Estimating Model Parameters

- Least Squares
- Estimating σ^2
- The Coefficient of Determination

A Linear Probabilistic Model

There are parameters β_0, β_1 and σ^2, such that for any fixed value of the independent variable x, the dependent variable is a random variable related to x through the **model equation**

$$Y = \beta_0 + \beta_1 x + \epsilon \tag{1}$$

The quantity ϵ in the model equation is a random variable, assumed to be normally distributed with

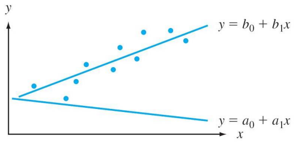
$$E(\epsilon) = 0, \quad Var(\epsilon) = \sigma^2.$$

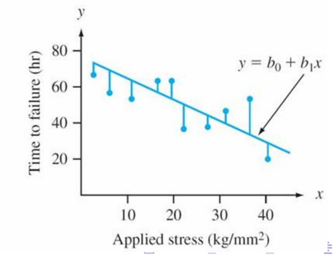
Parameter Estimation

- Sample data consists of n pairs $(x_1, y_1), \cdots, (x_n, y_n)$
- These observations are assumed to have been obtained independently of one another.
  That is, y_i is the observed value of Y_i, where
  $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ and the n deviations $\epsilon_1, \epsilon_2, \epsilon_n$ are independent rv's.
- The values of β_0, β_1, and σ^2 remain to be estimated.

Least Squares

- A good estimate for regression line should be the one that observed points are scattered rather closely about this line.
- A measure of closeness is the vertical distances (deviations) from the observed points to the line.
- The best-fit line is then the one having the smallest possible sum of squared deviations.





Principle of Least Squares (Ordinary Least Squares or OLS)

- The vertical deviation of the point (x_i, y_i) from the line $y = b_0 + b_1 x$ is

$$\text{height of point} - \text{height of line} = y_i - (b_0 + b_1 x_i)$$

- The sum of squared vertical deviations from the points $(x_1, y_1), \cdots, (x_n, y_n)$ to the line is then

$$f(b_0, b_1) = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2$$

- The point estimates of β_0 and β_1, denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$ and called the **least squares estimates**, are those values that minimize $f(b_0, b_1)$.
- The estimated regression line or least squares line is then the line whose equation is $y = \hat{\beta}_0 + \hat{\beta}_1 x$.

The minimizing values of b_0 and b_1 are found by taking partial derivatives of $f(b_0, b_1)$ with respect to both b_0 and b_1, equating them both to zero [analogously to $f'(b) = 0$ in univariate calculus], and solving the equations

$$\frac{\partial f(b_0, b_1)}{\partial b_0} = \sum 2(y_i - b_0 - b_1 x_i)(-1) = 0$$

$$\frac{\partial f(b_0, b_1)}{\partial b_1} = \sum 2(y_i - b_0 - b_1 x_i)(-x_i) = 0$$

Least Square Estimator

- The least squares estimate of the slope coefficient β_1 of the true regression line is
$$b_1 = \hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$
Computing formulas for the numerator and denominator of $\hat{\beta}_1$ are
$$S_{xy} = \sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)/n \quad S_{xx} = \sum x_i^2 - \left(\sum x_i\right)^2/n$$
- The least squares estimate of the intercept β_0 of the true regression line is
$$b_0 = \hat{\beta}_0 = \frac{\sum y_i - \hat{\beta}_1 \sum x_i}{n} = \bar{y} - \hat{\beta}_1 \bar{x}$$
- The computational formulas for S_{xy} and S_{xx} require only the summary statistics $\sum x_i, \sum y_i, \sum x_i^2$ and $\sum x_i y_i$.

Standard Error of $\hat{\beta}_1$

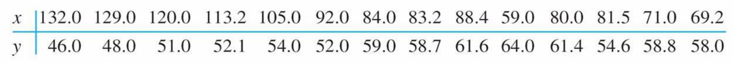
$$\begin{aligned}
Var\left(\hat{\beta}_1\right) &= Var\left(\frac{\sum(x_i-\bar{x})(Y_i-\bar{Y})}{\sum(x_i-\bar{x})^2}\right)\\
&= Var\left(\frac{\sum(x_i-\bar{x})Y_i}{S_{xx}}\right) = Var\left(\sum \frac{x_i-\bar{x}}{S_{xx}} Y_i\right)\\
&= \sum \frac{(x_i-\bar{x})^2}{S_{xx}^2} Var\left(Y_i\right) = \frac{S_{xx}}{S_{xx}^2}\sigma^2 = \frac{\sigma^2}{S_{xx}}
\end{aligned}$$

So the standard error of $\hat{\beta}_1$ is

$$s_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{S_{xx}}}$$

Example

Following data are on x = iodine value (g) and y = cetane number for a sample of 14 biofuels



x	132.0	129.0	120.0	113.2	105.0	92.0	84.0	83.2	88.4	59.0	80.0	81.5	71.0	69.2
y	46.0	48.0	51.0	52.1	54.0	52.0	59.0	58.7	61.6	64.0	61.4	54.6	58.8	58.0

Calculating the summary statistics

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

from which

$$S_{xx} = 128913.93 - (1307.5)^2/14 = 6802.7693$$
$$S_{xy} = 71347.30 - (1307.5)(799.2)/14 = -1424.41429$$

Example Cont'd

The estimated slope of the true regression line (i.e., the slope of the least squares line) is

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{-1424.41429}{6802.7693} = -0.209$$

We estimate that the expected change in true average cetane number associated with a 1g increase in iodine value is -0.209, i.e., a decrease of 0.209.

Since $\bar{x} = 93.392857$ and $\bar{y} = 55.657143$, the estimated intercept of the true regression line is

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 55.658 - (-0.209)(93.393) = 75.21$$

The equation of the estimated regression line (least squares line) is $y = 75.21 - 0.209x$.

Example Cont'd

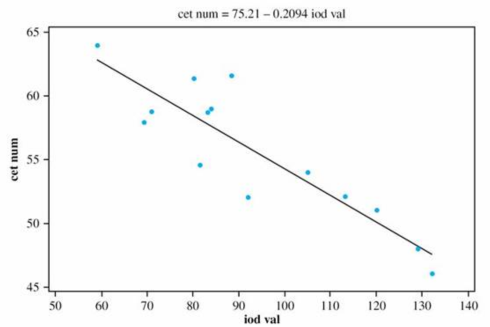


Figure: Scatter plot for data with least square line superimposed.

Remarks

- Be sure that the scatter plot shows a linear pattern with relatively homogenous variation before fitting the simple linear regression model.
- The least squares line should not be used make a prediction for an x value much beyond the range of observed data. The **danger of extrapolation** is that the fitted relationship may not be valid for such x values.

Estimating σ^2

The parameter σ^2 determines the amount of variability inherent in the regression model. A large value of σ^2 will lead to observed (x_i, y_i)'s that are quite spread out about the true regression line, whereas when σ^2 is small the observed points will tend to fall very close to the true line.

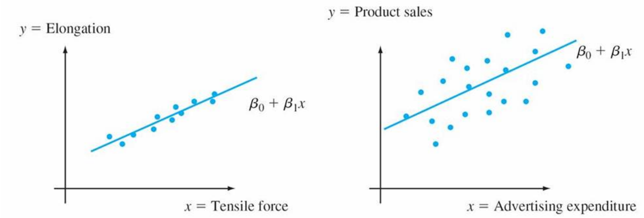


Figure: Typical sample for σ^2: (a)small; (b)large.

Estimating σ^2

An estimate of σ^2 will be used in confidence interval (CI) formulas and hypothesis-testing procedures.

Many large deviations (residuals) suggest a large value of σ^2, whereas deviations all of which are small in magnitude suggest that σ^2 is small.

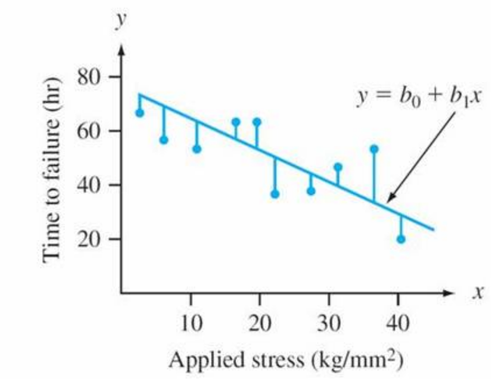
Definition

- The **fitted values** (or predicted values) $\hat{y}_1, \cdots, \hat{y}_n$ are obtained by successively substituting $x_1, \cdots, x_n$ into the equation of the estimated regression line:

$$\hat{y}_1 = \hat{\beta}_0 + \hat{\beta}_1 x_1, \cdot, \hat{y}_n = \hat{\beta}_0 + \hat{\beta}_1 x_n$$

- The **residuals** are the differences $y_1 - \hat{y}_1, y_2 - \hat{y}_2, \cdots, y_n - \hat{y}_n$ between the observed and fitted y values.

After fitting a regression line to the data, the residuals are identified by the vertical line segments from the observed points to the line.



Estimating σ^2

- The **error sum of squares** (equivalently, residual sum of squares), denoted by SSE, is

$$SSE = \sum (y_i - \hat{y}_i)^2 = \sum \left(y_i - (\hat{\beta}_0 - \hat{\beta}_1 x_i) \right)^2$$

- The estimate of σ^2 is

$$\hat{\sigma}^2 = s^2 = \frac{SSE}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2}$$

The $n-2$ df comes from the fact that when obtaining s^2, the two parameters β_0 and β_1 must first be estimated, which results in a loss of 2 df.

A Shortcut Formula to Calculate SSE

Computation of SSE from the defining formula involves much tedious arithmetic, because both the predicted values and residuals must first be calculated.

Use of the following computational formula does not require these quantities.

$$SSE = \sum y_i^2 - \hat{\beta}_0 \sum y_i - \hat{\beta}_1 \sum x_i y_i$$

Figure below shows three different scatter plots of bivariate data. In all three plots, the heights of the different points vary substantially, indicating that there is much variability in observed y values.

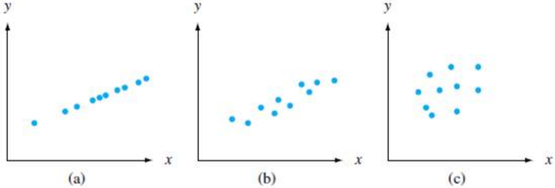


Figure: Using the model to explain y variation: (a) data for which all variation is explained; (b) data for which most variation is explained; (c) data for which little variation is explained.

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.

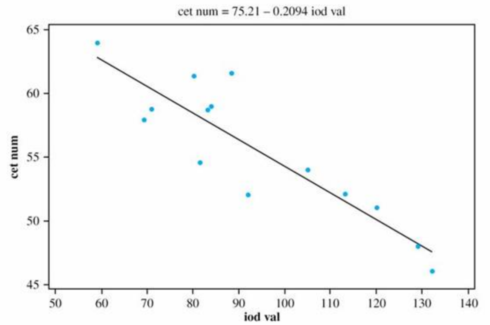


Figure: Scatter plot for data with least square line superimposed.

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.