The Coefficient of Determination

- The error sum of squares SSE can be interpreted as a measure of how much variation in y is left unexplained by the model—that is, how much cannot be attributed to a linear relationship.
- In (a), SSE $= 0$, and there is no unexplained variation, whereas unexplained variation is small for the data of (b) and much larger in (c).
- A quantitative measure of the total amount of variation in observed y values is given by the **total sum of squares**

$$SST = S_{yy} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \left(\sum y_i\right)^2 / n$$

The Coefficient of Determination

- The **coefficient of determination**, denoted by r^2, is given by

$$r^2 = 1 - \frac{SSE}{SST}$$

- It is interpreted as the proportion of observed y variation that can be explained by the simple linear regression model (attributed to an approximate linear relationship between y and x).
- r^2 is always between 0 and 1.
- The higher the value of r^2, the more successful is the simple linear regression model in explaining y variation.
- If r^2 is small, an analyst will usually want to search for an alternative model that can more effectively explain y variation.

Example Cont'd

The scatter plot of the iodine value-cetane number data in previous example portends a reasonably high r^2 value.

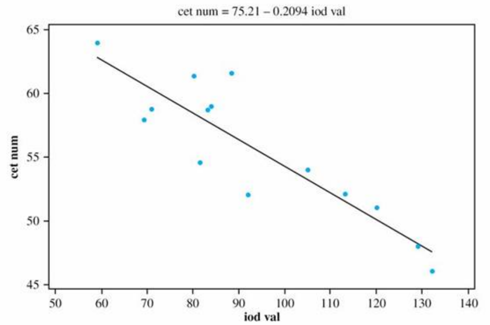


Figure: Scatter plot for data with least square line superimposed.

Example Cont'd

With

$$\sum x_i = 1307.5, \quad \sum y_i = 779.2,$$
$$\sum x_i^2 = 128913.93, \quad \sum x_i y_i = 71347.30, \quad \sum y_i^2 = 43745.22$$

we have

$$\hat{\beta}_0 = 75.212432 \quad \hat{\beta}_1 = -0.20938742$$

Further

$$SST = 43745.22 - (779.2)^2/14 = 377.174$$
$$SSE = 43745.22 - (75.212432)(779.2) - (-0.20938742)(71347.30) = 78.920$$

The coefficient of determination is then

$$r^2 = 1 - SSE/SST = 1 - (78.920)/(377.174) = 0.791$$

That is, 79.1% of the observed variation in cetane number can be explained by the simple linear regression relationship between cetane number and iodine value.

The Regression Sum of Squares

The coefficient of determination can be written in a slightly different way by introducing a third sum of squares—**regression sum of squares**, SSR—iven by

$$SSR = \sum(\hat{y}_i - \bar{y})^2 = SST - SSE.$$

Regression sum of squares is interpreted as the amount of total variation that is explained by the model.

Then we have

$$r^2 = 1 - SSE/SST = (SST - SSE)/SST = SSR/SST$$

the ratio of explained variation to total variation.

S1211Q Introduction to Statistics
Lecture 22

Wei Wang

August 7, 2012

Inferential Problems concerning β_1

- Based on Least Squares method, we have the estimator of slope β_1, $\hat{\beta}_1$, but this doesn't answer some of the most important inferential problems.

Inferential Problems concerning β_1

- Based on Least Squares method, we have the estimator of slope β_1, $\hat{\beta}_1$, but this doesn't answer some of the most important inferential problems.
- Is $\hat{\beta}_1$ unbiased?

Inferential Problems concerning β_1

- Based on Least Squares method, we have the estimator of slope β_1, $\hat{\beta}_1$, but this doesn't answer some of the most important inferential problems.
- Is $\hat{\beta}_1$ unbiased?
- What's the (estimated) standard error?

Inferential Problems concerning β_1

- Based on Least Squares method, we have the estimator of slope β_1, $\hat{\beta}_1$, but this doesn't answer some of the most important inferential problems.
- Is $\hat{\beta}_1$ unbiased?
- What's the (estimated) standard error?
- How to get Confidence Interval of β_1?

Inferential Problems concerning β_1

- Based on Least Squares method, we have the estimator of slope β_1, $\hat{\beta}_1$, but this doesn't answer some of the most important inferential problems.
- Is $\hat{\beta}_1$ unbiased?
- What's the (estimated) standard error?
- How to get Confidence Interval of β_1?
- How to perform Hypothesis Test and get *P*-value about null hypothesis $H_0 : \beta_1 = 0$

Sampling Distribution of $\hat{\beta}_1$

- The least squares estimator $\hat{\beta}_1$ is an unbiased estimator, which mean that $E(\hat{\beta}_1) = \beta_1$.
- Also we have shown yesterday that the variance of this estimator is σ^2/S_{xx}. The estimated standard error is $s_{\hat{\beta}_1} = \frac{s}{\sqrt{S_{xx}}}$.
- In particular, under the assumption that the noise terms are normally distributed, the $\hat{\beta}_1$ is also normally distributed

$$\hat{\beta}_1 \sim N(\beta_1, \sigma^2/S_{xx})$$

Confidence Interval of β_1

- The way to build confidence interval for β_1 is the classical procedure, standardizing the estimator by subtracting its mean and then dividing by its estimated standard error.
- It turns out that the standardized variable

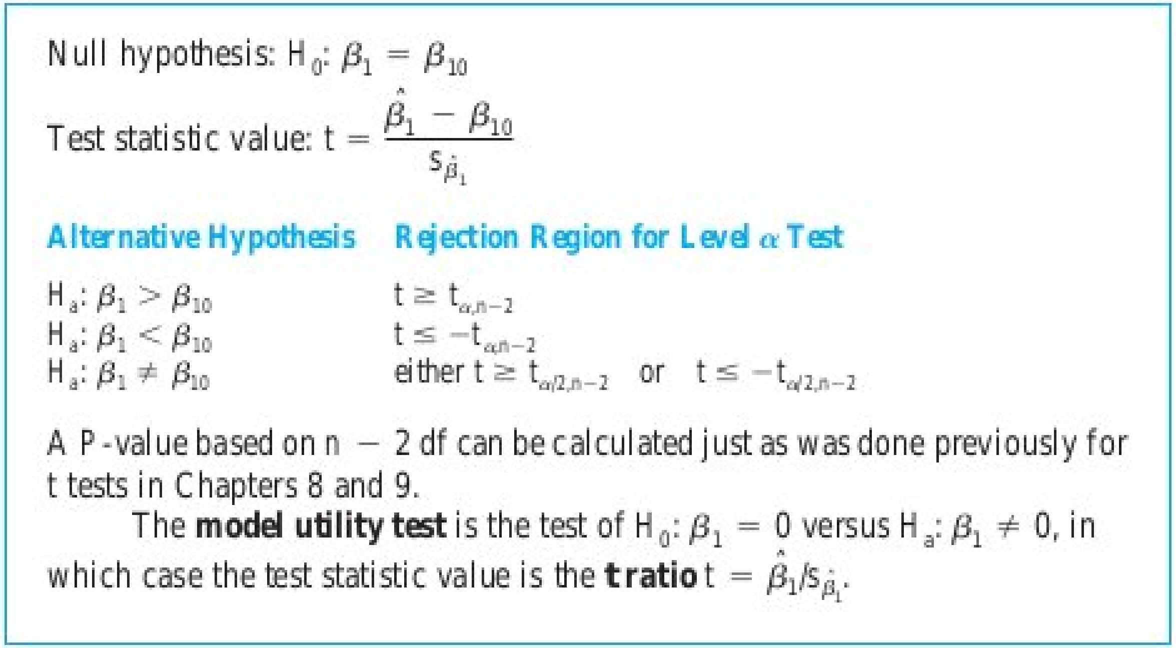
$$T = \frac{\hat{\beta}_1 - \beta_1}{S/\sqrt{S_{xx}}} = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}}$$

  follows a t distribution with df $n - 2$.

- So a $100(1 - \alpha)\%$ CI for the slope β_1 is

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2} \cdot s_{\hat{\beta}_1}$$

Hypothesis Testing



Null hypothesis: H_0: $\beta_1 = \beta_{10}$

Test statistic value: $t = \dfrac{\hat{\beta}_1 - \beta_{10}}{s_{\hat{\beta}_1}}$

Alternative Hypothesis	Rejection Region for Level α Test
H_a: $\beta_1 > \beta_{10}$	$t \geq t_{\alpha,n-2}$
H_a: $\beta_1 < \beta_{10}$	$t \leq -t_{\alpha,n-2}$
H_a: $\beta_1 \neq \beta_{10}$	either $t \geq t_{\alpha/2,n-2}$ or $t \leq -t_{\alpha/2,n-2}$

A P-value based on $n - 2$ df can be calculated just as was done previously for t tests in Chapters 8 and 9.

The **model utility test** is the test of H_0: $\beta_1 = 0$ versus H_a: $\beta_1 \neq 0$, in which case the test statistic value is the **t ratio** $t = \hat{\beta}_1/s_{\hat{\beta}_1}$.