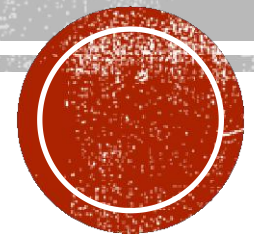
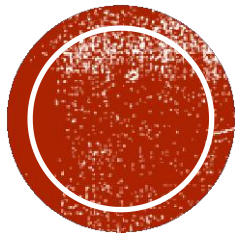
GOODNESS-OF-FIT AND TESTS FOR THE SLOPE





1. OVERVIEW OF LR AND LEAST SQUARE MEANS

Overview

In *simple linear regression*, we will study how *to relate* an *outcome* variable y to one *predictor* variable x, where x can be accurately measured/observed, but y may also depend on *unobserved error* e.

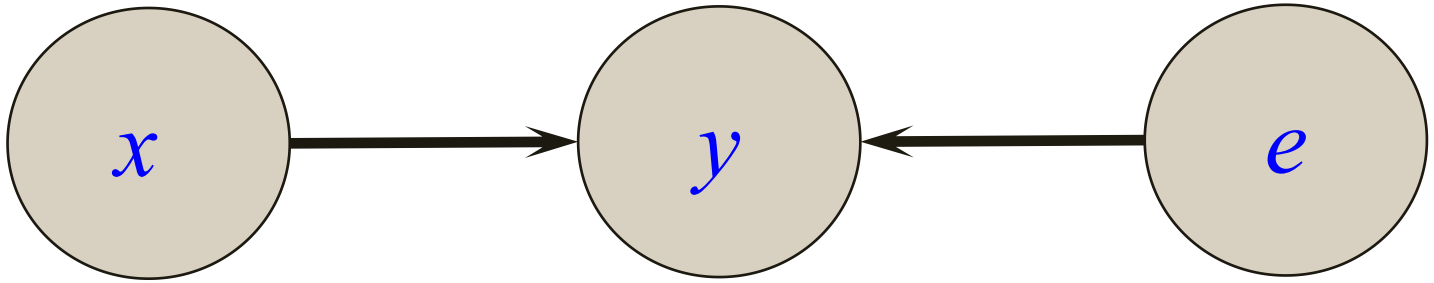


Fig. 1: Outcome y, predictor x and random error e.

EXAMPLE 1 (OBSTETRICS)

Obstetricians sometimes order tests to measure **estriol levels (x)** from 24-hour urine specimens taken from pregnant women who are near term, because level of estriol has been found to be related to infant **birthweight (y).**

The test can provide *indirect evidence* of an abnormally small fetus. The **relationship between estriol level and birthweight** can be quantified by fitting a ***regression line*** that relates the two variables.

Greene and Touchstone conducted a study to relate *birthweight* and *estriol level* in pregnant women. **Table 1** contains the actual data points. **Figure 1** is a plot of the data points of 31 pregnant women from the study.

Table 1: Sample data from the greene-touchstone study relating *birthweight* and *estriol level* in pregnant women near term.

i	Estriol (mg/24 hr) x_i	Birthweight (g/100) y_i	i	Estriol (mg/24 hr) x_i	Birthweight (g/100) y_i	i	Estriol (mg/24 hr) x_i	Birthweight (g/100) y_i
1	7	25	12	19	31	22	15	35
2	9	25	13	21	30	23	16	35
3	9	25	14	24	28	24	19	34
4	12	27	15	15	32	25	18	35
5	14	27	16	16	32	26	17	36
6	16	27	17	17	32	27	18	37
7	16	24	18	25	32	28	20	38
8	14	30	19	27	34	29	22	40
9	16	30	20	15	34	30	25	39
10	16	31	21	15	34	31	24	43
11	17	30						

Source: From Rosner, Bernard. Fundamentals of Biostatistics. 7th Edition, 2011 Duxbury, Brooks/Cole, Cengage Learning. Page 433.

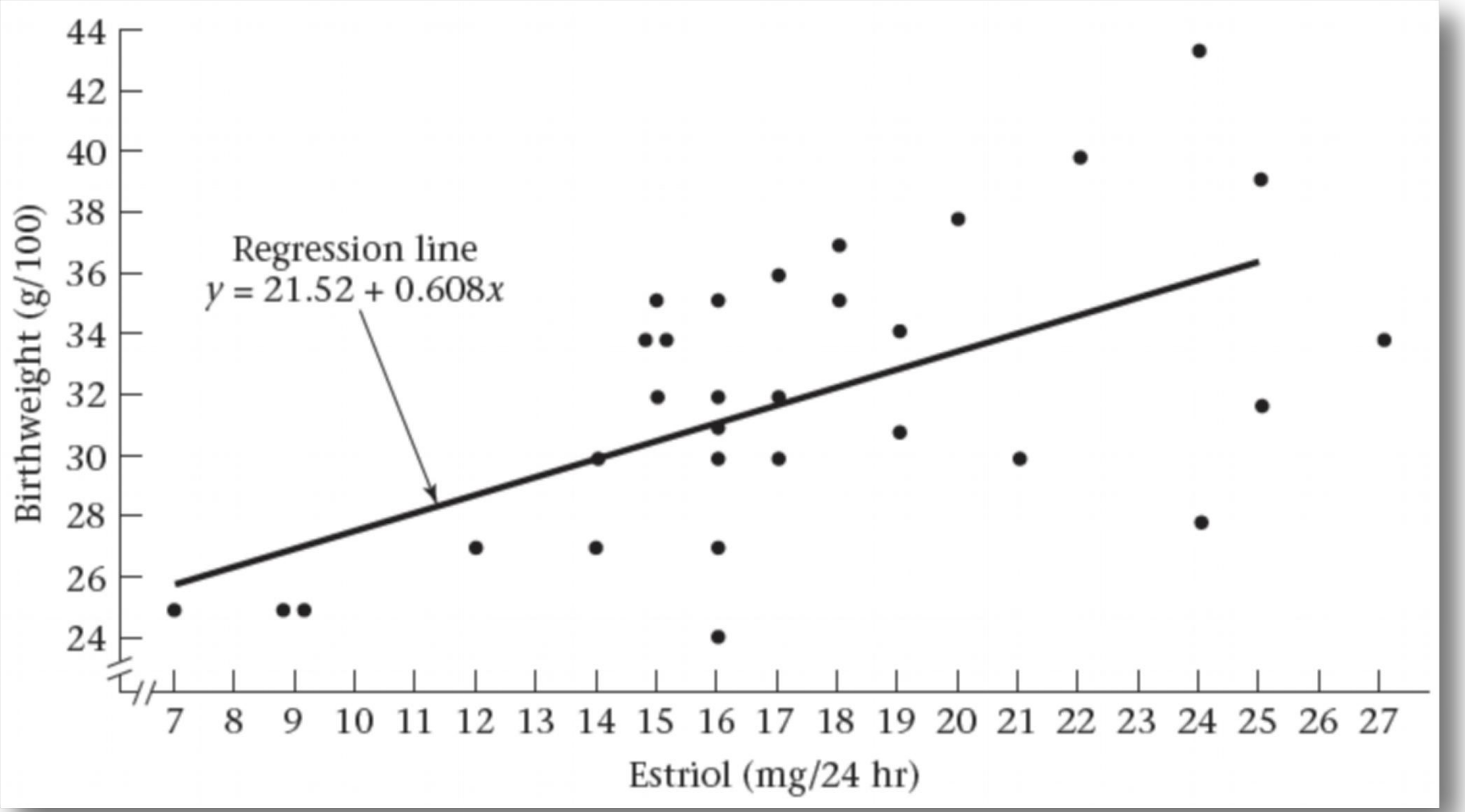


FIGURE 1: Data from the greene-touchstone study relating *birthweight* (y) and *estriol level* (x) in pregnant women near term.

Source: From Rosner, Bernard. *Fundamentals of Biostatistics*. 7th Edition, 2011 Duxbury, Brooks/Cole, Cengage Learning. Page 433.

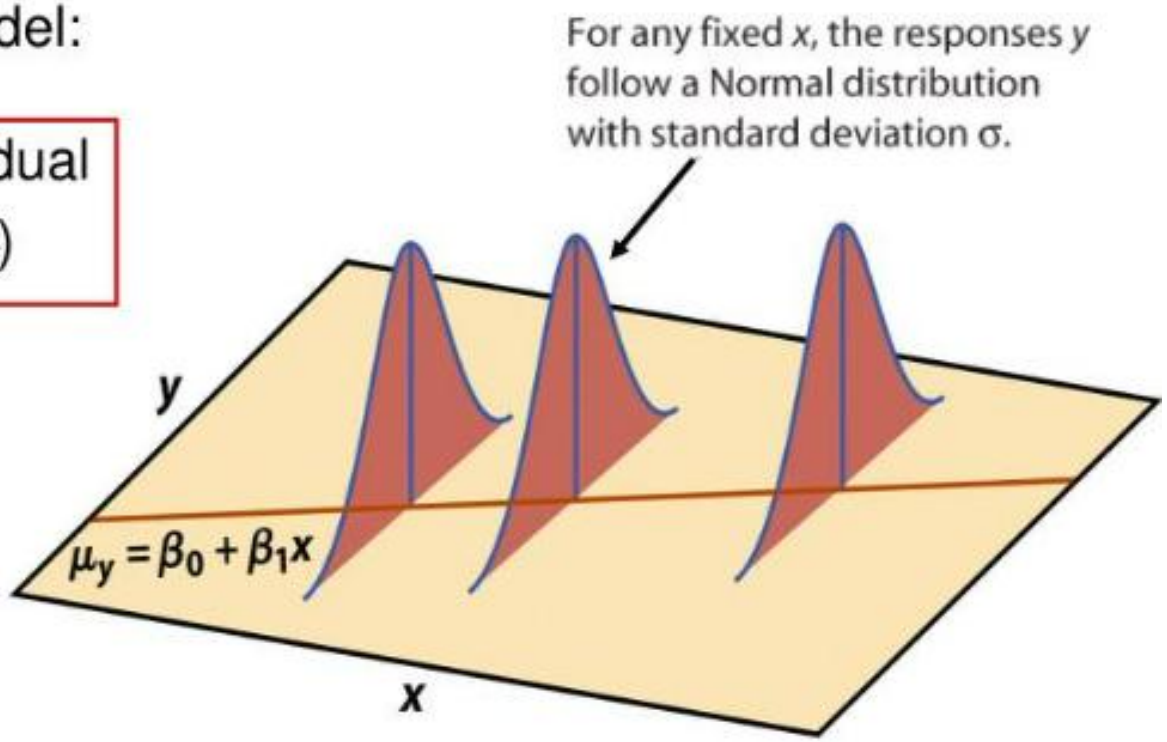
Simple linear regression model

In the population, the linear regression equation is $E(y) = \beta_0 + \beta_1 x$.

Sample data then fits the model:

$$\text{Data} = \boxed{\text{fit}} + \boxed{\text{residual}}$$

$$y_i = \boxed{(\beta_0 + \beta_1 x_i)} + \boxed{(\varepsilon_i)}$$



where the ε_i are **independent** and **Normally** distributed $N(0,\sigma)$.

Linear regression assumes **equal standard deviation of *y*** (σ is the same for all values of x).

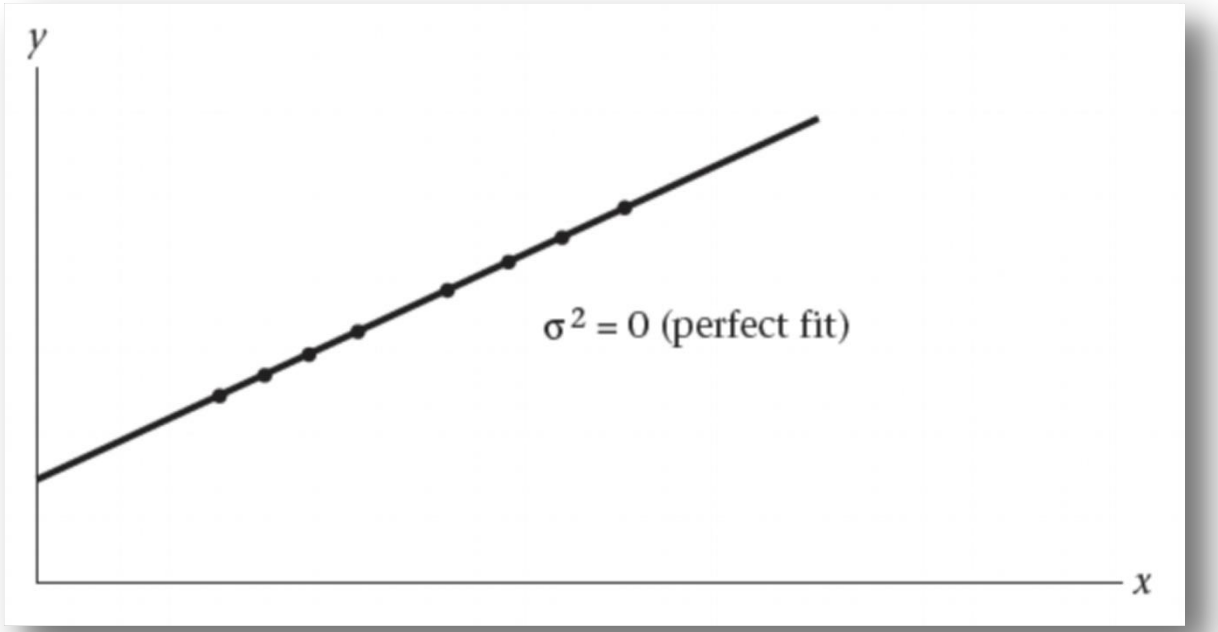
FULL LINEAR REGRESSION MODEL

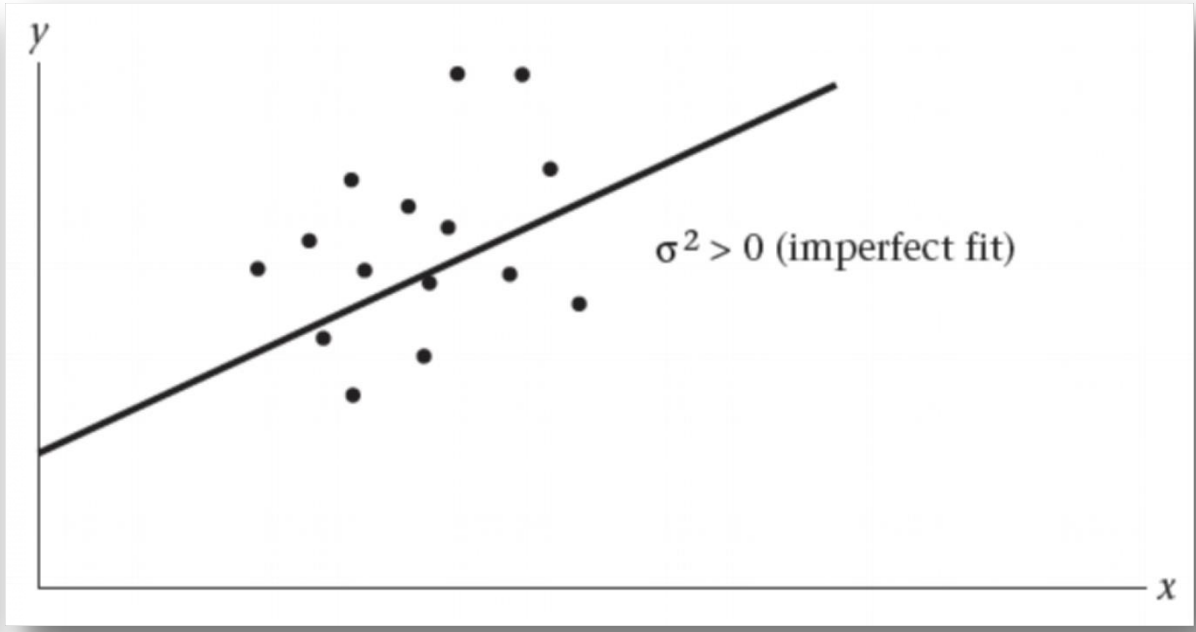
The ***full linear regression model*** then takes the following form: $y = \beta_0 + \beta_1 x + e$, where $e \sim N(0, \sigma^2)$, β_0, β_1 and σ^2 are unknown parameters.

For any linear regression model of the form $y = \beta_0 + \beta_1 x + e$, *y is called the* ***dependent variable*** *and x is called the* ***independent variable****, because we are trying to predict y as a function of x.*

Our tasks: (i) estimating the unknowns; (ii) performing significance tests for the regression, and (iii) predicting y for a new x value.

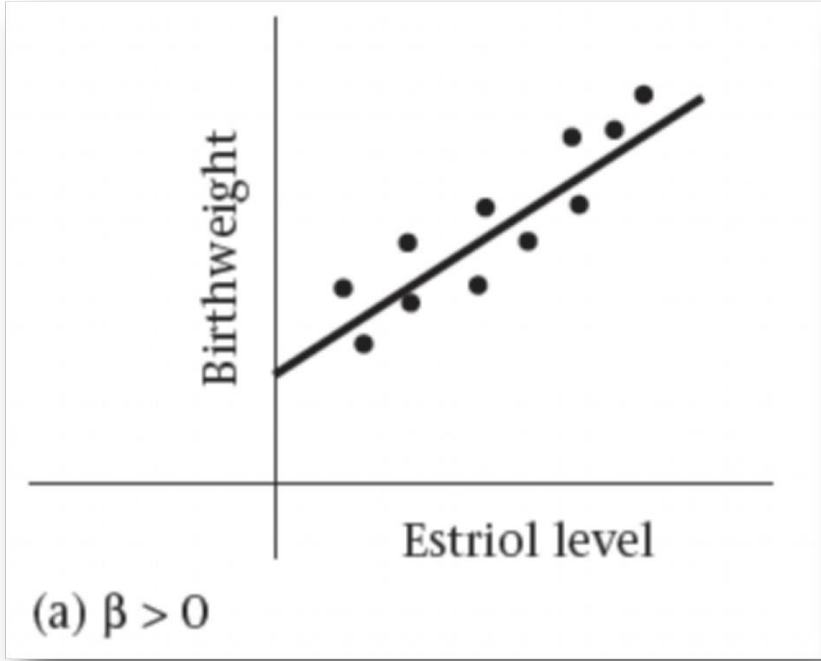
$$y|x \sim N(\beta_0 + \beta_1 x, \sigma^2).$$



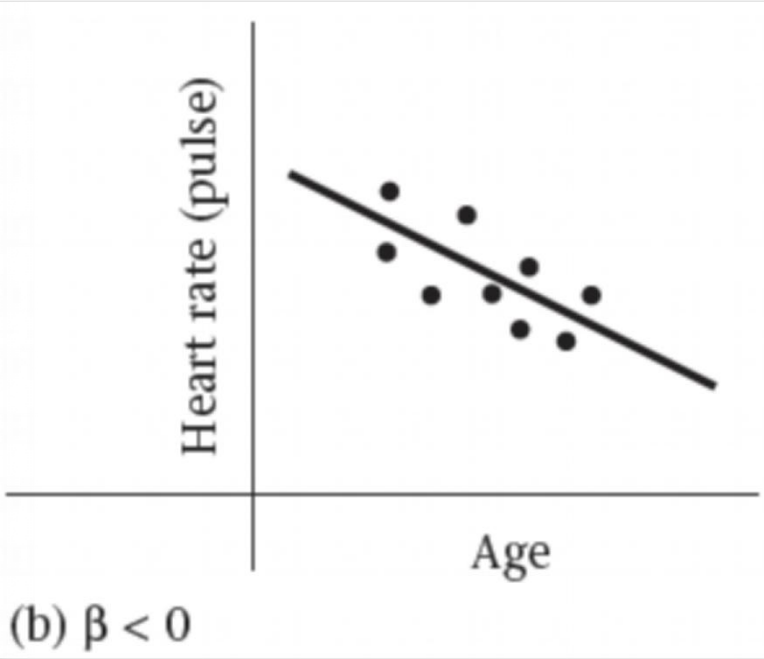


Perfect fit

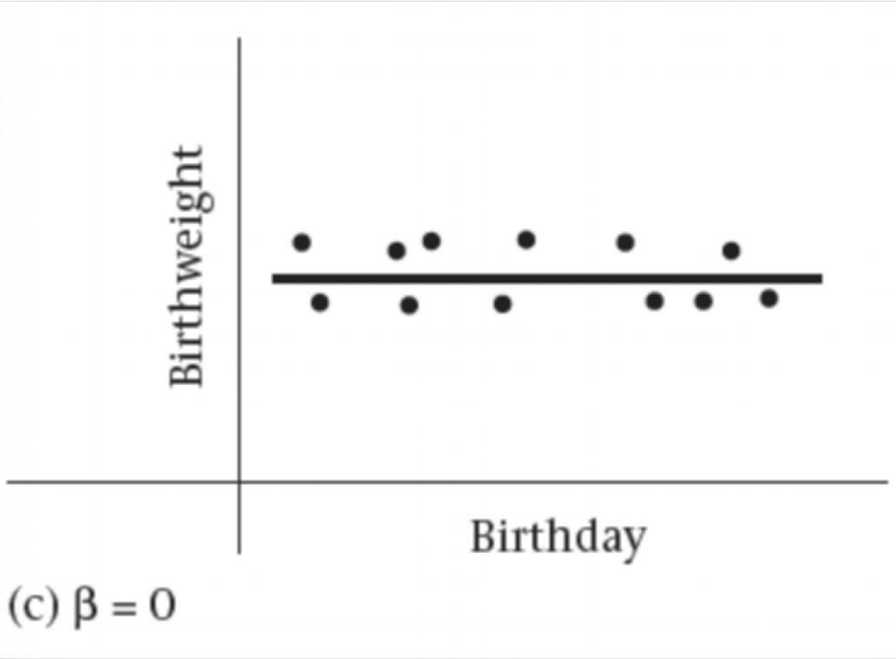
Imperfect fit



(a) $\beta > 0$



(b) $\beta < 0$



(c) $\beta = 0$

Fitting Regression Lines

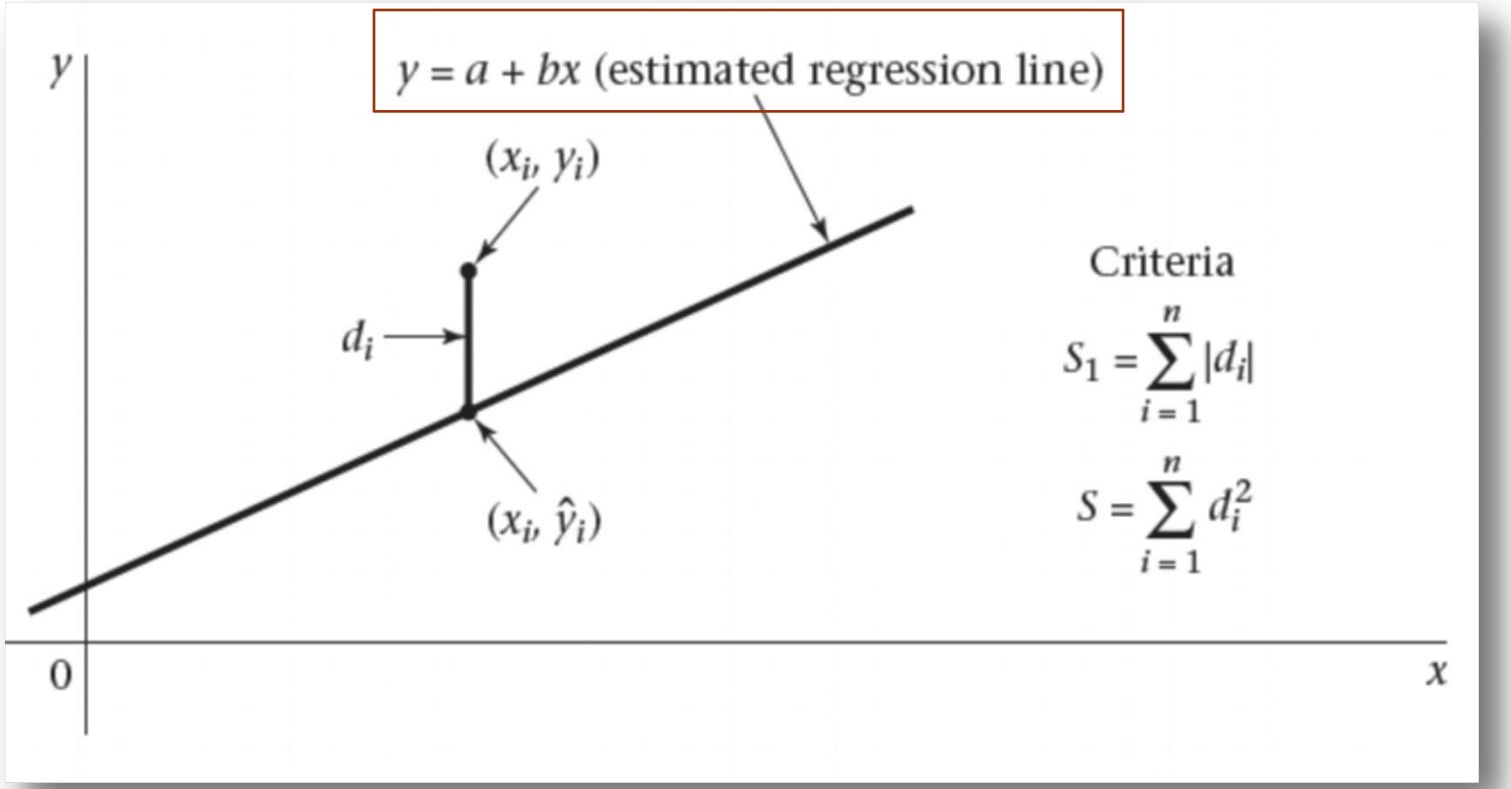


Fig 2: **Least square** criterion for judging the fit of a regression line.

Let $(x_i, \hat{y}_i) = (x_i, \hat{\alpha} + \hat{\beta}x_i)$ be the point on the estimated regression line at x_i, then this distance is given by

$$\boldsymbol{d_i = y_i - \hat{y}_i = y_i - \hat{a} - \hat{\beta}x_i.}$$

A good-fitting line would make S, the **sum of the squared distances**, as small as possible.

LEAST SQUARES ESTIMATES

Let $\bar{x} = \sum_{i=1}^{n} x_i/n = (x_1 + \ldots + x_n)/n$,

$$L_{x,x} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

$$= \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n},$$

and

$$L_{x,y} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

$$= \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}.$$

- Least Squares Estimates of $\hat{a}$ and $\hat{\beta}$ are

$$\hat{a} = \bar{y} - \bar{x}\hat{\beta} \text{ and } \hat{\beta} = \frac{L_{x,y}}{L_{x,x}},$$

respectively.

The point $(\bar{x}, \bar{y})$ always falls on the regression line!

- One unbiased estimator of σ^2 is given by

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - \hat{a} - \hat{\beta}x_i\right)^2.$$

- The ***Least Squares Line*** is given by

$$y = a + bx.$$

EXAMPLE 2: COMPUTE THE REGRESSION LINE IN EXAMPLE 1

STEP 1: COMPUTE THE MOMENTS

$$\sum_{i=1}^{31} x_i = 534,$$

$$\sum_{i=1}^{31} x_i^2 = 9876,$$

$$\bar{x} = \frac{534}{31} = 17.22851,$$

$$\sum_{i=1}^{31} y_i = 992,$$

$$\sum_{i=1}^{31} y_i^2 = 32419,$$

$$\bar{y} = \frac{992}{31} = 32,$$

$$\sum_{i=1}^{31} x_i y_i = 17500.$$

STEP 2: COMPUTE THE LS

$$L_{x,x} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$
$$= 9876 - 31\left(\frac{534}{31}\right)^2$$
$$= 677.41935,$$

$$L_{x,y} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$
$$= 17500 - 31\left(\frac{534}{31}\right)\left(\frac{992}{31}\right)$$
$$= 412.$$

Step 3: Compute the betas

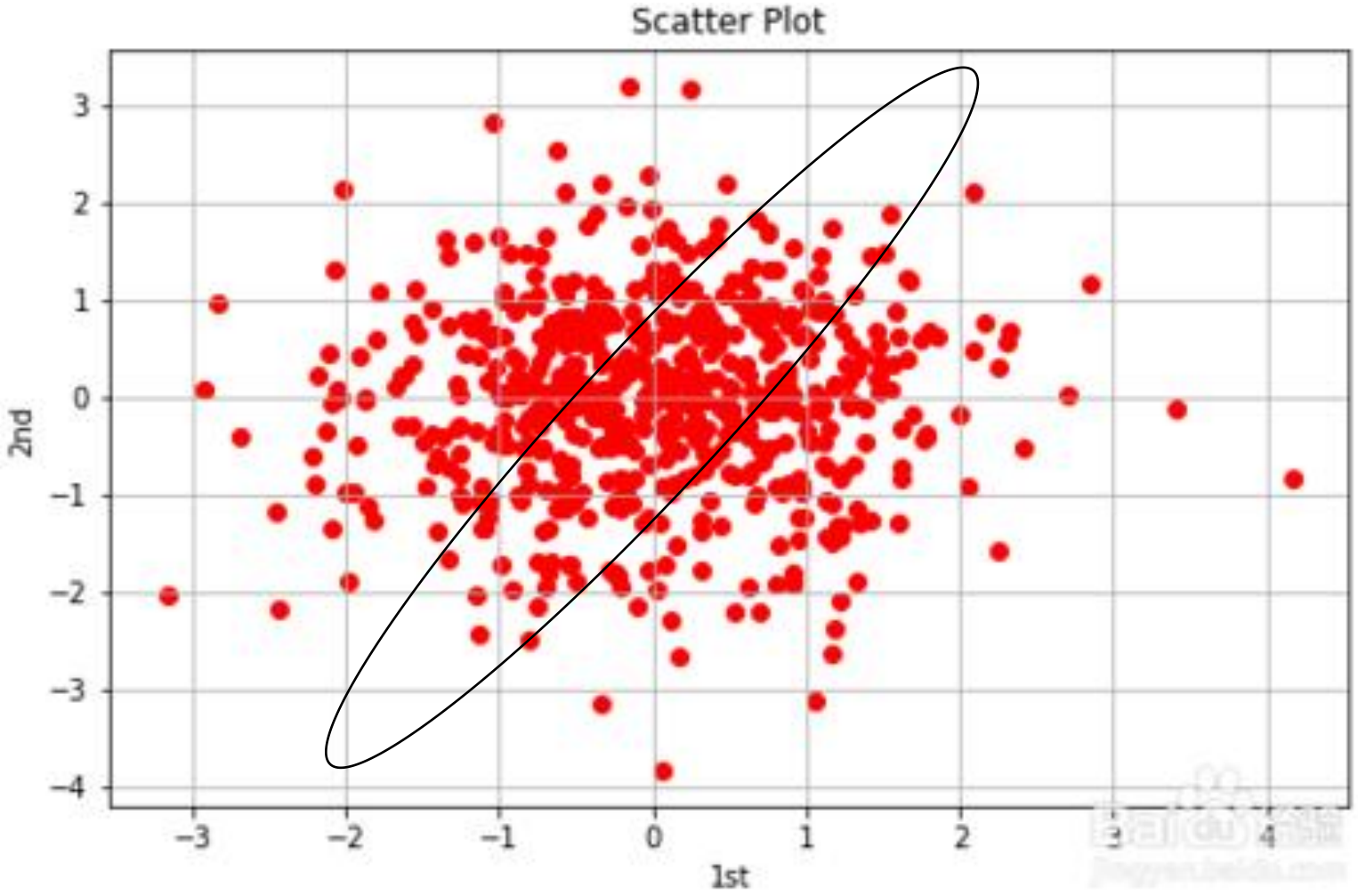
$$\hat{\beta} = \frac{L_{x,y}}{L_{x,x}} = \frac{412}{677.41935} = 0.60819,$$

$$\hat{a} = \bar{y} - \hat{\beta}\bar{x} = 32 - (0.60819)(17.2258) = 21.5234,$$

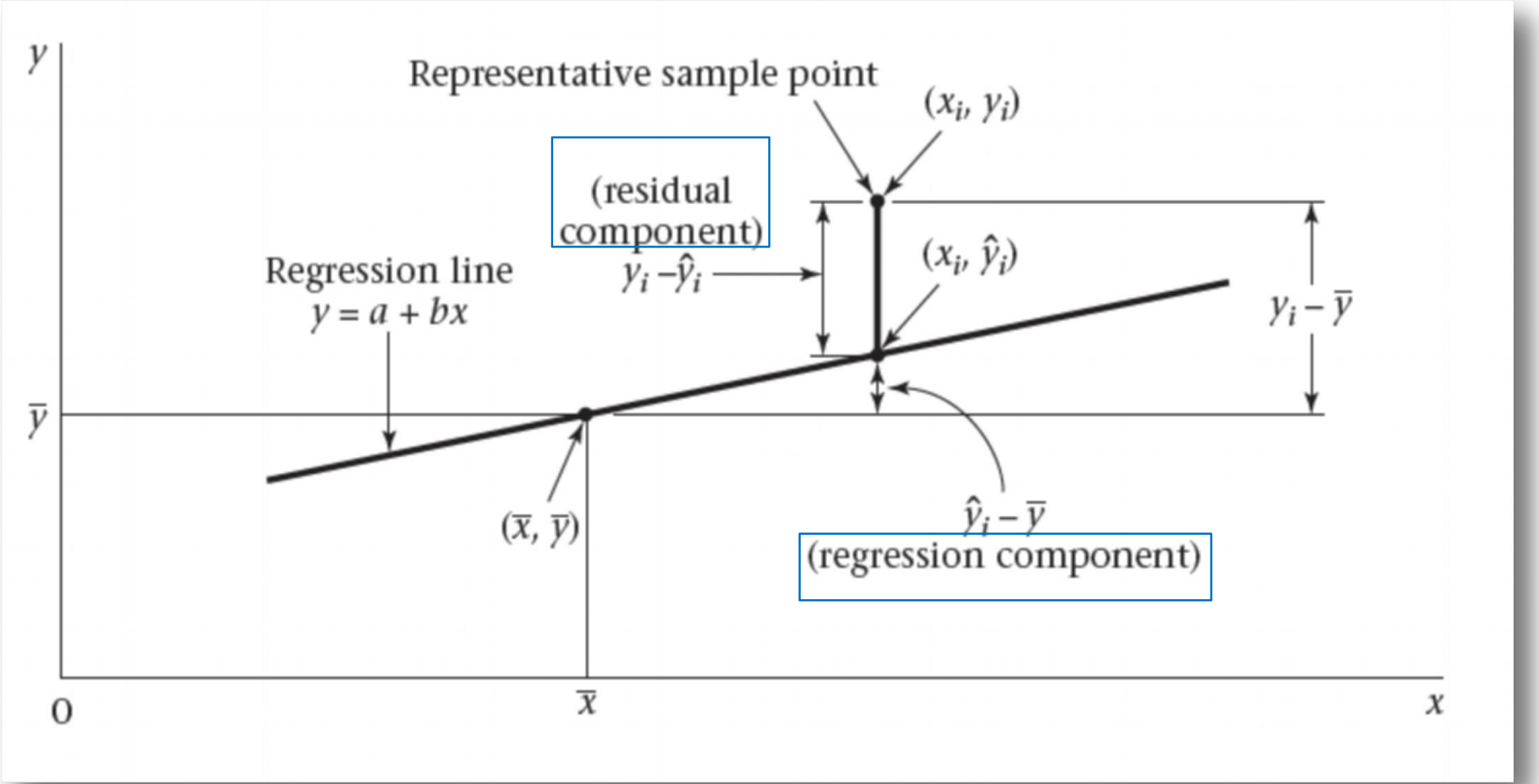
Regression line: $y = 21.5234 + 0.60819x$.

2. Inferences about parameters from regression lines

SAMPLING ERROR



2.1. Goodness-of-fit



The deviation $y_i - \bar{y}$ can be separated into residual $y_i - \hat{y}_i$ and regression $\hat{y}_i - \bar{y}$ components:

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}).$$

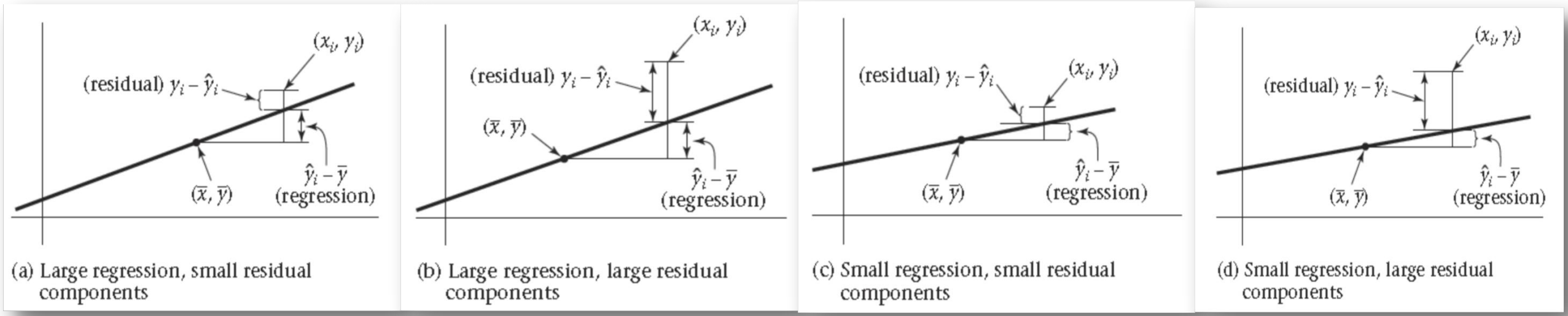


Fig. 3: regression lines with varying goodness-of-fit relationships

(A). The best situation for goodness of fit; B). An intermediate situation for goodness of fit

(C) Another intermediate situation for goodness of fit (d). The worst situation for goodness of fit

2.2. SOURCES OF VARIATION

- **Total Sum of Squares:**

$$\text{TotSS} = L_{y,y} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

measures the total amount of variation in the y data.

- **Regression Sum of Squares:**

$$\text{RegSS} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \hat{\beta}_1 L_{x,y}$$

measures the amount of total variation in the y data that *is accounted for* by the fitted model.

- **Residual Sum of Squares:**

$$\begin{aligned}\text{ResSS} &= \text{TotSS} - \text{RegSS} \\ &= L_{y,y} - \hat{\beta}_1 L_{x,y} \\ &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2\end{aligned}$$

measures the amount of total variation in the y data that is not accounted for by the fitted model.

- We have decomposition:

$$\textbf{Total SS} = \textbf{Reg SS} + \textbf{Res SS.}$$

- **Coefficient of Determination** is defined as

$$R^2 = \frac{\text{RegSS}}{\text{TotalSS}}$$

- It is the proportion of total variation in the y data that is accounted for by the fitted model.

- Clearly, $0 \leq R^2 \leq 1$. Larger R^2 implies better fit to the data.

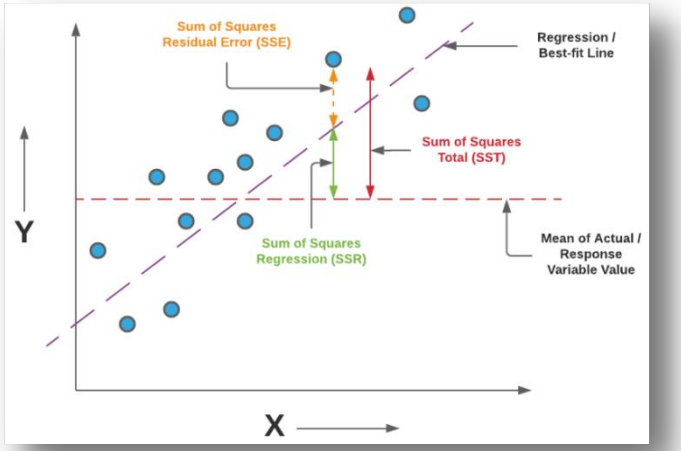
2.3. F TEST: (ANOVA)

H_0: $\beta = 0$ vs. H_1: $\beta \neq 0$

Table 2: Analysis of variance (ANOVA)

Source	DF	Sum of Squares	Mean Squares	F
Regression	1	RegSS	$\text{Reg MS} = \frac{\text{RegSS}}{1}$	$F = \frac{\text{RegMS}}{\text{ResMS}}$.
Residual	$n-2$	ResSS	$\text{ResMS} = \frac{\text{ResSS}}{n-2}$	Let $e \sim N(0, \sigma^2)$. If H_0 is true, then $F \sim F_{1,n-2}$.
Total	$n-1$	Total SS	Decision rule: If $F > F_{1-\alpha,1,n-2}$, then reject H_0.	

- *F* statistic:



$$F = \frac{(\text{RegSS})/1}{(\text{ResSS})/(n-2)} = \frac{\text{RegMS}}{\text{ResMS}}.$$

- Null distribution:If $H_0: \beta_1 = 0$ is true, then $F \sim F_{1,n-2}$, the centralized *F* distribution with $(1, n-2)$ degrees of freedom.

Short Computational Form for Regression and Residual SS

$$\text{Regression SS} = bL_{xy} = b^2 L_{xx} = L_{xy}^2 / L_{xx}$$

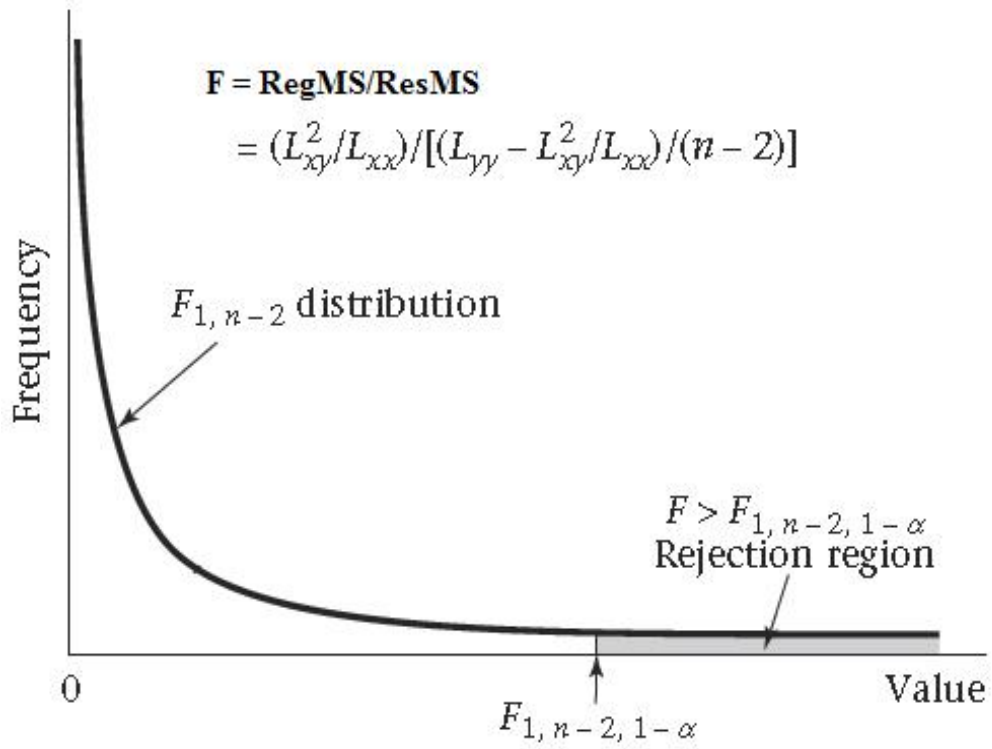
$$\text{Residual SS} = \text{Total SS} - \text{Regression SS} = L_{yy} - L_{xy}^2 / L_{xx}$$

DECISION RULE

Reject H_0 at nominal level $\alpha \in (0,1)$ if $F > F_{1,n-2,1-\alpha}$ **(Fig. 4).**

For example, when $\alpha = 0.05$ and $n = 31$, we have critical value $F_{1,n-2,1-\alpha} = F_{0.95,1,29} = 4.18$.

Figure 4: Rejection region for the simple linear-regression F test



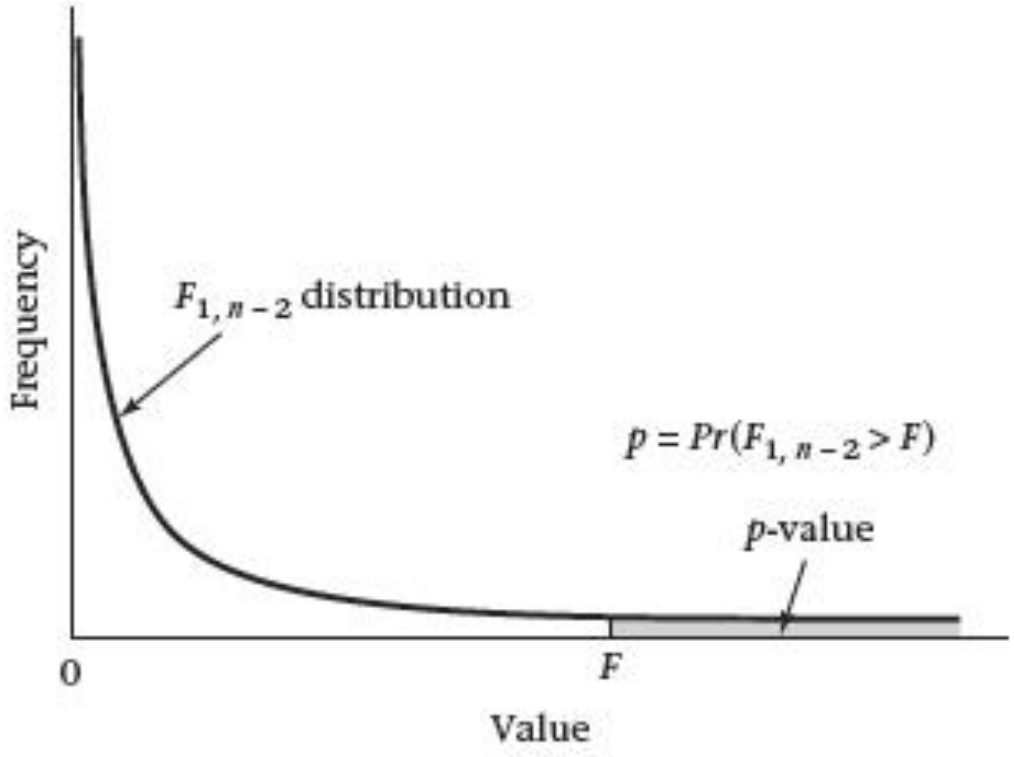
P-VALUE OF THE F STATISTIC

The exact p-value (**Fig. 5**) is given by

$$p = \Pr\left(F_{1,n-2} > F\right).$$

Reject H_0 at nominal level $\alpha \in (0,1)$ if $p < \alpha$.

Figure 5: Computation of the p-value for the simple linear-regression F test



- Recall $n = 31, L_{x,x} = 677.41935, L_{x,y} = 412, \widehat{\beta}_1 = 0.60819.$

We compute:

$$\text{Total SS} = L_{y,y} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

$$= 32418 - 31(32)^2 = 674,$$

$$\text{Reg SS} = \widehat{\beta}_1 L_{x,y}$$

$$= (0.60819)(412) = 250.57428.$$

Therefore,

$$R^2 = \frac{\text{Reg SS}}{\text{Total SS}} = \frac{250.57428}{674} = 0.37177.$$

- That is, 37.18% of the *total variation* in birth weight is accounted for by the linear relationship with estriol level.

- **For *F* test, we compute:**

$$\text{ResSS} = \text{TotalSS} - \text{RegSS} = 674 - 250.57428 = 423.42572,$$

and

$$\text{ResMS} = \frac{\text{Res SS}}{n-2} = \frac{423.42572}{29} = 14.60089.$$

Since

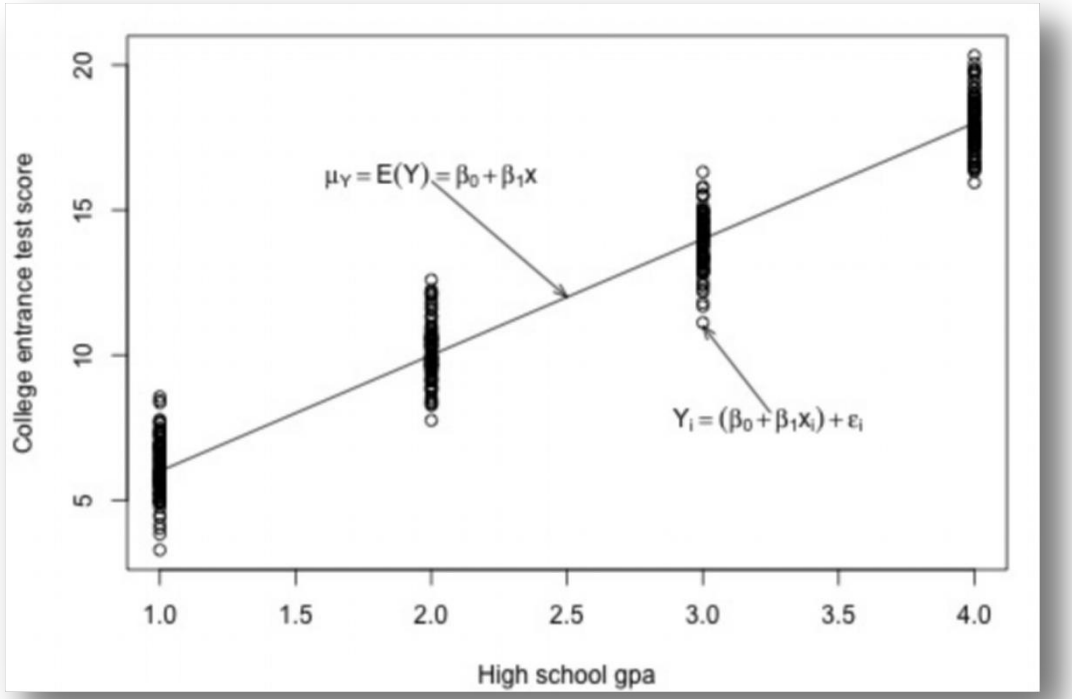
$$\text{Reg MS} = \text{Reg SS} = 250.57428.$$

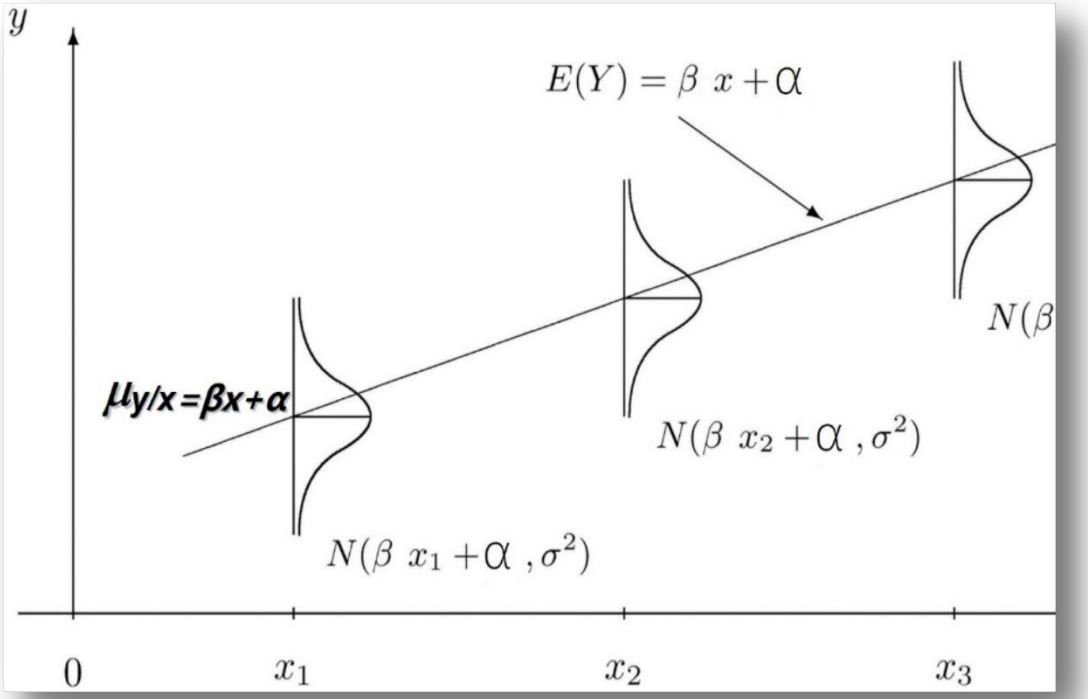
We have

$$F = \frac{\text{Reg MS}}{\text{Res MS}} = \frac{250.57428}{14.60089} = 17.16.$$

- Since the realized $F = 17.16 > 4.18$, the data provide sufficient evidence to reject the null hypothesis at significant level 0.05.

2.4. EQUIVALENT T TEST





$$\beta = \mu|_{X=X+1} - \mu|_{X=X}$$

proof

To understand why we use the t-distribution, you need to know what is the underlying distribution of $\widehat{\beta}$ and of the Residual sum of squares (RSS) as these two put together will give you the t-distribution.

The easier part is the distribution of $\widehat{\beta}$ which is a normal distribution - to see this note that $\widehat{\beta} = (X^T X)^{-1} X^T Y$ so it is a linear function of Y where $Y \sim N(X\beta, \sigma^2 I_n)$. As a result it is also normally distributed, $\widehat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1})$ - let me know if you need help deriving the distribution of $\widehat{\beta}$.

Additionally, $RSS \sim \sigma^2 \chi^2_{n-p}$, where n is the number of observations and p is the number of parameters used in your regression. The proof of this is a bit more involved, but also straightforward to derive (see proof here Why is RSS distributed chi square times n-p?).

Up until this point I have considered everything in matrix/vector notation, but let's for simplicity use $\widehat{\beta}_i$ and use its normal distribution which will give us:

$$\frac{\widehat{\beta}_i - \beta_i}{\sigma\sqrt{(X^T X)^{-1}_{ii}}} \sim N(0,1)$$

Additionally, from the chi-squared distribution of RSS we have that:

$$\frac{(n-p)s^2}{\sigma^2} \sim \chi^2_{n-p}$$

This was simply a rearrangement of the first chi-squared expression and is independent of the $N(0,1)$. Additionally, we define $s^2 = \frac{RSS}{n-p}$, which is an unbiased estimator for σ^2. By the definition of the t_{n-p} definition that dividing a normal distribution by an independent chi-squared (over its degrees of freedom) gives you a t-distribution (for the proof see: A normal divided by the $\sqrt{\chi^2(s)/s}$ gives you a t-distribution -- proof) you get that:

$$\frac{\widehat{\beta}_i - \beta_i}{s\sqrt{(X^T X)^{-1}_{ii}}} \sim t_{n-p}$$

Where $s\sqrt{(X^T X)^{-1}_{ii}} = SE(\widehat{\beta}_i)$.

- t statistic:

$$t = \frac{|b|}{se(b)} = \frac{|\widehat{\beta}|}{\sqrt{\frac{\text{ResMS}}{L_{x,x}}}}.$$

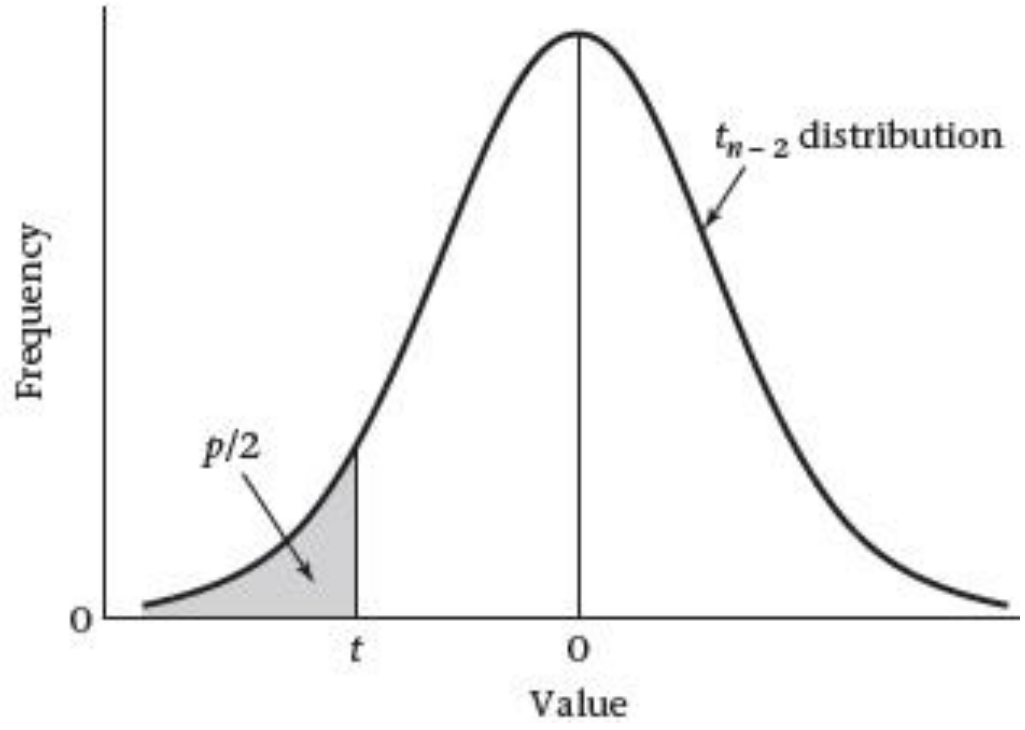
- Null distribution:If $H_0: \beta = 0$ is true, then $t \sim t_{n-2}$, the centralized t distribution with $n-2$ degrees of freedom.

P-VALUE OF A NEGATIVE T VALUE

If $t < 0$, the exact p-value (**Fig. 5a**) is given by

$p = 2 \times$(area to the left of t under a t_{n-2} distribution).

Fig. 5a: Computation of the p-value of a negative t value.

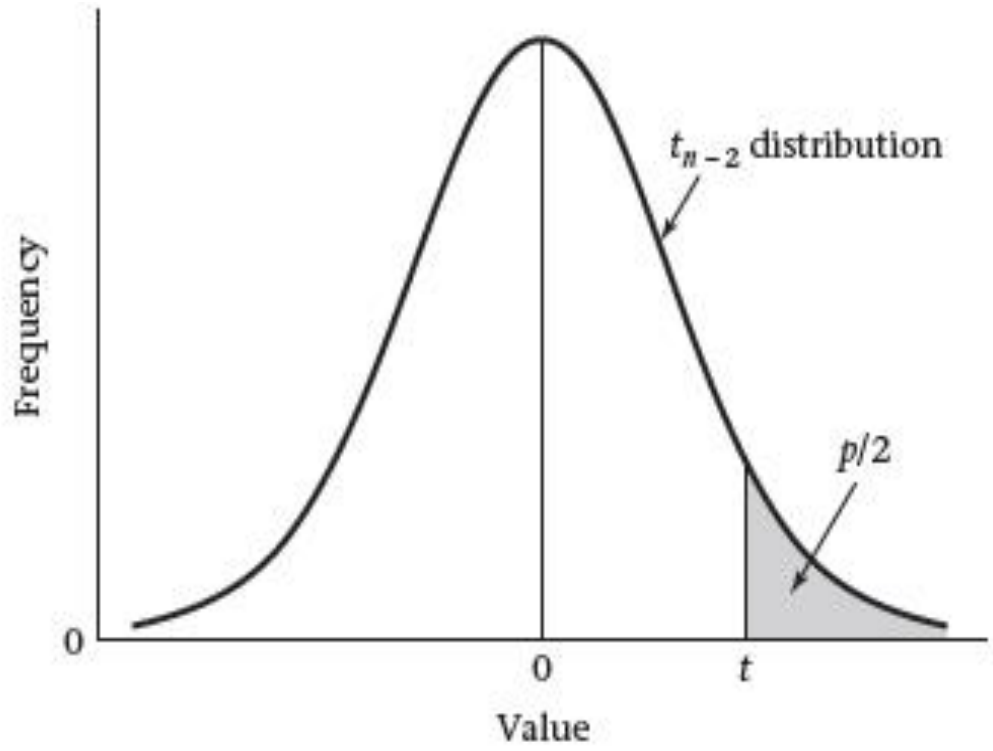


P-VALUE OF A POSITIVE *T* VALUE

If $t > 0$, the exact p-value (**Fig. 5b**) is given by

$p = 2 \times$ (area to the right of t under a t_{n-2} distribution).

Fig. 5b: Computation of the p-value of a positive t value.

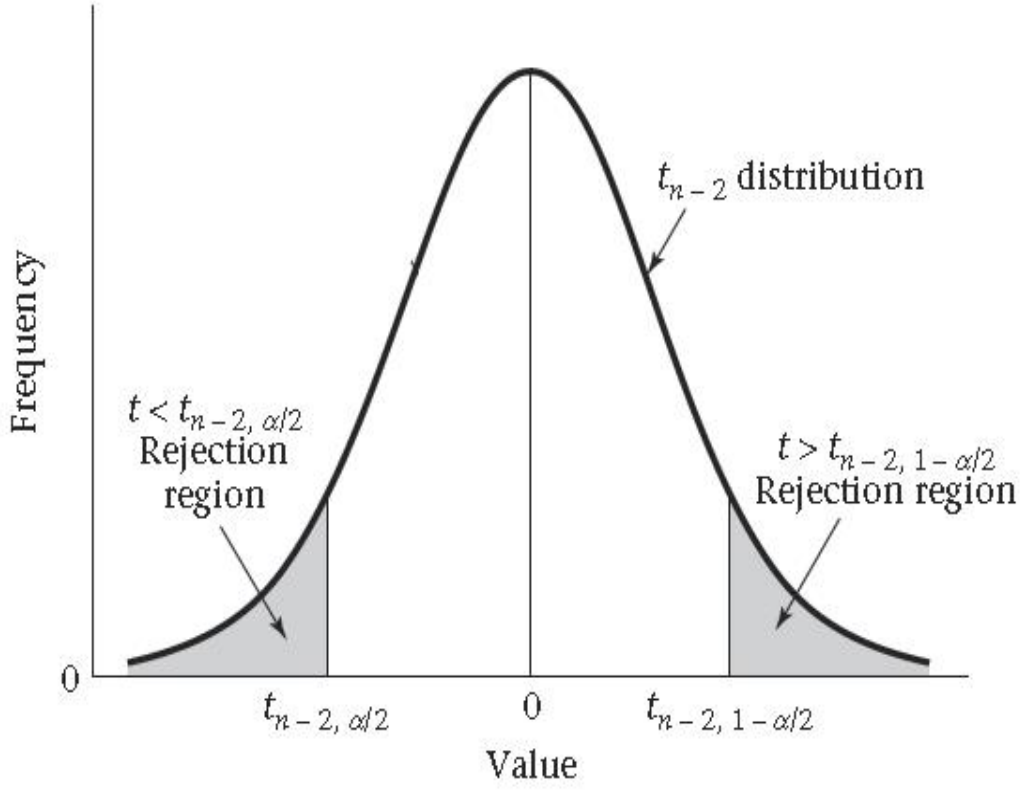


DECISION RULE

Reject H_0 at nominal level $\alpha \in (0,1)$ if $|t| > t_{1-\alpha/2,n-2}$ **(Fig. 6)**.

For example, when $\alpha = 0.05$ and $n = 31$, we have critical value $t_{1-\alpha/2,n-2} = t_{0.975,29} = 2.045$.

Fig. 6: Rejection regions for the t test for slope.



T TEST

- Recall $n = 31, L_{x,x} = 677.41935, L_{x,y} = 412, \widehat{\beta}_1 = 0.60819$ and $\hat{y} = 21.5234 + 0.60819x$.

$$t = \frac{\widehat{\beta}_1}{\sqrt{\frac{\text{Res MS}}{L_{x,x}}}} = \frac{0.60819}{\sqrt{\frac{14.60089}{677.41935}}} = 4.1426$$

- See! $t^2 = (4.1426)^2 = 17.16 = F$. Since the realized $t = 4.14 > 2.045$, the data provide sufficient evidence to reject the null hypothesis at significant level 0.05.

AFTER CLASS CHALLENGES (REQUESTED/OPTIONAL)

Under the full model $y = \alpha + \beta x + e$ with $e \sim N(0, \sigma^2)$, *mathematically prove* the following statements:

(i) b ($\widehat{\beta}$) is an unbiased estimator of β ;

(ii) $\widehat{\sigma}^2 = \text{ResMS}$ is an unbiased estimator of σ^2;

(iii) Statistic $t^2 = F$ and $t^2_{n-2,1-\alpha/2} = F_{1,n-2,1-\alpha}$.

HINTS

- **Proposition:** $t^2 = F$ and $t^2_{1-\alpha/2,n-2} = F_{1-\alpha,1,n-2}$.
- *Proof:*

(1) $t^2 = \dfrac{\hat{\beta}\hat{\beta}}{\dfrac{\text{Res MS}}{L_{x,x}}}$

$$= \frac{\hat{\beta}\dfrac{L_{x,y}}{L_{x,x}}}{\dfrac{\text{Res MS}}{L_{x,x}}}$$

$$= \frac{\hat{\beta}L_{x,y}}{\text{Res MS}}$$

$$= \frac{\text{Reg MS}}{\text{Res MS}} = F.$$

(2) Let $t \sim t_{n-2}$. Then $t^2 \sim F_{1,n-2}$. (Review the definition of t and F distributions). Since

$$\left(|t| > t_{1-\frac{\alpha}{2},n-2}\right) \Leftrightarrow \left(t^2 > t^2_{1-\alpha/2,n-2}\right),$$

we have

$$\Pr\left(t^2 > t^2_{1-\alpha/2,n-2}\right) = \Pr\left(|t| > t_{1-\alpha/2,n-2}\right).$$

- By the definition of $t_{1-\alpha/2,n-2}$, we observe that

$$\Pr\left(|t| > t_{1-\frac{\alpha}{2},n-2}\right) = \Pr\left(t > t_{1-\frac{\alpha}{2},n-2} \text{ or } t < -t_{1-\frac{\alpha}{2},n-2}\right)$$

$$= \Pr\left(t > t_{1-\frac{\alpha}{2},n-2}\right) + \Pr\left(t < -t_{1-\frac{\alpha}{2},n-2}\right)$$
$$= \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha.$$

- Therefore, $\Pr\left(t^2 > t^2_{1-\alpha/2,n-2}\right) = \alpha.$

- Since $t^2 \sim F_{1,n-2}$, we have $t^2_{1-\alpha/2,n-2} = F_{1-\alpha,\,1,\,n-2}$.