MA 576 HW 5

Benjamin Draves

```
#load necessary packages
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(ggplot2)
```

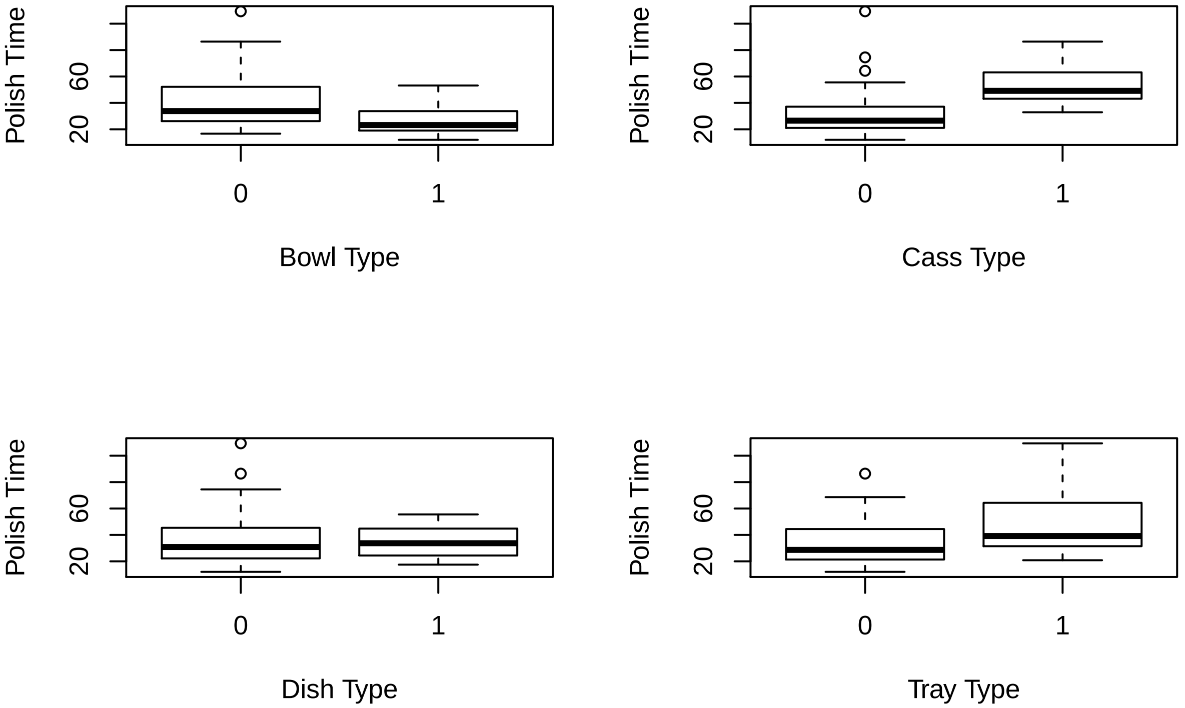
Exercise 3

(a)

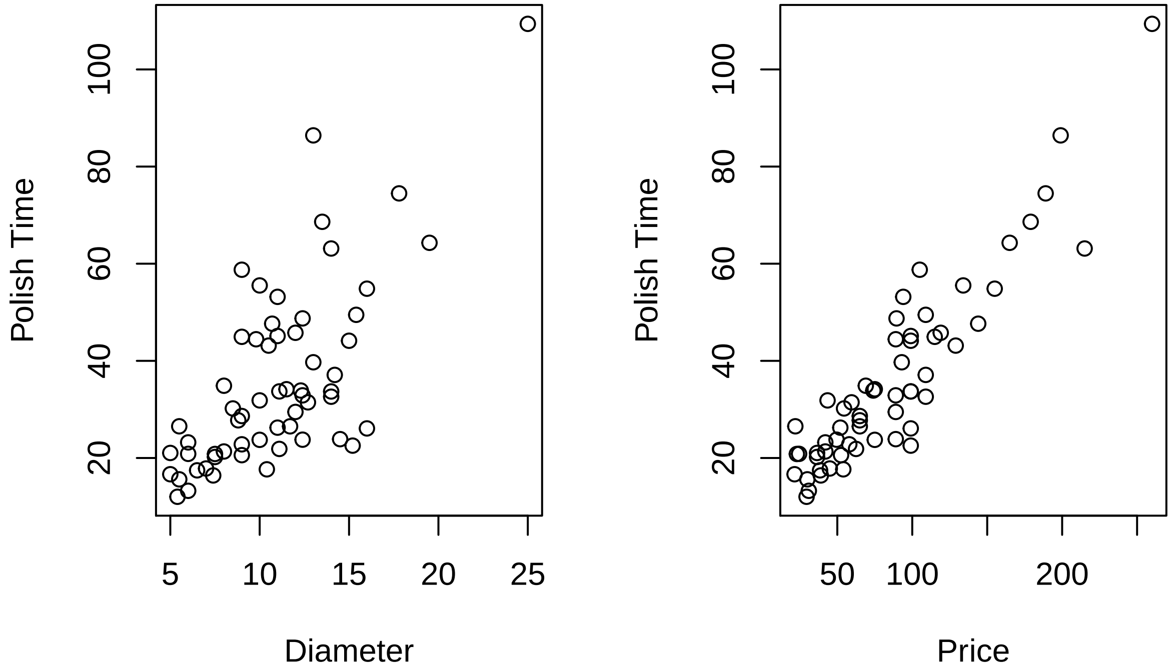
```
polish = read.table("~/Desktop/Courses/MA 576/data/polishing.dat", header = T)
head(polish)
```

```
##   BOWL CASS DISH TRAY DIAM  TIME PRICE
## 1    0    1    0    0 10.7 47.65   144
## 2    0    1    0    0 14.0 63.13   215
## 3    0    1    0    0  9.0 58.76   105
## 4    1    0    0    0  8.0 34.88    69
## 5    0    0    1    0 10.0 55.53   134
## 6    0    1    0    0 10.5 43.14   129
```

```
#categorical covariates
par(mfrow = c(2,2))
boxplot(TIME ~ BOWL, data = polish, xlab = "Bowl Type", ylab = "Polish Time")
boxplot(TIME ~ CASS, data = polish, xlab = "Cass Type", ylab = "Polish Time")
boxplot(TIME ~ DISH, data = polish, xlab = "Dish Type", ylab = "Polish Time")
boxplot(TIME ~ TRAY, data = polish, xlab = "Tray Type", ylab = "Polish Time")
```



```
#cont. covariates
par(mfrow = c(1,2))
plot(TIME~DIAM, data = polish, xlab = "Diameter", ylab = "Polish Time")
plot(TIME~PRICE, data = polish, xlab = "Price", ylab = "Polish Time")
```



(b)

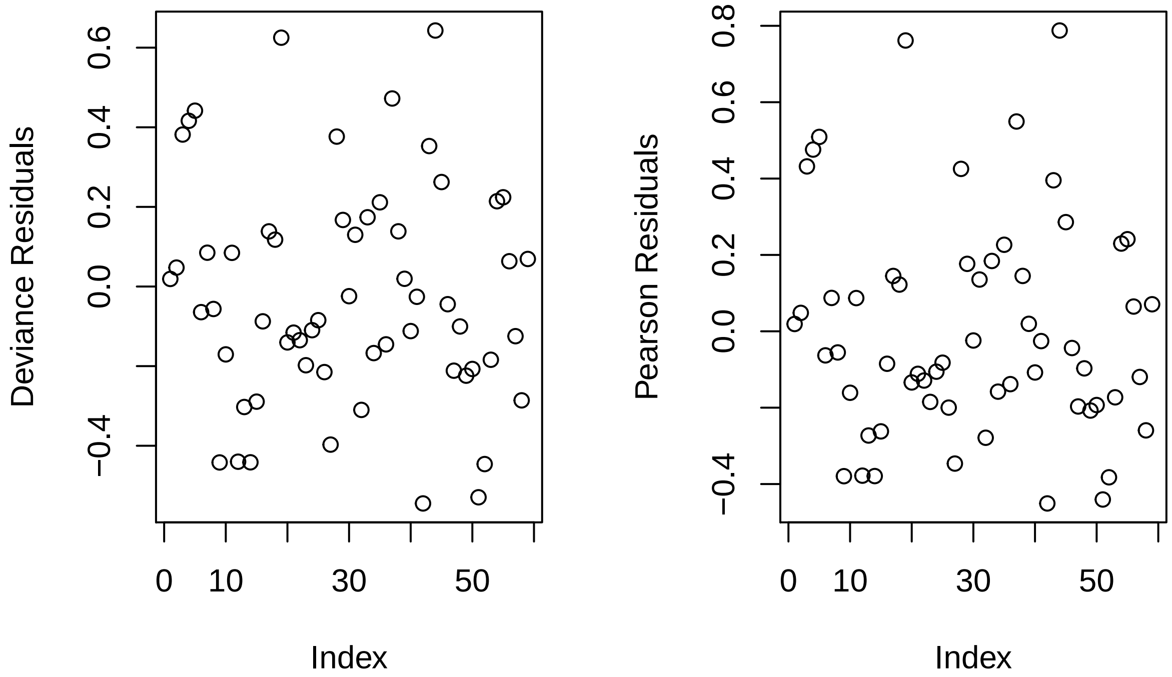
```
#build Gamma glm
m1 = glm(TIME ~ DIAM + BOWL + CASS + TRAY + DISH, data = polish, family = Gamma(link = log))
summary(m1)
```

```
##
## Call:
## glm(formula = TIME ~ DIAM + BOWL + CASS + TRAY + DISH, family = Gamma(link = log),
##     data = polish)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.54489  -0.20244  -0.06442   0.13852   0.64306
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.36106    0.16178  14.594  < 2e-16 ***
## DIAM         0.07671    0.01176   6.525 2.62e-08 ***
## BOWL         0.18791    0.11847   1.586  0.11864
## CASS         0.66308    0.13841   4.791 1.38e-05 ***
## TRAY         0.33264    0.14286   2.328  0.02373 *
## DISH         0.47731    0.15211   3.138  0.00278 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.08899162)
##
##     Null deviance: 14.0053  on 58  degrees of freedom
## Residual deviance:  4.5039  on 53  degrees of freedom
## AIC: 438.65
##
## Number of Fisher Scoring iterations: 4
```

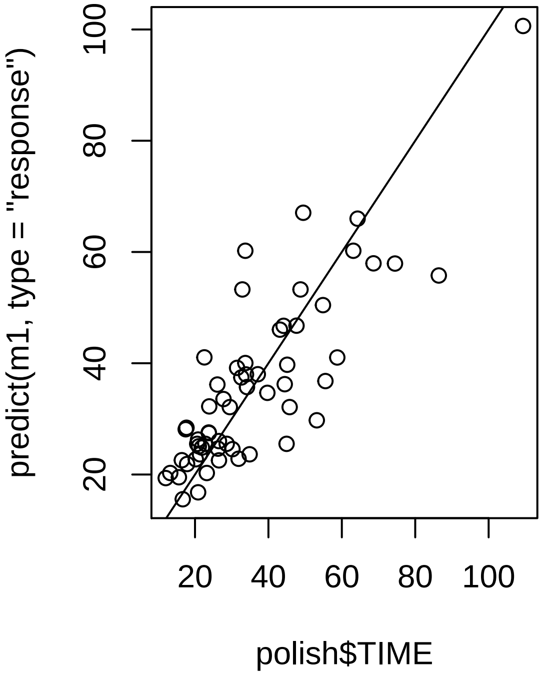
```
#test for underdispersion
X2 = sum(residuals(m1, type = "pearson")^2)/m1$df.residual
pval = pchisq(X2*m1$df.residual, m1$df.residual)
pval
```

```
## [1] 3.707619e-19
```

```
#plot residuals
par(mfrow  = c(1,2))
plot(residuals(m1, type ="deviance"), ylab = "Deviance Residuals")
plot(residuals(m1, type ="pearson"), ylab = "Pearson Residuals")
```



```r
#plots true versus predicted
plot(polish$TIME, predict(m1, type = "response"))
abline(a = 0, b = 1)
```



It appears that this model's residuals are quite appropriate. That is there does not appear to be any clear trend in the pearson or deviance residuals. Moreover they all apear to lay within the ±1 threshold. A test for underdispersion was completed and we have signficant evidence to suggest under dispersion in this model. Even with this underdispersion all coefficents save BOWL were significant at the $\alpha = 0.05$ level. It appears that our model's assumptions about the variance may be too strong.

(c)

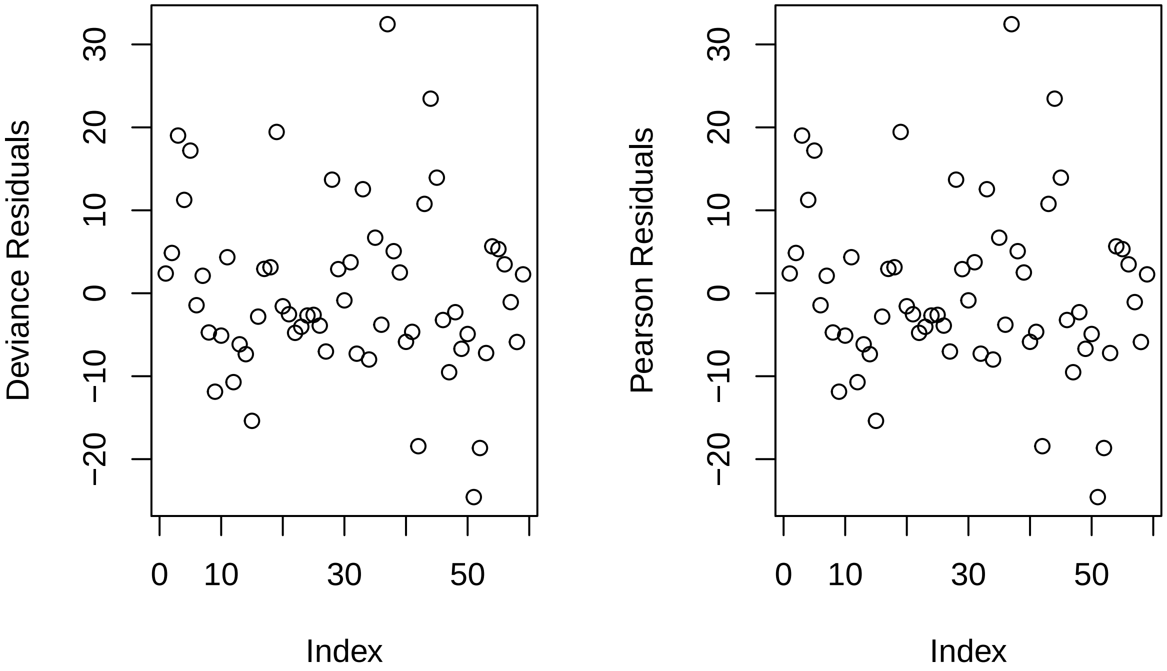
```
#build Gaussian glm
m2 = glm(TIME ~ DIAM + BOWL +CASS+TRAY+DISH, data = polish, family = gaussian(link = log))
summary(m2)
```

```
##
## Call:
## glm(formula = TIME ~ DIAM + BOWL + CASS + TRAY + DISH, family = gaussian(link = log),
##     data = polish)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -24.573   -5.865   -2.286    4.601   32.440
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.283473   0.193870  11.778  < 2e-16 ***
## DIAM        0.076517   0.009879   7.745 2.86e-10 ***
## BOWL        0.266724   0.171758   1.553  0.12640
## CASS        0.710421   0.163762   4.338 6.48e-05 ***
## TRAY        0.457883   0.175025   2.616  0.01156 *
## DISH        0.597666   0.190331   3.140  0.00276 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 116.5986)
##
##     Null deviance: 20974.7  on 58  degrees of freedom
## Residual deviance:  6179.7  on 53  degrees of freedom
## AIC: 455.87
##
## Number of Fisher Scoring iterations: 5
```

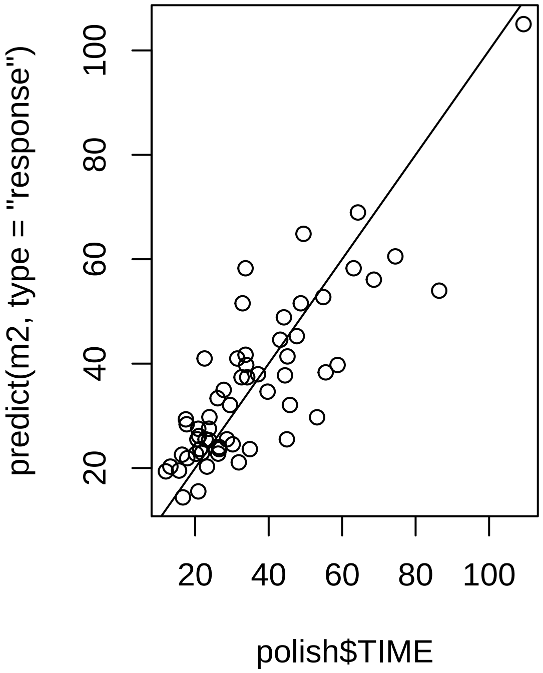
```
#test for overdispersion
X2 = sum(residuals(m2, type = "pearson")^2)/m2$df.residual
pval = 1 - pchisq(X2*m2$df.residual, m1$df.residual)
pval
```

```
## [1] 0
```

```
#plot residuals
par(mfrow  = c(1,2))
plot(residuals(m2, type ="deviance"), ylab = "Deviance Residuals")
plot(residuals(m2, type ="pearson"), ylab = "Pearson Residuals")
```



```r
#plots true versus predicted
plot(polish$TIME, predict(m2, type = "response"))
abline(a = 0, b = 1)
```



This model appears to fit the data very similarily expect for the clear error estimates. That is, the coefficents all appear similar, if not entirely unchanged. This model, however, appears to be extremely overdispersed. A test was completed and we have signfincant evidence to suggest over dispersion here. This is apparent in the deviance residuals that range from ±30.

It appears here that the models are near equivalent save the assumptions on the variance. In either case it appears that the variance doesn't change as a function of the mean. For this reason, I suggest using the Gaussian GLM with the log link and estimate the dispersion for testing purposes.

Exercise 4

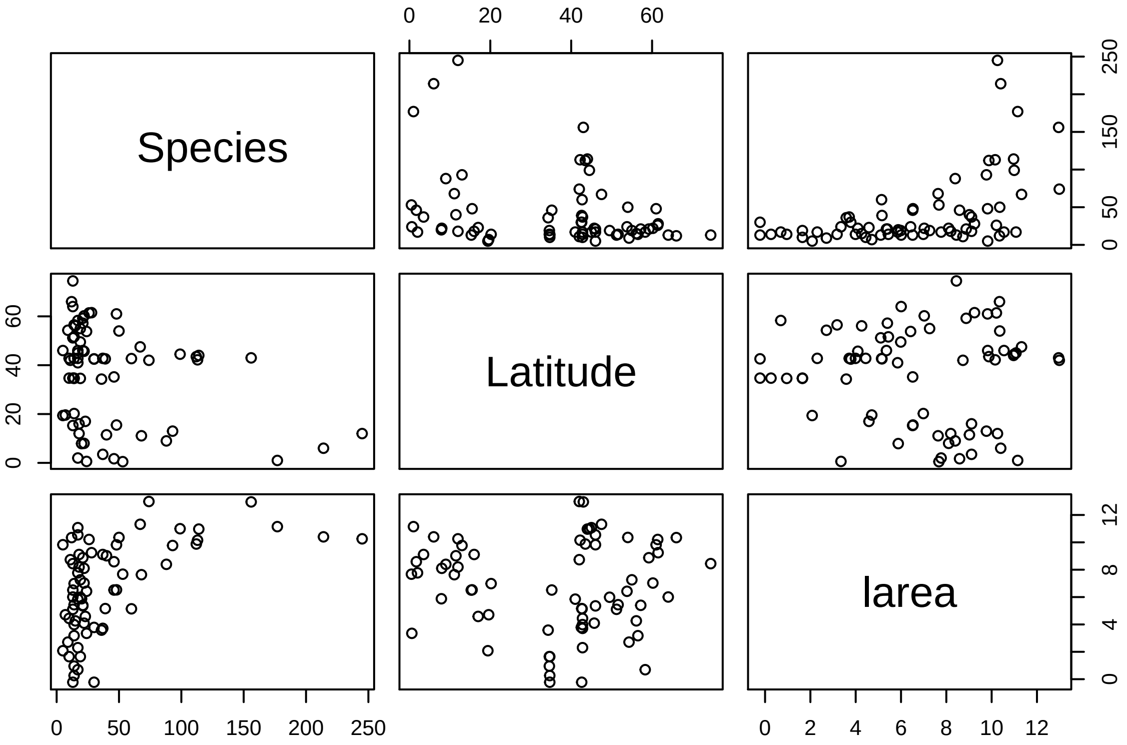
(a)

```
fish = read.table("~/Desktop/Courses/MA 576/data/fish.txt", header = T)
head(fish)
```

```
##   Obs      Name Species  Area Latitude
## 1   1    Albert      46  5346      1.7
## 2   2 Bangweulu      68  2072     11.1
## 3   3      Chad      93 17500     13.0
## 4   4    Chilwa      13   673     15.3
## 5   5    Edward      53  2150      0.5
## 6   6      Kivu      17  2370      2.0
```

```
#make new log area covariate
fish$larea = log(fish$Area)

#make plots
pairs(fish[,c(3,5:6)])
```



(b)

```
m3 = glm(Species~ Latitude + larea, data = fish, family = poisson())
summary(m3)
```

```
##
## Call:
## glm(formula = Species ~ Latitude + larea, family = poisson(),
##     data = fish)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -8.4720  -3.1214  -0.3955   2.0121  12.7145
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.5949644  0.0668158   38.84   <2e-16 ***
## Latitude    -0.0149475  0.0008788  -17.01   <2e-16 ***
## larea        0.2092975  0.0068808   30.42   <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: 2646.1  on 69  degrees of freedom
## Residual deviance: 1249.9  on 67  degrees of freedom
## AIC: 1616.4
##
## Number of Fisher Scoring iterations: 5
```

```
#test for over dispersion
X2 = sum(residuals(m3, type = "pearson")^2)/m3$df.residual
1 - pchisq(X2*m3$df.residual, m3$df.residual)
```

```
## [1] 0
```

We have evidence for overdispersion as evident by the asympototic hypothesis test.

(c)

```
m4 = glm(Species~ Latitude + larea, data = fish, family = quasipoisson())
summary(m4)
```

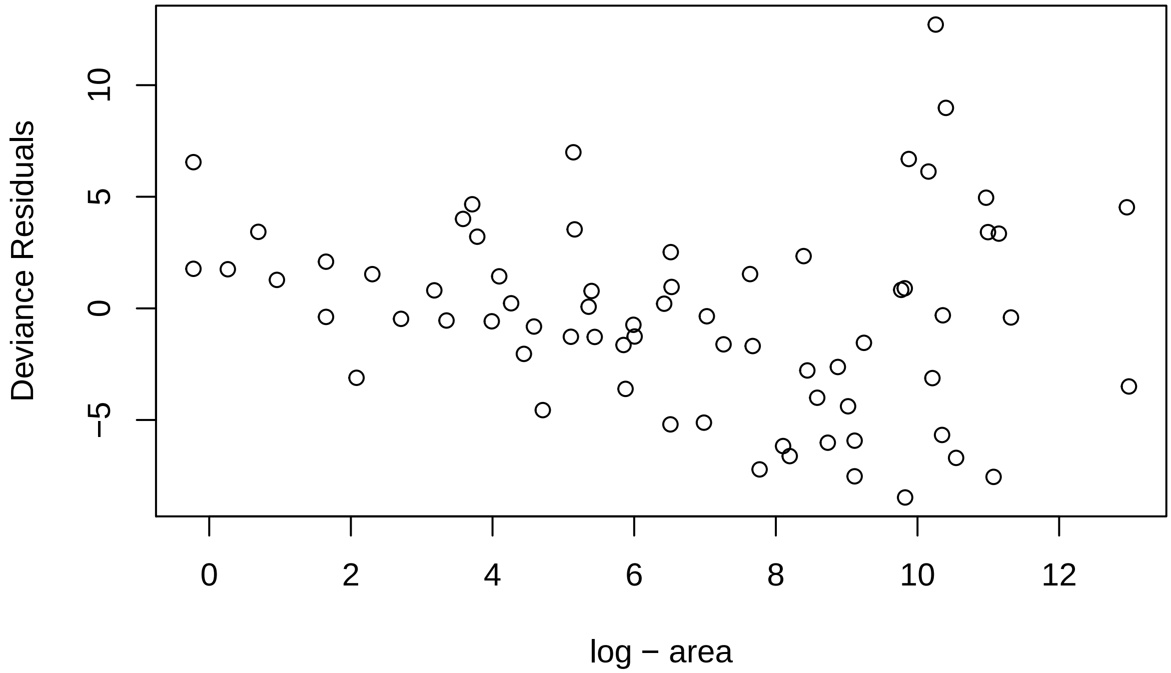
```
##
## Call:
## glm(formula = Species ~ Latitude + larea, family = quasipoisson(),
##     data = fish)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -8.4720  -3.1214  -0.3955   2.0121  12.7145
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.594964   0.293185   8.851 7.08e-13 ***
## Latitude    -0.014947   0.003856  -3.876 0.000244 ***
## larea        0.209298   0.030192   6.932 2.00e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 19.25417)
##
##     Null deviance: 2646.1  on 69  degrees of freedom
## Residual deviance: 1249.9  on 67  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

```
#test for over dispersion
X2 = sum(residuals(m4, type = "pearson")^2)/m4$df.residual
1 - pchisq(X2*m4$df.residual, m4$df.residual)
```

```
## [1] 0
```

```
#plot deviane residuals
plot(fish$larea, residuals(m4, type = "deviance"), xlab = "log - area", ylab = "Deviance Residu
```

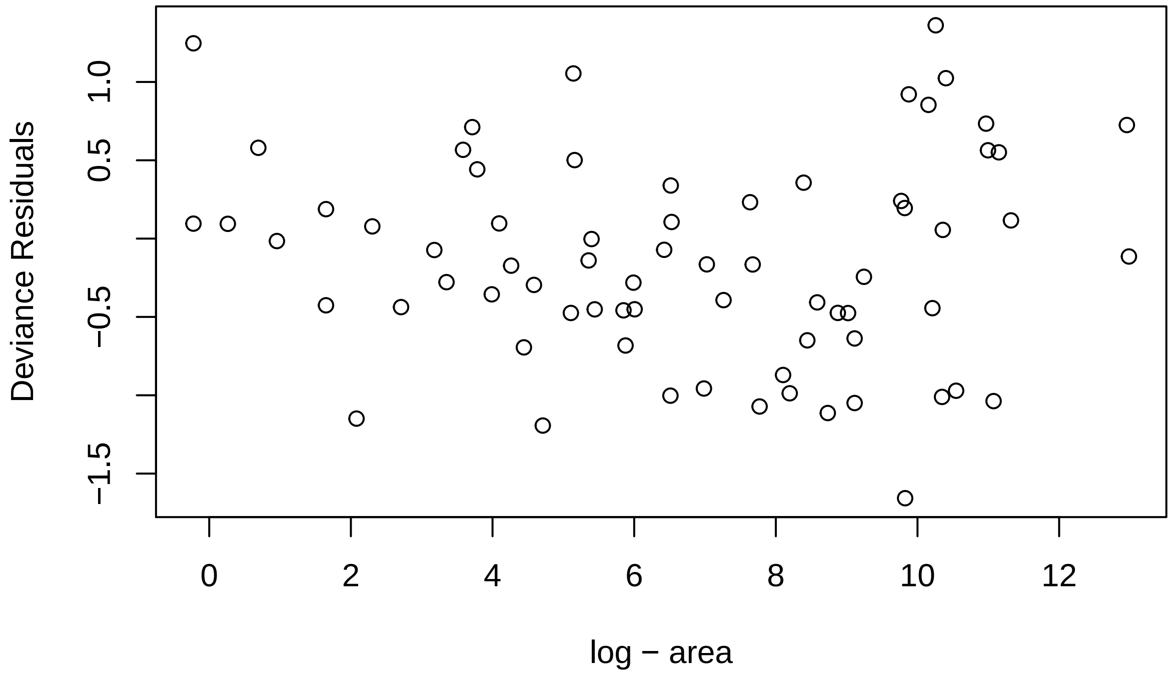


(d)

```
m5 = glm(Species~ Latitude + larea, data = fish, family = quasi(variance = "mu^2", link = "log
summary(m5)
```

```
##
## Call:
## glm(formula = Species ~ Latitude + larea, family = quasi(variance = "mu^2",
##     link = "log"), data = fish)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6571  -0.4751  -0.1648   0.2384   1.3616
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.958730   0.238858  12.387  < 2e-16 ***
## Latitude    -0.013061   0.004069  -3.210  0.00204 **
## larea        0.154812   0.024085   6.428 1.58e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasi family taken to be 0.4447408)
##
##     Null deviance: 59.894  on 69  degrees of freedom
## Residual deviance: 31.121  on 67  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 8
```

```
#plot deviane residuals
plot(fish$larea, residuals(m5, type = "deviance"), xlab = "log - area", ylab = "Deviance Resid
```



#### (e) The first model is very clearly not modeling all of the variance in the model. The assumption that this data has the variane - mean relation impossed by the poisson GLM is not a

good one.

The second model assumes the variance structure is given by $V(\mu) = \sigma^2\mu$ for some constant σ^2 that does not vary over the population. Well, plotting the deviance residuals of this model shows that the variance increases as the area of the lake also increases. Therefore this is not a good assumption.

The quasi-glm is the most appropriate fit for this data. It models the changing variance for each data point as evident in the last deviance residual plot where it appears that the linear space predictions have a constant variance after scaling by the associated estimates $\hat{V_i}$.

Using this last model to for inferential reason we see that the a lake at 0 latitude with area of 1 that the number of expected specicies is given by $\exp(2.958730) = 19.27348$. Moreover, holding area constant, for each additional degree of latitude away from the equator we expect the number of species to decrease by a factor of $exp(-0.013061) = 0.9870239$. Lastly, holding latidude constant, we expect the number of species to incease by a factor for $\exp(0.154812) = 1.167438$ for each additional unit increase in $\log(area)$.