Lecture 08 - Inference with Two Continuous Variables

Samuel DeWitt

Inference with Two Continuous Variables

I. Scatterplots
II. Correlation Coefficient
III. Coefficient of determination
IV. Regression equation
V. Comparison of correlation and regression
VI. R Tutorial - Coefficients & Slopes

Scatterplots

- Simple way to visualize the relationship between two continuous variables
  - X-axis is the independent variable (the *cause*)
  - Y-axis us the dependent variable (the *effect*)

- Two characteristics to note in scatterplots
  - Direction of the relationship
    - Is an imaginary trend line positive or negative? Strength of relationship
    - How tightly do the dots cluster around the trend line?

Example - Prior Record and Sentence Length

Well known that the strongest predictor of sentence length is a defendant's prior record.

Sample of 10 inmates convicted of burglary - sentence length and how many prior arrests?

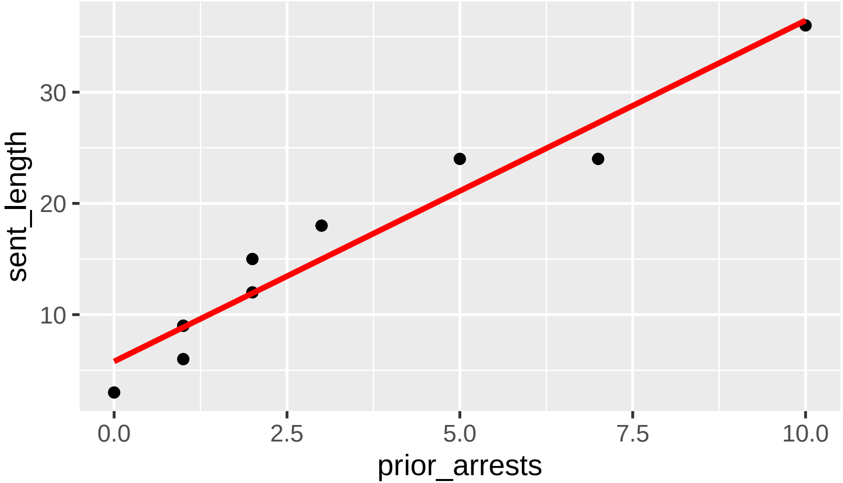
Prior Arrests	Sentence Length (Months)
0	3
1	6
1	9
1	9
2	12
2	15
3	18
5	24
7	24
10	36

Scatterplot - Prior Record and Sentence Length

A simple way to examine this relationship is by using a scatterplot.

As a means of review, a scatterplot places dots at each intersecting value of the independent (prior arrests) and dependent (sentence length) variables.

Scatterplot - Prior Record and Sentence Length

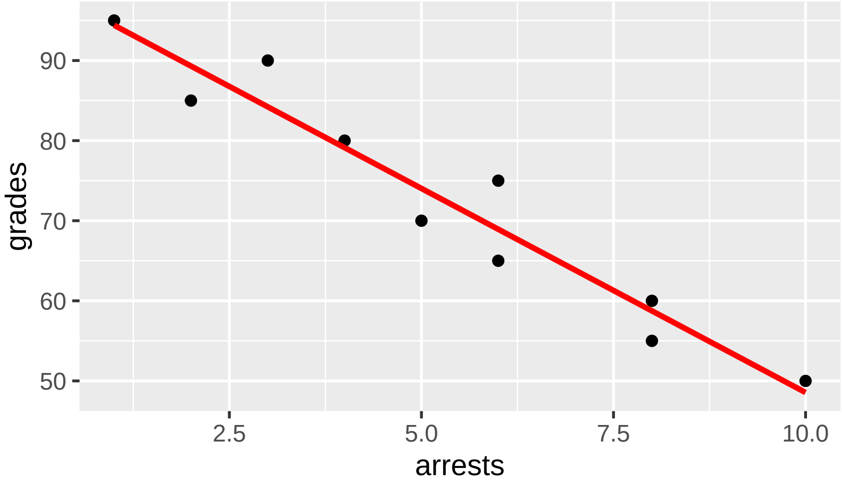


Another Example - Test Scores and Juvenile Arrest

We take another random sample of ten youth and examine the relationship between 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

Scatterplot - Test Scores and Juvenile Arrest



More on Scatterplots

- Advantages of scatterplots
  - Informative (*a picture tells a thousand words*)
    - Can usually tell the direction of the trend without too much trouble
    - Can also tell if the relationship is non-linear

- Disadvantages of scatterplots
  - Cannot consolidate information into a single numerical index; *doesn't give us a number*

Caveat - sometimes numbers are not all that helpful by themselves.

When Numbers Aren't Helpful - Anscombe's Quartet

Sometimes numbers can be misleading without plotting the data. A prominent example is Anscombe's quartet.

The quartet is comprised of 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:

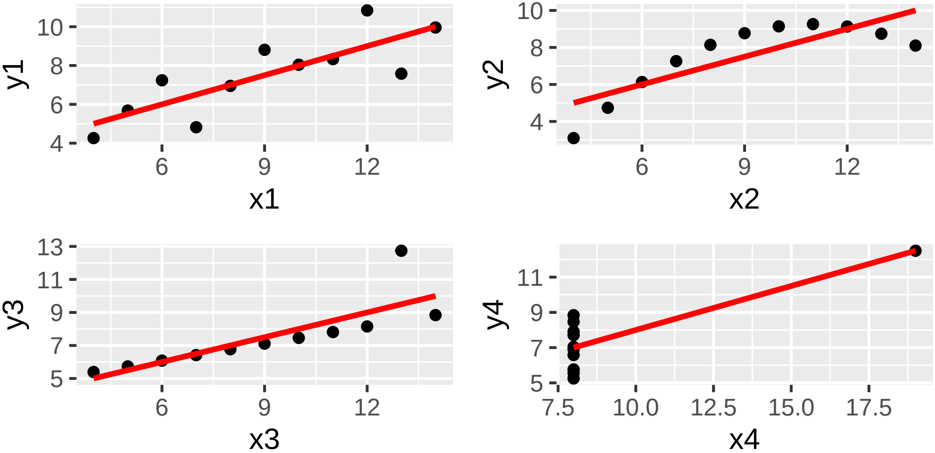
$$\overline{x}, \overline{y}, s_x^2, s_y^2, r_{xy}, \alpha, b_x, R^2$$

When Numbers Aren't Helpful - Anscombe's Quartet

Here are the Anscombe data (I will plot them on the following slide):

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

When Numbers Aren't Helpful - Anscombe's Quartet



Pearson Correlation Coefficient

- Quantifies nature of the linear relationship between two continuous variables
  - Pearson product-moment correlation coefficient
    - a.k.a., *Pearson's r*

- Pearon's r is the sample counterpart to the population correlation coefficient, ρ (rho)
  - Ranges from -1.0 to +1.0
  - 0.0 = no relationship and -1.0/+1.0 = perfect relationship
  - Same interpretive rules as ϕ (phi from χ^2) and η (eta, from ANOVA).

Computing Pearson's r

- Definitional formula

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

- Computational formula

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

Components of Pearson's r

- Numerator - $\sum((x - \overline{x}) * (y - \overline{y}))$
  - This is the **cross-product** of X and Y
  - If you divide the cross-product by degrees of freedom (n-2) you get an estimate of the population **covariance** between X and Y (σ_{XY})

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

Computing Pearson's r - Prior Record and Sentence Length

- Definitional formula for 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
$\overline{x} = 3.2$	$\overline{y} = 15.6$					

Computing Pearson's r - Prior Record and Sentence Length

- Definitional formula for Pearson's r:

$$r = \frac{\sum((x-\overline{x}) * (y-\overline{y}))}{\sqrt{\sum(x-\overline{x})^2 * \sum(y-\overline{y})^2}} = \frac{280.80}{\sqrt{91.60 * 914.40}} = \frac{280.80}{289.41} = 0.97$$

- Interpretation of Pearson's r
  - Magnitude = Strong relationship (>0.5)
  - Direction = Positive relationship
    - Longer prior record correlated with longer sentence length
    - More accurately, people who are above the mean on prior record tend to be above the mean on sentence length

Computing Pearson's r - Prior Record and Sentence Length

- Computational formula for Pearon's r

Prior Record	X^2	Sentence Length	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

Computing Pearson's r - Prior Record and Sentence Length

- Computational formula for Pearon's r

$$
\begin{aligned}
r &= \frac{\sum(x * y) - (n * \overline{x} * \overline{y})}{\sqrt{[\sum(x^2) - n * \overline{x}^2] * [\sum(y^2) - n * \overline{y}^2]}} && (1)\\
&= \frac{780 - (10)(3.2)(15.6)}{\sqrt{[194 - (10)(3.2^2)][3348 - (10)(15.6^2)]}} && (2)\\
&= \frac{280.8}{\sqrt{[91.6][914.4]}} && (3)\\
&= 0.97 && (4)
\end{aligned}
$$

- Same conclusions as before, I just wanted to show you where the numbers go.

Auxiliary Statistic for Pearson's r

- Coefficient of determination (r^2)
  - **Explained variance** interpretation
  - Same interpretation as eta-square (η^2)
  - Prorportion of the variation in y that is explained by variation in x

- Prior record and sentence length example
  - $r^2 = 0.97^2 = 0.941$
  - Prior record explains 94.1% of the variance in sentence length

Hypothesis Test for Prior Record and Sentence Length

- **Step 1: Formally state hypotheses**
  - $H_1 : \rho > 0$
  - $H_0 : \rho \leq 0$

- **Step 2: Choose a probability distribution**
  - t-distribution
  - $df = n - 2 = 10 - 2 = 8$
- **Step 3: Make decision rules**
  - $\alpha = .05$(one-tailed)
  - $t_{crit} = 1.860$
  - Reject H_0 if TS > 1.860

Hypothesis Test for Prior Record and Sentence Length

- **Step 4: Compute the test statistic**

$$TS = r\sqrt{\frac{n-2}{1-r^2}} = 0.97\sqrt{\frac{10-2}{1-0.97^2}} = 11.286$$

- **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.

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
$\overline{x} = 72.5$	$\overline{y} = 5.3$			

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

Hypothesis Test - Test Scores and Juvenile Arrest

- **Step 1: Formally state hypotheses**
  - $H_1 : \rho \neq 0$
  - $H_0 : \rho = 0$
- **Step 2: Choose a probability distribution**
  - t-distribution
  - $df = n - 2 = 10 - 2 = 8$
- **Step 3: Make decision rules**
  - $\alpha = .05$(two-tailed)
  - $t_{crit} = 2.306$
  - Reject H_0 if $|\text{TS}| > 2.306$

Hypothesis Test - Test Scores and Juvenile Arrest

- **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.

- Coefficient of determination
  - $r^2 = 0.966^2 = 0.933$

Regression Equation

- Another way to describe the relationship between two continuous variables
  - Advantage: can predict the value of the DV for given values of the IV

- Population (bivariate) regression equation
  - $y_i = \alpha + \beta x_i + \epsilon_i$
    - α and β are unknown parameters to be estimated from the sample data
    - i indexes individual cases, $i = 1, 2, ..., n$

- Sample (bivariate) regression equation
  - $y_i = a + bx_i + e_i$

Regression Equation (cont)

- Regression estimates describe a regression (trend) line
    - α is a **constant** or **y-intercept**
        - Identifies the location where the regression line crosses the y-axis (x=0)
    - β is a **slope**
        - Impact on y when x increases by one unit
    - e is an **error**, **residual**, or **disturbance**
        - Not all observations lie directly on the regression line, so there is error in predicting y from x

Formula for Regression Coefficients

- Definitional formula for a slope

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

- Computational formula for a slope

$$\frac{\sum XY - n\overline{XY}}{\sum X^2 - n\overline{X}^2}$$

- Formula for the constant

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

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

Definitional formula:

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

Bivariate Regression Equation - Prior Record and Sentence Length

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

Computational formula:

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

Bivariate Regression Equation - Prior Record and Sentence Length

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:

$$Y = 5.78 + 3.066X$$

Bivariate Regression Equation - Prior Record and Sentence Length

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 & 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:

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

Bivariate Regression Equation - School Performance & Juvenle Arrest

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.

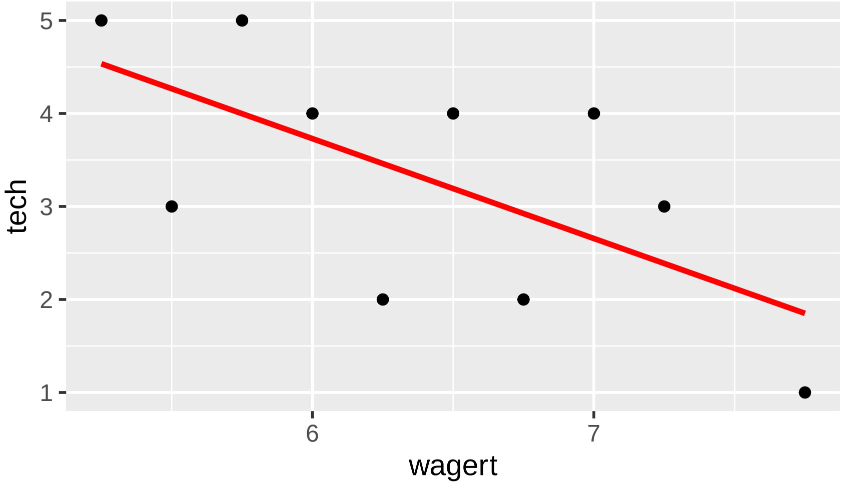
The bivariate regression equation is, then:

$$Y = 18.57 + -0.183X$$

Example - Wages and Parole Violation

Is there a negative relationship between a persons' wages and the number of technical violations while they are on parole?

First, a Scatterplot



Calculating the Slope - Wages and Parole Violation

Definitional formula for b

Hourly			Parole			$(x-\overline{x})$
Wages	$x-\overline{x}$	$(x-\overline{x})^2$	Violations	$y-\overline{y}$	$(y-\overline{y})^2$	$(y-\overline{y})$
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

$$\begin{aligned} b &= \frac{\sum (X-\overline{X})(Y-\overline{Y})}{\sum (X-\overline{X})^2} \tag{5} \\ &= \frac{-6.20}{5.775} \tag{6} \\ &= -1.07 \tag{7} \end{aligned}$$

Calculating the Slope - Wages and Parole Violation

Computational formula for b

Hourly Wages	X^2	Parole Violations	Y^2	XY
5.25	27.5625	5	25	26.25
5.50	30.2500	3	9	16.50
5.75	33.0625	5	25	28.75
6.00	36.0000	4	16	24.00
6.25	39.0625	2	4	12.50
6.50	42.2500	4	16	26.00
6.75	45.5625	2	4	13.50
7.00	49.0000	4	16	28.00
7.25	52.5625	3	9	21.75
7.75	60.0625	1	1	7.75
64.0	415.375	33	125	205.00

$$b = \frac{\sum XY - n\overline{XY}}{\sum X^2 - n\overline{X}^2} \tag{8}$$

$$= \frac{205.0 - (10)(6.4)(3.3)}{415.375 - (10)(6.4^2)} \tag{9}$$

$$= \frac{205.0 - 211.2}{415.375 - 409.6} \tag{10}$$

$$= \frac{-6.2}{5.775} \tag{11}$$

$$= -1.07 \tag{12}$$

Calculating the Intercept - Wages and Parole Violation

- Formula for the constant:

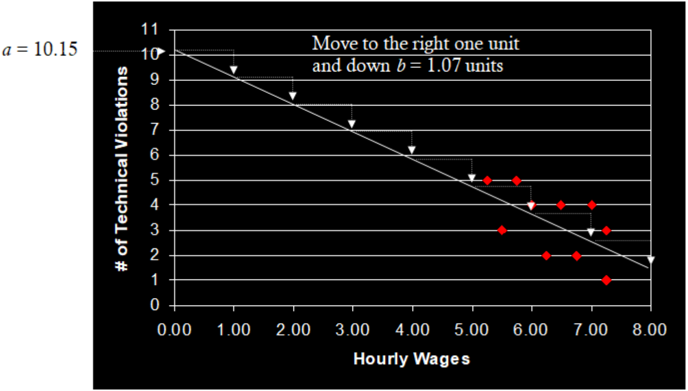
$$a = \overline{y} - b\overline{x} = 3.3 - (-1.07)(6.4) = 3.3 + 6.848 = 10.15$$

- Write the regression equation
  - $\hat{y} = a + bx + e$
  - $\hat{y} = 10.15 - 1.07x + e$

- Interpreting the slope ($b = -1.07$)
  - A one unit increase in x produces a b unit increase (or decrease) in y
  - Or...individuals with $1 more in wages per hour have 1.07 fewer parole violations, on average.

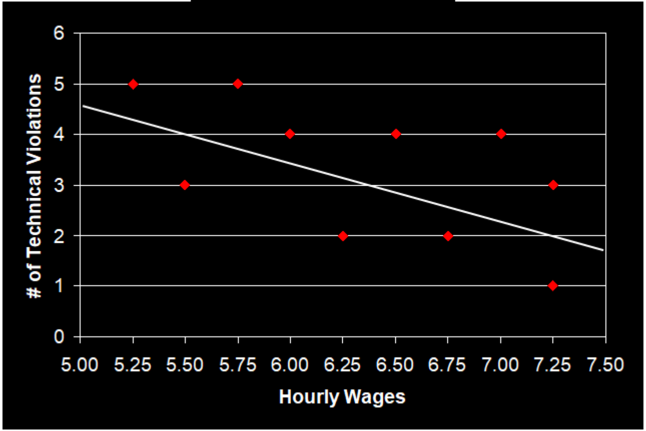
Drawing a Regression Line

$$\hat{y} = a + bx + e = 10.15 - 1.07x + e$$



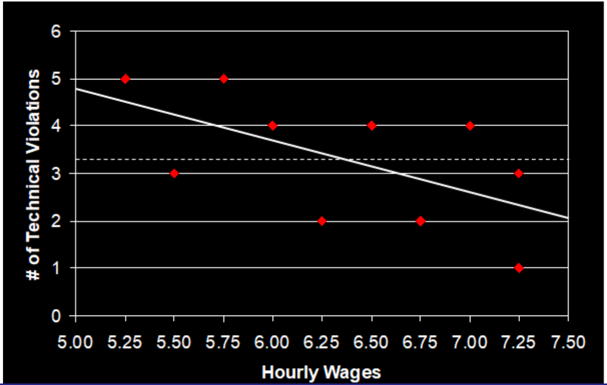
Drawing a Regression Line (cont)

$$y = 10.15 - 1.07x + e$$



Regression Line as An Adjusted or Conditional Mean

- Regression line is $\overline{y} \mid x$ compared to $\overline{y}$
  - Does knowledge about x improve our ability to predict y? I.e., is the error smaller?



$$\overline{y} = 3.3$$

$$\hat{y} = \overline{y} \big| x = 10.15 - 1.07x$$

Evaluating the Fit of a Regression Line

- Minimum squared error
  - Regression error (e) forms the basis for choosing values for a and b
  - Coefficients are those that minimize the sum of the squared errors around the regression line
    - Positive deviations cancel out negative deviations, so squared deviations are taken

- Regression has the property of least squares
  - *Ordinary least squares* (OLS) regression
  - Same property as the mean, only difference is that a regression line represents a **conditional mean**

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

Hypothesis Test for Wages and Parole Violation

- **Step 1**: Formally state hypotheses
  - $H_1 : \beta < 0$
  - $H_0 : \beta \geq 0$

- **Step 2**: Choose a probability distribution
  - t-distribution
  - df = n-2 = 10-2 = 8

- **Step 3**: Make decision rules
  - $\alpha = .05$ (one-tailed)
  - $t_{crit} = 1.860$
  - Reject H_0 is T.S. < -1.860

Hypothesis Test for Wages and Parole Violation (cont)

- **Step 4**: Compute the test statistic

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

  - Formula for the standard error of the slope (s_b)

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

  - We have everything we need except s_e^2, the mean squared error (i.e., error variance)

Hypothesis Test for Wages and Parole Violation (cont)

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

$$
\begin{aligned}
s_b &= \sqrt{\frac{[\sum(y-\hat{y})^2]/(n-2)}{\sum(x-\overline{x})^2}} \tag{13}\\
&= \sqrt{\frac{9.4439/(10-2)}{5.775}} \tag{14}\\
&= \sqrt{\frac{1.1805}{5.775}} \tag{15}\\
&= 0.4521 \tag{16}
\end{aligned}
$$

Hypothesis Test for Wages and Parole Violation (cont)

- **Step 4**: Compute the test statistic (finally...)
  - $TS = \frac{-1.07}{0.4521} = -2.367$

- **Step 5**: Make a decision about the null hypothesis
  - Reject H_0, conclude that higher wages are associated with significantly fewer parole violations.

Comparing Correlation and Regression Coefficients

- b and r are very closely related

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

  - Numerator is the same, only difference is the denominator.

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

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

Comparability of Correlation and Regression Coefficients

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.

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

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

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)
```

Manual Method - Intermediate Calculations

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)
```

Manual Method - Calculating the Coefficients

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
```

Manual Method - Calculating MSE for the Regression Line

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

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

```
## [1] 10.171
```

```
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 - Correlation Coefficient

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
```

Automatic Method - Intercept and Slope

```
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.

Two Questions

What are your two questions today?