14.2 Testing for Significance

Objective of this section:

To test if the intercept β_0, the slope β_1 or both parameters are significant.

From now on, assume

$$\varepsilon_i \sim N\left(0, \sigma^2\right),$$

that is ε_i come from a normal population with mean 0 and variance σ^2. The residual for the i'th observation is

$$e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i), i = 1, \ldots, n.$$

Motivating Example (continue):

In the pizza example, the fitted values and the residuals are

x_i	2	6	8	8	12	16	20	20	22	26
y_i	58	105	88	118	117	137	157	169	149	202
$\hat{y}_i = 60 + 5x_i$	70	90	100	100	120	140	160	160	170	190
$e_i = y_i - \hat{y}_i$	-12	15	-12	18	-3	-3	-3	9	-21	12

◆

Intuitively, the residuals can be thought as good "estimate" of ε_i since

$$e_i = y_i - \hat{y}_i = \beta_0 + \beta_1 x_i + \varepsilon_i - b_0 - b_1 x_i = (\beta_0 - b_0) + (\beta_1 - b_1) x_i + \varepsilon_i \approx \varepsilon_i.$$

Thus,

$$\frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum_{i=1}^{n} (e_i - \bar{e})^2}{n-2} \approx \frac{\sum_{i=1}^{n} (\varepsilon_i - \bar{\varepsilon})^2}{n-1} \equiv \text{the sample variance estimate of } \sigma^2$$

The mean square error (estimate of σ^2) is

$$s^2 = MSE = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2}, SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Motivating Example (continue):

$$s^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{(-12)^2 + 15^2 + \cdots + (-21)^2 + 12^2}{10-2} = 191.25$$

◆

Important Results:

Let b_0 and b_1 be the sample values of the random variables B_0 and B_1. Then,

$$Var(B_0) = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n s_{XX}}, Var(B_1) = \frac{\sigma^2}{s_{XX}}.$$

Confidence Intervals of β_0 and β_1:

The $100(1-\alpha)\%$ **confidence interval for** β_0 **is**

$$b_0 \pm \left(t_{n-2,\alpha/2} \cdot s.e.(B_0) \right) = b_0 \pm \left[t_{n-2,\alpha/2} \cdot \left(\frac{s^2 \sum_{i=1}^{n} x_i^2}{n s_{XX}} \right)^{1/2} \right]$$

$$\equiv \left[b_0 - t_{n-2,\alpha/2} \cdot \left(\frac{s^2 \sum_{i=1}^{n} x_i^2}{n s_{XX}} \right)^{1/2}, b_0 + t_{n-2,\alpha/2} \cdot \left(\frac{s^2 \sum_{i=1}^{n} x_i^2}{n s_{XX}} \right)^{1/2} \right]$$

and **The** $100(1-\alpha)\%$ **confidence interval for** β_1 **is**

$$b_1 \pm \left(t_{n-2,\alpha/2} \cdot s.e.(B_1) \right) = b_1 \pm \left[t_{n-2,\alpha/2} \cdot \left(\frac{s^2}{s_{XX}} \right)^{1/2} \right]$$

$$\equiv \left[b_1 - t_{n-2,\alpha/2} \cdot \left(\frac{s^2}{s_{XX}} \right)^{1/2}, b_1 + t_{n-2,\alpha/2} \cdot \left(\frac{s^2}{s_{XX}} \right)^{1/2} \right]$$

Motivating Example (continue):

The 95% confidence interval for β_0 is

$$b_0 \pm \left[t_{n-2,\alpha/2} \cdot \left(\frac{s^2 \sum_{i=1}^{n} x_i^2}{n s_{XX}} \right)^{1/2} \right] = 60 \pm \left(t_{8,0.025} \cdot \left(\frac{191.25 \cdot 2528}{10 \cdot 568} \right)^{1/2} \right) = 60 \pm (2.306 \cdot 9.226)$$

$$\equiv [38.72, 81.27]$$

and the 95% confidence interval for β_1 is

$$b_1 \pm \left[t_{n-2,\alpha/2} \cdot \left(\frac{s^2}{s_{XX}} \right)^{1/2} \right] = 5 \pm \left(t_{8,0.025} \cdot \left(\frac{191.25}{568} \right)^{1/2} \right) = 5 \pm (2.306 \cdot 0.58) \equiv [3.66, 6.33]$$

t-tests for β_0 and β_1:

To test $H_0 : \beta_0 = 0 \ v.s. \ H_a : \beta_0 \neq 0$,

$$t = \frac{b_0}{s.e.(B_0)} = \frac{b_0}{\left(\frac{s^2 \sum_{i=1}^{n} x_i^2}{n s_{XX}} \right)^{1/2}},$$

$$|t| > t_{n-2,\alpha/2} \Rightarrow \text{reject } H_0$$

$$|t| \leq t_{n-2,\alpha/2} \Rightarrow \text{not reject } H_0 .$$

$$\text{p-value} = P\left(|T_{n-2}| > |t| \right).$$

To test $H_0 : \beta_1 = 0 \ v.s. \ H_a : \beta_1 \neq 0$,

$$t = \frac{b_1}{s.e.(B_1)} = \frac{b_1}{\left(\frac{s^2}{s_{XX}} \right)^{1/2}},$$

$$|t| > t_{n-2,\alpha/2} \Rightarrow \text{reject } H_0$$

$$|t| \le t_{n-2,\alpha/2} \Rightarrow \text{not reject } H_0 \cdot$$

$$\text{p - value} = P\left(|T_{n-2}| > |t|\right).$$

Motivating Example (continue):

Assume $\alpha = 0.05$. To test $H_0 : \beta_0 = 0$ v.s. $H_a : \beta_0 \ne 0$,

$$t = \frac{b_0}{\left(\frac{s^2 \sum_{i=1}^{n} x_i^2}{n s_{XX}}\right)^{1/2}} = \frac{60}{\left(\frac{191.25 \cdot 2528}{10 \cdot 568}\right)^{1/2}} = \frac{60}{9.226} = 6.50,$$

$$|t| = 6.50 > 2.306 = t_{8,0.025} = t_{n-2,\alpha/2}$$

Therefore, we reject H_0. Also,

$$\text{p - value} = P\left(|T_{n-2}| > |t|\right) = P\left(|T_8| > |6.5|\right) = 0.0002,$$

we reject H_0 based on p-value.

To test $H_0 : \beta_1 = 0$ v.s. $H_a : \beta_1 \ne 0$,

$$t = \frac{b_1}{\left(\frac{s^2}{s_{XX}}\right)^{1/2}} = \frac{5}{\left(\frac{191.25}{568}\right)^{1/2}} = \frac{5}{0.58} = 8.61,$$

$$|t| = 8.61 > 2.306 = t_{8,0.025} = t_{n-2,\alpha/2}$$

Therefore, we reject H_0. Also,

$$\text{p - value} = P\left(|T_{n-2}| > |t|\right) = P\left(|T_8| > |8.61|\right) \approx 0,$$

we reject H_0 based on p-value.

Note:

The confidence intervals can also be used to test the above hypotheses. Since $0 \notin [38.72, 81.27]$**, we reject** $H_0 : \beta_0 = 0$ **. Similarly,** $0 \notin [3.66, 6.33]$**,**

we reject $H_0 : \beta_1 = 0$ **.**

The Analysis of Variance (F-test) :

$$H_0 : \beta_1 = 0 \ v.s. \ H_a : \beta_1 \neq 0$$

We have the following 2 models:

Horizontal: $y = \beta_0 + \varepsilon \quad \Rightarrow \hat{y} = \bar{y}$

Line : $y = \beta_0 + \beta_1 x + \varepsilon \quad \Rightarrow \hat{y} = b_0 + b_1 x$

Note:

The object function for the model 1 is $S(\beta_0) = \sum_{i=1}^{n} (y_i - \beta_0)^2$ **. Thus, the estimate of the parameter** β_0 **can be obtained by solving** $\frac{\partial S(\beta_0)}{\partial \beta_0} = 0$**.**

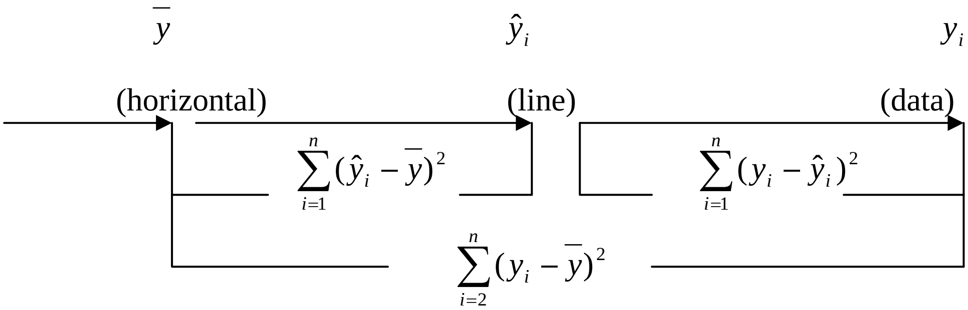
$\bar{y}$ **is the solution.** $\hat{y} = \bar{y}$**.**

Fundamental Equation:

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$\Leftrightarrow$

("distance" between data and horizontal line)=
("distance" between data and line)+("distance" between model line and horizontal line) .



[Derivation of Fundamental Equation]:

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(y_i-\hat{y}_i+\hat{y}_i-\bar{y})^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+2\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})$$

$$=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$$

since

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})=\sum_{i=1}^{n}\left[y_i-(\bar{y}+b_1(x_i-\bar{x}))\right]\left[\bar{y}+b_1(x_i-\bar{x})-\bar{y}\right]$$

$$=\sum_{i=1}^{n}\left[(y_i-\bar{y})-b_1(x_i-\bar{x})\right]b_1(x_i-\bar{x})$$

$$=b_1\left\{\sum_{i=1}^{n}\left[(y_i-\bar{y})-b_1(x_i-\bar{x})\right](x_i-\bar{x})\right\}$$

$$=b_1\left[\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})-b_1\sum_{i=1}^{n}(x_i-\bar{x})^2\right]=b_1(s_{XY}-b_1s_{XX})$$

$$=b_1(s_{XY}-\frac{s_{XY}}{s_{XX}}s_{XX})=b_1(s_{XY}-s_{XY})=0$$

The ANOVA (Analysis of Variance) table corresponding to the fundamental equation:

Source	df	SS	MS
Residual (Error)	*n-2*	$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$	$\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-2}$
Due to regression	*1*	$\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$	$\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$
Total (corrected)	*n-1*	$\sum_{i=1}^{n}(y_i-\bar{y})^2$	

Let

$$f=\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 \Big/ 1}{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \Big/ n-2}=\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{s^2},$$

the ratio of the mean sum of square due to the regression and mean residual sum of square. Intuitively, large *F* value might imply the difference between the line and the horizontal line is relatively large to the random variation reflected by the mean residual sum of square. That is, β_1 is so significant such that the difference between the line and the horizontal line are apparent. Therefore, the F value can provide important information about if $H_0:\beta_1=0$.

Next question to ask: **how large value of *F* can be considered to be large?**

To test $H_0:\beta_1=0 \; v.s. \; H_a:\beta_1\neq 0$ **,**

$$f > f_{1,n-2,\alpha} \Rightarrow \text{ reject } H_0$$
$$f \le f_{1,n-2,\alpha} \Rightarrow \text{ not reject } H_0$$

Note:

The sum of square due to the regression and the mean sum of square due to regression are

$$MSR = \frac{SSR}{1} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \, .$$

Thus, the *f* statistic is

$$f = \frac{MSR}{MSE} \, .$$

Note:

For ease of computation, the following equations can be used:

$$MSR = SSR = b_1 s_{XY} = b_1^2 s_{XX} \, .$$

Note:

Let *t* be the statistic for testing $H_0:\beta_1=0 \; v.s. \; H_a:\beta_1\neq 0$ **.**

Then,

$$f = t^2 .$$

Motivating Example (continue):

Assume $\alpha = 0.05$. To test $H_0 : \beta_1 = 0$ v.s. $H_a : \beta_1 \neq 0$, we have the following:

$$b_1 = 5, s_{XY} = 2840 , SSR = b_1 s_{XY} = 5 \cdot 2840 = 14200 ,$$

$$\sum_{i=1}^{10} (y_i - \bar{y})^2 = \sum_{i=1}^{10} y_i^2 - 10 \bar{y}^2 = 184730 - 10 \cdot 130^2 = 15730$$

$$\Rightarrow SSE = \sum_{i=1}^{10} (y_i - \bar{y})^2 - SSR = 15730 - 14200 = 1530$$

Thus, we have the following ANOVA table

Source	**df**	**SS**	**MS**	**f**
Residual (Error)	**n-2=8**	**SSE=1530**	$MSE = \frac{SSE}{8} = 191.25$	$f = \frac{MSR}{MSE} = \frac{14200}{191.25} = 74.1$
Regression	**1**	**SSR=14200**	$MSR = \frac{SSR}{1} = 14200$	
Total	**9**	**15730**		

Since

$$f = 74.1 > 5.32 = f_{1,8,0.05} ,$$

we reject $H_0 : \beta_1 = 0$. Note that

$$f = 74.1 = (8.61)^2 = t^2 .$$

Example 2 (continue):

Suppose the model is

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, \ldots, 20, \varepsilon_i \sim N(0, \sigma^2),$$

and

$$\sum_{i=1}^{20} x_i = 1330, \sum_{i=1}^{20} y_i = 1862.8, \sum_{i=1}^{20} x_i^2 = 90662,$$

$$\sum_{i=1}^{20} y_i^2 = 173554.26, \sum_{i=1}^{20} x_i y_i = 124206.9$$

(a)Provide an ANOVA table.

(b) Find the 95% confidence interval for β_1.and use the confidence interval to test $H_0 : \beta_1 = 0$.

(c) Use F statistic to test $H_0 : \beta_0 = \beta_1 = 0$ at $\alpha = 0.05$.

[solution:]

(a)

Since

$$s_{YY} = SST = \sum_{i=1}^{20} y_i^2 - 20 \cdot \bar{y}^2 = 173554.26 - 20 \cdot \left(\frac{1862.8}{20}\right)^2 = 53.06$$

$$SSR = b_1 s_{XY} = 0.149 \cdot 330.7 = 49.220$$

$$SSE = SST - SSR = 53.06 - 49.220 = 3.848$$

The ANOVA table is

Source	df	SS	MS
Residual (Error)	**n-2=18**	**SSE=3.848**	$MSE = \frac{SSE}{18} = 0.214$
Regression	**1**	**SSR=49.220**	$MSR = \frac{SSR}{1} = 49.220$
Total	**19**	**53.068**	

(b) The 95% confidence interval for β_1 is

$$b_1 \pm t_{n-2,\alpha/2}\left(\frac{s^2}{s_{XX}}\right)^{1/2} = 0.149 \pm t_{18,0.025} \cdot \left(\frac{0.214}{2217}\right)^{1/2} \equiv [0.128, 0.170].$$

Since $0 \notin [0.128, 0.170]$, we reject $H_0 : \beta_1 = 0$.

(c)

Since $\sum_{i=1}^{20} \hat{y}_i^2 = \sum_{i=1}^{20} y_i^2 - \sum_{i=1}^{20} (y_i - \hat{y}_i)^2 = 173554.26 - 3.848 = 173550.412$ and

$$f = \frac{\sum_{i=1}^{20} \hat{y}_i^2 \Big/ 2}{s^2} = \frac{86775.206}{0.214} = 405491.6 > 3.5546 = f_{2,18,0.05},$$

we reject $H_0 : \beta_0 = \beta_1 = 0$.

Online Exercise:

Exercise 14.2.1