Assignment 3: Predicting insurance charges by age and BMI

Your name and student ID

July 15, 2025

Run this chunk of code to load the autograder package!

Instructions

- 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] **Type one line of code to import these data into R. Assign the data to `insure_data`. Execute the code by hitting the green arrow and ensure the dataset 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.**

```r
insure_data <- read_csv("data/insurance.csv") # SOLUTION
```

```
## Rows: 1338 Columns: 7
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): sex, smoker, region
## dbl (4): age, bmi, children, charges
##
## 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.
```

```r
insure_data
```

```
## # A tibble: 1,338 x 7
##      age sex      bmi children smoker region    charges
##    <dbl> <chr>  <dbl>    <dbl> <chr>  <chr>       <dbl>
##  1    19 female  27.9        0 yes    southwest  16885.
##  2    18 male    33.8        1 no     southeast   1726.
##  3    28 male    33          3 no     southeast   4449.
##  4    33 male    22.7        0 no     northwest  21984.
##  5    32 male    28.9        0 no     northwest   3867.
##  6    31 female  25.7        0 no     southeast   3757.
##  7    46 female  33.4        1 no     southeast   8241.
##  8    37 female  27.7        3 no     northwest   7282.
##  9    37 male    29.8        2 no     northeast   6406.
## 10    60 female  25.8        0 no     northwest  28923.
## # i 1,328 more rows
```

```r
. = ottr::check("tests/p1.R")
```

```
## All tests passed!
```

Execute the functions below one line at a time to get to know your dataset.

```r
dim(insure_data)
```

```
## [1] 1338    7
```

```r
names(insure_data)
```

```
## [1] "age"      "sex"      "bmi"      "children" "smoker"   "region"   "charges"
```

```r
str(insure_data)
```

```
## spc_tbl_ [1,338 x 7] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ age     : num [1:1338] 19 18 28 33 32 31 46 37 37 60 ...
##  $ sex     : chr [1:1338] "female" "male" "male" "male" ...
##  $ bmi     : num [1:1338] 27.9 33.8 33 22.7 28.9 ...
##  $ children: num [1:1338] 0 1 3 0 0 0 1 3 2 0 ...
##  $ smoker  : chr [1:1338] "yes" "no" "no" "no" ...
##  $ region  : chr [1:1338] "southwest" "southeast" "southeast" "northwest" ...
##  $ charges : num [1:1338] 16885 1726 4449 21984 3867 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   age = col_double(),
##   ..   sex = col_character(),
##   ..   bmi = col_double(),
##   ..   children = col_double(),
##   ..   smoker = col_character(),
##   ..   region = col_character(),
##   ..   charges = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```r
head(insure_data)
```

```
## # A tibble: 6 x 7
##     age sex      bmi children smoker region    charges
##   <dbl> <chr>  <dbl>    <dbl> <chr>  <chr>       <dbl>
## 1    19 female  27.9        0 yes    southwest  16885.
## 2    18 male    33.8        1 no     southeast   1726.
## 3    28 male    33          3 no     southeast   4449.
## 4    33 male    22.7        0 no     northwest  21984.
## 5    32 male    28.9        0 no     northwest   3867.
## 6    31 female  25.7        0 no     southeast   3757.
```

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

```r
p2 <- nrow(insure_data) # SOLUTION
p2
```

```
## [1] 1338
```

```r
. = ottr::check("tests/p2.R")
```

```
## All tests passed!
```

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

```r
p3 <- c("sex", "smoker", "region") # SOLUTION
p3
```

```
## [1] "sex"    "smoker" "region"
```

```r
. = ottr::check("tests/p3.R")
```

```
## All tests passed!
```

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

```r
p4 <- 0  # SOLUTION
p4
```

```
## [1] 0
```

```r
. = ottr::check("tests/p4.R")
```

```
## All tests passed!
```

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

```r
# BEGIN SOLUTION
p5 <- c("bmi", "charges", "age")
p5 <- c("bmi", "charges") # also accepted
# END SOLUTION
p5
```

```
## [1] "bmi"     "charges"
```

```r
. = ottr::check("tests/p5.R")
```

```
## All tests passed!
```

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

```r
# BEGIN SOLUTION
p6 <- c("children")
p6 <- c("children", "age") # also accepted
# END SOLUTION
p6
```

```
## [1] "children" "age"
```

```r
. = ottr::check("tests/p6.R")
```

```
## All tests passed!
```

Run the following code. Remind yourself what the `mutate()` function does in general, and notice that a new function called `case_when()` is also being used.

```
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. What did the code above accomplish?

Type your answer here, replacing this text.

The above code created a new variable called `bmi_cat` that created four categories of BMI: underweight, normal, overweight, and obese, based on the continuous variable BMI.

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

```r
p8 <- 'ordinal' # SOLUTION
p8
```

```
## [1] "ordinal"
```

```r
. = ottr::check("tests/p8.R")
```

```
## All tests passed!
```

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' # SOLUTION
p9
```

```
## [1] "Smokers of normal BMI"
```

```
. = ottr::check("tests/p9.R")
```

```
## All tests passed!
```

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

```
insure_subset <- insure_data %>% filter(smoker == "yes" & bmi_cat == "Normal") # SOLUTION
insure_subset
```

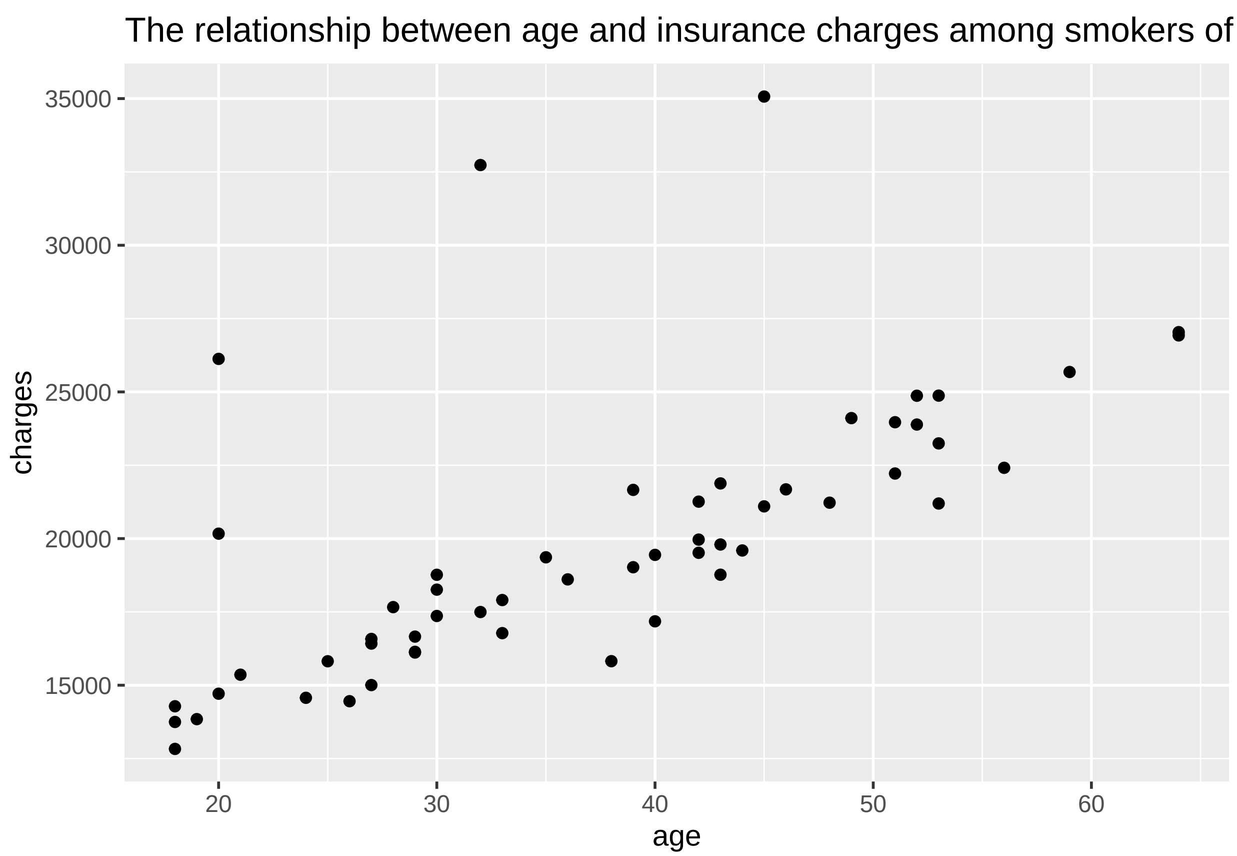
```
## # A tibble: 55 x 8
##      age sex      bmi children smoker region    charges bmi_cat
##    <dbl> <chr>  <dbl>    <dbl> <chr>  <chr>       <dbl> <chr>
##  1    53 female  22.9        1 yes    southeast  23245. Normal
##  2    20 female  22.4        0 yes    northwest  14712. Normal
##  3    28 male    24.0        3 yes    southeast  17663. Normal
##  4    27 female  24.8        0 yes    southeast  16578. Normal
##  5    45 male    22.9        2 yes    northwest  21099. Normal
##  6    56 male    20.0        0 yes    northeast  22413. Normal
##  7    38 male    19.3        0 yes    southwest  15821. Normal
##  8    32 female  17.8        2 yes    northwest  32734. Normal
##  9    42 female  23.4        0 yes    northeast  19965. Normal
## 10    48 male    24.4        0 yes    southeast  21224. Normal
## # i 45 more rows
```

```
. = ottr::check("tests/p10.R")
```

```
## All tests passed!
```

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 <- ggplot(insure_subset, aes(x = age, y = charges)) + geom_point() + labs(title = "The relationship
p11
```




```
. = ottr::check("tests/p11.R")
```

```
## All tests passed!
```

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 object `insure_mod` and uncomment the line of code below the model to tidy the output.**

```r
insure_model <- lm(formula = charges ~ age, data = insure_subset) # SOLUTION
tidy(insure_model)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic       p.value
##   <chr>          <dbl>     <dbl>     <dbl>         <dbl>
## 1 (Intercept)   10656.    1471.       7.24 0.00000000184
## 2 age             246.      37.4      6.58 0.0000000217
```

```r
. = ottr::check("tests/p12.R")
```

```
## All tests passed!
```

13. [**1 point**] **Interpret the slope parameter in the context of this problem.**

Type your answer here, replacing this text.

For every year increase in age, medical charges increase by $246.14. **14.** [**1 point**] **Interpret the intercept parameter.**

Type your answer here, replacing this text.

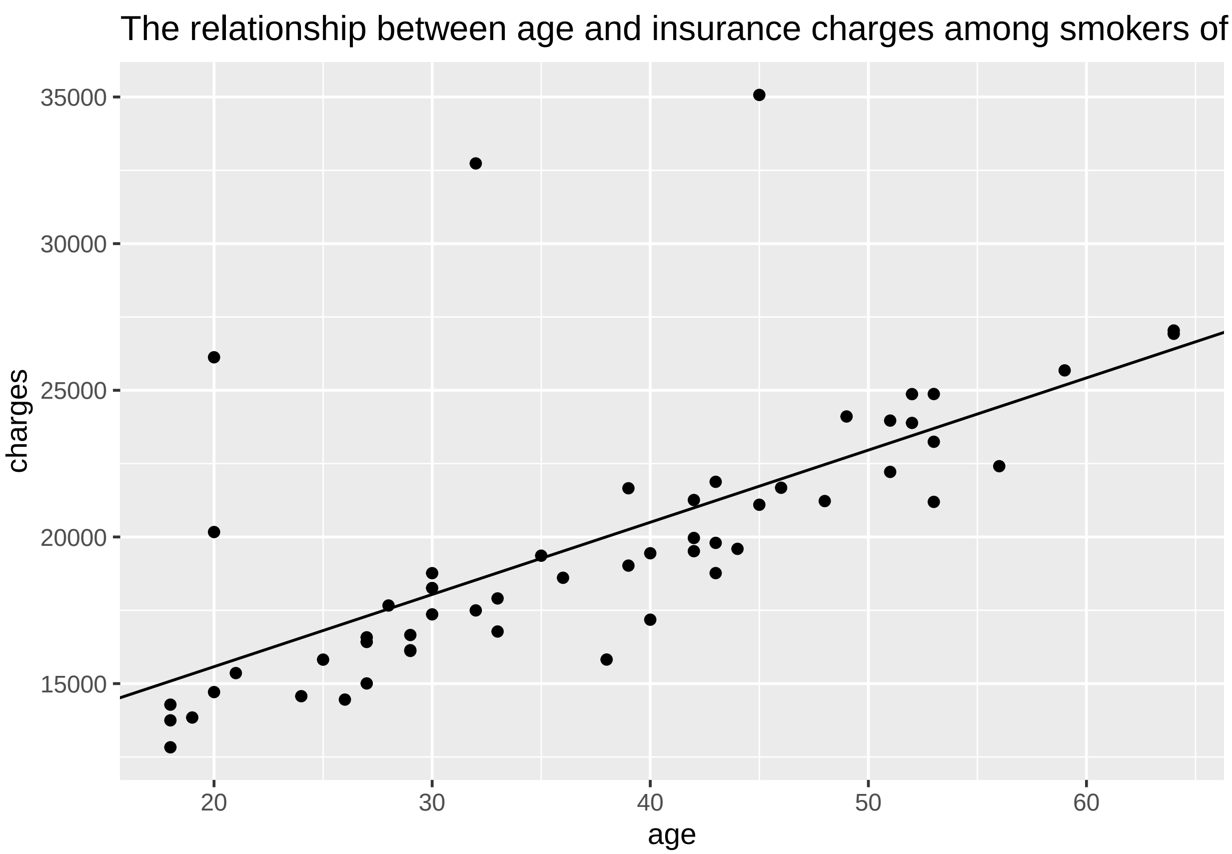
The model predicts that the insurance charged would be $10,656.14 for a person of aged 0. **15.** [**1 point**] **Does the intercept make sense in this context?**

Type your answer here, replacing this text.

No because being 0 years old is non sensical. Further, the minimum age in the dataset is 18, so extrapolation to 0 is not supported by the data. (student can say either of these items or both.)

16. [1 point] Add the line of best fit to your scatterplot by copying and pasting the plot's code from question 11 in the chunk below and adding a `geom` that can be used to add a regression line.

```
p16 <- ggplot(insure_subset, aes(x = age, y = charges)) + geom_point() + labs(title = "The relationship
p16
```




```
. = ottr::check("tests/p16.R")
```

```
## All tests passed!
```

17. [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?

Type your answer here, replacing this text.

The line seems high. There is a large proportion of points below the line. That's because there exists some notable outliers above the line which don't follow the linear trend of the data points.

Run the following `filter()` function in the chunk below.

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

18. [2 points] How many individuals were removed? Who were they?

Type your answer here, replacing this text.

Three individuals were removed. They were the "y outliers", the two people with the highest charges in the dataset and a third person who was 20 years old with a charge > $25,000.

19. [2 points] Run a regression model on `insure_smaller_subset` between `charges` and `age`. Assign the model to `insure_better_model` and analyze the output using the `tidy()` function.

```r
insure_better_model <- lm(formula = charges ~ age, data = insure_smaller_subset) # SOLUTION
tidy(insure_better_model)
```

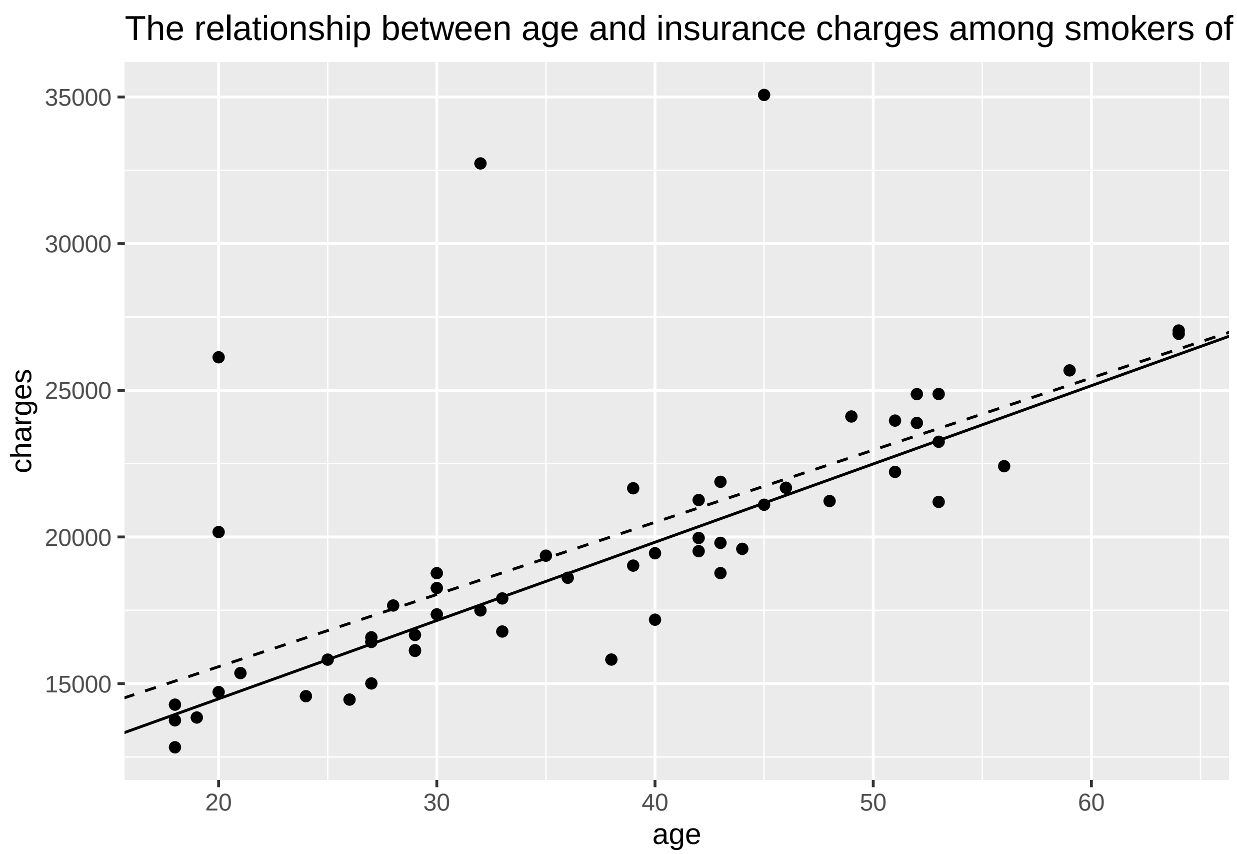
```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    9144.     633.       14.4 1.81e-19
## 2 age             267.      16.0      16.7 4.44e-22
```

```r
. = ottr::check("tests/p19.R")
```

```
## All tests passed!
```

20. [2 points] **Add the new regression line to your ggplot from question 16. Keep the original regression line on the plot for comparison. To distinguish the lines, change the color, line type, or line width of one of the lines.**

```
p20 <- ggplot(insure_subset, aes(x = age, y = charges)) + geom_point() + labs(title = "The relationship
p20
```




```
. = ottr::check("tests/p20.R")
```

```
## All tests passed!
```

21. [1 point] Calculate the r-squared value for `insure_model` and assign this value to `insure_model_r2`.

```r
insure_model_r2 <- glance(insure_model) %>% pull(r.squared) # SOLUTION
insure_model_r2
```

```
## [1] 0.449261
```

```r
. = ottr::check("tests/p21.R")
```

```
## All tests passed!
```

22. [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`.

```r
insure_better_model_r2 <- glance(insure_better_model) %>% pull(r.squared) # SOLUTION
insure_better_model_r2
```

```
## [1] 0.8477642
```

```r
. = ottr::check("tests/p22.R")
```

```
## All tests passed!
```

23. [2 points] **Calculate the correlation coefficient between `age` and `charges` using `insure_subset`. Also calculate the squared correlation coefficient. You should use `summarize()` to create a dataframe of these two values and name the two variables `corr` and `corr_sq`, respectively. What do you notice about the relationship between the correlation coefficient and r-squared values that you calculated earlier?**

```
p23 <- insure_subset %>% summarize(corr = cor(age, charges), corr_sq = corr^2) # SOLUTION
p23
```

```
## # A tibble: 1 x 2
##    corr corr_sq
##   <dbl>   <dbl>
## 1 0.670   0.449
```

```
. = ottr::check("tests/p23.R")
```

```
## All tests passed!
```

24. [2 points] **Calculate the correlation coefficient between `age` and `charges` using the smaller dataset `insure_smaller_subset`. Also calculate the squared correlation coefficient. You should use `summarize()` to create a dataframe of these two values and name the two variables `corr` and `corr_sq`, respectively. What do you notice about the relationship between the correlation coefficient and r-squared values that you calculated earlier?**

```
p24 <- insure_smaller_subset %>% summarize(corr = cor(age, charges), corr_sq = corr^2) # SOLUTION
p24
```

```
## # A tibble: 1 x 2
##    corr corr_sq
##   <dbl>   <dbl>
## 1 0.921   0.848
```

```
. = ottr::check("tests/p24.R")
```

```
## All tests passed!
```

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.

25. [1 point] Make a new dataframe called `insure_smokers` that includes smokers of any BMI from the original `insure_data` dataset.

```
insure_smokers <- insure_data %>% filter(smoker == "yes") # SOLUTION
insure_smokers
```

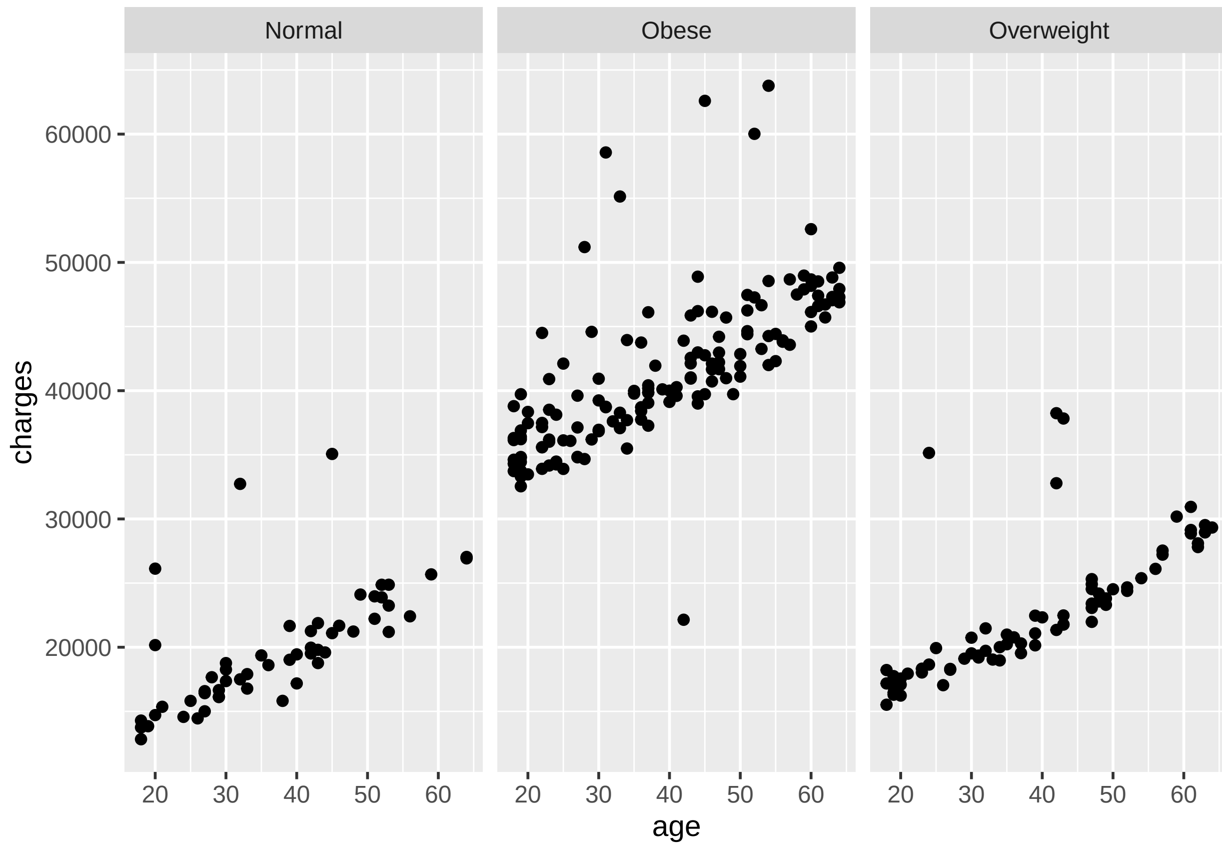
```
## # A tibble: 274 x 8
##      age sex      bmi children smoker region    charges bmi_cat
##    <dbl> <chr>  <dbl>    <dbl> <chr>  <chr>       <dbl> <chr>
##  1    19 female  27.9        0 yes    southwest  16885. Overweight
##  2    62 female  26.3        0 yes    southeast  27809. Overweight
##  3    27 male    42.1        0 yes    southeast  39612. Obese
##  4    30 male    35.3        0 yes    southwest  36837. Obese
##  5    34 female  31.9        1 yes    northeast  37702. Obese
##  6    31 male    36.3        2 yes    southwest  38711  Obese
##  7    22 male    35.6        0 yes    southwest  35586. Obese
##  8    28 male    36.4        1 yes    southwest  51195. Obese
##  9    35 male    36.7        1 yes    northeast  39774. Obese
## 10    60 male    39.9        0 yes    southwest  48173. Obese
## # i 264 more rows
```

```
. = ottr::check("tests/p25.R")
```

```
## All tests passed!
```

26. [1 point] Make a scatterplot that examines the relationship between `age` and `charges` for normal, overweight, and obese individuals in three side by side plots. A `facet_` command may help you.

```
p26 <- ggplot(insure_smokers, aes(x = age, y = charges)) + geom_point() + facet_wrap(~ bmi_cat) # SOLUT
p26
```




```
. = ottr::check("tests/p26.R")
```

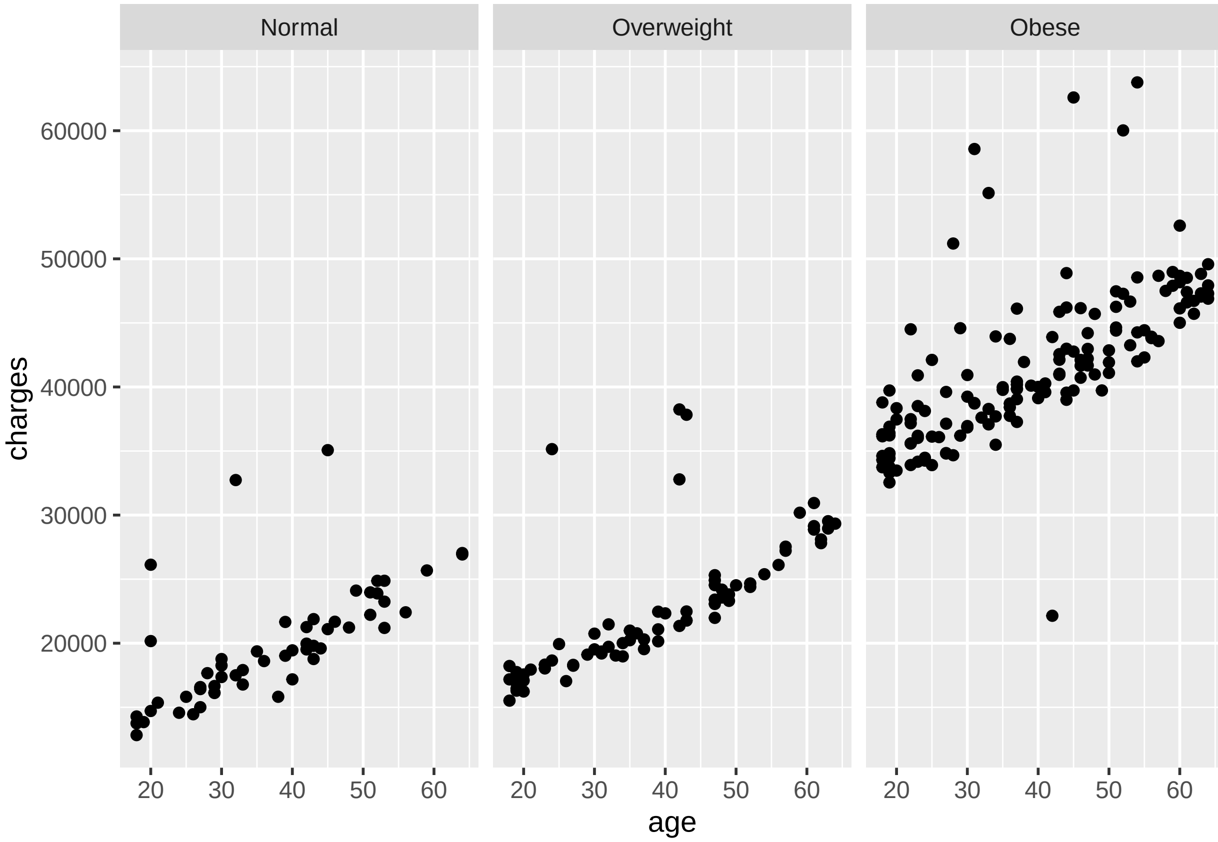
```
## All tests passed!
```

The plot above automatically displays the BMI categories alphabetically. Run the chunk below to assign a different order to the values of `bmi_cat`.

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

27. [1 point] Re-run your code from question 26, but facet using `bmi_cat_ordered`.

```
p27 <- ggplot(insure_smokers, aes(x = age, y = charges)) + geom_point() + facet_wrap(~bmi_cat_ordered) +
p27
```




```
. = ottr::check("tests/p27.R")
```

```
## All tests passed!
```

28. [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. Call your models `normal_mod`, `overweight_mod`, `obese_mod`. Use the `tidy()` function to display the output from each model.

```
insure_smokers_normal <- insure_smokers %>% filter(bmi_cat == "Normal") # SOLUTION
insure_smokers_overweight <- insure_smokers %>% filter(bmi_cat == "Overweight") # SOLUTION
insure_smokers_obese <- insure_smokers %>% filter(bmi_cat == "Obese") # SOLUTION

normal_mod <- lm(charges ~ age, data = insure_smokers_normal) # SOLUTION
overweight_mod <- lm(charges ~ age, data = insure_smokers_overweight) # SOLUTION
obese_mod <- lm(charges ~ age, data = insure_smokers_obese) # SOLUTION
tidy(normal_mod)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic       p.value
##   <chr>          <dbl>     <dbl>     <dbl>         <dbl>
## 1 (Intercept)   10656.    1471.       7.24 0.00000000184
## 2 age             246.      37.4      6.58 0.0000000217
```

```
tidy(overweight_mod)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   12400.    1176.      10.5  3.01e-16
## 2 age             264.      28.9      9.16 1.07e-13
```

```
tidy(obese_mod)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   30558.    1093.       28.0 7.96e-60
## 2 age             281.      26.2      10.7 5.05e-20
```

```
. = ottr::check("tests/p28.R")
```

```
## All tests passed!
```

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 models. Show your work for each calculation.

29. [1 point] **Predict the medical charges for a 20 year old with a normal BMI.**

```r
p29 <- 10656.1 + 246.1 * 20 # SOLUTION
# = $15578.1 # SOLUTION
p29
```

```
## [1] 15578.1
```

```r
. = ottr::check("tests/p29.R")
```

```
## All tests passed!
```

30. [1 point] **Predict the medical charges for a 20 year old with an overweight BMI.**

```r
p30 <- 12399.7 + 264.2 * 20 # SOLUTION
# = $17683.7 # SOLUTION
p30
```

```
## [1] 17683.7
```

```r
. = ottr::check("tests/p30.R")
```

```
## All tests passed!
```

31. [1 point] **Predict the medical charges for a 20 year old with an obese BMI.**

```r
p31 <- 30558.1 + 281.2 * 20 # SOLUTION
# = $36182.1 # SOLUTION
p31
```

```
## [1] 36182.1
```

```r
. = ottr::check("tests/p31.R")
```

```
## All tests passed!
```

32. [3 points] In three sentences maximum, comment on (1) the direction of the association, (2) how much the slopes vary across the BMI groups, and (3) how much the predicted medical charges for a 20-year old varies by BMI category.

Type your answer here, replacing this text.

There was a positive association between age and medical charges for normal, overweight, and obese individuals. The relationship was of similar magnitude for each BMI group, though the slope increased in magnitude for overweight and obese individuals, implying that a steeper relationship for overweight individuals, and even steeper for obese individuals vs. normal BMI individuals. For a given age, obese individuals had much higher charges than overweight and normal weight individuals. ## END