Regression Inferences
Sections 26.3, 26.4, 26.5
Lecture 51

Robb T. Koether

Hampden-Sydney College

Fri, Apr 22, 2016

Outline

Outline

Regression Assumptions

- For the regression model to be applicable, the following assumptions must hold.

Regression Assumptions

- For the regression model to be applicable, the following assumptions must hold.
- For each value of x,

Regression Assumptions

- For the regression model to be applicable, the following assumptions must hold.
- For each value of x,
  - The set of associated values of y are normally distributed.

Regression Assumptions

- For the regression model to be applicable, the following assumptions must hold.
- For each value of x,
  - The set of associated values of y are normally distributed.
  - The y values are independent of each other.

Regression Assumptions

- For the regression model to be applicable, the following assumptions must hold.
- For each value of x,
  - The set of associated values of y are normally distributed.
  - The y values are independent of each other.
  - The standard deviation of the associated y values is σ (the same value for all x).

Regression Assumptions

- For the regression model to be applicable, the following assumptions must hold.
- For each value of x,
  - The set of associated values of y are normally distributed.
  - The y values are independent of each other.
  - The standard deviation of the associated y values is σ (the same value for all x).
- Furthermore, for each x, $\hat{y}$ is an estimate of the *average y value,* μ_y, for that x value.

Regression Assumptions

- For the regression model to be applicable, the following assumptions must hold.
- For each value of x,
  - The set of associated values of y are normally distributed.
  - The y values are independent of each other.
  - The standard deviation of the associated y values is σ (the same value for all x).
- Furthermore, for each x, $\hat{y}$ is an estimate of the *average y value,* μ_y, for that x value.
- The final assumption is that the mean responses μ_y have a straight-line relationship with x.

Regression Assumptions

- Under these assumptions, we calculate the regression equation

$$\hat{y} = a + bx.$$

- But it is only an approximation to the "true" relation is given by

$$\mu_y = \alpha + \beta x.$$

- That is,

$$a \text{ is an estimate of } \alpha$$

and

$$b \text{ is an estimate of } \beta.$$

Example

Example (Quiz Average vs. Test Average)

- Is there a linear relation between quiz averages and test averages?

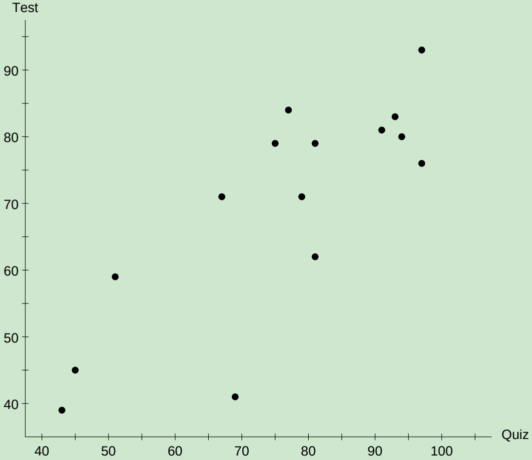
Quiz	Test
79	71
45	45
51	59
81	62
69	41

Quiz	Test
75	79
97	93
97	76
43	39
67	71

Quiz	Test
94	80
77	84
81	79
91	81
93	93

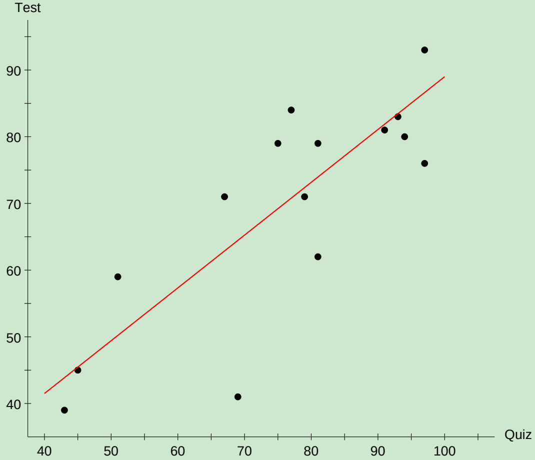
Example

Example (Quiz Average vs. Test Average)



Example

Example (Quiz Average vs. Test Average)



Outline

Testing the Significance of the Relation

- If there really is no relation between *x* and *y*, then the graph of the true equation would be a horizontal line (slope 0).
- If there really is a relation, then the slope will be nonzero.
- Therefore, to test of the existence of a relation, we test whether $\beta = 0$, using *b* as our estimator of β.

Testing the Significance of the Relation

- If there really is no relation between x and y, then the graph of the true equation would be a horizontal line (slope 0).
- If there really is a relation, then the slope will be nonzero.
- Therefore, to test of the existence of a relation, we test whether $\beta = 0$, using b as our estimator of β.
- To make a reasonable allowance for randomness, we must estimate the variability of the statistic b.

Testing the Significance of the Relation

- If there really is no relation between x and y, then the graph of the true equation would be a horizontal line (slope 0).
- If there really is a relation, then the slope will be nonzero.
- Therefore, to test of the existence of a relation, we test whether $\beta = 0$, using b as our estimator of β.
- To make a reasonable allowance for randomness, we must estimate the variability of the statistic b.
- Then, if b differs from 0 (which it most likely will), we can judge whether that difference is so great that it could not reasonably be attributed to randomness.

Testing the Significance of the Relation

- If there really is no relation between *x* and *y*, then the graph of the true equation would be a horizontal line (slope 0).
- If there really is a relation, then the slope will be nonzero.
- Therefore, to test of the existence of a relation, we test whether $\beta = 0$, using *b* as our estimator of β.
- To make a reasonable allowance for randomness, we must estimate the variability of the statistic *b*.
- Then, if *b* differs from 0 (which it most likely will), we can judge whether that difference is so great that it could not reasonably be attributed to randomness.
- The variability of *b* will be measured by the standard error of *b*, denoted SE_b.

Outline

The Regression Standard Error

- We will set up a procedure to test whether there really is a relation between x and y.
- The formulas for this procedure will require two additional standard errors.

The Regression Standard Error

- We will set up a procedure to test whether there really is a relation between x and y.
- The formulas for this procedure will require two additional standard errors.
- The regression standard error is

$$\text{SE}_{reg} = \sqrt{\frac{\sum (y - \hat{y})^2}{n - 2}}.$$

The Regression Standard Error

- We will set up a procedure to test whether there really is a relation between x and y.
- The formulas for this procedure will require two additional standard errors.
- The regression standard error is

$$\text{SE}_{reg} = \sqrt{\frac{\sum (y - \hat{y})^2}{n - 2}}.$$

- This measures the variability of y from the regression line.

The Regression Standard Error

- We will set up a procedure to test whether there really is a relation between x and y.
- The formulas for this procedure will require two additional standard errors.
- The regression standard error is

$$\mathrm{SE}_{reg} = \sqrt{\frac{\sum (y - \hat{y})^2}{n - 2}}.$$

- This measures the variability of y from the regression line.
- We will define the standard error of b soon.

Example

Example (Quiz Average vs. Test Average)

- In our example, we have the regression equation:

$$\hat{y} = 10.12 + 0.7905x.$$

- From this, we can calculate $\hat{y}$ for each (x, y) pair in the data set.
- Then we compute $(y - \hat{y})^2$ and then add them up to get

$$\sum (y - \hat{y})^2.$$

- Then calculate SE_{reg}.

Example

Example (Quiz Average vs. Test Average)

x	y	$\hat{y}$	$y - \hat{y}$	$(y - \hat{y})^2$
79	71			
45	45			
51	59			
81	62			
69	41			
75	79			
97	93			
97	76			
43	39			
67	71			
94	80			
77	84			
81	79			
91	81			
93	93			

Example

Example (Quiz Average vs. Test Average)

x	y	$\hat{y}$	$y - \hat{y}$	$(y - \hat{y})^2$
79	71	72.571		
45	45	45.696		
51	59	50.439		
81	62	74.152		
69	41	64.667		
75	79	69.410		
97	93	86.799		
97	76	86.799		
43	39	44.115		
67	71	63.086		
94	80	84.428		
77	84	70.990		
81	79	74.152		
91	81	82.057		
93	93	83.638		

Example

Example (Quiz Average vs. Test Average)

x	y	$\hat{y}$	$y - \hat{y}$	$(y - \hat{y})^2$
79	71	72.571	-1.571	
45	45	45.696	-0.696	
51	59	50.439	8.561	
81	62	74.152	-12.152	
69	41	64.667	-23.667	
75	79	69.410	9.590	
97	93	86.799	6.201	
97	76	86.799	-10.799	
43	39	44.115	-5.115	
67	71	63.086	7.914	
94	80	84.428	-4.428	
77	84	70.990	13.010	
81	79	74.152	4.848	
91	81	82.057	-1.057	
93	93	83.638	9.362	

Example

Example (Quiz Average vs. Test Average)

x	y	$\hat{y}$	$y - \hat{y}$	$(y - \hat{y})^2$
79	71	72.571	-1.571	
45	45	45.696	-0.696	
51	59	50.439	8.561	
81	62	74.152	-12.152	
69	41	64.667	-23.667	
75	79	69.410	9.590	
97	93	86.799	6.201	
97	76	86.799	-10.799	
43	39	44.115	-5.115	
67	71	63.086	7.914	
94	80	84.428	-4.428	
77	84	70.990	13.010	
81	79	74.152	4.848	
91	81	82.057	-1.057	
93	93	83.638	9.362	
			0.000	

Example

Example (Quiz Average vs. Test Average)

x	y	$\hat{y}$	$y - \hat{y}$	$(y - \hat{y})^2$
79	71	72.571	-1.571	2.469
45	45	45.696	-0.696	0.485
51	59	50.439	8.561	73.295
81	62	74.152	-12.152	147.677
69	41	64.667	-23.667	560.120
75	79	69.410	9.590	91.977
97	93	86.799	6.201	38.447
97	76	86.799	-10.799	116.628
43	39	44.115	-5.115	26.165
67	71	63.086	7.914	62.632
94	80	84.428	-4.428	19.608
77	84	70.990	13.010	169.248
81	79	74.152	4.848	23.501
91	81	82.057	-1.057	1.117
93	93	83.638	9.362	87.654
			0.000	

Example

Example (Quiz Average vs. Test Average)

x	y	$\hat{y}$	$y - \hat{y}$	$(y - \hat{y})^2$
79	71	72.571	-1.571	2.469
45	45	45.696	-0.696	0.485
51	59	50.439	8.561	73.295
81	62	74.152	-12.152	147.677
69	41	64.667	-23.667	560.120
75	79	69.410	9.590	91.977
97	93	86.799	6.201	38.447
97	76	86.799	-10.799	116.628
43	39	44.115	-5.115	26.165
67	71	63.086	7.914	62.632
94	80	84.428	-4.428	19.608
77	84	70.990	13.010	169.248
81	79	74.152	4.848	23.501
91	81	82.057	-1.057	1.117
93	93	83.638	9.362	87.654
			0.000	1421.022

Example

Example (Quiz Average vs. Test Average)

- Now calculate SE_{reg}:

Example

Example (Quiz Average vs. Test Average)

- Now calculate SE_{reg}:

$$\text{SE}_{reg} = \sqrt{\frac{\sum (y - \hat{y})^2}{n - 2}}$$

Example

Example (Quiz Average vs. Test Average)

- Now calculate SE_{reg}:

$$SE_{reg} = \sqrt{\frac{\sum(y - \hat{y})^2}{n - 2}}$$
$$= \sqrt{\frac{1421.022}{13}}$$

Example

Example (Quiz Average vs. Test Average)

- Now calculate SE_{reg}:

$$\text{SE}_{reg} = \sqrt{\frac{\sum(y-\hat{y})^2}{n-2}} = \sqrt{\frac{1421.022}{13}} = 10.455.$$

The Regression Standard Error

- The other standard error is the standard error of the slope, SE_b.
- The formula is

$$SE_b = \frac{SE_{reg}}{\sqrt{\sum (x - \overline{x})^2}}.$$

Example

Example (Quiz Average vs. Test Average)

- In out example, we calculated $\sum (x - \overline{x})^2 = 4586.0$.

Example

Example (Quiz Average vs. Test Average)

- In out example, we calculated $\sum (x - \overline{x})^2 = 4586.0$.
- We can now calculate SE_b:

Example

Example (Quiz Average vs. Test Average)

- In out example, we calculated $\sum(x - \overline{x})^2 = 4586.0$.
- We can now calculate SE_b:

$$\text{SE}_b = \frac{\text{SE}_{reg}}{\sqrt{\sum(x - \overline{x})^2}}$$

Example

Example (Quiz Average vs. Test Average)

- In out example, we calculated $\sum(x - \overline{x})^2 = 4586.0$.
- We can now calculate SE_b:

$$\text{SE}_b = \frac{\text{SE}_{reg}}{\sqrt{\sum(x - \overline{x})^2}} = \frac{10.455}{\sqrt{4586.0}}$$

Example

Example (Quiz Average vs. Test Average)

- In out example, we calculated $\sum (x - \overline{x})^2 = 4586.0$.
- We can now calculate SE_b:

$$\begin{aligned}\text{SE}_b &= \frac{\text{SE}_{reg}}{\sqrt{\sum (x - \overline{x})^2}} \\ &= \frac{10.455}{\sqrt{4586.0}} \\ &= 0.1544.\end{aligned}$$

Example

Example (Quiz Average vs. Test Average)

- In out example, we calculated $\sum(x - \overline{x})^2 = 4586.0$.
- We can now calculate SE_b:

$$\text{SE}_b = \frac{\text{SE}_{reg}}{\sqrt{\sum(x-\overline{x})^2}} = \frac{10.455}{\sqrt{4586.0}} = 0.1544.$$

- This number SE_b estimates the variability of our estimator b, which estimates the true slope of the linear relation.

Outline

The Testing Procedure

- Now we are ready to specify the testing procedure.
- The null hypothesis will state that no relation exists. That is, $\beta = 0$.
- Therefore, the null hypothesis is

$$H_0\colon \beta = 0.$$

- The alternative hypothesis is one of the following:

$$H_a\colon \beta \neq 0,$$
$$H_a\colon \beta < 0,$$
$$H_a\colon \beta > 0.$$

The Testing Procedure

- The test statistic is
$$t = \frac{b}{\text{SE}_b}$$
which measures by how many standard errors b deviates from 0.
- Because this is a t statistic, we use `tcdf()` to find the p-value and we need to know the number of degrees of freedom.
- The number of degrees of freedom is $n - 2$.

Outline

Testing the Significance of the Relation

- We will now perform this test to determine whether there is a relation between one's quiz average and one's test average.

Example (Quiz Average vs. Test Average)

(1) The hypotheses:

$$H_0 : \beta = 0$$
$$H_a : \beta \neq 0.$$

(2) Let $\alpha = 0.05$.

Testing the Significance of the Relation

Example (Quiz Average vs. Test Average)

(3) The test statistic is

$$t = \frac{b}{\text{SE}_b}.$$

(4) We have already found that

$$b = 0.7905,$$
$$\text{SE}_b = 0.1544.$$

So we calculate

$$t = \frac{0.7905}{0.1544} = 5.1198.$$

Testing the Significance of the Relation

Example (Quiz Average vs. Test Average)

(5) The p-value is

$$\begin{aligned} p\text{-value} &= 2 \times \texttt{tcdf(5.1198,E99,13)} \\ &= 2 \times 9.840 \times 10^{-5} \\ &= 1.968 \times 10^{-4}. \end{aligned}$$

(6) Reject H_0 and conclude that there really is a relation between one's quiz average and one's test average.

Outline

Another Example

- In the month of April (2015), is the afternoon high temperature predictable from the morning low temperature?

Another Example

- In the month of April (2015), is the afternoon high temperature predictable from the morning low temperature?

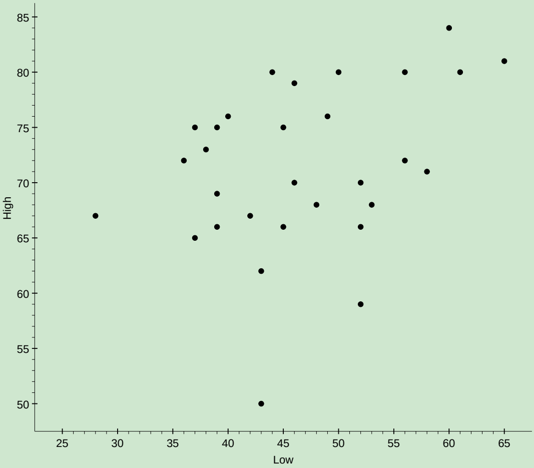
Date	Low	High
1	39	66
2	37	75
3	61	80
4	42	67
5	28	68
6	46	79
7	58	71
8	53	68
9	52	59
10	50	80

Date	Low	High
11	49	76
12	38	73
13	44	80
14	52	70
15	52	66
16	48	68
17	56	80
18	60	84
19	56	72
20	65	81

Date	Low	High
21	46	70
22	40	76
23	45	66
24	39	69
25	43	50
26	43	62
27	37	65
28	36	72
29	39	75
30	45	75

Example

Example (Morning Low vs. Afternoon High)



Example

Example (Morning Low vs. Afternoon High)

- We will test whether there is sufficient evidence, at the 5% level, indicating that there is a relation between the morning low temperature and the afternoon high temperature.

Example

Example (Morning Low vs. Afternoon High)

- We will test whether there is sufficient evidence, at the 5% level, indicating that there is a relation between the morning low temperature and the afternoon high temperature.
- Review the assumptions.

Example

Example (Morning Low vs. Afternoon High)

(1) The hypotheses are

$$H_0 : \beta = 0$$
$$H_a : \beta \neq 0$$

(2) Let $\alpha = 0.05$.

(3) The test statistic is

$$t = \frac{b}{\mathrm{SE}_b}.$$

Example

Example (Morning Low vs. Afternoon High)

(4) Enter the low temps in $\mathtt{L}_1$ and the high temps in $\mathtt{L}_2$.

(5) Use `LinRegTTest` on the TI-83.

(6) It reports

$$\begin{aligned} t &= 1.949, \\ p &= 0.0613, \text{ (i.e., } p\text{-value)} \\ df &= 28, \\ a &= 57.68, \\ b &= 0.2949, \\ s &= 7.052, \text{ ()i.e., } SE_{reg}) \\ r^2 &= 0.1195, \\ r &= 0.3457. \end{aligned}$$

Example

Example (Morning Low vs. Afternoon High)

(5) The p-value is 0.0613.

(6) We conclude that there is not (quite) sufficient evidence for a relation between the morning low and the afternoon high in April.

Outline

Assignment

Assignment

- Read Sections 26.3, 26.4, 26.5.
- Apply Your Knowledge: 2 (`LinRegTTest`), 4, 6.
- Check Your Skills: 19, 20, 21, 22.
- Exercises 25(c)(d), 26, 28, 31.