Final Project RMD

Marali Benitez, Youcef Djoudi, Kelly Thackery, Samina Issa, Samira Bechi

2025-04-09

##loading in packages and data  
library(wooldridge)  
library(AER)  
library(quantmod)  
library(PerformanceAnalytics)  
library(lmtest)  
  
#ceosal2 data  
ceosalary <- ceosal2  
  
#Murder rates data  
data("MurderRates")  
mr <- MurderRates

**Data: ceosal2**

**(a) Describe the data. Describe the variables; response, predictors, continuous, categorical variables and missing data**.

library(wooldridge)  
data("ceosal2")  
head(ceosal2)

## salary age college grad comten ceoten sales profits mktval lsalary lsales  
## 1 1161 49 1 1 9 2 6200 966 23200 7.057037 8.732305  
## 2 600 43 1 1 10 10 283 48 1100 6.396930 5.645447  
## 3 379 51 1 1 9 3 169 40 1100 5.937536 5.129899  
## 4 651 55 1 0 22 22 1100 -54 1000 6.478509 7.003066  
## 5 497 44 1 1 8 6 351 28 387 6.208590 5.860786  
## 6 1067 64 1 1 7 7 19000 614 3900 6.972606 9.852194  
## lmktval comtensq ceotensq profmarg  
## 1 10.051908 81 4 15.580646  
## 2 7.003066 100 100 16.961130  
## 3 7.003066 81 9 23.668638  
## 4 6.907755 484 484 -4.909091  
## 5 5.958425 64 36 7.977208  
## 6 8.268732 49 49 3.231579

*This dataset (ceosal2) includes information on CEOs and their companies. The analysis focuses on how various CEO and firm characteristics influence company profits*.

##### Variables:

**Response Variable:** - profits: Annual firm profits (in millions).  
- Type: **Continuous**.

| Variable | Description | Type |
| --- | --- | --- |
| salary | CEO salary (in thousands) | Continuous |
| age | CEO age | Continuous |
| college | 1 = undergraduate degree | Categorical (binary) |
| grad | 1 = graduate degree | Categorical (binary) |
| comten | Years with the company | Continuous |
| ceoten | Years as CEO | Continuous |
| sales | Company sales (in millions) | Continuous |
| mktval | Market value (in millions) | Continuous |
| lmktval | Log of market value | Continuous |
| comtensq | Square of company tenure | Continuous |
| ceotensq | Square of CEO tenure | Continuous |
| profmarg | Profit margin = profits/sales | Continuous |

**Excluded Predictors:** - lsalary, lsales: Log-transformed versions of salary and sales; excluded from final model.

#Check for missing values  
colSums(is.na(ceosal2))

## salary age college grad comten ceoten sales profits   
## 0 0 0 0 0 0 0 0   
## mktval lsalary lsales lmktval comtensq ceotensq profmarg   
## 0 0 0 0 0 0 0

*NO missing values in the data set*.

**(b) Build an optimal model. Print summary: discuss F-test, t-tests, Rsquare, S-square. Take out outliers**.

#preliminary model  
model <- lm(profits ~ . -lsalary -lmktval -lsales, data=ceosal2)  
summary(model)

##   
## Call:  
## lm(formula = profits ~ . - lsalary - lmktval - lsales, data = ceosal2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -595.72 -24.52 7.21 37.55 543.72   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 21.725679 119.334493 0.182 0.85576   
## salary 0.011629 0.021625 0.538 0.59146   
## age -0.067916 1.592252 -0.043 0.96603   
## college -47.570474 68.855153 -0.691 0.49061   
## grad -1.020730 23.409112 -0.044 0.96527   
## comten 1.261179 3.511631 0.359 0.71995   
## ceoten -4.988814 4.417210 -1.129 0.26037   
## sales 0.016682 0.002843 5.867 2.36e-08 \*\*\*  
## mktval 0.044778 0.002732 16.392 < 2e-16 \*\*\*  
## comtensq -0.014689 0.077158 -0.190 0.84925   
## ceotensq 0.133996 0.147531 0.908 0.36507   
## profmarg 1.900953 0.627057 3.032 0.00283 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 146.5 on 165 degrees of freedom  
## Multiple R-squared: 0.877, Adjusted R-squared: 0.8688   
## F-statistic: 107 on 11 and 165 DF, p-value: < 2.2e-16

*In the primary model, all variables were included except for lsalary, lmktval, and lsales. The R² value is 0.877, indicating that approximately 87.7% of the variation in profits is accounted for by the model. The F-test p-value is less than 2.2e-16, confirming that the model is statistically significant and not due to random chance. Based on the summary output, the predictors that show statistical significance are sales, market value (mktval), and profit margin (profmarg). The residual standard error is 146.5, suggesting that, on average, the model’s profit predictions differ from the actual values by about 146.5 units.*

# Optimal Model  
model1<-step(model)

## Start: AIC=1776.95

## Step: AIC=1763.32  
## profits ~ sales + mktval + profmarg  
##   
## Df Sum of Sq RSS AIC  
## <none> 3588346 1763.3  
## - profmarg 1 189835 3778181 1770.4  
## - sales 1 814705 4403051 1797.5  
## - mktval 1 6296009 9884355 1940.7

summary(model1)

##   
## Call:  
## lm(formula = profits ~ sales + mktval + profmarg, data = ceosal2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -617.41 -23.28 8.08 34.34 553.52   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -26.818425 13.237245 -2.026 0.04430 \*   
## sales 0.017124 0.002732 6.267 2.83e-09 \*\*\*  
## mktval 0.045084 0.002588 17.422 < 2e-16 \*\*\*  
## profmarg 1.852935 0.612487 3.025 0.00286 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 144 on 173 degrees of freedom  
## Multiple R-squared: 0.8754, Adjusted R-squared: 0.8732   
## F-statistic: 405 on 3 and 173 DF, p-value: < 2.2e-16

*The optimal modelincludes only sales, market value (mktval), and profit margin (profmarg) as predictors. This model achieves an R² of 0.8754, meaning it explains about 87.5% of the variation in profits. The F-test p-value is less than 2.2e-16, indicating the model is statistically significant. All three predictors are highly significant, with p-values less than 0.01. The residual standard error is 144, suggesting that on average, the model’s predictions deviate from actual profits by about 144 units.*

*The primary model included a wider range of predictors while still retaining sales, market value, and profit margin as key variables. In contrast, the optimal model selected only these three significant predictors using stepwise AIC. Despite having fewer variables, the optimal model maintains nearly the same explanatory power (R² ~87.5%).*

#taking out outliers  
cooks <- cooks.distance(model1)  
summary(cooks)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000000 0.000017 0.000095 0.102251 0.000630 11.768635

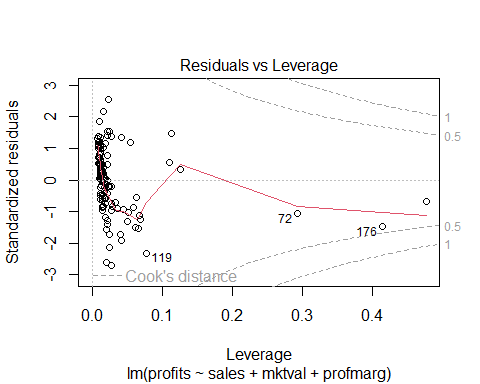
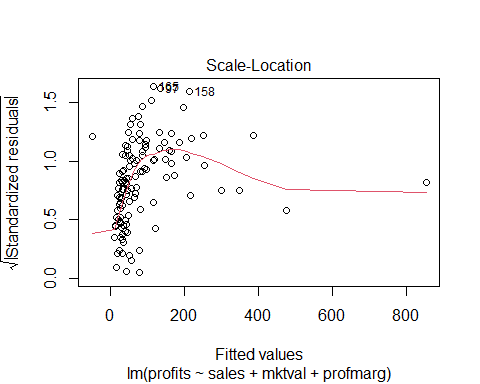
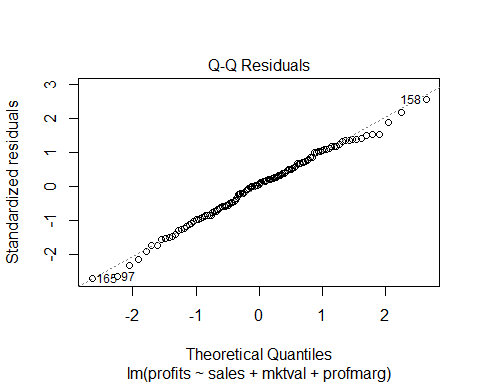
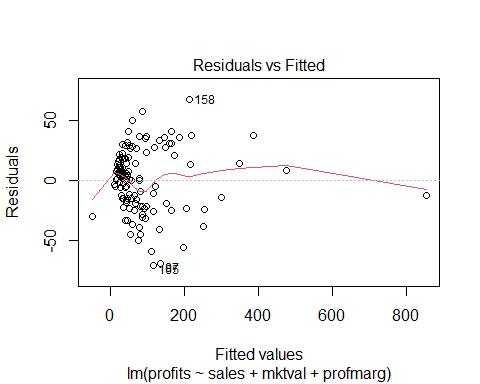
#creating reduced model with no outliers in model  
ceosal2.new <- ceosal2[ cooks <=0.000394 , ]  
model2 <- update(model1, data=ceosal2.new)  
summary(model2)

##   
## Call:  
## lm(formula = profits ~ sales + mktval + profmarg, data = ceosal2.new)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -70.531 -18.035 1.454 18.072 66.973   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -20.050318 5.024294 -3.991 0.000114 \*\*\*  
## sales 0.016492 0.002215 7.447 1.66e-11 \*\*\*  
## mktval 0.044318 0.002054 21.577 < 2e-16 \*\*\*  
## profmarg 2.023538 0.445915 4.538 1.37e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 26.59 on 119 degrees of freedom  
## Multiple R-squared: 0.9414, Adjusted R-squared: 0.9399   
## F-statistic: 637.4 on 3 and 119 DF, p-value: < 2.2e-16

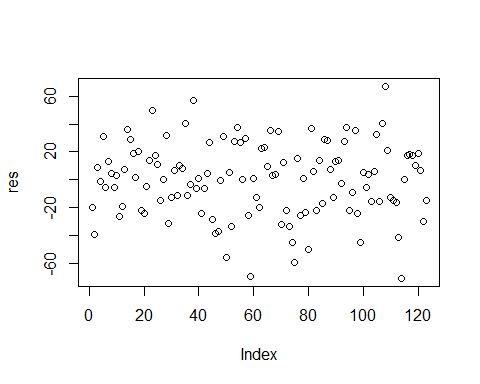
*After removing outliers, the refined model includes sales, market value (mktval), and profit margin (profmarg) as predictors. The R² value increased to 0.9414, indicating that the model now explains about 94.1% of the variation in profits—a substantial improvement. All predictors remain highly significant with p-values < 0.001. The residual standard error dropped to 26.59*

**(c) Briefly discuss residuals**.

res <- model2$residuals  
plot(model2)



plot(res)

 *Residuals vs Fitted Plot shows a fairly random scatter of residuals around zero, which suggests that the linearity assumption holds reasonably well. A slight curve at higher fitted values indicates some mild non-linearity.* *Q-Q Plot largely follow the 45-degree line, indicating that they are approximately normally distributed. There are a few deviations at the tails.* *The Scale-Location plot shows that the spread of the standardized residuals remains consistent across the range of fitted values.* *Residuals vs Leverage Plot’s most points fall within safe Cook’s distance lines.* *Index Plot confirms there’s no clear trend or clustering pattern in the residuals.* *These plots suggest that the reduced model fits the data well*

**(d) Make 2 predictions with CI’s. Interpret predictions if needed**.

# Hypothetical CEO 1 - using your final model (model2)  
new\_ceo1 <- data.frame(  
 sales = 4500,  
 mktval = 1800,  
 profmarg = 10  
)  
  
# Hypothetical CEO 2 - using your final model (model2)  
new\_ceo2 <- data.frame(  
 sales = 7000,  
 mktval = 3000,  
 profmarg = 15  
)  
  
#make predictions with 95% CI  
# Predictions for CEO 1  
predict(model2, newdata = new\_ceo1, interval = "confidence")

## fit lwr upr  
## 1 154.1698 140.1769 168.1627

# Predictions for CEO 2  
predict(model2, newdata = new\_ceo2, interval = "confidence")

## fit lwr upr  
## 1 258.6981 234.9522 282.444

*For CEO 1, the predicted profit is approximately 154.17 units, with a 95% confidence interval between 140.18 and 168.16, meaning the true average profit for similar companies is likely within this range.* *For CEO 2, the predicted profit is around 258.70 units, with a 95% confidence interval from 234.95 to 282.44, indicating a higher expected profit driven by greater sales, market value, and profit margin.*