International and Domestic Salary:

Approach and Internal Report

Michael Alex Force

Methodology:

When approaching prospective clientele and potential work, information is gathered surrounding the nature of the experimentation being conducted, the primary objectives, intricacies in data collection, and the relevancy of information provided. When consulting with a new client, building a strong rapport is critical in ensuring mutual understanding with what is being asked for and delivered. Through this, elaboration is required surrounding the client's primary goals of the work: what is the main goal of conducting said experiment, by what manner was the data collected, what is the hypothesis being investigated, and how is this study beneficial to you as the client; through these questions a greater understanding of the client's needs are being met.

From the "Salary" data set, the primary objective targets determining influential factors towards dictating annual earnings, the results of which serve to highlight potential controllable variables one could address in pursuing economic growth. Accordingly, the analysis focuses on decerning between numerous variables that are either within one's ability to control and other, intrinsic characteristics, that are outside of alteration but could potentially influence earning potential. The intricacies of the data set consist of 6884 rows and 9 columns corresponding to individuals sampled globally and their respective salary, age, gender, education level, job title, years of experience, country, race, and whether they hold a senior position or not; the methodology surrounding the data collection process including survey practices and time frame to which data was collected was not provided(Aboutalebi, A., 2024).

To properly address relevant questions posed by the client, various statistical techniques were utilized. To develop an understanding of the relationship between salary and the other presented variables, varying techniques could offer statistical insight including simple linear regression, multiple linear regression, analysis of variance (ANOVA), and two-sample t-Tests. There are assumptions accompanying said testing practices, all of which require preliminary work assessing the presence of either normality in data or normality in residuals (in the case of regression models) along with homogeneity of variance within groups or homoscedasticity for regression models. To narrow down theorized testing practices, initial analysis was conducted on the data set to understand data types allowing for an accurate assessment of which variables had the potential to be implemented in specific models. Before testing could be conducted, even to the extent of assessing prior assumptions before commencing said testing, the data set required slight modification ensuring the variable types like gender, country, senior position, race, job position and education were all treated as categorical data; denoted as factors within the R Studios environment. Additional standardizing was implemented towards shifting education level from a numbered system (I.E. 0,1,2,3) to text explicitly outlining education status (I.E. "High School

International and Domestic Salary:

Approach and Internal Report

Michael Alex Force

Diploma", "Bachelor's Degree", "Master's Degree", "PhD"). For the remaining data types, salary (USD), age, and years of service were already presented as metric data which could then be utilized in various testing practices. Once the data was in the correct data types, preliminary work could then be conducted assessing prior assumptions that could make education decisions when dictating higher level statistical testing aiding in addressing the pertinent questions of the client.

Internal Report:

Initial exploratory data analysis of the "salary" data set concluded abnormalities in the distribution of individuals obtaining higher education when compared to the statistical values of the U.S. via the U.S. Bureau of Labor Statistics with the United States having 37.9% of the population with a bachelor's degree or higher compared to the 93% within the "salary" data set (Bureau, U. C., 2022). Accordingly, key insights and findings may be misrepresentative to those other than individuals with a college education, or at the very least, might not scale properly on a tangible "dollar figure" basis and serve primarily in determining what is deemed as a legitimate contributing factor exclusively. Additional exploratory data analysis concluded that, based on average salary exclusively, corporate or managerial positions held the largest annual earnings.

In depth testing to assess assumptions surrounding the various testing methodologies concluded various concerns. Based on developing a multiple linear regression model with salary being dependent on age and years of experience, a high variance inflation factor (VIF) concluded the presence of multicollinearity within said independent variables which were then visualized via a scatter plot and supported through a Spearman rank correlation that was calculated after said data failed assumptions of normality. To approach these findings, a 70:30 split was applied with respect to a subset with variables salary, age, and years of experience. A ridge regression model was fitted to the training data set using a k-fold cross validation approach utilizing 10 folds which was then optimized by finding the lambda value that minimized the mean squared error. The R-squared value was calculated for said model at which point the testing data set was utilized with its respective R-squared value calculated to compare overall model adequacy and consistency.

For the non-dichotomous variables, relating to country, race, and education level, prior theorized ANOVA testing required assessing assumptions of linearity in residuals along with homoscedasticity and independence in sampling. There is strong theoretical evidence to support a lack of influence between variables, apart from sampled individuals having some genetic relation between one and other; homogeneity of variance was assessed through box plots to easily visualize variation. With respect to country and race, the box plots plausibly deduced ANOVA as a potential testing method; a finding that was supported via plotting residuals to fitted values. An assessment of normality from the residuals contradicted the testing assumptions relying on a non-parametric

alternative through Kruskal-Wallis concluding no statistical difference in central tendency for groups in either country or race. The preliminary work associated with education level concluding unequal variance leading to Kruskal-Wallis testing being applied from the starting yielding data to support a strong statistical difference in central tendency from education levels concluding different education levels are associated with different annual salaries.

The dichotomous variables senior and gender had initial propositions of two-sample t-Tests for their respective groups; the assumptions of which required normality in data and an assessment for equal or unequal variance. Both variables failed the assumption of normality via a visual assessment utilizing histograms, a non-parametric alternative was applied using Mann Whitney U test concluding statistically significant differences between male and female along with senior and non-senior positions.

The primary challenge encountered was the non-parametric alternatives that were consistently required throughout the analysis. Under the Central Limit Theorem (CLT), the distribution of the means of samples would approach normality, in the case of the presented data though a small sample size of individuals rather than a collection of means, particularly with respect to the overall population of any given country, resulting in non-normal data. While non-parametric testing yielded conclusive results, superior sample practices including additional data points would more aptly represent a population; bootstrapping or alternative sampling practices to effectively implement to theorized results of CLT would allow for conventional parametric testing which was initially proposed.

Author's Note:

The synopsis presented above explored the broad overhead view discussing the initial approach taken when understanding the intricacies of the data set, interpreting the primary objectives, and developing theorized testing towards solving issues. From the proposed investigation and analysis techniques, these were refined to specifically target the goals of the client; this was reflected in answering the underlying questions of the study including analyzing the influence each variable had on salary. With respect to the consultation report and video presentation, the primary components of the study were conveyed including the purpose of the study, relevant information regarding the data set, preliminary work, proposed testing methodologies including relevant issues, corresponding results, key insights in said findings and potential for future work. To effectively convey meaningful findings from the study, tables and charts were utilized to emphasize primary findings while also aiding in supporting the necessity for specific testing practices. Examples of which include boxplots and histograms to visualize trends in distributions and assumptions of variance necessary for specific statistical tests. To

properly convey significant findings in testing, and any additional assumptions accompanying said testing, brief elaboration was incorporated to illustrate proper practices, the necessity for said methodologies, and the meaningfulness in the results. Irrelevant complex statistical points and intricacies were left out so that only relevant components were conveyed to the client.

Appendix:

```
#------------------------------------------------------------------------
#Required packages
library(glmnet)
library(ggplot2)
library(nortest)
#------------------------------------------------------------------------
#General description of data frame
head(df)
dim(df)
str(df)
#------------------------------------------------------------------------
#Converting categorical variables from characters to factors
df$Gender <- as.factor(df$Gender)
df$Education.Level <- as.factor(df$Education.Level)
df$Job.Title <- as.factor(df$Job.Title)
df$Country <- as.factor(df$Country)
df$Race <- as.factor(df$Race)
df$Senior <- as.factor(df$Senior)
#------------------------------------------------------------------------
#Multiple linear regression model
ylm <- lm(Salary ~ Age + Years.of.Experience, df)
#variance influence factor
vif <- car::vif(ylm)
```

International and Domestic Salary:

Approach and Internal Report

Michael Alex Force

```
vif
#Anderson-Darling test
ad.test(df$Age)
ad.test(df$Years.of.Experience)
#Spearman rank
cor(df$Age, df$Years.of.Experience, method = "spearman")
#scatterplot for relationship between age and years of experience
df |>
  ggplot(aes(x = Age, y = Years.of.Experience)) +
  geom_point(colour = 'blue') +
  ggtitle("Age to Years of Experience: Salary Data Set")
#---------------------------------------------------------------------
#Splitting data set
df2 <- as.data.frame(cbind(df$Salary, df$Age, df$Years.of.Experience))
colnames(df2) <- c("Salary", "Age", "Years.of.Experience")

sample <- sample(c(TRUE, FALSE), nrow(df2), replace=TRUE, prob=c(0.7,0.3))
train  <- df2[sample, ]
test   <- df2[!sample, ]

y <- train$Salary
x <- data.matrix(train[,c('Age', 'Years.of.Experience')])
x
y

#Ridge regression model development
cv_model <- cv.glmnet(x,y, alpha = 0)
```

International and Domestic Salary:

Approach and Internal Report

Michael Alex Force

```
best_lambda <- cv_model$lambda.min
best_lambda
plot(cv_model)
#applying optimal lambda
best_model <- glmnet(x, y, alpha = 0, lambda = best_lambda)
coef(best_model)

#Using improved model to predict data
y_predicted <- predict(best_model, s = best_lambda, newx = x)

#find SST and SSE
sst <- sum((y - mean(y))^2)
sse <- sum((y_predicted - y)^2)

#find R-Squared
rsq <- 1 - sse/sst
rsq

#Applying test data set to trained model

y2 <- test$Salary
x2 <- data.matrix(test[,c('Age', 'Years.of.Experience')])

y_predicted <- predict(best_model, s = best_lambda, newx = x2)

sst <- sum((y2 - mean(y2))^2)
sse <- sum((y_predicted - y2)^2)
```

International and Domestic Salary:

Approach and Internal Report

Michael Alex Force

```
rsq <- 1 - sse/sst
rsq


model <- glmnet(x, y, alpha = 0)
summary(model)
plot(model, xvar = "lambda")


#------------------------------------------------------------------------
#Boxplot to visualize education and race
df |>
  ggplot(aes(x = Education.Level, y = Salary, fill = Education.Level)) +
  geom_boxplot()


df |>
  ggplot(aes(x = Race, y = Salary, fill = Race)) +
  geom_boxplot()
#ANOVA model
ylm_Race <- lm(Salary ~ Race, df)
summary(ylm_Race)
#Q-Q plot for race and Anderson-Darling test
qqnorm(ylm_Race$residuals)
qqline(ylm_Race$residuals)
ad.test(ylm_Race$residuals)
plot(ylm_Race, which = 1)
#Boxplot for country
df |>
```

International and Domestic Salary:

Approach and Internal Report

Michael Alex Force

```
  ggplot(aes(x = Country, y = Salary, fill = Country)) +
  geom_boxplot()
#ANOVA model, Q-Q plot, and Anderson-Darling
ylm_Country <- lm(Salary ~ Country, df)
summary(ylm_Country)
qqnorm(ylm_Country$residuals)
qqline(ylm_Country$residuals)
ad.test(ylm_Country$residuals)
plot(ylm_Country, which = 1)
#-----------------------------------------------------------------------
#Kruskal-Wallis tests for education, race, and country

kruskal.test(Salary ~ Education.Level, df)
kruskal.test(Salary ~ Race, df)
kruskal.test(Salary ~ Country, df)

#Histogram for gender followed by Mann Whitney U test and two sample t test
df |>
  ggplot() +
  geom_histogram(aes(x = Salary, fill = Gender)) +
  facet_wrap(~Gender)

M <- df |> filter(Gender == "Male")
F <- df |> filter(Gender == "Female")
wilcox.test(M$Salary, F$Salary, paired = FALSE)
```

International and Domestic Salary:

Approach and Internal Report

Michael Alex Force

#Histogram for senior variable followed by Mann Whitney U test and t test

```
df |>
  ggplot() +
  geom_histogram(aes(x = Salary, fill = Senior)) +
  facet_wrap(~Senior)

Syes <- df |> filter(Senior == 1)
Sno <- df |> filter(Senior == 0)
wilcox.test(Syes$Salary, Sno$Salary, paired = FALSE)
#------------------------------------------------------------------------
```

International and Domestic Salary:

Approach and Internal Report

Michael Alex Force

References

Aboutalebi, A. (2024, February 18). *Salary by job title and country*. Kaggle. https://www.kaggle.com/datasets/amirmahdiabbootalebi/salary-by-job-title-and-country

Bureau, U. C. (2022, February 24). *Census Bureau releases New Educational Attainment Data*. Census.gov. https://www.census.gov/newsroom/press-releases/2022/educational-attainment.html