Homework-5-Q1

BIOS507 Spring 2025 | Dr Lukemire | Elizabeth Nemeti Due: March 24 2025

Problem 1.

ABCD Corp is a mid-sized logistics company that prides itself on maintaining high employee satisfaction. Unfortunately, recent surveys indicate that employee job satisfaction levels vary widely depending on workload and job type. To better understand these dynamics the company has conducted a study to examine how weekly work hours influence job satisfaction for each type of employee in the company. The company has two broad categories of employees:

- **Office Workers**: Includes software developers, accountants, marketing specialists, and project managers. These roles are cognitively demanding, requiring problem-solving, meeting deadlines, and long hours of computer-based work.
- **Manual Laborers**: Includes warehouse workers, machine operators, and delivery personnel. These employees perform physically demanding tasks such as lifting, operating machinery, and moving goods across warehouses or delivery routes.

The company has collected data on number of hours worked, type of employee, and job satisfaction (0 to 100). The data set is ABCD_job_satisfaction.csv. Note that this is simulated data for this example homework problem and does not represent a real study.

- **Y (response)**: Job_Satisfaction (satisfaction score, categorical)
- **X1 (identifier)**: Employee_ID (just a unique ID, not a predictor)
- **X2 (predictor)**: Work_Hours (hours worked, continuous)
- **X3 (predictor)**: Job_Role (either Manual Laborer or Office Worker, categorical)

```
data_path = "/Users/elizabethnemeti/Documents/GitHub/BIOS507-Coursework/Homeworks/Homework-5/"
data_file <- file.path(data_path, "ABCD_job_satisfaction.csv")
ABCD_data <- read.csv(data_file, header = TRUE)

head(ABCD_data)
```

```
##   Employee_ID Work_Hours       Job_Role Job_Satisfaction
## 1           1   50.24097 Manual Laborer         39.26338
## 2           2   22.39232 Manual Laborer         57.05304
## 3           3   44.80320 Manual Laborer         46.67247
## 4           4   46.83140  Office Worker         18.64590
## 5           5   41.38649  Office Worker         24.31448
## 6           6   40.15516  Office Worker         22.66592
```

```
str(ABCD_data)
```

```
## 'data.frame':    500 obs. of  4 variables:
##  $ Employee_ID     : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Work_Hours      : num  50.2 22.4 44.8 46.8 41.4 ...
##  $ Job_Role        : chr  "Manual Laborer" "Manual Laborer" "Manual Laborer" "Office Worker" ...
##  $ Job_Satisfaction: num  39.3 57.1 46.7 18.6 24.3 ...
```

1. Fit a model that captures the relationship between job type and hours worked and the outcome of job satisfaction. Be sure to carry out all the usual steps (EDA, writing the model, etc). Interpret the coefficients, or, if easier, create a plot demonstrating the effects.

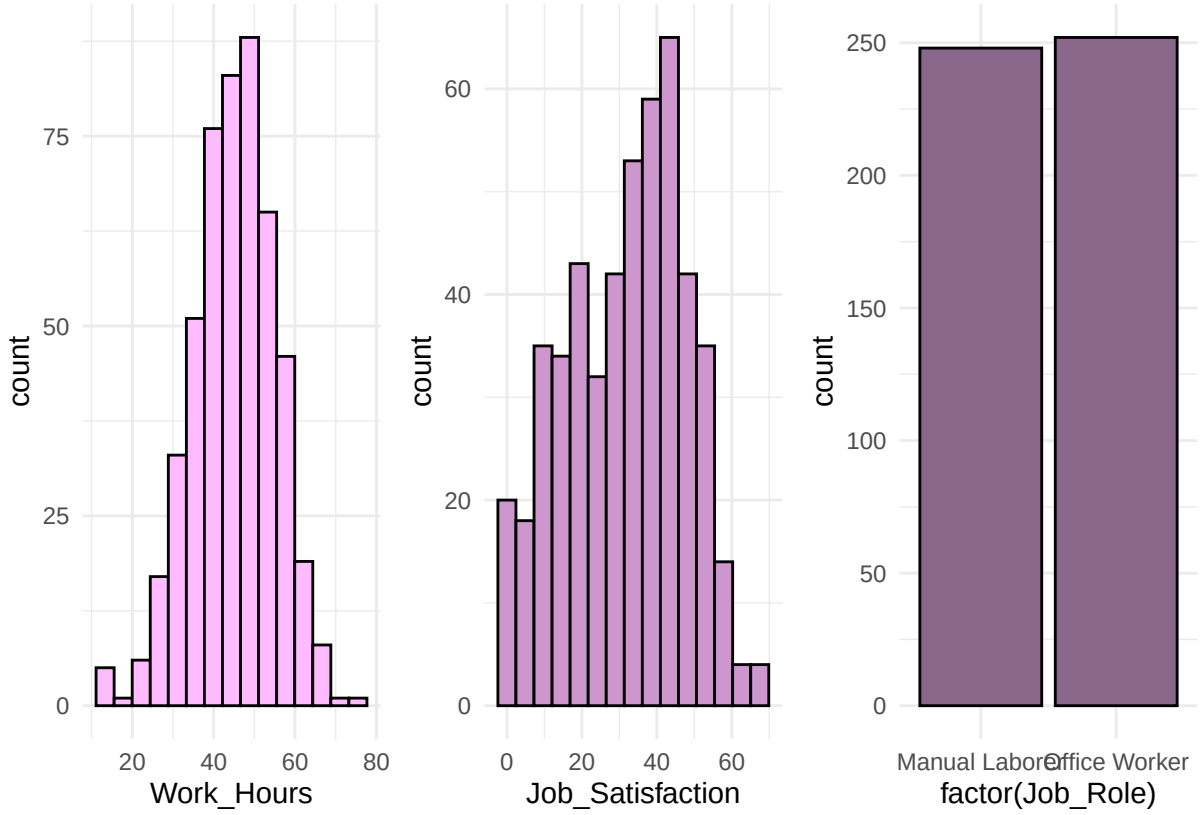
First, we will plot the variables of interest to check for anomalies. We have 1 continuous predictor variable (Work_Hours), and a continuous response variable (Job_Satisfaction) for which we will use a histogram, and we have 1 categorical variable (Job_Role) for which we would rather use a bar chart.

```
workhours_histogram <- ABCD_data %>%
  ggplot(aes(x = Work_Hours)) +
  geom_histogram(bins = 15, color = "black", fill = "plum1") +
  theme_minimal()

jobsatisfaction_histogram <- ABCD_data %>%
  ggplot(aes(x = Job_Satisfaction)) +
  geom_histogram(bins = 15, color = "black", fill = "plum3") +
  theme_minimal()

jobrole_barchart <- ABCD_data %>%
  ggplot(aes(x = factor(Job_Role))) +
  geom_bar(color = "black", fill = "plum4") +
  theme_minimal()

workhours_histogram + jobsatisfaction_histogram + jobrole_barchart
```




```
summary(ABCD_data$Work_Hours)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.38   38.52   45.29   44.99   51.84   75.52
```

```
summary(ABCD_data$Job_Satisfaction)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   19.84   34.40   31.72   44.20   67.45
```

At first glance, the data appears free of anomalies. The distribution of job roles is nearly balanced, with a similar number of Manual Laborers and Office Workers. This balanced representation allows us to make insights into both job roles without significant bias due to overrepresentation of one group. The recorded work hours range from 13 hours/week to 75 hours/week, with an average of ~45 hours/week. This average appears standard for the U.S., where a 40-hour work week is quite often exceeded. The data set does not disclose whether these higher work hours include overtime compensation. This could be a factor that while not included, significantly contributes to lower job satisfaction. The data also does not specify whether the <20 hours/week reports, are from part-time workers - which could skew perceptions of job satisfaction compared to full time workers - or admittance to "slacking off"/having no work to do. If we knew the conditions of the population reporting <20 hours/week, this could provide clearer insight on this key variable. Another factor worth considering is the geographical/cultural context of the company, which the data has not informed us of. For example, countries that emphasize a work-life balance, such as Spain or Italy, where their collectivist culture and structured break times for family and community are standard, may have much poorer job satisfaction scores with work weeks leading all the way up to 75 hours, compared to individualist cultures e.g. U.S., Japan. The job satisfaction scores range from 0-67, with an average of ~34, and most scores falling below 40, which is concerningly low for a company that prides itself on high employee satisfaction. This could suggest previous assessments were inflated for appearance, assessments were conducted too long ago to reflect the current population well, or the current survey variables do not account for critical variables contributing to job satisfaction, such as salary, employer benefits, insurance, or childcare coverage. Finally, with reports even at 0 for satisfaction, this indicates a need for a more comprehensive survey to capture employee satisfaction more meaningfully.

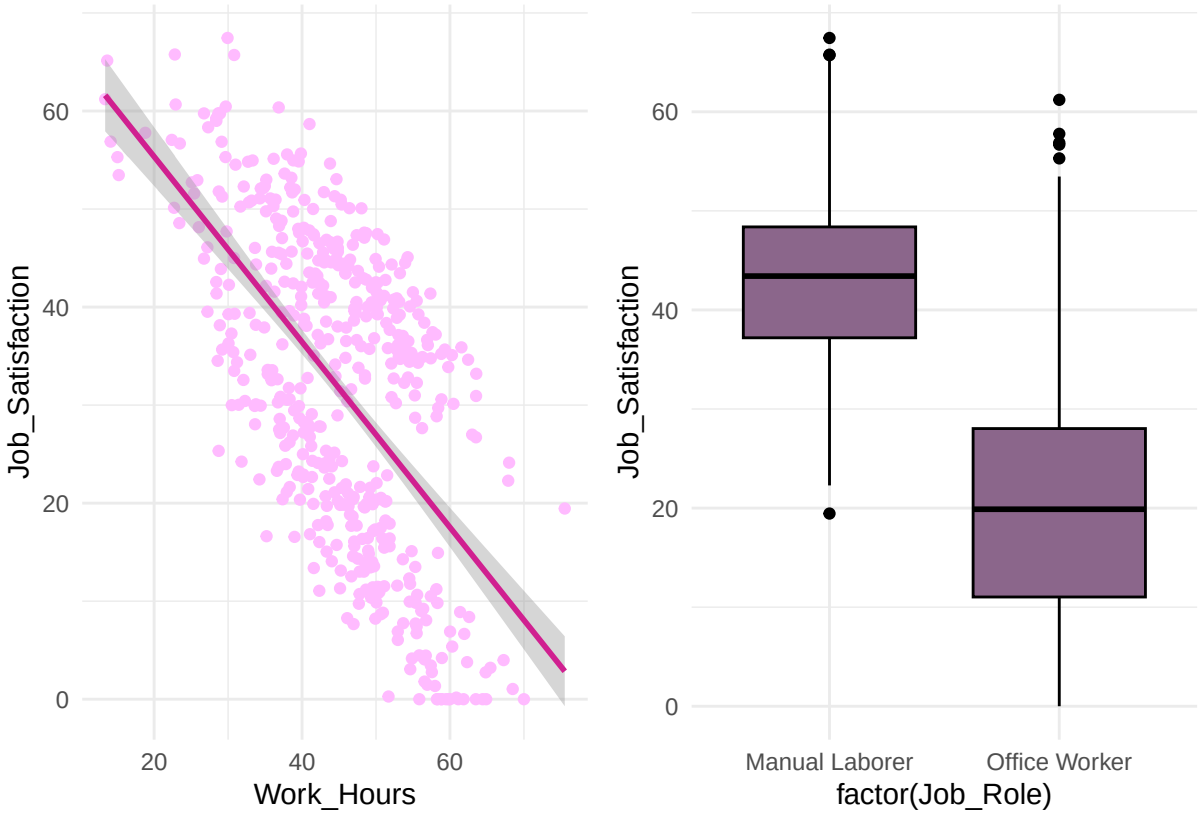
Next, we will generate scatterplots of our continuous predictor against the outcome of interest to indicate whether or not a linear relationship is appropriate. For our categorical predictor, we will use a box plot.

```
satisfaction_workhours_scatter <- ABCD_data %>%
  ggplot(aes(x = Work_Hours, y = Job_Satisfaction)) +
  geom_point(color = "plum1") +
  geom_smooth(method = "lm", color = "violetred") +
  theme_minimal()

satisfaction_jobrole <- ABCD_data %>%
  ggplot(aes(x = factor(Job_Role), y = Job_Satisfaction)) +
  geom_boxplot(color = "black", fill = "plum4") +
  theme_minimal()

satisfaction_workhours_scatter + satisfaction_jobrole
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



The scatterplot indicates a clear linear relationship between Work Hours and Job Satisfaction, where Job Satisfaction sharply decreases as Work Hours increase. While the downward trend is strong, the spread across the y axis again suggests other factors beyond work hours may likely contribute to job satisfaction.The boxplot indicates that Office Workers generally have lower Job Satisfaction compared to Manual Laborers. Nonetheless, both job roles have relatively low overall Job Satisfaction scores. There are some outliers in both groups falling closer to 60, indicating that there are some individuals with some satisfaction in their role at ABCD.

Our hypotheses are:

- H0: The effect of Work Hours on Job Satisfaction is the same for both Job Roles (Manual Laborer vs Office Worker). ßWork_Hours:Job_Role = 0
- HA: The effect of Work Hours on Job Satisfaction differs between Job Roles (Manual Laborer vs Office Worker). ßWork_Hours:Job_Role != 0

Our written models are:

- Full model: E[Job_Satisfaction] = ß0 + ß1(Work_Hours) + ß2(Job_Role) + ß3(Work_HoursxJob_Role)
- Reduced model: E[Job_Satisfaction] = ß0 + ß1(Work_Hours) + ß2(Job_Role)

```
# time to fit the models and run anova
full_model <- lm(Job_Satisfaction ~ Work_Hours + Job_Role + Work_Hours:Job_Role, data = ABCD_data)
reduced_model <- lm(Job_Satisfaction ~ Work_Hours + Job_Role, data = ABCD_data)

anova <- anova(reduced_model, full_model)
print(anova)
```

```
## Analysis of Variance Table
##
## Model 1: Job_Satisfaction ~ Work_Hours + Job_Role
## Model 2: Job_Satisfaction ~ Work_Hours + Job_Role + Work_Hours:Job_Role
```

```
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    497 14959
## 2    496 12751  1    2208.3 85.903 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(full_model)
```

```
##
## Call:
## lm(formula = Job_Satisfaction ~ Work_Hours + Job_Role + Work_Hours:Job_Role,
##     data = ABCD_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.7106  -3.3802   0.1646   2.9542  15.0794
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      74.43553    1.52941  48.669   <2e-16 ***
## Work_Hours                       -0.69275    0.03340 -20.743   <2e-16 ***
## Job_RoleOffice Worker            -3.91015    2.09003  -1.871    0.062 .
## Work_Hours:Job_RoleOffice Worker -0.42048    0.04537  -9.268   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.07 on 496 degrees of freedom
## Multiple R-squared:  0.898,  Adjusted R-squared:  0.8974
## F-statistic:  1456 on 3 and 496 DF,  p-value: < 2.2e-16
```

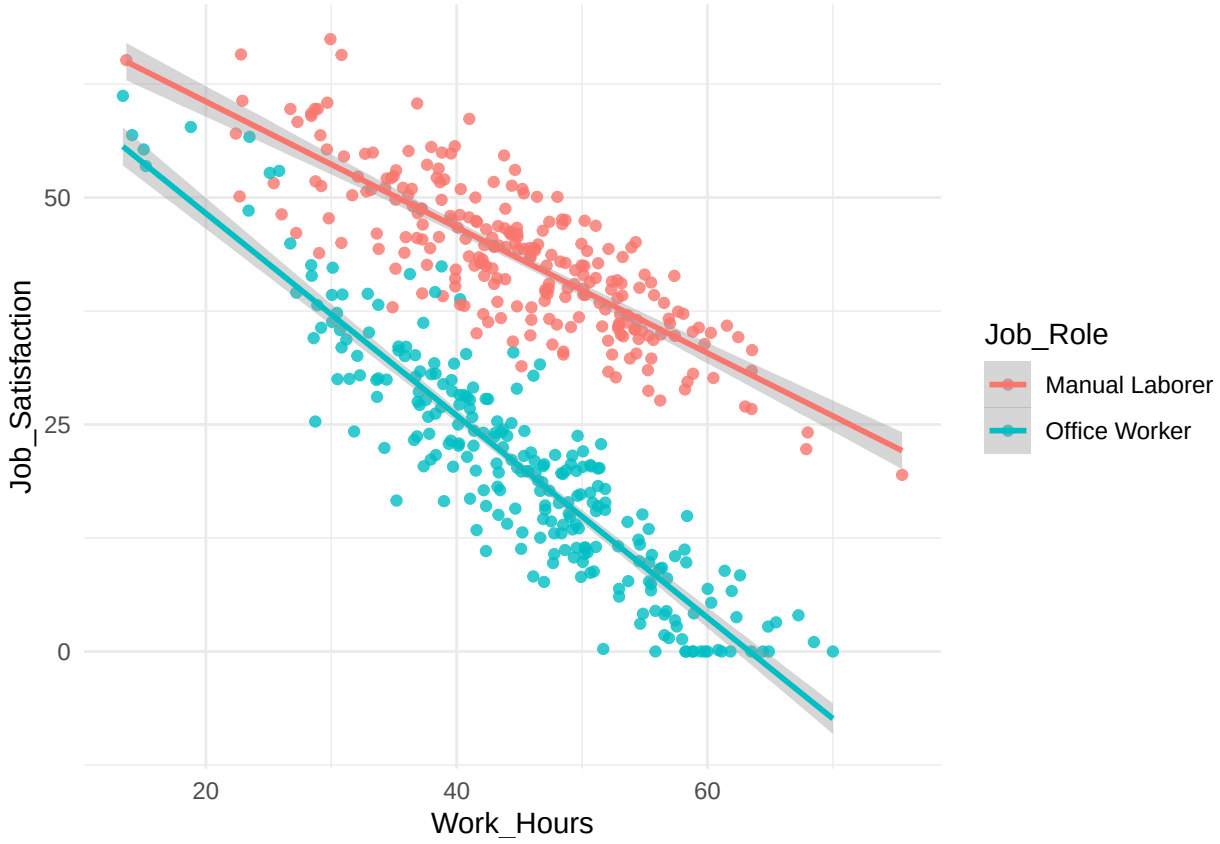
The p-value for the interaction term is <2.2e-16. Therefore there is strong evidence that the effect of Work Hours on Job Satisfaction differs depending on the Job Role (Manual Laborer vs Office Worker). Therefore, we reject H0 at a(alpha) = 0.05. The numerator df (df1) is 1. It's calculated as the difference between the residual df of the reduced model (497) and the full model (496). The df1 being 1 is because the full model has one additional parameter (the interaction term between Work_Hours and Job_Role) compared to the reduced model. The F-statistic is 85.903, indicating that including the interaction term in the model provides great improvement in model fit, relative to the unexplained variance.Therefore, the effect of Work Hours on Job Satisfaction does differ by Job Role.

To interpret the coefficients next, the intercept (ß0) is 74.43553. While no one is working 0 hours/week, ß0 represents the baseline Job Satisfaction score for a Manual Laborer (our reference group). In contrast, the baseline satisfaction score at 0 hours/week for Office Workers would be: 74.43553 + (-3.91015) = 70.52538. This indicates Office Workers being typically less satisfied than Manual Laborers, before even observing hours worked. The coefficient for Work_Hours (ß1) is -0.69, meaning that for Manual Laborers, Job Satisfaction decreases by approx. 0.69 points per hour worked. The interaction term's coefficient is -0.42, therefore Office workers' Job Satisfaction score decreases additionally by 0.42048 per hour worked than Manual Laborers, hence the effect of Work Hours on Job Satisfaction is even greater. Specifically, Office Workers' score decreases by 1.11, (-0.69 + (-0.42) = -1.11) per hour worked. This -ve effect is highly significant as the p-value < 2e-16. Below, we have a plot to visualize this -ve effect across Job Roles. Our interpretation of the coefficients aligns with the downward trend of Job Satisfaction decreasing as Work Hours increases in both Job Roles.

```
ggplot(ABCD_data, aes(x = Work_Hours,
                      y = Job_Satisfaction,
                      color = Job_Role)) +
  geom_point(alpha = 0.8) +
```

```
  geom_smooth(method = "lm") +
  labs(
    x = "Work_Hours",
    y = "Job_Satisfaction",
    color = "Job_Role") +
  theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



2. The company asks you if there is evidence that the job satisfaction for an office worker who works 40 hours per week is different than a manual laborer who works 35 hours per week. Assess this, being sure to provide a confidence interval for the quantity that you report.

First, let's check the Job Satisfaction of a Manual Laborer working 35 hours/week:

- prediction = ß0 + (Hours Worked x Work Hours Effect for Manual Laborers)
- prediction = 74.43553 + (35)(-0.69275)
- prediction = 50.18928
- If a Manual Laborer works 35 hours/week, their predicted Job Satisfaction will be approx. 50.19.

Next, let's check the Job Satisfaction of an Office Worker working 40 hours/week:

- prediction = ß0 + (Job Role Effect for Office Worker) + (Hours Worked x Work Hours Effect for Office Workers) + (Hours Worked x Interaction Term)
- prediction = ß0 + ß2 + (40 x ß1) + (40 x ß3)
- prediction = 74.43553 + (-3.91015) + (40)(-0.69275) + (40)(-0.42048))
- prediction = 25.99618
- If an Office Worker works 40 hours/week, their predicted Job Satisfaction will be approx. 25.99618.

Now, let's assess the difference between the two predictions:

For Manual Laborer at 35 hours/week:

- Intercept: 1
- Work_Hours: 35
- Job_Role: 0, since it's the reference category
- Interaction: (Work_Hours × Job_Role) = 35 × 0 = 0

For Office Worker at 40 hours/week:

- Intercept: 1
- Work_Hours: 40
- Job_Role: 1, since it's NOT the reference category
- Interaction: (Work_Hours × Job_Role) = 40 × 1 = 40

Difference:

- Intercept: 1 - 1 = 0
- Work Hours: 40 - 35 = 5
- Job Role: Office Worker - Manual Laborer = 1 - 0 = 1
- Interaction Term: 40 - 0 = 40

```
difference <- c(0, 5, 1, 40) # plug in our values (matching the order of coefs: Intercept, Work_Hours,

coefs <- coef(full_model) # get our coefs from the full model

point_estimate <- sum(coefs * difference)
point_estimate #-24.19302
```

```
## [1] -24.19302
```

```
vcov_matrix <- vcov(full_model) # get our variance covariance matrix of the coefs
vcov_matrix
```

```
##                                 (Intercept)   Work_Hours Job_RoleOffice Worker
## (Intercept)                      2.33910648 -0.049932538           -2.33910648
## Work_Hours                      -0.04993254  0.001115327            0.04993254
## Job_RoleOffice Worker           -2.33910648  0.049932538            4.36824059
## Work_Hours:Job_RoleOffice Worker  0.04993254 -0.001115327           -0.09255823
##                                 Work_Hours:Job_RoleOffice Worker
## (Intercept)                                        0.049932538
## Work_Hours                                        -0.001115327
## Job_RoleOffice Worker                             -0.092558232
## Work_Hours:Job_RoleOffice Worker                   0.002058158
```

```
# now we need to get the SE for our CI later
SE_difference <- sqrt(t(difference) %*% vcov_matrix %*% difference) # matrix multiplication here
SE_difference # 0.5811299
```

```
##           [,1]
## [1,] 0.5811299
```

```
# time to get the CI at 95% (specified by prof.)
#remember, 0.025 in lower tail, 0.025 in upper tail so, 0.975 for 95%
crit_t_val <- qt(0.975, df = df.residual(full_model)) # 0.5811299
crit_t_val
```

```
## [1] 1.964758
```

```
CI_lower <- point_estimate - (crit_t_val * SE_difference)
CI_lower # -25.3348
```

```
##           [,1]
## [1,] -25.3348
```

```
CI_upper <- point_estimate + (crit_t_val * SE_difference)
CI_upper # -23.05124
```

```
##           [,1]
## [1,] -23.05124
```

We calculated that a) a Manual Laborer working 35 hours/week, has a predicted Job Satisfaction of approx. 50.19 b) an Office Worker working 40 hours/week has a predicted Job Satisfaction of approx. 25.99618. The difference in predicted job satisfaction between these groups is -24.19302, indicating that Office Workers working 40 hours/week have lower job satisfaction compared to Manual Laborers working 35 hours/week, by approx. 24.19 points. We constructed a 95% CI which is: (-25.3348,-23.05124), where this is the range of plausible values for the true difference in job satisfaction between the job roles. Since the CI does not contain 0, we can reject the H0 that there is no difference in job satisfaction for an office worker who works 40 hours/week and a manual laborer who works 35 hours/week.

- (Note: The problem's wording didn't specify if the company had an H0 or HA specifically in mind when asking for this assessment. However, if the company's hypothesis was that there is no difference in job satisfaction between the job roles, then we would be rejecting it.)

3. The company asks you if there is evidence that the job satisfaction for an office worker who works 20 hours per week is different than a manual laborer who works 40 hours per week. Assess this, being sure to provide a confidence interval for the quantity that you report.

First, let's check the Job Satisfaction of a Manual Laborer working 40 hours/week:

- prediction = ß0 + (Hours Worked x Work Hours Effect for Manual Laborers)
- prediction = 74.43553 + (40)(-0.69275)
- prediction = 46.72553
- If a Manual Laborer works 40 hours/week, their predicted Job Satisfaction will be approx. 46.72.

Next, let's check the Job Satisfaction of an Office Worker working 20 hours/week:

- prediction = ß0 + Job Role Effect for Office Worker + (Hours Worked x Work Hours Effect for Office Workers) + (Hours Worked x Interaction Term)
- prediction = ß0 + ß2 + (20 x ß1) + (20 x ß3)
- prediction = 74.43553 + (-3.91015) + (20)(-0.69275)) + (20)(-0.42048))
- prediction = 48.26078
- If an Office Worker works 40 hours/week, their predicted Job Satisfaction will be approx. 48.26.

Now, let's assess the difference between the two predictions:

For Manual Laborer at 40 hours/week:

- Intercept: 1
- Work_Hours: 40
- Job_Role: 0, since it's the reference category
- Interaction: (Work_Hours × Job_Role) = 40 × 0 = 0

For Office Worker at 20 hours/week:

- Intercept: 1
- Work_Hours: 20
- Job_Role: 1, since it's NOT the reference category
- Interaction: (Work_Hours × Job_Role) = 20 × 1 = 20

Difference:

- Intercept: 1 - 1 = 0
- Work Hours: 20 - 40 = -20

- Job Role: Office Worker - Manual Laborer = 1 - 0 = 1
- Interaction Term: 20 - 0 = 20

```
difference <- c(0, -20, 1, 20) # plug in our values again

coefs <- coef(full_model)

point_estimate <- sum(coefs * difference) # new point estimate
point_estimate #1.535204
```

```
## [1] 1.535204
```

```
vcov_matrix <- vcov(full_model)
vcov_matrix
```

```
##                                 (Intercept)   Work_Hours Job_RoleOffice Worker
## (Intercept)                      2.33910648 -0.049932538           -2.33910648
## Work_Hours                      -0.04993254  0.001115327            0.04993254
## Job_RoleOffice Worker           -2.33910648  0.049932538            4.36824059
## Work_Hours:Job_RoleOffice Worker  0.04993254 -0.001115327           -0.09255823
##                                 Work_Hours:Job_RoleOffice Worker
## (Intercept)                                          0.049932538
## Work_Hours                                          -0.001115327
## Job_RoleOffice Worker                               -0.092558232
## Work_Hours:Job_RoleOffice Worker                     0.002058158
```

```
# get SE again for our CI
SE_difference <- sqrt(t(difference) %*% vcov_matrix %*% difference)
SE_difference # 0.9111892
```

```
##           [,1]
## [1,] 0.9111892
```

```
# time to get the CI at 95% (specified by prof.)
crit_t_val <- qt(0.975, df = df.residual(full_model)) # 1.964758
crit_t_val
```

```
## [1] 1.964758
```

```
CI_lower <- point_estimate - (crit_t_val * SE_difference)
CI_lower # -0.2550625
```

```
##            [,1]
## [1,] -0.2550625
```

```
CI_upper <- point_estimate + (crit_t_val * SE_difference)
CI_upper # 3.32547
```

```
##         [,1]
## [1,] 3.32547
```

We calculated that a) a Manual Laborer working 40 hours/week, has a predicted Job Satisfaction of approx. 46.72 b) an Office Worker working 20 hours/week has a predicted Job Satisfaction of approx. 48.26. The difference in predicted job satisfaction between job roles is 1.535204, indicating that Manual Laborers working 40 hours/week have lower job satisfaction compared to Office Workers working 20 hours/week, by approx. 1.54 points. We constructed a 95% CI which is: (-0.2550625, 3.32547), where this is the range of plausible values for the true difference in job satisfaction between the job roles. Since the CI contains 0, we fail to reject the H0 that there is no difference in job satisfaction for an office worker who works 20 hours/week and a manual laborer who works 40 hours/week. Since the true difference could be 0, their job satisfaction may

be the same.