STAT3015/4030/7030 Generalised Linear Modelling

Tutorial 10

1. **Emulating Jane Austen's Writing Style** (Ramsey and Schafer, 2013). When she died in 1817, the English novelist Jane Austen had not yet finished the novel *Sanditon*, but she did leave notes on how she intended to conclude the book. The novel was completed by a ghost writer, who attempted to emulate (that is, copy or imitate) Austen's style. In 1978, a researcher reported counts of some words found in chapters of books written by Austen and in chapters written by the ghost writer. The data is in the file `austen.txt` .

Word	**Book**			
	Sense and Sensibility	Emma	Sanditon I	Sanditon II
a	147	186	101	83
an	25	26	11	29
this	32	39	15	15
that	94	105	37	22
with	59	74	28	43
without	18	10	10	4

Was Jane Austen consistent in the three books in her relative uses of these words? Did the ghost writer do a good job in terms of matching the relative rates of occurrence of these six words? In particular, did the ghost writer match the relative rates that Austen used the words in the first part of *Sanditon*?

Solution: Poisson distribution is appropriate to model the response variable `COUNT` in this problem. Why?

```
> jau <- read.table("austen.txt", header=TRUE)
> jau

   COUNT              BOOK    WORD
1    147 Sense&Sensibility       a
2     25 Sense&Sensibility      an
3     32 Sense&Sensibility    this
4     94 Sense&Sensibility    that
```

```
5     59 Sense&Sensibility    with
6     18 Sense&Sensibility without
7    186              Emma       a
8     26              Emma      an
9     39              Emma    this
10   105              Emma    that
11    74              Emma    with
12    10              Emma without
13   101         SanditonI       a
14    11         SanditonI      an
15    15         SanditonI    this
16    37         SanditonI    that
17    28         SanditonI    with
18    10         SanditonI without
19    83        SanditonII       a
20    29        SanditonII      an
21    15        SanditonII    this
22    22        SanditonII    that
23    43        SanditonII    with
24     4        SanditonII without
```

```
> # Create new indicator variable for BOOK=SanditonII.
> # This is the book written by the ghost writer.
> jau$SAND <- as.numeric(jau$BOOK=="SanditonII")
```

(a) First we see whether Jane Austen was herself consistent by testing the interaction terms in her books only:

```
> # use the subset argument to restrict the data to Jane Austen's books only.
> jaglm <- glm(COUNT~BOOK*WORD, data=jau, family=poisson, subset=(SAND==0))
> jaglmr <- glm(COUNT~BOOK+WORD, data=jau, family=poisson, subset=(SAND==0))
> anova(jaglmr, jaglm, test="Chisq")

Analysis of Deviance Table

Model 1: COUNT ~ BOOK + WORD
Model 2: COUNT ~ BOOK * WORD
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        10     12.587
2         0      0.000 10   12.587   0.2477
```

The drop in deviance test for the significance of the interaction term has p-value $= 0.2477$. So there is no evidence that Austen herself was inconsistent with her use of words across her three books considered here.

(b) Next, we want to see if the ghost writer in the book *Sanditon II* matched Jane Austen's usage of the 6 words. To answer the question of interest, we create an indicator variable SAND taking value 1 if the book is Sanditon II, and zero otherwise. Then we fit the model:

$$\log \text{Count} = \text{Book} + \text{Word} + \text{Sand:Word}$$

and test if the interaction between Word and Sand is significant.

```
> # Create indicator variables for Sense&Sensibility and Emma.
> # This makes Sandition I the baseline JaneAusten book.
> First <- as.numeric(jau$BOOK=="Sense&Sensibility")
> Second <- as.numeric(jau$BOOK=="Emma")
> jglm <- glm(COUNT~First+Second+SAND+WORD+WORD:SAND,
                family=poisson, data=jau)
> summary(jglm)

Call:
glm(formula = COUNT ~ First + Second + SAND + WORD + WORD:SAND,
    family = poisson, data = jau)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7140  -0.2153   0.0000   0.3288   1.5512

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       4.45670    0.07919  56.277  < 2e-16 ***
First             0.61866    0.08728   7.088 1.36e-12 ***
Second            0.77851    0.08499   9.160  < 2e-16 ***
SAND             -0.03786    0.13535  -0.280  0.77970
WORDan           -1.94591    0.13577 -14.333  < 2e-16 ***
WORDthat         -0.60921    0.08088  -7.532 4.98e-14 ***
WORDthis         -1.61870    0.11803 -13.714  < 2e-16 ***
WORDwith         -0.99164    0.09228 -10.746  < 2e-16 ***
WORDwithout      -2.43546    0.16917 -14.396  < 2e-16 ***
SAND:WORDan       0.89437    0.25488   3.509  0.00045 ***
SAND:WORDthat    -0.71859    0.25307  -2.839  0.00452 **
SAND:WORDthis    -0.09209    0.30438  -0.303  0.76222
SAND:WORDwith     0.33400    0.20933   1.596  0.11059
SAND:WORDwithout -0.59709    0.53914  -1.107  0.26808
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

    Null deviance: 878.858  on 23  degrees of freedom
Residual deviance:  12.587  on 10  degrees of freedom
AIC: 169.18

Number of Fisher Scoring iterations: 4
```

```
> jglmr <- glm(COUNT~BOOK+WORD, data=jau, family=poisson)
> anova(jglmr, jglm, test="Chisq")

Analysis of Deviance Table

Model 1: COUNT ~ BOOK + WORD
Model 2: COUNT ~ First + Second + SAND + WORD + WORD:SAND
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1        15     44.324
2        10     12.587  5   31.737 6.699e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Whereby we see that the interaction terms are overwhelmingly significant (p-value < 0.001). This signifies a clear difference between the writing styles of the ghost writer and Jane Austen. From examining the coefficients of the full model we note that this is principally due to the authors' use of the words 'an' and 'that'.

(c) To address the question specific to the *Sanditon* volumes:

```
> # The subset command here resticts books to either SanditonI OR SanditonII
> jgglm <- glm(COUNT~BOOK*WORD, data=jau, family=poisson,
+              subset=((BOOK=="SanditonI")|(BOOK=="SanditonII")))
> summary(jgglm)

Call:
glm(formula = COUNT ~ BOOK * WORD, family = poisson, data = jau,
    subset = ((BOOK == "SanditonI") | (BOOK == "SanditonII")))

Deviance Residuals:
 [1]  0  0  0  0  0  0  0  0  0  0  0  0

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                 4.6151     0.0995  46.381  < 2e-16 ***
BOOKSanditonII             -0.1963     0.1482  -1.325  0.18522
WORDan                     -2.2172     0.3175  -6.983 2.88e-12 ***
WORDthat                   -1.0042     0.1922  -5.226 1.74e-07 ***
```

```
WORDthis                    -1.9071     0.2767  -6.892 5.50e-12 ***
WORDwith                    -1.2829     0.2136  -6.007 1.89e-09 ***
WORDwithout                 -2.3125     0.3315  -6.976 3.04e-12 ***
BOOKSanditonII:WORDan        1.1657     0.3839   3.037  0.00239 **
BOOKSanditonII:WORDthat     -0.3236     0.3073  -1.053  0.29232
BOOKSanditonII:WORDthis      0.1963     0.3941   0.498  0.61842
BOOKSanditonII:WORDwith      0.6253     0.2845   2.198  0.02794 *
BOOKSanditonII:WORDwithout  -0.7200     0.6099  -1.181  0.23777
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance:  2.5951e+02  on 11  degrees of freedom
Residual deviance: -6.6613e-16  on  0  degrees of freedom
AIC: 83.881

Number of Fisher Scoring iterations: 3
```

```
> jgglmr <- glm(COUNT~BOOK+WORD, data=jau, family=poisson,
+               subset=((BOOK=="SanditonI")|(BOOK=="SanditonII")))
> anova(jgglmr, jgglm, test="Chisq")

Analysis of Deviance Table

Model 1: COUNT ~ BOOK + WORD
Model 2: COUNT ~ BOOK * WORD
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         5     19.777
2         0      0.000  5   19.777 0.001376 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Restricting analysis to the *Sanditon* volumes, we see that the interaction term between `Word` and `Book` is significant, indicating that the ghost writer failed to match Austen's writing style from the first part of Sanditon.

2. **El Nino and Hurricanes** (Ramsey and Schafer, 2013). The data set `elnino.txt` contains data on the number of tropical storms and hurricanes each year from 1950 to 1997. Data is also recorded on whether the year was a cold, warm or neutral El Nino year; a constructed numerical variable `temperature` that takes on the values -1, 0, and 1, according to whether the El Nino temperature is cold, neutral or warm; and a variable indicating whether West Africa was wet or dry that year. It is thought that wet years in West Africa often bring more hurricanes, and that the warm phase of El Nino suppresses

hurricanes while a cold phase encourages them. Use a poisson log-linear regression to describe the distribution of (a) number of storms and (b) number of hurricanes as a function of `temperature` and West African wetness.

Solution:

(a) We start with the following model:

```
> eln <- read.table("elnino.txt", header=TRUE)
> eglm1 <- glm(storms~temperature+west.africa, data=eln, family=poisson)
> summary(eglm1)

Call:
glm(formula = storms ~ temperature + west.africa, family = poisson,
    data = eln)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3791  -0.5243  -0.1100   0.3908   1.9735

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.16880    0.06324  34.295  < 2e-16 ***
temperature -0.18158    0.06280  -2.891  0.00384 **
west.africa  0.14435    0.10243   1.409  0.15876
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 50.875  on 47  degrees of freedom
Residual deviance: 35.223  on 45  degrees of freedom
AIC: 235.14

Number of Fisher Scoring iterations: 4

> eglm1r <- glm(storms~west.africa, data=eln, family=poisson)
> anova(eglm1r, eglm1, test="Chisq")

Analysis of Deviance Table

Model 1: storms ~ west.africa
Model 2: storms ~ temperature + west.africa
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        46     43.662
2        45     35.223  1   8.4389 0.003673 **
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> 1-exp(coef(eglm1)[2]+c(1,0,-1)*1.96*sqrt(diag(vcov(eglm1))[2]))

[1] 0.05681291 0.16604882 0.26263349
```

Hence, it appears that El Nino temperature does affect the number of storms, even after accounting for West African wetness (p-value=0.004). As the temperature shifts from neutral to warm, or from cold to neutral we expect a drop of $1 - e^{-0.18}$ i.e. 17% in the mean number of storms (CI: (5.7%,26.2%)).

(b) Again we fit the basic model:

```
> eglm2 <- glm(hurricanes~temperature+west.africa, data=eln, family=poisson)
> summary(eglm2)

Call:
glm(formula = hurricanes ~ temperature + west.africa, family = poisson,
    data = eln)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.18542  -0.60539  -0.08583   0.38731   1.99350

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.64620    0.08208  20.056   <2e-16 ***
temperature -0.21729    0.08072  -2.692   0.0071 **
west.africa  0.20233    0.13045   1.551   0.1209
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 44.414  on 47  degrees of freedom
Residual deviance: 29.425  on 45  degrees of freedom
AIC: 205.26

Number of Fisher Scoring iterations: 4

> eglm2r <- glm(hurricanes~west.africa, data=eln, family=poisson)
> anova(eglm2r, eglm2, test="Chisq")

Analysis of Deviance Table
```

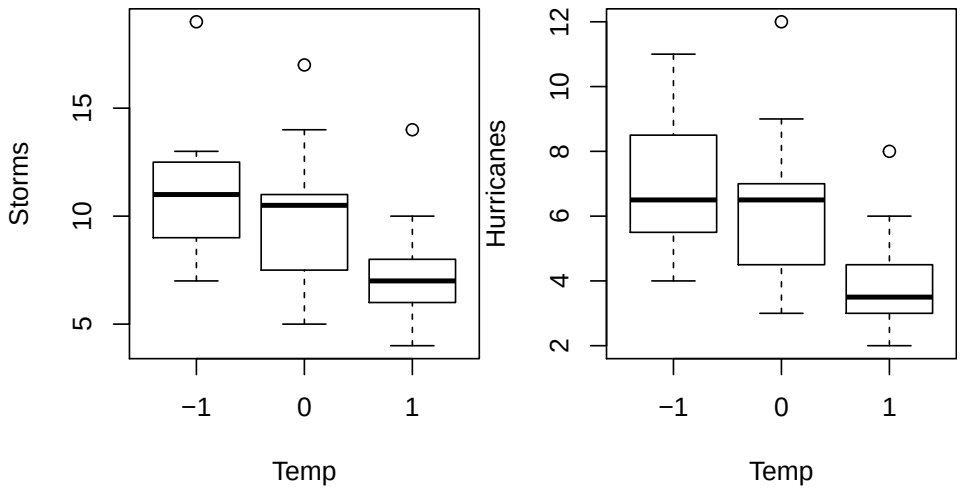
```
Model 1: hurricanes ~ west.africa
Model 2: hurricanes ~ temperature + west.africa
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        46     36.771
2        45     29.425  1   7.3461 0.006721 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> 1-exp(coef(eglm2)[2]+c(1,0,-1)*1.96*sqrt(diag(vcov(eglm2))[2]))
```

```
[1] 0.0573696 0.1953033 0.3130534
```

Hence, it appears that El Nino temperature does affect the number of hurricanes, even after accounting for West African wetness (p-value=0.007). As the temperature shifts from neutral to warm or from cold to neutral we expect a drop of $1 - e^{-0.22}$ i.e. 19.5% in the mean number of storms (CI: (5.7%, 31.3%)).

```
> par(mar=c(2, 4, 0, 0), mfrow=c(1, 2), pty="s")
> plot(factor(eln$temp), eln$storms, xlab="Temp", ylab="Storms")
> plot(factor(eln$temp), eln$hurricanes, xlab="Temp", ylab="Hurricanes")
```



3. Schriener, Gregoire and Lawrie (1962) conducted an experiment to examine the effect of supposedly inert gases on fungal growth. The molecular weight (MW) of each inert gas and the fungal growth rate (in millimeters per hour) for 10 samples are given below:

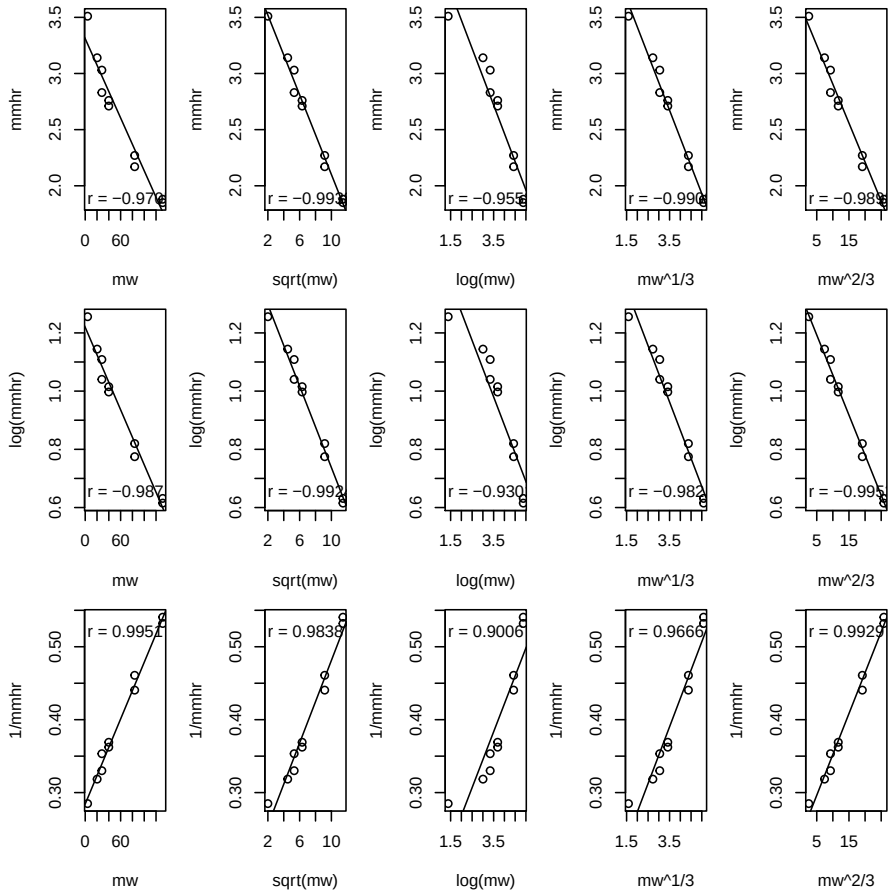
Gas	He	Ne	N_2	N_2	Ar	Ar	Kr	Kr	Xe	Xe
MW	4.0	20.2	28.2	28.2	39.9	39.9	83.8	83.8	131.3	131.3
mm/hr	3.51	3.14	3.03	2.83	2.71	2.76	2.27	2.17	1.88	1.85

Schriener et al. report a relationship of: $mm/hr = 3.87 - 0.1774\sqrt{MW}$.

(a) Plot each of three transformations of the response variable $(y, \log y, 1/y)$ versus each of five transformations of the predictor variable $(x, \sqrt{x}, \log x, x^{1/3}, x^{2/3})$ and include the correlation cofficient and normal linear regression line on each of the fifteen plots. Which of the relationships seem the most linear? Is the Schriener et al. model one of these?

Solution:

```
> mw <- c(4.0, 20.2, 28.2, 28.2, 39.9, 39.9, 83.8, 83.8, 131.3, 131.3)
> mmhr <- c(3.51, 3.14, 3.03, 2.83, 2.71, 2.76, 2.27, 2.17, 1.88, 1.85)
> xs <- cbind(mw, sqrt(mw), log(mw), mw^(1/3), mw^(2/3))
> # Assign names to columns of xs
> dimnames(xs)[[2]] <- c("mw", "sqrt(mw)", "log(mw)", "mw^1/3", "mw^2/3")
> ys <- cbind(mmhr, log(mmhr), 1/mmhr)
> # assign names to columns of ys
> dimnames(ys)[[2]] <- c("mmhr", "log(mmhr)", "1/mmhr")
> par(mfrow=c(3, 5), mar=c(4, 4, 1, 1))
> for(i in 1:3) {
+   for(j in 1:5) {
+     plot(xs[,j],ys[,i],xlab=dimnames(xs)[[2]][j],
+          ylab=dimnames(ys)[[2]][i])
+     abline(lsfit(xs[,j],ys[,i])$coef)
+     cr <- cor(xs[,j],ys[,i])
+     xt <- min(xs[,j])
+     yt <- ifelse(i==3,max(ys[,i])-0.02,min(ys[,i])+0.04)
+     text(xt,yt,paste("r","=",as.character(round(cr,4))),adj=0)
+   }
+ }
```



It appears that the best (i.e., most linear) two models are

$$\frac{1}{\texttt{mmhr}} = \beta_0 + \beta_1 \texttt{mw},$$
$$\log(\texttt{mmhr}) = \beta_0 + \beta_1 \texttt{mw}^{2/3}.$$

which does not include the Shriener et al. model. However, the Schreiner et al. model is the third best among those tried, and certainly seems rather linear.

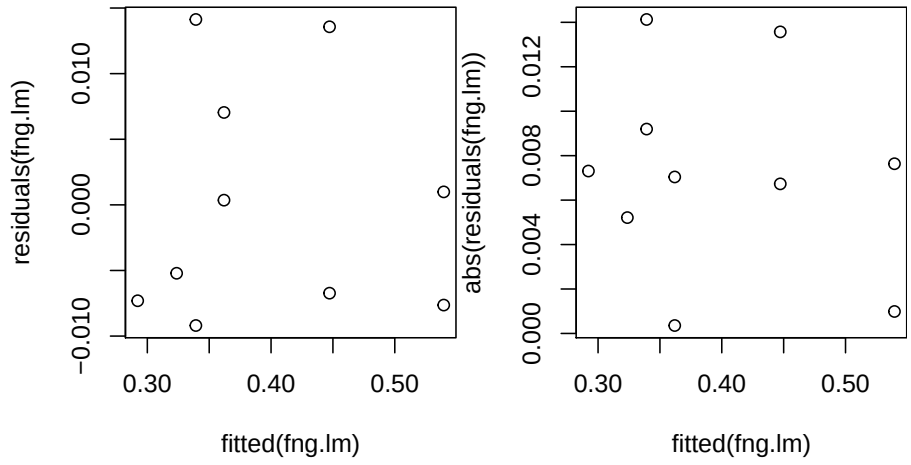
(b) Assess the suitability of the normal linear model which regresses the inverse of the rate $(1/y)$ on the untransformed molecular weight (x) using appropriate diagnostics.

Solution:

```
> fng.lm <- lm(1/mmhr ~ mw)
> summary(fng.lm)

Call:
lm(formula = 1/mmhr ~ mw)

Residuals:
      Min        1Q    Median        3Q       Max
```

```
-0.009196 -0.007164 -0.002428  0.005529  0.014128

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.2844365  0.0050063   56.82 1.02e-11 ***
mw          0.0019430  0.0000683   28.45 2.52e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.009376 on 8 degrees of freedom
Multiple R-squared:  0.9902,        Adjusted R-squared:  0.989
F-statistic: 809.2 on 1 and 8 DF,  p-value: 2.521e-09
```

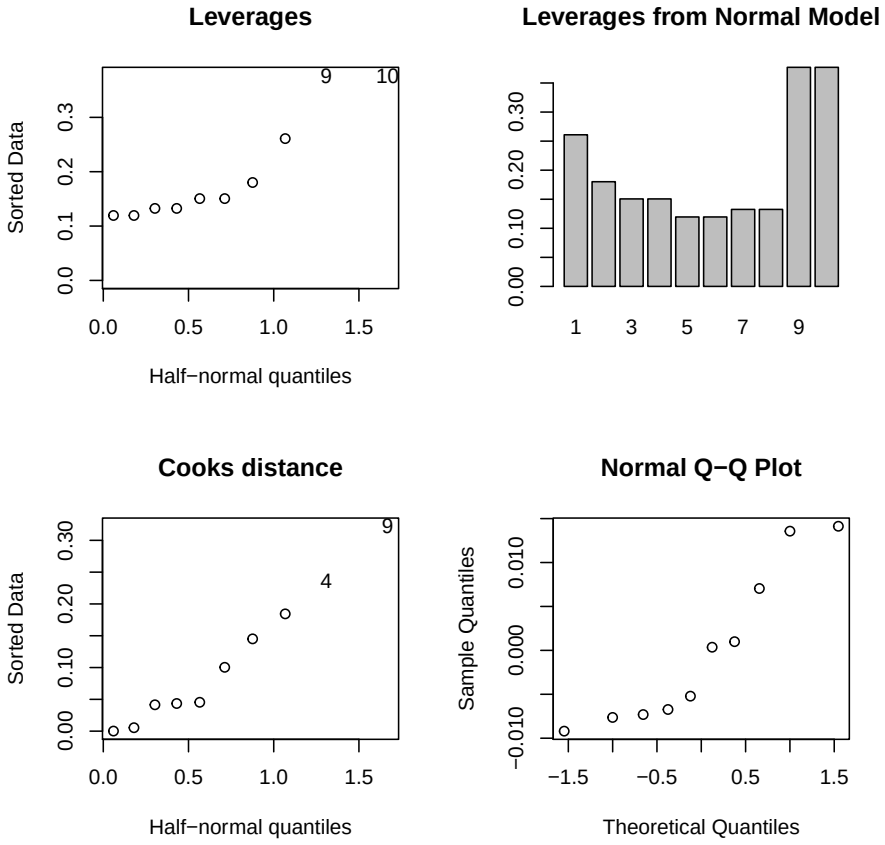
```
> par(mfrow=c(1, 2), pty="s", mar=c(2, 4, 0, 0))
> plot(fitted(fng.lm), residuals(fng.lm))
> plot(fitted(fng.lm), abs(residuals(fng.lm)))
```



Neither of these plots looks particularly problematic.

```
> library(faraway)
> fng.inf <- influence(fng.lm)
```

```
> par(mfrow=c(2, 2))
> halfnorm(fng.inf$hat, main="Leverages")
> barplot(fng.inf$hat, main="Leverages from Normal Model")
> halfnorm(cooks.distance(fng.lm), main="Cooks distance")
> -sort(-fng.inf$hat)[1:3]

        9        10         1
0.3769678 0.3769678 0.2608964

> qqnorm(residuals(fng.lm))
```



These plots do show some cause for concern. In particular, the normal q-q plot shows that the residuals have a distinct positive skew. Also, the leverages of the two points associated with the gas Xenon are close to the cut-off value of $2p/n = 2(2) = 10 = 0.4$ for this regression and may be having an undue influence. However, the Cooks' distance values are all < 1.

(c) The structure of the model in part (b) brings to mind a GLM using which distribution (HINT: think about canonical links)? Why would this model make sense for this data? Fit this GLM to the data and assess its suitability using appropriate diagnostics. Do you think this model is preferable to the normal model of part (b)?

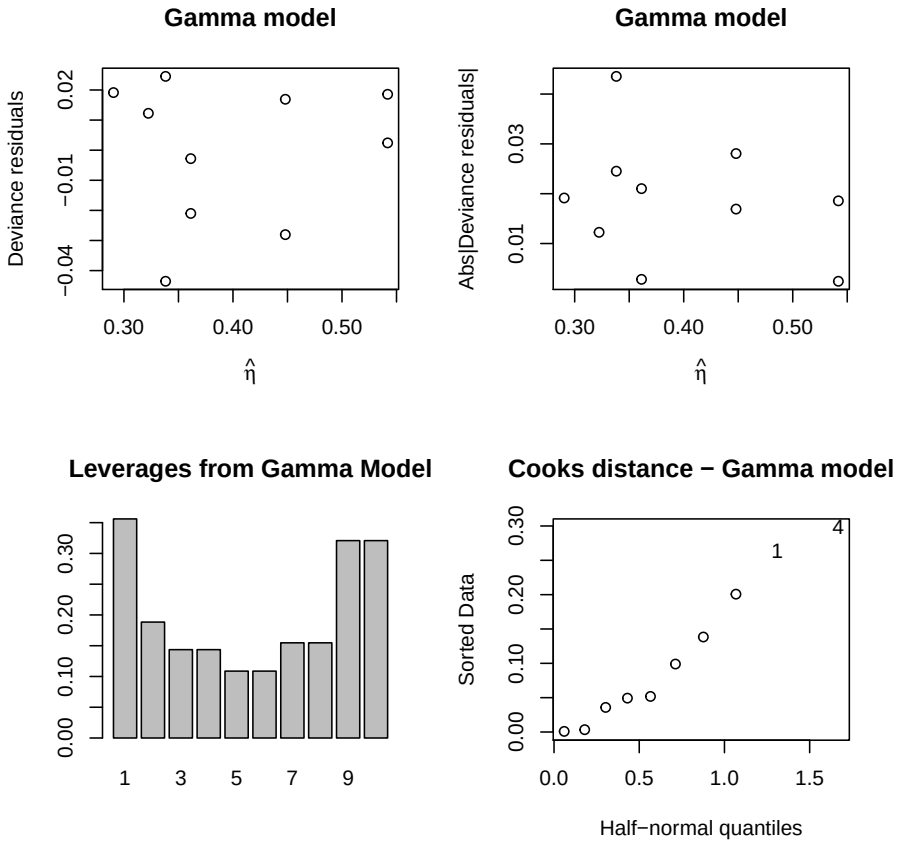
Solution: The obvious choice is a gamma GLM, which makes sense since gamma distributions are positively skewed, in line with the information from our normal q-q plot in part (b) and the canonical link in R is $1/\mu$.

```
> fng.glm <- glm(mmhr ~ mw,family=Gamma)
> summary(fng.glm)$coef
```

```
               Estimate   Std. Error  t value     Pr(>|t|)
(Intercept) 0.282485792 4.495728e-03 62.83427 4.575944e-12
mw          0.001975362 7.905526e-05 24.98710 7.040174e-09
```

```
> par(mfrow=c(2, 2))
> plot(residuals(fng.glm)~predict(fng.glm, type="link"),
+      xlab=expression(hat(eta)), ylab="Deviance residuals",
+      main="Gamma model")
> plot(abs(residuals(fng.glm))~predict(fng.glm, type="link"),
+      xlab=expression(hat(eta)), ylab="Abs|Deviance residuals|",
+      main="Gamma model")
> fng.inf.glm <- influence(fng.glm)
> barplot(fng.inf.glm$hat, main="Leverages from Gamma Model")
> halfnorm(cooks.distance(fng.glm), main="Cooks distance - Gamma model")
> -sort(-fng.inf.glm$hat)[1:3]
```

```
        1         9        10
0.3560708 0.3207883 0.3207883
```



So, the coefficient values are quite similar for both models. However, there is perhaps some concern now that the diagnostic plots are showing some incorrectness in the variance function, in that the deviance residuals may be decreasing in absolute size as the linear predictor increases, though it is certainly not conclusive. Also, the leverages seem to be a bit less of a concern than they were, though now the data point associated with the gas Helium is perhaps overly influential, but all Cooks' distance values are < 1.

Based on all of the above analyses, there is not really a definitive reason to choose either the gamma GLM or the normal linear model, though perhaps the GLM would be slightly favored due to the fact that growth rates must be positive numbers and normal distributions can yield negative outcomes. Of course, the plot of the absolute value of the deviance residuals versus the linear predictor values is a slight concern.

(d) The element Radon is an inert gas with a molecular weight of 222. Predict the fungal growth rate in the presence of Radon gas using the normal linear regression from part (b) as well as the GLM from part (c) and the Schriener et al. normal linear model. Also, find 95% confidence intervals for this fungal growth rate using each of the three models.

Solution:

```
> fng.sch <- lm(mmhr ~ sqrt(mw))
> mw <- 222
> prd1 <- predict(fng.lm,newdata=list(mw=222),se.fit=T)
> prd2 <- predict(fng.glm,newdata=list(mw=222),se.fit=T)
> prd3 <- predict(fng.sch,newdata=list(mw=222),se.fit=T)
> fts <- c(prd1$fit,prd2$fit,prd3$fit)
> sds <- c(prd1$se.fit,prd2$se.fit,prd3$se.fit)
> ests <- c(1/prd1$fit,1/prd2$fit,prd3$fit)
> lowerf <- fts - qt(0.975,8)*sds
> upperf <- fts + qt(0.975,8)*sds
> lower <- rep(0, 3)
> lower[1:2] <- 1/upperf[1:2]
> lower[3] <- lowerf[3]
> upper <- rep(0, 3)
> upper[1:2] <- 1/lowerf[1:2]
> upper[3] <- upperf[3]
> cbind(lower, ests, upper)

     lower     ests    upper
1 1.347091 1.397075 1.450911
1 1.325761 1.386932 1.454021
1 1.095670 1.231014 1.366358
```

So, it appears that the Schriener model gives a somewhat different prediction (which might be used to test its correctness by actually growing fungus in the presence of Radon gas), while the other two models (which have the same link structure but different error structures) give basically the same prediction and interval.

References

F. L. Ramsey and D. W. Schafer. The statistical sleuth: a course in methods of data analysis. Brooks/Cole, 2013.