

Correlation & Linear Regression

STAT 200 – Introductory Statistics

Torin Quinlivan

October 15, 2026



Today's Objective

After today's lecture, you should:

- Understand the theory behind testing the relationship between two numeric variables.
- Estimate a value of the dependent variable (and provide a confidence interval).
- Predict a value of the dependent variable (and provide a prediction interval).

Measures of Correlation

We now move beyond summary statistics about a single variable. Let us look at quantifying the *relationship* between variables.

All correlation measures indicate the strength of the relationship between two variables. The values range between -1 (perfect negative correlation) and $+1$ (perfect positive correlation).

Measures of Correlation

We now move beyond summary statistics about a single variable. Let us look at quantifying the *relationship* between variables.

All correlation measures indicate the strength of the relationship between two variables. The values range between -1 (perfect negative correlation) and $+1$ (perfect positive correlation).

There are three types of correlation I introduce here:

- r , the Pearson correlation coefficient.
- ρ , the Spearman rank correlation coefficient.
- τ , the Kendall rank correlation coefficient.

Pearson's Correlation Coefficient

The Pearson correlation coefficient, r , is the “usual” measure of correlation between two numeric variables. It was developed by Karl Pearson in 1895 from a related idea introduced by Francis Galton in the 1880s, and for which the mathematical formula was derived and published by Auguste Bravais in 1844.

The Pearson correlation coefficient:

- is *not* robust to outliers, so its value can be misleading if outliers are present.
- only quantifies the *linear* relationship between two variables.
- is appropriate for variables that are at least interval level.

Pearson's Correlation Coefficient

The formula for the Pearson correlation coefficient, r , is:

$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) = \frac{1}{n-1} \sum_{i=1}^n z_{x_i} z_{y_i}$$

The R function is `cor(x, y, method="pearson")`.

Spearman's Rank Correlation Coefficient

Spearman's rank correlation coefficient (Spearman's ρ) is a measure of the correlation between the *ranks* of two variables. Essentially, it measures how well the relationship between two variables can be described using a monotonic function. Spearman devised it in 1904 based on Pearson's earlier writings (1895).

The Spearman coefficient:

- *is* robust to outliers.
- is appropriate for variables that are at least ordinal level.

Spearman's Rank Correlation Coefficient

Spearman's ρ is the Pearson correlation coefficient on the ranks (as opposed to the values).

The R function is `cor(x, y, method="spearman")`.

Kendall's Rank Correlation Coefficient

The Kendall rank correlation coefficient, commonly referred to as Kendall's τ coefficient, measures the *ordinal* association between two measured quantities. Like Spearman's ρ , it is a measure of rank correlation. Maurice Kendall developed it in 1938, though Gustav Fechner had proposed a similar measure in the context of time series in 1897.

Kendall's τ coefficient:

- is robust to outliers.
- is appropriate for variables that are at least ordinal level.

The R function is `cor(x, y, method="kendall")`.

Correlation Example

Let us determine the correlation between the IQ of a person and how much television they watch per week.

This is the R code used for this analysis:

```
# The data
iq = c(106,100,86,101,99,103,97,113,112,110)
tv = c(7,27,2,50,8,29,20,12,6,17)

# The analysis
r  <- cor(iq, tv, method="pearson")
rho <- cor(iq, tv, method="spearman")
tau <- cor(iq, tv, method="kendall")
```

Correlation Example

The output:

```
> r = 0.000378362  
> rho = -0.030303030  
> tau = -0.022222222
```

Correlation Example

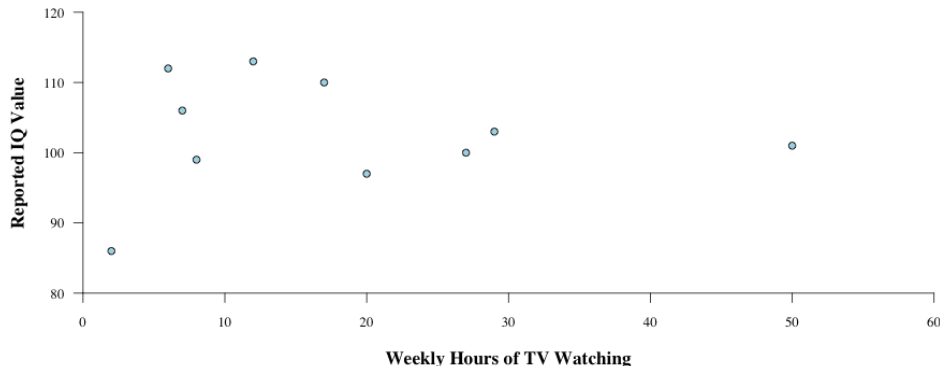
The output:

```
> r = 0.000378362  
> rho = -0.030303030  
> tau = -0.022222222
```

So, which of the three correlations is “most” appropriate for these two variables?

Correlation Example

Here is a graphic of the data. It may help us determine which of the three measures would be “best”:



Example #1: Violent Crime Rate in 1990 vs. 2000

Understanding the persistence of violent crime in cities can help determine which cities are being successful in reducing it and which cities are not.

What is the relationship between the violent crime rate in 1990 and in 2000?

Example #1: Violent Crime Rate in 1990 vs. 2000

Understanding the persistence of violent crime in cities can help determine which cities are being successful in reducing it and which cities are not.

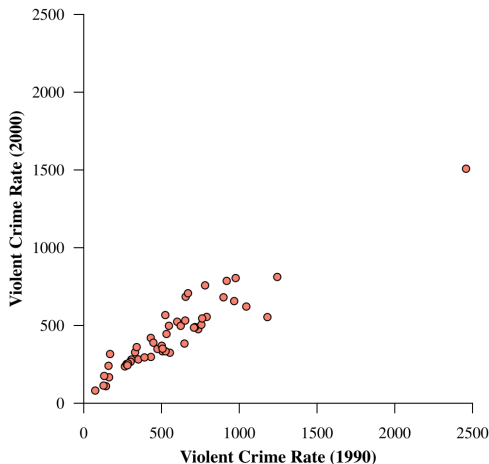
What is the relationship between the violent crime rate in 1990 and in 2000?

Note that we want to know *more* than just if there is a significant relationship. We want to:

- Specify how much one influences the other.
- Estimate violent crime rates in 2000, given the value in 1990.
- Predict violent crime rates in 2000, given the value in 1990.

Example #1: Violent Crime Rate in 1990 vs. 2000

What is the relationship between these two variables?



To summarize the relationship, we use a “line of best fit”:

$$\hat{y} = \beta_0 + \beta_1 x$$

To summarize the relationship, we use a “line of best fit”:

$$\hat{y} = \beta_0 + \beta_1 x$$

However, this raises the following question: What do we mean by the “best” fit?

The Theory

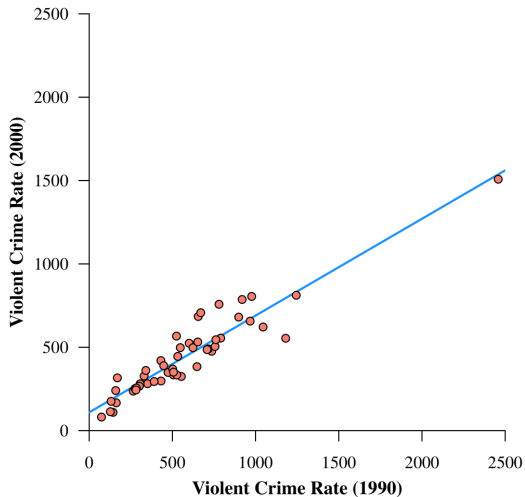
This is the most important question we can answer. Different answers lead to different fitting methods.

MATH/STAT 225 covers a few of the different methods. It also covers the easiest method in great detail.

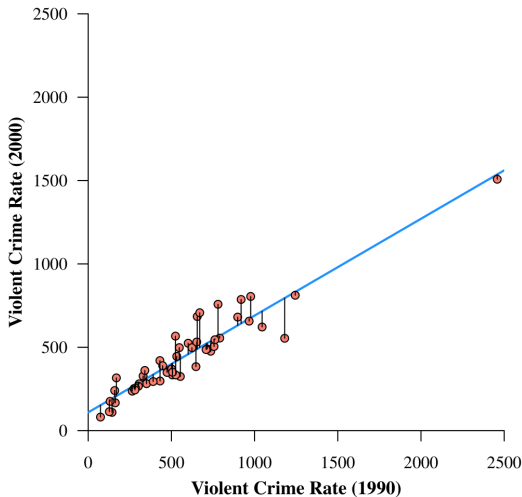
In this course, we will only look at the easiest method: ordinary least squares (OLS).

It is based on **minimizing the sum of the squared residual values**.

The Theory: The Line



The Theory: The Residuals



The Theory

Recall that we are defining “best fit” as the line that minimizes the sum of the squared errors (residuals). The process requires differential calculus. If you’ve not had calculus, you can space out for a moment. Here’s a great use for it, try to follow along.

The Theory (Space Out)

The first step in minimization is to determine the objective (target) function that needs to be minimized.

$$\begin{aligned} Q &= \sum_i e_i^2 \\ &= \sum_i (y_i - \hat{y}_i)^2 \\ &= \sum_i (y_i - (\beta_0 + \beta_1 x_i))^2 \end{aligned}$$

This is our objective function. To minimize it, we take its derivative with respect to each parameter, set each equal to 0, and solve for the parameter.

The Theory (Space Out)

$$\begin{aligned}\frac{\partial}{\partial \beta_0} Q &= \frac{\partial}{\partial \beta_0} \sum_i (y_i - \beta_0 - \beta_1 x_i)^2 \\ &= \sum_i 2(y_i - \beta_0 - \beta_1 x_i)(-1)\end{aligned}$$

$$\begin{aligned}0 &\stackrel{\text{set}}{=} -2 \left(\sum_i y_i - \sum_i b_0 - \sum_i b_1 x_i \right) \\ &= n\bar{y} - nb_0 - n\bar{x}b_1\end{aligned}$$

$$\Rightarrow b_0 = \bar{y} - b_1 \bar{x}$$

The Theory (Space Out)

Do not forget the definition of the sample mean:

$$\bar{x} = \frac{1}{n} \sum x_i \iff \sum x_i = n\bar{x}$$

The Theory (Space Out)

$$\begin{aligned}\frac{\partial}{\partial \beta_1} Q &= \frac{\partial}{\partial \beta_1} \sum_i (y_i - \beta_0 - \beta_1 x_i)^2 \\ &= \sum_i 2(y_i - \beta_0 - \beta_1 x_i)(-x_i)\end{aligned}$$

$$\begin{aligned}0 &\stackrel{\text{set}}{=} -2 \left(\sum_i x_i y_i - \sum_i b_0 x_i - \sum_i b_1 x_i^2 \right) \\ &= \sum_i x_i y_i - (\bar{y} - b_1 \bar{x}) \sum_i x_i - b_1 \sum_i x_i^2 \\ &= \sum_i x_i y_i - (\bar{y} - b_1 \bar{x}) n \bar{x} - b_1 \sum_i x_i^2 \\ &= \sum_i x_i y_i - n \bar{x} \bar{y} + b_1 n \bar{x}^2 - b_1 \sum_i x_i^2\end{aligned}$$

The Theory (Space Out)

$$0 = \sum_i x_i y_i - n\bar{x}\bar{y} + b_1 n\bar{x}^2 - b_1 \sum_i x_i^2$$

$$b_1 \sum_i x_i^2 - b_1 n\bar{x}^2 = \sum_i x_i y_i - n\bar{x}\bar{y}$$

$$b_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}$$

The Theory (Wake Up!)

Thus, our estimators of the y -intercept and slope (effect) are:

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$b_1 = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sum x_i^2 - n \bar{x}^2}$$

The Theory (Space Out)

By the way, the second formula can also be written as:

$$\begin{aligned} b_1 &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \\ &= r \left(\frac{s_y}{s_x} \right) \end{aligned}$$

This gives some insight into the slope, the correlation, and the relationship between them.

Formulas for Linear Regression

To estimate the line of best fit, we have the formula below:

$$\hat{y} = b_0 + b_1x$$

And we can calculate b_0 and b_1 with the formulas:

$$b_1 = r \left(\frac{s_y}{s_x} \right)$$

$$b_0 = \bar{y} - b_1\bar{x}$$

Remember, we have to find the slope, b_1 , first before we can find the y -intercept, b_0 .

Example #2: Linear Regression

Suppose we have two variables, x and y , with the following summary statistics:

	x	y
Mean	8	26
Std. Dev.	2	8
Correlation	0.75	

Find the line of best fit for predicting y using x .

Example #2: Linear Regression

The average STAT 200 student studies for 7 hours in the week before the midterm exam, with a variance of 9 hours squared, and sleeps for an average of 45 hours that week, with a variance of 144 hours squared. The correlation between the hours spent studying and the hours spent sleeping is $r = -0.60$.

Using the fact that Sam studied for 8 hours the week before the exam, predict how many hours he slept that week.

The Theory: R

Of course, for any real data set, it is unreasonable to do calculations by hand.

The calculations can be done in R, of course:

- The first step is to fit the model.
- The second step is to summarize the results.

We will return to this later in the course.