cars

speed dist

1 4 2

2 4 10

3 7 4

4 7 22

5 8 16

6 9 10

7 10 18

The cars dataset gives Speed and Stopping Distances of Cars. This dataset is a data frame with 50 rows and 2 variables. The rows refer to cars and the variables refer to speed (the numeric Speed in mph) and dist (the numeric stopping distance in ft.).

Linear Regression Model

create a linear model using lm(FORMULA, DATA)

model1=lm(dist~speed.c)

// lm- linear model (dependent var ~ independent variables)

summary(model1)

Residuals:

Min 1Q Median 3Q Max

-29.069 -9.525 -2.272 9.215 43.201

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 42.9800 2.1750 19.761 < 2e-16 \*\*\*

speed.c 3.9324 0.4155 9.464 1.49e-12 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.38 on 48 degrees of freedom

Multiple R-squared: 0.6511, Adjusted R-squared: 0.6438

F-statistic: 89.57 on 1 and 48 DF, p-value: 1.49e-12

Dist = 42.98 + 3.93speed

Result Interpretation:-

1. Residuals:

Min 1Q Median 3Q Max

-29.069 -9.525 -2.272 9.215 43.201

Residuals are essentially the difference between the actual value and predicted values (response values that the model predict (distance to stop car in our case)).

When assessing how well the model fit the data, you should look for a symmetrical distribution across these points on the mean value zero (0). In our example, we can see that the distribution of the residuals does not appear to be strongly symmetrical. That means that the model predicts certain points that fall far away from the actual observed points. We could take this further consider plotting the residuals to see whether this normally distributed, etc. but will skip this for this example.

2. **Coefficients**

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 42.9800 2.1750 19.761 < 2e-16 \*\*\*

speed.c 3.9324 0.4155 9.464 1.49e-12 \*\*\*

Dist = 42.98 + 3.93speed

In simple linear regression, the coefficients are two unknown constants that represent the intercept and slope terms in the linear model. If we wanted to predict the Distance required for a car to stop given its speed, we would get a training set and produce estimates of the coefficients to then use it in the model formula.

every 1 mph increase in the speed of a car, the required distance to stop goes up by 3.9324088 feet

State the Hypotheses

The [null hypothesis](http://stattrek.com/Help/Glossary.aspx?Target=Null%20hypothesis) states that the slope is equal to zero, and the alternative hypothesis states that the slope is not equal to zero.

H0: Β1 = 0

H0: The slope of the regression line is equal to zero.

Ha: Β1 ≠ 0

Ha: The slope of the regression line is *not* equal to zero.

Note:

If there is a significant linear relationship between the independent variable *X* and the dependent variable *Y*, the slope will *not* equal zero.

*Aim: Try to discard/reject null hypothesis*

**3. Standard Error [ideal close to 0]**

**(Standard Error of the Estimate)**

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 42.9800 2.1750 19.761 < 2e-16 \*\*\*

speed.c 3.9324 0.4155 9.464 1.49e-12 \*\*\*

Figure 1 shows two regression examples. You can see that in Graph A, the points are closer to the line than they are in Graph B. Therefore, the predictions in Graph A are more accurate than in Graph B.

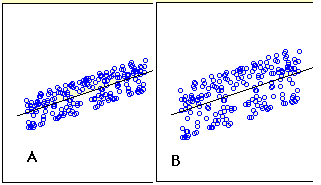
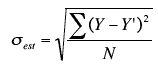


Figure 1. Regressions differing in accuracy of prediction.

The standard error of the estimate is a measure of the accuracy of predictions. The standard error of the estimate is defined below:



Recall that the regression line is the line that minimizes the sum of squared deviations of prediction (also called the *sum of squares error*). The standard error of the estimate is closely related to this quantity.

Where σest is the standard error of the estimate, Y is an actual score, Y' is a predicted score, and N is the number of pairs of scores. The numerator is the sum of squared differences between the actual scores and the predicted scores.

Assume the data in Table 1 are the data from a population of five X, Y pairs.

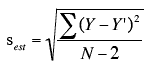
Table 1. Example data.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **X** | **Y** | **Y'** | **Y-Y'** | **(Y-Y')2** |
|  | 1.00 | 1.00 | 1.210 | -0.210 | 0.044 |
|  | 2.00 | 2.00 | 1.635 | 0.365 | 0.133 |
|  | 3.00 | 1.30 | 2.060 | -0.760 | 0.578 |
|  | 4.00 | 3.75 | 2.485 | 1.265 | 1.600 |
|  | 5.00 | 2.25 | 2.910 | -0.660 | 0.436 |
| Sum | 15.00 | 10.30 | 10.30 | 0.000 | 2.791 |

The last column shows that the sum of the squared errors of prediction is 2.791. Therefore, the standard error of the estimate is

http://onlinestatbook.com/2/regression/graphics/se_est_example.gif

Similar formulas are used when the standard error of the estimate is computed from a sample rather than a population. The only difference is that the denominator is N-2 rather than N. The reason N-2 is used rather than N is that two parameters (the slope and the intercept) were estimated in order to estimate the sum of squares. Formulas for a sample comparable to the ones for a population are shown below.



http://onlinestatbook.com/2/regression/graphics/se_estimate_sample_example.gif

***Interpretation (Std. Error of Est.):***

* *Std. Error of Est* . represents average observed values fall from the estimate.
* Conveniently, it tells you how wrong the regression model is on average using the units of the response variable.
* Smaller values are better because it indicates that the observations are closer to the fitted line.
* The Standard Error measures the average amount that the estimates vary. We’d ideally want a lower number relative to its coefficients.
* In our example, we’ve previously determined that for every 1 mph increase in the speed of a car, the required distance to stop goes up by 3.9324088 feet. we can say that the required distance for a car to stop can vary by **0.4155128** feet from actual (value given by line) (3.932).
* The Standard Errors can also be used to compute confidence intervals.

*4.* **t value [ideal high value/ far away from 0]**

(Significance Test for the Slope ‘b’)

Estimate Std. Error t value Pr(>|t|)

(Intercept) 42.9800 2.1750 19.761 < 2e-16 \*\*\*

speed.c 3.9324 0.4155 9.464 1.49e-12 \*\*\*

**t = coefficient of slope ‘b’ / standard error of the slope, Sb**

The estimated standard error (std. error of est.) of b is computed using the following formula:

http://onlinestatbook.com/2/regression/graphics/seb.gif

**sb = sqrt [ Σ(yi - ŷi)2 / (n - 2) ] / sqrt [ Σ(xi – xmean)2 ]**

Where sb is the estimated standard error of b, sest is the standard error of the estimate, and SSX is the sum of squared deviations of X from the mean of X. SSX is calculated as

http://onlinestatbook.com/2/regression/graphics/ssx.gif

where Mx is the mean of X.

These formulas are illustrated with the data shown in Table 1.  The column X has the values of the *predictor variable* and the column Y has the values of the *criterion variable*. The third column, x, contains the differences between the values of column X and the mean of X. The fourth column, x2, is the square of the x column. The fifth column, y, contains the differences between the values of column Y and the mean of Y. The last column, y2, is simply square of the y column.

Table 1.

Example data.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **X** | **Y** | **x** | **x2** | **y** | **y2** |
|  | 1.00 | 1.00 | -2.00 | 4 | -1.06 | 1.1236 |
|  | 2.00 | 2.00 | -1.00 | 1 | -0.06 | 0.0036 |
|  | 3.00 | 1.30 | 0.00 | 0 | -0.76 | 0.5776 |
|  | 4.00 | 3.75 | 1.00 | 1 | 1.69 | 2.8561 |
|  | 5.00 | 2.25 | 2.00 | 4 | 0.19 | 0.0361 |
| Sum | 15.00 | 10.30 | 0.00 | 10.00 | 0.00 | 4.5970 |

The computation of the standard error of the estimate (sest) for these data is 0.964.

sest = 0.964

SSX is the sum of squared deviations from the mean of X. It is, therefore, equal to the sum of the x2 column and is equal to 10.

SSX = 10.00

We now have all the information to compute the standard error of b:

http://onlinestatbook.com/2/regression/graphics/seb_ex.gif

As shown previously, the slope (b) is 0.425. Therefore,

http://onlinestatbook.com/2/regression/graphics/t_b.gif

***Interpretation (t-value):***

The t-value is a measure of how many standard deviations our coefficient estimate is far away from 0. We want it to be far away from zero as this would indicate we could reject the null hypothesis - that is, we could declare a relationship between speed and distance exist.

**5. P-Value-Coefficient - Pr(>t) [ideal close to 0 / lest than 5% (alpha level=0.05) acceptable]**

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 42.9800 2.1750 19.761 < 2e-16 \*\*\*

speed.c 3.9324 0.4155 9.464 1.49e-12 \*\*\*

The [p-value](http://statisticsbyjim.com/glossary/p-value/) for each independent variable tests the [null hypothesis](http://statisticsbyjim.com/glossary/null-hypothesis/) that the variable has no [correlation](http://statisticsbyjim.com/glossary/correlation/) with the dependent variable. If there is no correlation, there is no association between the changes in the independent variable and the shifts in the dependent variable. In other words, there is no [effect](http://statisticsbyjim.com/glossary/effect/).

If the p-value for a variable is less than your [significance level](http://statisticsbyjim.com/glossary/significance-level/), your [sample](http://statisticsbyjim.com/glossary/sample/) data provide enough evidence to reject the null hypothesis for the entire [population](http://statisticsbyjim.com/glossary/population/). Your data favor the hypothesis that there *is* a non-zero correlation. Changes in the independent variable *are* associated with changes in the response at the population level. This variable is statistically significant and probably a worthwhile addition to your regression model.

On the other hand, a p-value that is greater than the significance level indicates that there is insufficient evidence in your sample to conclude that a non-zero correlation exists.

The p-value for each term tests the null hypothesis that the coefficient is equal to zero (no effect)…*probability.*

What is desired: probability should be minimum for coefficient equal to zero

A low p-value (< 0.05) indicates that you can reject the null hypothesis. In other words, a predictor that has a low p-value is likely to be a meaningful addition to your model because changes in the predictor's value are related to changes in the response variable.

Conversely, a larger (insignificant) p-value suggests that changes in the predictor are not associated with changes in the response.

The Pr(>t) acronym found in the model output relates to the probability of observing any value equal or larger than t. Typically, a p-value of 5% or less is a good cut-off point.

Pr(>t)  should close to 0

Probability that model will deviate more than t value– It should be zero

Pr(<t)  should close to 1

Probability that model will not deviate more than t value– It should be 1 (100%)

*6.* **Residual Standard Error [ideal close to 0]**

Residual standard error: 15.38 on 48 degrees of freedom

Multiple R-squared: 0.6511, Adjusted R-squared: 0.6438

F-statistic: 89.57 on 1 and 48 DF, p-value: 1.49e-12

* Residual Standard Error is measure of the quality of a linear regression fit. Theoretically, every linear model is assumed to contain an error term E. Due to the presence of this error term, we are not capable of perfectly predicting our response variable (dist) from the predictor (speed) one.
* The Residual Standard Error is the average amount that the response (dist) will deviate from the true regression line.
* In our example, the actual distance required to stop can deviate from the true regression line by approximately **15.3795867** feet, on average. In other words, given that the mean distance for all cars to stop is **42.98** and that the Residual Standard Error is **15.3795867**, we can say that the percentage error is (any prediction would still be off by) **35.78%**.
* It’s also worth noting that the Residual Standard Error was calculated with 48 degrees of freedom. Simplistically, degrees of freedom are the number of data points that went into the estimation of the parameters used after taking into account these parameters (restriction). In our case, we had 50 data points and two parameters (intercept and slope).
* “**Residual Standard Error**” is the  standard deviation of the error  with a slight twist.  [Standard deviation](http://www.learnbymarketing.com/definitions/average-variance-and-standard-deviation/) is the square root of variance.  Standard Error is very similar to RMSE.  The only difference is that instead of dividing by n-1, you subtract n minus 1 + # of variables involved.

sqrt(SSE/(n-1-k)) #Residual Standard Error

K= number of coefficient

SSE=Total sum of squared error, Σ(yi - ŷi)2

**7. F-Value: F-Statistic and P-Value**

Residual standard error: 15.38 on 48 degrees of freedom

Multiple R-squared: 0.6511, Adjusted R-squared: 0.6438

F-statistic: 89.57 on 1 and 48 DF, p-value: 1.49e-12

The F-Statistic is a “global” test that checks if at least one of your coefficients are nonzero.

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9 | #F-Statistic  #Ho: All coefficients are zero  #Ha: At least one coefficient is nonzero  SSE;-sum(model$residuals\*\*2)  SSyy;-sum((y-mean(y))\*\*2)  K= # coefficient  **F= ((SSyy-SSE)/k) / (SSE/(n-(k+1)))**  It’s similar to a [T statistic](http://www.statisticshowto.com/t-statistic/)from a [T-Test](http://www.statisticshowto.com/probability-and-statistics/t-test/); A-T test will tell you if a single variable is [statistically significant](http://www.statisticshowto.com/what-is-statistical-significance/) and an F test will tell you if a group of variables are jointly significant. |