

## Review of Linear Regression

In Chapter 2, we used linear regression to describe linear relationships. The setting for this is a bivariate data set (i.e., a list of cases/subjects for which two variables have been measured for each case) with both variables being quantitative. We use graphs (scatterplots and residual plots), using the horizontal axis for our explanatory ( $x$ ) variable and the vertical axis for the response ( $y$ ) variable) to help us discern whether there is, in fact, a linear relationship  $y = b_0 + b_1x$  between the variables. The correlation  $r$  provides a quantitative (objective) measure of whether there is such a relationship, with  $r$  close to ( $\pm 1$ ) indicating a strong linear relationship, and  $r \approx 0$  indicating little to know linear relationship. While we can draw scatterplots and calculate the sample statistics  $b_0$ ,  $b_1$  and  $r$  by hand, one rarely does so, instead relying on a software package to take care of such tedium.

Now we want to think of the regression line we computed from our sample ( $y = b_0 + b_1x$ ) as an estimate of the true regression line in the population ( $y = \beta_0 + \beta_1x$ ), just as  $\bar{x}$  in univariate quantitative data has served as an estimate of the true (population) mean  $\mu$ . So we will learn how to compute CI and significance tests for  $\beta_0$  and  $\beta_1$ . We will also learn to compute confidence intervals for predictions we make.

**Example:** A tax consultant studied the current relation between the selling price and assessed valuation of one-family dwellings in a large tax district. He obtained a random sample of recent sales of one-family dwellings and, after plotting the data, found a regression line of:

$$(\text{selling price in } \$1000\text{s}) = 4.98 + 2.40 (\text{assessed valuation in } \$1000).$$

At the same time, a second tax consultant obtained a second random sample of recent sales of one-family dwellings, and after plotting the data, found a regression line of:

$$(\text{selling price in } \$1000\text{s}) = 1.89 + 2.63 (\text{assessed valuation in } \$1000).$$

Both consultants attempted to model the true linear relationship they believed to exist between price and valuation. The regression lines they found were sample estimates of the true (but unknown) population relationship between selling price ( $y$ ) and assessed valuation ( $x$ ).

## The Simple Linear Model and Related Assumptions

We write the true linear relationship between the sampled variables  $x$  and  $y$  as  $y = \beta_0 + \beta_1x + \epsilon$ , where  $\epsilon$  is a random error component.

Assumptions for the simple linear model:

1. At each fixed explanatory ( $x$ ) value, the (population) distribution of the responses ( $y$ -values) is normal, with mean  $\mu = \beta_0 + \beta_1 x$  and standard deviation  $\sigma$ .

Note: The mean generally changes for different  $x$  values, while the standard deviation is the same for at every  $x$ .

2. If we denote by  $(x_i, y_i)$  the individual observed data pairs, and write  $\hat{y}_i$  for the predicted value of the response at  $x = x_i$ , then the residuals

$$e_i = y_i - \hat{y}_i = y_i - (\beta_0 + \beta_1 x_i)$$

are independent—that is, the size and sign of a particular residual are not influenced by the size and sign of the preceding residuals.

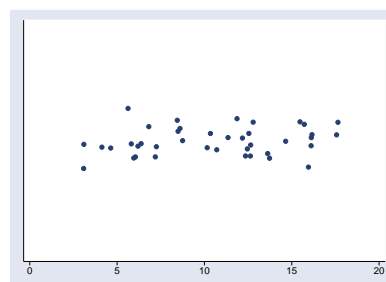
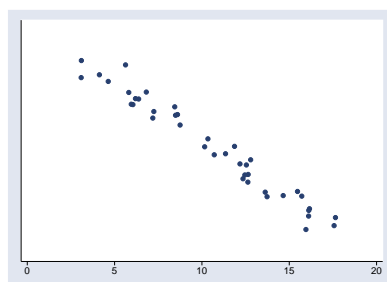
Software output for linear regression includes the values of  $b_0$  and  $b_1$  which are used to estimate  $\beta_0$  and  $\beta_1$ , respectively. Since  $b_0$  and  $b_1$  are statistics (i.e., computed from the sampled coordinate pairs  $(x_i, y_i)$ ), they exhibit the kind of random fluctuation from one sample to the next that is to be expected from all random variables. In the real-estate example above, the first tax consultant would estimate  $\beta_0$  and  $\beta_1$  to be 4.98 and 2.4 respectively, while the second would estimate them to be 1.89 and 2.63.

Other quantities provided by software output for linear regression include the value of  $r^2$  (or  $R^2$ ), which is the square of the correlation, and tells us how much of the observed variation (in the  $y$ -values) is explained by the regression line, and  $s$  (called the *root mean square error*), which serves as an estimate for  $\sigma$ . From the value of  $r^2$  we may determine the sample correlation  $r$  by taking the square root (choosing the negative/positive sign based on whether the line falls or rises from left to right), which itself is sample statistic estimating the value of the true correlation  $\rho$ .

## Testing Model Utility

The reason for doing regression is generally to make predictions. So, one should ask in which situations the resulting estimated regression line  $y = b_0 + b_1 x$  is useful. Put another way, given that there is a linear relationship between  $x$  and  $y$ , in what situations would it be valid to say that knowledge of the  $x$ -value is not helpful in predicting a corresponding  $y$ ?

Choose one:



Answer: When the value of  $\beta_1$  (or, equivalently, of  $\rho$ ) is zero.

It is entirely possible, however, that  $\beta_1 = 0$  while  $b_1$  (our estimate for  $\beta_1$ ) is nonzero due just to random fluctuations between samples. Thus, we require an inference procedure. The hypotheses for the model utility test (which is a test of significance) are

$$H_0 : \beta_1 = 0, \quad H_a : \beta_1 \neq 0.$$

To carry out this test of significance, one would only need to know the sampling distribution for  $b_1$ . Under the assumptions for simple linear regression stated above,  $b_1 \sim N(\beta_1, \sigma_{b_1})$ , where

$$\sigma_{b_1} := \sqrt{\frac{\sigma}{\sum(x_i - \bar{x})^2}}.$$

As before, our inability to know the value of  $\sigma$  requires us to use  $SE_{b_1}$  in its place:

$$SE_{b_1} := \sqrt{\frac{s}{\sum(x_i - \bar{x})^2}},$$

and this leads to a test statistic

$$t = \frac{b_1}{SE_{b_1}},$$

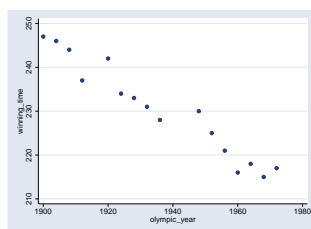
having  $(n - 2)$  degrees of freedom (since *two* parameters  $\beta_0$  and  $\beta_1$  have been estimated to get this far). One may look up a  $P$ -value in Table D, much as with other  $t$ -tests, comparing it to a specified significance level  $\alpha$ . Alternatively, one might construct a  $100 \times (1 - \alpha)\%$  confidence interval for  $\beta_1$ :

$$b_1 \pm (\text{critical } t\text{-value}) \cdot (SE_{b_1}),$$

and draw the same conclusion as for the test of significance by observing whether zero is in this confidence interval. Here the critical  $t$ -value is determined using  $(n - 2)$  degrees of freedom.

If you're interested in the  $P$ -value associated with model utility, or if you want the 95% confidence interval for either  $\beta_0$  or  $\beta_1$ , these are included in *Stata* output for linear regression. And, even if you want a different level CI, you don't have to compute  $SE_{b_1}$ , because once again, it is included next to  $b_1$  in the column marked "Std. Err."

**Example:** Is the winning time (in seconds) for the men's 1500-meter run in the Olympics decreasing significantly over time? Below is a scatterplot of the winning time vs. year including those years between 1900 and 1972 when there were summer olympics, as well as the result from regression analysis.



Source	SS	df	MS			
Model	1654.64697	1	1654.64697	Number of obs	=	16
Residual	98.2170282	14	7.01550201	F( 1, 14)	=	235.86
Total	1752.864	15	116.8576	Prob > F	=	0.0000
				R-squared	=	0.9441
				Adj R-squared	=	0.9401
				Root MSE	=	2.6481

winning_time	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
olympic_year	-.43772	.0285019	-15.36	0.000	-.4988504 - .3765896
_cons	1077.663	55.19781	19.52	0.000	959.2759 1196.051

What is the result of the linear regression on this data?

We get the line  $\hat{y} = 1077.663 - 0.438x$ .

Is this a useful model? (Hypothesis Test)

Yes. The  $P$ -value for the model utility test is essentially 0.

Construct a 95% confidence interval for  $\beta_0$ .

There is no need. It is provided as (959.276, 1196.051).

Construct a 99% CI for  $\beta_1$ .

It is  $-0.438 \pm (2.977)(0.0285)$ , or  $(-0.523, -0.353)$ .

## Using the Model to Estimate and Predict

While the regression line is generally used for estimation/prediction, always remember that extrapolation (prediction at  $x$ -values beyond the scope of your data) is always more tenuous than interpolation (prediction at  $x$ -values within the range of your data). What winning time would our linear model predict for the 1500 m in the year 300 B.C.?

**Example:** A tax consultant studied the current relation between the selling price and assessed valuation of single-family dwellings in a large tax district. He obtained a random sample of recent sales of such homes and, after plotting the data, found a regression line of

(selling price in \$1000's) =  $2.58 + 2.64$  (assessed valuation in \$1000's)    or     $\hat{y} = 2.58 + 2.64x$ .

- (a) Estimate the average selling price of single-family residential homes in this district which have an assessed valuation of \$30,000.

\$81,780

- (b) Predict the selling price of the house at 1993 Calvin Ave., a single-family residential home in this district, which has an assessed valuation of \$30,000.

\$81,780

- (c) Which number, if either, do you think may have the greater error (greater variability, wider confidence interval)?

The prediction of selling price ( $y$ ) for a single home at  $x = 30$  will have more variability than the average selling price at  $x = 30$ .

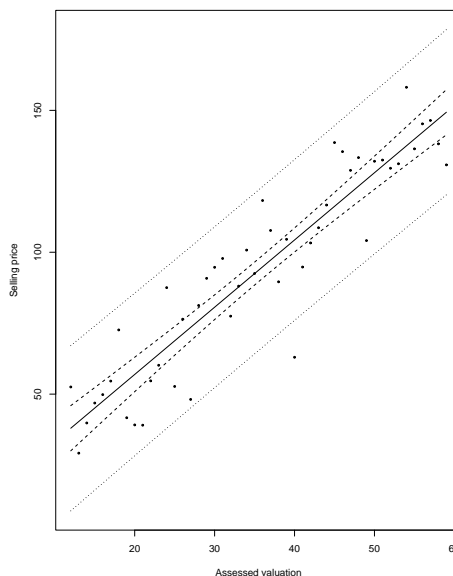
**Level C Confidence Interval for  $\mu_{x_0}$**

$$\hat{y} \pm (t \text{ crit. val.}) \cdot s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

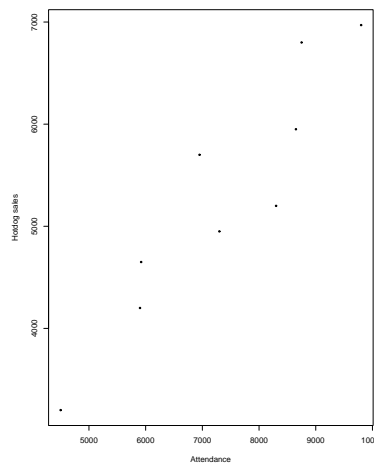
**Level C Prediction Interval for  $\mu_{x_0}$**

$$\hat{y} \pm (t \text{ crit. val.}) \cdot s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

Confidence intervals are at their narrowest when the chosen  $x_0$  is the same as  $\bar{x}$ .



**Example:** A scatterplot of the daily attendance and the number of hotdog sales for a sample of 9 games of a minor league baseball team suggests that the relationship between attendance and sales may be linear. The regression output (from *Minitab*, a different piece of statistical software) is as follows:



The regression equation is  
 hotdog sales = 386 + .671 Attendance

Predictor	Coef	SE Coef	T	P
Constant	386.1	728.7	0.53	0.613
Attendance	0.67095	0.09667	6.94	0.000

S = 467.4      R-Sq = 87.3%      R-Sq(adj) = 85.5%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	10523992	10523992	48.18	0.000
Residual Error	7	1529106	218444		
Total	8	12053098			

Predicted Values for New Observations

New Obs	Fit	SE Fit	95.0% CI	95.0% PI
1	5083	160	( 4705, 5461)	( 3914, 6251)

Values of Predictors for New Observations

New Obs	Attendance
1	7000