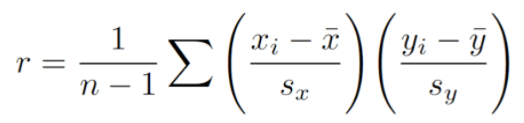
**Linear Regression and Correlation Coefficients**

**The correlation coefficient** (17.2 in ebook, Chapter 5 of Hawkes)

The correlation coefficient is defined for a list of *n* pairs of data points (xi,yi):



Here, sx is the standard deviation of x.

The distance each point xi is away from the mean () is xi.

So each term in parentheses is the deviation from the mean,

measured in standard deviations.

Why does this represent correlation between x and y?

* r indicates to what degree 2 variables “move together” (co-vary)

Go to <http://guessthecorrelation.com/>

In RStudio, under *Tools*, use the “*Install Packages*…” feature to install HistData. Then load the packages:

library(tidyverse)

library(HistData)

You can see all the datasets available by using the command.

help(HistData)

We’re going to work with the dataset GaltonFamilies.

data("GaltonFamilies") #loads GaltonFamilies as a dataframe

* Inspect GaltonFamilies using the head() function

Now suppose the goal of your final project was to determine what influence the parents’ heights have on their son’s height. Your two independent variables are mother’s height and father’s height, and the dependent variable will be: \_\_\_\_\_\_\_\_\_\_\_\_\_

We are going to do separate linear regressions for each independent variable, and determine (a) which has a stronger correlation with the son’s height, and (b) to what extent we can predict the son’s heights given either of the parent’s heights.

#filter out the female children

m\_heights <- GaltonFamilies %>% filter(gender == "male")

* Inspect m\_heights

#randomly choose one son from each family to put into our dataset

m\_heights <- m\_heights %>%

group\_by(family) %>%

sample\_n(1) %>% ungroup()

* Inspect m\_heights

# select just the headers that we want and rename childHeight to son

m\_heights <- m\_heights %>%

select(father, mother, childHeight) %>%

rename(son = childHeight)

* Inspect m\_heights

**How to scatter-plot with dataframes (requires tidyverse)**

m\_heights %>% ggplot(aes(father, son)) +

geom\_point()

Calculate the correlation coefficient using the cor() function

x <- m\_heights$father

y <- m\_heights$son

r <- cor(x,y)

* Determine which is more strongly correlated to the son’s height: the father’s height, or the mother’s?

Now we will calculate the slope and intercept of the regression line.

mu\_x <- mean(x)

mu\_y <- mean(y)

s\_x <- sd(x)

s\_y <- sd(y)

m\_heights %>%

ggplot(aes(father, son)) + geom\_point() +

geom\_abline(slope = r\*s\_y/s\_x, intercept = mu\_y - r\*s\_y/s\_x \*mu\_x)

One can also ask a different question: How much of the variation in the son’s height can be explained by your independent variable of choice (eg, the father’s height)? Or to put it another way, if you knew only one piece of information (the height of the father), how well can you predict the height of the son?

The degree to which the father’s height determines the son’s height is captured in the R2 value, which is included in the summary() function:

fit <- lm(y~x)

summary(fit)

Look for the output that specifies *Multiple R-squared: 0.1611*

(Since we are only dealing with one independent variable at a time, ignore the *Adjusted R-Squared*.)

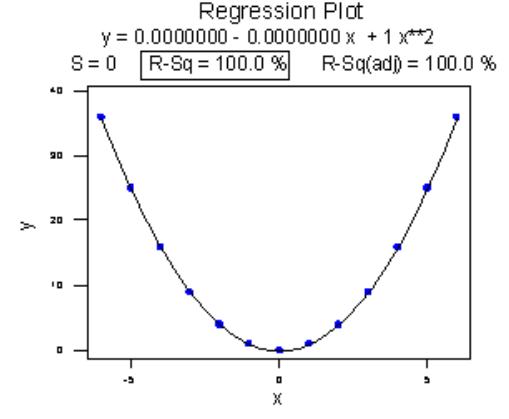
This tells us that 16% of the variation in the sons’ heights is determined by their fathers’ heights.

* Which is more predictive of the son’s height: the father’s height, or the mother’s? use the lm() function.

**The difference between *correlation coefficient* (r) and the *coefficient of determination* (**R2**)**

You might have noticed that for the son/father fit, and also for the mother/son fit, if you square the r value, you get the R2 value. This doesn’t happen in all situations for all models, but it happens here because the x and y data are roughly normally distributed (“bivariate normal”), so the dependent variable can be predicted through a linear model.

Caution: The correlation coefficient from linear regression tells you only the degree of *linear* correlation.



This graph was generated by plotting a parabolic dataset, y = x^2. Linear regression would create a line of best fit that is horizontal, with r = 0. And yet, the curved fit shown in the model is completely predictive of y (R2 = 100%). In this situation, r is telling you that the data have no linear relationship. Photo credit: https://online.stat.psu.edu/stat462/node/98/

**Linear Regression the Easy Way**

You might notice that lm() does all the fitting for you – its calculation of the slope and intercept should be the same as the linear model that you put into geom\_abline. Check this.

Since lm() does the fitting for you, you can skip all the typing into ggplot and simply do this:

#must specify first what x and y are

ggplot(m\_heights, aes(x, y)) +

geom\_point() + stat\_smooth(m\_heights = "lm", col = "red")

You don’t even have to define the means and standard deviations; lm() does all of that for you. Just specify what x and y mean and you are good to go.

**Model Validation with Q-Q Plots**

In the following, we are going to use a built-in R function that helps you determine if data is normally distributed. A “Q-Q Plot” will be a straight line if the data has a normal distribution. Let’s use the father’s heights from Galton families as an example:

qqnorm(m\_heights$father) #is this a straight line?

This qqnorm() function will be useful for the next question:

**Diagnostics**: **is a linear, single-variable model *justified*?**

You can always fit data to a line, as we did with father/son heights. But Is the child’s height *actually predicted by a linear model?*

There are 4 or 5 common ways to test the validity of a linear fit. Here, I just present the Q-Q Plot method. Let’s suppose that your R2 is 60%.

This means that 60% of the variance in y is:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

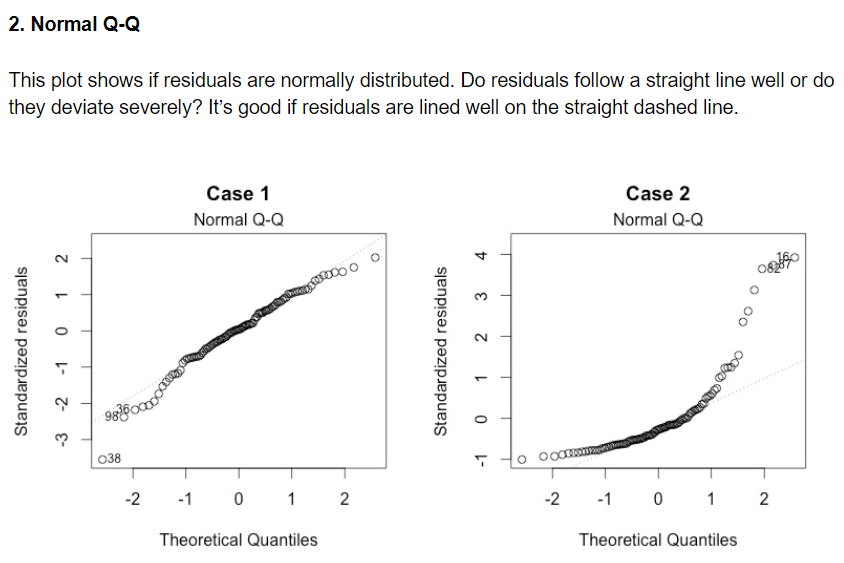
If the other 40% of the variation (the **residuals**) in y is explained by random fluctuation, then your linear model is valid. But if there exists another variable that explains a significant amount of that missing 40%, then your data should not be described by a single-variable linear model.

Or, if you could explain a lot more of the variance with a non-linear model, then the linear model is not a valid description of the relationship between the variables.

How can you tell if the residuals are due to another variable, or due to random chance?

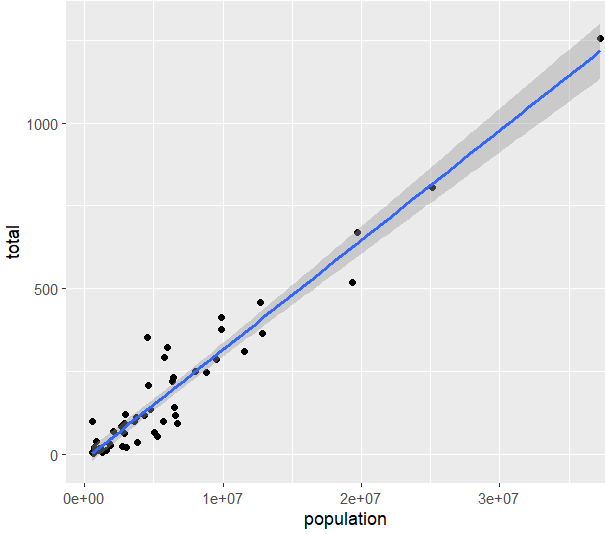
We know what random variation looks like. It looks normal. So we should capture these “residuals” using resid() and see if they are normally distributed.

# assumes you have already defined *fit* using a lm() function

resid\_list= resid(fit)

qqnorm(resid\_list)

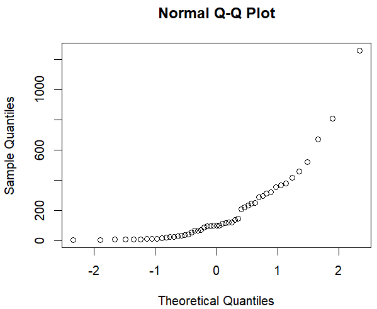
The following image is borrowed from <https://data.library.virginia.edu/diagnostic-plots/>

DSA 225|O’Neil Linear Regression: Prediction

The number of murders in a year plotted against the population of a state looks like this:

Note that the x-axis is using e+07 to mean 10,000,000. So we see markings for 10 million, 20 million, 30 million.

1. Suppose a state had a population of 1.5e+07, or in other words 15 million people. Use the graph to predict the number of murders per year.
2. Write an equation that relates the total number of murders (t) to the population (p). Hint: y = mx+ b. Read off m and b from the graph above. Recall that slope is equal to rise over run.

3. Linear regression is valid when the data is normally distributed. Normally distributed data looks like a straight line on a QQ plot (called qqnorm in R). Looking at the qqnorm() plot at right for the number of murders, is linear regression going to produce reliable predictions?

Let’s look at the output of doing a fit, assuming y is t (total number of murders) and x is p (population).

fit<- lm(y ~ x) #could also do lm(t ~ p)

summary(fit)

The output from RStudio is as follows:

Residuals:

Min 1Q Median 3Q Max

-112.889 -25.656 -3.687 25.505 217.780

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.713e+01 1.198e+01 -1.43 0.159

x 3.316e-05 1.315e-06 25.23 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 63.77 on 49 degrees of freedom

Multiple R-squared: 0.9285, Adjusted R-squared: 0.9271

F-statistic: 636.5 on 1 and 49 DF, p-value: < 2.2e-16

1. What hypothesis is the p-value testing?
2. Is there sufficient evidence at the .05 level to reject the null?
3. What is the value of R^2?

What do you conclude from this?

1. What is the value of the y-intercept?
2. Does this match the “b” you found previously?
3. What is the value of the slope coefficient [called ‘x’ in the output]?
4. Does this match the slope you found previously?