STAT3015/4030/7030 Generalised Linear Modelling

Tutorial 9

1. The data file `Leuk.txt` contains the survival time in weeks and the initial white blood cell count for leukemia patients. It also contains an indicator of whether the patient was positive (AG = 1) or negative (AG = 2) for a particular blood factor called AG.

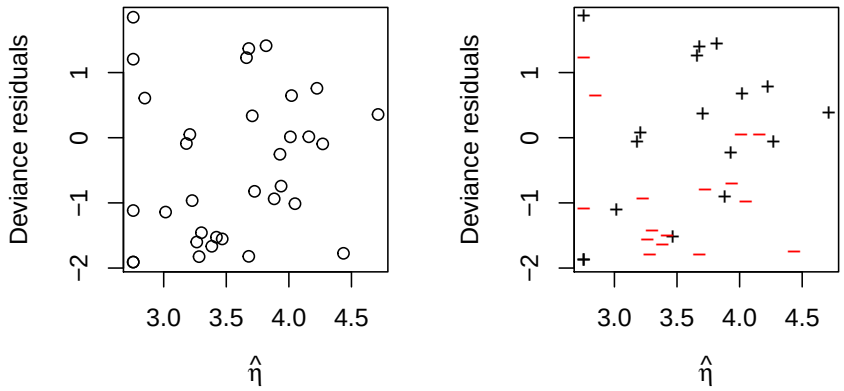
   (a) To model survival times in weeks, fit a gamma GLM with a logarithmic link and the logarithm of the white blood cell count as the predictor. Plot the deviance residuals versus the linear predictor values. Do you see any pattern to be concerned about? Replot the deviance residuals using different symbols for the residuals from the positive and the negative patients. Do you see any pattern to be concerned about now?

   Solution:

```
> leuk <- read.table("Leuk.txt", header=TRUE)
> attach(leuk)
> names(leuk)

[1] "surv" "wbc"  "ag"

> leuk.glm <- glm(surv~log(wbc), family=Gamma(link=log))
> par(pty="s")
> par(mfrow=c(1, 2))
> plot(residuals(leuk.glm)~predict(leuk.glm, type="link"),
+      xlab=expression(hat(eta)), ylab="Deviance residuals")
> plot(residuals(leuk.glm)~predict(leuk.glm, type="link"), type="n",
+      xlab=expression(hat(eta)), ylab="Deviance residuals")
> points(predict(leuk.glm, type="link")[ag==1], residuals(leuk.glm)[ag==1],
+        pch="+")
> points(predict(leuk.glm, type="link")[ag==2], residuals(leuk.glm)[ag==2],
+        pch="-", col=2)
```



The first plot does not appear to show any problematic pattern. The second plot indicates that blood factor may be a significant predictor, as the positive and negative blood factor residuals are not evenly spread.

(b) Include the blood factor as a categorical predictor. Is it significant? Again, plot the deviance residuals using separate symbols for residuals from the positive and negative patients. Do you see any patterns of concern?

Solution:

```
> agind <- ifelse(ag==2, 1, 0)
> leuk.glm1 <- glm(surv~log(wbc)+agind, family=Gamma(link=log))
> anova(leuk.glm1, test="Chisq")

Analysis of Deviance Table

Model: Gamma, link: log

Response: surv

Terms added sequentially (first to last)
```

```
         Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                        32     58.138
log(wbc)  1  10.3149        31     47.823 0.002080 **
agind     1   7.4955        30     40.328 0.008683 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> summary(leuk.glm1)$dispersion
```

```
[1] 1.088374
```

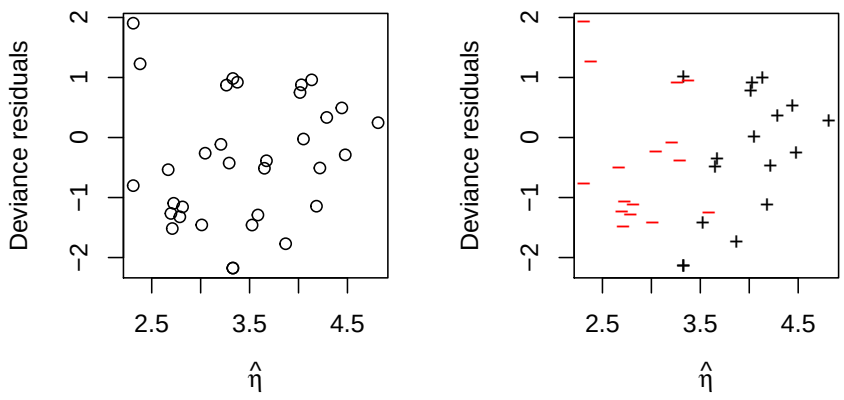
```
> summary(leuk.glm1)$coef
```

```
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept)  6.8297872  1.3216634  5.167569 1.451308e-05
log(wbc)    -0.3040454  0.1375460 -2.210500 3.483896e-02
agind       -1.0180652  0.3643135 -2.794476 8.973475e-03
```

```
> 2*(1-pt(2.794474, leuk.glm1$df.res))
```

```
[1] 0.008973509
```

```
> par(pty="s")
> par(mfrow=c(1, 2))
> plot(residuals(leuk.glm1)~predict(leuk.glm1, type="link"),
+      xlab=expression(hat(eta)), ylab="Deviance residuals")
> plot(residuals(leuk.glm1)~predict(leuk.glm1, type="link"), type="n",
+      xlab=expression(hat(eta)), ylab="Deviance residuals")
> points(predict(leuk.glm1, type="link")[ag==1], residuals(leuk.glm1)[ag==1],
+        pch="+")
> points(predict(leuk.glm1, type="link")[ag==2], residuals(leuk.glm1)[ag==2],
+        pch="-", col=2)
```



Clearly, the AG blood factor is highly significant. As for the residual plot, it does not show any pattern to be overly concerned about (note that the fact that all the "−"s are on the left-hand side is not a problem, it just indicates that the linear predictor values for such patients are smaller than those of the other patients, which is just another statement that the factor is significant).

(c) Include an interaction term between the blood factor and the logarithm of the white blood cell count in the model. What does the inclusion of this interaction term allow in the model structure? Test whether the interaction is significant.

Solution: The inclusion of an interaction term in the model allows the "slope" associated with the logarithm of the white blood cell count to be different for patients with a positive blood factor than it is for patients with a negative blood factor. More precisely, since the model structure is:

$$\log\{E(\text{surv})\} = \beta_0 + \beta_1 \log(\text{wbc}) + \beta_2 z + \beta_3 z \log(\text{wbc}) \rightarrow E(\text{surv}) = \alpha \text{wbc}^{\beta_1 + \beta_3 z},$$

where $\alpha = \exp(\beta_0 + \beta_2 z)$ and z is the indicator of for the AG blood factor, we see that the inclusion of the interaction term allows the exponent of `wbc` in the "power model" to be different for the two different AG blood factor groups. Now, to test the significance of the interaction:

```
> leuk.glm2 <- glm(surv~log(wbc)*agind, family=Gamma(link=log))
> anova(leuk.glm2, test="Chisq")

Analysis of Deviance Table

Model: Gamma, link: log

Response: surv

Terms added sequentially (first to last)


              Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                             32     58.138
log(wbc)       1  10.3149        31     47.823 0.002375 **
agind          1   7.4955        30     40.328 0.009584 **
log(wbc):agind 1   1.7709        29     38.557 0.207973
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So, the interaction is not significant, and we can reasonably assume that the exponent in the "power model" is the same for both levels of the blood factor.

2. The data file `Transaction.txt` contains data regarding the audit of a large Australian corporation. The audit process incurred different transaction costs for examining each of two different transaction types. The available data consist of the total time (in minutes) spent on examining both types of transactions in 261 different branch offices of the enterprise as well as the total number of transactions of the two different types occuring at each branch during the 1985-86 financial year. The aim is to develop a model for the cost associated with dealing with each transaction type. Since cost is related to time, we can treat time as the response and try to model its relationship to the number of different transactions. Moreover, since all the time is spent on one or the other of the two transaction types, the response should be additively related to the explanatory variables and ought to go through the origin.

   (a) Fit an ordinary least-squares regression to this data. Do you think that an intercept of zero is reasonable based on this analysis? Plot the residuals versus fitted values, as well as a normal Q-Q plot of the residuals. Do you think that a normal linear model is reasonable for these data based on these plots?

   Solution:

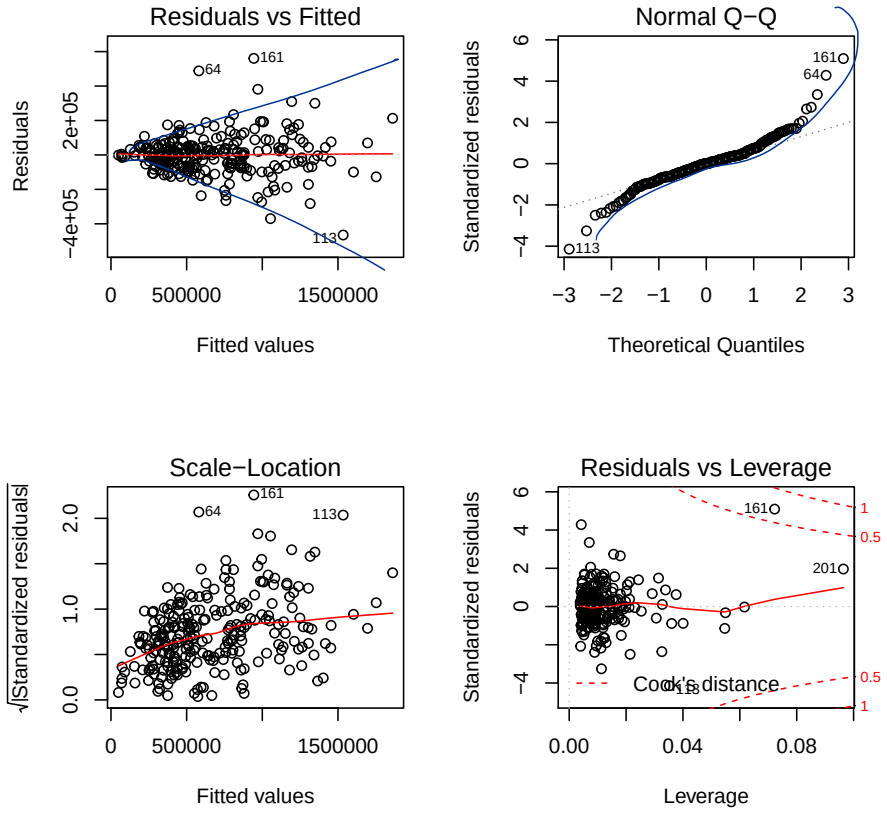
```
> trns <- read.table("Transaction.txt", header=TRUE)
> attach(trns)
> names(trns)
```

```
[1] "Time"  "Type1" "Type2"

> trns.reg <- lm(Time~Type1+Type2)
> summary(trns.reg)$coef

                Estimate   Std. Error    t value     Pr(>|t|)
(Intercept) 14464.076220 1.705289e+04  0.8481892 3.971188e-01
Type1           5.462514 4.332375e-01 12.6085914 1.016170e-28
Type2           2.034394 9.432790e-02 21.5672565 1.120348e-59

> par(mfrow=c(2, 2))
> plot(trns.reg)
```



So, it appears that the intercept is not significant for this model. However, the diagnostic plots clearly show that there is heteroscedasticity and non-normality in this dataset.

(b) Generally speaking random times are asymmetrically distributed, and indeed often exponentially distributed. Thus, the distribution of the total of a number of random times often falls into the gamma distribution family. Fit a gamma generalised linear model to the transaction data, maintaining the additive structure of the relationship

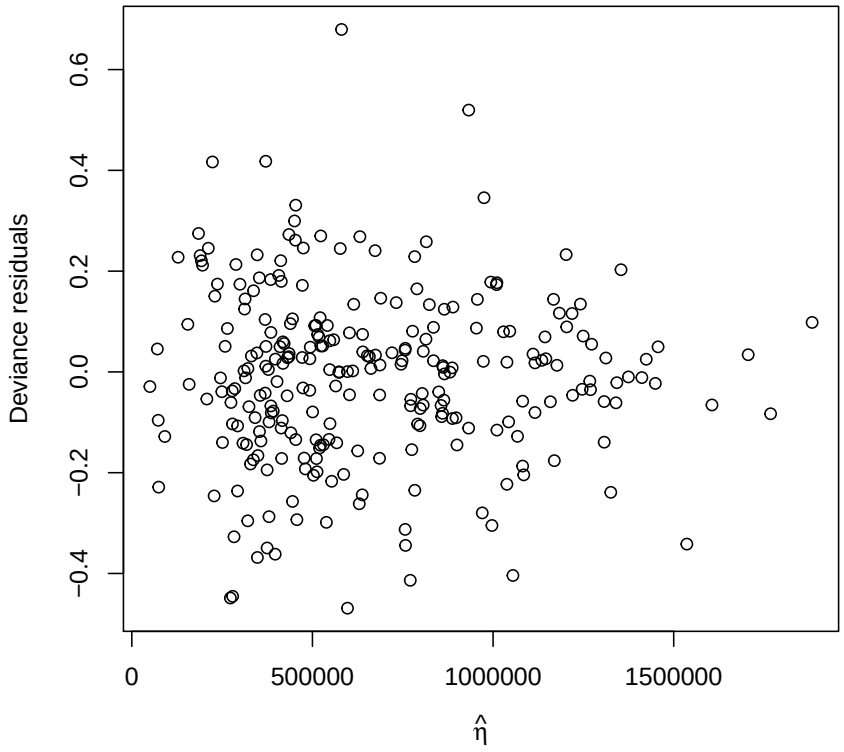
between the mean response and the predictors (i.e., retain the identity link function). Plot the deviance residuals versus the linear predictor values. Comment on the plot. Do you think that a model without intercept is justifiable for this model?

Solution:

```
> trns.glm <- glm(Time~Type1+Type2, family=Gamma(link=identity))
> summary(trns.glm)$coef

                Estimate   Std. Error   t value     Pr(>|t|)
(Intercept) 15359.563935 5.182508e+03  2.963732 3.323430e-03
Type1           5.705443 4.257404e-01 13.401225 1.939182e-31
Type2           2.006855 5.803102e-02 34.582463 7.378258e-99

> plot(residuals(trns.glm)~predict(trns.glm, type="link"),
+      xlab=expression(hat(eta)), ylab="Deviance residuals")
> trns.glm1 <- glm(Time~Type1+Type2-1, family=Gamma(link=identity))
```



There may be an overdispersion problem (the data with low values of the linear predictor seem slightly more variable than those with large linear predictor values), but it is not as severe as it was for the normal model. Also, it appears that the intercept is significant for this model.

(c) Suppose that we believe the model without intercept is appropriate and that a new branch is surveyed. If it is found that the 100,000 transactions of each type were recorded for this branch during the 1985-86 financial year, estimate the amount of time that will be necessary to complete the audit process for this branch. In addition, estimate and find a 95% confidence interval for the proportion of this total audit time spent to examine transactions of the first type.

Note: the proportion of time spent on the first transaction will be estimated by a non-linear function of the parameters $h(\beta)$. To find the confidence interval, use the following approximation for the variance:

$$Var(h(\hat{\beta})) \approx \frac{\partial h(\beta)}{\partial \beta}^T Var(\hat{\beta}) \frac{\partial h(\beta)}{\partial \beta} = \phi \frac{\partial h(\beta)}{\partial \beta}^T (X^T W X)^{-1} \frac{\partial h(\beta)}{\partial \beta}$$

This is the delta-approximation to the variance of a function of a random quantity. Note that $Var(\hat{\beta}) = \phi(X^T W X)^{-1}$. From R, we obtain the estimate $\hat{\phi}$ with the command `m1$dispersion` , and we obtain an estimate of the matrix $(X^T W X)^{-1}$ with the command `m1$cov.unscaled` (for the `glm` object named `m1`).

Solution:

```
> summary(trns.glm1)$coef

      Estimate Std. Error  t value      Pr(>|t|)
Type1 5.471466 0.42913384 12.75002  3.175370e-29
Type2 2.121886 0.04609572 46.03217 1.055674e-126

> summary(trns.glm1)$dispersion

[1] 0.03053187
```

So, the estimate of the time (in minutes) needed for the new branch is:

$$100000\hat{\beta}_1 + 100000\hat{\beta}_2 = 100000(5.471466 + 2.121886) = 759335.$$

The proportion of time spent on Type 1 transactions is:

$$h(\beta) = \frac{100000\beta_1}{100000\beta_1 + 100000\beta_2} = \frac{\beta_1}{\beta_1 + \beta_2},$$

so that $h(\hat{\beta}) = 0.72056$ and

$$\frac{\partial h(\beta)}{\partial \beta} = \begin{pmatrix} \frac{\beta_2}{(\beta_1+\beta_2)^2} \\ -\frac{\beta_1}{(\beta_1+\beta_2)^2} \end{pmatrix}$$

So, the required confidence interval is:

```
> cf <- coef(trns.glm1)
> est <- cf[1]/(cf[1]+cf[2])
> dh <- c(cf[2],-cf[1])/((cf[1]+cf[2])^2)
> sd <- sqrt(summary(trns.glm1)$dispersion)
> sd <- sd*sqrt(t(dh)%*%summary(trns.glm1)$cov.unscaled%*%dh)
> upper <- est + qt(0.975,259)*sd
> lower <- est - qt(0.975,259)*sd
> as.vector(c(lower, est, upper))

[1] 0.6822893 0.7205600 0.7588308
```