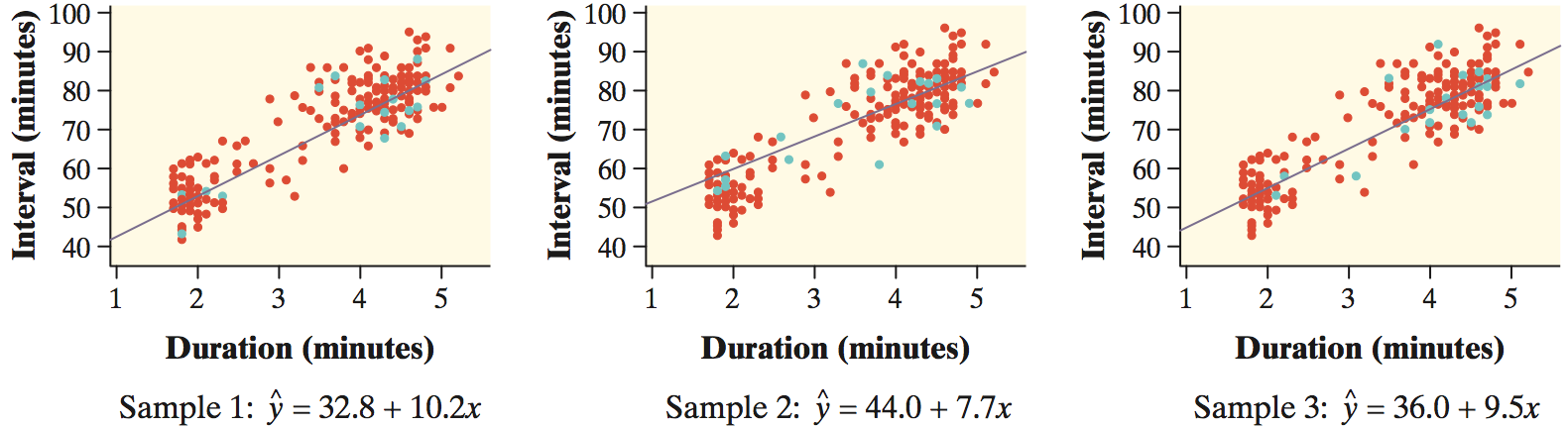
**Section 12.1 - Inference for Linear Regression** (pp. 739-765)

Background - In Chapter 3, we examined data on eruptions of the Old Faithful geyser. Below is a scatterplot of the duration and interval of time until the next eruption for all 222 recorded eruptions in a single month. The least-squares regression line for this population of data has been added to the graph. It has slope 10.36 and y-intercept 33.97. We call this the population regression line (or true regression line) because it uses all the observations that month.

|  |  |
| --- | --- |
|  | Suppose we take an SRS of 20 eruptions from the population and calculate the least-squares regression line for the sample data. How does the slope of the **sample regression line** (also called the *estimated regression line*) relate to the slope of the population regression line? |

**Sampling Distribution of *b***

The figures below show the results of taking three different SRSs of 20 Old Faithful eruptions in this month. Each graph displays the selected points and the LSRL for that sample.  
  


Notice that the slopes of the sample regression lines – 10.2, 7.7, and 9.5 – vary quite a bit from the slope of the population regression line, 10.36.

The pattern of variation in the slope *b* is described by its sampling distribution.

Confidence intervals and significance tests about the slope of the population regression line are based on the sampling distribution of b, the slope of the sample regression line.

|  |  |
| --- | --- |
| Screen shot 2010-12-06 at 8.29.50 AM.png | Fathom software was used to simulate choosing 1000 SRSs of n = 20 from the Old Faithful data, each time calculating the equation of the LSRL for the sample.  The values of the slope *b* for the 1000 sample regression lines are plotted. Describe this approximate sampling distribution of *b*. |

**Shape:**

**Center:**

**Spread:**

**Conditions for Regression Inference**

The slope *b* and intercept *a* of the least-squares line are statistics. That is, we calculate them from the sample data. These statistics would take somewhat different values if we repeated the data production process. To do inference, think of *a* and *b* as estimates of unknown parameters α and β that describe the population of interest.

Suppose we have n observations on an explanatory variable x and a response variable y. Our goal is to study or predict the behavior of y for given values of x.

* **Linear** - The (true) relationship between x and y is linear. For any fixed value of x, the mean response µy falls on the population (true) regression line µy = α + βx. The slope β and intercept α are usually unknown parameters.
* **Independent** - Individual observations are independent of each other.
* **Normal** - For any fixed value of x, the response y varies according to a Normal distribution.
* **Equal** **variance** - The standard deviation of y (call it σ) is the same for all values of x. The common standard deviation σ is usually an unknown parameter.
* **Random** - The data come from a well-designed random sample or randomized experiment.

|  |  |
| --- | --- |
| Screen shot 2010-12-06 at 9.15.04 AM.png | For each possible value of the explanatory variable x, the mean of the responses µ(y | x) moves along this line.  The Normal curves show how y will vary when x is held fixed at different values. All the curves have the same standard deviation σ, so the variability of y is the same for all values of x.  The value of σ determines whether the points fall close to the population regression line (small σ) or are widely scattered (large σ). |

Let’s suppose the conditions for regression are met for the data set, that the population regression line is , and that the spread around the line is given by σ = 6. Let’s focus on the eruption that lasted x = 2 minutes. For this “subpopulation”:

* The average interval of time to the next eruption is
* The intervals of time until the next eruption follow a Normal distribution with
* For about 95% of these eruptions, the interval of time until the next eruption is between

**How to Check Conditions for Regression Inference**• **Linear** - Examine the scatterplot to check that the overall pattern is roughly linear. Look for curved patterns in the residual plot. Check to see that the residuals center on the “residual = 0” line at each x-value in the residual plot.

• **Independent** - Look at how the data were produced. Random sampling and random assignment help ensure the independence of individual observations. If sampling is done without replacement, remember to check that the population is at least 10 times as large as the sample (10% condition).

• **Normal -**  Make a stemplot, histogram, or Normal probability plot of the residuals and check for clear skewness or other major departures from Normality.

• **Equal** **variance** - Look at the scatter of the residuals above and below the “residual = 0” line in the residual plot. The amount of scatter should be roughly the same from the smallest to the largest x-value.

• **Random** - See if the data were produced by random sampling or a randomized experiment.

**Example** - Mrs. Barrett’s class did a variation of the helicopter experiment on page 738. Students randomly assigned 14 helicopters to each of five drop heights: 152 centimeters (cm), 203 cm, 254 cm, 307 cm, and 442 cm. Teams of students released the 70 helicopters in a predetermined random order and measured the flight times in seconds. The class used Minitab to carry out a least-squares regression analysis for these data. A scatterplot, residual plot, histogram, and Normal probability plot of the residuals are shown below.

|  |  |  |
| --- | --- | --- |
| Screen shot 2010-12-06 at 9.27.28 AM.png | Screen shot 2010-12-06 at 9.27.43 AM.png | Screen shot 2010-12-06 at 9.28.00 AM.png |
| Screen shot 2010-12-06 at 9.28.17 AM.png | **Linear** - The scatterplot shows a clear linear form. For each drop height used in the experiment, the residuals are centered on the horizontal line at 0. The residual plot shows a random scatter about the horizontal line. | |

**Independent** - Because the helicopters were released in a random order and no helicopter was used twice, knowing the result of one observation should give no additional information about another observation.

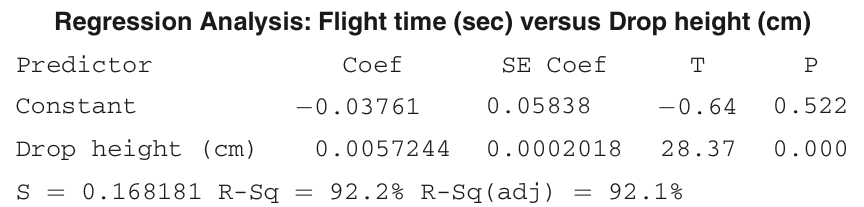
**Normal** - The histogram of the residuals is single-peaked, unimodal, and somewhat bell-shaped. In addition, the Normal probability plot is very close to linear.

**Equal variance** - The residual plot shows a similar amount of scatter about the residual = 0 line for the 152, 203, 254, and 442 cm drop heights. Flight times (and the corresponding residuals) seem to vary more for the helicopters that were dropped from a height of 307 cm.

**Estimating the Parameters** - When the conditions are met, we can perform inference about the regression model . The first step is to estimate the unknown parameters.

* If we calculate the least squares regression line, then
* The remaining parameter is the standard deviation σ, which describes

Because σ is the standard deviation of responses about the population regression line, we estimate it by the standard deviation of the residuals:

**Example**  


**The Sampling Distribution of *b***

We will now return to the Old Faithful example. For all 222 eruptions in a single month, the population regression line for predicting the interval of time until the next eruption *y* from the duration of the previous eruption *x* is *μY = 33.97 + 10.36x* . The standard deviation of responses about this line is σ = 6.159.

|  |  |
| --- | --- |
| Screen shot 2010-12-06 at 10.24.22 AM.png | If we take all possible SRSs of 20 eruptions from the populations, we get the actual sampling distribution of *b*.  **Shape:**  **Center:**  **Spread:** |

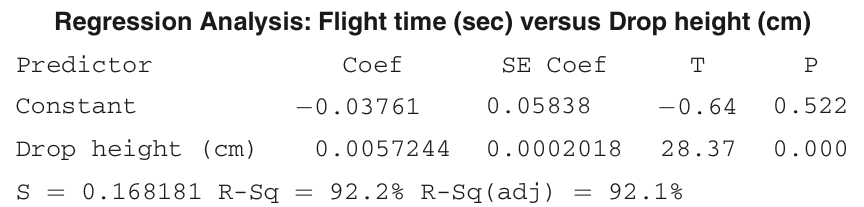
In practice, we do not know σ for the population regression line. We will estimate it with the standard deviation of the residuals, *s*. Then we estimate the spread of the sampling distribution of *b* with the **standard error of the slope:**

If we transform the values of *b* by standardizing them, since the sampling distribution of *b* is Normal, the statistic has the Normal distribution.

Replacing the standard deviation σb of the sampling distribution with its standard error gives

which has a *t* distribution with *n-2* degrees of freedom.

|  |
| --- |
| **Constructing a Confidence Interval for a Least-Squares Regression Line**  When the conditions for regression inference are met, a level C confidence interval for the slope β of the population (true) regression line is  b ± *t\**SEb  In this formula, the standard error of the slope is    and *t\** is the critical value for the *t* distribution with df = n-2 having area between -*t\** and *t\*.* |

**Example**

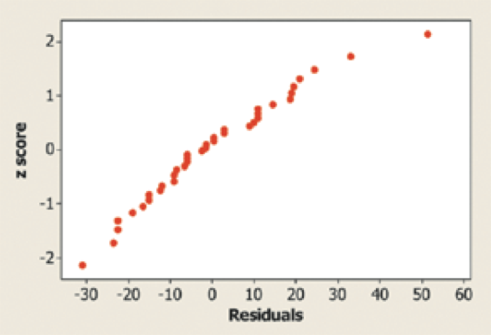
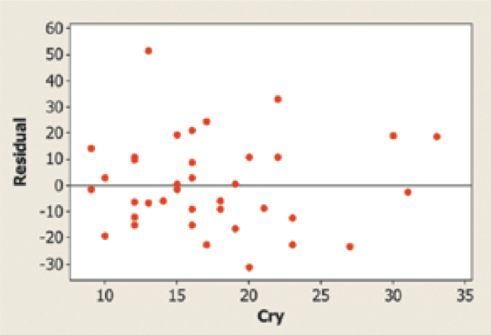
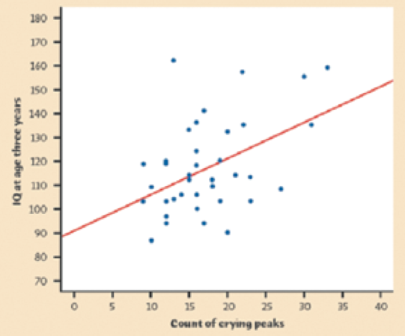
**Application** - Do CYU on pp. 750-751.

**Performing a Significance Test for Slope**

|  |
| --- |
| ***t* Test for the Slope of the Population Regression Line**  Suppose the conditions for inference are met. To test the hypothesis *H0: β = hypothesized value* , compute the test statistic:  Find the *P-*value by calculating the probability of getting a *t*-statistic this large or larger in the direction specified by the alternative hypothesis *Ha*. Use the *t-*distribution with df = n - 2. |

If sample data suggest a linear relationship between two variables, how can we determine whether this happened just by chance or whether there is actually a linear relationship between x and y in the population?

|  |  |
| --- | --- |
| **Example**: | Screen shot 2010-12-06 at 11.19.50 AM.png |



**Confidence Intervals Give More Information**

Let’s construct a 90% confidence interval for the slope β in the crying and IQ example.

**Technology**

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |

HW: Read Sec 12.1; do problems 3, 5, 7, 9, 11, 13, 15 on pp. 759-762.