problem 1

Ericka Smith

11/15/2020

```
seizure <- read_table2("seizure.txt")
```

Load data

```
## Parsed with column specification:
## cols(
##   y1 = col_double(),
##   y2 = col_double(),
##   y3 = col_double(),
##   y4 = col_double(),
##   trt = col_double(),
##   base = col_double(),
##   age = col_double()
## )
```

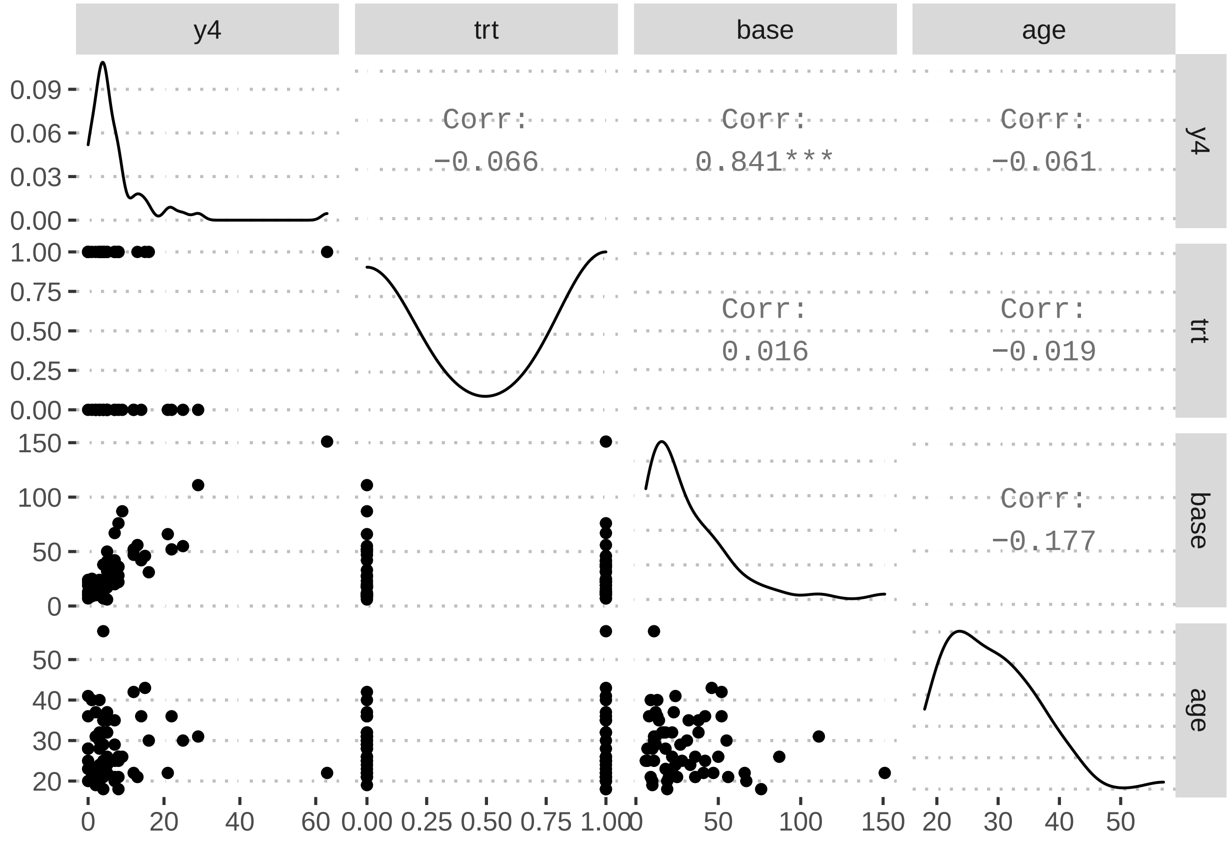
Notes on data: * 0 is placebo, 1 is progabride * y1, y2, y3, y4 are counts for seizures for each successive two weeks * base is number of seizures during baseline 8 wk period

Part A

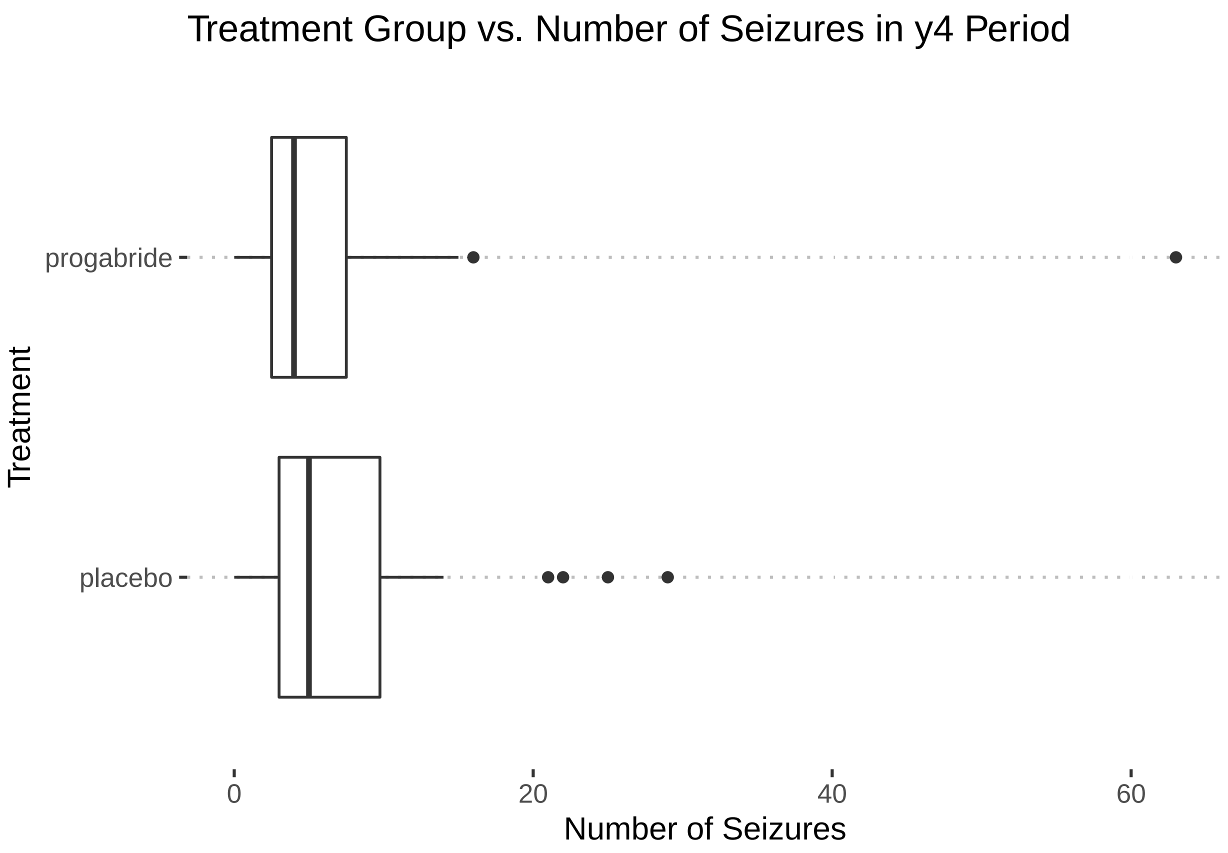
```
mod <- glm(y4 ~ age+trt+base, family = "poisson", data = seizure)
summary(mod)
```

```
##
## Call:
## glm(formula = y4 ~ age + trt + base, family = "poisson", data = seizure)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1813  -0.9729  -0.2089   0.5932   3.9145
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.728587   0.245920   2.963  0.00305 **
## age          0.015612   0.007159   2.181  0.02921 *
## trt         -0.278916   0.098866  -2.821  0.00479 **
## base         0.022153   0.001092  20.293  < 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: 473.79  on 58  degrees of freedom
## Residual deviance: 144.57  on 55  degrees of freedom
## AIC: 340.83
##
## Number of Fisher Scoring iterations: 5
```

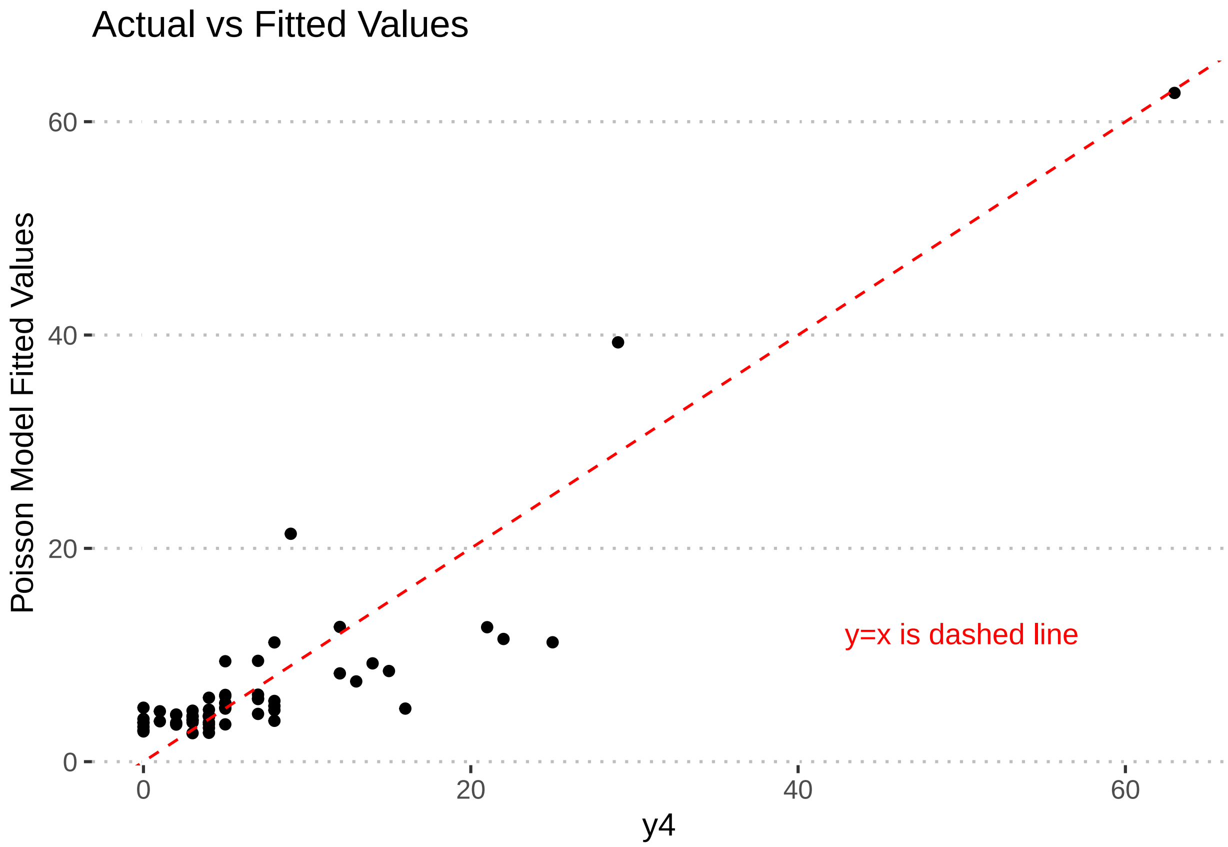
```
ggpairs(seizure[,4:7])+theme_pubclean()
```



```
ggplot(seizure, aes(y = as.factor(trt), x = y4, group=as.factor(trt)))+
  geom_boxplot()+
  theme_pubclean()+
  labs(title = "Treatment Group vs. Number of Seizures in y4 Period",
       x = "Number of Seizures",
       y = "Treatment")+
  scale_y_discrete(labels=c("0" = "placebo", "1" = "progabride"))
```



```
ggplot()+
  geom_point(aes(x = seizure$y4, mod$fitted.values))+
  geom_abline(slope = 1, intercept = 0, color = "red", linetype= "dashed")+
  theme_pubclean()+
  labs(title = "Actual vs Fitted Values",
       x = "y4",
       y = "Poisson Model Fitted Values")+
  annotate("text", x=50, y=12, label="y=x is dashed line", color = "red")
```



My plots indicate that base is highly correlated with y4, which makes sense intuitively because frequency of seizures should at least somewhat rely on prior frequency of seizures. I was surprised to see that trt is significant (alpha = 0.05) in the model despite the fact that the boxplot indicates there isn't a difference in means between the treatment and placebo groups. There a a bunch of reasons this could happen though, and so I defer to the model in this instance. The model gives age, trt, and base all as significant predictors of y4 (p=0.02921, p=0.00479, p< 2e-16, respectively). Based on the coefficients it appears that y4 increases with both age and base, but decreases with the treatment. Again though, just because it's a significant predictor in the model does not clearly state that it is causing a decrease in seizures.

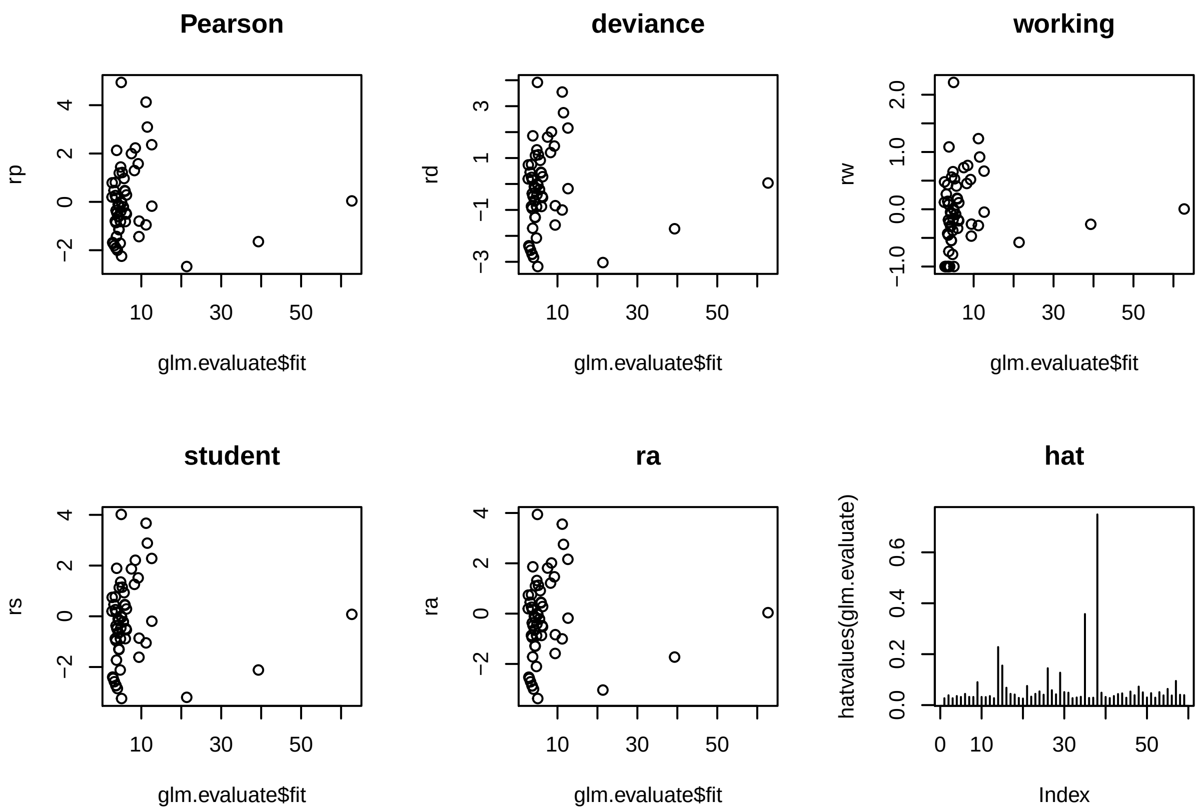
Part B

Residual Analysis

```
glm.evaluate=mod

rp=resid(glm.evaluate, "pearson")
rd=resid(glm.evaluate, "deviance")
rw=resid(glm.evaluate, "working")
rs=rstudent(glm.evaluate)
ra=3*(glm.evaluate$y^{2/3}-glm.evaluate$fit^{2/3})/glm.evaluate$fit^{1/6}/2

par(mfrow=c(2,3))
plot(glm.evaluate$fit,rp, main="Pearson")
plot(glm.evaluate$fit,rd, main = "deviance")
plot(glm.evaluate$fit,rw, main="working")
plot(glm.evaluate$fit,rs, main="student")
plot(glm.evaluate$fit,ra, main="ra")
plot(hatvalues(glm.evaluate), type="h", main="hat")
```



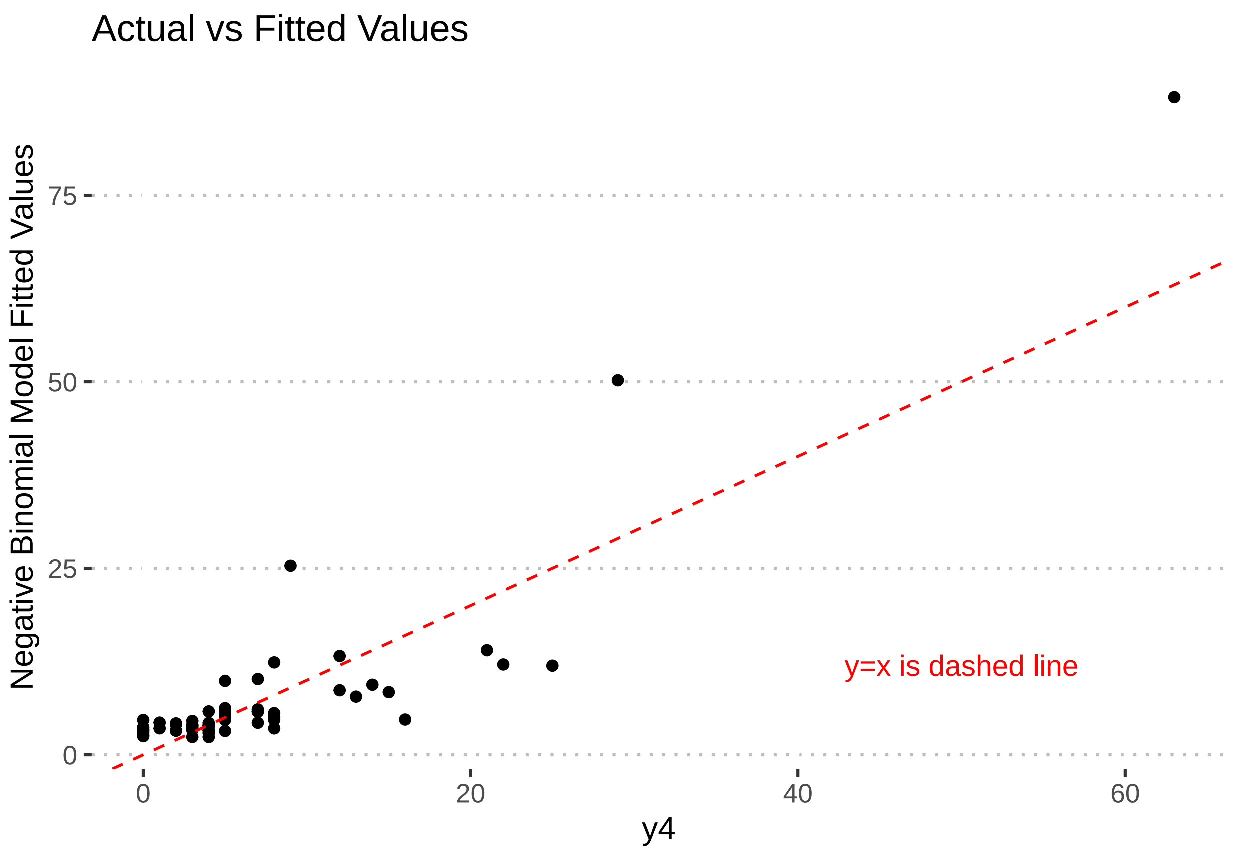
The residual plots indicate that we have a few outliers. Without removing them it's hard to say whether or not the rest of the residuals appear evenly and randomly distributed.

Part C

```
mod_nb <- glm.nb(y4 ~ age+trt+base, data=seizure)
summary(mod_nb)
```

```
##
## Call:
## glm.nb(formula = y4 ~ age + trt + base, data = seizure, init.theta = 4.919155396,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5616  -0.7157  -0.1236   0.5451   2.5743
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.636997   0.378339   1.684   0.0922 .
## age          0.014921   0.011134   1.340   0.1802    
## trt         -0.317674   0.163840  -1.939   0.0525 .  
## base         0.025374   0.002658   9.546   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for Negative Binomial(4.9192) family taken to be 1)
## 
##     Null deviance: 181.952  on 58  degrees of freedom
## Residual deviance:  71.074  on 55  degrees of freedom
## AIC: 314.27
## 
## Number of Fisher Scoring iterations: 1
## 
## 
##               Theta:  4.92 
##           Std. Err.:  1.77 
## 
##  2 x log-likelihood:  -304.268
```

```r
ggplot()+
  geom_point(aes(x = seizure$y4, mod_nb$fitted.values))+
  geom_abline(slope = 1, intercept = 0, color = "red", linetype= "dashed")+
  theme_pubclean()+
  labs(title = "Actual vs Fitted Values",
       x = "y4",
       y = "Negative Binomial Model Fitted Values")+
  annotate("text", x=50, y=12, label="y=x is dashed line", color = "red")
```

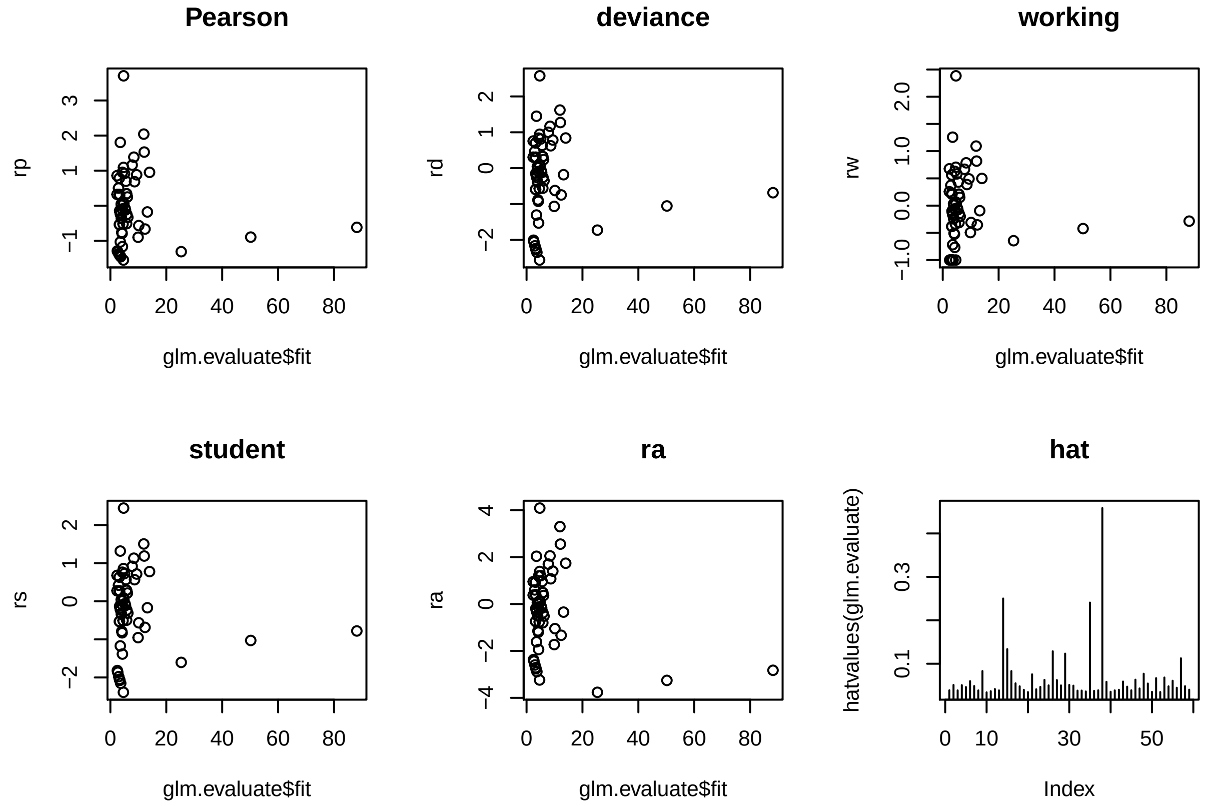


Residual Analysis

```
glm.evaluate=mod_nb

rp=resid(glm.evaluate, "pearson")
rd=resid(glm.evaluate, "deviance")
rw=resid(glm.evaluate, "working")
rs=rstudent(glm.evaluate)
ra=3*(glm.evaluate$y^{2/3}-glm.evaluate$fit^{2/3})/glm.evaluate$fit^{1/6}/2

par(mfrow=c(2,3))
plot(glm.evaluate$fit,rp, main="Pearson")
plot(glm.evaluate$fit,rd, main = "deviance")
plot(glm.evaluate$fit,rw, main="working")
plot(glm.evaluate$fit,rs, main="student")
plot(glm.evaluate$fit,ra, main="ra")
plot(hatvalues(glm.evaluate), type="h", main="hat")
```



Based on the fitted values the model isn't perfect, of course, but it provides a decent fit. The residuals indicate a few outliers so it's difficult to stringently evaluate them, but overall they look okay. The model itself only names base as a significant predictor (p<2e-16). Trt is nearly significant at alpha = 0.05, with p= 0.0525. Age however is not even close, with p=0.1802. Base has a positive coefficient so we expect that more seizures at baseline is associated with more seizures during the y4 period. Age is the same (though not a good predictor), and trt is the opposite (though again, not significant predictor).

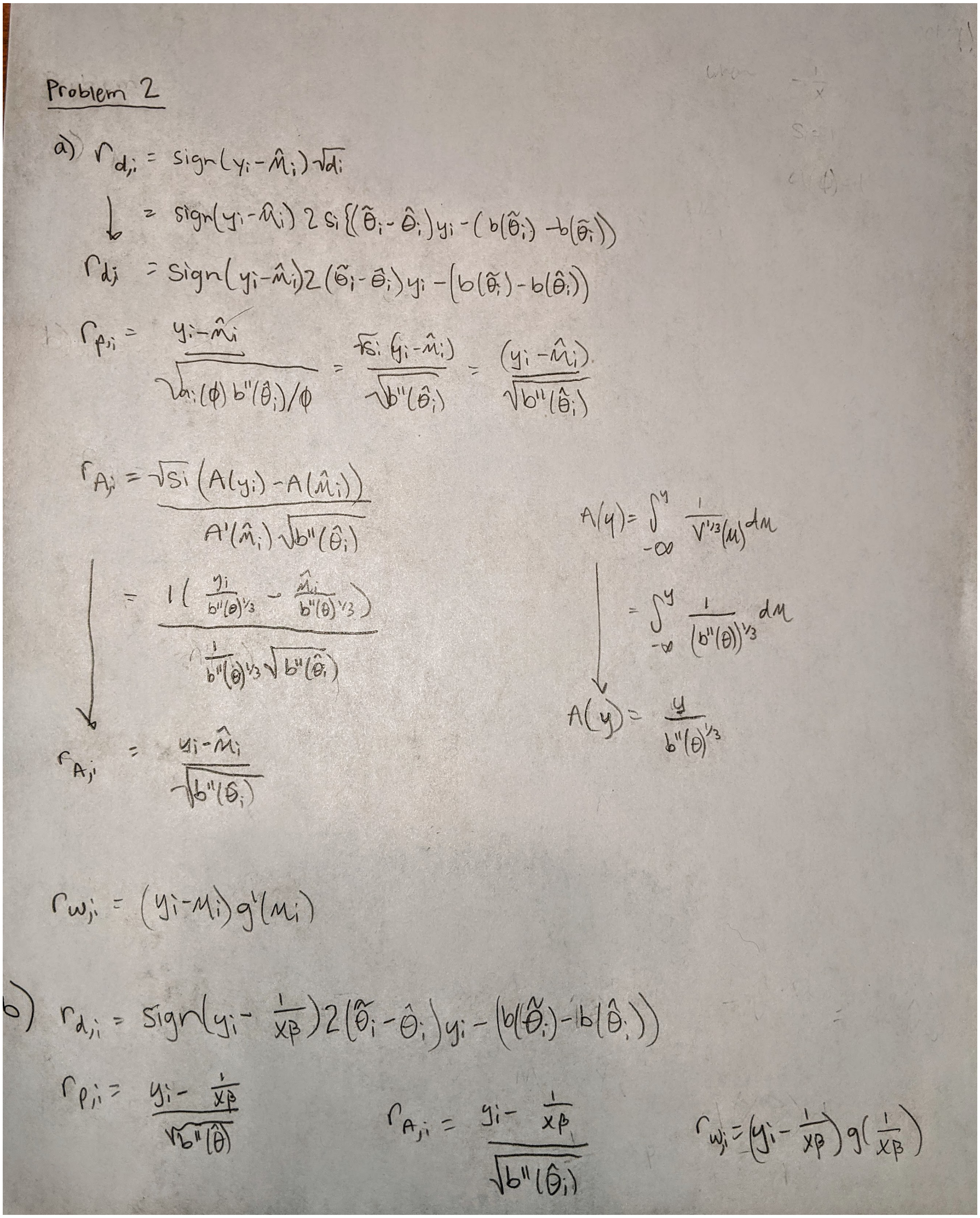
Part D

I believe the negative binomial regression does provide a better fit to the data. It has a lower AIC value. That makes sense, since it has two less predictors than the poisson model does. The AIC for the negative binomial model is 314.27, while it is 340.83 for the poisson model. It should also be noted that the Null deviance and Residual deviance are both smaller in the negative binomial model.

I do think it would be worth trying to look at this data in a way that controls for baseline number of seizures better, so as to get a more clear idea of the efficacy of the treatment. That is the major benefit the poisson model has over the negative binomial model, is that it at least includes the trt variable.

Regardless I believe the negative binomial model is the better fit.

Problem 2



Problem 2

a) $r_{d,i} = \text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i}$

$\downarrow \; = \text{sign}(y_i - \hat{\mu}_i)\, 2\, s_i \{(\tilde{\theta}_i - \hat{\theta}_i) y_i - (b(\tilde{\theta}_i) - b(\tilde{\theta}_i))$

$r_{d,i} = \text{sign}(y_i - \hat{\mu}_i) 2 (\tilde{\theta}_i - \hat{\theta}_i) y_i - (b(\tilde{\theta}_i) - b(\hat{\theta}_i))$

$$r_{p,i} = \frac{y_i - \hat{\mu}_i}{\sqrt{a_i(\phi)\, b''(\hat{\theta}_i)/\phi}} = \frac{\sqrt{s_i}\,(y_i - \hat{\mu}_i)}{\sqrt{b''(\hat{\theta}_i)}} = \frac{(y_i - \hat{\mu}_i)}{\sqrt{b''(\hat{\theta}_i)}}$$

$$r_{A,i} = \frac{\sqrt{s_i}\,(A(y_i) - A(\hat{\mu}_i))}{A'(\hat{\mu}_i)\sqrt{b''(\hat{\theta}_i)}}$$

$$= \frac{1\left(\frac{y_i}{b''(\theta)^{1/3}} - \frac{\hat{\mu}_i}{b''(\theta)^{1/3}}\right)}{\frac{1}{b''(\theta)^{1/3}}\sqrt{b''(\hat{\theta}_i)}}$$

$$\downarrow$$

$$r_{A,i} = \frac{y_i - \hat{\mu}_i}{\sqrt{b''(\hat{\theta}_i)}}$$

$$A(y) = \int_{-\infty}^{y} \frac{1}{V^{1/3}(\mu)}\, d\mu$$

$$\downarrow \; = \int_{-\infty}^{y} \frac{1}{(b''(\theta))^{1/3}}\, d\mu$$

$$A(y) = \frac{y}{b''(\theta)^{1/3}}$$

$r_{w,i} = (y_i - \mu_i)\, g'(\mu_i)$

b) $r_{d,i} = \text{sign}\left(y_i - \frac{1}{X\beta}\right) 2 (\tilde{\theta}_i - \hat{\theta}_i) y_i - (b(\tilde{\theta}_i) - b(\hat{\theta}_i))$

$$r_{p,i} = \frac{y_i - \frac{1}{X\beta}}{\sqrt{b''(\hat{\theta})}} \qquad r_{A,i} = \frac{y_i - \frac{1}{X\beta}}{\sqrt{b''(\hat{\theta}_i)}} \qquad r_{w,i} = \left(y_i - \frac{1}{X\beta}\right) g\left(\frac{1}{X\beta}\right)$$

Figure 1: 2