

Week Three: Introduction to Regression

Week Three Goals

- Learn about scatterplots, response variable and predictor variable
- Learn about correlation values
- Interpret scatterplots and correlation values
- Learn about the least-squares regression line
- Understand and Interpret slope and y-intercept
- Learn about coefficient of determination
- Use regression equations for predicting
- Calculate a residual
- Learn about residual plots
- Use StatHelper or Minitab 19 for regression
- Examples

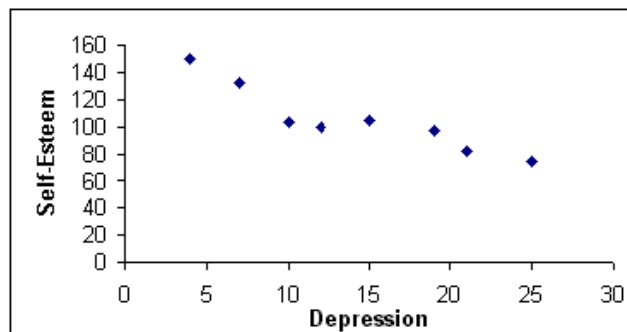
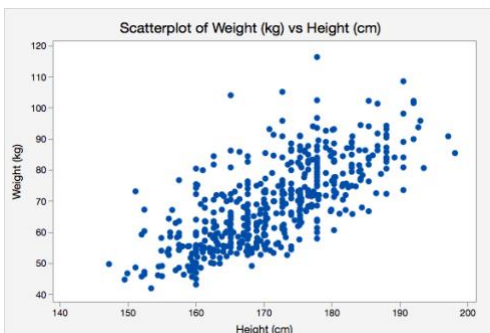
Scatterplot

In this unit, we will study 2 numerical variables and look to see if a possible linear relationship exists. We will begin by building and interpreting a scatterplot. A scatterplot is a graph that uses dots to represent the values obtained for two different variables - one plotted along the x-axis and the other plotted along the y-axis.

For example, this scatterplot below (left) shows the height and weight of a sample of American adults. Each dot represents one person with their height (in cm) placed along the x-axis and their weight (in kg) placed along the y-axis. As the height increases, the weight increases. There is a positive linear relationship.

The scatterplot on the right uses a depression level to try and predict a self-esteem level. As the depression level increases, the self-esteem level decreases. This is a negative relationship.

In this unit, we purposely pick the numerical variable that is independent to be placed on the x-axis (the horizontal axis) and the dependent variable to be on the y-axis (the vertical axis). Scatter plots are used when you want to show the relationship between two variables. Scatter plots are sometimes called correlation plots because they show how two variables are correlated.



The Response Variable

The y-variable is called the RESPONSE VARIABLE. It is also called the dependent variable. This is the variable that is being explained by the predictor variable. This is the variable we are trying to predict. The response variable is placed on the vertical axis (the y-axis).

The Predictor Variable

The x-variable is called the PREDICTOR VARIABLE. It is also called the explanatory variable, independent variable or INPUT VARIABLE. This is the variable that is being *used* to explain or predict the response variable. The predictor variable is placed on the horizontal axis (the x-axis).

GOAL: Read the problem carefully to determine which variable is x (predictor variable) and which variable is y (response variable).

Use this sentence template to help you find x and y:

“ PREDICTS ”
predictor variable response variable

For example:

“The **weight of the car** PREDICTS the **miles per gallon**”.

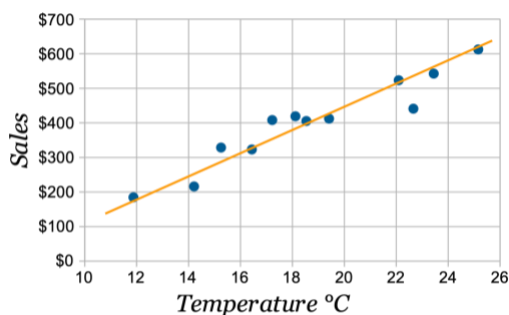
If you are not sure which variable is the predictor/input variable (x) and which is the response variable (y), read through the problem and questions completely and you will eventually find a question/sentence that tells you WHAT is being predicted. What is being predicted is ALWAYS your RESPONSE (y-variable).

For example, if the problem reads: *...using the regression line, predict the weight of a baby given the mother's age.*

Based on this statement, I can see that we are trying to **predict baby weight**. Therefore, I can determine that:

- Weight of the baby is the y-variable (response variable)
- Mother's age is the x-variable (predictor/input variable)

In the following scatterplot, the graph is not just a simple log of the noon temperatures (°C) and ice cream sales (\$) for a sample of summer days, but it also visualizes the relationship between temperature and sales by including the line of best fit. Viewing the scatterplot with the line helps you visualize the linear relationship; i.e. there appears to be a positive linear relationship between the noon temperature and the ice cream sales. As the temperature increases, the sales appear to increase (see [this link for a discussion](#) of this example). A scatterplot with a regression line on it is also called a **Fitted Line Plot**.



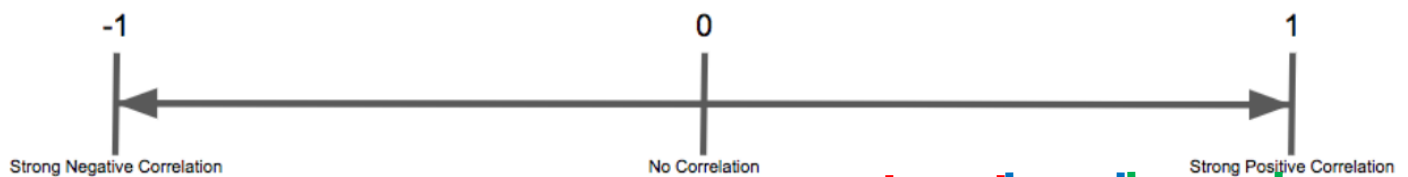
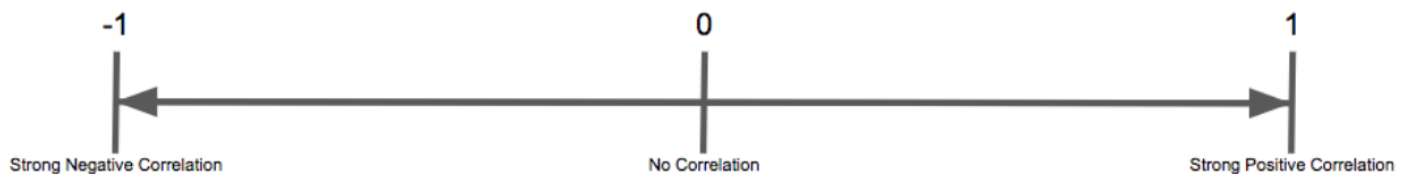
Correlation

The Correlation Value (or Correlation Coefficient) is a widely used method for determining the strength and direction of the linear relationship between two numbers or two sets of numbers. This coefficient is calculated using the following formula and is called Pearson's Correlation Coefficient. We will not be using the formula. We will use technology to calculate Pearson's r-value.

Correlation Coefficient Formula

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

The correlation value is denoted by the letter r or the Greek letter rho (ρ) and is a number between negative one and positive one, -1 being the strongest possible negative correlation; +1 being the strongest possible positive correlation.



If $r = -1$, then there is a perfect, negative linear relationship.

If r is between -1 and -0.80, then there is a very strong, negative linear relationship.

If r is between -0.80 and -0.50, then there is a moderate, negative linear relationship.

If r is between -0.50 and -0.25, then there is a weak, negative linear relationship.

If r is between -0.25 and +0.25, then there is a negligible linear association.

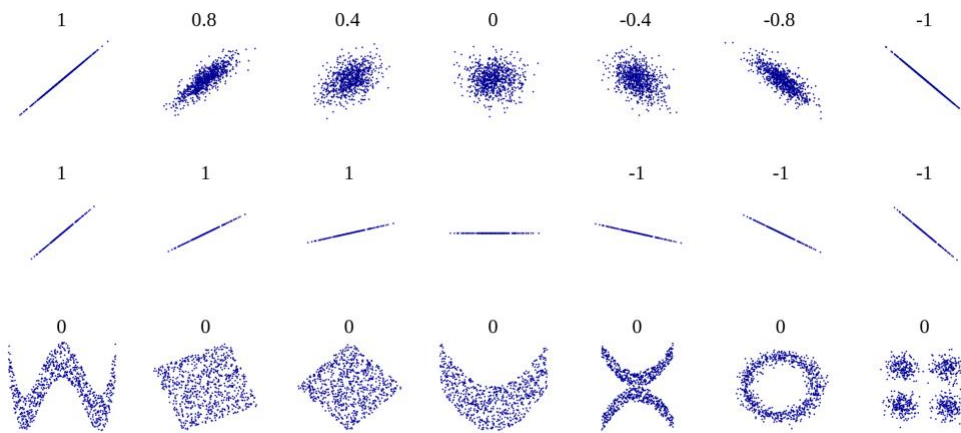
If r is between +0.25 and +0.50, then there is a weak, positive linear relationship.

If r is between +0.50 and +0.80, then there is a moderate, positive linear relationship.

If r is between +0.80 and +1, then there is a very strong, positive linear relationship.

Whether or not the outcome of the second number is CAUSED by the first is not being determined here, just that the outcomes of the two numbers happen in concert with each other. If the correlation is zero, then there is NO linear correlation between the two sets of numbers. There may be a different relationship. For example, a parabola will result in a linear correlation equal to zero.

Examples of Correlation Values and Scatterplots

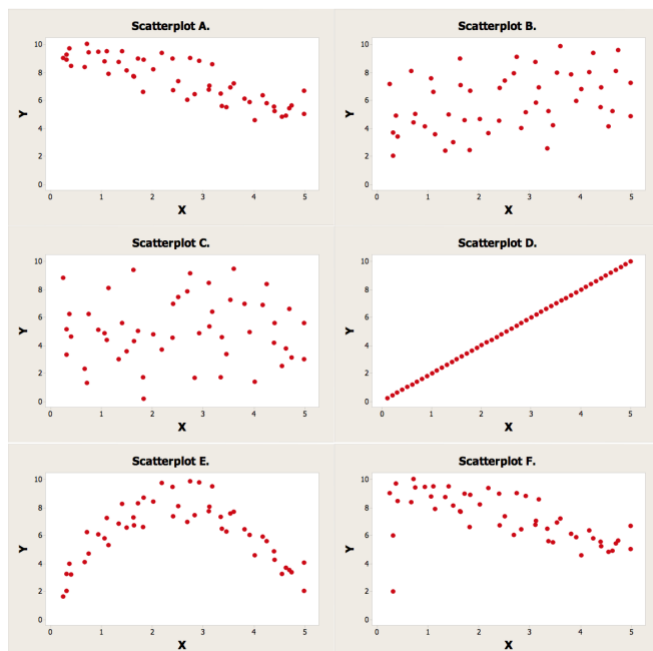


Range of r-values	Strength of the Linear Association
0	No linear association
0 to ± 0.25	Negligible linear association
± 0.25 to ± 0.50	Weak linear association
± 0.50 to ± 0.80	Moderate (or somewhat strong) linear association
± 0.80 to ± 1	Very Strong linear association
± 1	Perfect linear association

Example 1

Match the scatterplot with its correlation value. The answers are at the bottom.

$r = -0.9$ $r = -0.6$ $r = 0$ $r = 0$ $r = +0.4$ $r = +1$



Answers are on the next page...

Answers:

Matching

A. $r = -0.9$

B. $r = +0.4$

C. $r = 0$

D. $r = +1$

E. $r = 0$

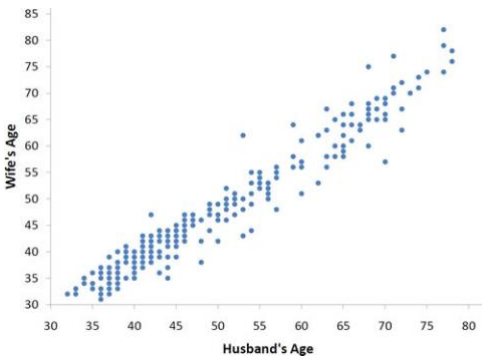
F. $r = -0.6$

Interpret a Scatterplot and/or Correlation Value

If you are asked to interpret a scatterplot or interpret the correlation value, please use a complete sentence and use the context of the problem. Describe the strength (strong/moderate/weak) and direction (positive/negative) of the linear relationship by using a template sentence such as:

“There appears to be a _____, _____ linear relationship between _____ and _____.”
(very) strong/moderate/weak positive/negative x y

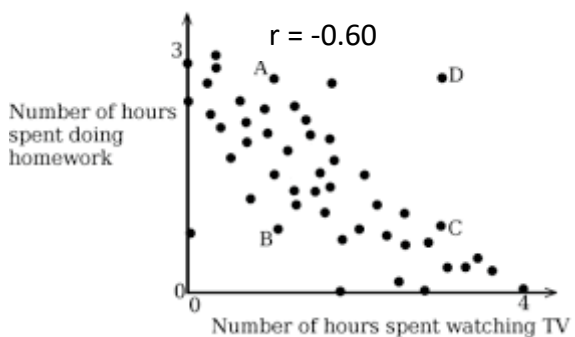
Example 2



Interpretation: There appears to be a strong, positive linear relationship between Husband's Age and Wife's Age.

In other words, as the husband's age increases, the wife's age also increases. The line of best fit will have positive slope.

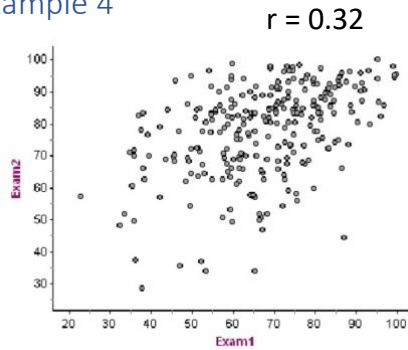
Example 3



Interpretation: There appears to be a moderately strong, **negative** relationship between # hours spent watching TV and # hours spent doing homework.

In other words, as the number of hours spent watching TV increases, the number of hours spent doing homework decreases. Note: The sign of the correlation value is negative. The slope of the line of best fit will be **negative**.

Example 4



Interpretation: There appears to be a weak, **positive** linear relationship between Exam 1 and Exam 2 scores.

In other words, as the Exam 1 scores increase, the exam 2 scores increase. However, it does not appear to be a strong relationship.

Example 5 (with data)

Manatees (also called “sea cows”) are sometimes killed by powerboats. The table below gives the number of registered powerboats and the number of powerboat-related manatee deaths in the state of Florida from 1991 through 2000. I used StatHelper to produce a scatterplot.

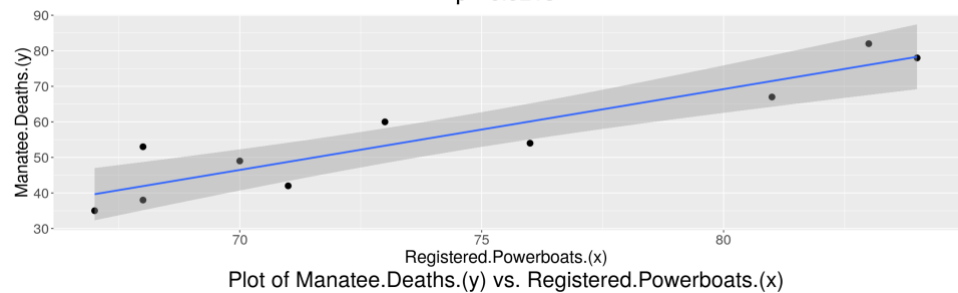
Year	Registered Powerboats (X)	Manatee Deaths (Y)
1991	68	53
1992	68	38
1993	67	35
1994	70	49
1995	71	42
1996	73	60
1997	76	54
1998	81	67
1999	83	82
2000	84	78

Technology results in a correlation value $r = .9215$.

The scatterplot is given as:

$$Y_i = -112.71 + 2.274X_i$$

$$r = 0.9215$$



Interpretation: There appears to be a very strong, **positive** linear relationship between the number of registered powerboats and the number of manatee deaths.

In other words, as the number of registered powerboats increases, the number of manatee deaths also increases. The slope of the line of best fit will be **positive**.

Video discussing scatterplots and correlation from Prof. Coffey

Here is a short video discussing scatterplots and correlation: <https://youtu.be/8jEYsEtVspQ>

Here is the same video with interpreting: <https://youtu.be/umLBa57WjNQ>

Least-Squares Regression Line

Unless a scatterplot has perfect linear correlation, we cannot produce **one** line that will go through all of the data points. We will use technology to produce this 'best-fitting' line. This line of best fit is called the Least-Squares Regression line or the regression model.

The least-squares regression line is the line that makes the vertical distance from the data points to the regression line as small as possible. It's called a "least squares" because the best line of fit is one that **minimizes** the variance. In other words, the best fitting line minimizes the sum of the squares of the error terms (residuals). The line of best fit always goes through the point that is the mean of the x-values and the mean of the y-values (\bar{x}, \bar{y}).

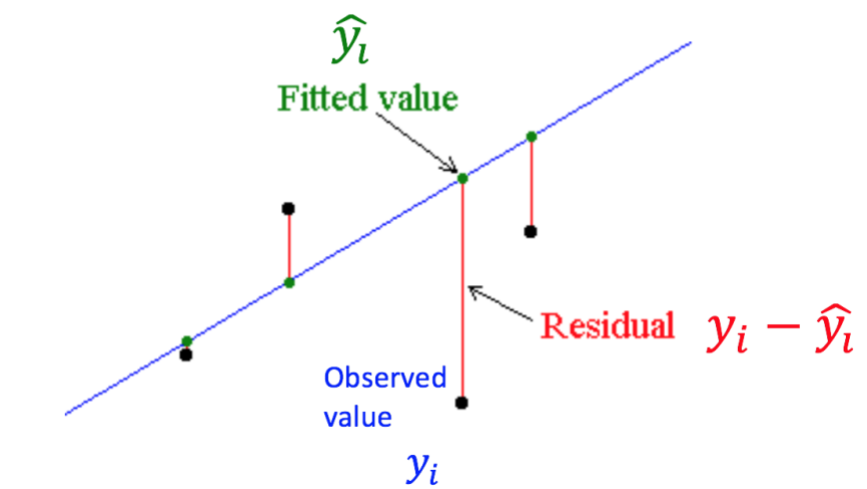
Residuals (or Error Terms)

The difference between the observed value (y) and the predicted value (\hat{y}) is called the residual (e). Another word for residual is 'error'. Each data point has one residual.

Residual = Observed y-value - Predicted y-value

$$e = y - \hat{y}$$

When a least-squares regression line is calculated, both the sum and the mean of the residuals are equal to zero. That is, $\sum e = 0$ and $\bar{e} = 0$.



Observed Value: The observed value is the y-value that you collected. We use the symbol y_i .

Fitted Value: The fitted value is the estimated y-value that you get out when you use the regression equation at a specific x-value. We use the symbol \hat{y} to represent the fitted value and it is pronounced y-hat.

A residual is the difference between these two values.

The least-squares regression model is given by:

$$y_i = \beta_1 x_i + \beta_0 \quad \text{or} \\ y_i = \beta_0 + \beta_1 x_i$$

The estimated regression equation is given by the formula:

$$\hat{y} = b_0 + b_1 x$$

The constant The slope

$$\hat{y} = a + bx$$

The slope b_1 is calculated by:

$$b_1 = \frac{\sum(xy) - \frac{\sum(x)\sum(y)}{n}}{\sum(x^2) - \frac{[\sum(x)]^2}{n}}$$

And the intercept b_0 can be found with:

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

We will use technology to calculate the slope and y-intercept!

Understand and Interpret Slope

The slope of the least-squares regression line represents the average change in the response variable for a one-unit increase in the predictor (x) variable.

If the slope is positive, then there is an increase in the response variable, with a one-unit increase in x.

If the slope is negative, then there is a decrease in the response variable, with a one-unit increase in x.

Think about it...

- If the scatterplot shows an increasing, positive trend, then we would expect the slope to be positive.
- If the scatterplot shows a decreasing, negative trend, then we would expect the slope to be negative.
- If the correlation is positive, then the slope of the line is positive.
- If the correlation is negative, then the slope of the line is negative.

You may use a template sentence to interpret the slope, if needed.

"For each unit increase in (predictor variable), the (response variable) increases/decreases by (slope), on average."

or

"(slope) represents the average change in (response variable) for each unit increase in (predictor variable)."

Understand and Interpret the Constant

The constant is another way of saying the y-intercept. It represents the value of the response variable when the predictor variable is equal to zero. NOTE: It does not always make sense to interpret the y-intercept. Only interpret the y-intercept if having an x-value equal to zero is in the scope of the data or makes sense for the data you are studying.

You may use a template sentence to interpret the y-intercept of the regression equation:

“(y-intercept value) represents the (response variable) when the (predictor variable) is zero.”

Video discussing regression equations and interpreting slope from Prof. Coffey

Here is a short video discussing equations and interpreting slope: https://youtu.be/7_nUkxzVIXg

Here is the same video with interpreting: <https://youtu.be/PMqgMdwaYk4>

Coefficient of Determination (R^2)

The Coefficient of Determination, R^2 , is the proportion of variability (fluctuation) in the response variable that can be attributed to the least-squares regression model. In other words, R^2 is the proportion of variation in the response variable that is being explained by the regression line. **We will use technology to calculate R-squared.**

$$R^2 = 1 - \frac{\sum(\text{residuals})^2}{\sum(y - \bar{y})^2} \qquad R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \qquad R^2 = \frac{\text{explained variation}}{\text{total variation}}$$

In fact, we can say $R^2 = (r)^2$, where r is the correlation between the response and predictor variables. The higher the R^2 , the more useful the model is in making predictions (i.e. it has more explanatory ability). As the explanatory ability of the line decreases, $R^2 = (r)^2$, decreases.

If you are provided a value for R-squared and wonder what the correlation value is, you could take the square root of R-squared to get r . Then, you would have to look at a scatterplot or the sign of the slope to determine if r should be positive or negative. Remember, R-squared is always positive; it was calculated with squared values.

Interpret the Coefficient of Determination, R^2

R^2 gives the proportion of the total variation in the response variable that is being explained by the predictor variable and the regression equation. You may use this template to interpret R^2 :

“ _____% of the variation in (response variable) is being explained by the regression model.”

If R^2 is high, then a high proportion of variation in (response variable) is being explained by the model.

The coefficient of determination can be used to:

- Determine how well the linear equation fits the data. Note: If $R^2 \geq 75\%$, then we would say it fits the data well.
- Determine if we should use the regression line to make predictions. Note: If $R^2 \geq 75\%$, then we could use the regression line to make predictions.
- Determine if the model seems good to you. Note: If $R^2 \geq 75\%$, then we would consider the linear model to be 'good'.
- Determine if the regression model is useful in a practical sense (meaning we would want to use the model for predictions). Note: If $R^2 \geq 75\%$, then we would consider the regression model useful in a practical sense. We would go ahead and use the line to make predictions.

If $R^2 \geq 75\%$, then we consider there to be a reasonable amount of variation in y that is being explained. 75% is a loose threshold; There is wiggle room. If you are happy with a coefficient of determination, R^2 , of 70% then state this. Once you determine if R^2 is high enough, then you can say that you feel the model **does** fit the data well and you **would** use the line to make predictions and it **does** seem good to you and the model **is practically useful**. Keep in mind, you do not have to say all of these things. I am just addressing all the bullet points above.

Video discussing Coefficient of Determination (R^2) from Prof. Coffey

Here is a short video discussing R-squared: <https://youtu.be/2WOHaw6lxac>

Here is the same video with interpreting: https://youtu.be/9_26dgYVNRo

Predict/Estimate with Regression Equation

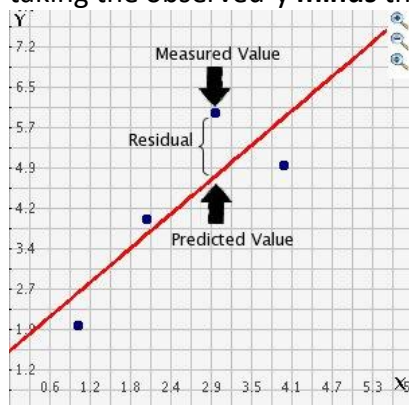
A regression equation can be used to predict the response variable when given a value for the predictor/input variable. To do this by hand, you would get a calculator handy and...

1. Sub in the given x -value into the equation and calculate the fitted y -value (\hat{y}).
2. If you have raw data, make sure this x -value is in the scope of the data you used to generate the linear equation. If it is not, then do not use the prediction/estimate. This is called extrapolation.
3. Interpret this fitted y -value as a prediction for y or an estimate for the average y -value.

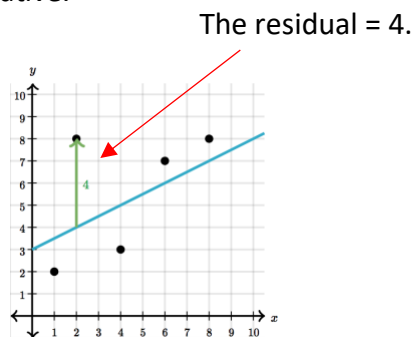
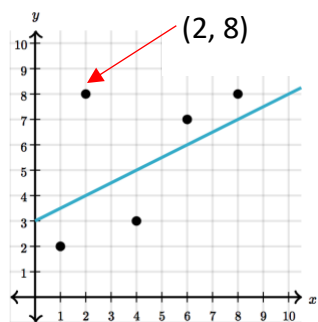
Calculate a Residual

There is error when calculating the least-squares regression line. We call that error a **residual**.

A residual is a measure of how well a line fits an individual data point and the residual can be calculated by taking the observed-y **minus** the predicted-y.



Consider this simple data set with a line of fit drawn through it and notice how point (2,8) is 4 units above the line. This vertical distance is known as a **residual**. For data points above the line, the residual is positive. For data points below the line, the residual is negative.

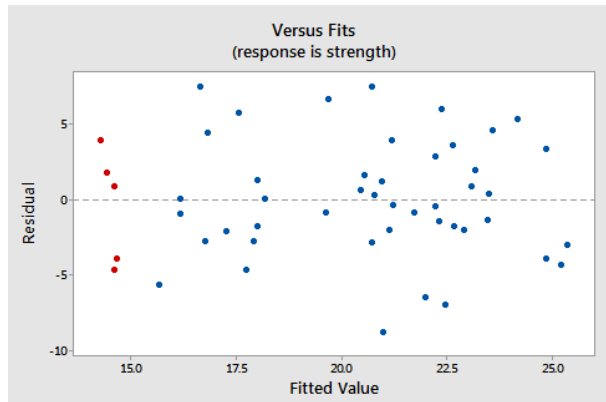
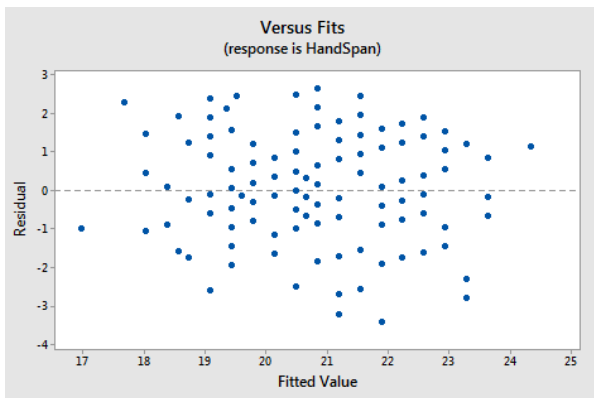


For example, the residual for the point (4,3) is -2. The closer a data point's residual is to 0, the better the fit. In this case, the line fits the point (4,3) better than it fits the point (2,8).

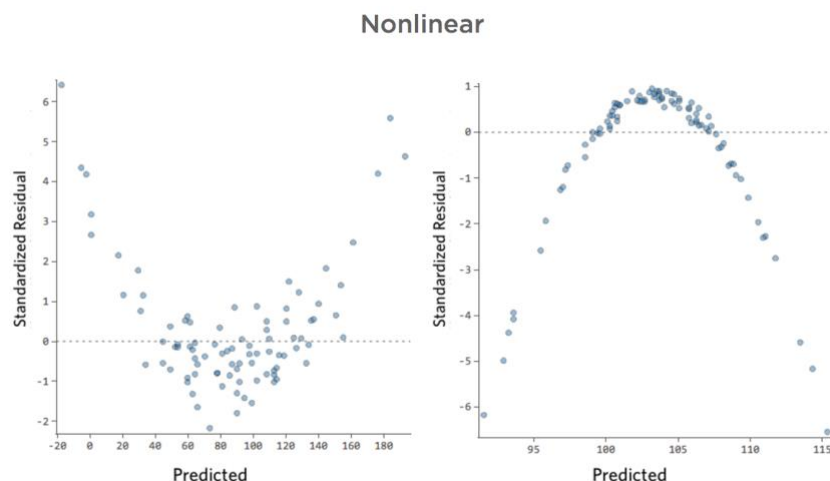
Residual Plots

A residual plot is a graph that shows the residuals on the vertical axis and the predictor variable on the horizontal axis. If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a non-linear model is more appropriate.

Here are two examples of residual plots that show residuals that are **randomly** dispersed (some residuals are above the zero-residual line, some are below; it appears to be a cloud of dots with NO PATTERN. When we see residual plots with no pattern to them, then we know that a line was a good way to model our data. A linear model is valid for these data.



Here are two examples of residual plots that show residuals that are **NOT randomly** dispersed...there is a pattern to the residuals. This indicates that the residuals are not independent of one another and that a line may not be the best model for these data. In both cases, the residuals are forming a quadratic pattern. Our conclusion, based on the residual plots, would be that the data are nonlinear. We could say that a linear model is not a valid model for these data.



Video discussing Predicting, Calculating a Residual and Residual Plots from Prof. Coffey

Here is a short video: https://youtu.be/h8z_17P4dVs

Here is the same video with interpreting: <https://youtu.be/EBYvAKfMNzU>

Use StatHelper for Regression

1. Choose Regression and Correlation from the menu on the left.
2. Press start to upload the .xlsx file.
3. Choose the correct sheet, response (y) variable and input (predictor, x) variable.

The screenshot shows the StatHelper web application interface. On the left is a dark sidebar menu with options: Welcome, Descriptive Statistics, Probability Distributions, Hypothesis Tests, Confidence Intervals, Sample Size, Contingency Tables, ANOVA, and Regression and Correlation (highlighted). The main area is divided into sections: 'Select a file to upload' with a 'Browse...' button and a file named '3_Week 3 STAT 14!' selected; 'Select the Data Sheet' with a dropdown menu showing 'Manatee Deaths'; 'Select the Response Variable' with a dropdown menu showing 'Manatee.Deaths.(y)'; and 'Select the Input Variable' with a dropdown menu showing 'Registered.Powerboats.(x)'. At the bottom is a horizontal tab bar with tabs: Instructions, Description, Example, Data View, Work (active), ANOVA Table, ScatterPlot, and Interpretation. Two blue arrows point from the text below to the 'Work' and 'Interpretation' tabs.

The correlation value is calculated for you under the **Work** tab and under the **Interpretation** tab.

The scatterplot is found under the **ScatterPlot** tab.

The slope and y-intercept are calculated for you in the **Work** tab. Use these two values to produce the least-squares regression equation.

The coefficient of determination is calculated for you in the **Work** tab.

Any predictions will be done outside of StatHelper using a calculator.

Video showing Regression using StatHelper from Prof. Coffey

Here is a short video discussing Regression with StatHelper: <https://youtu.be/iJtzomCoBC0>

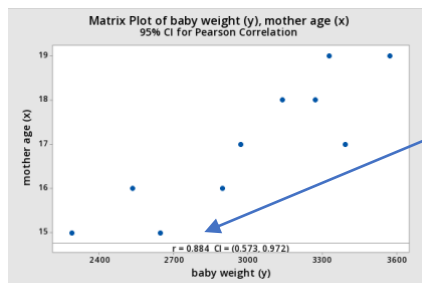
Use Minitab 19 for Regression

Minitab keystrokes for getting Pearson's correlation value:

Stat→basic statistics→correlation

Bring in the response variable and the predictor variable in any order.

The order of the two variables does not matter.



The correlation value can be found here and here.

Correlations

Correlations
baby weight (y)
mother age (x) 0.884

Minitab keystrokes for generating the regression line and coefficient of determination

Stat→regression→regression→Fit regression model

Response: Bring in the y-variable.

Continuous Predictor: Bring in the x-variable.

BABYWEIGHT.MTW

Regression Analysis: baby weight (y) versus mother age (x)

Regression Equation

baby weight (y) = -1163 + 245.1 mother age (x)

least-squares regression line

Coefficients

Term	Coef	SE	Coef	T-Value	P-Value	VIF
Constant	-1163	783	-1.49	0.176		
mother age (x)	245.1	45.9	5.34	0.001	1.00	

If you need more decimal places for the slope and y-intercept, find them here; right click on them to get more decimal places.

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
205.308	78.09%	75.35%	66.77%

The coefficient of determination (R^2)

Note: If you take the square root of R^2 , convert to a decimal, and use the sign (+/-) of your slope, you have the correlation value, r.

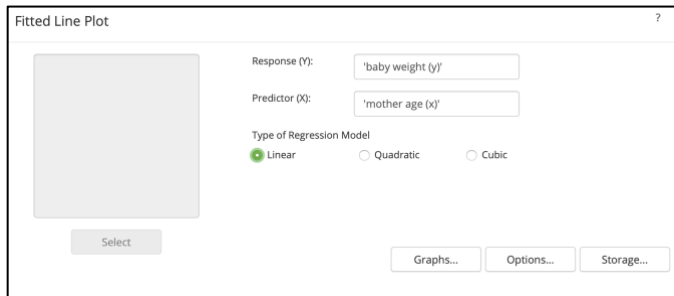
Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	1201970	1201970	28.52	0.001
mother age (x)	1	1201970	1201970	28.52	0.001
Error	8	337212	42152		
Lack-of-Fit	3	78683	26228	0.51	0.694
Pure Error	5	258530	51706		
Total	9	1539183			

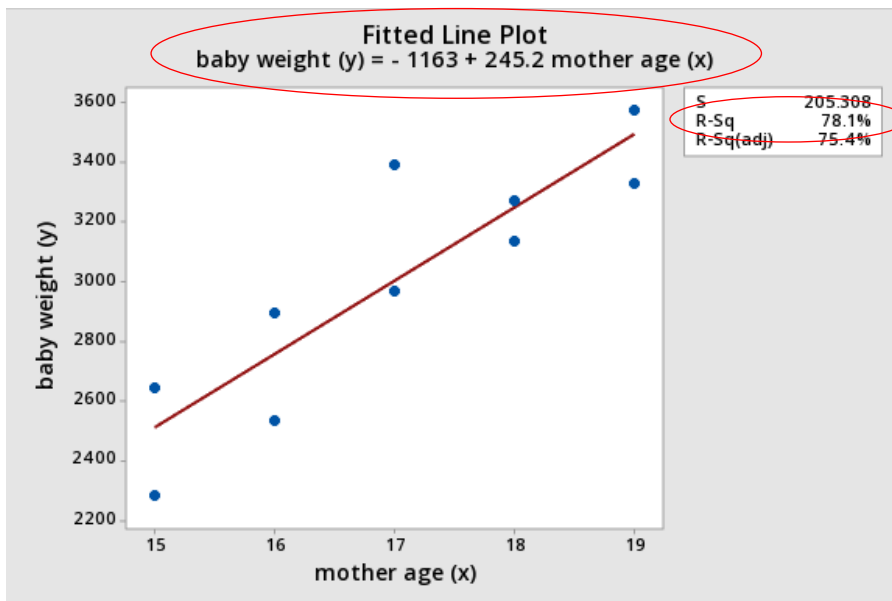
Minitab keystrokes for a Fitted Line Plot (scatterplot with regression line):

Stat→Regression→Fitted Line Plot

Select the appropriate variable for the response (y) and for the predictor (x).



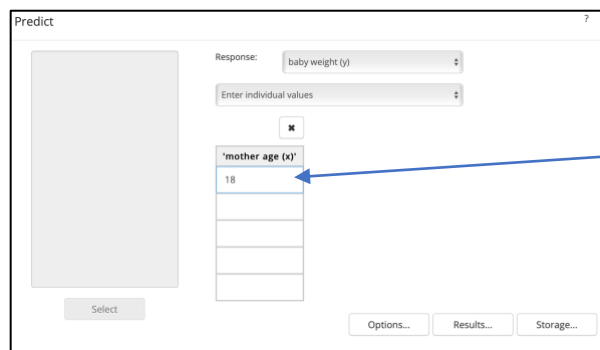
The 'Fitted Line Plot' dialog box in Minitab. It has a 'Response (Y):' field with 'baby weight (y)' entered. The 'Predictor (X):' field has 'mother age (x)' entered. Under 'Type of Regression Model', the 'Linear' radio button is selected. There are 'Select', 'Graphs...', 'Options...', and 'Storage...' buttons.



Minitab keystrokes for making a prediction/estimate:

First, there must be a regression model already built in Minitab. Then...

stat→regression→regression→predict



The 'Predict' dialog box in Minitab. The 'Response:' field has 'baby weight (y)' entered. The 'Enter individual values' field is active. Below it, the 'mother age (x)' variable is listed, and the value '18' is entered in the input field. There are 'Select', 'Options...', 'Results...', and 'Storage...' buttons.

Enter the x-value here.

NOTE: You do not need to use Minitab. You can manually sub the x-value into the regression equation and solve for y.

Prediction for baby weight (y)

Regression Equation

baby weight (y) = -1163 + 245.1 mother age (x)

Settings

Variable	Setting
mother age (x)	18

Prediction

Fit	SE Fit	95% CI	95% PI
3249.25	79.5156	(3065.89, 3432.61)	(2741.54, 3756.96)

The y-value that corresponds to the x-value that you used, can be found here.

Video showing Regression using Minitab 19 from Prof. Coffey

Here is a short video discussing Regression with Minitab 19: <https://youtu.be/RDvL6ZzXe4M>

Example 7

Data were collected on the depth of a dive of penguins and the duration of the dive. The following linear model is a fairly good summary of the data, where t is the duration of the dive in minutes and d is the depth of the dive in yards. The equation for the linear model is $d = 0.015 + 2.915 t$

- A. Interpret the slope.

Answer: For each unit increase in the duration of the dive, the depth of the dive increases by approximately 2.915 yards, on average.

- B. Interpret the y-intercept.

Answer: If the duration of the dive is 0 seconds, then we predict the depth of the dive is 0.015 yards??

NOTE: A dive that lasts $t = 0$ minutes makes no sense in the context of this problem. This is an example of a y-intercept that makes no sense to interpret.

Example 8

When cigarettes are burned, one by-product in the smoke is carbon monoxide. Data is collected to determine whether the carbon monoxide emission can be predicted by the nicotine level of the cigarette. It is determined that the relationship is approximately linear when we predict carbon monoxide, C , from the nicotine level, N . Both variables are measured in milligrams. The formula for the linear model is $C = 3.0 + 10.3 N$

- A. Interpret the slope.

Answer: For each unit increase in the amount of nicotine, the amount of carbon monoxide in the smoke increases by 10.3 mg, on average.

- B. Interpret the y-intercept:

Answer: If the amount of nicotine is zero, then we predict that the amount of carbon monoxide in the smoke will be about 3.0 mg.

Example 9

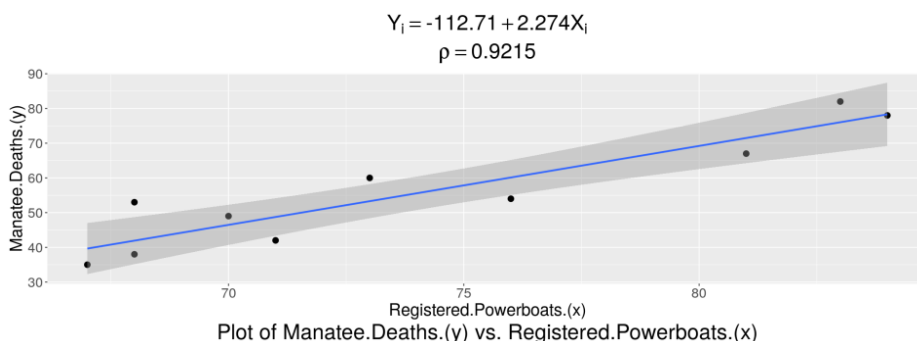
Manatees (also called “sea cows”) are sometimes killed by powerboats. The table below gives the number of registered powerboats and the number of powerboat-related manatee deaths in the state of Florida from 1991 through 2000. The data are found in the file: **3_Week 3 STAT 145 Data.xlsx**, under the sheet labeled **Manatee Deaths**.

Year	Registered Powerboats (X)	Manatee Deaths (Y)
1991	68	53
1992	68	38
1993	67	35
1994	70	49
1995	71	42
1996	73	60
1997	76	54
1998	81	67
1999	83	82
2000	84	78

- A. Using technology, produce a scatterplot, calculate the correlation value and describe/interpret the relationship between the variables.

Answer: Technology results in a correlation value $r = .9215$.

Interpretation: There appears to be a very strong, **positive** linear relationship between the number of registered powerboats and the number of manatee deaths.



- B. Using technology, calculate the least-squares regression line (the regression equation). Interpret the slope and y-intercept in the context of the problem.

Answer: (The equation is given in the output of StatHelper and Minitab): $y = -112.71 + 2.274x$

- C. Using technology, calculate and interpret the coefficient of determination (R-squared) in the context of the problem.

Answer: R-squared = 84.92%

Interpretation: 84.92% of the variability in Manatee Deaths is being explained by the regression line.

The coefficient of determination is:

$$R^2 = (r^2) \times 100$$

$$R^2 = 0.9215^2 \times 100$$

$$R^2 = 0.8492 \times 100$$

$$R^2 = 84.92$$

- D. Using the regression equation, predict the number of manatee deaths when 76 power boats are registered.

Answer: $y = -112.71 + 2.274(76)$

$$y = -112.71 + 172.824$$

$$y = 60.114$$

The regression equation predicts 60.114 manatee deaths when 76 power boats are registered.

- E. Using the regression equation, predict the number of manatee deaths when 89 power boats are registered.

Answer: 89 power boats is not in the scope of our original data. We should not use the equation to make predictions outside the scope of the data (this is called extrapolation).

- F. Calculate the residual (error term) when there are 76 power boats registered.

Answer: We know from part D that the predicted number of manatee deaths is 60.114.

Looking at the raw data, we see that when 76 boats are registered, 54 deaths were observed.

$$\text{Residual} = 60.114 - 54 = 6.114.$$

1997	76	54
------	----	----

Example 10 Multiple Choice Practice

1. A researcher examined the linear relationship between a person's age (X) and the amount of retirement money they have saved (Y) and found $R^2 = 25\%$. What can you conclude?
 - a. It can be concluded that there is a negative association between the two variables
 - b. 75% of the variation in people's retirement income is explained by the linear relationship with age
 - c. 25% of the variation in people's retirement income is explained by the linear relationship with age
 - d. 0.25% of the variation in people's retirement income is explained by the linear relationship with age
2. A restaurant predicts its daily sales (Y) based on the number of customers for that day (X), using the regression equation $\hat{y} = 12x + 5$. What can you conclude from this equation?
 - a. The average daily sales amount is \$12
 - b. An increase of one customer is associated with a \$12 increase in daily sales, on average
 - c. When the number of customers is zero, daily sales is \$12
 - d. Each increase of one customer causes a \$5 increase in daily sales
3. The coefficient of determination is the _____ of the linear correlation coefficient.
 - a. square root
 - b. square
 - c. opposite
 - d. reciprocal
4. An academic advisor uses a random sample of recent graduates to predict starting salary based on GPA (see the information below). Give a practical interpretation of $R^2 = 0.66$.

$$\text{Salary} = -92040 + 228(\text{GPA}); R^2 = 0.66; r = 0.81$$

 - a. 66% of the differences in SALARY are caused by differences in GPA.
 - b. 66% of the variation in SALARY can be explained by the regression line that uses GPA.
 - c. 66% of the variation in GPA can be explained by the regression line that uses SALARY.
 - d. We can predict SALARY correctly 66% of the time using the regression line.
 - e. We estimate SALARY to increase \$.66 for every 1-point increase in GPA.
5. An academic advisor uses a random sample of recent graduates to predict starting salary based on GPA (see the information below). What can we say about the relationship between salary and GPA?

$$\text{Salary} = 228(\text{GPA}) - 92040; R^2 = 0.66; r = 0.81$$

 - a. As salary increases, GPA decreases
 - b. There is a weak linear relationship between GPA and salary
 - c. There is a negative linear relationship between GPA and salary
 - d. As GPA increases, salary increases

ANSWERS:

1. C—this is how you interpret R-squared
2. B—the slope of the line is 12, this is the correct interpretation of slope.
3. B—the relationship between coefficient of determination (R-squared) is that it is literally the correlation value squared.
4. B—sometimes R-squared is given as a decimal; it is always converted and interpreted as a percentage. It represents the proportion of variation in the y-variable that is being explained by the line (that uses the x variable as a predictor).
5. D—the slope of the line is positive and the correlation value is positive and shows a strong linear relationship. The scatterplot would show an increasing trend. As the x-variable increases, the y-variable increases. This is expressed in D.