12.1 The simple linear Regression Model

1. A linear probabilistic model

For any fixed value of the independent variable x, the dependent variable

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

ε is a RV, assume randomly distributed with

$E(\varepsilon) = 0$, $V(\varepsilon) = \sigma^2$

12.2. Estimating Model Parameters

1. least square

$$y_i = b_0 + b_1 x_i$$

$$b_1 = \hat{\beta_1} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

$$S_{xy} = \sum x_i y_i - (\sum x_i)(\sum y_i)/n.$$

$$S_{xx} = \sum x_i^2 - (\sum x_i)^2/n$$

$$b_0 = \beta_0 = \bar{y} - \beta_1 \bar{x}$$

2. Estimating σ^2 and σ

- Error sum of square

$$SSE = \sum [y_i - (\hat{\beta_0} + \hat{\beta_1} x_i)]^2$$

Estimate of σ^2

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

$$SSE = S_{yy} - \text{slope} \cdot S_{xy}$$

— The coefficient of determination.

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

It's interpreted as the proportion of observed y variance that can be explained by the model

12.3. Inference About the slope parameter β_1

— Proposition:

1. The mean value of $\hat{\beta}_1$ is $E(\hat{\beta}_1) = \mu_{\hat{\beta}_1} = \hat{\beta}_1$
2. Variance and standard deviation of $\hat{\beta}_1$

$$V(\hat{\beta}_1) = \sigma^2_{\hat{\beta}_1} = \frac{\sigma^2}{S_{xx}}$$

$$\sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{S_{xx}}} \qquad S_{xx} = \sum (x_i - \bar{x})^2$$

3. Estimate $\hat{\beta}_1$ has normal distribution.

— Theorm

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

has t distribution with $df = n-2$

— Confidence Interval for β_1 slope of line.

· A $100(1-\alpha)\%$ CI for β_1

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

— Model utility test

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

Model utility test

$H_0: \beta_1 = 0$

$H_a: \beta_1 \neq 0$

2. Test statistic

$$t = \frac{\hat{\beta}_1 - \beta_{10}}{S_{\hat{\beta}_1}} \quad df = n-2 \qquad t = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}}$$

3. Alternative Hypothesis — P-value

$H_a: \beta_1 > \beta_{10}$ — Right of t_{n-2}.

$H_a: \beta_1 < \beta_{10}$ — Left of t_{n-2}.

$H_a: \beta_1 \neq \beta_{10}$ — 2 · right of $|t_{n-2}|$

12.4. Inference Concerning $\mu_{Y \cdot x^*}$ and the

1. Prediction of future Y values.

- Proposition

  • Let $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$, where x^* is some fixed value of x.

  • The mean value of $\hat{Y}$ is.

$$E(\hat{Y}) = E(\hat{\beta}_0 + \hat{\beta}_1 x^*)$$

$$= \mu_{\hat{\beta}_0 + \hat{\beta}_1 x^*} = \beta_0 + \beta_1 x^*.$$

Thus, $\hat{\beta}_0 + \hat{\beta}_1 x^*$ is an unbiased estimate for $\beta_0 + \beta_1 x^*$ ($\mu_{Y \cdot x^*}$).

• The variance of $\hat{Y}$

$$V(Y) = \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right]$$

$$S_{\hat{y}} = s \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

• $\hat{Y}$ has a normal distribution.

2. Inference Concerning $\mu_{Y \cdot x^*}$

- The variable T

$$T = \frac{\hat{Y} - (\beta_0 + \beta_1 x^*)}{S_{\hat{Y}}}$$

has t-distribution with $n-2$ df.

- A $100(1-\alpha)\%$ CI for $\mu_{Y \cdot x^*}$, the expected value of Y when $x = x^*$, is

$$\hat{y} \pm t_{\alpha/2, n-2} \cdot S_{\hat{Y}}.$$

3. A Prediction Interval for a Future Value of Y.

A $100(1-\alpha)\%$ PI for a future Y observation to be made when $x = x^*$ is

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2, n-2} \cdot S \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

$$= \hat{y} \pm t_{\alpha/2, n-2} \cdot \sqrt{S^2 + S_{\hat{Y}}^2}$$

12.1 Correlation.

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

$$= \sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{n}$$

- Correlation

$$r = \frac{S_{xy}}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}}$$

$$= \frac{S_{xy}}{\sqrt{S_{xx}} \sqrt{S_{yy}}}$$

- Properties of r
  1. the value of r is independent of unit
  2. $-1 \le r \le 1$
  3. $r = 1$ iff all points lie on a straight line with positive slope.

2. Absense of Correlation.
- Let R denote sample correlation coefficient.
- When $H_0: \rho = 0$ is true.

$$T = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \quad \text{with } n-2 \text{ d.f.}$$

Alter Hypothesis.	P-value
$H_a: \rho > 0$	Right of T curve.
$H_a: \rho < 0$	left of T curve
$H_a: \rho \neq 0$.	2·(right of $\|t\|$.)

3. When $(X_1, Y_2) \cdots (X_n, Y_n)$ is a sample from a bivariate normal distribution.

$$V = \frac{1}{2}\ln\left(\frac{1+R}{1-R}\right).$$

has approximately a normal distribute.

$$\mu_V = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) \quad \sigma_V^2 = \frac{1}{n-3}.$$

4. A $(1-\alpha)$% CI for ρ is

$$\left(\frac{e^{2c_1}-1}{e^{2c_1}+1}, \frac{e^{2c_2}-1}{e^{2c_2}+1}\right).$$

Where C_1, C_2 are left, right endpoints.

1. The most of softwares provide regression output such as least squares estimates, r^2, regression sum of squares and other information upon request.

Use $\hat{y}$ to write fitted line.

```
The regression equation is
cet num = 75.2 - 0.209 iod val

Predictor        Coef    SE Coef       T       P
Constant       75.212      2.984   25.21   0.000
iod val     -0.20939     0.03109   -6.73   0.000

s = 2.56450  R-sq = 79.1%  R-sq(adj) = 77.3%

Analysis of Variance

SOURCE        DF       SS       MS       F       P
Regression     1   298.25   298.25   45.35   0.000
Error         12    78.92     6.58
Total         13   377.17
```

($\hat{\beta}_0$, $\hat{\beta}_1$; $SE\hat{\beta}_1$; $100r^2$; SSE; SST; $t = \frac{\hat{\beta}_1}{SE\hat{\beta}_1}$.)

   1. The F test gives exactly the same result as the model utility t test because $t^2 = f$ and $t^2_{\alpha/2,n-2} = F_{\alpha,1,n-2}$.

Source of Variation	df	Sum of Squares	Mean Square	f
Regression	1	SSR	SSR	$\frac{\text{SSR}}{\text{SSE}/(n-2)}$
Error	$n-2$	SSE	$s^2 = \frac{\text{SSE}}{n-2}$	
Total	$n-1$	SST		

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

Assumption of Regression line.

① linearity between x and y.
Check by scatter plot of x and Y.

② Y's have normal distribution $N(\mu_{xi}, \sigma^2)$
Normal Probability Plot.

③ The variance of the residual is the same for all x's.
By scatter residual vs. fit.