Section I
Linear Regression

So far we have discussed describing and summarizing one variable, but very often we want to know if two or more variables are related and if they are related, we want to describe that relationship. One way to analyze the relationship between two or more variables is a method called linear regression, specifically least squares linear regression. In this section, we will only look at the relationship between two variables. Note: least squares linear regression is not the only type of regression analysis, but it is the only type discussed in this course.

An **ordered pair** consists of values of two variables for each individual in the data set.

Data that consist of ordered pairs is called **bivariate data.**

Response variable (Dependent variable) is the variable whose value can be explained by the value of the **explanatory** or **predicator variable (independent variable)**.

Example: GPA depends on Number of Hours Studied, Height depends on Shoe Size

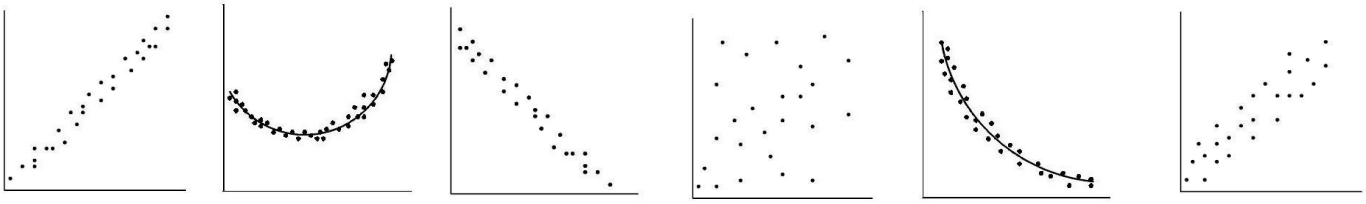
Scatter plot is a graph that shows the relationship between two quantitative variables, measured on the same individual.

Explanatory variable is on the horizontal axis (x-axis)

Response variable is on the vertical axis (y-axis)

Note: cannot always be sure which is which (does weight depend on height? or does height depend on weight?)

Determine whether a linear, nonlinear or no relationship exists:



Positive Linear | **Nonlinear** | **Negative Linear** | **No Relationship** | **Nonlinear** | **Positive Linear**

We don't just want to look at a scatter plot to determine if there is a relationship we also want to determine how strong that linear relationship is, therefore we want to find the **linear correlation coefficient**.

The **linear correlation coefficient** is a measure of the strength and direction of the linear relation between two quantitative variables. We use the Greek letter ρ (rho) to represent the population correlation coefficient and r to represent the sample correlation coefficient. The following is the formula for the sample correlation coefficient:

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right)}\sqrt{\left(\sum y^2 - \frac{(\sum y)^2}{n}\right)}}$$

Please note: you will NOT be using this formula to calculate the linear correlation coefficient, you will be learning how to use your calculator and reading a Minitab printout to find the linear correlation coefficient.

Properties of the Linear Correlation Coefficient

1) $-1 \le r \le 1$
2) If r = +1, then a perfect positive linear relation exists between the two variables.
3) If r = − 1, then a perfect negative linear relation exists between the two variables.
4) The closer r is to +1, the stronger is the evidence of positive linear association between the two variables.
5) The closer r is to −1, the stronger is the evidence of negative linear association between the two variables.
6) If r is close to 0, then little or no evidence exists of a linear relation between the two variables.
   Note: the linear correlation coefficient is a measure of the strength of the linear relation, r close to 0 does not imply no relation, just no linear relation.
7) The linear correlation coefficient is a unitless measure of association.
8) The correlation coefficient is not resistant. Therefore, an observation that does not follow the overall pattern of the data could affect the value of the linear correlation coefficient.

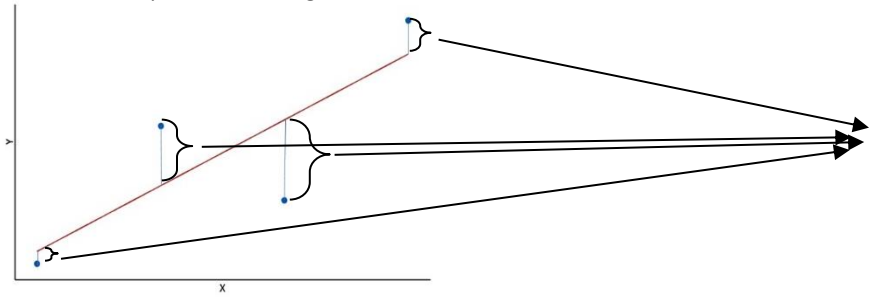
Note: Correlation is not the same as causation. In general, when two variables are correlated we cannot conclude that changing the value of one variable will cause a change in the value of the other.

Least-Squares Regression

Now we know that two variables have a linear relation, we want to find a line that best fits the points. One way to do this is to pick two points that appear to be a good fit of the data and find the line through those points. But is this the best line? Meaning will the predictions made be accurate.

The method we will be using to find the line that best fits the data is called **least-squares regression**. The line found by using this method is the line in which the sum of the squared vertical distances from the observed value and the line is as small as possible. This line is called the **least-squares regression line.** The least-squares regression line is written as a linear equation containing two variables, x and $\hat{y}$ and an equal sign.

Least-Squares Linear Regression Model



The sum of these distances squared would be the smallest for all the lines draw to fit these points.

Finding the Least-Squares Regression Line

Given ordered pairs (*x, y*), with means $\bar{x}$ and $\bar{y}$, sample standard deviations s_x and s_y, and correlation coefficient *r*, the equation of the least-squares regression line for predicting *y* from *x* is

$\hat{y} = b_0 + b_1 x$ where $b_1 = r * \frac{s_y}{s_x}$ is the **slope** and $b_0 = \bar{y} - b_1 \bar{x}$ is the **y-intercept**

In general, the variable we want to predict is call the **response variable** (dependent variable) and the variable we are given is called the **explanatory variable** or **predictor variable** (independent variable).

Please note: you will NOT be using these formulas to calculate the slope or y-intercept, you will be learning how to use your calculator and reading a Minitab printout to find the slope and y-intercept.

Note: The least-squares regression line goes through the **point of averages** $(\bar{x}, \bar{y})$.

Note: If r is positive, then the slope is positive. If r is negative, then the slope is negative.

Interpretation of Slope: (change in y)/(change in x) The slope of the best-fit line tells us how the dependent variable (y) changes for every one unit increase in the independent (x) variable, on average.

Example: For a line whose slope is 1.35, if x increases by 1, y will increase by 1.35.
If a line whose slope is –3, if x increases by 1, y will decrease by 3.

Interpretation of y-intercept: If the y-intercept is near the observed values, then the y-intercept is the value of the predicted y value when the x value is zero. If the y-intercept is not near the observed values then the y-intercept does not have a useful interpretation.

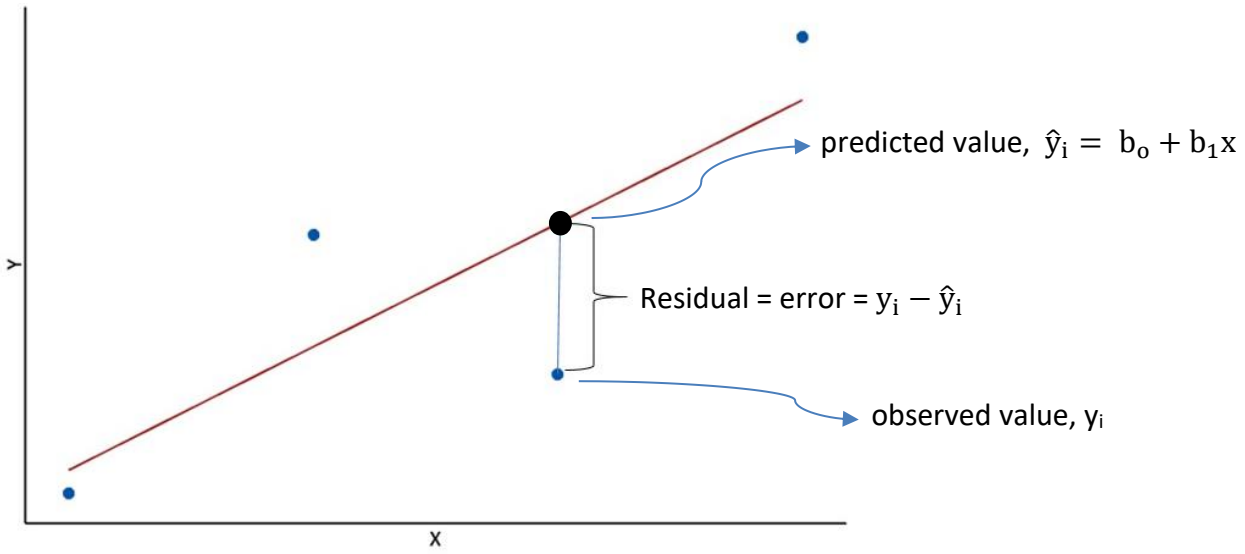
Diagnostics on the Least-Squares Regression Line

You don't want to use the least squares regression line to make predictions of the explanatory variable (x-values) that are much larger or much smaller than those observed. We don't know what happens outside the scope of the observed values, therefore you should not use the regression model to make predictions outside the scope of the model. Making predictions for values for the explanatory variable that are outside the range of the data is called **extrapolation.**

Residual Analysis is used to determine whether a linear model is appropriate to describe the relation between the explanatory and response variables.

Given a point (x, y) on a scatterplot, and the least-squares regression line $\hat{y} = b_o + b_1 x$, the **residual** for the point (x, y) is the difference between the observed value of y and the predicted value $\hat{y}$.
Residual = error = $y - \hat{y}$

Least-Squares Linear Regression Model



For example, if the least squares regression equation is found to be $\hat{y} = 10 + 6x$ and one of the observed points from the data set was (3, 25.75).

Then the predicted value when x = 3 would be $\hat{y} = 10 + 6(3) = 28$.

So the residual for this observation would be, **residual** = 25.75 – 28 = **–2.25**.

Note: The residuals are positive for points above the line and negative for points below the line.

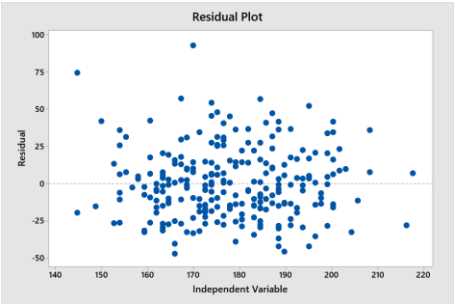
The least-squares regression line satisfies the least-squares property. This means that the sum of the squared residuals is less for the least-squares regression line than for any other line.

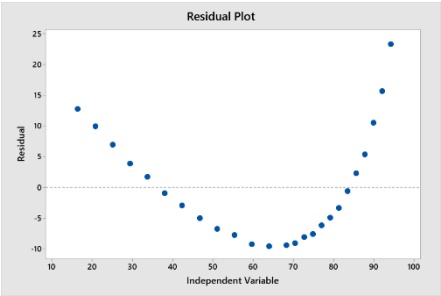
A **residual plot** is a plot in which the residuals are plotted against the values of the explanatory variable x.

Note: 1)When a residual plot exhibits a noticeable pattern, the variables do not have a linear relationship, and the least-squares regression line should not be used.

2) When a residual plot exhibits no noticeable pattern, the least-squares line may be used to describe the relationship between the variables.

In which plot below would a least-squares line be used to describe the relationship between two variables? Why?

A) 

B) 

Graph A, because the residual plot appears to be random, while in graph B there appears to be a pattern

You cannot just rely on the correlation coefficient to determine whether two variables have a linear relationship; even when the correlation is close to 1 or −1, the relationship may not be linear, a residual should be constructed to determine whether two variables have a linear relationship.

Determining Outliers and Influential Points in a Regression Model

An **outlier** is an observation that does not fit the overall pattern of the data. An outlier can be determined by a residual plot, a boxplot of the residuals or using a Minitab printout. An outlier has a standard residual that is either greater than 2 or less than -2.

An **influential point** is a point that, when included in a scatterplot, strongly affects the position of the least-squares regression line. i.e. an influential point is an observation that significantly affects the value of the slope and/or y-intercept of the least-squares regression line and the value of the correlation coefficient.

Please note you will using a Minitab printout to determine outliers and/or influential observations.

Coefficient of Determination, r^2, measures the proportion of total variation in the response variable that is explained by the least-squares regression line. Since r^2 is a proportion, it can never be negative or greater than 1. $(0 \le r^2 \le 1)$

$r^2 = 0$ means the least-squares regression line has no explanatory value

$r^2 = 1$ means the least-squares regression line explains 100% of the variation in the response variable (i.e. the closer r^2 is to 1, the closer the predictions made by the least-squares regression line are to the actual values, on average.)

The coefficient of determination is a measure of how well the least-square regression line describes the relation between the explanatory and response variable. The closer r^2 is to 1, the better the line describes how changes in the explanatory variable affect the value of the response variable.

Note: Squaring the linear correlation coefficient to obtain the coefficient of determination works only for the least-squares linear regression model line $\hat{y} = b_o + b_1 x$

Summary

The coefficient of determination r^2 measures the proportion of the variation in the outcome variable that is explained by the least-squares regression line.

The larger the value of r^2, the closer the predictions made by the least-squares regression line are to the actual values, on average.

To compute the coefficient of determination, first compute the correlation coefficient, then square it to obtain r^2. (Only works for least-square linear regression model.)

Regression Examples using the calculator

1) A random sample of 11 statistics students produced the following data, where *x* is the third exam score out of 80, and *y* is the final exam score out of 200. Can you predict the final exam score of a random student if you know the third exam score?

Third exam score	65	67	71	71	66	75	67	70	71	69	69
Final exam score	175	184	187	185	170	198	183	182	185	180	178

a) What is the explanatory variable? **Third exam score** What is the response variable? **Final exam score**

b) What is the least squares regression equation? $\hat{y} = 2.14x + 34.07$

c) Interpret the meaning of the slope. **For every increase of 1 point on the 3rd exam score, the final exam score increases by 2.14 points**

d) What is the correlation coefficient, r? Interpret this result.

r = 0.85 **Moderate, Positive, Linear relationship**

.

e) What is the coefficient of determination, r^2? Interpret this result

$r^2 = (0.85)^2 = 0.7225$ **72.25% of the variation in final exam score can be explained by the third exam score.**

f) Suppose you received a score of 72 on the third exam, predict the score on the final exam.

$\hat{y} = 2.14(72) + 34.07 = 188.15$

2) SCUBA divers have maximum dive times they cannot exceed when going to different depths. The table below shows different depths with the maximum dive times in minutes.

Depth in feet	50	60	70	80	90
Maximum dive time	80	55	45	35	25

a) What is the explanatory variable? **Depth in feet** What is the response variable? **Maximum dive time**

b) What is the least squares regression equation? $\hat{y} = -1.3x + 139$

c) Interpret the meaning of the slope. **For every increase of 1foot in depth, the maximum dive time decreases by 1.3 minutes.**

d) What is the correlation coefficient, r? Interpret this result. **r = –0.97 Very strong, negative, linear relationship**

e) What is the coefficient of determination, r^2? Interpret this result. **$r^2 = (-0.97)^2 = 0.94$**
94% of the variation in maximum dive time can be explained by depth.

f) Predict the maximum dive time for a) 85 feet and b) 110 feet.
a) $\hat{y}$ = -1.3(85) + 139 = 28.5min b) can not predict since 110ft is not in the scope of the data

3) The following table shows the life expectancy for an individual born in the United States in certain years.

Year of Birth	1930	1940	1950	1965	1973	1982	1987	1992	2010
Life Expectancy	59.7	62.9	70.2	69.7	71.4	74.5	75	75.7	78.7

a) What is the explanatory variable? **Year of Birth** What is the response variable? **Life Expectancy**

b) What is the least squares regression equation? $\hat{y} = 0.227x - 377.24$

c) What is the correlation coefficient, r? Interpret this result.

r = 0.96 Very Strong, positive, linear relationship

.
d) What is the coefficient of determination, r^2? Interpret this result.

$r^2 = (0.96)^2 = 0.92$ 92% of the variation in life expectancy can be explained by year of birth.

e) Predict the life expectancy of an individual born in the United States in the year 2000.

$\hat{y} = 0.227(2000) - 377.24 = 76.76$ **years**

4) The following table gives the gold medal times for every other Summer Olympics for the women's 100-meter freestyle (swimming).

Year	1912	1924	1932	1952	1960	1968	1976	1984	1992	2000	2008	2016
Time (seconds)	82.2	72.4	66.8	66.8	61.2	60.0	55.65	55.92	54.64	53.8	53.1	52.7

a) What is the explanatory variable? **Year** What is the response variable? **Time in seconds**

b) What is the least squares regression equation? $\hat{y} = -0.257x + 567.44$

c) What is the correlation coefficient, r? Interpret this result.

r = –0.95 very strong, negative, linear relationship

.d) What is the coefficient of determination, r^2? Interpret this result.

$r^2 = (-0.95)^2 = 0.90$

90% of the variation in time can be explained by year

Minitab Regression Examples

1) The Kelley Blue Book provides information on wholesale and retail prices of cars. The Minitab printout below presents the years old the car is and the price in dollars. Determine the following:

a) What is the explanatory variable? **Age** What is the response variable? **Price**

b) What is the least squares regression equation? $\hat{y} = 20550 - 1580x$

c) What is the coefficient of determination, r^2? Interpret this result.

r^2 = 61.46% 61.46% of the variation in price can be explained by age

d) What is the correlation coefficient, r? Interpret this result.

r = $\sqrt{0.6146}$ = –0.784 (slope is negative so r must be negative)

Moderate, negative, linear relationship

e) Are there any outliers or influential observations? If so,
which observations are outliers? **Obs #40, 42, 81**
which are influential? **Obs #10, #27**

f) Suppose the age of the cars ranged from 1 to 11 years.

What would be the predicted price if the car that is 3.75 years old? $\hat{y} = 20550 - 1580(3.75) = 14625$

What would be the predicted price if the car if it was 15 years old? **Cannot predict, 15 is not in the scope of the data**

Regression Analysis: Price versus Age

Regression Equation

Price = 20550 - 1580 Age

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	20550	639	32.17	0.000	
Age	-1580	135	-11.71	0.000	1.00

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
2942.12	61.46%	61.01%	59.61%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	1187289816	1187289816	137.16	0.000
Age	1	1187289816	1187289816	137.16	0.000
Error	86	744423496	8656087		
Lack-of-Fit	9	157482746	17498083	2.30	0.024
Pure Error	77	586940750	7622607		
Total	87	1931713312			

Fits and Diagnostics for Unusual Observations

Obs	Price	Fit	Resid	Std Resid	
10	7987	3170	4817	1.74	X
27	4995	4750	245	0.09	X
40	24675	14230	10445	3.57	R
42	4321	11070	-6749	-2.32	R
81	22995	9490	13505	4.66	R

R Large residual
X Unusual X

2) Is the average a teacher gets pay and the amount spent per student in each of the 50 states and the District of Columbia have a linear relationship? The Minitab printout below presents the average teacher salary and the average amount spent per student. Determine the following:

a) What is the explanatory variable? **Spend** What is the response variable? **Pay**

b) What is the least squares regression equation? $\hat{y} = 22074 + 1.318x$

c) What is the coefficient of determination, r^2? Interpret this result.

r^2 = 70.14% 70.14% of the variation in pay can be explained by spend.

d) What is the correlation coefficient, r? Interpret this result.

r = $\sqrt{0.7014}$ = 0.837 moderate, positive, linear relationship

e) Are there any outliers or influential observations? If so,
which observations are outliers? **Obs #9,12,33**
which are influential? **Obs #31,33**

f) Suppose the amount spent per student was between $8,097 and $22,366. What would the predict school teacher annual salary to be:

1) if the amount spent per student was $24,000? **cannot predict, since $24,000 is not in the scope of the data**

2) if the amount spent per student was $12,500? $\hat{y} = 22074 + 1.318(12500) = 38549$

Regression Analysis: Average Pay of Teacher versus spending per student

Regression Equation

Average Pay of Teacher = 22074 + 1.318 spending per student

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	22074	1635	13.50	0.000	
spending per student	1.318	0.123	10.73	0.000	1.00

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
2852.39	70.14%	69.53%	65.17%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	936569507	936569507	115.11	0.000
spending per student	1	936569507	936569507	115.11	0.000
Error	49	398670200	8136127		
Total	50	1335239707			

Fits and Diagnostics for Unusual Observations

Obs	Average Pay of Teacher	Fit	Resid	Std Resid		
9	55209	47334	7875	2.90	R	
12	46790	40199	6591	2.34	R	
31	51443	50916	527	0.20		X
33	45589	52880	-7291	-2.90	R	X

R Large residual
X Unusual X