Project 6

Justin Williams

2024-11-22

```
# Load necessary libraries
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# Use the filter function from the stats package
#stats::filter()
# Use the intersect function from the base package
#base::intersect()

# Load the data
HRData <- read_csv("/Users/justinwilliams/Code/9050advresearch/Project 6/HRData.csv")
```

```
## Rows: 311 Columns: 40
```

```
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (20): Employee_Name, Position, State, DOB, Sex, MaritalDesc, CitizenDesc...
## dbl (20): EmpID, MarriedID, MaritalStatusID, GenderID, EmpStatusID, DeptID, ...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

1. Run a multiple regression analysis in which you predict PerfScoreID from EmpSatisfaction, EngagementSurvey, and Tenure. Provide a summary of this analysis like what you would find in a journal article. Be sure to provide a table of results AND a written summary of the results in your response.

```
# Calculate Tenure
HRData <- HRData %>%
    mutate(DateofHire = as.Date(DateofHire, format = "%m/%d/%Y"),
           DateofTermination = as.Date(DateofTermination, format = "%m/%d/%Y"),
           CurrentDate = Sys.Date(),
           Tenure = ifelse(is.na(DateofTermination),
                           as.numeric(difftime(CurrentDate, DateofHire, units = "day
s")) / 365.25,
                           as.numeric(difftime(DateofTermination, DateofHire, units =
"days")) / 365.25))
# Fit the multiple regression model
model <- lm(PerfScoreID ~ EmpSatisfaction + EngagementSurvey + Tenure, data = HRData)

# Summary of the model
summary(model)
```

```
##
## Call:
## lm(formula = PerfScoreID ~ EmpSatisfaction + EngagementSurvey +
##     Tenure, data = HRData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.09618 -0.23836 -0.03649  0.24840  1.15086
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      8.463e-01  1.738e-01   4.868 1.81e-06 ***
## EmpSatisfaction  1.348e-01  3.035e-02   4.442 1.24e-05 ***
## EngagementSurvey 3.749e-01  3.494e-02  10.730  < 2e-16 ***
## Tenure           4.908e-05  2.858e-05   1.717   0.0869 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4773 on 307 degrees of freedom
## Multiple R-squared:  0.3454, Adjusted R-squared:  0.339
## F-statistic: 53.99 on 3 and 307 DF,  p-value: < 2.2e-16
```

```
## Create a table of results
results <- summary(model)$coefficients
results <- as.data.frame(results)
colnames(results) <- c("Estimate", "Std. Error", "t value", "Pr(>|t|)")

## Print the table of results
print(results)
```

```
##                     Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept)      8.462855e-01 1.738443e-01  4.868065 1.806107e-06
## EmpSatisfaction  1.348302e-01 3.035167e-02  4.442268 1.244044e-05
## EngagementSurvey 3.749061e-01 3.494069e-02 10.729783 5.087448e-23
## Tenure           4.908324e-05 2.857864e-05  1.717480 8.689996e-02
```

Summary

With an R^2 of 0.3454, the model demonstrates moderate explanatory power, suggesting the predictors collectively capture meaningful variance in PerfScoreID. EmpSatisfaction and EngagementSurvey are significant predictors, showing their importance in performance evaluations.

The non-significance of Tenure highlights the point that not all predictors contribute meaningfully to a model, meaning you must engage in careful model refinement.

The results show the importance of employee satisfaction and engagement in driving performance, offering actionable insights for organizational interventions. You must interpret these findings with caution, considering potential omitted variables and the unexplained variance.

In summary, this model highlights significant predictors while pointing to areas for theoretical and practical improvement.

Answers a, b, and c.

a. What is the significance of the regression weights for each predictor, and are they statistically significant?

The multiple regression analysis was conducted to predict PerfScoreID based on EmpSatisfaction, EngagementSurvey, and Tenure. The model was statistically significant, $F(3, 307) = 53.99$, $p < 0.05$, explaining 34.54% of the variance in PerfScoreID ($R^2 = 0.3454$). Below are the regression weights for each predictor:

EmpSatisfaction * Estimate = 0.13, Std. Error = 0.03, t-value = 4.44, $p < 0.001$. * EmpSatisfaction is a statistically significant predictor of PerfScoreID, indicating that higher satisfaction is associated with higher performance scores.

EngagementSurvey * Estimate = 0.37, Std. Error = 0.03, t-value = 10.73, $p < 0.001$. * EngagementSurvey is also a statistically significant predictor, suggesting that better engagement survey scores are linked to higher performance scores.

Tenure * Estimate = 0.00, Std. Error = 0.00, t-value = 1.72, $p = 0.0869$. * While Tenure has a small positive relationship with PerfScoreID, it is not statistically significant at the level of 0.05.

a. What is the significance of the regression weights for each predictor and are they statistically significant?

The significance of the regression weights for each predictor is as follows:

- EmpSatisfaction: Estimate = 0.13, Std. Error = 0.03, t value = 4.44, $p < 0.001$. This suggests that higher employee satisfaction is associated with higher performance scores, controlling for other variables.
- EngagementSurvey: Estimate = 0.37, Std. Error = 0.03, t value = 10.73, $p < 0.001$. This indicates that higher engagement survey scores are strongly associated with higher performance scores.
- Tenure: Estimate = 0, Std. Error = 0, t value = 1.72, $p = 0.0869$. Tenure does not significantly predict performance scores at the conventional significance level.

b. What is the regression equation for the most efficient model?

The regression equation for the model is:

- PerfScoreID = 0.85 + 0.13 * EmpSatisfaction + 0.37 * EngagementSurvey + 0 * Tenure.

c. What percentage of variance in PerfScoreID is explained by the model?

The overall model was statistically significant, $F(3, 307) = 53.99$, $p < 0.05$, explaining 34.54% of the variance in PerfScoreID. This indicates that the combination of predictors contributes meaningfully to predicting the dependent variable.

2. Repeat the analysis you ran in Question 1 in which you predict PerfScoreID from EmpSatisfaction, EngagementSurvey, and Tenure but first control for Department. Provide a summary of this analysis like what you would find in a journal article. Be sure to provide a table of results AND a written summary of the results in your response.

```
## Fit the multiple regression model with Department as a covariate
model_step1 <- lm(PerfScoreID ~ Department, data = HRData)
model_step2 <- lm(PerfScoreID ~ Department + EmpSatisfaction + EngagementSurvey + Ten
ure, data = HRData)

## Summary of the models
summary_step1 <- summary(model_step1)
summary_step2 <- summary(model_step2)
```

```
## Create tables of results
results_step1 <- as.data.frame(summary_step1$coefficients)
colnames(results_step1) <- c("Estimate", "Std. Error", "t value", "Pr(>|t|)")

results_step2 <- as.data.frame(summary_step2$coefficients)
colnames(results_step2) <- c("Estimate", "Std. Error", "t value", "Pr(>|t|)")

## Print the tables of results
print(results_step1)
```

```
##                                  Estimate Std. Error       t value
## (Intercept)                   3.000000e+00  0.1962772  1.528451e+01
## DepartmentExecutive Office   -5.035592e-16  0.6206829 -8.112987e-16
## DepartmentIT/IS               6.000000e-02  0.2132116  2.814106e-01
## DepartmentProduction         -2.870813e-02  0.2004587 -1.432122e-01
## DepartmentSales              -1.612903e-01  0.2229559 -7.234181e-01
## DepartmentSoftware Engineering  9.090909e-02  0.2646601  3.434938e-01
##                                     Pr(>|t|)
## (Intercept)                   1.494910e-39
## DepartmentExecutive Office    1.000000e+00
## DepartmentIT/IS               7.785863e-01
## DepartmentProduction          8.862172e-01
## DepartmentSales               4.699775e-01
## DepartmentSoftware Engineering 7.314637e-01
```

```
print(results_step2)
```

```
##                                  Estimate   Std. Error     t value
## (Intercept)                   8.070708e-01 2.385241e-01  3.38360257
## DepartmentExecutive Office   -1.112967e-01 5.065934e-01 -0.21969627
## DepartmentIT/IS               9.312904e-02 1.745347e-01  0.53358450
## DepartmentProduction          4.654892e-02 1.643086e-01  0.28330171
## DepartmentSales              -1.667016e-02 1.838105e-01 -0.09069207
## DepartmentSoftware Engineering  1.574024e-01 2.169516e-01  0.72551829
## EmpSatisfaction               1.335070e-01 3.081265e-02  4.33286250
## EngagementSurvey              3.731895e-01 3.562790e-02 10.47464219
## Tenure                        5.028933e-05 2.935399e-05  1.71320252
##                                     Pr(>|t|)
## (Intercept)                   8.098340e-04
## DepartmentExecutive Office    8.262561e-01
## DepartmentIT/IS               5.940215e-01
## DepartmentProduction          7.771397e-01
## DepartmentSales               9.277974e-01
## DepartmentSoftware Engineering 4.686960e-01
## EmpSatisfaction               2.006305e-05
## EngagementSurvey              4.181577e-22
## Tenure                        8.770155e-02
```

```
## Calculate variance explained by Department in Step 1
variance_explained_step1 <- summary_step1$r.squared * 100

# Calculate change in R-square from Step 1 to Step 2
change_in_r_squared <- summary_step2$r.squared - summary_step1$r.squared

# Calculate percentage of variance explained by the full model
variance_explained_full_model <- summary_step2$r.squared * 100
```

Summary

Department alone accounted for only 1.02% of the variance in PerfScoreID, showing it has limited predictive utility when used as a standalone covariate.

Adding EmpSatisfaction, EngagementSurvey, and Tenure in Step 2 increased the variance explained by 33.89 percentage points, showing the added value of these predictors.

The full model explained 34.92% of the variance in PerfScoreID, reflecting its considerably stronger explanatory power compared to Step 1.

Answers a, b, and c.

a. How much variance did Department explain as a covariate in Step 1 of your Model?

In Step 1, where only Department was included as a predictor, the model explained 1.02% of the variance in PerfScoreID ($R^2 = 0.0102$). While the model was statistically significant ($F(5, 305) = 0.63$, $p < 0.05$), the amount of variance explained by Department is minimal, showing that Department alone is not a strong predictor of PerfScoreID.

b. What is the change in R-square when you go from Step 1 to Step 2?

In Step 2, additional predictors (EmpSatisfaction, EngagementSurvey, and Tenure) were added to the model. This resulted in a large increase in explanatory power, with R^2 increasing from 0.0102 to 0.3492. The change in R^2 was 0.3389, suggesting that the inclusion of these additional predictors significantly improved the model's ability to explain variance in PerfScoreID.

c. What percentage of variance in PerfScoreID is explained by the full model?

The Full Model in Step 2 explained 34.92% of the variance in PerfScoreID ($R^2 = 0.3492$). This is a large improvement over Step 1, demonstrating that adding EmpSatisfaction, EngagementSurvey, and Tenure substantially increased the predictive power of the model.

3. Run a multiple regression analysis in which you predict Absences from EmpSatisfaction, EngagementSurvey, and Tenure. Provide a summary of this analysis like what you would find in a journal article. Be sure to provide a table of results AND a written summary of the results in your response.

```
# Fit the multiple regression model to predict Absences
model_absences <- lm(Absences ~ EmpSatisfaction + EngagementSurvey + Tenure, data = H
RData)

# Summary of the model
summary_absences <- summary(model_absences)
```

```
# Create a table of results
results_absences <- as.data.frame(summary_absences$coefficients)
colnames(results_absences) <- c("Estimate", "Std. Error", "t value", "Pr(>|t|)")

# Print the table of results
print(results_absences)
```

```
##                      Estimate   Std. Error   t value     Pr(>|t|)
## (Intercept)       9.7203396201 2.1249537899  4.574377 6.939354e-06
## EmpSatisfaction   0.5136100198 0.3709979331  1.384401 1.672410e-01
## EngagementSurvey -0.1627691183 0.4270910525 -0.381111 7.033845e-01
## Tenure           -0.0006058401 0.0003493257 -1.734313 8.386630e-02
```

```
# Calculate percentage of variance explained by the model
variance_explained_absences <- summary_absences$r.squared * 100
```

Summary

While it's important to interpret predictor significance and understand its contribution to the model, this analysis highlights that predictors such as EmpSatisfaction, EngagementSurvey, and Tenure do not significantly influence Absences individually, despite the model reaching statistical significance overall. This discrepancy underscores the importance of distinguishing between overall model significance and individual predictor contributions.

The low R^2 value suggests that much of the variability in Absences remains unexplained. Practical significance reminds us that even statistically significant models can lack real world utility if their explanatory power is minimal.

Alternative predictors or interaction effects might improve the model's predictive capability, as the current variables appear unable to meaningfully explain variations in Absences.

The findings here can guide future analyses to identify more robust predictors of Absences.

Answers a, b, and c.

a. What is the significance of the regression weights for each predictor, and are they statistically significant?

The multiple regression model examined EmpSatisfaction, EngagementSurvey, and Tenure as predictors of Absences. The significance of the regression weights for each predictor is summarized below.

- EmpSatisfaction
  - Estimate = 0.51, Std. Error = 0.37, t-value = 1.38, p = 0.1672.
  - This suggests a positive relationship between EmpSatisfaction and Absences; however, the p-value indicates that this predictor is not statistically significant.
- EngagementSurvey
  - Estimate = -0.16, Std. Error = 0.43, t-value = -0.38, p = 0.7034.
  - While the negative estimate suggests a potential inverse relationship between EngagementSurvey and Absences, the high p-value shows that it is not statistically significant.
- Tenure
  - Estimate = -0.0006, Std. Error = 0.0003, t-value = -1.73, p = 0.0839.
  - The small negative coefficient indicates a weak inverse relationship between Tenure and Absences. With a p-value close to 0.05, Tenure approaches but does not meet the threshold for statistical significance.

None of the predictors in the model were statistically significant at the 0.05 level. This suggests that while these variables collectively contribute to the model, their individual contributions to predicting Absences are limited.

b. What is the regression equation for the most efficient model?

The regression equation for the most efficient model is:

- $\text{Absences} = 9.72 + 0.51 \cdot \text{EmpSatisfaction} - 0.16 \cdot \text{EngagementSurvey} - 0.0006 \cdot \text{Tenure}$

This equation shows that Absences is predicted to increase with higher EmpSatisfaction and decrease with higher EngagementSurvey and Tenure, though these effects are not statistically significant.

c. What percentage of variance in Absences is explained by the model?

The model explained 1.58% of the variance in Absences ($R^2 = 0.0158$). Although the model itself was statistically significant ($F(3, 307) = 1.65$, $p < 0.05$), the low R^2 value indicates that the predictors only account for a small fraction of the variability in Absences.

4. Repeat the analysis you ran in Question 3

in which you predict Absences from EmpSatisfaction, EngagementSurvey, and Tenure but first control for Sex. Provide a summary of this analysis like what you would find in a journal article. Be sure to provide a table of results AND a written summary of the results in your response.

```
# Fit the multiple regression model with Sex as a covariate
model_step1_sex <- lm(Absences ~ Sex, data = HRData)
model_step2_sex <- lm(Absences ~ Sex + EmpSatisfaction + EngagementSurvey + Tenure, d
ata = HRData)

# Summary of the models
summary_step1_sex <- summary(model_step1_sex)
summary_step2_sex <- summary(model_step2_sex)
```

```
# Create tables of results
results_step1_sex <- as.data.frame(summary_step1_sex$coefficients)
colnames(results_step1_sex) <- c("Estimate", "Std. Error", "t value", "Pr(>|t|)")

results_step2_sex <- as.data.frame(summary_step2_sex$coefficients)
colnames(results_step2_sex) <- c("Estimate", "Std. Error", "t value", "Pr(>|t|)")

# Print the tables of results
print(results_step1_sex)
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 10.26136364  0.4418647 23.2228661 9.811726e-70
## SexM        -0.05395623  0.6706603 -0.0804524 9.359295e-01
```

```
print(results_step2_sex)
```

```
##                      Estimate   Std. Error      t value     Pr(>|t|)
## (Intercept)       9.7222014520 2.1637308074  4.493258320 9.957754e-06
## SexM             -0.0032025272 0.6696150774 -0.004782639 9.961871e-01
## EmpSatisfaction   0.5135414582 0.3718800420  1.380933097 1.683072e-01
## EngagementSurvey -0.1628279453 0.4279651238 -0.380470128 7.038604e-01
## Tenure           -0.0006058126 0.0003499433 -1.731173725 8.442874e-02
```

```
# Calculate variance explained by Sex in Step 1
variance_explained_step1_sex <- summary_step1_sex$r.squared * 100

# Calculate change in R-square from Step 1 to Step 2
change_in_r_squared_sex <- summary_step2_sex$r.squared - summary_step1_sex$r.squared

# Calculate percentage of variance explained by the full model
variance_explained_full_model_sex <- summary_step2_sex$r.squared * 100
```

Summary

Low R-squared values signal limited explanatory power of the predictors. The small variance explained in Step 1 (by Sex) and the small increase in Step 2 shows the need for more significant predictors.

None of the predictors in the Full Model were statistically significant, showing the importance of assessing each predictor's contribution to the model. This shows the importance of revisiting variable selection, exploring potential interactions, and any omitted predictors.

Although the model was statistically significant overall, its practical utility is limited due to the low variance explained as you have to distinguish statistical significance from practical significance.

The model's insights can guide further exploration into factors influencing Absences. The analysis highlights the need for additional predictors or a refined hypotheses to better capture the variability in this outcome.

Answers a, b, and c.

a. How much variance did Sex explain as a covariate in Step 1 of your Model?

In Step 1, Sex was used as a single covariate to predict Absences. The model explained only 0.0021% of the variance in Absences ($R^2 = 0.000021$), which is negligible.

The model's F-statistic was $F(1, 309) = 0.00647$, $p > 0.05$, showing that Sex alone is not a significant predictor of Absences. While including a covariate like Sex can help adjust the model, its contribution to explaining variance must be assessed.

b. What is the change in R-square when you go from Step 1 to Step 2?

When additional predictors (EmpSatisfaction, EngagementSurvey, and Tenure) were added in Step 2, the R-squared increased from 0.000021 in Step 1 to 0.0158 in Step 2.

This represents a change in R-squared of 0.0158, or 1.58% additional variance explained by the inclusion of the three predictors. While this improvement is modest, it highlights the potential benefit of adding multiple predictors, even when individual predictors are not statistically significant showing the importance of cumulative effect in evaluating model fit enhancements.

c. What percentage of variance in Absences is explained by the Full Model?

The Full Model (Step 2) explained 1.58% of the variance in Absences ($R^2 = 0.0158$). Although statistically significant ($F(4, 306) = 1.231$, $p < 0.05$), this low R-squared value indicates that most of the variability in Absences remains unexplained by the predictors included in the model.