of Fisher Scoring iterations: 4
> plot(bike.fit1, which = 1)
```



```
> 1-pchisq(245.37,46)
[1] 0


> bike.fit2 <- glm(Hopetoun.Street ~ HighBrook*factor(Weekend),
+                  family=quasipoisson,data=bike.df)
> summary(bike.fit2)

Call:
glm(formula = Hopetoun.Street ~ HighBrook * factor(Weekend),
    family = quasipoisson, data = bike.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.4450  -1.3334  -0.2217   0.7398   5.0512

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                  4.859711   0.103014  47.175  < 2e-16 ***
HighBrook                    0.011769   0.002501   4.705 2.35e-05 ***
factor(Weekend)1            -0.849313   0.189217  -4.489 4.78e-05 ***
HighBrook:factor(Weekend)1 -0.006515   0.004595  -1.418    0.163
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 5.426052)

    Null deviance: 1868.72  on 49  degrees of freedom
Residual deviance:  245.37  on 46  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```

```
> bike.fit3 <- glm(Hopetoun.Street ~ HighBrook+factor(Weekend),
+                  family=quasipoisson,data=bike.df)
> summary(bike.fit3)

Call:
glm(formula = Hopetoun.Street ~ HighBrook + factor(Weekend),
    family = quasipoisson, data = bike.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.2574  -1.3972  -0.0873   0.5820   4.8613

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       4.938377   0.087638  56.350  < 2e-16 ***
HighBrook         0.009778   0.002107   4.642 2.79e-05 ***
factor(Weekend)1 -1.098004   0.077285 -14.207  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 5.514623)

    Null deviance: 1868.72  on 49  degrees of freedom
Residual deviance:  256.44  on 47  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4


> plot(bike.fit3, which = 1)
```



```
> exp(confint(bike.fit3))
                       2.5 %      97.5 %
(Intercept)      117.5258690 165.7037415
HighBrook          1.0056357   1.0139743
factor(Weekend)1   0.2858264   0.3870313
```

```
> 100*(exp(confint(bike.fit3))-1)
                         2.5 %       97.5 %
(Intercept)      11652.5868978 16470.374150
HighBrook            0.5635692     1.397426
factor(Weekend)1   -71.4173645   -61.296869
```

```
> 100*(exp(confint(bike.fit3)[2,]*10)-1)
    2.5 %    97.5 %
 5.780786 14.886577
```

Appendix B Coronary heart disease data

A sample of 100 subjects from an at-risk population were tested for presence of significant coronary heart disease (CHD).

The data were grouped prior to analysis, so that subjects of the same age form a group. The data frame contains the following variables:

age	The age of subjects in a particular group in years.
n	The number of subjects in a particular age group.
y	The number of subjects in a particular age group with significant CHD.
p	The proportion of subjects in a particular age group with significant CHD (y/n).

It was of interest to see how age affected the risk of significant coronary heart disease for this at-risk population.

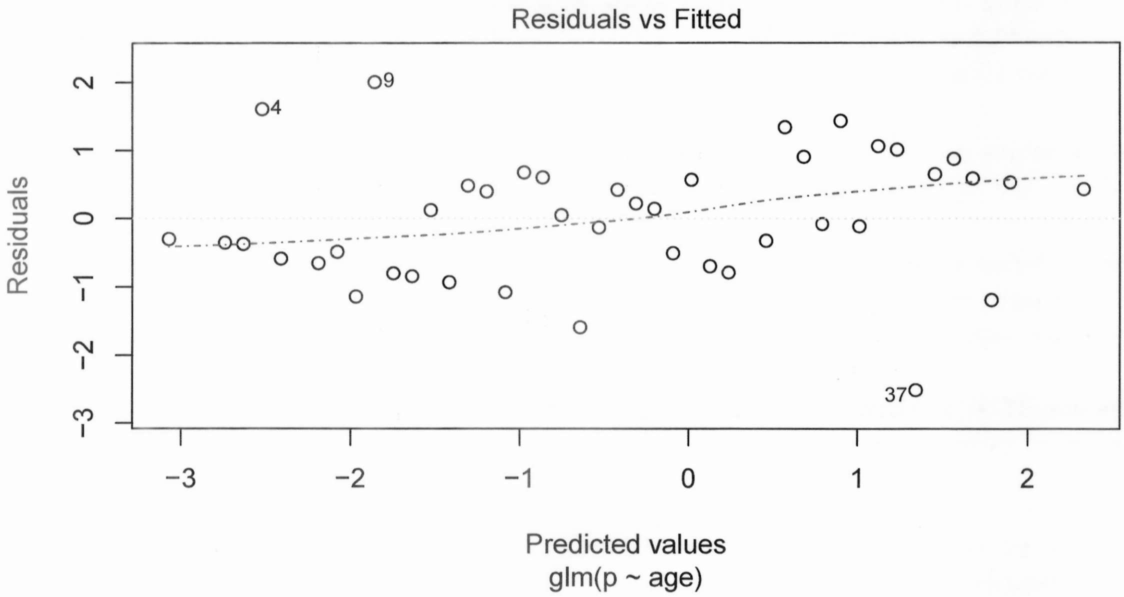
These data are taken from the textbook Hosmer and Lemeshow (2013), Applied Logistic Regression.

```
> head(CHD.df, 8)
  age y n   p
1  20 0 1 0.0
2  23 0 1 0.0
3  24 0 1 0.0
4  25 1 2 0.5
5  26 0 2 0.0
6  28 0 2 0.0
7  29 0 1 0.0
8  30 0 5 0.0
```

```
> tail(CHD.df, 8)
   age y n   p
36  59 2 2 1.0
37  60 0 2 0.0
38  61 1 1 1.0
39  62 2 2 1.0
40  63 1 1 1.0
41  64 1 2 0.5
42  65 1 1 1.0
43  69 1 1 1.0
```

```
> CHD.glm = glm(p ~ age,family=binomial, weight=n, data = CHD.df)
> summary(CHD.glm)

Call:
glm(formula = p ~ age, family = binomial, data = CHD.df, weights = n)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.50855  -0.61905   0.05056   0.59488   2.00169

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.27844    1.13053  -4.669 3.03e-06 ***
age          0.11032    0.02402   4.593 4.36e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 63.958  on 42  degrees of freedom
Residual deviance: 34.976  on 41  degrees of freedom
AIC: 69.31

Number of Fisher Scoring iterations: 4


> res.dev = summary(CHD.glm)$deviance
> dfr = summary(CHD.glm)$df.residual
> p.value = 1 - pchisq(res.dev, dfr)
> p.value
[1] 0.7344482
```

```
> plot(CHD.glm, which = 1, lty = 4)
```




```
> confint(CHD.glm)
                  2.5 %     97.5 %
(Intercept) -7.68684591 -3.2196472
age          0.06638542  0.1612921
```

```
> exp(confint(CHD.glm))
                   2.5 %     97.5 %
(Intercept) 0.0004588231 0.03996916
age         1.0686385171 1.17502820
```

```
> 100 * (exp(confint(CHD.glm)[2, ]) - 1)
    2.5 %    97.5 %
 6.863852 17.502820
```