Chapter 6

Correlation and Simple Linear Regression

6.1 Relationships between two quantitative variables

The independence test in Chapter 5 provided a technique for assessing evidence of a relationship between two categorical variables. The terms ***relationship*** and ***association*** are synonyms that, in statistics, imply that particular values on one variable tend to occur more often with some other values of the other variable or that knowing something about the level of one variable provides information about the patterns of values on the other variable. These terms are not specific to the "form" of the relationship – any pattern (strong or weak, negative or positive, easily described or complicated) satisfy the definition. There are two other aspects to using these terms in a statistical context. First, they are not directional – an association between x and y is the same as saying there is an association between y and x. Second, they are not causal unless the levels of one of the variables are randomly assigned in an experimental context. We add to this terminology the idea of correlation between variables x and y. ***Correlation***, in most statistical contexts, is a measure of the specific type of relationship between the variables: the **linear relationship between two quantitative variables**[1]. So as we start to review these ideas from your previous statistics course, remember that associations and relationships are more general than correlations and it is possible to have no correlation where there is a strong relationship between variables. "Correlation" is used colloquially as a synonym for relationship but we will work to reserve it for its more specialized usage here to refer specifically to the linear relationship.

Assessing and then modeling relationships between quantitative variables drives the rest of the chapters, so we should get started with some motivating examples to start to think about what relationships between quantitative variables "look like"... To motivate these methods, we will start with a study of the effects of beer consumption on blood alcohol levels (BAC, in grams of alcohol per deciliter of blood). A group of $n = 16$ student volunteers at The Ohio State University drank a randomly assigned number of beers[2]. Thirty minutes later, a police officer measured their BAC. Your instincts, especially as well-educated college students with some chemistry knowledge, should inform you about the direction of this relationship – that there is a ***positive relationship*** between `Beers` and `BAC`. In other words, **higher values of one variable are associated with higher values of the other**. Similarly, lower values of one are associated with lower values of the other. In fact there are online calculators that tell you how much your BAC increases for each extra beer consumed (for example: http://www.craftbeer.com/beer-studies/blood-alcohol-content-

[1] There are measures of correlation between categorical variables but when statisticians say correlation they mean correlation of quantitative variables. If they are discussing correlations of other types, they will make that clear.

[2] Some of the details of this study have been lost, so we will assume that the subjects were randomly assigned and that a beer means a regular sized can of beer and that the beer was of regular strength. We don't know if any of that is actually true. It would be nice to repeat this study to know more details and possibly have a larger sample size but I doubt if our institutional review board would allow students to drink as much as 9 beers.

calculator if you plug in 1 beer). The increase in y (`BAC`) for a 1 unit increase in x (here, 1 more beer) is an example of a ***slope coefficient*** that is applicable if the relationship between the variables is linear and something that will be fundamental in what is called a ***simple linear regression model***. In a simple linear regression model (simple means that there is only one explanatory variable) the slope is the expected change in the mean response for a one unit increase in the explanatory variable. You could also use the *BAC* calculator and the models that we are going to develop to pick a total number of beers you will consume and get a predicted *BAC*, which employs the entire equation we will estimate.

Before we get to the specifics of this model and how we measure correlation, we should graphically explore the relationship between `Beers` and `BAC` in a scatterplot. Figure 6.1 shows a ***scatterplot*** of the results that display the expected positive relationship. Scatterplots display the response pairs for the two quantitative variables with the explanatory variable on the x-axis and the response variable on the y-axis. The relationship between `Beers` and `BAC` appears to be relatively linear but there is possibly more variability than one might expect. For example, for students consuming 5 beers, their *BAC*s range from 0.05 to 0.10. If you look at the online *BAC* calculators, you will see that other factors such as weight, sex, and beer percent alcohol can impact the results. We might also be interested in previous alcohol consumption. In Chapter 8, we will learn how to estimate the relationship between `Beers` and `BAC` after correcting or controlling for those "other variables" using ***multiple linear regression***, where we incorporate more than one quantitative explanatory variable into the linear model (somewhat like in the 2-Way ANOVA). Some of this variability might be hard or impossible to explain regardless of the other variables available and is considered unexplained variation and goes into the residual errors in our models, just like in the ANOVA models. To make scatterplots as in Figure 6.1, you could use the base R function `plot`, but we will want to again access the power of `ggplot2` so will use `geom_point` to add the points to the plot at the "x" and "y" coordinates that you provide in `aes(x = ..., y = ...)`.

```
library(readr)
BB <- read_csv("http://www.math.montana.edu/courses/s217/documents/beersbac.csv")
```

```
BB |> ggplot(mapping = aes(x = Beers, y = BAC)) +
  geom_point() +
  theme_bw()
```

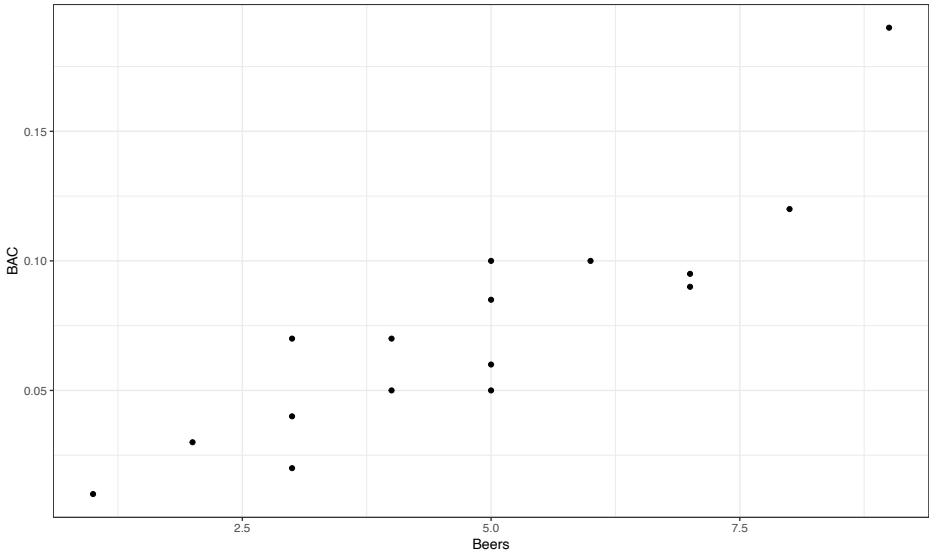

Figure 6.1: Scatterplot of *Beers* consumed versus *BAC*.

There are a few general things to look for in scatterplots:

1. **Assess the direction of the relationship** – is it positive or negative?
2. **Consider the strength of the relationship**. The general idea of assessing strength visually is about how hard or easy it is to see the pattern. If it is hard to see a pattern, then it is weak. If it is easy to see, then it is strong.
3. **Consider the linearity of the relationship**. Does it appear to curve or does it follow a relatively straight line? Curving relationships are called ***curvilinear*** or ***nonlinear*** and can be strong or weak just like linear relationships – it is all about how tightly the points follow the pattern you identify.
4. **Check for unusual observations – outliers** – by looking for points that don't follow the overall pattern. Being large in x or y doesn't mean that the point is an outlier. Being unusual relative to the overall pattern makes a point an outlier in this setting.
5. **Check for changing variability** in one variable based on values of the other variable. This will tie into a constant variance assumption later in the regression models.
6. **Finally, look for distinct groups** in the scatterplot. This might suggest that observations from two populations, say males and females, were combined but the relationship between the two quantitative variables might be different for the two groups.

Going back to Figure 6.1 it appears that there is a moderately strong linear relationship between `Beers` and `BAC` – not weak but with some variability around what appears to be a fairly clear to see straight-line relationship. There might even be a hint of a nonlinear relationship in the higher beer values. There are no clear outliers because the observation at 9 beers seems to be following the overall pattern fairly closely. There is little evidence of non-constant variance mainly because of the limited size of the data set – we'll check this with better plots later. And there are no clearly distinct groups in this plot, possibly because the # of beers was randomly assigned. These data have one more interesting feature to be noted – that subjects managed to consume 8 or 9 beers. This seems to be a large number. I have never been able to trace this data set to the original study so it is hard to know if (1) they had this study approved by a human subjects research review board to make sure it was "safe", (2) every subject in the study was able to consume their randomly assigned amount, and (3) whether subjects were asked to show up to the study with BACs of 0. We also don't know the exact alcohol concentration of the beer consumed or volume. So while this is a fun example to start these methods with, a better version of this data set would be nice...

In making scatterplots, there is always a choice of a variable for the x-axis and the y-axis. It is our convention to put explanatory or independent variables (the ones used to explain or predict the responses) on the x-axis. In studies where the subjects are randomly assigned to levels of a variable, this is very clearly an explanatory variable, and we can go as far as making causal inferences with it. In observational studies, it can be less clear which variable explains which. In these cases, make the most reasonable choice based on the observed variables but remember that, when the direction of relationship is unclear, you could have switched the axes and thus the implication of which variable is explanatory.

–>

```
# natural log transformation of area burned
mtfires <- mtfires |> mutate(loghectares = log(hectares))
```

```
library(alr4)
data(ais)
library(tibble)
ais <- as_tibble(ais)
aisR <- ais |>
  select(Ht, Hc, Bfat)
summary(aisR)
```

```
##        Ht              Hc              Bfat
##  Min.   :148.9   Min.   :35.90   Min.   : 5.630
##  1st Qu.:174.0   1st Qu.:40.60   1st Qu.: 8.545
##  Median :179.7   Median :43.50   Median :11.650
##  Mean   :180.1   Mean   :43.09   Mean   :13.507
##  3rd Qu.:186.2   3rd Qu.:45.58   3rd Qu.:18.080
##  Max.   :209.4   Max.   :59.70   Max.   :35.520
```

```
aisR |> ggpairs() + theme_bw()
```

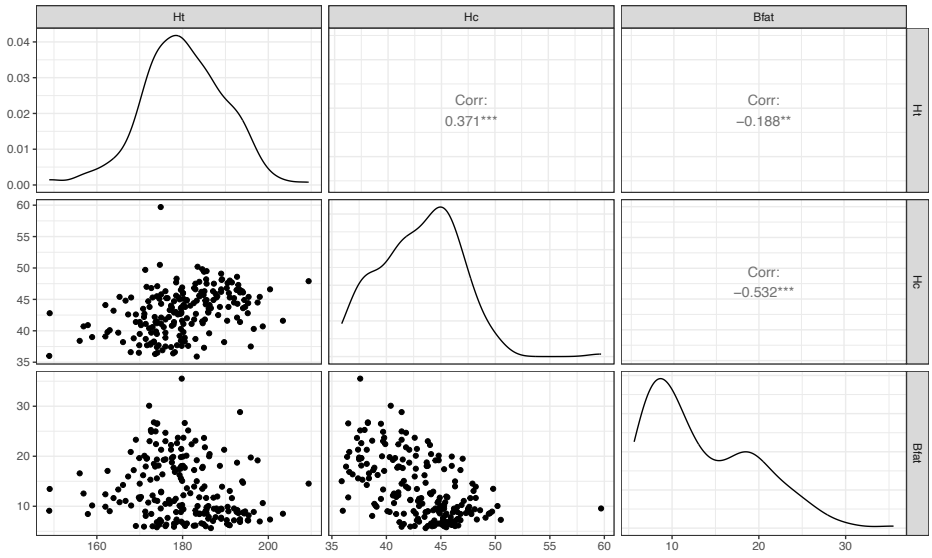

Figure 6.2: (ref:fig6-4)

–>

```
aisR2 <- aisR |> slice(-56, -166) #Removes observations in rows 56 and 166
aisR2 |> ggpairs() + theme_bw()
```

–>

–>

```
library(spuRs) #install.packages("spuRs")
data(ufc)
ufc <- as_tibble(ufc)

ufc |> ggplot(mapping = aes(x = dbh.cm, y = height.m)) +
  geom_point() +
  geom_smooth(method = "lm") +
  theme_bw()
```

6.2 Describing relationships with a regression model

When the relationship appears to be relatively linear, it makes sense to estimate and then interpret a line to represent the relationship between the variables. This line is called a ***regression line*** and involves finding a

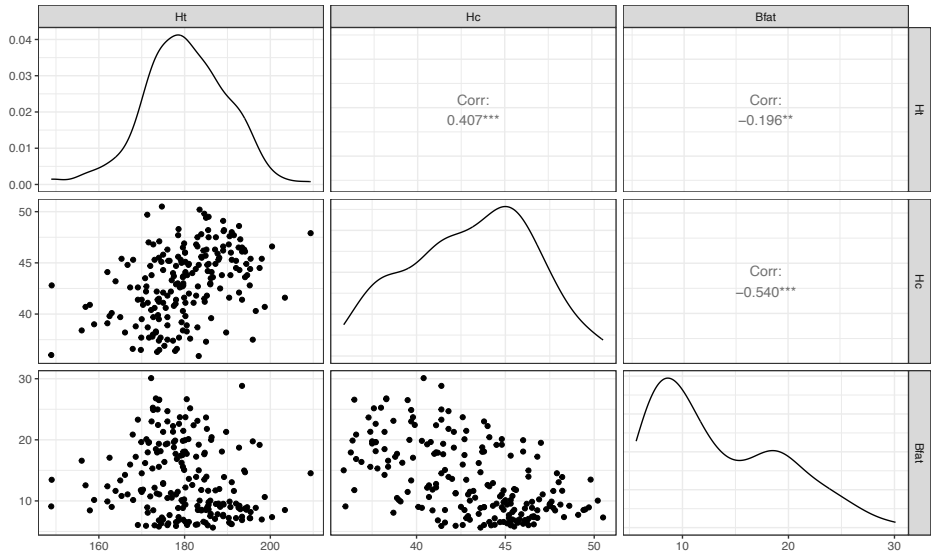

Figure 6.3: (ref:fig6-5)

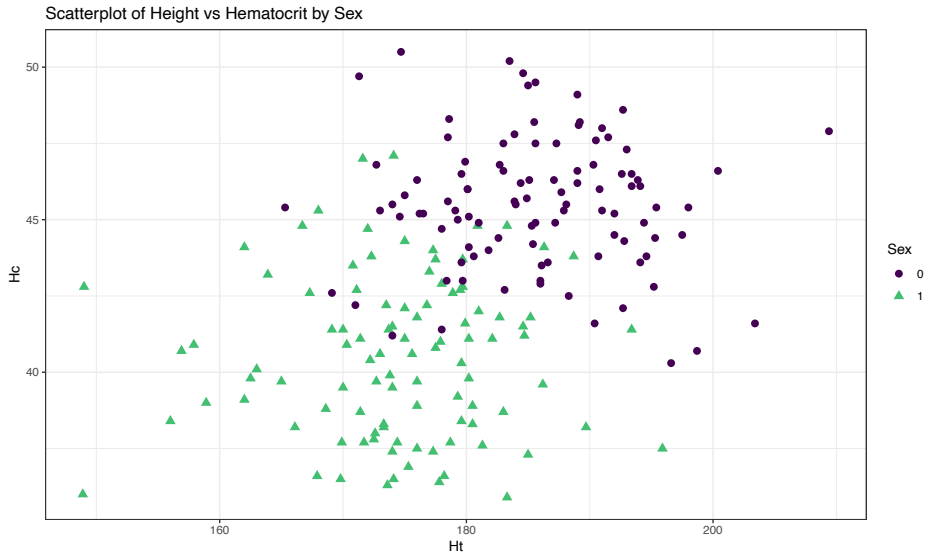

Figure 6.4: (ref:fig6-7)

line that best fits (explains variation in) the response variable for the given values of the explanatory variable. For regression, it matters which variable you choose for x and which you choose for y – for correlation it did not matter. This regression line describes the "effect" of x on y and also provides an equation for predicting values of y for given values of x. The *Beers* and *BAC* data provide a nice example to start our exploration of regression models. The beer consumption is a clear explanatory variable, detectable in the story because (1) it was randomly assigned to subjects and (2) basic science supports beer consumption amount being an explanatory variable for *BAC*. In some situations, this will not be so clear, but look for random assignment or scientific logic to guide your choices of variables as explanatory or response[3].

[3]Even with clear scientific logic, we sometimes make choices to flip the model directions to facilitate different types of analyses. In Vsevolozhskaya et al. [2014] we looked at genomic differences based on obesity groups, even though we were really interested in exploring how gene-level differences explained differences in obesity.

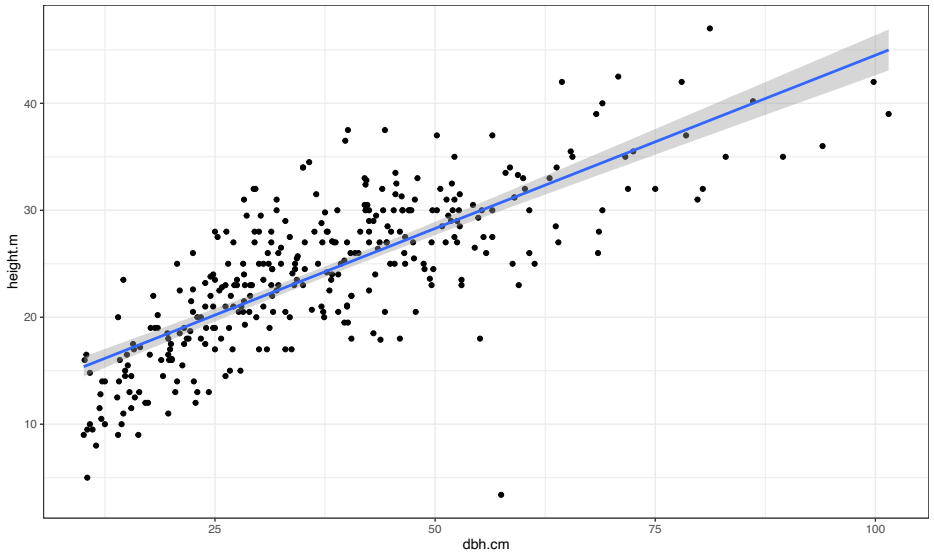

Figure 6.5: (ref:fig6-11)

```
BB |> ggplot(mapping = aes(x = Beers, y = BAC)) +
  geom_smooth(method = "lm", col = "cyan4") +
  geom_point() +
  theme_bw() +
  geom_segment(aes(y = 0.05914, yend = 0.05914, x = 4, xend = 0), col = "blue",
               lty = 2, arrow = arrow(length = unit(.3, "cm"))) +
  geom_segment(aes(x = 4, xend = 4, y = 0, yend = 0.05914),
               arrow = arrow(length = unit(.3, "cm")), col = "blue") +
  geom_segment(aes(y = 0.0771, yend = 0.0771, x = 5, xend = 0), col = "forestgreen",
               lty = 2, arrow = arrow(length = unit(.3, "cm"))) +
  geom_segment(aes(x = 5, xend = 5, y = 0, yend = 0.0771),
               arrow = arrow(length = unit(.3, "cm")), col = "forestgreen")
```

The equation for a line is $y = a + bx$, or maybe $y = mx + b$. In the version $mx + b$ you learned that m is a slope coefficient that relates a change in x to changes in y and that b is a y-intercept (the value of y when x is 0). In Figure 6.6, extra lines are added to help you see the defining characteristics of the line. The slope, whatever letter you use, is the change in y for a one-unit increase in x. Here, the slope is the change in `BAC` for a 1 beer increase in `Beers`, such as the change from 4 to 5 beers. The y-values (dashed lines with arrows) for `Beers` $= 4$ and 5 go from 0.059 to 0.077. This means that for a 1 beer increase ($+1$ unit change in x), the `BAC` goes up by $0.077 - 0.059 = 0.018$ ($+0.018$ unit change in y). We can also try to find the y-intercept on the graph by looking for the `BAC` level for 0 `Beers` consumed. The y-value (`BAC`) ends up being around -0.01 if you extend the regression line to `Beers` $= 0$. You might assume that the `BAC` should be 0 for `Beers` $= 0$ but the researchers did not observe any students at 0 `Beers`, so we don't really know what the `BAC` might be at this value. We have to use our line to ***predict*** this value. This ends up providing a prediction below 0 – an impossible value for BAC. If the y-intercept were positive, it would suggest that the students have a BAC over 0 even without drinking.

The numbers reported were very accurate because we weren't using the plot alone to generate the values – we were using a linear model to estimate the equation to describe the relationship between `Beers` and `BAC`. In statistics, we estimate "m" and "b". We also write the equation starting with the y-intercept and use slightly different notation that allows us to extend to more complicated models with more variables. Specifically, the estimated regression equation is $\widehat{y} = b_0 + b_1 x$, where

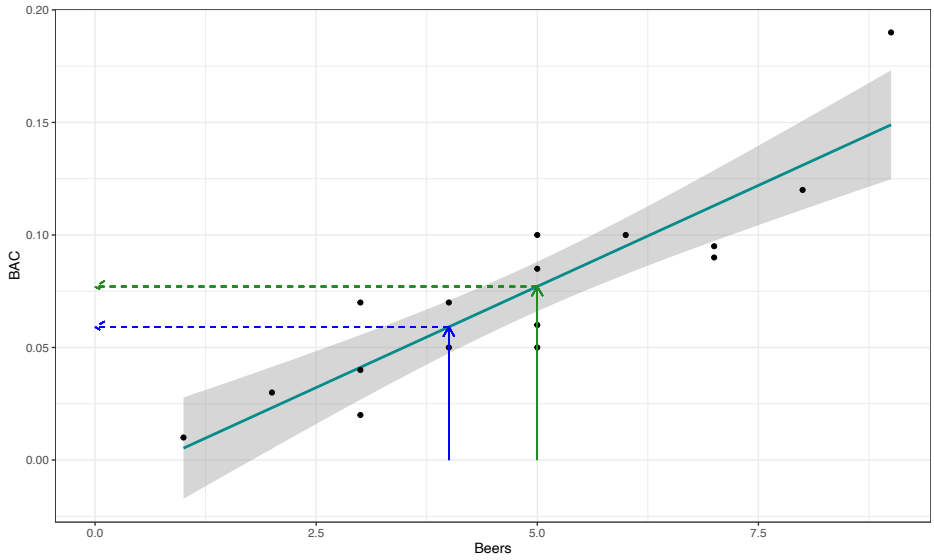

Figure 6.6: Scatterplot with estimated regression line (solid line) for the *Beers* and *BAC* data. The solid arrows indicate the predictor variable values of 4 and 5 beers and the dashed lines illustrate the predicted mean BAC for 4 and 5 beers consumed based on the SLR model.

- $\widehat{y}$ is the estimated value of y for a given x,
- b_0 is the estimated y-intercept (predicted value of y when x is 0),
- b_1 is the estimated slope coefficient, and
- x is the explanatory variable.

One of the differences between when you learned equations in algebra classes and our situation is that the line is not a perfect description of the relationship between x and y – it is an "on average" description and will usually leave differences between the line and the observations, which we call residuals ($e = y - \widehat{y}$). We worked with residuals in the ANOVA[4] material. The residuals describe the vertical distance in the scatterplot between our model (regression line) and the actual observed data point. The lack of a perfect fit of the line to the observations distinguishes statistical equations from those you learned in math classes. The equations work the same, but we have to modify interpretations of the coefficients to reflect this.

We also tie this estimated model to a theoretical or ***population regression model***:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where:

- y_i is the observed response for the i^{th} observation,
- x_i is the observed value of the explanatory variable for the i^{th} observation,
- $\beta_0 + \beta_1 x_i$ is the true mean function evaluated at x_i,
- β_0 is the true (or population) y-intercept,
- β_1 is the true (or population) slope coefficient, and
- the deviations, ε_i, are assumed to be independent and normally distributed with mean 0 and standard deviation σ or, more compactly, $\varepsilon_i \sim N(0, \sigma^2)$.

[4]The residuals from these methods and ANOVA are the same because they all come from linear models but are completely different from the standardized residuals used in the Chi-square material in Chapter 5.

This presents another version of the linear model from Chapters 2, 3, and 4, now with a quantitative explanatory variable instead of categorical explanatory variable(s). This chapter focuses mostly on the estimated regression coefficients, but remember that we are doing statistics and our desire is to make inferences to a larger population. So, estimated coefficients, b_0 and b_1, are approximations to theoretical coefficients, β_0 and β_1. In other words, b_0 and b_1 are the statistics that try to estimate the true population parameters β_0 and β_1, respectively.

To get estimated regression coefficients, we use the `lm` function and our standard `lm(y ~ x, data = ...)` setup. This is the same function used to estimate our ANOVA models and much of this will look familiar. In fact, the ties between ANOVA and regression are deep and fundamental but not the topic of this section. For the *Beers* and *BAC* example, the ***estimated regression coefficients*** can be found from:

```
m1 <- lm(BAC ~ Beers, data = BB)
m1
```

```
##
## Call:
## lm(formula = BAC ~ Beers, data = BB)
##
## Coefficients:
## (Intercept)        Beers
##    -0.01270      0.01796
```

More often, we will extract these from the ***coefficient table*** produced by a model `summary`:

```
summary(m1)
```

```
##
## Call:
## lm(formula = BAC ~ Beers, data = BB)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.027118 -0.017350  0.001773  0.008623  0.041027
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.012701   0.012638  -1.005    0.332
## Beers        0.017964   0.002402   7.480 2.97e-06
##
## Residual standard error: 0.02044 on 14 degrees of freedom
## Multiple R-squared:  0.7998, Adjusted R-squared:  0.7855
## F-statistic: 55.94 on 1 and 14 DF,  p-value: 2.969e-06
```

From either version of the output, you can find the estimated y-intercept in the `(Intercept)` part of the output and the slope coefficient in the `Beers` part of the output. So $b_0 = -0.0127$, $b_1 = 0.01796$, and the ***estimated regression equation*** is

$$\widehat{\text{BAC}}_i = -0.0127 + 0.01796 \cdot \text{Beers}_i.$$

This is the equation that was plotted in Figure 6.6. In writing out the equation, it is good to replace x and y with the variable names to make the predictor and response variables clear. **If you prefer to write all equations with x and y, you need to define x and y or else these equations are not clearly defined.**

There is a general interpretation for the slope coefficient that you will need to master. In general, we interpret the slope coefficient as:

- **Slope interpretation (general):** For a 1 [***unit of X***] increase in $\boldsymbol{X}$, we expect, *on average*, a $\boldsymbol{b_1}$ [***unit of Y***] change in $\boldsymbol{Y}$.

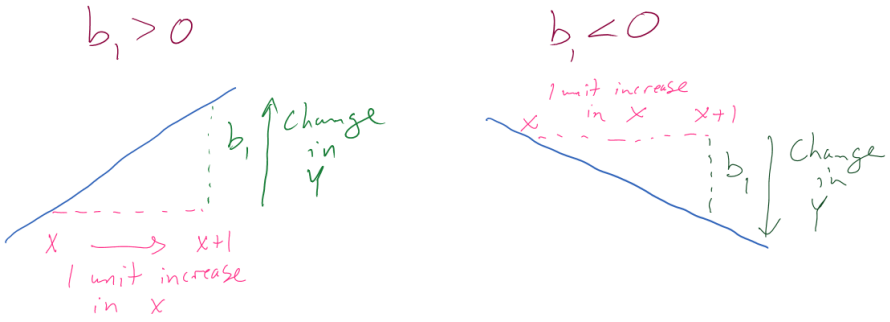

Figure 6.7: Diagram of interpretation of slope coefficients.

Figure 6.7 can help you think about the different sorts of slope coefficients we might need to interpret, both providing changes in the response variable for 1 unit increases in the predictor variable.

Applied to this problem, for each additional 1 beer consumed, we expect a 0.018 gram per dL change in the *BAC on average*. Using "change" in the interpretation for what happened in the response allows you to use the same template for the interpretation even with negative slopes – be careful about saying "decrease" when the slope is negative as you can create a double-negative and end up implying an increase... Note also that you need to carefully incorporate the units of x and the units of y to make the interpretation clear. For example, if the change in *BAC* for 1 beer increase is 0.018, then we could also modify the size of the change in x to be a 10 beer increase and then the estimated change in *BAC* is $10 * 0.018 = 0.18$ g/dL. Both are correct as long as you are clear about the change in x you are talking about. Typically, we will just use the units used in the original variables and only change the scale of "change in x" when it provides an interpretation we are particularly interested in.

Similarly, the general interpretation for a y-intercept is:

- ***Y*-intercept interpretation (general):** For $\boldsymbol{X} = 0$ [***units of X***], we expect, on average, $\boldsymbol{b_0}$ [***units of Y***] **in** $\boldsymbol{Y}$.

Again, applied to the *BAC* data set: For 0 beers for *Beers* consumed, we expect, on average, -0.012 g/dL *BAC*. The y-intercept interpretation is often less interesting than the slope interpretation but can be interesting in some situations. Here, it is predicting average *BAC* for `Beers` $= 0$, which is a value outside the scope of the x's (*Beers* was observed between 1 and 9). Prediction outside the scope of the predictor values is called ***extrapolation***. Extrapolation is dangerous at best and misleading at worst. That said, if you are asked to interpret the y-intercept you should still interpret it, but it is also good to note if it is outside of the region where we had observations on the explanatory variable. Another example is useful for practicing how to do these interpretations.

In the Australian Athlete data, we saw a weak negative relationship between *Body Fat* (% body weight that is fat) and *Hematocrit* (% red blood cells in the blood). The scatterplot in Figure 6.8 shows just the results for the female athletes along with the regression line which has a negative slope coefficient. The estimated regression coefficients are found using the `lm` function:

```
m2 <- lm(Hc ~ Bfat, data = aisR2 |> filter(Sex == 1)) #Results for Females
summary(m2)
```

```
##
## Call:
```

```
## lm(formula = Hc ~ Bfat, data = filter(aisR2, Sex == 1))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2399 -2.2132 -0.1061  1.8917  6.6453
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 42.01378    0.93269  45.046   <2e-16
## Bfat        -0.08504    0.05067  -1.678   0.0965
##
## Residual standard error: 2.598 on 97 degrees of freedom
## Multiple R-squared:  0.02822,    Adjusted R-squared:  0.0182
## F-statistic: 2.816 on 1 and 97 DF,  p-value: 0.09653
```

```
aisR2 |> filter(Sex == 1) |> ggplot(mapping = aes(x = Bfat, y = Hc)) +
  geom_point() +
  geom_smooth(method = "lm") +
  theme_bw() +
  labs(title = "Scatterplot of Body Fat vs Hematocrit for Female Athletes",
       y = "Hc (% blood)", x = "Body fat (% weight)")
```

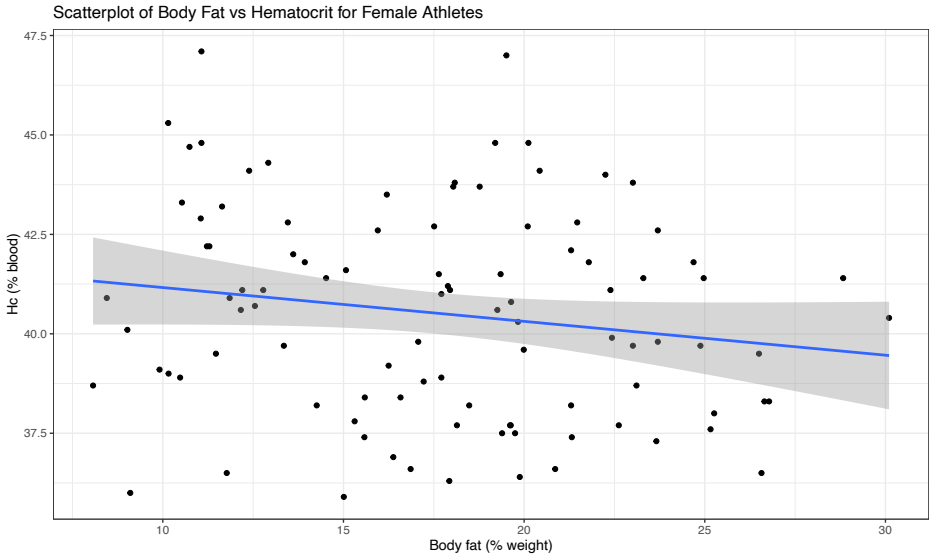

Figure 6.8: Scatterplot of Hematocrit versus Body Fat for female athletes. Note how the `filter` was used to pipe the subset of the data set to the plot.

Based on these results, the estimated regression equation is $\widehat{\text{Hc}}_i = 42.014 - 0.085 \cdot \text{BodyFat}_i$ with $b_0 = 42.014$ and $b_1 = 0.085$. The slope coefficient interpretation is: For a one percent increase in body fat, we expect, on average, a -0.085% (blood) change in Hematocrit for Australian female athletes. For the y-intercept, the interpretation is: For a 0% body fat female athlete, we expect a Hematocrit of 42.014% on average. Again, this y-intercept involves extrapolation to a region of x's that we did not observed. None of the athletes had body fat below 5% so we don't know what would happen to the hematocrit of an athlete that had no body fat except that it probably would not continue to follow a linear relationship.

6.3 Least Squares Estimation

The previous results used the `lm` function as a "black box" to generate the estimated coefficients. The lines produced probably look reasonable but you could imagine drawing other lines that might look equally plausible. Because we are interested in explaining variation in the response variable, we want a model that in some sense minimizes the residuals ($e_i = y_i - \widehat{y}_i$) and explains the responses as well as possible, in other words has $y_i - \widehat{y}_i$ as small as possible. We can't just add these e_i's up because it would always be 0 (remember why we use the variance to measure spread from introductory statistics?). We use a similar technique in regression, we find the regression line that minimizes the squared residuals $e_i^2 = (y_i - \widehat{y}_i)^2$ over all the observations, minimizing the ***Sum of Squared Residuals*** $= \boldsymbol{\Sigma e_i^2}$. Finding the estimated regression coefficients that minimize the sum of squared residuals is called ***least squares estimation*** and provides us a reasonable method for finding the "best" estimated regression line of all the possible choices.

For the *Beers* vs *BAC* data, Figure 6.9 shows the result of a search for the optimal slope coefficient between values of 0 and 0.03. The plot shows how the sum of the squared residuals was minimized for the value that `lm` returned at 0.018. The main point of this is that if any other slope coefficient was tried, it did not do as good **on the least squares criterion** as the least squares estimates.

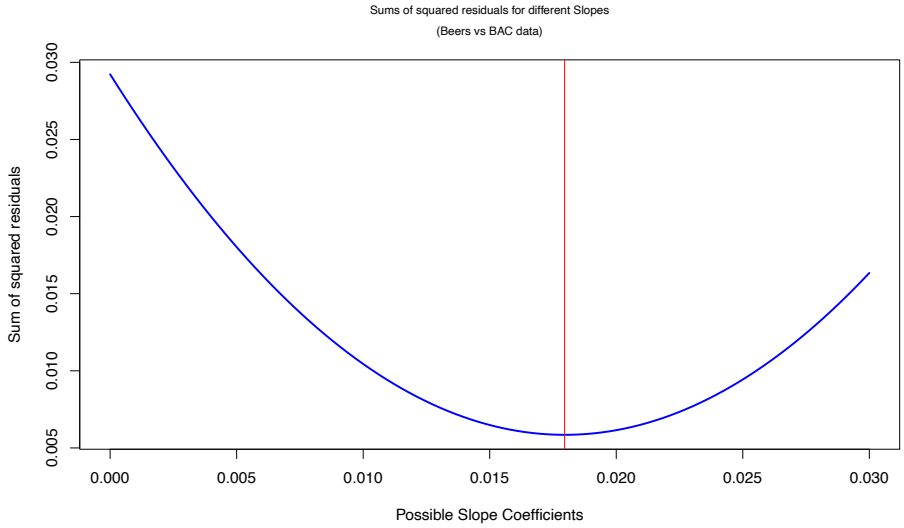

Figure 6.9: Plot of sum of squared residuals vs possible slope coefficients for *Beers* vs *BAC* data, with vertical line for the least squares estimate that minimizes the sum of squared residuals.

Sometimes it is helpful to have a go at finding the estimates yourself. If you install and load the `tigerstats` [Robinson and White, 2020] and `manipulate` [Allaire, 2014] packages in RStudio and then run `FindRegLine()`, you get a chance to try to find the optimal slope and intercept for a fake data set. Click on the "sprocket" icon in the upper left of the plot and you will see something like Figure 6.10. This interaction can help you see how the residuals are being measuring in the y-direction and appreciate that `lm` takes care of this for us.

```
> library(tigerstats)
> library(manipulate)
> FindRegLine()

Equation of the regression line is:
y = 4.34 + -0.02x

Your final score is 13143.99
```

```
Thanks for playing!
```

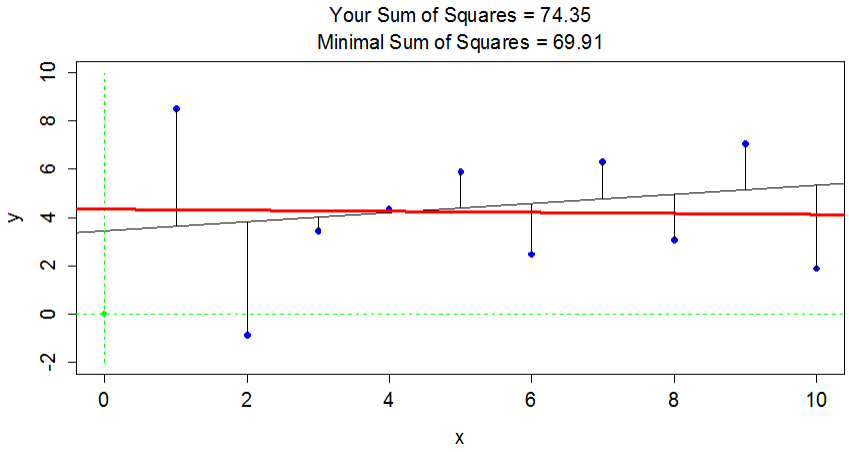

Figure 6.10: Results of running `FindRegLine()` where I didn't quite find the least squares line. The correct line is the bold (red) line and produced a smaller sum of squared residuals than the guessed thinner (black) line.

It ends up that the least squares criterion does not require a search across coefficients or trial and error – there are some "simple" equations available for calculating the estimates of the y-intercept and slope:

$$b_1 = \frac{\Sigma_i(x_i - \bar{x})(y_i - \bar{y})}{\Sigma_i(x_i - \bar{x})^2} = r\frac{s_y}{s_x} \text{ and } b_0 = \bar{y} - b_1\bar{x}.$$

You will never need to use these equations but they do inform some properties of the regression line. The slope coefficient, b_1, is based on the variability in x and y and the correlation between them. If $\boldsymbol{r = 0}$, then the slope coefficient will also be 0. The intercept is a function of the means of x and y and what the estimated slope coefficient is. **If the slope coefficient, $\boldsymbol{b_1}$, is 0, then $\boldsymbol{b_0 = \bar{y}}$** (which is just the mean of the response variable for all observed values of x – this is a very boring model!). The slope is 0 when the correlation is 0. So when there is no linear relationship between x and y ($r = 0$), the least squares regression line is a horizontal line with height $\bar{y}$, and the line produces the same fitted values for all x values. You can also think about this as when there is no relationship between x and y, the best prediction of y is the mean of the y-values and it doesn't change based on the values of x. It is less obvious in these equations, but they also imply that **the regression line ALWAYS goes through the point $(\bar{x}, \bar{y})$.** It provides a sort of anchor point for all regression lines.

For one more example, we can revisit the Montana wildfire areas burned (log-hectares) and the average summer temperature (degrees F), which had $\boldsymbol{r = 0.81}$. The interpretations of the different parts of the regression model follow the least squares estimation provided by `lm`:

```
fire1 <- lm(loghectares ~ Temperature, data = mtfires)
summary(fire1)
```

```
##
## Call:
## lm(formula = loghectares ~ Temperature, data = mtfires)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -3.0822 -0.9549  0.1210  1.0007  2.4728
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -69.7845    12.3132  -5.667 1.26e-05
## Temperature   1.3884     0.2165   6.412 2.35e-06
##
## Residual standard error: 1.476 on 21 degrees of freedom
## Multiple R-squared:  0.6619, Adjusted R-squared:  0.6458
## F-statistic: 41.12 on 1 and 21 DF,  p-value: 2.347e-06
```

- Regression Equation (Completely Specified):
  - Estimated model: $\widehat{\log(\text{Ha})} = -69.78 + 1.39 \cdot \text{Temp}$
  - Or $\widehat{y} = -69.78 + 1.39x$ with **Y = log(Ha) and X = Temperature**
- Response Variable: Yearly *log* Hectares burned by wildfires
- Explanatory Variable: Average Summer Temperature
- Estimated y-Intercept (b_0): -69.78
- Estimated slope (b_1): 1.39
- Slope Interpretation: For a 1 degree Fahrenheit increase in Average Summer Temperature we would expect, **on average**, a 1.39 log(Hectares) *change* in log(Hectares) burned in Montana.
- Y-intercept Interpretation: If temperature were 0 degrees F, we would expect -69.78 log(Hectares) burned **on average** in Montana.

One other use of regression equations is for prediction. It is a trivial exercise (or maybe not – we'll see when you try it!) to plug an x-value of interest into the regression equation and get an estimate for y at that x. Basically, the regression lines displayed in the scatterplots show the predictions from the regression line across the range of x's. Formally, ***prediction*** involves estimating the response for a particular value of x. We know that it won't be perfect but it is our best guess. Suppose that we are interested in predicting the log-area burned for a summer that had an average temperature of 59°F. If we plug 59°F into the regression equation, $\widehat{\log(\text{Ha})} = -69.78 + 1.39 \bullet \text{Temp}$, we get

$$\begin{aligned}\widehat{\log(\text{Ha})} &= -69.78 \text{ log-hectares} + 1.39 \text{ log-hectares}/^\circ\text{F} \bullet 59^\circ\text{F} \\ &= -69.78 \text{ log-hectares} + 1.39 \text{ log-hectares}/\cancel{^\circ\text{F}} \bullet 59\cancel{^\circ\text{F}} \\ &= 12.23 \text{ log-hectares}\end{aligned}$$

We did not observe any summers at exactly $x = 59$ but did observe some nearby and this result seems relatively reasonable.

Now suppose someone asks you to use this equation for predicting Temperature = $65^\circ F$. We can run that through the equation: $-69.78 + 1.39 * 65 = 20.57$ log-hectares. But can we trust this prediction? We did not observe any summers over 60 degrees F so we are now predicting outside the scope of our observations – performing ***extrapolation***. Having a scatterplot in hand helps us to assess the range of values where we can reasonably use the equation – here between 54 and 60 degrees F seems reasonable.

```
mtfires |> ggplot(mapping = aes(x = Temperature, y = loghectares)) +
  geom_point(aes(color = Year), size = 2.5) +
  geom_smooth(method = "lm") +
  theme_bw() +
  scale_color_viridis() +
```

```
  labs(title = "Scatterplot with regression line for Area burned vs
       Temperature, colored by year")
```

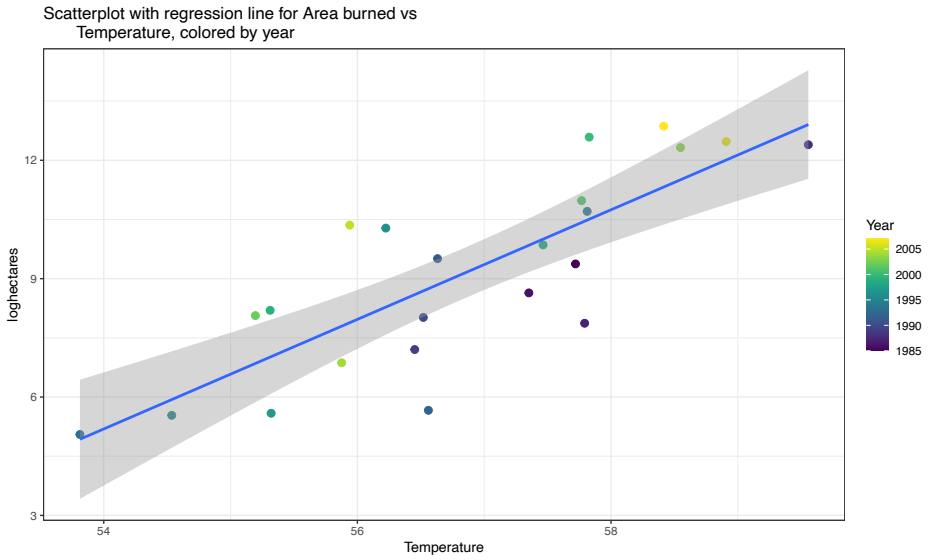

Figure 6.11: Scatterplot of log-hectares burned versus temperature with estimated regression line. Information on the year of each observation is added using a local aesthetic inside `geom_point` to color the points on a color gradient based on `Year`.

6.4 Measuring the strength of regressions: R^2

At the beginning of the chapter, we used the correlation coefficient to measure the strength and direction of the linear relationship. The regression line provides an even more detailed description of the direction of the linear relationship than the correlation provided; in regression we addressed the question of "for a unit change in x, what sort of change in y do we expect, on average?" whereas the correlation just addressed whether the relationship was positive or negative. However, the **regression line tells us nothing about the strength of the relationship**. Consider the three scatterplots in Figure 6.12: the left panel is the original BAC data and the two right panels have fake data that generated exactly the same estimated regression model with a weaker (middle panel) and then a stronger (right panel) linear relationship between `Beers` and `BAC`. This suggests that the regression line is a useful but incomplete characterization of relationships between variables – we need a measure of strength of the relationship to go with the equation.

We could use the correlation coefficient, $\boldsymbol{r}$, again to characterize strength but it is somewhat redundant to report a measure that contains direction information. It also will not extend to multiple regression models where we have more than one predictor variable in the same model.

In regression models, we use the ***coefficient of determination*** (symbol: $\mathbf{R^2}$) to accompany our regression line and describe the strength of the relationship and assess the quality of the model fit. It can either be scaled between 0 and 1 or 0 to 100% and has "units" of the proportion or percentage of the variation in y that is explained by the model that includes x (and later more than one x). For example, an $\mathbf{R^2}$ of 0% corresponds to explaining 0% of the variation in the response with our model (worst possible fit) and $\boldsymbol{R^2} = 100\%$ means that all the variation in the response was explained by the model (best possible fit). In between, it provides a nice summary of how much of the total variability in the response we can account for with our model including x (and, in Chapter 8, including multiple predictor variables).

The $\mathbf{R^2}$ is calculated using the sums of squares we encountered in the ANOVA methods. We once again have some total amount of variability that is attributed to the variation based on the model fit, here we call it

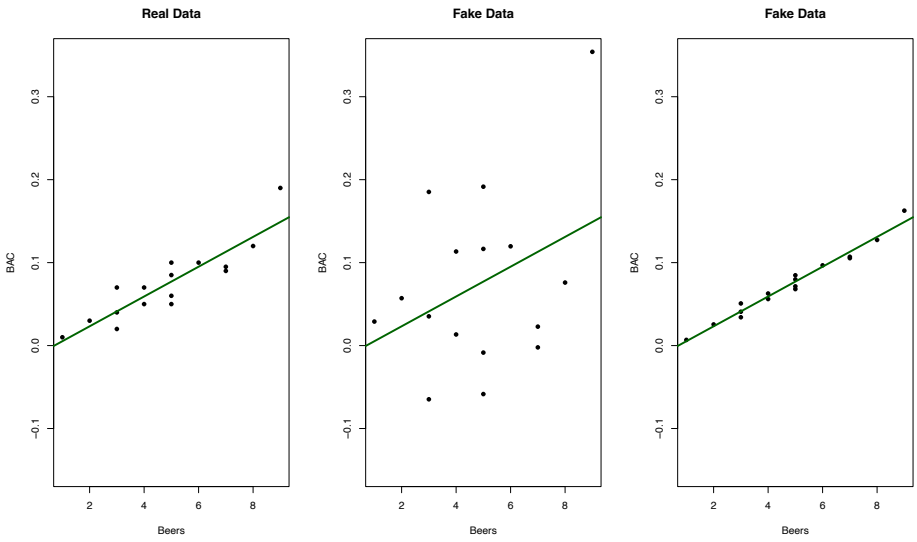

Figure 6.12: Three scatterplots with the same estimated regression line.

$SS_{regression}$, and the residual variability, still $SS_{error} = \Sigma(y - \widehat{y})^2$. The $SS_{regression}$ is most easily calculated as $SS_{regression} = SS_{Total} - SS_{error}$, the difference between the total variability and the variability not explained by the model under consideration. Using these quantities, we calculate the portion of the total variability that the model explains as

$$\boldsymbol{R^2} = \frac{SS_{regression}}{SS_{Total}} = 1 - \frac{SS_{error}}{SS_{Total}}.$$

It also ends up that the coefficient of determination for models with one predictor is the correlation coefficient ($\boldsymbol{r}$) squared ($\boldsymbol{R^2 = r^2}$). So we can quickly find coefficients of determination if we know correlations in simple linear regression models. In the real *Beers* and *BAC* data, $\boldsymbol{r} = 0.8943$. So $\boldsymbol{R^2} = 0.79998$ or approximately 0.80. So 80% of the variation in *BAC* is explained by *Beer* consumption. That leaves 20% of the variation in the responses to be unexplained by our model. In this case much of the unexplained variation is likely attributable to differences in physical characteristics (that were not measured) but the statistical model places that unexplained variation into the category of "random errors". We don't actually have to find $\boldsymbol{r}$ to get coefficients of determination – the result is part of the regular summary of a regression model that we have not discussed. We repeat the full `lm` model summary below – note that a number is reported for the "`Multiple R-squared`" in the second to last line of the output. It is reported as a proportion and it is your choice whether you want to report and interpret it as a proportion or percentage, just make that clear in how you discuss it.

```
m1 <- lm(BAC ~ Beers, data = BB)
summary(m1)
```

```
##
## Call:
## lm(formula = BAC ~ Beers, data = BB)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.027118 -0.017350  0.001773  0.008623  0.041027
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.012701   0.012638  -1.005    0.332
## Beers        0.017964   0.002402   7.480 2.97e-06
##
## Residual standard error: 0.02044 on 14 degrees of freedom
## Multiple R-squared:  0.7998, Adjusted R-squared:  0.7855
## F-statistic: 55.94 on 1 and 14 DF,  p-value: 2.969e-06
```

In this output, be careful because there is another related quantity called ***Adjusted* R-squared** that we will discuss later. This other quantity is not a measure of the strength of the relationship but will be useful.

We could also revisit the ANOVA table for this model to verify the source of the $\mathbf{R^2}$ of 0.80 based on $\text{SS}_{\text{regression}} = 0.02337$ and $\text{SS}_{\text{Total}} = 0.02337 + 0.00585$. This provides 0.80 from 0.02337/0.02922.

```
anova(m1)
```

```
## Analysis of Variance Table
##
## Response: BAC
##           Df    Sum Sq   Mean Sq F value    Pr(>F)
## Beers      1 0.0233753 0.0233753  55.944 2.969e-06
## Residuals 14 0.0058497 0.0004178
```

```
SStotal <- 0.0233753 + 0.0058497
SSregression <- 0.0233753
SSregression/SStotal
```

```
## [1] 0.7998392
```

In Figure 6.12, there are three examples with the same regression model, but different strengths of relationships. In the real data set $\boldsymbol{R^2} = 80\%$. For the first fake data set (middle panel), the $\mathbf{R^2}$ drops to 13.8% and for the second fake data set (right panel), $\mathbf{R^2}$ is 97.3%. As a summary, $\mathbf{R^2}$ provides a natural scale to understand "how good" each model is at explaining the responses. We can revisit some of our previous models to get a little more practice with using this summary of strength or quality of regression models.

For the Montana fire data, $\boldsymbol{R^2} = 66.2\%$. So the proportion of variation of log-area burned that is explained by average summer temperature is 0.662. This is "good" but also leaves quite a bit of unexplained variation in the responses. There is a long list of reasons why this explanatory variable leaves a lot of variation in the response unexplained. Note that we were careful about using the scaling of the response variable (log(area burned)) in the interpretation – this is because we would get a much different answer if area burned vs temperature was considered.

```
fire1 <- lm(loghectares ~ Temperature, data = mtfires)
summary(fire1)
```

```
##
## Call:
## lm(formula = loghectares ~ Temperature, data = mtfires)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0822 -0.9549  0.1210  1.0007  2.4728
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -69.7845    12.3132  -5.667 1.26e-05
## Temperature   1.3884     0.2165   6.412 2.35e-06
##
## Residual standard error: 1.476 on 21 degrees of freedom
## Multiple R-squared:  0.6619, Adjusted R-squared:  0.6458
## F-statistic: 41.12 on 1 and 21 DF,  p-value: 2.347e-06
```

For the model for female Australian athletes that used *Body fat* to explain *Hematocrit*, the estimated regression model was $\widehat{\text{Hc}}_i = 42.014 - 0.085 \cdot \text{BodyFat}_i$ and $\boldsymbol{r} = -0.168$. The coefficient of determination is $\boldsymbol{R^2} = (-0.168)^2 = 0.0282$. So *body fat* explains 2.8% of the variation in *Hematocrit* in these women. That is not a very good regression model with over 97% of the variation in *Hematocrit* unexplained by this model. The scatterplot showed a fairly weak relationship but this provides numerical and interpretable information

that drives that point home.

```
m2 <- lm(Hc ~ Bfat, data = aisR2 |> filter(Sex == 1)) #Results for Females
summary(m2)
```

```
##
## Call:
## lm(formula = Hc ~ Bfat, data = filter(aisR2, Sex == 1))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2399 -2.2132 -0.1061  1.8917  6.6453
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 42.01378    0.93269  45.046   <2e-16
## Bfat        -0.08504    0.05067  -1.678   0.0965
##
## Residual standard error: 2.598 on 97 degrees of freedom
## Multiple R-squared:  0.02822,    Adjusted R-squared:  0.0182
## F-statistic: 2.816 on 1 and 97 DF,  p-value: 0.09653
```

6.5 Outliers: leverage and influence

In the review of correlation, we loosely considered the impacts of outliers on the correlation. We removed unusual points to see both the visual changes (in the scatterplot) as well as changes in the correlation coefficient in Figures 6.2 and 6.3. In this section, we formalize these ideas in the context of impacts of unusual points on our regression equation. In regression, it is possible for a single point to have a big impact on the overall regression results but it is also possible to have a clear outlier that has little impact on the results. We call an observation ***influential*** if its removal causes a "big" change in the regression line, specifically in terms of impacting the slope coefficient. Points that are on the edges of the x's (far from the mean of the x's) have the potential for more impact on the line as we will see in some examples shortly.

You can think of the regression line being balanced at $\bar{x}$ and the further from that location a point is, the more a single point can move the line. We can measure the distance of points from $\bar{x}$ to quantify each observation's potential for impact on the line using what is called the ***leverage*** of a point. Leverage is a positive numerical measure with larger values corresponding to more leverage. The scale changes depending on the sample size (n) and the complexity of the model so all that matters is which observations have more or less relative leverage in a particular data set. The observations with x-values that provide higher leverage have increased potential to influence the estimated regression line. Along with measuring the leverage, we can also measure the influence that each point has on the regression line using ***Cook's Distance*** or ***Cook's D***. It also is a positive measure with higher values suggesting more influence. The rule of thumb is that Cook's D values over 1.0 correspond to clearly influential points, values over 0.5 have some influence and values lower than 0.5 indicate points that are not influential on the regression model slope coefficients. One part of the regular diagnostic plots we will use for regression models displays the leverages on the x-axis, the standardized residuals on the y-axis, and adds contour lines for Cook's Distances in a panel that is labeled "Residuals vs Leverage". This allows us to see the potential for impact of a point (leverage), how far it's observation was from the regression line (residual), and to see a measure of that point's influence (Cook's D).

To extract the level of Cook's D on the "Residuals vs Leverage" plot, look for contours to show up on the upper and lower right of the plot. They show increasing levels of influence going to the upper and lower right corners as you combine higher leverage (x-axis) and larger residuals (y-axis) – the two ingredients required to be influential on the line. The contours are displayed for Cook's D values of 0.5 and 1.0 if there are points

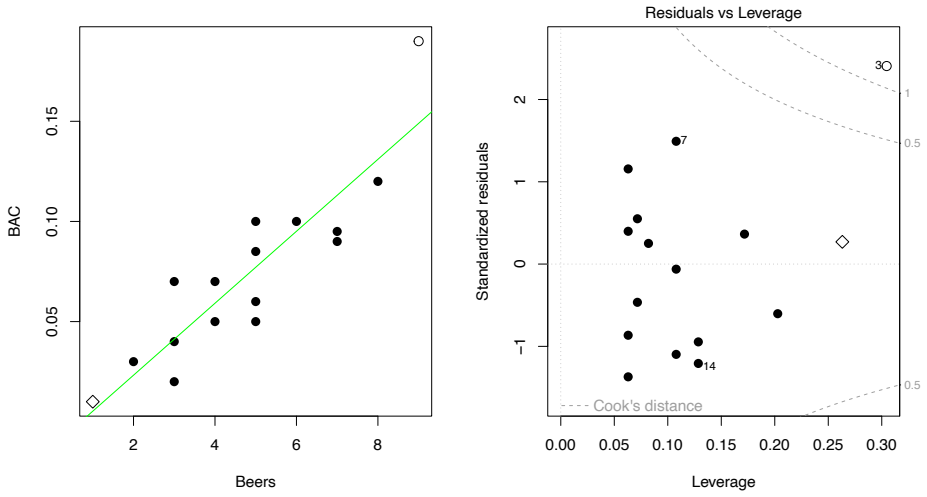

Figure 6.13: Scatterplot and Residuals vs Leverage plot for the real BAC data. Two high leverage points are flagged, with only one that has a Cook's D value over 1 ("○") and is indicated as influential.

near or over those levels. The Cook's D values come from a topographical surface of values that is a sort of U-shaped valley in the middle of the plot centered at $y = 0$ with the lowest contour corresponding to Cook's D values below 0.5 (no influence). As you move to the upper right or lower right corners, the influence increases and the edges of the valley get steeper. If you do not see any contours in the plot, then no points were even close to being influential based on Cook's D.

To illustrate these concepts, the original *Beers* and *BAC* data are used again. In the scatter plot in Figure 6.13, two points are plotted with different characters. The point for 1 *Beer* and *BAC* of 0.010 is displayed as a "◇" and the 9 *Beer* and *BAC* 0.19 observation is displayed with a "○". These two points are the furthest from the mean of the of the x's ($\overline{\text{Beers}} = 4.8$) but show two different levels of influence on the line. The "◇" point has a leverage of 0.27 and the 9 *Beer* observation ("○") had a leverage of 0.30. The 1 *Beer* observation was close to the pattern defined by the other points, had a small residual, and a Cook's D value below 0.5 (it did not exceed the first of the contours). So even though it had high leverage, it was not an influential point. The 9 *Beer* observation had the highest leverage in the data set and was quite a bit above the pattern defined by the other points and ends up being an influential point with a Cook's D over 1. We might want to consider fitting this model without that observation to get a better estimate of the effects of beer consumption on BAC or revisit our assumption that the relationship is really linear here.

To further explore influence, we will add a point to the original data set and move it around so you can see how those changes impact the results. For each scatterplot in Figure 6.14, the Residuals vs Leverage plot is displayed to its right. The original data are "•" and the original regression line is the dashed line in Figure 6.14. First, a fake observation at 11 *Beers* and 0.1 *BAC* is added, at (11, 0.1), in the top panels of the figure. This observation is clearly an outlier and heavily impacts the slope of the regression line (so is clearly influential). This added point drops the $\mathbf{R^2}$ from 0.80 in the original data to 0.24. The accompanying Residuals vs Leverage plot shows that this point has extremely high leverage and a Cook's D over 1 – it is a clearly influential point. However, **having high leverage does not always make points influential.** Consider the second row of plots with an added point of (11, 0.19). The regression line barely changes and $\mathbf{R^2}$ increases a little. This point has the same leverage as in the first example since it is the same set of x's and the distance to the mean of the x's is unchanged. But it is not influential since its Cook's D value is less than 0.5. This occurred because it followed the overall pattern of observations even though it was "far away" from the other observations in the x-direction. The last two rows of plots show what happens when low leverage outliers are encountered. If observations are near the center of the x's, it ends up that to be influential the points have to be very far from the pattern of the other observations. The (5, 0.19)

example almost attains a Cook's D of 0.5 but has little impact on the regression line, especially the slope coefficient. It does impact the y-intercept and drops the R-squared value to 0.57. The same result occurs if the observation is noticeably lower than the other points.

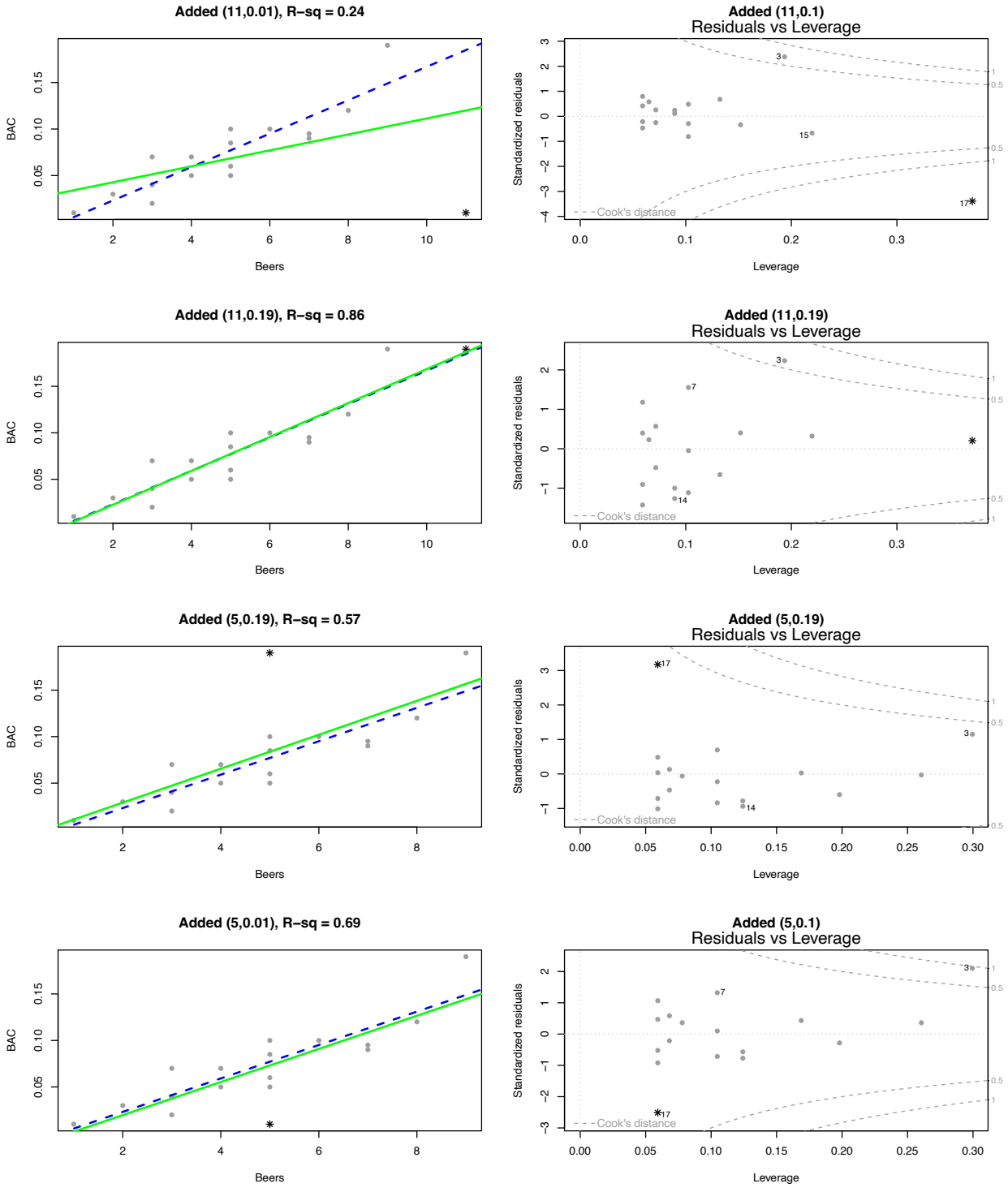

Figure 6.14: Plots exploring the impacts of moving a single additional observation in the BAC example. The added point is indicated with * and the original regression line is the dashed line in the left column.

When we are doing regressions, we get very worried about points "at the edges" having an undue influence on the results. When we start using multiple predictors, say if we had body weight data on these subjects as well as beer consumption, it becomes harder to "see" if the points are "far away" from the other observations and we will trust the Residuals vs Leverage plots to help us identify the influential points. These techniques work the same in the multiple regression models in Chapter 8 as they do in these simpler, single predictor regression models.

6.6 Residual diagnostics – setting the stage for inference

Influential points are not the only potential issue that can cause us to have concerns about our regression model. There are two levels to these considerations. The first is related to issues that directly impact the least squares regression line and cause concerns about whether a line is a reasonable representation of the relationship between the two variables. These issues for regression model estimation have been discussed previously (the same concerns in estimating correlation apply to regression models). The second level is whether the line we have will be useful for making inferences for the population that our data were collected from and whether the data follow our assumed model. Our window into problems of both types is the residuals ($e_i = y_i - \widehat{y}_i$). By exploring patterns in how the line "misses" the responses we can gain information about the reasonableness of using the estimated regression line and sometimes information about how we might fix problems. The validity conditions for doing inference in a regression setting (Chapter 7) involve two sets of considerations, those that are assessed based on the data collection and measurement process and those that can be assessed using diagnostic plots. The first set is:

- **Quantitative variables condition**
  - We'll discuss using categorical predictor variables later – to use simple linear regression both the explanatory and response variables need to quantitative.
- **Independence of observations**
  - As in the ANOVA models, linear regression models assume that the observations are collected in a fashion that makes them independent.
  - This will be based on the "story" of the data. Consult a statistician if your data violate this assumption as there are more advanced methods that adjust for dependency in observations but they are beyond the scope of this material.

The remaining assumptions for getting valid inferences from regression models can be assessed using diagnostic plots:

- **Linearity of relationship**
  - We should not report a linear regression model if the data show a curve (curvilinear relationship between x and y).
  - Examine the initial scatterplot to assess the potential for a curving relationship.
  - Examine the Residuals vs Fitted (top left panel of diagnostic plot display) plot:
    - If the model missed a curve in the relationship, the residuals often will highlight that missed pattern and a curve will show up in this plot.
    - Try to explain or understand the pattern in what is left over. If we have a good model, there shouldn't be much left to "explain" in the residuals (i.e., no patterns left over after accounting for x).
- **Equal (constant) variance**
  - We assume that the variation is the same for all the observations and especially that the variability does not change in the responses as a function of our predictor variables or the fitted values.
  - There are three plots to help with this:

    - Examine the original scatterplot and look at the variation around the line and whether it looks constant across values of x.
    - Examine the Residuals vs Fitted plot and look for evidence of changing spread in the residuals, being careful to try to separate curving patterns from non-constant variance (and look for situations where both are present as you can violate both conditions simultaneously).
    - Examine the "Scale-Location" plot and look for changing spread as a function of the fitted values.
        - The y-axis in this plot is the square-root of the absolute value of the standardized residual. This scale flips the negative residuals on top of the positive ones to help you better assess changing variability without being distracted by whether the residuals are above or below 0.
        - Because of the absolute value, curves in the Residuals vs Fitted plot can present as sort of looking like non-constant variance in the Scale-Location plot – check for nonlinearity in the residuals vs fitted values before using this plot. If nonlinearity is present, just use the Residuals vs Fitted and original scatterplot for assessing constant variance around the curving pattern.
  - If there are patterns of increasing or decreasing variation (often described as funnel or cone shapes), then it might be possible to use a transformation to fix this problem (more later). It is possible to have decreasing and then increasing variability and this also is a violation of this condition.

- **Normality of residuals**
  - Examine the Normal QQ-plot for violations of the normality assumption as in Chapters 3 and 4.
    - Specifically review the discussion of identifying skews in different directions and heavy vs light tailed distributions.
    - Skewed and heavy-tailed distributions are the main problems for our inferences, especially since both kinds of distributions can contain outliers that can wreak havoc on the estimated regression line.
    - Light-tailed distributions cause us no real inference issues except that the results are conservative so you should note when you observe these situations but feel free to proceed with using your model results.
    - Remember that clear outliers are an example of a violation of the normality assumption but some outliers may just influence the regression line and make it fit poorly and this issue will be more clearly observed in the residuals vs fitted than in the QQ-plot.

- **No influential points**
  - Examine the Residuals vs Leverage plot as discussed in the previous section.
  - Consider removing influential points (one at a time) and focusing on results without those points in the data set.

To assess these later assumptions, we will use the four residual diagnostic plots that R provides from `lm` fitted models. They are similar to the results from ANOVA models but the Residuals vs Leverage plot is now interesting as was discussed in Section 6.5. Now we can fully assess the potential for trusting the estimated regression models in a couple of our examples:

- **Beers vs BAC:**
  - Quantitative variables condition:
    - Both variables are quantitative.

- Independence of observations:
  - We can assume that all the subjects are independent of each other. There is only one measurement per student and it is unlikely that one subject's beer consumption would impact another's BAC. Unless the students were trading blood it isn't possible for one person's beer consumption to change someone else's BAC.

```
m1 <- lm(BAC ~ Beers, data = BB)
par(mfrow = c(2,2))
plot(m1, add.smooth = F, main = "Beers vs BAC", pch = 16)
```

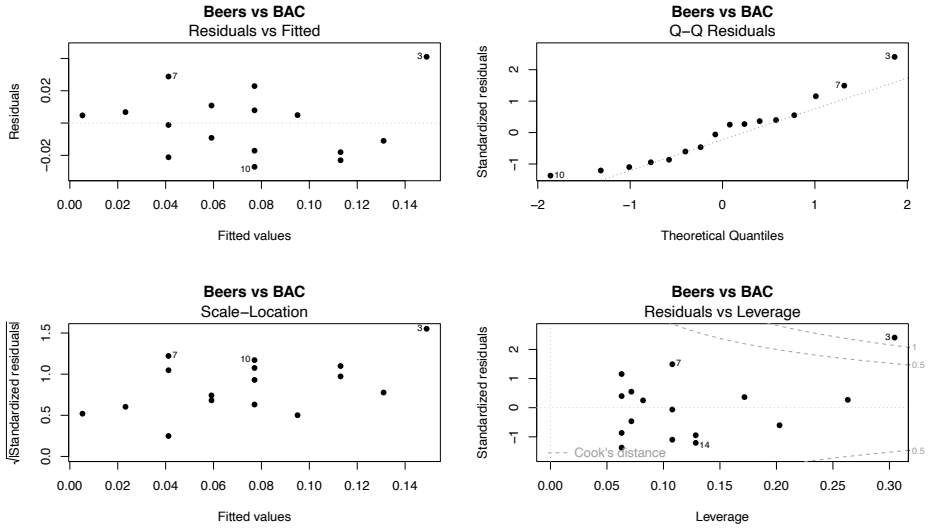

Figure 6.15: Full suite of diagnostics plots for *Beer* vs *BAC* data.

- Linearity, constant variance from Residuals vs Fitted:
  - We previously have identified a potentially influential outlier point in these data. Consulting the Residuals vs Fitted plot in Figure 6.15, if you trust that influential point, shows some curvature with a pattern of decreasing residuals as a function of the fitted values and then an increase at the right. Or, if you do not trust that highest BAC observation, then there is a mostly linear relationship with an outlier identified. We would probably suggest that it is an outlier, should be removed from the analysis, and inferences constrained to the region of beer consumption from 1 to 8 beers since we don't know what might happen at higher values.
- Constant variance from Scale-Location:
  - There is some evidence of increasing variability in this plot as the spread of the results increases from left to right, however this is just an artifact of the pattern in the original residuals and not real evidence of non-constant variance. Note that there is little to no evidence of non-constant variance in the Residuals vs Fitted.
- Normality from Normal QQ Plot:
  - The left tail is a little short and the right tail is a little long, suggesting a slightly right skewed distribution in the residuals. This also corresponds to having a large positive outlying value. But we would conclude that there is a minor issue with normality in the residuals here.
- Influential points from Residuals vs Leverage:

- Previously discussed, this plot shows one influential point with a Cook's D value over 1 that is distorting the fitted model and is likely the biggest issue here.

- **Tree height and tree diameter** (suspicious observation already removed):
  - Quantitative variables: Met
  - Independence of observations:
    - There are multiple trees that were measured in each plot. One problem might be that once a tree is established in an area, the other trees might not grow as tall. The other problem is that some sites might have better soil conditions than others. Then, all the trees in those rich soil areas might be systematically taller than the trees in other areas. Again, there are statistical methods to account for this sort of "clustering" of measurements but this technically violates the assumption that the trees are independent of each other. So this assumption is violated, but we will proceed with that caveat on our results – the precision of our inferences might be slightly over-stated due to some potential dependency in the measurements.
  - Linearity, constant variance from Residuals vs Fitted in Figure 6.16.
    - There is evidence of a curve that was missed by the linear model.
    - There is also evidence of increasing variability AROUND the curve in the residuals.
  - Constant variance from Scale-Location:
    - This plot actually shows relatively constant variance but this plot is misleading when curves are present in the data set. **Focus on the Residuals vs Fitted to diagnose non-constant variance in situations where a curve was missed.**
  - Normality in Normal QQ plot:
    - There is no indication of any problem with the normality assumption.
  - Influential points?
    - The Cook's D contours do not show up in this plot so none of the points are influential.

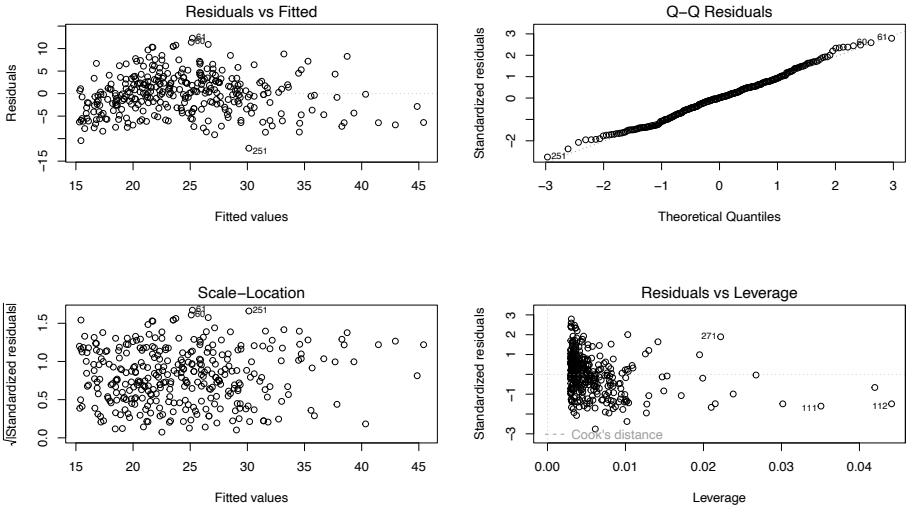

Figure 6.16: Diagnostics plots for tree height and diameter simple linear regression model.

So the main issues with this model are the curving relationship and non-constant variance. We'll revisit this example later to see if we can find a model on transformed variables that has better diagnostics. Reporting

the following regression model that has a decent R^2 of 62.6% would be misleading since it does not accurately represent the relationship between tree diameter and tree height.

```
tree1 <- lm(height.m ~ dbh.cm, data = ufc |> slice(-168))
summary(tree1)
```

```
##
## Call:
## lm(formula = height.m ~ dbh.cm, data = slice(ufc, -168))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.1333  -3.1154   0.0711   2.7548  12.3076
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.98364    0.57422   20.87   <2e-16
## dbh.cm       0.32939    0.01395   23.61   <2e-16
##
## Residual standard error: 4.413 on 333 degrees of freedom
## Multiple R-squared:  0.626,  Adjusted R-squared:  0.6249
## F-statistic: 557.4 on 1 and 333 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2,2))
plot(tree1, add.smooth = F)
```

6.7 Old Faithful discharge and waiting times

A study in August 1985 considered time for Old Faithful and how that might relate to *waiting time* for the next eruption (Ripley and Venables [2024], Azzalini and Bowman [1990]). This sort of research provides the Yellowstone National Park (YNP) staff a way to show tourists a predicted time to next eruption so they can quickly see it erupt and then get back in their cars, not wasting too much time in the outdoors. Or, less cynically, the opportunity to study the behavior of the eruption of a geyser. Both variables are measured in minutes and the scatterplot in Figure 6.17 shows a moderate to strong positive and relatively linear relationship. We added a ***smoothing line*** (dashed line) using `geom_smooth` to this plot – this is actually the default choice in `geom_smooth` and we have to use `geom_smooth(method = "lm")` to get the regression (straight) line. Smoothing lines provide regression-like fits but are performed on local areas of the relationship between the two variables and so can highlight where the relationships change, especially highlighting curvilinear relationships. They can also return straight lines just like the regression line if that is reasonable. The technical details of regression smoothing are not covered here but they are a useful graphical addition to help visualize nonlinearity in relationships and a topic you can explore further based on the sources related to the `mgcv` R package [Wood, 2023], which is being used by `geom_smooth`.

In these data, there appear to be two groups of eruptions (shorter length, shorter wait and longer length, longer wait) – but we don't know enough about these data to assume that there are two groups. The smoothing line does help us to see if the relationship appears to change or stay the same across different values of the explanatory variable, `Duration`. The smoothing line suggests that the upper group might have a less steep slope than the lower group as it sort of levels off for observations with `Duration` of over 4 minutes. It also indicates that there is one point for an eruption under 1 minute in `Duration` that might be causing some problems for both the linear fit and the smoothing line. The story of these data involve some measurements during the night that were just noted as being short, medium, and long – and they were re-coded as 2, 3, or 4 minute duration eruptions. You can see responses stacking up at 2 and 4 minute durations and this is obviously a problematic aspect of these data. We'll see if our diagnostics detect some

of these issues when we fit a simple linear regression to try to explain waiting time based on duration of prior eruption.

```
data(geyser, package = "MASS")
geyser <- as_tibble(geyser)
# Aligns the duration with time to next eruption
G2 <- tibble(Waiting = geyser$waiting[-1], Duration = geyser$duration[-299])

G2 |> ggplot(mapping = aes(x = Duration, y = Waiting)) +
  geom_point() +
  geom_smooth(method = "lm") +
  geom_smooth(lty = 2, col = "red", lwd = 1.5, se = F) + #Add smoothing line
  theme_bw() +
  labs(title = "Scatterplot with regression and smoothing line,
       Waiting Time vs Duration")
```

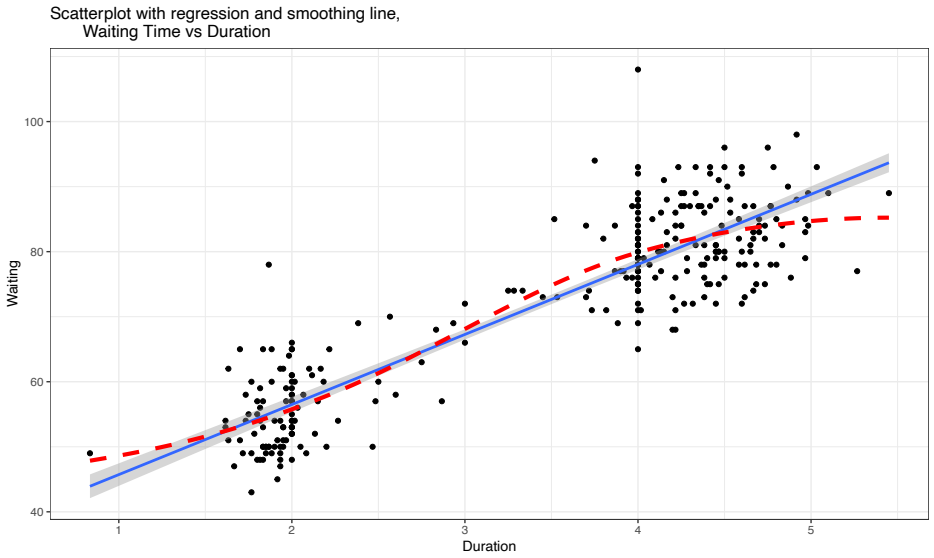

Figure 6.17: Scatterplot of Old Faithful waiting times to next eruption (minutes) and duration of prior eruption (minutes) with smoothing line (dashed) and regression line (solid).

An initial concern with these data is that the observations are likely not independent. Since they were taken consecutively, one waiting time might be related to the next waiting time – violating the independence assumption. As noted above, there might be two groups (types) of eruptions – short ones and long ones. The Normal QQ-Plot in Figure 6.18 also suggests a few observations creating a slightly long right tail. Those observations might warrant further exploration as they also show up as unusual in the Residuals vs Fitted plot. There are no highly influential points in the data set with all points having Cook's D smaller than 0.5 (contours are not displayed because no points are near or over them), so these outliers are not necessarily moving the regression line around. There are two distinct groups of observations but the variability is not clearly changing so we do not have to worry about non-constant variance here. So these results might be relatively trustworthy if we assume that the same relationship holds for all levels of duration of eruptions.

```
OF1 <- lm(Waiting ~ Duration, data = G2)
summary(OF1)
```

```
##
## Call:
```

```
## lm(formula = Waiting ~ Duration, data = G2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.6940  -4.4954  -0.0966   3.9544  29.9544
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  34.9452     1.1807   29.60   <2e-16
## Duration     10.7751     0.3235   33.31   <2e-16
##
## Residual standard error: 6.392 on 296 degrees of freedom
## Multiple R-squared:  0.7894, Adjusted R-squared:  0.7887
## F-statistic:  1110 on 1 and 296 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2,2))
plot(OF1)
```

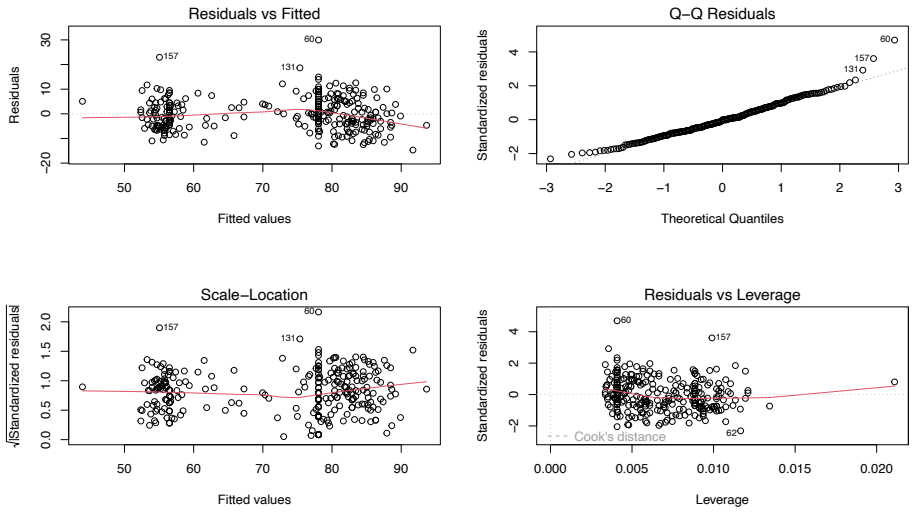

Figure 6.18: Diagnostic plots for Old Faithful waiting time model.

The estimated regression equation is $\widehat{\text{WaitingTime}}_i = 34.95 + 10.78 \cdot \text{Duration}_i$, suggesting that for a 1 minute increase in eruption `Duration` we would expect, on average, a 10.78 minute change in the `WaitingTime`. This equation might provide a useful tool for the YNP staff to predict waiting times. The $\mathbf{R^2}$ is fairly large: 78.9% of the variation in *waiting time* is explained by the *duration* of the previous eruption. But maybe this is more about two types of eruptions/waiting times? We could consider the relationship within the shorter and longer eruptions but since there are observations residing between the two groups, it is difficult to know where to split the explanatory variable into two groups. Maybe we really need to measure additional information that might explain why there are two groups in the responses...

6.8 Chapter summary

The correlation coefficient (***r*** or Pearson's Product Moment Correlation Coefficient) measures the strength and direction of the linear relationship between two quantitative variables. Regression models estimate the impacts of changes in x on the mean of the response variable y. Direction of the assumed relationship (which variable explains or causes the other) matters for regression models but does not matter for correlation. Regression lines only describe the pattern of the relationship; in regression, we use the coefficient of determination to describe the strength of the relationship between the variables as a percentage of the response variable that is explained by the model. If we are choosing between models, we prefer them to have higher R^2 values for obvious reasons, but we will discover in Chapter 8 that maximizing the coefficient of determination is not a good way to pick a model when we have multiple candidate options.

In this chapter, a wide variety of potential problems were explored when using regression models. This included a discussion of the conditions that will be required for using the models to perform trustworthy inferences in the remaining chapters. It is important to remember that correlation and regression models only measure the **linear** association between variables and that can be misleading if a nonlinear relationship is present. Similarly, influential observations can completely distort the apparent relationship between variables and should be assessed before trusting any regression output. It is also important to remember that regression lines should not be used outside the scope of the original observations – extrapolation should be checked for and avoided whenever possible or at least acknowledged when it is being performed.

Regression models look like they estimate the changes in y that are caused by changes in x, especially when you use x to predict y. This is not true unless the levels of x are randomly assigned and only then we can make causal inferences. Since this is not generally true, you should initially always assume that any regression equation describes the relationship – if you observe two subjects that are 1 unit of x apart, you can expect their mean to differ by b_1 – you should not, however, say that changing x causes a change in the mean of the responses. Despite all these cautions, regression models are very popular statistical methods. They provide detailed descriptions of relationships between variables and can be extended to situations where we are interested in multiple predictor variables. They also share ties to the ANOVA models discussed previously. When you are running R code, you will note that all the ANOVAs and the regression models are estimated using `lm`. The assumptions and diagnostic plots are quite similar. And in the next chapter, we will see that inference techniques look similar. People still like to distinguish among the different types of situations, but the underlying ***linear models*** are actually exactly the same...

6.9 Summary of important R code

The main components of the R code used in this chapter follow with the components to modify in lighter and/or ALL CAPS text where `y` is a response variable, `x` is an explanatory variable, and the data are in `DATASETNAME`.

- **DATASETNAME |> ggpairs()**
  - Requires the `GGally` package.
  - Makes a scatterplot matrix that also displays the correlation coefficients.
- **cor(y ~ x, data = DATASETNAME)**
  - Provides the estimated correlation coefficient between x and y.
- **plot(y ~ x, data = DATASETNAME)**
  - Provides a base R scatter plot.
- **DATASETNAME |> ggplot(mapping = aes(x = x, y = y) +**
  geom_point() +
  geom_smooth(method = "lm")
  - Provides a scatter plot with a regression line.

  - Add **color = groupfactor** to the **aes()** to color points and get regression lines based on a grouping (categorical) variable.
  - Add **+ geom_smooth(se = F, lty = 2)** to add a smoothing line to the scatterplot as a dashed line.

- **MODELNAME <- lm(y ~ x, data = DATASETNAME)**
  - Estimates a regression model using least squares.
- **summary(MODELNAME)**
  - Provides parameter estimates and R-squared (used heavily in Chapter 7 and 8 as well).
- **par(mfrow = c(2, 2)); plot(MODELNAME)**
  - Provides four regression diagnostic plots in one plot.

6.10 Practice problems

6.1. **Treadmill data analysis** These questions revisit the treadmill data set from Chapter 1. Researchers were interested in whether the run test variable could be used to replace the treadmill oxygen consumption variable that is expensive to measure. The following code loads the data set and provides a scatterplot matrix using `ggpairs` for all variables except for the subject identifier variable that was in the first column and was removed by `select(-1)`.

```
treadmill <- read_csv("http://www.math.montana.edu/courses/s217/documents/treadmill.csv")
library(psych)
treadmill |> select(-1) |> ggpairs()
```

6.1.1. First, we should get a sense of the strength of the correlation between the variable of primary interest, `TreadMillOx`, and the other variables and consider whether outliers or nonlinearity are going to be major issues here. Which variable is it most strongly correlated with? Which variables are next most strongly correlated with this variable?

6.1.2. Fit the SLR using `RunTime` as explanatory variable for `TreadMillOx`. Report the estimated model.

6.1.3. Predict the treadmill oxygen value for a subject with a run time of 14 minutes. Repeat for a subject with a run time of 16 minutes. Is there something different about these two predictions?

6.1.4. Interpret the slope coefficient from the estimated model, remembering the units on the variables.

6.1.5. Report and interpret the y-intercept from the SLR.

6.1.6. Report and interpret the R^2 value from the output. Show how you can find this value from the original correlation matrix result.

6.1.7. Produce the diagnostic plots and discuss any potential issues. What is the approximate leverage of the highest leverage observation and how large is its Cook's D? What does that tell you about its potential influence in this model?