Case 4: Descriptive Analysis and Employee Retention

Dr. Avery Haviv
University of Rochester
MKT436 and MKT436R

Winter 2020

Introduction

The purpose of this in class case is to use and practice descriptive, and causal analysis. We will be using a dataset posted on Kaggle that concerns human resources and employee retention. This data set is quite popular and has generated a lot of discussions. You can read some other more here if you are interested: https://www.kaggle.com/ludobenistant/hr-analytics/kernels

Throughout our dependent variable will be `left`, which indicates if an employee ended up leaving the company.

First, load the data and see the various columns:

```
hrData = read.csv('D:/Dropbox/Teaching Lectures/Employee Retention Case/Employee Retention Dataset.csv'

#Check the first few rows to understand the data
head(hrData)
```

```
##   satisfaction_level last_evaluation number_project average_montly_hours
## 1               0.58            0.74              4                  215
## 2               0.82            0.67              2                  202
## 3               0.45            0.69              5                  193
## 4               0.78            0.82              5                  247
## 5               0.49            0.60              3                  214
## 6               0.36            0.95              3                  206
##   time_spend_company Work_accident left promotion_last_5years sales salary
## 1                  3             0    0                     0 sales    low
## 2                  3             0    0                     0 sales    low
## 3                  3             0    0                     0 sales    low
## 4                  3             0    0                     0 sales    low
## 5                  2             0    0                     0 sales    low
## 6                  4             0    0                     0 sales    low
```

Module 1: Descriptive Analysis with Correlation and Linear Regression

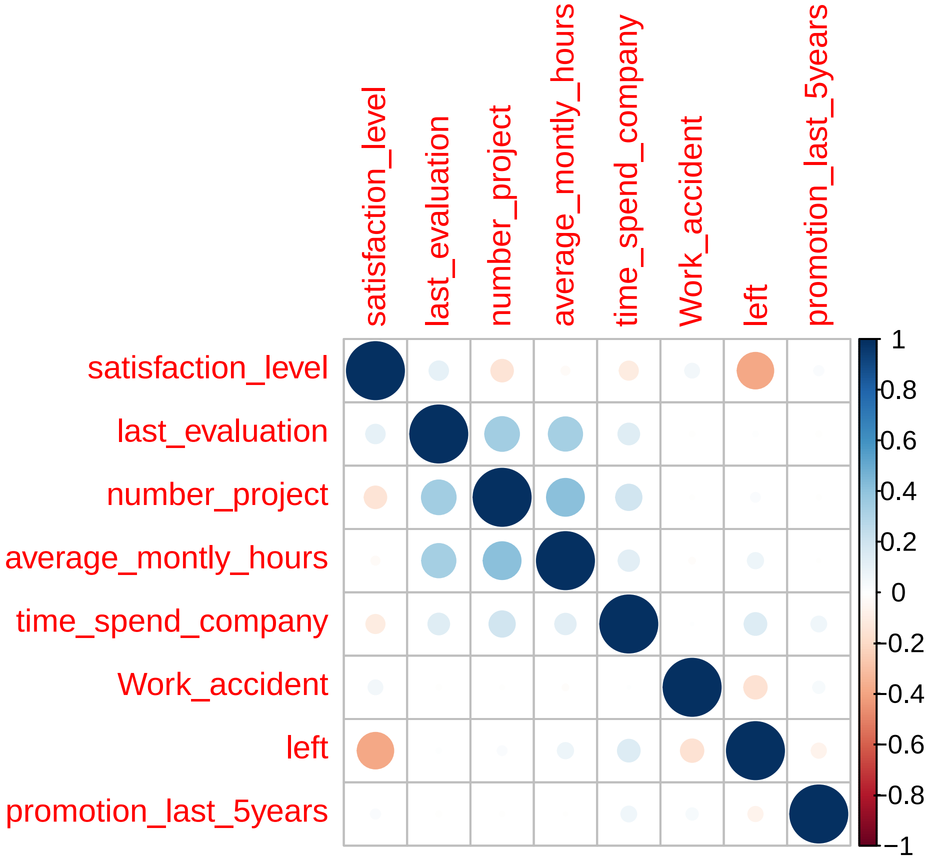
First, we can check the correlations. We can do this by either looking directly at the correlations or by using a correlation plot.

```
install.packages('corrplot',repos='http://cran.us.r-project.org')
```

```
## Installing package into 'C:/Users/owner/Documents/R/win-library/3.5'
## (as 'lib' is unspecified)
```

```
## Warning: package 'corrplot' is in use and will not be installed
```

```r
library('corrplot')
corrplot(cor(hrData[,1:8]),method='circle')
```



Second, lets run a simple regression using all the available variables. By writing the formula `left .`, you will control for all available variables in the dataframe. We will use both `summary` and `anova` functions to evaluate statistical errors.

```r
summary(lm(left~.,data=hrData))
```

```
##
## Call:
## lm(formula = left ~ ., data = hrData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8074 -0.2544 -0.1055  0.1679  1.1438
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            0.3271046  0.0249239  13.124  < 2e-16 ***
## satisfaction_level    -0.6439672  0.0128127 -50.260  < 2e-16 ***
## last_evaluation        0.0872994  0.0200734   4.349 1.38e-05 ***
## number_project        -0.0340501  0.0029113 -11.696  < 2e-16 ***
## average_montly_hours   0.0006413  0.0000698   9.187  < 2e-16 ***
## time_spend_company     0.0364441  0.0021909  16.634  < 2e-16 ***
## Work_accident         -0.1554197  0.0087907 -17.680  < 2e-16 ***
## promotion_last_5years -0.1120016  0.0217360  -5.153 2.60e-07 ***
## saleshr                0.0346008  0.0194701   1.777  0.07557 .  
## salesIT               -0.0254637  0.0173887  -1.464  0.14311    
## salesmanagement       -0.0623505  0.0206611  -3.018  0.00255 ** 
## salesmarketing        -0.0024222  0.0187803  -0.129  0.89738    
## salesproduct_mng      -0.0240979  0.0185540  -1.299  0.19403    
## salesRandD            -0.0752722  0.0191734  -3.926 8.68e-05 ***
## salessales            -0.0051736  0.0148529  -0.348  0.72760    
## salessupport           0.0059326  0.0158191   0.375  0.70765    
## salestechnical         0.0073094  0.0154432   0.473  0.63600    
## salarylow              0.1989519  0.0119150  16.698  < 2e-16 ***
## salarymedium           0.1204051  0.0119618  10.066  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.3775 on 14980 degrees of freedom
## Multiple R-squared:  0.2155, Adjusted R-squared:  0.2146 
## F-statistic: 228.6 on 18 and 14980 DF,  p-value: < 2.2e-16
```

```
anova(lm(left~.,data=hrData))
```

```
## Analysis of Variance Table
## 
## Response: left
##                          Df  Sum Sq Mean Sq  F value    Pr(>F)    
## satisfaction_level        1  410.39  410.39 2880.165 < 2.2e-16 ***
## last_evaluation           1    6.17    6.17   43.297 4.859e-11 ***
## number_project            1    7.79    7.79   54.645 1.521e-13 ***
## average_montly_hours      1   14.19   14.19   99.611 < 2.2e-16 ***
## time_spend_company        1   30.90   30.90  216.886 < 2.2e-16 ***
## Work_accident             1   46.93   46.93  329.336 < 2.2e-16 ***
## promotion_last_5years     1    8.05    8.05   56.466 6.042e-14 ***
## sales                     9   13.04    1.45   10.172 8.962e-16 ***
## salary                    2   48.85   24.43  171.417 < 2.2e-16 ***
## Residuals             14980 2134.49    0.14                       
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The `anova` function differs from `summary` in that it tests for the significance of categorical variables *jointly*, rather than considering each dummy variable independently. This shows that satisfaction level is the most important variable and that high satisfaction makes someone less likely to leave. The presence of a work accident and time spent are the next most important.

Discussion Questions

1. What is the interpretation of the regression coefficient of satisfaction_level?
2. What is the interpretation of the statistical significance of satisfaction_level?
3. Do the variables that are not significant effect 'left'?
4. Of the different job types (HR, IT, etc.), what are the top 3 most likely to leave this firm?

Module 2: Interaction Effects

Do we think that the impact satisfaction might depend on salary? Let's check by adding an interaction between satisfaction and type of job:

```
summary(lm(left~.+satisfaction_level*salary,data=hrData))
```

```
## 
## Call:
## lm(formula = left ~ . + satisfaction_level * salary, data = hrData)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -0.8224 -0.2483 -0.0995  0.1523  1.1349 
## 
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                      6.785e-02  3.822e-02   1.775 0.075884 .  
## satisfaction_level              -2.370e-01  4.734e-02  -5.006 5.62e-07 ***
## last_evaluation                  8.888e-02  2.001e-02   4.442 8.98e-06 ***
## number_project                  -3.431e-02  2.902e-03 -11.822  < 2e-16 ***
## average_montly_hours             6.305e-04  6.959e-05   9.061  < 2e-16 ***
## time_spend_company               3.716e-02  2.185e-03  17.007  < 2e-16 ***
## Work_accident                   -1.554e-01  8.763e-03 -17.730  < 2e-16 ***
## promotion_last_5years           -1.136e-01  2.167e-02  -5.241 1.62e-07 ***
## saleshr                          3.430e-02  1.941e-02   1.767 0.077242 .  
## salesIT                         -2.557e-02  1.733e-02  -1.475 0.140114    
## salesmanagement                 -6.508e-02  2.060e-02  -3.159 0.001584 ** 
## salesmarketing                  -1.858e-03  1.872e-02  -0.099 0.920940    
## salesproduct_mng                -2.299e-02  1.849e-02  -1.243 0.213877    
## salesRandD                      -7.334e-02  1.911e-02  -3.837 0.000125 ***
## salessales                      -5.984e-03  1.481e-02  -0.404 0.686074    
## salessupport                     4.372e-03  1.577e-02   0.277 0.781616    
## salestechnical                   7.032e-03  1.539e-02   0.457 0.647820    
## salarylow                        5.072e-01  3.397e-02  14.931  < 2e-16 ***
## salarymedium                     3.588e-01  3.447e-02  10.409  < 2e-16 ***
## satisfaction_level:salarylow    -4.886e-01  5.031e-02  -9.712  < 2e-16 ***
## satisfaction_level:salarymedium -3.735e-01  5.097e-02  -7.328 2.45e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.3763 on 14978 degrees of freedom
## Multiple R-squared:  0.2207, Adjusted R-squared:  0.2196 
## F-statistic: 212.1 on 20 and 14978 DF,  p-value: < 2.2e-16
```

Interpreting the interaction effect, satisfaction has a larger effect on those with low salary than it does on those with high salary.

Lets check with another interaction, this time between satisfaction and work accidents:

```
summary(lm(left~.+satisfaction_level*Work_accident,data=hrData))
```

```
## 
## Call:
## lm(formula = left ~ . + satisfaction_level * Work_accident, data = hrData)
## 
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.83162 -0.24771 -0.09713  0.13936  1.07293 
## 
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       3.620e-01  2.502e-02  14.469  < 2e-16
## satisfaction_level               -6.968e-01  1.362e-02 -51.173  < 2e-16
## last_evaluation                   9.344e-02  2.000e-02   4.672 3.01e-06
## number_project                   -3.457e-02  2.900e-03 -11.923  < 2e-16
## average_montly_hours              6.185e-04  6.955e-05   8.893  < 2e-16
## time_spend_company                3.680e-02  2.182e-03  16.865  < 2e-16
## Work_accident                    -4.229e-01  2.560e-02 -16.518  < 2e-16
## promotion_last_5years            -1.108e-01  2.165e-02  -5.116 3.16e-07
## saleshr                           3.335e-02  1.939e-02   1.720 0.085454
## salesIT                          -2.716e-02  1.732e-02  -1.568 0.116840
## salesmanagement                  -6.589e-02  2.058e-02  -3.202 0.001369
## salesmarketing                   -4.333e-03  1.870e-02  -0.232 0.816829
## salesproduct_mng                 -2.622e-02  1.848e-02  -1.419 0.155925
## salesRandD                       -7.257e-02  1.910e-02  -3.800 0.000145
## salessales                       -5.690e-03  1.479e-02  -0.385 0.700514
## salessupport                      4.999e-03  1.576e-02   0.317 0.751004
## salestechnical                    6.735e-03  1.538e-02   0.438 0.661449
## salarylow                         1.980e-01  1.187e-02  16.688  < 2e-16
## salarymedium                      1.188e-01  1.191e-02   9.973  < 2e-16
## satisfaction_level:Work_accident  4.158e-01  3.740e-02  11.117  < 2e-16
## 
## (Intercept)                      ***
## satisfaction_level               ***
## last_evaluation                  ***
## number_project                   ***
## average_montly_hours             ***
## time_spend_company               ***
## Work_accident                    ***
## promotion_last_5years            ***
## saleshr                          .  
## salesIT                             
## salesmanagement                  ** 
## salesmarketing                      
## salesproduct_mng                    
## salesRandD                       ***
## salessales                          
## salessupport                        
## salestechnical                      
## salarylow                        ***
## salarymedium                     ***
## satisfaction_level:Work_accident ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.3759 on 14979 degrees of freedom
## Multiple R-squared:  0.2219, Adjusted R-squared:  0.2209 
## F-statistic: 224.8 on 19 and 14979 DF,  p-value: < 2.2e-16
```

Overall, accidents make people less likely to leave. If they are satisfied, then the effect of an accident is 0.

Going through potentially important interactions manually is too difficult. In module 3 will use what we learned in descriptive analysis to figure out the best regression with 2-degree interactions.

Discussion Questions:

1) How do you model an interaction effect in R?
2) What is the interpretation of the coefficients on the interaction effect?

Module 3: Descriptive analysis Using 'leaps'

Including all potential interactions in a linear regression is a complete mess and hard to interpret. This makes descriptive analysis difficult because there will be too many coefficients. This will also increase variance in terms of the bias/variance trade-off since we will be controlling for so many variables.

```
#Run this at your own risk - there are 135 coefficients! Very difficult to interpret that.
summary(lm('left~.^2',data=hrData))
```

Instead let's find the best, simple regression we can by sorting through a number of potential candidates. We will use the 'leaps' package to find a simpler model so we can see clearly which interactions are most important.

The default is for 'method' to be set to 'exhaustive', which tries every single possible regression. Because of the large number of variables, the computational time would be very large. Instead, set method to 'forward' to speed up the function, which will make it sequentially add variables one at a time.

```
install.packages('leaps',repos='http://cran.us.r-project.org')
```

```
## Installing package into 'C:/Users/owner/Documents/R/win-library/3.5'
## (as 'lib' is unspecified)
```

```
library('leaps')
leapsModels = regsubsets(left~.^2,data=hrData,method='forward')
subsetSummary = summary(leapsModels)
```

We only want one regression, we can use the AIC or the BIC to sort through the models returned by `regsubsets` We want a smaller model so let's get the coefficients that correspond to the best BIC.

```
bestBIC = which.min(subsetSummary$bic)
```

Next, we can extract coefficients, and round them, so they are readable, and therefore easier to interpret.

```
round(coef(leapsModels,bestBIC),3)
```

```
##                             (Intercept)
##                                   4.051
##                      satisfaction_level
##                                  -1.231
##                         last_evaluation
##                                  -3.603
##                    average_montly_hours
##                                  -0.016
##                      time_spend_company
##                                  -0.308
##   satisfaction_level:time_spend_company
##                                   0.203
##    last_evaluation:average_montly_hours
##                                   0.018
## average_montly_hours:time_spend_company
##                                   0.001
##        time_spend_company:Work_accident
##                                  -0.035
##            time_spend_company:salarylow
##                                   0.025
```

We've gone from 135 coefficients to 10. This is easier to interpret, and we can also see the most important interactions. Keep in mind you can have `regsubsets` give you an even smaller model if you would prefer! You would simply set the second argument to the number of coefficients you want.

Discussion Questions:

1. Can you interpret the resulting coefficients from this analysis as causal? Why or why not?
2. According to this model, which types of employees are affected by a low salary level?

Module 4: Using MARS to find non-linear relationships

So far we've only found linear relationships in the data, as that is all correlations and regressions do by default. We know from our discussion of functional form assumptions that relationships might be non-linear. We currently haven't used our knowledge of functional form assumptions to look for non-linear relationships. A simple way to do this is using MARS since it is 'non-parametric'. Let's use code for MARS and see if the relationships are all linear.

```
install.packages('earth',repos='http://cran.us.r-project.org')
```

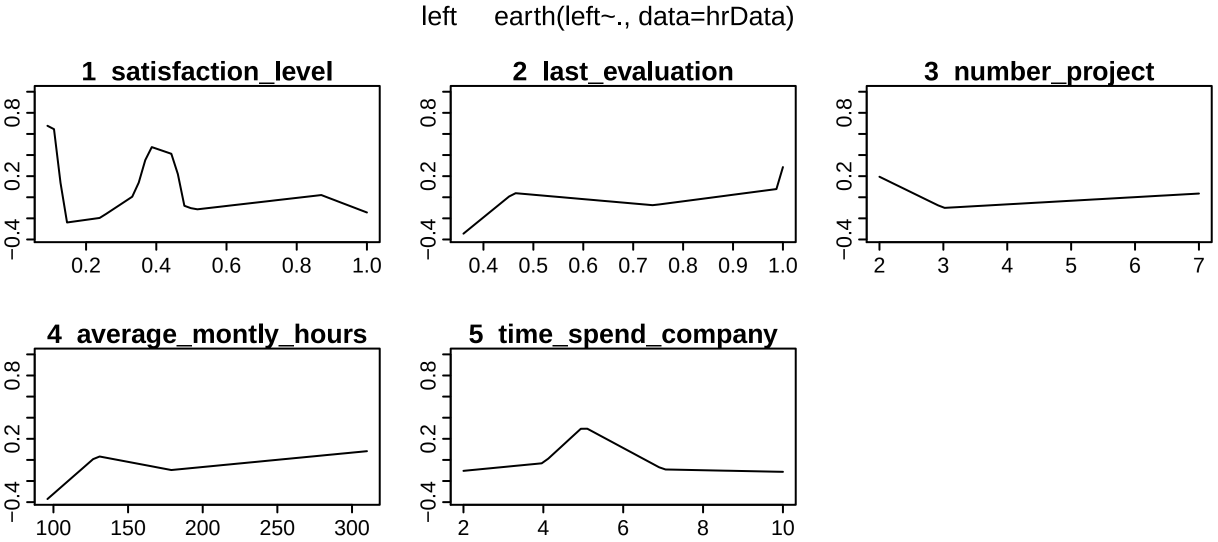
```
## Installing package into 'C:/Users/owner/Documents/R/win-library/3.5'
## (as 'lib' is unspecified)
```

```
##
##   There is a binary version available but the source version is
##   later:
##       binary source needs_compilation
## earth  5.1.2  5.3.0              TRUE
##
##   Binaries will be installed
```

```
## Warning: package 'earth' is in use and will not be installed
```

```
library('earth')
earthFit = earth(left~.,data=hrData)
plotmo(earthFit)
```

```
##  plotmo grid:    satisfaction_level last_evaluation number_project
##                                0.64            0.72              4
##  average_montly_hours time_spend_company Work_accident
##                   200                  3             0
##  promotion_last_5years sales salary
##                      0 sales    low
```



Some very interesting non-linear relationships! Particularly in satisfaction, number of projects, and time spent at the firm. Having a last evaluation of exactly 100 makes someone 20% more likely to leave when compared to 99. Satisfaction has a strange increase from 0.14 to 0.38. Correlation would not detect these relationships, because correlation only looks at LINEAR associations.

This worked well as a descriptive analysis. It summarized a number of potentially meaningful relationships in the data that we could explore further. Using a more complex model makes it harder to detect these relationships.

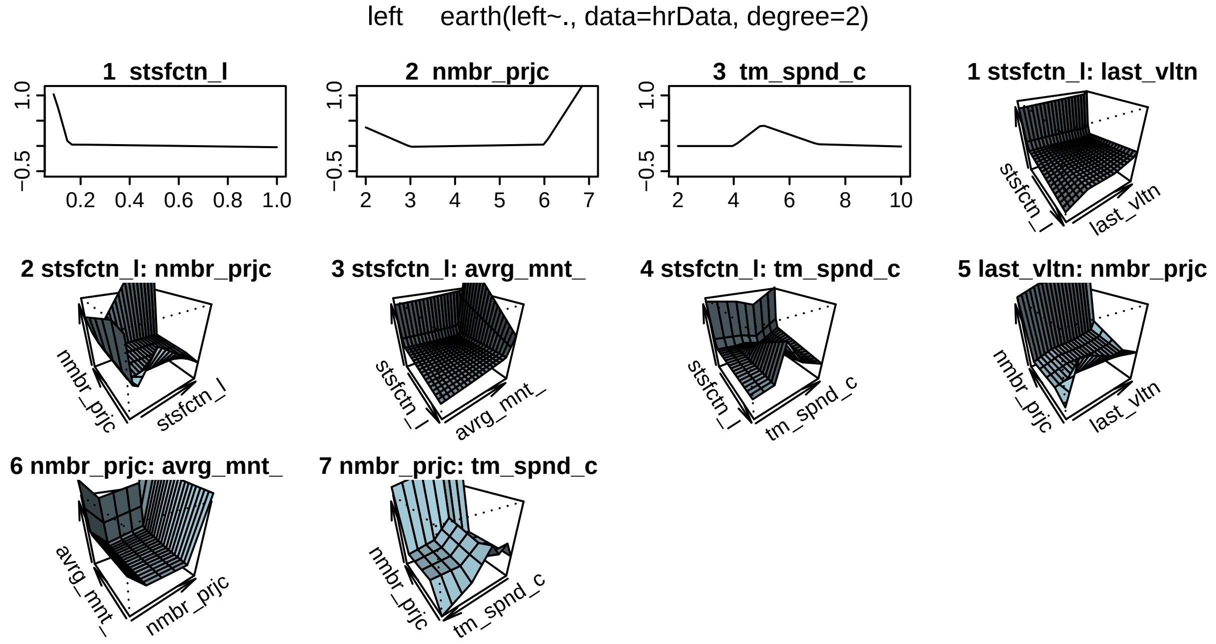
Interactions and MARS

If we want a deeper, but harder to interpret analysis, we can add interactions to the MARS model. This will require us to interpret 3-D plots, which is harder[1]

```
earthFit = earth(left~.,data=hrData,degree=2)
plotmo(earthFit)
```

```
##  plotmo grid:    satisfaction_level last_evaluation number_project
##                                0.64            0.72              4
##  average_montly_hours time_spend_company Work_accident
##                   200                  3             0
##  promotion_last_5years sales salary
##                      0 sales    low
```

[1] This is also very helpful



Interpreting this figure requires zooming in on the plots closely, or generating the model in R[2]. We can observe the following relationships based on these interactions:

1. From the second interaction, those with no projects and who are somewhat satisfied are more likely to leave
2. From the fourth interaction, those who are highly satisfied and have spent 3 years at the company are likely to leave
3. From the fifth interaction, those with no projects and a decent last evaluation are more likely to leave

Discussion Questions

1. What can we learn from the 7th interaction term? Keep in mind that `plotmo` deletes vowels to make the variable names more readable.

Module 5: Causal

This module draws on course materials from Topic 3. Suppose the causal question for the firm is how employee evaluations affect retention. The initial regression is as follows:

```
summary(lm(left~.,data=hrData))
```

```
##
## Call:
```

[2] You might need to expand your plot window to fit all these figures

```
## lm(formula = left ~ ., data = hrData)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -0.8074 -0.2544 -0.1055  0.1679  1.1438 
## 
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)    
## (Intercept)            0.3271046  0.0249239  13.124  < 2e-16 ***
## satisfaction_level    -0.6439672  0.0128127 -50.260  < 2e-16 ***
## last_evaluation        0.0872994  0.0200734   4.349 1.38e-05 ***
## number_project        -0.0340501  0.0029113 -11.696  < 2e-16 ***
## average_montly_hours   0.0006413  0.0000698   9.187  < 2e-16 ***
## time_spend_company     0.0364441  0.0021909  16.634  < 2e-16 ***
## Work_accident         -0.1554197  0.0087907 -17.680  < 2e-16 ***
## promotion_last_5years -0.1120016  0.0217360  -5.153 2.60e-07 ***
## saleshr                0.0346008  0.0194701   1.777  0.07557 .  
## salesIT               -0.0254637  0.0173887  -1.464  0.14311    
## salesmanagement       -0.0623505  0.0206611  -3.018  0.00255 ** 
## salesmarketing        -0.0024222  0.0187803  -0.129  0.89738    
## salesproduct_mng      -0.0240979  0.0185540  -1.299  0.19403    
## salesRandD            -0.0752722  0.0191734  -3.926 8.68e-05 ***
## salessales            -0.0051736  0.0148529  -0.348  0.72760    
## salessupport           0.0059326  0.0158191   0.375  0.70765    
## salestechnical         0.0073094  0.0154432   0.473  0.63600    
## salarylow              0.1989519  0.0119150  16.698  < 2e-16 ***
## salarymedium           0.1204051  0.0119618  10.066  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.3775 on 14980 degrees of freedom
## Multiple R-squared:  0.2155, Adjusted R-squared:  0.2146 
## F-statistic: 228.6 on 18 and 14980 DF,  p-value: < 2.2e-16
```

The coefficients say that all is being equal, a higher evaluation increases the likelihood that someone leaves. If you were to interpret the coefficients in a causal way, the coefficient says that increasing an evaluation makes someone more likely to leave a firm.

Intuitively this makes little sense, as the evaluation itself would likely lead to positive outcomes at the firm in the future. This regression suffers from an omitted variable bias that reversed the sign of the coefficient. Regression gives you the 'correlation', but not the 'causation'.

Predictive models also suffer from omitted variable bias. To confirm this, we will investigate the effect of salary within a simple neural network, which I have estimated and stored as `nnet1`[3]. Since neural networks are difficult to interpret, we cannot look at coefficient estimates as we would in a regression or MARS analysis. Instead, we will see how the prediction of the probability of leaving the firm changes when we reduce the evaluation for the first employee in the dataset:

First, let's check the predicted leaving probability for the first employee:

```
#Make predictions using the predict function.
#We will discuss how to use this function when discussing predictive methods.
newData = trainingData[1,]
#Predicted leaving probability when at their actual evaluation
```

[3] I have not included the code to estimate the neural network as it is beyond the scope of this class. However, interested students can email me for it.

```
predict(nnet1,newData)
```

```
##          [,1]
## 1 0.06959988
```

Second, we change last_evaluation to zero, and get the new predicted leaving probability:

```
#Predicted leaving probability when their evaluation is reduced to zero
newDataLowEval = newData
newDataLowEval$last_evaluation = 0
predict(nnet1,newDataLowEval)
```

```
##           [,1]
## 1 -0.08067422
```

As in regression, a higher evaluation raises the probability of leaving. Omitted variable bias affects **ALL** models, not just regression. In any case, when you are making a decision about an independent variable, you need to consider omitted variable bias.

Discussion Questions:

1. Why do you think we are observing a consistent positive relationship between evaluations and leaving? What omitted variables might cause that?
2. How do you decide what variables to include in a causal model? How does this differ from how you select variables in a predictive model?
3. This analysis did not control for the overall level of benefits that an employee received. In what way would you expect this omitted variable to bias the coefficient on salary?
4. Can the firm determine whether salary increases lead to an improvement in employee retainment from this dataset alone?