# Stat 230 HW 8

Name: Victor Huang

Homework 8 is due **by 3pm Thursday, Nov 18**. Please complete the assignment in this Markdown document, filling in your answers and R code below. I didn't create answer and R chunk fields like I did with homework 1, but please fill in your answers and R code in the same manner as hw 1. Submit a hard copy of the **compiled pdf or word doc** either

- in class
- in drop-in office hours
- in the paper holder outside my CMC 222 office door

Tips for using Markdown with homework sets:

- Work through a problem by putting your R code into R chunks in this .Rmd. Run the R code to make sure it works, then knit the .Rmd to verify they work in that environment.
  - Make sure you load your data in the .Rmd and include any needed `library` commands.
- Feel free to edit or delete questions, instructions, or code provided in this file when producing your homework solution.
- For your final document, you can change the output type from `html_document` to `word_document` or `pdf_document`. These two to output types are better formatted for printing.
  - on maize: you may need to allow for pop-ups from this site
- If you want to knit to pdf while running Rstudio from your computer (*not* from maize), you will need a LaTeX compiler installed on your computer. This could be MiKTeX, MacTeX (mac), or TinyTex. The latter is installed in R: first install the R package `tinytex`, then run the command `tinytex::install_tinytex()` to install this software.
  - If you are using maize, you don't need to install anything to knit to pdf!

---

## Problem 1: USGS Rake data

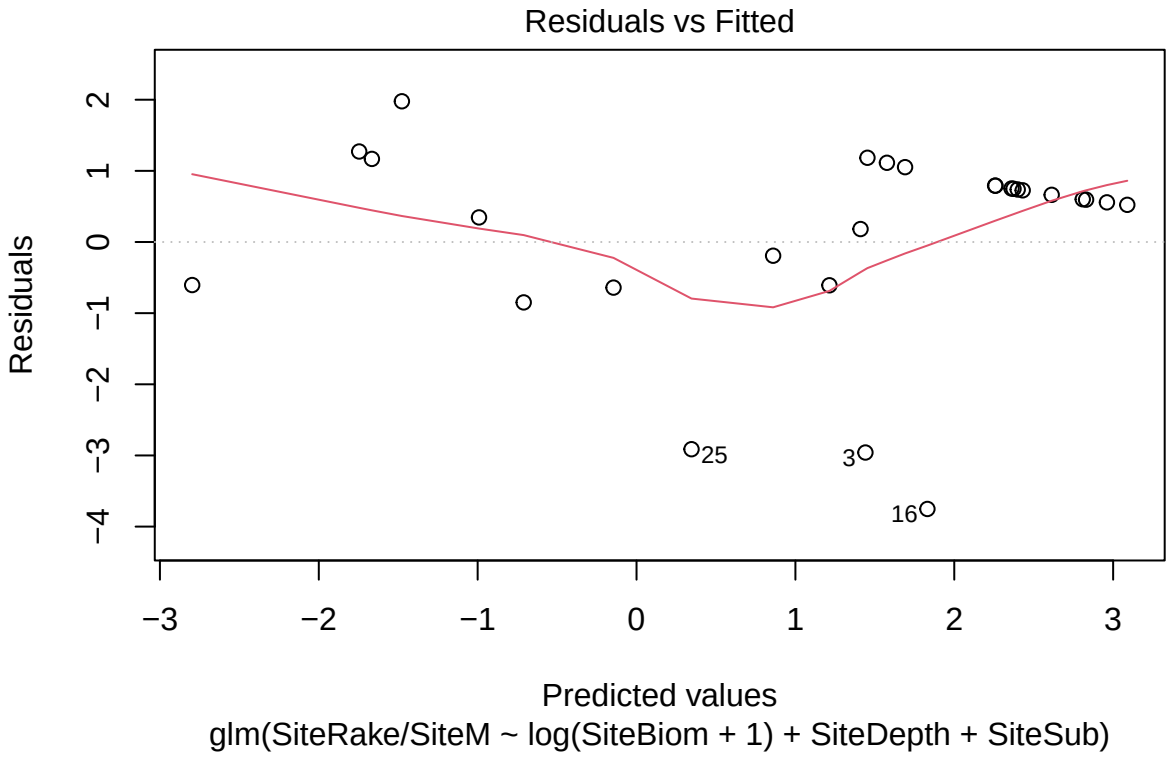
Consider the rake data used in the day 23 quasi-binomial worksheet.

```
> RakeData <- read.csv("http://people.carleton.edu/~kstclair/data/RakeData.csv")
> summary(RakeData)
       X            SiteRake         SiteM      SiteBiom
 Min.   : 1.00   Min.   :0.000   Min.   :6   Min.   :   0.0162
 1st Qu.: 8.50   1st Qu.:2.000   1st Qu.:6   1st Qu.:  53.9410
 Median :16.00   Median :6.000   Median :6   Median : 225.2125
 Mean   :15.85   Mean   :4.185   Mean   :6   Mean   : 395.5006
 3rd Qu.:23.50   3rd Qu.:6.000   3rd Qu.:6   3rd Qu.: 717.8724
```

```
 Max.   :30.00   Max.   :6.000   Max.   :6   Max.   :1293.2130
  SiteDepth        SiteSub
 Min.   :0.1000   Length:27
 1st Qu.:0.5000   Class :character
 Median :0.6333   Mode  :character
 Mean   :0.6660
 3rd Qu.:0.8000
 Max.   :1.3000
```

**(1a) Fit the regular (no quasi-) binomial model that was fit in the day 23 quasi-binomial worksheet: the binomial regression of $y$ on log of biomass, site substrate and site depth. Plot the deviance residuals against fitted values for this model and give the row numbers of the three cases with the most extreme residual deviance values. What are the response values for these cases? Are these cases over- or underestimated?**

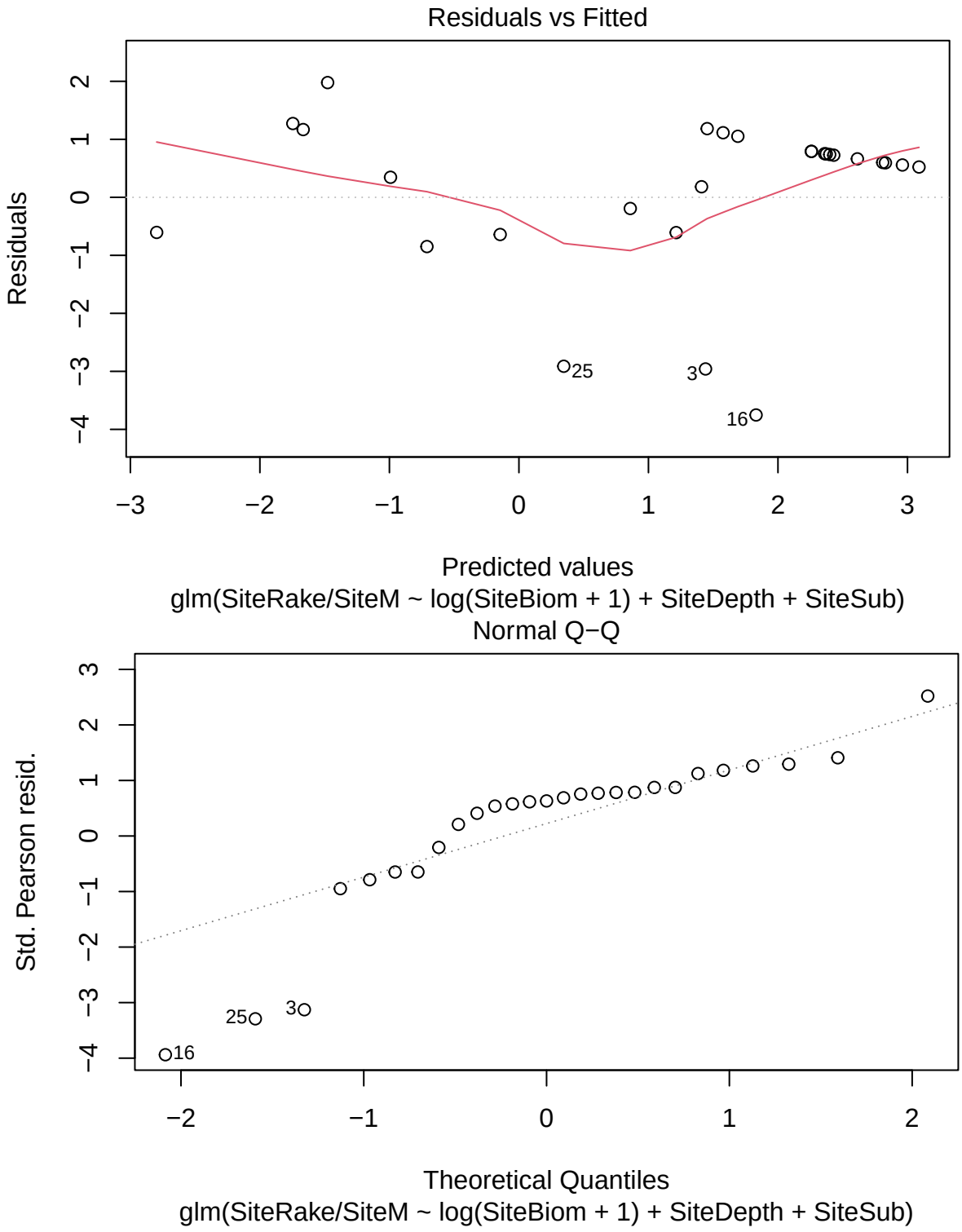
```
> rake_glm_binom <- glm(SiteRake/SiteM ~ log(SiteBiom + 1) + SiteDepth + SiteSub, family = binomial, we
> rake_aug <- augment(rake_glm_binom, type.residuals = "deviance")
> rake_aug %>% arrange(desc(.resid))
# A tibble: 27 x 11
   `SiteRake/SiteM` `log(SiteBiom + 1)` SiteDepth SiteSub `(weights)` .fitted
              <dbl>               <dbl>     <dbl> <chr>         <int>   <dbl>
 1            0.5                 2.04       1.3  silt              6   -1.48
 2            1                   5.26       0.5  sand              6    1.45
 3            1                   5.42       0.5  sand              6    1.58
 4            1                   4.78       0.4  silt              6    1.69
 5            0.333               1.48       0.8  sand              6   -1.75
 6            1                   5.66       0.1  sand              6    2.26
 7            1                   6.76       1.13 silt              6    2.26
 8            0.333               0.957      0.8  silt              6   -1.67
 9            1                   6.97       0.8  sand              6    2.36
10            1                   5.69       0.4  silt              6    2.37
# ... with 17 more rows, and 5 more variables: .resid <dbl>, .std.resid <dbl>,
#   .hat <dbl>, .sigma <dbl>, .cooksd <dbl>
> rake_aug %>% slice_max(abs(.resid), n = 3)
# A tibble: 3 x 11
  `SiteRake/SiteM` `log(SiteBiom +~ SiteDepth SiteSub `(weights)` .fitted .resid
             <dbl>            <dbl>     <dbl> <chr>         <int>   <dbl>  <dbl>
1            0                4.55     0.967 sand              6   0.346  -3.25
2            0.333            5.36     0.633 silt              6   1.83   -2.97
3            0.333            4.78     0.6   silt              6   1.44   -2.54
# ... with 4 more variables: .std.resid <dbl>, .hat <dbl>, .sigma <dbl>,
#   .cooksd <dbl>
> plot(rake_glm_binom, which = 1)
```

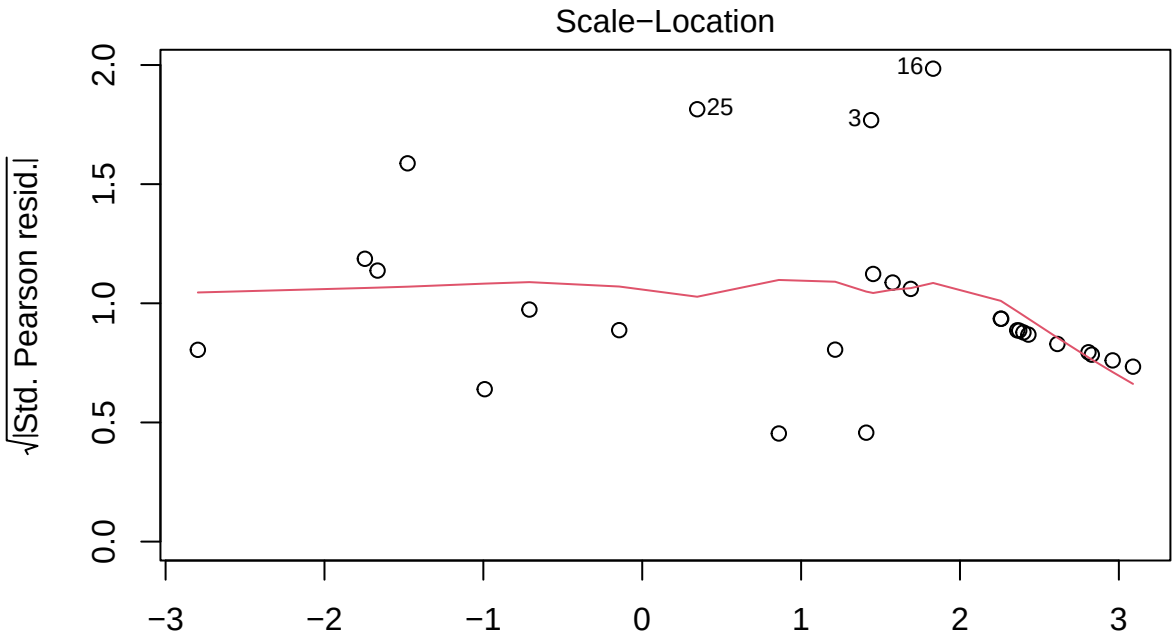


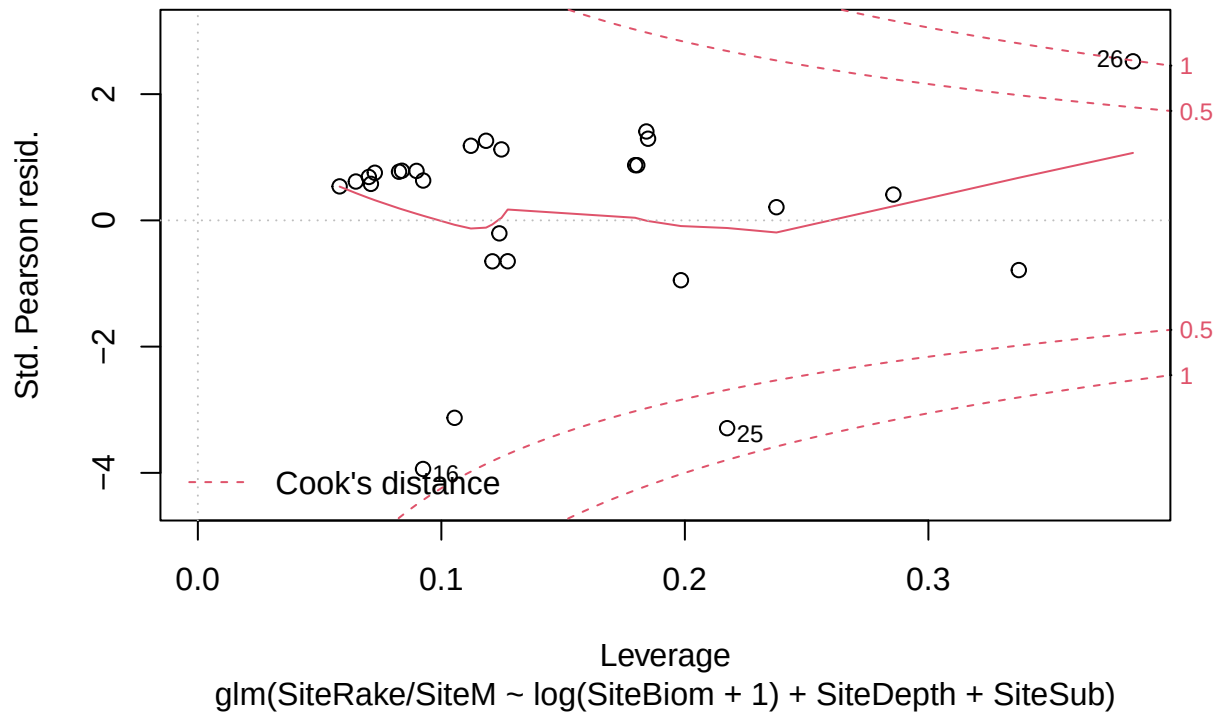

The three most extreme cases have rows 25, 3, and 16. With response values 0, 0.333, and 0.333. They are all underestimated

**(1b) Look at the Cook's distance measure for the regular binomial model. Which case has the highest value? Determine why these cases have high Cook's distance values.**

```
> plot(rake_glm_binom)
```


Residuals vs Fitted
Residuals
Predicted values
glm(SiteRake/SiteM ~ log(SiteBiom + 1) + SiteDepth + SiteSub)
Normal Q−Q
Std. Pearson resid.
Theoretical Quantiles
glm(SiteRake/SiteM ~ log(SiteBiom + 1) + SiteDepth + SiteSub)




```
> resid(rake_glm_binom)
         1          2          3          4          5          6          7
 1.0928964  1.4253688 -2.5371589  1.0403453  0.1861771  1.5017582  0.7310881
         8          9         10         11         12         13         14
 1.1290686 -0.6490350 -0.5819784  1.0341519  0.8296458 -0.9007835  1.0059172
        15         16         17         18         19         20         21
```

```
 0.8381422 -2.9652024  1.5877209  0.7783168  0.9214129 -0.8429167 -0.1905666
        22         23         24         25         26         27
 1.0916332  1.0501179  0.3382906 -3.2516121  1.7354841  1.0210740
> rake_aug %>% arrange(desc(.cooksd))
# A tibble: 27 x 11
   `SiteRake/SiteM` `log(SiteBiom + 1)` SiteDepth SiteSub `(weights)` .fitted
              <dbl>               <dbl>     <dbl> <chr>         <int>   <dbl>
 1            0.5                 2.04      1.3   silt              6  -1.48
 2            0                   4.55      0.967 sand              6   0.346
 3            0.333               5.36      0.633 silt              6   1.83
 4            0.333               4.78      0.6   silt              6   1.44
 5            0.333               1.48      0.8   sand              6  -1.75
 6            0.333               0.957     0.8   silt              6  -1.67
 7            0.333               2.16      0.3   silt              6  -0.145
 8            0.167               2.70      0.7   sand              6  -0.710
 9            1                   5.26      0.5   sand              6   1.45
10            1                   4.78      0.4   silt              6   1.69
# ... with 17 more rows, and 5 more variables: .resid <dbl>, .std.resid <dbl>,
#   .hat <dbl>, .sigma <dbl>, .cooksd <dbl>
> rake_aug %>% slice_max(.cooksd, n = 1)
# A tibble: 1 x 11
  `SiteRake/SiteM` `log(SiteBiom +~ SiteDepth SiteSub `(weights)` .fitted .resid
             <dbl>            <dbl>     <dbl> <chr>         <int>   <dbl>  <dbl>
1              0.5             2.04       1.3 silt              6   -1.48   1.74
# ... with 4 more variables: .std.resid <dbl>, .hat <dbl>, .sigma <dbl>,
#   .cooksd <dbl>
```

The highest cook's distance is the 26 row. This can be caused by it having a high standard residual.

**(1c) Refit the regular binomial model without the *three cases from part (a)*. Run the goodness-of-fit test. Explain why the results of this test change compared to the results of the GOF test with all cases (done in the day 22 markdown)?**

```
> rake_new_glm <- update (rake_glm_binom, subset = c(-3, -25, -16))
> p1 <- 1-pchisq(15.557, df = 20)
```

Ater removing the three cases we get a high p-value. This means that our original model was good and the change is caused by the removing the three outliers.

**(1d) Refit the regular binomial model without the *one case from part (b)* but including the three cases from part (c). Run the goodness-of-fit test. Explain why the GOF test from (c) suggests that the model is adequate (with the 3 unusual residual cases removed) but the test in (d) is not adequate (with the highest Cook's distance case removed).**

```
> rake_new_glm2 <- update (rake_glm_binom, subset = c(-26))
> p2 <- 1-pchisq(44.886, df = 22)
```

**Since the p-value is significantly small. The original model proves to be inadequate due to the three outliers being removed in the last question**

## Problem 2: Galapagos: ch. 22 exercise 18

The data set is `ex1220`. **In addition** to parts (a)-(c):

- use a **deviance residual plot** to verify your GOF conclusions in part (a) (remember that most residuals are between $+/-2$ if the model fits well)
- and use a **quasi-Poisson** model for (b)-(c) if your GOF test suggests that it is needed.

```
> ex1220 <- ex1220
> ex1220_glm <- glm(Native ~ log(Area) + log(Elev) + log(DistNear) + log(AreaNear), family = quasipoiss
> summary(ex1220_glm)

Call:
glm(formula = Native ~ log(Area) + log(Elev) + log(DistNear) +
    log(AreaNear), family = quasipoisson(), data = ex1220)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.4515  -1.6623   0.2330   0.7056   3.6582

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.22136    0.88621   2.507 0.019059 *
log(Area)     0.24788    0.05718   4.335 0.000209 ***
log(Elev)     0.07663    0.17979   0.426 0.673576
log(DistNear) -0.06046    0.04071  -1.485 0.150036
log(AreaNear) -0.05163    0.02089  -2.471 0.020623 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 3.720004)

    Null deviance: 700.717  on 29  degrees of freedom
Residual deviance:  95.764  on 25  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
> 1 -pchisq(95.764, df = 25)
[1] 3.210995e-10
> ex1220_glm2 <- glm(Native ~ log(Area) + log(DistNear) + log(AreaNear), family = quasipoisson(), data =
> summary(ex1220_glm2)

Call:
glm(formula = Native ~ log(Area) + log(DistNear) + log(AreaNear),
    family = quasipoisson(), data = ex1220)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.4748  -1.6666   0.1745   0.6280   3.8165
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.59389    0.13782  18.821  < 2e-16 ***
log(Area)     0.27002    0.02388  11.306 1.55e-11 ***
log(DistNear) -0.06099    0.04001  -1.524   0.1395
log(AreaNear) -0.05023    0.02041  -2.461   0.0208 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 3.623204)

    Null deviance: 700.717  on 29  degrees of freedom
Residual deviance:  96.448  on 26  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
> 1 -pchisq(96.448, df = 26)
[1] 4.965645e-10
> ex1220_glm3<- glm(Native ~ log(Area) + log(Elev) + log(AreaNear), family = quasipoisson(), data = ex1
> summary(ex1220_glm3)

Call:
glm(formula = Native ~ log(Area) + log(Elev) + log(AreaNear),
    family = quasipoisson(), data = ex1220)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3153  -1.6185  -0.2067   1.0482   3.6110

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.07440    0.89221   2.325 0.028145 *
log(Area)     0.25376    0.05783   4.388 0.000169 ***
log(Elev)     0.08592    0.18142   0.474 0.639740
log(AreaNear) -0.05098    0.02152  -2.369 0.025528 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 4.00664)

    Null deviance: 700.72  on 29  degrees of freedom
Residual deviance: 104.08  on 26  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
> 1 -pchisq(104.08, df = 26)
[1] 2.665745e-11
```

**We get p-values of 3.210995e-10. We find that we can get rid of log(Elev) and log(DistNear). If we double area, it would ahve a multiplicative change of 0.966, holding all other terms constant.**

### Problem 3: El Nino and Hurricanse: ch. 22 exercise 21

Data is `ex1028`. In addition to answering the questions for this exercise, add the following:

- For both models (a) and (b), interpret the effect of the El Nino temperature on the response as it changes from cold to neutral and cold to warm and explain whether these effects are significant. Be careful to use the `ElNino` variable in data set `ex1028` rather than `Temperature`.
- You should also recode the `WestAfrica` variable (0 = dry and 1 = wet) to make it a factor variable with wet/dry levels.

```
> ex1028 <- ex1028
> ex1028$WestAfrica <- fct_recode(factor(ex1028$WestAfrica),
+                                 "wet" = "1",
+                                 "dry" = "0")
> #a
>
> ex1028_glm<- glm(Storms ~ ElNino, family = quasipoisson(), data = ex1028)
> summary(ex1028_glm)

Call:
glm(formula = Storms ~ ElNino, family = quasipoisson(), data = ex1028)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.73385  -0.63641  -0.03755   0.33129   2.27212

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.40919    0.06847  35.184  < 2e-16 ***
ElNinoneutral -0.11288    0.09969  -1.132  0.26350
ElNinowarm    -0.44559    0.10959  -4.066  0.00019 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 0.8346076)

    Null deviance: 50.875  on 47  degrees of freedom
Residual deviance: 35.990  on 45  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
> 1 -pchisq(35.990, df = 47)
[1] 0.8787009
>
> #b
> ex1028_glm<- glm(Hurricanes ~ ElNino + WestAfrica, family = quasipoisson(), data = ex1028)
> summary(ex1028_glm)
```

```
Call:
glm(formula = Hurricanes ~ ElNino + WestAfrica, family = quasipoisson(),
    data = ex1028)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.312  -0.500  -0.274   0.480   1.859

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.80344    0.10423  17.303  < 2e-16 ***
ElNinoneutral -0.04463    0.11584  -0.385  0.70189
ElNinowarm    -0.46206    0.13511  -3.420  0.00136 **
WestAfricawet  0.21913    0.10529   2.081  0.04327 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 0.6513443)

    Null deviance: 44.414  on 47  degrees of freedom
Residual deviance: 27.322  on 44  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
> 1 -pchisq(27.322, df = 44)
[1] 0.9770889
```

Looking at the model coefficients, for (a) we see that going from cold to neutral will result in a multiplicative change of0.8932578. And going from cold to warm will result in a multiplicative change of 0.6404463. For the second model we get cold to neutral and cold to warm as 0.9563513 and 0.6299845 respectively. Since we get a high p-value, we now that the model is adequate However, we can see that cold to neutral is not significant, but cold to warm is significant.