

Inferences about Regression

Raja Almukkahal
Larry Ottman
Danielle DeLancey
Addie Evans
Ellen Lawskey
Brenda Meery

To access the online version of this FlexBook
click the link below:

https://www.ck12.org/c/statistics/inferences-about-regression/lesson/Inferences-about-Regression-ADV-PST/?referrer=concept_details

To access a customizable version of this book, as well as other interactive content, visit www.ck12.org

CK-12 Foundation is a non-profit organization with a mission to reduce the cost of textbook materials for the K-12 market both in the U.S. and worldwide. Using an open-source, collaborative, and web-based compilation model, CK-12 pioneers and promotes the creation and distribution of high-quality, adaptive online textbooks that can be mixed, modified and printed (i.e., the FlexBook® textbooks).

Copyright © 2021 CK-12 Foundation, www.ck12.org

The names “CK-12” and “CK12” and associated logos and the terms “**FlexBook®**” and “**FlexBook Platform®**” (collectively “CK-12 Marks”) are trademarks and service marks of CK-12 Foundation and are protected by federal, state, and international laws.

Any form of reproduction of this book in any format or medium, in whole or in sections, must be attributed according to our attribution guidelines.

<https://www.ck12info.org/about/attribution-guidelines>

Except as otherwise noted, all CK-12 Content (including CK-12 Curriculum Material) is made available to Users in accordance with the CK-12 Curriculum Materials License

<https://www.ck12info.org/curriculum-materials-license>



Complete terms for use for the CK-12 website can be found at:
<http://www.ck12info.org/terms-of-use/>

Printed: November 22, 2021 (PST)



AUTHORS

Raja Almukkahal
Larry Ottman
Danielle DeLancey
Addie Evans
Ellen Lawsky
Brenda Meery



Inferences about Regression

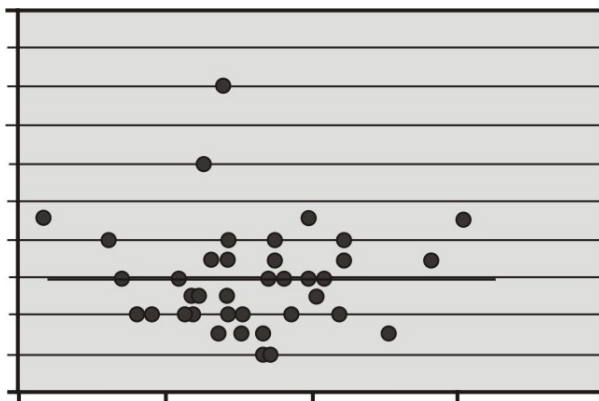
Linear regression models used for testing null hypothesis, predicting probability and for building confidence intervals

Inferences about Regression

Inferences and Assumptions about Linear Regression

Hypothesis Testing for Linear Relationships

Let's think for a minute about the relationship between correlation and the **linear regression model**. As we learned, if there is no correlation between the two variables X and Y , then it would be nearly impossible to fit a meaningful regression line to the points on a scatterplot graph. If there was no correlation, and our correlation value, or r -value, was 0, we would always come up with the same predicted value, which would be the mean of all the **predicted values**, or the mean of $\hat{Y}$. The figure below shows an example of what a regression line fit to variables with no correlation ($r = 0$) would look like. As you can see, for any value of X , we always get the same predicted value of Y .



Using this knowledge, we can determine that if there is no relationship between X and Y , constructing a regression line doesn't help us very much, because, again, the predicted score would always be the same. Therefore, when we estimate a linear regression model, we want to ensure that the **regression coefficient**, β , for the population does not equal zero. Furthermore, it is beneficial to test how strong (or far away) from zero the regression coefficient must be to strengthen our prediction of the Y scores.

In hypothesis testing of linear regression models, the null hypothesis to be tested is that the regression coefficient, β , equals zero. Our alternative hypothesis is that our regression coefficient does not equal zero.

$$H_0 : \beta = 0$$

$$H_a : \beta \neq 0$$

The test statistic for this hypothesis test is calculated as follows:

$$t = \frac{b - \beta}{s_b}$$

where $s_b = \frac{s}{\sqrt{\sum (x - \bar{x})^2}} = \frac{s}{\sqrt{SS_X}}$,

$$s = \sqrt{\frac{SSE}{n - 2}}, \text{ and}$$

$$SSE = \text{sum of residual error squared}$$

Testing the Accuracy of a Regression Equation

Let's say that a football coach is using the results from a short physical fitness test to predict the results of a longer, more comprehensive one. He developed the regression equation $Y = 0.635X + 1.22$, and the standard error of estimate is 0.56. The summary statistics are as follows:

Summary statistics for two foot ball fitness tests.

$n = 24$	$\sum xy = 591.50$
$\sum x = 118$	$\sum y = 104.3$
$\bar{x} = 4.92$	$\bar{y} = 4.35$
$\sum x^2 = 704$	$\sum y^2 = 510.01$
$SS_X = 123.83$	$SS_Y = 56.74$

Using $\alpha = 0.05$, test the null hypothesis that, in the population, the regression coefficient is zero, or $H_0 : \beta = 0$.

We use the t -distribution to calculate the test statistic and find that the critical values in the t -distribution at 22 degrees of freedom are 2.074 standard scores above and below the mean. Also, the test statistic can be calculated as follows:

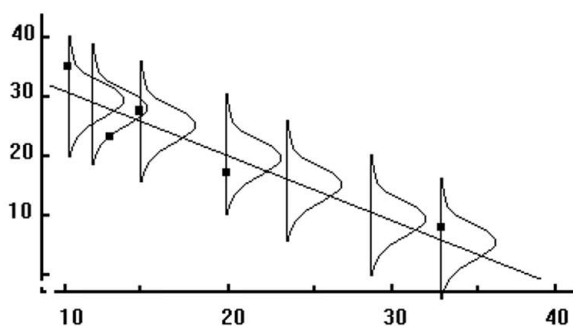
$$s_b = \frac{0.56}{\sqrt{123.83}} = 0.05$$

$$t = \frac{0.635 - 0}{0.05} = 12.70$$

Since the observed value of the test statistic exceeds the critical value, the null hypothesis would be rejected, and we can conclude that if the null hypothesis were true, we would observe a regression coefficient of 0.635 by chance less than 5% of the time.

Making Inferences about Predicted Scores

As we have mentioned, a regression line makes predictions about variables based on the relationship of the existing data. However, it is important to remember that the regression line simply infers, or estimates, what the value will be. These predictions are never accurate 100% of the time, unless there is a perfect correlation. What this means is that for every predicted value, we have a normal distribution (also known as the **conditional distribution**, since it is conditional on the X value) that describes the likelihood of obtaining other scores that are associated with the value of the **predictor variable**, X .



If we assume that these distributions are normal, we are able to make inferences about each of the predicted scores. We can ask questions like, “If the predictor variable, X , equals 4, what percentage of the distribution of Y scores will be lower than 3?”

The reason why we would ask questions like this depends on the scenario. Suppose, for example, that we want to know the percentage of students with a 5 on their short physical fitness test that have a predicted score higher than 5 on their long physical fitness test. If the coach is using this predicted score as a cutoff for playing in a varsity match, and this percentage is too low, he may want to consider changing the standards of the test.

To find the percentage of students with scores above or below a certain point, we use the concept of standard scores and the standard normal distribution.

Since we have a certain predicted value for every value of X , the Y values take on the shape of a normal distribution. This distribution has a mean (the regression line) and a standard error, which we found to be equal to 0.56. In short, the conditional distribution is used to determine the percentage of Y values above or below a certain value that are associated with a specific value of X .

Calculating Probability

Using our example above, if a student scored a 5 on the short test, what is the probability that he or she would have a score of 5 or greater on the long physical fitness test?

From the regression equation $Y = 0.635X + 1.22$, we find that the predicted score when the value of X is 5 is 4.40. Consider the conditional distribution of Y scores when the value of X is 5. Under our assumption, this distribution is normally distributed around the predicted value 4.40 and has a standard error of 0.56.

Therefore, to find the percentage of Y scores of 5 or greater, we use the general formula for a z -score to calculate the following:

$$z = \frac{Y - \hat{Y}}{s} = \frac{5 - 4.40}{0.56} = 1.07$$

Using the z -distribution table, we find that the area to the right of a z -score of 1.07 is 0.1423. Therefore, we can conclude that the proportion of predicted scores of 5 or greater given a score of 5 on the short test is 0.1423, or 14.23%.

Prediction Intervals

Similar to hypothesis testing for samples and populations, we can also build a confidence interval around our regression results. This helps us ask questions like “If the predictor variable, X , is equal to a certain value, what are the likely values for Y ?” A confidence interval gives us a range of scores that has a certain percent probability of including the score that we are after.

We know that the standard error of the predicted score is smaller when the predicted value is close to the actual value, and it increases as X deviates from the mean. This means that the weaker of a predictor that the regression line is, the larger the standard error of the predicted score will be. The formulas for the standard error of a predicted score and a confidence interval are as follows:

$$s_{\hat{Y}} = s \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x - \bar{x})^2}}$$

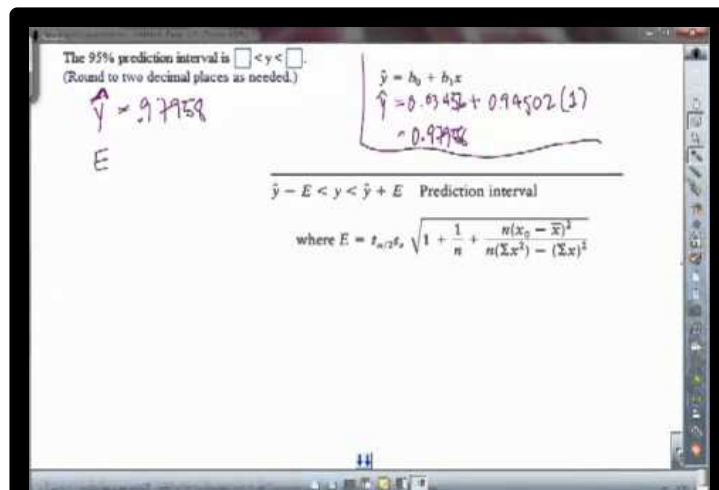
$$CI = \hat{Y} \pm t s_{\hat{Y}}$$

where:

$\hat{Y}$ is the predicted score.

t is the critical value for $n - 2$ degrees of freedom.

$s_{\hat{Y}}$ is the standard error of the predicted score.



Developing Confidence Intervals

Develop a 95% confidence interval for the predicted score of a student who scores a 4 on the short physical fitness exam.

We calculate the standard error of the predicted score using the formula as follows:

$$s_{\hat{Y}} = s \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x - \bar{x})^2}} = 0.56 \sqrt{1 + \frac{1}{24} + \frac{(4 - 4.92)^2}{123.83}} = 0.57$$

Using the general formula for a confidence interval, we can calculate the answer as shown:

$$\begin{aligned} CI &= \hat{Y} \pm t s_{\hat{Y}} \\ CI_{0.95} &= 3.76 \pm (2.074)(0.57) \\ CI_{0.95} &= 3.76 \pm 1.18 \\ CI_{0.95} &= (2.58, 4.94) \end{aligned}$$

Therefore, we can say that we are 95% confident that given a student's short physical fitness test score, X , of 4, the interval from 2.58 to 4.94 will contain the student's score for the longer physical fitness test.

Regression Assumptions

We make several assumptions under a linear regression model, including:

At each value of X , there is a distribution of Y . These distributions have a mean centered at the predicted value and a standard error that is calculated using the sum of squares.

Using a regression model to predict scores only works if the regression line is a good fit to the data. If this relationship is non-linear, we could either transform the data (i.e., a logarithmic transformation) or try one of the other regression equations that are

available with Excel or a graphing calculator.

The standard deviations and the variances of each of these distributions for each of the predicted values are equal. This is called homoscedasticity.

Finally, for each given value of X , the values of Y are independent of each other.

Example

The following example uses data from the previous section:

Verbal SAT scores were used to predict the GPA of students. From the data, we found this least squares regression line:

$$\hat{Y} = 0.0055X + 0.097$$

We also found that $SSE = 0.26$ for the $n = 7$ samples.

Example 1

Suppose a student scores a 650 on the verbal SAT. Assuming the data is normally distributed, what is the probability that they will have a GPA of at least 3.8?

Using the least squares regression line, we will predict the GPA for a verbal SAT score of 650:

$$\hat{Y} = 0.0055(650) + 0.097 = 3.575 + 0.097 = 3.672$$

This means that an SAT score of 650 predicts that the student will have a GPA of 3.672. They could have a higher, or a lower GPA though, so now we will look at the probability that a student with a GRE score of 650 has a GPA of at least 3.8. First we have to find:

$$s = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{0.26}{7-2}} = \sqrt{\frac{0.26}{5}} \approx 0.228$$

$$z = \frac{Y - \hat{Y}}{s} = \frac{3.8 - 3.672}{0.228} \approx 0.56$$

Now we simply look at the z -table to find the probability of getting a z -score of 0.56 or higher.

$$P(z > 0.56) = 0.288.$$

The probability of having a GPA of at least 3.8 when scoring a 650 on the verbal SAT is 0.288.

Review

For 1-10, a college counselor is putting on a presentation about the financial benefits of further education and takes a random sample of 120 parents. Each parent was asked a number of questions, including the number of years of education that he or she has (including college) and his or her yearly income (recorded in the thousands of dollars). The summary data for this survey are as follows:

$$n = 120 \quad r = 0.67$$

$$\sum x = 1,782 \quad \sum y = 1,854$$

$$s_X = 3.6 \quad s_Y = 4.2$$

$$s_{XY} = 3.12 \quad SS_X = 1542$$

1. What is the predictor variable? What is your reasoning behind this decision?
2. Do you think that these two variables (income and level of formal education) are correlated? If so, please describe the nature of their relationship.
3. What would be the regression equation for predicting income, $\hat{Y}$, from the level of education, X ?
4. Using this regression equation, predict the income for a person with 2 years of college (13.5 years of formal education).
5. Test the null hypothesis that in the population, the regression coefficient for this scenario is zero.
 - a. First develop the [null and alternative hypotheses](#).
 - b. Set the critical value to $\alpha = 0.05$.
 - c. Compute the test statistic.
 - d. Make a decision regarding the null hypothesis.
6. For those parents with 15 years of formal education, what is the percentage who will have an annual income greater than \$18,500?
7. For those parents with 12 years of formal education, what is the percentage who will have an annual income greater than \$18,500?
8. Develop a 95% confidence interval for the predicted annual income when a parent indicates that he or she has a college degree (i.e., 16 years of formal education).
9. If you were the college counselor, what would you say in the presentation to the parents and students about the relationship between further education and salary? Would you encourage students to further their education based on these analyses? Why or why not?

10. Using the same [null and alternative hypotheses](#), and test statistics as you did in question 5, make a decision at the significance level of $\alpha = 0.01$.

Review (Answers)

To view the Review answers, open this [PDF file](#) and look for section 9.3.

Vocabulary

Language: **English** ▼

Term	Definition
homoscedasticity	Homoscedasticity occurs when the standard deviations and the variances of each of these distributions for each of the predicted values are equal.
linear regression model	When we estimate a linear regression model, we want to ensure that the regression coefficient for the population, beta, does not equal zero.
predicted values	The positive predictive value is the probability that a test gives a true result for a true statistic. The negative predictive value is the probability that a test gives a false result for a false statistic
predictor variable	A predictor variable is a variable that can be used to predict the value of another variable (as in statistical regression).
regression coefficient	The regression coefficient (or the slope) explains the nature of the relationship between the two variables.

1.0 REFERENCES

Image	Attributions
	
	