Workshop 4: solutions

1. Let us start by calculating the expected value of the vector $\mathbf{z}$. We have that

$$
\begin{aligned}
\mathbb{E}(z_i) &= \mathbb{E}[g'(\mu_i)(y_i - \mu_i) + \mathbf{X}_i\boldsymbol{\beta}] \\
&= g'(\mu_i)[\mathbb{E}(y_i) - \mu_i] + \mathbf{X}_i\boldsymbol{\beta} \\
&= g'(\mu_i)(\mu_i - \mu_i) + \mathbf{X}_i\boldsymbol{\beta} \\
&= \mathbf{X}_i\boldsymbol{\beta},
\end{aligned}
$$

and so $\mathbb{E}(\mathbf{z}) = \mathbf{X}\boldsymbol{\beta}$. From standard results on transformation of variances we have that

$$
\begin{aligned}
\mathrm{var}(z_i) &= \mathrm{var}(g'(\mu_i)(y_i - \mu_i) + \mathbf{X}_i\boldsymbol{\beta}) \\
&= g'(\mu_i)^2 \mathrm{var}(y_i) \\
&= g'(\mu_i)^2 V(\mu_i)\phi.
\end{aligned}
$$

Since the y_i are mutually independent, so are the z_i. Hence the covariance matrix for $\mathbf{z}$ is a diagonal matrix whose ith entry is given by $g'(\mu_i)^2 V(\mu_i)\phi$. Using the notation from the lecture notes (section 11.2), $g'(\mu_i)^2 V(\mu_i) = w_i^{-1}$, and hence we can also write the covariance matrix for $\mathbf{z}$ as $\mathbf{W}^{-1}\phi$ where $\mathbf{W}$ is a diagonal matrix, with $W_{ii} = w_i$.

2. (a)

$$
\frac{\mathrm{d}}{\mathrm{d}\beta}\sum_{i=1}^{n} w_i(z_i - \beta x_i)^2 = -2\sum_i w_i(z_i - x_i\beta)x_i
$$

Setting this to zero and re-arranging yields:

$$
\hat{\beta} = \frac{\sum_i w_i z_i x_i}{\sum_i w_i x_i^2}.
$$

Differentiating the weighted sum of squares again shows $\hat{\beta}$ *minimizes* it.

(b) The link function is the identity function, $g(\mu_i) = \mu_i$, and so $g'(\mu_i) = 1$. For the Poisson distribution we have $V(\mu_i) = \mu_i$ (see section 11.1 of the notes). Hence, at each IRLS iteration we will have

$$
w_i = (g'(\mu_i)^2 V(\mu_i))^{-1} = \frac{1}{\mu_i},
$$

and

$$
z_i = g'(\mu_i)(y_i - \mu_i) + x_i\beta = (y_i - \mu_i) + x_i\beta = (y_i - x_i\beta) + x_i\beta = y_i.
$$

Note that $g(\mu_i) = \eta_i = x_i\beta$ but because g is the identity function, $g(\mu_i) = \mu_i = \eta_i = x_i\beta$.

So the IRLS steps will proceed as follows.

- **Initialize** Set $\boldsymbol{\mu}^{(0)} = (1, 1, 3, 7, 4)^{\mathsf{T}}$ (i.e. to $\mathbf{y}$, except when $y_i = 0$, since $\mu_i > 0$). $\boldsymbol{\eta}^{(0)} = \boldsymbol{\mu}^{(0)}$, since the model has an identity link.

- **Iteration 1** First find the iterative weights and working responses:

$$w_i^{(0)} = (\mu_i^{(0)})^{-1} \Rightarrow (w_1^{(0)}, w_2^{(0)}, w_3^{(0)}, w_4^{(0)}, w_5^{(0)}) = \left(1, 1, \frac{1}{3}, \frac{1}{7}, \frac{1}{4}\right),$$

$$z_i^{(0)} = y_i \Rightarrow (z_1^{(0)}, z_2^{(0)}, z_3^{(0)}, z_4^{(0)}, z_5^{(0)}) = (0, 1, 3, 7, 4).$$

Using part (a) we have that $\beta^{(1)} = \frac{\sum_{i=1}^{5} w_i^{(0)} z_i^{(0)} x_i}{\sum_i w_i^{(0)} x_i^2}$. I will do this calculation in R.

```
omega_0 <- c(1, 1, 1/3, 1/7, 1/4)
z_0 <- c(0, 1, 3, 7, 4)
x <- c(1, 2, 3, 4, 5)
beta_1 <- sum(omega_0*z_0*x)/sum(omega_0*(x^2))
beta_1
## [1] 0.8466523
```

- **Iteration 2** We have that $w_i^{(1)} = (\mu_i^{(1)})^{-1} = (x_i \beta^{(1)})^{-1}$ and $z_i^{(1)} = y_i$. We can then calculate $\beta^{(2)}$ as in the first iteration.

```
mu_1<- x*beta_1
omega_1 <- 1/mu_1
z_1 <- c(0, 1, 3, 7, 4)
beta_2 <- sum(omega_1*z_1*x)/sum(omega_1*(x^2))
beta_2
## [1] 1
```

Actually, $\hat{\beta} = 1$ is the MLE — in this case the IRLS has converged in 2 steps. Let's confirm this.

```
x <- c(1, 2, 3, 4, 5)
y <- c(0, 1, 3, 7, 4)
res_toy <- glm(y ~ x-1, family = poisson(link = "identity"))
res_toy$coefficients

## x
## 1
```

(c) From general theory, we know that (asymptotic distribution of the maximum likelihood estimate $\hat{\beta}$) the covariance matrix for a GLM with scale parameter 1 is $(\mathbf{X}^\mathsf{T} \mathbf{W} \mathbf{X})^{-1}$. For the current single-parameter model this becomes $\left(\sum_i w_i x_i^2\right)^{-1}$, with the square root of this giving the standard error for $\hat{\beta}$. To evaluate it we need the final estimates for the w_i:

$$\hat{\mu}_i = \hat{\eta}_i = x_i \hat{\beta} = x_i$$

	1	2	3	4	5
$w_i = \hat{\mu}_i^{-1}$	1	1/2	1/3	1/4	1/5

which results in a standard error estimate for $\hat{\beta}$ of 0.258. We can check this is indeed the correct value by comparing to the value returned by the glm function.

```
sqrt(vcov(res_toy))

##           x
## x 0.2581989
```

3. Suppose that the original GLM has p parameters. From the generalized likelihood ratio test distributional result, under the null hypothesis that the simpler model (the one with p predictors) is correct, we have that

$$D_p^* - D_{p+1}^* \sim \chi_{1(=p+1-p)}^2.$$

The expected value for a χ_1^2 r.v. is 1, so this is the expected deviance decrease on adding an irrelevant predictor to the model (assuming the original model is a reasonable one).

4. From workshop 3, we have that for an Inverse Gaussian distribution

$$\theta = -\frac{1}{2\mu^2}, \quad b(\theta) = -\frac{1}{\mu}, \quad \omega_i = 1,$$

and hence

$$\hat{\theta}_i = -\frac{1}{2\hat{\mu}_i^2}, \quad b(\hat{\theta}_i) = -\frac{1}{\hat{\mu}_i}, \quad \tilde{\theta}_i = -\frac{1}{2y_i^2}, \quad b(\tilde{\theta}_i) = -\frac{1}{y_i}.$$
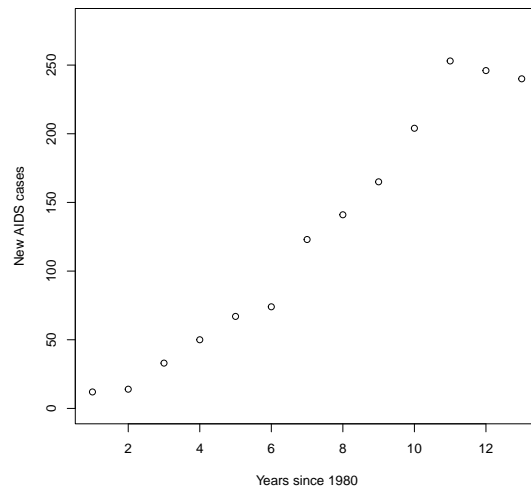
We know from the lecture notes that

$$D = \sum_{i=1}^{n} 2\omega_i \{y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)\},$$

and so replacing leads to

$$\begin{aligned}
D &= \sum_{i=1}^{n} 2 \left\{ y_i \left(-\frac{1}{2y_i^2} + \frac{1}{2\hat{\mu}_i^2} \right) + \frac{1}{y_i} - \frac{1}{\hat{\mu}_i} \right\} \\
&= \sum_{i=1}^{n} \left(\frac{1}{y_i} + \frac{y_i}{2\hat{\mu}_i^2} - \frac{2}{\hat{\mu}_i} \right) \\
&= \sum_{i=1}^{n} \left(\frac{\hat{\mu}_i^2 + y_i^2 - 2y_i\hat{\mu}_i}{y_i\hat{\mu}_i^2} \right) \\
&= \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{y_i\hat{\mu}_i^2}.
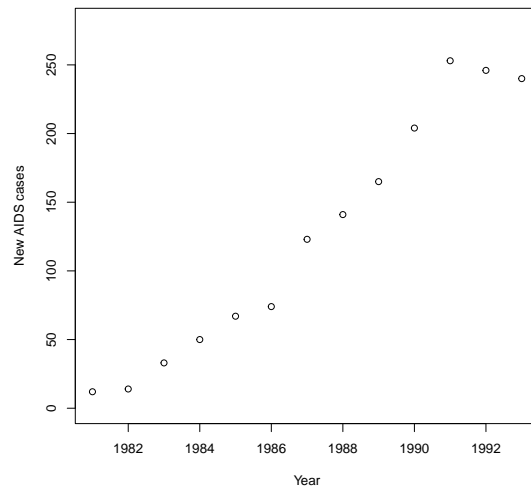\end{aligned}$$

5. (a) One possible way to plot the data is as follows.

```r
y <- c(12,14,33,50,67,74,123,141,165,204,253,246,240)
t <- 1:13
plot(t, y, xlab = "Years since 1980", ylab = "New AIDS cases",
     ylim = c(0, 280))
```

We could change the x-axis to be the year instead of the year since 1980.

```r
plot(t + 1980, y, xlab = "Year", ylab = "New AIDS cases",
     ylim = c(0, 280))
```


(b) Recall that the model we want to fit is the following one:

$$\mathbb{E}(y_i) \equiv \mu_i = \gamma e^{\alpha t_i},$$

which could be turned into a GLM by using the log link so that

$$\log(\mu_i) = \log(\gamma) + \alpha t_i = \beta_1 + \beta_2 t_i, \quad \beta_1 = \log(\gamma), \quad \beta_2 = \alpha.$$

Because the response variable y is a count (number of new AIDS cases), a Poisson distribution is a sensible distribution to try. Let's fit the model in R.

```
am0 <- glm(y ~ t,family = poisson(link = "log"))
# alternatively, because the log link is the canonical one
# we could just write
#am0 <- glm(y ~ t,family = poisson)
```

(c) First, remember that for the Poisson distribution the scaled deviance coincides with the deviance (because $\phi = 1$). Let's print the summary output of the fitted model so that we can access the deviance.

```
summary(am0)

##
## Call:
## glm(formula = y ~ t, family = poisson(link = "log"))
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 3.140590   0.078247   40.14   <2e-16 ***
## t           0.202121   0.007771   26.01   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 872.206  on 12  degrees of freedom
## Residual deviance:  80.686  on 11  degrees of freedom
## AIC: 166.37
##
## Number of Fisher Scoring iterations: 4
```

The residual deviance is the deviance of the fitted model. Note that we could access it directly by using the command `deviance`:

```
deviance(am0)

## [1] 80.68649
```

We know that under the null hypothesis that the model is ok, the scaled deviance should follow approximately, in the large-sample limit, a chi-squared distribution with 11 degrees of freedom (why 11? there are 13 observations and the model has two parameters). We can then calculate an approximate p-value as follows:

```
pchisq(deviance(am0), df = 11, lower.tail = FALSE)

## [1] 1.086916e-12

#equivalent to
#1 - pchisq(deviance(am0), df = 11)
```

The p-value is extremely low and so there is very strong evidence to reject the null hypothesis that the model is correct. Note that the sample size is low and the result is only approximate, but still, this is a very low p-value. Out of curiosity, just by looking at the deviance, we should

have an idea whether the model is fitting well or not. Recall that if $Z \sim \chi_r^2$, then $E(Z) = r$ and $\text{var}(Z) = 2r$. In this case, we have a deviance of roughly 80, whereas the limiting distribution under the assumption that the model is correct should have a mean of 11 and a standard deviation of $\sqrt{22} = 4.7$.

(d) From part(c) we should already expect to see some problematic residual plots.

```
par(mfrow = c(2,2))
plot(am0)
```
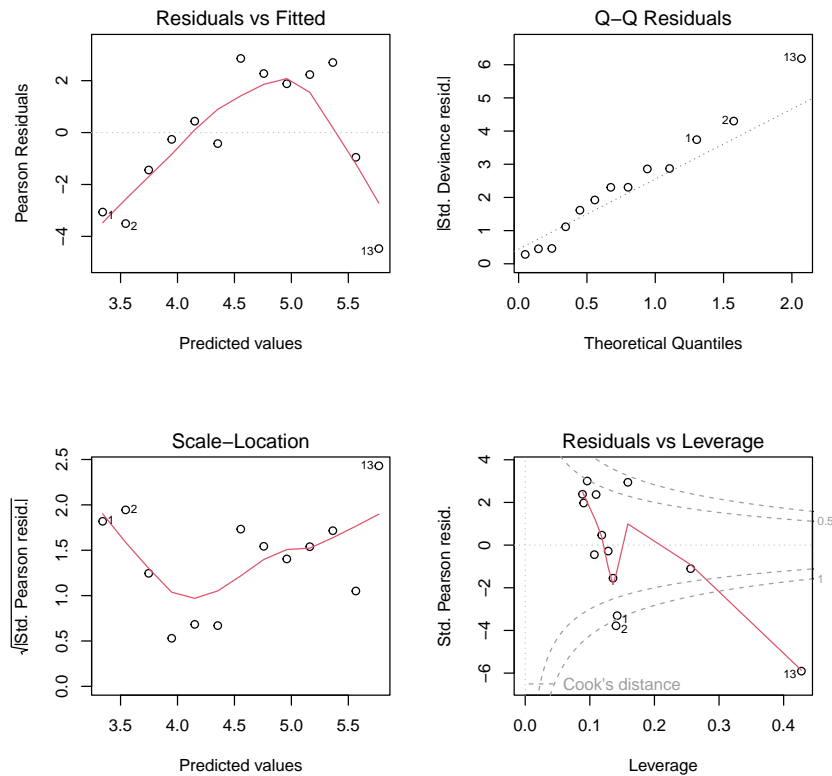

Indeed, the residual plots are worrying. There is a clear pattern in the mean of the residuals, when plotted against the predicted values (and the same would be true if the predicted values were replaced by the fitted values). If we plot the residuals against time we get the same sort of pattern.

```
plot(t, residuals(am0, type = "pearson"), ylab = "Pearson residuals")
lowess_smooth <- lowess(t, residuals(am0, type = "pearson"))
lines(lowess_smooth, col = "red")
abline(h = 0, lty = 3, col = "gray80")
```
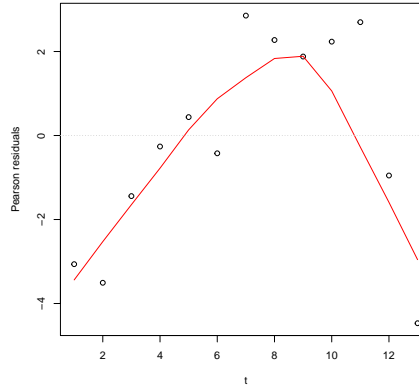
It is also worrying the very high influence of the final year's data, evident in the residuals versus leverage plot.

(e) From the residual plots in (d) it seems sensible to amend the model by adding a quadratic term to obtain:

$$\log(\mu_i) = \beta_1 + \beta_2 t_i + \beta_3 t_i^2.$$

The following code fits this model.

```
am1 <-  glm(y ~ t + I(t^2), family = poisson(link = "log"))
```

(f) Let's start by printing the summary output of this new model that includes a quadratic effect of time.

```
summary(am1)

##
## Call:
## glm(formula = y ~ t + I(t^2), family = poisson(link = "log"))
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.901459   0.186877  10.175  < 2e-16 ***
## t             0.556003   0.045780  12.145  < 2e-16 ***
## I(t^2)       -0.021346   0.002659  -8.029 9.82e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 872.2058  on 12  degrees of freedom
## Residual deviance:   9.2402  on 10  degrees of freedom
## AIC: 96.924
##
## Number of Fisher Scoring iterations: 4
```
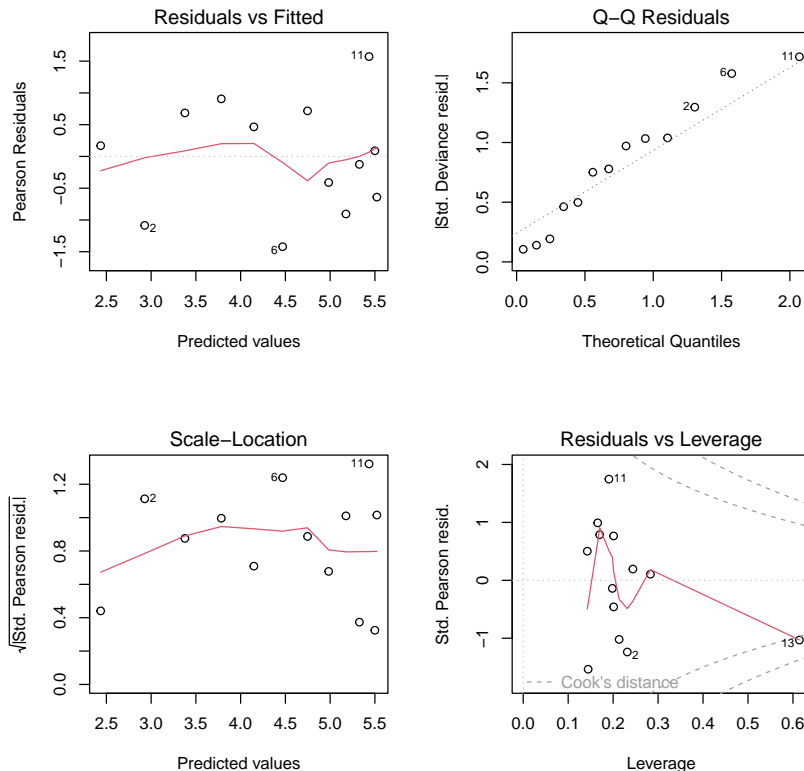
We can notice that the deviance of the fitted model is now much lower and close to what is expected for a χ_{10}^2 random variable. Indeed, the approximate p-value does not lead to a rejection of the null hypothesis (that the model is ok).

7

```
pchisq(deviance(am1), df = 10, lower.tail = FALSE)
```

```
## [1] 0.5094652
```

Let's inspect the residual plots.

```
par(mfrow = c(2,2))
plot(am1)
```


The residual plots are now much improved: the clear trend in the mean has gone, the (vertical) spread of the residuals is reasonably even, the influence of point 13 is much reduced.

Regarding the significance of the terms, the reported p-values are very low, i.e., for each parameter there is clear evidence that it is not zero.

(g) We have the AIC included in the summary of each model but we can also extract it easily using the function AIC:

```
AIC(am0, am1)
```

```
##     df       AIC
## am0  2 166.36982
## am1  3  96.92358
```

AIC is much lower for model am1 and so based on this criterion this should be the preferred model.

(h) First, residual plots are much better for model am1, i.e., for the model with a quadratic trend. It would not have been reliable to base any conclusions on model am0 in light of its residual

plots. Second, all coefficients are firmly significant in model `am1`, including β_3, providing clear evidence that `am1` is preferable to `am0`. The AIC also prefers model `am1` to model `am0`.

As an example, although it was not needed here, in order to check if model `am1` should be preferred to model `am0`, we could have used a generalized likelihood ratio test as follows.

```
anova(am0, am1, test = "Chisq")

## Analysis of Deviance Table
##
## Model 1: y ~ t
## Model 2: y ~ t + I(t^2)
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1        11     80.686
## 2        10      9.240  1   71.446 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Obviously, the conclusion is the same as before: the tiny p-value indicates that `am0` should be firmly rejected in favour of `am1`. The `test = "Chisq"` argument to anova is justified because the scale parameter (ϕ) is known for this model: had it been estimated it would be preferable to set test to `"F"`.

(i) From the summary output the quadratic coefficient (β_3) is negative, so exponential rise is slowing with time. We can also confirm this with a plot where we include the underlying AIDS case rate, with 95% confidence limits.

```
new.t <- seq(1, 13, length = 100)
fv <- predict(am1, data.frame(t = new.t), se = TRUE)
qz <- qnorm(0.975)

plot(t, y, xlab = "Years since 1980", ylab = "New AIDS cases",
     ylim = c(0, 280))
lines(new.t, exp(fv$fit))
lines(new.t, exp(fv$fit - qz*fv$se.fit), lty = 2)
lines(new.t, exp(fv$fit + qz*fv$se.fit), lty = 2)
```
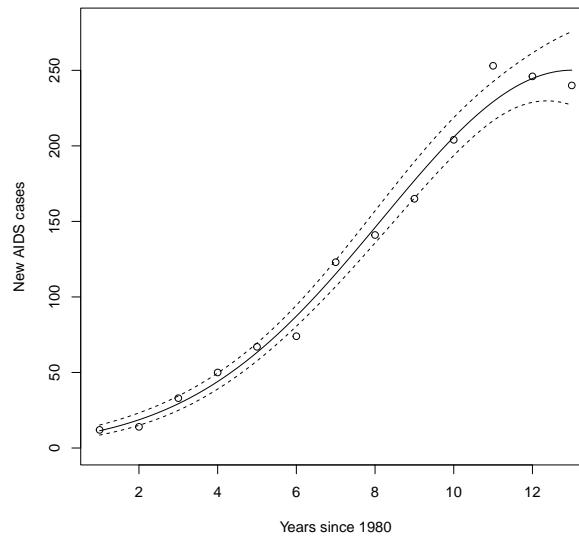
By the end of the data there is evidence that the rate of increase is slowing. Notice that by default the `predict` function when passed a `glm` object predicts on the scale of the linear predictor: we have to apply the inverse of the link function to get back onto the original response scale.

6. (a) One possibility to reproduce the plot is as follows.

```
load("orings.rda")
plot(orings$temperature, orings$damaged, ylim = c(0, 6),
     xlim = c(30, 85), xlab = "Temperature", ylab ="Damaged")
```
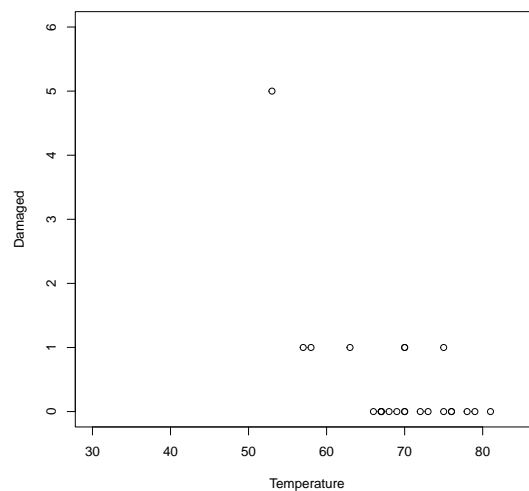


```
#another possibility
#with(orings, plot(temperature, damaged,
#                  xlim = c(30, 85), ylim = c(0, 6)))
```

(b) Let's fit the suggested logistic regression model.

```
oringm <- glm(cbind(damaged, undamaged) ˜ temperature,
              binomial(link = "logit"), data = orings)
oringm

##
## Call:  glm(formula = cbind(damaged, undamaged) ˜ temperature, family = binomial(
##     data = orings)
##
## Coefficients:
## (Intercept)   temperature
##     11.6630       -0.2162
##
## Degrees of Freedom: 22 Total (i.e. Null);   21 Residual
## Null Deviance:      38.9
## Residual Deviance: 16.91   AIC: 33.67
```

An alternative way of fitting this model is as follows.

```
oringm_alt <- glm(damaged/(damaged + undamaged) ˜ temperature,
                  binomial(link = "logit"),
                  weights = damaged + undamaged, data = orings)
oringm_alt

##
## Call:  glm(formula = damaged/(damaged + undamaged) ˜ temperature, family = binom:
##     data = orings, weights = damaged + undamaged)
##
## Coefficients:
## (Intercept)   temperature
##     11.6630       -0.2162
##
## Degrees of Freedom: 22 Total (i.e. Null);   21 Residual
## Null Deviance:      38.9
## Residual Deviance: 16.91   AIC: 33.67
```

Once we have fitted the model, and even though it was not specifically asked in the question, how should we interpret the parameters in a logistic regression model? In particular, how should we interpret β_2? Using the logit link function with the binomial distribution produces a useful interpretation. To understand this interpretation, the concept of *odds* first must be understood. If event A has probability p of occurring, then the odds of event A occurring is the ratio of the probability that A occurs to the probability that A does not occur: $p/(1-p)$. For example, if the probability of o-ring damage is 0.6, the odds of o-ring damage is $0.6/(1-0.6) = 1.5$. Clearly, using the logit link function in a binomial GLM is equivalent to modelling the logarithm of the odds (or the 'log-odds'). In this specific case, the logistic regression models the log odds of o-ring

damage as a function of the temperature. We have the following for a given temperature i, $\texttt{temp}_i$

$$\log\left(\frac{p_i}{1-p_i}\right) = \log\left(\frac{p_{\texttt{temp}_i}}{1-p_{\texttt{temp}_i}}\right) = \beta_1 + \beta_2\texttt{temp}_i,$$

$$\log\left(\frac{p_{\texttt{temp}_i+1}}{1-p_{\texttt{temp}_i+1}}\right) = \beta_1 + \beta_2(\texttt{temp}_i + 1).$$

Therefore,

$$\log\left(\frac{p_{\texttt{temp}_i+1}}{1-p_{\texttt{temp}_i+1}}\right) - \log\left(\frac{p_{\texttt{temp}_i}}{1-p_{\texttt{temp}_i}}\right) = \beta_1 + \beta_2(\texttt{temp}_i + 1) - (\beta_1 + \beta_2\texttt{temp}_i) = \beta_2.$$

Thus, β_2 represents the change in log odds of oring damage associated with a unit increase in temperature. Further note that,

$$\log\left(\frac{\frac{p_{\texttt{temp}_i+1}}{1-p_{\texttt{temp}_i+1}}}{\frac{p_{\texttt{temp}_i}}{1-p_{\texttt{temp}_i}}}\right) = \beta_2 \Rightarrow \frac{\frac{p_{\texttt{temp}_i+1}}{1-p_{\texttt{temp}_i+1}}}{\frac{p_{\texttt{temp}_i}}{1-p_{\texttt{temp}_i}}} = e^{\beta_2} \Rightarrow \frac{p_{\texttt{temp}_i+1}}{1-p_{\texttt{temp}_i+1}} = e^{\beta_2}\frac{p_{\texttt{temp}_i}}{1-p_{\texttt{temp}_i}}.$$

So the odds of oring damage at $\texttt{temp}_i + 1$ is e^{β_2} times the odds at $\texttt{temp}_i$. In this particular example $\hat{\beta}_2 = -0.216 \Rightarrow e^{-0.212} = 0.806$ and so the odds of oring damage at $\texttt{temp}_i + 1$ is 0.806 times the odds at $\texttt{temp}_i$ (so a unit temperature increase leads to a decrease in the odds of oring damage). We can also easily attach a confidence interval to this statement.

```
exp(confint.default(oringm)[2,])
```

```
##      2.5 %      97.5 %
## 0.7258160  0.8940366
```

That is, the odds of oring damage at $\texttt{temp}_i + 1$ is 0.806 $(0.726, 0.894)$ times the odds at $\texttt{temp}_i$ Note that if $\beta_2 = 0$ then we would have $e^{\beta_2} = 1$ and there would be no effect on the odds of oring damage.

(c) As observed from the plot in (a), the number of damaged orings at each launch temperature is basically either zero or one (there is only one temperature for which the number of damaged orings is five) and therefore the residual basically can take only two values given a fixed linear predictor. As a result, the residuals will not be smoothly distributed but will show a patterned, non-continuous distribution. To convince ourselves.

```
par(mfrow = c(2, 2))
plot(oringm)
```
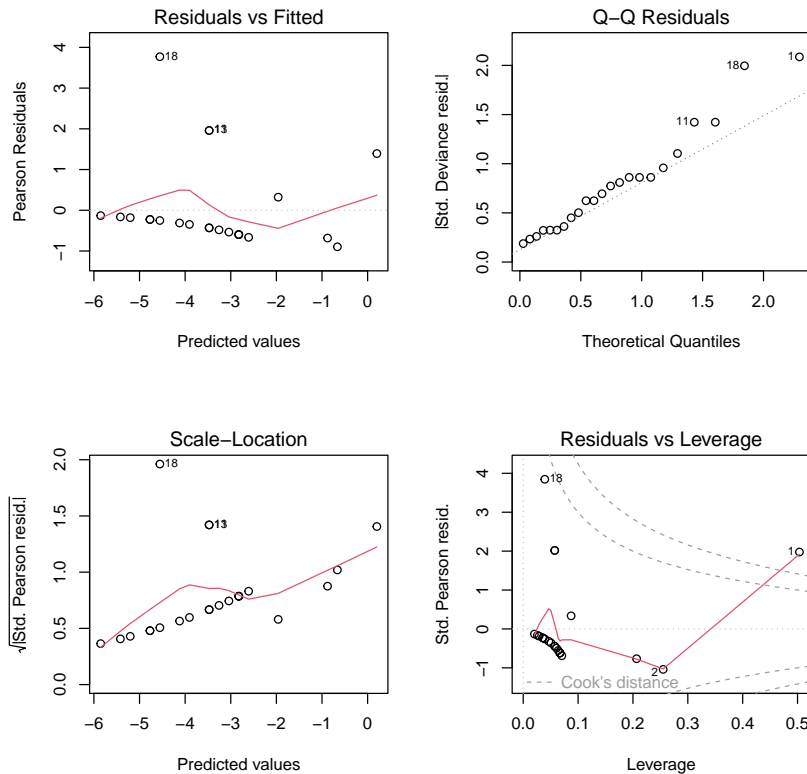
(d) Let's construct the plot asked in the question. Remember that predictions will be on the scale of the linear predictor, so that we need to convert back to the probability scale by applying the inverse of the logit function. We were told in the exercise sheet how this can be easily conducted in R. Note that this is implementing nothing more than

$$p_i = \frac{e^{\beta_1 + \beta_2 \texttt{temp}_i}}{1 + e^{\beta_1 + \beta_2 \texttt{temp}_i}} = \frac{1}{1 + e^{-\beta_1 - \beta_2 \texttt{temp}_i}}.$$

Further note that $\mathbb{E}(y_i) = 6 \times p_i$.

```r
new_temperature <- 30:85
fv <- predict(oringm, data.frame(temperature = new_temperature),
              se = TRUE)
qz <- qnorm(0.975)
ilogit <- binomial()$linkinv

plot(orings$temperature, orings$damaged, ylim = c(0, 6),
     xlim = c(30, 85), xlab = "Temperature", ylab ="Damaged")
lines(new_temperature, ilogit(fv$fit)*6, col = 1)
uc <- fv$fit + qz*fv$se.fit
lines(new_temperature, ilogit(uc)*6, col = 1, lty = 2)
lc <- fv$fit - qz*fv$se.fit
lines(new_temperature, ilogit(lc)*6, col = 1, lty = 2)
```
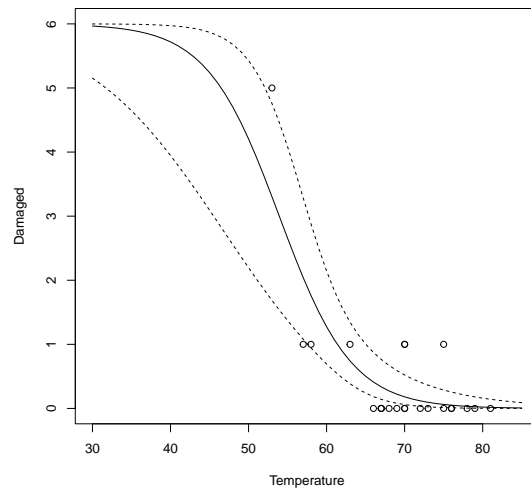
As mentioned before, instead of the `ilogit` function we could have just implemented everything 'manually'.

```
probs <- exp(fv$fit)/(1 + exp(fv$fit))
probs_l <- exp(fv$fit - qz*fv$se.fit)/(1 + exp(fv$fit - qz*fv$se.fit))
probs_u <- exp(fv$fit + qz*fv$se.fit)/(1 + exp(fv$fit + qz*fv$se.fit))

plot(orings$temperature, orings$damaged, ylim = c(0, 6),
     xlim = c(30, 85), xlab = "Temperature", ylab ="Damaged")
lines(new_temperature, probs*6, col = 1)
lines(new_temperature, probs_l*6, col = 1, lty = 2)
lines(new_temperature, probs_u*6, col = 1, lty = 2)
```
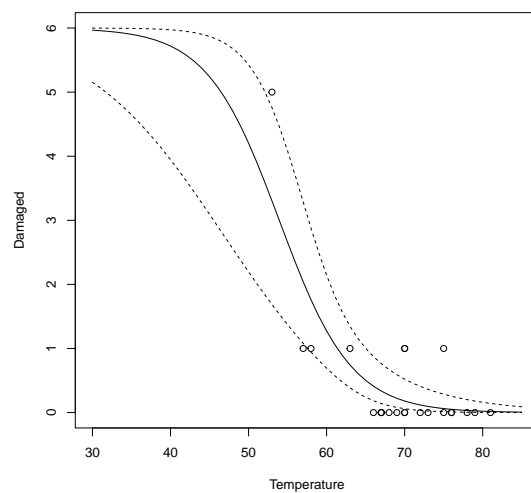

The fit seems reasonable. Further we have the deviance is 16.912 on 21 degrees of freedom which seems acceptable. Let's confirm by computing the approximate p-value.

```r
pchisq(deviance(oringm), df = 21, lower.tail = FALSE)
```

```
## [1] 0.7164099
```

Indeed, there is a quite high probability of a χ^2_{21} r.v. being as large as 16.912.

(e) The expected number of o-ring failures (95%CI) at the Challenger launch temperature of 31 degrees F is 5.958 $(5.067, 5.998)$. This suggest, in fact, a disaster, with 6 (lower bound of 5) out of 6 of the o-rings failing. Note however that we are extrapolating way beyond the observed data (the minimum launch temperature in the data is 53 and most launching temperatures are above 65).

```r
6*ilogit(fv$fit[2])
```

```
##        2
## 5.958205
```

```r
6*ilogit(fv$fit[2] - qz*fv$se.fit[2])
```

```
##        2
## 5.066826
```

```r
6*ilogit(fv$fit[2] + qz*fv$se.fit[2])
```

```
##        2
## 5.998397
```

(f) We know that observations with a Cook's distance above one have the potential to impact the results. Let's check the exact value of Cook's distance for the first observation.

```r
cooks.distance(oringm)[1]
```

```
##        1
## 1.980532
```

A Cook's distance of almost 2! We should remove this observation from the dataset, rerun the analysis, and check if the conclusions are maintained. In particular, does the model still predict a disaster at 31F? I will not repeat everything here but you can try it!!

7. (a) Let's start by fitting fit the independence model as required.

```r
al <- data.frame(y = c(435, 147, 375, 134),
                 gender = as.factor(c("F","F","M","M")),
                 belief = as.factor(c(1, 0, 1, 0)))
al
```

```
##     y gender belief
## 1 435      F      1
## 2 147      F      0
## 3 375      M      1
## 4 134      M      0
```

```r
alm0 <- glm(y ~ gender + belief, family = poisson, data = al)
model.matrix(alm0)
```

```
##    (Intercept) genderM belief1
## 1            1       0       1
## 2            1       0       0
## 3            1       1       1
## 4            1       1       0
## attr(,"assign")
## [1] 0 1 2
## attr(,"contrasts")
## attr(,"contrasts")$gender
## [1] "contr.treatment"
##
## attr(,"contrasts")$belief
## [1] "contr.treatment"


alm0

##
## Call:  glm(formula = y ~ gender + belief, family = poisson, data = al)
##
## Coefficients:
## (Intercept)       genderM       belief1
##       5.010        -0.134         1.059
##
## Degrees of Freedom: 3 Total (i.e. Null);  1 Residual
## Null Deviance:      272.7
## Residual Deviance: 0.162  AIC: 35.41
```

We know that the model in the worksheet is not identifiable and R by default, as confirmed by the model matrix, set females (F comes before M) and believers to the baseline. The deviance of this independence model is 0.162 which should be compared against a chi-squared distribution with one degree of freedom. The p-value (computed below) does not offer any evidence to reject the null hypothesis that the model is fitting the data adequately.

```
pchisq(deviance(alm0), df = 1, lower.tail = FALSE)

## [1] 0.6873263
```

We can also look at the fitted values (which are at the scale of the response) and compare against the entries of the contingency table.

```
fitted(alm0)

##       1       2       3       4
## 432.099 149.901 377.901 131.099
```

The fit appears to be quite close!

(b) In order to check a possible association between gender and after life belief we need to fit a model with an interaction between these two factors and compare against the independence model. The interaction model can be written as

$$\log(\mu_i) = \tau + \beta_k + \delta_j + \zeta_{kj} \text{ if } y_i \text{ is gender } k \text{ and faith } j,$$

where ζ_{kj} is the 'interaction parameter' (of course we know that this model is not identifiable, but we know how to identify it and, in particular, what R will do). Let's now implement this interaction model.

```
alm1 <- glm(y ~ gender*belief, family = poisson, data = al)
alm1

##
## Call:  glm(formula = y ~ gender * belief, family = poisson, data = al)
##
## Coefficients:
##     (Intercept)            genderM            belief1   genderM:belief1
##         4.99043           -0.09259            1.08491          -0.05583
##
## Degrees of Freedom: 3 Total (i.e. Null);  0 Residual
## Null Deviance:      272.7
## Residual Deviance: 1.021e-14   AIC: 37.25
```

To test whether there is evidence for an interaction between gender and believe the null hypothesis that model `alm0` is correct is tested against the more general alternative that `alm1` is correct, using analysis of deviance.

```
anova(alm0, alm1, test = "Chisq")

## Analysis of Deviance Table
##
## Model 1: y ~ gender + belief
## Model 2: y ~ gender * belief
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1         1      0.162
## 2         0      0.000  1    0.162    0.6873
```

A p-value of 0.69 suggests that there is no evidence to reject model `alm0` and the hypothesis of no association between gender and belief in the afterlife. Notice that, in fact, the model with the interaction is the saturated model, which is why its deviance is numerically zero, and there was not really any need to fit it and compare it with the independence model explicitly. In this case we could just as well have examined the deviance of the independence model.