Regression Analysis Project 1

Group 10: Brennan Chan, Tim McKinley, Connor Kelly

Fit a multiple regression model using Earnings as the dependent variable, and Events and Score as the independent variable

First lets load in the packages and data.

```
# Core packages
suppressPackageStartupMessages(library(tidyverse))
suppressPackageStartupMessages(library(tidymodels))
golf = read_csv("Golf.csv", show_col_types = FALSE)

# Peek
summary(golf)
```

```
    Earnings           Events          Score      
 Min.   :  21250   Min.   :16.00   Min.   :68.98  
 1st Qu.: 402589   1st Qu.:23.00   1st Qu.:70.31  
 Median :1156517   Median :26.00   Median :70.92  
 Mean   :1726394   Mean   :25.76   Mean   :71.14  
 3rd Qu.:2359507   3rd Qu.:30.00   3rd Qu.:71.75  
 Max.   :5787225   Max.   :36.00   Max.   :75.01  
```

Now we can fit the multiple regression model.

```
golf_recipe = recipe(Earnings ~ Events + Score, data = golf)

lm_model = linear_reg() |>
  set_engine("lm")

lm_workflow = workflow() |>
  add_recipe(golf_recipe) |>
```

```
  add_model(lm_model)

lm_fit = lm_workflow |>
  fit(data = golf)

#coeffecients
lm_fit |> tidy()
```

```
# A tibble: 3 x 5
  term         estimate std.error statistic    p.value
  <chr>           <dbl>     <dbl>     <dbl>      <dbl>
1 (Intercept) 73365196. 11192809.      6.55 0.00000369
2 Events       -116049.    43500.     -2.67 0.0157
3 Score        -965038.   162452.     -5.94 0.0000127
```

```
#F-test p-value & r squared
lm_fit |> glance()
```

```
# A tibble: 1 x 12
  r.squared adj.r.squared   sigma statistic    p.value    df logLik   AIC   BIC
      <dbl>         <dbl>   <dbl>     <dbl>      <dbl> <dbl>  <dbl> <dbl> <dbl>
1     0.777         0.752 926971.      31.3 0.00000137     2  -317.  641.  646.
# i 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

```
# analysis of variance table
lm_fit |> extract_fit_engine() |> anova()
```

```
Analysis of Variance Table

Response: ..y
          Df     Sum Sq    Mean Sq F value    Pr(>F)
Events     1 2.3514e+13 2.3514e+13  27.365 5.645e-05 ***
Score      1 3.0323e+13 3.0323e+13  35.289 1.273e-05 ***
Residuals 18 1.5467e+13 8.5927e+11
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comments

Individual Slope Tests

```
#coeffecients
lm_fit |> tidy()
```

```
# A tibble: 3 x 5
  term         estimate std.error statistic    p.value
  <chr>           <dbl>     <dbl>     <dbl>      <dbl>
1 (Intercept) 73365196. 11192809.      6.55 0.00000369
2 Events       -116049.    43500.     -2.67 0.0157
3 Score        -965038.   162452.     -5.94 0.0000127
```

For each predictor, we test $(H_0 : \beta_j = 0)$ versus $(H_a : \beta_j \neq 0)$.

Predictor	Estimate	t-value	p-value	Decision
Events	$-116{,}049$	-2.668	0.0157	Significant
Score	$-965{,}039$	-5.940	1.27×10^{-5}	Significant

- For Events: $(t = -2.67,\ p = 0.0157 < 0.05)$. We reject (H_0) and conclude that Events has a significant linear relationship with Earnings when Score is held constant.
- For Score: $(t = -5.94,\ p < 0.001)$. We reject (H_0) and conclude that Score is also a significant linear predictor of Earnings when Events is held constant.

Thus, both predictors make significant contributions to explaining variation in Earnings.

F-Test

Hypotheses:

$$H_0 : \beta_1 = \beta_2 = 0 H_a : \text{At least one } \beta_j \neq 0$$

$$F(2, 18) = 31.33, \quad p = 1.374 \times 10^{-6}$$

Because the p-value is less than 0.05, we reject H_0. There is sufficient evidence that at least one of the slope coefficients differs from zero. Therefore, the regression model provides useful linear predictive information for Earnings.

Coefficient of Determination

The coefficient of determination is ($R^2 = 0.7768$). This indicates that approximately 77.68% of the variability in Earnings is explained by the linear relationship with Events and Score.

Give the equation of the model and the equation of the fitted model

The regression model is given by:

$$\widehat{\text{Earnings}} = 73{,}365{,}196 - 116{,}049(\text{Events}) - 965{,}039(\text{Score})$$

Where:

- β_0 = Intercept = 73,365,196
- β_1 = Coefficient for Events = -116,049
- β_2 = Coefficient for Score = -965,039

The fitted model is given by:

$$E(Earnings \mid Events, Score) = 73,365,196 - 116,049(Events) - 965,039(Score)$$

Which of the independent variables, Events or Score, would you choose if you want only one predictor?

We think score makes more sense to use because a higher score in golf just means you played worse. That should impact the earnings as a lower score means one is golfing better, and in turn probably earned more.

Fit a model with the independent variable chosen (Score) from the previous step

```
fit_score = lm(Earnings ~ Score, data = golf)
summary(fit_score)
```

```
Call:
lm(formula = Earnings ~ Score, data = golf)

Residuals:
     Min       1Q   Median       3Q      Max
-1089495  -657702  -507083   282169  2653186

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 81760015   12350070   6.620 2.47e-06 ***
Score       -1125076     173581  -6.482 3.28e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1066000 on 19 degrees of freedom
Multiple R-squared:  0.6886,    Adjusted R-squared:  0.6722
F-statistic: 42.01 on 1 and 19 DF,  p-value: 3.279e-06
```

Comments

F-test

Hypotheses:

$$H_0 : \beta_1 = 0 \quad \text{vs} \quad H_a : \beta_1 \neq 0.$$

From the summary output:

$$F(1, 19) = 42.01, \quad p = 3.279 \times 10^{-6}.$$

Because the p-value is less than 0.05, we reject (H_0). There is sufficient evidence that the model provides useful linear predictive information for Earnings.

t-test

For the slope of Score:

$$t = -6.482, \quad p = 3.28 \times 10^{-6}.$$

Since ($p < 0.05$), we reject (H_0) and conclude that Score is a significant linear predictor of Earnings.

Coefficient of Determination

$$R^2 = 0.6886$$

Thus, approximately 68.86% of the variation in Earnings is explained by its linear relationship with Score.

Give the equation of the model and the equation of the fitted model.

Model Form: $E(\text{Earnings} \mid \text{Score}) = \beta_0 + \beta_1(\text{Score})$

Fitted Model: $\widehat{\text{Earnings}} = 81,760,015 - 1,125,076(\text{Score})$

Where: - β_0 = Intercept = 81,760,015 - β_1 = Coefficient for Score = -1,125,076

Compare the model that uses both predictors and the model that just uses one predictor.

Model	Predictors	Adjusted (R^2)
Multiple Regression	Events and Score	0.752
Simple Regression	Score only	0.672

The multiple regression model with Events and Score explains a larger proportion of the variation in Earnings (adjusted ($R^2 = 0.752$)) than the simple regression model using only Score (adjusted ($R^2 = 0.672$)).

Because the increase in explained variability is meaningful and both predictors are significant in the multiple model, we prefer the multiple regression model. This model provides greater explanatory power while still maintaining a relatively simple structure with only two predictors.