**Honey Production in the United States: 1998-2012**

**Sneh Koul & Faith Arnold**

Advanced Statistical Methods for IS&T

ISQA 4150-001

Fall 2018

**Introduction**

Pollination is an incredibly important part of crop production and for increasing the health and yield of almost every kind of plant. Because of their heavy role in pollination, the populations and health of bee colonies is vital for the continuation of production. In 2006, there was a dramatic decline in honeybee populations due to colony collapse. While no specific underlying cause has been identified, analyzing the health of the bee colonies can be useful in determining if the decline has a heavy impact on the work of the bees. Although since bees are small and have a pretty short lifespan and a complicated hive and colony system, it’s tough to count bees. But by measuring the production of honey, one can easily infer the health and relative population of the colonies, since unhealthy colonies won’t have very good production.

The purpose of this analysis is to use the data from honey production to infer the health of the bees across different states and over time. A lot of new changes in agricultural fertilization and pest control can have negative effects on the health of the colonies, and these changes could be much more detrimental to the overall production of crops and of plants worldwide. Honey production and bee colony health are important indicators of crop pollination and are vital signs of agricultural production.

**Dataset description**

The data used for our analysis was retrieved from Kaggle. The data provided contains 626 rows with 8 columns, which include data from 43 states over 14 years (1998-2012).

* *state:* Abbreviated state name.
* *numcol*: Number of honey-producing colonies
* *yieldpercol*: Honey yield per colony (lb)
* *totalprod*: Total production (lb)
* *stocks*: Stocks held by producers (lb)
* *priceperlb*: Refers to mean price per pound ($)
* *prodvalue*: Value of production ($)
* *year*: Year the data is from.

Honey production data was published by the National Agricultural Statistics Service (NASS) of the U.S. Department of Agriculture.

**Descriptive Statistics**

The figure below is the descriptive statistics for the honey production dataset combining the data from all the states. For each state, data is collected for each variable below from the year of 1998 to 2012. Some states were not fully represented from 1998 to 2012 so they were removed. In total there are 40 states represented in the analysis of this dataset.

****

**Research questions / Hypothesis development**

The following are some research questions we would like to address in this project:

1. How has honey production changed from 1998-2012 in each state in the United States? Is there any significant difference in mean honey production among the states? (ANOVA analysis)
   1. Hypothesis
      1. H0: µ1 = µ2 = µ3 … = µn where µ1 is mean honey production for Alabama, µ2 is mean honey production for Arizona, µ3 is mean honey production for Arkansas….(this can be applied to the other states as well)
      2. Ha: There is a state population where the average honey production is different.
2. Is there a relationship between honey production and the cost of honey? (Simple Linear Regression analysis)
   1. Hypothesis
      1. We believe there is a significant effect of honey production on the cost of honey.
3. Does the state and the honey production influence the price of honey per pound? (Multiple Regression)
   1. Hypothesis
      1. We believe the state and honey production do influence the price of honey per pound.

**Planned statistical methods**

*ANOVA*

This method will be used to address research question one. We specifically plan on using a randomized block design to answer this question. The blocking variable will be the years, 1998-2012 while the treatment groups will be the states. Additionally, we will use Fisher’s LSD test to determine where the difference lies between the states.

*Simple Regression*

This method will be used to address research questions 2 and 3. The independent and dependent variable for question 2 would be honey production and cost of honey, respectively.

The independent and dependent variable for question 3 would be the number of colonies and cost of honey, respectively. The correlation coefficient will be able to determine the strength of association between the independent and dependent variable. We will also use an F test to test for significance of the model. Model assumptions will also be examined.

*Multiple Regression*

This method will be used to address research question 4. The independent variables will be states and honey production. The dependent variable will be the price of honey per pound. The same idea described above for simple regression will also be used for multiple regression.

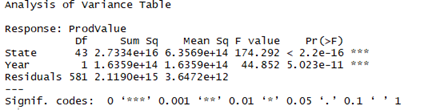
**Results**

**ANOVA Analysis**

ANOVA analysis was conducted on the honey production dataset to see if the honey

production varied among states and if year also had an influence of the honey production.

The aov and anova functions were run to generate the following ANOVA table.



We observe that p value for the treatment is significant implying that there is at least a

state where the average honey production is different. The blocking variable is also

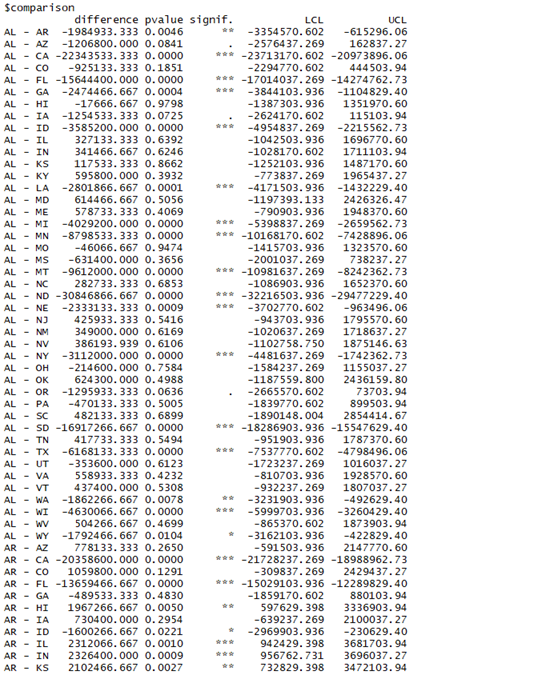
considered significant with a p value below 0.05. This also shows that the honey

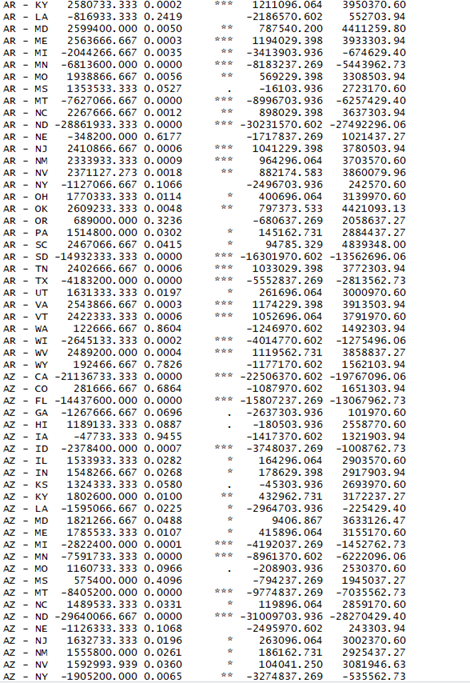
production changes by year as well. To determine where the differences lie between

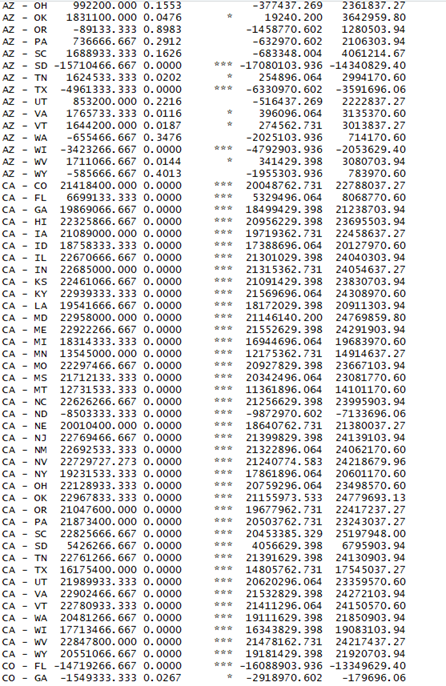
states for honey production Fisher’s LSD test was conducted. Shown below are all the comparisons tested. There are 587 differences observed between states and they include the following:

|  |
| --- |
| AL-AR,AL-CA,AL-FL,AL-GA,AL-ID,AL-LA,AL-MI,AL-MN,AL-MT,AL-ND,AL-NE,AL-NY,AL-SD,AL-TX,AL-WA,AL-WI,AL-WY,AR-CA,AR-FL,AR-HI,AR-ID,AR-IL,AR-IN,AR-KS,AR-KY,AR-MD,AR-ME,AR-MI,AR-MN,AR-MO,AR-MT,AR-NC,AR-ND,AR-NJ,AR-NM,AR-NV,AR-OH,AR-OK,AR-PA,AR-SC,AR-SD,AR-TN,AR-TX,AR-UT,AR-VA,AR-VT,AR-WI,AR-WV,AZ-CA,AZ-FL,AZ-ID,AZ-IL,AZ-IN,AZ-KY,AZ-LA,AZ-MD,AZ-ME,AZ-MI,AZ-MN,AZ-MT,AZ-NC,AZ-ND,AZ-NJ,AZ-NM,AZ-NV,AZ-NY,AZ-OK,AZ-SD,AZ-TN,AZ-TX,AZ-VA,AZ-VT,AZ-WI,AZ-WV,CA-CO,CA-FL,CA-GA,CA-HI,CA-IA,CA-ID,CA-IL,CA-IN,CA-KS,CA-KY,CA-LA,CA-MD,CA-ME,CA-MI,CA-MN,CA-MO,CA-MS,CA-MT,CA-NC,CA-ND,CA-NE,CA-NJ,CA-NM,CA-NV,CA-NY,CA-OH,CA-OK,CA-OR,CA-PA,CA-SC,CA-SD,CA-TN,CA-TX,CA-UT,CA-VA,CA-VT,CA-WA,CA-WI,CA-WV,CA-WY,CO-FL,CO-GA,CO-ID,CO-KY,CO-LA,CO-ME,CO-MI,CO-MN,CO-MT,CO-ND,CO-NE,CO-NY,CO-SD,CO-TX,CO-VA,CO-WI,CO-WV,FL-GA,FL-HI,FL-IA,FL-ID,FL-IL,FL-IN,FL-KS,FL-KY,FL-LA,FL-MD,FL-ME,FL-MI,FL-MN,FL-MO,FL-MS,FL-MT,FL-NC,FL-ND,FL-NE,FL-NJ,FL-NM,FL-NV,FL-NY,FL-OH,FL-OK,FL-OR,FL-PA,FL-SC,FL-TN,FL-TX,FL-UT,FL-VA,FL-VT,FL-WA,FL-WI,FL-WV,FL-WY,GA-HI,GA-IL,GA-IN,GA-KS,GA-KY,GA-MD,GA-ME,GA-MI,GA-MN,GA-MO,GA-MS,GA-MT,GA-NC,GA-ND,GA-NJ,GA-NM,GA-NV,GA-OH,GA-OK,GA-PA,GA-SC,GA-SD,GA-TN,GA-TX,GA-UT,GA-VA,GA-VT,GA-WI,GA-WV,HI-ID,HI-LA,HI-MI,HI-MN,HI-MT,HI-ND,HI-NE,HI-NY,HI-SD,HI-TX,HI-WA,HI-WI,HI-WY,IA-ID,IA-IL,IA-IN,IA-KS,IA-KY,IA-LA,IA-MD,IA-ME,IA-MI,IA-MN,IA-MT,IA-NC,IA-ND,IA-NJ,IA-NM,IA-NV,IA-NY,IA-OK,IA-SD,IA-TN,IA-TX,IA-VA,IA-VT,IA-WI,IA-WV,ID-IL,ID-IN,ID-KS,ID-KY,ID-MD,ID-ME,ID-MN,ID-MO,ID-MS,ID-MT,ID-NC,ID-ND,ID-NJ,ID-NM,ID-NV,ID-OH,ID-OK,ID-OR,ID-PA,ID-SC,ID-SD,ID-TN,ID-TX,ID-UT,ID-VA,ID-VT,ID-WA,ID-WV,ID-WY,IL-LA,IL-MI,IL-MN,IL-MT,IL-ND,IL-NE,IL-NY,IL-OR,IL-SD,IL-TX,IL-WA,IL-WI,IL-WY,IN-LA,IN-MI,IN-MN,IN-MT,IN-ND,IN-NE,IN-NY,IN-OR,IN-SD,IN-TX,IN-WA,IN-WI,IN-WY,KS-LA,KS-MI,KS-MN,KS-MT,KS-ND,KS-NE,KS-NY,KS-OR,KS-SD,KS-TX,KS-WA,KS-WI,KS-WY,KY-LA,KY-MI,KY-MN,KY-MT,KY-ND,KY-NE,KY-NY,KY-OR,KY-SD,KY-TX,KY-WA,KY-WI,KY-WY,LA-MD,LA-ME,LA-MN,LA-MO,LA-MS,LA-MT,LA-NC,LA-ND,LA-NJ,LA-NM,LA-NV,LA-OH,LA-OK,LA-OR,LA-PA,LA-SC,LA-SD,LA-TN,LA-TX,LA-UT,LA-VA,LA-VT,LA-WI,LA-WV,MD-MI,MD-MN,MD-MT,MD-ND,MD-NE,MD-NY,MD-OR,MD-SD,MD-TX,MD-WA,MD-WI,MD-WY,ME-MI,ME-MN,ME-MT,ME-ND,ME-NE,ME-NY,ME-OR,ME-SD,ME-TX,ME-WA,ME-WI,ME-WY,MI-MN,MI-MO,MI-MS,MI-MT,MI-NC,MI-ND,MI-NE,MI-NJ,MI-NM,MI-NV,MI-OH,MI-OK,MI-OR,MI-PA,MI-SC,MI-SD,MI-TN,MI-TX,MI-UT,MI-VA,MI-VT,MI-WA,MI-WV,MI-WY,MN-MO,MN-MS,MN-NC,MN-ND,MN-NE,MN-NJ,MN-NM,MN-NV,MN-NY,MN-OH,MN-OK,MN-OR,MN-PA,MN-SC,MN-SD,MN-TN,MN-TX,MN-UT,MN-VA,MN-VT,MN-WA,MN-WI,MN-WV,MN-WY,MO-MT,MO-ND,MO-NE,MO-NY,MO-SD,MO-TX,MO-WA,MO-WI,MO-WY,MS-MT,MS-ND,MS-NE,MS-NY,MS-SD,MS-TX,MS-WI,MT-NC,MT-ND,MT-NE,MT-NJ,MT-NM,MT-NV,MT-NY,MT-OH,MT-OK,MT-OR,MT-PA,MT-SC,MT-SD,MT-TN,MT-TX,MT-UT,MT-VA,MT-VT,MT-WA,MT-WI,MT-WV,MT-WY,NC-ND,NC-NE,NC-NY,NC-OR,NC-SD,NC-TX,NC-WA,NC-WI,NC-WY,ND-NE,ND-NJ,ND-NM,ND-NV,ND-NY,ND-OH,ND-OK,ND-OR,ND-PA,ND-SC,ND-SD,ND-TN,ND-TX,ND-UT,ND-VA,ND-VT,ND-WA,ND-WI,ND-WV,ND-WY,NE-NJ,NE-NM,NE-NV,NE-OH,NE-OK,NE-PA,NE-SC,NE-SD,NE-TN,NE-TX,NE-UT,NE-VA,NE-VT,NE-WI,NE-WV,NJ-NY,NJ-OR,NJ-SD,NJ-TX,NJ-WA,NJ-WI,NJ-WY,NM-NY,NM-OR,NM-SD,NM-TX,NM-WA,NM-WI,NM-WY,NV-NY,NV-OR,NV-SD,NV-TX,NV-WA,NV-WI,NV-WY,NY-OH,NY-OK,NY-OR,NY-PA,NY-SC,NY-SD,NY-TN,NY-TX,NY-UT,NY-VA,NY-VT,NY-WI,NY-WV,OH-SD,OH-TX,OH-WA,OH-WI,OH-WY,OK-OR,OK-SD,OK-TX,OK-WA,OK-WI,OK-WY,OR-SD,OR-TN,OR-TX,OR-VA,OR-VT,OR-WI,OR-WV,PA-SD,PA-TX,PA-WA,PA-WI,SC-SD,SC-TX,SC-WI,SD-TN,SD-TX,SD-UT,SD-VA,SD-VT,SD-WA,SD-WI,SD-WV,SD-WY,TN-TX,TN-WA,TN-WI,TN-WY,TX-UT,TX-VA,TX-VT,TX-WA,TX-WI,TX-WV,TX-WY,UT-WA,UT-WI,UT-WY,VA-WA,VA-WI,VA-WY,VT-WA,VT-WI,VT-WY,WA-WI,WA-WV,WI-WV,WI-WY,WV-WY |

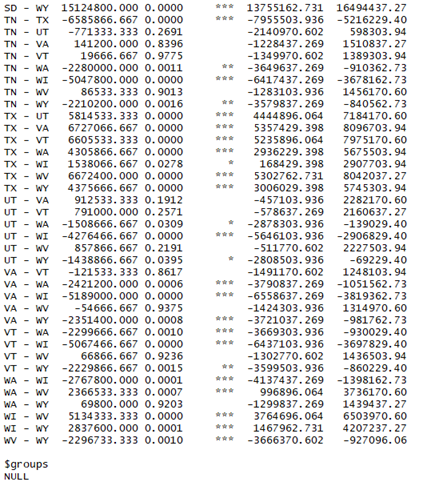
**Fisher’s LSD Test (Not all output data shown):**







…..



**Simple Linear Regression**

Now we would like to look at the relationship between honey production and the cost of honey through generating a linear model. Our dataset was split state by state to examine this relationship. It is important to note that due to this our sample size for each linear model was reduced to 13 data points associated with each variable. Below you will find the linear model, the plot of the model, and the assumptions associated with a simple linear regression model for each state.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

*Alabama:* From the plot we see that there is negative relationship between honey production and the price per lb of honey. The adjusted R-squared is 0.8112 which shows that there really is a strong relationship between the honey production vs. the price per pound for the state of Alabama. The F-value is considered statistically significant as the p value is 2.863x10-6 showing that the relationship would be considered significant. Now let’s look at the assumptions for this simple linear regression model. From the residuals vs. fitted plot we see a non-constant pattern meaning that another model would be better to represent this data. We would also say that this normally distributed as many points fall on the expected line. There are also definitely some high leverage points that should be further evaluated which include 7, 12, 14.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

*Arkansas:* From the plot we see that there is negative relationship between honey production and the price per lb of honey. The adjusted R-squared is 0.5443 which shows that there is a somewhat strong relationship between the honey production vs. the price per pound for the state of Arkansas. The F-value is considered statistically significant as the p value is 0.00102 showing that the relationship would be considered significant. Now let’s look at the assumptions for this simple linear regression model. From the residuals vs. fitted plot we see a non-constant pattern meaning that another model would be better to represent this data. We would also say that this normally distributed as many points fall on the expected line. There are also definitely some high leverage points that should be further evaluated which include 5, 15.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

*Arizona:* From the plot we see that there is negative relationship between honey production and the price per lb of honey. The adjusted R-squared is 0.6554 which shows that there really is a strong relationship between the honey production vs. the price per pound for the state of Arizona. The F-value is considered statistically significant as the p value is 0.0001551 showing that the relationship would be considered significant. Now let’s look at the assumptions for this simple linear regression model. From the residuals vs. fitted plot we see a don’t see a non-constant pattern meaning that the linear model is a good representation of this data. We would also say that this normally distributed as many points fall on the expected line. There are also definitely some high leverage points that should be further evaluated which include 6, 13, 15.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

*California:* From the plot we see that there is negative relationship between honey production and the price per lb of honey. The adjusted R-squared is 0.3037 which shows that there is a weak relationship between the honey production vs. the price per pound for the state of California. The F-value is considered statistically significant as the p value is 0.01943 showing that the relationship would be considered significant. Now let’s look at the assumptions for this simple linear regression model. From the residuals vs. fitted plot we see a don’t see a non-constant pattern meaning that the linear model is a good representation of this data. We would also say that this normally distributed as many points fall on the expected line. There are also definitely some high leverage points that should be further evaluated which include 6, 10, 15.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

*Colorado:* From the plot we see that there is negative relationship between honey production and the price per lb of honey. The adjusted R-squared is 0.07642 which shows that there really isn’t a relationship between the honey production vs. the price per pound for the state of Colorado. The F-value is not considered statistically significant as the p value is 0.1656 showing that the relationship would not be considered significant. Now let’s look at the assumptions for this simple linear regression model. From the residuals vs. fitted plot we see a do see a non-constant pattern meaning that the linear model is not good representation of this data rather another model would be better. We would also say that this normally distributed as many points fall on the expected line. There are also definitely some high leverage points that should be further evaluated which include 9, 15.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

*Florida:* From the plot we see that there is negative relationship between honey production and the price per lb of honey. The adjusted R-squared is 0.5768 which shows that there is a strong relationship between the honey production vs. the price per pound for the state of Florida. The F-value is considered statistically significant as the p value is 0.0006186 showing that the relationship would be considered significant. Now let’s look at the assumptions for this simple linear regression model. From the residuals vs. fitted plot we see a do see a non-constant pattern meaning that the linear model is not good representation of this data rather another model would be better. We would also say that this normally distributed as many points fall on the expected line. There are also definitely some high leverage points that should be further evaluated which include 8,10, 15.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

*Georgia:* From the plot we see that there is slight negative relationship between honey production and the price per lb of honey. The adjusted R-squared is -0.04661 which shows that there really isn’t a relationship between the honey production vs. the price per pound for the state of Georgia. The F-value is not considered statistically significant as the p value is 0.5501 showing that the relationship would not be considered significant. Now let’s look at the assumptions for this simple linear regression model. From the residuals vs. fitted plot we see a do see a non-constant pattern meaning that the linear model is not good representation of this data rather another model would be better. We would also say that this normally distributed as many points fall on the expected line. There are also definitely some high leverage points that should be further evaluated which include 1, 11, 15.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

*Hawaii:* From the plot we see that there is negative relationship between honey production and the price per lb of honey. The adjusted R-squared is 0.005341 which shows that there really isn’t a relationship between the honey production vs. the price per pound for the state of Hawaii. The F-value is not considered statistically significant as the p value is 0.3187 showing that the relationship would not be considered significant. Now let’s look at the assumptions for this simple linear regression model. From the residuals vs. fitted plot we see a do see a non-constant pattern meaning that the linear model is not good representation of this data rather another model would be better. We would also say that this normally distributed as many points fall on the expected line. There are also definitely some high leverage points that should be further evaluated which include 2, 14, 15.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

*Idaho:* From the plot we see that there is negative relationship between honey production and the price per lb of honey. The adjusted R-squared is 0.2635 which shows that there is a relationship between the honey production vs. the price per pound for the state of Idaho. The F-value is considered statistically significant as the p value is 0.0291 showing that the relationship would be considered significant. Now let’s look at the assumptions for this simple linear regression model. From the residuals vs. fitted plot we see a do see a non-constant pattern meaning that the linear model is not good representation of this data rather another model would be better. We would also say that this normally distributed as many points fall on the expected line. There are also definitely some high leverage points that should be further evaluated which include 5, 7, 8.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

*Illinois:* From the plot we see that there is negative relationship between honey production and the price per lb of honey. The adjusted R-squared is 0.2467 which shows that there is a weak relationship between the honey production vs. the price per pound for the state of Illinois. The F-value is considered statistically significant as the p value is 0.03436 showing that the relationship would be considered significant. Now let’s look at the assumptions for this simple linear regression model. From the residuals vs. fitted plot we see a do see a non-constant pattern meaning that the linear model is not good representation of this data rather another model would be better. We would also say that this normally distributed as many points fall on the expected line. There are also definitely some high leverage points that should be further evaluated which include 14 and 15.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

*Indiana:* From the plot we see that there is negative relationship between honey production and the price per lb of honey. The adjusted R-squared is 0.3487 which shows that there is a strong relationship between the honey production vs. the price per pound for the state of Indiana. The F-value is considered statistically significant as the p value is 0.01206 showing that the relationship would be considered significant. Now let’s look at the assumptions for this simple linear regression model. From the residuals vs. fitted plot we see a do see a non-constant pattern meaning that the linear model is not good representation of this data rather another model would be better. We would also say that this normally distributed as many points fall on the expected line. There are also definitely some high leverage points that should be further evaluated which include 6,13, 15.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

*Iowa:* From the plot we see that there is negative relationship between honey production and the price per lb of honey. The adjusted R-squared is 0.2165 which shows that there is a weak relationship between the honey production vs. the price per pound for the state of Iowa. The F-value is considered statistically significant as the p value is 0.0459 showing that the relationship would be considered significant. Now let’s look at the assumptions for this simple linear regression model. From the residuals vs. fitted plot we see a do see a non-constant pattern meaning that the linear model is not good representation of this data rather another model would be better. We would also say that this is not normally distributed as some points don’t fall on the expected line. There are also definitely some high leverage points that should be further evaluated which include 15.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

*Kansas:* From the plot we see that there is negative relationship between honey production and the price per lb of honey. The adjusted R-squared is 0.506 which shows that there is a strong relationship between the honey production vs. the price per pound for the state of Kansas. The F-value is considered statistically significant as the p value is 0.001769 showing that the relationship would be considered significant. Now let’s look at the assumptions for this simple linear regression model. From the residuals vs. fitted plot we see a do see a non-constant pattern meaning that the linear model is not good representation of this data rather another model would be better. We would also say that this normally distributed as many points fall on the expected line. There are also definitely some high leverage points that should be further evaluated which include 7,10, 15.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

*Kentucky:* From the plot we see that there is no relationship between honey production and the price per lb of honey. The adjusted R-squared is -0.07488 which shows that there isn’t a relationship between the honey production vs. the price per pound for the state of Kentucky. The F-value is not considered statistically significant as the p value is 0.8776 showing that the relationship would not be considered significant. Now let’s look at the assumptions for this simple linear regression model. From the residuals vs. fitted plot we see a don’t see a non-constant pattern meaning that the linear model is good representation of this data. We would also say that this normally distributed as many points fall on the expected line. There are also definitely some high leverage points that should be further evaluated which include 2, 13, 14.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

*Louisiana:* From the plot we see that there is negative relationship between honey production and the price per lb of honey. The adjusted R-squared is 0.2915 which shows that there is not a very strong relationship between the honey production vs. the price per pound for the state of Louisiana. The F-value is considered statistically significant as the p value is 0.02201 showing that the relationship would be considered significant. Now let’s look at the assumptions for this simple linear regression model. From the residuals vs. fitted plot we see a relatively constant pattern meaning that this is a good representation of the data. We would say that this could be normally distributed as many points fall close on the expected line. There are also definitely some high leverage points that should be further evaluated which include 14 & 15.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

*Maine:* From the plot we see that there is negative relationship between honey production and the price per lb of honey. The adjusted R-squared is 0.1037 which shows that there is not a very strong relationship between the honey production vs. the price per pound for the state of Maine. The F-value is not considered statistically significant as the p value is 0.1295 showing that the relationship would not be considered significant. Now let’s look at the assumptions for this simple linear regression model. From the residuals vs. fitted plot we see a non-constant pattern meaning that another model would be better to represent this data. We would say that this could be normally distributed as many points fall close on the expected line. There are also definitely some high leverage points that should be further evaluated which include 5, 12, & 15.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

*Michigan:* From the plot we see that there is negative relationship between honey production and the price per lb of honey. The adjusted R-squared is 0.2715 which shows that there is not a very strong relationship between the honey production vs. the price per pound for the state of Michigan. The F-value is considered statistically significant as the p value is 0.02692 showing that the relationship would be considered significant. Now let’s look at the assumptions for this simple linear regression model. From the residuals vs. fitted plot we see a non-constant pattern meaning that another model would be better to represent this data. We would say that this is normally distributed as many points fall on or close on the expected line. There are also definitely some high leverage points that should be further evaluated which include 3, 15.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

*Minnesota:* From the plot we see that there is negative relationship between honey production and the price per lb of honey. The adjusted R-squared is 0.6069 which shows that there is a relationship between the honey production vs. the price per pound for the state of Minnesota. The F-value is considered statistically significant as the p value is 0.0003159 showing that the relationship would be considered significant. Now let’s look at the assumptions for this simple linear regression model. From the residuals vs. fitted plot we see a non-constant pattern meaning that another model would be better to represent this data. We would say that this could be normally distributed as many points fall close on the expected line. There are also definitely some high leverage points that should be further evaluated which include 3, 14, & 15.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

*Mississippi:* From the plot we see that there is positive relationship between honey production and the price per lb of honey. The adjusted R-squared is 0.4958 which shows that there might not be a very strong relationship between the honey production vs. the price per pound for the state of Mississippi. The F-value is considered statistically significant as the p value is 0.002035 showing that the relationship would be considered significant. Now let’s look at the assumptions for this simple linear regression model. From the residuals vs. fitted plot we see a non-constant pattern meaning that another model would be better to represent this data. We would say that this could be normally distributed as many points fall close on the expected line. There are also definitely some high leverage points that should be further evaluated which include 14.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

*Missouri:* From the plot we see that there is negative relationship between honey production and the price per lb of honey. The adjusted R-squared is 0.6665 which shows that there is a relationship between the honey production vs. the price per pound for the state of Missouri. The F-value is considered statistically significant as the p value is 0.0001247 showing that the relationship would be considered significant. Now let’s look at the assumptions for this simple linear regression model. From the residuals vs. fitted plot we see a non-constant pattern meaning that another model would be better to represent this data. We would say that this could be normally distributed as many points fall close on the expected line. There are also definitely some high leverage points that should be further evaluated which include 3 & 15.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

*Montana:* From the plot we see that there is slight negative relationship between honey production and the price per lb of honey. The adjusted R-squared is -0.002727 which shows that there is not a very strong relationship at all between the honey production vs. the price per pound for the state of Montana. The F-value is not considered statistically significant as the p value is 0.3446 showing that the relationship would not be considered significant. Now let’s look at the assumptions for this simple linear regression model. From the residuals vs. fitted plot we see a non-constant pattern meaning that another model would be better to represent this data. We would say that this could be normally distributed as many points fall close on the expected line. There are also definitely some high leverage points that should be further evaluated which include 2, 14, & 15.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

*North Carolina:* From the plot we see that there is slight positive relationship between honey production and the price per lb of honey. The adjusted R-squared is -0.01573 which shows that there is not a very strong relationship at all between the honey production vs. the price per pound for the state of North Carolina. The F-value is not considered statistically significant as the p value is 0.3924 showing that the relationship would not be considered significant. Now let’s look at the assumptions for this simple linear regression model. From the residuals vs. fitted plot we see a non-constant pattern meaning that another model would be better to represent this data. We would say that this could be normally distributed as many points fall close on the expected line. There are also definitely some high leverage points that should be further evaluated which include 11 & 15.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

*North Dakota:* From the plot we see that there is positive relationship between honey production and the price per lb of honey. The adjusted R-squared is 0.04785 which shows that there is not a very strong relationship between the honey production vs. the price per pound for the state of North Dakota. The F-value is not considered statistically significant as the p value is 0.2144 showing that the relationship would be considered significant. Now let’s look at the assumptions for this simple linear regression model. From the residuals vs. fitted plot we see a non-constant pattern meaning that another model would be better to represent this data. We would say that this could be normally distributed as many points fall close on the expected line. There are also definitely some high leverage points that should be further evaluated which include 5 & 15.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

*Nebraska:* From the plot we see that there is negative relationship between honey production and the price per lb of honey. The adjusted R-squared is 0.1916 which shows that there is not a very strong relationship between the honey production vs. the price per pound for the state of Nebraska. The F-value is not considered statistically significant as the p value is 0.0581 showing that the relationship would be considered not significant. Now let’s look at the assumptions for this simple linear regression model. From the residuals vs. fitted plot we see a non-constant pattern meaning that another model would be better to represent this data. We would say that this could be normally distributed as many points fall close on the expected line. There are also definitely some high leverage points that should be further evaluated which include 3, 4, & 15.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

*New Jersey:* From the plot we see that there is hardly any relationship between honey production and the price per lb of honey. The adjusted R-squared is -0.0732 which shows that there is not a very strong relationship between the honey production vs. the price per pound for the state of New Jersey. The F-value is not considered statistically significant as the p value is 0.8351 showing that the relationship would not be considered significant. Now let’s look at the assumptions for this simple linear regression model. From the residuals vs. fitted plot we see a non-constant pattern meaning that another model would be better to represent this data. We would say that this could be normally distributed as many points fall close on the expected line. There are also definitely some high leverage points that should be further evaluated which include 1 & 14.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

*New Mexico:* From the plot we see that there is negative relationship between honey production and the price per lb of honey. The adjusted R-squared is 0.3001 which shows that there is not a very strong relationship between the honey production vs. the price per pound for the state of New Mexico. The F-value is considered statistically significant as the p value is 0.02015 showing that the relationship would be considered significant. Now let’s look at the assumptions for this simple linear regression model. From the residuals vs. fitted plot we see a non-constant pattern meaning that another model would be better to represent this data. We would say that this could be normally distributed as many points fall very close to the expected line. There are also definitely some high leverage points that should be further evaluated which include 3 & 15.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

*New York:* From the plot we see that there is negative relationship between honey production and the price per lb of honey. The adjusted R-squared is 0.4913 which shows that there is not a very strong relationship between the honey production vs. the price per pound for the state of New York. The F-value is considered statistically significant as the p value is 0.002162 showing that the relationship would be considered significant. Now let’s look at the assumptions for this simple linear regression model. From the residuals vs. fitted plot we see a non-constant pattern meaning that another model would be better to represent this data. We would say that this is probably not normally distributed as many points do not fall on the expected line. There are also definitely some high leverage points that should be further evaluated which include 3, 5, & 15.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

*Ohio:* From the plot we see that there is negative relationship between honey production and the price per lb of honey. The adjusted R-squared is 0.3227 which shows that there is not a very strong relationship between the honey production vs. the price per pound for the state of Ohio. The F-value is considered statistically significant as the p value is 0.01593 showing that the relationship would be considered significant. Now let’s look at the assumptions for this simple linear regression model. From the residuals vs. fitted plot we see a non-constant pattern meaning that another model would be better to represent this data. We would say that this is not normally distributed as many points do not fall on the expected line. There are also definitely some high leverage points that should be further evaluated which include 12 & 15.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

*Oregon:* From the plot we see that there is negative relationship between honey production and the price per lb of honey. The adjusted R-squared is 0.08101 which shows that there is not a very strong relationship between the honey production vs. the price per pound for the state of Oregon. The F-value is not considered statistically significant as the p value is 0.1589 showing that the relationship would not be considered significant. Now let’s look at the assumptions for this simple linear regression model. From the residuals vs. fitted plot we see a non-constant pattern meaning that another model would be better to represent this data. We would say that this could be normally distributed as many points fall close on the expected line. There are also definitely some high leverage points that should be further evaluated which include 8 & 15.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

*Pennsylvania:* From the plot we see that there is negative relationship between honey production and the price per lb of honey. The adjusted R-squared is 0.4881 which shows that there is not a very strong relationship between the honey production vs. the price per pound for the state of Pennsylvania. The F-value is considered statistically significant as the p value is 0.002258 showing that the relationship would be considered significant. Now let’s look at the assumptions for this simple linear regression model. From the residuals vs. fitted plot we see a relatively constant pattern meaning that this model is a good representation of the data. We would say that this is not normally distributed as many points do not fall on the expected line. There are also definitely some high leverage points that should be further evaluated which include 14 & 15.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

*South Dakota:* From the plot we see that there is negative relationship between honey production and the price per lb of honey. The adjusted R-squared is 0.1287 which shows that there is not a very strong relationship between the honey production vs. the price per pound for the state of South Dakota. The F-value is not considered statistically significant as the p value is 0.1034 showing that the relationship would be considered not significant. Now let’s look at the assumptions for this simple linear regression model. From the residuals vs. fitted plot we see a non-constant pattern meaning that another model would be better to represent this data. We would say that this could be normally distributed as many points fall close on the expected line. There are also definitely some high leverage points that should be further evaluated which include 9 & 15.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

*Tennessee:* From the plot we see that there is negative relationship between honey production and the price per lb of honey. The adjusted R-squared is 0.008957 which shows that there is not at all a very strong relationship between the honey production vs. the price per pound for the state of Tennessee. The F-value is not considered statistically significant as the p value is 0.3078 showing that the relationship would not be considered significant. Now let’s look at the assumptions for this simple linear regression model. From the residuals vs. fitted plot we see a non-constant pattern meaning that another model would be better to represent this data. We would say that this could be normally distributed as many points fall close on the expected line. There are also definitely some high leverage points that should be further evaluated which include 6, 13, & 14.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

*Texas:* From the plot we see that there is negative relationship between honey production and the price per lb of honey. The adjusted R-squared is 0.2438 which shows that there is not a very strong relationship between the honey production vs. the price per pound for the state of Wyoming. The F-value is considered statistically significant as the p value is 0.0536 showing that the relationship would not be considered significant. Now let’s look at the assumptions for this simple linear regression model. From the residuals vs. fitted plot we see a non-constant pattern meaning that another model would be better to represent this data. We would say that this could be normally distributed as many points fall close on the expected line. There are also definitely some high leverage points that should be further evaluated which include 6, 14, & 15.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

*Utah:* From the plot we see that there is negative relationship between honey production and the price per lb of honey. The adjusted R-squared is 0.007596 which shows that there is not a very strong relationship between the honey production vs. the price per pound for the state of Wyoming. The F-value is not considered statistically significant as the p value is 0.3119 showing that the relationship would be considered not significant. Now let’s look at the assumptions for this simple linear regression model. From the residuals vs. fitted plot we see a non-constant pattern meaning that another model would be better to represent this data. We would say that this could be normally distributed as many points fall close on the expected line. There are also definitely some high leverage points that should be further evaluated which include 1, 4, &15.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

*Vermont:* From the plot we see that there is negative relationship between honey production and the price per lb of honey. The adjusted R-squared is 0.2659 which shows that there is not a very strong relationship between the honey production vs. the price per pound for the state of Vermont. The F-value is considered statistically significant as the p value is 0.02845 showing that the relationship would be considered significant. Now let’s look at the assumptions for this simple linear regression model. From the residuals vs. fitted plot we see a non-constant pattern meaning that another model would be better to represent this data. We would say that this could be normally distributed as many points fall close on the expected line. There are also definitely some high leverage points that should be further evaluated which include 6 & 13.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

*Virginia:* From the plot we see that there is negative relationship between honey production and the price per lb of honey. The adjusted R-squared is 0.4418 which shows that there is not a very strong relationship between the honey production vs. the price per pound for the state of Virginia. The F-value is considered statistically significant as the p value is 0.004102 showing that the relationship would be considered significant. Now let’s look at the assumptions for this simple linear regression model. From the residuals vs. fitted plot we see a non-constant pattern meaning that another model would be better to represent this data. We would say that this could normally distributed as many points fall on the expected line. There are also definitely some high leverage points that should be further evaluated which include 4 & 14.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

*Washington:* From the plot we see that there is a slight negative relationship between honey production and the price per lb of honey. The adjusted R-squared is -0.03566 which shows that there is not a very strong relationship between the honey production vs. the price per pound for the state of Washington. The F-value is considered statistically significant as the p value is 0.4844 showing that the relationship could be considered significant. Now let’s look at the assumptions for this simple linear regression model. From the residuals vs. fitted plot we see a non-constant pattern meaning that another model would be better to represent this data. We would say that this is normally distributed as many points fall on the expected line. There are also definitely some high leverage points that should be further evaluated which include 6 & 15.

|  |  |
| --- | --- |
|  |  |
|  |  |
| w |  |

*Wisconsin:* From the plot we see that there is negative relationship between honey production and the price per lb of honey. The adjusted R-squared is 0.6162 which shows that there is a relationship between the honey production vs. the price per pound for the state of Wisconsin. The F-value is considered statistically significant as the p value is 0.0003198 showing that the relationship would be considered significant. Now let’s look at the assumptions for this simple linear regression model. From the residuals vs. fitted plot we see a non-constant pattern meaning that another model would be better to represent this data. We would say that this is normally distributed as many points do fall on the expected line. There are also definitely some high leverage points that should be further evaluated which include 5 & 15.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

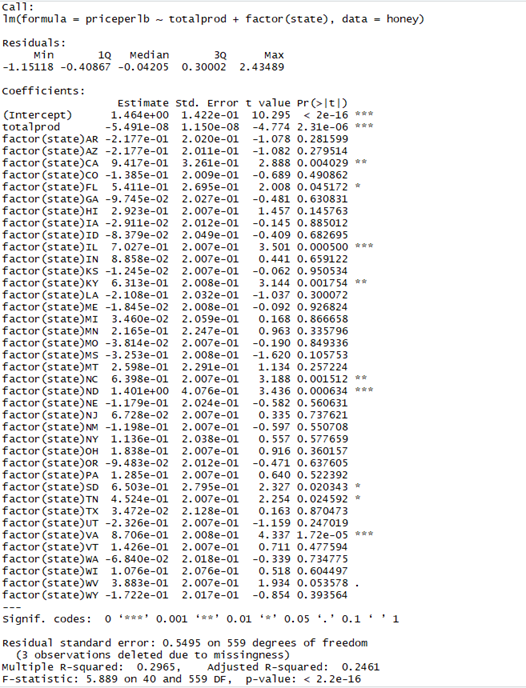
*West Virginia:* From the plot we see that there is negative relationship between honey production and the price per lb of honey. The adjusted R-squared is 0.362 which shows that there is not a very strong relationship between the honey production vs. the price per pound for the state of West Virginia. The F-value is considered statistically significant as the p value is 0.0104 showing that the relationship would be considered significant. Now let’s look at the assumptions for this simple linear regression model. From the residuals vs. fitted plot we see a relatively constant pattern meaning that this model is a good representation of the data. We would say that this could be normally distributed as many points fall close on the expected line. There are also definitely some high leverage points that should be further evaluated which include 1 & 15.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

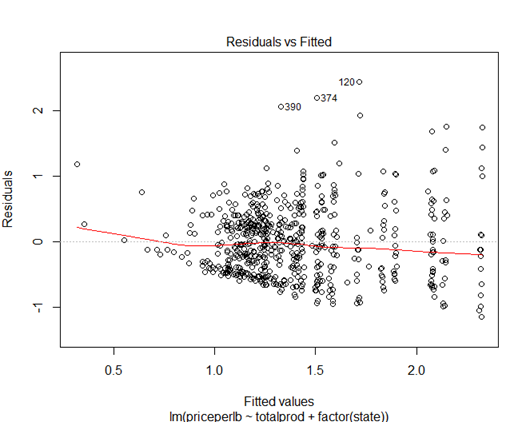
*Wyoming*: From the plot we see that there is negative relationship between honey production and the price per lb of honey. The adjusted R-squared is 0.2801 which shows that there is not a very strong relationship between the honey production vs. the price per pound for the state of Wyoming. The F-value is considered statistically significant as the p value is 0.0247 showing that the relationship would be considered significant. Now let’s look at the assumptions for this simple linear regression model. From the residuals vs. fitted plot we see a relatively constant pattern meaning that this model is a good representation of the data. We would say that this is not normally distributed as many points do not fall on the expected line. There are also definitely some high leverage points that should be further evaluated which include 15.

**Multiple Regression**

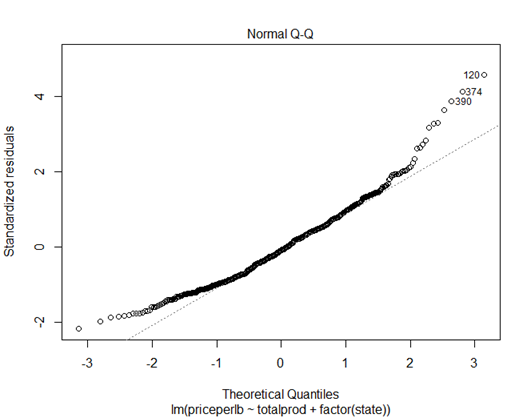
A model was developed to examine the relationship between state and honey production vs. price of honey per pound. Since state is a categorical variable it was considered the dummy variable.



We can observe that of all the states only 9 are considered significant and honey production is also considered significant. The adjusted R-squared is 0.2461 which shows that there isn’t really a strong relationship between the state and honey production vs. the price per pound. On the contrary, the F-value is considered statistically significant as the p value is < 2.2e-16 showing that the relationship would be considered significant. Given this information it can’t be concluded if the relationship would be considered linear. Now let’s check the assumptions for this multiple regression model. From the residuals vs. fitted plot we see that this is a random plot meaning that a linear model would be accurate for this dataset.



From the normal probability plot we see that there are quite a few values not found near the expected line, so we would say that this isn’t normally distributed.



**Conclusion**

In conclusion, honey production is significantly different among states and between the years of 1998-2012 as well. When examining if there is a linear relationship between honey production and the cost of honey, we hypothesized that there is a relationship. We see that various states show a negative relationship but about 3 states (North Dakota, North Carolina, and Mississippi) showed a slightly positive relationship with a non-significant R-squared value. It is important to note that all these models may not be the best representation for this dataset due to low sample size. We wanted to further examine if the states would also influence this linear relationship through a multiple regression model and determined that there was a weak relationship, meaning that state and the amount of honey produced could affect the price of honey per pound. Further analysis can be conducted to determine what the best model for this dataset would be.