Patrick Keener
SID: 1385832
DSC 423 – Assignment 2

Assignment 2

Honor statement: I have completed this work independently. The solutions given are entirely my own work.

1) Short Essay

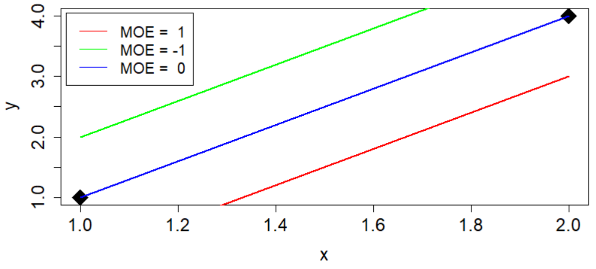
1a) When building a model, what are the assumptions about the residuals and how does one verify them?

When building a model, four assumptions are made about the residuals (the difference between the model's projection and the observation): The Mean of errors is 0, homoscedasticity, normality, and independence[1].

Mean of Errors

The first assumption is that the Mean of the Errors ("MOE") is 0. If the mean of the errors is 0 then it means that the model is plotting a line that goes exactly through the center of the data.

Figure 1 Lines fit to Black Diamonds

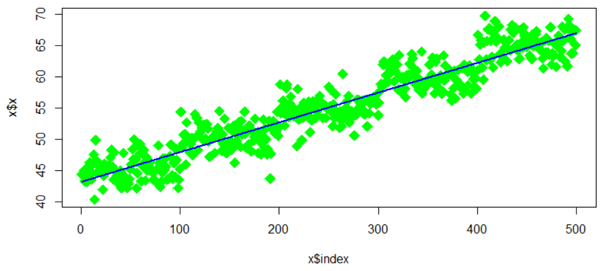


Figure[2] 1 shows 2 diamonds, which represent the underlying data, and 3 lines. The red line has an MOE of 1, which means it is underestimating the data. The Green line has an MOE of -1, which means it is overestimating the data. Finally, the Blue line has an MOE of 0, meaning it fits the data perfectly.

An Ordinary Least Squares ("OLS") model will always fit a line with an MOE of 0, however for other models the MOE can be calculated by subtracting the forecasted points (red/green/blue lines) from the observed points (black diamonds) then averaging the results.

Homoscedasticity

Figure 2 Homoscedastic Data

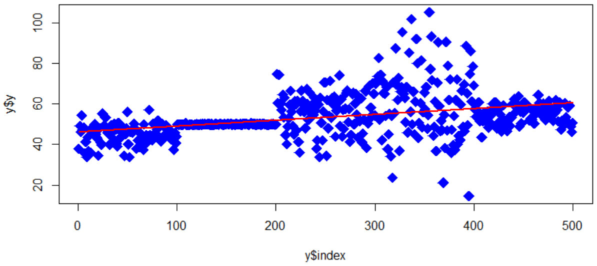


The second assumption around residuals is Homoscedasticity, which means that the variance (how much we expect the magnitude of observations to vary) is stable across time. Figure 2 shows a homoscedastic data set. The most important feature to distinguish homoscedasticity is that the *size* of the residuals doesn't change over time. The challenge of having a data set that is not homoscedastic (called "heteroscedastic") such as in figure 3, is that the accuracy of the forecast changes depending on the variance at that particular point in the graph, resulting in confidence intervals that are too wide or too narrow[3]. Further, since the variance is inconsistent, tests

[1] CSC 423 slides, Module 2, Lecture 3, Slide 2.
[2] Unless otherwise noted, graphs were generated for this exercise by me (Patrick Keener) in R.
[3] Duke University's Regression Diagnostics Web Page: http://people.duke.edu/~rnau/testing.htm

that rely on variance and therefore standard deviation are no longer useful measures for assessing a model.

Figure 3 Heteroscedastic Data

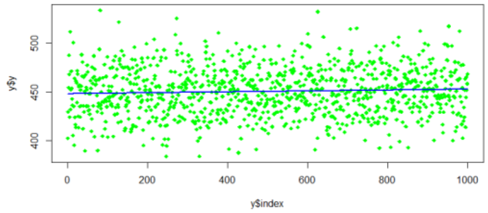


Since this characteristic has such important implications it falls to the modeler to confirm the assumption. Several tests[4] may be performed to this end, including Bartlett's Test, which tests whether variance across samples is equal, and Levene's test, which gives the likelihood that the variance is stable across time.

Normality

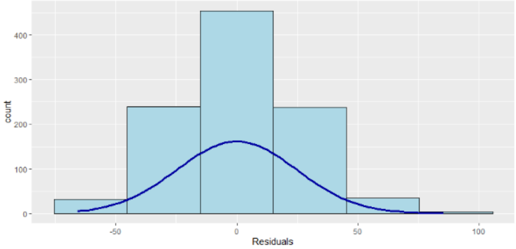
The next assumption is normality. This doesn't mean that the whole data set is normally distributed, but rather the residuals themselves are normally distributed. This assumption is important for testing whether a variable is useful, as well as calculating confidence intervals. Violation of this assumption can result in extreme points changing the forecast significantly.

Figure 4 Normally Distributed Residuals



Normally distributed residuals can be identified in several ways. If a clear relationship exists then a simple histogram can be used to confirm the relationship. The expectation is that the bars in the graph have a similar shape to the solid blue line, which in figure[5] 5 is the case. Alternatively, several tests exist to aid in this. The Shapiro-Wilk[6] test directly tests whether the sample came from a normally distributed population. The Kolmogorov-Smirnov test will give the likelihood that the distribution of a sample (in this case, the residuals) is equal to the distribution of another sample (for example, a normal distribution).

Figure 5 Distribution of Residuals from Figure 4



For contrast, figures 6 and 7 display residuals that are not normally distributed. The histogram does not resemble a normal distribution.

Figure 7 Residuals Not Normally Distributed

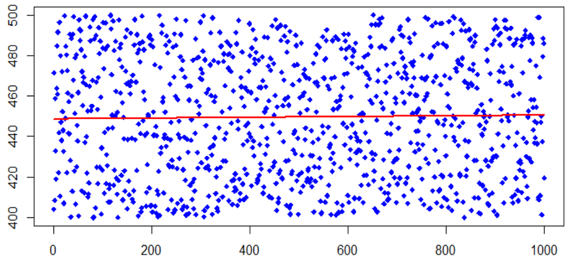
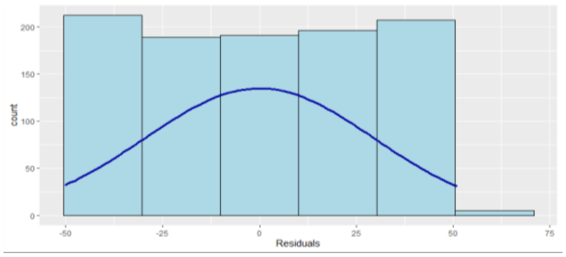


Figure 6 Distribution of Residuals from Figure 7



[4] Both tests for heteroscedasticity were found on the CRAN web page: https://bit.ly/3eTivBM

[5] Figures 5 and 7 were generated using the OLSRR package in R

[6] The tests for normality were found on the CRAN web page: https://bit.ly/2xbon8F

Independence[7]

The final assumption is that residuals are independent, which means that the residuals don't influence each other. If independence is violated it may mean a few things: for time series data it implies that there is more information left in the data that can be modeled. In other types of models, it may indicate that some variables are interacting in a way that isn't modeled.

Independence[8] can be tested by looking at autocorrelation (whether movement in the residual is correlated with the same residual at a different point) and multicollinearity (the tendency of variables to behave similarly to each other), which can be measured using a Variance Inflation Factor ("VIF") test, which will help identify potential areas for further investigation.

1b) Define interaction term

An interaction term is a term where two variables directly influence each other. For example, maybe both the weight and torque of a vehicle independently influence acceleration, but together they likely influence it more than individually- a "greater than the sum of their parts" type situation.

Interaction terms often lend a convex or concave shape to an otherwise straight/linear forecast, and the severity and direction of the shape is dictated by the beta of the interaction term.

[7] No graphs here- couldn't figure out how to effectively show independence or non-independence of residuals visually

[8] Duke University's Regression Diagnostics Web Page: http://people.duke.edu/~rnau/testing.htm

2) Banking

2a) nothing to submit

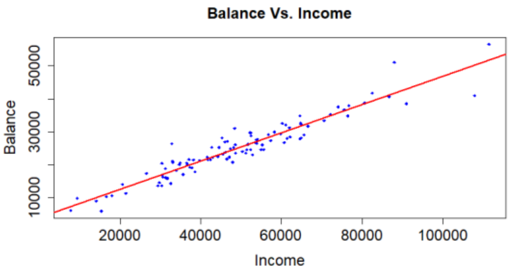
2b) nothing to submit.

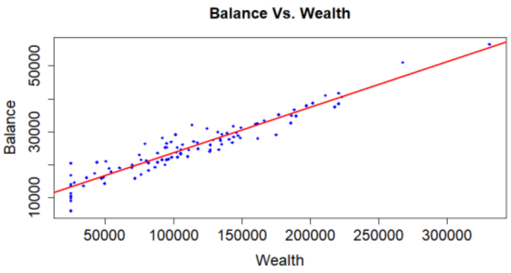
102 obs., 6 variables

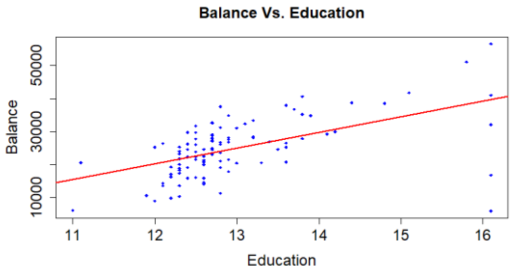
	Age	Education	Income	HomeVal	Wealth	Balance
Min.	19.50	11.00	7,741	40,313	24,999	5,956
1stQu.	33.92	12.40	35,078	83,017	70,263	20,036
Median	36.10	12.70	47,656	97,744	102,348	24,661
Mean	35.45	12.98	48,811	106,845	109,026	24,888
3rdQu.	37.58	13.20	60,157	121,791	142,518	29,180
Max.	43.10	16.10	111,548	276,139	331,009	56,569

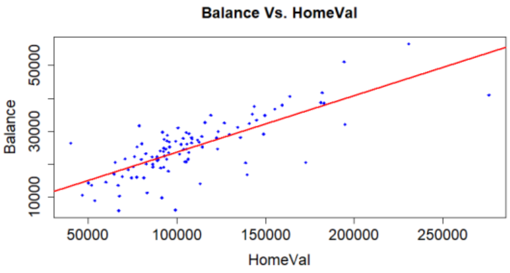
2c) Create scatterplots between bank balance and the other 5 variables

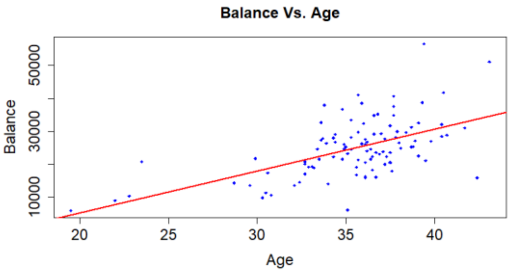
Initial thoughts on seeing the graphs are that wealth and income are the best predictors- points seem close to the linear regression line (red).


Balance Vs. Income


Balance Vs. Wealth


Balance Vs. Education

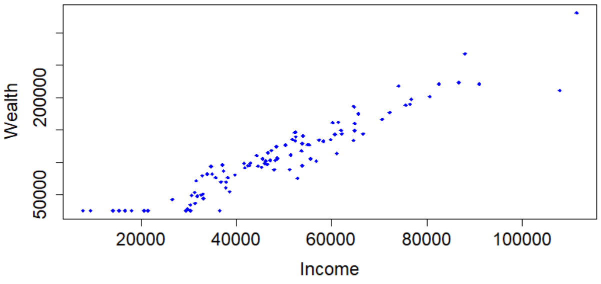

Balance Vs. HomeVal


Balance Vs. Age

2d) Compute Correlations found in bank data and include in submission. Describe strongly associated variables.

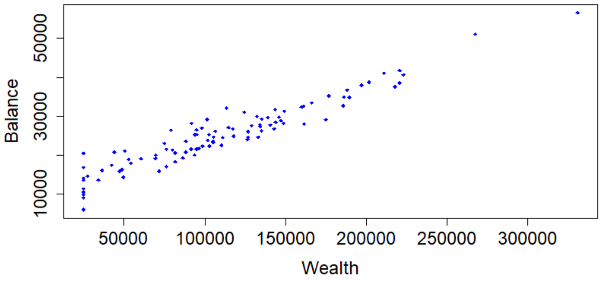
The correlation matrix for the Banking data set can be found in Table 1 (below), color-coded from highest correlation (green) to lowest correlation (red). The most strongly correlated variables are the correlations between income and wealth (Figure 8), between wealth and balance (Figure 9), and income and balance (Figure 10). All of these variables have correlations between .94 and .95, and are correlated with each other. This suggests the possibility that the process that generates one or more of these variables includes one or more of the other variables, resulting in the potential multicollinearity.

Figure 8 Income & Wealth, Correlation .9467



Logically, these variables are closely related, and while they each may be part of the others' process, there may still be enough additional information to warrant retention of the variables. The collinearity should be evaluated using a VIF test to give context to the decision.

Figure 9 Wealth & Balance, Correlation .9487



Wealth generates income and a certain amount of wealth is likely to be kept as a balance. Income, under proper stewardship, leads to wealth. Finally, a higher income would drive a higher balance, even if it's used just as a place to park the cash until it can be put to more productive use.

Figure 10 Income & Balance, Correlation .9517

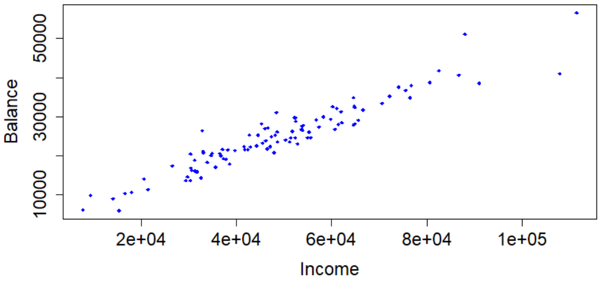


Table 1 Correlation Matrix, Banking Data Set

	Age	Education	Income	HomeVal	Wealth	Balance
Age						
Education	0.1735					
Income	0.4771	0.5731				
HomeVal	0.3865	0.7489	0.7954			
Wealth	0.4681	0.4681	0.9467	0.6985		
Balance	0.5655	0.5522	0.9517	0.7664	0.9487	

2e) Fit a regression model of balance vs the other 5 variables. Present the regression & evaluate it.

This model is a 1st order linear regression model which consists of 5 predictor variables (Age, Education, Income, HomeVal, and Wealth) and attempts to predict Balance. This model, referred to as Model 1, has significant predictive power.

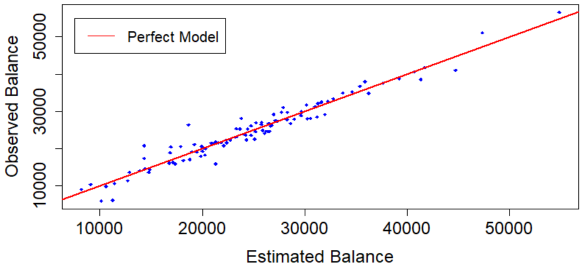
Model-level analysis (table 2) shows an adjusted R-Squared of .944, which suggests the model explains 94.4% of the variance in the data. Further, the f-statistic is 341.4 leading to a p-value approaching 0 (2e-16). Finally, the residual standard error is around 2,000. Since balance ranges from around 6,000 up to 56,000, the distribution of the errors is important- an error of 2,000 when the balance is 6,000 is meaningful, whereas a value of 2,000 when the balance is 56,000 is much less concerning.

Table 2 Model Output Analysis, Model 1

Adjusted R-Squared	0.944		Min	-5365.5
F-Statistic	341.4		1Q	-1102.6
P-value	0		Median	-85.9
RSE	2059		3Q	868.9
			Max	7746.5

A comparison of the model estimates and the observed balances (Figure 11) shows normally distributed residuals. The importance of homoscedasticity in the residuals is highlighted here: if variance is constant throughout the model then we are just as likely to see a variance of 30% when the value is 6,000 as we are to see a variance of 2% when the value is 56,000; the model itself becomes less reliable on smaller balances. This does not preclude the model from use, but is merely a weakness that users must be attentive to.

Figure 11 Estimated Balance vs Observed Balance



2f) Which of the five predictors have a significant effect on balance? Explain. (α = .05)

The method used for variable evaluation (table 3) will be a two-step process: first, disqualify p-values that don't meet the criteria (α = .05), next calculate the beta weights to determine which variables have significant effects. Beta weights were chosen because they calculate a standardized impact on the model for each variable. This is accomplished by multiplying the beta by one standard deviation: $\text{ß}_i * \sigma_i$ where i is the subject variable. Beta weights do suffer from one significant weakness in that they don't account for suppressing effects of variables.

Table 3 Variable Analysis, Model 1

	Estimate	Std. Error	t value	Pr (>|t|)
(Intercept)	-10331.37219	4219.459063	-2.449	0.01616
banking$Age	317.458452	61.037332	5.201	0.00000
banking$Education	590.281539	315.121208	1.873	0.06409
banking$Income	0.146844	0.040832	3.596	0.00051
banking$HomeVal	0.009864	0.010989	0.898	0.37159
banking$Wealth	0.074142	0.0112	6.62	0.00000

Table 3is truncated; however, there are 3 variables with sufficient p-values in the current model[9]. The candidate variables are: Age (1.12e-06), Wealth (2.06e-09), and Income (5.12e-04). Comparing p-values alone we might rank importance as: Wealth, Age, then Income; however, this only accounts for the likelihood that a relationship exists, not necessarily the importance of the relationship. To help correct for this, the next step is to calculate the beta weights (Table 4).

Table 4 Beta Weight Calculation for Model 1 Candidate Variables

	ß	σ	Beta Weight
banking$Age	317.458	3.886	1233.795
banking$Income	0.147	19361.88	2843.175
banking$Wealth	0.074	59836.6	4436.405

Note: values are calculated using 16 decimal precision, but displayed at 3

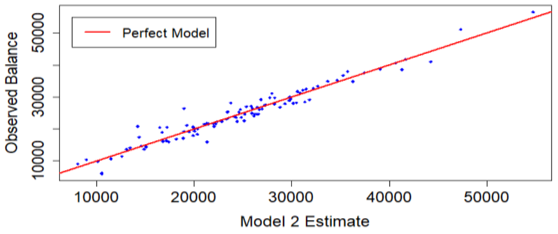
Using beta weights, the importance of each variable is ordered as: Wealth, Income, Age (most to least).

<u>2g) Remove the variable with the largest p-value > α * refit model. Present new model.</u>

The next step is to continue fitting the model. The highest p-value belongs to Homeval (.37159). Removing this variable creates Model 2, and has the effect of *increasing* the adj-R^2 from .944 to .9441, and the residual standard error falls from 2059 to 2057, a minimal change but in-line with the miniscule changes to adj-R^2. Further, the p-values for wealth, age, and income go down, while their betas all increase.

A visual analysis of the new model output (Figure 12) shows almost no difference, which was to be expected. While it appears that almost no change has occurred, removal of this variable increased the computational efficiency of this model while removing suppressing effects of the variable and potential complications caused from multicollinearity[10].

Figure 12 Estimated balance vs Observed Balance, Model 2

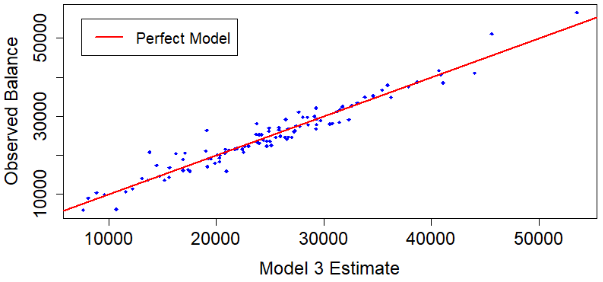


[9] Typically, p-values would be iteratively re-assessed each time a variable is removed, however that wasn't done here in an effort to stick to the assignment.
[10] VIF test shows that there was no collinearity.

2h) Analyze if all four predictors have a significant association with balance (α = .05). If not, continue to remove one insignificant variable at a time until all predictors are significant.

Continuing down the model building road[11], no additional variables fail the t-test. Education went from a p-value of .064 to .004 after removing HomeVal and so met the alpha threshold. However, with a standard deviation of 1.01 and a beta of 750, the beta weight is only 756.9 which is much lower than the beta weight the other variables. The model was tested without Education and the adj-R^2 fell from .9441 to .9399. In an effort to build a parsimonious model, a reduction in accuracy of .42% was deemed reasonable: The Education variable will be removed which will create Model 3.

Figure 13 Estimated balance vs Observed Balance, Model 3



It should be noted that in this new model (Model 3) the p-value of income, which had been the highest of the 3 original significant variables, is now the *lowest* of the 3. Furthermore, its beta increased by approximately 30% from .16 to .21 with corresponding effects on beta weight. It should also be noted that the f-statistic of the model increased from 427.4 to 527.7.

Continuing with the process, the variable with the next lowest beta weight is Age. In the new model, Age's beta weight is 1173.223 vs a beta weight of 4103.556 for Income and 3818.173 for Wealth. Since the beta weight is high it is expected that the model will have significant degradation in accuracy and therefore will not warrant removal of the variable.

The removal of the Age variable results in the adj-R^2 value falling from .9399 to **.9262**. Given that so little accuracy is lost while removing 1/3 of the variables, it makes sense to remove this variable as well. At this point only two variables remain: Income and Wealth, which is model 4. For the final step of variable analysis beta weights will be ignored since there are only two remaining variables. Instead, Income and Wealth will be fit to a single-variable linear model for comparison, creating models 5a and 5b, respectively.

Figure 15 Estimated vs Observed Balance, Model 5a (Income)

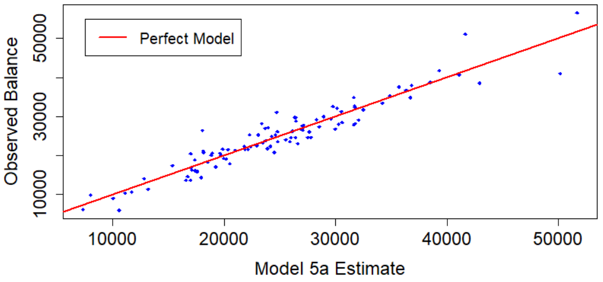
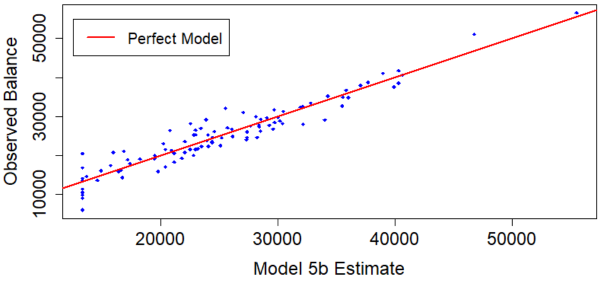


Figure 14 Estimated vs Observed Balance, Model 5b (Wealth)



The model containing only income has an adj-R^2 of .90 vs the wealth model which has an adj-R^2 of .89. It will cost 2.6% accuracy to drop down to a single variable (income). *This makes sense*, given our understanding of money flows. We assume that the simplified model of *balance* is inflows (income) – outflows (expenses) – investments (wealth), summed since the account opened. It makes sense that we would predict the balance by predicting the factor that forms the upper bound of the balance. Likewise,

[11] P-values, beta-weights, and adj-R^2 statistics are always recalculated after each variable is removed.

wealth is the accumulation of the surplus of income – expenses removed from the account. It makes sense that the cumulative excess income is a strong predictor of balance.

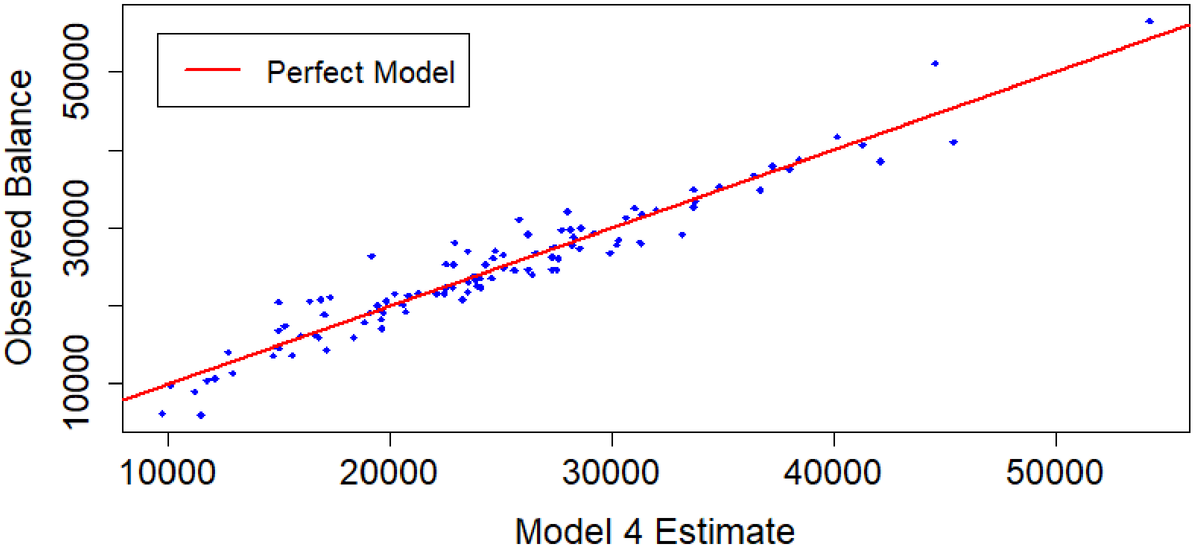
The choice of whether to reduce to a single variable model driven by income will depend primarily on whether there is a clear line of where the marketing is targeted to. For example, if only balances above 30,000 will be marketed to, the additional accuracy is unnecessary and the income-only model can be used. If instead the cut off is around 27,000 where a dense cluster of accounts exists additional accuracy may warrant.

2j) Discuss the adj-R^2 for the final model. (I have swapped 2i and 2j since the R^2 was the focus of the analysis at this point in the narrative).

Model 4 will be the subject of the remainder of the analysis, which contains Income and Wealth, has an R^2 of .9262, and a f-statistic of 635.1, indicating that the model explains approximately 92.6% of the variance and it is almost certain that at least one variable has a beta that is not 0.

2i) Interpret each of the regression coefficients for the final model.

Figure 16 Estimated Balance vs Observed Balance, Model 4



The final model has an intercept of 6,280, indicating that the lowest account balance on record is $6,280; this forms the model baseline. The Income beta is .232, which means that for each dollar of income, 23.2 cents is held as balance on average. This may occur for several reasons:

- Their income exceeds their expenses and investments by 23.2% and is held in their bank account
- A large portion of their income is held in their account prior to being removed for expenses

This relationship could be explored (in a separate analysis) for additional insights.

The wealth beta is .067, which means that for each dollar of wealth, 6.7 cents is held as balance in the account. This may be indicative of an explicit investment strategy that includes cash, is a transitory amount that is held between when the income enters into the account and the income leaves the account to go into investments, or a combination of both. Table 5 is a summary of the variable level results.

Table 5 Model 4 Summary

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6,280.000	758.6000	8.279	6.05E-13
banking$Income	0.232	0.0377	6.152	1.63E-08
banking$Wealth	0.067	0.0122	5.488	3.14E-07

In section 2d the possibility of collinearity between income and wealth was alluded to, and in section 2j the relationship between income, wealth, and balance was described, implying a relationship where each variable contributed to the other, which raises the possibility of multicollinearity. To measure this, a Variable Inflation Factor ("VIF") test can be used to gauge the relative likelihood of collinearity for each variable.

In the financial industry the standard for VIF is as follows: a VIF > 5 should be investigated, while a VIF > 10 indicates significant collinearity and the variable should be evaluated for omission.

When a VIF test is conducted on this model, Income and Wealth score a 9.631 (since there are only 2 variables, they share a score). On model 2, which is the largest model where all variables pass the t-test, the VIF of wealth is 10.5, and the VIF of income comes in at 12.5.

Personal Note: After conducting this analysis I have concluded that I should have explored collinearity and other assumptions before deciding on a model. Had I done this, I would have discarded model 4 in favor of a new model (model 6) which contains income and age and has an adj-R^2 of .9202 which is slightly worse than model 4, however the VIF score is only 1.29.

2k) Are there any influential points in the dataset? Explain the potential impact.[12]

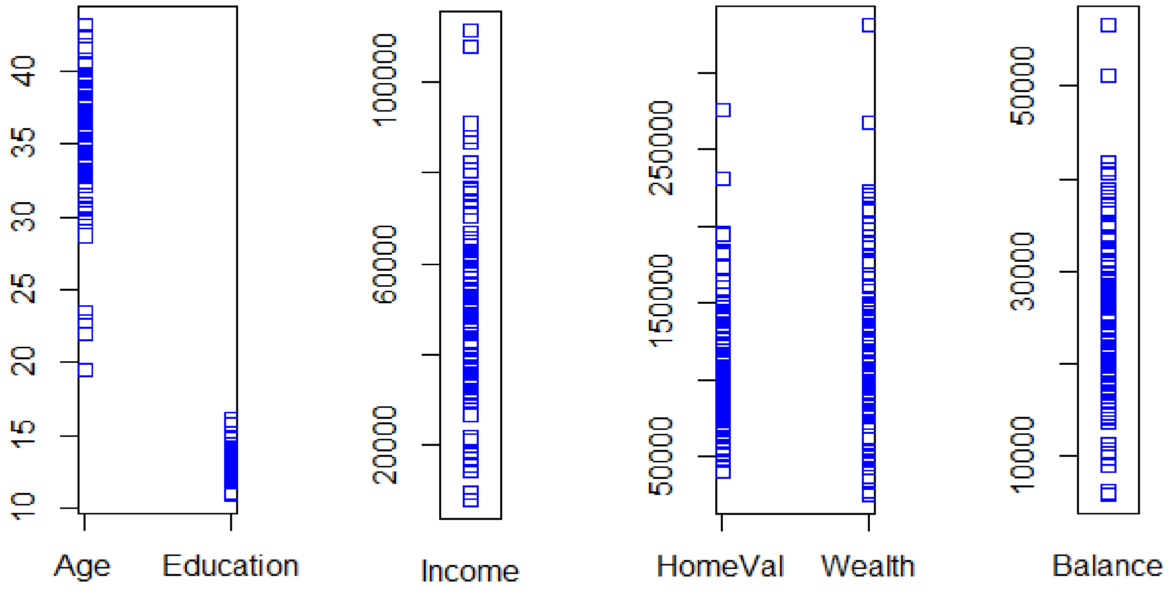
Every data point has some impact on the overall fit of the model. Even a miniscule deviation from the mean while fitting the model will cause an adjustment to the slope (beta) for that variable. Since, when generating a model, the slopes of all variables are derived in concert, every observation impacts the beta of every other variable to some degree. This process implies several things:

- First, more observations reduce the impact that any individual observation has on the model, which implies that the larger the dataset, the more extreme the variable must be relative to its peers to influence the model.
- Second, it is possible in a model with multiple variables to experience suppression or exacerbation of betas due to the data of unrelated variables, and influential values may magnify this effect.

[12] Definitions of outliers and leverage points: Penn State website, https://online.stat.psu.edu/stat462/node/170/

One of the ways to identify outliers is through the use of strip charts (Figure 17). The presence of extreme values in the dataset can be easily identified through a quick look at a strip chart, which will help target the analysis.

Figure 17 Strip Charts for the Banking Dataset



For each variable:

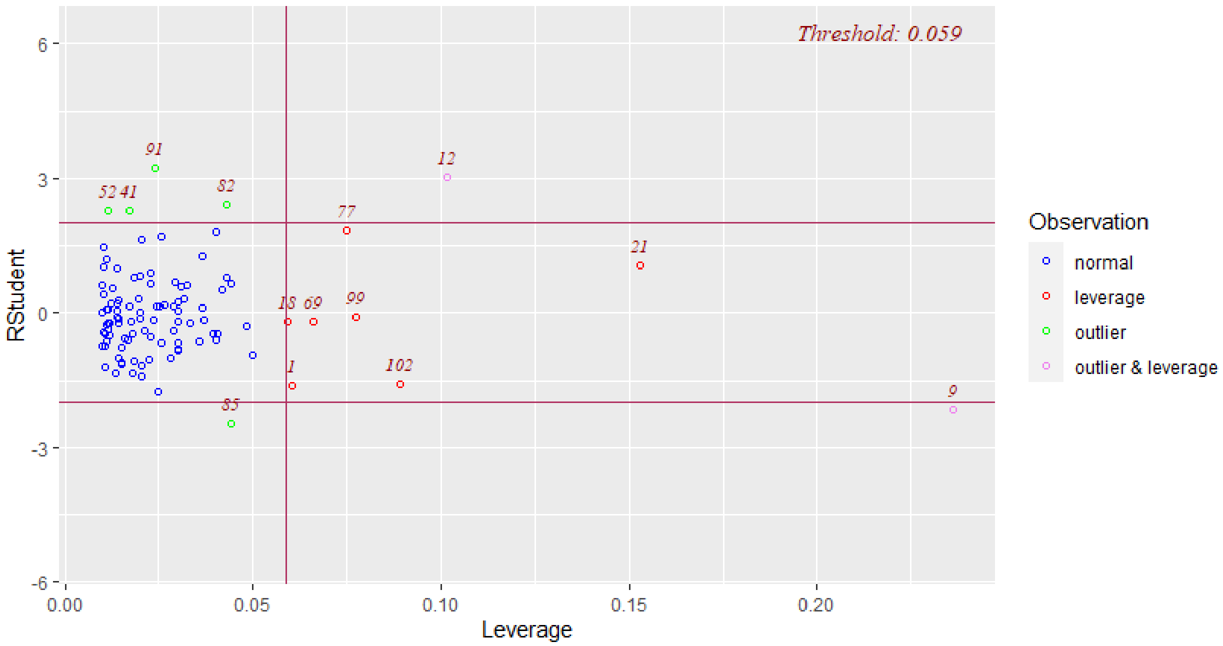
- <u>Age:</u> Age appears contiguous between 27 and around 45 years old, however there is another cluster of around 4 down between 19 and 24. One method of assessing these is to look at a chart of banking vs balance. An observation shows that these are in fact outliers, however they do follow the trend of smaller age = smaller balance, which indicates the impact will be less pronounced. This can be confirmed by removing the observations from the dataset and examining the impact on adj-R^2.
- <u>Education:</u> Education is tightly clustered, which is to be expected given the small number of choices. No further analysis here is needed.
- <u>Income:</u> Income is one of the most predictive variables (by weighted beta), so these observations within this variable will have a larger relative impact on the model. There are several income clusters but they are largely contiguous except for the top two extreme values which are about 15% higher than the others. Due to the importance of this variable, additional analysis is warranted.
- <u>HomeVal:</u> HomeVal has two extreme values at around the $250,000 and $300,000 mark. Home value has a relatively low weighted beta and is not in the final model so it may make sense to run a quick analysis by excluding those values and checking the impact of adding HomeVal to the final model. If it doesn’t become significantly predictive or increase adj-R^2 by a reasonable amount, no further analysis is needed.

- Wealth: Wealth has two outliers at the top, which fall in the $250,00-$350,000 range. Wealth, like Income, is in the final model and has consistently had one of the highest weighted betas which implies that an extreme value here may influence the model, especially if there are extreme values in other variables within this observation.
- Balance: Finally, Balance contains two outliers in the $50,000-$70,000 range. If these correspond to the other extreme values, we have seen they may not reduce the explainability of the model.

Multivariate Analysis: Model 4

The identified extreme values may or may not cause problems within the model. To analyze the impact of the model a targeted approach is warranted, which may be facilitated with the use of the Outlier and Leverage Diagnostics from the OLSRR library in R (Figure 18). This chart plots the observations' distance from the model in standard-deviations on the y-axis, and its leverage score, which is a measure of how extreme the data point is. The y-axis threshold is ±2-standard deviations, and the x-axis threshold is a leverage score of .059.

Figure 18 Outlier and Leverage Diagnostics



Looking at Figure 18 we can see some values with high leverage, particularly observation 21, as well as potential outliers, primarily observation 91. In particular, however, observations 9 and 12 are both a leverage point *and* an outlier, detailed in Table 6.

Table 6 Influential Term Values

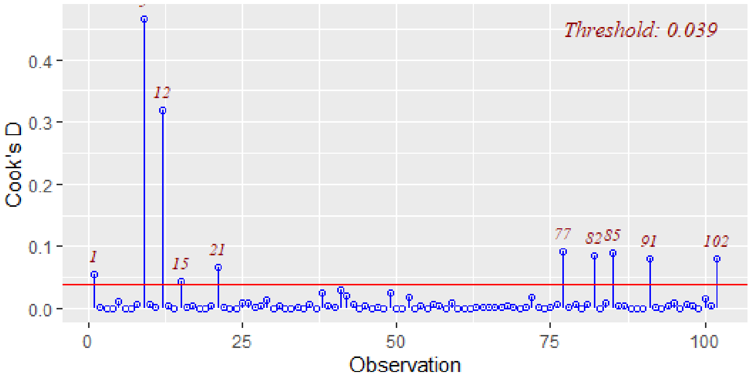
Observation	Age	Education	Income	HomeVal	Wealth	Balance
9	35.7	16.1	107,935	276,139	211,085	41,032
12	43.1	15.8	88,041	194,369	267,556	51,107
21	39.4	16.1	111,548	230,893	331,009	56,569
91	33.9	12.1	32,813	40,313	79,167	26,405

The leverage points (red) are good to be aware of, however their impact on the model is likely contributing positively to its adj-R^2. To confirm, remove them then recalibrate the model and check.

The outliers are important to note because they deviate significantly from the model and are contributing to the unexplained variance. If many of these occur or if there is a pattern in them there may be additional information that needs to be captured in the model, or the model may significantly violate the normality assumption.

Observations that are both leverage points and outliers have a significant impact on unexplained variance and should be analyzed. Cook's D Chart can be generated to give an idea of the impact on the model if the observation is removed (Figure 19). The highest value (cut off) is observation number 9, and the second highest is 12, which are points that are both leveraged and outliers. These are significantly higher than the others, and may warrant excluding them, regenerating the model, and comparing.

Figure 19 Cook's D Chart



When refit, the model's R^2 increases from .9262 to .9268 which is an inconsequential amount: the data's outliers do not have a significant impact on model efficacy and therefore do not preclude the model from use.

3) Wateroil

a.

The purpose of the model is to predict the necessary Voltage to separate the water from the mixture of water and oil. The Data Set ("WATEROIL") contained 19 columns: 3 rows of required voltage with all independent variables set to 0, and 16 rows of experimental results where the voltage was tested given varying levels of the independent variables.

It is unclear and likely not relevant to the model how the levels for each variable were arrived at, however it is important to understand that each variable contained 3 levels: the ground state (0) as well as two other levels. It's important to understand the data and purpose of the model: These variables are continuous, even though 3 values for each variable are present. Therefore, these variables should *not* be treated as categorical.

The WATEROIL dataset contains 8 variables:

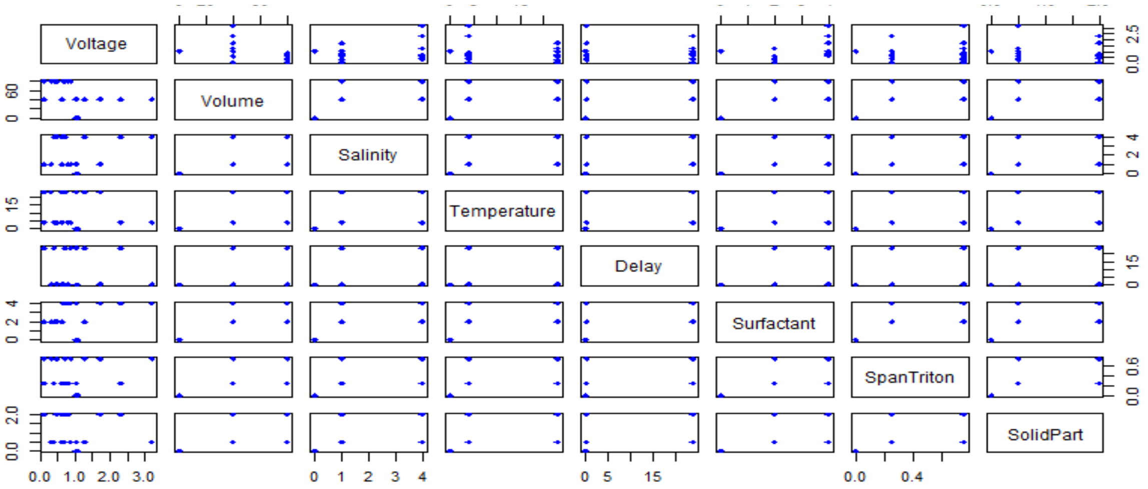
y: Voltage (kw/cm)
X_1: Disperse Phase Volume (%)

X_2: Salinity (%)
X_3: Temperature (°C)
X_4: Time Delay (hours)
X_5: Surfactant Concentration (%)
X_6: Span: Triton
X_7: Solid Particles (%)

Visual Inspection

Variable selection often starts with univariate analysis, which itself is often begun with a visual inspection of the data (Figure 20). The univariate plots for the WATEROIL data set appear segmented to the point of giving the appearance of being categorical; for expediency, the values of each variable will be referred to as low, medium, and high based on their distance from 0. Upon closer inspection, only two potential correlations with Voltage: Surfactant (Figure 21) and Volume (Figure 22).

Figure 20 Univariate Plot of WATEROIL Dataset



The relationship between Surfactant and Voltage seen in figure 18 does not appear strong. The difference between the medium and high levels is very small, and is primarily distinguished by the high value's distribution of voltages extending all the way to the highest voltages, while the medium value's distribution of voltages is concentrated in lower values.

Figure 21 Voltage vs Surfactant, Scatterplot

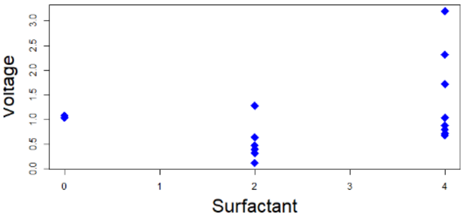
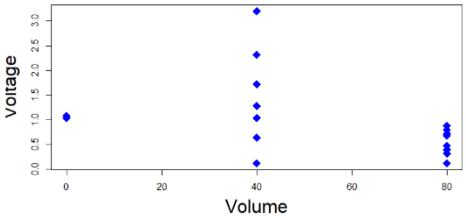


Figure 22 Voltage vs Volume, Scatterplot



Similarly, the relationship between Volume and Voltage can best be described by the medium value's more uniform distribution of voltages whereas the high value's voltages are concentrated in lower voltages.

Correlation Analysis

The next step is to build a correlation matrix to determine whether a linear relationship can be detected statistically (Table 7). Voltage does not have any strong correlations however, the highest ones are Volume (-.369), Surfactant (.328), and Salinity (.236). Further, Volume has modest correlations with Surfactant (.587), SpanTriton (.477), and Salinity (.423), and Surfactant has additional modest correlations with SpanTriton(.477) and Salinity (.423). These correlations that exclude voltage will be investigated as potential interaction terms.

Table 7 WATEROIL Correlation Matrix

	Voltage	Volume	Salinity	Temperature	Delay	Surfactant	SpanTriton
Volume	(0.369)						
Salinity	0.236	0.423					
Temperature	(0.183)	0.377	0.271				
Delay	(0.135)	0.288	0.208	0.185			
Surfactant	0.328	0.587	0.423	0.377	0.288		
SpanTriton	0.094	0.477	0.344	0.306	0.234	0.477	
SolidPart	(0.128)	0.423	0.305	0.271	0.208	0.423	0.344

Additional correlation matrices were constructed using transformed variables. The transformations explored were squaring (Table 8) and log scaling (Table 9).

Table 8 WATEROIL Correlation Matrix of Transformed (squared) and First-Order Variables

	Voltage	Volume	Salinity	Temperature	Delay	Surfactant	SpanTriton
vol.sq	(0.4548)						
sal.sq	0.2678	0.3146					
temp.sq	(0.1868)	0.2980	0.2147				
delay.sq	(0.1352)	0.2830	0.2039	0.1816			
surf.sq	0.4484	0.4230	0.3049	0.2715	0.2076		
span.sq	0.1215	0.3408	0.2456	0.2187	0.1672	0.3408	
solid.sq	(0.1295)	0.3146	0.2267	0.2019	0.1544	0.3146	0.2555

Table 9 WATEROIL Correlation Matrix of Transformed (log scaled) and First-Order Variables

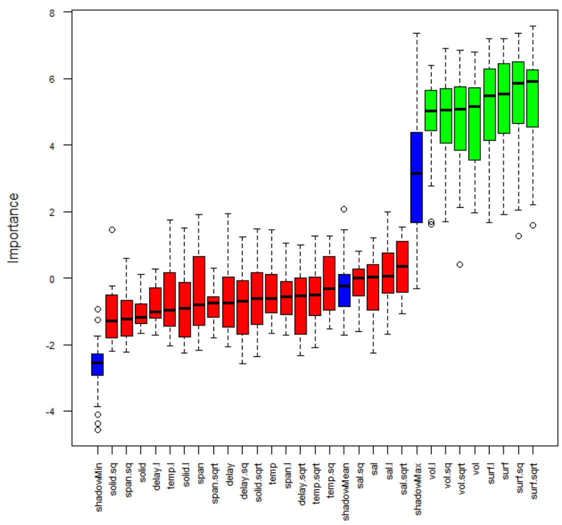
	Voltage	Volume	Salinity	Temperature	Delay	Surfactant	SpanTriton	solid
vol.l	-0.15718		0.540226	0.48107	0.367835	0.749575	0.608726	0.540226
sal.l	0.187379	0.541341		0.347427	0.26565	0.541341	0.43962	0.390149
temp.l	-0.15557	0.591099	0.426011		0.290067	0.591099	0.480029	0.426011
delay.l	-0.13518	0.318188	0.229321	0.20421		0.318188	0.258399	0.229321
surf.l	0.197943	0.692105	0.498807	0.444187	0.339633		0.562055	0.498807
span.l	0.083035	0.520055	0.374808	0.333766	0.255204	0.520055		0.374808
solid.l	-0.12404	0.500294	0.360567	0.321084	0.245507	0.500294	0.406287	

Additional correlation matrices were inspected that included correlation between log-scaled and squared transformations.

Algorithmic Variable Selection

Next, using StepAIC, forward and backward selection was used to build models in order to determine which variables those processes chose. On balance, the variables selected corresponded to the variables identified in the visual and correlation analysis. Additionally, a library called Boruta, which uses an algorithm based on Random Forests to determine the most important factors, was invoked. Boruta's results (Figure 23) corresponded to the analysis already conducted. Finally, the library Caret was used in a similar manner to Boruta, and it also confirmed the importance of the variables previously identified.

Figure 23 Variable Importance, Boruta Output



Model Fitting

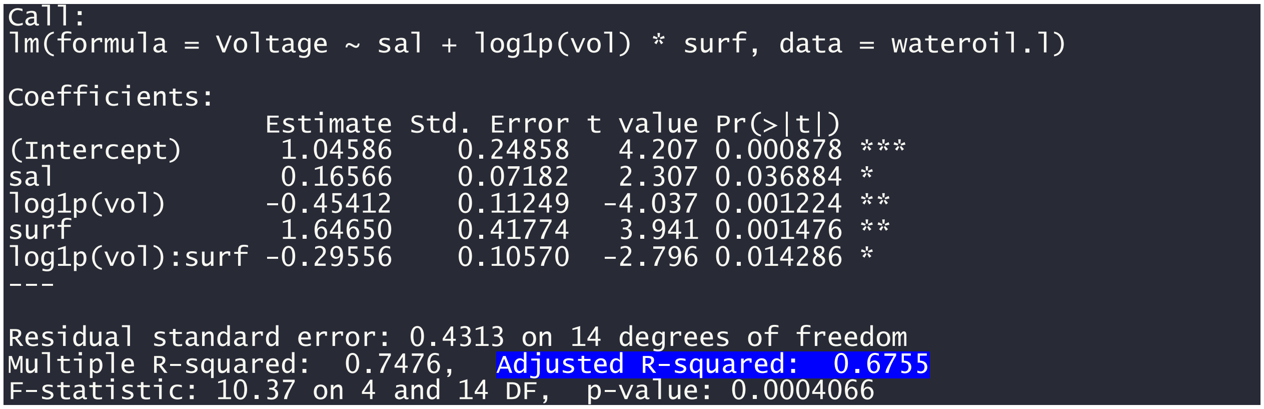
Finally, model fitting was conducted. Beginning with first-order variables, the variables with the highest correlations were added first: first-order, interaction, and second-order terms. Ultimately, about 100 models were tested, most of which were unusable, however the final model includes Salinity, Volume, and Surfactant.

2b) Final Model Specification

$$y_i = 1.04586 + .16566(x_1) + 1.6465(x_2) - .45412(\log(x_3)) - .29556(x_2 * \log(x_3))$$

Where:

$x_1 = Salinity\ (\%)$
$x_2 = Surfactant\ Concentration\ (\%)$
$x_3 = Disperse\ Phase\ Volume\ (\%)$

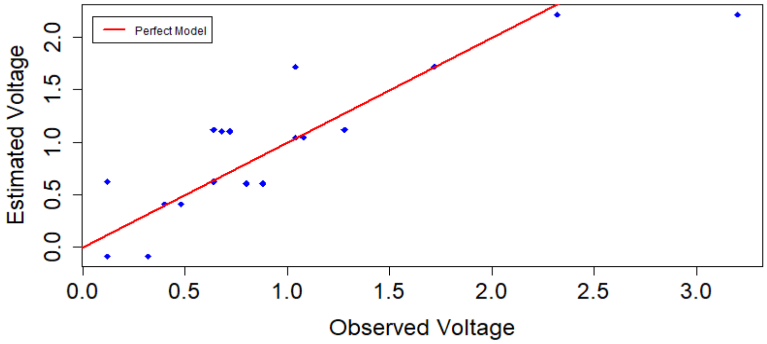


```
Call:
lm(formula = Voltage ~ sal + log1p(vol) * surf, data = wateroil.l)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.04586    0.24858   4.207 0.000878 ***
sal              0.16566    0.07182   2.307 0.036884 *
log1p(vol)      -0.45412    0.11249  -4.037 0.001224 **
surf             1.64650    0.41774   3.941 0.001476 **
log1p(vol):surf -0.29556    0.10570  -2.796 0.014286 *
---

Residual standard error: 0.4313 on 14 degrees of freedom
Multiple R-squared:  0.7476,	Adjusted R-squared:  0.6755
F-statistic: 10.37 on 4 and 14 DF,  p-value: 0.0004066
```

2c) Describe the Model

The model contains 5 variables including the intercept and interaction terms, and consists of salinity and the interaction of log-scaled volume and surfactant (and their main effects). The adj-R^2 was .6755 with an f-statistic of 10.37 (p-value of .0004). The highest variable-level p-value was .03. Finally, the residual standard error was .4313. Figure 24 plots the model's estimate against the observed values.

Figure 24 Estimated vs. Observed Voltage



The model describes the voltage necessary to separate water and oil given a range of variables explored throughout this analysis. Many of the variables had little if any predictive power. Generally speaking, higher surfactant and salinity levels tended to increase the voltage required, while higher disperse phase volume tended to lower voltage requirements, though using the natural log to gain the best fit implies that the marginal reduction of voltage decreases as volume increases (a non-linear relationship).

Ultimately, this model explains a relatively small amount of the variance (67.55%). One issue was the very low number of observations (19, of which 3 were all 0s). Additionally, lack of domain knowledge likely resulted in overlooked relationships that would be clear to experts.