DATA 1204 - Assignment5

Chinedu Onyeka

12/11/2021

Load Libraries

```
# Load libraries
library(tidyverse)
```

Hypothesis Statement

$H_0 : \beta = 0$, coefficient β of the dividend is zero and not statistically significant
$H_1 : \beta \neq 0$, co-efficient β of dividend is not equal to zero and is statistically significant

Read the file

```
url <- "https://raw.githubusercontent.com/chinedu2301/DC_Analytics/main/ols_stock.csv"
stock <- read_csv(url)
```

Check the head

```
# check the head
head(stock)
```

```
## # A tibble: 6 x 6
##   stock_return dividend earnings_ranking debt_to_equity marketcap
##          <dbl>    <dbl>            <dbl>          <dbl>     <dbl>
## 1          691        0               44           0.07       185
## 2         2038        0               28           0.09       207
## 3          371        0               48           0.12       288
## 4          515        0               45           0.17       545
## 5          752        0               43           0.23       241
## 6          433        0               46           0.31       665
## # ... with 1 more variable: stock_return_scaled <dbl>
```

Check summary of the data

```
# check the summary
summary(stock)
```

```
##   stock_return     dividend      earnings_ranking debt_to_equity
##  Min.   : 202   Min.   :0.0000   Min.   : 1       Min.   :0.0700
##  1st Qu.:1596   1st Qu.:0.0000   1st Qu.:13       1st Qu.:0.5000
##  Median :2095   Median :0.0000   Median :25       Median :1.0500
```

```
##  Mean   :2510   Mean   :0.4898   Mean   :25       Mean   :0.9829
##  3rd Qu.:3606   3rd Qu.:1.0000   3rd Qu.:37       3rd Qu.:1.4300
##  Max.   :4796   Max.   :1.0000   Max.   :49       Max.   :2.0000
##    marketcap    stock_return_scaled
##  Min.   : 185   Min.   : 30.38
##  1st Qu.: 872   1st Qu.:125.15
##  Median :1172   Median :161.30
##  Mean   :1468   Mean   :198.40
##  3rd Qu.:2221   3rd Qu.:230.95
##  Max.   :2997   Max.   :984.54
```

```
# fit the model
linear_model <- lm(stock_return_scaled ~ dividend, data = stock)

# check the summary of the model
summary(linear_model)
```

```
##
## Call:
## lm(formula = stock_return_scaled ~ dividend, data = stock)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -174.38  -71.47  -36.62   26.19  779.78
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   204.76      29.29   6.991 8.43e-09 ***
## dividend      -12.97      41.85  -0.310    0.758
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 146.4 on 47 degrees of freedom
## Multiple R-squared:  0.002041,   Adjusted R-squared:  -0.01919
## F-statistic: 0.09611 on 1 and 47 DF,  p-value: 0.7579
```

```
# get the model co-efficients
summary(linear_model)$coefficient
```

```
##              Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 204.75755   29.28741  6.9913162 8.425152e-09
## dividend    -12.97319   41.84787 -0.3100085 7.579250e-01
```

Insights

The p-value for the dividend predictor is about $0.757925 > 0.05$ (level of significance), therefore, we do not reject the null hypothesis. i.e we accept the null hypothesis that the co-efficient β of the dividend predictor is zero and not statistically significant

From the model summary and coefficients, we can see that the intercept is about 204.7576, while the slope of the model is -12.9732. Therefore, the equation of the model is $Y(StockReturnScaled) = 204.7576 - 12.9732 * X(dividend)$

It can be seen that the model shows a negative relationship between stock_return_scaled and dividend.

Also, the R-Squared is about 0.2% which indicates a poor model.

```r
# get the R-squared

summary(linear_model)$r.squared
```

```
## [1] 0.00204062
```

From the R-Squared value, we see that the R square is 0.00204062 (0.2%) which indicates a poor model.

```r
# Try a prediction
newdata <- data.frame(dividend = 1000) # wrap the parameter
predict(linear_model, newdata) # apply predict
```

```
##         1
## -12768.44
```

Variables to Include to help increase accuracy of the model
Sometimes a single predictor may not be enough to predict the target variable. To improve accuracy, more features or variables may have to be added. In this case, I will add the marketcap, and earnings_ranking. We can run a model that includes those extra two features to see if the model accuracy will improve.

```r
# fit the model with two additional variables
linear_model_more_variables <- lm(stock_return_scaled ~
                                    dividend + marketcap + earnings_ranking, data = stock)

# get the summary of the new model
summary(linear_model_more_variables)
```

```
##
## Call:
## lm(formula = stock_return_scaled ~ dividend + marketcap + earnings_ranking,
##     data = stock)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -96.62 -42.56 -20.08  22.07 569.79
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      730.00534   85.17649   8.571 5.17e-11 ***
## dividend          -0.78782   68.53077  -0.011    0.991
## marketcap         -0.19470    0.03233  -6.022 2.90e-07 ***
## earnings_ranking  -9.81975    2.04602  -4.799 1.79e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 100.8 on 45 degrees of freedom
## Multiple R-squared:  0.5473, Adjusted R-squared:  0.5171
## F-statistic: 18.13 on 3 and 45 DF,  p-value: 7.378e-08
```

From the model summary, we can see that the R-Squared for the model has increased drastically from 0.2% to about 54.73%. Also, from the p-values, `marketcap` and `earnings_ranking` are statistically significant predictors of the dependent variable (stock_return_scaled)