Influences of farm demography on U.S. agricultural subsidies

Becca Cutforth

2023-12-15

Abstract
U.S. agricultural subsidy policies are contentious and highly influential. In order to better understand how these policies affect farmers, I analyzed how federal subsidies are influenced by various demographic variables. My results indicate that subsidies favor large, high-income farms that grow crops, especially corn. Furthermore, they indicate that subsidies may disfavor female farmers. My analyses indicate that farms in the Midwest receive more subsidies than all other U.S. regions, and that this effect is likely linked to differences in other analyzed farm characteristics.

Introduction
For this project, I chose to study federal agricultural subsidies in the U.S. Subsidies form a critical link between government policy and farm profitability. In order to better understand how subsidies favor or disfavor different types of farms, I asked the research question: How do farm demographic characteristics in a county influence the amount of government subsidies farms in that county receive?

Based on developing research about 'subsidy gaps' between White and Black farmers, I hypothesized that counties with higher proportions of female-owned and minority-owned farms receive lower subsidies (EWG, 2007). I hypothesized that counties with higher proportions of farms with internet access receive higher subsidies, since internet access likely helps farmers access information about subsidy programs. I also hypothesized that counties with higher average farm sizes and higher average farm incomes receive more subsidies, since research has shown that subsidies are biased towards large, conventional farms (Edwards, 2023). I further tested the effects of U.S. region on subsidies, and hypothesized that counties in the Midwest would receive more subsidies than counties in other regions, since the Midwest received much higher total subsidies compared to other US regions between 1995 and 2021 (EWG, 2021). I also expected that subsidies would be targeted towards counties with high proportions of crop-producing farms and especially corn-producing farms (Edwards, 2023).

Methods
My dataset for this project is sourced from the 2017 United States Census of Agriculture. This dataset represents all U.S. states except Alaska. Because it is census data, it is likely highly representative of practicing farmers in the U.S., although it may underrepresent farms that are less willing to respond to or are less accessible to census workers. This dataset is also limited by the fact that information is grouped by county. In order to control for county size in this analysis, I standardized all variables by the number of farm operations in a county, and therefore all my analyses predict outcomes for an average farm in a given county.

My response variable is the average federal subsidy amount in $ that a farm operation receives in a given county ("govsub"). I analyze the categorical variable "region", with regional levels as sourced from U.S. census data: Pacific, Mountain West, Midwest, Northeast, South, and Southeast. I analyze the proportion of crop operations in a county as a categorical variable ("crop_group") alongside region using two-factor ANOVA analysis.

I also predict "govsub" in a multiple linear regression analysis using "region" in addition to 7 quantitative predictors. I analyze the proportion of farm operations in a county that have principal producers that are female or Hispanic ("female_prop" and "hisp_prop"). Unfortunately, principle producer data for other racial minority groups was too sparse to analyze. I also analyze the proportion of operations in a county that have internet access ("internet"), average farm acreage in a county ("farm_acres"), average farm income ("farm_income"), proportion of operations in a county that grow crops ("crop_prop_op"), and the proportion of operations growing crops that grow corn ("corn_prop_crop").

Results

The initial combining, and wrangling of the 2017 ag census data set can be found at https://github.com/evanmacarthur/Stat230Proj1. I selected variables of interest from this dataset and divided county-wide variables by number of operations in a county. I created the 'region' variable from state name data (Fig. A1). Before completing EDA for each analysis, I removed NAs from my variables of interest. This resulted in a dataset with 2306 counties for my linear regression analysis, and 2813 counties for my ANOVA analysis. Before completing ANOVA EDA, I then also split my crop proportion data into three equally sized groups and designated these groups as "high", "medium", or "low" cropping (Fig. A2).

Two-factor ANOVA

EDA

For my untransformed data, the normality and equal variance of different groups looked questionable - the largest standard deviation ratio between groups was 48.614 (Fig. A5). I tried a log transformation, and I used the transformTukey function to find the most normalizing power transformation, which turned out to be a tenth root transformation.

I was inclined to use a log transformation, since it did the best job improving equal variance of groups (Fig. A4) (max:min sd ratio = 3.34 for 10th root and = 2.70 for log). However, the 10th root transformation did improve the overall normality of govsub the most (Fig. A3) ((mean, median, IQR) = (5640, 2420, 6431) -> (2.19, 2.18, 0.509) for 10th root and (7.70, 7.81, 2.34) for log). The 10th root transformation may be more appropriate in my linear regression. 2.70 was still higher than the max:min sd ratio cutoff of 2, and many groups were still not normally distributed and had outliers, so I proceeded with model interpretation with great caution.

It looked like counties classified in higher crop groups received more subsidies, and that there could be some regional differences (Fig. A4). High crop group counties in the Midwest (8.87), Mountainwest (9.95), and South (9.14) had higher medians than in the Southeast (6.15), Northeast (7.22), and Pacific (7.48) (Fig. A5). It also looked like there could be an interaction between region and proportion of operations growing crops. In all regions except the Southeast, counties with higher proportions of crop operations seemed to receive more subsidies. In an interplot (Fig. 1), the slopes of line segments between govsublog ~ region groups differed for each crop group.

Modeling

I assessed the ANOVA model (govsublog ~ region*crop_group) (Fig. A6). This model somewhat fit ANOVA conditions, with similar residual variance in all but two groups, slightly left skewed residuals, and a few outliers (Fig. A7). The interaction term was significant (F=7.55, p = 5.6e-12), so I kept this as my final model. Because I had so many levels and interactions, I created letters from my Tukey post-hoc testing results to efficiently compare groups.

I was also curious about the individual effects of region type on govsub, so I ran a quick one factor ANOVA. The conditions for this model looked ok - residuals exhibited equal variance, but deviated from normality at extreme values and had a few outliers (Fig. A7). I found the following significant group differences (Fig. 1):

Midwest » Mountainwest » South » Southeast, Pacific, Northeast

Farms in counties in the Midwest received significantly higher subsidies than farms in counties in every other region. Most extreme difference: I am 95% confident that average farms in counties in the Midwest received between 3.9325 and 7.5592 times as many subsidies as average farms in counties in the Northeast. The estimated difference between these regions is 5.45 times as many subsidies received in the Midwest than the Northeast, an effect that is 22% larger than typical county to county variation.

When I added interactive effects of crop proportion, these differences became more complicated (Fig. 1). For the high crop group, average farms in the Midwest and Mountainwest received significantly more subsidies than average farms in the Northeast and Southeast; average farms in the Pacific and South were not significantly different from either group. For the medium crop group, average farms in the Midwest received significantly more subsidies than farms in all other regional groups. Average farms in the Northeast received

significantly lower subsidies than all other groups. For the low crop group, average farms in the Mountainwest received significantly more subsidies than average farms in the Northeast and Pacific, otherwise there were few significant differences between groups.

The effect of crop proportion differed substantially by region (Fig. 1). In the Midwest, average farms in higher crop group counties received significantly more subsidies. In the Mountainwest, average farms in high crop group counties received significantly more subsidies than medium and low crop group counties. The Northeast and Pacific showed no significant differences between crop groups. In the South, average farms in higher crop group counties received significantly more subsidies. In the Southeast, average farms in medium crop_group counties received significantly higher subsidies than high and low crop group counties.

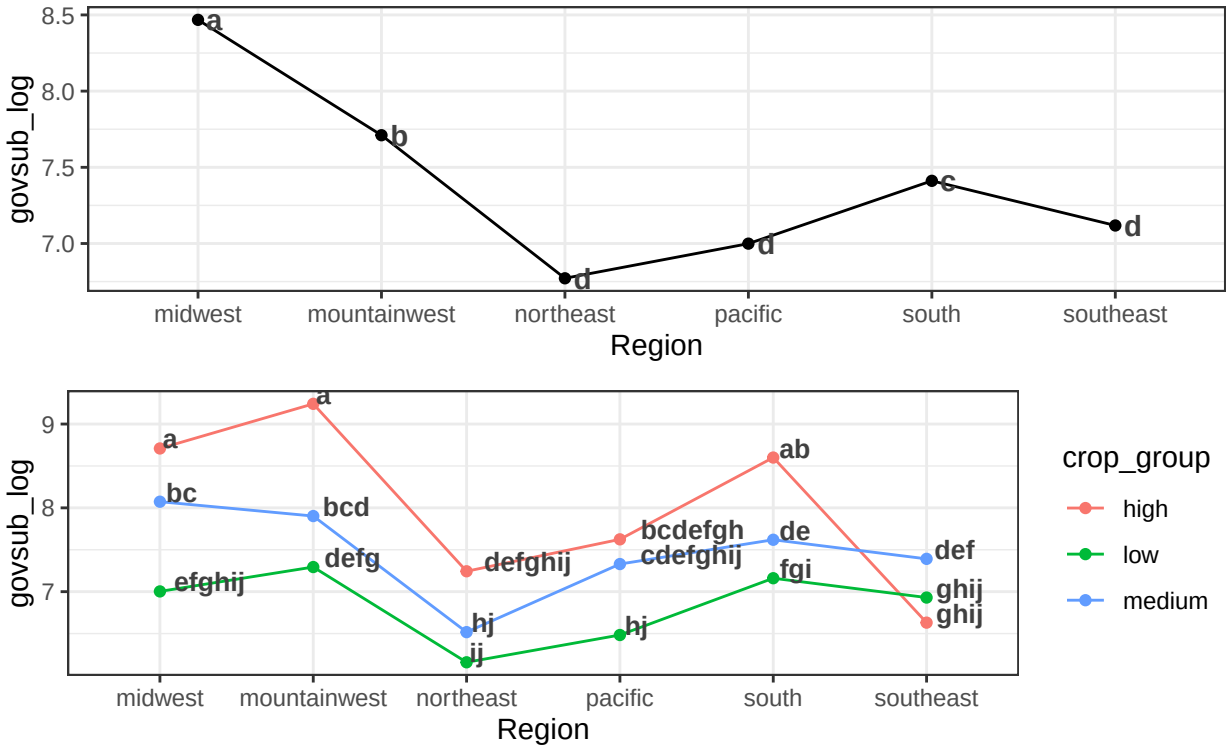

Figure 1. Group differences for one-factor (top) and two factor (bottom) ANOVA models predicting log(government subsidies). Significantly different groups have non-overlapping letters.

Linear regression

EDA

For my linear regression, I decided to analyze cropping proportion as a quantitative variable instead of using the categorical variable I created for the ANOVA analysis, since I expected that this would make it a much stronger predictor. In initial scatterplots, predictor-response relationships all looked questionably linear (Fig. A8). Since a 10th root transformation was most helpful for normalizing govsub in my ANOVA analysis, I tried it here as well. I did notice that this transformation made the distribution of govsub slightly bimodal, however (Fig. A3), so I was careful to check linearity of predictor relationships.

In addition to normalizing govsub, as described above, the 10th root transformation did a good job improving linearity of govsub ~ predictor relationships, although farm_income, farm_acres, and hisp_prop were still very right skewed. corn_prop_crop also looked right skewed and questionably linear. I transformed these variables to further improve linearity. I expected that log transformations would be very helpful for farm_income and farm_acres since they looked highly right skewed, but I tried both log and square root transformations for hisp_prop and corn_prop_crop since they were proportional data.

Log transformations greatly improved the normality of farm_income, (mean, median, IQR = (5963169, 3794000, 5768000)->(15.088, 15.149, 1.432)) farm_acres, (mean, median, IQR = (635.05, 271, 375)->(5.7895, 5.6021, 1.1962)) and hisp_prop (mean, median, IQR = (0.034566, 0.014403, 0.022885)->(-4.1802,-4.2403, 1.4461)). Corn_prop_crop was a bit normalized by a square root transformation (mean, median, IQR = (0.17396, 0.076923, 0.28318)->(0.3313, 0.27735, 0.42993)), although it notably looked right skewed and bimodal (Fig. 2).

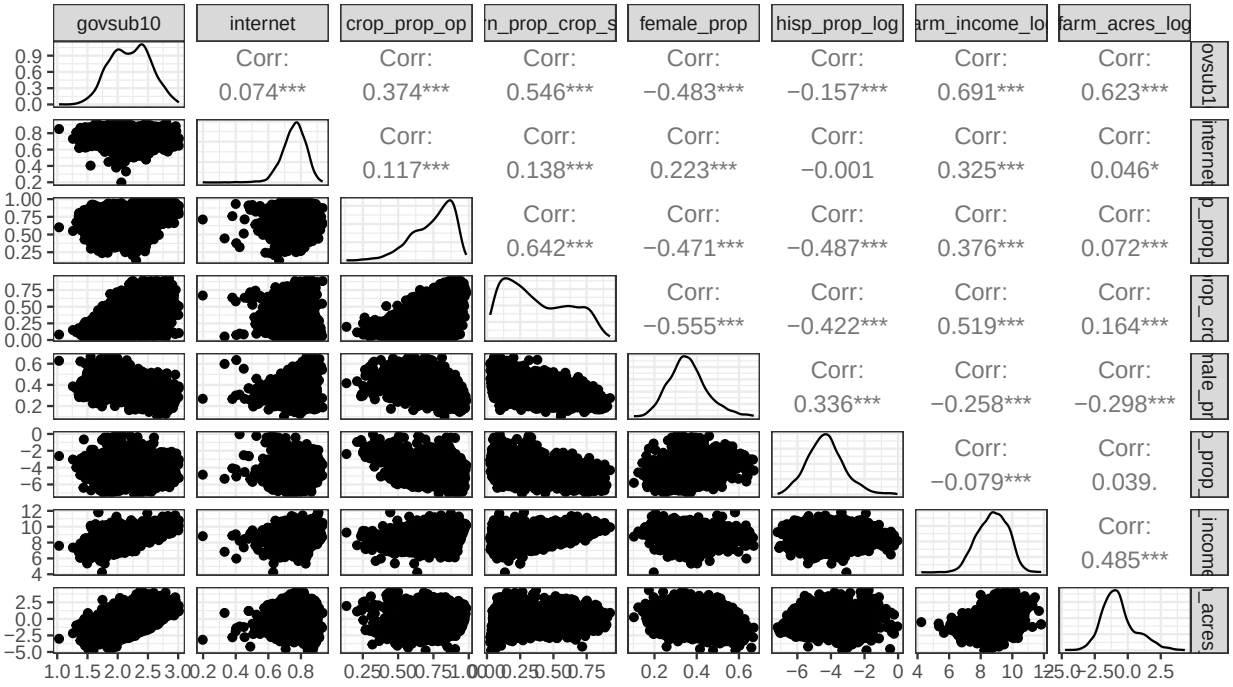

Figure 2. Scatterplots and correlations for relationships between all quantitative predictor and response variables

Although the transformation of corn_prop_crop was a bit questionable, the relationship between govsub and sqrt(corn_prop_crop) looked much more linear (Fig. A8). Every other predictor looked like they had fairly linear relationships with govsub (Fig.2, Fig. A8). After transforming my variables, all correlations between predictors and govsub were significant (Fig. 2). Internet, crop_prop_op, sqrt(corn_prop_prop), log(farm_income), and log(farm_area) were positively correlated with $govsub^{0.1}$, and female_prop and log(hisp_prop) were negatively correlated with $govsub^{0.1}$ (Fig. 2). I felt confident that these were interesting and appropriately-fitting predictors, so I proceeded with model selection with these transformed variables.

Modeling

I started by assessing conditions of a basic model containing all of my variables of interest (model 1) (Fig. A10). The residuals were slightly heteroscedastic, with decreased variance at higher fitted values. They also deviated from normality, appearing slightly left skewed. Both plots showed a few outliers with low residual values. I was not majorly concerned about these residuals, so I proceeded with model selection and interpretation.

Before comparing different models, I also assessed collinearity between predictors. A lot of predictors were pretty highly (r=0.2-0.64) correlated with each other (Fig. 2). Corn_prop_crop had the highest correlations with other variables. However, VIF scores for all predictors were below 2 (Fig. A9), indicating that no problematic multicollinearity was present. I assumed that my model coefficients accurately took into account the effects of other predictors and that multicollinearity did not give me a reason to remove any predictors.

Model Selection Best subsets selection could not handle the region variable (split it into 6 variables, but only showed up to 8 subsets), so I performed automated variable selection using stepwise and backwards

regression. Both stepwise and backwards regression supported keeping all predictors except hisp_prop. This model (model 2) had the lowest AIC value, -7958.5. However, since the model with hisp_prop (model 1) had a very similar AIC value, -7957.5, I double checked the potential significance of this predictor before dropping it.

The coefficient for the hisp_prop predictor was insignificant (t=-1.00, p=0.318), which was supported by insignificant nested F test results between the model 1 and model 2 (F=1, p=0.32). Furthermore, hisp_prop did not improve model strength (adj R^2 = 0.705 for both). Conditions looked about the same for both models (Fig. A10). I saw no reason to include hisp_prop, and therefore removed it to simplify my model.

Interactions Because I saw a significant interaction between region and crop_prop_op in my ANOVA analysis, I was inclined to test out regional interactions in my linear regression as well. After adding the interaction (model 3), Adj R^2 increased from 0.705 to 0.707; RSE decreased from 0.178 to 0.177. Interactions between crop_prop_op and regionmidwest (p=0.033), regionnortheast (p=0.048), and regionsoutheast (p=0.028) were significant. Residuals and conditions looked similar (Fig. A10). As discussed below, model 3 had many more influential points, which was concerning. Because of this, and because the interaction did not substantially improve model strength, I chose not to include it.

Outliers and influential points Model 3 had more issues with potentially influential points than model 2 (Fig. A11). Although all Cook's Distance values for both models were below 0.5, model 3 had 9 points above 0.015, whereas model 2 only had 4 points. Potential influence of high leverage points can be seen in the standardized residual vs leverage plots in Fig. A11, which have an upwards sloping best fit line in model 3 and a generally flat best fit line in model 2. Both models had similar numbers of concerning outliers: model 3 had 17 points with |standardized residuals| > 3 and model 2 had 16 points. I chose not to remove these points because I was not overly concerned about their influence on the model.

Final model $\widehat{govsub^{0.1}} = 1.100 + 0.007(regionmountainwest) - 0.054(regionmidwest) + 0.076(regionsouth) - 0.085(regionsoutheast) + 0.210(crop\_prop\_op) + 0.370(corn\_prop\_crop\_sqrt) + 0.134(farm\_income\_log) + 0.089(farm\_acres\_log) - 0.290(female\_prop) - 0.224(internet)$

My final linear regression model was model 2, which included all of my chosen predictors except the proportion of farmers in a county that are hispanic and had no interaction terms. All predictors in my model, except regionMountainwest, were significant (Fig. A12). I chose not to remove this region level by combining it with the baseline level, since this would affect how all other levels differed from the baseline. Overall, this model was significant (F=502, p<2e-16) and explained 70.5% of the variation in $govsub)^{0.1}$.

Interpretation of models

Interpretation of both of my models was constrained by deviation from parametric conditions for ANOVA and linear regression. I do not have to worry about random selection of my data, since it was collected as a census. However, this data almost certainly violates independence, since farm demographic variables in one county likely influence surrounding counties. Data for my ANOVA model also did not exhibit sufficient equal variance between groups, and residuals for both models exhibited mild left skew.

Interestingly, after including the additional predictors in my linear model, differences in government subsidies received by region changed substantially compared to my ANOVA analysis. My one-factor ANOVA results indicated that regions received significantly different subsidies, from highest to lowest, in the order:

Midwest » Mountainwest » South » Southeast, Pacific, Northeast

My linear regression, on the other hand, suggested the following order:

South > Mountainwest » Pacific » Midwest > Southeast > Northeast

where all regions except Mountainwest significantly differed from Pacific in the given direction. The greatest change was that the effect of being in the Midwest changed from significantly increasing subsidies relative to all other regions to significantly decreasing subsidies relative to the South, Mountainwest, and Pacific when all other chosen factors were controlled. This could mean that the Midwest has different characteristics related to other factors in my model (farm size, farm income, farm crop choice, gender of farm owner, internet access) that cause it to receive more subsidies than other regions.

Although internet access was slightly positively correlated with government subsidies (r=0.074), when all chosen factors were controlled, it showed a significantly negative effect on subsidies. For every 0.1 increase in the proportion of farms with internet access in a county, average (10th root of) government subsidy $ received by farms was predicted to decrease by 0.0224. Counties with low internet access may tend to have characteristics related to farm size, income, cropping proportion, etc that decrease subsidies received. Although it was negatively correlated with government subsidies (r=-0.157), the proportion of farms in a county owned by Hispanic producers was not significantly associated with average government subsidies received in that county after controlling for other predictor effects.

All other variables retained the same directional relationships with govsub in the final model as in raw correlations. As in my ANOVA analysis, average farms in counties with a greater share of farms growing crops were predicted to receive more subsidies, even when all other predictors were held constant. For every 0.1 increase in crop proportion, average (10th root of) government subsidy $ received by farms was predicted to increase by 0.0210. Similarly, average farms in counties with a greater share of farms growing corn (0.37 increase in $govsub^{0.1}$ for every 0.1 increase in sqrt(corn proportion)), larger farms (0.0085 increase in $govsub^{0.1}$ for every 10% increase in average farm size in acres), and farms making more income (0.013 increase in $govsub^{0.1}$ for every 10% increase in average farm income in $) were predicted to receive significantly more government subsidies with all other predictors held constant. Counties with a greater share of farms owned by a female producers were predicted to receive significantly less government subsidies, holding all other predictors constant (0.029 decrease in $govsub^{0.1}$ for every 0.1 increase in the proportion of farms in a county owned by female producers).

Conclusions

My results supported most of my hypotheses. I did not find support for the hypothesis that goverment subsidies disfavored counties with more Hispanic farmers, but my results did support the hypothesis that subsidies disfavored counties with more female producers, even after accounting for farm size, income, crop-growing, corn-growing, region, and internet access. Contrary to my internet access hypothesis, my linear regression model indicated that farms with more internet access receive fewer subsidies. Since the raw correlation between internet access and subsidies was positive, this result only makes sense in the context of controlling for the effects of my other variables. My results supported my hypotheses that counties with more crop producers, corn producers, and higher average farm size and income received more subsidies. My results somewhat supported my hypothesis that the Midwest would receive more subsidies than all other U.S. regions. Midwestern counties received greater subsidies than all other regions in my one-factor ANOVA model, but received low relative subsidies after accounting for farm size, income, crop-growing, corn-growing, female farm management, and internet access.

These results were limited by my county-scale analysis, and do not necessarily reflect outcomes for individual farms. Furthermore, my ANOVA analysis violated parametric statistical conditions, and would be better studied using nonparametric methods. Future analyses could address these concerns and could also focus on variables I was not able to find workable data for, such as organic farming practices and farm ownership by other minority groups.

References

EWG, 2021: https://farm.ewg.org/progdetail.php?fips=00000&progcode=total&page=states

EWG, 2007: https://www.ewg.org/research/short-crop

Edwards, 2023: https://www.cato.org/briefing-paper/cutting-federal-farm-subsidies

U.S. Census of Agriculture: https://www.nass.usda.gov/AgCensus/

U.S. Census Regions: https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf

Sample code for Tukey post-hoc letters: https://stackoverflow.com/questions/75816862/how-do-i-add-tukeyhsd-results-from-multcompletters-function-to-an-interaction-pl

Appendices

```
df1$region = df1$state_lower
df1$region<-as.factor(df1$region)

levels(df1$region) <- list("pacific" = c("washington", "oregon", "california", "hawaii"),
                           "mountainwest" = c("montana", "idaho", "wyoming", "nevada",
                                              "utah", "colorado", "arizona",
                                              "new mexico"),
                           "midwest" = c("north dakota", "south dakota", "nebraska",
                                         "kansas", "minnesota", "iowa", "missouri",
                                         "wisconsin", "illinois", "michigan", "indiana",
                                         "ohio"),
                           "northeast" = c("maine", "connecticut", "new hampshire",
                                           "vermont", "massachusetts", "rhode island",
                                           "new york", "new jersey", "pennsylvania"),
                           "south" = c("texas", "oklahoma", "arkansas", "louisiana",
                                       "mississippi", "alabama", "tennessee",
                                       "kentucky"),
                           "southeast" = c("west virginia", "maryland", "delaware",
                                           "virginia", "north carolina",
                                           "south carolina", "georgia", "florida"))
```

Figure A1. Creation of regional groups from state names.

```
df3$crop_group <- as.factor(cut_number(df3$crop_prop_op, n=3))
#cutoffs: [0.0685,0.673], (0.673,0.826], (0.826,0.983]

df3 <- df3 %>%
  mutate(crop_group = factor(crop_group,
                     labels = c("low", "medium", "high")))
```

Figure A2. Creation of crop group categorical variable.

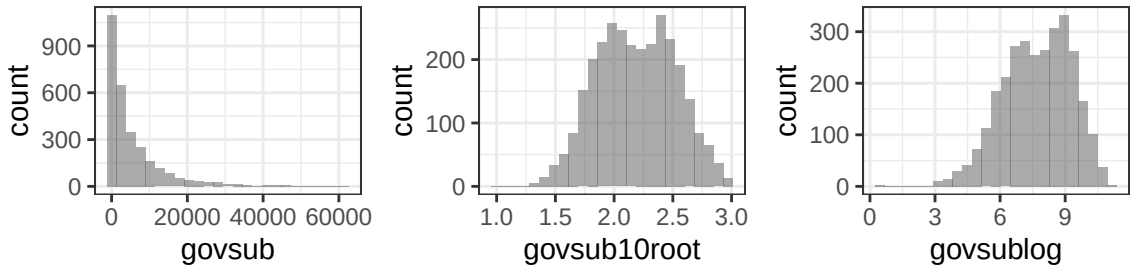

Figure A3. Distributions of untransformed, 10th root transformed, and log transformed government farm subsidy data.

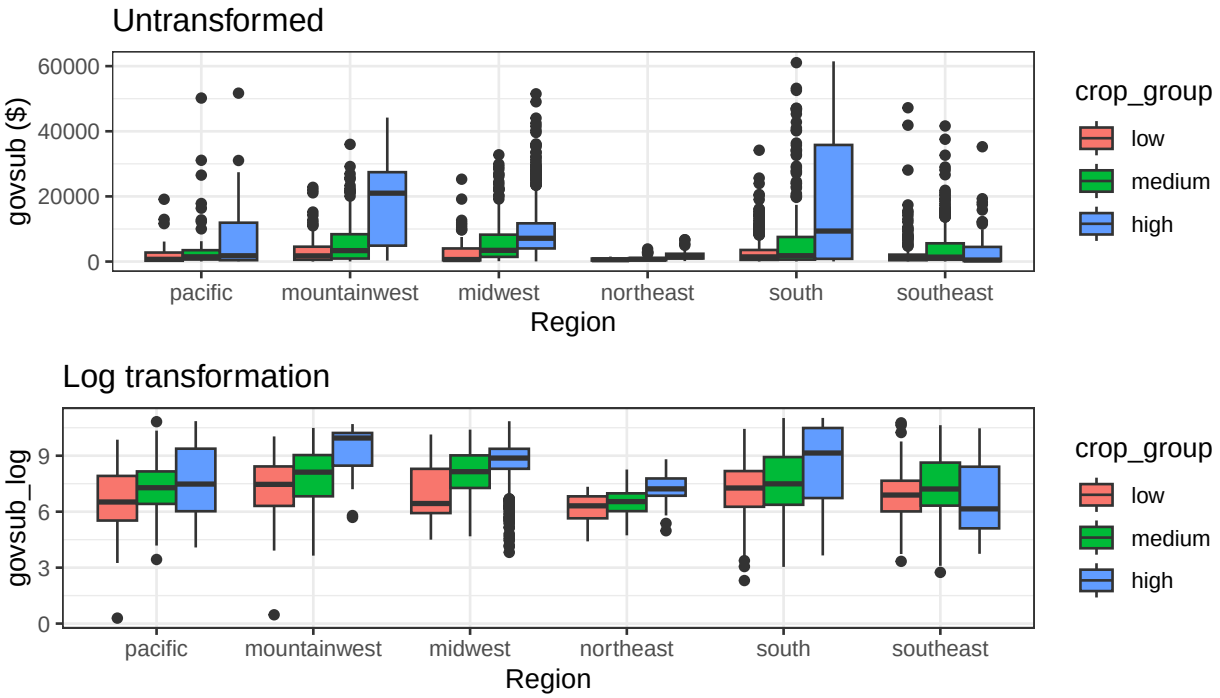

Figure A4. Boxplots showing differences in untransformed and log transformed government farm subsidies by region and crop proportion.

Table 1: Summary of log transformed government farm subsidies

region.crop_group	min	Q1	median	Q3	max	mean	sd	n	missing
pacific.low	0.293	5.53	6.52	7.92	9.86	6.48	1.808	58	0
mountainwest.low	0.472	6.31	7.46	8.43	10.03	7.29	1.494	136	0
midwest.low	4.498	5.92	6.44	8.30	10.14	7.00	1.514	55	0
northeast.low	4.408	5.64	6.31	6.82	7.34	6.16	0.782	23	0
south.low	2.307	6.26	7.27	8.18	10.44	7.16	1.418	505	0
southeast.low	3.337	6.01	6.89	7.66	10.76	6.93	1.297	212	0
pacific.medium	3.437	6.42	7.28	8.16	10.82	7.33	1.600	49	0
mountainwest.medium	3.644	6.82	8.12	9.04	10.49	7.90	1.542	104	0
midwest.medium	4.679	7.27	8.14	9.02	10.40	8.07	1.233	248	0
northeast.medium	4.732	6.03	6.54	6.98	8.26	6.52	0.805	89	0
south.medium	3.040	6.37	7.49	8.93	11.02	7.62	1.784	243	0
southeast.medium	2.746	6.32	7.21	8.63	10.64	7.39	1.597	255	0
pacific.high	4.081	6.02	7.48	9.38	10.85	7.62	2.055	22	0
mountainwest.high	5.694	8.47	9.95	10.22	10.70	9.24	1.446	24	0
midwest.high	3.818	8.30	8.87	9.37	10.85	8.71	1.073	740	0
northeast.high	4.978	6.85	7.22	7.78	8.81	7.24	0.762	78	0
south.high	3.654	6.73	9.14	10.49	11.03	8.60	2.098	64	0
southeast.high	3.741	5.11	6.15	8.41	10.47	6.63	1.861	61	0

Figure A5. Numerical summary of log(government farm subsidies) for each region x crop

proportion group.

```
##                      Df Sum Sq Mean Sq F value  Pr(>F)
## crop_group            2    971     486  249.99 < 2e-16 ***
## region                5    488      98   50.27 < 2e-16 ***
## crop_group:region    10    147      15    7.55 5.6e-12 ***
## Residuals          2948   5728       2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure A6. Summary of two-way non-additive ANOVA model predicting log(government farm subsidies).

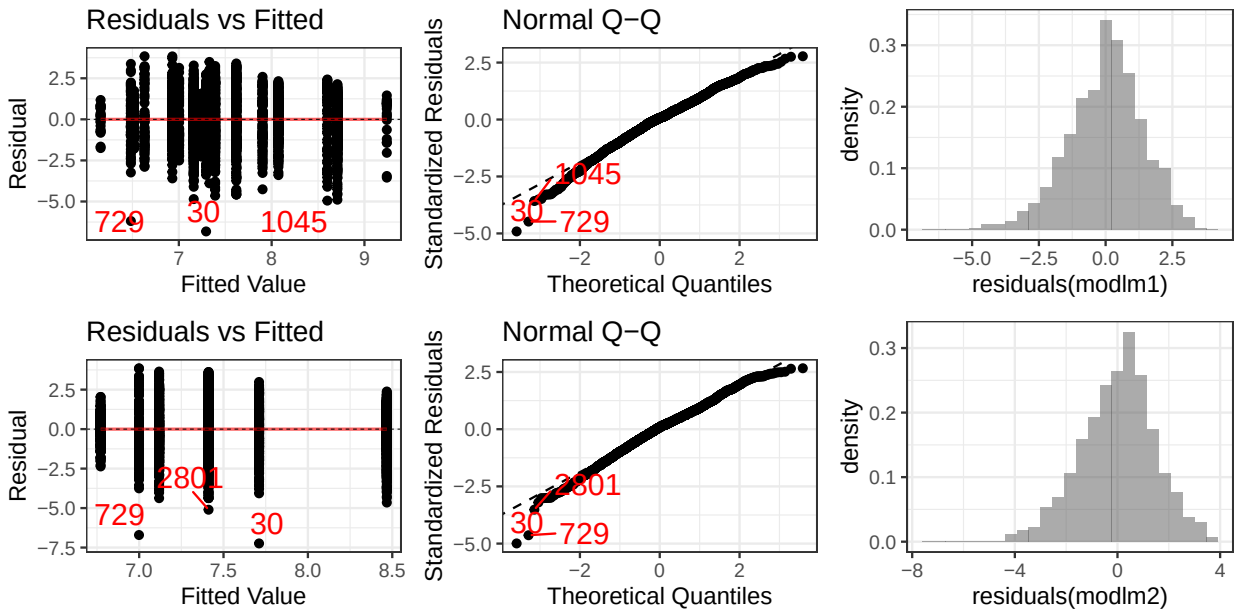

Figure A7. Residual diagnostics for two-way non-additive ANOVA model (top) and one-way ANOVA model (bottom) predicting log(government farm subsidies) by crop proportion and region (top) and region (bottom).

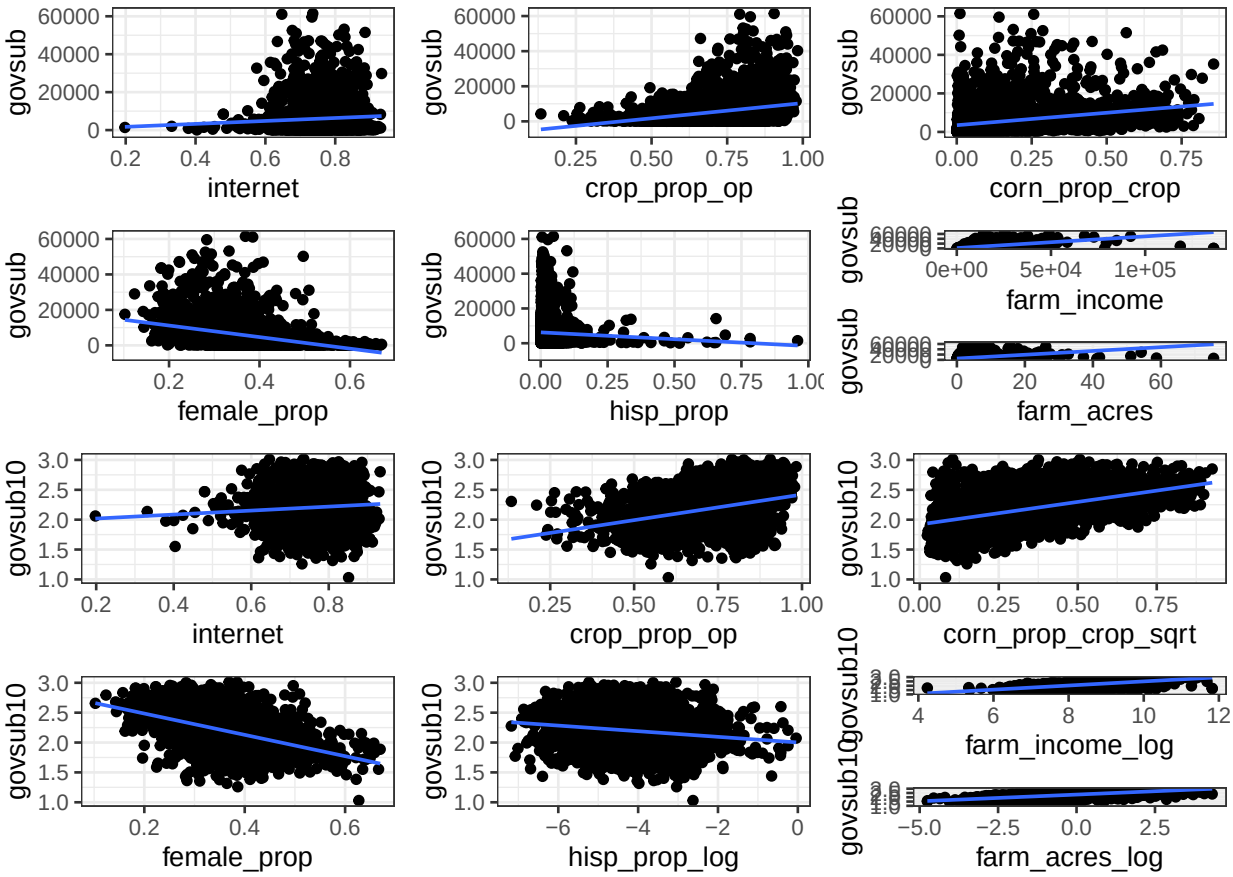

Figure A8. Relationships between untransformed (top) or 10th root transformed government farm subsidies (bottom) and transformed predictors. Linear regression lines shown in blue.

```
##                     GVIF Df GVIF^(1/(2*Df))
## region              4.47  5            1.16
## crop_prop_op        2.50  1            1.58
## corn_prop_crop_sqrt 3.42  1            1.85
## farm_income_log     2.27  1            1.51
## farm_acres_log      1.73  1            1.31
## female_prop         2.24  1            1.50
## hisp_prop_log       1.50  1            1.22
## internet            1.38  1            1.17
```

Figure A9. Non-adjusted and adjusted VIF scores for all linear regression predictors.

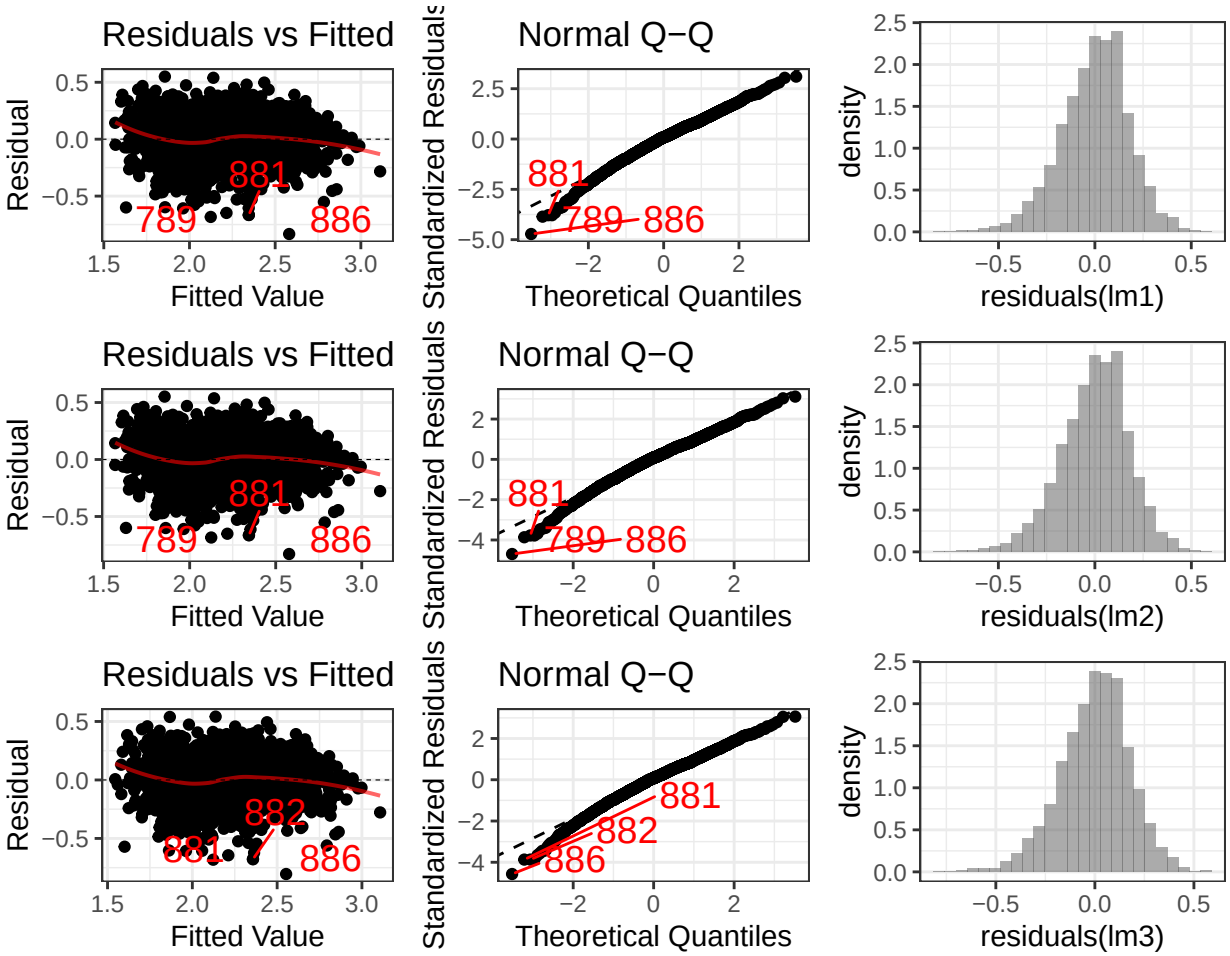

Figure A10. Residual diagnostics for model 1 (top), model 2 (middle), model 3 (bottom)

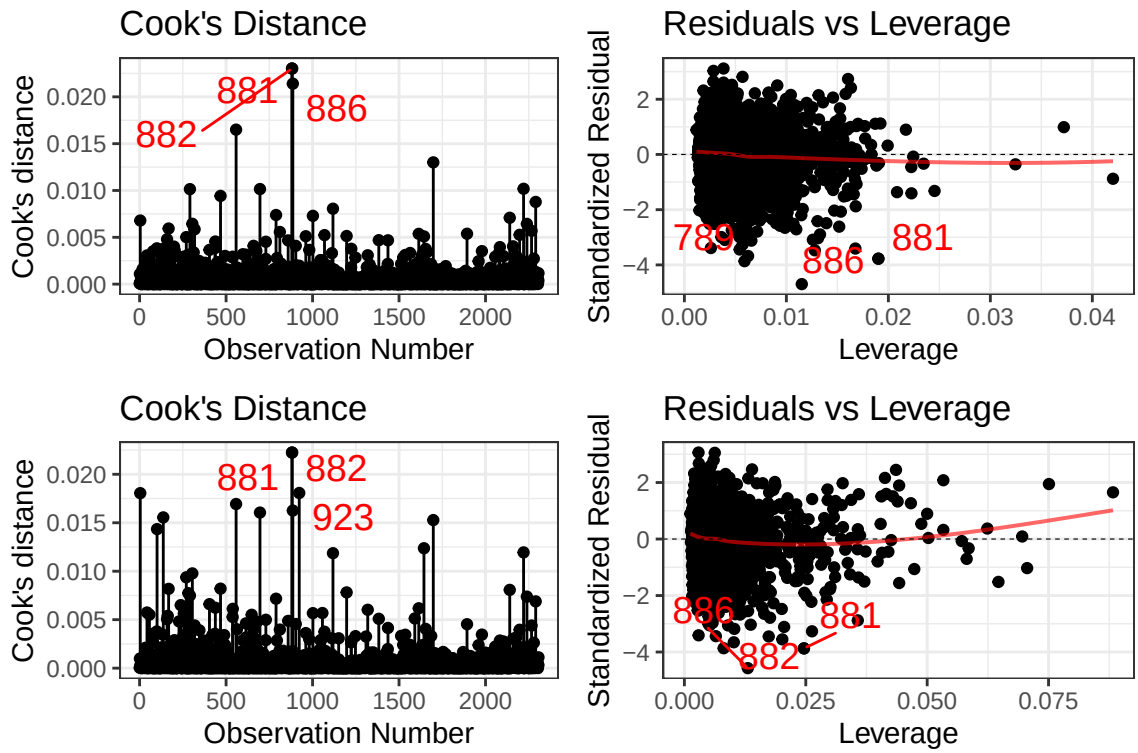

Figure A11. Cook's distance values and standardized residual vs leverage values for model 2 (top) and model 3 (bottom)

```
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.10010    0.07607   14.46  < 2e-16 ***
## regionmountainwest    0.00697    0.02667    0.26  0.79376
## regionmidwest        -0.05419    0.02517   -2.15  0.03142 *
## regionnortheast      -0.14182    0.02573   -5.51  3.9e-08 ***
## regionsouth           0.07610    0.02436    3.12  0.00181 **
## regionsoutheast      -0.08458    0.02436   -3.47  0.00053 ***
## crop_prop_op          0.20971    0.03923    5.35  9.9e-08 ***
## corn_prop_crop_sqrt   0.37000    0.02885   12.82  < 2e-16 ***
## farm_income_log       0.13356    0.00586   22.79  < 2e-16 ***
## farm_acres_log        0.08919    0.00373   23.94  < 2e-16 ***
## female_prop          -0.28989    0.06243   -4.64  3.6e-06 ***
## internet             -0.22453    0.06006   -3.74  0.00019 ***
##
## Residual standard error: 0.178 on 2294 degrees of freedom
## Multiple R-squared:  0.707,  Adjusted R-squared:  0.705
## F-statistic:  502 on 11 and 2294 DF,  p-value: <2e-16
```

Figure A12. Summary of final linear regression model.