STAT230 Work Team Project: Proposal & EDA

Work Team 7: Evan, Becca, Laith

due Thursday, November 2nd

Introduction & Motivation

We are interested in how we can attempt to predict the profitability of farms by different farming practices. Specifically we are curious about sustainable agriculture decisions and decisions about crop types. From a conservationist perspective, it is important that sustainable choices are economically viable for farmers. https://www.nass.usda.gov/AgCensus/

Our data is sourced from the United States Census of Agriculture, between the years 2009 and 2017, along with county demographic information from Wikipedia, sourced from the U.S Census. The data is collected by county, often summing the a variable across all operations in a given county. This limits our specificity, but we are still able to estimate what is typical for farms across the county by looking at averages for each county.

Our response variable in reported farm income, and we are planning to predict it with a number of variables (population density, farm acres, bee colonies, grazing rotation, fertilizer use).

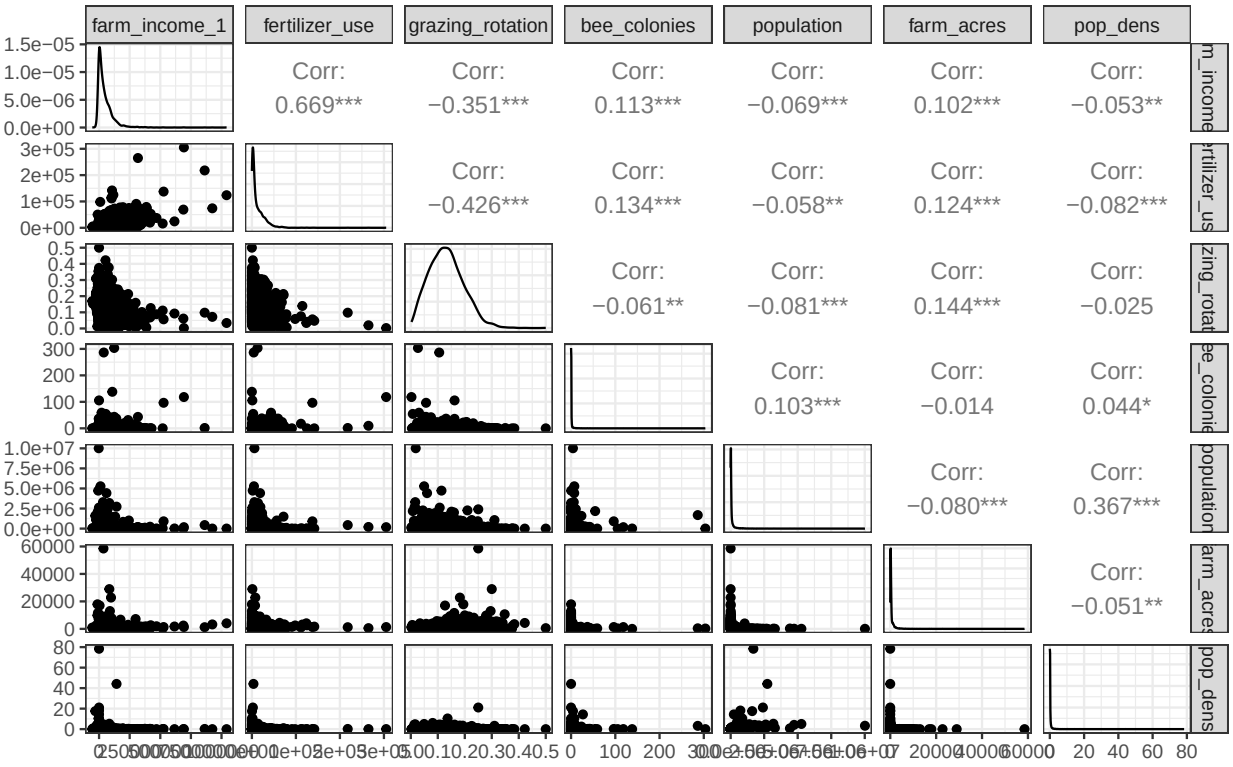
grazing rotation refers to the number of farming operations in a county that reported using either crop cycling or alternating patterns of grazing to minimize impact. Farm acres and bee colonies are counts that are summed across all farms within a state. Fertilizer use is the total dollar amount spent on fertilizers in that year by all farming operations in a given county. Population density was estimated based on county sizes and populations based on the U.S census via Wikipedia. We feel that this combination of predictors is appropriate because it examines many dimensions of farming practices. It includes information about context of the farming (population density, farm size) as well as the types of methods of sustainable or unsustainable practices carried out in each of these counties. It is valuable to understand how economically viable sustainable farming is in the short term in order to access if further supportive measures should be taken in order to ensure both sufficient food production, as well as producing that food in ways that will limit harm to the Earth.

Dataset & Wrangling

Exploratory Data Analysis (EDA)

Explore distributions and associations graphically and numerically.

```
ggpairs(df1, columns = 1:7, progress = FALSE)
```




```
#looks like farm income could maybe use a transformation to improve linearity.
```

There are significant correlations with all variables, but the linearity looks pretty bad, will try some transformations.

```
gf_histogram(~farm_income_1, data=df1)
```

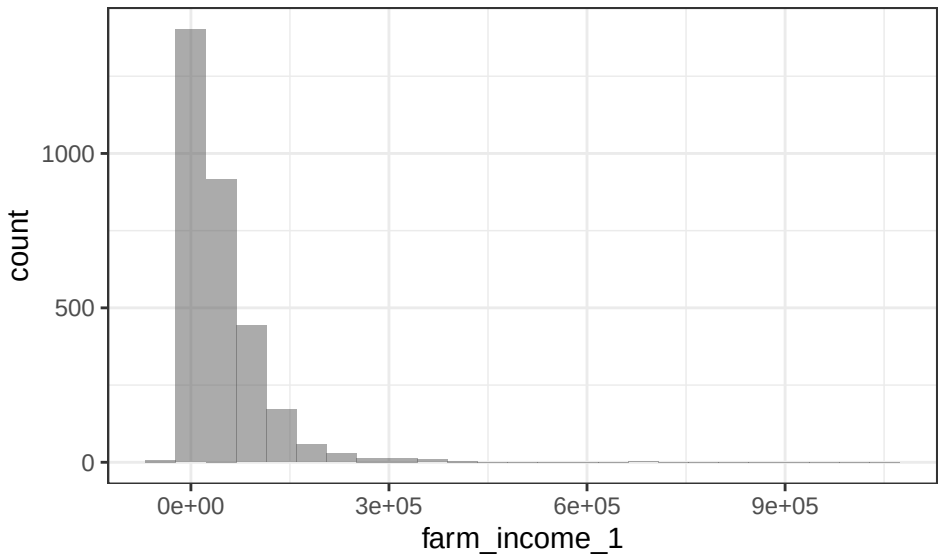


Table 1: Summary of Farm Income

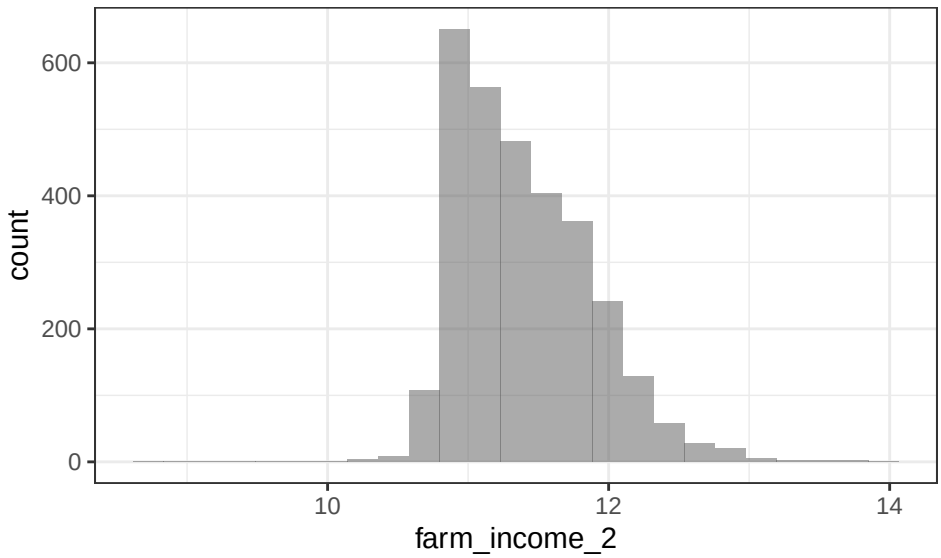
min	Q1	median	Q3	max	mean	sd	n	missing
-55150	5494.5	27177	66913	1041275	46857	67968	3076	0

```
favstats(~farm_income_1, data=df1) |>
  kable(booktabs=TRUE, align = "c", caption = "Summary of Farm Income") |>
  kable_styling(position = "center")
```

The mean is 46857, the median is 27177, we can also tell it looks right skewed from the histogram, and the IQR is 61419

We log-transposed+transformed farm income such that it better matched a normal distribution.

```
df2 <- df1 |>
  mutate(farm_income_2 = log(farm_income_1+55152))

#we had one value that was very low because it was by far the smallest, even after we transposed the da
df2 <- df2 |>
  filter(farm_income_2 > 5)
gf_histogram(~farm_income_2, data=df2)
```



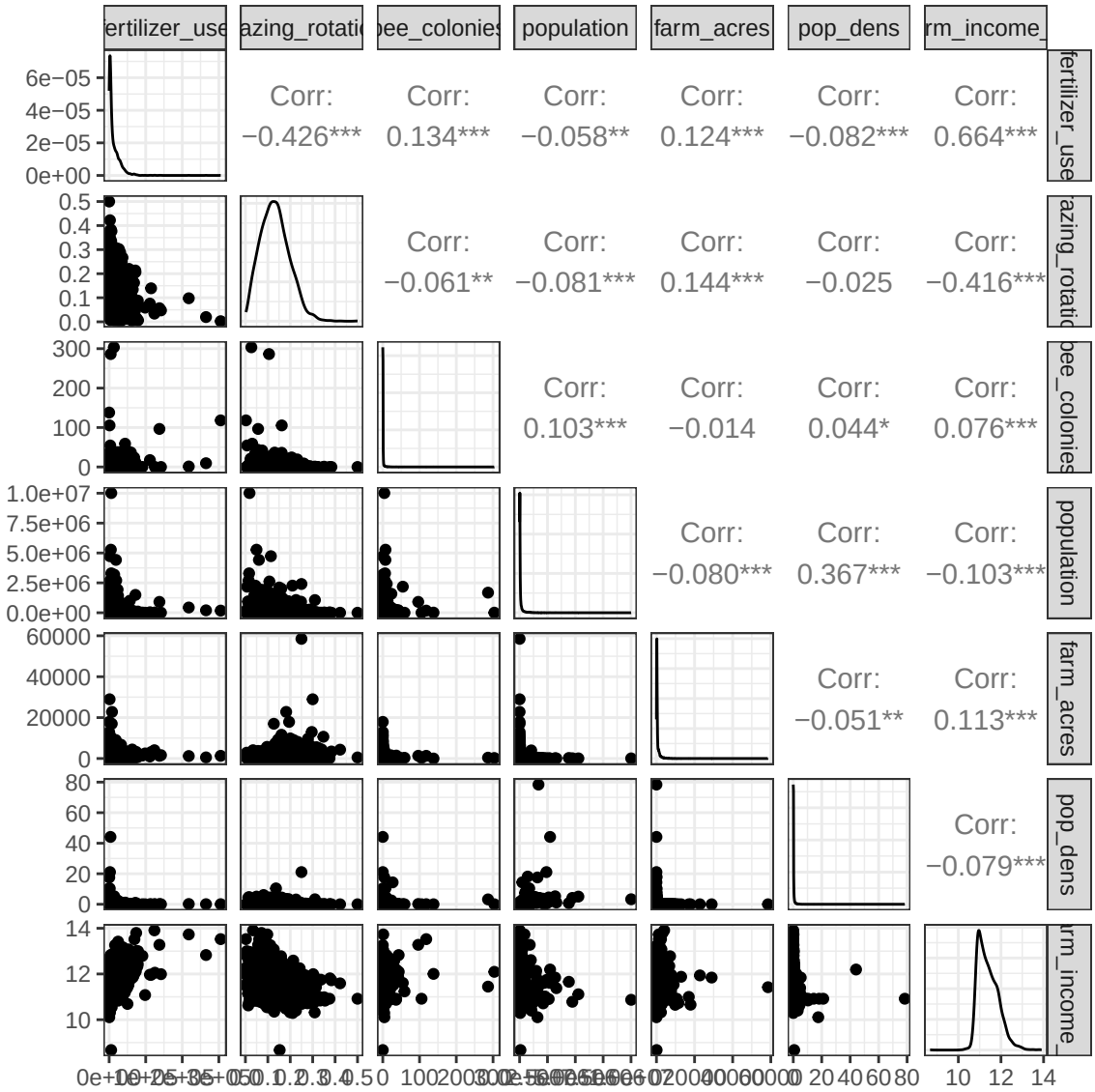

```
favstats(~farm_income_2, data=df2) |>
  kable(booktabs = TRUE, align = 'c', caption = "Summary of Transformed Farm Income") |>
  kable_styling(position = "center")
```

Looks much improved, the mean is 11.4, and median is 11.318, and the histogram still is not ideal, but better. The IQR = 0.699

Table 2: Summary of Transformed Farm Income

min	Q1	median	Q3	max	mean	sd	n	missing
8.6733	11.013	11.319	11.712	13.908	11.4	0.48011	3075	0

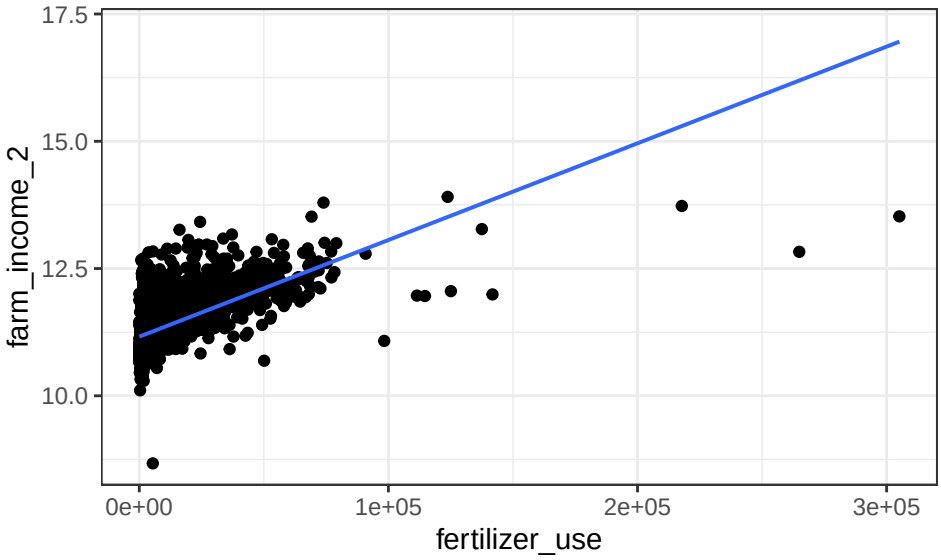
```
ggpairs(df2, columns = 2:8, progress = FALSE)
```



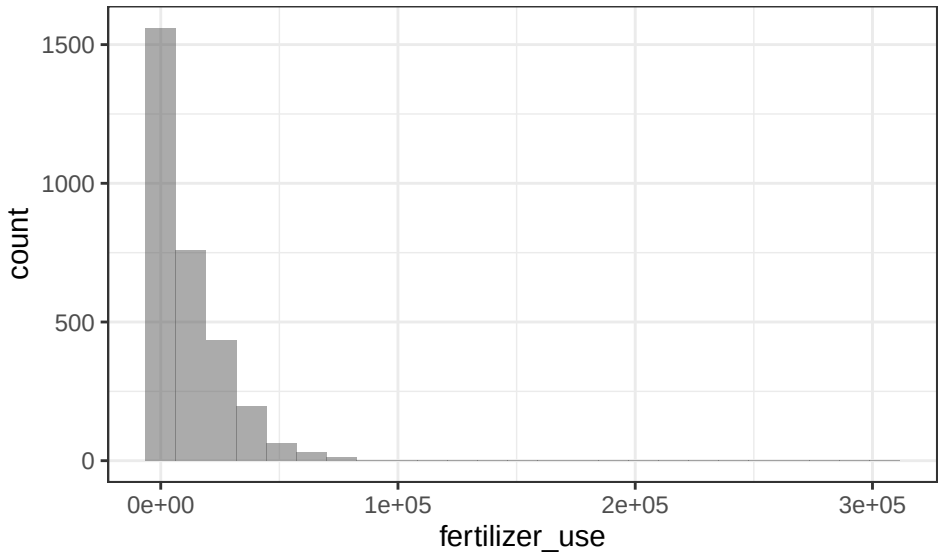
This improved linearity slightly, but still questionable, lets look at predictors.

```
gf_point(data = df2, farm_income_2 ~ fertilizer_use) |>
  gf_lm()
```

min	Q1	median	Q3	max	mean	sd	n	missing
0	2105.3	6182	18511	305114	12600	16789	3069	6




```
gf_histogram(~fertilizer_use, data=df2)
```

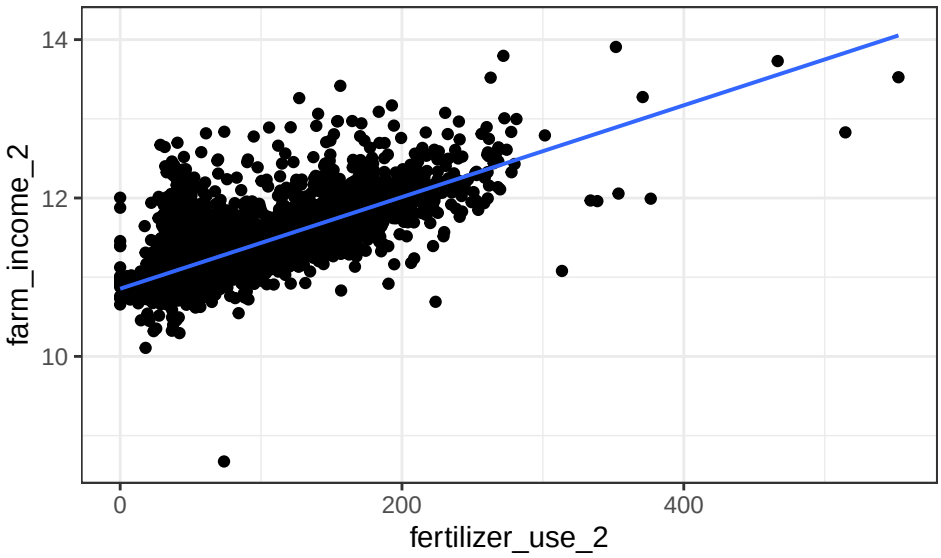



```
favstats(~fertilizer_use, data=df2) |>
  kable(booktabs = TRUE, align = 'c') |>
  kable_styling(position = "center")
```

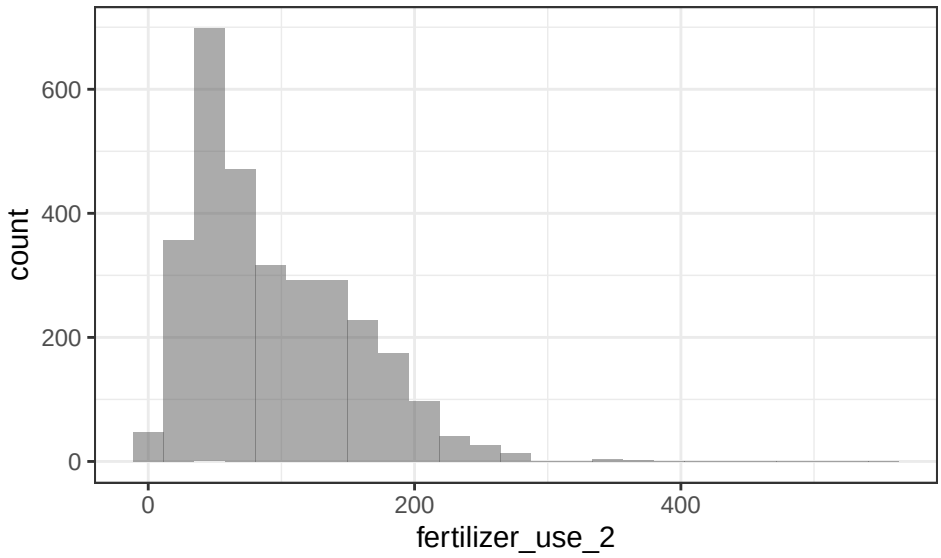
```
#fertilizer use is super right skewed and not linear w/ farm income, so trying a transformation
```

This does not look great, heavily skewed distribution and it is difficult to assess the scatterplot, the mean is 12600, the median is 6182, and the IQR is 16406. Will try transforming.

```r
df3 <- df2 |>
  mutate(fertilizer_use_2 = sqrt(fertilizer_use))

gf_point(data = df3, farm_income_2 ~ fertilizer_use_2) |>
  gf_lm()
```




```r
gf_histogram(~fertilizer_use_2, data=df3)
```

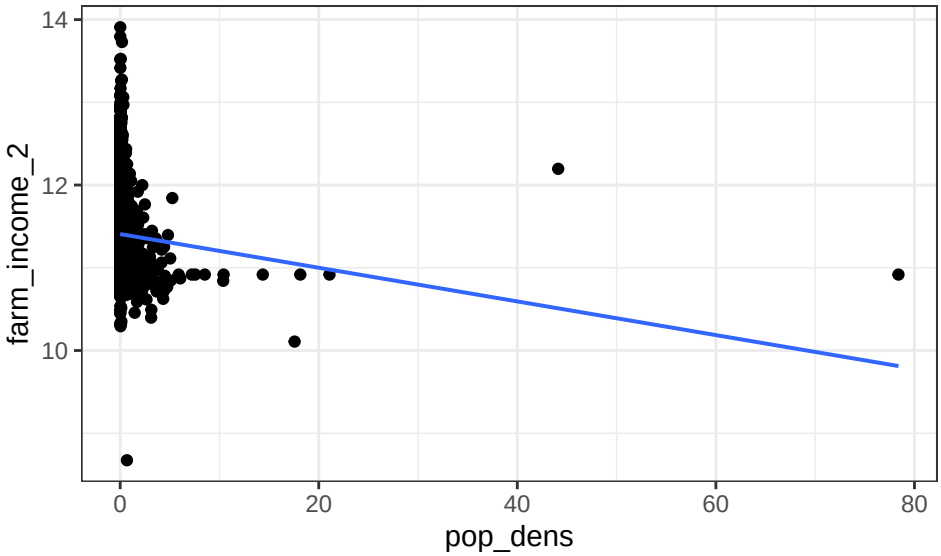



```r
favstats(~fertilizer_use_2, data=df3) |>
  kable(booktabs = TRUE, align = 'c') |>
  kable_styling(position = "center")
```

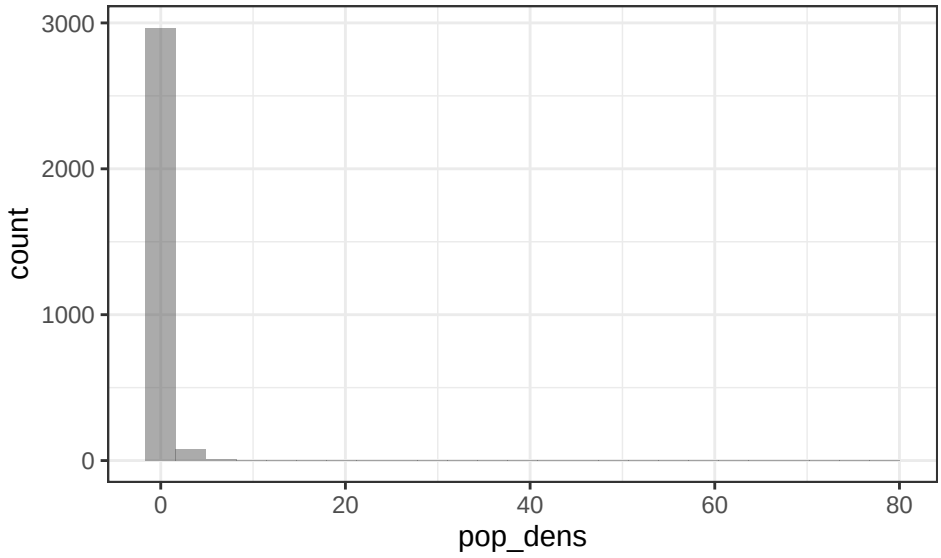
Once again, this improves how the data looks. The distribution is still somewhat right skewed, with a mean of 94.397 and a median of 78.628, and an IQR of 90.167, but this is an improvement, but linearity looks better based on the scatterplot.

min	Q1	median	Q3	max	mean	sd	n	missing
0	45.883	78.626	136.05	552.37	94.397	60.751	3069	6

```
gf_point(data = df3, farm_income_2 ~ pop_dens) |>
  gf_lm()
```




```
gf_histogram(~pop_dens, data=df3)
```



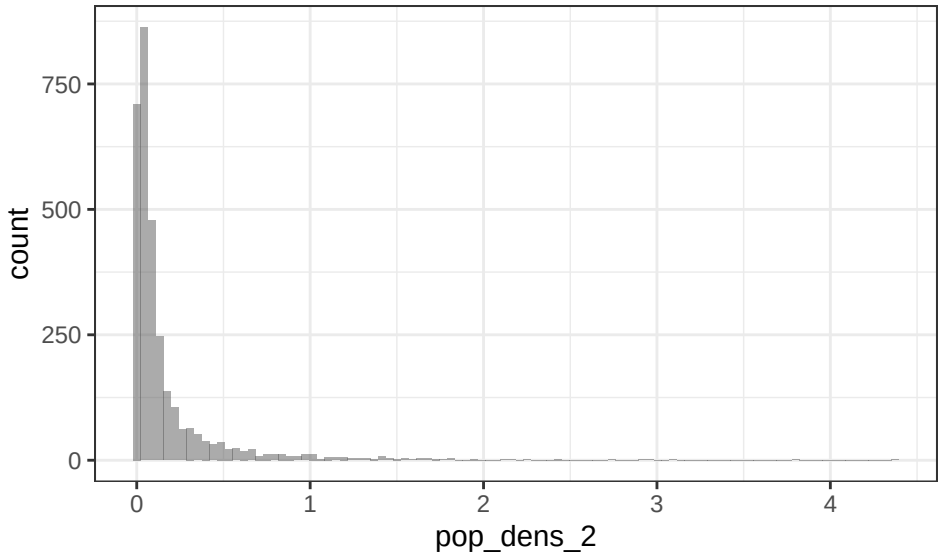

```
favstats(~pop_dens, data=df3) |>
  kable(booktabs = TRUE, align = 'c') |>
  kable_styling(position = "center")
```

min	Q1	median	Q3	max	mean	sd	n	missing
0.00015	0.02472	0.06443	0.16653	78.391	0.29922	1.8532	3061	14

```
#population density is super right skewed and not linear w/ farm income, so trying a transformation
```

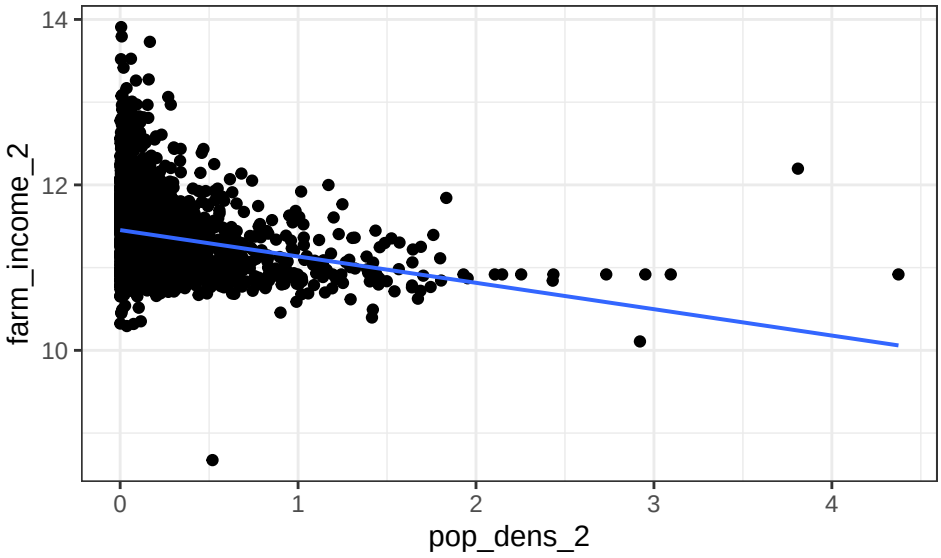
This one is also not great, scatterplot not very useful, the histogram is right skewed, the mean is 0.29922 and median is 0.064434, and the IQR is 0.14181.

```
df3 <- df3 |>
  mutate(pop_dens_2 = log(pop_dens+1))
#again looks a lot better

#looks like we still have a value that is inf because there is somehow 0 acres in the county, just goin

df3 <- df3 |>
  filter(!is.infinite(pop_dens_2))

gf_histogram(data = df3, ~ pop_dens_2, bins = 100)
```




```
gf_point(data = df3, farm_income_2 ~ pop_dens_2) |>
  gf_lm()
```

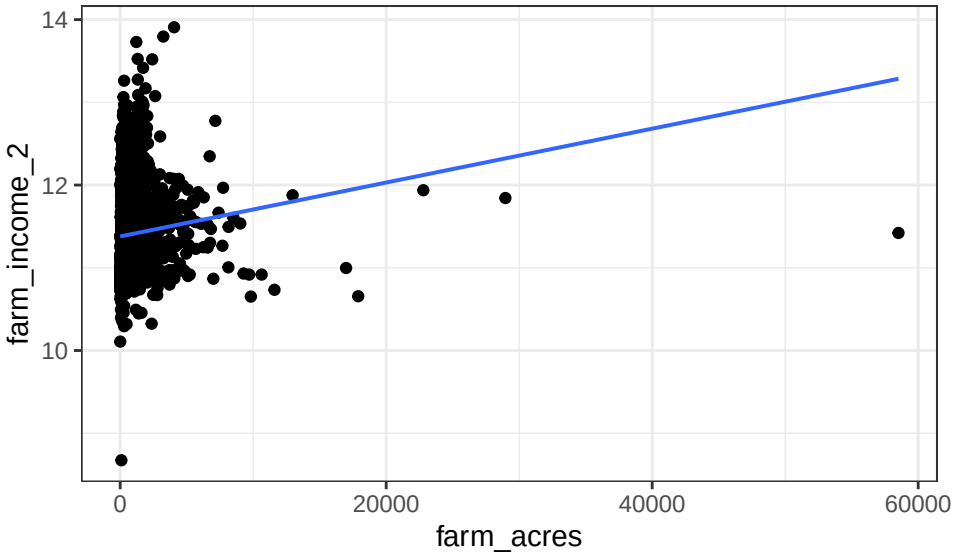
min	Q1	median	Q3	max	mean	sd	n	missing
0.00015	0.02442	0.06244	0.15404	4.3744	0.16477	0.30615	3061	14




```
favstats(~pop_dens_2, data=df3) |>
  kable(booktabs = TRUE, align = 'c') |>
  kable_styling(position = "center")
```

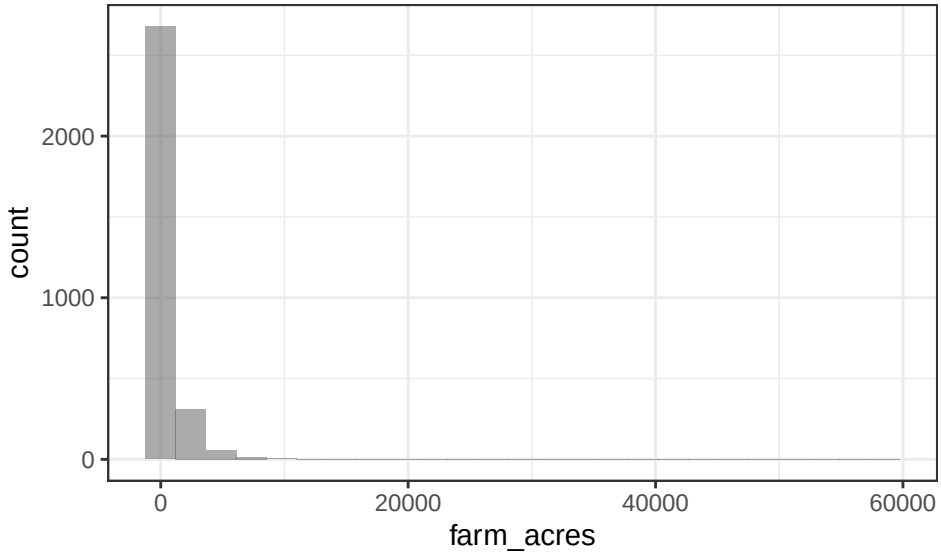
This one still does not look great after transforming, but it probably helped. The graph is still right skewed, but not as strongly, with a mean of 0.16477 and a median of 0.062443, and IQR of 0.12962, and the relationship based on the scatter plot is not very useful.

```
gf_point(data = df3, farm_income_2 ~ farm_acres) |>
  gf_lm()
```



min	Q1	median	Q3	max	mean	sd	n	missing
0	156.87	267.43	527.67	58518	653.95	1663.4	3075	0

```
gf_histogram(data = df3, ~ farm_acres)
```



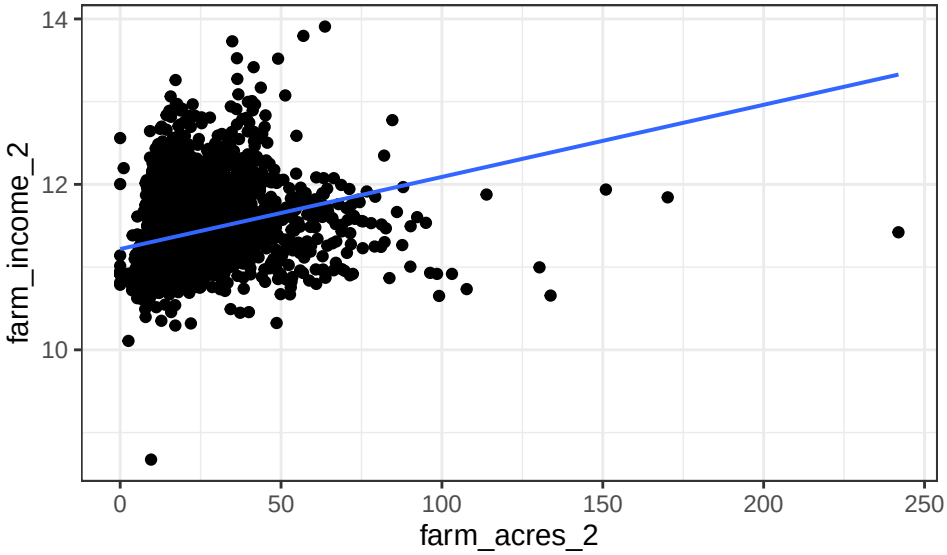

```
favstats(~farm_acres, data=df3) |>
  kable(booktabs = TRUE, align = 'c') |>
  kable_styling(position = "center")
```

```
#farm acres is super right skewed and not linear w/ farm income, so trying a transformation
```

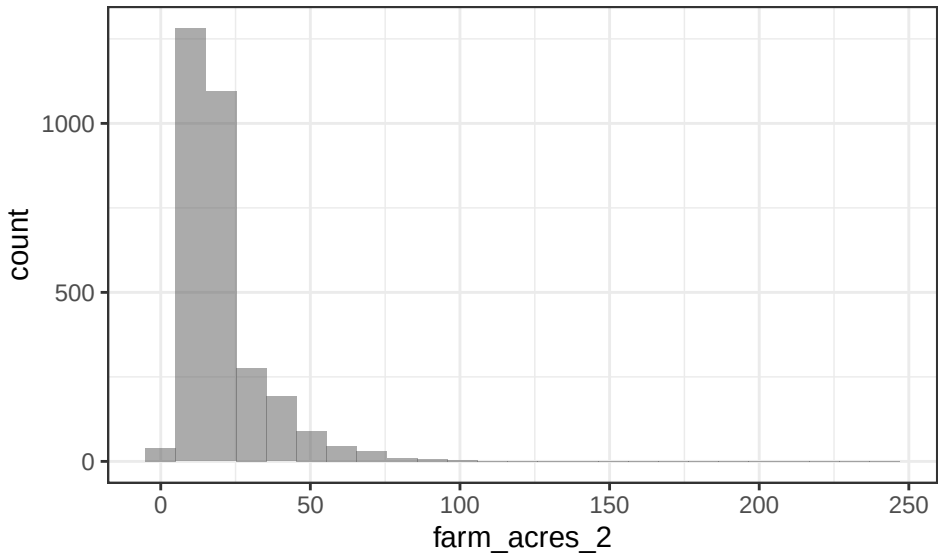
Once again, unhelpful scatterplot, right skewed distribution, mean of 291893, median of 179306 and IQR of 282196

```
df3 <- df3 |>
  mutate(farm_acres_2 = sqrt(farm_acres))

gf_point(data = df3, farm_income_2 ~ farm_acres_2) |>
  gf_lm()
```

min	Q1	median	Q3	max	mean	sd	n	missing
0	12.525	16.353	22.971	241.9	20.833	14.833	3075	0




```
gf_histogram(data = df3, ~ farm_acres_2)
```



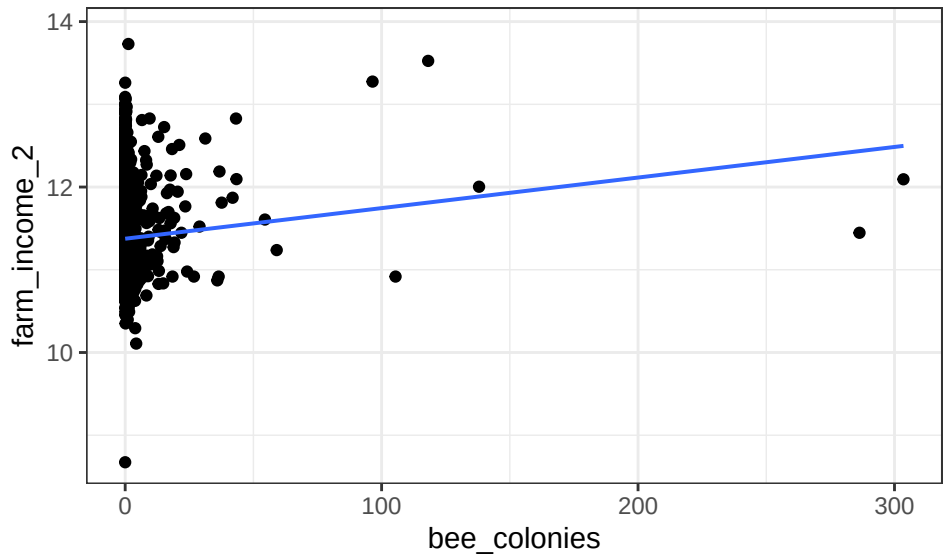

```
favstats(~farm_acres_2, data=df3) |>
  kable(booktabs = TRUE, align = 'c') |>
  kable_styling(position = "center")
```

```
#looks better, but still pretty questionable
```

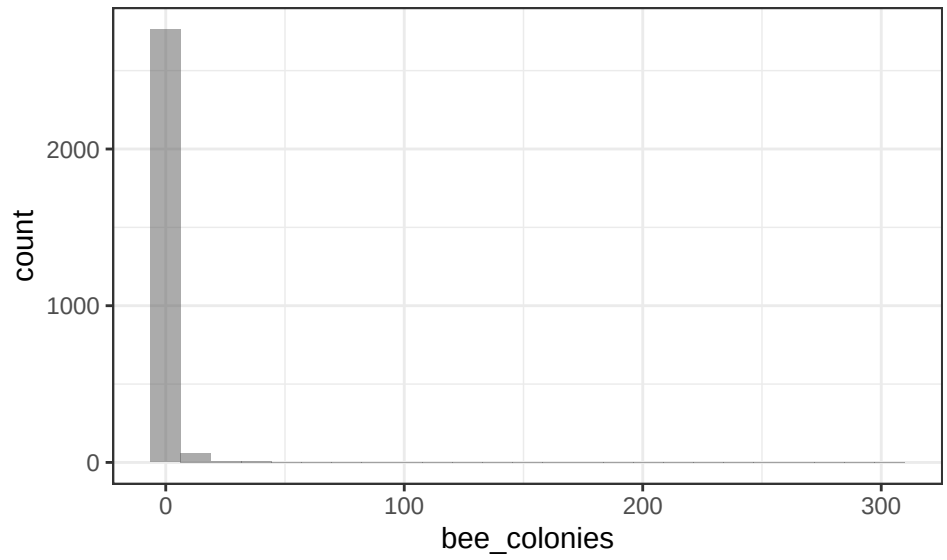
This transformation improved it somewhat, the histogram is less right skewed, but the relationship does not look very linear based on the scatter plot, the mean is 20.833 and the median is 16.353, and the IQR is 10.446

min	Q1	median	Q3	max	mean	sd	n	missing
0	0.04734	0.16083	0.40523	303.5	1.1837	9.5011	2849	226

```
gf_point(data = df3, farm_income_2 ~ bee_colonies) |>
  gf_lm()
```




```
gf_histogram(data = df3, ~ bee_colonies)
```

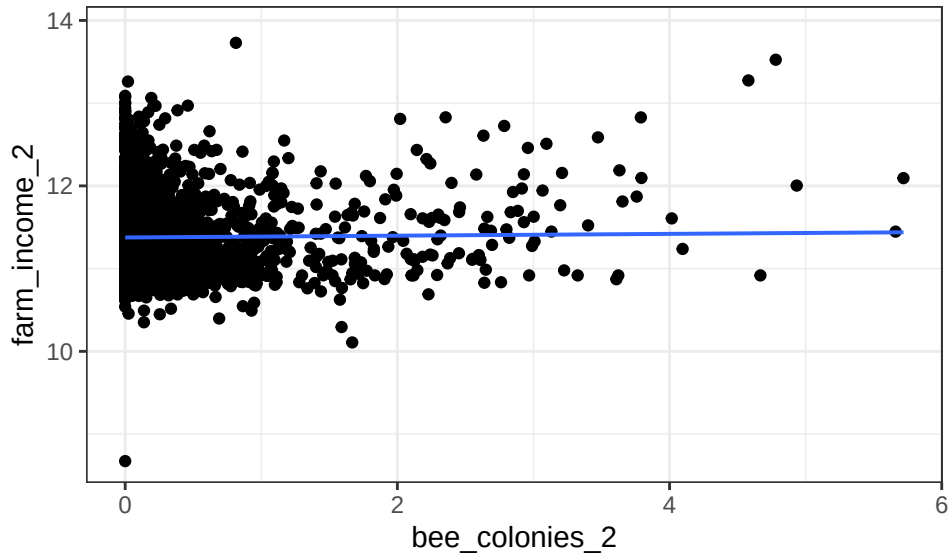



```
favstats(~bee_colonies, data=df3)|>
  kable(booktabs = TRUE, align = 'c') |>
  kable_styling(position = "center")
```

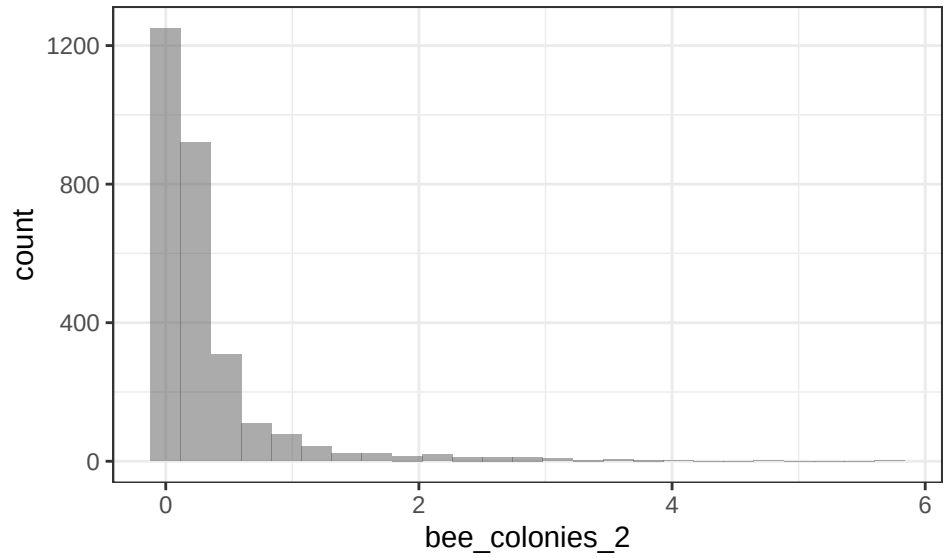
```
#bee colonies is super right skewed and not linear w/ farm income, so trying a transformation
#again, much improved
```

Very strongly right skewed histogram, scatterplot is not useful for assessing linearity, the mean is 1.1837, the median is 0.16083, and the IQR is 0.35789

```
df3 <- df3 |>
  mutate(bee_colonies_2 = log(bee_colonies+1))

gf_point(data = df3, farm_income_2 ~ bee_colonies_2) |>
  gf_lm()
```




```
gf_histogram(data = df3, ~ bee_colonies_2)
```

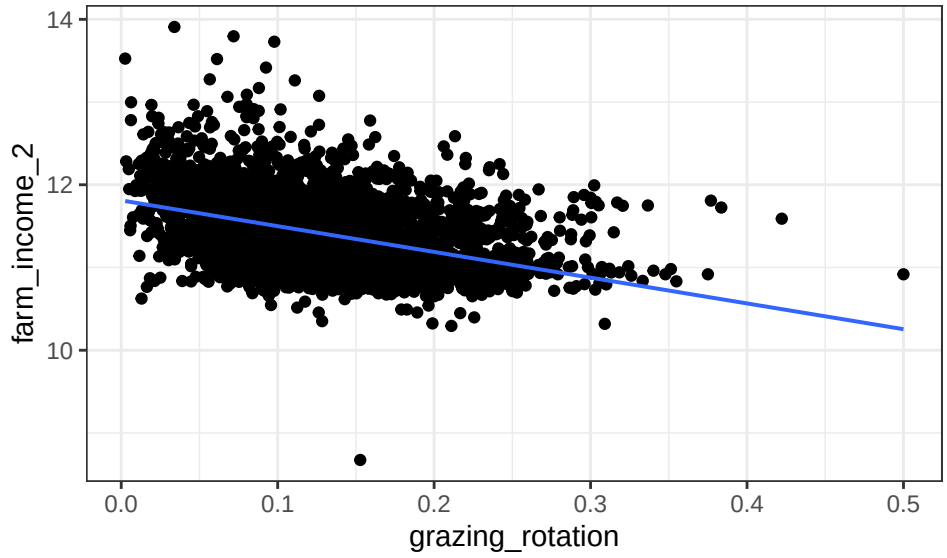


min	Q1	median	Q3	max	mean	sd	n	missing
0	0.04625	0.14913	0.3402	5.7187	0.32371	0.56401	2849	226

```
favstats(~bee_colonies_2, data=df3)|>
  kable(booktabs = TRUE, align = 'c') |>
  kable_styling(position = "center")
```

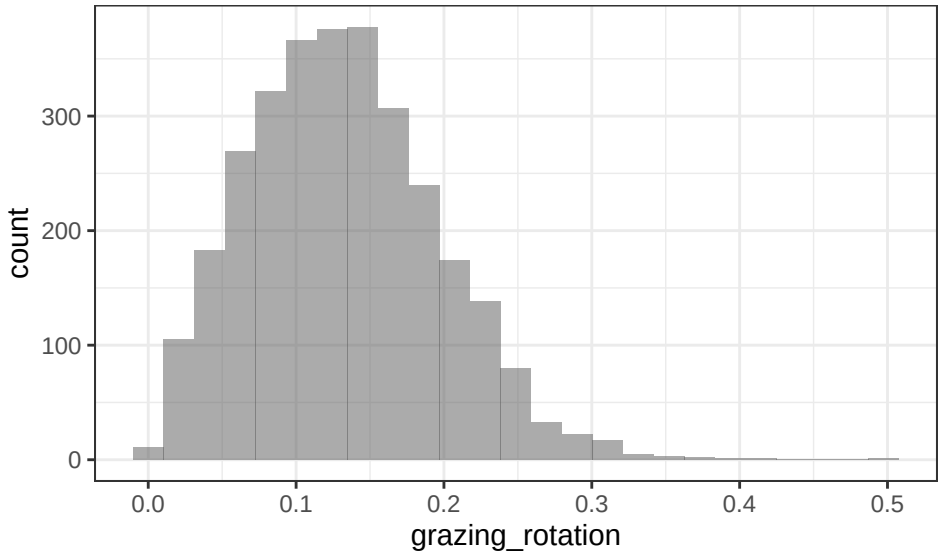
After transforming the data, the histogram looks somewhat better, the main part of the distribution is very normal, but we have introduced a large number of 0s into the dataset, which represent real values, so probably shouldn't be removed. the mean is 4.1501 the median is 4.5433 and the IQR is 2.4136. The scatter plot similarlly seems to show a fairly linear relationship aside from a large number of 0 values.

```
gf_point(data = df3, farm_income_2 ~ grazing_rotation) |>
  gf_lm()
```




```
gf_histogram(data = df3, ~ grazing_rotation)
```

min	Q1	median	Q3	max	mean	sd	n	missing
0.00253	0.08572	0.12806	0.17316	0.5	0.13218	0.06355	3034	41



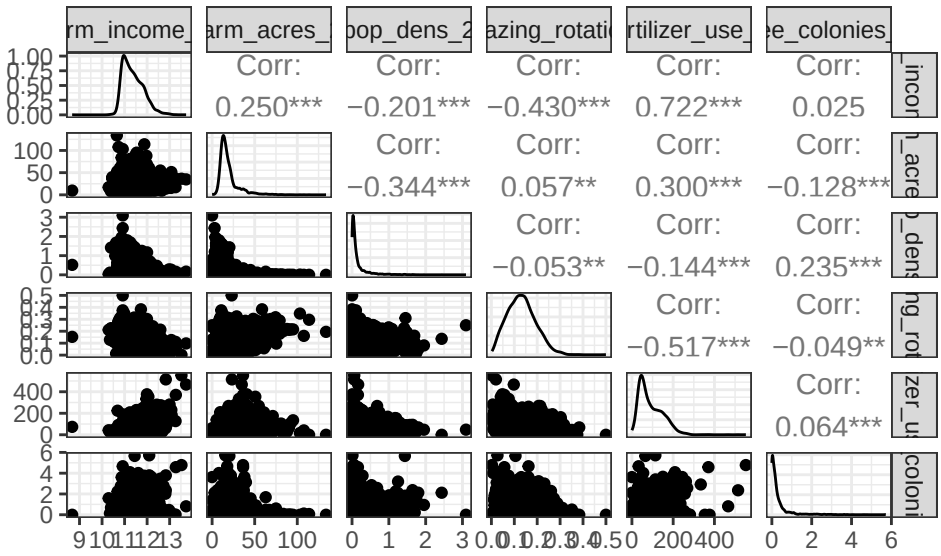

```
favstats(~grazing_rotation, data=df3)|>
  kable(booktabs = TRUE, align = 'c') |>
  kable_styling(position = "center")
```

```
#looks pretty good actually
```

This distribution looks pretty good. Relatively normal histogram, and the scatter plot shows a relatively linear relationship. mean is 0.13218, median is 0.12806 and IQR is 0.08744

After we have tried to improve all of the response variables, lets take a look at the ggpairs again

```
df4 <- df3 |>
  dplyr::select(c(farm_income_2, farm_acres_2, pop_dens_2, grazing_rotation, fertilizer_use_2, bee_colo
df4 <- na.omit(df4)
df3<-na.omit(df3)
ggpairs(df4, columns = 1:6, progress = FALSE)
```



These generally look much better, we still have a lot of 0 values in the population density, but they are real, so the data is not actually normally distributed. The scatter plots generally look much more linear than before transforming the variables, and the correlations are similar to before transforming. Bee colonies is no longer correlated though I removed missing values to make my following tests and operations easier.

Model selection

```
best.sub <- regsubsets(farm_income_2 ~ farm_acres_2 + pop_dens_2 + grazing_rotation + fertilizer_use_2 +
names(msummary(best.sub))
```

```
## [1] "which"  "rsq"    "rss"    "adjr2"  "cp"     "bic"    "outmat" "obj"
```

```
with(msummary(best.sub), data.frame(adjr2, cp, bic, rss, outmat)) %>%
  kable(digits = 3, booktabs = TRUE, align = 'c') %>%
  row_spec(row = 0, angle = 90)
```

	adjr2	cp	bic	rss	farm_acres_2	pop_dens_2	grazing_rotation	fertilizer_use_2	bee_colonies_2
1 (1)	0.521	100.154	-2053.1	283.48				*	
2 (1)	0.530	44.510	-2101.5	277.86		*		*	
3 (1)	0.536	5.694	-2134.1	273.89		*	*	*	
4 (1)	0.537	4.283	-2129.6	273.56	*	*	*	*	
5 (1)	0.537	6.000	-2121.9	273.53	*	*	*	*	*

```
mod.small <- lm(farm_income_2 ~ 1, data = df4)
mod.all <- lm(farm_income_2 ~ farm_acres_2 + pop_dens_2 + grazing_rotation + fertilizer_use_2 + bee_colo
stepAIC(mod.small, scope = list(lower = mod.small, upper = mod.all),
        direction = "both", trace = FALSE)$anova
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## farm_income_2 ~ 1
##
## Final Model:
## farm_income_2 ~ fertilizer_use_2 + pop_dens_2 + grazing_rotation +
##     farm_acres_2
##
##
##                 Step Df  Deviance Resid. Df Resid. Dev     AIC
## 1                                      2812     591.50 -4384.4
## 2 + fertilizer_use_2  1 308.01801      2811     283.48 -6451.4
## 3       + pop_dens_2  1   5.61712      2810     277.86 -6505.7
## 4 + grazing_rotation  1   3.97730      2809     273.89 -6544.3
## 5     + farm_acres_2  1   0.33237      2808     273.55 -6545.7
```

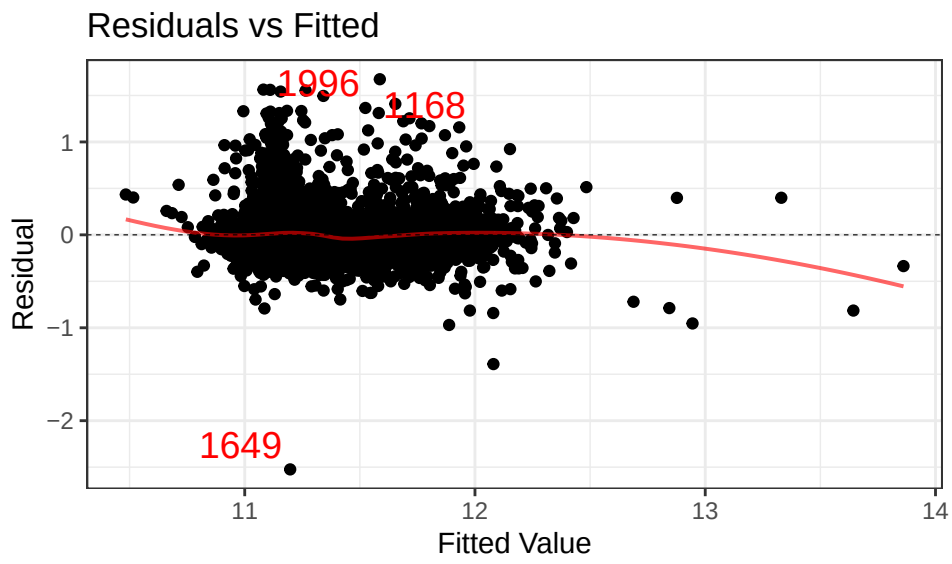
```
stepAIC(mod.all, direction = "backward", trace = FALSE)$anova
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## farm_income_2 ~ farm_acres_2 + pop_dens_2 + grazing_rotation +
##     fertilizer_use_2 + bee_colonies_2
##
## Final Model:
## farm_income_2 ~ farm_acres_2 + pop_dens_2 + grazing_rotation +
##     fertilizer_use_2
##
##
##               Step Df Deviance Resid. Df Resid. Dev     AIC
## 1                                   2807     273.53 -6544.0
## 2 - bee_colonies_2  1 0.027553      2808     273.55 -6545.7
```

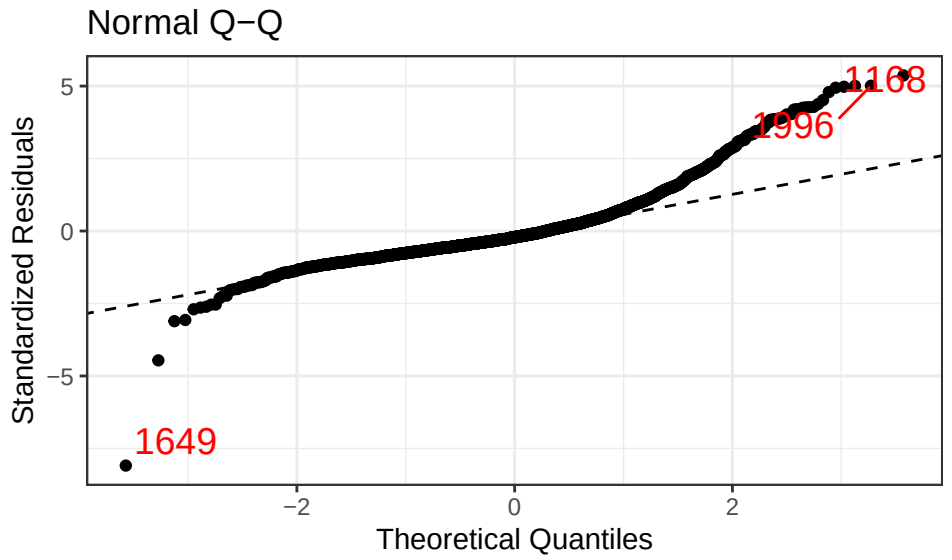
The final model suggested by best subsets regression is the 4-predictor model with pop_dens_2, grazing_rotation, fertilizer_use_2, and farm_acres_2 without bee colonies. This model had the lowest C_p score of 4.2828. This model was also recommended by stepwise regression and backwards elimination, with the lowest AIC score of -6545.7. However, since the 5-predictor model including bee_colonies_2 had similar AIC, and C_p scores, we will manually compare the model strength and significance using the tools we have learned in STAT-230.

```
fmA <- lm(farm_income_2 ~ pop_dens_2 + grazing_rotation + fertilizer_use_2 + farm_acres_2, data = df4)
mplot(fmA, which = 1)
```

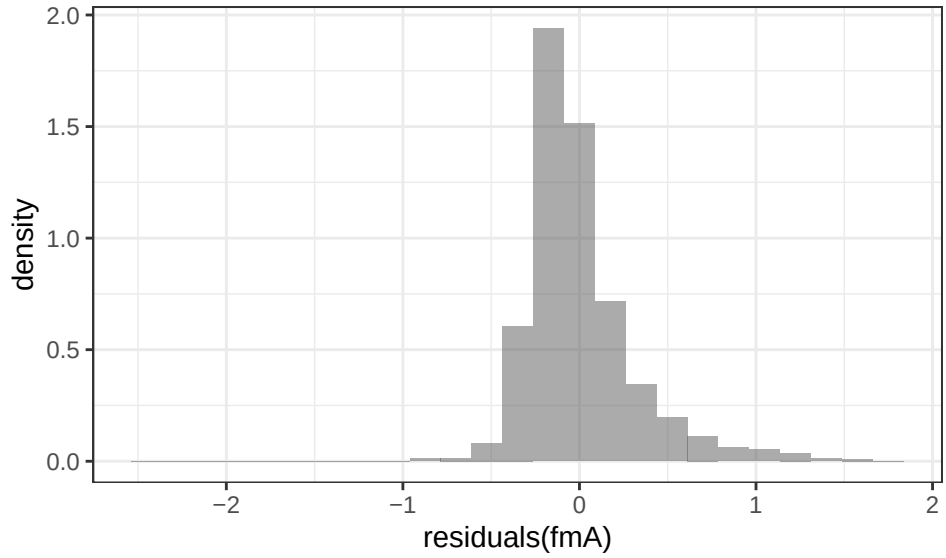
```
## `geom_smooth()` using formula = 'y ~ x'
```




```
mplot(fmA, which = 2)
```

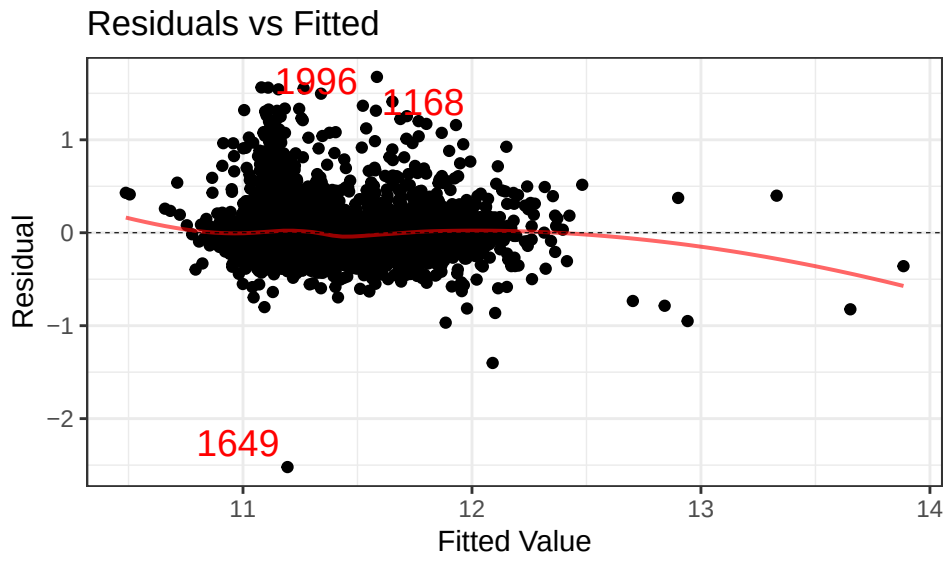



```
gf_dhistogram(~ residuals(fmA))
```

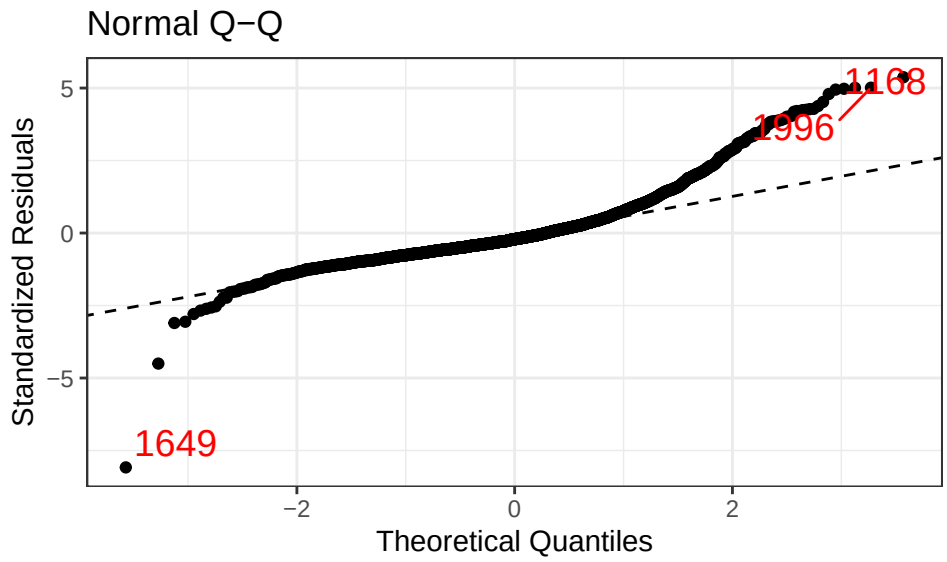



```
fmB <- lm(farm_income_2 ~ farm_acres_2 + pop_dens_2 + grazing_rotation + fertilizer_use_2 + bee_colonies
mplot(fmB, which = 1)
```

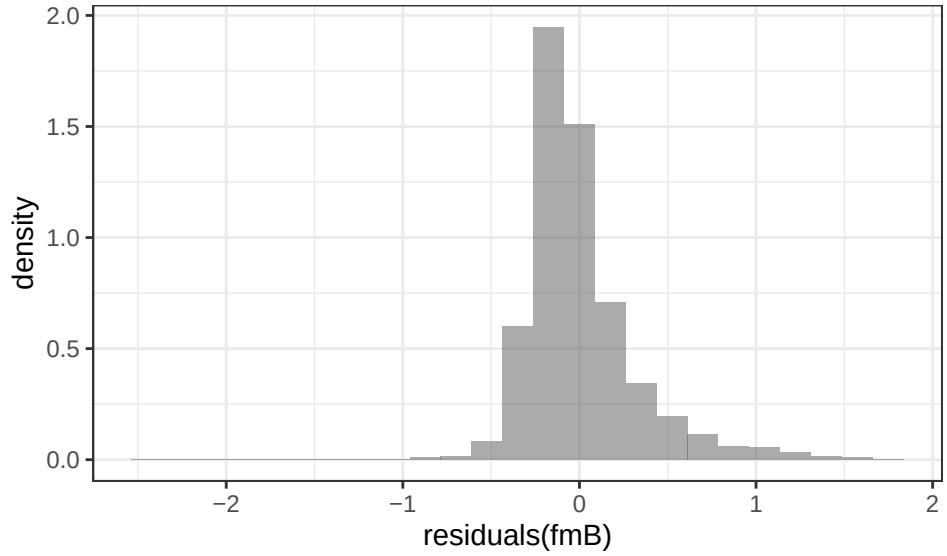
```
## `geom_smooth()` using formula = 'y ~ x'
```




```
mplot(fmB, which = 2)
```




```
gf_dhistogram(~ residuals(fmB))
```



The residuals vs fitted plot, QQ plot, and histogram of the residuals of both models are practically identical and will not serve as a point of difference. The Linearity condition is mostly met as there is mostly symmetry over the line residual=0 on the residuals vs fitted values plot, and there does not seem to be major bending or curving of points. The Equal Variance condition is questionable as there seems to be funneling in of the spread of points on the residual vs fitted value plot as you move to higher fitted values. The Normality condition is also questionable as, though the QQ-plot is mostly a straight line with mostly no bending, there does seem to be a very large and bent tail at the high end. Furthermore, the histogram of the residuals, while it is unimodal and centered close to 0 (a little below 0 actually), there does seem to be a significant right skew. We will assume independence and randomness and proceed. Finally, there appear to be several outliers on both models.

Compare conditions and fit to untransformed model

Conditions are definitely worse in the untransformed model. The data is even more skewed, with many extreme residual values that disrupt normality and equal variance. Therefore, it looks like our transformations were helpful, even if they didn't completely fix the data to fit modeling conditions.

```
msummary(fmA)
```

```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      11.023637   0.024966  441.55  < 2e-16 ***
## pop_dens_2       -0.179728   0.023776   -7.56  5.5e-14 ***
## grazing_rotation -0.772752   0.116352   -6.64  3.7e-11 ***
## fertilizer_use_2  0.005092   0.000129   39.57  < 2e-16 ***
## farm_acres_2      0.001030   0.000558    1.85    0.065 .
##
## Residual standard error: 0.312 on 2808 degrees of freedom
## Multiple R-squared:  0.538,  Adjusted R-squared:  0.537
## F-statistic:  816 on 4 and 2808 DF,  p-value: <2e-16
```

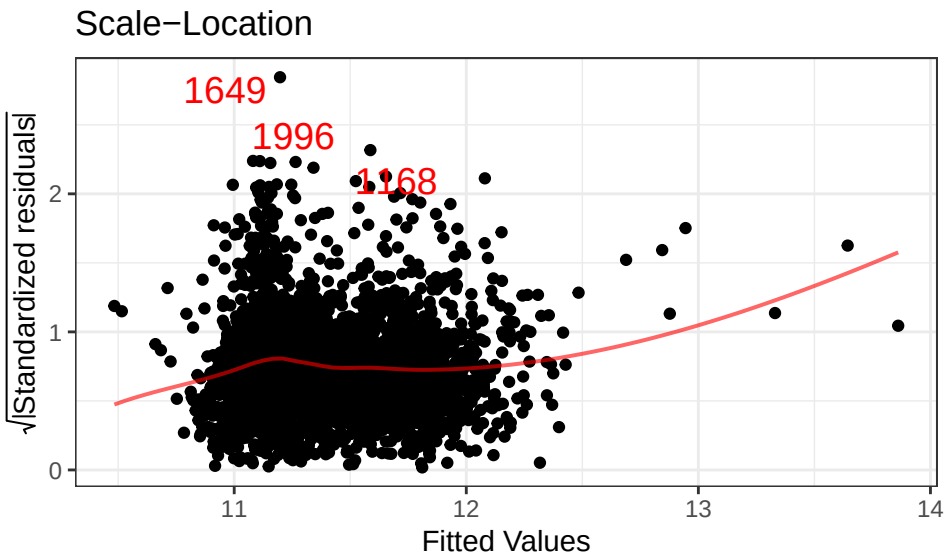
```
msummary(fmA2)
```

```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.16e+04   2.50e+03   12.64  < 2e-16 ***
## pop_dens         -5.19e+03   1.23e+03   -4.22  2.6e-05 ***
## grazing_rotation -1.12e+05   1.51e+04   -7.45  1.2e-13 ***
## fertilizer_use    2.28e+00   5.82e-02   39.23  < 2e-16 ***
## farm_acres        1.83e+00   9.06e-01    2.02    0.044 *
##
## Residual standard error: 43100 on 2808 degrees of freedom
## Multiple R-squared:  0.47,   Adjusted R-squared:  0.469
## F-statistic:  621 on 4 and 2808 DF,  p-value: <2e-16
```

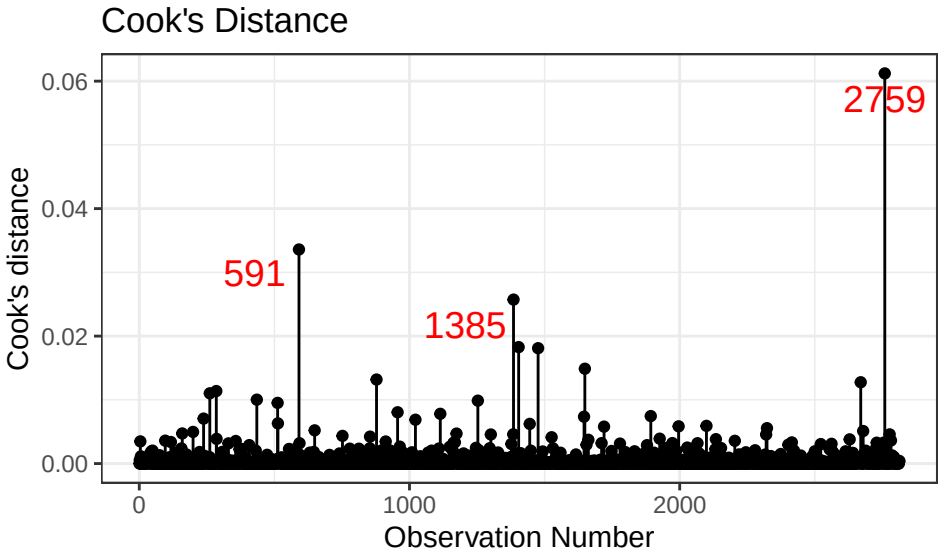
The transformed model also has a much higher R^2 adjusted, higher by 6.8%. This verifies that our transformations were helpful for improving model strength.

```
mplot(fmA, which = 3)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

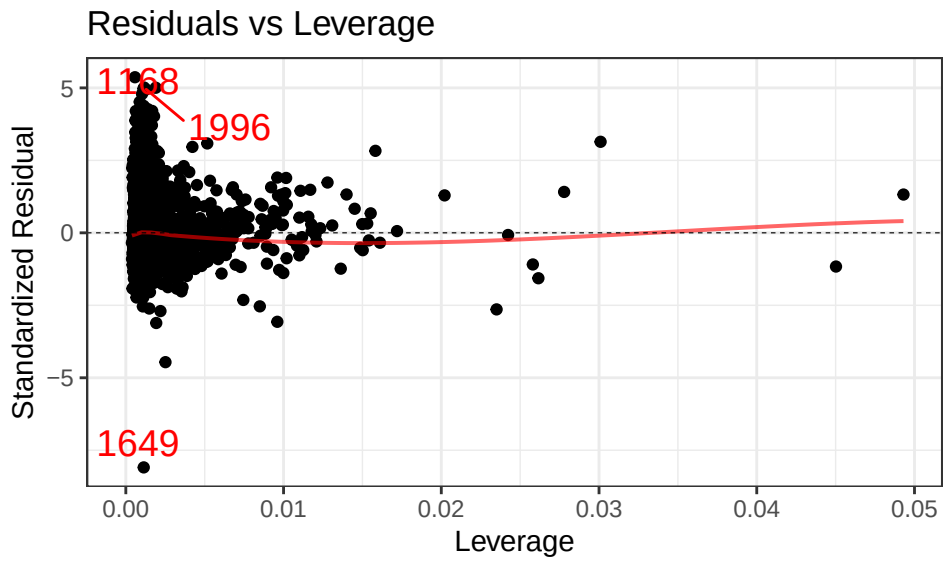



```
mplot(fmA, which = 4)
```



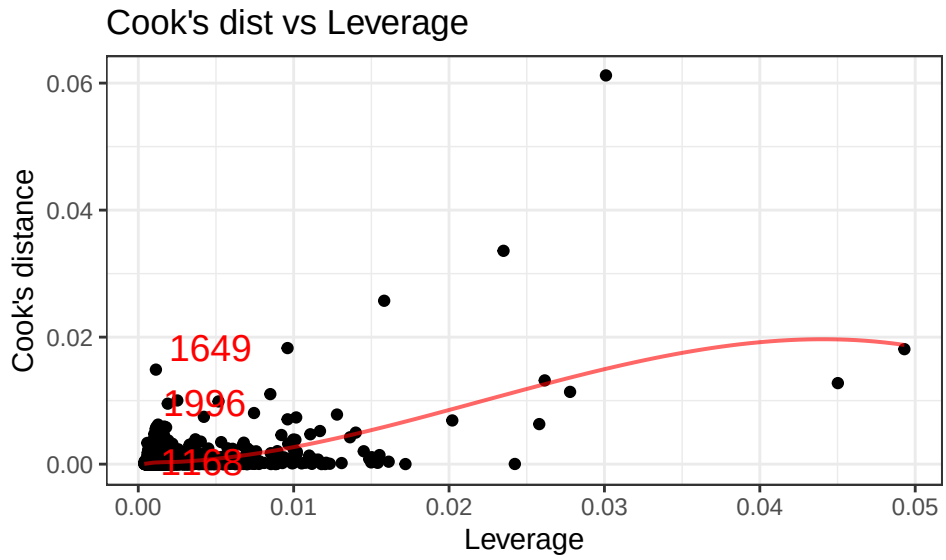

```
mplot(fmA, which = 5)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
mplot(fmA, which = 6)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
fmA_leverage <- hatvalues(fmA)
high_fmA_leverage <- which(fmA_leverage > 0.0035549)
length(high_fmA_leverage)
```

```
## [1] 236
```

```
problematic_fmA_leverage <- which(fmA_leverage >0.0053324)
length(problematic_fmA_leverage)
```

```
## [1] 136
```

```
fmA_stdresiduals <- rstandard(fmA)
bad_fmA_outlier <- which (abs(fmA_stdresiduals) > 2)
length(bad_fmA_outlier)
```

```
## [1] 153
```

```
really_bad_fmA_outlier <- which (abs(fmA_stdresiduals) > 3)
length(really_bad_fmA_outlier)
```

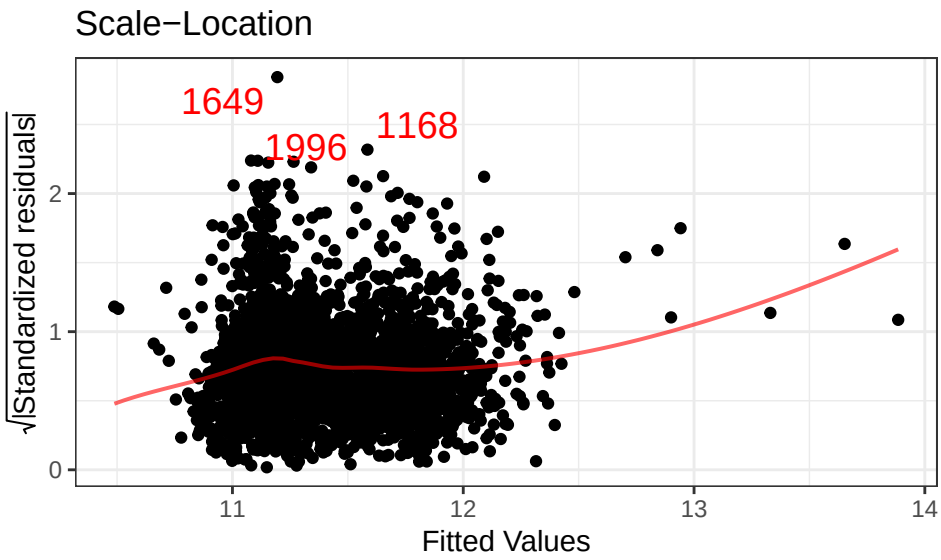
```
## [1] 62
```

```
fmA_cdistance <- cooks.distance(fmA)
influential_fmA <- which(fmA_cdistance > 0.5)
length(influential_fmA)
```

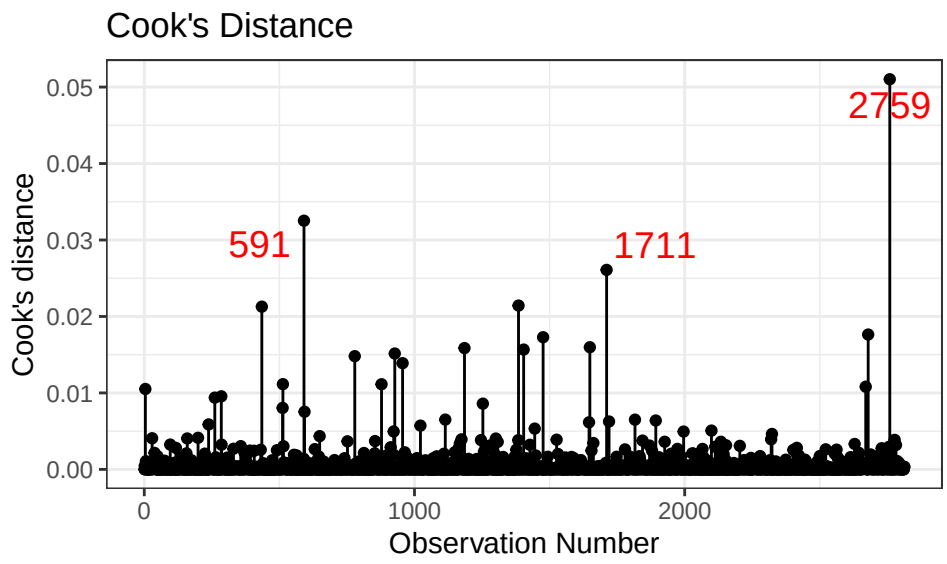
```
## [1] 0
```

```
mplot(fmB, which = 3)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

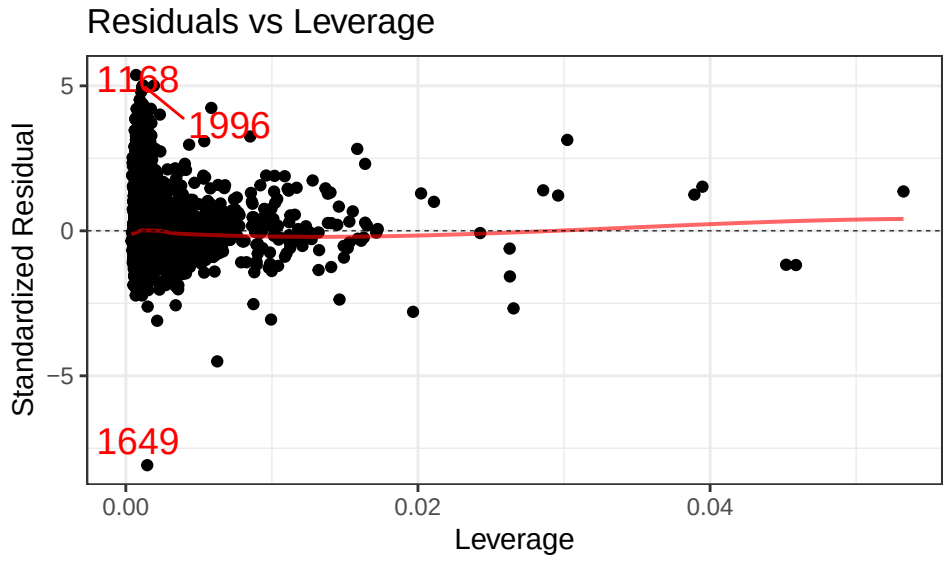



```
mplot(fmB, which = 4)
```



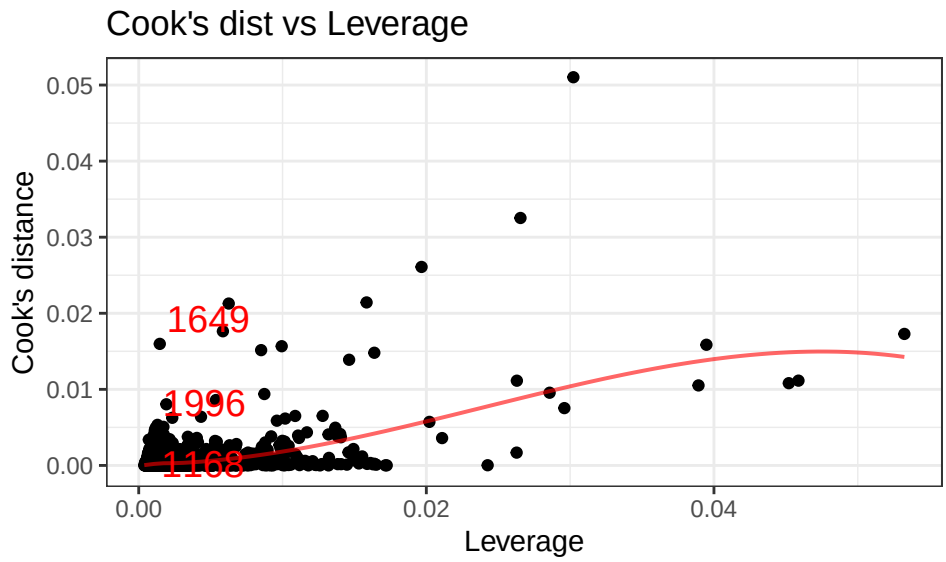

```
mplot(fmB, which = 5)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
mplot(fmB, which = 6)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
fmB_leverage<- hatvalues(fmB)
high_fmB_leverage <- which(fmB_leverage > 0.0035549)
length(high_fmB_leverage)
```

```
## [1] 337
```

```
problematic_fmB_leverage <- which(fmB_leverage > 0.0053324)
length(problematic_fmB_leverage)
```

```
## [1] 207
```

```
fmB_stdresiduals <- rstandard(fmB)
bad_fmB_outlier <- which (abs(fmB_stdresiduals) > 2)
length(bad_fmB_outlier)
```

```
## [1] 153
```

```
really_bad_fmB_outlier <- which (abs(fmB_stdresiduals) > 3)
length(really_bad_fmB_outlier)
```

```
## [1] 62
```

```
fmB_cdistance <- cooks.distance(fmB)
influential_fmB <- which(fmB_cdistance > 0.5)
fmB_cdistance[influential_fmB]
```

```
## named numeric(0)
```

```
dim(df4)
```

```
## [1] 2813    6
```

```
(2*(4+1))/2813
```

```
## [1] 0.0035549
```

```
(3*(4+1))/2813
```

```
## [1] 0.0053324
```

```
207-136
```

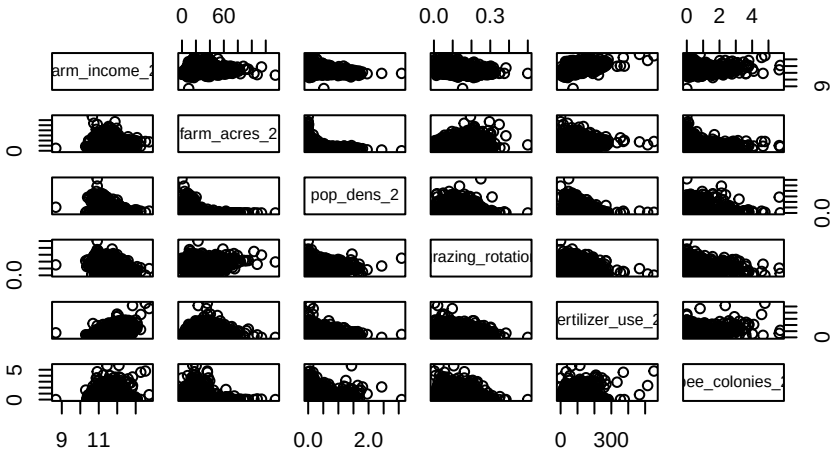
```
## [1] 71
```

fmA: High leverage would be 2(k+1)/n = 2(4+1)/2813 = 0.0035549 and problematic leverage would be 3(k+1)/n = 0.0053324 There are 236 points with high leverages, 136 of which have problematic leverages. There are 91 points with the absolute value of their standardized residuals above 2 but not 3, making them outliers. There are 62 points with an absolute value of their standardized residuals above 3, making them big outliers. There do not appear to be many influential points, as there are no points with Cook's distances above or near 0.5. fmB: High leverage would be 2(k+1)/n = 2(5+1)/2813 = 0.0035549 and problematic leverage would be 3(k+1)/n = 0.0053324. There are 337 points with high leverages, 207 of which have problematic leverages. There are 91 points with the absolute value of their standardized residuals above 2 but not 3, making them outliers. There are 62 points with an absolute value of their standardized residuals above 3, making them big outliers. There do not appear to be many influential points, as there are no points with Cook's distances above or near 0.5. So, fmA seems to have 71 less points with problematic leverages. They have the same number of big outliers, and neither have influential points.

```
cor(df4)
```

```
##                 farm_income_2 farm_acres_2 pop_dens_2 grazing_rotation
## farm_income_2        1.000000     0.250045  -0.200636        -0.430220
## farm_acres_2         0.250045     1.000000  -0.343740         0.056543
## pop_dens_2          -0.200636    -0.343740   1.000000        -0.052888
## grazing_rotation    -0.430220     0.056543  -0.052888         1.000000
## fertilizer_use_2     0.721624     0.299724  -0.144407        -0.517474
## bee_colonies_2       0.024889    -0.128397   0.235198        -0.049328
##                 fertilizer_use_2 bee_colonies_2
## farm_income_2           0.721624      0.024889
## farm_acres_2            0.299724     -0.128397
## pop_dens_2             -0.144407      0.235198
## grazing_rotation       -0.517474     -0.049328
## fertilizer_use_2        1.000000      0.063654
## bee_colonies_2          0.063654      1.000000
```

```
plot(df4)
```




```
vif(fmA)
```

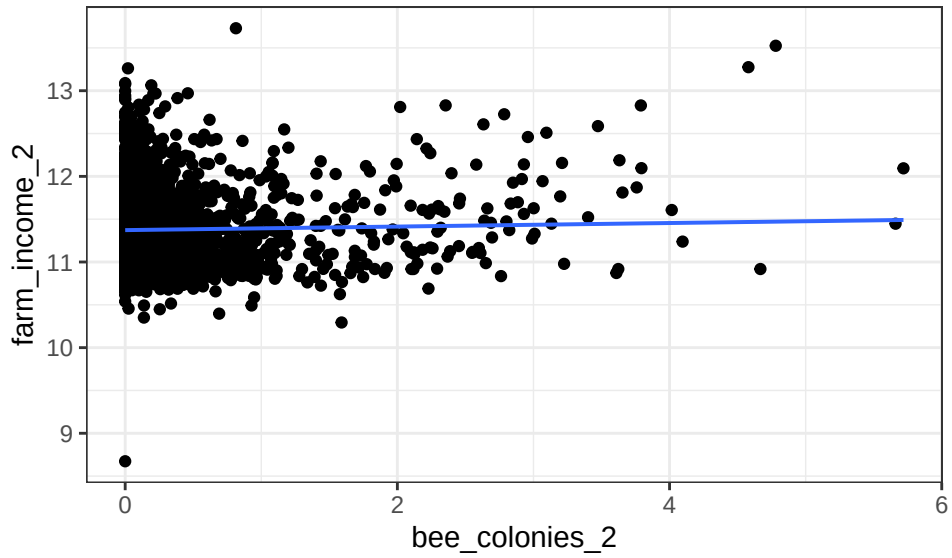
```
##       pop_dens_2 grazing_rotation fertilizer_use_2     farm_acres_2
##           1.1428           1.4723           1.6138           1.2880
```

```
vif(fmB)
```

```
##     farm_acres_2       pop_dens_2 grazing_rotation fertilizer_use_2
##           1.2984           1.1975           1.4745           1.6378
##   bee_colonies_2
##           1.0786
```

Thankfully, for both models, none of the VIF values are above 5 which would strongly indicate multicollinearity. The inclusion of bee_colonies_2 in fmB has little impact on VIF. The correlation matrix for the bee_colonies_2 pairs isn't too concerning with correlation coefficients of 0.23 or lower.

```
gf_point(farm_income_2 ~ bee_colonies_2, data = df4) %>% gf_lm()
```



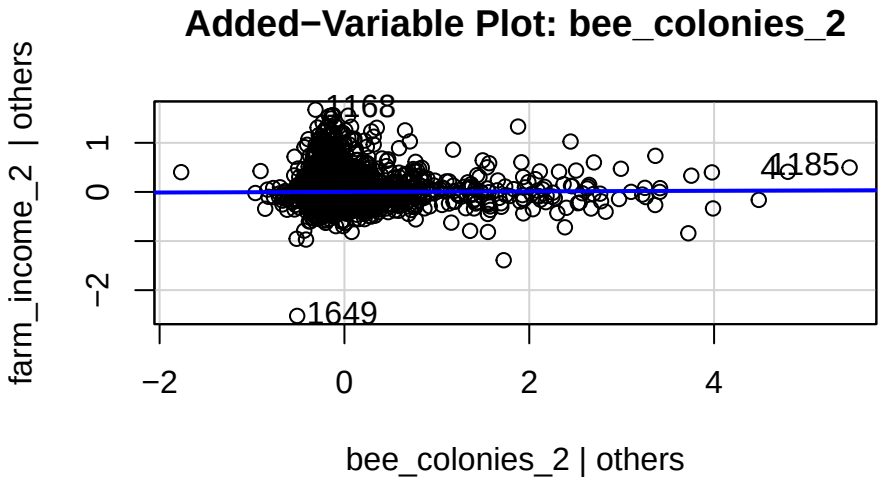

```
cor(farm_income_2 ~ bee_colonies_2, data = df4)
```

```
## [1] 0.024889
```

```
fm_bee_colonies_2 <- lm(farm_income_2 ~ bee_colonies_2, data = df4)
msummary(fm_bee_colonies_2)
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    11.37228    0.00998 1139.36   <2e-16 ***
## bee_colonies_2  0.02070    0.01568    1.32     0.19
##
## Residual standard error: 0.459 on 2811 degrees of freedom
## Multiple R-squared:  0.000619,   Adjusted R-squared:  0.000264
## F-statistic: 1.74 on 1 and 2811 DF,  p-value: 0.187
```

```
avPlot(fmB, "bee_colonies_2")
```



> Looking at a simple

linear regression model of the relationship between farm_income_2 and bee_colonies_2, I see a weak, positive, unclearly linear relationship between farm_income_2 and farm_acres_2 with several outliers and a correlation coefficient of 0.024889. The AV plot of the additional predictor of fmB, bee_colonies_2, shows a horizontal band that is curvilinear. This indicates that bee_colonies_2 contains very little additional information about farm_income_2, beyond that from the other predictors. Both of these indicate that bee_colonies_2 may not be a significant or strong predictor for farm_income_2.

```
anova(fmA, fmB)
```

```
## Analysis of Variance Table
##
## Model 1: farm_income_2 ~ pop_dens_2 + grazing_rotation + fertilizer_use_2 +
##     farm_acres_2
## Model 2: farm_income_2 ~ farm_acres_2 + pop_dens_2 + grazing_rotation +
##     fertilizer_use_2 + bee_colonies_2
##   Res.Df RSS Df Sum of Sq    F Pr(>F)
## 1   2808 274
## 2   2807 274  1    0.0276 0.28   0.59
```

$H_0 : \beta_{farm,acres,2} = 0$ $H_A : \beta_{farm,acres,2} \neq 0$

> Since the p-value of 0.59 is large, we fail to reject the null hypothesis and do not have enough evidence to conclude the alternative hypothesis that the predictor bee_colonies_2 is significant.

```
coef.actual<-fmB$coefficients["bee_colonies_2"]

#bootstrap
set.seed(230) #get same result every time you knit file
bootstrap <- do(1000)*lm(farm_income_2 ~ farm_acres_2 + pop_dens_2 + grazing_rotation + fertilizer_use_2
names(bootstrap)
```

```
##  [1] "Intercept"        "farm_acres_2"     "pop_dens_2"       "grazing_rotation"
##  [5] "fertilizer_use_2" "bee_colonies_2"   "sigma"            "r.squared"
##  [9] "F"                "numdf"            "dendf"            ".row"
## [13] ".index"
```

```
std.bts <- ((bootstrap$bee_colonies_2-mean(bootstrap$bee_colonies_2))/sd(bootstrap$bee_colonies_2))

qtU <- qdata(std.bts, p = 0.975)
qtU
```

```
##  97.5%
## 2.0036
```

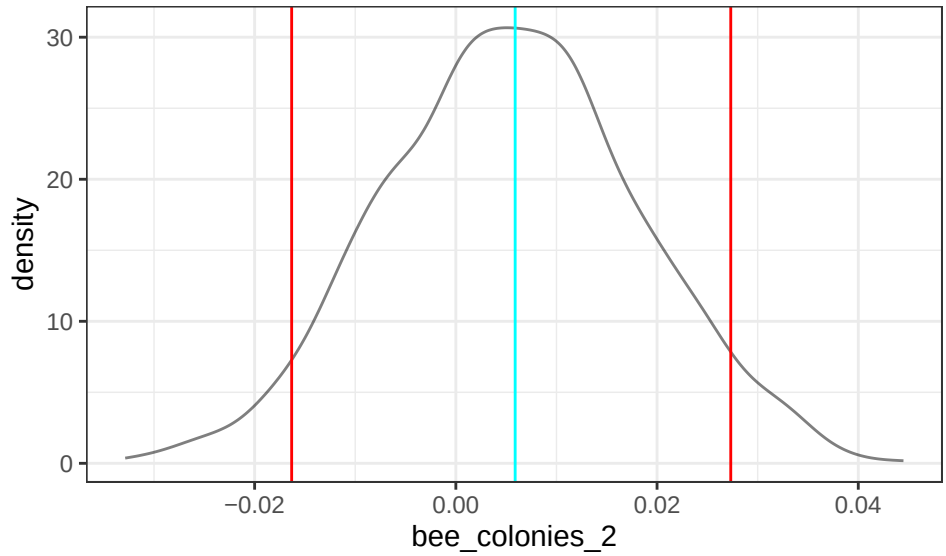
```
qtL <- qdata(std.bts, p = 0.025)
qtL
```

```
##    2.5%
## -1.9326
```

```r
SE <- msummary(fmB)$coefficients["bee_colonies_2","Std. Error"]
c(coef.actual - qtU*SE, coef.actual - qtL*SE)
```

```
## bee_colonies_2 bee_colonies_2
##      -0.016319       0.027324
```

```r
gf_dens(~ bee_colonies_2, data = bootstrap) %>%
  gf_vline(xintercept = coef.actual, color = "cyan") |>
  gf_vline(xintercept = c(coef.actual - qtU*SE, coef.actual - qtL*SE), color = "red")
```



>Since the normality condition was not met for either model, we will test the significance of bee_colonies_2 using bootstrapping. The resulting 95% confidence interval for the coefficient of farm_acres_2 is (-0.016319, 0.027324). Since this interval includes 0,we are not 95% confident that the bee_colonies_2 is not 0, and conclude that this predictor is not significant.

Final model comparison

```r
msummary(fmA)
```

```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      11.023637   0.024966  441.55  < 2e-16 ***
## pop_dens_2       -0.179728   0.023776   -7.56  5.5e-14 ***
## grazing_rotation -0.772752   0.116352   -6.64  3.7e-11 ***
## fertilizer_use_2  0.005092   0.000129   39.57  < 2e-16 ***
## farm_acres_2      0.001030   0.000558    1.85    0.065 .
##
## Residual standard error: 0.312 on 2808 degrees of freedom
## Multiple R-squared:  0.538,  Adjusted R-squared:  0.537
## F-statistic:  816 on 4 and 2808 DF,  p-value: <2e-16
```

```r
msummary(fmB)
```

```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      11.02278    0.02502  440.55  < 2e-16 ***
## farm_acres_2      0.00106    0.00056    1.89    0.059 .
## pop_dens_2       -0.18250    0.02434   -7.50  8.7e-14 ***
## grazing_rotation -0.77515    0.11646   -6.66  3.4e-11 ***
## fertilizer_use_2  0.00508    0.00013   39.21  < 2e-16 ***
## bee_colonies_2    0.00590    0.01109    0.53    0.595
##
## Residual standard error: 0.312 on 2807 degrees of freedom
## Multiple R-squared:  0.538,  Adjusted R-squared:  0.537
## F-statistic:  653 on 5 and 2807 DF,  p-value: <2e-16
```

Conclude that we should use fmA: in addition to all the reasons above (bee colonies not significant in bootstrap test, no clear relationship with farm income in AV plot, insig nested f test, fewer points with high leverage, similar conditions), fmB, the model with bee colonies has the same adjusted R squared as fmA, the model without it.

##Outlier removal > One outlier, 1649, stands out in our equal variance plot and has the largest standard residual value in fmA. Let's try removing it and reassessing conditions.

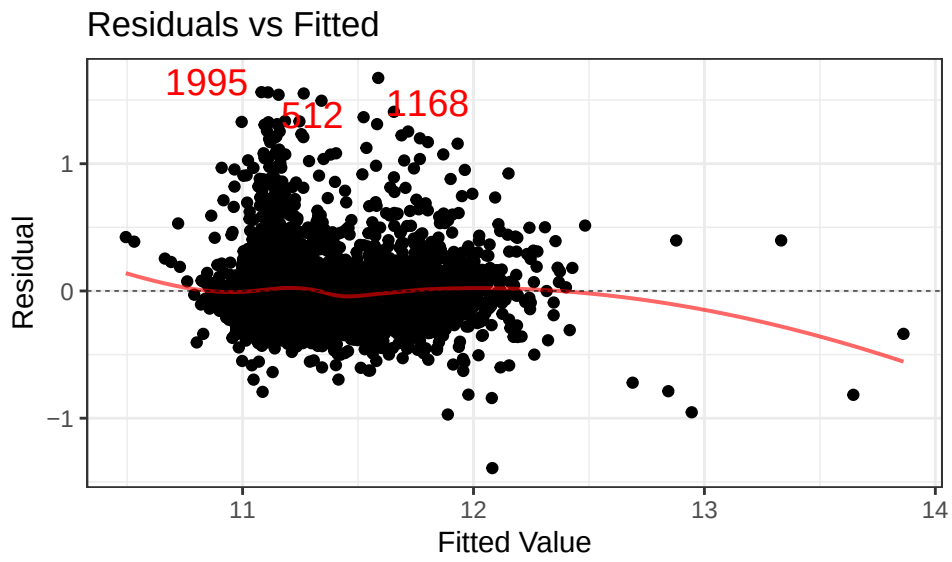
```
stresid <- rstandard(fmA)
which.max(abs(stresid))
```

```
## 1649
## 1649
```

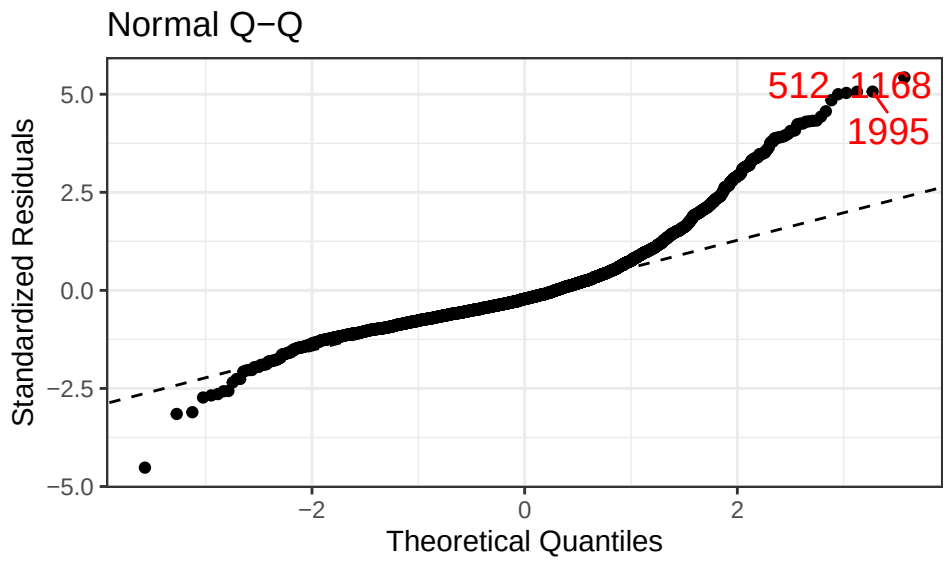
```
df5 <- df4[-which.max(abs(stresid)),]
```

```
fmA3 <- lm(farm_income_2 ~ pop_dens_2 + grazing_rotation + fertilizer_use_2 + farm_acres_2, data = df5)
mplot(fmA3, which = 1)
```

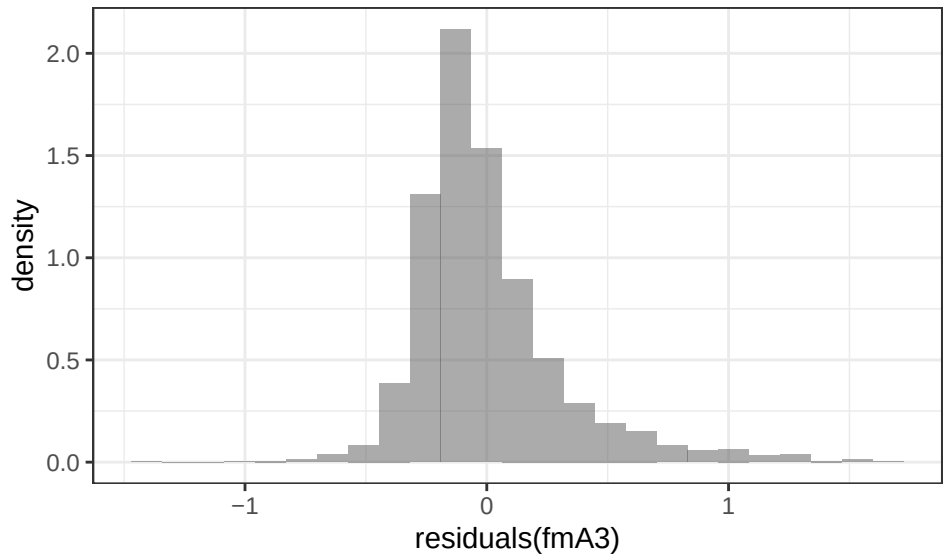
```
## `geom_smooth()` using formula = 'y ~ x'
```



```
mplot(fmA3, which = 2)
```




```
gf_dhistogram(~ residuals(fmA3))
```



>Normality and equal variance look better, although still very questionable

```
msummary(fmA)
```

```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      11.023637   0.024966  441.55  < 2e-16 ***
## pop_dens_2       -0.179728   0.023776   -7.56  5.5e-14 ***
## grazing_rotation -0.772752   0.116352   -6.64  3.7e-11 ***
## fertilizer_use_2  0.005092   0.000129   39.57  < 2e-16 ***
## farm_acres_2      0.001030   0.000558    1.85    0.065 .
##
```

```
## Residual standard error: 0.312 on 2808 degrees of freedom
## Multiple R-squared:  0.538,  Adjusted R-squared:  0.537
## F-statistic:  816 on 4 and 2808 DF,  p-value: <2e-16
```

```
msummary(fmA3)
```

```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      11.022920   0.024677  446.68  < 2e-16 ***
## pop_dens_2       -0.175508   0.023507   -7.47  1.1e-13 ***
## grazing_rotation -0.763733   0.115014   -6.64  3.7e-11 ***
## fertilizer_use_2  0.005098   0.000127   40.07  < 2e-16 ***
## farm_acres_2      0.000991   0.000551    1.80    0.072 .
##
## Residual standard error: 0.309 on 2807 degrees of freedom
## Multiple R-squared:  0.543,  Adjusted R-squared:  0.542
## F-statistic:  833 on 4 and 2807 DF,  p-value: <2e-16
```

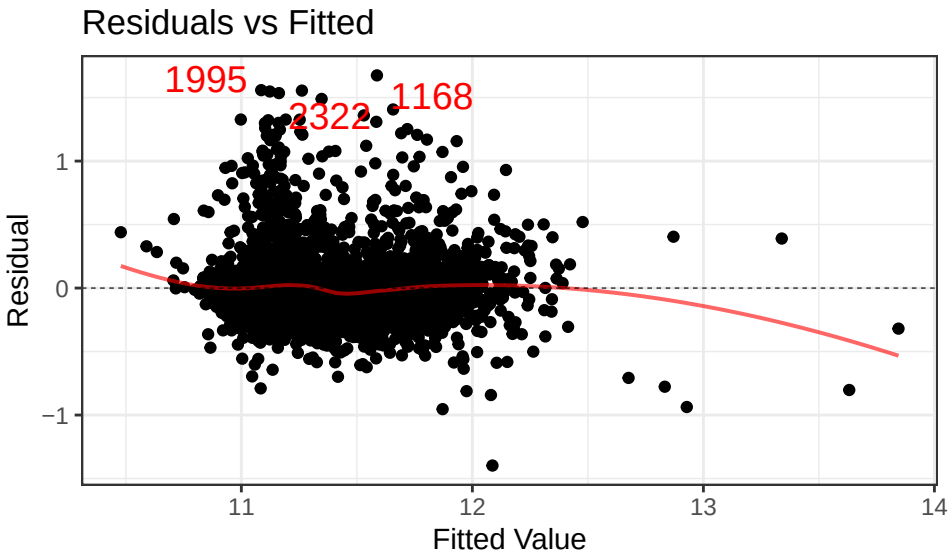
adjusted R^2 increases by 0.5%

Conclusion: we should remove this outlier

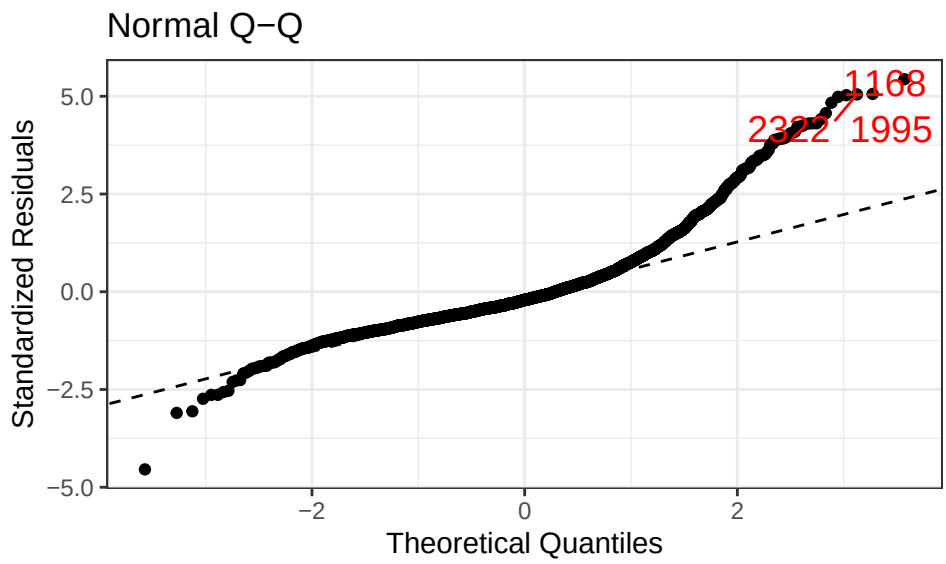
Interactions

```
fmC <- lm(farm_income_2 ~ pop_dens_2 + grazing_rotation + fertilizer_use_2 + farm_acres_2 + pop_dens_2:g
mplot(fmC, which = 1)
```

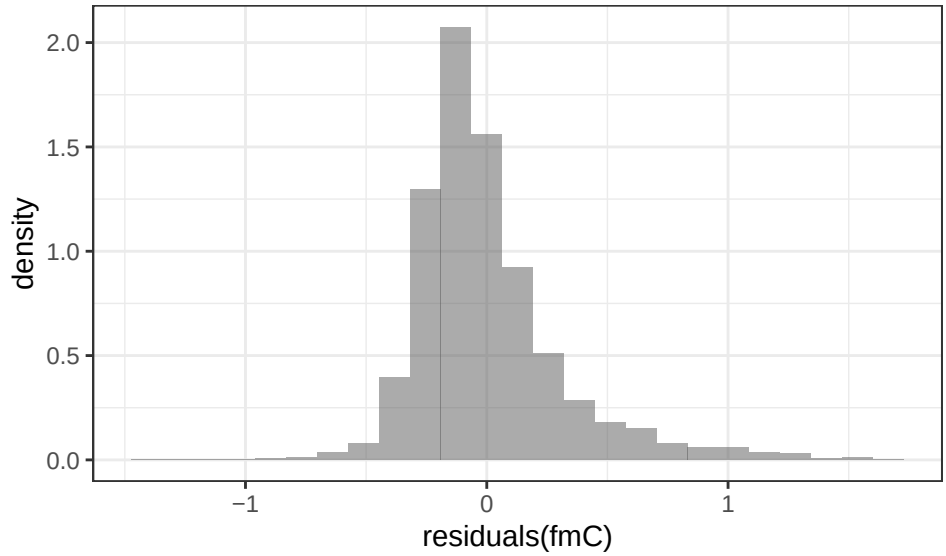
```
## `geom_smooth()` using formula = 'y ~ x'
```




```
mplot(fmC, which = 2)
```




```
gf_dhistogram(~ residuals(fmC))
```



> Conditions look about the same for the interaction model.

```
msummary(fmA3)
```

```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      11.022920   0.024677  446.68  < 2e-16 ***
## pop_dens_2       -0.175508   0.023507   -7.47  1.1e-13 ***
## grazing_rotation -0.763733   0.115014   -6.64  3.7e-11 ***
## fertilizer_use_2  0.005098   0.000127   40.07  < 2e-16 ***
## farm_acres_2      0.000991   0.000551    1.80    0.072 .
##
## Residual standard error: 0.309 on 2807 degrees of freedom
## Multiple R-squared:  0.543,  Adjusted R-squared:  0.542
## F-statistic:  833 on 4 and 2807 DF,  p-value: <2e-16
```

```
msummary(fmC)
```

```
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 11.049113   0.026524  416.56  < 2e-16 ***
## pop_dens_2                  -0.249320   0.077285   -3.23   0.0013 **
## grazing_rotation            -0.888741   0.129364   -6.87  7.9e-12 ***
## fertilizer_use_2             0.004965   0.000144   34.48  < 2e-16 ***
## farm_acres_2                 0.001302   0.000561    2.32   0.0203 *
## pop_dens_2:grazing_rotation  0.729898   0.381994    1.91   0.0561 .
## pop_dens_2:fertilizer_use_2  0.001493   0.000677    2.21   0.0274 *
## pop_dens_2:farm_acres_2     -0.012375   0.005313   -2.33   0.0199 *
##
## Residual standard error: 0.308 on 2804 degrees of freedom
## Multiple R-squared:  0.544,  Adjusted R-squared:  0.543
## F-statistic:  478 on 7 and 2804 DF,  p-value: <2e-16
```

Ideally, adj R^2 should increase by 2% per additional term, but it only increases by 0.001 when adding three interaction terms. Even though all of these interaction terms are significant, we choose not to include them because they do not notably improve model conditions or model strength, and they will make interpretation much more difficult.

Final model: fmA3!

##Final Conclusion:

Overall, our final model fmA3, indicates that 54.2% of the variability in farm_income_2 can be explained by the multiple linear relationship between farm_income_2 and pop_dens_2, grazing_rotation, fertilizer_use_2, and farm_acres_2.

This model indicates that average farm income in a county increases with increasing population density, proportion of farms that use grazing rotations, average fertilizer use, and average farm size. Specifically, average farm income increases by $0.31 for every 1 person/acre density increase; farm income increases by $0.05 for every 10% increase in the percent of farms that use grazing rotations; farm income increases by $1.00 for every $1 increase in average farm fertilizer expenditure; and farm income increases by $1.00 for every 1 acre increase in average farm size. That is, counties with bigger farms that use more fertilizer, counties with more farms that use grazing rotations, and counties with large population densities are likely to contain more profitable farms.

```
#population density
coef.pop<-fmA3$coefficients["pop_dens_2"]
exp(coef.pop)/exp(1)
```

```
## pop_dens_2
##    0.30866
```

```
#grazing
coef.graze<-fmA3$coefficients["grazing_rotation"]
exp(coef.graze)/1
```

```
## grazing_rotation
##          0.46592
```

```r
#fertilizer
coef.fert<-fmA3$coefficients["fertilizer_use_2"]
exp(coef.fert)/(1^2)
```

```
## fertilizer_use_2
##           1.0051
```

```r
#acreage
coef.acre<-fmA3$coefficients["farm_acres_2"]
exp(coef.acre)/(1^2)
```

```
## farm_acres_2
##        1.001
```

Reverse transformations log(farm income) sqrt(fertilizer) log(pop density) sqrt(farm area)

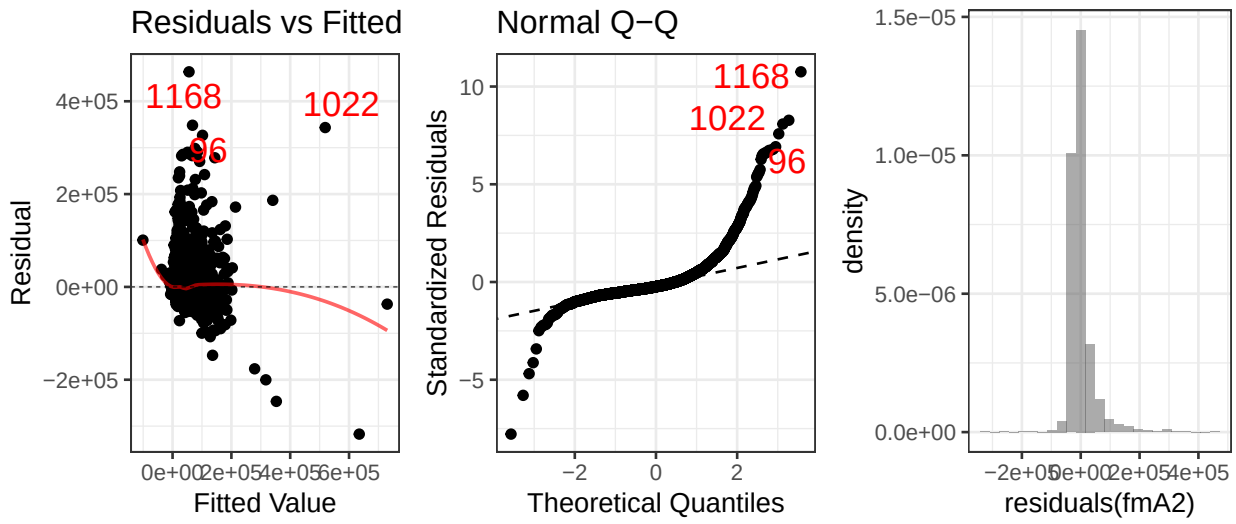
Appendices

Appendix 1. Plots of observed vs theoretical values, and Q-Q plots for the initial three models

```r
Pa1 <- mplot(fmA2, which = 1)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```r
Pa2 <- mplot(fmA2, which = 2)
Pa3 <- gf_dhistogram(~ residuals(fmA2))
grid.arrange(Pa1, Pa2, Pa3, ncol = 3)
```



Appendix 2

```
Pa4 <- mplot(fmA, which = 1)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
Pa5 <- mplot(fmA, which = 2)
Pa6 <- gf_dhistogram(~ residuals(fmA))

Pa7 <- mplot(fmB, which = 1)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
Pa8 <- mplot(fmB, which = 2)
Pa9 <- gf_dhistogram(~ residuals(fmB))

grid.arrange(Pa4, Pa5, Pa6, Pa7, Pa8, Pa9, ncol = 3,  nrow = 2)
```