DOM 207: IDSBRP
MINI PROJECT 3

INTRODUCTION

We, Aditya Arora (2010110038) and Sanjana Nair (2010110764) under Dr. Jaideep Ghosh as part of the DS Project Team for the Consulting firm CX, are presenting our findings in our project contracted out by SEBI. The data provided contains 9 different parameters for evaluating the salary scales of CEOs of Indian Firms. We have the CEO data from 177 Indian companies. Our work is to analyze the various variables to see which have a direct impact or subsidiary influence and to determine which variables to include and which to not. Using regression analysis, we can successfully analyze our data and give the best inference from it.

METHODOLOGY

In our data “SalaryData” we have one dependent variable, which is “Salary” and 9 Independent variables -
“Age”, “College”, “ Grad”, “Tenure”, “CeoTen”, “Sales”, “Profits”, “Mktval”, “Profmarg”.
We first consider the simplest regression model possible and observe what significance levels, adjusted R^2 value, etc, are obtained.
The code for the initial regression model is as follows:

```
> salarys1<-lm(Salary~College+Grad+Age+Tenure+CeoTen+Sales+Profits+Mktval+Profmarg,data=salary)
> summary(salarys1)
```

And the output is:

```
Call:
lm(formula = Salary ~ College + Grad + Age + Tenure + CeoTen + 
    Sales + Profits + Mktval + Profmarg, data = salary)

Residuals:
    Min      1Q  Median      3Q     Max 
-1108.9  -272.7  -104.8   212.1  4485.7 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)  
(Intercept)  711.48819  401.30075   1.773   0.0781 .
College     -132.53517  250.59622  -0.529   0.5976  
Grad         -56.05080   84.82088  -0.661   0.5096  
Age            3.17751    5.67010   0.560   0.5760  
Tenure        -5.27596    3.91034  -1.349   0.1791  
CeoTen        13.75820    6.17652   2.228   0.0273 *
Sales          0.01606    0.01133   1.417   0.1582  
Profits        0.10527    0.28315   0.372   0.7105  
Mktval         0.02115    0.01606   1.316   0.1898  
Profmarg      -1.83252    2.33473  -0.785   0.4336  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 534.4 on 167 degrees of freedom
Multiple R-squared:  0.2153,    Adjusted R-squared:  0.173 
F-statistic: 5.091 on 9 and 167 DF,  p-value: 4.383e-06
```

With the above output, we notice that most of the variables are not statistically significant. Moreover, we also see how the Adjusted R^2 is 0.173, and our goal will be to increase the adjusted R^2 to be as optimum as possible.

NORMALIZING THE VARIABLES

Now, to increase the efficiency of our regression model, our first step is to normalize as many variables as possible. We normalize to have most of the variables occupying the same scale. To do this we use the Shapiro-Wilk Test to first see which of the variable columns are already normally distributed, and if they are not normally distributed, we use log, square root, and other methods to make them normal. Here, the Null Hypothesis is that the input variable is normally distributed and so if the p-value is less than 0.05 and then we can reject the Null Hypothesis and the variable is not normally distributed. We do not check normality for our control variables: College, Grad as their range is discrete i.e.{0,1} and normal distribution is continuous.

AGE:

```
> shapiro.test(salary$Age)

        Shapiro-Wilk normality test

data:  salary$Age
W = 0.98753, p-value = 0.1195
```

Here we cannot reject the fact that the variable is not normally distributed.

TENURE:

```
> shapiro.test(salary$Tenure)

        Shapiro-Wilk normality test

data:  salary$Tenure
W = 0.95425, p-value = 1.669e-05

> shapiro.test(log(salary$Tenure))

        Shapiro-Wilk normality test

data:  log(salary$Tenure)
W = 0.89055, p-value = 4.03e-10

> shapiro.test(1/salary$Tenure)

        Shapiro-Wilk normality test

data:  1/salary$Tenure
W = 0.65948, p-value < 2.2e-16

> #Tenure is not normal and is not transforming to Normal
```

We see that when we check the normality of tenure we can reject the null hypothesis and we know that the Tenure variable is not normally distributed. So, we transform the variable to log to check and see the same output. Further, we try the multiplicative inverse, but do not get a different output. We conclude with the fact that Tenure is not transforming to normal.

CEOTEN:

```
> shapiro.test(salary$CeoTen)

        Shapiro-Wilk normality test

data:  salary$CeoTen
W = 0.84332, p-value = 1.653e-12

> shapiro.test(log(salary$CeoTen+1))

        Shapiro-Wilk normality test

data:  log(salary$CeoTen + 1)
W = 0.98127, p-value = 0.01757

> #CeoTen initially rejected as not normal has been transformed to normal.
```

After analyzing for CeoTen and noticing the variable is rejected as being normally distributed. So, we transform with log and see that the p-value is higher than 0.05 and we cannot reject it as normally distributed. Some of the entries in CeoTen are 0, so we take the log of CeoTen+1 as the log is not defined for 0, and adding 1 to each entry does not affect their relative ordering.

SALES:

```
> shapiro.test(salary$Sales)

        Shapiro-Wilk normality test

data:  salary$Sales
W = 0.54873, p-value < 2.2e-16

> shapiro.test(log(salary$Sales))

        Shapiro-Wilk normality test

data:  log(salary$Sales)
W = 0.99568, p-value = 0.8936

> #Salary initially rejected as not normal has been transformed to normal.
```

Similar to CeoTen, after rejecting the null hypothesis for Salary, we are unable to reject it for the log of Salary.

PROFITS:

```
> shapiro.test(salary$Profits)

        Shapiro-Wilk normality test

data:  salary$Profits
W = 0.59733, p-value < 2.2e-16

> shapiro.test(log(salary$Profits+463+1))

        Shapiro-Wilk normality test

data:  log(salary$Profits + 463 + 1)
W = 0.51807, p-value < 2.2e-16

> shapiro.test(sqrt(salary$Profits+463+1))

        Shapiro-Wilk normality test

data:  sqrt(salary$Profits + 463 + 1)
W = 0.7135, p-value < 2.2e-16

> shapiro.test((salary$Profits+463+1)^2)

        Shapiro-Wilk normality test

data:  (salary$Profits + 463 + 1)^2
W = 0.387, p-value < 2.2e-16

> #Profits is not normal and is not transforming to Normal
```

Some of the entries in Profits are negative, so we take the log of CeoTen+463+1 since 463 is the most negative entry and 1 is added if any entry has 0 profit value. Also, this does not affect their relative ordering hence we are allowed to do this transformation.

MKTVAL:

```
> shapiro.test(salary$Mktval)

        Shapiro-Wilk normality test

data:  salary$Mktval
W = 0.51092, p-value < 2.2e-16

> shapiro.test(log(salary$Mktval))

        Shapiro-Wilk normality test

data:  log(salary$Mktval)
W = 0.9201, p-value = 2.909e-08

> shapiro.test(sqrt(salary$Mktval))

        Shapiro-Wilk normality test

data:  sqrt(salary$Mktval)
W = 0.74218, p-value = 2.684e-16

> shapiro.test((salary$Mktval)^2)

        Shapiro-Wilk normality test

data:  (salary$Mktval)^2
W = 0.24186, p-value < 2.2e-16

> shapiro.test(log10(salary$Mktval))

        Shapiro-Wilk normality test

data:  log10(salary$Mktval)
W = 0.9201, p-value = 2.909e-08

> #Mktval is not normal and is not transforming to Normal
```

We fail to not reject the Null hypothesis that Mktval is normally distributed since the p-value is less than 0.05 and hence we are not able to transform the Mktval to a random variable.We decide to use this variable in our model nonetheless.

PROFMARG:

```
> shapiro.test(salary$Profmarg)

        Shapiro-Wilk normality test

data:  salary$Profmarg
W = 0.35145, p-value < 2.2e-16

> shapiro.test(log(salary$Profmarg))

        Shapiro-Wilk normality test

data:  log(salary$Profmarg)
W = 0.9531, p-value = 2.14e-05

Warning message:
In log(salary$Profmarg) : NaNs produced
> shapiro.test(log(salary$Profmarg+1))

        Shapiro-Wilk normality test

data:  log(salary$Profmarg + 1)
W = 0.98101, p-value = 0.02055

Warning message:
In log(salary$Profmarg + 1) : NaNs produced
> shapiro.test(log(salary$Profmarg+1+203.0769))

        Shapiro-Wilk normality test

data:  log(salary$Profmarg + 1 + 203.0769)
W = 0.091748, p-value < 2.2e-16

> shapiro.test(log10(salary$Profmarg+1+203.0769))

        Shapiro-Wilk normality test

data:  log10(salary$Profmarg + 1 + 203.0769)
W = 0.091748, p-value < 2.2e-16

> shapiro.test((salary$Profmarg+1+203.0769)^2)

        Shapiro-Wilk normality test

data:  (salary$Profmarg + 1 + 203.0769)^2
W = 0.60444, p-value < 2.2e-16
```

Some of the entries in Profmarg are negative, so we take the log of Profmarg +203.0769+1 since 203.0769 is the most negative entry and 1 is added if any entry has 0 Profmarg value. Also, this does not affect their relative ordering hence we are allowed to do this transformation, just like we did for Profits.

Hence, our model after checking for normality for each variable and transforming, we get:

```
> salarys1<-lm(log(Salary)~College+Grad+Age+Tenure+log(CeoTen+1)+log(Sales)+Mktval+Profmarg+Profits,data=salary)
> summary(salarys1)
```

```
Call:
lm(formula = log(Salary) ~ College + Grad + Age + Tenure + log(CeoTen +
    1) + log(Sales) + Mktval + Profmarg + Profits + Age, data = salary)

Residuals:
     Min       1Q   Median       3Q      Max
-2.37434 -0.26274 -0.00572  0.26824  1.91833

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.988e+00  4.220e-01  11.819  < 2e-16 ***
College         -5.373e-02  2.298e-01  -0.234  0.81542
Grad            -9.180e-02  7.786e-02  -1.179  0.24010
Age             -1.282e-04  5.152e-03  -0.025  0.98018
Tenure          -1.064e-02  3.583e-03  -2.968  0.00344 **
log(CeoTen + 1)  1.922e-01  4.907e-02   3.917  0.00013 ***
log(Sales)       2.133e-01  3.364e-02   6.340 2.06e-09 ***
Mktval           1.255e-05  1.469e-05   0.855  0.39403
Profmarg        -2.642e-03  2.121e-03  -1.246  0.21461
Profits          3.623e-05  2.418e-04   0.150  0.88107
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4891 on 167 degrees of freedom
Multiple R-squared:  0.3821,    Adjusted R-squared:  0.3488
F-statistic: 11.48 on 9 and 167 DF,  p-value: 6.426e-14
```

We can observe that the adjusted R^2 value has gone up from 0.173 to 0.3564, and the strength of the significance levels for the intercept and CeoTen has increased, and while the variables Tenure, Profits, Mktval, Profmarg are not normal we decide to keep them in our model as our dependent variable is normalized and we do not want to lose the information provided by these variables.

CHECKING FOR MULTICOLLINEARITY AND SUBSIDIARY INFLUENCES:

Now, to check if any independent variables in our model have any subsidiary influences on other independent variables, we take the following steps:

We first compute the Variation Inflation Factors for each variable;

```
> vif(salarys1)
       College           Grad            Age         Tenure log(CeoTen + 1)      log(Sales)
      1.072611       1.117324       1.385097       1.428040       1.133139       1.707516
        Mktval       Profmarg        Profits
      6.587716       1.055960       7.036851
```

Vif value of a variable indicates the effect its dependency (on other independent variables) has on its variance. So a high VIFvalue implies that there is high multicollinearity in the model. This would mean that at least one of the independent variables can be explained as dependent on the other independent variables in the model which would give incorrect values for the beta coefficients in the model.

The VIF values for Profits and Mktval are greater than 5 which is the standard lower bound for a high VIF value.

```
> cor(salary)
                Id       Salary          Age      College         Grad       Tenure       CeoTen        Sales
Id       1.000000000 -0.057845626 -0.07402231  0.08475763  0.084198608 -0.007918751 -0.051436404 -0.16019364
Salary  -0.057845626  1.000000000  0.11538394 -0.06702522 -0.002999832  0.037698187  0.142947678  0.38022387
Age     -0.074022308  0.115383944  1.00000000 -0.17806227 -0.123163323  0.479413536  0.338741704  0.12713402
College  0.084757626 -0.067025223 -0.17806227  1.00000000  0.181445273 -0.157109257 -0.106288424 -0.02149227
Grad     0.084198608 -0.002999832 -0.12316332  0.18144527  1.000000000 -0.228334613 -0.102806453  0.07632622
Tenure  -0.007918751  0.037698187  0.47941354 -0.15710926 -0.228334613  1.000000000  0.315121243  0.10439983
CeoTen  -0.051436404  0.142947678  0.33874170 -0.10628842 -0.102806453  0.315121243  1.000000000 -0.06771469
Sales   -0.160193645  0.380223875  0.12713402 -0.02149227  0.076326224  0.104399833 -0.067714685  1.00000000
Profits -0.130370075  0.393927574  0.11474310 -0.04598209  0.097825529  0.143737237 -0.021606750  0.79828723
Mktval  -0.103375898  0.406307097  0.10717932 -0.02757797  0.122976062  0.136095997  0.006609425  0.75466160
Profmarg -0.159147412 -0.028935405  0.01467792 -0.01753082 -0.015395217  0.047173891  0.048804685 -0.01735349


               Profits       Mktval    Profmarg
Id       -0.13037008 -0.103375898 -0.15914741
Salary    0.39392757  0.406307097 -0.02893541
Age       0.11474310  0.107179325  0.01467792
College  -0.04598209 -0.027577967 -0.01753082
Grad      0.09782553  0.122976062 -0.01539522
Tenure    0.14373724  0.136095997  0.04717389
CeoTen   -0.02160675  0.006609425  0.04880468
Sales     0.79828723  0.754661598 -0.01735349
Profits   1.00000000  0.918127962  0.12547925
Mktval    0.91812796  1.000000000  0.06701876
Profmarg  0.12547925  0.067018756  1.00000000
```

We also check the correlation among the variables to further investigate which combinations of variables may be contributing to the multicollinearity of the model. The correlation between Profits-Sales, Mkvtval-Sales, and Mkvtval-Profits is higher than 0.5 which is set as our lower bound for a high correlation value.

Now we test the effect on the VIF values of the model when we remove Profits, Mktval, and both Profits and Mktval:

Removing Profits and MktVal:

```
> salarys1<-lm(log(Salary)~College+Grad+Age+Tenure+log(CeoTen+1)+log(Sales)+Profmarg,data=salary)
> summary(salarys1)
```

```
Call:
lm(formula = log(Salary) ~ College + Grad + Age + Tenure + log(CeoTen +
    1) + log(Sales) + Profmarg, data = salary)

Residuals:
     Min       1Q   Median       3Q      Max
-2.32472 -0.27159 -0.02226  0.28330  1.87286

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.7475855  0.4065570  11.678  < 2e-16 ***
College        -0.0511207  0.2311453  -0.221 0.825233
Grad           -0.0774860  0.0779752  -0.994 0.321777
Age            -0.0001839  0.0051841  -0.035 0.971746
Tenure         -0.0105326  0.0036059  -2.921 0.003966 **
log(CeoTen + 1) 0.1939470  0.0491859   3.943 0.000118 ***
log(Sales)      0.2517115  0.0273325   9.209  < 2e-16 ***
Profmarg       -0.0021940  0.0020876  -1.051 0.294781
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4922 on 169 degrees of freedom
Multiple R-squared:  0.3666,    Adjusted R-squared:  0.3403
F-statistic: 13.97 on 7 and 169 DF,  p-value: 3.117e-14
```

```
> vif(salarys1)
       College           Grad            Age         Tenure log(CeoTen + 1)      log(Sales)
      1.071371       1.106089       1.384592       1.427629       1.123668       1.112898
      Profmarg
      1.009868
```

Removing MktVal:

```
> salarys1<-lm(log(Salary)~College+Grad+Age+Tenure+log(CeoTen+1)+log(Sales)+Profits+Profmarg,data=salary)
> summary(salarys1)
```

```
Call:
lm(formula = log(Salary) ~ College + Grad + Age + Tenure + log(CeoTen +
    1) + log(Sales) + Mktval + Profmarg, data = salary)

Residuals:
    Min      1Q  Median      3Q     Max
-2.3803 -0.2535 -0.0041  0.2689  1.9189

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.982e+00  4.186e-01  11.900  < 2e-16 ***
College        -5.487e-02  2.290e-01  -0.240 0.810936
Grad           -9.224e-02  7.758e-02  -1.189 0.236116
Age            -1.143e-04  5.136e-03  -0.022 0.982264
Tenure         -1.064e-02  3.573e-03  -2.978 0.003329 **
log(CeoTen + 1) 1.916e-01  4.874e-02   3.931 0.000123 ***
log(Sales)      2.144e-01  3.261e-02   6.576 5.85e-10 ***
Mktval          1.448e-05  7.059e-06   2.051 0.041821 *
Profmarg       -2.582e-03  2.077e-03  -1.243 0.215468
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4876 on 168 degrees of freedom
Multiple R-squared:  0.382,     Adjusted R-squared:  0.3526
F-statistic: 12.98 on 8 and 168 DF,  p-value: 1.726e-14
```

```
> vif(salarys1)
       College           Grad            Age         Tenure log(CeoTen + 1)      log(Sales)
      1.071457       1.110487       1.384603       1.427737       1.124218       1.702110
       Profits       Profmarg
      1.635268       1.040174
```

Removing Profits:

```
> salarys1<-lm(log(Salary)~College+Grad+Age+Tenure+log(CeoTen+1)+log(Sales)+Mktval+Profmarg,data=salary)
> summary(salarys1)
```

```
Call:
lm(formula = log(Salary) ~ College + Grad + Age + Tenure + log(CeoTen +
    1) + log(Sales) + Mktval + Profmarg, data = salary)

Residuals:
    Min      1Q  Median      3Q     Max
-2.3803 -0.2535 -0.0041  0.2689  1.9189

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.982e+00  4.186e-01  11.900  < 2e-16 ***
College         -5.487e-02  2.290e-01  -0.240 0.810936
Grad            -9.224e-02  7.758e-02  -1.189 0.236116
Age             -1.143e-04  5.136e-03  -0.022 0.982264
Tenure          -1.064e-02  3.573e-03  -2.978 0.003329 **
log(CeoTen + 1)  1.916e-01  4.874e-02   3.931 0.000123 ***
log(Sales)       2.144e-01  3.261e-02   6.576 5.85e-10 ***
Mktval           1.448e-05  7.059e-06   2.051 0.041821 *
Profmarg        -2.582e-03  2.077e-03  -1.243 0.215468
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4876 on 168 degrees of freedom
Multiple R-squared:  0.382,     Adjusted R-squared:  0.3526
F-statistic: 12.98 on 8 and 168 DF,  p-value: 1.726e-14
```

```
> vif(salarys1)
       College           Grad            Age         Tenure log(CeoTen + 1)      log(Sales)
      1.071440       1.115686       1.384652       1.427935       1.124291       1.614114
        Mktval       Profmarg
      1.530895       1.018324
```

After analyzing we observe the following:

- Correlations
  After seeing how Profits-Sales(0.79828723), Mkvtval-Sales(0.75466160), Mkvtval-Profits (0.918127962) have a higher correlation than 0.5, we notice how the correlations of variables which we relate to Profits are higher than with the other two variables.
- Adjusted R^2
  With the three different R^2 values, we notice that the drop in adjusted R^2 is the least when only Profits are removed from the regression model
- Vifs
  Moreover, by calculating the Vifs of all the versions, we notice a significant dip in the VIF for Mktval which is desirable i.e. it fell below 5. Hence we can keep Mktval and use its information, and we don’t need to remove both Mktval and Profits.

So from these, we infer that the variable Profits are not positively contributing to the regression model, and we conclude that Profits should not be in the regression model.

CHECKING FOR POSSIBLE IMPROVEMENT IN MODEL BY REMOVING ANY NON-STATISTICALLY SIGNIFICANT IV'S:

From the last regression model, we see that apart from our Control variables- College and Grad, Age and ProfMarg are the two independent variables that are still not statistically significant. So to check the impact of these two variables on our model, we observe what happens when we remove either one of the variables and if the adjusted R^2 and the statistical significance of the IV change.

REMOVING PROFMARG:

```
> salarys1<-lm(log(Salary)~College+Grad+Tenure+log(CeoTen+1)+log(Sales)+Mktval+Age,data=salary)
> summary(salarys1)

Call:
lm(formula = log(Salary) ~ College + Grad + Tenure + log(CeoTen +
    1) + log(Sales) + Mktval + Age, data = salary)

Residuals:
     Min       1Q   Median       3Q      Max
-2.39177 -0.25239 -0.00783  0.26323  1.92541

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.951e+00  4.186e-01  11.828  < 2e-16 ***
College         -5.265e-02  2.294e-01  -0.230 0.818715
Grad            -9.150e-02  7.770e-02  -1.178 0.240638
Tenure          -1.080e-02  3.576e-03  -3.020 0.002920 **
log(CeoTen + 1)  1.868e-01  4.866e-02   3.838 0.000175 ***
log(Sales)       2.170e-01  3.259e-02   6.659 3.71e-10 ***
Mktval           1.368e-05  7.041e-06   1.943 0.053721 .
Age              3.625e-05  5.143e-03   0.007 0.994385
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4884 on 169 degrees of freedom
Multiple R-squared:  0.3764,    Adjusted R-squared:  0.3505
F-statistic: 14.57 on 7 and 169 DF,  p-value: 8.914e-15
```

REMOVING AGE:

```
> salarys1<-lm(log(Salary)~College+Grad+Tenure+log(CeoTen+1)+log(Sales)+Mktval+Profmarg,data=salary)
> summary(salarys1)

Call:
lm(formula = log(Salary) ~ College + Grad + Tenure + log(CeoTen +
    1) + log(Sales) + Mktval + Profmarg, data = salary)

Residuals:
    Min      1Q  Median      3Q     Max
-2.3804 -0.2531 -0.0043  0.2680  1.9188

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.977e+00  3.425e-01  14.529  < 2e-16 ***
College         -5.437e-02  2.272e-01  -0.239  0.81119
Grad            -9.222e-02  7.735e-02  -1.192  0.23479
Tenure          -1.067e-02  3.274e-03  -3.259  0.00135 **
log(CeoTen + 1)  1.914e-01  4.759e-02   4.021 8.70e-05 ***
log(Sales)       2.144e-01  3.236e-02   6.624 4.46e-10 ***
Mktval           1.448e-05  7.038e-06   2.057  0.04120 *
Profmarg        -2.581e-03  2.070e-03  -1.247  0.21417
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4862 on 169 degrees of freedom
Multiple R-squared:  0.382,     Adjusted R-squared:  0.3564
F-statistic: 14.93 on 7 and 169 DF,  p-value: 4.264e-15
```

From the above two regression models, we see the following:

- The Adjusted R^2 is higher when we remove Age than when we remove ProfMarg.The adjusted R^2 of the model when we remove Age is also higher than the model when we had both Age and Profmarg but not Profits.
- The statistical significance of MktVal is lesser when ProfMarg is removed than when Age is removed.

Due to the above two points, we have decided to remove Age, and our analysis tells us removing Age improves adjusted R^2 without hindering the significance values of the IVS and is most optimum for our regression model. Moreover, practically thinking about this in real life as “Age is just a number”, Age should not have an effect on a CEO’s Salary.

DOING HYPOTHESIS TESTING FOR INDEPENDENT VARIABLES IN THE MODEL

College and Grad are the control variables, and hence we do not need to do any hypothesis testing for them.

Tenure:

NULL: The coefficient of Tenure in the model is 0

ALTERNATIVE: The coefficient of Tenure in the model is not 0

Since the p-value(0.00135) <0.01 We reject the Null hypothesis, hence the coefficient of Tenure in the model is not 0

log(CeoTen+1):

NULL: The coefficient of log(CeoTen+1) in the model is 0

ALTERNATIVE: The coefficient of log(CeoTen+1) in the model is not 0

Since the p-value(8.70e-05) <0.001 We reject the Null hypothesis, hence the coefficient of Tenure in the model is not 0

log(Sales):

NULL: The coefficient of log(Sales) in the model is 0

ALTERNATIVE: The coefficient of log(Sales) in the model is not 0

Since the p-value(4.46e-10) <0.001 We reject the Null hypothesis, hence the coefficient of log(Sales) in the model is not 0

Mktval:

NULL: The coefficient of Mktval in the model is 0

ALTERNATIVE: The coefficient of Mktval in the model is not 0

Since the p-value(0.0418) <0.05 We reject the Null hypothesis , hence the coefficient of Mktval in the model is not 0

Profmarg:

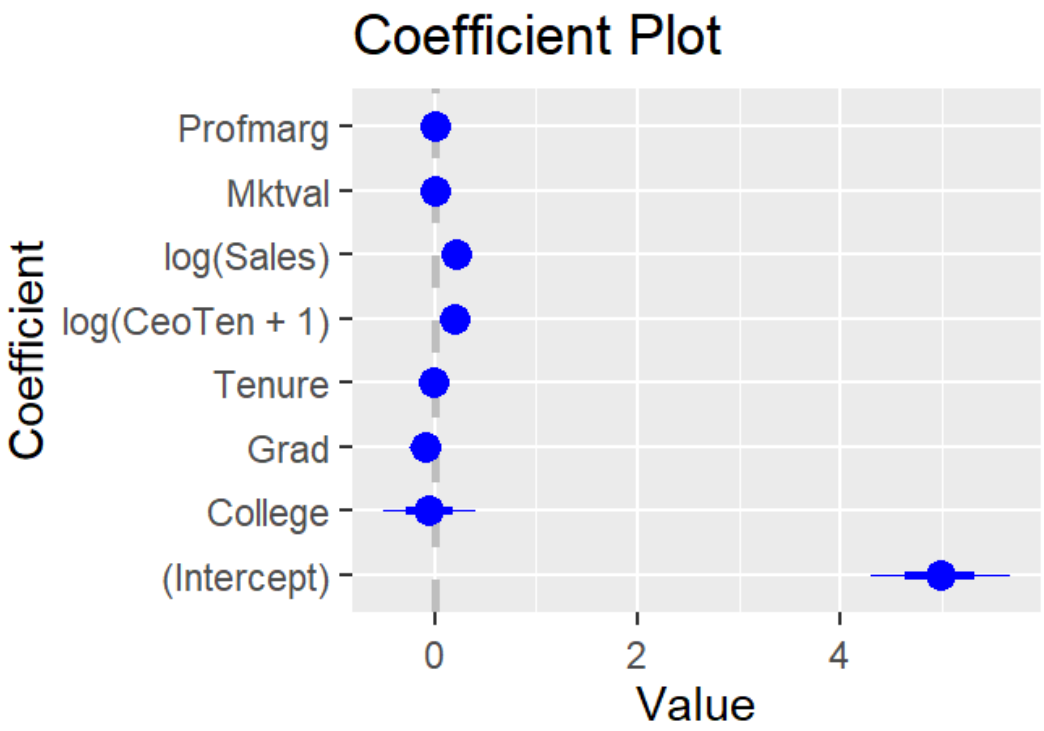
NULL: The coefficient of Profmargin the model is 0

ALTERNATIVE: The coefficient of Profmarg in the model is not 0

Since the p-value(0.21) >0.05 We cannot reject the Null hypothesis , hence the coefficient of Tenure in the model cannot be conclude to not be 0

But we decide to include Profmarg in the model because most of the independent variables are statistically significant and the overall p value for the model is 4.26e-15<0.05. Also, when only Profmarg has been removed from the model, the adjusted R^2 drops and the significance level of Mktval is increased from 0.05 to 0.1. Hence we decided that Mktval should be kept in the model.

```
> coef(salarys1)
    (Intercept)         College            Grad          Tenure
   4.976537e+00   -5.437054e-02   -9.222344e-02   -1.067105e-02
log(CeoTen + 1)      log(Sales)          Mktval         Profmarg
   1.913731e-01    2.143726e-01    1.447998e-05   -2.581052e-03
```



Profmarg is not statistically significant, which matches with the coefficient plot where the confidence interval contains 0. We know Mktval, log(Sales), log(CeoTen+1), and Tenure are statistically significant from the summary of the model. We can also observe that log(Sales) and log(CeoTen+1) lie to the right of 0, and since they are statistically significant, their confidence intervals will not contain 0. Tenure and Grad similarly lie to the left of 0. Though Mktval seems to lie on 0 since it shows to be statistically significant through the code, it should have a very small confidence interval which does not contain 0.

RESULTS AND DISCUSSION

Our final regression model is

```
> salarys1<-lm(log(Salary)~College+Grad+Tenure+log(CeoTen+1)+log(Sales)+Mktval+Profmarg,data=salary)
> summary(salarys1)

Call:
lm(formula = log(Salary) ~ College + Grad + Tenure + log(CeoTen +
    1) + log(Sales) + Mktval + Profmarg, data = salary)

Residuals:
    Min      1Q  Median      3Q     Max
-2.3804 -0.2531 -0.0043  0.2680  1.9188

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.977e+00  3.425e-01  14.529  < 2e-16 ***
College         -5.437e-02  2.272e-01  -0.239  0.81119
Grad            -9.222e-02  7.735e-02  -1.192  0.23479
Tenure          -1.067e-02  3.274e-03  -3.259  0.00135 **
log(CeoTen + 1)  1.914e-01  4.759e-02   4.021 8.70e-05 ***
log(Sales)       2.144e-01  3.236e-02   6.624 4.46e-10 ***
Mktval           1.448e-05  7.038e-06   2.057  0.04120 *
Profmarg        -2.581e-03  2.070e-03  -1.247  0.21417
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4862 on 169 degrees of freedom
Multiple R-squared:  0.382,     Adjusted R-squared:  0.3564
F-statistic: 14.93 on 7 and 169 DF,  p-value: 4.264e-15
```

Discussion of the coefficients:

Variable	**Coefficient**	**Significance level**
Intercept	4.977	0.001
College(Control)	-5.437e^-02	Not applicable
Grad(Control)	-9.222e^-02	Not applicable
Tenure	-1.067e^-02	0.01
log(CeoTen+1)	1.914e^-01	0.001
log(Sales)	2.144e^-01	0.001
MktVal	1.44e^-05	0.05
ProfMarg	2.581e^-03	Not significant

R^2	Adjusted R^2
0.382	0.3564

Since in the regression model our LHS is log(Salary) to get the change in only Salary when the other independent variables are kept constant, in the LHS we would shift log to RHS and convert it to the exponential of each variable's coefficient. As we know, the value of an exponential is always positive. Therefore, College, Grad, Tenure, CeoTen, Sales, MktVal, and ProfMarg all have a positive impact on a CEO's Salary. This positive impact makes perfect sense in the real world as the given variables should intuitively have a positive impact on a CEO's Salary.

To elaborate on the above information:
For the variables where the coefficient is negative, the exponential value overall will be positive however less than 1. So the impact for College, Grad, and Tenure is positive but less than 1, hence not as impactful.
For the variables where the coefficient is positive, the exponential value is positive and greater than 1. So for CeoTen, Sales, MktVal and Profmarg are positive and greater than 1, so quite impactful.

CONCLUSION

After our Regression analysis, we see how the variables we determined after transforming and optimizing all have a positive impact on the Salary of a CEO. We observe that some variables need to be normalized. Also, we find out that from the original 9 variables- Age, College, Grad, Tenure, CeoTen, Sales, Profits, Mktval, and Profmarg; we remove Age and Profits as these two variables bring in unnecessary data not optimally affecting our regression model.
When we relate our results to real life, we understand that the impact of College, Grad, and Tenure should not determine the CEO's salary as much as other IV's which means that the past does not reflect the worth of the CEO as strongly as what they do after they become a CEO that is, the Tenure of a CEO, the Sales, MktVal, and Profit Margin show the true worth of a CEO and determines their salary.