Lecture 08 - Inference with Two Continuous Variables

Data Analysis in CJ (CJUS 6103)

Outline

I. Scatterplots
II. Correlation coefficient (Pearson's r)
III. Bivariate regression equation
IV. Evaluating the fit of the regression equation
V. Comparability of correlation and regression coefficients
VI. R Tutorial - Coefficients & Slopes

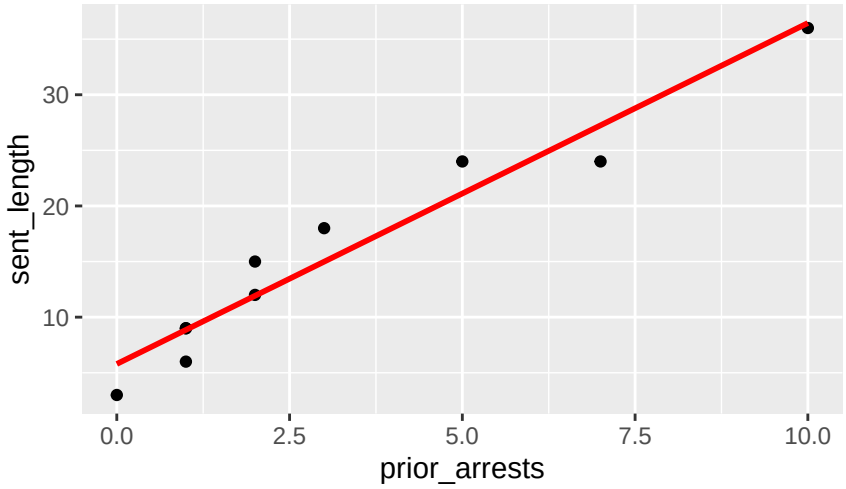
Scatterplots

Correlation and regression are used to assess the relationship between two continuous variables. One variable is defined as the dependent variable (which we denote Y), and the other is defined as the independent variable (which we denote X).

One simple way to assess the relationship between two variables is to use a scatterplot, or graphical display that summarizes the nature of the association between a continuous independent and dependent variable. The X-axis is the independent variable, and the Y-axis is the dependent variable. Each observation receives a dot at its respective X- and Y-values.

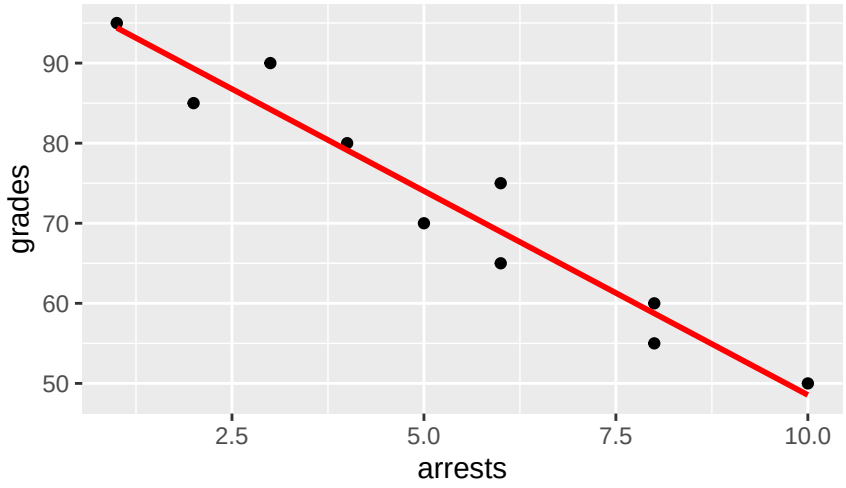
Let's start with an example. It is well known that one of the strongest predictors of sentence length is a defendant's prior record. We collect data from a sample of 10 inmates convicted of burglary, and ask them how many months they received in their sentence (Y) and how many arrests they had prior to conviction (X). We obtain the following data.

Prior Arrests	Sentence Length (Months)
0	3
1	6
1	9
1	9
2	12
2	15
3	18
5	24
7	24
10	36



Let's say that we also gathered data from 10 youths in juvenile detention on average school performance and the number of juvenile arrests.

Test Scores	Number of Arrests
50	10
55	8
60	8
65	6
70	5
75	6
80	4
85	2
90	3
95	1



We see that as school performances increases, juvenile arrests decrease. Thus, there is a negative relationship between school performance and juvenile arrest. In other words, youths with high school performance tend to have a low number of arrests.

Scatterplots thus tell us something about the direction of the association between two variables. We can add a trend line to the scatterplot to aid in the interpretation of the direction of association, or what we will refer to later as a regression line. A second piece of information that we can obtain is the amount of variation there is around the regression line, which is an indication of the strength of the association. The closer the dots are to the regression line, the stronger the association between X and Y.

An advantage offered by scatterplot is the ability to identify outliers. The disadvantage of using a scatterplot to summarize the relationship between two variables is that it is not very precise. We can only determine that prior record and sentence length are positively related, and we can only "eyeball" the regression line. Our ultimate goal is to be a little more exact in describing the nature of the relationship between X and Y. There are two statistics that we can compute to be more precise: a correlation coefficient or a regression equation.

Caveat - sometimes numbers are not all that helpful by themselves. To illustrate this I will introduce a classic example called Anscombe's Quartet in the following section.

When Numbers Aren't Helpful - Anscombe's Quartet

Sometimes numbers can be misleading without plotting the data. A prominent example is Anscombe's Quartet. Francis Anscombe created these data in 1973 to demonstrate the influences of outliers on the statistical properties of a data set. The quartet comprises four data sets with eleven observations each. They were designed so that the following statistics were exactly the same or similar to the 2nd/3rd decimal place:

$$\overline{x}, \overline{y}, s_x^2, s_y^2, r_{xy}, \alpha, b_x, R^2$$

Below are the four data sets. The differences between them are obvious. Though some X and Y values are the same across the series, there are enough differences to suspect that the data would yield at least some detectable differences in descriptive statistics, but in practice there are little to no differences.

Group 1		Group 2		Group 3		Group 4	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

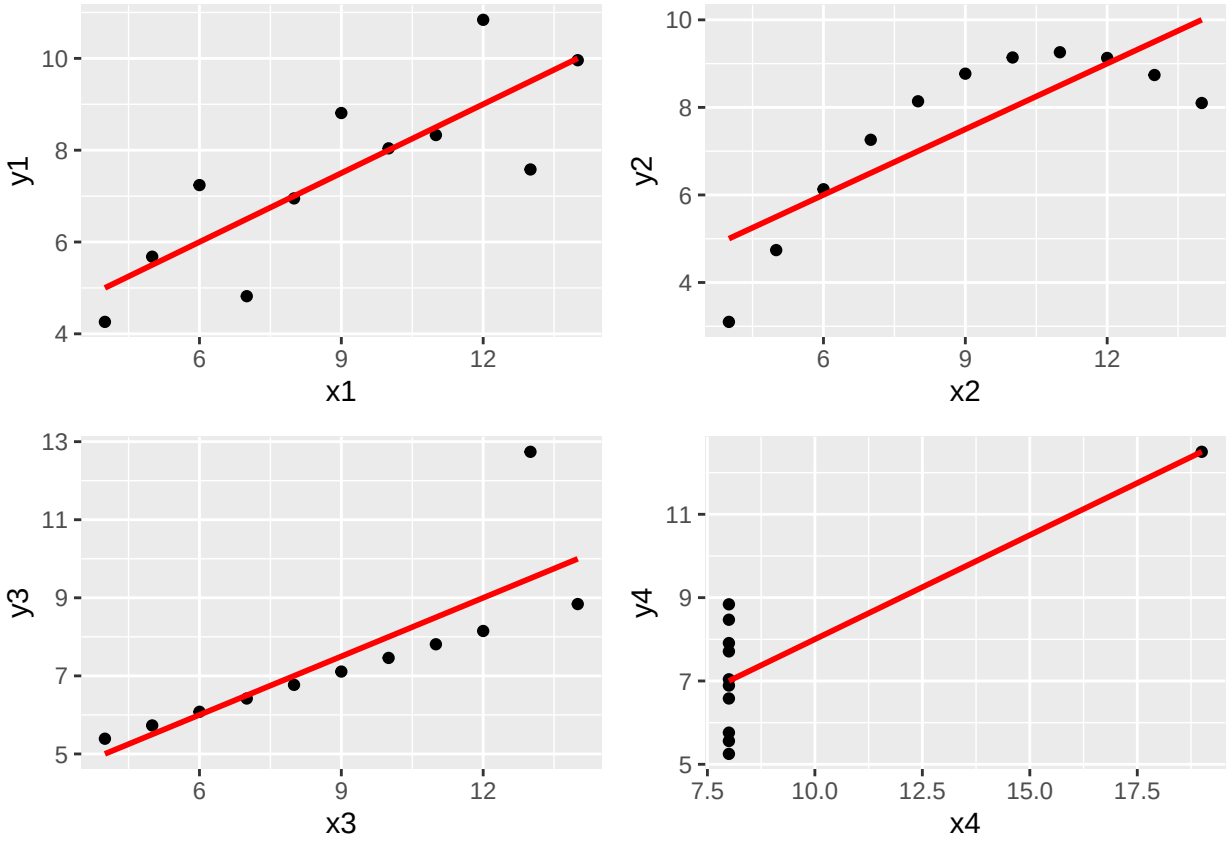
Now, to enter the data and create a plot to visualize each data set.

```
group1<-data.frame(x1<-c(10.0,8.0,13.0,9.0,11.0,14.0,6.0,4.0,12.0,7.0,5.0),
                   y1<-c(8.04,6.95,7.58,8.81,8.33,9.96,7.24,4.26,10.84,4.82,5.68))

group2<-data.frame(x2<-c(10.0,8.0,13.0,9.0,11.0,14.0,6.0,4.0,12.0,7.0,5.0),
                   y2<-c(9.14,8.14,8.74,8.77,9.26,8.10,6.13,3.10,9.13,7.26,4.74))

group3<-data.frame(x3<-c(10.0,8.0,13.0,9.0,11.0,14.0,6.0,4.0,12.0,7.0,5.0),
                   y3<-c(7.46,6.77,12.74,7.11,7.81,8.84,6.08,5.39,8.15,6.42,5.73))

group4<-data.frame(x4<-c(8.0,8.0,8.0,8.0,8.0,8.0,8.0,19.0,8.0,8.0,8.0),
                   y4<-c(6.58,5.76,7.71,8.84,8.47,7.04,5.25,12.50,5.56,7.91,6.89))
```

In the next step, I need to create four individual scatter plots with best fit lines, store them in separate objects, then arrange them together using the ggarrange() function from the **ggpubr** package.

```
g1<-ggplot(group1, aes(x=x1, y=y1))+geom_point()+
  geom_smooth(method="lm", color="red", se=FALSE)

g2<-ggplot(group2, aes(x=x2, y=y2))+geom_point()+
  geom_smooth(method="lm", color="red", se=FALSE)

g3<-ggplot(group3, aes(x=x3, y=y3))+geom_point()+
  geom_smooth(method="lm", color="red", se=FALSE)

g4<-ggplot(group4, aes(x=x4, y=y4))+geom_point()+
```

```
  geom_smooth(method="lm", color="red", se=FALSE)

ggarrange(g1, g2, g3, g4, nrow=2, ncol=2)
```



The slopes are exactly the same (or very close to it). However, it's clear from each individual graph that the nature of the relationship between X and Y differs. The graph for Group 1 (upper left) depicts a traditional linear, positive relationship - as X increases, so does Y. Group 2 (upper right) is not typical; as X increases, Y increases until an **inflection point** around an X value of 10.5, then begins to decrease. We would call this a nonlinear relationship. Group 3 (lower left) depicts a relationship affected by a single influential observation that pulls the regression line upward so the slope appears larger than it really is. Finally, Group 4 (lower right) depicts a relationship affected by an extreme outlier; there is no relationship between X and Y but it appears that there is one because of one outlying value in the upper right quadrant of the graph.

Anscombe's quartet encourages us to not rely solely on descriptive statistics to summarize the relationship between two variables, as the very same statistics can come from very different X->Y relationships.

Correlation Coefficient

A correlation "standardizes" the association between two variables. We need to calculate the variance of X, the variance of Y, and the crossproduct (or covariance) of X and Y. The correlation coefficient we will be discussing today is called Pearson's r. Pearson's r is the sample counterpart to the population correlation coefficient, ρ (rho)

More generally, correlation coefficients range from -1.0 to +1.0 and have similar interpretive rules as ϕ from χ^2 and η from ANOVA. The direction of the correlation coefficients indicates the direction of the relationship between two variables. If it is negative, we expect an increase in one variable to be associated with a decrease in the other. By contrast, if the coefficient is positive, an increase in one variable is associated with an increase in the other variable. Further, values closer to an absolute value of 1 indicate a stronger relationship, while correlation coefficient values closer to 0 indicate a lack of a relationship between the variables.

Computing Pearson's r

Definitional formula

$$r = \frac{\sum((x - \overline{x}) * (y - \overline{y}))}{\sqrt{\sum(x - \overline{x})^2 * \sum(y - \overline{y})^2}}$$

The numerator - $\sum((x - \overline{x}) * (y - \overline{y}))$ - is the **cross-product** of X and Y. If you divide the cross-product by degrees of freedom (n-2, because we use a sample standard deviation to estimate a population standard deviation for both X and Y) you get an estimate of the population **covariance** between X and Y (σ_{XY})

The denominator - $\sum(x - \overline{x})^2$ and $\sum(y - \overline{y})^2$ - is the sum of squares for each individual variable.As before, if you divide each individual sum of squares by each variable's degrees of freedom (n-1) to obtain an estimate for the population **variance** of X or Y (σ_x^2, σ_Y^2).

- Computational formula

$$r = \frac{\sum(x * y) - (n * \overline{x} * \overline{y})}{\sqrt{[\sum(x^2) - n * \overline{x}^2] * [\sum(y^2) - n * \overline{y}^2]}}$$

Computing Pearon's r - Prior Record and Sentence Length

Definitional formula for Pearson's r

X	Y	$(X-\overline{X})$	$(X-\overline{X})^2$	$(Y-\overline{Y})$	$(Y-\overline{Y})^2$	$(X-\overline{X})*(Y-\overline{Y})$
0	3	-3.2	10.24	-12.6	158.76	40.32
1	6	-2.2	4.84	-9.6	92.16	21.12
1	9	-2.2	4.84	-6.6	43.56	14.52
1	9	-2.2	4.84	-6.6	43.56	14.52
2	12	-1.2	1.44	-3.6	12.96	4.32
2	15	-1.2	1.44	-0.6	0.36	0.72
3	18	-0.2	0.04	2.4	5.76	-0.48
5	24	1.8	3.24	8.4	70.56	15.12
7	24	3.8	14.44	8.4	70.56	31.92
10	36	6.8	46.24	20.4	416.16	138.72
32	156		91.6		914.14	280.80
$\overline{x}=3.2$	$\overline{y}=15.6$					

$$r=\frac{\sum((x-\overline{x})*(y-\overline{y}))}{\sqrt{\sum(x-\overline{x})^2*\sum(y-\overline{y})^2}}=\frac{280.80}{\sqrt{91.60*914.40}}=\frac{280.80}{289.41}=0.970$$

Computational formula for Pearson's r

X	X^2	Y	Y^2	XY
0	0	3	9	0
1	1	6	36	6
1	1	9	91	9
1	1	9	81	9
2	4	12	144	24
2	4	15	225	30
3	9	18	324	54
5	25	24	576	120
7	49	24	576	168
10	100	36	1296	360
32	194	156	3348	780

$$r=\frac{\sum(x*y)-(n*\overline{x}*\overline{y})}{\sqrt{[\sum(x^2)-n*\overline{x}^2]*[\sum(y^2)-n*\overline{y}^2]}}=\frac{780-(10)(3.2)(15.6)}{\sqrt{[194-(10)(3.2^2)][3348-(10)(15.6^2)]}}=\frac{280.8}{\sqrt{[91.6][914.4]}}=0.970$$

Based upon the value of Pearson's r here, we can conclude that the relationship between prior record and sentence length is strong (>.50) and it is a positive relationship. So, as prior record increases, we expect sentence length to also increase. More accurately, people who are above the mean on prior record tend to be above the mean on sentence length. A note of caution, though - this cannot be interpreted as causal, only correlational.

Computing Pearon's r - School Performance and Juvenile Arrest

X	Y	XY	X^2	Y^2
50	10	500	2500	100
55	8	440	3025	64
60	8	480	3600	64
65	6	390	4225	36
70	5	350	4900	25
75	6	450	5625	36
80	4	320	6400	16
85	2	170	7225	4
90	3	270	8100	9
95	1	95	9025	1
725	53	3465	54625	355
$\overline{x} = 72.5$	$\overline{y} = 5.3$			

$$r = \frac{\sum(x * y) - (n * \overline{x} * \overline{y})}{\sqrt{[\sum(x^2) - n * \overline{x}^2] * [\sum(y^2) - n * \overline{y}^2]}} = \frac{3465 - (10)(72.5)(5.3)}{\sqrt{[54625 - (10)(72.5^2)][355 - (10)(5.3^2)]}} = \frac{-377.5}{\sqrt{[2062.5][74.1]}} = -0.966$$

The association between school performance and juvenile arrests is thus strong (it is very close to -1) and negative. We can interpret the correlation coefficient to mean that youth who are above the mean on school performance tend to be below the mean on number of arrests (and vice versa).

Auxiliary Statistics for Pearson's r

A useful property of the correlation coefficient is that when we square it, we can use an "explained variance" interpretation. So, prior record explains $0.970^2 = 0.941 = 94.1\%$ of the variance in sentence length. School performance explains $\breve{}0.966^2 = 0.933 = 93.3\%$ of the variance in juvenile arrest.

Hypothesis Test for Pearson's r - Prior Record and Sentence Length

How about if we want to conduct a hypothesis test? We want to know if a linear relationship exists between prior record and sentence length in the population, or if our estimate of the correlation is the result of sampling error. Let's go through the five steps.

Step 1. State Hypotheses

Our research hypothesis is this: Does having a prior record increase sentence length? The population parameter we are trying to estimate is ρ, the population correlation coefficient, and its sample analog is r. The null and alternative hypotheses are $H_0 : \rho = 0$ and $H_1 : \rho > 0$.

Step 2. Obtain a Probability Distribution

The probability distribution for correlation coefficients is the t-distribution, with $df = n\breve{}2$. For this test, we have $10 - 2 = 8$ degrees of freedom.

Step 3. Make Decision Rules

We will use α=.05. We will reject the null hypothesis if TS > 1.860.

Step 4. Calculate the Test Statistic

The test statistic for a correlation coefficient is:

$$TS = r\sqrt{\frac{n-2}{1-r^2}} = 0.970\sqrt{\frac{10-2}{1-(0.970)^2}} = 0.970\sqrt{135.36} = 11.285$$

Step 5. Make a Decision about the Null Hypothesis

Reject H_0, conclude that having a longer prior record is correlated with a significantly longer sentence length.

Hypothesis Test for Pearson's r - Academic Performance and Juvenile Arrest

Step 1. State Hypotheses

Our research hypothesis is this: Are school performance and juvenile arrests correlated with one another? This is a two-tailed hypothesis, because we are not implying a direction to any correlation we may observe between these variables (though we already have some inference that it's probably strongly negative).

$H_1 : \rho \neq 0$; $H_0 : \rho = 0$

Step 2. Obtain a Probability Distribution

As before, we will be using the t-distribution. We have $10 - 2 = 8$ degrees of freedom for this test.

Step 3. Make Decision Rules

We will use $\alpha = .05$ (two-tailed), making our critical value equal to 2.306. Therefore, we will reject the null hypothesis is $|\text{TS}| > 2.306$.

Step 4. Calculate the Test Statistic

$$TS = r\sqrt{\frac{n-2}{1-r^2}} = -0.966\sqrt{\frac{10-2}{1-(-0.966^2)}} = -0.966\sqrt{119.68} = -10.568$$

Step 5. Make a Decision about the Null Hypothesis

Reject H_0, school performance is significantly associated with juvenile arrest.

Bivariate Regression Equation

Another way to assess the relationship between two continuous variables is by estimating a regression equation. When a scatterplot indicates that two variables are more or less linearly related, it is convenient to draw a straight line through the middle of the data points. When we estimate the regression line (as opposed to trying to eyeball it), it can be shown that it is the **best-fitting** line, which means that the line falls as close to every data point as possible.

A regression equation in the population is of the form:

$$Y = \alpha + \beta X$$

In this equation, α and β are the population parameters that summarize the association between X and Y. A regression equation using sample data to estimate these parameters is of the form:

$$Y = a + bX$$

In this equation, a is the y-intercept (or constant), and b is the slope. Once these two parameters are estimated, we can substitute any value for X to determine the best guess of Y for that particular value of X. The estimate b tells us the expected change in Y if X increases by one unit. It is calculated by the formula:

$$b = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sum (X - \overline{X})^2} = \frac{\sum XY - n\overline{XY}}{\sum X^2 - n\overline{X}^2}$$

The y-intercept, a, is very simply the value of Y when $X = 0$. The purpose of the intercept is to **anchor** the regression line to the y-axis. It simply tells us the point at which the regression line crosses the y-axis. Once we know b, we can solve for the intercept using this equation:

$$a = \overline{Y} - b\overline{X}$$

Bivariate Regression Equation - Prior Record and Sentence Length

Let's estimate the regression equation for our prior record and sentence length example.

X	Y	$(X-\overline{X})$	$(X-\overline{X})^2$	$(Y-\overline{Y})$	$(Y-\overline{Y})^2$	$(X-\overline{X})*(Y-\overline{Y})$
0	3	-3.2	10.24	-12.6	158.76	40.32
1	6	-2.2	4.84	-9.6	92.16	21.12
1	9	-2.2	4.84	-6.6	43.56	14.52
1	9	-2.2	4.84	-6.6	43.56	14.52
2	12	-1.2	1.44	-3.6	12.96	4.32
2	15	-1.2	1.44	-0.6	0.36	0.72
3	18	-0.2	0.04	2.4	5.76	-0.48
5	24	1.8	3.24	8.4	70.56	15.12
7	24	3.8	14.44	8.4	70.56	31.92
10	36	6.8	46.24	20.4	416.16	138.72
32	156		91.6		914.14	280.80
$\overline{x}=3.2$	$\overline{y}=15.6$					

Definitional formula:

$$b=\frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sum(X-\overline{X})^2}=\frac{280.80}{91.60}=3.066$$

Computational formula:

X	X^2	Y	Y^2	XY
0	0	3	9	0
1	1	6	36	6
1	1	9	91	9
1	1	9	81	9
2	4	12	144	24
2	4	15	225	30
3	9	18	324	54
5	25	24	576	120
7	49	24	576	168
10	100	36	1296	360
32	194	156	3348	780

$$b=\frac{\sum XY-n\overline{XY}}{\sum X^2-n\overline{X}^2}=\frac{780-(10*3.2*15.6)}{194-(10*3.2^2)}=3.066$$

We would interpret this value to mean that every additional arrest is associated with 3.066 additional months sentenced to prison, on average.

Now we can compute the intercept so we can finish the bivariate regression equation:

$$a=\overline{Y}-b\overline{X}=15.6-(3.07)*(3.2)=5.78$$

The intercept value means that the expected value of Y (months sentenced) is equal to 5.78 when the value of X is 0 (i.e., no prior arrests).

Our regression equations is, then:

$$Y=5.78+3.066X$$

This regression equation comes in handy when we want to calculate predicted values of Y for given values of X. Let's pick a few values of X to illustrate.

$\hat{y} = 5.78 + 3.07X$

X	Equation	$\hat{y}$
0	$\hat{y} = 5.78 + (3.066 * 0)$	5.78
2	$\hat{y} = 5.78 + (3.066 * 2)$	11.91
4	$\hat{y} = 5.78 + (3.066 * 4)$	18.04
6	$\hat{y} = 5.78 + (3.066 * 6)$	24.18
8	$\hat{y} = 5.78 + (3.066 * 8)$	30.31
10	$\hat{y} = 5.78 + (3.066 * 10)$	36.44

Bivariate Regression Equation - School Performance and Juvenle Arrest

X	Y	XY	X^2	Y^2
50	10	500	2500	100
55	8	440	3025	64
60	8	480	3600	64
65	6	390	4225	36
70	5	350	4900	25
75	6	450	5625	36
80	4	320	6400	16
85	2	170	7225	4
90	3	270	8100	9
95	1	95	9025	1
725	53	3465	54625	355
$\overline{x} = 72.5$	$\overline{y} = 5.3$			

Computational formula for the slope:

$$b = \frac{\sum XY - n\overline{XY}}{\sum X^2 - n\overline{X}^2} = \frac{3465 - (10 * 72.5 * 5.3)}{54625 - (10 * 72.5^2)} = -0.183$$

The slope value of -0.183 means that, for every one point increase in test score, we expect the number of arrests to decrease by 0.183, on average.

Now, for the intercept:

$$a = \overline{Y} - b\overline{X} = 5.3 - (-0.183) * (72.5) = 18.57$$

The intercept value means that the expected number of juvenile arrests fr a youth who has a score of 0 is 18.57.

The bivariate regression equation is, then:

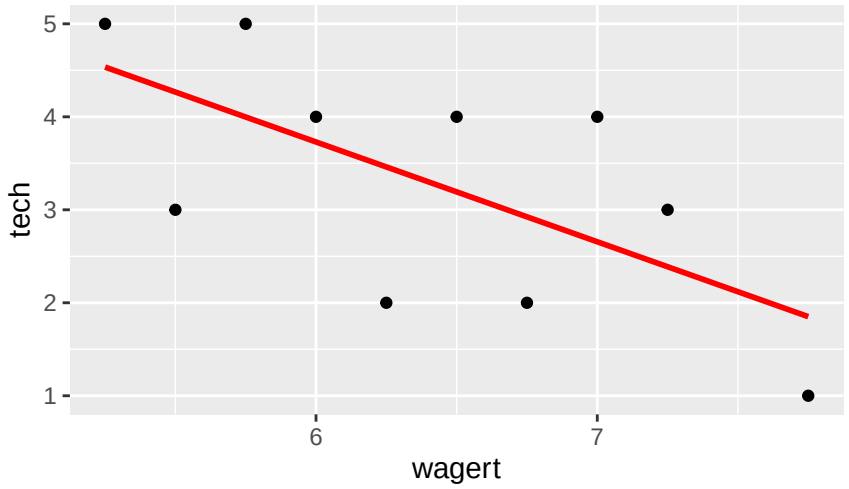
$$Y = 18.57 + -0.183X$$

Hypothesis Testing and Bivariate Slopes

Like the hypothesis test we conduct for the correlation coefficient, we may also conduct a hypothesis test for a slope. This requires some additional calculations, however, and is much more complicated than the simple test for the correlation coefficient.

We will conduct a hypothesis test using a new example - hourly wages and parole violations.

First, we will examine a scatterplot of the data:



It appears that the relationship is negative, but the points are not clustered very tightly around the best fit line. If we had to guess at this point, we might say that we expect a negative slope with a decent level of variability around a regression equation prediction.

Definitional formula for b

Hourly Wages	$X - \overline{X}$	$(X - \overline{X})^2$	Parole Violations	$Y - \overline{Y}$	$(Y - \overline{Y})^2$	$(X - \overline{X})(Y - \overline{Y})$
5.25	-1.15	1.3225	5	1.7	2.89	-1.955
5.50	-0.90	0.8100	3	-0.30	0.09	0.270
5.75	-0.65	0.4225	5	1.70	2.89	-1.105
6.00	-0.40	0.1600	4	0.70	0.49	-0.280
6.25	-0.15	0.0225	2	-1.3	1.69	-0.455
6.50	0.10	0.0100	4	0.70	0.49	0.070
6.75	0.35	0.1225	2	-1.30	1.69	-0.455
7.00	0.60	0.3600	4	0.70	0.49	0.420
7.25	0.85	0.7225	3	-0.30	0.09	-0.255
7.75	1.35	1.8225	1	-2.30	5.29	-3.105
64.0	0	5.775	33	0	16.1	-6.20

$$b = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sum (X - \overline{X})^2} = \frac{-6.20}{5.775} = -1.07$$

Now, we calculate the intercept:

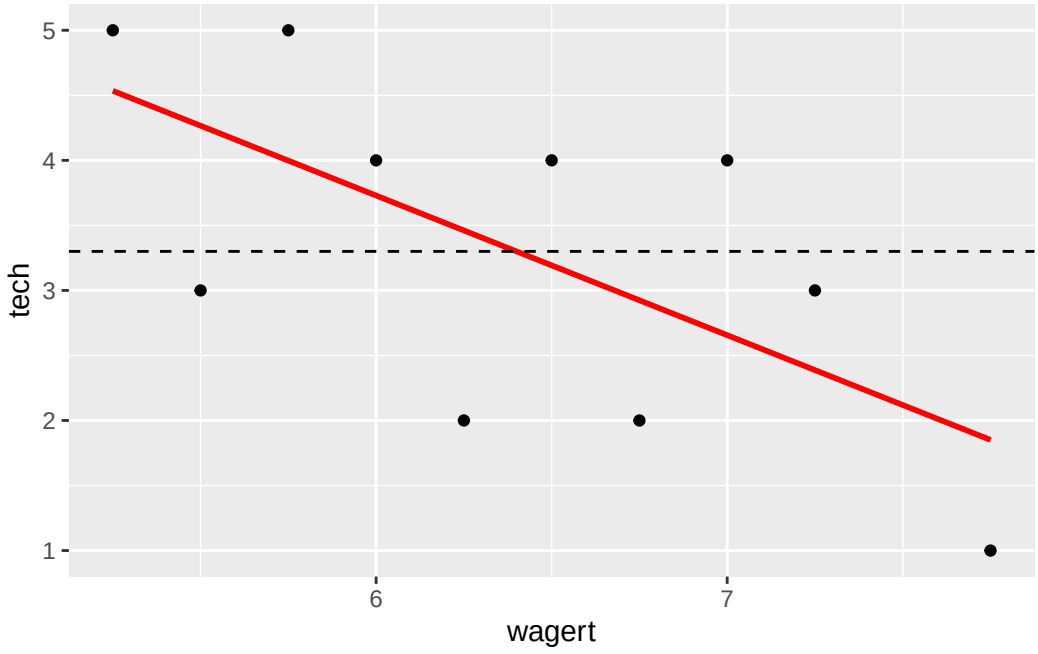
$$a = \overline{y} - b\overline{x} = 3.3 - (-1.07)(6.4) = 10.15$$

Our regression equation is:

$$Y = 10.15 + -1.07X$$

The results of the bivariate slope calculation confirm the negative relationship - as wages increase by $1, we expect the number of parole violations to decrease by 1.07, on average. Further, the expected number of parole violations at zero wages is 10.15.

We're not ready for the hypothesis test just yet, as we have to go over some of the qualities of the regression line before we calculate a test statistic. Namely, that the regression line represents the expected average of y ($\overline{Y}$) given values of X. The question we ask about this line is if it improves our ability to predict values of Y beyond simply using the unconditional mean of Y to predict values of Y.



By design, the equations we use to provide estimates for a and b result in what is known as **minimum squared error**. That is, estimates for the intercept and slope are those that minimize the sum of the squared errors around the regression line - compared to any other value, that intercept and slope result in the minimum value for the sum of the squared error. This is why we often refer to regression analysis as *Ordinary least squares* (OLS). It's the same property that a mean has (least squared differences) but the difference is that the regression line represents a **conditional mean**

$$\sum e^2 = \sum (y - \hat{y})^2 = \sum (y - a - bx)^2 = \text{minimum}$$

It was important to discuss the error about the regression line because it plays an important role in the test statistic for a slope hypothesis test.

Step 1. State Hypotheses

Hypotheses for a slope test are similar to those from a correlation coefficient test - i.e., both may be bidirectional and both reference a population parameter we think our sample statistic represents. In this case, we are trying to say something about the true value of β in the population as we may infer from our estimate for it, b. For this test, we are interested in knowing if higher wages decreases parole violations (so, we expect β to be negative). It is worth noting that regression functions in statistical programs generally provide results from two-tailed tests by default (as this makes it harder to reject the null hypothesis and is more conservative re: statistical significance).

$H_1 : \beta < 0$ $H_0 : \beta \geq 0$

Step 2. Obtain a Probability Distribution

The probability distribution for a slope hypothesis test is the same as for a correlation coefficient test - the t-distribution with $n-2$ degrees of freedom. Here, the are $10-2=8$ degrees of freedom.

Step 3. Make Decision Rules

We will use $\alpha = .05$ one-tailed, making our critical score equivalent to -1.86. Therefore, we will reject the null hypothesis if our TS < -1.86.

Step 4. Calculate the Test Statistic

The formula for the TS is:

$$TS = \frac{b-\beta}{s_b} \Rightarrow TS = \frac{b}{s_b} \text{ recall under } H_0, \beta = 0$$

Which requires us to calculate a new value, s_b which stands for the **standard error of the slope** which requires us to compute the mean squared error about the predicted regression line (I said we would come back to why error around the regression line is important!). We calculate s_b using the following formula:

$$s_b = \sqrt{\frac{s_e^2}{SS_X}} = \sqrt{\frac{\sum e^2/(n-2)}{\sum (x-\overline{x})^2}} = \sqrt{\frac{[\sum (y-\hat{y})^2]/(n-2)}{\sum (x-\overline{x})^2}}$$

We have everything we need, except for $\sum e^2$, which requires us to create a new table.

Hourly Wages	Parole Violations	$\hat{y}$	$y-\hat{y}$	$(y-\hat{y})^2$
5.25	5	4.5325	0.4675	0.2186
5.50	3	4.265	-1.265	1.6002
5.75	5	3.9975	1.0025	1.005
6.00	4	3.73	0.27	0.0729
6.25	2	3.4625	-1.4625	2.1389
6.50	4	3.195	0.805	0.648
6.75	2	2.9275	-0.9275	0.8603
7.00	4	2.66	1.34	1.7956
7.25	3	2.3925	0.6075	0.3691
7.75	1	1.8575	-0.8575	0.7353
64.0	33	33.02	-0.02	9.4439

$\hat{Y}$ is calculated using the regression equation () for all observed values of X (wages). We then subtract $\hat{Y}$ from the **observed** values of Y to calculate the error around the regression line, square that value, and sum it (the sum of squares for the regression line). Dividing by our degrees of freedom then produces a **mean squared error** about the regression line.

$$s_b = \sqrt{\frac{[\sum (y-\hat{y})^2]/(n-2)}{\sum (x-\overline{x})^2}} = \sqrt{\frac{9.4439/(10-2)}{5.775}} = \sqrt{\frac{1.1805}{5.775}} = 0.4521$$

And, finally, our test statistic:

$$TS = \frac{-1.07}{0.4521} = -2.367$$

Step 5. Make a Decision about the Null Hypothesis

Reject H_0, conclude that higher wages are associated with significantly fewer parole violations.

Comparability of Correlation and Regression Coefficients

Judging by their respective formulas, you may guess that correlation and slope coefficients are closely related.

$$b = \frac{\sum((x-\overline{x})*(y-\overline{y}))}{\sum(x-\overline{x})^2};\ r = \frac{\sum((x-\overline{x})*(y-\overline{y}))}{\sqrt{\sum(x-\overline{x})^2 * \sum(y-\overline{y})^2}}$$

It's actually quite simple to compute one from the other:

$$b = r\frac{s_y}{s_x} = r\left(\frac{\sqrt{\sum(y-\overline{y})^2/n}}{\sqrt{\sum(x-\overline{x})^2/n}}\right) = r\frac{\sqrt{\sum(y-\overline{y})^2}}{\sqrt{\sum(x-\overline{x})^2}}$$

$$r = b\frac{s_x}{s_y} = b\left(\frac{\sqrt{\sum(x-\overline{x})^2/n}}{\sqrt{\sum(y-\overline{y})^2/n}}\right) = b\frac{\sqrt{\sum(x-\overline{x})^2}}{\sqrt{\sum(y-\overline{y})^2}}$$

Let's test these out with the prior record and sentence length example. Recall that the correlation for that relationship was 0.970, the slope was 3.066, the sum of squares for X was 91.6, and the sum of squares for Y was 914.14.

$$r = b\frac{\sqrt{\sum(x-\overline{x})^2}}{\sqrt{\sum(y-\overline{y})^2}} = 3.066\frac{\sqrt{91.60}}{\sqrt{914.40}} = 0.970$$

Given their similarity, what does regression add to correlation? Let's use the prior record and sentence length example to illustrate. Consider the case where we have no information other than a person's sentence length. Our best guess for any particular individual, absent any further information, is the mean sentence length for the sample, $\overline{Y} = 15.6$. Once we include an additional variable, X (number of prior arrests), we can incorporate the extra information that this variable contributes by estimating a regression line and predicting the value of Y that lies at any point on the line. So, whenever two variables are significantly correlated, we can make better predictions about Y when we take into account the value of X than if we simply calculated the mean of Y.

R Tutorial - Coefficients & Slopes

In this section, I will show you how to compute correlation and slope coefficients in R and then check that your calculations are correct using automatic R functions.

Manual Method

Entering the data is exactly as we have done before when entering variable values. Nothing to see here, really.

```
wagert<-c(5.25,5.50,5.75,6.00,6.25,6.50,6.75,7.00,7.25,7.75)
tech<-c(5,3,5,4,2,4,2,4,3,1)
```

Next, we need to calculate averages and squared deviations for X and Y as well as their cross-product:

```
X_avg<-sum(wagert)/length(wagert)
Y_avg<-sum(tech)/length(tech)

X_sqrdev<-(wagert-X_avg)^2
Y_sqrdev<-(tech-Y_avg)^2

XY_cross<-(wagert-X_avg)*(tech-Y_avg)
```

With that information in hand, we can compute both the correlation and slope coefficients:

```
XY_corr<-(sum(XY_cross)/sqrt(sum(X_sqrdev)*sum(Y_sqrdev)))
XY_slope<-sum(XY_cross)/sum(X_sqrdev)

XY_corr
```

```
## [1] -0.6429878
```

```
XY_slope
```

```
## [1] -1.073593
```

And now, we need to compute the mean squared error about the regression line:

```
Y_int<-Y_avg-(XY_slope*X_avg)

Y_hat<-Y_int+(XY_slope*wagert)
Yhat_sqrdev<-(tech-Y_hat)^2
reg_MSE<-sum(Yhat_sqrdev)/(length(tech)-2)
slope_std_err<-sqrt(reg_MSE/sum(X_sqrdev))
slope_std_err
```

```
## [1] 0.4521168
```

At this point, we have everything we need to conduct both hypothesis tests we reviewed in this lecture! The values are slightly different than those we calculated by hand but only slightly - the intercept is off by only 2 hundredths of a point.

Automatic Method

Computing correlation coefficients and slopes in R is very straightforward and only requires using two commands the cor.test() function and the lm() function.

```
cor.test(wagert,tech)
```

```
##
##  Pearson's product-moment correlation
##
## data:  wagert and tech
## t = -2.3746, df = 8, p-value = 0.04492
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.9058770 -0.0224504
## sample estimates:
##        cor
## -0.6429878
```

```
lm(tech~wagert)
```

```
##
## Call:
## lm(formula = tech ~ wagert)
##
## Coefficients:
## (Intercept)       wagert
##      10.171       -1.074
```

There's a lot more to both functions than I will cover right now, especially the lm() command. We will review the lm() function much more when we discuss multivariate regression.