Hypothesis Testing

Regression

Objectives

1. Perform a hypothesis test on the slope of the regression equation
2. Perform hypothesis tests on the linear correlation coefficient
3. Create a prediction interval for a regression equation

Slope of a Line

Recall that the slope of a line tells us how much the y-coordinates change by increasing each x-coordinate by 1.

Slope of a Line

Recall that the slope of a line tells us how much the y-coordinates change by increasing each x-coordinate by 1.

Previously, we looked at calculating regression equations in the form

$$\hat{y} = mx + b$$

which we can use to predict data values both inside and outside our dataset.

Slope of a Line

Recall that the slope of a line tells us how much the y-coordinates change by increasing each x-coordinate by 1.

Previously, we looked at calculating regression equations in the form

$$\hat{y} = mx + b$$

which we can use to predict data values both inside and outside our dataset.

Also, recall that the coefficient of determination, r^2, tells us the prediction error decrease we get by using the linear regression equation, as opposed to the mean of the y-coordinates.

Using the y-Intercept

However, we might also be able to use the y-intercept of our linear regression equation as a predictor:

$$y = b$$

Using the y-Intercept

However, we might also be able to use the y-intercept of our linear regression equation as a predictor:

$$y = b$$

In this case, the slope of the equation, m, is 0.

Using the y-Intercept

However, we might also be able to use the y-intercept of our linear regression equation as a predictor:

$$y = b$$

In this case, the slope of the equation, m, is 0.

We can then test whether or not the linear regression model we obtained contributes to predicting values of y.

Hypotheses Regarding Slope

For this hypothesis, testing, we are operating under the assumption that the following null hypothesis is true:

The slope of the regression equation does not help in predicting values of y.

Hypotheses Regarding Slope

For this hypothesis, testing, we are operating under the assumption that the following null hypothesis is true:

The slope of the regression equation does not help in predicting values of y.

To put it mathematically:

$$H_0 : m = 0$$

Hypotheses Regarding Slope

For this hypothesis, testing, we are operating under the assumption that the following null hypothesis is true:

The slope of the regression equation does not help in predicting values of y.

To put it mathematically:

$$H_0 : m = 0$$

Our alternative hypothesis will be one of the forms

Left-Tailed	**Right-Tailed**	**Two-Tailed**
$m < 0$	$m > 0$	$m \neq 0$

Hypotheses Regarding Slope

For this hypothesis, testing, we are operating under the assumption that the following null hypothesis is true:

The slope of the regression equation does not help in predicting values of y.

To put it mathematically:

$$H_0 : m = 0$$

Our alternative hypothesis will be one of the forms

Left-Tailed	**Right-Tailed**	**Two-Tailed**
$m < 0$	$m > 0$	$m \neq 0$

H_A states that x is not helpful in predicting y.

Test Statistic

Since the population standard deviation of our regression slope is likely not to be known, we can estimate it by using the **estimated standard error of the least squares slope**.

Test Statistic

Since the population standard deviation of our regression slope is likely not to be known, we can estimate it by using the **estimated standard error of the least squares slope**.

$$\frac{s}{\sqrt{SS_{xx}}}$$

where

$$SS_{xx} = \text{total sum of squares} = \sum x^2 - \frac{1}{n}\left(\sum x\right)^2$$

Test Statistic

Since the population standard deviation of our regression slope is likely not to be known, we can estimate it by using the **estimated standard error of the least squares slope**.

$$\frac{s}{\sqrt{SS_{xx}}}$$

where

$$SS_{xx} = \text{total sum of squares} = \sum x^2 - \frac{1}{n}\left(\sum x\right)^2$$

Our test statistic, t, is thus

$$t = \frac{\hat{m}}{s/\sqrt{SS_{xx}}}$$

where $\hat{m}$ is the slope of our regression equation.

p-Value

The area under the curve either to the left, right, or outside of our test statistic(s), i.e. the p-value, can be found by taking $n-2$ degrees of freedom with that test statistic.

Confidence Intervals

We can construct confidence intervals as

$$\hat{m} \pm t_{\alpha/2}\left(s/\sqrt{SS_{xx}}\right)$$

Confidence Intervals

We can construct confidence intervals as

$$\hat{m} \pm t_{\alpha/2}\left(s/\sqrt{SS_{xx}}\right)$$

Remember, for one-sided confidence intervals, use t_α instead of $t_{\alpha/2}$.

Residuals Review

Recall that the **residual** of a data point is the distance the y-coordinate of that data point is from the corresponding y-coordinate on the regression equation.

Residuals Review

Recall that the **residual** of a data point is the distance the y-coordinate of that data point is from the corresponding y-coordinate on the regression equation.

We usually denote the residual of data point (x_i, y_i) as ϵ_i, and it can be found by

$$\epsilon_i = y_{\text{observed}} - y_{\text{predicted}}$$

Residuals Review

Recall that the **residual** of a data point is the distance the y-coordinate of that data point is from the corresponding y-coordinate on the regression equation.

We usually denote the residual of data point (x_i, y_i) as ϵ_i, and it can be found by

$$\epsilon_i = y_{\text{observed}} - y_{\text{predicted}}$$

When performing hypothesis testing regarding the slope of the regression equation, we operate under the following assumptions:

You Know What Happens When You Assume, Right?

Assumptions:

You Know What Happens When You Assume, Right?

Assumptions:

- The mean of all values of ϵ is 0.

You Know What Happens When You Assume, Right?

Assumptions:

- The mean of all values of ϵ is 0.
- The variance of all values of ϵ is constant for all values of x.

You Know What Happens When You Assume, Right?

Assumptions:

- The mean of all values of ϵ is 0.
- The variance of all values of ϵ is constant for all values of x.
- The distribution of ϵ is normal.

You Know What Happens When You Assume, Right?

Assumptions:

- The mean of all values of ϵ is 0.
- The variance of all values of ϵ is constant for all values of x.
- The distribution of ϵ is normal.
- Values of ϵ associated with any two values of y are independent.

Example 1

The table below lists the total study times x, in hours, and the score on a statistics exam y for 12 students.

(a) Find the least squares regression equation for the data and interpret the slope.

Time	1.9	1.3	1.7	1.2	1.2	2.7	1.7	2.7	2.0	1.8	1.4	2.4
Score	84	81	86	82	77	97	93	96	89	89	88	95

Example 1

The table below lists the total study times x, in hours, and the score on a statistics exam y for 12 students.

(a) Find the least squares regression equation for the data and interpret the slope.

Time	1.9	1.3	1.7	1.2	1.2	2.7	1.7	2.7	2.0	1.8	1.4	2.4
Score	84	81	86	82	77	97	93	96	89	89	88	95

$$\hat{y} = 10.3158x + 69.1711$$

Example 1

The table below lists the total study times x, in hours, and the score on a statistics exam y for 12 students.

(a) Find the least squares regression equation for the data and interpret the slope.

Time	1.9	1.3	1.7	1.2	1.2	2.7	1.7	2.7	2.0	1.8	1.4	2.4
Score	84	81	86	82	77	97	93	96	89	89	88	95

$$\hat{y} = 10.3158x + 69.1711$$

$$\text{Score} = 10.3158\,(\text{Time}) + 69.1711$$

Example 1

The table below lists the total study times x, in hours, and the score on a statistics exam y for 12 students.

(a) Find the least squares regression equation for the data and interpret the slope.

Time	1.9	1.3	1.7	1.2	1.2	2.7	1.7	2.7	2.0	1.8	1.4	2.4
Score	84	81	86	82	77	97	93	96	89	89	88	95

$$\hat{y} = 10.3158x + 69.1711$$

$$\text{Score} = 10.3158\,(\text{Time}) + 69.1711$$

For every increased hour in study time, the predicted score increases by about 10.3

Example 1

(b) At the 5% significance level, test the claim whether study time is useful as a predictor of test score.

Example 1

(b) At the 5% significance level, test the claim whether study time is useful as a predictor of test score.

$H_0: \ m = 0$ (study time is not useful)

Example 1

(b) At the 5% significance level, test the claim whether study time is useful as a predictor of test score.

$H_0: \ m = 0$ (study time is not useful)
$H_A: \ m \neq 0$ (study time is useful)

Example 1

(b) At the 5% significance level, test the claim whether study time is useful as a predictor of test score.

$H_0:\ m = 0$ (study time is not useful)
$H_A:\ m \neq 0$ (study time is useful)

t = 5.5365 (critical values: $\pm$2.2281)
p-value: 0.0002 ($\alpha = 0.05$)
95.0% confidence interval for population slope: (6.1643, 14.4673)

Example 1

(b) At the 5% significance level, test the claim whether study time is useful as a predictor of test score.

$H_0: \ m = 0$ (study time is not useful)
$H_A: \ m \neq 0$ (study time is useful)

t $= 5.5365$ (critical values: ± 2.2281)
p-value: 0.0002 ($\alpha = 0.05$)
95.0% confidence interval for population slope: (6.1643, 14.4673)

Reject the null hypothesis

Example 1

(b) At the 5% significance level, test the claim whether study time is useful as a predictor of test score.

$H_0: \ m = 0$ (study time is not useful)
$H_A: \ m \neq 0$ (study time is useful)

t $=$ 5.5365 (critical values: $\pm$2.2281)
p-value: 0.0002 ($\alpha = 0.05$)
95.0% confidence interval for population slope: (6.1643, 14.4673)

Reject the null hypothesis

> *At the 5% significance level, we have sufficient evidence to reject the claim that study time is not useful in predicting the score on this statistics exam.*

Objectives

1. Perform a hypothesis test on the slope of the regression equation
2. Perform hypothesis tests on the linear correlation coefficient
3. Create a prediction interval for a regression equation

Objectives

1. Perform a hypothesis test on the slope of the regression equation
2. Perform hypothesis tests on the linear correlation coefficient
3. Create a prediction interval for a regression equation