# Chapter 15

## Inference for Regression

**Objectives:**

- Identify the conditions necessary to do inference for regression.
- Given a set of data, check that the conditions for doing inference for regression are present.
- Explain what is meant by the standard error about the least-squares regression line.
- Compute a confidence interval for the slope of the regression line.
- Conduct a test of the hypothesis that the slope of the regression line is zero (or that the correlation is 0) in the population

### Case Closed: Three pointers in College Basketball

The three-point line was installed in college basketball for the 1986-1987 season and forever changed the nature of the game. The three-point line is a painted arc 19 feet, 9 inches from the center of the basket rim. It is so called because successful shots from beyond the line are awarded one point more than the two points awarded for "field goals" made from closer to the basket. Closer shots have always been higher-percentage shots, and that hasn't changed. But it can be argued that since the three-point line was introduced, there have been more upsets, more near upsets, more dramatic comebacks, and more parity among teams, hence making college basketball a more exciting game to watch.

A look back at the effect of this new feature 20 years later reveals some interesting results. You might expect, for example, that as the average number of three-point shots taken per game increases, the average number of three-point shots made also increases—a positive association. This appears to be borne out by the data:[1]

| Season | Average number taken | Average number made | Percent made |
|---|---|---|---|
| 1986–87 | 9.2 | 3.5 | 38.4 |
| 1987–88 | 10.4 | 4.0 | 38.3 |
| 1988–89 | 11.8 | 4.4 | 37.8 |
| 1989–90 | 12.8 | 4.7 | 36.8 |
| 1990–91 | 13.8 | 5.0 | 36.2 |
| 1991–92 | 14.0 | 5.0 | 35.6 |
| 1992–93 | 14.9 | 5.3 | 35.4 |
| 1993–94 | 16.5 | 5.7 | 34.5 |
| 1994–95 | 17.2 | 5.9 | 34.5 |
| 1995–96 | 17.1 | 5.9 | 34.3 |
| 1996–97 | 17.1 | 5.8 | 34.1 |
| 1997–98 | 17.4 | 6.0 | 34.4 |
| 1998–99 | 17.4 | 5.9 | 34.2 |
| 1999–00 | 17.7 | 6.1 | 34.4 |
| 2000–01 | 17.7 | 6.1 | 34.6 |
| 2001–02 | 18.3 | 6.3 | 34.6 |
| 2002–03 | 18.1 | 6.3 | 34.8 |
| 2003–04 | 18.3 | 6.3 | 34.6 |
| 2004–05 | 18.3 | 6.4 | 34.7 |

But now look at the percent of shots made versus the number of shots taken. This is clearly a negative association. This says that as more three-point shots are taken, the lower the percent of successful shots. Is this a contradiction? Is it possible that knowing the number of three-point shots taken can help us determine the percent of shots made? If so, how confident can we be in making our predictions? After you have studied this chapter, you will be better able to answer interesting questions like these.

Example 1: (Review) Infants who cry easily may be more easily stimulated than others. This may be a sign of higher IQ. Child development researchers explored the relationship between the crying of infants four to ten days old and their later IQ test scores. A snap of a rubber band on the sole of the foot caused the infants to cry. The researchers recorded the crying and measured its intensity by the number of peaks in the most active 20 seconds. They later measured the children's IQ at age three years using the Stanford-Binet IQ test.

L4 = residuals

| Crying (L5) | IQ (L6) | Crying | IQ | Crying | IQ | Crying | IQ |
|---|---|---|---|---|---|---|---|
| 10 | 87 | 20 | 90 | 17 | 94 | 12 | 94 |
| 12 | 97 | 16 | 100 | 19 | 103 | 12 | 103 |
| 9 | 103 | 23 | 103 | 13 | 104 | 14 | 106 |
| 16 | 106 | 27 | 108 | 18 | 109 | 10 | 109 |
| 18 | 109 | 15 | 112 | 18 | 112 | 23 | 113 |
| 15 | 114 | 21 | 114 | 16 | 118 | 9 | 119 |
| 12 | 119 | 12 | 120 | 19 | 120 | 16 | 124 |
| 20 | 132 | 15 | 133 | 22 | 135 | 31 | 135 |
| 16 | 136 | 17 | 141 | 30 | 155 | 22 | 157 |
| 33 | 159 | 13 | 162 | | | | |

a. Make a scatterplot of the data. Use the *Crying* count as the explanatory and the IQ as the response variable. Describe what you see.

Moderate positive linear relationship

r = .455

b. Find the LSRL of the data. How well does the line fit the data?

$\hat{IQ}$ = 1.4929(Crying Count) + 91.27

Residuals: Residuals are scattered so linear model is a good choice

$r^2$ = .207   20.7% of variation in IQ is explained by LSRL on Crying Count so not very well.

## Conditions for the Regression Model

The slope $b$ and intercept $a$ of the least-squares line are statistics. That is, we calculate them from the sample data. In the setting of the infants, *these statistics would take somewhat different values if we repeated the study with different infants.* To do formal inference, we think of $a$ and $b$ as estimates of unknown parameters. The parameters appear in a mathematical model of the process that produces our data.

Conditions for Regression Inference:

We have n observations of an explanatory variable x and a response variable y. Our goal is to predict the values of y for given values of x.

1. Repeated responses of y are independent.
2. The mean response μ has a straight line relationship with x: $\mu_y = \alpha + \beta x$. The slope β and y-intercept α are unknown.
3. The standard deviation of y, σ, is the same for all values if x and is unknown.
4. For any fixed value of x, the response variable y varies according to a Normal distribution.

The heart of this model is that there is an "on the average" straight-line relationship between $y$ and $x$. The ***true regression line*** $\mu_y = \alpha + \beta x$ says that the mean response $\mu_y$ moves along a straight line as the explanatory variable $x$ changes. We can't observe the true regression line. The values of $y$ that we do observe vary about their means according to a Normal distribution. If we hold $x$ fixed and take many observations of $y$, the Normal pattern will eventually appear in a stemplot or histogram. In practice, we observe $y$ for many different values of $x$, so that we see an overall linear pattern formed by points scattered about the true line. The standard deviation $\sigma$ determines whether the points fall close to the true regression line (small $\sigma$) or are widely scattered (large $\sigma$).

## Checking Conditions for Regression

You can fit a least-squares line to any set of explanatory-response data when both variables are quantitative. If the scatterplot doesn't show a roughly linear pattern, the fitted line may be almost useless. But it is still the line that fits the data best in the least-squares sense. To use regression inference, however, the data must satisfy the regression model conditions. Before we do inference, we must check these conditions one by one.

· **The observations are independent.** In particular, repeated observations on the same individual are not allowed. So we can't use ordinary regression to make inferences about the growth of a single child over time, for example.

· **The true relationship is linear.** We can't observe the true regression line, so we will almost never see a perfect straight-line relationship in our data. Look at the scatterplot to check that the overall pattern is roughly linear. A plot of the residuals against $x$ magnifies any unusual pattern. Draw a horizontal line at zero on the residual plot to orient your eye. Because the sum of the residuals is always zero, zero is also the mean of the residuals.

· **The standard deviation of the response about the true line is the same everywhere.** Look at the scatterplot again. The scatter of the data points about the line should be roughly the same over the entire range of the data. A plot of the residuals against $x$, with a horizontal line at zero, makes this easier to check. It is quite common to find that as the response $y$ gets larger, so does the scatter of the points about the fitted line. Rather than remaining fixed, the standard deviation $\sigma$, which measures the variability of responses, may be changing with $x$ as the mean response changes with $x$. You cannot safely use our inference procedures when this happens.

· **The response varies Normally about the true regression line.** We can't observe the true regression line. We can observe the least-squares line and the residuals, which show the variation of the response about the fitted line. The residuals estimate the deviations of the response from the true regression line, so they should follow a Normal distribution. Make a histogram or stemplot of the residuals and check for clear skewness or other major departures from Normality. It turns out that inference for regression is not very sensitive to a minor lack of Normality, especially when we have many observations. Do beware of influential observations, which move the regression line and can greatly affect the results of inference.

Fortunately, it is not hard to check for gross violations of the conditions for regression inference. Checking conditions uses the residuals. Most regression software will calculate and save the residuals for you

Example 2: Check the condition for the infant example.

* Each infant measured once so we have independence
* linear - Scatterplot shows a moderately linear, residuals have no pattern

Sum of the residuals is zero

Residuals
* Scatterplot shows roughly the same scatter

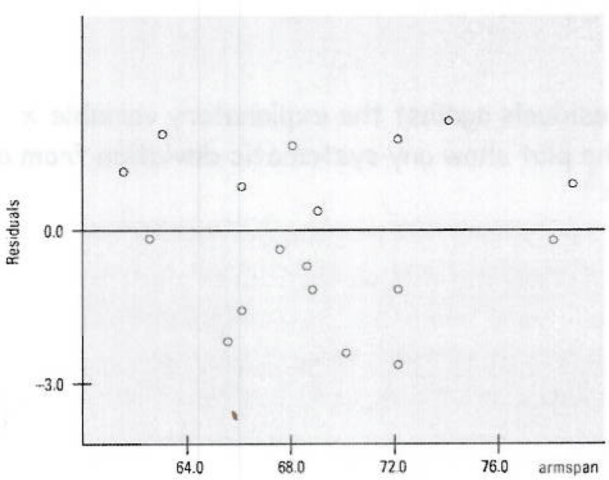
* Slight skew to residuals but not enough to doubt Normality

Example 3: The students in Mr. Shenk's class measured their arm spans and heights (see Activity 15), entered their results into a Minitab worksheet, requested least-squares regression of height on arm span (both in inches), and obtained the following output:

| Predictor | Coef | Stdev | t-ratio | p |
|---|---|---|---|---|
| Constant | 11.547 | 5.600 | 2.06 | 0.056 |
| armspan | 0.84042 | 0.08091 | 10.39 | 0.000 |

s = 1.613    R-sq = 87.1%    R-sq(adj) = 86.3%

A residual plot for the data is shown below.



(a) Determine the equation of the least-squares regression line from the "Coef" column in the printout.

$\hat{height} = .84042(\text{armspan}) + 11.547$

(b) In your opinion, is the least-squares line an appropriate model for the data? Would you be willing to predict a student's height if you knew that his arm span is 76 inches? Explain. Yes - the $r^2$ value is high + the residual plot does not have a pattern. 76 is in data range so we are not extrapolating

Example 4: One of nature's patterns connects the percent of adult birds in a colony that return from the previous year and the number of new adults that join the colony. Here are data for 13 colonies of sparrowhawks:

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Percent return, $x$: | 74 | 66 | 81 | 52 | 73 | 62 | 52 | 45 | 62 | 46 | 60 | 46 | 38 |
| New adults, $y$: | 5 | 6 | 8 | 11 | 12 | 15 | 16 | 17 | 18 | 18 | 19 | 20 | 20 |

This is an example of data that satisfy the conditions for regression inference well. Here are the residuals for the 13 colonies.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Percent return: | 74 | 66 | 81 | 52 | 73 | 62 | 52 |
| Residual: | −4.44 | −5.87 | 0.69 | −5.13 | 2.26 | 1.92 | −0.13 |
| Percent return: | 45 | 62 | 46 | 60 | 46 | 38 | |
| Residual: | −1.25 | 4.92 | 0.05 | 5.31 | 2.05 | −0.38 | |

(a) **Independent observations.** Can we assume that the 13 observations are independent?

Yes they are 13 independent colonies

(b) **Linear relationship.** Make a residual plot. A plot of the residuals against the explanatory variable $x$ magnifies the deviations from the least-squares line. Does the plot show any systematic deviation from a roughly linear pattern? NO- Fairly Scattered

(c) **Spread about the line stays the same.** Does your plot in (b) show any systematic change in spread as $x$ changes? NO - maybe slightly wider in middle but nothing to worry about

(d) **Normal variation about the line.** Make a histogram of the residuals. With only 13 observations, no clear shape emerges. Do strong skewness or outliers suggest lack of Normality?

Normality is OK - No strong outliers or skew

## Estimating the Parameters

The first step in inference is to estimate the unknown parameters $\alpha$, $\beta$, and $\sigma$. When the regression model describes our data and we calculate the least-squares line $\hat{y} = a + bx$, the slope $b$ of the least-squares line is an unbiased estimator of the true slope $\beta$, and the intercept $a$ of the least-squares line is an unbiased estimator of the true intercept $\alpha$.

para. $\alpha$ = true y int of pop — statistic $a$

$\beta$ = true slope of pop — $b$

Example 5: What was the slope and y-intercept of the infant example? Interpret these in context.

$b = 1.4929$

For each increase of one in crying score, the IQ increased by 1.4929

$a = 91.27$

No meaning

The remaining parameter of the model is the standard deviation σ, which describes the variability of the response *y* about the true regression line. The least-squares line estimates the true regression line. So the residuals estimate how much *y* varies about the true line. The residuals are the vertical deviations of the data points from the least-squares line:

$$\text{residual} = \text{observed } y - \text{predicted } y$$
$$= y - \hat{y}$$

There are *n* residuals, one for each data point. Because σ is the standard deviation of responses about the true regression line, we estimate it by *a sample standard deviation of the residuals.* We call this sample standard deviation a *standard error* to emphasize that it is estimated from data. The residuals from a least-squares line always have mean zero. That simplifies their standard error.

Standard Error about the LSRL:

$$s = \sqrt{\frac{\sum \text{residuals}^2}{n-2}} = \sqrt{\frac{\sum (y-\hat{y})^2}{n-2}}$$

Because we use the standard error about the line so often in regression inference, we just call it *s*. Notice that $s^2$ is an average of the squared deviations of the data points from the line, so it qualifies as a variance. We average the squared deviations by dividing by *n* - 2, the number of data points less 2. It turns out that if we know *n* - 2 of the *n* residuals, the other two are determined. That is, *n* - 2 is the ***degrees of freedom*** of *s*. We first met the idea of degrees of freedom in the case of the ordinary sample standard deviation of *n* observations, which has *n* - 1 degrees of freedom. Now we are observing two variables rather than one, and the proper degrees of freedom is *n* - 2 rather than *n* - 1.

Calculating *s* begins with finding the predicted response for each *x* in your data set, then the residuals, and then *s*. In practice you will use technology that does this arithmetic instantly.

<u>Example 6:</u> Find the standard error about the line for the infant example.

L4

$$s = \sqrt{\frac{11023.3888}{38-2}} = \sqrt{306.21} = 17.5$$

# Confidence Intervals for the Regression Slope

The slope β of the true regression line is usually the most important parameter in a regression problem. The slope is the rate of change of the mean response as the explanatory variable increases. We often want to estimate β. The slope *b* of the least-squares line is an unbiased estimator of β. A confidence interval is more useful because it shows how accurate the estimate *b* is likely to be. The confidence interval for β has the familiar form

$$\text{estimate} \pm t^{*}SE_{\text{estimate}}$$

Because *b* is our estimate, the confidence interval becomes

$$b \pm t^{*}SE_b$$

**Confidence Interval for Regression Slope**

A level C confidence interval for the slope β of the true regression line is

$$b \pm t^{*}SE_b$$

In this expression, the standard error of the least-squares slope *b* is

$$SE_b = \frac{s}{\sqrt{\sum (x-\bar{x})^2}}$$

and $t^*$ is the critical value for the $t(n-2)$ density curve with area C between $-t^*$ and $t^*$.

Example 7: Use the computer printout to find the 95% confidence interval for the slope of the line in the infants problem.

```
Regression Analysis
The regression equation is
IQ = 91.3 + 1.49 Crycount

Predictor      Coef      StDev       T       P
Constant     91.268      8.934   10.22   0.000
Crycount     1.4929     0.4870    3.07   0.004

S = 17.50          R-Sq = 20.7%
```

Slope

estimates $\sigma$

$SE_b$

We usually ignore this part.

30 df

**Figure 15.5** ***Minitab regression output for the crying and IQ data, for Example 15.5.***

(.5052, 2.4806)

We are 95% confident the mean IQ raises by .5052 + 2.4806 points for each additional peak in crying

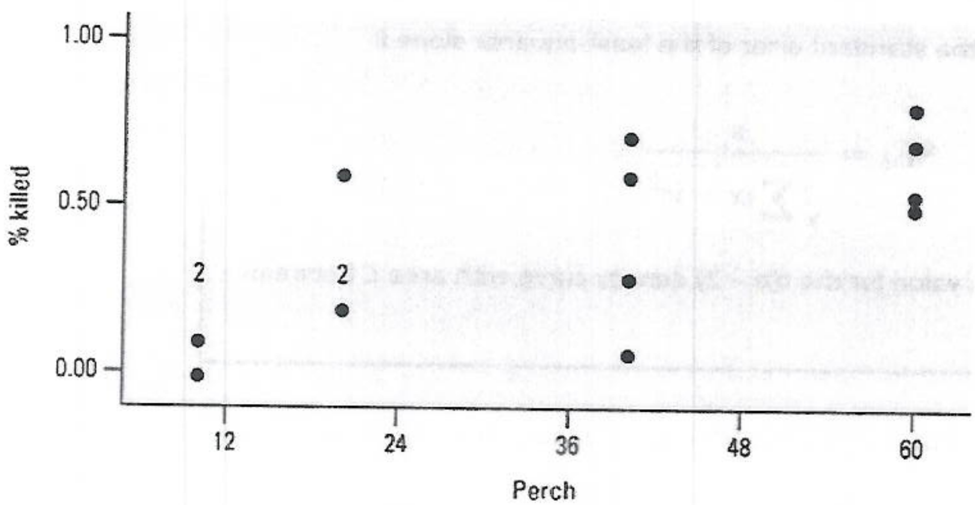
$1.4929 \pm 2.028(.4870)$

$1.4929 \pm .9877$

Example 8: Here is one way in which nature regulates the size of animal populations: high population density attracts predators, who remove a higher proportion of the population than when the density of the prey is low. One study looked at kelp perch and their common predator, the kelp bass. The researcher set up four large circular pens on sandy ocean bottoms off the coast of southern California. He chose young perch at random from a large group and placed 10, 20, 40, and 60 perch in the four pens. Then he dropped the nets protecting the pens, allowing bass to swarm in, and counted the perch left after 2 hours. Here are data on the proportions of perch eaten in four repetitions of this setup:[4]

| Number of perch | Proportion killed | | | |
|---|---|---|---|---|
| 10 | 0.0 | 0.1 | 0.3 | 0.3 |
| 20 | 0.2 | 0.3 | 0.3 | 0.6 |
| 40 | 0.075 | 0.3 | 0.6 | 0.725 |
| 60 | 0.517 | 0.55 | 0.7 | 0.817 |

The explanatory variable is the number of perch (the prey) in a confined area. The response variable is the proportion of perch killed by bass (the predator) in 2 hours when the bass are allowed access to the perch. A scatterplot shows a linear relationship. Minitab output for regression is also shown.




| Predictor | Coef | Stdev | t-ratio | p |
|---|---|---|---|---|
| Constant | 0.12049 | 0.09269 | 1.30 | 0.215 |
| Perch | 0.008569 | 0.002456 | 3.49 | 0.004 |

s = 0.1886    R-sq = 46.5%    R-sq(adj) = 42.7%

(a) What is the equation of the least-squares line for predicting proportion killed from count of perch? What part of this equation shows that more perch does result in a higher proportion being killed by bass? the Slope is positive

$\widehat{\text{percent Killed}} = .008569(\text{perch}) + .12049$

(b) What is the regression standard error $s$?

.1886 estimates $\sigma$

(c) Do the data support the principle that "more prey attract more predators, who drive down the number of prey"? Follow the Data Analysis Toolbox.

Scatterplot shows positive linear relationship

Mean proportion killed: .175, .35, .425, .646 proportions increase as # of perch

yes it supports

(d) Construct and interpret a 95% confidence interval for the slope of the true regression line.

14 df

$.008569 \pm 2.145(.002456)$

$.008569 \pm .00527$

$(.003299, .013839)$

We are 95% confident that the proportion of perch killed increases on average from .003299 to .013839 for each additional perch added to the pen.

Example 9:

Microsoft Excel

| | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | SUMMARY OUTPUT | | | | | | |
| 2 | | | | | | | |
| 3 | *Regression Statistics* | | | | | | |
| 4 | Multiple R | 0.6021 | | | | | |
| 5 | R Square | 0.4652 | | | | | |
| 6 | Adjusted R Square | 0.4270 | | | | | |
| 7 | Standard Error | 0.1886 | | | | | |
| 8 | Observations | 16.0000 | | | | | |
| 9 | | | | | | | |
| 10 | | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
| 11 | Intercept | 0.1205 | 0.0927 | 1.2999 | 0.2146 | -0.0783 | 0.3193 |
| 12 | Perch | 0.0086 | 0.0025 | 3.4899 | 0.0036 | 0.0033 | 0.0138 |
| 13 | | | | | | | |

Sheet5 / Sheet1 / Sheet2 / Sheet2

(a) Excel gives a 95% confidence interval for the slope of the population regression line. What is it? Interpret this interval.

.0033 to .0138

(b) Starting from the Minitab values of the least-squares slope $b$ and its standard error, verify this confidence interval.

What?

(c) Give a 90% confidence interval for the population slope. As usual, this interval is shorter than the 95% interval.

$.0086 \pm 1.7613(.0025)$

$.0086 \pm .0044$

$(.0042, .013)$

## Testing the Hypothesis of No Linear Relationship

The most common hypothesis about the slope is

$$H_0: \beta = 0$$

A regression line with slope 0 is horizontal. That is, the mean of $y$ does not change at all when $x$ changes. So this $H_0$ says that there is *no true linear relationship* between $x$ and $y$. Put another way, $H_0$ says that *straight-line dependence on $x$ is of no value for predicting $y$*. Put yet another way, $H_0$ says that there is *no correlation* between $x$ and $y$ in the population from which we drew our data. You can use the test for zero slope to test the hypothesis of zero correlation between any two quantitative variables. That's a useful trick. **Do notice that testing correlation makes sense only if the observations are a random sample.** That is often not the case in regression settings, where researchers may fix in advance the values of $x$ they want to study.

The test statistic is just the standardized version of the least-squares slope $b$. It is another $t$ statistic. Here are the details.

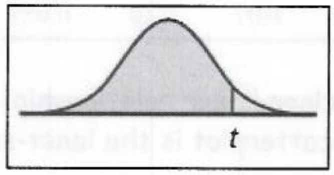
## Significance Tests for Regression Slope

To test the hypothesis $H_0$: $\beta = 0$, compute the ***t* statistic**

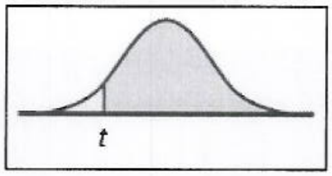
$$t = \frac{b}{\mathrm{SE}_b}$$

In terms of a random variable $T$ having the $t(n-2)$ distribution, the $P$-value for a test of $H_0$ against

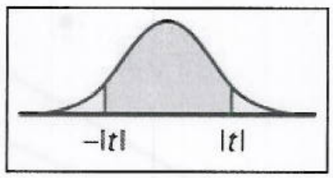
$H_a$: $\beta > 0$ is $P(T \geq t)$




$H_a$: $\beta < 0$ is $P(T \leq t)$




$H_a$: $\beta \neq 0$ is $2P(T \geq |t|)$




This test is also a test of the hypothesis that the correlation is 0 in the population.

Example 10: Interpret the results of the infants' crying example.

$H_o$: $\beta = 0$

$H_a$: $\beta > 0$

$t = 3.07$ Verify by 'formula'

$P = .004$

Very significant

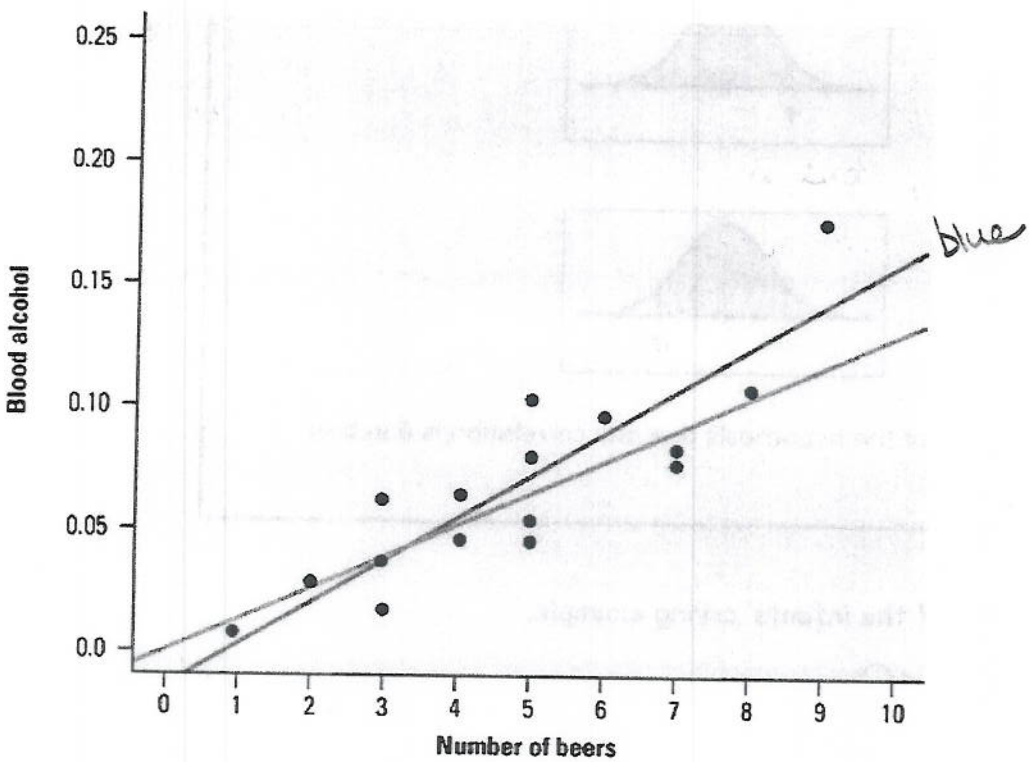
Very strong evidence crying + IQ are correlated

<u>Example 11:</u> An earlier example looked at how well the number of beers a student drinks predicts his or her blood alcohol content (BAC). Sixteen student volunteers at Ohio State University drank a randomly assigned number of cans of beer. Thirty minutes later, a police officer measured their BAC. Here are the data:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Student: | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Beers: | 5 | 2 | 9 | 8 | 3 | 7 | 3 | 5 |
| BAC: | 0.10 | 0.03 | 0.19 | 0.12 | 0.04 | 0.095 | 0.07 | 0.06 |
| Student: | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| Beers: | 3 | 5 | 4 | 6 | 5 | 7 | 1 | 4 |
| BAC: | 0.02 | 0.05 | 0.07 | 0.10 | 0.085 | 0.09 | 0.01 | 0.05 |

The scatterplot in Figure 15.7 shows a clear linear relationship. Figure 15.8 gives part of the Minitab regression output. The blue line on the scatterplot is the least-squares line

$$\hat{y} = -0.0127 + 0.0180x$$




```
The regression equation is
BAC = - 0.0127 + 0.0180 Beers


Predictor        Coef       StDev         T         P
Constant     -0.01270     0.01264     -1.00     0.332
Beers        0.017964    0.002402      7.48     0.000

S = 0.02044      R-Sq = 80.0%
```

Now we're ready to perform formal inference. Once again, we are guided by the Inference Toolbox.

Population: Students
parameter: true β of BAC vs Beers

$H_o: \beta = 0$ $H_a: \beta > 0$

Indepent: each student participated once
Linear Relationship: scatterplot shows linear relationship
Spread: Assume that spread remains same
Normality: Assume

$t = 7.48$ p value = 0

Very sig, reject the null

It appears there is a positive correlation between # of beers consumed + BAC

**Example 12:** Coffee is a leading export from several developing countries. When coffee prices are high, farmers often clear forest to plant more coffee trees. Here are five years' data on prices paid to coffee growers in Indonesia and the percent of forest area lost in a national park that lies in a coffee-producing region:[6]

| | | | | | |
|---|---|---|---|---|---|
| Price (cents per pound) | 29 | 40 | 54 | 55 | 71 |
| Forest lost (percent) | 0.49 | 1.59 | 1.69 | 1.82 | 3.10 |

After entering these data into your calculator, a regression analysis reports the following results:

$$\text{forest lost} = -1.0134 + 0.0552 \times \text{price}$$

$$r^2 = 0.91 \qquad P = 0.0125$$

(a) Explain in words what the slope β of the population regression line would tell us if we knew it. Based on the data, what are the estimates of β and the intercept α of the population regression line?

for each increase of one cent the percent of forest lost increases by .0552.

b= .0552 a= -1.0134

(b) What does $r^2 = 0.91$ add to the information given by the equation of the least-squares line? Interpret $r^2$.

91% of variation in percent of forest lost is explained by the LSRL on price. It is a very strong linear relationship

(c) To what null and alternative hypotheses does the *P*-value refer? What does this *P*-value tell you?

$H_0: \beta = 0$

$H_a: \beta \neq 0$

Very significant result Strong evidence against null. There is strong association between coffee price + percent of forest lost

(d) Calculate the residuals for the five data points. Check that their sum is 0 (up to roundoff error). Use the residuals to estimate the standard deviation σ of percents of forest lost about the means given by the population regression line.

Sum = 0

$S = \sqrt{\frac{\Sigma \text{resids}^2}{n-2}} = \sqrt{\frac{.3215}{3}} = .3274$

(e) Do you think that coffee price will allow good prediction of forest lost? Explain.

Yes, high $r^2$, No ~~sca~~ pattern to residuals

Example 13: In the casting of metal parts, molten metal flows through a "gate" into a die that shapes the part. The gate velocity (the speed at which metal is forced through the gate) plays a critical role in die casting. A firm that casts cylindrical aluminum pistons examined 12 types formed from the same alloy. What is the relationship between the cylinder wall thickness (inches) and the gate velocity (feet per second) chosen by the skilled workers who do the casting? If there is a clear pattern, it can be used to direct new workers or to automate the process. Here are the data:

| Thickness | Velocity | Thickness | Velocity | Thickness | Velocity |
|---|---|---|---|---|---|
| 0.248 | 123.8 | 0.524 | 228.6 | 0.697 | 145.2 |
| 0.359 | 223.9 | 0.552 | 223.8 | 0.752 | 263.1 |
| 0.366 | 180.9 | 0.628 | 326.2 | 0.806 | 302.4 |
| 0.400 | 104.8 | 0.697 | 302.4 | 0.821 | 302.4 |

**CrunchIt!**

Simple Linear Regression

**Simple linear regression results:**
Dependent Variable: veloc
Independent Variable: thick
veloc = 70.436874 + 274.7821 thick
Sample size: 12
R (correlation coefficient) = 0.7019
R-sq = 0.49266976
Estimate of error standard deviation: 56.364124

**Parameter estimates:**

| Parameter | Estimate | Std. Err. | DF | T-Stat | P-Value |
|---|---|---|---|---|---|
| Intercept | 70.436874 | 52.898945 | | | |
| Slope | 274.7821 | 88.17712 | | | |

**Predicted values:**

| X value | Pred. Y | s.e.(Pred.y) | 95% C.I. | 95% P.I. |
|---|---|---|---|---|
| 0.5 | 207.82793 | 17.428537 | (168.99474, 246.66113) | (76.374, 339.28186) |

(a) Make a scatterplot suitable for predicting gate velocity from thickness. Give the value of $r^2$ and the equation of the least-squares line. Draw the line on your plot.

Strong positive Linear relationship

$\hat{\text{Velocity}} = 274.7821(\text{Thickness}) + 70.436874$

$r^2 = .4926$

(b) Based on the information given, test the hypothesis that there is no straight-line relationship between thickness and gate velocity. Give a test statistic, its approximate P-value using a table, and your conclusion.

$H_0: \beta = 0$ $\quad t = 3.1163$ $\quad d.f = 10$

$H_a: \beta \neq 0$ $\quad$ Sig $\quad$ P-Value .0109

There appears to be a Significant Linear Relationship between velocity & gate thickness

Example 14: The rate at which an icicle grows depends on temperature, water flow, and wind. The data below are for an icicle grown in a cold chamber at -11°C with no wind and a water flow of 11.9 milligrams per second:

| Time (min): | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 |
|---|---|---|---|---|---|---|---|---|---|
| Length (cm): | 0.6 | 1.8 | 2.9 | 4.0 | 5.0 | 6.1 | 7.9 | 10.1 | 10.9 |
| Time (min): | 100 | 110 | 120 | 130 | 140 | 150 | 160 | 170 | 180 |
| Length (cm): | 12.7 | 14.4 | 16.6 | 18.1 | 19.9 | 21.0 | 23.4 | 24.7 | 27.8 |

We want to predict length from time.

Minitab

```
Session

Regression Analysis: Length versus Time

The regression equation is
--omitted--

Predictor      Coef       SE Coef         T        P
Constant     -2.3948       0.3963     -6.04    0.000
Time        0.158483     0.003661     43.29    0.000

S = 0.8059   R-Sq = 99.2%        R-Sq(adj) = 99.1%
```

(a) Make a scatterplot suitable for predicting length from time. The pattern is very linear. What is the coefficient of determination $r^2$? Time explains almost all of the change in length.

$r^2 = 99.2$

(b) Use the computer output to estimate the three parameters α, β, and σ.

$\alpha \approx -2.3948$ $\quad \beta = .158483 \quad \sigma = .8059$

(c) What is the equation of the least-squares regression line of length on time? Add this line to your plot.

$\widehat{\text{Length}} = .158483(\text{time}) - 2.3948$