

LINEAR REGRESSION

Simple Linear Regression

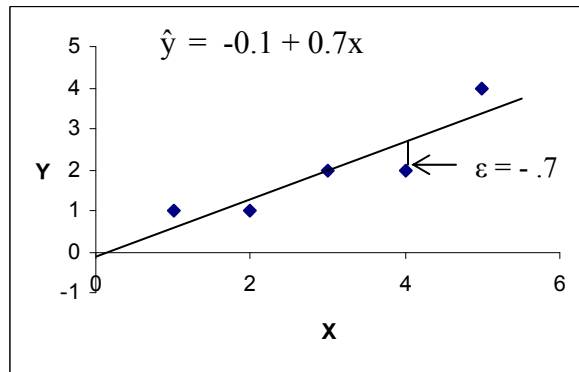
Probabilistic models include both a deterministic element and a random error element – individual data values may fall around a regression model line in a random pattern and not always directly on the line.

Straight line model: $y = \beta_0 + \beta_1x + \varepsilon$ $\beta_0 =$ y-intercept of the line; $\beta_1 =$ slope
 $\varepsilon =$ random error

For the data:

X	Y
1	1
2	1
3	2
4	2
5	4

The x-y scatterplot with a model fitted line looks like \longrightarrow



The y-intercept of the line is -0.1 and
 The slope is 0.7 .

The model line *predicts* or *estimates* values of y ; predicted values are indicated by \hat{y} .
 The *vertical* distances from the actual data points to the line are the errors of prediction.
 For example, for an x value of 4, the model predicts a y value of $-0.1 + 0.7(4) = 2.7$.
 The actual y value from the table is 2, and the error is (observed – predicted) = $(2 - 2.7) = -.7$.

The **least squares regression line** method minimizes the sum of the *squares* of the errors of prediction. There may be many ways to get the sum of the errors to equal zero, since there are both positive and negative errors, but there is only one line that minimizes the sum of the squares of the errors.

In the **least squares line**: a) the sum of the errors (SE) is zero; b) the sum of squared errors (SSE) is less than any other straight-line model. The point (\bar{x}, \bar{y}) is always on the least squares line.

Formulas

$$\text{Slope } \hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} \qquad \text{y-intercept } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{\sum y_i}{n} - \hat{\beta}_1 \frac{\sum x_i}{n}$$

$$SS_{xy} = \sum_{i=1}^n x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} \qquad SS_{xx} = \sum_{i=1}^n x_i^2 - \frac{(\sum x_i)^2}{n}$$

$$= \sum (x_i - \bar{x})(y_i - \bar{y}) \qquad = \sum (x_i - \bar{x})^2$$

Estimates of y should only be made over *the observed range of x* – the model may not be accurate outside of that range. Note that the equation predicting y from x cannot be used to predict x from y. A new regression line must be calculated for that.

Estimated standard error

The variance σ^2 of the random error ϵ produces errors in the model parameters and the prediction value \hat{y} . Deviations from predicted values are also called residuals.

The sum of squares of the deviations (errors) $SSE = SS_{yy} - \hat{\beta}_1 SS_{xy} = \sum (y_i - \hat{y})^2$

$$SS_{yy} = \sum_{i=1}^n y_i^2 - \frac{(\sum y_i)^2}{n} = \sum (y_i - \bar{y})^2$$

$$s^2 = \frac{SSE}{n-2} = \text{Mean Square for Error (MSE)}$$

$s = \sqrt{s^2}$ = estimated standard error (SE) of the regression model \hat{y} values

Since ϵ is assumed to be normally distributed, 95% of observed y values will lie within 2s of their predicted values.

The slope β_1

The slope tells us how many units the y variable is expected to change for one unit of change in the x variable. If the x and y values are unrelated, then the slope of the regression line would be zero – changes in x would be useless in predicting changes in y.

We can estimate the standard deviation of $\hat{\beta}_1$ by:

$$s_{\hat{\beta}_1} = \frac{s}{\sqrt{SS_{xx}}} = \text{the estimated standard error of } \hat{\beta}_1$$

To test the hypothesis that $\beta_1 = 0$, use a **t-test** with **n-2** degrees of freedom.

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 > 0 \quad \text{or} \quad \beta_1 < 0 \quad \text{or} \quad \beta_1 \neq 0$$

Rejection region: $t > t_\alpha$ or $t < -t_\alpha$ or $|t| > t_{\alpha/2}$

$$\text{Test } t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$$

The alternative hypotheses depend on the nature of the proposed relationship between x and y . A positive linear relationship means the slope is positive (> 0); y would *increase* as x *increases*. A negative linear relationship means the slope is negative (< 0); y would *decrease* as x *increases*. For $\beta_1 \neq 0$, the direction of the slope of the line is not necessarily important, just that it is not zero (that is, some kind of relationship exists).

The p-value (observed significance level) given in software output is the two-tailed p-value. If doing a one-tail test, divide this p-value by 2 to get the one-tail p-value, but *only if* the test t is appropriate (> 0 or < 0) for the tail of the alternative hypothesis.

Otherwise: if H_a is that $\beta_1 > 0$ and $t < 0$, then $p = 1 - p/2$
 if H_a is that $\beta_1 < 0$ and $t > 0$, then $p = 1 - p/2$

The confidence interval for β_1 is: $\hat{\beta}_1 \pm t_{\alpha/2} s_{\hat{\beta}_1}$

The endpoint values for this interval are also listed on computer printouts.

The correlation coefficient r

The correlation coefficient r measures the *strength* of the linear relationship between x and y . It is positive if the slope is positive and negative if the slope is negative, but ranges only from -1 (perfect negative relationship) to $+1$ (perfect positive relationship). An r value near 0 indicates little or no *linear* relationship, but there could be a strong nonlinear relationship. Also, *a high correlation does not mean there's a causal relationship*.

$$\text{sample } r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}} \quad (\rho = \text{population correlation coefficient})$$

Coefficient of determination r^2

The coefficient of determination r^2 is the *proportion* of the total sample variability around \bar{y} that is *explained* by the linear relationship; it ranges from 0 to 1.

An r^2 of .70 means that 70% of the variability of y about the mean \bar{y} is explained by the linear relationship between y and x ; or, the total variability of y about their predicted values is reduced by 70% by using the model \hat{y} to predict y instead of the mean.

It can be computed by squaring the correlation coefficient r , or by the following:

$$r^2 = \frac{SS_{yy} - SSE}{SS_{yy}} = \frac{\sum (y_i - \bar{y})^2 - \sum (y_i - \hat{y})^2}{\sum (y_i - \bar{y})^2} = \frac{\sum (\hat{y} - \bar{y})^2}{\sum (y - \bar{y})^2}$$

$$= \frac{\text{explained variation}}{\text{total variation}}$$

SS_{yy} is the sum of the squared deviations about the mean, while SSE is the sum of the squared deviations about the \hat{y} values predicted by the regression equation. If there were no difference in these sums, then the regression line would be the same as the horizontal mean line, the numerator would equal zero, and r^2 would be 0.

The test t statistic can also be computed using r and r^2 : $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$

Estimation and prediction

The derived least squares line is an estimation of the true mean y values for particular values of x . The equation for the true mean $E(y)$ is: $y = \beta_0 + \beta_1x$ (without the 'hats'). The regression equation can be used to *estimate* a mean y value for a particular x or used to *predict* a new y value for a particular x . The values are the same, but the errors are different.

An estimation of the mean value of y involves the error $\hat{y} - E(y)$. A prediction of a y value includes two errors – the error of estimating the mean and the random error ϵ . Confidence intervals are around $E(y)$ estimations; prediction intervals are around predicted values. The prediction interval is always wider than the confidence interval.

Confidence interval around \hat{y} as the mean value of y for an x at a certain point (x_p):

$$\hat{y} \pm t_{\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

Prediction interval around \hat{y} as the predicted new value of y for x_p :

$$\hat{y} \pm t_{\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

s = standard error from the computer program
 $t_{\alpha/2}$ is based on $(n - 2)$ degrees of freedom

These intervals are not straight lines paralleling the least squares line – the widths depend on x_p . The errors of both estimation and prediction will be smallest when $x_p = \bar{x}$ ($x_p - \bar{x} = 0$), and get bigger as you move away from the mean.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.9036961
R Square	0.8166667
Adjusted R Square	0.7555556
Standard Error	0.6055301

correlation coefficient r
 coefficient of determination - proportion of explained variation (explained/total) = 4.9/6

SE of predicted y values = \sqrt{MSE}

Observations	5
--------------	---

ANOVA					
	df	SS	MS	F	Significance F
Regression	1	4.9	4.9	13.36364	0.035352847
Residual	3	1.1	0.3667		
Total	4	6			

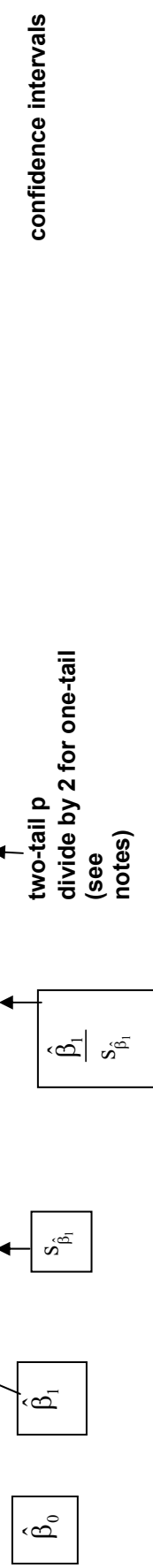
Regression SS = SS_{yy} - SSE = $\sum (\hat{y} - \bar{y})^2$
 (explained variability)

Residual SS = SSE = $\sum (y_i - \hat{y})^2$
 (unexplained variability)

Total SS = SS_{yy} = $\sum (y_i - \bar{y})^2$

Residual MS = MSE or s^2

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-0.1	0.635085296	-0.157459164	0.884884	-2.121124854	1.921124854	-2.121124854	1.921124854
X Variable 1	0.7	0.191485422	3.655630775	0.035353	0.090607928	1.309392072	0.090607928	1.309392072



Multiple Linear Regression

Multiple regression models involve several independent variables to estimate or predict one dependent variable. Although the independent variables can be higher order terms (like x^2) or qualitative terms, the linear regression model includes only quantitative, first-order terms.

The general model for k number of predictors is:

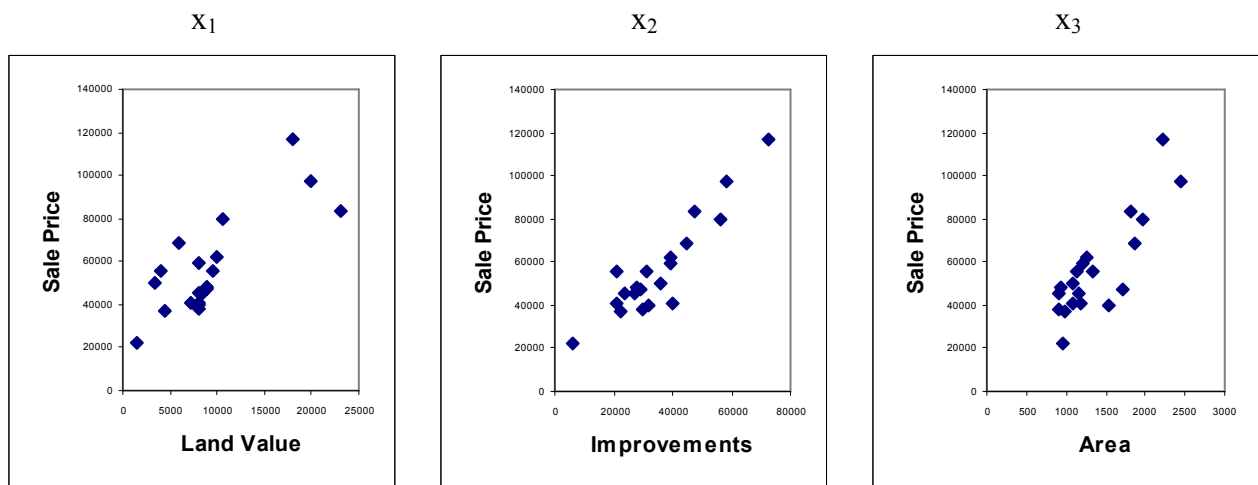
$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_kx_k + \varepsilon$$

β_i is the slope of the line relating y to x_i when all other x's are held constant.

The method of least squares is used to produce the model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1x_1 + \dots + \hat{\beta}_kx_k$ that minimizes the sum of squares of errors $SSE = \sum (y - \hat{y})^2$.

Since producing the model may involve solving a large system of simultaneous linear equations, computer programs will be used.

The real estate example in the text looks at the influence of three factors on the sales price of a house and yields scatter plots of:



with a regression equation of: $\hat{y} = 1470 + .8145x_1 + .8204x_2 + 13.53x_3$

The β coefficients mean:

- A \$1 increase in $x_1 \rightarrow$ \$.8145 increase in y when x_2 & x_3 are held constant
- A \$1 increase in $x_2 \rightarrow$ \$.8204 increase in y when x_1 & x_3 are held constant
- A 1 sq ft increase in $x_3 \rightarrow$ \$13.53 increase in y when x_1 & x_2 are held constant

In general, $\hat{\beta}_0$ has no practical interpretation, unless it makes sense to set all the predictors equal to zero.

The variance of the random error ε is estimated by:

$$s^2 = \frac{\text{SSE}}{n - \# \text{ of } \beta \text{ parameters}} = \frac{\sum (y_i - \hat{y}_i)^2}{n - (k + 1)}$$

n = number of observations of the y variable

k = number of independent (x) variables

$k + 1$ = number of β parameters

\hat{y} = y value predicted by the model

s^2 = the **mean square for error** (MSE)

SSE is found under Residual SS in the printout; MSE is under Residual MS.

Confidence intervals can be constructed for each β separately by:

$$\hat{\beta}_i \pm t_{\alpha/2} s_{\hat{\beta}_i} \quad \text{where } t_{\alpha/2} \text{ is based on degrees of freedom } n - (k + 1)$$

$s_{\hat{\beta}_i}$ is listed as “standard error” in the Excel table

Tests on β

To test the hypothesis that an individual $\beta_i = 0$, use a **t-test** with $n - (k+1)$ degrees of freedom.

$$H_0: \beta_i = 0$$

$$H_a: \beta_i > 0 \quad \text{or} \quad \beta_i < 0 \quad \text{or} \quad \beta_i \neq 0$$

$$\text{Rejection region: } t > t_{\alpha} \quad \text{or} \quad t < -t_{\alpha} \quad \text{or} \quad |t| > t_{\alpha/2}$$

$$\text{Test } t = \frac{\hat{\beta}_i}{s_{\hat{\beta}_i}}$$

If you reject the null hypothesis, then you can conclude that the independent variable y does have a linear relationship to x_i *when the other x values are held constant*.

If you do not reject the null, then either: 1) there is no relationship or the relationship is non-linear; or 2) a Type II error has occurred. So, the test is really testing for evidence of a *linear* relationship to individual x variables.

Multiple coefficient of determination

Conducting a series of t tests on each β parameter is a risky way of deciding which independent variables to keep and which to discard (concerning predictive value). For example, a set of t tests on 10 β 's, all of which are about zero, will result in a Type I error in at least one of the tests about 40% of the time.

The multiple coefficient of determination R^2 has the same formula as for single linear regression:

$$R^2 = \frac{SS_{yy} - SSE}{SS_{yy}} = \frac{\sum (\hat{y} - \bar{y})^2}{\sum (y - \bar{y})^2} = \frac{\text{explained variation}}{\text{total variation}}$$

R^2 is a measure of how well the model fits the data, that is, how *useful* the model is, as long as the sample contains substantially more data points than the number of β parameters.

The **adjusted multiple coefficient of determination** R_a^2 adjusts for sample size n and number of β parameters and is always smaller (more conservative) than R^2 .

$$R_a^2 = 1 - \left[\frac{(n-1)}{n-(k+1)} \right] \left(\frac{SSE}{SS_{yy}} \right) = 1 - \left[\frac{(n-1)}{n-(k+1)} \right] (1 - R^2)$$

The global F-test

Judging the global usefulness, or overall significance, of the model is done using the following hypotheses:

$$H_0: \beta_1 = \beta_2 = \beta_3 = \dots = \beta_k = 0$$

$$H_a: \text{At least one } \beta \neq 0$$

The test F-statistic is:
$$F = \frac{(SS_{yy} - SSE)/k}{SSE/[n-(k+1)]} = \frac{R^2/k}{(1-R^2)/[n-(k+1)]}$$

$$= \frac{\text{Mean Square, Regression}}{\text{Mean Square, Residual (Error)}}$$

F_α is determined using k degrees of freedom in the numerator and $n-(k+1)$ in the denominator.

In Excel, the p-value (observed significance) of the global F-test is listed as “Significance F”.

Rejecting the null means that the model is statistically useful, not that it’s the best model in terms of prediction and reliability. If the null is rejected, you can do t-tests on β parameters of the most interest to test for individual significance.

Confidence intervals for y

PHStat in Excel calculates confidence (estimation) intervals and prediction intervals around \hat{y} in an interactive worksheet that asks for values for the independent variables. As in simple linear regression, the prediction interval is always wider than the estimation interval.

Regression Analysis

<u>Regression Statistics</u>	
Multiple R	0.94732611
R Square	0.897426758

Correlation coefficient R
Coefficient of determination: (SSRegression/SSTotal)

Adjusted R Square	0.878194275
Standard Error	7919.482541
Observations	20

SE of predicted y values = $\sqrt{\text{MSE}}$

Regression SS = SSyy - SSE = $\sum (\hat{y} - \bar{y})^2$ (explained variability)
Residual SS = SSE = $\sum (y_i - \hat{y})^2$ (unexplained variability)
Total SS = SSyy = $\sum (y_i - \bar{y})^2$
Residual MS = MSE or s^2

<u>ANOVA</u>					
	df	SS	MS	F	Significance F
Regression	3	8779676741	2926558914	46.66203335	3.8999E-08
Residual	n-(k+1) 16	1003491259	62718203.71		
Total	n-1 19	9783168000			overall p-value

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	$\hat{\beta}_0$ 1470.275919	5746.324583	0.2558663709	0.801316449	-10711.38794	13651.93977
LandValue	$\hat{\beta}_1$ 0.814490116	0.512218713	1.590121751	0.131369541	-0.27136504	1.900345273
Improvements	$\hat{\beta}_2$ 0.820444695	0.211184936	3.884958409	0.001314786	0.372752632	1.268136758
Area	$\hat{\beta}_3$ 13.52864993	6.585680064	2.05425253	0.056664442	-0.432368042	27.489666791

