12 Simple Linear Regression and Correlation

12.3 Inferences About the Slope Parameter β_1

Inferences About the Slope Parameter β_1

In virtually all of our inferential work thus far, the notion of sampling variability has been pervasive.

In particular, properties of sampling distributions of various statistics have been the basis for developing confidence interval formulas and hypothesis-testing methods.

The key idea here is that the value of any quantity calculated from sample data—the value of any statistic—will vary from one sample to another.

Example 10

The following data is representative of that reported in the article “An Experimental Correlation of Oxides of Nitrogen Emissions from Power Boilers Based on Field Data” (*J. of Engr. for Power*, July 1973: 165–170), x = burner-area liberation rate (MBtu/hr-ft^2) and y = NO_x emission rate (ppm).

There are 14 observations, made at the x values 100, 125, 125, 150, 150, 200, 200, 250, 250, 300, 300, 350, 400, and 400, respectively.

Example 10 cont'd

Suppose that the slope and intercept of the true regression line are $\beta_1 = 1.70$ and $\beta_0 = -50$, with $\sigma = 35$ (consistent with the values $\hat{\beta}_1 = 1.7114$, $\hat{\beta}_0 = -45.55$, $s = 36.75$).

We proceeded to generate a sample of random deviations $\tilde{\epsilon}_1, \ldots, \tilde{\epsilon}_{14}$ from a normal distribution with mean 0 and standard deviation 35 and then added $\tilde{\epsilon}_i$ to $\beta_0 + \beta_1 x_i$ obtain 14 corresponding y values.

Regression calculations were then carried out to obtain the estimated slope, intercept, and standard deviation.

Example 10 cont'd

This process was repeated a total of 20 times, resulting in the values given in Table 12.1.

	$\hat{\beta}_1$	$\hat{\beta}_0$	s		$\hat{\beta}_1$	$\hat{\beta}_0$	s
1.	1.7559	−60.62	43.23	11.	1.7843	−67.36	41.80
2.	1.6400	−49.40	30.69	12.	1.5822	−28.64	32.46
3.	1.4699	−4.80	36.26	13.	1.8194	−83.99	40.80
4.	1.6944	−41.95	22.89	14.	1.6469	−32.03	28.11
5.	1.4497	5.80	36.84	15.	1.7712	−52.66	33.04
6.	1.7309	−70.01	39.56	16.	1.7004	−58.06	43.44
7.	1.8890	−95.01	42.37	17.	1.6103	−27.89	25.60
8.	1.6471	−40.30	43.71	18.	1.6396	−24.89	40.78
9.	1.7216	−42.68	23.68	19.	1.7857	−77.31	32.38
10.	1.7058	−63.31	31.58	20.	1.6342	−17.00	30.93

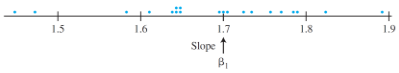
Simulation Results for Example 10

Table 12.1

There is clearly variation in values of the estimated slope and estimated intercept, as well as the estimated standard deviation.

Example 10 cont'd

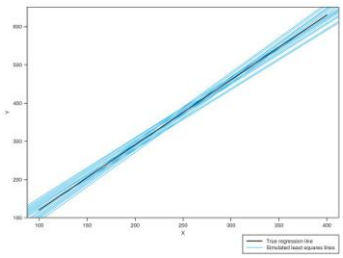
The equation of the least squares line thus varies from one sample to the next. Figure 12.13 shows a dotplot of the estimated slopes as well as graphs of the true regression line and the 20 sample regression lines.



(a) dotplot of estimated slopes
Simulation results from Example 10

Figure 12.13

Example 10 cont'd



(b) graphs of the true regression line and 20 least squares lines (from S-Plus)
Simulation results from Example 10

Figure 12.13

Inferences About the Slope Parameter β_1

The slope β_1 of the population regression line is the true average change in the dependent variable *y* associated with a 1-unit increase in the independent variable *x*.

The slope of the least squares line, $\hat{\beta}_1$, gives a point estimate of β_1. In the same way that a confidence interval for μ and procedures for testing hypotheses about μ were based on properties of the sampling distribution of $\overline{X}$, further inferences about β_1 are based on thinking of $\hat{\beta}_1$ as a statistic and investigating its sampling distribution.

The values of the x_i's are assumed to be chosen before the experiment is performed, so only the Y_i's are random.

9

Inferences About the Slope Parameter β_1

The estimators (statistics, and thus random variables) for β_0 and β_1 are obtained by replacing y_i by Y_i in (12.2) and (12.3):

$$\hat{\beta}_1 = \frac{\Sigma(x_i - \bar{x})(Y_i - \bar{Y})}{\Sigma(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \frac{\Sigma Y_i - \hat{\beta}_1 \Sigma x_i}{n}$$

Similarly, the estimator for σ^2 results from replacing each y_i in the formula for s^2 by the rv Y_i:

$$\hat{\sigma}^2 = S^2 = \frac{\Sigma Y_i^2 - \hat{\beta}_0 \Sigma Y_i - \hat{\beta}_1 \Sigma x_i Y_i}{n - 2}$$

10

Inferences About the Slope Parameter β_1

The denominator of $\hat{\beta}_1$, $S_{xx} = \Sigma(x_i - \bar{x})^2$, depends only on the x_i's and not on the Y_i's, so it is a constant. Then because $\Sigma(x_i - \bar{x})\bar{Y} = \bar{Y}\Sigma(x_i - \bar{x}) = \bar{Y} \cdot 0 = 0$, the slope estimator can be written as

$$\hat{\beta}_1 = \frac{\Sigma(x_i - \bar{x})Y_i}{S_{xx}} = \Sigma c_i Y_i \quad \text{where} \quad c_i = (x_i - \bar{x})/S_{xx}$$

That is, $\hat{\beta}_1$ is a linear function of the independent rv's $Y_1, Y_2, \ldots, Y_n$, each of which is normally distributed.

11

Inferences About the Slope Parameter β_1

Invoking properties of a linear function of random variables as discussed earlier, leads to the following results.

Proposition

1. The mean value of $\hat{\beta}_1$ is $E(\hat{\beta}_1) = \mu_{\hat{\beta}_1} = \beta_1$, so $\hat{\beta}_1$ is an unbiased estimator of β_1 (the distribution of $\hat{\beta}_1$ is always centered at the value of β_1).

2. The variance and standard deviation of β_1 are

$$V(\hat{\beta}_1) = \sigma^2_{\hat{\beta}_1} = \frac{\sigma^2}{S_{xx}} \qquad \sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{S_{xx}}} \qquad \textbf{(12.4)}$$

where $S_{xx} = \Sigma(x_i - \bar{x})^2 = \Sigma x_i^2 - (\Sigma x_i)^2/n$.

12

Inferences About the Slope Parameter β_1

Replacing σ by its estimate s gives an estimate for $\sigma_{\hat{\beta}_1}$ (the estimated standard deviation, i.e., estimated standard error, of $\hat{\beta}_1$):

$$s_{\hat{\beta}_1} = \frac{s}{\sqrt{S_{xx}}}$$

(This estimate can also be denoted by $\hat{\sigma}_{\hat{\beta}_1}$.)

3. The estimator $\hat{\beta}_1$ has a normal distribution (because it is a linear function of independent normal rv's).

13

Inferences About the Slope Parameter β_1

According to (12.4), the variance of $\hat{\beta}_1$ equals the variance σ^2 of the random error term—or, equivalently, of any Y_i, divided by $\sum(x_i - \bar{x})^2$. This denominator is a measure of how spread out the x_i's are about $\bar{x}$.

We conclude that making observations at x_i values that are quite spread out results in a more precise estimator of the slope parameter (smaller variance of $\hat{\beta}_1$), whereas values of x_i all close to one another imply a highly variable estimator.

Of course, if the x_i's are spread out too far, a linear model may not be appropriate throughout the range of observation.

14

Inferences About the Slope Parameter β_1

Many inferential procedures discussed previously were based on standardizing an estimator by first subtracting its mean value and then dividing by its estimated standard deviation.

In particular, test procedures and a CI for the mean μ of a normal population utilized the fact that the standardized variable $(\bar{X} - \mu)/(S/\sqrt{n})$—that is $(\bar{X} - \mu)/S_{\bar{X}}$—had a *t* distribution with $n - 1$ df.

A similar result here provides the key to further inferences concerning β_1.

15

Inferences About the Slope Parameter β_1

Theorem

The assumptions of the simple linear regression model imply that the standardized variable

$$T = \frac{\hat{\beta}_1 - \beta_1}{S/\sqrt{S_{xx}}}$$

$$= \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}}$$

has a *t* distribution with $n - 2$ df.

16

A Confidence Interval for β_1

A Confidence Interval for β_1

As in the derivation of previous CIs, we begin with a probability statement:

$$P\left(-t_{\alpha/2,n-2} < \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} < t_{\alpha/2,n-2}\right) = 1 - \alpha$$

Manipulation of the inequalities inside the parentheses to isolate β_1 and substitution of estimates in place of the estimators gives the CI formula.

A 100(1 – α)% **CI for the slope β_1** of the true regression line is

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2} \cdot s_{\hat{\beta}_1}$$

A Confidence Interval for β_1

This interval has the same general form as did many of our previous intervals.

It is centered at the point estimate of the parameter, and the amount it extends out to each side depends on the desired confidence level (through the *t* critical value) and on the amount of variability in the estimator $\hat{\beta}_1$ (through $s_{\hat{\beta}_1}$, which will tend to be small when there is little variability in the distribution of $\hat{\beta}_1$ and large otherwise).

Example 11

Variations in clay brick masonry weight have implications not only for structural and acoustical design but also for design of heating, ventilating, and air conditioning systems.

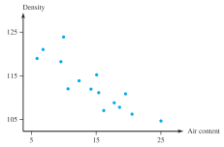
The article “Clay Brick Masonry Weight Variation” (*J. of Architectural Engr.*, 1996: 135–137) gave a scatter plot of *y* = mortar dry density (lb/ft^3) versus *x* = mortar air content (%) for a sample of mortar specimens, from which the following representative data was read:

x	5.7	6.8	9.6	10.0	10.7	12.6	14.4	15.0	15.3
y	119.0	121.3	118.2	124.0	112.3	114.1	112.2	115.1	111.3
x	16.2	17.8	18.7	19.7	20.6	25.0			
y	107.2	108.9	107.8	111.0	106.2	105.0			

Example 11

cont'd

The scatter plot of this data in Figure 12.14 certainly suggests the appropriateness of the simple linear regression model; there appears to be a substantial negative linear relationship between air content and density, one in which density tends to decrease as air content increases.



Scatter plot of the data from Example 11

Figure 12.14

21

Example 11

cont'd

The values of the summary statistics required for calculation of the least squares estimates are

$\Sigma x_i = 218.1$ $\Sigma y_i = 1693.6$ $\Sigma x_i y_i = 24{,}252.54$ $\Sigma x_i^2 = 3577.01$
$\Sigma y_i^2 = 191{,}672.90$

from which $S_{xy} = -372.404$, $S_{xx} = 405.836$, $\hat{\beta}_1 = -.917622$, $\hat{\beta}_0 = 126.248889$, SST = 454.163, SSE = 112.4432, and $r^2 = 1 - 112.4432/454.1693 = .752$.

22

Example 11

cont'd

Roughly 75% of the observed variation in density can be attributed to the simple linear regression model relationship between density and air content. Error df is 15 – 2 = 13, giving $s^2 = 112.4432/13 = 8.6495$ and $s = 2.941$.

The estimated standard deviation of $\hat{\beta}_1$ is

$$s_{\hat{\beta}_1} = \frac{s}{\sqrt{S_{xx}}} = \frac{2.941}{\sqrt{405.836}} = .1460$$

A confidence level of 95% requires $t_{.025,13} = 2.160$. The CI is

$-.918 \pm (2.160)(.1460) = -.918 \pm .315 = (-1.233, -.603)$

23

Example 11

cont'd

With a high degree of confidence, we estimate that an average decrease in density of between .603 lb/ft^3 and 1.233 lb/ft^3 is associated with a 1% increase in air content (at least for air content values between roughly 5% and 25%, corresponding to the x values in our sample).

The interval is reasonably narrow, indicating that the slope of the population line has been precisely estimated.

Notice that the interval includes only negative values, so we can be quite confident of the tendency for density to decrease as air content increases.

24

Example 11

cont'd

Looking at the SAS output of Figure 12.15, we find the value $s_{\hat{\beta}_1}$ of under Parameter Estimates as the second number in the Standard Error column.

Dependent Variable: DENSITY

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob > F
Model	1	341.72606	341.72606	39.508	0.0001
Error	13	112.44327	8.64948		
C Total	14	454.16933			

Root MSE	2.94100	R-square	0.7524
Dep Mean	112.90667	Adj R-sq	0.7334
C.V.	2.60481		

SAS output for the data of Example 11

Figure 12.15

25

Example 11

cont'd

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob >|T|
INTERCEP	1	126.248889	2.25441683	56.001	0.0001
AIRCONT	1	-0.917622	0.14598888	-6.286	0.0001

Obs	Dep Var DENSITY	Predict Value	Residual
1	119.0	121.0	-2.0184
2	121.3	120.0	1.2909
3	118.2	117.4	0.7603
4	124.0	117.1	6.9273
5	112.3	116.4	-4.1303
6	114.1	114.7	-0.5869
7	112.2	113.0	-0.8351
8	115.1	112.5	2.6154
9	111.3	112.2	-0.9093
10	107.2	111.4	-4.1834
11	108.9	109.9	-1.0152
12	107.8	109.1	-1.2894
13	111.0	108.2	2.8283
14	106.2	107.3	-1.1459
15	105.0	103.3	1.6917
Sum of Residuals			0
Sum of Squared Residuals			112.4433
Predicted Resid SS (Press)			146.4144

SAS output for the data of Example 11

Figure 12.15

26

Example 11

cont'd

All of the widely used statistical packages include this estimated standard error in output.

There is also an estimated standard error for the statistic $\hat{\beta}_0$ from which a CI for the intercept β_0 of the population regression line can be calculated.

27

Hypothesis-Testing Procedures

28

Hypothesis-Testing Procedures

As before, the null hypothesis in a test about β_1 will be an equality statement. The null value (value of β_1 claimed true by the null hypothesis) is denoted by β_{10} (read "beta one nought," *not* "beta ten").

The test statistic results from replacing β_1 by the null value β_{10} in the standardized variable T—that is, from standardizing the estimator of β_1 under the assumption that H_0 is true.

The test statistic thus has a t distribution with $n - 2$ df when H_0 is true, so the type I error probability is controlled at the desired level α by using an appropriate t critical value.

29

Hypothesis-Testing Procedures

The most commonly encountered pair of hypotheses about β_1 is H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$. When this null hypothesis is true, $\mu_{Y \cdot x} = \beta_0$ independent of x. Then knowledge of x gives no information about the value of the dependent variable.

A test of these two hypotheses is often referred to as the *model utility test* in simple linear regression. Unless n is quite small, H_0 will be rejected and the utility of the model confirmed precisely when r^2 is reasonably large.

30

Hypothesis-Testing Procedures

The simple linear regression model should not be used for further inferences (estimates of mean value or predictions of future values) unless the model utility test results in rejection of H_0 for a suitably small α.

Null hypothesis: H_0: $\beta_1 = \beta_{10}$

Test statistic value: $t = \dfrac{\hat{\beta}_1 - \beta_{10}}{s_{\hat{\beta}_1}}$

31

Hypothesis-Testing Procedures

Alternative Hypothesis	**Alternative Hypothesis**
H_a: $\beta_1 > \beta_{10}$	$t \geq t_{\alpha,n-2}$
H_a: $\beta_1 < \beta_{10}$	$t \leq -t_{\alpha,n-2}$
H_a: $\beta_1 \neq \beta_{10}$	either $t \geq t_{\alpha/2,n-2}$ or $t \leq -t_{\alpha/2,n-2}$

A P-value based on $n - 2$ can be calculated just as was done previously for t tests.

The **model utility test** is the test of H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$, in which case the test statistic value is the ***t* ratio** $t = \hat{\beta}_1 / s_{\hat{\beta}_1}$.

32

Example 12

Mopeds are very popular in Europe because of cost and ease of operation. However, they can be dangerous if performance characteristics are modified. One of the features commonly manipulated is the maximum speed.

The article "Procedure to Verify the Maximum Speed of Automatic Transmission Mopeds in Periodic Motor Vehicle Inspections" (*J. of Automotive Engr.*, 2008: 1615–1623) included a simple linear regression analysis of the variables x = test track speed (km/h) and y = rolling test speed.

Example 12

cont'd

Here is data read from a graph in the article:

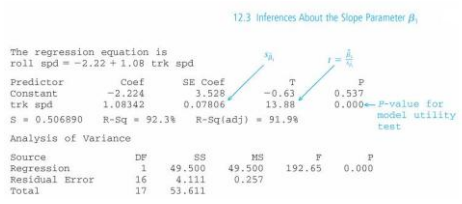
x	42.2	42.6	43.3	43.5	43.7	44.1	44.9	45.3	45.7
y	44	44	44	45	45	46	46	46	47
x	45.7	45.9	46.0	46.2	46.2	46.8	46.8	47.1	47.2
y	48	48	48	47	48	48	49	49	49

A scatter plot of the data shows a substantial linear pattern.

Example 12

cont'd

The Minitab output in Figure 12.16 gives the coefficient of determination as $r^2 = .923$, which certainly portends a useful linear relationship.



Minitab output for the moped data of Example 12

Figure 12.16

Example 12

cont'd

Let's carry out the model utility test at a significance level $\alpha = .01$.

The parameter of interest is β_1, the expected change in rolling track speed associated with a 1 km/h increase in test speed.

The null hypothesis H_0: $\beta_1 = 0$ will be rejected in favor of the alternative H_0: $\beta_1 \neq 0$ if the t ratio $t = \hat{\beta}_1 / s_{\hat{\beta}_1}$ satisfies either $t \geq t_{\alpha/2,n-2} = t_{.005,16} = 2.921$ or $t \leq -2.921$.

Example 12

cont'd

From Figure 12.16, $\hat{\beta}_1 = 1.08342$, $s_{\hat{\beta}_1} = .07806$, and

$$t = \frac{1.08342}{.07806} = 13.88 \text{ (also on output)}$$

12.3 Inferences About the Slope Parameter β_1

The regression equation is
roll spd = −2.22 + 1.08 trk spd

Predictor	Coef	SE Coef	T	P
Constant	−2.224	3.528	−0.63	0.537
trk spd	1.08342	0.07806	13.88	0.000

S = 0.506890 R-Sq = 92.3% R-Sq(adj) = 91.9%

← P-value for model utility test

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	49.500	49.500	192.65	0.000
Residual Error	16	4.111	0.257		
Total	17	53.611			

Minitab output for the moped data of Example 12

Figure 12.16

37

Example 12

cont'd

Clearly this *t* ratio falls well into the upper tail of the two-tailed rejection region, so H_0 is resoundingly rejected.

Alternatively, the *P*-value is twice the area captured under the 16 df *t* curve to the right of 13.88. Minitab gives *P*-value = .000.

Thus the null hypothesis of no useful linear relationship can be rejected at any reasonable significance level. This confirms the utility of the model, and gives us license to calculate various estimates and predictions.

38

Regression and ANOVA

39

Regression and ANOVA

The decomposition of the total sum of squares $\sum(y_i - \bar{y})^2$ into a part SSE, which measures unexplained variation, and a part SSR, which measures variation explained by the linear relationship, is strongly reminiscent of one-way ANOVA.

40

Regression and ANOVA

In fact, the null hypothesis H_0: $\beta_1 = 0$ can be tested against H_a: $\beta_1 \neq 0$ by constructing an ANOVA table (Table 12.2) and rejecting H_0 if $f \geq F_{\alpha,1,n-2}$.

Source of Variation	df	Sum of Squares	Mean Square	f
Regression	1	SSR	SSR	$\frac{\text{SSR}}{\text{SSE}/(n-2)}$
Error	$n-2$	SSE	$s^2 = \frac{\text{SSE}}{n-2}$	
Total	$n-1$	SST		

ANOVA Table for Simple Linear Regression

Table 12.2

Regression and ANOVA

The F test gives exactly the same result as the model utility t test because $t^2 = f$ and $t^2_{\alpha/2,n-2} = F_{\alpha,1,n-2}$. Virtually all computer packages that have regression options include such an ANOVA table in the output.

For example, Figure 12.15 shows SAS output for the mortar data of Example 11.

Dependent Variable: DENSITY

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob > F
Model	1	341.72606	341.72606	39.508	0.0001
Error	13	112.44327	8.64948		
C Total	14	454.16933			

Root MSE	2.94100	R-square	0.7524
Dep Mean	112.90667	Adj R-sq	0.7334
C.V.	2.60481		

SAS output for the data of Example 11

Figure 12.15

Regression and ANOVA cont'd

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob >|T|
INTERCEP	1	126.248889	2.25441683	56.001	0.0001
AIRCONT	1	−0.917622	0.14598888	−6.286	0.0001

Obs	Dep Var DENSITY	Predict Value	Residual
1	119.0	121.0	−2.0184
2	121.3	120.0	1.2909
3	118.2	117.4	0.7603
4	124.0	117.1	6.9273
5	112.3	116.4	−4.1303
6	114.1	114.7	−0.5869
7	112.2	113.0	−0.8351
8	115.1	112.5	2.6154
9	111.3	112.2	−0.9093
10	107.2	111.4	−4.1834
11	108.9	109.9	−1.0152
12	107.8	109.1	−1.2894
13	111.0	108.2	2.8283
14	106.2	107.3	−1.1459
15	105.0	103.3	1.6917
Sum of Residuals			0
Sum of Squared Residuals			112.4433
Predicted Resid SS (Press)			146.4144

SAS output for the data of Example 11

Figure 12.15

Regression and ANOVA

The ANOVA table at the top of the output has $f = 39.508$ with a P-value of .0001 for the model utility test.

The table of parameter estimates gives $t = -6.286$, again with $P = .0001$ and $(-6.286)^2 = 39.51$.