INTRODUCTION

Our group is focusing on the Iowa House File dataset. In this data set our intended population is all single-family housing units in Iowa. The actual population is all single-family housing units in Iowa in 2015. These units were randomly selected from Iowa Housing data of all housing units that were single-family homes. In the original dataset, one of our concerns was the amount of missing data for our variables. Another concern for our data set was the confusing variable names. However, we then found a dictionary variable for this dataset. Our response variable is monthly rent (RNTP). Our three explanatory variables are lot size (ACR), household income in the past 12 months (HINCP), and vehicles (VEH). We are using our explanatory variables to predict monthly rent. Lot size (ACR) is our categorical variable and is an ordinal variable. Household income (HINCP) is a quantitative variable and is a continuous variable. Vehicles (VEH) is a quantitative variable and is an ordinal variable. Our categorical variable (ACR) has a scale of 1 through 3 with 1 representing a house on less than one acre. Two representing a house on one to ten acres and 3 representing a house on greater than ten acres.

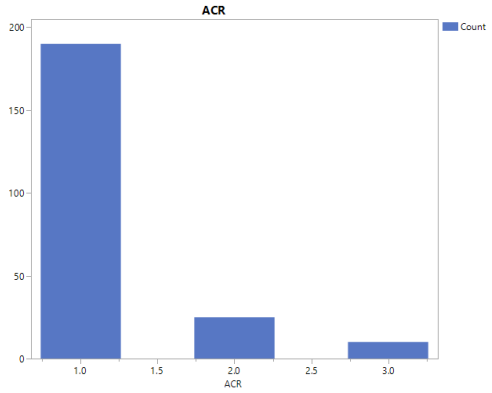
METHODS

We used JMP 15.0 software to perform our statistical analysis. This includes the Analyze Distribution and Fit Model functions. Our regression was fitted with 223 rows of complete data. Rows with missing data were omitted. We do not know why some values, especially ACR, are missing. ACR 2 is an indicator variable for acreage type 2 and ACR 3 is an indicator variable for acreage type 3. ACR type 1 was coded as 0. We will interpret the JMP output in the results section.

RESULTS

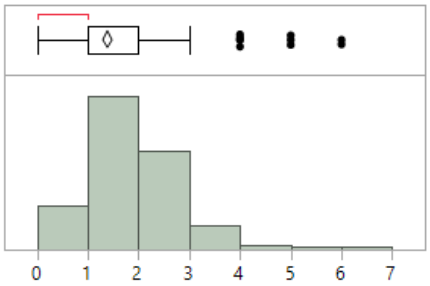
See the following pages. There are Summary Statistics, Scatterplot Matrix, Fitted Regression, Assessment of Assumptions, and Conclusion sections.

Summary Statistics



These graphs are the summary statistics and graphs for RNT, ACR, VEH, and HINCP. All variables are clearly right-skewed. There are many outliers.

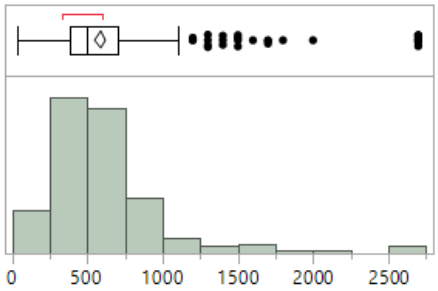
VEH



Quantiles		
100.0%	maximum	6
99.5%		5.595
97.5%		3
90.0%		2
75.0%	quartile	2
50.0%	median	1
25.0%	quartile	1
10.0%		0
2.5%		0
0.5%		0
0.0%	minimum	0

Summary Statistics	
Mean	1.3916667
Std Dev	0.9345677
Std Err Mean	0.042657
Upper 95% Mean	1.4754846
Lower 95% Mean	1.3078487
N	480

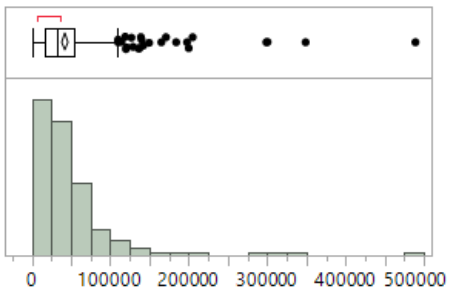
RNTP



Quantiles		
100.0%	maximum	2700
99.5%		2700
97.5%		1647.5
90.0%		900
75.0%	quartile	700
50.0%	median	500
25.0%	quartile	380
10.0%		250
2.5%		125.25
0.5%		40
0.0%	minimum	30

Summary Statistics	
Mean	584.1
Std Dev	383.01435
Std Err Mean	17.128922
Upper 95% Mean	617.7537
Lower 95% Mean	550.4463
N	500

HINCP

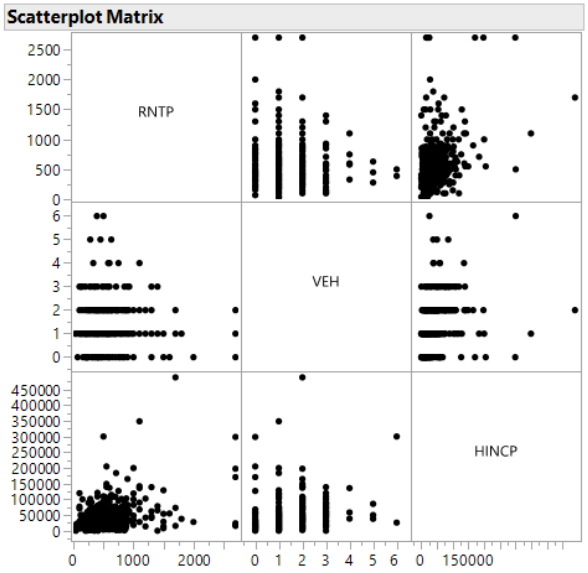


Quantiles		
100.0%	maximum	489000
99.5%		329357.5
97.5%		141950
90.0%		83000
75.0%	quartile	53950
50.0%	median	32000
25.0%	quartile	16800
10.0%		8800
2.5%		1707.5
0.5%		0
0.0%	minimum	0

Summary Statistics	
Mean	42815.79
Std Dev	44539.352
Std Err Mean	2032.934
Upper 95% Mean	46810.36
Lower 95% Mean	38821.219
N	480

Scatterplot and Correlation Matrix

There are no obvious correlations from the scatterplot matrix. However, HINCP seems positively correlated with RNTP. We will refer to the correlation matrix for a more precise analysis. We see a weak correlation between HINCP and RNTP. There seems to be no correlation between VEH and RNTP and HINCP and VEH.



Fitted Regression

Our untransformed multivariate regression line is,

$$\hat{y} = 479.060 \; - \; 156.127 \bullet ACR\,2 - 117.800 \bullet ACR\,3 - 22.361 \bullet VEH + 0.00239 \bullet HINCP$$

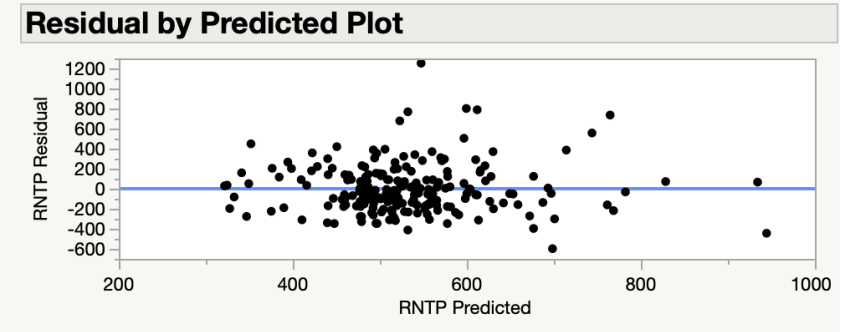
where acreage type 1 is our base group, ACR 2 is an indicator variable for acreage type 2 and ACR 3 is an indicator variable for acreage type 3.

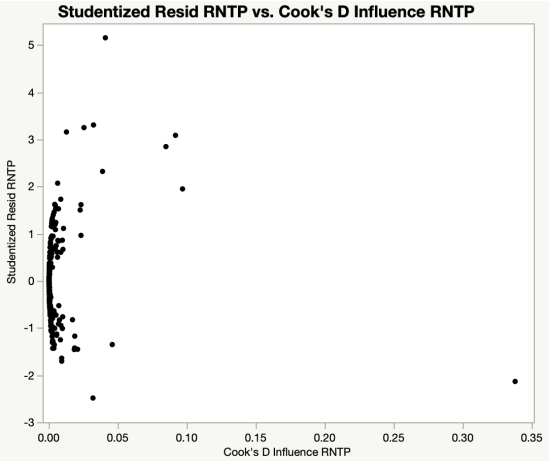
From our equation, we have that units of acreage type 2 are predicted to be $156.13 cheaper than units of type 1. For units of type 3, we predict that they will have a predicted price of $117.80 cheaper than that of type 1. We also have that for every additional car that a family has, the predicted rental price decreases by $22.36. Finally, for every additional dollar of income that a household makes per year, we expect the rental price to increase by 2.39 cents.

Assessment of Assumptions

Residuals

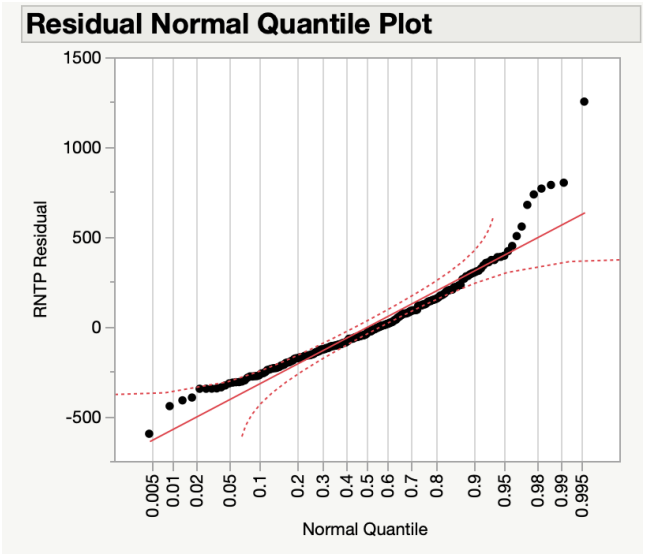
According to our residual plot, we appear to have equal variances across all of the data. The variances appear to have outliers around the cluster of points. Additionally, there is not an obvious pattern in the residuals, so we conclude that the data is linear.





Studentized Residuals, Cook's D

From our studentized residual Cook's D Influence graph, these appear to be outliers with studentized values greater than 3. From this graph, we can also conclude that we have no influential points since all Cook's D influence measures are under 1.



Normal Quantiles

We do appear to have mostly normal data though as shown in the normal quantile plot.

Independence

Since the data was a random subset of a larger data set, selected from rows with non-null RNTP values, we assume independence.

We conclude that we do not need to perform any transformations, and our assumptions are met.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	1746840	582280	9.7294
Error	219	13106623	59848	Prob > F
C. Total	222	14853464		<.0001*

Summary of Fit

RSquare	0.117605
RSquare Adj	0.105517
Root Mean Square Error	244.6377
Mean of Response	530.4036
Observations (or Sum Wgts)	223

Analysis of Variance

We can conclude that at least one of our coefficients is nonzero since we obtained a p-value of <0.0001 in our multivariate F-test.

Summary of Fit

RMSE is 244. 11.76% of the total variability in the data is explained by the model.

Conclusions

We conclude that the model is not a great fit for the data, as only 11.76% of the total variability in the data is explained by the model. It also has a large Root Mean Square Error of 244, which is highly significant considering that the IQR for response RNTP is only 320.

We know that the model is valid because the prerequisite assumptions of linearity, independence, equal variances, and normality were met. Additionally, the F-Test gave a p-value of <0.0001, meaning the overall model is statistically significant. The estimated parameters are the following: Intercept of 570 with a standard error of 48, ACR slope of -93 with a standard error of 34, VEH slope of -24 with a standard error of 19, and HINCP slope of 0.0024 with a standard error of 0.0005. Confidence intervals for each term can be found in the table below. All estimates were statistically significant, that is, likely to not be zero, except for VEH which had a p-value of 0.2186.

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%
Intercept	569.68741	47.87663	11.90	<.0001*	475.32951	664.04532
ACR	-93.4286	33.66199	-2.78	0.0060*	-159.7715	-27.08568
VEH	-23.80037	19.28953	-1.23	0.2186	-61.81725	14.216509
HINCP	0.0024166	0.000498	4.86	<.0001*	0.001436	0.0033972