HW 2 - Predictive Modeling in Finance and Insurance

Dennis Goldenberg

2024-02-01

1. Nursing Home Utilization

```
# import packages
library(ggplot2)
library(magrittr)
```

```
# read in data
# WNH <- read.csv(file)
WNH <- read.csv('WiscNursingHome.csv', header = TRUE)
WNH$CRYEAR <- factor(WNH$CRYEAR)
WNH <- WNH[WNH$CRYEAR == 2001,]
```

1a) Estimation of Coefficients

```
#Generate variables to analyze
WNH$LOGTPY <- log(WNH$TPY)
WNH$LOGNUMBED <- log(WNH$NUMBED)
```

Using the generated variables, I calculate $x^T x$, adding in a column for the intercept:

```
x <- cbind(1,WNH$LOGNUMBED)
xTx <- t(x) %*% x
xTx
```

```
##          [,1]     [,2]
## [1,]  355.000 1582.334
## [2,] 1582.334 7138.724
```

Then, I find $\left(x^T x\right)^{-1}$:

```
xTxInv <- solve(xTx)
xTxInv
```

```
##            [,1]        [,2]
## [1,]  0.2343245 -0.05193920
## [2,] -0.0519392  0.01165267
```

Finally, I find $x^T y$:

```
y <- WNH$LOGTPY
xTy <- t(x) %*% y
xTy
```

```
##          [,1]
## [1,] 1550.747
```

```
## [2,] 6999.582
```

Using the formula for linear regression that $\beta = (x^T x)^{-1} x^T y$:

```
beta <- xTxInv %*% xTy
beta
```

```
##              [,1]
## [1,] -0.1746945
## [2,]  1.0192307
```

1b. The prediction Matrix

Since $\hat{y} = x\hat{\beta}$, and $\beta = (x^T x)^{-1} x^T y$, the prediction matrix $H = x\left(x^T x\right)^{-1} x^T$, so:

$$\hat{y} = x(x^T x)^{-1} x^T y = Hy$$

I find the diagonals of said matrix H and store them in "leverages" variable, as they represent the leverage of each data point; the first 6 outputs are shown below to verify with the Excel document:

```
H <- x %*% xTxInv %*% t(x)
leverages <- diag(H)
head(leverages)
```

```
## [1] 0.031426544 0.006281299 0.005372343 0.004351815 0.003224867 0.002906796
```

1c. Making Predictions

Since $\hat{y} = Hy$, I calculate and store in the "pred" variable, showing the first 6 predicted values for verification with excel:

```
pred <- H %*% y
head(pred)
```

```
##          [,1]
## [1,] 2.771261
## [2,] 3.812560
## [3,] 3.891001
## [4,] 3.998387
## [5,] 4.559011
## [6,] 4.278781
```

1d. Calculating Summary Statistics

The R^2 value is the proportion of variation explained by the regression. R^2_{adj} is adjusted for the number of predictors; its formula is:

$$R^2_{adj} = 1 - \frac{\frac{SSE}{n-p-1}}{\frac{SST}{n-1}} = 1 - \frac{\frac{SSE}{n-2}}{\frac{SST}{n-1}}$$

Then, the F statistic measures the significance of the regression; its formula is:

$$F_{stat} = \frac{\frac{SST-SSE}{p}}{\frac{SSE}{N-(p+1)}} = \frac{SST - SSE}{\frac{SSE}{N-2}}$$

The p-value is simply the probability that so much variation was observed by a model with no predictive power:

$$p = \mathbb{P}(F \geq F_{stat}), \text{ where } F \sim \text{F-dist}(1, N-2)$$