STAT511HW5

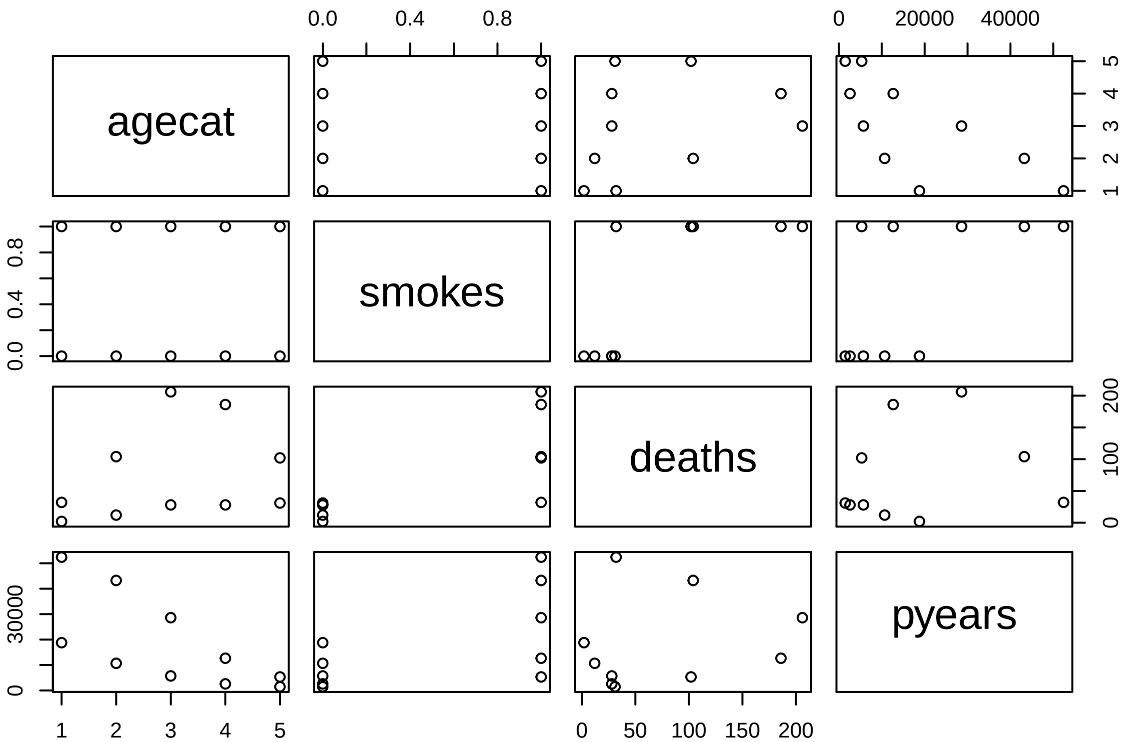
Ben Straub

October 27, 2015

SMOKING

Exploratory Data Analysis

```
'data.frame':   10 obs. of  4 variables:
 $ agecat: int  1 2 3 4 5 1 2 3 4 5
 $ smokes: int  1 1 1 1 1 0 0 0 0 0
 $ deaths: int  32 104 206 186 102 2 12 28 28 31
 $ pyears: int  52407 43248 28612 12663 5317 18790 10673 5710 2585 1462
```



OBSERVATIONS:

- It looks like there are two relationships going on between agecategories and person years
- Several of the variables have relationships with each other that look to have two different categories.
- Inspection of the observations I see that the age categories coded as smokers have a higher amount of deaths than the age categories who do not smoke.
- This makes sense as scientifically you would expect to find people who do not smoke to have a different life trend then people who do smoke.
- I will attempt to fit a model that has deaths as the response variable and agecats, smokes and pyears as predictor variables. I will report one Linear Model, but I did investigate several variations of the Linear Model. I will also code Agecats and smokes as factor variables.

- I learned later that coding them as factors was a bad move, especially as age categories has an ordering property to it, which is very important!

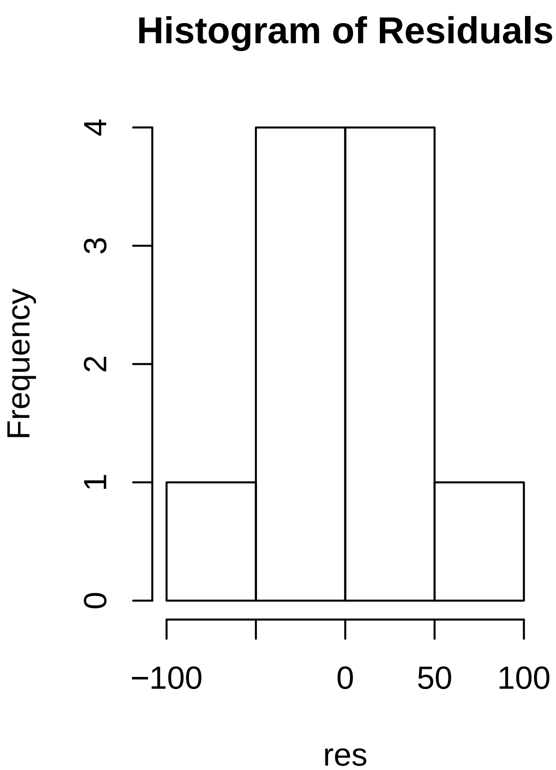
```
fit = lm(deaths~factor(agecat)+factor(smokes)+pyears, data=smoking)
# summary(fit)
fit = lm(deaths~factor(agecat)+smokes+pyears, data=smoking)
# summary(fit)
fit = lm(deaths~agecat+factor(smokes)+pyears, data=smoking)
kable(summary(fit)$coeff, digits = 5)
```

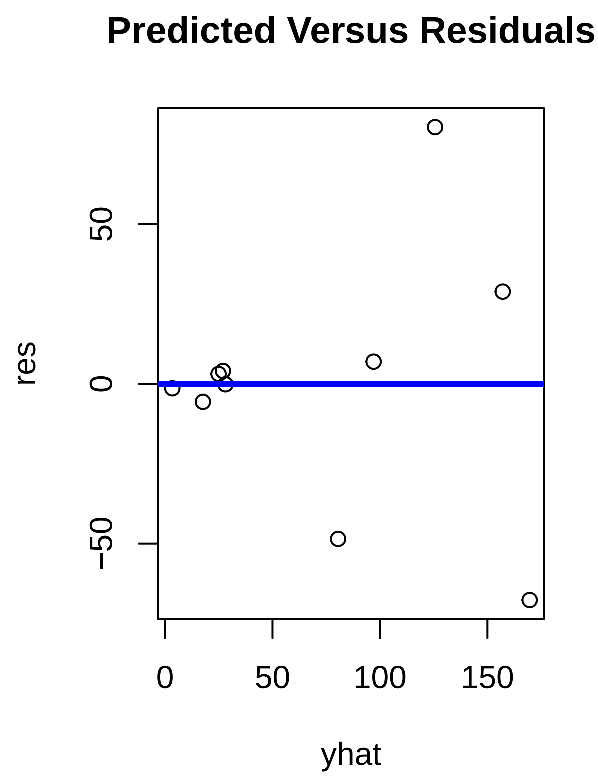
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	48.44218	92.05640	0.52622	0.61761
agecat	-3.65328	23.79286	-0.15355	0.88300
factor(smokes)1	151.19901	60.48944	2.49959	0.04655
pyears	-0.00220	0.00252	-0.87313	0.41616

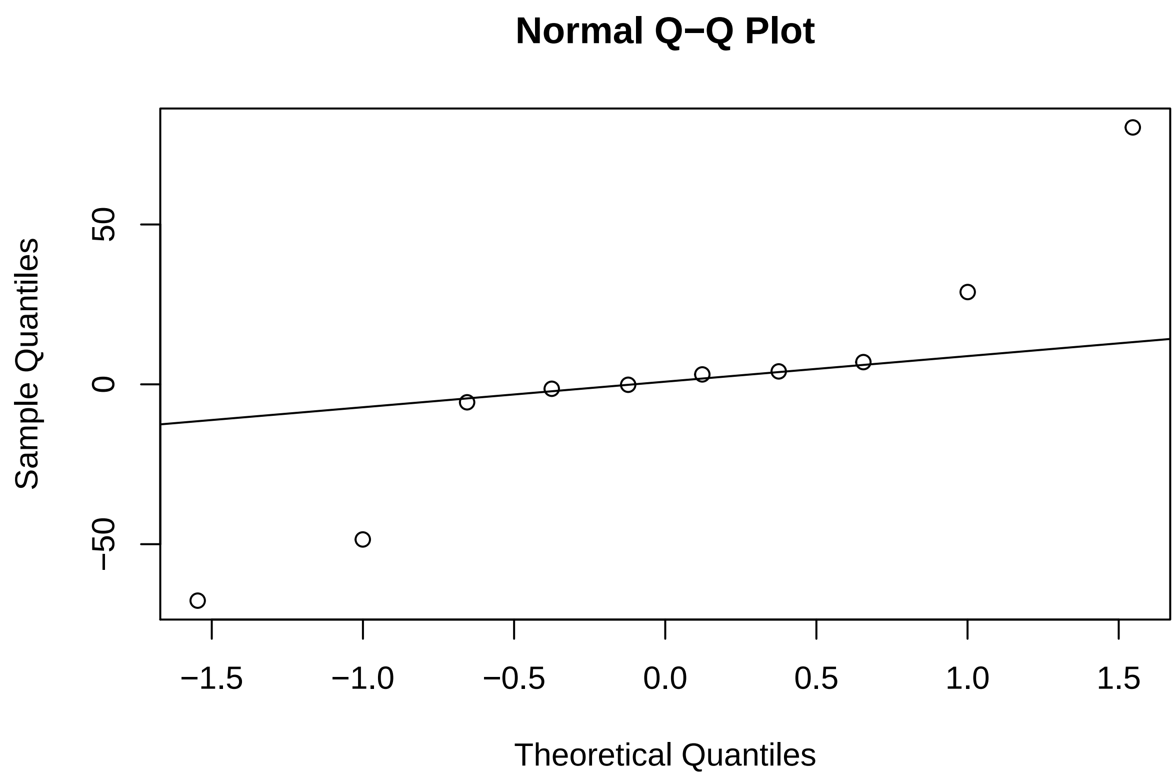
I tried to fit 3 Linear Models and found the third model to have the most variables with significant values, i.e. one variable smokes.

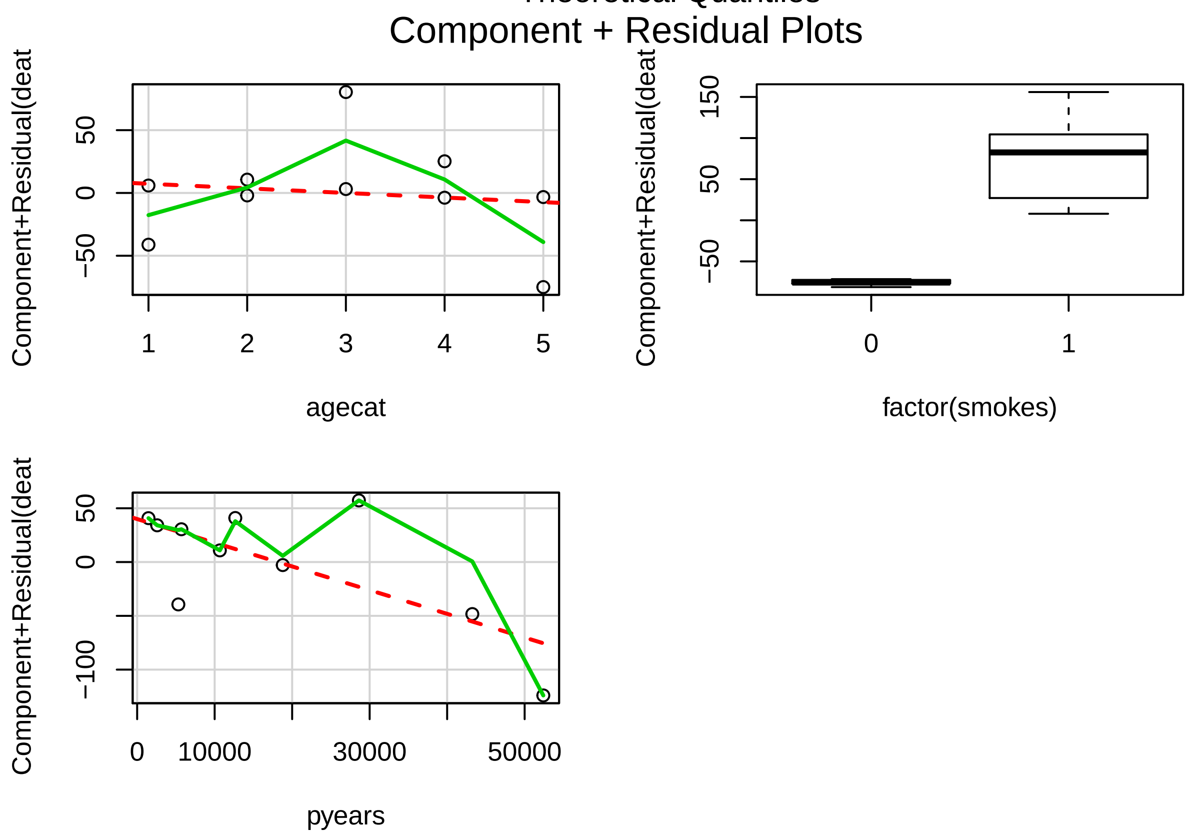
$$ModelEquation:$$

Residuals Diagnostics of Linear Model






Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−50
0
50
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5


Component + Residual Plots
Component+Residual(deat
agecat
factor(smokes)
pyears
−50
0
50
150
1
2
3
4
5
0
1
−100
10000
30000
50000

CONCLUSION: **The Linear Model is not a good fit for the Smoking Data Set. We do see a nice normally distributed histogram of residuals, but the residuals plotted against the yhats shows heteroscadacity and the QQ-plot violates our normal modeling assumptions with its tails.. Also, the Residulas plotted against the Yhats shows that there is a clustering of data, which is an incidation that something else is going on in the Data. The component Residuals Plots also show us a non-linear relationship between the variables. Therefore, we can definitively conclude that the Linear Model is a poor fit for the Smoking Data. Onward to GLMs!**

Generalized Linear Model for the Smoking Data Set.

$$ModelEquation:$$

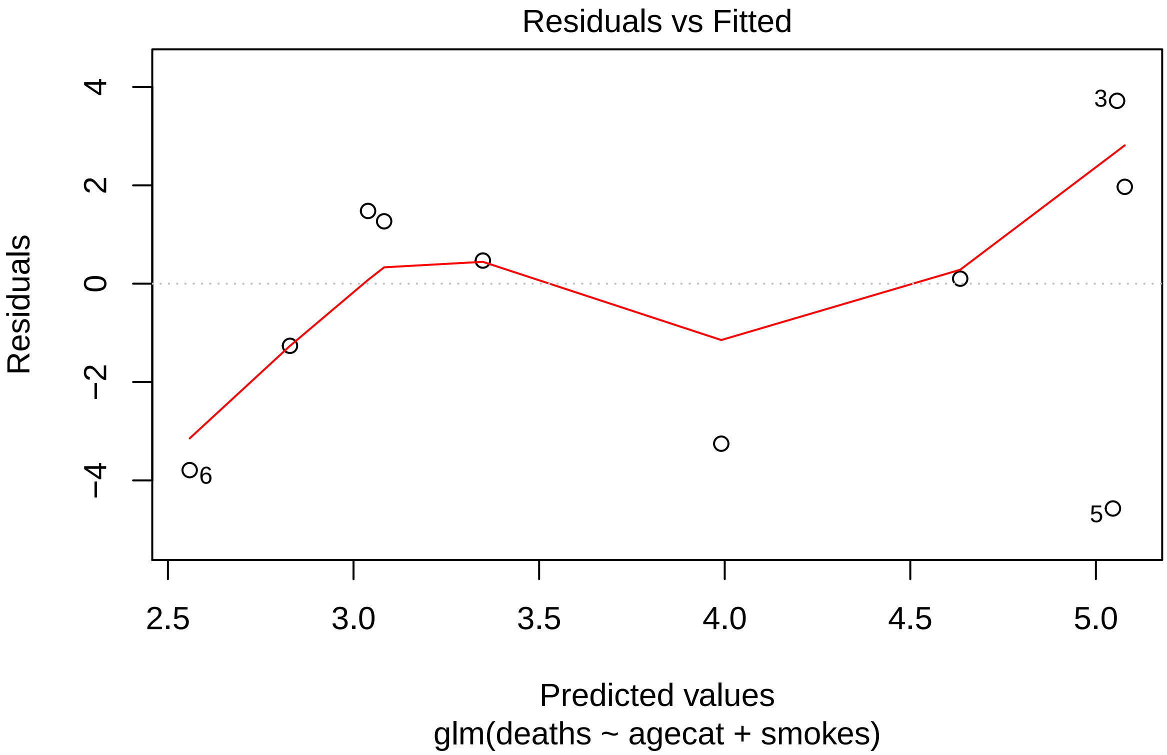
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-8.11833	0.13929	-58.28206	0.00000
agecat	0.83583	0.02904	28.77722	0.00000
smokes	0.40637	0.10720	3.79093	0.00015

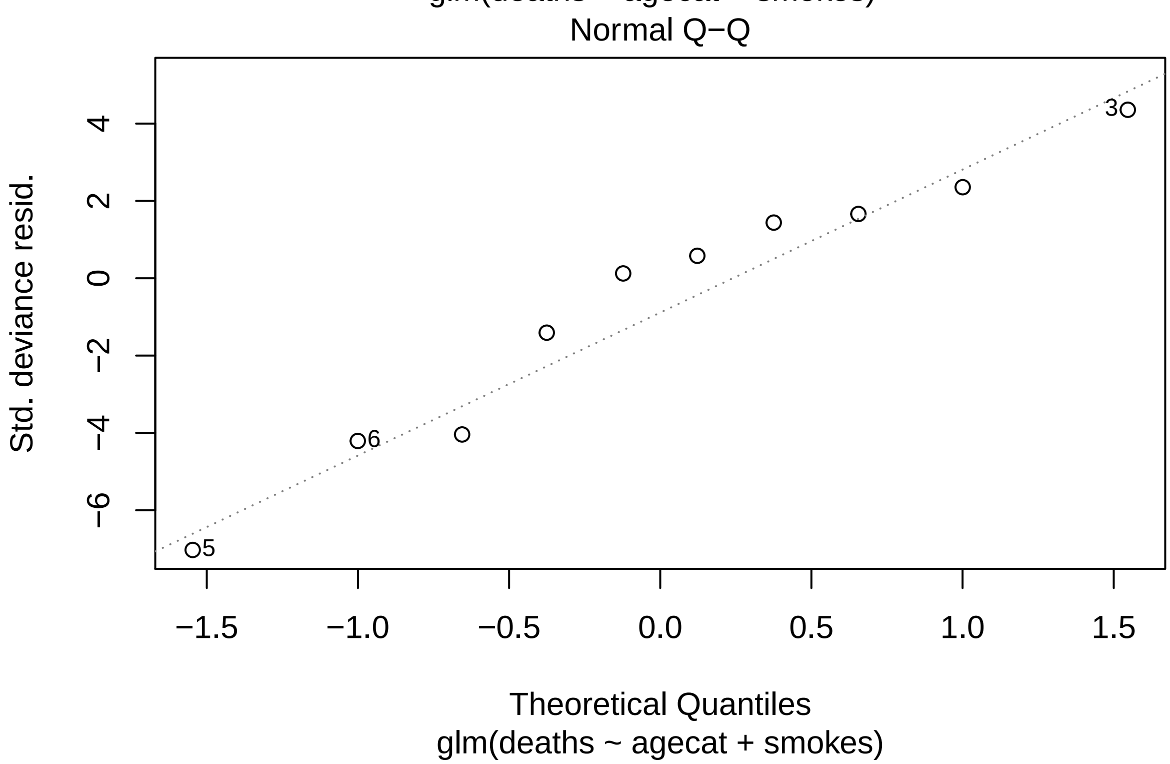
OBSERVATIONS:

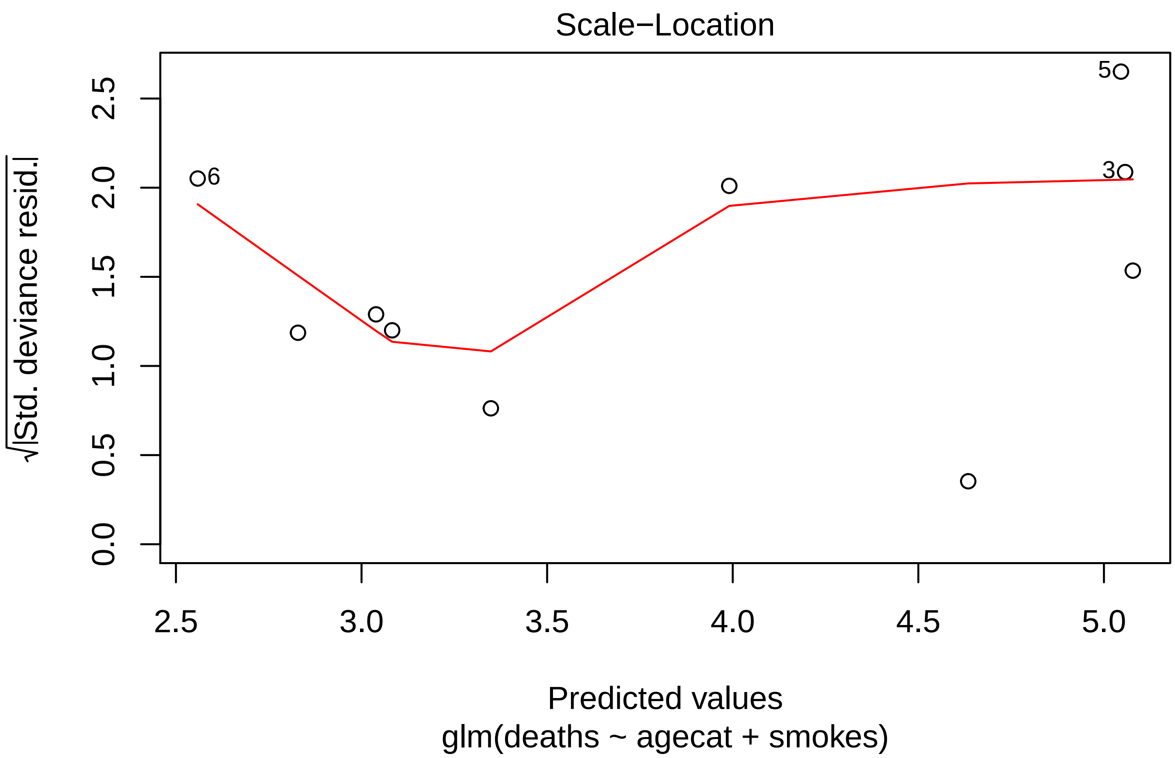
- The new Generalized Linear Model is an excellent fit for the Data. The model's summary of z-scores shows that the variables agecats and smokes are good predictors for the response variable death.

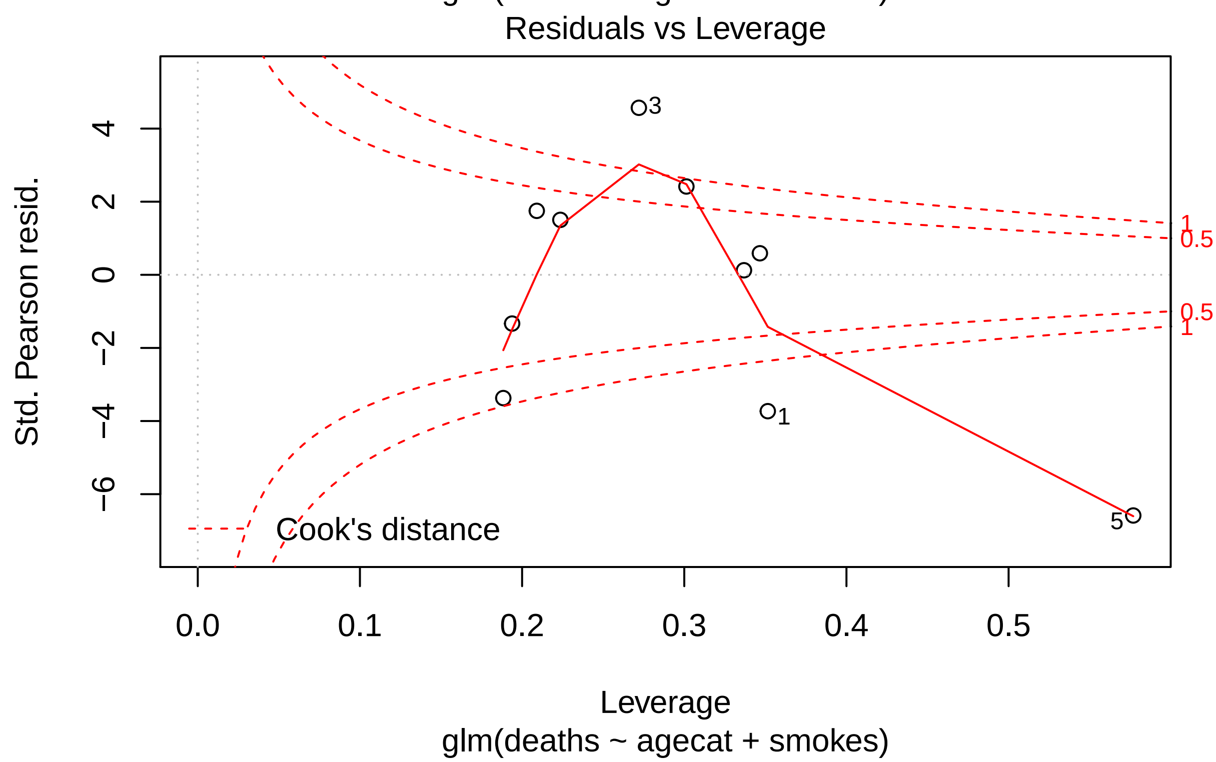
- We can now interpret the model. The GLM shows us that as a person ages and smokes a multiplcative effect occurs in the model on their chance of death. A person who does not smoke still has a multiplicative effect, but to a lesser degree as smoking does not increase their chance of dying in the model. Sadly, the older you get the chances of you dying increases whether you smoke or don't smoke… but smoking increases that rate!

Residual Diagnostics for GLM of Smoking Data

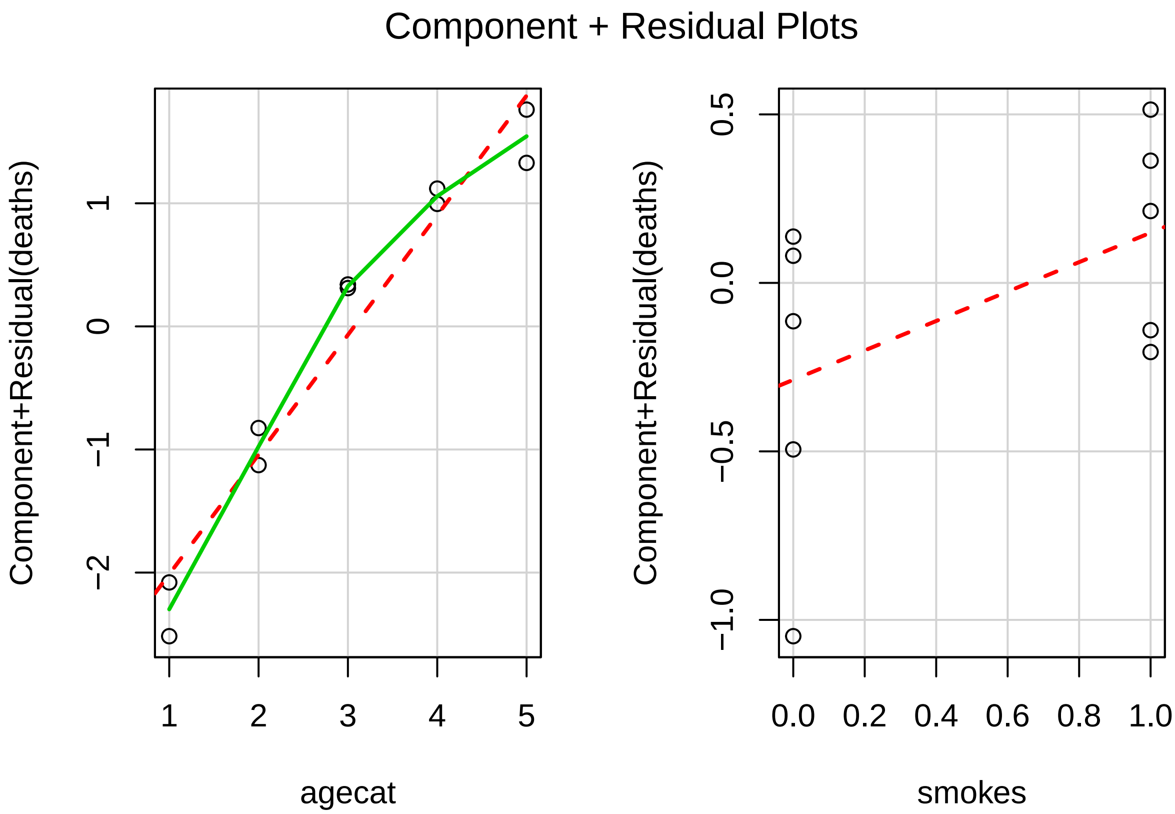






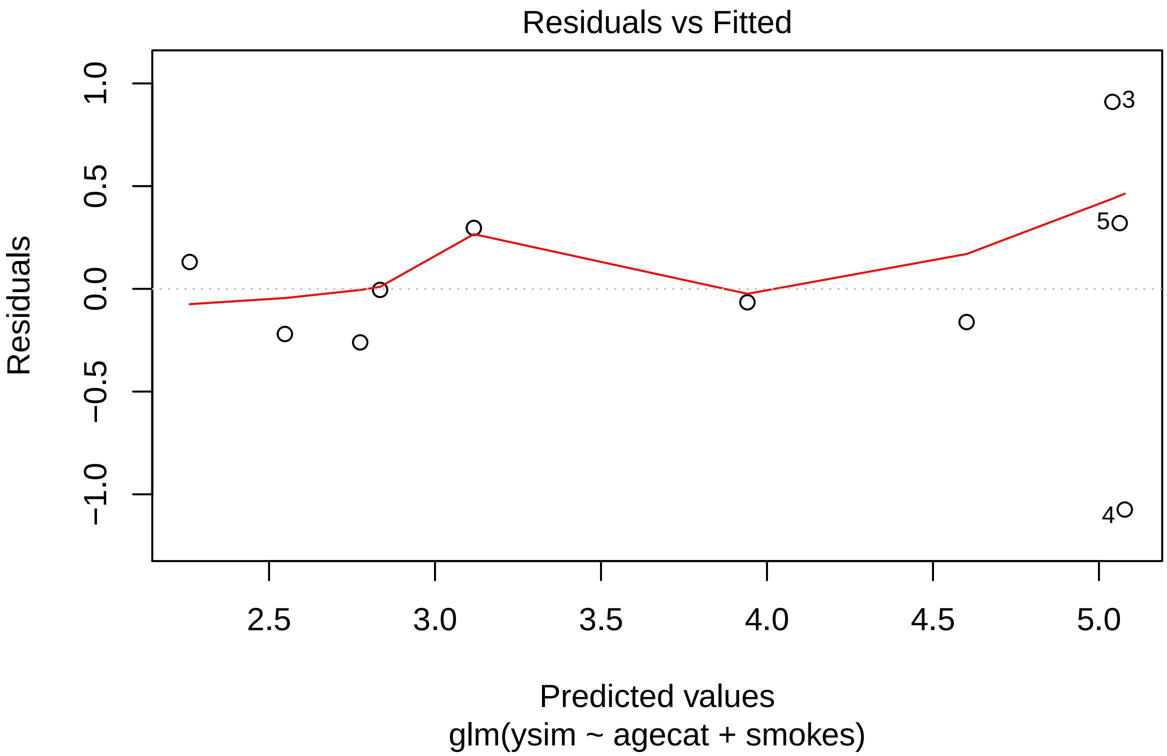


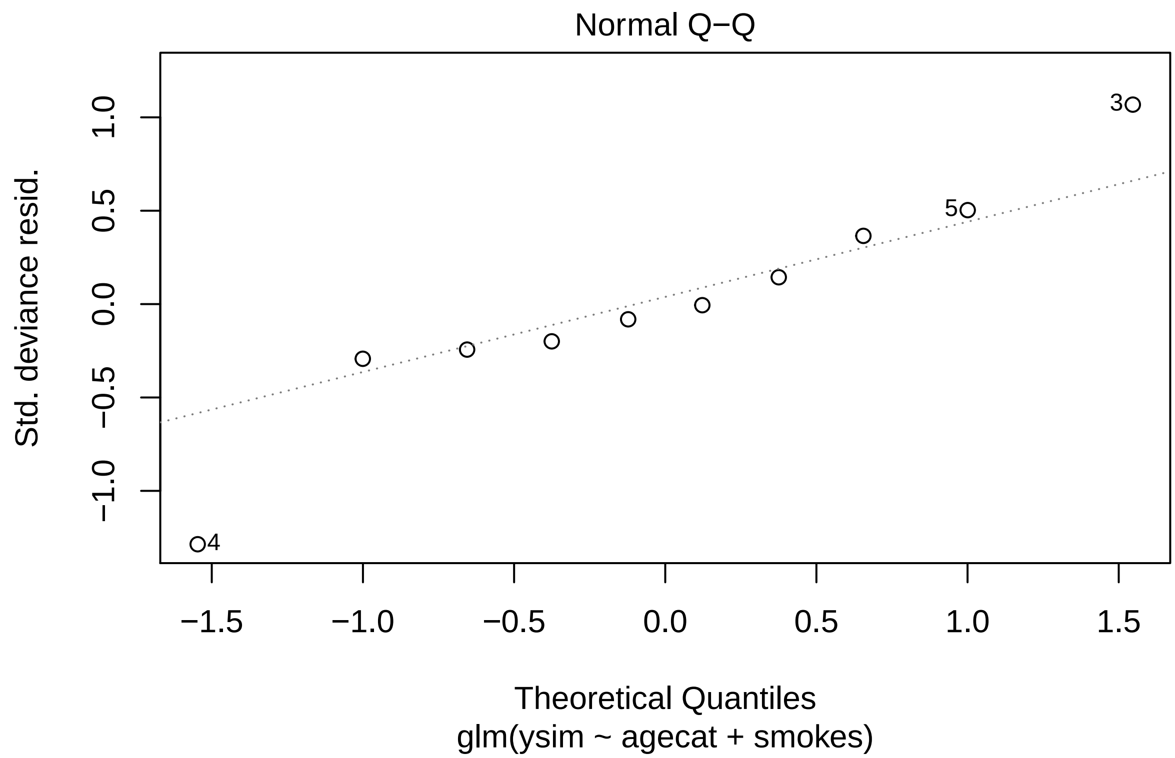

```
## Warning in smoother(.x, partial.res[, var], col = col.lines[2], log.x =
## FALSE, : could not fit smooth
```


Component + Residual Plots
Component+Residual(deaths)
agecat
1
2
3
4
5
−2
−1
0
1
Component+Residual(deaths)
smokes
0.0
0.2
0.4
0.6
0.8
1.0
−1.0
−0.5
0.0
0.5

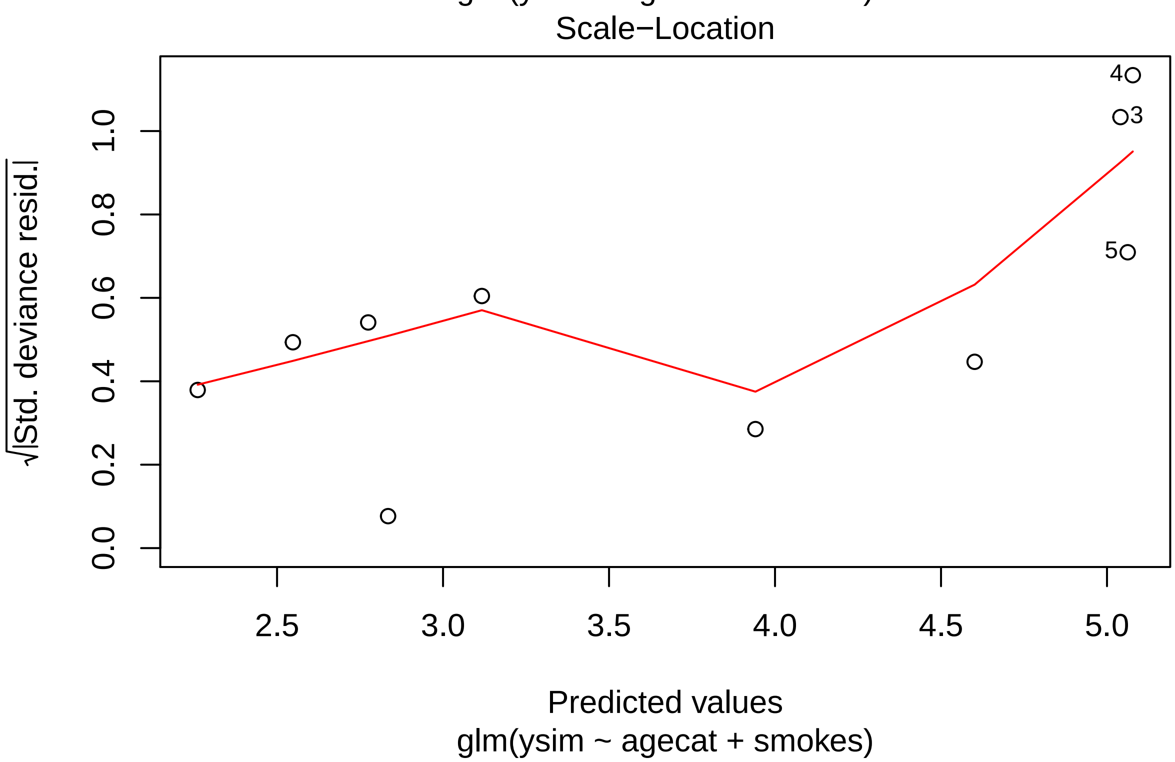
OBSERVATIONS: **The QQ-Plot looks much better for the GLM then my previous LM and the graph of the residuals vs fitted does not show heteroscadacity nor does it show the clustering of data seen in the previous LM. The Component Residuals of the Data show a linear trend as well between the variables and residuals. We will compare a simulation of data versus our smoking data set to see if we can gain any more insight into the current GLM model.**

Model Simulation

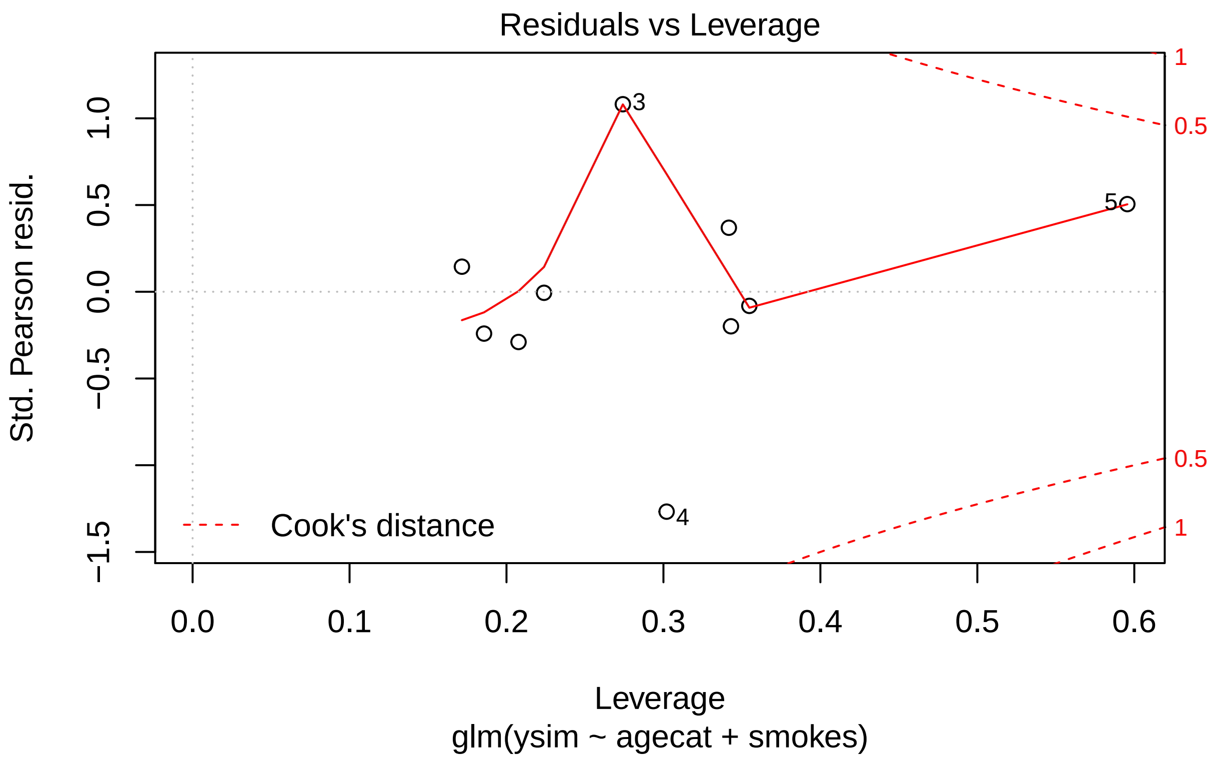




Normal Q−Q
Std. deviance resid.
Theoretical Quantiles
glm(ysim ~ agecat + smokes)
−1.0
−0.5
0.0
0.5
1.0
−1.5
1.5
3
5
4



Scale−Location
√|Std. deviance resid.|
Predicted values
glm(ysim ~ agecat + smokes)
0.0
0.2
0.4
0.6
0.8
1.0
2.5
3.0
3.5
4.0
4.5
5.0
4
3
5



OBSERVATIONS: **The Simulated Model's Residuals are comparable to the GLM that I created for the smoking data.**

Confidence Intervals for Generalized Linear Model of Smoking Data

```
  pyears agecat smokes
1   1000      3      1

   CI.down    CI.up
1 1.621751 1.784844

   CI.down    CI.up
1 5.061945 5.958652

##   pyears agecat smokes
## 1   1000      3      0

##    CI.down    CI.up
## 1 1.100851 1.493004

##    CI.down    CI.up
## 1 3.006724 4.450444
```

OBSERVATIONS: **Upon examination of the data and the confidence intervals I calculated for my model I noticed something strange. The confidence intervals for death is 5.06 to 5.96, which seems off. Taking the total pyears for a person who smokes in agecat 3 and dividing it by 1000 we get 28.612. Taking the corresponding total deaths, 206 and dividing it by 28.612 we get 7.19, which should lie in our confidence interval, but sadly 7.19 is not in our confidence interval. The same logic applies for the confidence interval for Non-Smokers in agecat 3. Something is wrong with the model as these values should be inside our confidence intervals. Perhaps, using a Polynomial would have made the model a better fit and procduced better confidence intervals.**

ISLAND SCRUB JAY

```
'data.frame':   5625 obs. of  6 variables:
 $ birds : num  0 0 0 0 0 0 0 0 0 0 ...
 $ x     : num  234870 237083 235732 237605 234239 ...
 $ y     : num  3767154 3766804 3766717 3766719 3766570 ...
 $ elev  : num  151 562 407 563 440 582 586 285 795 671 ...
 $ forest: num  0.02 0 0 0.26 0.01 0.1 0.12 0 0.17 0 ...
 $ chap  : num  0.29 0.49 0.72 0.25 0.01 0.48 0.57 0.03 0.12 0 ...
```

Exploratory Data Analysis

- The Island Scrub Jay Data Set has 5,265 observations and 6 Variables
- Variables: isj, x, y, elev, forest, chap
- I renamed the Variable isj as birds.
- The birds Variable is a coded as 0 for abscence of birds and 1 for presence of birds.
- There appears to be a lot of NAs in this data set!!

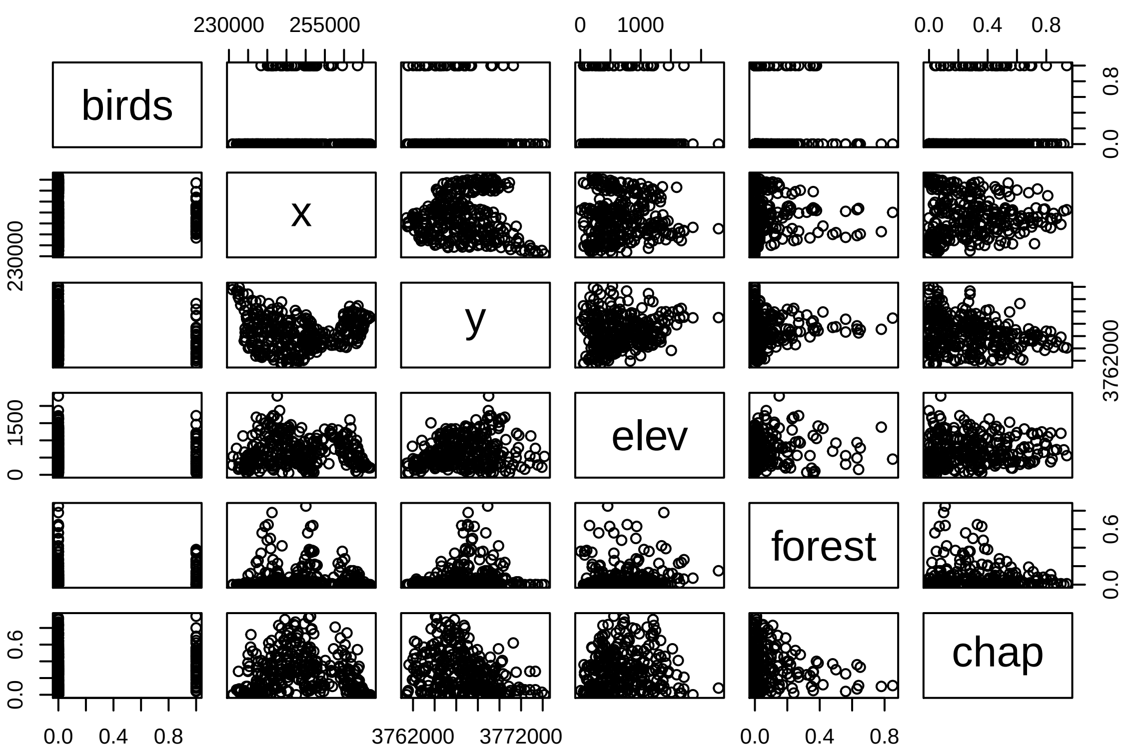
```
##   birds        x       y elev forest chap
## 1     0 234870.1 3767154  151   0.02 0.29
## 2     0 237083.0 3766804  562   0.00 0.49
## 3     0 235732.0 3766717  407   0.00 0.72
## 4     0 237605.0 3766719  563   0.26 0.25
## 5     0 234239.1 3766570  440   0.01 0.01
## 6     0 235005.1 3766420  582   0.10 0.48
```

```
##      birds        x       y elev forest chap
## 5616    NA 264336.7 3761124   NA     NA   NA
## 5617    NA 264636.7 3761124   NA     NA   NA
## 5618    NA 264936.7 3761124   NA     NA   NA
## 5619    NA 265236.7 3761124   NA     NA   NA
## 5620    NA 265536.7 3761124   NA     NA   NA
## 5621    NA 265836.7 3761124   NA     NA   NA
## 5622    NA 266136.7 3761124   NA     NA   NA
## 5623    NA 266436.7 3761124   NA     NA   NA
## 5624    NA 266736.7 3761124   NA     NA   NA
## 5625    NA 267036.7 3761124   NA     NA   NA
```

birds	x	y	elev	forest	chap
Min. :0.000	Min. :229837	Min. :3761124	Min. : 0.0	Min. :0.000	Min. :0.0000
1st Qu.:0.000	1st Qu.:239137	1st Qu.:3764424	1st Qu.: 375.5	1st Qu.:0.000	1st Qu.:0.0600
Median :0.000	Median :248437	Median :3767724	Median : 655.0	Median :0.000	Median :0.2000
Mean :0.124	Mean :248437	Mean :3767725	Mean : 717.8	Mean :0.064	Mean :0.2469
3rd Qu.:0.000	3rd Qu.:257737	3rd Qu.:3771024	3rd Qu.:1004.5	3rd Qu.:0.060	3rd Qu.:0.3900
Max. :1.000	Max. :267037	Max. :3774324	Max. :2289.0	Max. :1.000	Max. :0.9800
NA's :5318	NA	NA	NA's :2838	NA's :2838	NA's :2838

```
## Warning in if (drop) {: the condition has length > 1 and only the first
## element will be used
```

- Upon further examination of the data set I found that there was only 303 complete cases of Data, i.e. 303 cases with elevation, chap, forest, x, y and birds filled out.
- The Data has 38 entries for the presences of birds.
- The Data has 265 entries for the absence of birds.
- I also noticed that there is a relationship between the data entries with the NAs.
- It looks like the island data has entries for forest, elevation and chapral and the water around the island has NAs for forest, elevation and chapral in it. Makes sense.



- I thought the pairs plot would be useful. It seems like some of the data in the pairs plots have strange peaks in it and some have negative linear trends in it. I'm unsure if this is a useful graph.

Simple Linear Model Fit

- I did a simple linear model to just gain some intuition of the data set.

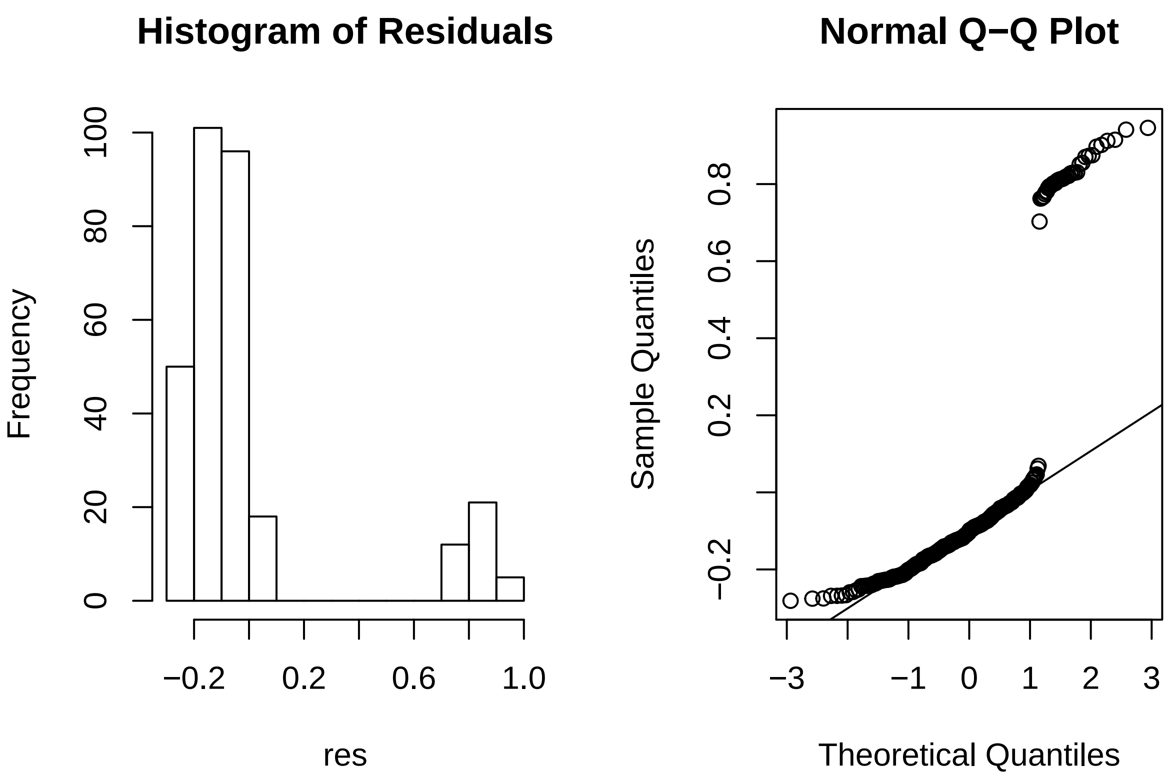
- I included all variables in the data set as well the entire data set.

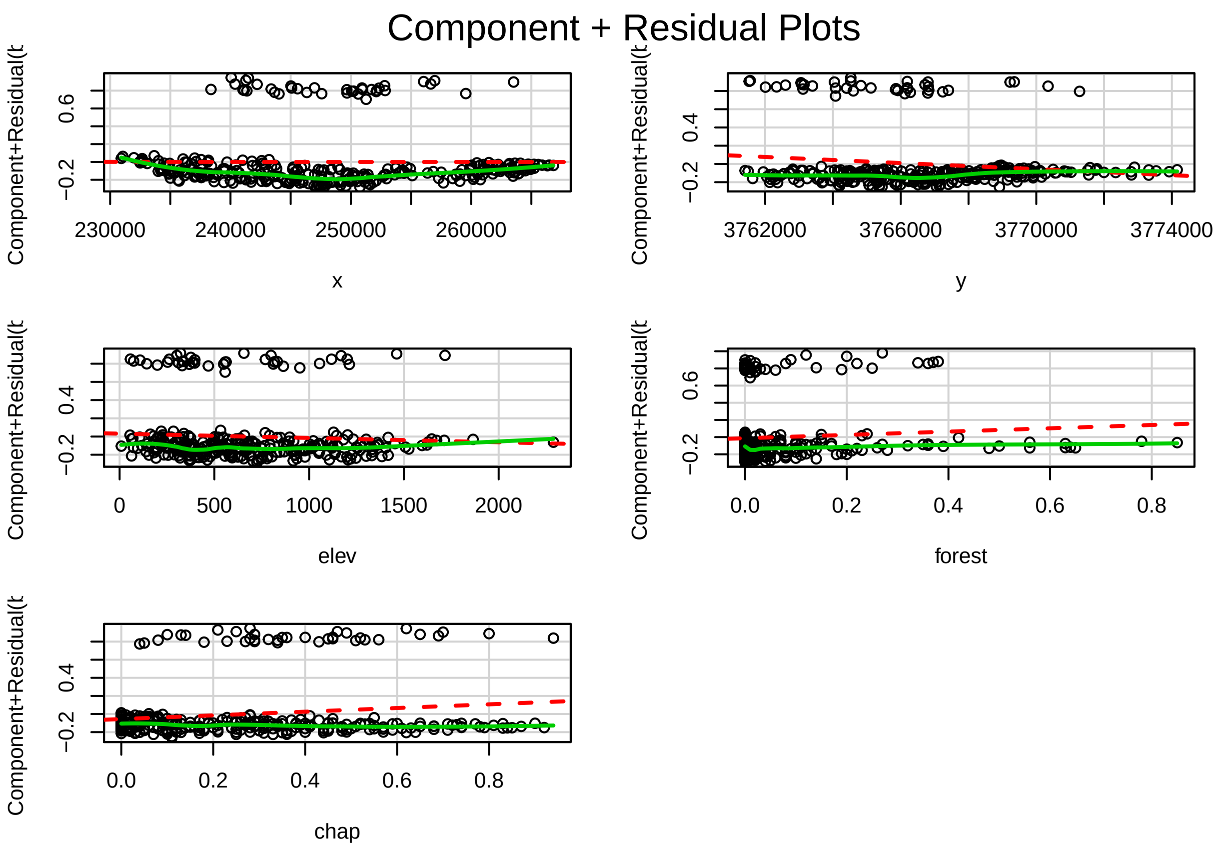
$$ModelEquation:$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	62.74130	28.42933	2.20692	0.02808
x	0.00000	0.00000	-0.01920	0.98470
y	-0.00002	0.00001	-2.20524	0.02820
elev	-0.00005	0.00005	-1.00379	0.31629
forest	0.19257	0.14007	1.37477	0.17024
chap	0.20322	0.08473	2.39843	0.01708

- The Linear Model gives us two variables of significance y and chap.
- The Scrub Jay likes the Chaprel bush, especially during mating season, but this current linear model does not seem to be a good fit. It would also be hard to make accurate predictions with such a weak model.

Residual Diagnostics of Linear Model





- The histogram of the residuals are bimodal and do not demonstrate normailty. However, it does demonstrate there are two seperate distribution occurring in the data.
- The QQ-Plot also shows us two distributions.
- I noticed that in the component residuals plots of the varibles that there is a lot of clustering of data points. . . perhaps the absence and presence of birds!!
- A General Linear Model would be a better fit than the Linear Model we have been using.

General Linear Model for Island Scrub Jay Date Set

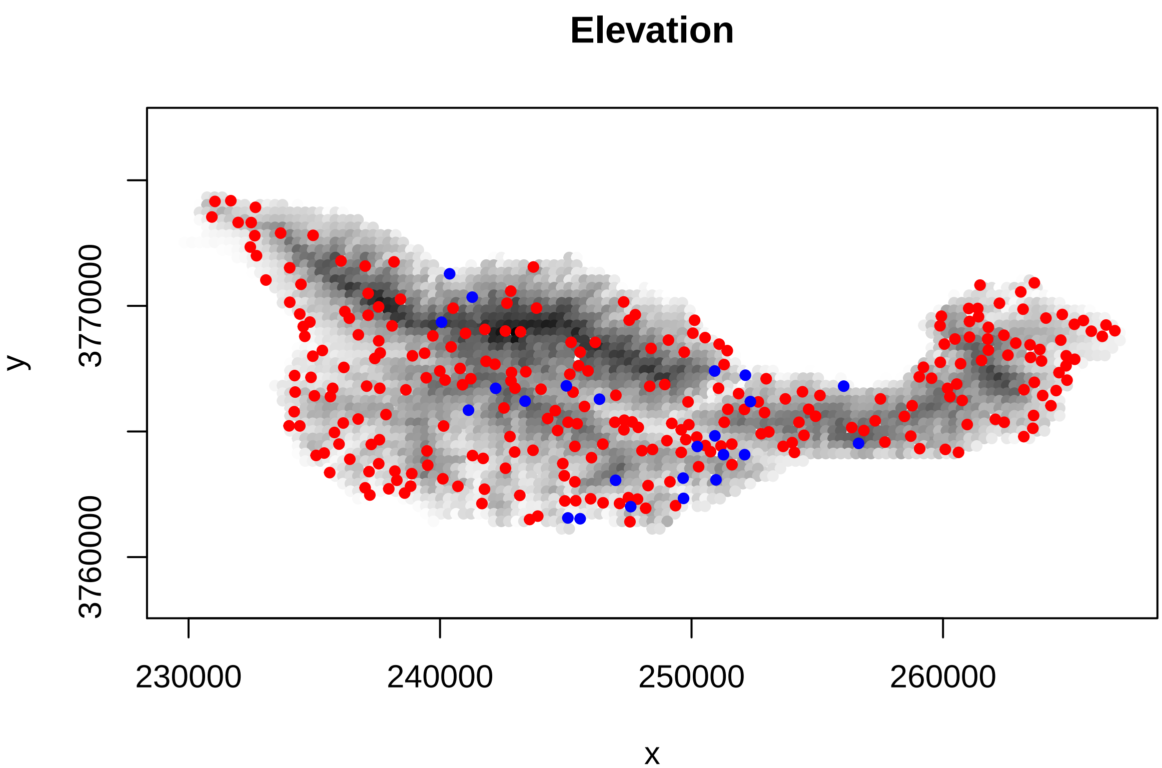
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	820.19836	343.84351	2.38538	0.01706
x	0.00001	0.00002	0.34245	0.73201
y	-0.00022	0.00009	-2.39182	0.01677
elev	-0.00043	0.00053	-0.80475	0.42096
forest	2.20219	1.20557	1.82668	0.06775
chap	1.83883	0.75349	2.44041	0.01467

OBSERVATIONS: **The chap, forest and y variables are significant from this current general linear model. The Results I got from the Residuals of my Simple Linear Model showed some clusering of the data set. I decided to include an interaction effect between x and y to see if this changed the model variables significance.**

General Linear Model with Interaction Term

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-22041.27447	9540.12621	-2.31038	0.02087
x	0.09361	0.03935	2.37850	0.01738
y	0.00585	0.00253	2.30990	0.02089
elev	-0.00049	0.00054	-0.90930	0.36319
forest	2.61434	1.21990	2.14308	0.03211
chap	1.69114	0.77111	2.19313	0.02830
x:y	0.00000	0.00000	-2.37826	0.01739

OBSERVATIONS: **The Model shows that chap, forest and y variables still being significant, but also has the x and xy varibles being significant. This makes sense as x and y are longitude and latitude coordinates for the island. The model with the Interaction Term is a marked improvment from the previous model. However, the map below has the island with blue dots representing observed sitings of scrub jay and red dots showing the absence of scrub jays. I noticed that some of the observed data points might have a polynomial trend to them. I'm unsure if I can go off a gut hunch or if I needed to prove it using prp plots??? On my hunch and some insight provided by Ben Lee, I decided to build another model with chaparrel as a polynomial term of degree 2 to see if anything of significane changes.**



General Liner Model with Interaction Term and Polynomial Term.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-23064.37505	10062.12423	-2.29220	0.02189
x	0.09741	0.04142	2.35176	0.01868
y	0.00612	0.00267	2.29182	0.02192
elev	-0.00053	0.00055	-0.96473	0.33468
forest	2.25519	1.27281	1.77182	0.07642
poly(chap, 2)1	9.71198	3.83497	2.53248	0.01133
poly(chap, 2)2	-7.33993	3.46899	-2.11587	0.03436
x:y	0.00000	0.00000	-2.35156	0.01869

OBSERVATIONS: **This new model has the same level of significane for each variable except that the forest variable has now become significant (although it is weak). I believe (from my limited knowledge) that this is the best model to make a predicted probablility map for the Island Scrub Jay Data Set. This map will show that the presence of chapparel, forest and depending on the xy coordinates will increase the likelihood of finding island scrub jay on the island.**

Predicted Probability of Island Scrub Jay

- The triangles indicate where sightings of scrub jay occurred.
- The darker the gray the more likely you will find an island scrub jay.