

Chapter 14
Inference for Regression
Section 1
Inference about the Model

Example 14.1: Crying and IQ

Infants who cry easily may be more easily stimulated than others and 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.

TABLE 14.1 Infants' crying and IQ scores

Crying	IQ	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				

Source: Samuel Karelitz et al., "Relation of crying activity in early infancy to speech and intellectual development at age three years," *Child Development*, 35 (1964), pp. 769-777.

Least regression line equation:

$$\hat{y} = a + bx = 91 + 1.493x$$

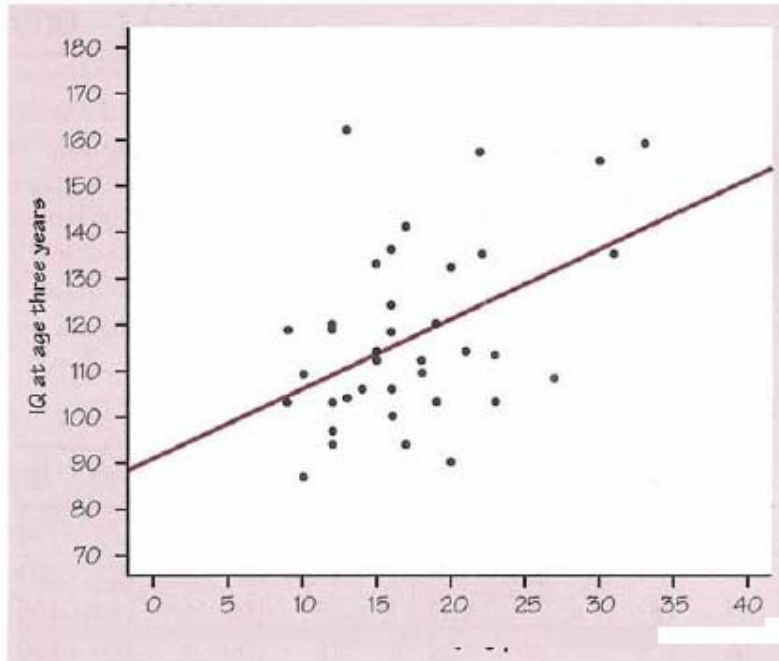
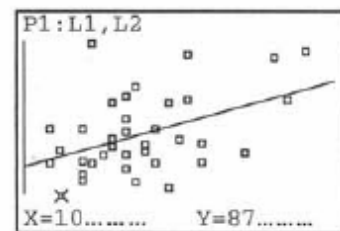
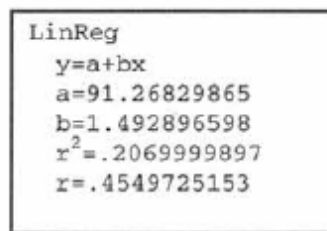
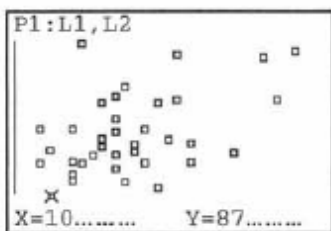


FIGURE 14.1 Scatterplot of the IQ score of infants at age three years against the intensity of their crying soon after birth, for Example 14.1.

The \hat{y} 's or predicted values for the IQ, attempts to estimate the actual values of the population's IQ measured several years later. The predictions usually won't agree exactly with the actual values.



- According to the r^2 , only about 21% of the variation in IQ scores is explained by the crying intensity.
- The prediction will not be very accurate.
- We need to do inference procedures on our statistics to assess the confidence level of our prediction and to test how statistically significant our findings are based on a significance level.
-

The regression Model

Statistics: slope b and a intercept calculated from the *sample*, both unbiased estimators.

What are our statistics, slope, b and y-intercept, a trying to estimate?

Parameters: slope β and α intercept calculated from the *population*.

In other words: The sample regression line estimates the population regression line.

- We could take more samples, calculate more regression lines and would end up with many \hat{y} 's, one per sample.
- Then, take the “average” of all the \hat{y} 's and compute the sampling distribution of all the \hat{y} 's.
- This sampling distribution's *mean* would attempt to estimate the μ_y of the population regression line whose slope and y intercept are β and α respectively.

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 on y , the normal pattern will eventually appear.

Let's say for $x=20$

$$\hat{y}_1 = 91.27 + 1.49(20) = 121.07$$

$$\hat{y}_2 = 83.91 + 1.49(20) = 115.51$$

$$\hat{y}_3 = 114.72 + 1.49(20) = 120.92$$

$$\hat{y}_4 = 94.78 + 1.49(20) = 121.18$$

$$\hat{y}_5 = 86.82 + 1.49(20) = 125.82$$

$$\hat{y}_6 = 120.72 + 1.49(20) = 125.52$$

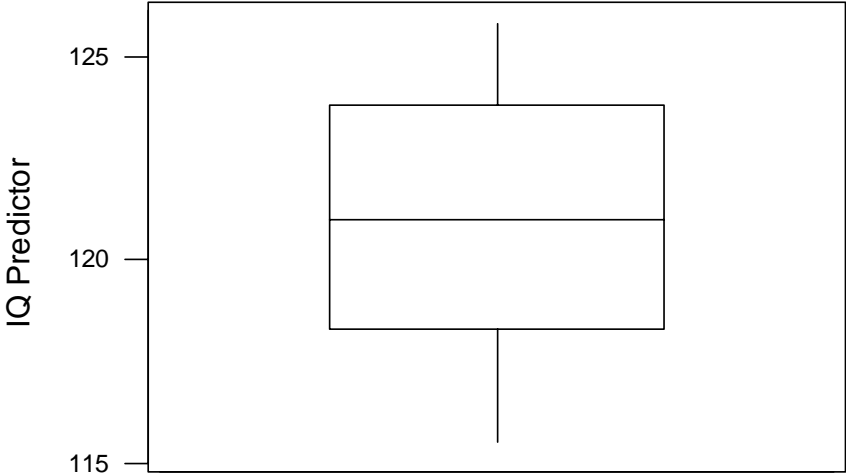
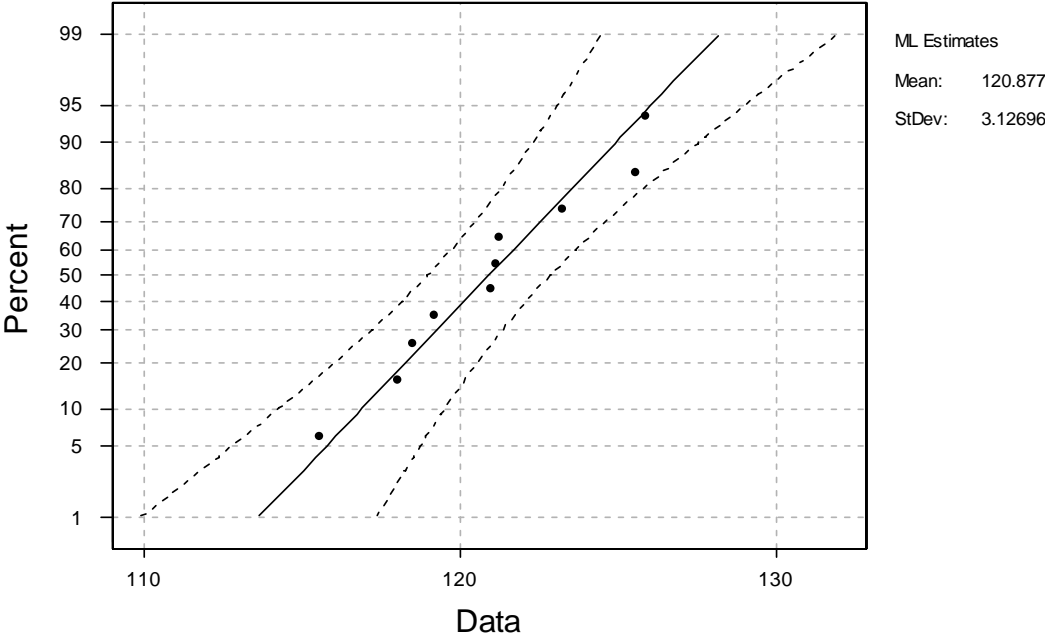
$$\hat{y}_7 = 108.19 + 1.49(20) = 117.99$$

$$\hat{y}_8 = 89.43 + 1.49(20) = 118.43$$

$$\hat{y}_9 = 91.41 + 1.49(20) = 123.21$$

$$\hat{y}_{10} = 115.52 + 1.49(20) = 119.12$$

Normal Probability Plot for IQ Predictor



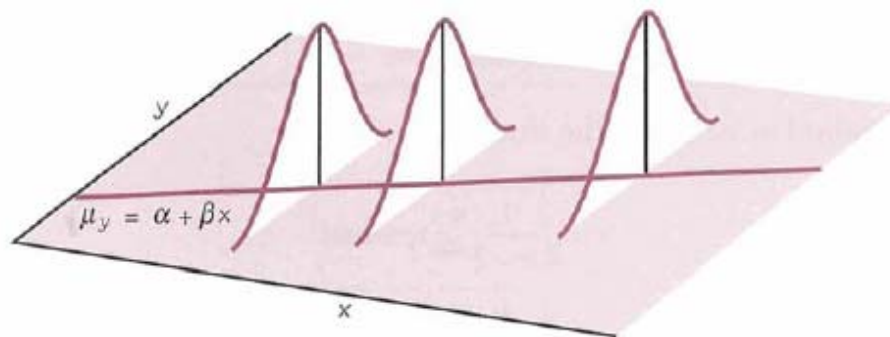
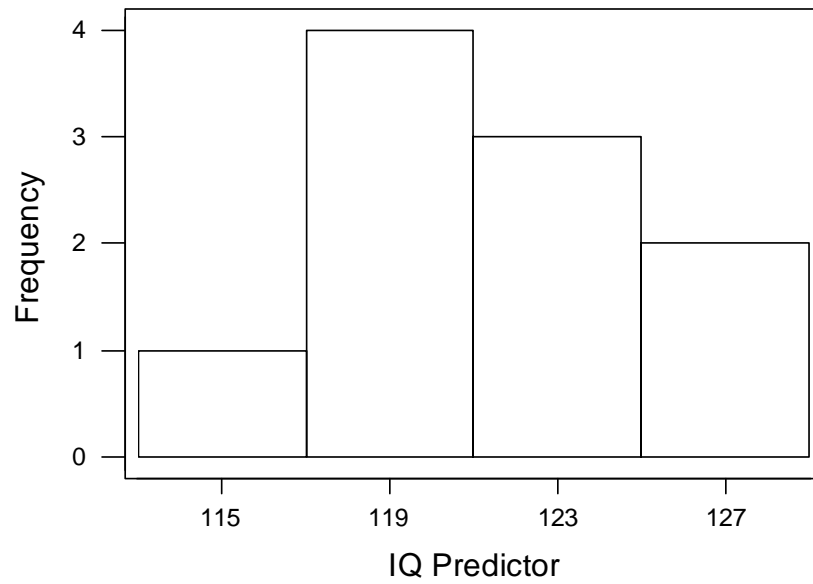


FIGURE 14.2 The regression model. The line is the true regression line, which shows how the mean response μ_y changes as the explanatory variable x changes. For any fixed value of x , the observed response y varies according to a normal distribution having mean μ_y .

CONDITIONS FOR REGRESSION INFERENCE

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 .

- For any fixed value of x , the response y varies according to a normal distribution. Repeated responses y are independent of each other.
- The mean response μ_y has a straight-line relationship with x :

$$\mu_y = \alpha + \beta x$$

The slope β and intercept α are unknown parameters.

- The standard deviation of y (call it σ) is the same for all values of x . The value of σ is unknown.

Inference using the first data sample:

Because $\mathbf{b}=1.493$ estimates the unknown β , we estimate that on the average IQ is about 1.5 points higher for each added crying peak.

We need the intercept $\mathbf{a}=91.27$ to draw the line, but it has no statistical meaning in this example.

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 LSRL estimates the true regression line. So the residuals estimate how much y varies about the true line.

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

Because σ is the standard deviation of responses about the true regression line, we estimate it by a sample standard deviation of the residuals.

STANDARD ERROR ABOUT THE LEAST-SQUARES LINE

The standard error about the line is

$$\begin{aligned} s &= \sqrt{\frac{1}{n-2} \sum \text{residual}^2} \\ &= \sqrt{\frac{1}{n-2} \sum (y - \hat{y})^2} \end{aligned}$$

Use s to estimate the unknown σ in the regression model.

Since we observe two variables rather than one, $n-2$ is the degrees of freedom.

Table 14.1 shows that the first infant studied had 10 crying peaks and a later IQ of 87. The predicted IQ for $x = 10$ is

$$\begin{aligned}\hat{y} &= 91.27 + 1.493x \\ &= 91.27 + 1.493(10) = 106.2\end{aligned}$$

The residual for this observation is

$$\begin{aligned}\text{residual} &= y - \hat{y} \\ &= 87 - 106.2 = -19.2\end{aligned}$$

That is, the observed IQ for this infant lies 19.2 points below the least-squares line on the scatterplot.

Repeat this calculation 37 more times, once for each subject. The 38 residuals are

-19.20	-31.13	-22.65	-15.18	-12.18	-15.15	-16.63	-6.18
-1.70	-22.60	-6.68	-6.17	-9.15	-23.58	-9.14	2.80
-9.14	-1.66	-6.14	-12.60	0.34	-8.62	2.85	14.30
9.82	10.82	0.37	8.85	10.87	19.34	10.89	-2.55
20.85	24.35	18.94	32.89	18.47	51.32		

The variance about the line

$$\begin{aligned}s^2 &= \frac{1}{n-2} \sum \text{residual}^2 \\ &= \frac{1}{38-2} [(-19.20)^2 + (-31.13)^2 + \dots + 51.32^2] \\ &= \frac{1}{36} (11,023.3) = 306.20\end{aligned}$$

and the standard error $s = \sqrt{306.20} = 17.50$

Technology tip: enter your x and y values in list one and two and perform the least-square regression. The calculator updates a list name RESID (2nd/STAT). Store RESID in a list and find the 1-Var Stats.

The $\sum x^2$ is the sum of the squares of the residuals. Divide this number by (n-2) to get s^2 , the estimator of σ .

Confidence Intervals *for the regression slope*

The slope β of the true regression line is usually the most important parameter in a regression problem. A confidence interval is more useful because it shows how accurate the estimate \mathbf{b} is likely to be. The confidence interval for β has the form

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

$$\mathbf{b} \pm t^* \cdot SE_{\mathbf{b}}$$

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 recipe, the standard error of the least-squares slope b is

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

and t^* is the upper $(1 - C)/2$ critical value from the t distribution with $n - 2$ degrees of freedom.

Take a look at the denominator of SE_b on the formula above.

From chapter 1

$$s = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2} = \sqrt{\frac{1}{n-1}} \sqrt{\sum (x_i - \bar{x})^2}$$

Once again, you can use 1-Var Stats, this time on the x -values on list one, to find out $\sqrt{\sum(x - \bar{x})^2}$. Take the S_x and multiply it by $\sqrt{n-1}$.

$$\sqrt{\sum(x - \bar{x})^2} = S_x \times \sqrt{n-1}$$

Regression Output from *minitab*:

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

Predictor	Coef	StDev	T	P
Constant	91.268	8.934	10.22	0.00
Crycount	1.4929	0.4870	3.07	0.004

S=17.50 estimate σ R-sq = 20.7 % SEb

$$\begin{aligned} b \pm t^* \cdot SE_b &= 1.4929 \pm (2.042)(0.4870) \\ &= 1.4929 \pm 0.9944 \\ &= 0.4985 \text{ to } 2.4873 \end{aligned}$$

We are 95 % confident that mean IQ increases by between 0.5 and 2.5 points for each additional peak in crying.

You can find a confidence interval for the intercept α of the true regression line in the same way, using a and SE_a from the

“Constant” line of the printout. We rarely estimate α .

Testing the hypothesis of no linear relationship

$$H_0 : \beta = 0$$

- A regression line with slope 0 is horizontal.
- The mean of y does not change at all when x changes.
- H_0 says that there is no true linear relationship between x and y
- H_0 says that straight-line dependence on x is of no value for predicting y .
- H_0 says that there is no correlation between x and y in the population from which we drew out data.
- You can use the test for zero slope to test the hypothesis of zero correlation

between any two quantitative variables.

- Testing correlation only make sense if the observations are a random sample. That is often not the case in regression settings, where researchers may fix the values of \mathbf{x} they want to study.
- The test statistic is just the standardized version of the least-squares slope \mathbf{b} .

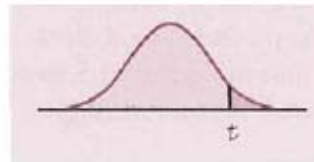
SIGNIFICANCE TESTS FOR REGRESSION SLOPE

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

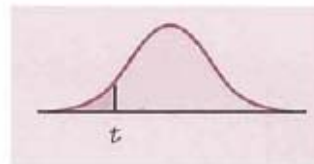
$$t = \frac{b}{SE_b}$$

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

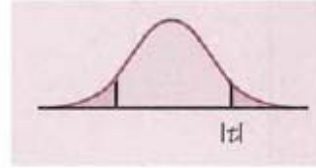
$H_a: \beta > 0$ is $P(T \geq t)$



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



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



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

Warning: Regression output from statistical software usually gives t and its two-sided P-value. For a one-sided test, divide the P-value in the output by 2.

In the example, the hypothesis $H_0 : \beta = 0$ says that crying has no straight –line relationship with IQ. However, the computer output above shows that there is relationship, $t=3.07$ with two-sided P-value 0.004. There is very strong evidence that IQ is correlated with crying.

Section 2

Predictions and Conditions

When predicting the response to a particular value of the explanatory variable, we would like to give a confidence interval that describes how accurate the prediction is. Let's go back to the Crying-IQ example for crying = 20.

- Do we want to predict the mean IQ for all babies whose crying peak is 20?
- or
- Do we want to predict the IQ for one individual baby whose crying peak is 20?

The actual prediction is the same, $\hat{y} = 121.07$. But the margin of error is different for the two kinds of prediction.

- Individual babies whose crying peak is 20 don't all have the same IQ.
- So we need a larger margin of error to pin down one baby's IQ than to estimate the mean IQ for all babies whose crying peak is 20.

Let's identify x^* as the given x value of the explanatory variable. In this case $x^* = 20$.

The only distinction between predicting a single outcome and predicting the mean of all outcomes when $x = x^*$ determines what margin of error is correct.

- To estimate the mean response μ_y , we use a ***confidence interval***.
- To estimate an individual response y , we use a ***prediction interval***.

Meaning of these two intervals

A **95% confidence interval**: as usual, is right 95% of the time in repeated use.

A **95% prediction interval**: is right 95% of the time in repeated use. "repeated use" now means that we take an observation on y for each of the n values of x in the original data, and then take one more observation y with $x = x^*$, then see if it covers the one more y . It will in 95% of all repetitions.

Confidence Intervals for Regression Response

A level C ***confidence interval for the mean response*** μ_y when x takes the value x^* is

$$\hat{y} \pm t^* \times SE_{\hat{\mu}}$$

The standard error $SE_{\hat{\mu}}$ is

$$SE_{\hat{\mu}} = s \times \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x - \bar{x})^2}}$$

The sum runs over all the observations on the explanatory variable x.

t^* is the upper $(1-C)/2$ critical value of the t distribution with $n-2$ degrees of freedom.

Prediction Intervals for Regression Response

A level C *prediction interval for a single observation* on y when x takes the value x^* is

$$\hat{y} \pm t^* \times SE_{\hat{y}}$$

The standard error for prediction $SE_{\hat{y}}$ is

$$SE_{\hat{y}} = s \times \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x - \bar{x})^2}}$$

t^* is the upper $(1-C)/2$ critical value of the t distribution with $n-2$ degrees of freedom.

Let's predict a **prediction interval** for the mean response IQ when $x^* = 20$

Solution:

For a 95% C-level $t^* = 2.042$

For $n=38$ $df = 36$

From output $s = 17.50$

From 1-Var-Stat $\bar{x} = 17.395$ and $S_x = 5.91$

on x variable $S_x^2 = 34.93$

Calculate $\sum (x - \bar{x})^2 = S_x^2 \times (n - 1)$

$$\sum (x - \bar{x})^2 = 34.93 \times 37 = 1,292.41$$

Compute $SE_{\hat{y}}$

$$\begin{aligned} SE_{\hat{y}} &= 17.50 \times \sqrt{1 + \frac{1}{38} + \frac{(2.61)^2}{1,292.41}} \\ &= 17.50 \times 1.0157 = 17.776 \end{aligned}$$

Compute \hat{y}

$$\hat{y} = 91.27 + 1.49(20) = 121.07$$

Put it all together: $\hat{y} \pm t^* \times SE_{\hat{y}}$

$$121.07 \pm 2.042 \times 17.776$$

$$121.07 \pm 36.2986$$

or $(84.7714, 157.3686)$ Quite a range!!

Conclusion: We are 95% confident that when x (crying intensity) = 20, the corresponding value of y (IQ) will be between 84.77 and 157.37.

Checking the regression conditions:

1. The observations are independent.
Repeated observations on the same individual are not allowed.
2. The true relationship is linear.
Look for an overall linear pattern.
3. The standard deviation of the response about the true line is the same everywhere.
The scatter of the data points about the line should be roughly the same over the entire range of the data. Plot the residuals.
4. The response varies normally about the true regression line.
Make a histogram or stemplot of the residuals and check for clear skewness or other major departures from normality. Like other t procedures, inference for regression is (with one exception) not very sensitive to minor lack of normality, especially when we have many observations. The exception is the prediction interval for a single response y .