**Section 3.2 - Least-Squares Regression**

In the previous section we examined scatterplots for linear relationships. Correlation measures the direction and strength of these relationships. When the plot shows a linear relationship, we would like to summarize the overall pattern by drawing a line on the scatterplot. This is called a **Regression Line**. In order to do this we must have an *explanatory* and a *response variable*.

|  |
| --- |
| Definition: A **Regression line** is a line that describes how a response variable y changes as an explanatory variable x changes. We often use a regression line to predict the value of y for a given value of x. |

Consider the example of p. 165.

**1. Interpreting a Regression Line** – A regression line is a *model* for the data, much like the density curves we considered in Chapter 2. It gives a compact mathematical description of the relationship between the response variable y and the explanatory variable x.

|  |
| --- |
| In this equation:  (read y-hat) is the **predicted value** of the response variable y for a given value of the explanatory variable x.b is the slope (rate of change), the amount by which y is *predicted* to change when x increases by one unit.a is the y-intercept, the *predicted* value of y when x = 0.Note: on the AP Exam formula sheet the regression equation is written . (Regardless of notation, the coefficient with x is the slope.) |

**Example:** Suppose there is a strong negative relationship between the miles driven and the advertised price of used Ford F-150 SuperCrew 4 x 4 trucks and that the regression equation is where is in dollars and x is in thousands of miles driven. Identify the slope and y-intercept of the regression line and interpret each in context.

**2. Prediction** – The regression line can be used to predict the response variable for a specific value of the explanatory variable x.

**Example**: Predict the price of a used F-150 with 100,000 miles on it.

|  |
| --- |
| **Definition: Extrapolation** is the use of a regression line for prediction far outside the interval of values of the explanatory variable x used to obtain the line. **Such predictions are often not accurate**. |

 **Team Work:** Complete CYU on p. 168.

**3. Residuals and the Least-Squares Regression Line** – In most cases, no line will pass exactly through all the points in a scatterplot. The predicted values (y-hat) will not be the actual values of the response variable y. *A good regression line makes the vertical distance between the actual points from the line as small as possible.*

|  |
| --- |
| **Definition:** A **residual** is the difference between an observed value of the response variable and the value predicted by the regression line. That isResidual = Observed y – Predicted y = y - |

|  |  |
| --- | --- |
| C:\Users\colin.mayo\AppData\Local\Microsoft\Windows\INetCache\Content.Word\Starnes5e_fig_03_09.jpg | **Example**: Find and interpret the residual for the F-150 that had 70,583 miles driven and a price of $21,994. |

In order to create the **Least-Squares regression line** we will choose the line that makes the *sum of the squared residuals as small as possible*.

|  |  |
| --- | --- |
| C:\Users\colin.mayo\AppData\Local\Microsoft\Windows\INetCache\Content.Word\Starnes5e_fig_03_10a.jpg | C:\Users\colin.mayo\AppData\Local\Microsoft\Windows\INetCache\Content.Word\Starnes5e_fig_03_10b.jpg |

**Technology** – The least-squares regression line can be found by using your graphing calculator. Details are listed on p. 171 of the text as well as p. 32 of NTA.

**Team Work**: Use your calculator to complete CYU on p. 172.

**4. How Well the Line Fits the Data: Residual Plots** - Because the residuals show how far the data fall from our regression line, examining the residuals helps us assess how well the line describes the data. It should be noted that the *mean of the least-squares residuals is always zero*.

A **residual plot** is a scatterplot of the residuals against the explanatory variable, x. They help us assess how well a regression line fits the data.

|  |  |
| --- | --- |
| C:\Users\colin.mayo\AppData\Local\Microsoft\Windows\INetCache\Content.Word\Starnes5e_fig_03_11a.jpg | C:\Users\colin.mayo\AppData\Local\Microsoft\Windows\INetCache\Content.Word\Starnes5e_fig_03_11b.jpg |

**Team Work**: Complete CYU on p. 174.

**Examining Residual Plots**

1. The residual plot should show *no obvious pattern*.

* A curved pattern shows that the relationship is not linear.
* A pattern that gets increasing larger says that the regression line will not be accurate for larger values of x.

2. The residuals shouldbe *relatively small in size*.

* To decide what “small” means, consider the size of the typical error with respect to the data points.

**Technology** - Using the calculator to graph residuals is covered on p. 178 of the text and p. 33 of NTA. To find the standard deviation of the residuals, divide the sum of the squared residuals by n-2 and take the square root.

**Team Work:** Complete CYU on p. 176.

**5. How Well the Line Fits the Data: The Role of r2 in Regression**

**Standard Deviation of the Residuals (s)** - To find out how far off the predictions are using the residuals, we can compute the Standard Deviation of the Residuals:

This value gives us the **approximate size of a “typical” or “average” predicted error (residual).**

**Example**. For the used F-150s, the standard deviation of the residuals is $5740. So when we use the number of miles to estimate the price, we will be off by an average of $5740.

**Coefficient of Determination.** While the standard deviation of the residuals, s, gives us a numerical estimate of the average size of our prediction errors from the regression line, there is another numerical quantity that tells us how well the least squares regression line predicts values of the response variable, y. It is called the **coefficient of determination, r2**.

Suppose we want to estimate the advertised price of a used F-150 from CarMax but do not know the number of miles. The mean price of the other used F-150s would be a reasonable guess.

|  |  |
| --- | --- |
| C:\Users\colin.mayo\AppData\Local\Microsoft\Windows\INetCache\Content.Word\Starnes5e_fig_03_15a.jpg | C:\Users\colin.mayo\AppData\Local\Microsoft\Windows\INetCache\Content.Word\Starnes5e_fig_03_15b.jpg |
| The first scatterplot shows the mean price line.The sum of the squared prediction errors when using the mean price y-bar is called the **total sum of squares (SST)**. | The second scatterplot shows the least-squares regression line.The sum of the squared prediction errors when using the least-squares regression line is called the **sum of squared errors (SSE).** |

We can use the SST and SSE to find the variation in asking price that is *unaccounted* for by the least-squares regression line.

Using this we can find the variability in advertised price that *is* accounted for by the least-squares regression line.

|  |
| --- |
| **Definition**: The **coefficient of determination, r2** is the fraction of the variation in the values of the response variable y that is accounted for by the least squares regression line of y on x. We can calculate **r2** usingWhere and  |

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*
“\_\_\_\_\_ % of the variation in the [response variable name] is accounted for by the regression line.” \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

**6. Example** – pp. 180-181

**6. Interpreting Computer Output** - When looking at computer output, always look for *slope, y-intercept,* and the values of *s* and *r2*.



**Example** - Refer to the example on pp. 181-182 of the text.

**7. Regression to the Mean**

It is possible to calculate the equation of the least-squares regression line using only the means and standard deviations of the two variables and their correlations.

|  |
| --- |
|       |

Example p. 183.

**8. Putting it all together** – Example on pp. 185-186

**9. Correlation and Regression Wisdom**

* *The distinction between explanatory and response variables is important in regression.*
	+ This is not true for correlation. Switching x and y will not affect the value of r.
	+ Switching x and y will give a different regression line.
* *Correlation and regression lines describe only linear relationships.*
	+ You can calculate correlation and the regression line for any relationship between quantitative variables but the results are only useful if the scatterplot shows a linear relationship.
	+ **ALWAYS PLOT YOUR DATA!**
* *Correlation and least-squares regression lines are not resistant.*
	+ One unusual point can change the correlation, r.
	+ Least-squares regression makes the sum of the squares of the vertical distances to the points from the line as small as possible. A point that is extreme in the x direction with no other points near it pulls the line toward itself. This type of point is called ***influential***.

|  |
| --- |
| **Definition**: An **outlier** is an observation that lies outside the overall pattern of the other observations. Points that are outliers in the y direction but not the x direction of a scatterplot have large residuals. Other outliers may not have large residuals.An observation is **influential** for a statistical calculation if removing it would markedly change the result of the calculation. Points that are outliers in the x direction of a scatterplot are often influential for the least-squares regression line. |

* *Association does not imply causation*.
	+ A strong association between two variables is not enough to draw conclusions about cause and effect.
	+ We will learn how to establish causation in Chapter 4.