Homework 5

Sarah Adilijiang

Problem 1

(a)

Set $Z = \frac{Y-\mu}{\mu}$, so $Y = \mu(1+Z)$

$$\Rightarrow \quad E(Z) = \frac{E(Y)-\mu}{\mu} = 0, \quad Var(Z) = \frac{Var(Y)}{\mu^2} = \frac{1}{\mu}, \quad E(Z^2) = Var(Z) + (E(Z))^2 = \frac{1}{\mu}$$

Set $\sqrt{Y} = \sqrt{\mu(1+Z)} = g(Z)$, where $g'(Z) = \frac{\sqrt{\mu}}{2\sqrt{1+Z}}$ and $g''(Z) = -\frac{\sqrt{\mu}}{4(1+Z)^{3/2}}$

From Taylor expansion, we have:

$$g(Z) = g(0) + \frac{g'(0)}{1!}(Z-0) + \frac{g''(0)}{2!}(Z-0)^2 + ... + \frac{g^{(k)}(0)}{k!}(Z-0)^k + ...$$

$$i.e. \quad \sqrt{Y} = \sqrt{\mu} + \frac{\sqrt{\mu}}{2}Z - \frac{\sqrt{\mu}}{8}Z^2 + O(Z^3)$$

$$\Rightarrow \quad E(\sqrt{Y}) \approx \sqrt{\mu} + \frac{\sqrt{\mu}}{2}E(Z) - \frac{\sqrt{\mu}}{8}E(Z^2) = \sqrt{\mu} - \frac{\sqrt{\mu}}{8\mu} = \sqrt{\mu}(1 - \frac{1}{8\mu})$$

$$Var(\sqrt{Y}) \approx Var(\sqrt{\mu} + \frac{\sqrt{\mu}}{2}Z) = \frac{\mu Var(Z)}{4} = \frac{1}{4}$$

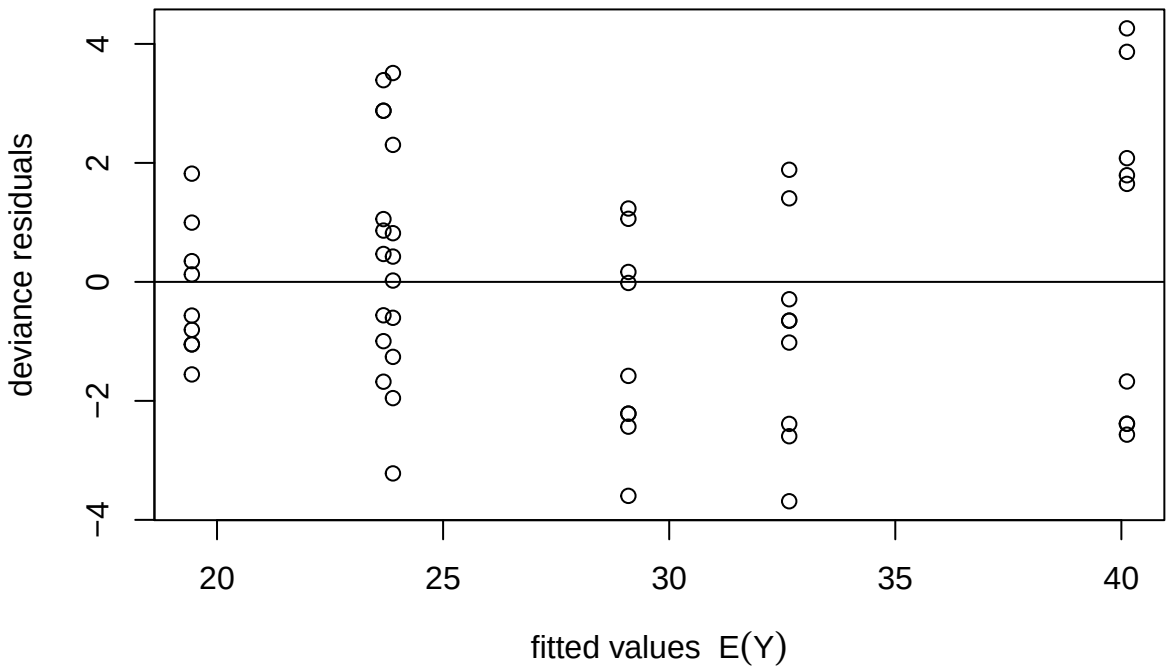
(b)

The GLM model for Poisson (count) response here is:
likelihood: $P(breaks_i = y_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}, \ y_i = 0, 1, 2, ...$
linear predictor: $\eta_i = \beta_0 + \beta_1 1_{wool_i=B} + \beta_2 1_{tension_i=M} + \beta_3 1_{tension_i=H} + \epsilon_i$
link function (log-link): $\eta_i = \log \mu_i$

```
data(warpbreaks)

# fit the Poisson model
mod1 = glm(breaks~wool+tension, warpbreaks, family=poisson)

# plot residuals ~ fitted values
plot(residuals(mod1)~predict(mod1, type="response"),
     ylab="deviance residuals", xlab=expression(paste("fitted values  ", E(Y))))
abline(h=0)
```




```
# model summary
summary(mod1)
```

```
##
## Call:
## glm(formula = breaks ~ wool + tension, family = poisson, data = warpbreaks)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6871  -1.6503  -0.4269   1.1902   4.2616
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.69196    0.04541  81.302  < 2e-16 ***
## woolB       -0.20599    0.05157  -3.994 6.49e-05 ***
## tensionM    -0.32132    0.06027  -5.332 9.73e-08 ***
## tensionH    -0.51849    0.06396  -8.107 5.21e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 297.37  on 53  degrees of freedom
## Residual deviance: 210.39  on 50  degrees of freedom
## AIC: 493.06
##
```

```
## Number of Fisher Scoring iterations: 4
```

```
# significance of predictors
drop1(mod1, test="Chi")
```

```
## Single term deletions
##
## Model:
## breaks ~ wool + tension
##         Df Deviance    AIC    LRT  Pr(>Chi)
## <none>       210.39 493.06
## wool     1   226.43 507.09 16.039 6.206e-05 ***
## tension  2   281.33 560.00 70.942 3.938e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# interpretations of coefficients
exp(coef(mod1))
```

```
## (Intercept)       woolB    tensionM    tensionH
##  40.1235380   0.8138425   0.7251908   0.5954198
```

Answer:

The plot of "residuals ~ fitted values $E(Y)$" shows no significant abnormalities and looks like normally distributed with constant variance, which indicates that the assumptions of the model are not violated and the model structure is fine.

In the summary of the Poisson model, all the coefficients are very significant. However, the residual deviance is 210.39 on 50 degrees of freedom, which indicates a dispersion problem in this Possion model.

If we ignore the dispersion problem and test the significance of the two predictors with chi-square difference-in-deviance LRT test, we see that the pedictors "wool" and "tension" are both very significant.

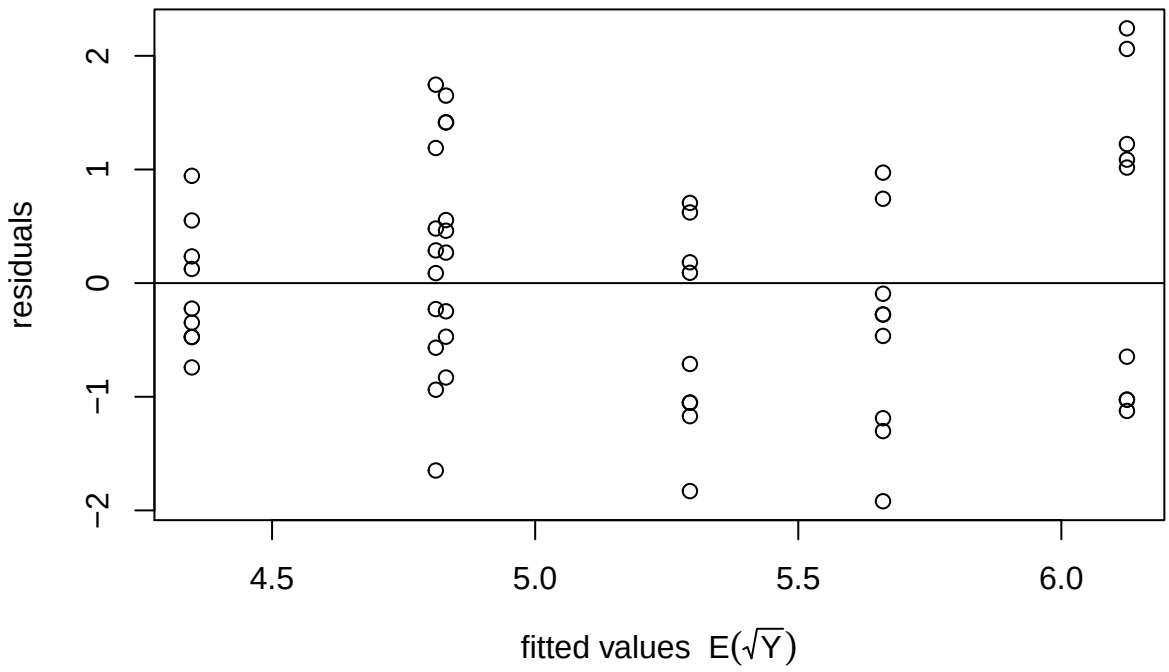
To better interpret the effects of the predictors, I calculated the exponentials of the coefficients. We can see that the number of breaks of wool type B is 0.8138425 times as much as that of wool type A when controlling for "tension". And the number of breaks at tension level M is 0.7251908 times as much as that at tension level L when controlling for "wool", while the number of breaks at tension level H is 0.5954198 times as much as that at tension level L when controlling for "wool".

(c)

The linear model with sqrt transformed response here is:
linear model: $\sqrt{breaks_i} = \beta_0 + \beta_1 1_{wool_i=B} + \beta_2 1_{tension_i=M} + \beta_3 1_{tension_i=H} + \epsilon_i$

```
# fit the sqrt transformed linear model
mod2 = lm(sqrt(breaks)~wool+tension, warpbreaks)

# plot residuals ~ fitted values
plot(residuals(mod2)~predict(mod2),
     ylab="residuals", xlab=expression(paste("fitted values  ", E(sqrt(Y)))) )
abline(h=0)
```




```
# model summary & interpretations of coefficients
summary(mod2)
```

```
##
## Call:
## lm(formula = sqrt(breaks) ~ wool + tension, data = warpbreaks)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9194 -0.7343 -0.1588  0.6849  2.2419
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.1247     0.2810  21.798  < 2e-16 ***
## woolB        -0.4636     0.2810  -1.650 0.105197
## tensionM     -0.8306     0.3441  -2.414 0.019504 *
## tensionH     -1.3136     0.3441  -3.817 0.000373 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.032 on 50 degrees of freedom
## Multiple R-squared:  0.2607, Adjusted R-squared:  0.2164
## F-statistic: 5.878 on 3 and 50 DF,  p-value: 0.001615
```

```
# significance of predictors
drop1(mod2, test="F")
```

```
## Single term deletions
##
## Model:
## sqrt(breaks) ~ wool + tension
##         Df Sum of Sq    RSS     AIC F value   Pr(>F)
## <none>               53.291  7.2859
## wool     1    2.9019 56.193  8.1492  2.7227 0.105197
## tension  2   15.8916 69.182 17.3790  7.4552 0.001467 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer:

The plot of "residuals ~ fitted values $E(\sqrt{Y})$" shows no significant abnormalities and looks like normally distributed with constant variance, which indicates that the assumptions of the model are not violated and the model structure is fine.

In the summary of the sqrt transformed linear model, the coefficients of "tension's" are significant, but the coefficient of "wool" is not significant. And the p-value of the significance of regression F-test is 0.001615, which indicates a good fit of the model.

When we test the significance of the two predictors with F-test, we see that the pedictor "tension" is very significant while the "wool" is not.

From the values of the estimated coefficients, we can see that the absolute value of $\sqrt{breaks}$ of wool type B is 0.4636 less than that of wool type A when controlling for "tension". And the absolute value of $\sqrt{breaks}$ at tension level M is 0.8306 less than that at tension level L when controlling for "wool", while the absolute value of $\sqrt{breaks}$ at tension level H is 1.3136 less than that at tension level L when controlling for "wool".

```r
# Compare the fitted values of mod1 and mod2
mod1_fitted_Y = unique(predict(mod1, type="response"))
mod2_fitted_Y = unique(predict(mod2))^2
rbind(mod1_fitted_Y, mod2_fitted_Y)
```

```
##                   [,1]     [,2]     [,3]     [,4]     [,5]     [,6]
## mod1_fitted_Y 40.12354 29.09722 23.89035 32.65424 23.68056 19.44298
## mod2_fitted_Y 37.51171 28.02770 23.14676 32.04743 23.33358 18.90052
```

Answer:

Here I calculated the $E(Y)$ in the sqrt transformed linear model, which is the square of the fitted values $E(\sqrt{Y})$. Though there are some differences between the two groups of fitted values, their absolute values are quite similar with each other. Thus the sqrt transformed linear model can be used as a good approximation of the Possion model in this case.