Correlation and Regression

[Put your name here]

In this assignment we will learn how to run correlation and regression analyses. Correlation measures the strength of the linear relationship between two numerical variables; regression provides a model for that linear relationship.

Instructions

Presumably, you have already created a new project and downloaded this file into it. Please knit the document (probably to HTML as you're working) and work back and forth between this R Markdown file and the knit output to answer the following questions.

When you are finished with the assignment, knit to PDF, export the PDF file to your computer, and then submit to the corresponding assignment in Canvas. Be sure to submit before the deadline!

Sometimes you will be asked to add your own R code. That will appear in this document as a code chunk with a request for you to add your own code, like so:

```
## Add code here to [do some task]...
```

When you see that in a code chunk, you need to type some R code to complete a task.

Sometimes you will be asked to type up your thoughts. Instructions to do that will be labeled as follows. If you are currently reading this in the knit output, please look back at the R Markdown file to see the following text:

(When you knit the document, you can't see the text from the line above. That's because the crazy notation surrounding that text marks it as a "comment", and therefore it doesn't appear in the output.) In these areas of the assignment, please use contextually meaningful full sentences/paragraphs (unless otherwise indicated) and proper spelling, grammar, punctuation, etc. This is not R code (although it may include inline R code when necessary), but rather a free response section where you talk about your analysis and conclusions.

Getting started

Make sure you're in a project

If you're looking at this document, you should have already created a project and uploaded this R Markdown file to that project folder.

Save your file!

The first thing we **always** do is save our file. You'll probably want to save this under a new name. Go to the "File" menu and then "Save As". Once you've saved the file with the new name, from them on it's easier to just hit Ctrl-S (or Cmd-S on a Mac) to keep saving it periodically.

Remember that file names should not have any spaces in them. (In fact, you should avoid other kinds of special characters as well, like periods, commas, number signs, etc. Stick to letters and numerals and you should be just fine.) If you want a multiword file name, I recommend using underscores like this: `this_filename_has_spaces_in_it`.

Load Packages

We load the standard `mosaic` package as well as the `reshape2` package for the `tips`, the `OIdata` package for the `state` data, and the `MASS` package for the `Rubber` data.

```
library(mosaic)
library(reshape2)
library(OIdata)
data(state)
library(MASS)
```

We also want to get rid of scientific notation:

```
options(scipen = 10)
```

Correlation

The word correlation describes a linear relationship between two numerical variables. As long as certain conditions are met, we can calculate a statistic called the Pearson correlation coefficient, usually denoted R. This value will be some number between -1 and 1. Coefficients close to zero indicate little or no correlation, coefficients close to 1 indicate strong positive correlations, and coefficients close to -1 indicate strong negative correlations. In between, we often use words like weak, moderately weak, moderate, and moderately strong. There are no exact cutoffs for when such words apply. You must learn from experience how to judge scatterplots and R values to make such determinations.

Let's examine a data set called `tips`. (Be careful! There is a data set called `tips` in the `openintro` package as well. It won't be a problem in this assignment since the `openintro` package isn't used here in the R Markdown document. However, in RStudio, you may have loaded the `openintro` package in a previous session. Make sure you are using the `reshape2` package version!) These 244 observations were collected by one waiter over a period of a few months working in a restaurant. One simple (and somewhat obvious) question we can ask is if there is a correlation between the size of a bill and the size of the tip.

If all we wanted was the value of R, we can find it by using the `cor` command:

```
cor(tips$total_bill, tips$tip)
```

```
## [1] 0.6757341
```

However, without checking conditions, this is very dangerous. R will gladly compute the correlation coefficient for any data, whether appropriate or not. Therefore, we will follow our inferential rubric to decide if there is a statistically significant relationship between the bill and the corresponding tip. In truth, the entire inferential rubric is probably overkill for such a simple question. Nevertheless, the rubric does ensure that we take care to identify our hypotheses and check conditions.

Hypotheses

Identify the sample and a reasonable population of interest.

The sample consists of 244 meals a waiter served over the course of several months working at a restaurant. The population is presumably all meals this waiter might ever serve at this restaurant. (It would not make sense to include other waiters or other restaurants in this population as bills and tips vary widely from person to person and restaurant to restaurant.)

Express the null and alternative hypotheses as contextually meaningful full sentences.

H_0: There is no correlation between the total bill and the tip.

H_A: There is a correlation between the total bill and the tip.

Express the null and alternative hypotheses in symbols.

The sample correlation coefficient is called R, but this is an estimate of the true correlation coefficient that would be obtained if we had access to the entire population. This population correlation coefficient is called ρ.

$H_0 : \rho = 0$

$H_A : \rho \neq 0$.

Keep in mind that although we are performing a two-sided test here, one could perform a one-sided test if the question of interest was about a positive or a negative correlation specifically.

Model

Identify the correct sampling distribution model.

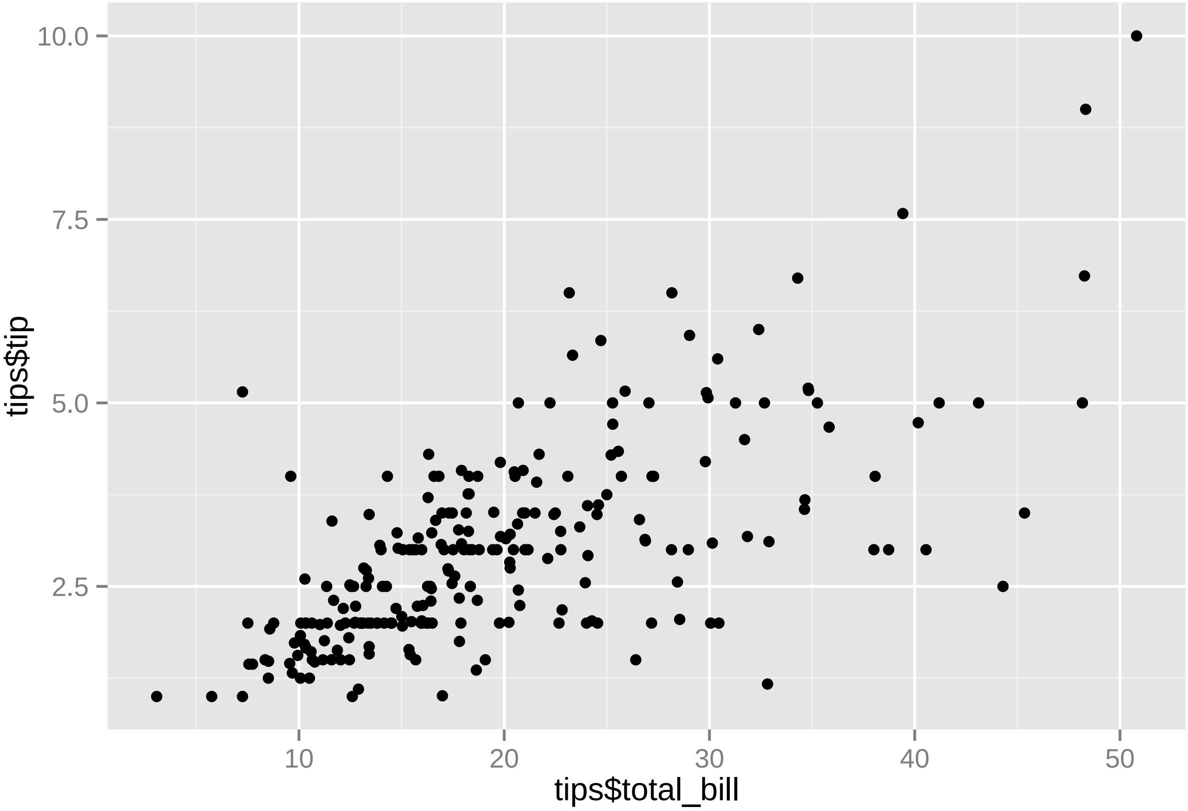
It turns out that sample correlation coefficients follow a t model with $n - 2$ degrees of freedom. Since there are 244 meals here, there are 242 degrees of freedom.

Check the relevant conditions to ensure that the model assumptions are met.

In addition to the standard "Random" and "10%", we introduce three new conditions. First, we need to make sure we have two numerical variables. The Pearson correlation coefficient makes no sense if the data is not numerical. Next, we need to know that the relationship is linear. Nonlinear associations can exist, but, again, the R value makes no sense for such relationship. Finally, we need to check for outliers. The last two conditions mentioned here should be checked by looking at a scatterplot.

- Random
    - This is not a random sample, but over several months, it seems reasonable that this is representative of this waiter's experiences at this restaurant.
- 10%
    - Assuming the waiter works at this restaurant for several year, 244 meals is probably less than 10% of all meals he will serve.
- Two numerical variables
    - `total_bill` and `tip` are numerical variables.
- "Straight enough"
    - Consider the scatterplot below:

```r
qplot(tips$total_bill, tips$tip)
```



No data will ever line up in a perfect straight line. The "straight enough" condition is meant to suggest that the "cloud of dots" should be more or less in a straight pattern moving across the plot. We are most concerned here with checking that the pattern does not curve substantially, and this does not appear to.

- Outliers
  - We don't see any significant outliers. There are a few dots here and there that are a little far from the main cloud, but nothing that worries us too much, especially given the large sample size.

Mechanics

Compute the test statistic.

Just for your edification, the formula for the test statistic is

$$t = \frac{R - \rho}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{R}{\sqrt{\frac{1-r^2}{n-2}}}.$$

(The last step takes into account the fact that the null value for ρ is zero.)

We'll never have to do this by hand. Here is the R code that computes our test.

```
test_cor <- cor.test(tips$total_bill, tips$tip)
t <- test_cor$statistic
t
```

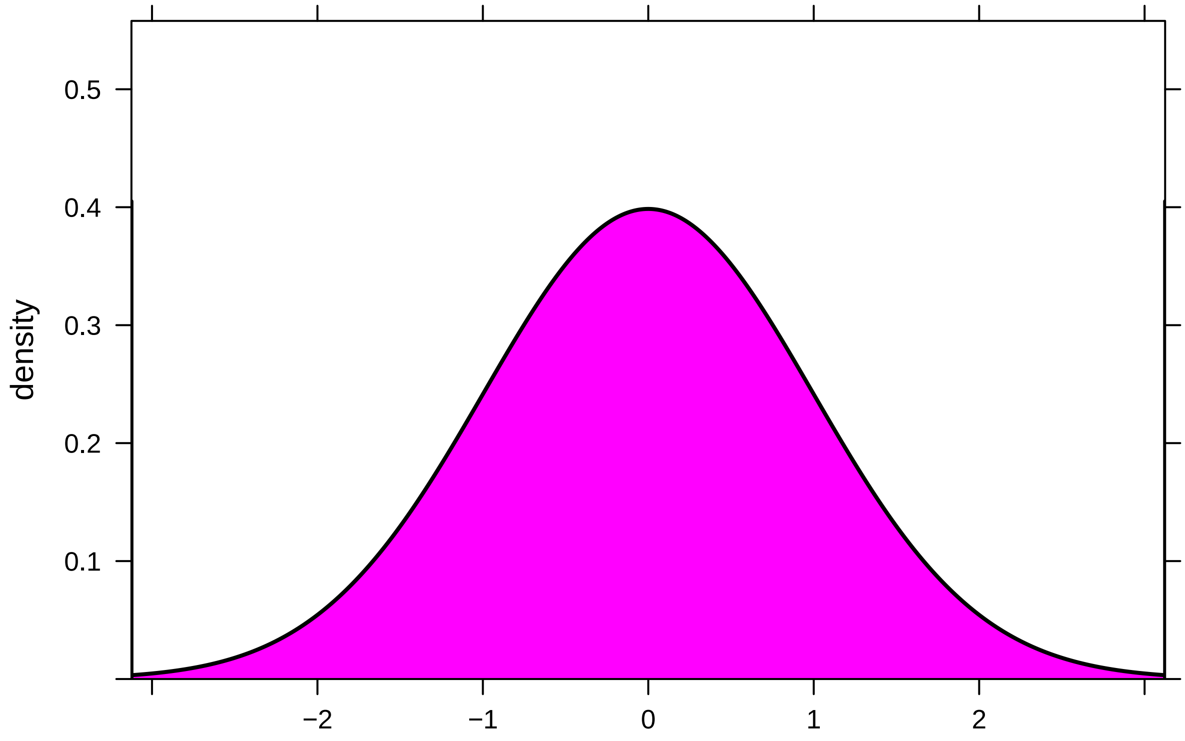
```
##        t
## 14.26035
```

Be aware that when the correlation is "obvious" to the eye, that usually means that R is far from zero, so we will likely see t scores here much, much larger than we're used to.

Plot the null distribution.

This step is fairly pointless as the t score is so crazy large, but we include it for completeness

```
pdist(dist = "t", df = test_cor$parameter, q = c(-t, t))
```



```
##            t            t
## 3.346235e-34 1.000000e+00
```

Calculate the P-value.

```
test_cor$p.value
```

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

Keep in mind that a P-value can never truly be zero. We will report this as $P < 0.001$

Conclusion

State the statistical conclusion.

We reject the null hypothesis.

State (but do not overstate) a contextually meaningful conclusion.

There is sufficient evidence to suggest that there is a correlation between the total bill and the tip (for this waiter at this restaurant).

Identify the possibility of either a Type I or Type II error and state what making such an error means in the context of the hypotheses.

There is the possibility of a Type I error. In that case, there would be no correlation between the total bill and the tip, even though we found such a correlation in our sample.

Confidence interval

Conditions

All relevant conditions have already been checked.

Calculation

```
test_cor$conf.int
```

```
## [1] 0.6011647 0.7386372
## attr(,"conf.level")
## [1] 0.95
```

Conclusion

We are 95% confident that the true correlation between the total bill and the tip (for this waiter at this restaurant) is captured in the interval (0.6011647, 0.7386372).

Your turn!

The `state` data from the `OIdata` package has a number of variables collected from various sources. There are 51 rows, representing the 50 states and the District of Columbia. Run a correlation test to determine if the median household income in each state is correlated with the percentage of the state's population that smokes.

There is something unusual about this example that you will need to consider in your answer. The sample is 50 states and the District of Columbia. The population is tricky though because these states don't represent some larger groups of states; we already have all the states in our data. One can think of this data, though, as a snapshot of what was true in each state at one point in time. Therefore, the population can be thought of as similar measurements taken at other times.

This also makes it difficult to check conditions. We do not have a random sample of states (as we have all of them), but remember that we're thinking of this as a random sample across a number of years in which we might have gathered this data. Having said that, I imagine that median income goes up every year with inflation, so it may or may not be representative of other years. The 10% condition also requires some thought.

Regression

When we have a linear relationship between two numerical variables, it's helpful to model this relationship with an actual straight line. Such a line is called a regression line, or a best-fit line, or sometimes a least-squares line.

The mathematics involved in figuring out what this line should be is more complicated than we cover in this course. But R will do all the complicated calculations for us.

Eventually, we will consider inference on the slope of the regression line. But let's take a moment to talk about some of the other aspects of regression that we can analyze and understand.

Since you've already done a correlation analysis on the `state` data, it makes sense to follow up with a regression analysis.

Getting the regression information from R

To run a regression analysis, we run the `lm` command. We should not trust this output until we have checked conditions below. However, we need some of this output in order to check the conditions, so we'll run the command now.

```
income_smoke <- lm(state$smoke ~ state$med_income)
```

Again, I want to emphasize that until the conditions are checked, we should not use the regression line in any way. In particular, we will avoid graphing the line until after we know that is okay to do so.

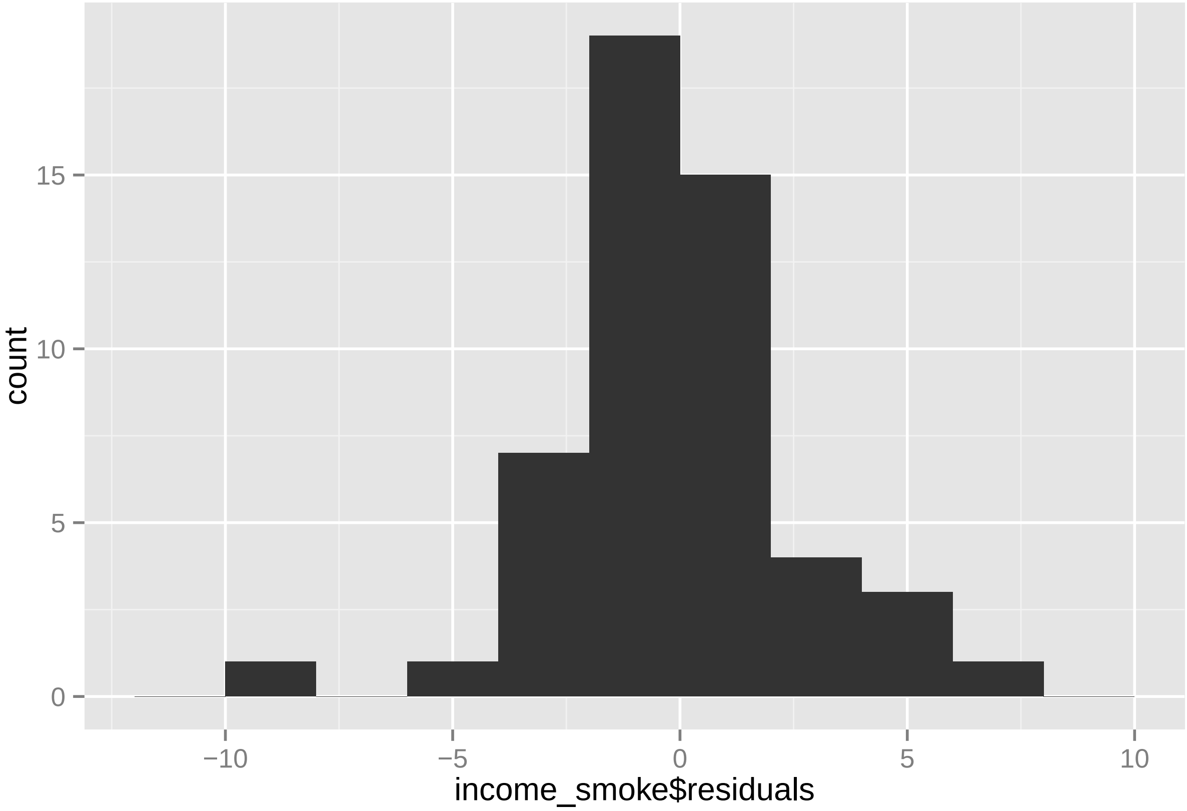
Conditions

The conditions for running a regression analysis include all the conditions you checked above for a correlation analysis. However, there is one important additional condition to check to ensure that our regression model is appropriate. I call this the "patterns in the residuals" condition. Your textbook breaks this up into a bunch of smaller conditions, but they all deal with analyzing residuals.

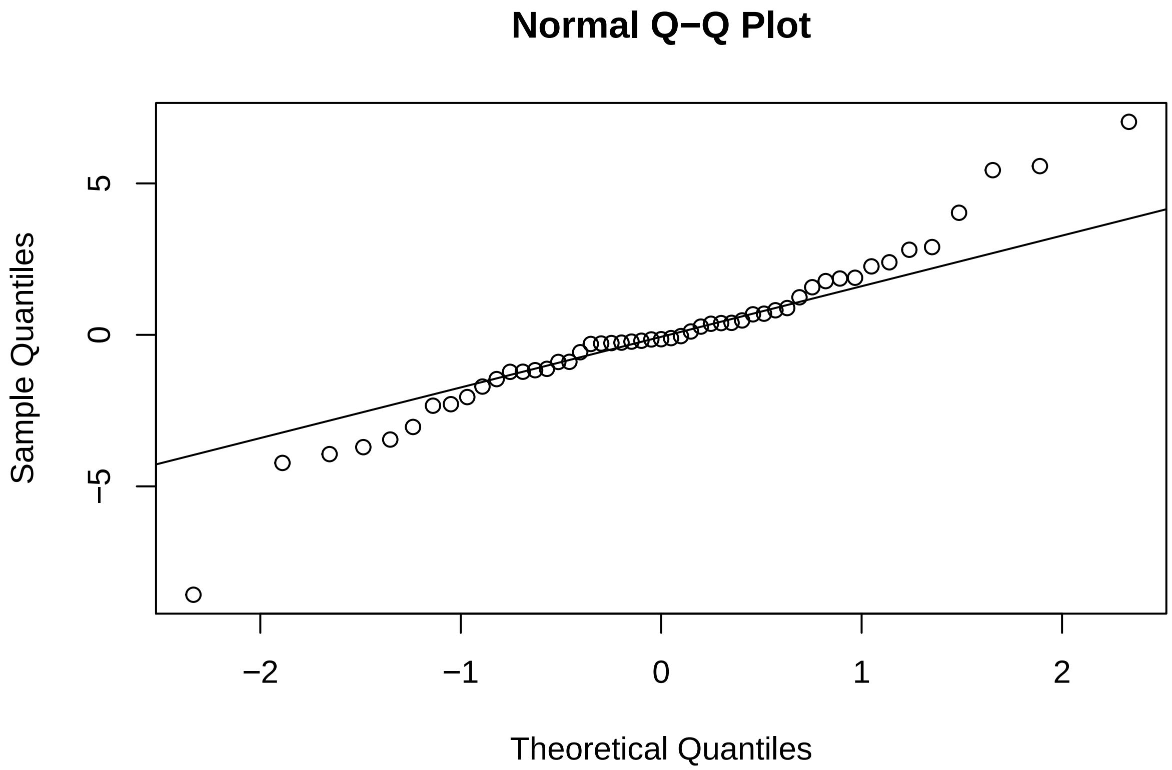
Recall that residuals are the distances from each data point to the regression line. We know that some of the points are going to lie above the line (positive residuals) and some of the points will lie below the line (negative residuals). And we also know that there will be about the same number of points above and below the line. What we need is for there not to be any pattern among these residuals.

To check for such patterns in the residuals, we use several different plots. First, we want our residuals to be normally distributed. We check this with a histogram and a QQ-plot, as always.

```
qplot(income_smoke$residuals, binwidth = 2)
```



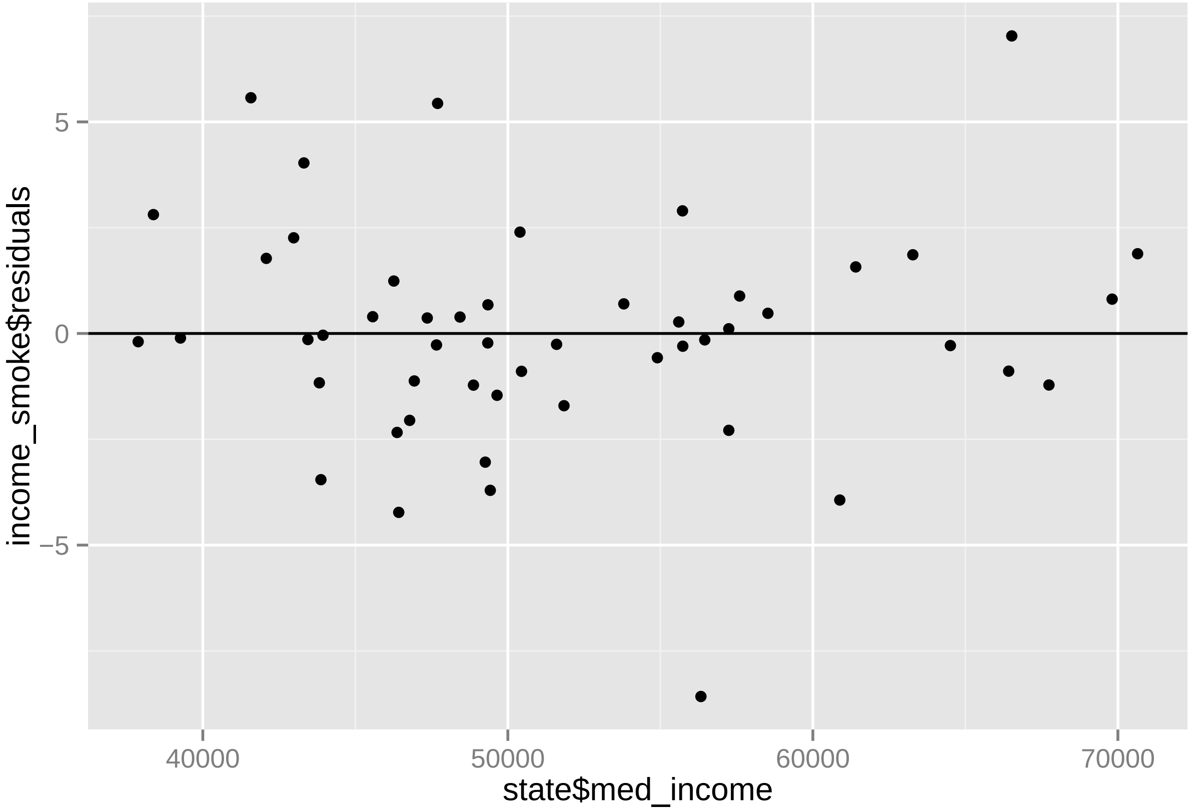
```
qqnorm(income_smoke$residuals)
qqline(income_smoke$residuals)
```



We see that the shape is mostly normal. The tails are not quite shaped the way we expect (they appear to be too thin, especially on the left where there is an outlier), but it's not an extreme issue.

We should also create a *residual plot*, which looks at the residuals above each value along the x-axis. (In the command below, we also add a horizontal reference line so that it is clear which points have positive or negative residuals.)

```r
qplot(state$med_income, income_smoke$residuals) + geom_hline(y = 0)
```

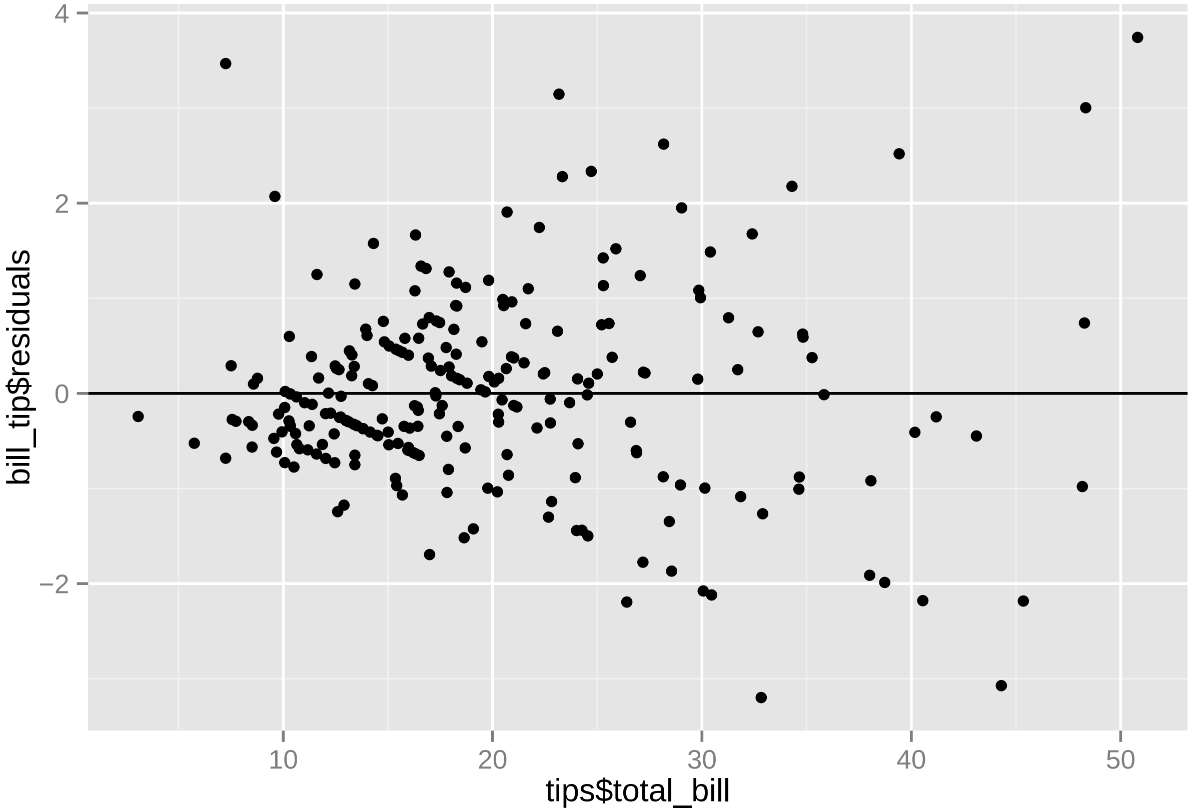


Other than the one outlier, this looks good. There is no systematic patterns in the residuals. A residual plot should look like the most boring plot you've ever seen.

Residual patterns that are problematic often involve curved data (where the dots follow a curve around the horizontal reference line instead of spreading evenly around it) and heteroskedasticity, which is a fanning out pattern.

As an example of the latter, let's look at the residual plot for the tipping example from the beginning of this assignment.

```r
bill_tip <- lm(tips$tip ~ tips$total_bill)
qplot(tips$total_bill, bill_tip$residuals) + geom_hline(y = 0)
```

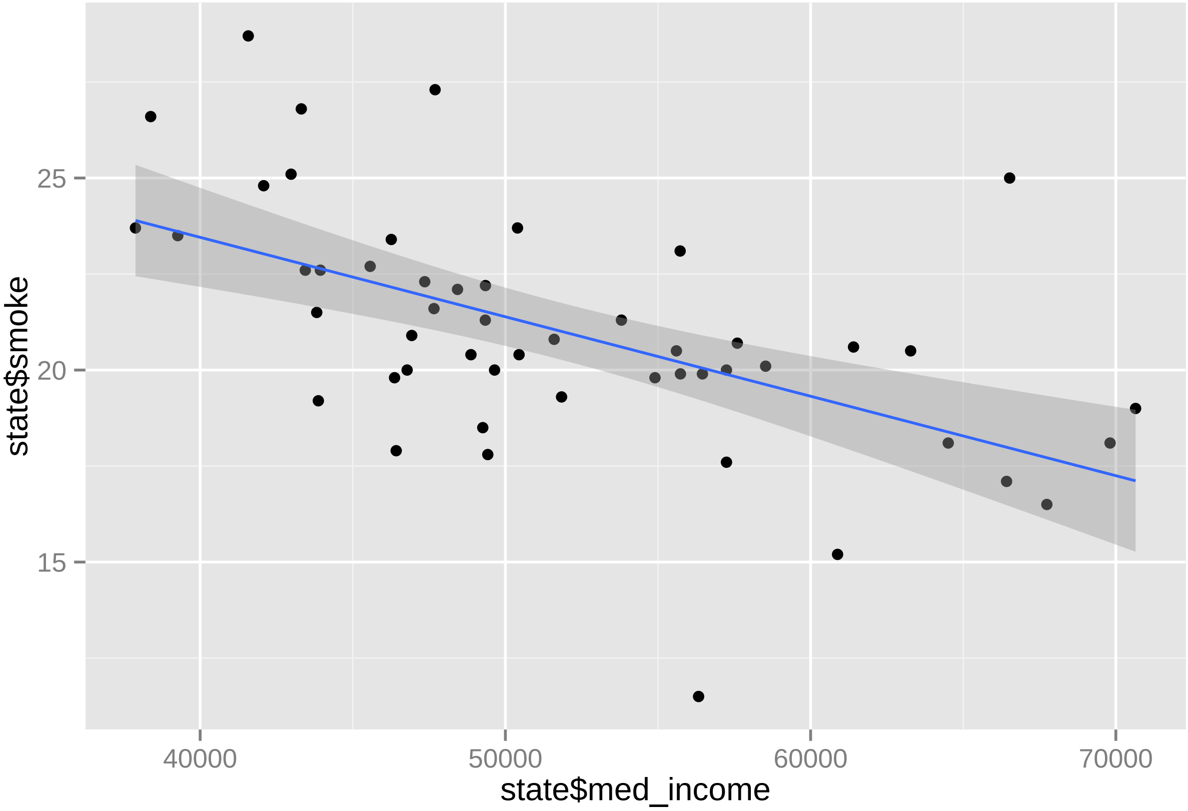


The residuals are quite small toward the left end of the residual plot, and then spread out and get larger toward the right end. This is a violation of the "patterns in the residuals" condition, and is the reason why we are not pursuing a regression analysis of the tip data here.

Plotting the regression line

Now that all the conditions are met, we can look at the regression line. We superimpose it on the scatterplot you produced above.

```
qplot(state$med_income, state$smoke) + geom_smooth(method = "lm")
```



The blue line is the regression line. The shaded region around the regression line is called a confidence band. It's like a 95% confidence interval for where we hope the "true" population regression line might lie.

Interpreting the coefficients.

The output of the `lm` command gives us the slope and intercept for the model.

```
income_smoke
```

```
##
## Call:
## lm(formula = state$smoke ~ state$med_income)
##
## Coefficients:
##      (Intercept)  state$med_income
##       31.7310844        -0.0002069
```

For inline code, these can be extracted with `income_smoke$coefficients`.

The intercept is 31.7310844 and the slope is -0.0002069.

The slope is always interpretable. The model predicts that one unit of increase in the x-axis corresponds to a change of -0.0002069 units in the y-direction (in other words, a decrease). Of course, one unit on the x-axis is a dollar, which isn't very meaningful on the scale of the data. So let's phrase it this way:

> The model predicts that an increase of $10,000 in median income corresponds to a decrease of 2.0687955 percentage points in the proportion of the state's population that smokes.

The intercept is a different story. There is always a literal interpretation:

> The model predicts that a state with median income $0 will have 31.7310844 percent smokers.

It is true that the model makes that prediction, but that prediction is nonsensical. Aside from the fact that it is impossible for the median income of a state to be $0, this is extrapolation—in other words, a prediction outside the range of the data.

The regression equation

When we report the equation of the regression line, we typically use words instead of x and y to make the equation more interpretable in the context of the problem. For example, for this data, we would write the equation as

$$\widehat{smoke} = 31.73 - 0.0002income$$

Interpreting R^2

The correlation coefficient R is of limited utility. The number doesn't have any kind of intrinsic meaning; it can only be judged by how close it is to zero or one in conjunction with a scatterplot to give you a sense of the strength of the correlation. In particular, some people try to interpret R as some kind of percentage, but it's not.

On the other hand, R^2 can be interpreted as a percentage. It represents the percent of variation in the y variable that can be explained by variation in the x variable.

The value of R^2 is not recorded in the output of the `lm` command directly. Instead, it appears in the `summary` of the `lm` output. We will store the summary information and then grab the R^2 value.

```
income_smoke_summ <- summary(income_smoke)
income_smoke_summ$r.squared
```

```
## [1] 0.3050746
```

We will word it this way:

> 30.5074627% of the variability in the percentage of smokers in a state can be explained by variability in the median income.

Thus, R^2 is a measure of the fit of the model. High values of R^2 mean that the line predicts the data values closely, whereas lower values of R^2 mean that there is still a lot of variability left in the residuals (presumably due to other factors that are not measured here).

Inference for the regression slope

We have already given an interpretation for the slope of the regression line. As with correlation, inference for the regression slope is a little overkill. Nevertheless, the rubric forces us to be careful to identify our sample and population, check conditions, and state conclusions.

Hypotheses

Identify the sample and a reasonable population of interest.

We discussed this above when you were introduced to this data set for your correlation analysis.

Express the null and alternative hypotheses as contextually meaningful full sentences.

H_0: There is no relationship between median income and the percentage of smokers in the state.

H_A: There is a relationship between median income and the percentage of smokers in the state.

Express the null and alternative hypotheses in symbols.

$H_0 : \beta_1 = 0$

$H_A : \beta_1 \neq 0$

Model

Identify the correct sampling distribution model.

The sampling distribution model for the slope is a t model with $n - 2$ degrees of freedom. So for this example, there are 49 degrees of freedom.

Check the relevant conditions to ensure that the model assumptions are met.

You checked all the conditions when you ran the correlation test except for the one new condition, the "patterns in the residual" condition, which we checked above as well.

Mechanics

Compute the test statistic.

There is a formula for the t score, but it's a little messy. We'll just let R tell us everything we need to know.

The summary we stored above as `income_smoke_summ` has a lot of information.

```
income_smoke_summ
```

```
##
## Call:
## lm(formula = state$smoke ~ state$med_income)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.5776 -1.1913 -0.1432  1.0623  7.0307
##
## Coefficients:
##                      Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)      31.73108444  2.33880701  13.567   < 2e-16 ***
## state$med_income -0.00020688  0.00004461  -4.638 0.0000264 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.636 on 49 degrees of freedom
## Multiple R-squared:  0.3051, Adjusted R-squared:  0.2909
## F-statistic: 21.51 on 1 and 49 DF,  p-value: 0.00002644
```

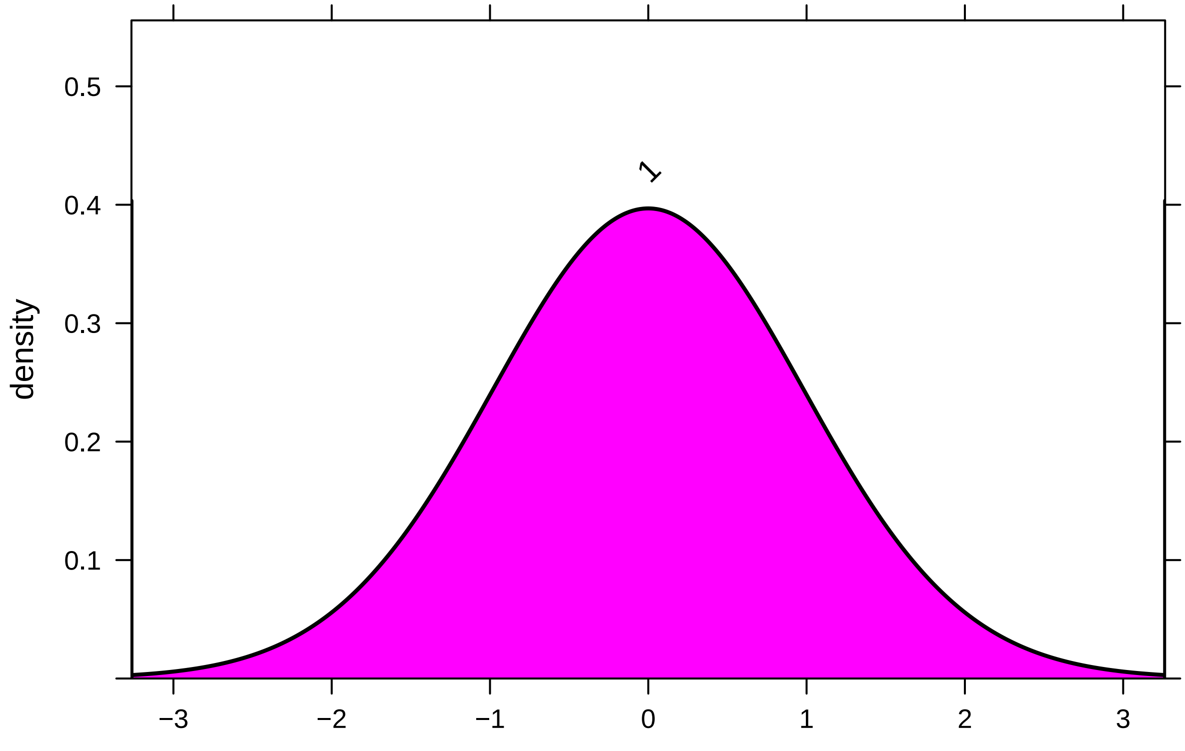
If we need to extract just the t score, it's a bit unwieldy:

```
t <- income_smoke_summ$coefficients["state$med_income",  "t value"]
t
```

```
## [1] -4.638013
```

Plot the null distribution.

```
pdist(dist = "t", df = income_smoke$df.residual, q = c(-t, t))
```



```
## [1] 0.99998678039 0.00001321961
```

Calculate the P-value.

This is also just as unwieldy as the t score, regrettably.

```
income_smoke_summm$coefficients["state$med_income",  "Pr(>|t|)"]
```

```
## [1] 0.00002643922
```

Conclusion

State the statistical conclusion.

We reject the null hypothesis.

State (but do not overstate) a contextually meaningful conclusion.

We have sufficient evidence to suggest that there is a relationship between median income in a state and the percentage of smokers in that state.

Identify the possibility of either a Type I or Type II error and state what making such an error means in the context of the hypotheses.

If we have made a Type I error, that means that there is actually no relationship between median income and smoking, but our sample shows such a relationship.

Confidence interval

Conditions

There are no additional conditions to check.

Calculation

Unlike our previous hypothesis tests, the confidence interval for the slope is not found in the output of the test. Instead, we use a special function `confint` that takes regression output and calculates a confidence interval for us.

```
confint(income_smoke)
```

```
##                          2.5 %       97.5 %
## (Intercept)      27.0310757828 36.431093103
## state$med_income -0.0002965171 -0.000117242
```

You can observe above that `confint` will calculate confidence intervals for the intercept and the slope. As we don't care about the intercept, we just need to grab the second line of this output.

```
confint(income_smoke)["state$med_income",]
```

```
##         2.5 %        97.5 %
## -0.0002965171 -0.0001172420
```

Conclusion

We are 95% confident that the true slope of the linear relationship between median income and smoking percentage is captured in the interval (-0.0002965, -0.0001172).

Your turn!

The `Rubber` data set contains data on the testing of tires. (Since it was a British study, they tested "tyres".)

Explore the relationship between the hardness of the tire (measured in something called Shore units—Google it if you want to know more) and the loss of tire material in an abrasion test (measured in grams per hour).

Your answer should walk through the steps in the same order as listed above:

- Run the `lm` command and store the regression output in a variable.
- Check conditions for regression (all of them, including the correlation conditions since you are not running a separate correlation analysis for this problem).
- Plot the regression line on the scatterplot.
- Interpret the coefficients: interpret the slope, give a literal interpretation of the intercept, and then comment on the appropriateness of that interpretation.
- Write the regression equation.
- Interpret R^2.
- Run the full rubric for inference on the slope parameter.