Simple Linear Regression

Ronet Bachman, Ph.D.

Presented by
Justice Research and Statistics Association

11/10/2016

Ordinary Least Squares (OLS) Regression

Dependent Variable (y) = interval/ratio

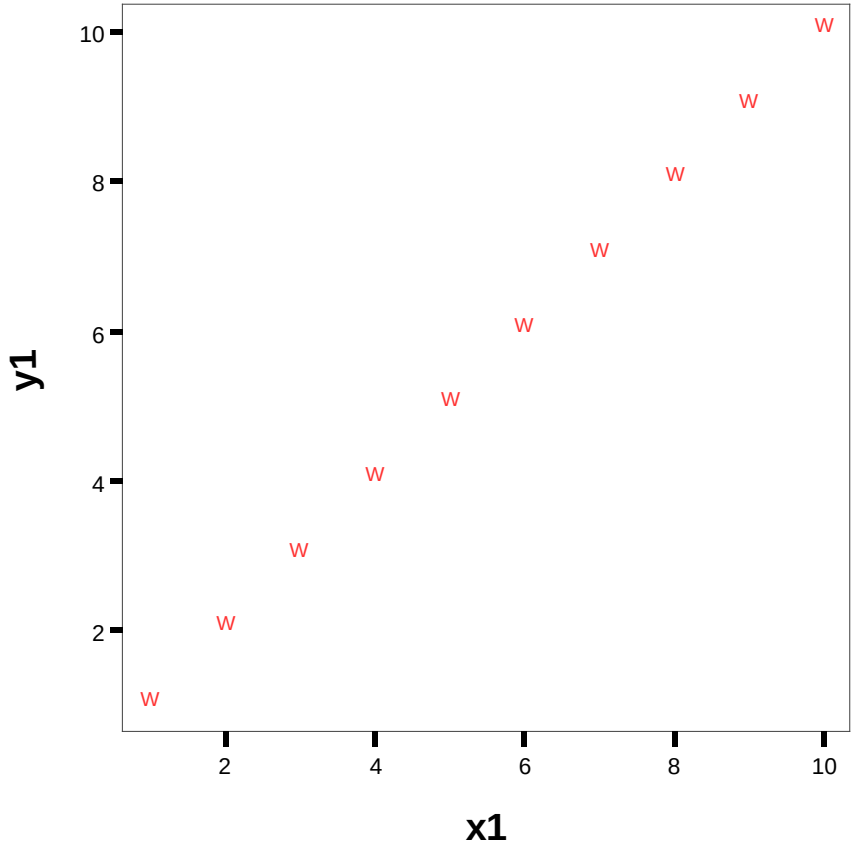
Independent Variable (x) = interval/ratio or dichotomy (coded 0,1)

Presented by Ronet Bachman, PhD
University of Delaware

We are going to Start with cases in with both the IV (x) and DV (y) are measured at the interval ratio level. Suppose we have data like this:

x1	y1
3	3
5	5
2	2
4	4
8	8
10	10
1	1
7	7
6	6
9	9

A scatterplot, where x is plotted on the horizontal axis and y is plotted on the vertical axis would graphically capture the bivariate relationship between x and y:

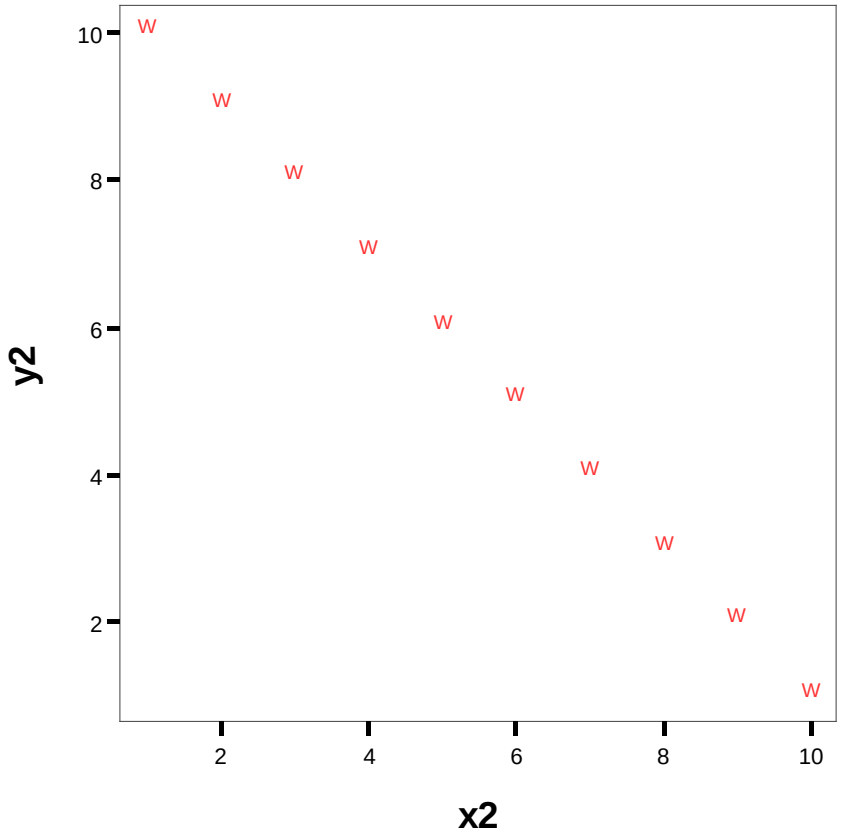


This graphically depicts a relationship where y increases as x increases – this is known as a positive relationship.

How about these two variables:

x2	y2
2	9
4	7
9	2
7	4
8	3
1	10
5	6
6	5
10	1
3	8

A scatterplot, where x is plotted on the horizontal axis and y is plotted on the vertical axis would graphically capture the bivariate relationship between x and y:

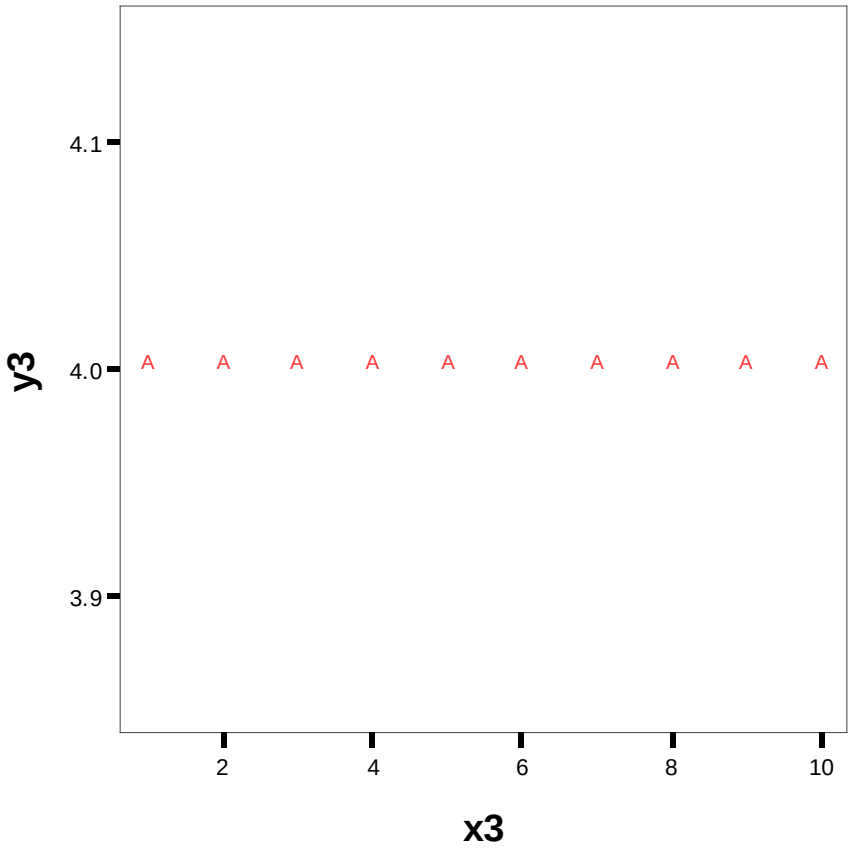


This graphically depicts a relationship where y decreases as x increases – whenever x and y go in opposite directions, this is known as a negative relationship.

How about these two variables:

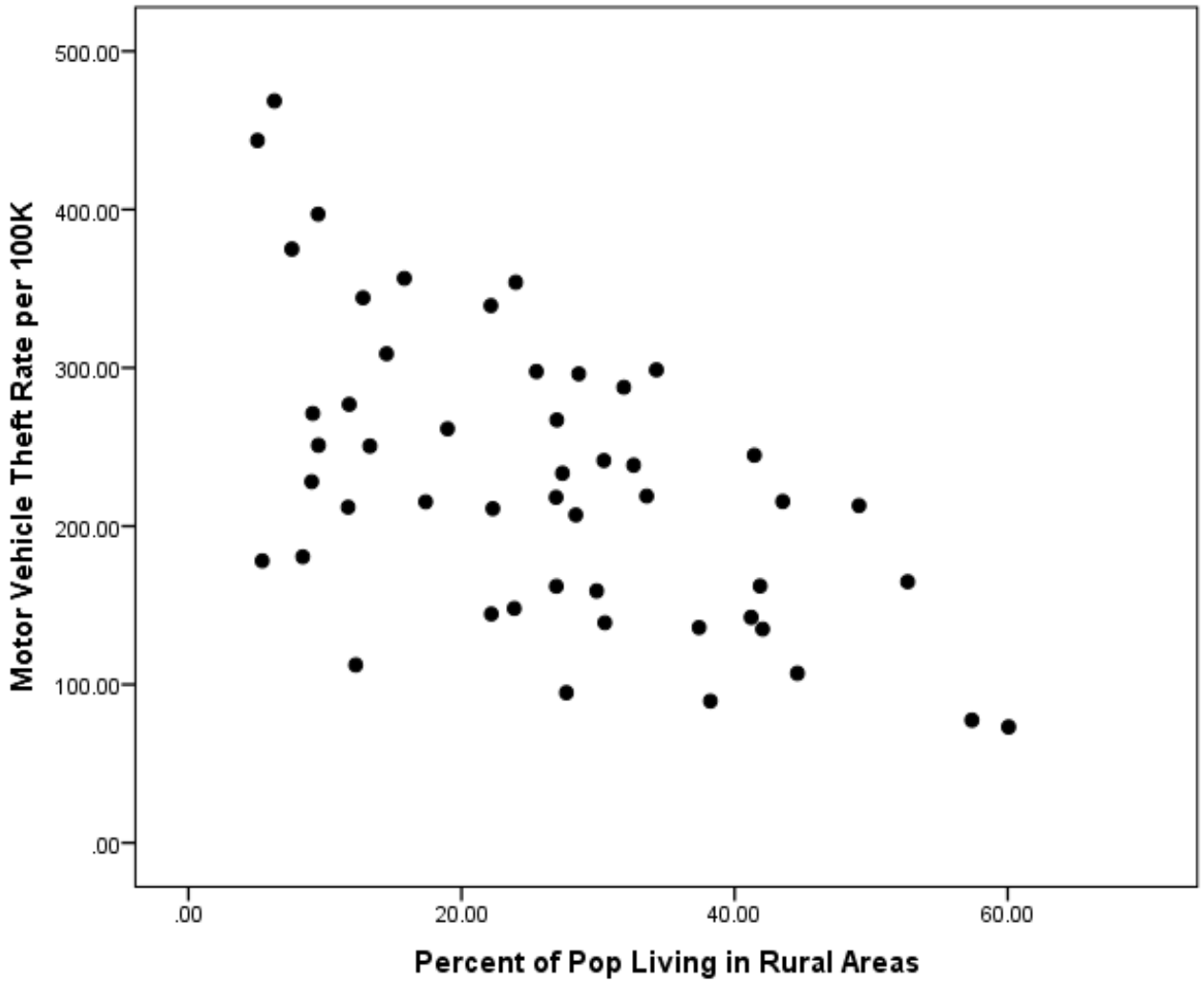
x3	y3
6	4
9	4
2	4
7	4
3	4
4	4
1	4
8	4
5	4
10	4

A scatterplot, where x is plotted on the horizontal axis and y is plotted on the vertical axis would graphically capture the bivariate relationship between x and y:



This graphically depicts a relationship where y does not change at all as x increases –this illustrates no relationship between the IV and DV.

In reality, of course, we don't have such perfect positive or negative relationships. Real scatterplots resemble a dart board rather than data points falling in a straight line.



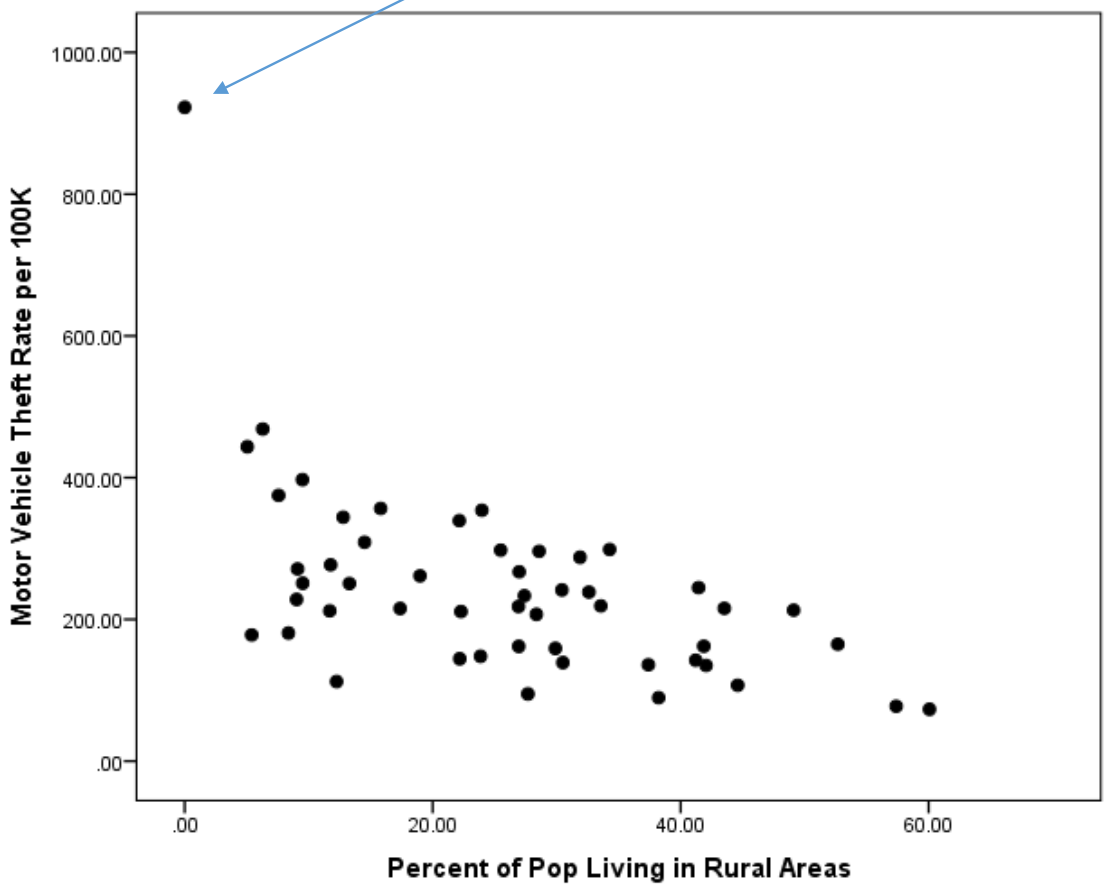
This is real state level data (without DC) illustrating a negative relationship, that is, as the percent rural population in a state increases, state motor vehicle rates decreases.

When we examine scatterplots, we are looking for several things:

- How close do the data points fall on a straight line – the strength of the relationship
- Whether the relationship is positive or negative - the direction of the relationship –
- If there are any bivariate outliers, or values that do not conform with the other data points.

What is a bivariate outlier?

This is a bivariate outlier – it is DC in this scatterplot of state-level data – it will bias estimates of statistics that attempt to quantify the relationship between these two variables!

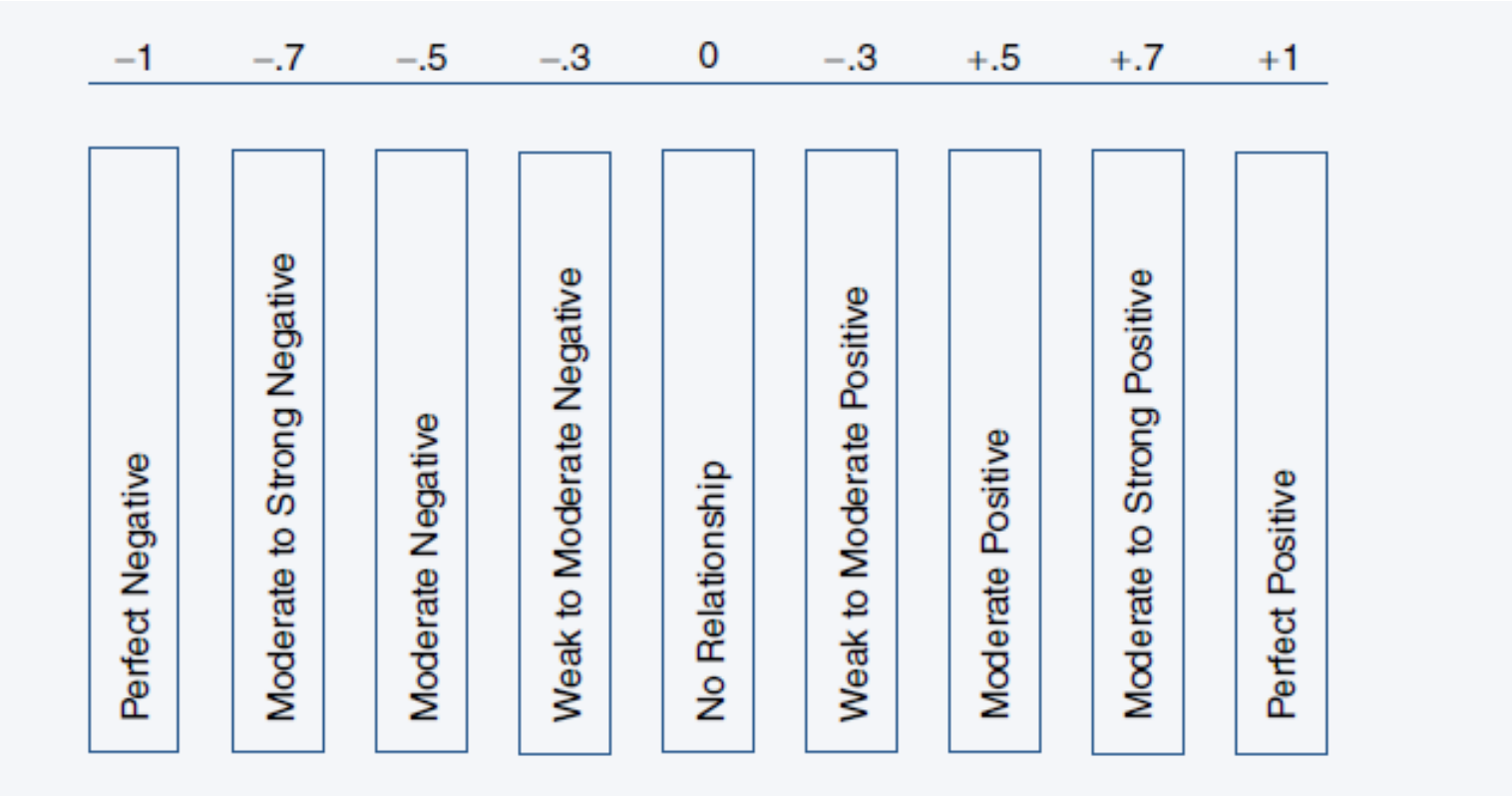


One statistic that quantifies the linear relationship between x and y is called the Pearson Correlation Coefficient (*r*)

$$r = \frac{\Sigma(x - \bar{X})(y - \bar{Y})}{\sqrt{[\Sigma(x - \bar{X})^2][\Sigma(y - \bar{Y})^2]}}$$

I won't go into the math for calculating *r,* but as you can see, it is essentially measuring the covariation between x and y! A covariation of 0 implies no relationship, while positive and negative signs indicate the direction of the relationship. The correlation coefficient is also standardized by the denominator!

Pearson's *r* Values Closer to Positive or Negative 1 Indicate Stronger Relationships



SPSS correlation matrix output

Correlations

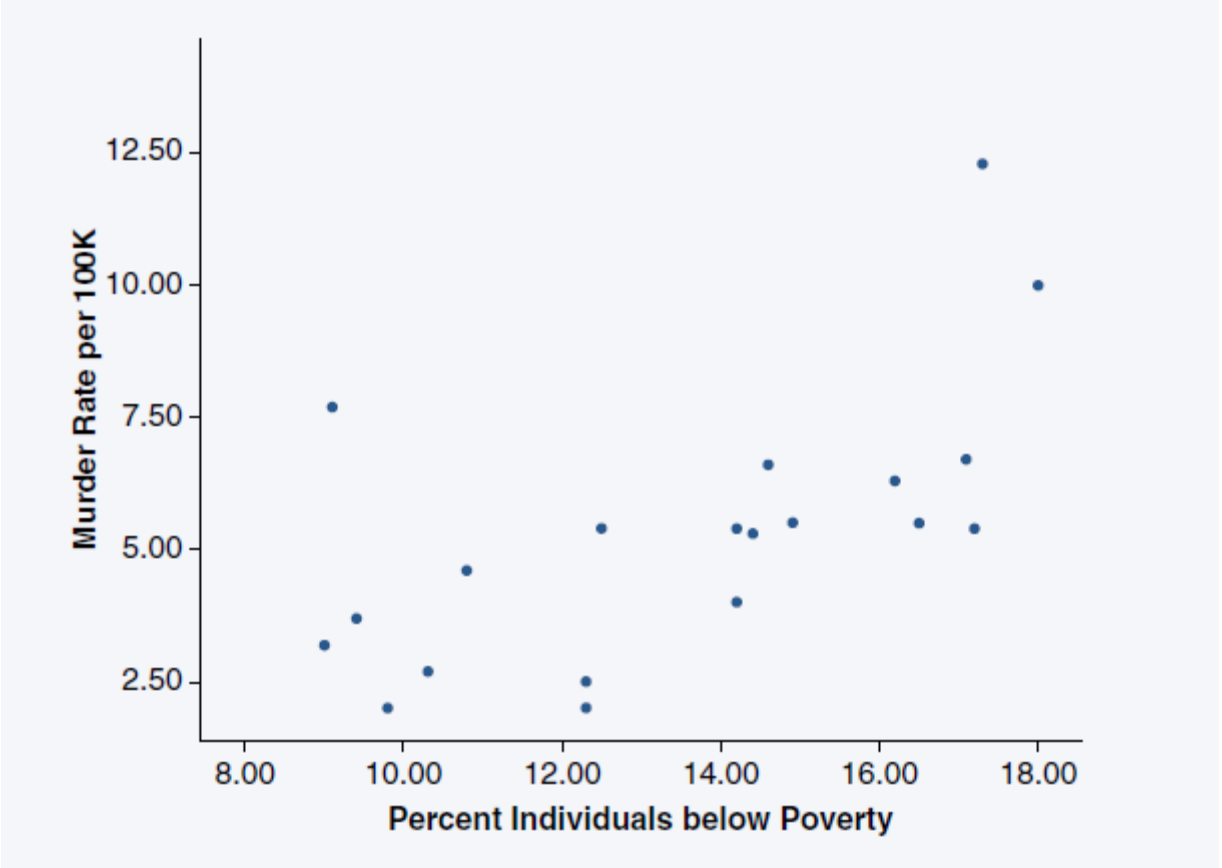
Correlations

		Murder Rate per 100K	Percent Individuals below poverty	Robbery Rate per 100K	Percent of Pop Living in Rural Areas	BurglaryRt	Divorces per 1K population
Murder Rate per 100K	Pearson Correlation	1	.621**	.450*	-.108	.738**	-.185
	Sig. (2-tailed)		.003	.046	.651	.000	.434
	N	20	20	20	20	20	20
Percent Individuals below poverty	Pearson Correlation	.621**	1	.118	.039	.749**	.004
	Sig. (2-tailed)	.003		.620	.869	.000	.986
	N	20	20	20	20	20	20
Robbery Rate per 100K	Pearson Correlation	.450*	.118	1	-.663**	.309	-.405
	Sig. (2-tailed)	.046	.620		.001	.185	.077
	N	20	20	20	20	20	20
Percent of Pop Living in Rural Areas	Pearson Correlation	-.108	.039	-.663**	1	-.014	.505*
	Sig. (2-tailed)	.651	.869	.001		.953	.023
	N	20	20	20	20	20	20
BurglaryRt	Pearson Correlation	.738**	.749**	.309	-.014	1	.055
	Sig. (2-tailed)	.000	.000	.185	.953		.817
	N	20	20	20	20	20	20
Divorces per 1K population	Pearson Correlation	-.185	.004	-.405	.505*	.055	1
	Sig. (2-tailed)	.434	.986	.077	.023	.817	
	N	20	20	20	20	20	20

**. Correlation is significant at the 0.01 level (2-tailed).

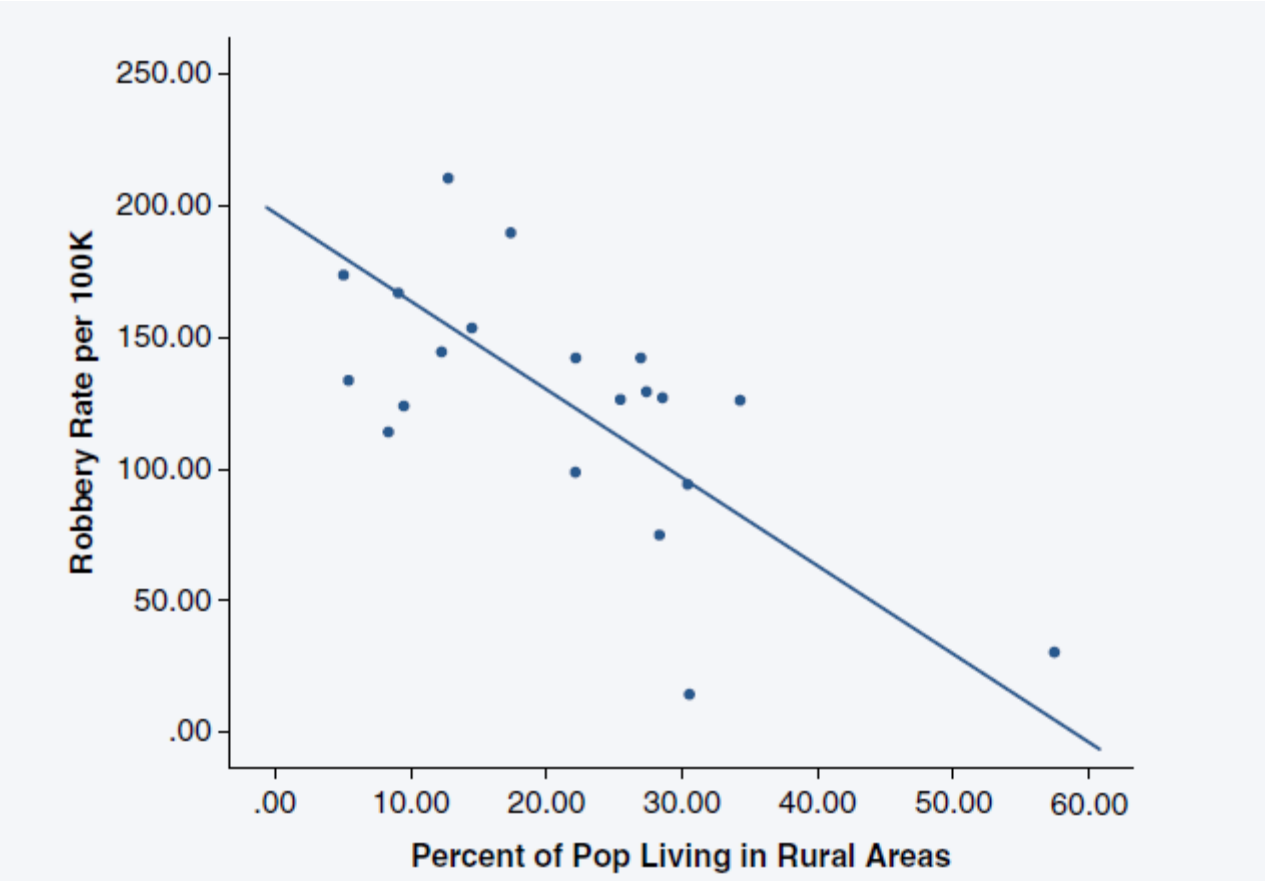
*. Correlation is significant at the 0.05 level (2-tailed).

Scatterplot between Murder Rate in State (y) and Poverty Rate (x), *n* = 20 States



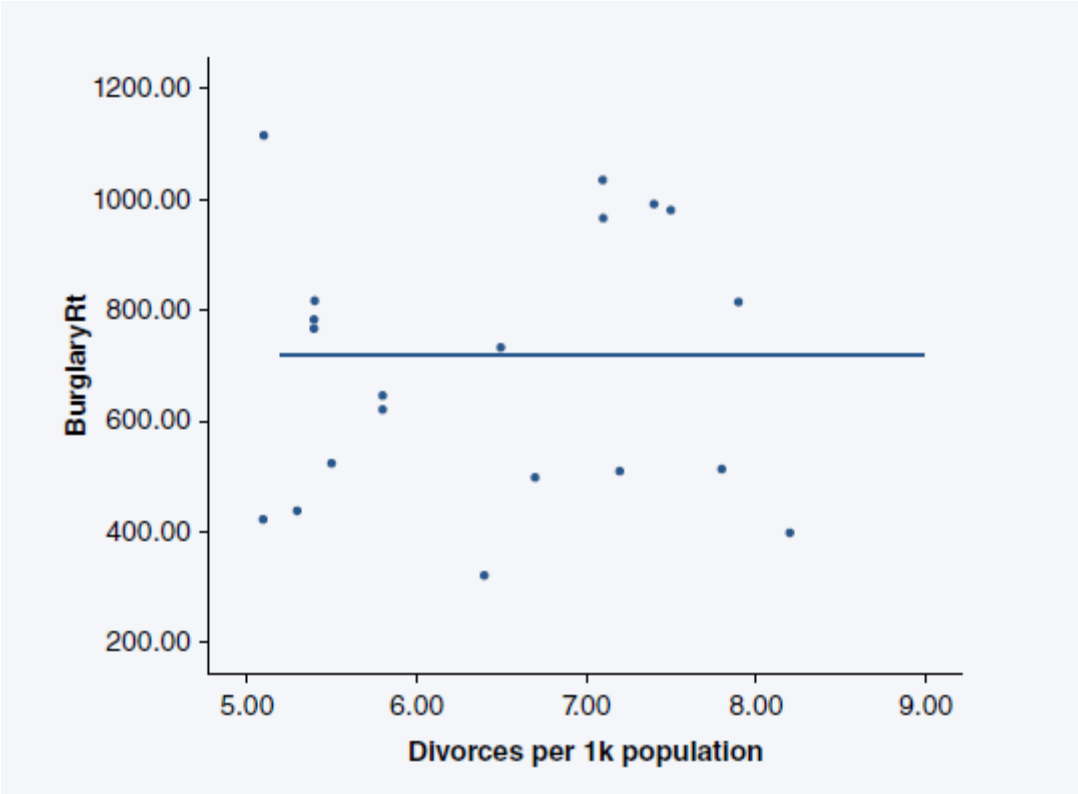
r = .621
Sig. = .003

Scatterplot between Robbery Rate in States (y) and Percent living in Rural Areas (x), *n* = 20 States



r = -.663
Sig. = .001

Scatterplot between Burglarly Rate in States (y) and Divorce Rate (x), *n* = 20 States



r = .055
Sig. = .817

A more precise way to interpret *r*

The Coefficient of Determination – r^2

r^2 = The proportion of the variation in y that is being explained by x.

	r	r^2
Rates of murder (y) and poverty (x) in states	.62	.38
Rates of robbery (y) and percent rural (x)	-.66	.44
Rates of burglary (y) and divorce rate (x)	.05	.02

So 38% of the variation in murder rates in states can be explained by poverty rates, and less than 1% of the variation in burglary rates in states can be explained by the divorce rate.

Ordinary Least Squares (OLS) Linear Regression -

Not only tell us the strength and the direction of the relationship between x and y, but it also tells us exactly how y changes with every one-unit increase in x – this allows us to make predictions about y!

Why the name 'least squares" – because it is calculated using the 'difference scores' of each *x* value from the mean of *x*, which you recall from the formula for the standard deviation must be squared to quantify the variation:

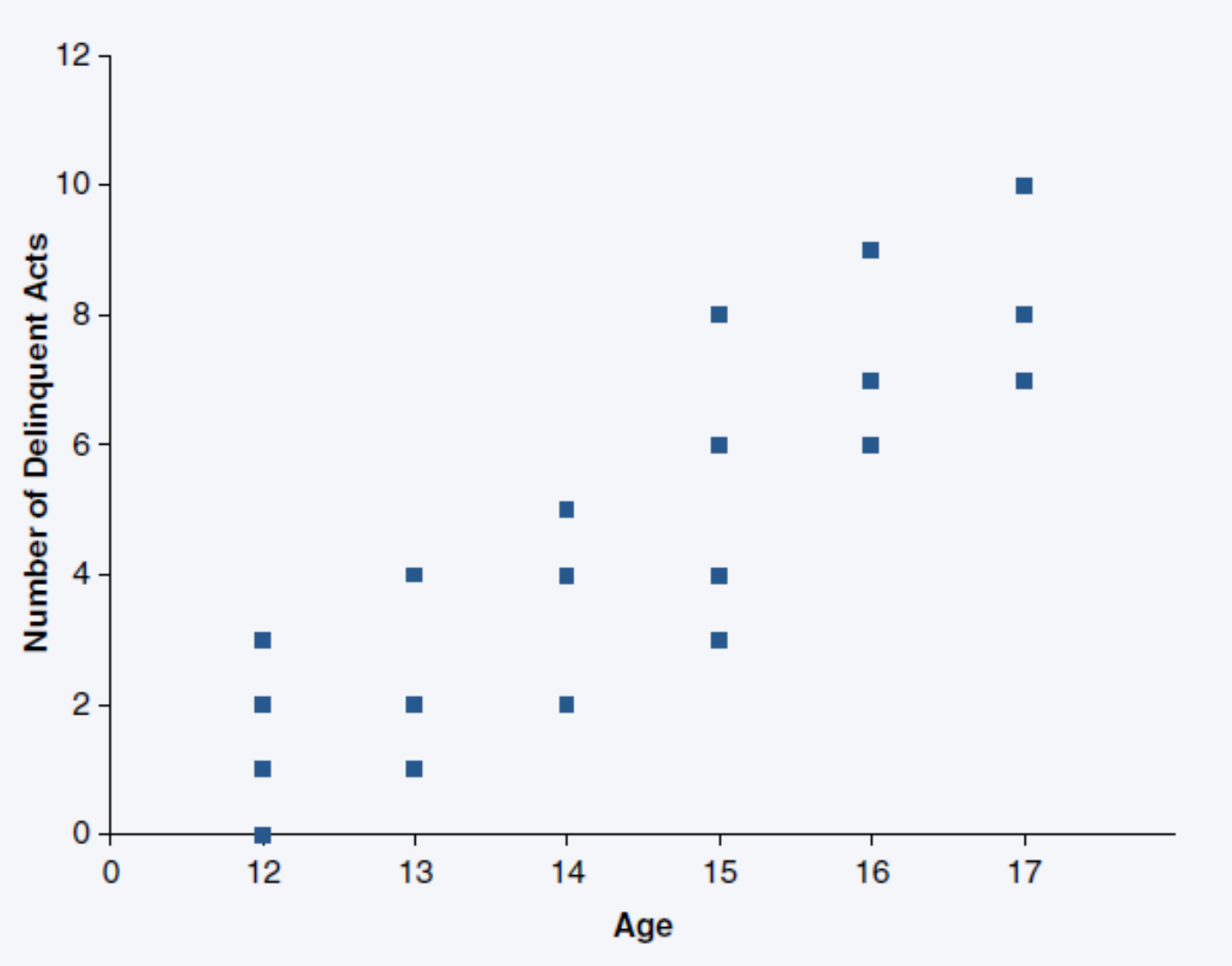
$$\Sigma(x - \bar{X}) = 0$$

$$\Sigma(x - \bar{X})^2 = \textit{Minimum Variance}$$

Assume we have these data for age (x) and delinquency scores (y)

Student	Age x	Self-Reported Delinquency y
1	12	0
2	12	2
3	12	1
4	12	3
5	13	4
6	13	2
7	13	1
8	14	2
9	14	5
10	14	4
11	15	3
12	15	4
13	15	6
14	15	8
15	16	9
16	16	7
17	16	6
18	17	8
19	17	10
20	17	7

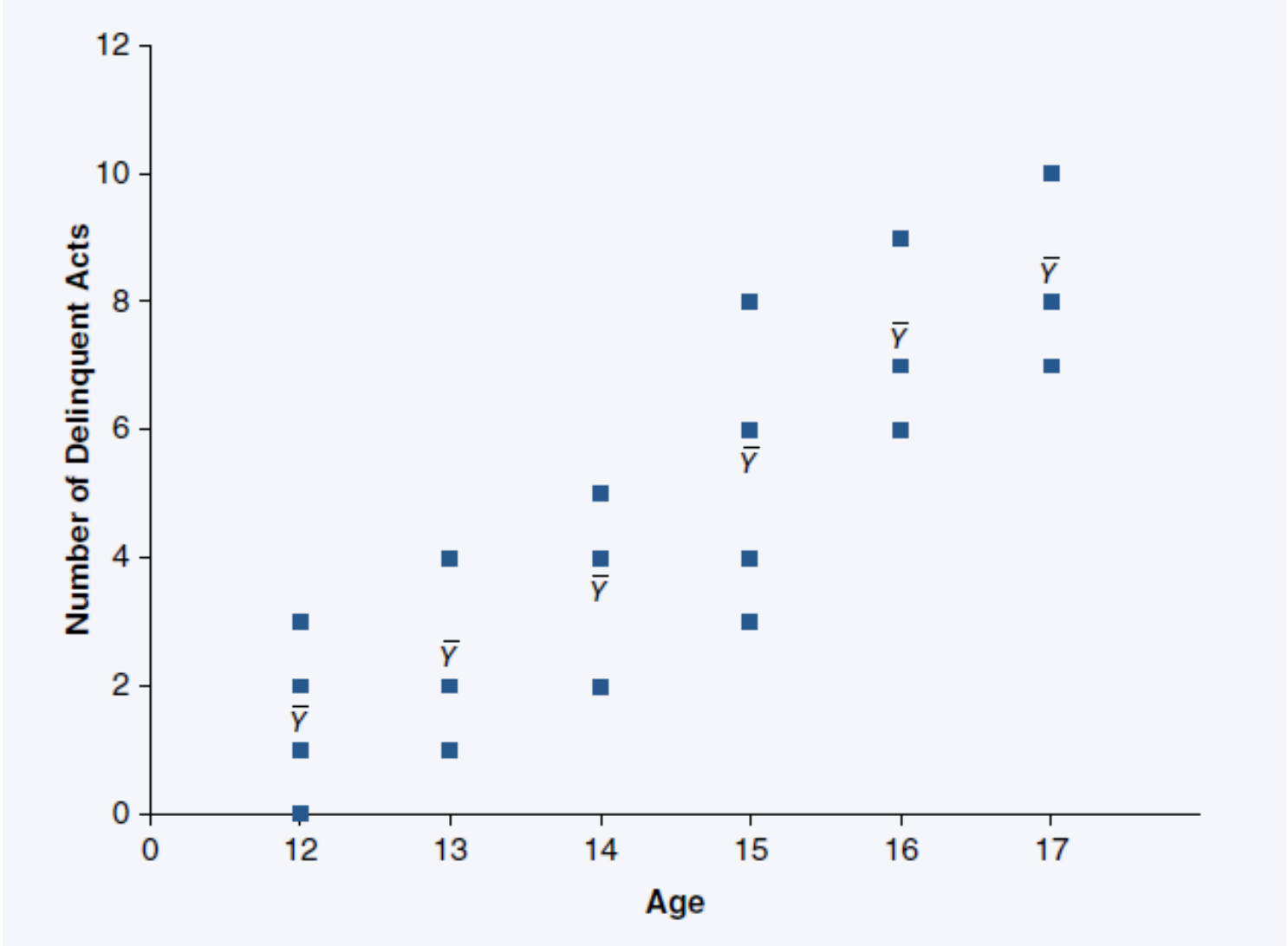
Scatterplot of Age (x) and Delinquency Rate (y)



If we calculate the mean delinquency score at each age value (x), and then draw a line through the scatterplot using these 'conditional means,' it would be the 'best fitting line' we could estimate mathematically because all the x values would fall closest to these conditional means, and hence to the line, compared to any other value

Age	*y Scores*	*Conditional* $\bar{Y}$
12	0, 1, 2, 3	1.5
13	4, 2, 1	2.3
14	2, 5, 4	3.7
15	3, 4, 6, 8	5.2
16	9, 7, 6	7.3
17	8, 10, 7	8.3

Visualize the line going through these conditional means of y at every value of x



The Specific Equation for the Ordinary Least Squares Regression Line:

$$y = \alpha + \beta x$$

---

where

y = the score on the y variable

α = the y intercept or constant

β = the slope of the regression line

x = the score on the x variable

OLS Equation for Sample Data:

$y = a + bx$

Regression

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Percent Individuals below poverty[b]	.	Enter

a. Dependent Variable: Murder Rate per 100K
b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.621[a]	.386	.352	2.0824

a. Predictors: (Constant), Percent Individuals below poverty

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	49.054	1	49.054	11.312	.003[b]
	Residual	78.054	18	4.336		
	Total	127.108	19			

a. Dependent Variable: Murder Rate per 100K
b. Predictors: (Constant), Percent Individuals below poverty

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-1.828	2.182		-.838	.413
	Percent Individuals below poverty	.531	.158	.621	3.363	.003

a. Dependent Variable: Murder Rate per 100K

a = intercept

b = slope

H_0: No relationship between IV and DV or slope = 0

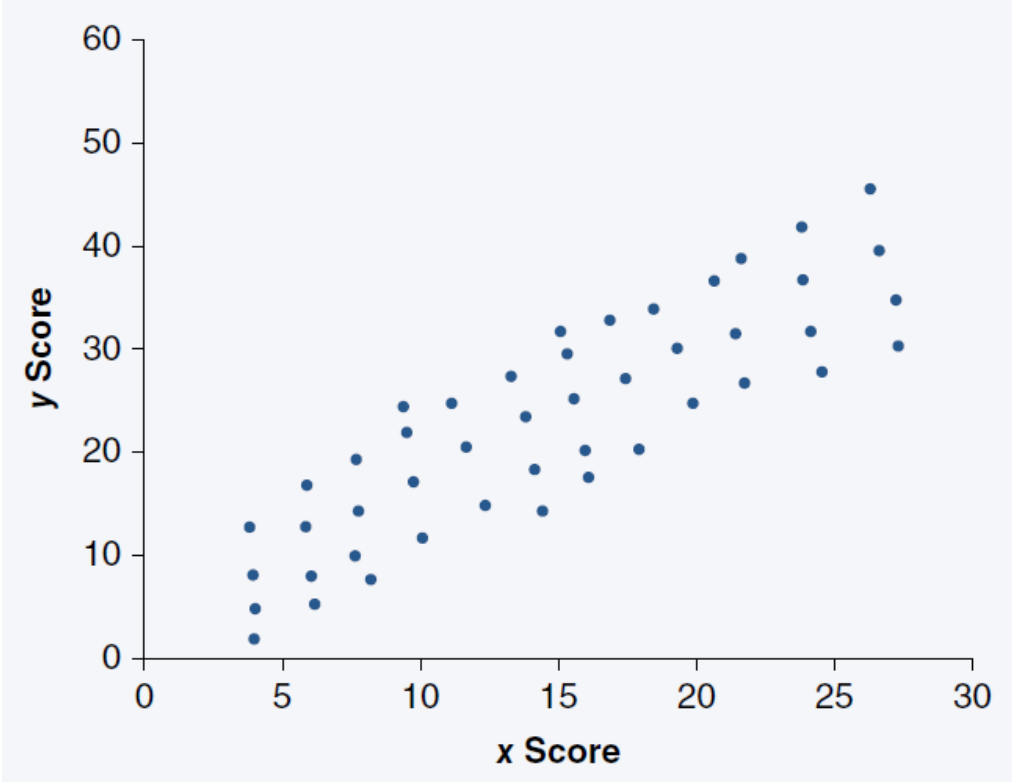
$$\text{Murder Rate } (y) = -1.828 + .53\ (x_{\text{poverty}})$$

Assumptions Necessary to Test Null Hypotheses (H0) for OLS Regression and Correlation Coefficients in the Population (β and ρ)

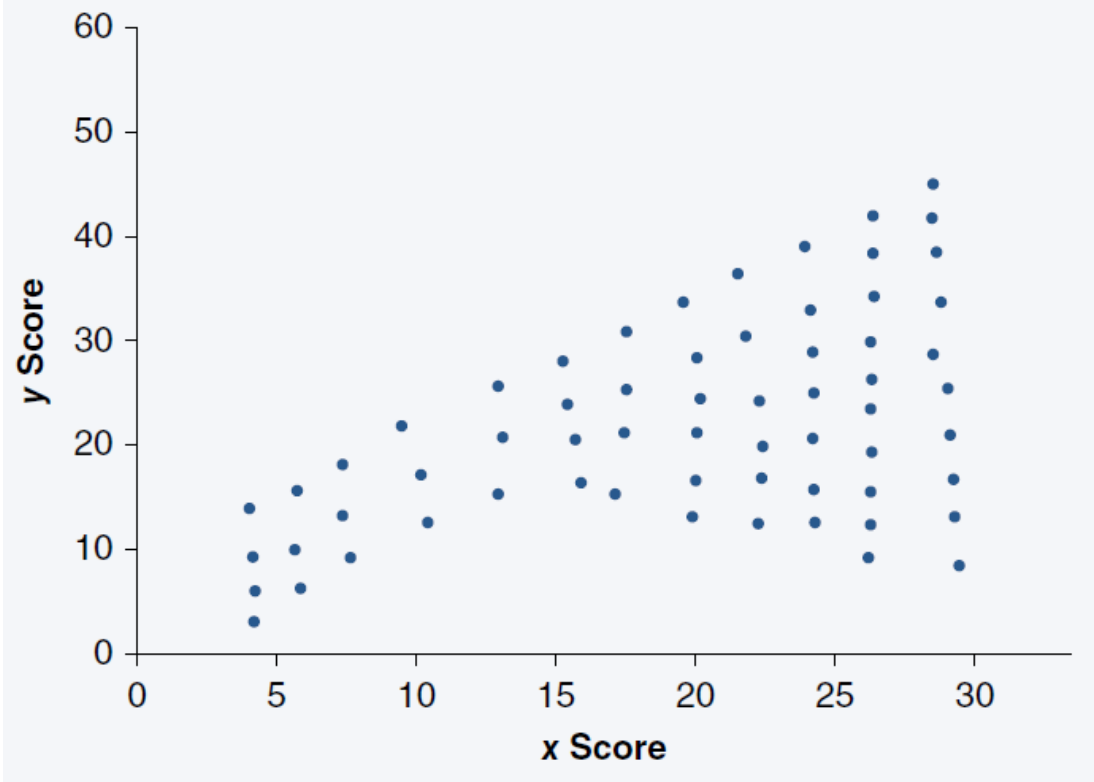
1. The observations were randomly selected.
2. Both variables have normal distributions.
3. The two variables are measured at the interval/ratio level (except for dichotomous IVs)
4. The variables are related in a linear form.
5. The error component is independent of and therefore uncorrelated with the independent or *x* variable, is normally distributed, has an expected value of zero, and has a constant variance across all levels of *x* (assumption of homoscedasticity).

Testing the Homoscedasticity Assumption – plotting residuals

ASSUMPTION NOT VIOLATED – RESIDUALS HAVE A CONSTANT VARIANCE ACROSS X VALUES



ASSUMPTION IS VIOLATED – RESIDUALS DO NOT HAVE A CONSTANT VARIANCE ACROSS X VALUES



Let's use the Regression equation for Poverty and Murder for Prediction:

$$\text{Murder rates } (y) = -1.828 + .53(x_{\text{Poverty}})$$

Recap: The correlation between rates of murder and poverty in states was .62, indicating that 38% of the variation in murder was being explained by rates of poverty in states. The slope indicates that as the poverty rate increases by one unit in states, rates of murder also increase by .53 units. This relationship was significant at the .003 level.

Predicted Murder Rate for State with Low Poverty ($x = 4$)

$$\text{Murder rate } (\hat{y}) = -1.828 + .53(4)$$

$$\text{Murder rate } (\hat{y}) = -1.828 + 2.12 = .29$$

Predicted Murder Rate for State with High Poverty ($x = 24$)

$$\text{Murder rate } (\hat{y}) = -1.828 + .53(24)$$

$$\text{Murder rate } (\hat{y}) = -1.828 + 12.72 = 10.89$$

Predicting State Level Robbery Rates (y) Using Percent of Population Living in Rural Areas (x)

Regression

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Percent of Pop Living in Rural Areas[b]	.	Enter

a. Dependent Variable: Robbery Rate per 100K
b. All requested variables entered.

The correlation in regression output is ALWAYS positive – it does not reflect the direction of the relationship! Why? Because when other IVs are added to the model, the slope coefficients will be both positive and negative!

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.663[a]	.440	.409	36.5968

a. Predictors: (Constant), Percent of Pop Living in Rural Areas

The correlation is moderate; 44% of the variation in robbery rates in states can be explained by rurality (percent living in rural areas).

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	18956.524	1	18956.524	14.154	.001[b]
	Residual	24107.888	18	1339.327		
	Total	43064.412	19			

a. Dependent Variable: Robbery Rate per 100K
b. Predictors: (Constant), Percent of Pop Living in Rural Areas

This F test is redundant at the bivariate level with the t test for the slope coefficient below

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	179.468	16.441		10.916	.000
	Percent of Pop Living in Rural Areas	-2.507	.666	-.663	-3.762	.001

a. Dependent Variable: Robbery Rate per 100K

Robbery (y) = 179.468 + -2.507 (x_{Rural})

When percent rural in a state increases by 1 unit, the robbery rate decreases by 2.507 units

H_0: $\beta=0$

We can reject the null at the alpha .01 level ($\alpha=.01$) and conclude that states with higher rates of rural population also have lower rates of robbery.

Predicting Burglary Rates (y) with the Divorce Rate (x)

Regression

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Divorces per 1K population[b]	.	Enter

a. Dependent Variable: BurglaryRt
b. All requested variables entered.

The correlation shows a very weak relationship, with less than 1 percent of the variation in burglary rates in states being explained by divorce rates.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.055[a]	.003	-.052	245.7588

a. Predictors: (Constant), Divorces per 1K population

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3330.000	1	3330.000	.055	.817[b]
	Residual	1087153.105	18	60397.395		
	Total	1090483.105	19			

a. Dependent Variable: BurglaryRt
b. Predictors: (Constant), Divorces per 1K population

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	613.778	353.465		1.736	.100
	Divorces per 1K population	12.751	54.303	.055	.235	.817

a. Dependent Variable: BurglaryRt

y (burglary rates) = 613.778 + 12.751 ($x_{divorce}$)

For every one unit increase in the divorce rate in states, burglary rates increase by 12.75 units. This is relationship is not significant!

OLS Can Also Handle IV's that are dichotomous and coded 0 and 1

For example, when predicting violent crime rates, the regional indicator of southern location is always important to examine as states in the South generally have higher rate of violent crime than states in the non-South.

In the following SPSS output, a variable called "South" is coded 1 for all states in the South and 0 for all states in the Non-South.

This dichotomous variable (South) is used as the independent variable (x) predicting murder rates (y) in states.

Predicting Murder Rates in States (y) Using Southern Dichotomous Indicator (x)

Regression

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	State in South[b]	.	Enter

a. Dependent Variable: Murder Rate per 100K
b. All requested variables entered.

The correlation is weak/moderate; 19.3% of the variation in state rates of murder can be explained by regional location, e.g. whether the state is located in the South versus Non-South

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.440[a]	.193	.149	2.3867

a. Predictors: (Constant), State in South

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	24.578	1	24.578	4.315	.052[b]
	Residual	102.530	18	5.696		
	Total	127.108	19			

a. Dependent Variable: Murder Rate per 100K
b. Predictors: (Constant), State in South

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	4.614	.638		7.234	.000
	State in South	2.419	1.165	.440	2.077	.052

a. Dependent Variable: Murder Rate per 100K

$$y \text{ (murder rates)} = 4.61 + 2.419\ (x_{\text{South}})$$

Interpretation of Dichotomous IV Continued:

$$\text{Murder rates } (y) = 4.61 + 2.419(x_{\text{South}})$$

When you interpret the slope coefficient for a dichotomy, you must do so relative to what is coded 0 and 1. If the coefficient (b) is positive, it indicates that *y increases* when *x* goes from 0 to 1. If *b* is negative, it indicates that *y decreases* as *x* goes from 0 to 1.

This coefficient indicates that, compared to states in the NonSouth (coded 0), murder rates in the South (coded 1) increase by 2.4 units.

You can see this mathematically when you predict murder rates using the equation:

Predicting murder rate (*y*) for States in the NonSouth:

$$\text{Murder rates } (y) = 4.61 + 2.419(0) = 4.61$$

Predicting murder rate (*y*) for States in the South:

$$\text{Murder rates } (y) = 4.61 + 2.419(1) = 7.029$$

One More Example: DV = Sentence Length Received (in days) by Murder Defendants
IV: Type of Adjudication: 1 = Jury Trial, 0 = Plea

Regression

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	jurytrial[b]	.	Enter

a. Dependent Variable: incarceration term in days
b. All requested variables entered.

Note: Individual level data generally always have more measurement error and variability, and as a result, typically have very small correlation and r^2 values - so this is typical!

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.273[a]	.075	.074	14091.541

a. Predictors: (Constant), jurytrial

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	16115320630.000	1	16115320630.000	81.156	.000[b]
	Residual	199762963300.000	1006	198571534.100		
	Total	215878283900.000	1007			

a. Dependent Variable: incarceration term in days
b. Predictors: (Constant), jurytrial

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	8947.369	907.716		9.857	.000
	jurytrial	9374.401	1040.597	.273	9.009	.000

a. Dependent Variable: incarceration term in days

$$y\ (\text{sentence length}) = 8947.369 + 9374.4\ (x_{\text{JuryTrial}})$$

Sentence Length in Days (y) = 8947.369 + 9374.4 ($x_{JuryTrial}$)

This coefficient indicates that, compared to murder defendants who plead guilty (coded 0), sentence length for defendants who were convicted as a result of a jury trial (coded 10), increase by 9374.4 units, or about 26 years!

You can see this mathematically when you predict murder rates using the equation:

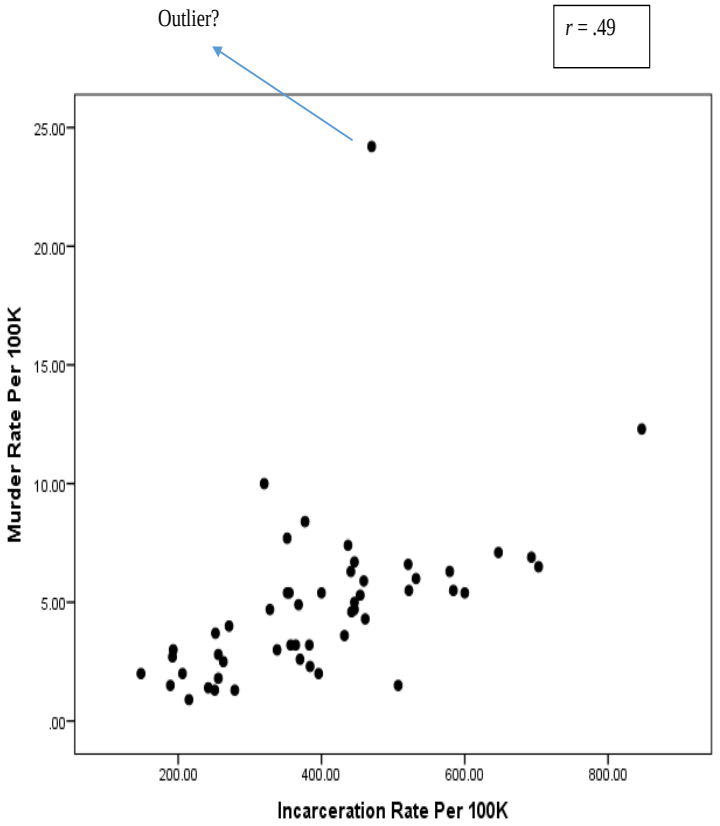
Predicting sentence length (y) for defendants who plead guilty:

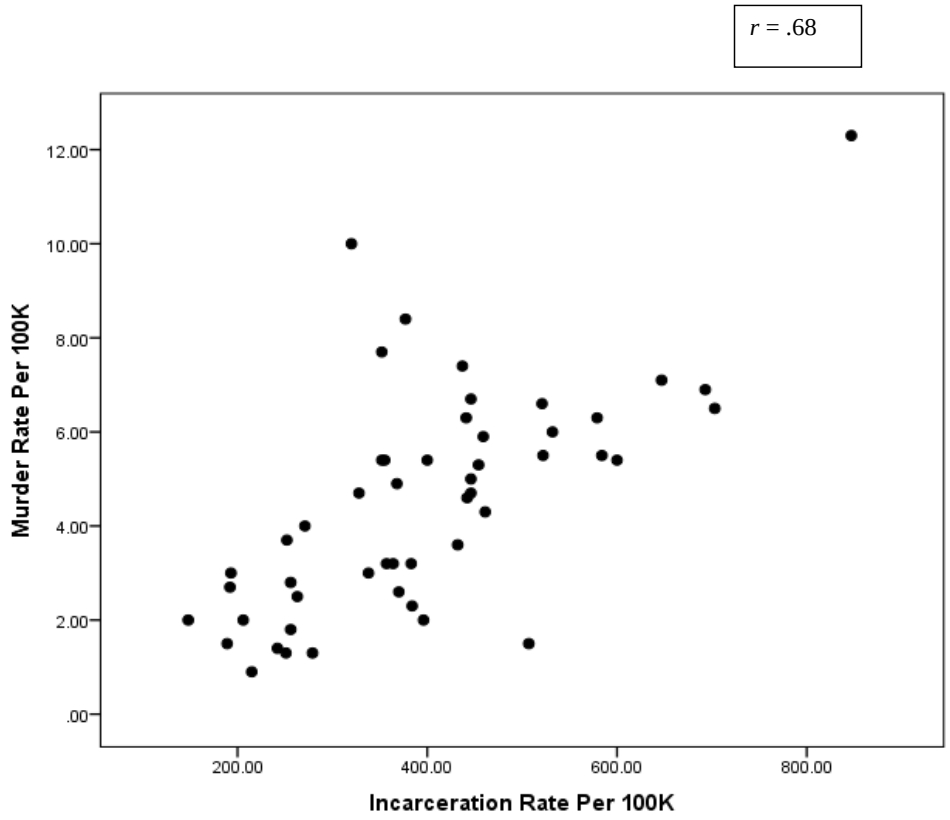
Sentence Length in Days (y) = 8947.369 + 9374.4 (0) = 8947.369 or about 24.5 years

Predicting sentence length (y) for defendants who went to trial:

Sentence Length in Days (y) = 8947.369 + 9374.4 (1) = 18321.77 or about 50.2 years

Another Word about Bivariate Outliers: Do Incarceration Rates affect Murder Rates?





Always Examine your data!

Now that we understand OLS Bivariate Regression, let's do some practice problems using SPSS!