STAT3015/4030/7030 Generalised Linear Modelling

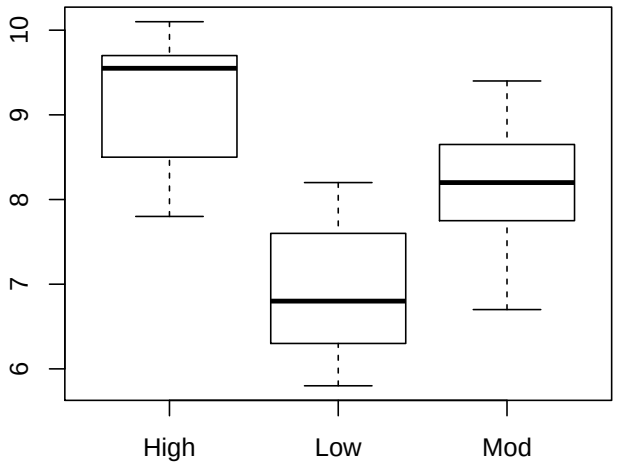
Tutorial 2

1. The file `productivity.csv` contains data regarding improvements in productivity for 27 business firms. Each firm was classifed according to whether their average expenditure for research and development in the past three years was high, moderate or low. In addition, the productivity improvement (measured on a scale of 0 to 100) was recorded for each firm.

   (a) Plot the productivity improvement scores versus the level of R&D expenditure. Do you think that heteroscedasticity will be a problem?

   Solution:

   ```
   > prd <- read.table("productivity.csv", header=T, sep=",")
   > attach(prd)
   > names(prd)
   > plot(RandD, prodscre)
   ```

   Notice that when given a categorical variable, the `plot(RandD, prodscre)` function creates side-by-side boxplots. However, we can still use this plot to investigate the possibility of heteroscedasticity, and it appears that the spread of each of the three boxplots is very similar, implying that homoscedasticity is a reasonable assumption.

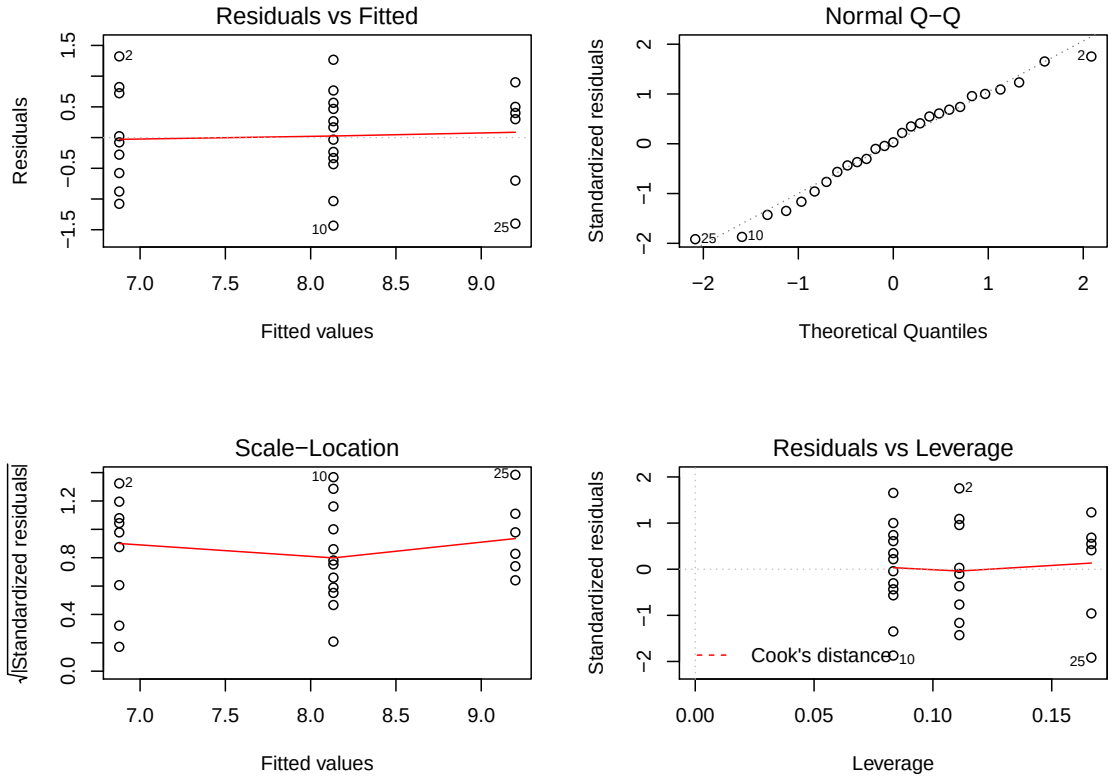
   (b) Fit a one-way analysis of variance model and test whether there is a difference between the expected productivity increases for each of the three factor levels.

Construct a normal q-q plot to investigate the suitability of the normal error assumption.

Solution:

```
> m1 <- lm(prodscre~RandD, data=prd)
> anova(m1)
Analysis of Variance Table

Response: prodscre
          Df Sum Sq Mean Sq F value    Pr(>F)
RandD      2 20.125 10.0626   15.72 4.331e-05 ***
Residuals 24 15.362  0.6401
> par(mfrow=c(2,2))
> plot(m1)
```



Clearly, there is a signficant difference in the productivity improvement among firms with different levels of research and development funding (p-value $= 4.331 \times 10^{-5}$). Also, the normal q-q plot indicates that our basic assumptions of normality and homoscedasticity appear to be reasonable.

(c) Suppose we wish to test whether the difference in productivity improvement between firms with low and moderate levels of research and development funding is the same as the difference in productivity improvement for firms with moderate and

high R&D funding (note that this is a sort of analog to linearity in the ANOVA setting). Write down the appropriate null hypothesis using parameters defined for your model. Test the null hypothesis at the $\alpha = 0.01$ level.

Solution:

The baseline level fit in R is High (you can see this by typing in the command `model.matrix(m1)`; also note from an alphabetical ordering of the factor levels, R takes the baseline level to be the first level by default). It may be more intuitive to recode R&D='Low' to be the baseline level. We are asked to test the null hypothesis $H_0 : \mu_{Mod} - \mu_{Low} = \mu_{High} - \mu_{Mod}$. This is equivalent to $H_0 : \mu_{High} - 2\mu_{Mod} + \mu_{Low} = 0$. In terms of regression coefficients, the null hypothesis is $H_0 : \beta_0 + \beta_1 - 2(\beta_0 + \beta_2) + \beta_0 = 0$. That is $H_0 : \beta_1 - 2\beta_2 = 0$. The R code is:

```
> p.high<-ifelse(RandD=="High",1,0)
> p.mod<-ifelse(RandD=="Mod",1,0)
> m1<-lm(prodscre~p.high+p.mod)
> anova(m1)
Analysis of Variance Table

Response: prodscre
          Df  Sum Sq Mean Sq F value    Pr(>F)
p.high     1 12.0179 12.0179  18.775 0.0002265 ***
p.mod      1  8.1073  8.1073  12.666 0.0015921 **
Residuals 24 15.3622  0.6401

> coef(m1)
(Intercept)      p.high       p.mod
   6.877778    2.322222    1.255556
> h <- c(0,1,-2)
> est<-t(h)%*%coefficients(m1)
> est
           [,1]
[1,] -0.1888889
> sigma<-summary(m1)$sigma
> Xmat <- cbind(1,p.high,p.mod)
> XtXi <- solve(t(Xmat)%*%Xmat)
> sd<-sigma*sqrt(t(h)%*%XtXi%*%h)
> sd
         [,1]
[1,] 0.625434
> qt(0.975,m1$df.residual)
[1] 2.063899
>
> upper <- est + (qt(0.975,m1$df.residual)*sd)
> lower <- est - (qt(0.975,m1$df.residual)*sd)
```

```
> c(lower,est,upper)
[1] -1.4797212 -0.1888889  1.1019435
> pval <- 2*(1-pt(abs(est/sd),m1$df.residual))
> pval
          [,1]
[1,] 0.7652442
```

The p-value of the test is 0.76, so, we cannot reject the null hypothesis. In other words, it is plausible that the increases in productivity across "consecutive" levels of R & D funding are equal. We can also see this conclusion from the 95% confidence interval estimate which contains the value zero.

Instead of performing the test using the coefficient estimates of the regression model using indicator variables, we can calculate the treatment effect estimates directly from the data, because the least squares estimates of the treatment effects are simply the cell means. The R-code is

```
> m1<-lm(prodscre~RandD,data=prd)
> Ybar <- mean(prodscre)
> ni <- tapply(prodscre,RandD,length)
> Ybari <- tapply(prodscre,RandD,mean)
> Ybari
    High      Low      Mod
9.200000 6.877778 8.133333
> MSE <- summary(m1)$sigma^2
>
> h <- c(1,1,-2)
> est <- as.vector(t(h)%*%Ybari)
> sd <- sqrt(MSE)*sqrt(sum((h^2)/ni))
> pval <- 2*(1-pt(abs(est/sd),24))
> cbind(est,sd,pval)
            est       sd      pval
[1,] -0.1888889 0.625434 0.7652442.
```

2. There are 8 distinct blood types among human beings, classified according to whether certain proteins are present or not. The blood types are: O-, O+, A-, A+, B-, B+, AB-, AB+. The appearance of an A in the name of the blood type indicates that the individual's blood contains the A-antigen, while the appearance of a B in the name indicates that the blood contains the B-antigen, and a + in the name indicates the existence of the so-called Rhesus (or Rh) factor in the blood. Thus, a person with AB- blood has both antigens in their blood stream, but no Rh factor, while a person with O+ blood has neither the A-antigen or the B-antigen, but does have the Rh factor. Suppose that a random sample of people is gathered and each person's blood type as well as the value for some quantitative biological trait are measured. The data are:

A- B- Rh
0 0 0
0 0 1
1 0 0
1 0 1
0 1 0
0 1 1
1 1 0
1 1 1

Blood Type	Responses			
O-	9	11		
O+	20	19	23	19
A-	12	10		
A+	17	18	21	20
B-	16			
B+	24	28	25	
AB-	15			
AB+	25			

(a) Fit a one-way ANOVA model to this data and test whether there is any difference in the response variable for individuals of different blood types.

Solution:

```
> resp <- c(9,11,20,19,23,19,12,10,17,18,21,20,16,24,28,25,15,25)
> btyp <- c(rep("O-",2),rep("O+",4),rep("A-",2),rep("A+",4),"B-",
    rep("B+",3),"AB-","AB+")
> btyp.aov <- aov(resp ~ btyp)
Warning message:
In model.matrix.default(mt, mf, contrasts) :
  variable 'btyp' converted to a factor
> summary(btyp.aov)
            Df Sum Sq Mean Sq F value   Pr(>F)
btyp         7  485.0   69.29   20.73 3.41e-05 ***
Residuals   10   33.4    3.34
```

So, there is a significant difference in response for the various blood groups (p-value $= 3.41 \times 10^{-5}$)

(b) Construct a linear combination of the μ_i's to see whether the ability to produce the A- antigen has any effect on the response variable. Repeat this process for the B-antigen and the Rh factor.

Solution:
The desired linear combination of the μ_i's is just the average of all the blood types containing that antigen, minus the average of those without.

```
> ni <- tapply(resp,btyp,length)
> lvl.mns <- tapply(resp,btyp,mean)
> lvl.mns
      A-       A+      AB-      AB+       B-       B+       O-
11.00000 19.00000 15.00000 25.00000 16.00000 25.66667 10.00000
      O+
20.25000
> h <- c(1/4,1/4,1/4,1/4,-1/4,-1/4,-1/4,-1/4) ##test A-antigen
> MSE <- sum((resp-fitted(btyp.aov))^2)/btyp.aov$df.residual
> est <- as.vector(t(h)%*%lvl.mns)
> sd <- sqrt(MSE)*sqrt(sum((h^2)/ni))
> pval <- 2*(1-pt(abs(est/sd),10))
> cbind(est,sd,pval)
            est      sd      pval
[1,] -0.4791667 1.00472 0.6436729
>
> h <- c(-1/4,-1/4,1/4,1/4,1/4,1/4,-1/4,-1/4) ##test B-antigen
> est <- as.vector(t(h)%*%lvl.mns)
> sd <- sqrt(MSE)*sqrt(sum((h^2)/ni))
> pval <- 2*(1-pt(abs(est/sd),10))
> cbind(est,sd,pval)
          est      sd        pval
[1,] 5.354167 1.00472 0.000333429
>
> h <- c(-1/4,1/4,-1/4,1/4,-1/4,1/4,-1/4,1/4) ##test Rh-factor
> est <- as.vector(t(h)%*%lvl.mns)
> sd <- sqrt(MSE)*sqrt(sum((h^2)/ni))
> pval <- 2*(1-pt(abs(est/sd),10))
> cbind(est,sd,pval)
          est      sd         pval
[1,] 9.479167 1.00472 2.702608e-06
```

So, it seems that the A-antigen is not significantly related to the response variable (p-value = 0.64), but the B-antigen (p-value = 0.00033) and Rh factor are (p-value $= 2.7 \times 10^{-6}$).

(c) Suppose that we now believe that the ability to make the A-antigen is not important in explaining variation in the response. Create a new factor based on the original one which has only 4 levels and ignores the ability to make the A-antigen. Refit an ANOVA model using this new factor and test whether there is any overall significance in explaining the response variation. Also, construct constrasts for this

new model to test whether the ability to make the B-antigen and the presence of the Rh factor have the same explanatory power.

Solution:

```
> btyp1 <- btyp
> btyp1 <- ifelse(btyp1=="A-","O-",btyp1)
> btyp1 <- ifelse(btyp1=="A+","O+",btyp1)
> btyp1 <- ifelse(btyp1=="AB-","B-",btyp1)
> btyp1 <- ifelse(btyp1=="AB+","B+",btyp1)
> btyp1.aov <- aov(resp ~ btyp1)
Warning message:
In model.matrix.default(mt, mf, contrasts) :
  variable 'btyp1' converted to a factor
> summary(btyp1.aov)
            Df Sum Sq Mean Sq F value  Pr(>F)
btyp1        3  480.1  160.02   58.38 3.7e-08 ***
Residuals   14   38.4    2.74
```

So, clearly there is still a significant relationship (p-value $= 3.7 \times 10^{-8}$). To investigate whether both factors have the same effect level test

$$H_0 : \frac{(\mu_{B+} + \mu_{O+})}{2} - \frac{(\mu_{B-} + \mu_{O-})}{2} = \frac{(\mu_{B+} + \mu_{B-})}{2} - \frac{(\mu_{O+} + \mu_{O-})}{2}$$

or simplifying

$$H_0 : \mu_{O+} - \mu_{B-} = 0$$

```
> ni <- tapply(resp,btyp1,length)
> lvl.mns <- tapply(resp,btyp1,mean)
> lvl.mns
    B-     B+     O-     O+
15.500 25.500 10.500 19.625
> MSE <- sum((resp-fitted(btyp1.aov))^2)/btyp1.aov$df.residual
> h <- c(-1,0,0,1)
> est <- as.vector(t(h)%*%lvl.mns)
> sd <- sqrt(MSE)*sqrt(sum((h^2)/ni))
> pval <- 2*(1-pt(abs(est/sd),14 ))
> cbind(est,sd,pval)
       est       sd        pval
[1,] 4.125 1.308881 0.007069867
```

And thus it does appear that the two blood factors have different level effects (p-value $= 0.007069867$)