Assignment 3: Predicting insurance charges by age and BMI

Your name and student ID

Today's date

Instructions

- Solutions will be released on Tuesday, September 15.
- This semester, homework assignments are for practice only and will not be turned in for marks.

Helpful hints:

- Every function you need to use was taught during lecture! So you may need to revisit the lecture code to help you along by opening the relevant files on Datahub. Alternatively, you may wish to view the code in the condensed PDFs posted on the course website. Good luck!
- Knit your file early and often to minimize knitting errors! If you copy and paste code for the slides, you are bound to get an error that is hard to diagnose. Typing out the code is the way to smooth knitting! We recommend knitting your file each time after you write a few sentences/add a new code chunk, so you can detect the source of knitting errors more easily. This will save you and the GSIs from frustration! **You must knit correctly before submitting.**
- It is good practice to not allow your code to run off the page. To avoid this, have a look at your knitted PDF and ensure all the code fits in the file. If it doesn't look right, go back to your .Rmd file and add spaces (new lines) using the return or enter key so that the code runs onto the next line.

```r
library(readr)
library(dplyr)
library(ggplot2)
library(broom)
library(forcats)
```

Predicting insurance charges by age and BMI

Problem: Medical insurance charges can vary according to the complexity of a procedure or condition that requires medical treatment. You are tasked with determining how these charges are associated with age, for patients who have a body mass index (bmi) in the "normal" range (bmi between 16 and 25) who are smokers.

Plan: You have chosen to use tools to examine relationships between two variables to address the problem. In particular, scatter plots and simple linear regression.

Data: You have access to the dataset `insurance.csv`, a claims dataset from an insurance provider.

Analysis and Conclusion: In this assignment you will perform the analysis and make a conclusion to help answer the problem statement.

1. [1 point] Please type one line of code below to import these data into R. Assign the data to `insure_data`. Execute the code by hitting the green arrow and ensure the data set has been saved by looking at the environment tab and viewing the data set by clicking the table icon to the right of its name.

```
insure_data <- "<<<<YOUR CODE HERE>>>>"

# ------- REMINDER -------
# The checks on this homework are only provided as sanity checks.
# They do not guarantee correctness.
# ------------------------

check_problem1()
```

```
## [1] "Checkpoint 1 Error: Incorrect. Try again."
## [1] "Checkpoint 2 Error: Please use the version of this data-reading function with _ instead of ."
##
## Problem 1
## Checkpoints Passed: 0
## Checkpoints Errored: 2
## 0% passed
## --------
## Test: FAILED
```

Use the space below to use the functions from lecture to get to know your dataset. Execute these functions line by line so you can look at their output, and examine these data.

```
dim(insure_data)
```

```
## NULL
```

```
names(insure_data)
```

```
## NULL
```

```
str(insure_data)
```

```
##  chr "<<<<YOUR CODE HERE>>>>"
```

```
head(insure_data)
```

```
## [1] "<<<<YOUR CODE HERE>>>>"
```

2. [1 point] How many individuals are in the dataset? Assign this number to p2.

```
p2 <- "<<<<YOUR CODE HERE>>>>"


check_problem2()
```

```
## [1] "Checkpoint 1 Error: Please assign p2 to a numeric."
## [1] "Checkpoint 2 Error: Please answer the correct number."
##
## Problem 2
## Checkpoints Passed: 0
## Checkpoints Errored: 2
## 0% passed
## --------
## Test: FAILED
```

3. [1 point] What are the nominal variables in the dataset? Assign the names of these variables to a vector of strings, p3.

```
p3 <- "<<<<YOUR CODE HERE>>>>"


check_problem3()
```

```
## [1] "Checkpoint 1 Passed: p3 is correctly assigned as a vector."
## [1] "Checkpoint 2 Passed: Variables are correctly specified as strings."
## [1] "Checkpoint 3 Error: Please include the required variables."
## [1] "Checkpoint 4 Error: Please check the length of your input"
##
## Problem 3
## Checkpoints Passed: 2
## Checkpoints Errored: 2
## 50% passed
## --------
## Test: FAILED
```

4. [1 point] How many ordinal variables are in the dataset? Assign the number of ordinal variables to p4.

```
p4 <- "<<<<YOUR CODE HERE>>>>"


check_problem4()
```

```
## [1] "Checkpoint 1 Error: Make sure p4 is numeric."
## [1] "Checkpoint 2 Passed: The number of selected variables are correct."
##
## Problem 4
## Checkpoints Passed: 1
## Checkpoints Errored: 1
## 50% passed
## --------
## Test: FAILED
```

5. [1 point] Are there continuous variables in the dataset? Assign the names of these variables to a vector of strings, p5.

```
p5 <- "<<<<YOUR CODE HERE>>>>"


check_problem5()
```

```
## [1] "Checkpoint 1 Passed: p5 is assigned as a vector."
## [1] "Checkpoint 2 Error: Please enter your variable names as strings (ie in quotations)"
##
## Problem 5
## Checkpoints Passed: 1
## Checkpoints Errored: 1
## 50% passed
## --------
## Test: FAILED
```

6. [1 point] What are the discrete variables in the dataset? Assign the names of these variables to a vector of strings, p6.

```
p6 <- "<<<<YOUR CODE HERE>>>>"


check_problem6()
```

```
## [1] "Checkpoint 1 Passed: p6 is assigned to a vector of strings."
## [1] "Checkpoint 2 Error: Please reconsider which variables are discrete."
##
## Problem 6
## Checkpoints Passed: 1
## Checkpoints Errored: 1
## 50% passed
## --------
## Test: FAILED
```

Run the following code by removing the "#" symbol in front of each of the six lines and executing the code chunk. Remind yourself what the `mutate()` function does in general, and notice that a new function `case_when()` is also being used.

##STOP: Remove eval = F from this chunk header before proceeding

```
insure_data <- insure_data %>%
  mutate(bmi_cat = case_when(bmi < 16 ~ "Underweight",
                             bmi >= 16 & bmi < 25 ~ "Normal",
                             bmi >= 25 & bmi < 30 ~ "Overweight",
                             bmi >= 30 ~ "Obese")
                 )
```

7. [1 point] What did the above code achieve?:

[TODO: YOUR ANSWER HERE]

8. [1 point] What type of variable is `bmi_cat`? Uncomment one of the choices below.

```
# p8 <- "ordinal"
# p8 <- "nominal"
# p8 <- "continuous"
# p8 <- "discrete"

# This only checks that you've selected an answer, not its correctness.
check_problem8()
```

```
## [1] "Checkpoint 1 Error: Incorrect choice."
##
## Problem 8
## Checkpoints Passed: 0
## Checkpoints Errored: 1
## 0% passed
## --------
## Test: FAILED
```

9. [1 point] Read the problem statement proposed at the beginning of this exercise. Who belongs to the population of interest? Uncomment one of the choices below.

```
# p9 <- "Smokers of normal BMI"
# p9 <- "Smokers of overweight BMI"
# p9 <- "Smokers who have abnormal BMI"
# p9 <- "All people at risk of high medical charges"

# This only checks that you've selected an answer, not its correctness.
check_problem9()
```

```
## [1] "Checkpoint 1 Error: Incorrect choice."
##
## Problem 9
## Checkpoints Passed: 0
## Checkpoints Errored: 1
## 0% passed
## --------
## Test: FAILED
```

10. [1 point] Using a dplyr function, make a new dataset called `insure_subset` containing the population of interest.

```
insure_subset <- "<<<<YOUR CODE HERE>>>>"


check_problem10()
```

```
## [1] "Checkpoint 1 Error: Please ensure insure_subset is a date frame."
## [1] "Checkpoint 1 Error: The number of observation fails to lie in the current range."
##
## Problem 10
## Checkpoints Passed: 0
## Checkpoints Errored: 2
## 0% passed
## --------
## Test: FAILED
```

11. [3 points] Make a scatter plot of the relationship between age and insurance charges for the population of interest. Give your plot an informative title.

```
p11 <- "<<<<YOUR CODE HERE>>>>"
p11
```

```
## [1] "<<<<YOUR CODE HERE>>>>"
```

```
check_problem11()
```

```
## [1] "Checkpoint 1 Error: You did not define a ggplot."
## [1] "Checkpoint 2 Error: Did you use the right dataset?"
## [1] "Checkpoint 3 Error: Choose the correct variable for x axis."
## [1] "Checkpoint 4 Error: Choose the correct variable for y axis."
## [1] "Checkpoint 5 Error: Did you use the scatter plot?"
## [1] "Checkpoint 6 Error: Did you forget to add a title?"
##
## Problem 11
## Checkpoints Passed: 0
## Checkpoints Errored: 6
## 0% passed
## --------
## Test: FAILED
```

12. [2 points] Run a linear regression model on the relationship between `age` and `charges`. Think about which variable is explanatory (X) and which is response (Y). Assign the regression model to the name `insure_mod`. Then type `tidy(insure_mod)` below the model's code and execute both lines.

```
insure_model <- "<<<<YOUR CODE HERE>>>>"
# <<<<YOUR CODE HERE>>>>


check_problem12()
```

```
## [1] "Checkpoint 1 Error: You didn't fit a linear regression model."
## [1] "Checkpoint 2 Error: A variable required is missing in the model."
## [1] "Checkpoint 3 Error: A variable required is missing in the model."
##
## Problem 12
## Checkpoints Passed: 0
## Checkpoints Errored: 3
## 0% passed
## --------
## Test: FAILED
```

13a. [1 point] Interpret the slope parameter:

[TODO: YOUR ANSWER HERE]

13b. [1 point] Interpret the intercept parameter:

[TODO: YOUR ANSWER HERE]

13c. [1 point] Does the intercept make sense in this context?:

[TODO: YOUR ANSWER HERE]

14. [1 point] Add the line of best fit to your scatterplot by copying and pasting the plot's code from Problem 11 into the chunk below and adding a `geom` that can be used to add a regression line:

```
p14 <- "<<<<YOUR CODE HERE>>>>"
p14
```

```
## [1] "<<<<YOUR CODE HERE>>>>"
```

```
check_problem14()
```

```
## [1] "Checkpoint 1 Error: You did not define a ggplot."
## [1] "Checkpoint 2 Error: Did you use the right dataset?"
## [1] "Checkpoint 3 Error: Please select the correct variable for x axis."
## [1] "Checkpoint 4 Error: Please select the correct variable for y axis."
## [1] "Checkpoint 5 Error: Did you use the scatter plot?"
## [1] "Checkpoint 6 Error: Did you add a title?"
## [1] "Checkpoint 7 Error: Did you plot the regression line?"
##
## Problem 14
## Checkpoints Passed: 0
## Checkpoints Errored: 7
## 0% passed
## --------
## Test: FAILED
```

15. [2 points] What do you notice about the fit of the line in terms of the proportion of points above vs. below the line. Why do you think that is?:

[TODO: YOUR ANSWER HERE]

Run the following `filter()` function by removing the "#" symbol in front of the two lines of code and executing the code chunk.

STOP: Remove eval = F from this chunk header before proceeding

```
insure_smaller_subset <- insure_subset %>%
  filter(charges < 30000 & ! (charges > 25000 & age == 20))
```

16. [2 points] How many individuals were removed? Who were they?:

[TODO: YOUR ANSWER HERE]

17. [2 points] Run a regression model on `insure_smaller_subset` between `charges` and `age`. Assign it to insure_better_model and look at the output using the `tidy()` function, as was done with the previous linear model.

```
insure_better_model <- "<<<<YOUR CODE HERE>>>>"
# "<<<<YOUR CODE HERE>>>>"


check_problem17()
```

```
## [1] "Checkpoint 1 Error: Please use linear regression model."
## [1] "Checkpoint 2 Error: Please include the required variables in the model."
## [1] "Checkpoint 3 Error: Please include the required variables in the model."
## [1] "Checkpoint 4 Error: The number of observations is incorrect. Please check the dataset."
##
## Problem 17
## Checkpoints Passed: 0
## Checkpoints Errored: 4
## 0% passed
## --------
## Test: FAILED
```

18. [2 points] Add the new regression line to your ggplot from Problem 14. Keep the older regression line on the plot for comparison. To distinguish them, change the color, line type, or line width of the newly-added line.

```
p18 <- "<<<<YOUR CODE HERE>>>>"
p18
```

```
## [1] "<<<<YOUR CODE HERE>>>>"
```

```
check_problem18()
```

```
## [1] "Checkpoint 1 Error: You did not define a ggplot."
## [1] "Checkpoint 2 Error: Did you use the right dataset?"
## [1] "Checkpoint 3 Error: Please use the required variable for x axis."
## [1] "Checkpoint 4 Error: Please use the required variable for y axis."
## [1] "Checkpoint 5 Error: Did you use a scatter plot?."
## [1] "Checkpoint 6 Error: Did you add a title?"
## [1] "Checkpoint 7 Error: Please add the original regression line."
## [1] "Checkpoint 8 Error: Please add the new regression line."
##
## Problem 18
## Checkpoints Passed: 0
## Checkpoints Errored: 8
## 0% passed
## --------
## Test: FAILED
```

19. [1 point] Calculate the r-squared value for `insure_model` using a function learned in class. Assign this value to insure_model_r2.

```
insure_model_r2 <- "<<<<YOUR CODE HERE>>>>"


check_problem19()
```

```
## [1] "Checkpoint 1 Error: Please assign a numeric to insure_model_r2"
## [1] "Checkpoint 2 Error: Not a valid r-squared value."
## [1] "Checkpoint 3 Error: Check your value. It should be less than 0.5."
##
## Problem 19
## Checkpoints Passed: 0
## Checkpoints Errored: 3
## 0% passed
## --------
## Test: FAILED
```

20. [1 point] Calculate the r-squared value for `insure_better_model` using a function learned in class. Assign this value to insure_better_model_r2.

```
insure_better_model_r2 <- "<<<<YOUR CODE HERE>>>>"


check_problem20()
```

```
## [1] "Checkpoint 1 Error: Please assign a numeric to insure_better_model_r2"
## [1] "Checkpoint 2 Error: Not a valid r-square value."
## [1] "Checkpoint 3 Error: Check your value. It should be greater than 0.5."
##
## Problem 20
## Checkpoints Passed: 0
## Checkpoints Errored: 3
## 0% passed
## --------
## Test: FAILED
```

21. [2 points] Calculate the correlation between `age` and `charges` using the subset `insure_subset`. Also calculate correlation squared. You should use `summarize()` and name the two new columns `corr` and `corr_sq`. What do you notice about the relationship between the correlation and r-squared values that you calculated earlier?

```
p21 <- "<<<<YOUR CODE HERE>>>>"


check_problem21()
```

```
## [1] "Checkpoint 1 Error: Please assign a numeric to insure_better_model_r2."
## [1] "Checkpoint 2 Error: Your input should be a double variable."
## [1] "Checkpoint 3 Error: Your input should be a double variable."
##
## Problem 21
## Checkpoints Passed: 0
## Checkpoints Errored: 3
## 0% passed
## --------
## Test: FAILED
```

22. [2 points] Calculate the correlation between `age` and `charges` using the smaller dataset `insure_smaller_subset`. Also calculate correlation squared. You should use `summarize()` and name the two new columns `corr` and `corr_sq`. What do you notice about the relationship between the correlation and r-squared values that you calculated earlier?

```
p22 <- "<<<<YOUR CODE HERE>>>>"


check_problem22()
```

```
## [1] "Checkpoint 1 Error: Please assign a numeric to insure_better_model_r2."
## [1] "Checkpoint 2 Error: Your input should be a double variable."
## [1] "Checkpoint 3 Error: Your input should be a double variable."
##
## Problem 22
## Checkpoints Passed: 0
## Checkpoints Errored: 3
## 0% passed
## --------
## Test: FAILED
```

PART B

Your supervisor asks you to extend your analysis to consider other smokers with BMIs classified as overweight or obese. In particular, she wanted to know if the relationship between age and medical charges is different for different BMI groups. You can use data visualization coupled with your skills in linear regression to help answer this question.

23. [1 point] Make a new dataset called `insure_smokers` that includes smokers of any BMI.

```
insure_smokers <- "<<<<YOUR CODE HERE>>>>"


check_problem23()
```

```
## [1] "Checkpoint 1 Error: The result is not a data frame."
## [1] "Checkpoint 1 Error: The number of observations falls in an incorrect range."
##
## Problem 23
## Checkpoints Passed: 0
## Checkpoints Errored: 2
## 0% passed
## --------
## Test: FAILED
```

24. [1 point] Make a scatter plot that examines the relationship between `age` and `charges` separately for normal, overweight, and obese individuals. A `facet_` command may help you.

```
p24 <- "<<<<YOUR CODE HERE>>>>"
p24
```

```
## [1] "<<<<YOUR CODE HERE>>>>"
```

```
check_problem24()
```

```
## [1] "Checkpoint 1 Error: You did not define a ggplot."
## [1] "Checkpoint 2 Error: Did you use the right dataset?"
## [1] "Checkpoint 3 Error: Incorrect x variable."
## [1] "Checkpoint 4 Error: Incorrect y variable."
## [1] "Checkpoint 5 Error: You didn't use the scatter plot."
## [1] "Checkpoint 6 Error: Did you wrap the variables using facet_wrap?"
##
## Problem 24
## Checkpoints Passed: 0
## Checkpoints Errored: 6
## 0% passed
## --------
## Test: FAILED
```

Is there something out of order with your plot you just made? The issue is that the plot is automatically displayed by listing the BMI categories alphabetically. Uncomment and run the following code chunk to assign an ordering to the values of `bmi_cat`.:

STOP: Remove eval = F from this chunk header before proceeding

```
insure_smokers <- insure_smokers %>%
  mutate(bmi_cat_ordered = forcats::fct_relevel(bmi_cat, "Normal", "Overweight", "Obese"))
```

25. [1 point] Re-run your plot code, but this time, facet using `bmi_cat_ordered`.

```
p25 <- "<<<<YOUR CODE HERE>>>>"
p25
```

```
## [1] "<<<<YOUR CODE HERE>>>>"
```

```
check_problem25()
```

```
## [1] "Checkpoint 1 Error: You did not define a ggplot."
## [1] "Checkpoint 2 Error: Did you use the right dataset?"
## [1] "Checkpoint 3 Error: Please use the required variable for x axis."
## [1] "Checkpoint 4 Error: Please use the required variable for y axis."
## [1] "Checkpoint 5 Error: Please use scatter plot."
## [1] "Checkpoint 6 Error: Did you wrap the data points into three facets for regression?"
##
## Problem 25
## Checkpoints Passed: 0
## Checkpoints Errored: 6
## 0% passed
## --------
## Test: FAILED
```

26. [3 points] Run a separate linear model for each BMI group. To do this, you will need to subset your data into the three groups of interest first. Call your models `normal_mod`, `overweight_mod`, `obese_mod`. Use the `tidy()` function to display the output from each model.

```
# "<<<<YOUR CODE HERE>>>>"
# "<<<<YOUR CODE HERE>>>>"
# "<<<<YOUR CODE HERE>>>>"

normal_mod <- "<<<<YOUR CODE HERE>>>>"
overweight_mod <- "<<<<YOUR CODE HERE>>>>"
obese_mod <- "<<<<YOUR CODE HERE>>>>"

# "<<<<YOUR CODE HERE>>>>"
# "<<<<YOUR CODE HERE>>>>"
# "<<<<YOUR CODE HERE>>>>"


check_problem26()
```

```
## [1] "Checkpoint 1 Error: Please use linear regression model."
## [1] "Checkpoint 2 Error: Please select the required variables."
## [1] "Checkpoint 3 Error: Please select the required variables"
## [1] "Checkpoint 4 Error: Please use linear regression model."
## [1] "Checkpoint 5 Error: Please select the required variables"
## [1] "Checkpoint 6 Error: Please select the required variables"
## [1] "Checkpoint 7 Error: Please use linear regression model."
## [1] "Checkpoint 8 Error: Please select the required variables"
## [1] "Checkpoint 9 Error: Please select the required variables"
##
## Problem 26
## Checkpoints Passed: 0
## Checkpoints Errored: 9
## 0% passed
## --------
## Test: FAILED
```

For the next three problems, use the models to predict medical charges for a 20-year old by weight category. You don't need an R function to make these predictions, just the output from the model. Show your work for each calculation by writing the mathematical expression in and round to the nearest dollar.

27. [1 point] **. . . among normal BMI group:**

```
p27 <- "<<<<YOUR CODE HERE>>>>"


check_problem27()
```

```
## [1] "Checkpoint 1 Error: The result should be numeric."
## [1] "Checkpoint 1 Error: The result should fall in the correct range."
##
## Problem 27
## Checkpoints Passed: 0
## Checkpoints Errored: 2
## 0% passed
## --------
## Test: FAILED
```

28. [1 point] **. . . among overweight BMI group:**

```
p28 <- "<<<<YOUR CODE HERE>>>>"


check_problem28()
```

```
## [1] "Checkpoint 1 Error: The result should be numeric."
## [1] "Checkpoint 1 Error: The result should fall in the correct range."
##
## Problem 28
## Checkpoints Passed: 0
## Checkpoints Errored: 2
## 0% passed
## --------
## Test: FAILED
```

29. [1 point] **. . . among obese BMI group:**

```
p29 <- "<<<<YOUR CODE HERE>>>>"


check_problem29()
```

```
## [1] "Checkpoint 1 Error: The result should be numeric."
## [1] "Checkpoint 1 Error: The result should fall in the correct range."
##
## Problem 29
## Checkpoints Passed: 0
## Checkpoints Errored: 2
## 0% passed
## --------
## Test: FAILED
```

30. [3 points] In three sentences maximum, (1) comment on the direction of the association, (2) comment on how much the slopes vary across the BMI groups, and (3) how much the prediction for a 20-year old varies.

[TODO: YOUR ANSWER HERE]

Check your score

Click on the middle icon on the top right of this code chunk (with the downwards gray arrow and green bar) to run all your code in order. Then, run this chunk to check your score.

```
# Just run this chunk.
total_score()
```

```
##                      Test Points_Possible          Type
## Problem 1          FAILED               1    autograded
## Problem 2          FAILED               1    autograded
## Problem 3          FAILED               1    autograded
## Problem 4          FAILED               1    autograded
## Problem 5          FAILED               1    autograded
## Problem 6          FAILED               1    autograded
## Problem 7  NOT YET GRADED               1 free-response
## Problem 8          FAILED               1    autograded
## Problem 9          FAILED               1    autograded
## Problem 10         FAILED               1    autograded
## Problem 11         FAILED               3    autograded
## Problem 12         FAILED               2    autograded
## Problem 13 NOT YET GRADED               3 free-response
## Problem 14         FAILED               1    autograded
## Problem 15 NOT YET GRADED               2 free-response
## Problem 16 NOT YET GRADED               2 free-response
## Problem 17         FAILED               2    autograded
## Problem 18         FAILED               2    autograded
## Problem 19         FAILED               1    autograded
## Problem 20         FAILED               1    autograded
## Problem 21         FAILED               2    autograded
## Problem 22         FAILED               2    autograded
## Problem 23         FAILED               1    autograded
## Problem 24         FAILED               1    autograded
## Problem 25         FAILED               1    autograded
## Problem 26         FAILED               3    autograded
## Problem 27         FAILED               1    autograded
## Problem 28         FAILED               1    autograded
## Problem 29         FAILED               1    autograded
## Problem 30 NOT YET GRADED               3 free-response
```