Statistics 213 – 3.2: Regression

UNIVERSITY OF CALGARY

Textbook:

11.1, 11.2, 11.3, 11.5

Objectives:

- Be able to create and interpret a ***scatterplot***
- Be able to calculate and interpret the ***correlation coefficient***
- Be able to calculate and interpret the ***coefficient of determination***
- Be able to run a regression analysis in R and appropriately interpret the output
- Be able to make predictions using a regression equation
- Be able to calculate ***residuals***
- Be able to check regression model assumptions

Motivation:

In the online module (the 3.1 notes), we focused in part on exploratory data analysis (EDA). The techniques we learned in that set of notes are used to describe the behavior of a variable in a given sample.

Examples (from the 3.1 notes)

- Letter grades on a midterm for a sample of $n = 107$ students
- Eye colors for a sample of $n = 208$ students
- Heights (in inches) for a sample of $n = 500$ MLB players

In the above examples, notice how we only focused on one variable at a time—letter grade, eye color, or height. While it is important to be able to describe individual variables in a given sample, there are many situations in which the relationship between two variables might be of more interest.

Examples

- How can we describe the relationship between height and weight in high school students?
- To what degree are temperature and rainfall levels related?
- Is there an association between the duration of an eruption of the Old Faithful geyser and the amount of time between eruptions?

In this unit, we will discuss three main ways of describing and summarizing the relationship between two variables: using scatterplots, computing the correlation coefficient, and determining the regression equation.

Bivariate Data

When we refer to ***bivariate data***, we're referring to information collected on two variables from the same set or sample of observations.

Example

Suppose we were interested in the relationship between height and weight in high school students. We take a sample of $n = 43$ students and measure each students' height and weight.

In this case, we have two variables of interest: height and weight. Each individual in our sample is measured on both of these variables. We could keep track of this information in a spreadsheet as follows:

Student	Height	Weight
A	73	195
B	69	135
C	70	145
D	72	170
E	73	172
F	69	168
G	68	155

Notice that these measurements of height and weight are "paired" in the sense that there is one height and one weight for each of the n students.

Explanatory vs. Response Variables

As we saw in the 3.1 notes, when researchers are interested in the relationship between two variables, they often suspect that one variable may be responsible (at least in part) for some of the change in the other variable. Thus, when we're looking at the relationship between two variables, we often consider one as the "explainer" and one as the "responder." Recall these definitions from the 3.1 notes:

An ***explanatory variable*** (***independent variable***, ***predictor variable***, ***x-variable***) is one that may explain or cause some degree of change in another variable.

A ***response variable*** (***dependent variable***, ***y-variable***) is a variable that changes—at least in part—due to the changes in the explanatory variable.

We'll return to explanatory and response variables in a little bit. The rest of the notes will focus on three main ways of describing and summarizing the relationship between two variables, starting with scatterplots.

Scatterplots

As we saw in the last set of notes, the quickest and easiest way to get an idea of the behavior of a variable is to generate some sort of picture or graph of it.

A ***scatterplot*** is a two-dimensional plot with one variables' values plotted along the horizontal axis (x-axis) and the other variables' values plotted along the vertical axis (y-axis). It is a good way to visualize the relationship between two variables.

It is accepted practice to plot the explanatory variable (x-variable) along the x-axis and the response variable (y-variable) along the y-axis.

Example 3.2.1

The heights (in inches) and weights (in pounds) were recorded for a sample of $n = 43$ high school students. The following R code shows how these values are read into R and then displayed as a data frame called "students":

RStudio

```
heights = c(73, 69, 70, 72, 73, 69, 68, 71, 71, 68, 69, 67, 66, 67, 72, 68, 75, 68, 73, 72, 72,
72, 72, 74, 68, 73, 68, 70, 72, 70, 67, 67, 71, 72, 73, 68, 72, 68, 67, 70, 71, 70, 67)

weights = c(195, 135, 145, 170, 172, 168, 155, 185, 175, 158, 185, 146, 135, 150, 160, 155, 230,
149, 240, 170, 198, 163, 230, 170, 151, 220, 145, 130, 160, 210, 145, 185, 237, 205, 147, 170, 1
81, 150, 150, 200, 175, 155, 167)

students = data.frame(heights, weights)
head(students)
```

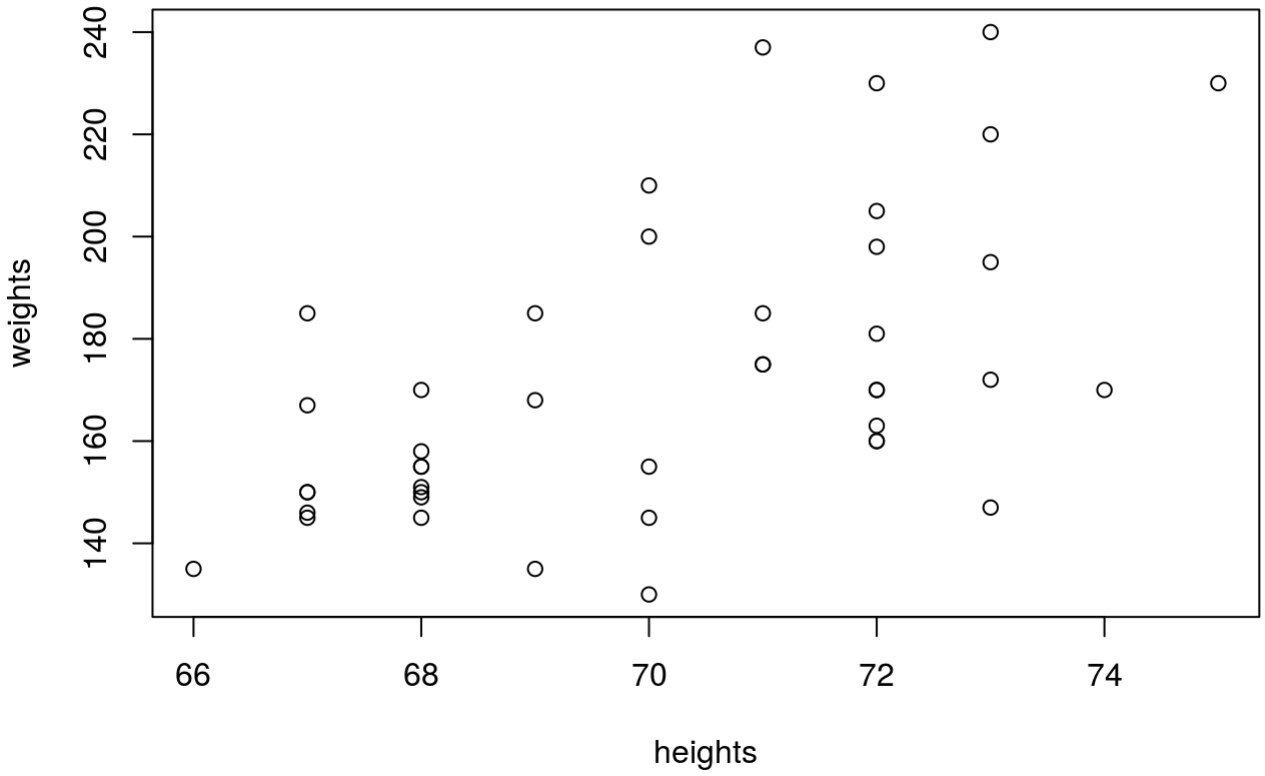
```
##   heights weights
## 1      73     195
## 2      69     135
## 3      70     145
## 4      72     170
## 5      73     172
## 6      69     168
```

To make a scatterplot of the data, we use the R function `plot(y~x)`.

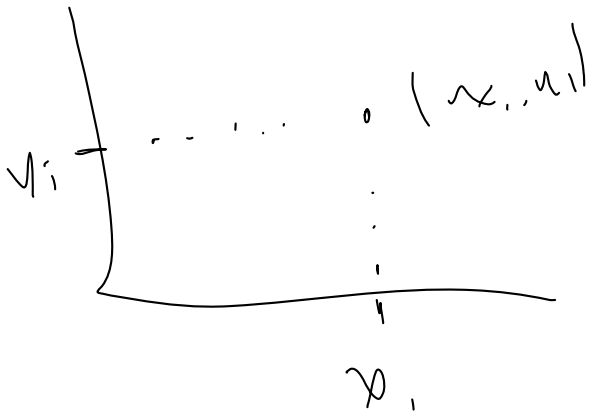
where `y` is the variable whose values are to be plotted along the y-axis and `x` is the variable whose values are to be plotted along the x-axis.

RStudio

```
plot(weights~heights)
```



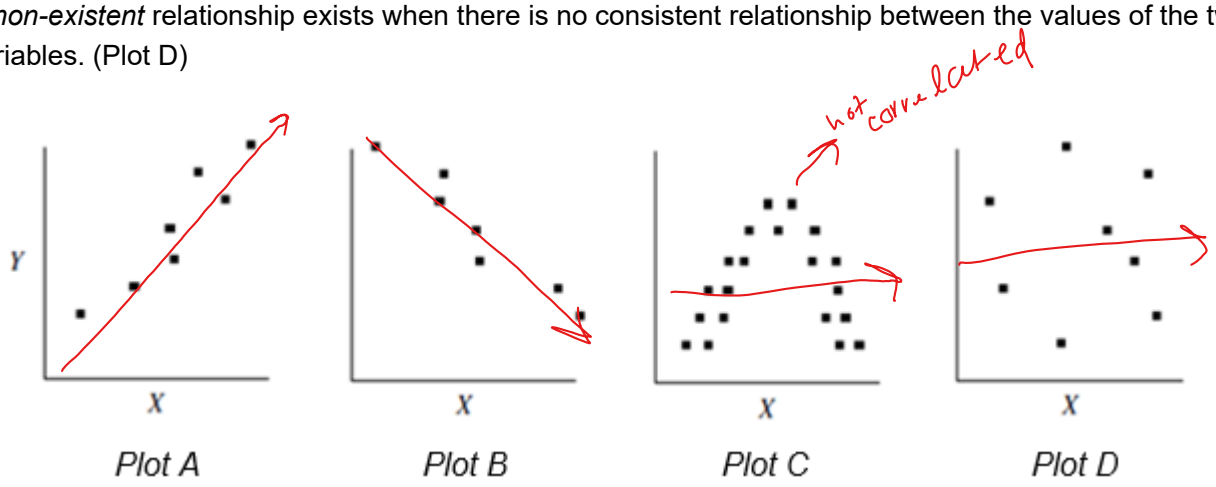
Notice that there is a point on the scatterplot for each of the $n = 43$ students. The points represent each pair of measurements (height, weight) for each student.



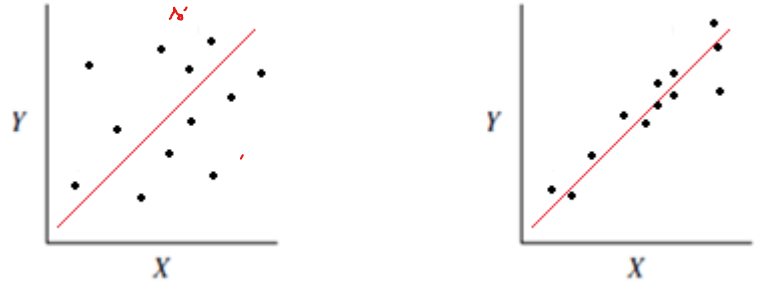
As you can see, a scatterplot can give a general idea of the relationship between the two variables. In general, we like to describe this relationship in terms of both its *direction* and its *strength*.

The *direction* of a relationship can generally be described as positive, negative, curvilinear, or non-existent.

- A *positive* linear relationship exists when, as the values of one of the variables increase, the values of the other generally increase as well. (Plot A)
- A *negative* linear relationship exists when, as the values of one of the variables increase, the values of the other generally decrease. (Plot B)
- A *curvilinear* relationship exists when the relationship between the variables' values can best be descried by a curve (rather than a line). (Plot C)
- A *non-existent* relationship exists when there is no consistent relationship between the values of the two variables. (Plot D)



The *strength* (or *magnitude*) of a relationship is based on how tightly the "cloud" of points is clustered about the trend line (the line/curve that best describes the direction of the relationship. The more tightly clustered the cloud the stronger the relationship is between the two variables.



Example 3.2.1 (revisited)

The relationship between height and weight in the example with $n = 43$ high school students appears to be a moderately strong positive linear relationship.

Correlation

Using a scatterplot is a good way to get a quick general idea of the relationship between two variables, but what if we wanted some way to quantify this relationship beyond the subjective interpretation of a scatterplot?

Pearson's correlation coefficient is a measure of the direction and strength of a linear relationship between two quantitative variables x and y. The sample correlation coefficient is usually denoted r and is computed as:

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{x}\,\overline{y}}{(n-1)s_x s_y}$$

where x_i is the i^{th} observation of the x variable, y_i is the i^{th} observation of the y variable, $\overline{x}$ and $\overline{y}$ are the means of the x and y variables, respectively, s_x and s_y are the standard deviations of the x and y variables, respectively, and n is the sample size.

The following are important features of the correlation coefficient:

- The distinction between the explanatory variable and the response variable doesn't matter in the calculation or interpretation of r
- $-1 \leq r \leq 1$
- The sign of r indicates the direction of the linear relationship
  - $+r$ indicates a positive relationship
  - $-r$ indicates a negative relationship
- Values of r closer to -1 or 1 suggest a strong linear relationship (with $r = 1$ and $r = -1$ representing perfect positive and negative correlations, respectively); values of r closer to 0 suggest a weak linear relationship

To compute correlation in R, we use the R function `cor(y~x)`.

Example 3.2.1 (Revisited)

The heights (in inches) and weights (in pounds) were recorded for a sample of $n = 43$ high school students. Use R to compute the correlation coefficient for the heights and weights of these students. Interpret this value.

RStudio

```
cor(weights~heights)
```

```
## [1] 0.5684901
```

Example 3.2.2

The annual rainfall (in mm) and maximum daily temperature (in Celsius) were recorded for $n = 11$ different locations in Mongolia.

```
rain = c(196, 196, 179, 197, 149, 112, 125, 99, 125, 84, 115)

temperature = c(5.7, 5.7, 7, 8, 8.5, 10.7, 11.4, 10.9, 11.4, 11.4, 11.4)
```

Use R to compute the correlation coefficient for the rainfall and temperatures of these locations. Interpret this value.

RStudio

```
cor(rain~temperature)
```

```
## [1] -0.918617
```

Coefficient of Determination

Another way to examine a bivariate relationship is to measure the contribution of the explanatory variable in predicting the value of the response variable.

The ***coefficient of determination*** is the squared correlation coefficient (r^2 or R^2) and represents the proportion of the total sample variation in the response variable that can be explained by its linear relationship with the explanatory variable.

Note that since $-1 \leq r \leq 1, 0 \leq r^2 \leq 1$. Also note that you need to know which variable is being treated as the explanatory variable and which variable is being treated as the response variable in order to interpret r^2.

Example 3.2.3

In the height/weight example involving the $n = 43$ high school students, what percentage of variation in weight can be explained by its linear relationship with height?

Regression

We can get even more detailed than correlation when describing the relationship between two variables. ***Regression*** is a technique that allows us to not only summarize a linear relationship but to make predictions about future values that have yet to be observed.

In this class, we will focus on ***simple linear regression***, which is a predictive model that describes the relationship between our explanatory variable and our response variable. Simple linear regression fits (or models) the prediction of the response variable from its relationship with the explanatory variable. This is done by observing measures on both an explanatory variable and a response variable and "fitting" a line that best describes the relationship between the two variables.

The ***regression equation*** is what defines the straight line that best describes how the values of the response variable are related, on average, to the values of one explanatory variable. The ***regression line*** itself is just the name of the line defined by the regression equation—the line of "best fit."

The Regression Equation and Regression Line

Let's look at the heights/weights example with the $n = 43$ high school students again. Suppose we wanted an equation that tells us how best to predict weight (y) given a specific height (x). In this case, we've got a set of data that is made up of 43 (x, y) points. We can use this data to come up with an equation of a line that "best fits" the relationship between height and weight.

Recall the equation for a straight line:

$$y = mx + b$$

where:

y = a value of the y-variable m = slope b = y-intercept x = a value of the x-variable

Rearrange the right-hand side (and change the symbols) and you have our regression equation:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

where:

$\hat{y}$ = the predicted mean y value for a given x $\hat{\beta}_0$ = an estimate of the y-intercept based on the data $\hat{\beta}_1$ = an estimate of the slope based on the data x = a value of our explanatory variable

The value of the ***y-intercept*** in a regression context has the same meaning as the value of the y-intercept in a general math context. That is, it is the average value of the y-variable when $x = 0$. In some scenarios, the y-intercept has a meaningful interpretation (e.g., the growth rate of a tree when temperature = 0). However, in other cases, the interpretation is meaningless and/or nonsensical (e.g., the weight of a child when height = 0). It depends on the variables!

A much more meaningful value in the context of regression is the ***slope***. In regression, the value of the slope has a slightly more specific meaning than it does in a general math context. In math, the slope is defined as the change in the y-variable compared to the change in the x-variable. Some may be familiar with the phrase “rise over run.”

In regression, it’s the same idea—however, it’s more specific. The regression slope is defined as the average change in the y-variable for every unit change in the x-variable. Think of it as “rise over *one*.”

While we’ll mostly be relying on Minitab to do our regression calculations for us, the following are the equations used to obtain the sample slope and y-intercept values.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = r\frac{s_y}{s_x}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

To run a regression analysis on an explanatory variable x and a response variable y in r, we use the function `lm(y~x)`. *Make sure your response variable is listed first, otherwise you will not get the correct regression output!*

Example 3.2.4

The annual rainfall (in mm) and maximum daily temperature (in Celsius) were recorded for $n = 11$ different locations in Mongolia.

```
rain = c(196, 196, 179, 197, 149, 112, 125, 99, 125, 84, 115)

temperature = c(5.7, 5.7, 7, 8, 8.5, 10.7, 11.4, 10.9, 11.4, 11.4, 11.4)
```

A regression analysis was performed to express annual rainfall (“rain”) as a linear function of maximum daily temperature (“temperature”).

RStudio

```
fit = lm(rain~temperature)

fit
```

```
##
## Call:
## lm(formula = rain ~ temperature)
##
## Coefficients:
## (Intercept)  temperature
##      295.25       -16.36
```

1. Write down the regression equation. Interpret the slope and the y-intercept.

2. What percentage of variation in annual rainfall is explained by its linear relationship with temperature?

Example 3.2.5

Old Faithful is a popular geyser in Yellowstone National Park that is famous for its consistent eruptions. The duration (length) of the geyser's eruptions (in minutes) as well as the amount of time spent waiting after the previous eruption (in minutes) were recorded for $n = 144$ eruptions.

```
duration = c(3.3, 3.3, 3.3, 3.4, 3.5, 3.5, 3.7, 3.7, 3.7, 3.7, 3.8, 3.8, 3.8, 3.9, 3.9, 3.9, 3.9
, 3.9, 3.9, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4.1, 4.1, 4.1, 4.1, 4.1, 4.1, 4.1, 4.2, 4.2, 4.2, 4.2, 4.
2, 4.2, 4.2, 4.2, 4.2, 4.2, 4.2, 4.2, 4.2, 4.2, 4.3, 4.3, 4.3, 4.3, 4.3, 4.3, 4.3, 4.3, 4.3, 4.3
, 4.3, 4.3, 4.3, 4.3, 4.3, 4.4, 4.4, 4.4, 4.4, 4.4, 4.4, 4.4, 4.4, 4.4, 4.4, 4.4, 4.4, 4.4, 4.5,
4.5, 4.5, 4.5, 4.5, 4.5, 4.5, 4.5, 4.5, 4.5, 4.5, 4.5, 4.5, 4.5, 4.5, 4.5, 4.5, 4.5, 4.5, 4.5,
4.5, 4.6, 4.6, 4.6, 4.6, 4.6, 4.6, 4.6, 4.6, 4.6, 4.6, 4.7, 4.7, 4.7, 4.7, 4.7, 4.7, 4.7, 4.7,
4.7, 4.7, 4.7, 4.7, 4.7, 4.8, 4.8, 4.8, 4.8, 4.8, 4.8, 4.8, 4.8, 4.8, 4.8, 4.9, 4.9, 4.9, 4.9, 5
, 5, 5, 5, 5, 5, 5.1, 5.3, 5.5)

waiting = c(74, 74, 74, 73, 85, 73, 74, 71, 84, 73, 71, 94, 82, 69, 77, 76, 77, 84, 77, 71, 92,
79, 75, 88, 74, 87, 89, 76, 78, 86, 80, 76, 77, 80, 85, 84, 91, 80, 73, 81, 71, 83, 88, 76, 68,
93, 82, 86, 68, 77, 87, 89, 81, 87, 79, 72, 72, 87, 78, 84, 89, 89, 87, 93, 86, 73, 81, 75, 78,
79, 92, 87, 87, 76, 93, 75, 87, 76, 75, 80, 84, 79, 80, 88, 88, 80, 81, 72, 90, 80, 81, 93, 91,
79, 80, 84, 86, 96, 93, 92, 78, 80, 85, 82, 78, 87, 72, 73, 83, 81, 74, 82, 75, 87, 85, 80, 75,
83, 84, 84, 78, 81, 82, 87, 93, 96, 78, 84, 78, 85, 87, 90, 76, 88, 98, 85, 93, 79, 89, 84, 83,
89, 77, 89)
```

A regression analysis was performed to express eruption duration ("duration") as a linear function of waiting time ("waiting").

RStudio

```
fit = lm(duration~waiting)

fit
```

```
##
## Call:
## lm(formula = duration ~ waiting)
##
## Coefficients:
## (Intercept)      waiting
##     2.78244      0.01948
```

1. Write down the regression equation. Interpret the slope.

2. Use R to compute the correlation. interpret this value.

```
cor(duration~waiting)
```

```
## [1] 0.3294965
```

Predicting Using the Regression Equation

We can get a general summary of our bivariate relationship by examining the regression equation and knowing how to interpret the y-intercept and the slope. We can also use our regression equation to predict the value of the response variable (y) for a specific value of the explanatory variable (x).

A predicted i^{th} value of our response variable, denoted $\hat{y}_i$, can be obtained by plugging in the specific i^{th} explanatory variable value (x_i) into the regression equation and solving for $\hat{y}_i$.

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_i x_i$$

Note: be wary of extrapolation! ***Extrapolation*** involves using the regression equation to predict y values for x values outside of the original range of the data set.

Example 3.2.6

The heights (in inches) and weights (in pounds) were recorded for a sample of $n = 43$ high school students.

```
heights = c(73, 69, 70, 72, 73, 69, 68, 71, 71, 68, 69, 67, 66, 67, 72, 68, 75, 68, 73, 72, 72,
72, 72, 74, 68, 73, 68, 70, 72, 70, 67, 67, 71, 72, 73, 68, 72, 68, 67, 70, 71, 70, 67)

weights = c(195, 135, 145, 170, 172, 168, 155, 185, 175, 158, 185, 146, 135, 150, 160, 155, 230,
149, 240, 170, 198, 163, 230, 170, 151, 220, 145, 130, 160, 210, 145, 185, 237, 205, 147, 170, 1
81, 150, 150, 200, 175, 155, 167)
```

A regression analysis was performed to express weight ("weights") as a linear function of height ("heights").

RStudio

```
fit = lm(weights~heights)

fit
```

```
##
## Call:
## lm(formula = weights ~ heights)
##
## Coefficients:
## (Intercept)      heights
##    -317.919        6.996
```

1. Predict the weight of a student who is 70 inches tall.

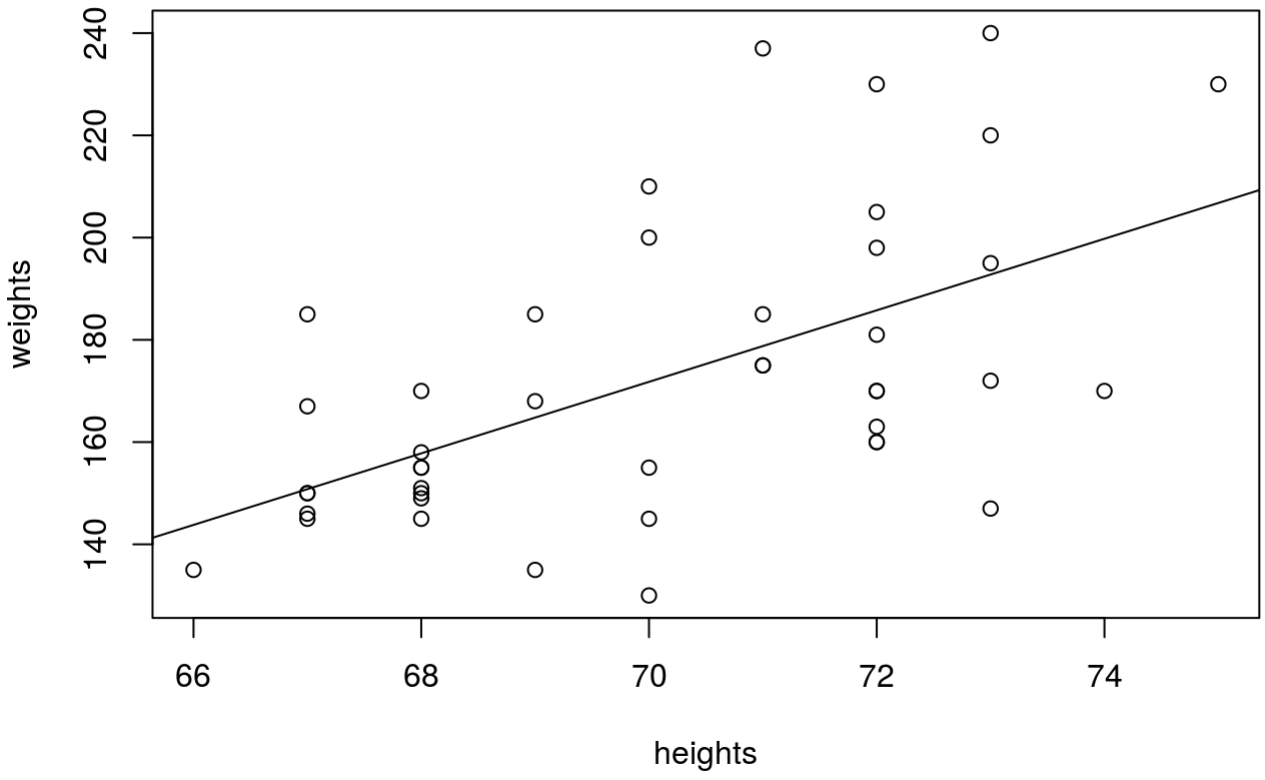
2. Predict the weight of a student who is 62 inches tall.

Residuals

Let's look again at the scatterplot of heights and weights for the n = 43 high school students, except this time let's also plot the regression line.

R Studio

```
plot(weights~heights)
abline(fit)
```



Notice that the regression line isn't perfect! For example, our calculation for the weight of a student who is 70 inches tall is not the same as any of the weights of the 70-inch-tall students in our actual sample.

In other words, there exist prediction errors in our estimation. A ***prediction error*** (or ***residual***) is the difference between the observed y value (y) and the predicted y value ($\hat{y}$) for any given x value. For the i^{th} x value, the residual e_i is calculated as:

$$e_i = y_i - \hat{y}_i$$

Example 3.2.7

In our height and weight example, a 70-inch tall student weighed 130 pounds. Compute this student's residual.

So you may be wondering: "if the regression line isn't perfect, why is there a unique regression equation/line for any given problem?"

Our method of calculating the regression line involves the ***method of least squares***. The regression line is the line that minimizes the total (summed) squared residuals. In other words, it is the line for which

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

is minimized. The regression line is therefore the *best possible linear explanation* for the relationship between the two variables.

Regression Model Assumptions

As you can see, regression is quite powerful. It allows us to describe the linear relationship between two variables and make predictions. However, in order for a regression analysis to be accurate, a few assumptions must be met:

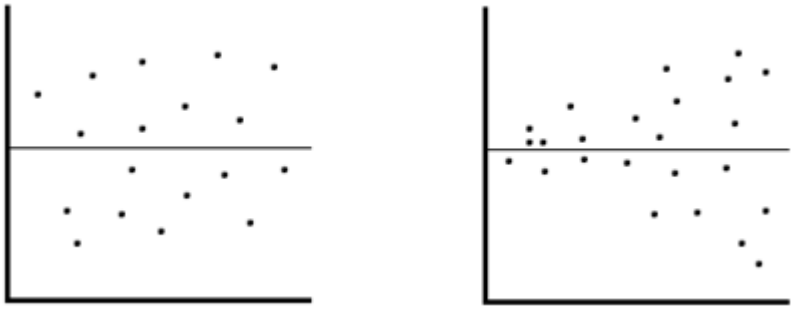
1. The assumption of ***constant variance*** (***homoscedasticity***): the variance of the residuals is constant. That is, the variance e_i values is the same regardless of the value of x_i.
2. The assumption of ***normality***: the distribution of the residuals is normal.

Checking Model Assumptions

In this class, we can focus on how to check these assumptions using plots.

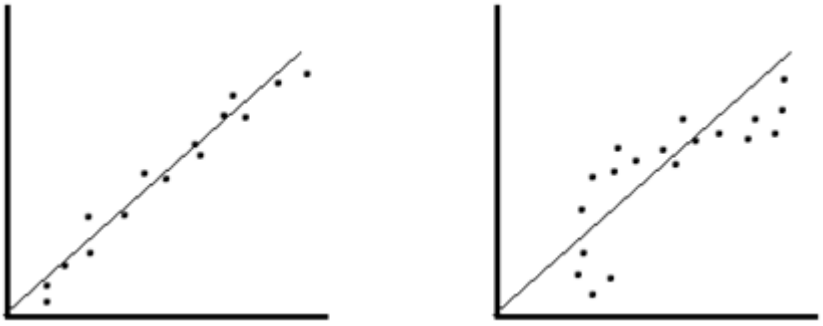
Checking for constant variance (homoscedasticity): look at a residual (or “residuals vs. fits”) plot.

- Good: The points appear fairly uniformly scattered about the flat dotted line. This suggests homoscedasticity.
- Bad: The points either gradually fan out from the line or gradually condense about the line. This suggests that the variance is not constant (***heteroscedasticity***).



Checking for normality: look at a normal probability plot.

- Good: The points appear in a straight (or near-straight) line and follow the diagonal line. This suggests normality.
- Bad: The points deviate from the diagonal line. This suggests non-normality.



Suppose you have a regression analysis that you have named `fit`. To create the normal probability plot and residual plot, you would use the R function `plot(fit)`.

Example 3.2.8

Recall the heights (in inches) and weights (in pounds) that were recorded for a sample of $n = 43$ high school students.

```
heights = c(73, 69, 70, 72, 73, 69, 68, 71, 71, 68, 69, 67, 66, 67, 72, 68, 75, 68, 73, 72, 72,
72, 72, 74, 68, 73, 68, 70, 72, 70, 67, 67, 71, 72, 73, 68, 72, 68, 67, 70, 71, 70, 67)

weights = c(195, 135, 145, 170, 172, 168, 155, 185, 175, 158, 185, 146, 135, 150, 160, 155, 230,
149, 240, 170, 198, 163, 230, 170, 151, 220, 145, 130, 160, 210, 145, 185, 237, 205, 147, 170, 1
81, 150, 150, 200, 175, 155, 167)
```

A regression analysis was performed to express weight ("weights") as a linear function of height ("heights").

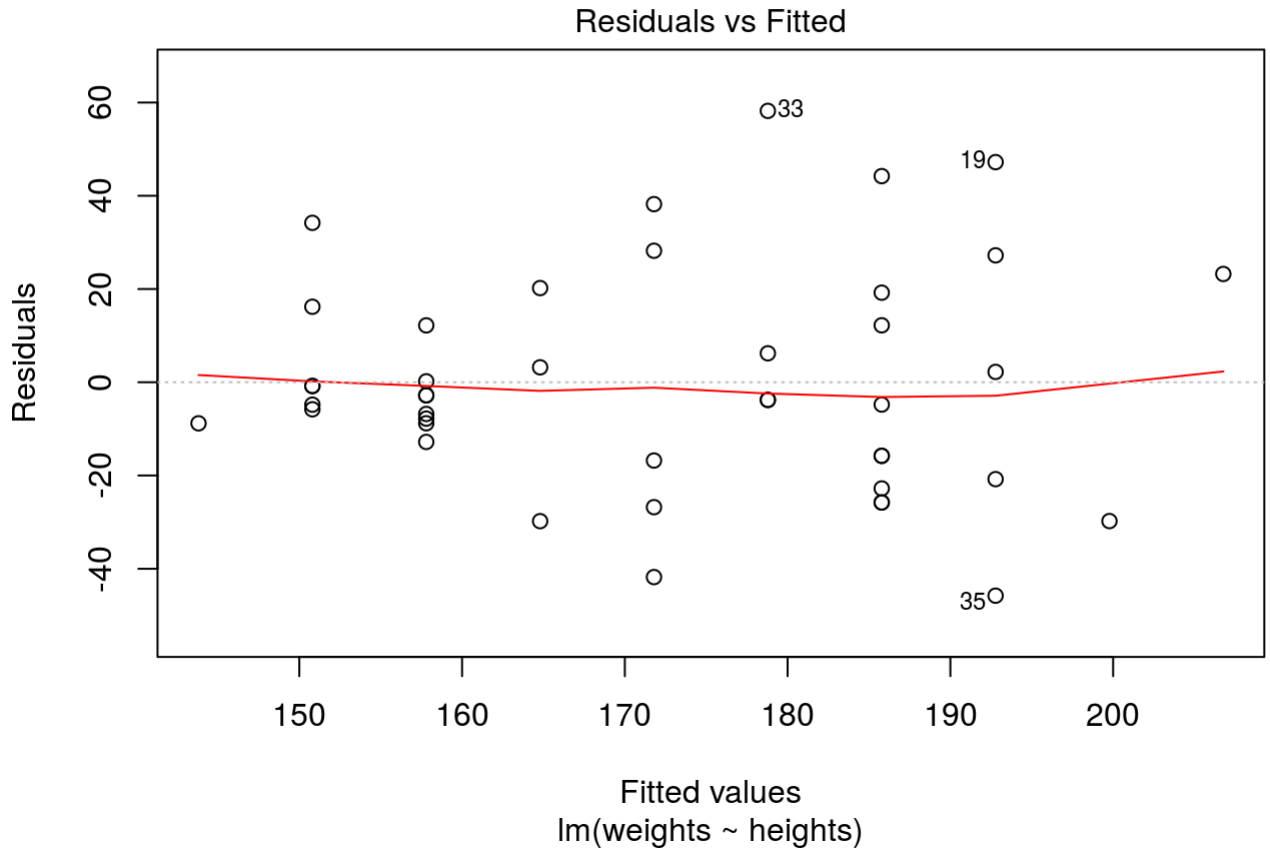
RStudio

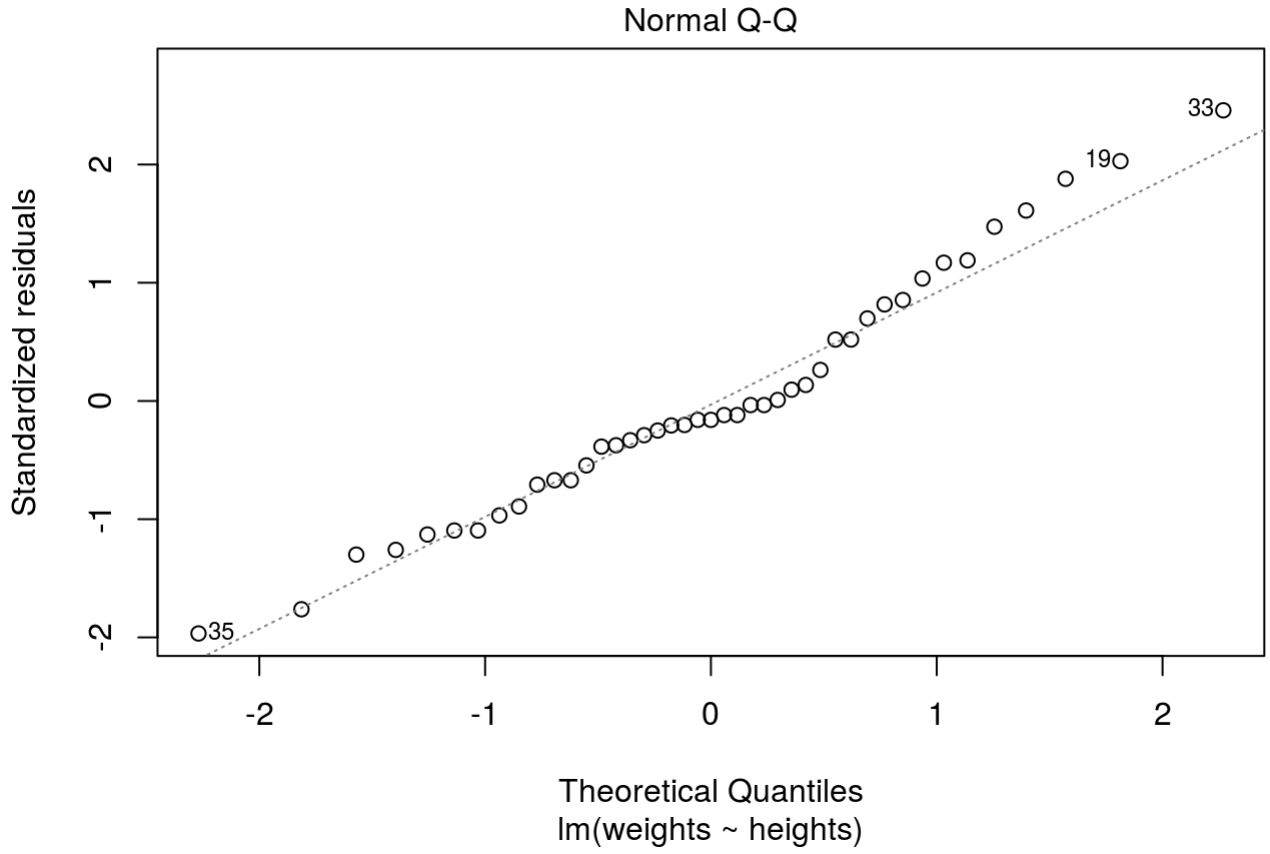
```
fit = lm(weights~heights)

fit
```

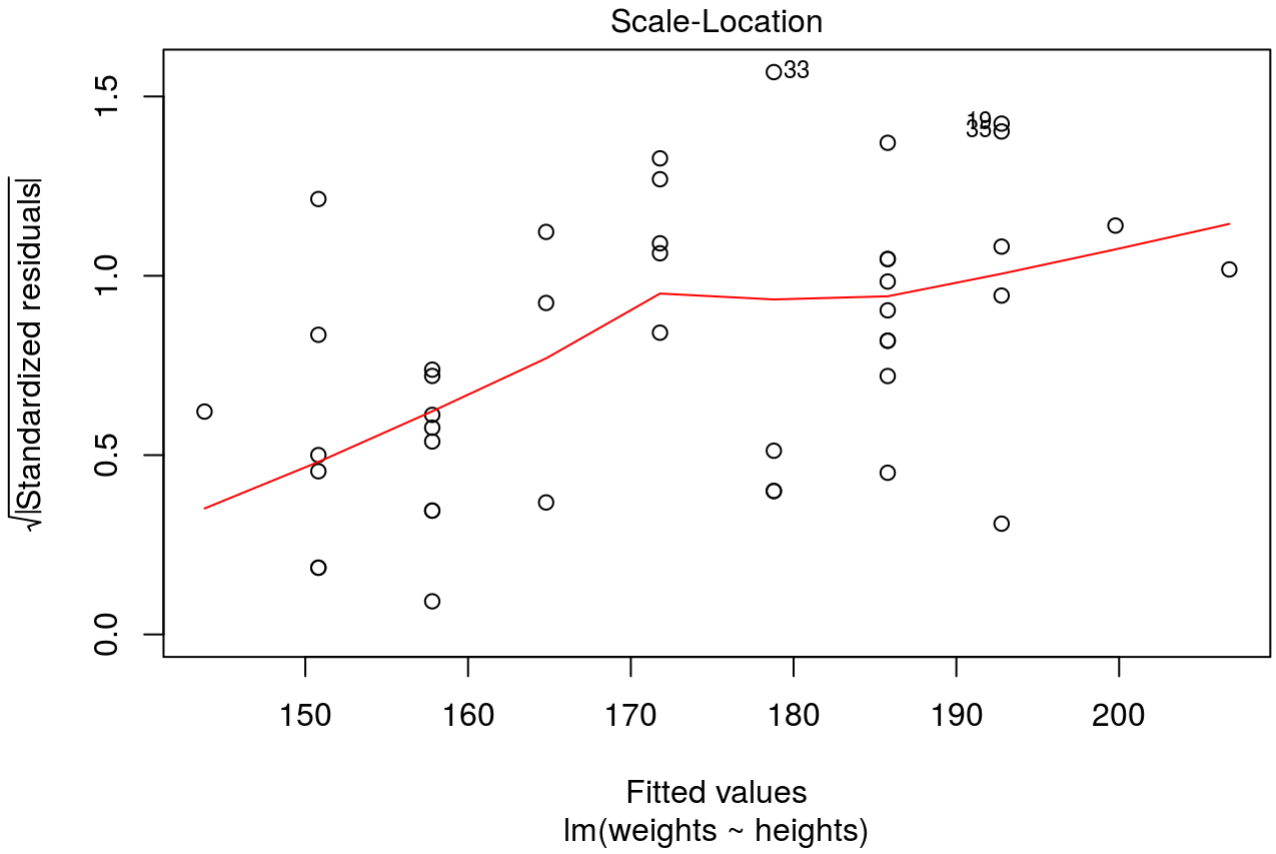
```
##
## Call:
## lm(formula = weights ~ heights)
##
## Coefficients:
## (Intercept)      heights
##    -317.919        6.996
```

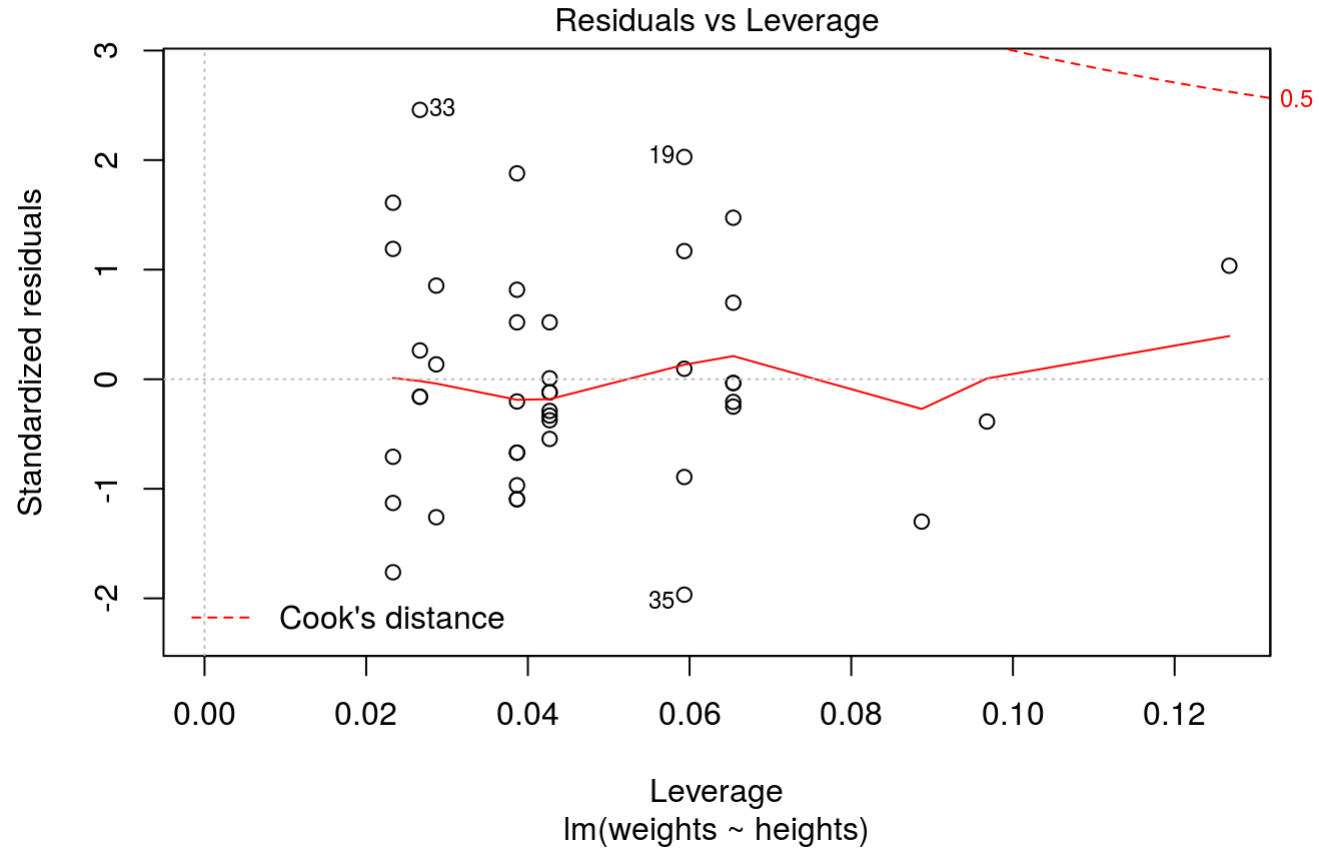
Use the residual plot and the normal probability plot to assess the assumptions of homoscedasticity and normality.

```
plot(fit)
```


Residuals vs Fitted
Residuals
Fitted values
lm(weights ~ heights)
33
19
35
60
40
20
0
-20
-40
150
160
170
180
190
200


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(weights ~ heights)
33
19
35
2
1
0
-1
-2


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
33
19
35
150
160
170
180
190
200
Fitted values
lm(weights ~ heights)


Residuals vs Leverage
Standardized residuals
-2
-1
0
1
2
3
33
19
35
0.5
Cook's distance
0.00
0.02
0.04
0.06
0.08
0.10
0.12
Leverage
lm(weights ~ heights)