BUDA 525: Team 4 Final Project

Ryan Antonini, Danny Germain, Joshua Meadows, Josh Nelson, Bill Robertson

9/27/2019

# Problem 1

# Problem 2

## Intro

When we are comparing models, there are several methods we can use. Some methods are more simple in that we only look at a calculation based on the specific model output itself and others use sampling methods that compare our prediction between models. The following article will compare six methods by describing how each works, advantages or when to use, and the disadvantages or restrictions of each.

## R-squared

R-squared, called the coefficient of determination, is used to explain the proportion of the dependent variable that is explainable in predicting the independent (y/response) variable.

In order to calculate R-squared, we need a regression model, which produces a prediction line. The numerator or “explained variance” is calculated by subtracting each of the predicted values from the actual values, square them, and sum across all data points. The denominator or “total variance” is calculated similarly, but instead of subtracting the actual value from each predicted value, we subtract the average of the actual values from the predicted value. We do the division and then subtract from 1 to get the R-squared. The output will be a number between 0 and 1, which can be thought of as a percent of the variation that can be explained by the model.

One advantage of evaluating the R-squared is that it is a simple view to get a feel for a single model. Another advantage is that we can explain the variation. However, we do not necessarily know if our model needs a transformation in the response/predictor variables or if our sampling is sufficient. R-squared should not be used to compare models with a different number of predictors, because adding predictors can only help the R-squared value. This method also should not be used to compare models where the response has been transformed because this number doesn’t account for the scaling that occurs in these instances.

## Anova

Anova or “analysis of variance” or “F-tests” is a way of comparing models that are a subset of one another (I.e. y~x+z vs. y~x). This test produces an F-statistic. If the F-statistic is effectively zero, we conclude the models basically account for the same amount of variance in the residuals, and thus Occam’s razor says we should prefer the simpler model. There are a few types of ANOVA test. The most common are type 1 and type 2. The difference between these two is that type 1 depends on the order the predictors are in whereas type 2 does not. If the predictors are orthogonal, these two tests will produce the same values.

## Akaike Information Criterion (AIC)

AIC can be used when models we want to compare have the same response (I.e. sqrt(y) vs. sqrt(y)). With these tests, we prefer the model with the lowest score. These quality estimators are nice because they consider the goodness of fit but add a penalty to models with a higher number of predictors.

## Bayesian Information Criterion (BIC)

BIC is similar to AIC in that we can compare models that have the same response (I.e. sqrt(y) vs. sqrt(y)). However, BIC penalizes the number of predictors more harshly, so we pay more attention to this score when we want a simpler model for explanation purposes, or we have cause to believe overfitting might be an issue. Again we want to see models with lower scores.

## K-Fold Cross-Validation

The K-Fold Cross Validation method works by splitting the data into groups. Once the groups are established, most commonly in five or ten groupings, we loop through the data one time for each grouping. We hold out one of the groups for prediction and the rest for modeling. We then calculate the residual sum of squares (RSS) by subtracting the modeled value from the actual value. We run this process for multiple models and then pick the model with the lowest RSS.

This method has some advantages and disadvantages that will impact whether or not to go this route. We can use it to compare models where the response is any transformation of the same variable (I.e. sqrt(y) vs. log(y)). The issue with using K-Fold is that it can be somewhat time-consuming, which is further compounded by the fact that it works best for large data sets. Consequently, these tests should be used sparingly. It is best to use other methods to narrow down the set of candidate models as much as possible before using these methods.

By leveraging parallel computing, we can somewhat reduce the run time of these methods, but it is important to note that we cannot reduce the time complexity of these algorithms. In general, the higher k is, the more refined your comparison is, but this also increases the run-time.

## Random Splitting

Random Splitting is similar to K-Fold Cross-Validation in many ways. They both also allow for model comparison with any transformation of the same response variable. They also both work well with large datasets. The main difference is that instead of splitting the entire dataset into sections and only testing once through each of those, we can select a much larger number of random samples. We loop through the dataset randomly sampling the training and validation set.Random splitting produces a much more refined comparison than k-fold cross-validation, but it also takes much more time

# Problem 3

We are asked to investigate a set of credit data to see what influences credit card balance. We start by looking at a summary of the data.

## ID Income Limit Rating   
## Min. : 1.0 Min. : 10.35 Min. : 855 Min. : 93.0   
## 1st Qu.:100.8 1st Qu.: 21.01 1st Qu.: 3088 1st Qu.:247.2   
## Median :200.5 Median : 33.12 Median : 4622 Median :344.0   
## Mean :200.5 Mean : 45.22 Mean : 4736 Mean :354.9   
## 3rd Qu.:300.2 3rd Qu.: 57.47 3rd Qu.: 5873 3rd Qu.:437.2   
## Max. :400.0 Max. :186.63 Max. :13913 Max. :982.0   
## Cards Age Education Gender Student   
## Min. :1.000 Min. :23.00 Min. : 5.00 Male :193 No :360   
## 1st Qu.:2.000 1st Qu.:41.75 1st Qu.:11.00 Female:207 Yes: 40   
## Median :3.000 Median :56.00 Median :14.00   
## Mean :2.958 Mean :55.67 Mean :13.45   
## 3rd Qu.:4.000 3rd Qu.:70.00 3rd Qu.:16.00   
## Max. :9.000 Max. :98.00 Max. :20.00   
## Married Ethnicity Balance   
## No :155 African American: 99 Min. : 0.00   
## Yes:245 Asian :102 1st Qu.: 68.75   
## Caucasian :199 Median : 459.50   
## Mean : 520.01   
## 3rd Qu.: 863.00   
## Max. :1999.00

We see that can be zero. We keep this information in mind as we proceed.

Next, we fit a linear model with as the response and all other variables as predictors to see how we should proceed.

##   
## Call:  
## lm(formula = Balance ~ ID + Income + Limit + Rating + Cards +   
## Education + Gender + Student + Married + Ethnicity, data = Credit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -181.59 -76.09 -8.09 56.05 328.39   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -519.43345 33.62383 -15.448 < 2e-16 \*\*\*  
## ID 0.03583 0.04356 0.823 0.4112   
## Income -7.88809 0.23231 -33.955 < 2e-16 \*\*\*  
## Limit 0.19170 0.03293 5.821 1.23e-08 \*\*\*  
## Rating 1.13245 0.49323 2.296 0.0222 \*   
## Cards 17.48400 4.36003 4.010 7.28e-05 \*\*\*  
## Education -1.16397 1.60521 -0.725 0.4688   
## GenderFemale -9.98325 10.02803 -0.996 0.3201   
## StudentYes 427.59524 16.79416 25.461 < 2e-16 \*\*\*  
## MarriedYes -7.04296 10.38311 -0.678 0.4980   
## EthnicityAsian 18.65478 14.16076 1.317 0.1885   
## EthnicityCaucasian 10.23033 12.29024 0.832 0.4057   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 99.26 on 388 degrees of freedom  
## Multiple R-squared: 0.9547, Adjusted R-squared: 0.9534   
## F-statistic: 743 on 11 and 388 DF, p-value: < 2.2e-16

Next, we fit a model that does not include and , do to ethical problems surrounding using these variables as predictors.

## Analysis of Variance Table  
##   
## Model 1: Balance ~ ID + Income + Limit + Rating + Cards + Education +   
## Gender + Student + Married + Ethnicity  
## Model 2: Balance ~ ID + Income + Limit + Rating + Cards + Education +   
## Student + Married  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 388 3822621   
## 2 391 3849031 -3 -26410 0.8935 0.4445

The results of our type 2 ANOVA tell us that we can successfully exclude and without reducing the quality of our model. Next, we see if there are any other variables we can exclude based on AIC.

## Start: AIC=3686.75  
## Balance ~ ID + Income + Limit + Rating + Cards + Education +   
## Student + Married  
##   
## Df Sum of Sq RSS AIC  
## - Married 1 2803 3851834 3685.0  
## - Education 1 5163 3854194 3685.3  
## - ID 1 8764 3857795 3685.7  
## <none> 3849031 3686.7  
## - Rating 1 48485 3897515 3689.8  
## - Cards 1 164802 4013832 3701.5  
## - Limit 1 344450 4193481 3719.0  
## - Student 1 6432087 10281117 4077.7  
## - Income 1 11365025 15214056 4234.5  
##   
## Step: AIC=3685.04  
## Balance ~ ID + Income + Limit + Rating + Cards + Education +   
## Student  
##   
## Df Sum of Sq RSS AIC  
## - Education 1 5676 3857510 3683.6  
## - ID 1 8585 3860419 3683.9  
## <none> 3851834 3685.0  
## - Rating 1 46613 3898447 3687.8  
## - Cards 1 168181 4020015 3700.1  
## - Limit 1 354462 4206296 3718.3  
## - Student 1 6507806 10359640 4078.8  
## - Income 1 11372463 15224297 4232.8  
##   
## Step: AIC=3683.63  
## Balance ~ ID + Income + Limit + Rating + Cards + Student  
##   
## Df Sum of Sq RSS AIC  
## - ID 1 8581 3866091 3682.5  
## <none> 3857510 3683.6  
## - Rating 1 49255 3906765 3686.7  
## - Cards 1 168303 4025814 3698.7  
## - Limit 1 349841 4207351 3716.4  
## - Student 1 6516737 10374247 4077.3  
## - Income 1 11367422 15224933 4230.8  
##   
## Step: AIC=3682.52  
## Balance ~ Income + Limit + Rating + Cards + Student  
##   
## Df Sum of Sq RSS AIC  
## <none> 3866091 3682.5  
## - Rating 1 48967 3915058 3685.6  
## - Cards 1 166410 4032502 3697.4  
## - Limit 1 350520 4216611 3715.2  
## - Student 1 6508348 10374439 4075.4  
## - Income 1 11358921 15225012 4228.8

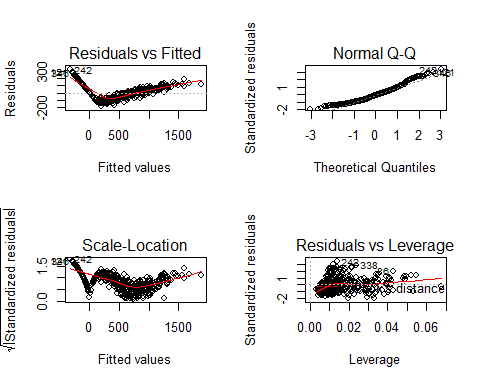
##   
## Call:  
## lm(formula = Balance ~ Income + Limit + Rating + Cards + Student,   
## data = Credit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -185.08 -77.12 -7.19 53.63 325.13   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -526.15552 19.74661 -26.645 < 2e-16 \*\*\*  
## Income -7.87492 0.23145 -34.024 < 2e-16 \*\*\*  
## Limit 0.19441 0.03253 5.977 5.10e-09 \*\*\*  
## Rating 1.08790 0.48700 2.234 0.026 \*   
## Cards 17.85173 4.33489 4.118 4.66e-05 \*\*\*  
## StudentYes 426.85015 16.57403 25.754 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 99.06 on 394 degrees of freedom  
## Multiple R-squared: 0.9542, Adjusted R-squared: 0.9536   
## F-statistic: 1640 on 5 and 394 DF, p-value: < 2.2e-16

## Analysis of Variance Table  
##   
## Model 1: Balance ~ ID + Income + Limit + Rating + Cards + Education +   
## Gender + Student + Married + Ethnicity  
## Model 2: Balance ~ Income + Limit + Rating + Cards + Student  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 388 3822621   
## 2 394 3866091 -6 -43470 0.7354 0.6214

We see that based on AIC and type 2 ANOVA we can use only the predictors , , , , and .

Next, we run diagnostics on this model that uses only 5 predictors.

## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 112.9874, Df = 1, p = < 2.22e-16

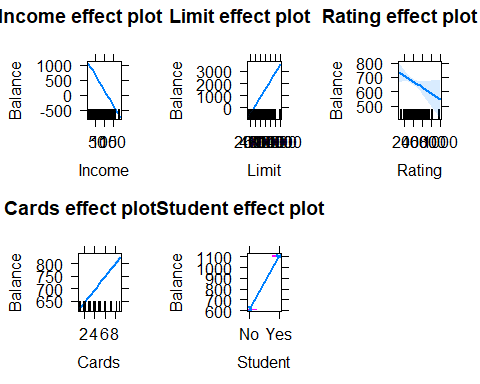
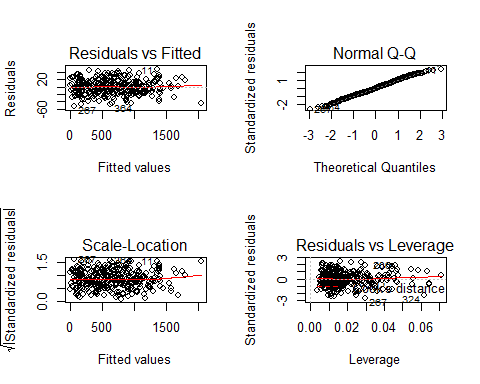


The results of our NCV test and the residual plot tell us we have NCV. By inspection, we see that this is possibly due to values where is zero. Next, we exclude these values to see if we can get better results.

##   
## Call:  
## lm(formula = Balance ~ Income + Limit + Rating + Cards + Student,   
## data = Credit2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -51.88 -13.25 -0.86 13.43 47.11   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -7.493e+02 5.255e+00 -142.579 <2e-16 \*\*\*  
## Income -1.013e+01 5.500e-02 -184.230 <2e-16 \*\*\*  
## Limit 3.410e-01 7.501e-03 45.466 <2e-16 \*\*\*  
## Rating -2.057e-01 1.109e-01 -1.854 0.0647 .   
## Cards 2.539e+01 9.467e-01 26.823 <2e-16 \*\*\*  
## StudentYes 5.027e+02 3.507e+00 143.346 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 19.95 on 304 degrees of freedom  
## Multiple R-squared: 0.9977, Adjusted R-squared: 0.9977   
## F-statistic: 2.655e+04 on 5 and 304 DF, p-value: < 2.2e-16

Notice that our R-squared value is 0.998. We would only expect t an R-squared value this high if we have discovered the results of a mathematical calculation. Upon further reflection, we see in the help file for this data set that this is a simulated dataset, so we have likely found something very close to the formula used to create the non-zero values in this dataset.

## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 0.1662148, Df = 1, p = 0.6835



We notice that has a profound effect on , with students on average having about a $500 higher balance compared to non-students, when everything else is accounted for.

Holding all else constant, the following trends are significant at a 0.01 alpha level: *For each additional credit card that a person has, their balance goes up, on average, about $25* For each additional $10,000 a person makes, their balance goes down about $10 \*For each additional dollar that a person’s credit limit goes up, their Balance increases $0.34

## Dealing with zero balance

There are 90 values where is zero, which is nearly a fourth of our data. In this section, we will try to find if there are factors that affect whether or not the balance is zero. This section reflects work that was done before we learned that we could exclude zero balance from our analysis and should be thought of as a “bonus section.”

We create a new variable as our response which takes the value 1 if the person has a balance, and 0 if they do not have a balance. We then create a model that test the 5 relevant predictors above and then we use AIC values to remove unnecessary predictors.

## Start: AIC=-992.33  
## BalanceF ~ Limit + Student + Rating + Cards + Education + Income  
##   
## Df Sum of Sq RSS AIC  
## - Rating 1 0.0229 32.341 -994.05  
## - Education 1 0.0231 32.341 -994.05  
## - Cards 1 0.1048 32.423 -993.04  
## <none> 32.318 -992.33  
## - Limit 1 0.5668 32.885 -987.38  
## - Student 1 2.2785 34.597 -967.08  
## - Income 1 9.8291 42.148 -888.12  
##   
## Step: AIC=-994.05  
## BalanceF ~ Limit + Student + Cards + Education + Income  
##   
## Df Sum of Sq RSS AIC  
## - Education 1 0.0200 32.361 -995.80  
## - Cards 1 0.0829 32.424 -995.02  
## <none> 32.341 -994.05  
## - Student 1 2.2565 34.598 -969.07  
## - Income 1 9.9392 42.281 -888.86  
## - Limit 1 31.0614 63.403 -726.78  
##   
## Step: AIC=-995.8  
## BalanceF ~ Limit + Student + Cards + Income  
##   
## Df Sum of Sq RSS AIC  
## - Cards 1 0.0872 32.448 -996.72  
## <none> 32.361 -995.80  
## - Student 1 2.2377 34.599 -971.06  
## - Income 1 9.9254 42.287 -890.80  
## - Limit 1 31.0585 63.420 -728.68  
##   
## Step: AIC=-996.72  
## BalanceF ~ Limit + Student + Income  
##   
## Df Sum of Sq RSS AIC  
## <none> 32.448 -996.72  
## - Student 1 2.2175 34.666 -972.28  
## - Income 1 10.0220 42.470 -891.06  
## - Limit 1 31.2380 63.686 -729.00

##   
## Call:  
## lm(formula = BalanceF ~ Limit + Student + Income, data = Credit3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.56004 -0.24852 0.01882 0.23604 0.52837   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.425e-01 3.408e-02 4.181 3.57e-05 \*\*\*  
## Limit 1.987e-04 1.018e-05 19.525 < 2e-16 \*\*\*  
## StudentYes 2.484e-01 4.775e-02 5.202 3.17e-07 \*\*\*  
## Income -7.373e-03 6.667e-04 -11.059 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2863 on 396 degrees of freedom  
## Multiple R-squared: 0.5348, Adjusted R-squared: 0.5313   
## F-statistic: 151.7 on 3 and 396 DF, p-value: < 2.2e-16

We find that we only need the 3 variables , , and for this model. Next, we test the interactions between these variables and see if we can remove any interaction terms using AIC.

## Start: AIC=-1065.42  
## BalanceF ~ Student \* Limit \* Income  
##   
## Df Sum of Sq RSS AIC  
## - Student:Limit:Income 1 0.037525 26.824 -1066.9  
## <none> 26.787 -1065.4  
##   
## Step: AIC=-1066.86  
## BalanceF ~ Student + Limit + Income + Student:Limit + Student:Income +   
## Limit:Income  
##   
## Df Sum of Sq RSS AIC  
## <none> 26.824 -1066.9  
## - Student:Income 1 1.1339 27.958 -1052.3  
## - Student:Limit 1 2.8001 29.624 -1029.2  
## - Limit:Income 1 3.2015 30.026 -1023.8

We see that we do not need the interaction between all 3 variables, and we can just use the pairwise interactions.

Notice that what we have done in this section falls outside the normal use of linear regression. We are using the predictors to generate a continuous response, even though our response can only have values zero or one in reality. We must now test whether our model will actually allow us to predict whether or not a person has a balance. To do this, we will use random-splitting, but instead of concerning ourselves with the RSS of the test set, we will predict if the person has a balance in the following way: if the fitted value of the model is greater than 0.5, we predict the person has a balance, and we predict they do not otherwise. We then calculated the percentage we predict correctly over 1000 randomly selected test sets of size 100, using the rest of the data to fit the model.

## [1] 5

## Loading required package: rngtools

## Loading required package: pkgmaker

## Loading required package: registry

##   
## Attaching package: 'pkgmaker'

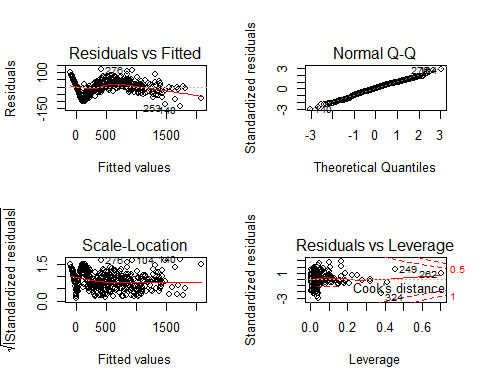
## The following object is masked from 'package:base':  
##   
## isFALSE

## [1] 978.26

We see that we predict correctly 97.8% of the time. We could first use this model to predict whether or not the person will have a balance, then predict the balance when appropriate, but instead, we try to use the information we learned to fit a model that effectively does all of this in one go by testing as the predictors of . We then use AIC to remove unnecessary terms, although this analysis is left out for brevity.

##   
## Call:  
## lm(formula = Balance ~ Student + Limit + Income + log(Rating) +   
## Cards + Student:Limit + Student:Income + Limit:Income + Student:log(Rating) +   
## Limit:log(Rating) + Limit:Cards + Income:log(Rating) + Income:Cards +   
## Student:Limit:Income + Student:Limit:log(Rating) + Limit:Income:Cards,   
## data = Credit3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -134.73 -26.93 1.94 33.31 121.17   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.774e+03 1.998e+02 8.878 < 2e-16 \*\*\*  
## StudentYes -2.150e+03 4.850e+02 -4.432 1.22e-05 \*\*\*  
## Limit -6.070e-01 6.460e-02 -9.396 < 2e-16 \*\*\*  
## Income 1.729e+01 4.307e+00 4.013 7.22e-05 \*\*\*  
## log(Rating) -3.967e+02 4.550e+01 -8.717 < 2e-16 \*\*\*  
## Cards 4.823e+00 7.308e+00 0.660 0.509669   
## StudentYes:Limit 8.563e-01 1.860e-01 4.604 5.64e-06 \*\*\*  
## StudentYes:Income -5.685e+00 1.108e+00 -5.130 4.62e-07 \*\*\*  
## Limit:Income -3.291e-04 1.107e-04 -2.974 0.003129 \*\*   
## StudentYes:log(Rating) 4.278e+02 1.140e+02 3.752 0.000203 \*\*\*  
## Limit:log(Rating) 1.491e-01 8.742e-03 17.055 < 2e-16 \*\*\*  
## Limit:Cards 4.415e-03 1.622e-03 2.722 0.006783 \*\*   
## Income:log(Rating) -4.063e+00 7.982e-01 -5.090 5.61e-07 \*\*\*  
## Income:Cards 3.992e-02 1.636e-01 0.244 0.807319   
## StudentYes:Limit:Income 8.556e-04 1.605e-04 5.329 1.69e-07 \*\*\*  
## StudentYes:Limit:log(Rating) -1.382e-01 2.596e-02 -5.323 1.74e-07 \*\*\*  
## Limit:Income:Cards -3.031e-05 1.756e-05 -1.726 0.085195 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 45.86 on 383 degrees of freedom  
## Multiple R-squared: 0.9904, Adjusted R-squared: 0.99   
## F-statistic: 2482 on 16 and 383 DF, p-value: < 2.2e-16

## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 3.921739, Df = 1, p = 0.047666



This resulting model is a little tedious to interpret but deals with the NCV, making it a good forecasting model, but not a good model to explain to a client.

# Problem 4

## rank discipline yrs.since.phd yrs.service sex   
## AsstProf : 67 A:181 Min. : 1.00 Min. : 0.00 Female: 39   
## AssocProf: 64 B:216 1st Qu.:12.00 1st Qu.: 7.00 Male :358   
## Prof :266 Median :21.00 Median :16.00   
## Mean :22.31 Mean :17.61   
## 3rd Qu.:32.00 3rd Qu.:27.00   
## Max. :56.00 Max. :60.00   
## salary   
## Min. : 57800   
## 1st Qu.: 91000   
## Median :107300   
## Mean :113706   
## 3rd Qu.:134185   
## Max. :231545

##   
## Call:  
## lm(formula = salary ~ rank + discipline + yrs.since.phd + yrs.service +   
## sex, data = Salaries)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -65248 -13211 -1775 10384 99592   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 65955.2 4588.6 14.374 < 2e-16 \*\*\*  
## rankAssocProf 12907.6 4145.3 3.114 0.00198 \*\*   
## rankProf 45066.0 4237.5 10.635 < 2e-16 \*\*\*  
## disciplineB 14417.6 2342.9 6.154 1.88e-09 \*\*\*  
## yrs.since.phd 535.1 241.0 2.220 0.02698 \*   
## yrs.service -489.5 211.9 -2.310 0.02143 \*   
## sexMale 4783.5 3858.7 1.240 0.21584   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 22540 on 390 degrees of freedom  
## Multiple R-squared: 0.4547, Adjusted R-squared: 0.4463   
## F-statistic: 54.2 on 6 and 390 DF, p-value: < 2.2e-16