Xinjing Zhu
Fall 2011
Regression Analysis Student Project

Introduction

For my student project for the NEAS Regression Analysis course, I looked at executive pay, company performance and industry type. I will regress the executive pay (in thousands of dollars) on four explanatory variables (X_1, X_2, X_3, X_4) to determine which variables are good indicators of executive pay.

Data

I used the data file EXECPAY, which contains the following variables measured on 308 companies:
Column1 = consecutive ID numbers.
Column2 = industry codes: 1 = industrial products, 2 = consumer products, 3 = financial services, 4 = retail and service, 5 = metals and mining, 6 = energy, 7 = utilities
Column3 = sales in millions of dollars.
Column4 = pay in thousands of dollars.
Column5 = percentage change in pay from 1985.
Column6 = year end value of a $100 investment made 3 years earlier
Column8 = company's average percentage return on equity over 3 years

Variables and Regression Equation

I chose pay in thousands of dollars (PAY86) as my response variable Y. The explanatory variables are as follows:

- X_1: sales in million of dollars (SALES)
- X_2: percentage change in pay from 1985 (Pct.inc.pay)
- X_3: year end value of a $100 investment made 3 years earlier (INVEST.100)
- X_4: company's average percentage return on equity over 3 years (AVG.pct.ROE)

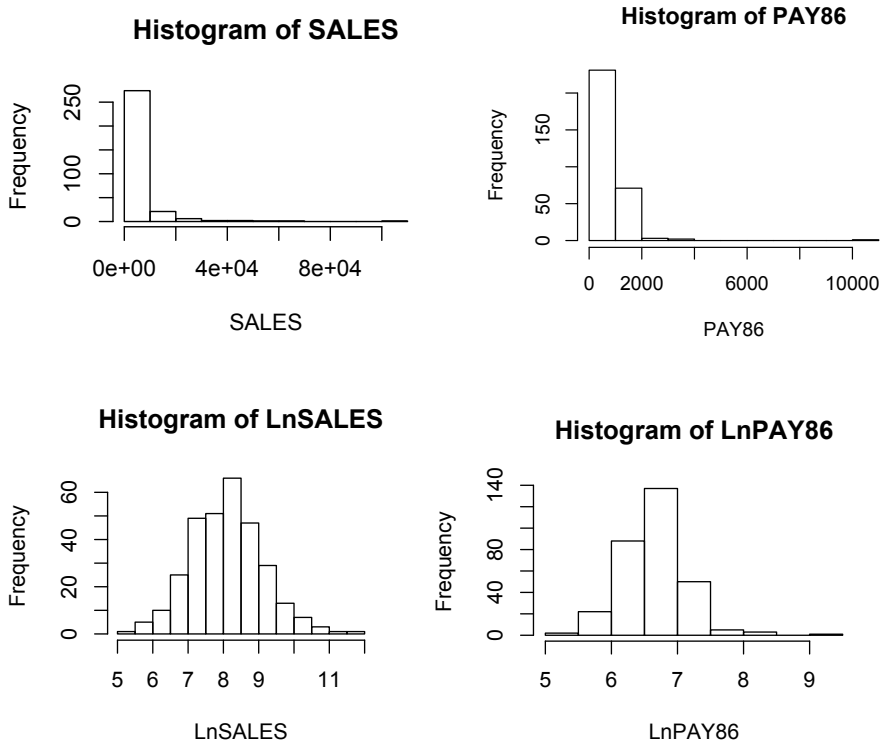
My regression equation will be: $Y = A + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4$

Hypothesis

My hypothesis is that all coefficients B_i (i = 1, …, 4) = 0.

Regression Analysis

I plotted the histograms of PAY86, SALES, Pct.inc.pay, INVEST.100, AVG.pct.ROE to check their distributions and found that PAY86 and SALES are not normally distributed. Then I used logarithmic transformations to transform PAY86 and SALES and rename them as LnPAY86 and LnSALES.



Model #1:

For my first model, I will run a regression on all 4 variables using R. Below is the results:

```
Call: lm(formula = LnPAY86 ~ LnSALES + Pct.inc.pay + INVEST.100 + AVG.pct.ROE)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.5349046  0.1851256  24.496  < 2e-16 ***
LnSALES     0.2237953  0.0210304  10.641  < 2e-16 ***
Pct.inc.pay 0.0035029  0.0004967   7.052  1.2e-11 ***
INVEST.100  0.0005534  0.0004011   1.380  0.16873
AVG.pct.ROE 0.0087652  0.0028650   3.059  0.00242 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3672 on 303 degrees of freedom
Multiple R-squared: 0.394,     Adjusted R-squared: 0.386
F-statistic: 49.24 on 4 and 303 DF,  p-value: < 2.2e-16 Multiple R-squared: 0.5488,
```

The regression equation using all 4 explanatory variables becomes:
$Y = 4.5349 + 0.2238X_1 + 0.0035X_2 + 0.0006X_3 + 0.0088X_4$

An R squared value of 0.394 indicates that using all explanatory variables is an OK predictor of the executive pay. The F statistic for this model is 49.24. I will compare this with the F statistics from models later.

Looking at each variable closer I can see that the INVEST.100 variable has a much larger p-value compared to the other variables. The absolute value of the t-statistic is also the smallest. I will remove this from my model and run a regression.

Model #2:

After removing the INVEST.100 variable, I ran a regression with the remaining and got the following results.

```
Call: lm(formula = LnPAY86 ~ LnSALES + Pct.inc.pay + AVG.pct.ROE)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.6179362  0.1753290  26.339  < 2e-16 ***
LnSALES     0.2218574  0.0210146  10.557  < 2e-16 ***
Pct.inc.pay 0.0035760  0.0004946   7.230 3.94e-12 ***
AVG.pct.ROE 0.0107398  0.0024855   4.321 2.11e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3678 on 304 degrees of freedom
Multiple R-squared: 0.3901,   Adjusted R-squared: 0.3841
F-statistic: 64.83 on 3 and 304 DF,  p-value: < 2.2e-16
```

The regression equation using 3 variables becomes:
$Y = 4.6179 + 0.2219X_1 + 0.0036X_2 + 0.0107X_4$

An R squared value of 0.3901 indicates that using all explanatory variables is an OK predictor of the executive pay. The F statistic for this model is 64.83. In this model, all variables are significant.

Conclusion

The following chart summarizes R squared and F statistics for the 2 models I created:

	Model # 1	Model # 2
R Squared	39.4%	39%
F Statistic	49.24	64.83

The model I would choose is Model #2. The R squared of 39% is quite close to that of model #1 (39.4%), However, Model #2 has higher F statistic (64.83) compared to Model #1 (49.24). In addition, all 3 variables LnSALES, Pct.inv.pay and AVG.pct.ROE in Model #2 are also statistically significant on a 95% confidence level.

Therefore, the model I am choosing is $Y = 4.6179 + 0.2219X_1 + 0.0036X_2 + 0.0107X_4$ and I can reject my null hypothesis. The response variables that are used in this model are LnSALES, Pct.inv.pay and AVG.pct.ROE. This means that those are the variables that affect the executive pay the most.