Assumptions for OLS

Laura Johnson

9/21/2020

Final Model: Mussel Abundance VS Drainage Basin Area

```
# Set working directory and bring in libraries
setwd("/Users/williamjohnson/Desktop/Laura/Hallett_Lab/Repositories/thesis-mussels/site_DATAexplore")
library(tidyverse)
```

```
## Warning: As of rlang 0.4.0, dplyr must be at least version 0.8.0.
## * dplyr 0.7.7 is too old for rlang 0.4.7.
## * Please update dplyr with `install.packages("dplyr")` and restart R.
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.7
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0
```

```
## -- Conflicts ------------------------------------------------------------------------------------ t
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
#Bring in needed files:
finalmodel <- as.tibble(read.csv("FinalModel.csv"), colnames = TRUE) #finalmodel spreadsheet has all ca
#model w/o zinc cr
finalmodel <- finalmodel%>%
  filter(!site_id == "ZINCCMP")

#Bring in site distance / drainage basin data
distarea <- as.tibble(read.csv("dist_area_final.csv"), colnames = TRUE)
# Need to link drainage basin areas with site id
drainmodel <- finalmodel %>%
  inner_join(distarea, by = "obs_id")
```

```
## Warning: `chr_along()` is deprecated as of rlang 0.2.0.
## This warning is displayed once per session.
```

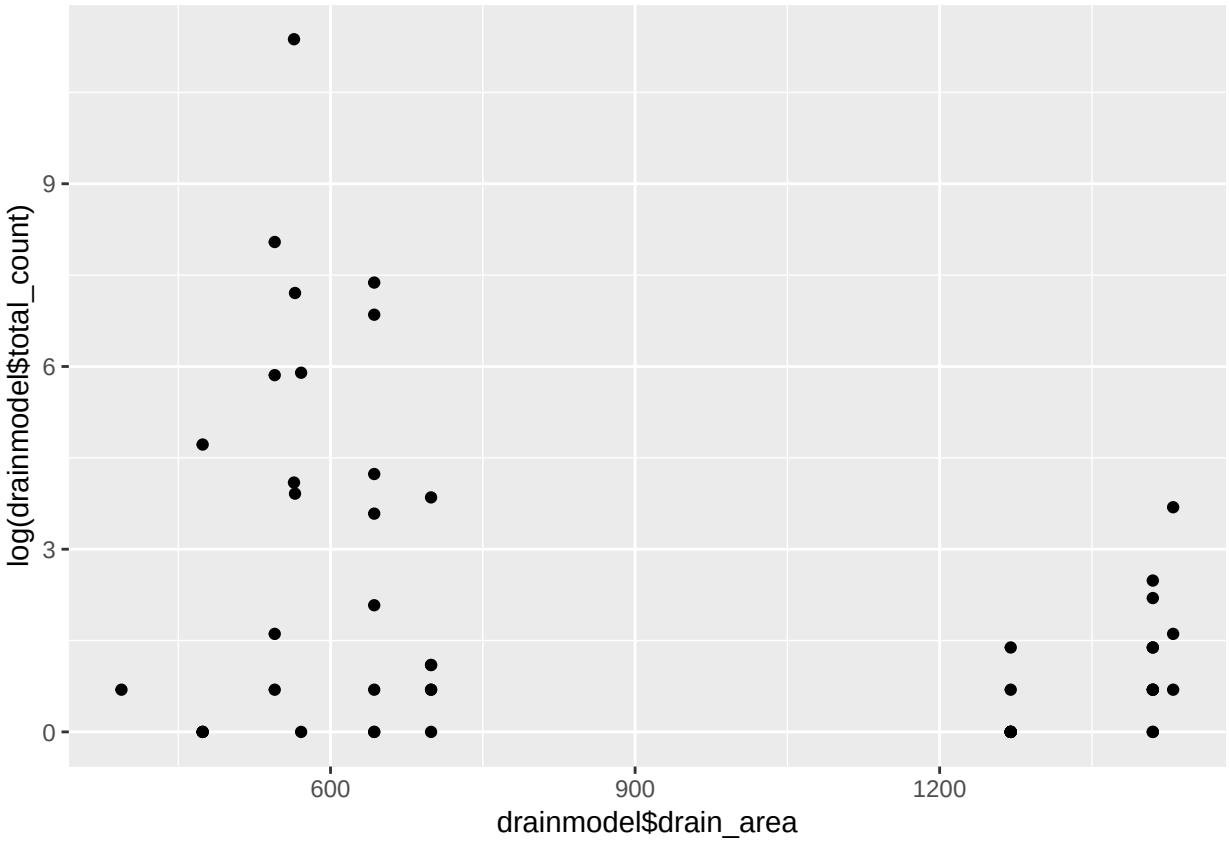
```
## Warning: Column `obs_id` joining factors with different levels, coercing to
## character vector
```

```
#Simple linear regression predicting mussel abundance by drainage basin area of aggregation
drain_model <- lm(log(drainmodel$total_count + .01) ~ drainmodel$drain_area)
summary(drain_model)
```

```
##
## Call:
## lm(formula = log(drainmodel$total_count + 0.01) ~ drainmodel$drain_area)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2266 -1.5550 -0.4878  1.1934  8.3687
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             4.4576441  0.9165864   4.863 1.33e-05 ***
## drainmodel$drain_area -0.0025761  0.0009325  -2.763  0.00816 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.502 on 47 degrees of freedom
## Multiple R-squared:  0.1397, Adjusted R-squared:  0.1214
## F-statistic: 7.632 on 1 and 47 DF,  p-value: 0.00816
```

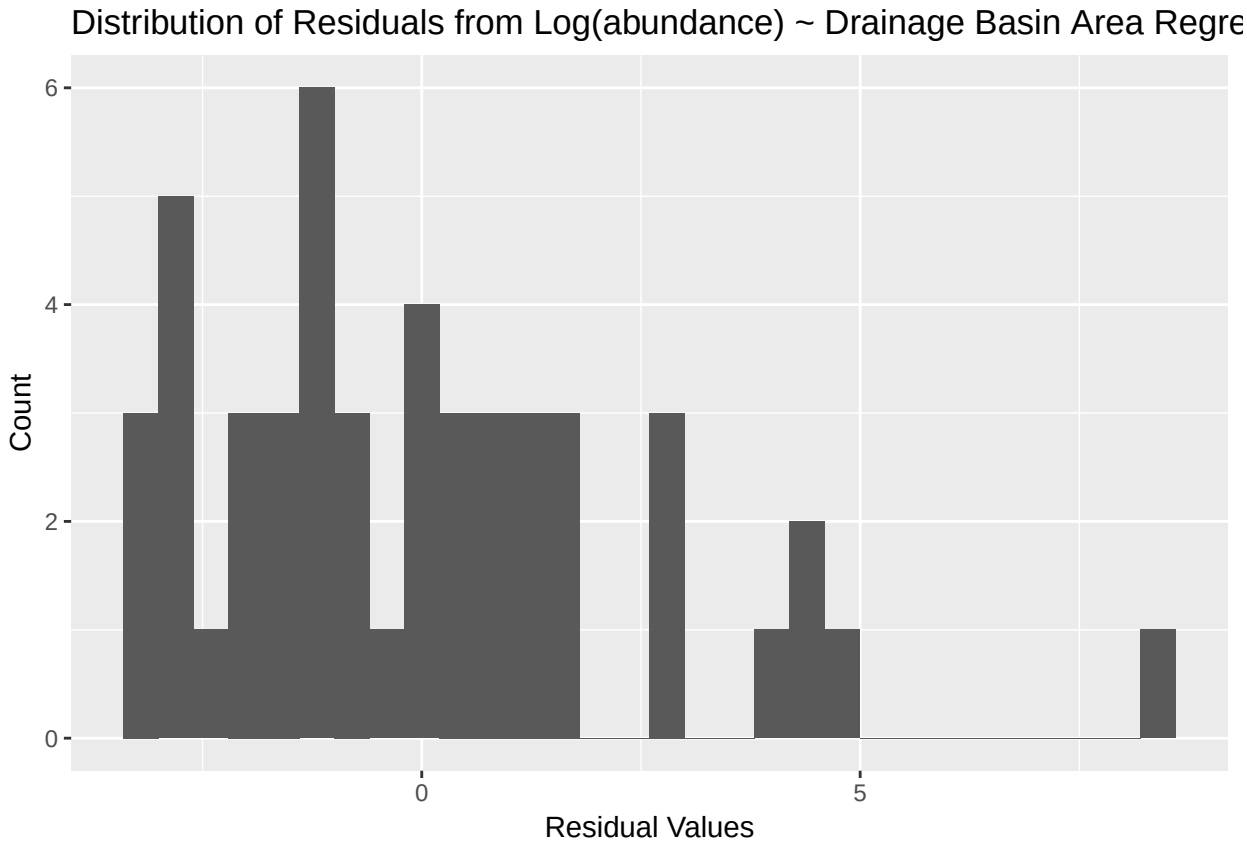
Assess Assumptions of Normality for use of OLS Regression for Abundance ~ Drainage Basin Area

```
#Assumption of linearity
ggplot(drainmodel) + geom_point(aes(x = drainmodel$drain_area, y = log(drainmodel$total_count)))
```



```
#Examine histogram of the residual values to determine if they are normally distributed
ggplot(drain_model) + geom_histogram(aes(drain_model$residuals)) + labs(x = "Residual Values", y = "Cou
  ggtitle("Distribution of Residuals from Log(abundance) ~ Drainage Basin Area Regression")
```

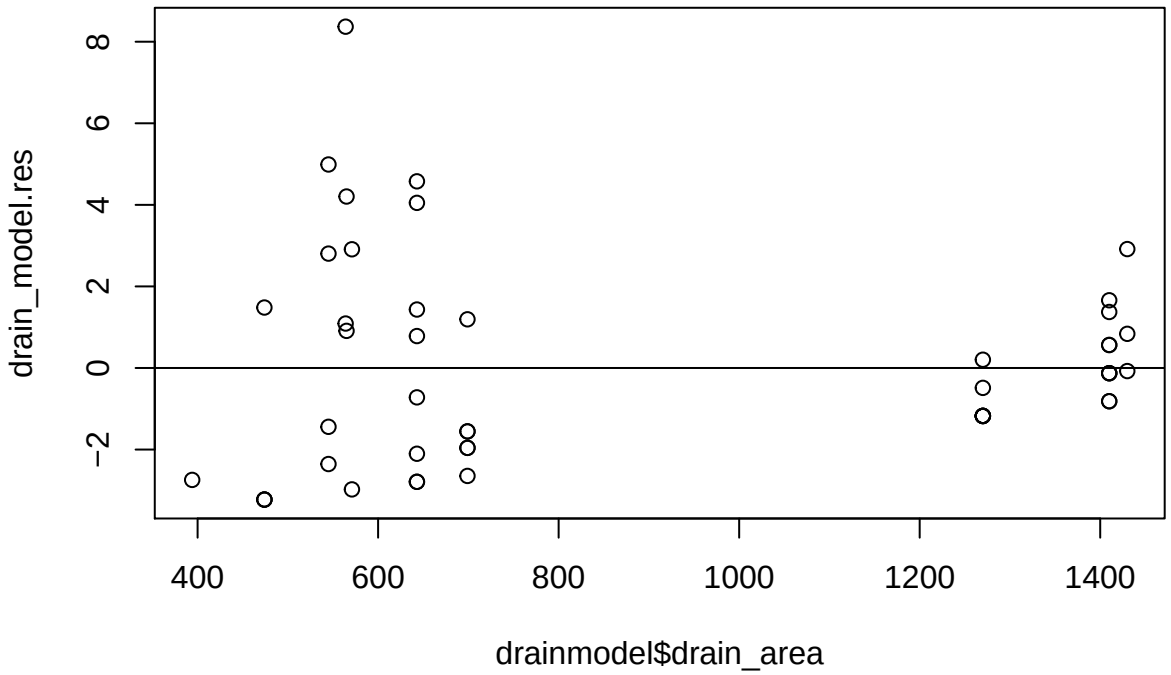
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
# Determine mean of residual values
print(mean(drain_model$residuals)) # YES the mean of residuals = 0!!!
```

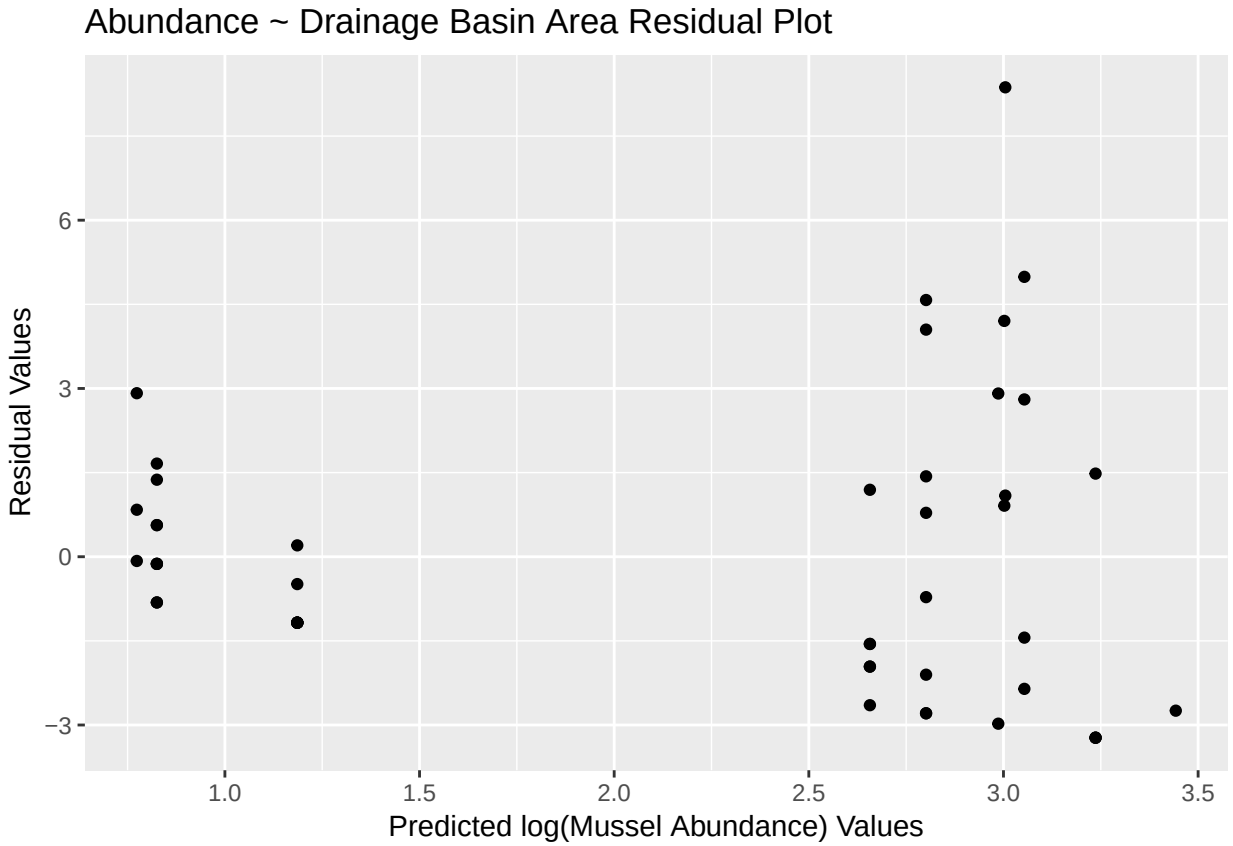
```
## [1] 4.411154e-17
```

```
# Assess whether residuals meet assumption of homoscedasticity
drain_model.res <- resid(drain_model)
plot(drainmodel$drain_area, drain_model.res)
abline(0,0) #Looks like greater variance at lower drainage basin areas than larger drainage basin areas
```




```
             # the sites where there were lots of low abundance aggregations but also high abundances,
            # abundances in the larger drainage basin area sites were much more similar to each other (
            # aggregation abundance)

# Residual Plot
ggplot(drain_model) + geom_point(aes(x = drain_model$fitted.values, y = drain_model$residuals)) +
  labs(x = "Predicted log(Mussel Abundance) Values", y = "Residual Values") + ggtitle("Abundance ~ Drai
```



Results:

1. Assumption of linearity:

Generally met, although it would be helpful to increase the sample size of sites with drainage basins b/w 750 - 1200 square miles... This is unfortunately not possible at this point!

2. Assumption of Zero Mean of Residuals:

MET!!!

3. Assumption of normality of residual error:

Not met- distribution does not look normal to me

4. Assumption of homoscedasticity:

Not met... there is greater variance of residual error as basin area size decreases. This is also where there was a large range in abundances of aggregations... lots of aggregations with low numbers, but also dense beds.

Final Model: Abundance ~ Land Use & Stream Power

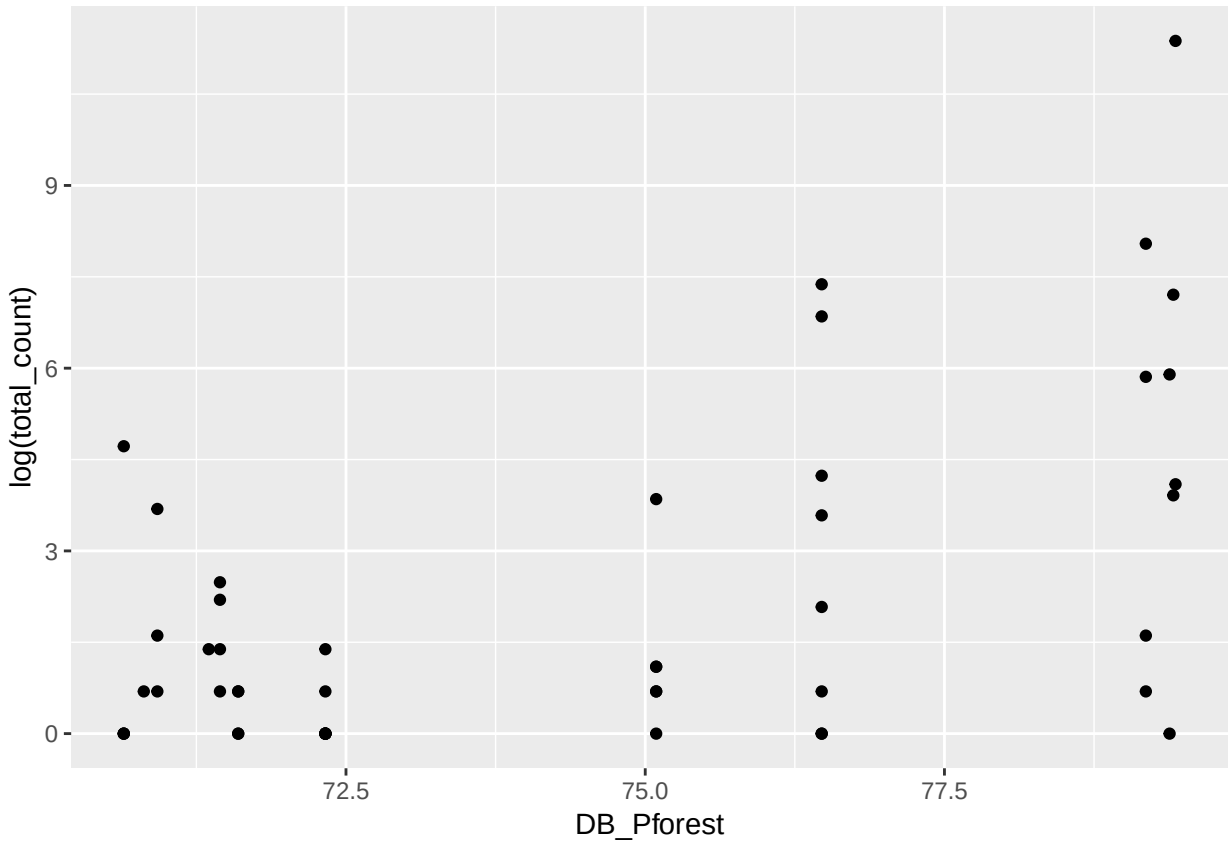
Stepwise regression was used to determine final model, which includes Drainage Basin % Forest, 10-year specific stream power, and the % timber harvest within the HUC12 as independent explanatory variables

```
mod1 <- lm(log(finalmodel$total_count + .01) ~ finalmodel$DB_Pforest + finalmodel$SP_10yr + finalmodel$
summary(mod1)
```

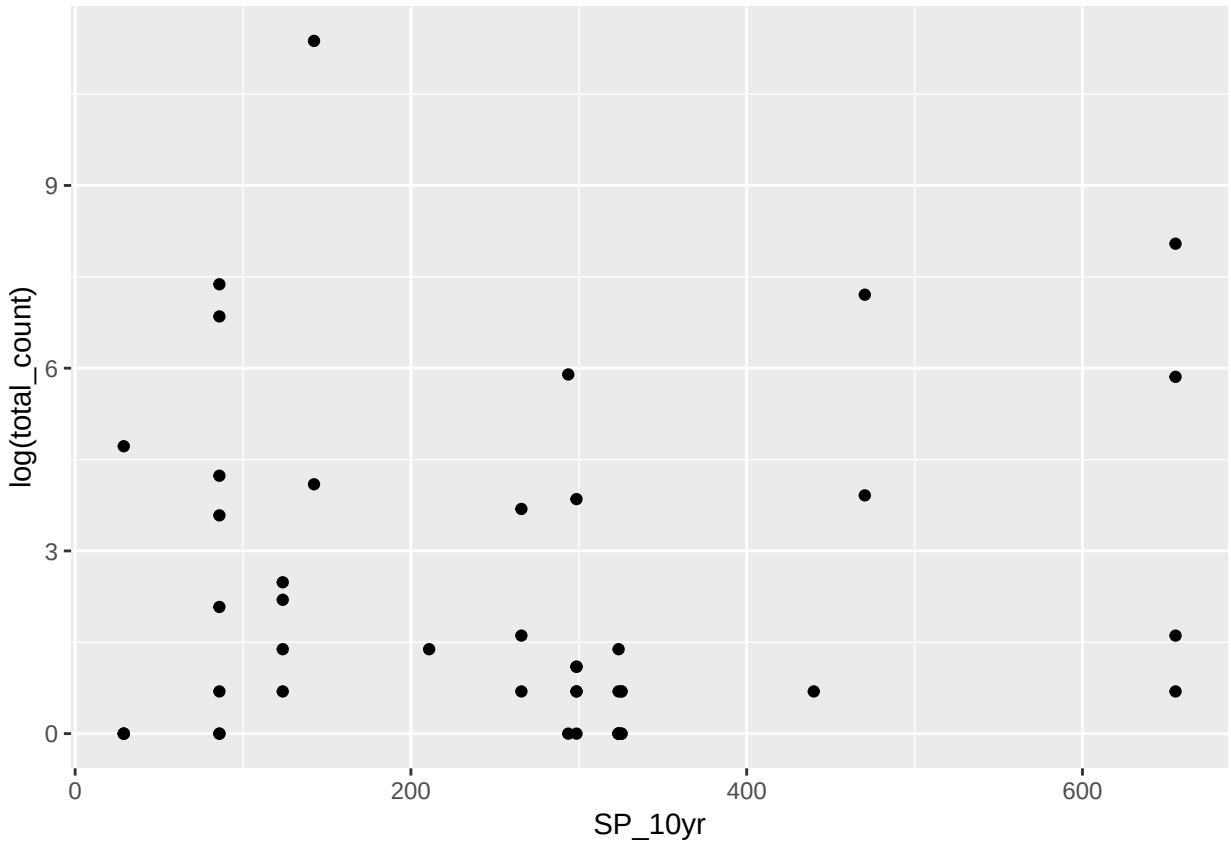
```
##
## Call:
## lm(formula = log(finalmodel$total_count + 0.01) ~ finalmodel$DB_Pforest +
##     finalmodel$SP_10yr + finalmodel$HUC_Pth)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4400 -0.8470 -0.2744  0.7227  5.2374
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -35.394983   7.476663  -4.734 2.22e-05 ***
## finalmodel$DB_Pforest   0.470889   0.103395   4.554 3.99e-05 ***
## finalmodel$SP_10yr     -0.004376   0.002129  -2.056   0.0457 *
## finalmodel$HUC_Pth      0.104668   0.045805   2.285   0.0271 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.144 on 45 degrees of freedom
## Multiple R-squared:  0.3956, Adjusted R-squared:  0.3553
## F-statistic: 9.818 on 3 and 45 DF,  p-value: 4.242e-05
```

Assess Assumptions of Normality for use of OLS Regression for Abundance ~ Land Use/ Stream Power Variables

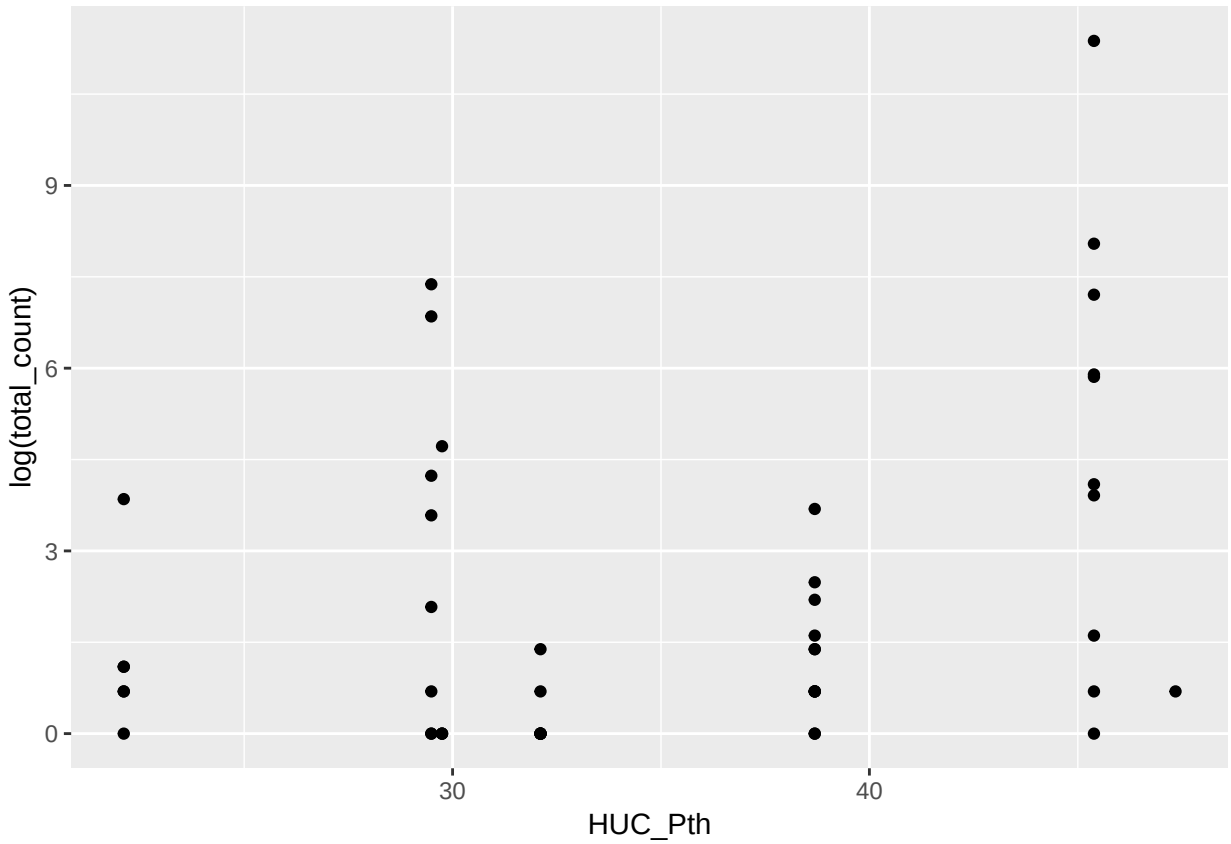
```
#Assumption of linearity
ggplot(finalmodel) + geom_point(aes(x = DB_Pforest, y = log(total_count))) #Yes
```




```
ggplot(finalmodel) + geom_point(aes(x = SP_10yr, y = log(total_count))) #Not much linearity here!!!
```

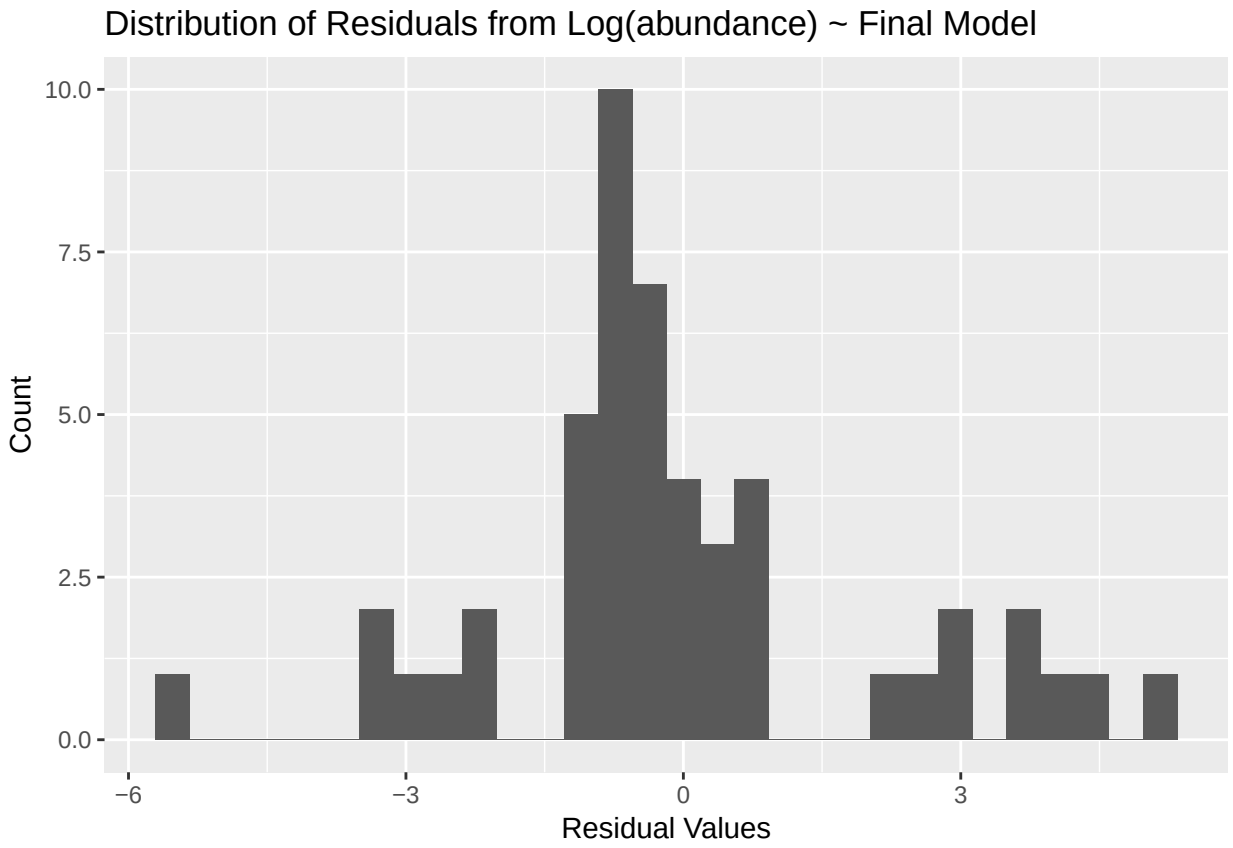



```
ggplot(finalmodel) + geom_point(aes(x = HUC_Pth, y = log(total_count))) #Hard to tell b/c of low sample
```




```
#Examine histogram of the residual values to determine if they are normally distributed
ggplot(mod1) + geom_histogram(aes(mod1$residuals)) + labs(x = "Residual Values", y = "Count") +
  ggtitle("Distribution of Residuals from Log(abundance) ~ Final Model")
```

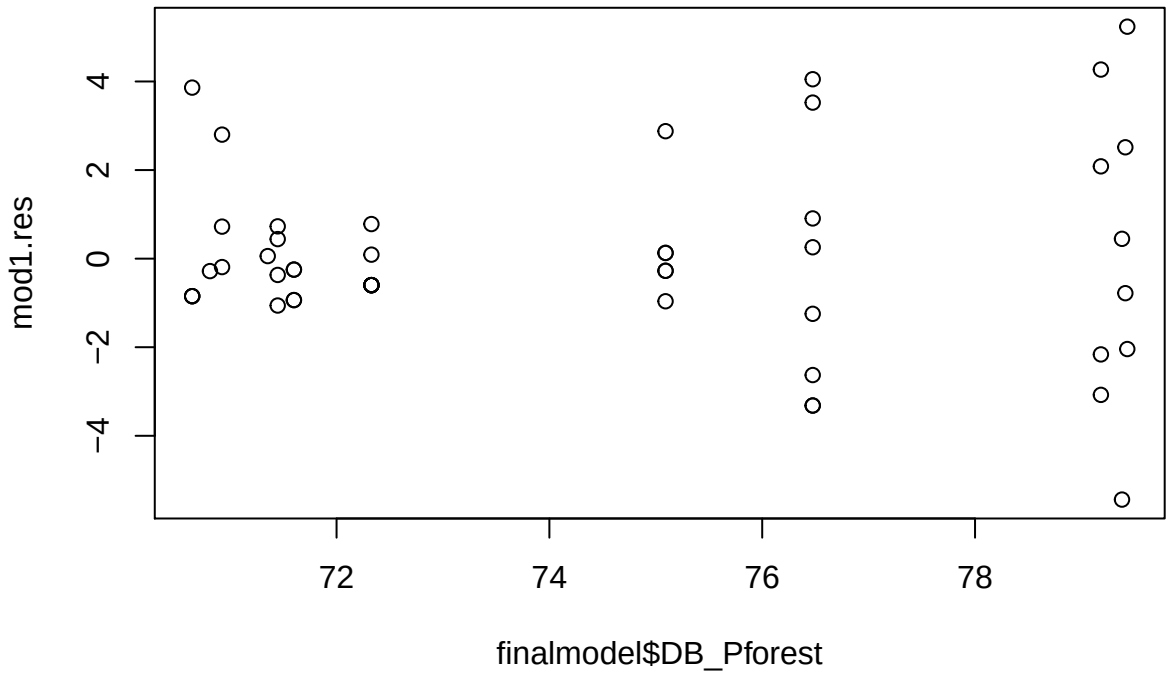
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



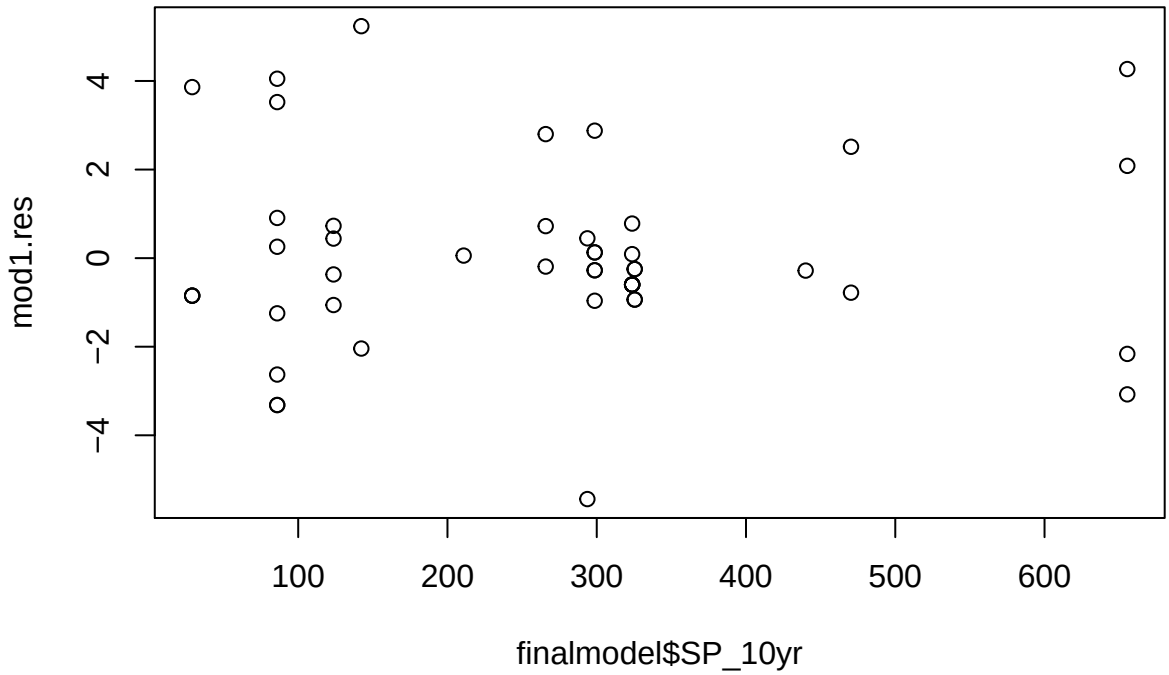

```
# Determine mean of residual values
print(mean(mod1$residuals)) # YES the mean of residuals = 0!!!
```

```
## [1] 8.160115e-17
```

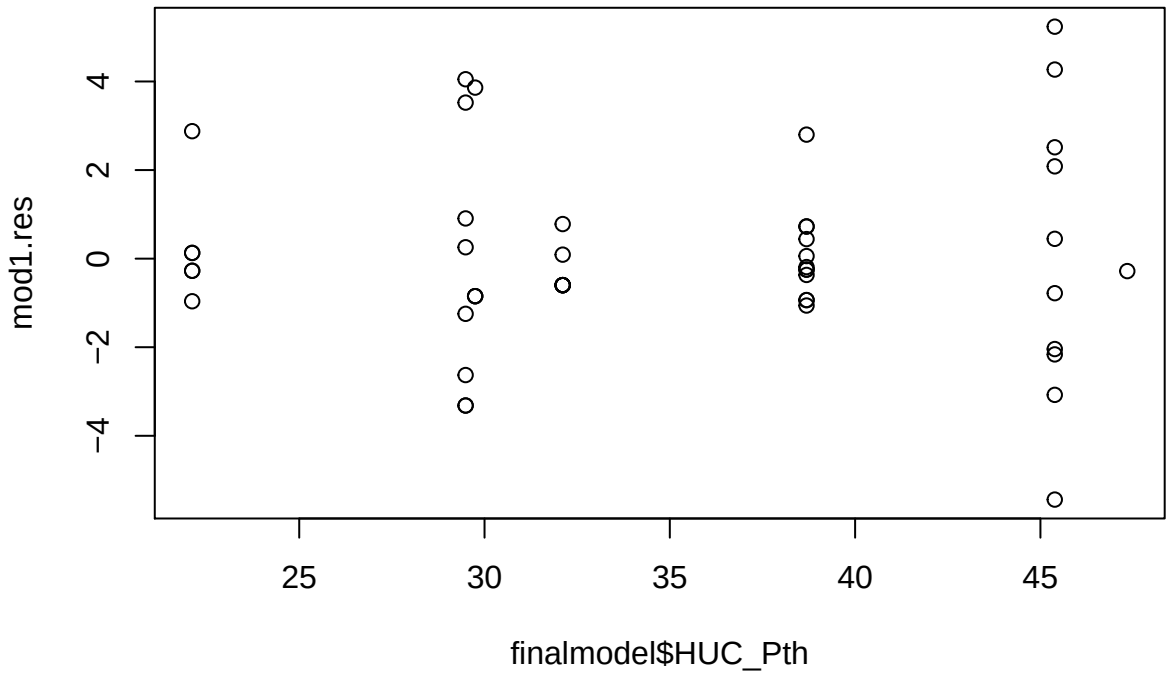
```
# Assumption of homoscedasticity of residuals
mod1.res <- resid(mod1)
plot(finalmodel$DB_Pforest, mod1.res) #No... greater variance at higher % forest
```




```
plot(finalmodel$SP_10yr, mod1.res) # Ok... still unequal vertical variance but better horizontal varian
```




```
plot(finalmodel$HUC_Pth, mod1.res) # Less variance of errors at low percent of timber harvest in the HU
```



##

Results:

1. Assumption of linearity:

Generally met for the drainage basin % forest variable. Seems not met for the stream power variable, but mostly because there doesn't appear to be any relationship. The linear relationship between mussel abundance and HUC12 timber harvest is unclear due to low sample size (not enough sites in enough HUC12 units)

2. Assumption of Zero Mean of Residuals:

MET!!!

3. Assumption of normality of residual error:

MET!!! The distribution of residual error appears normally distributed.

4. Assumption of homoscedasticity:

Not met... there is greater variance of residual error as the percent of forest in the drainage basin increases. There is greater variance of residual error as the percent of timber harvest in the HUC12 unit increases.

#My Thoughts: It is interesting that variance of residuals increases with decreasing drainage basin (DB) size and increasing DB % forest cover and HUC12 % timber harvest.... this corresponds with areas where the range of mussel abundances in aggregations was much greater.

##Potential Solution #1: Remove outliers

Not sure that mussel aggregations of only 1 mussel are ecologically relevant. These isolated mussels are likely linked to different processes than what is structuring larger aggregations of mussels. Try removing all aggregations of only 1 mussel.

```r
#filter to only consider aggregations with greater than 1 mussel
finalmodel2 <- finalmodel%>%
  filter(!site_id == "ZINCCMP") %>%
  filter(total_count > 1)

drainmodel2 <- finalmodel2 %>%
  inner_join(distarea, by = "obs_id")
```

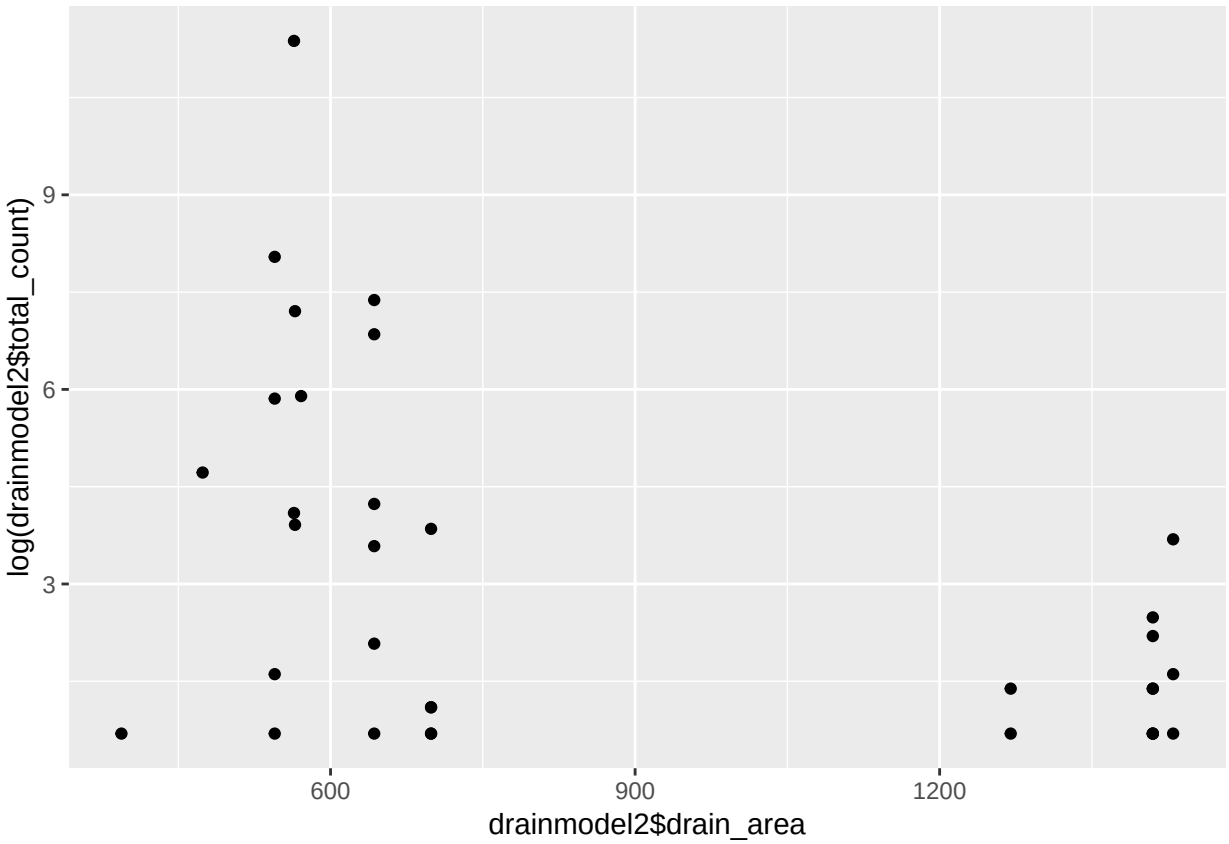
```
## Warning: Column `obs_id` joining factors with different levels, coercing to
## character vector
```

Repeat prior steps for Mussel Abundance ~ Drainage Basin and Assess any Differences

```r
#Simple linear regression predicting mussel abundance by drainage basin area of aggregation
drain_model2 <- lm(log(drainmodel2$total_count + .01) ~ drainmodel2$drain_area)
summary(drain_model2)
```

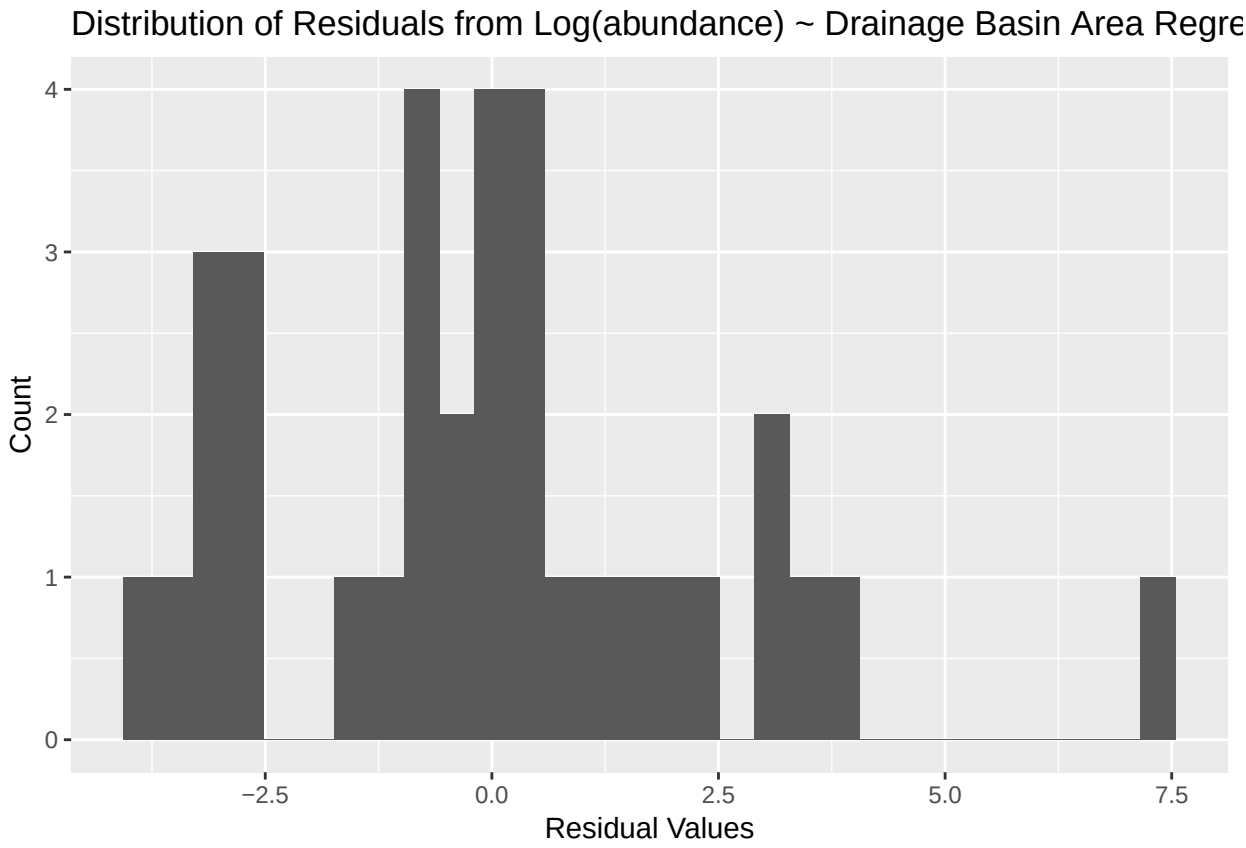
```
##
## Call:
## lm(formula = log(drainmodel2$total_count + 0.01) ~ drainmodel2$drain_area)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.909 -1.581 -0.068  1.040  7.307
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               5.861060   1.056534   5.547 4.04e-06 ***
## drainmodel2$drain_area -0.003182   0.001100  -2.893  0.00681 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.465 on 32 degrees of freedom
## Multiple R-squared:  0.2073, Adjusted R-squared:  0.1826
## F-statistic: 8.371 on 1 and 32 DF,  p-value: 0.006809
```

```r
#Assumption of linearity
ggplot(drainmodel2) + geom_point(aes(x = drainmodel2$drain_area, y = log(drainmodel2$total_count)))
```




```
#Examine histogram of the residual values to determine if they are normally distributed
ggplot(drain_model2) + geom_histogram(aes(drain_model2$residuals)) + labs(x = "Residual Values", y = "C
  ggtitle("Distribution of Residuals from Log(abundance) ~ Drainage Basin Area Regression")
```

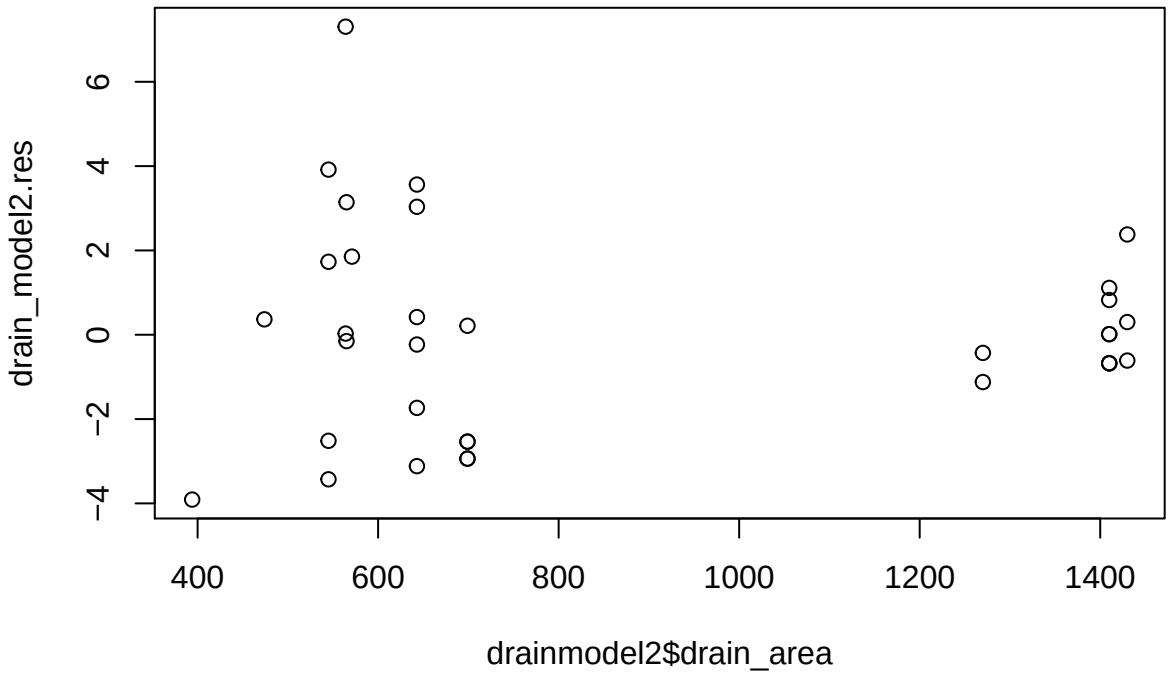
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



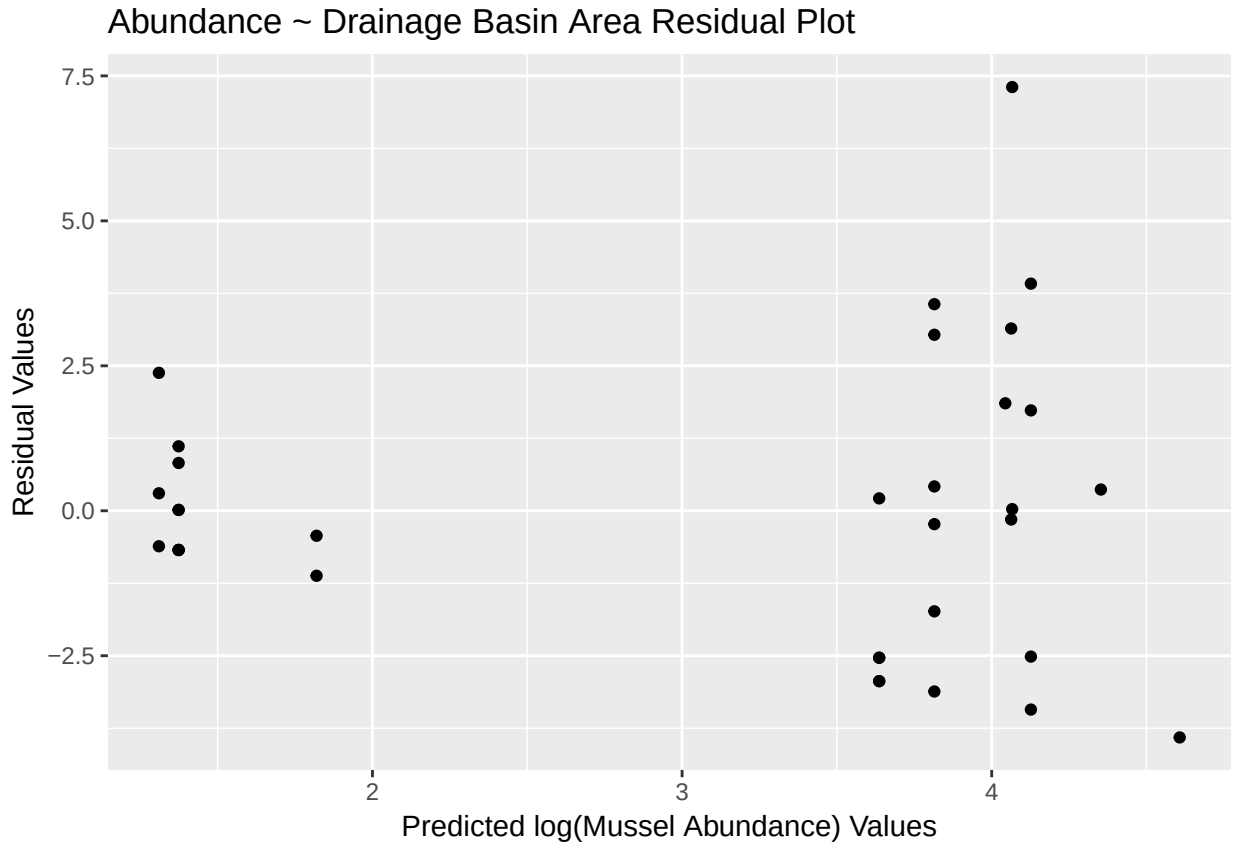

```
# Determine mean of residual values
print(mean(drain_model2$residuals)) # YES the mean of residuals = 0!!!
```

```
## [1] -5.230319e-17
```

```
# Assess whether residuals meet assumption of homoscedasticity
drain_model2.res <- resid(drain_model2)
plot(drainmodel2$drain_area, drain_model2.res) #Looks like greater variance at lower drainage basin are
```




```
# Residual Plot
ggplot(drain_model2) + geom_point(aes(x = drain_model2$fitted.values, y = drain_model2$residuals)) +
  labs(x = "Predicted log(Mussel Abundance) Values", y = "Residual Values") + ggtitle("Abundance ~ Drai
```



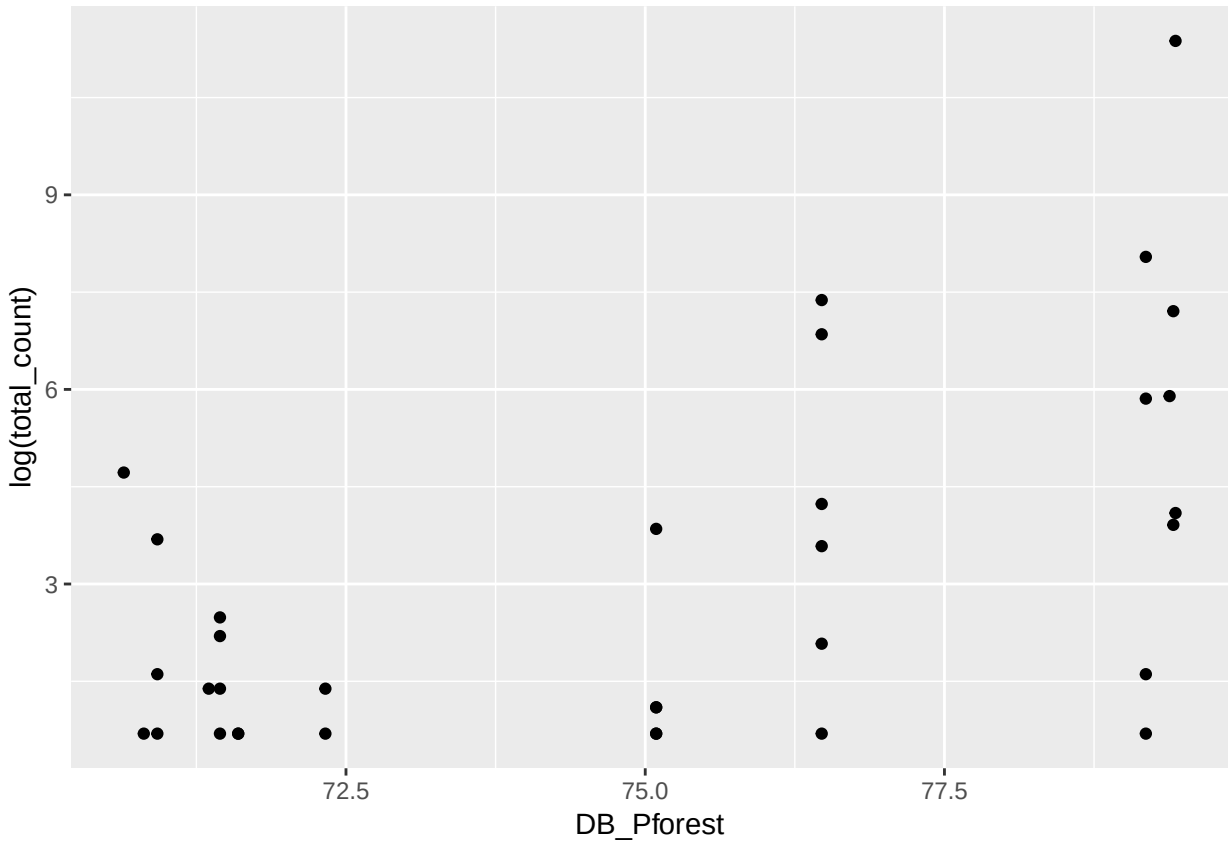
Differences:

Repeat prior steps for Mussel Abundance ~ Land Use/Stream Power and Assess any Differences

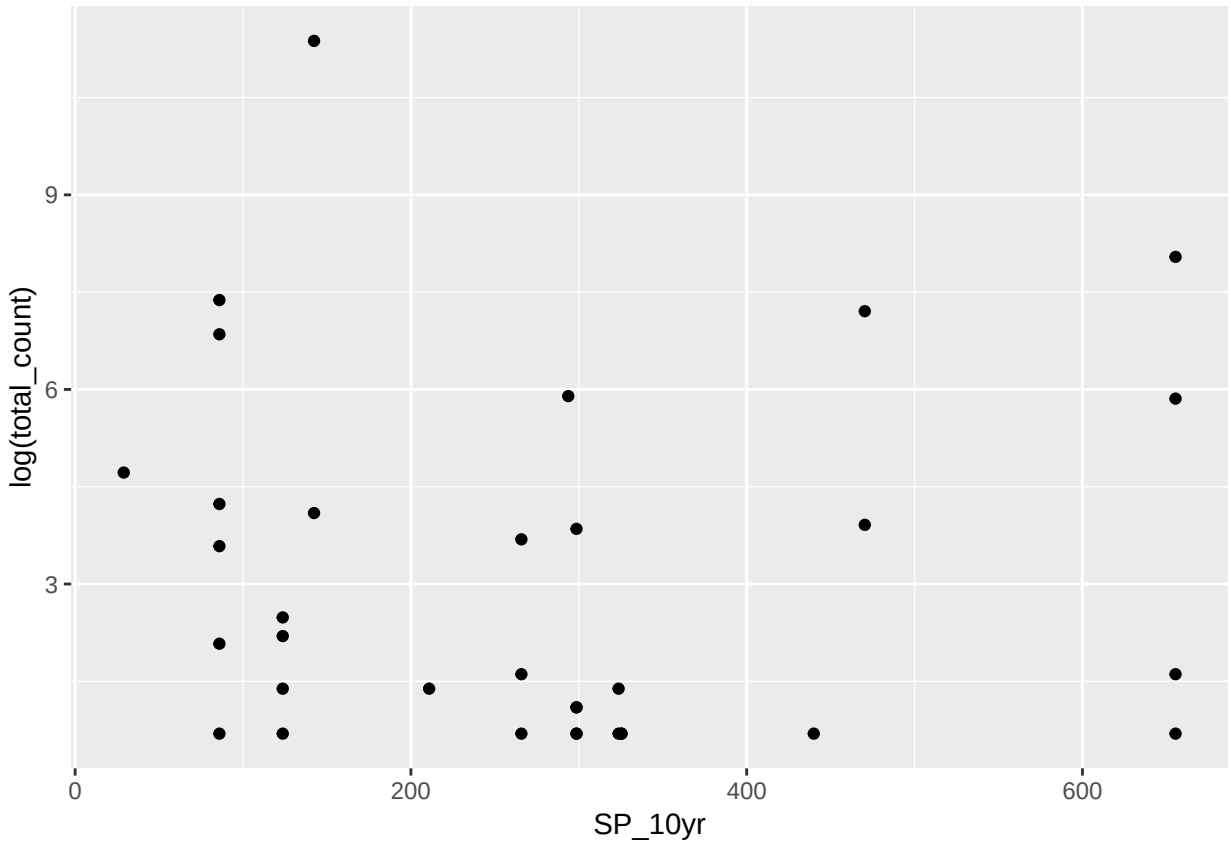
```
mod2 <- lm(log(finalmodel2$total_count + .01) ~ finalmodel2$DB_Pforest + finalmodel2$SP_10yr + finalmod
summary(mod2)
```

```
##
## Call:
## lm(formula = log(finalmodel2$total_count + 0.01) ~ finalmodel2$DB_Pforest +
##     finalmodel2$SP_10yr + finalmodel2$HUC_Pth)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6870 -0.8754 -0.4328  1.5192  4.1142
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -38.754248   8.705396  -4.452 0.000109 ***
## finalmodel2$DB_Pforest   0.530962   0.118320   4.488 9.86e-05 ***
## finalmodel2$SP_10yr     -0.006163   0.002368  -2.602 0.014258 *
## finalmodel2$HUC_Pth      0.103899   0.049377   2.104 0.043848 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.097 on 30 degrees of freedom
## Multiple R-squared:  0.4622, Adjusted R-squared:  0.4084
## F-statistic: 8.594 on 3 and 30 DF,  p-value: 0.0002872
```

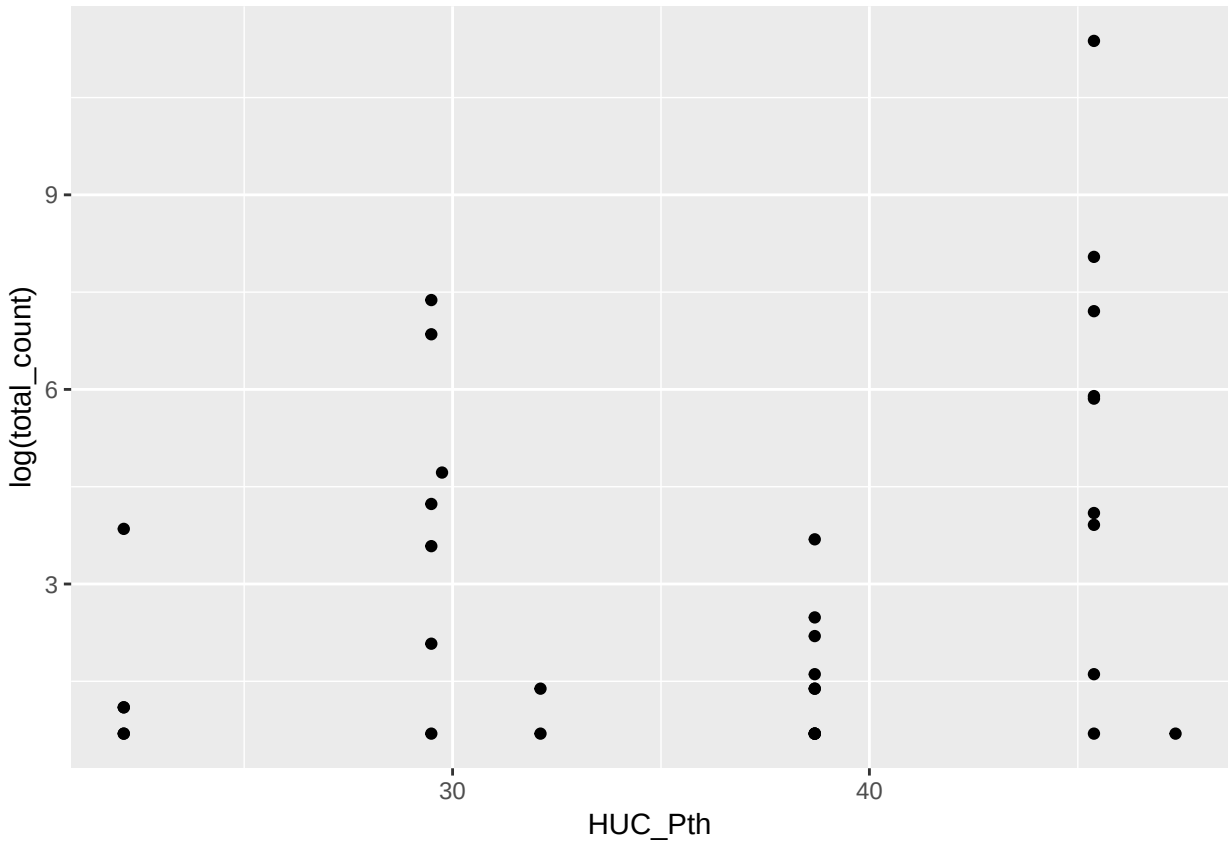
```
#Assumption of linearity
ggplot(finalmodel2) + geom_point(aes(x = DB_Pforest, y = log(total_count))) #Yes
```




```
ggplot(finalmodel2) + geom_point(aes(x = SP_10yr, y = log(total_count))) #Not much linearity here!!!
```

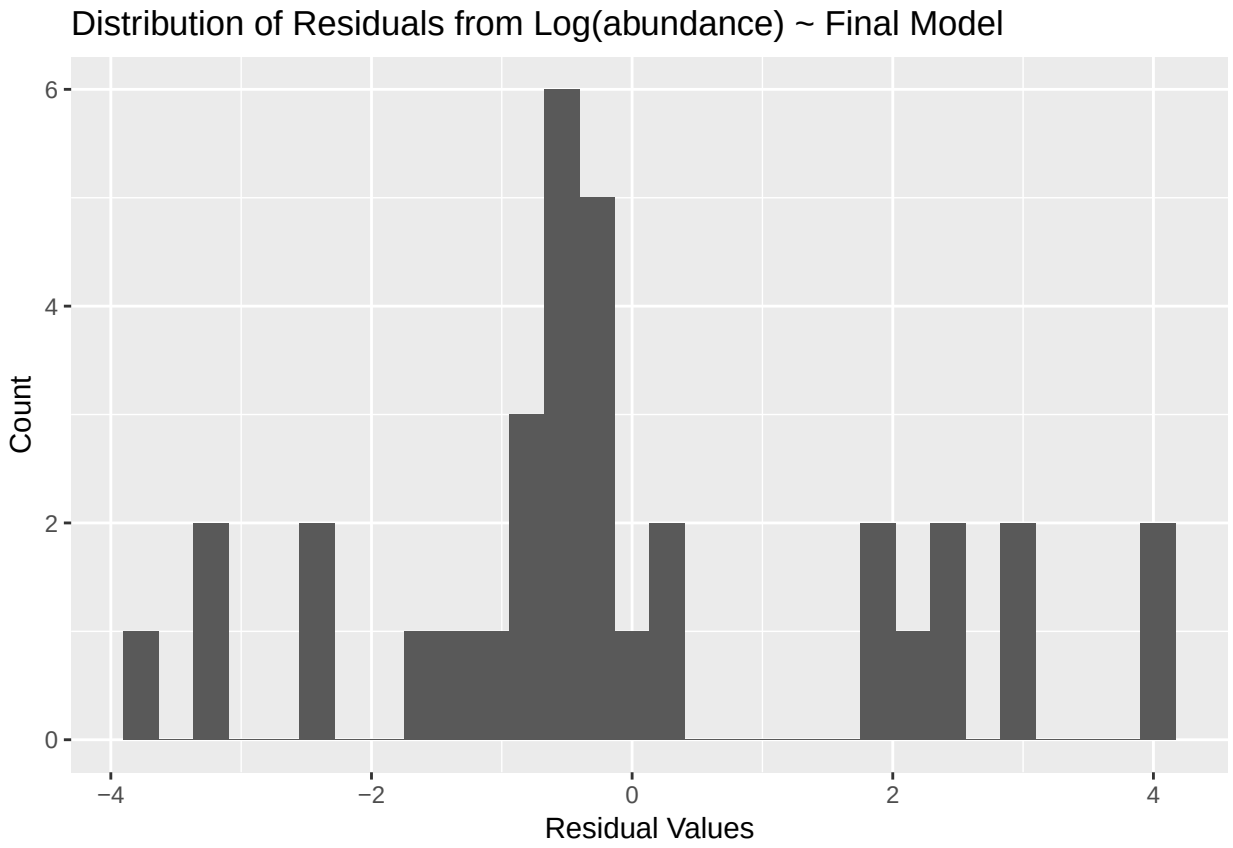



```
ggplot(finalmodel2) + geom_point(aes(x = HUC_Pth, y = log(total_count))) #Hard to tell b/c of low sampl
```




```
#Examine histogram of the residual values to determine if they are normally distributed
ggplot(mod2) + geom_histogram(aes(mod2$residuals)) + labs(x = "Residual Values", y = "Count") +
  ggtitle("Distribution of Residuals from Log(abundance) ~ Final Model")
```

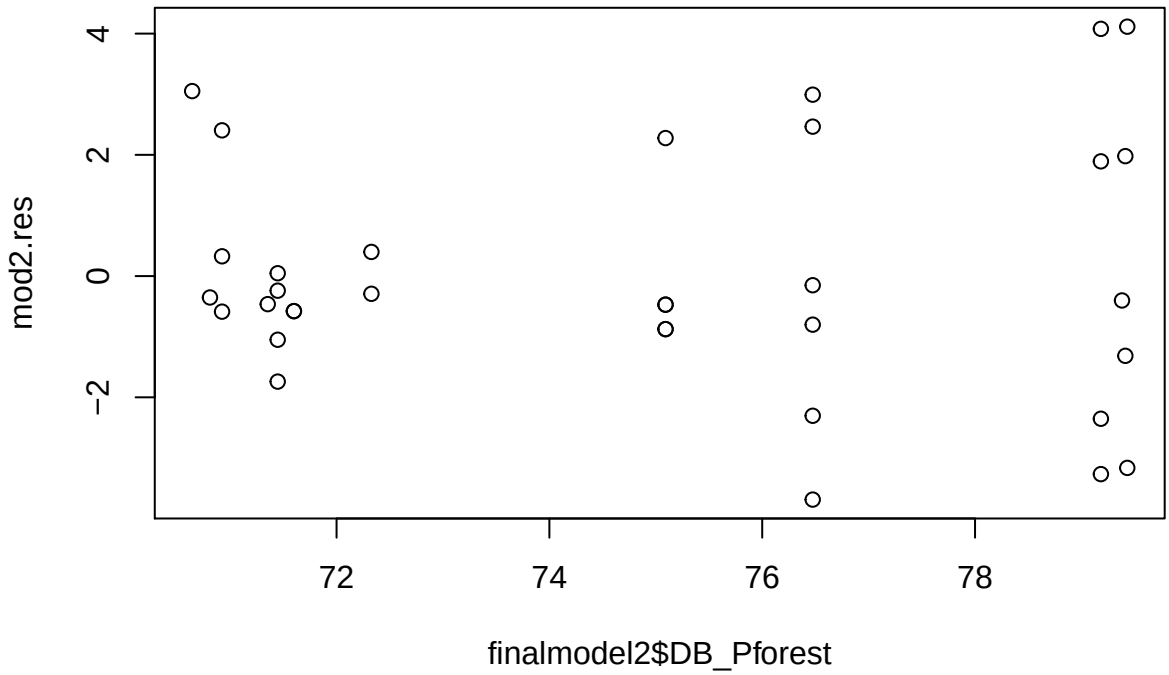
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



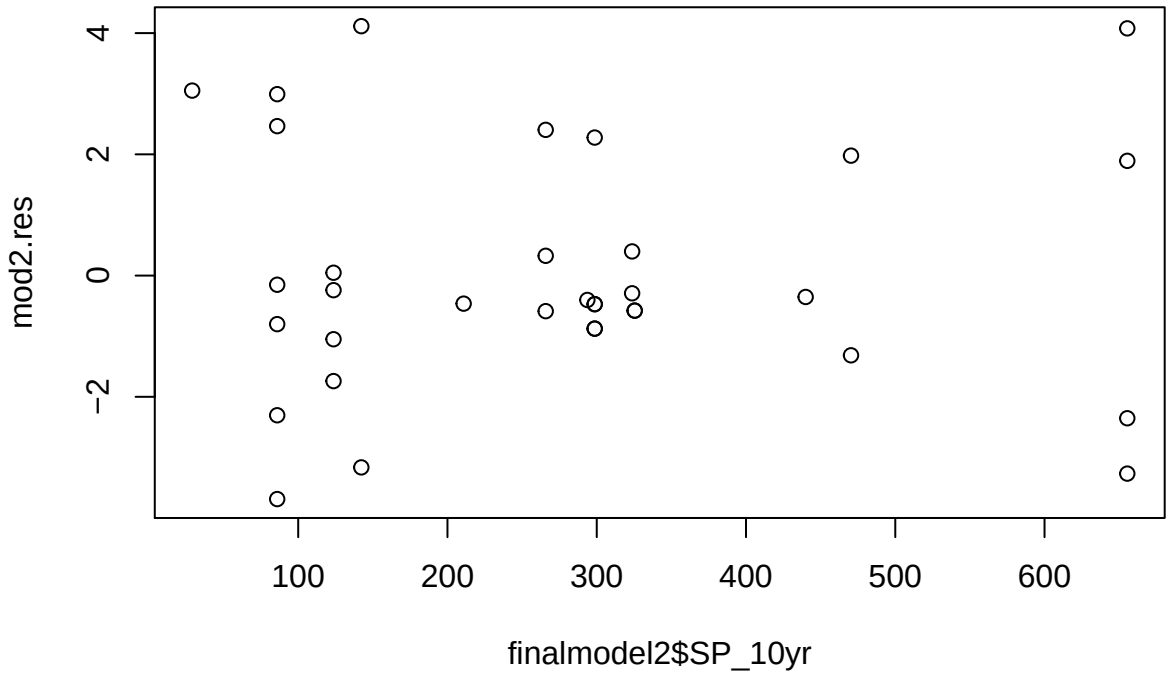

```
# Determine mean of residual values
print(mean(mod2$residuals)) # YES the mean of residuals = 0!!!
```

```
## [1] -4.571586e-17
```

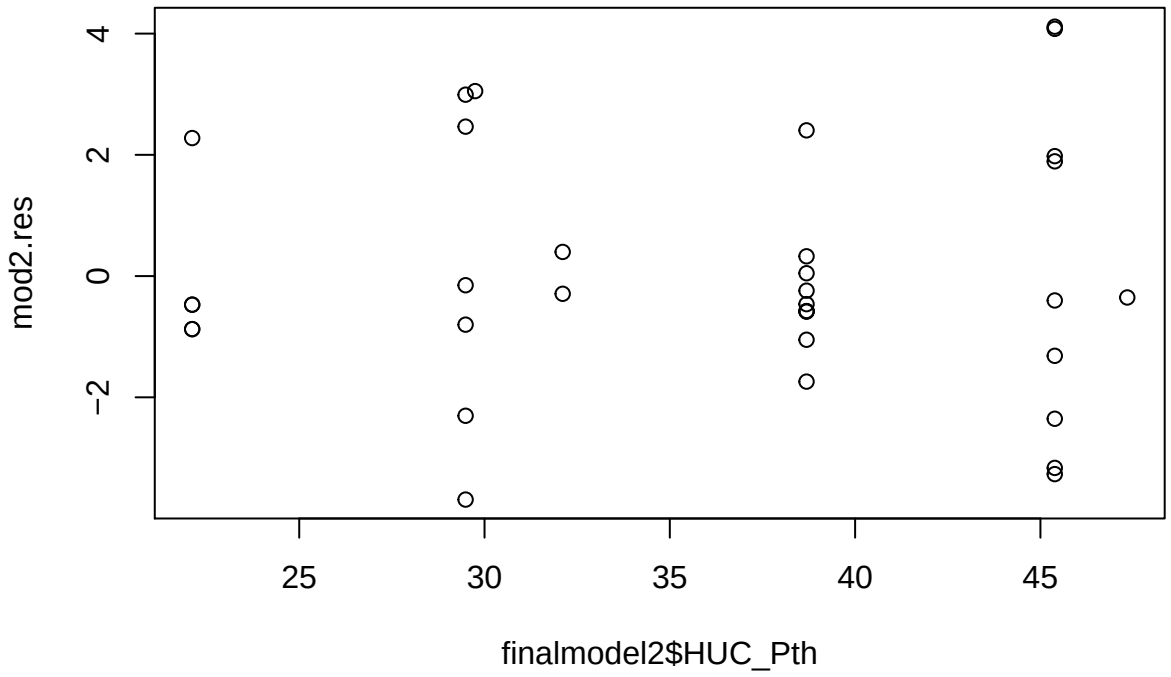
```
# Assumption of homoscedasticity of residuals
mod2.res <- resid(mod2)
plot(finalmodel2$DB_Pforest, mod2.res) #No... greater variance at higher % forest
```




```
plot(finalmodel2$SP_10yr, mod2.res) # Ok... still unequal vertical variance but better horizontal varia
```




```
plot(finalmodel2$HUC_Pth, mod2.res) # Less variance of errors at low percent of timber harvest in the H
```



Results:

Abundance ~ Basin Area:

Removing aggregations of only 1 mussel from dataset resulted in a more normal distribution of residual error values and a better vertical homoscedastic balance of residual errors but DID NOT fix the horizontal heteroscedasticity of the residiual errors. Removing aggregations of 1 mussel also resulted in slightly different variable coefficients in the model, marginally increased standard error for model coefficients, but reduced the residual standard error of the model as a whole. The model as a whole had a increased R^2 value, increased F statistic value, and decreased p-value.

Abundance ~ Land Use/ Stream Power

Removing aggregations of only 1 mussel drom the dataset resulted in slightly better vertical homoscedastic balance in the residual errors (lower bound of y axis went from -4 to -2), but did not do much for the horizontal heteroscedasticity problem. Removing aggregations of 1 mussel also resulted in slightly different variable coefficients in the model, marginally increased standard error for model coefficients, but reduced the residual standard error of the model as a whole. The model as a whole had a increased R^2 value, decreased F statistic value, and increased p-value (although model was still significant).