Chapter 2. Least Square Regression (I)

Reading Assignment

Chapte 11.

Simple Linear Regression Models

1. Example (mid-term and final grades).

Graph.

$$\text{final exam grade} = 27.60 + 0.51 * \text{mid-term grade}$$

2. Example (heights of family members).

Graph.

$$\text{son's height} = 33.73 + 0.516 * \text{father's height}$$

Why the name "Regression"?

Law of universal regression. F.Galton (1889). *Natural Inheritance*. Macmillan, London.

"Every peculiarity in a man is shared by his kinsman, but *on the average* in a less degree."

The mathematical model for simple regression

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

1. Y: dependent variable.
2. x: independent (or, explanatory) variable.
3. ε: random error, with $E[\varepsilon] = 0$ and $\mathrm{Var}[\varepsilon] = \sigma^2$.
4. β_0, β_1: population parameter, unknown.

$$E[Y|x] = \beta_0 + \beta_1 x$$

The method of least square

Suppose the data is $(x_1, Y_1), \ldots, (x_n, Y_n)$.

Find $(\hat{\beta}_0, \hat{\beta}_1)$ that minimize the sum of squares for error

$$\text{SSE} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} \left[Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right]^2 .$$

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

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

Formulae for simple regression:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

Theorem. $\hat{\beta}_i$ is an unbiased estimator for β_i.

Justification for the method of least square

Gauss-Markov Theorem. $\hat{\beta}_i$ has the minimal variance among all the linear unbiased estimators of β_i. If we further assume that ε_i has a normal distribution, then $\hat{\beta}_i$ is indeed the MVUE for β_i.

Properties of Regression coefficients

1. $\hat{\beta}_i$ is an unbiased estimator for β_i.

2. The variance matrix

$$\text{Var}\begin{bmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{bmatrix} = \sigma^2\begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2\end{bmatrix}^{-1} = \frac{\sigma^2}{nS_{xx}}\begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n\end{bmatrix}$$

3. An unbiased estimate for σ^2 is

$$S^2 \doteq \frac{1}{n-2}\text{SSE} = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \frac{1}{n-2}\left[S_{yy} - \hat{\beta}_1 S_{xy}\right]$$

4. If we further assume that $\varepsilon \sim N(0, \sigma^2)$, then

(a) $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)'$ is normally distributed.

(b) $S^2 \sim \chi^2(n-2)$.

(c) S^2 and $\hat{\beta}$ are independent.

[The proof will be given for the general case of multiple linear regressions.]

Inference of Regression coefficients

Assumption: $\varepsilon \sim N(0, \sigma^2)$.

Recall

$$\text{Var}\begin{bmatrix}\hat{\beta}_0 \\ \hat{\beta}_1\end{bmatrix} = \frac{\sigma^2}{nS_{xx}}\begin{bmatrix}\sum x_i^2 & -\sum x_i \\ -\sum x_i & n\end{bmatrix} = \sigma^2\begin{bmatrix}C_{00} & C_{01} \\ C_{10} & C_{11}\end{bmatrix}$$

Proposition.

$$T = \frac{\hat{\beta}_i - \beta_i}{S\sqrt{C_{ii}}} \sim t(n-2)$$

More generally, for $\theta = a_0\beta_0 + a_1\beta_1$, an unbiased estimator is

$$\hat{\theta} = a_0\hat{\beta}_0 + a_1\hat{\beta}_1,$$

with variance

$$\mathrm{Var}[\hat{\theta}] = \sigma^2[a_0, a_1] \begin{bmatrix} C_{00} & C_{01} \\ C_{10} & C_{11} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} \doteq \sigma^2 a' C a$$

PROPOSITION.

$$T = \frac{\hat{\theta} - \theta}{S\sqrt{a'Ca}} \sim t(n-2)$$

Estimation and Prediction

Two problems:

1. Estimate the population mean of Y given $x = x^*$.
2. Predict an individual Y given $x = x^*$.

Answer:

1.
$$\hat{\mu}_Y = \hat{\beta}_0 + \hat{\beta}_1 x^*.$$

2.
$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x^*.$$

Are they the "same"?

1.

$$\hat{\mu}_Y = \hat{\beta}_0 + \hat{\beta}_1 x^*.$$

$$T = \frac{\hat{\mu}_Y - \mu_Y}{S\sqrt{a'Ca}} \sim t(n-2)$$

where

$$a = [1, x^*]' \quad \text{and} \quad a'Ca = \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}$$

2.

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x^*.$$

$$\text{error} = \hat{Y} - Y = \hat{\beta}_0 + \hat{\beta}_1 x^* - (\beta_0 + \beta_1 x^* + \varepsilon)$$

$$\text{error} \sim N\left(0, \sigma^2[1 + a'Ca]\right)$$

$$T = \frac{\hat{Y} - Y}{S\sqrt{1 + a'Ca}} \sim t(n-2)$$

where

$$a = [1, x^*]' \quad \text{and} \quad 1 + a'Ca = 1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}$$

Analysis of Variance for Simple Regression

Total variation of Y.

$$\text{SST} \doteq \sum_i (Y_i - \bar{Y})^2.$$

- Variation accounted for by linear regression.

$$\text{SSR} \doteq \sum_i (\hat{Y}_i - \bar{Y})^2.$$

- Variation NOT accounted for by linear regression.

$$\text{SSE} \doteq \sum_i (Y_i - \hat{Y}_i)^2.$$

$$\text{SST} = \text{SSE} + \text{SSE},$$

$$\text{SSR} = \frac{S_{xy}^2}{S_{xx}} = (\hat{\beta}_1)^2 S_{xx}, \quad \text{SSE} = (n-2)S^2$$

Source of Variation	d.f.	Sum of Squares	Mean Square
Regression	1	SSR	$\text{MSR} = S_{xy}^2/S_{xx}$
Error	$n-2$	SSE	$\text{MSE} = S^2$
Total	$n-1$	SST	

MSR and MSE are independent.

An F-test

$$H_0 : \beta_1 = 0, \quad H_a : \beta_1 \neq 0.$$

Under H_0,

$$\text{MSR} \sim \chi^2(1), \quad \text{MSE} \sim \chi^2(n-2)$$

and

$$F = \frac{\text{MSR}}{\text{MSE}} \sim F(1, n-2).$$

Rejection Region with significance level $\alpha = \{F > F_\alpha\}$.

This is equivalent to the t-test!

Examples

1. The winning speed (mph) for Indianapolis 500 auto races from 1962 to 1971 were as follows. To save space, 100 mph is subtracted from each speed Y.

Year x	62	63	64	65	66	67	68	69	70	71
Speed Y	40.3	43.1	47.4	51.4	44.3	51.2	52.9	56.9	55.7	57.7

$$\bar{x} = 1966.5, \qquad \bar{Y} = 150.09$$

$$S_{xx} = 82.5, \quad S_{xy} = 151.85, \quad S_{yy} = 332.07$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = 1.841, \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} = -3469.46$$

Regression Line:

$$\hat{Y} = -3469.46 + 1.841x.$$

Error estimate:

$$S^2 = \frac{1}{n-2}\text{SSE} = \frac{1}{n-2}\left[S_{yy} - \hat{\beta}_1 S_{xy}\right] = 6.572,$$

(a) What's your estimate for the increase of speed per year?

$$\hat{\beta}_1 = 1.841$$

and its $(1-\alpha)$ confidence interval

$$1.841 \pm t_{\alpha/2}(8) S \sqrt{C_{11}} = 1.841 \pm t_{\alpha/2}(8) S \sqrt{\frac{1}{S_{xx}}}$$

or

$$1.841 \pm t_{\alpha/2}(8) \cdot 0.282.$$

(b) What's your prediction for the winning speed at 1974, and 95% interval?

$$\hat{Y} = -3469.46 + 1.841 \cdot 1974 = 164.6$$

The 95% interval is

$$\hat{Y} \pm t_{0.025}(8)\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}} = 164.6 \pm 2.306 \cdot 1.335$$

or

$$[161.5, 167.7]$$

The actual winning speed was 158.6 mph.

Discussion of extrapolation. (Predict 2006: 223.59 mph)

2. The data show the initial weights and gains in weight (grams) of 15 female rats on a high protein diet from 24 to 84 days of age. The point of interest is whether the gain is related to the initial body weight. If so, feeding experiments on female rats can be made more precise by taking account of the initial weights, either by pairing on initial weight or by adjusting the analysis for the difference in initial weight. What is your conclusion?

Rat number	1	2	3	4	5	6	7	8	9
Initial weight x	50	64	76	64	74	60	69	68	56
Gain Y	128	159	158	119	133	112	96	126	132

Rat number	10	11	12	13	14	15
Initial weight x	48	57	59	46	45	65
Gain Y	118	107	106	82	103	104

$$\bar{x} = 60.07, \quad \bar{y} = 118.87$$

$$S_{xx} = 1344.93, \quad S_{yy} = 6357.73, \quad S_{xy} = 1431.13$$

Regression line:

$$\hat{Y} = 54.950 + 1.064x$$

$$S^2 = \frac{1}{n-2}\text{SSE} = \frac{1}{n-2}\left[S_{yy} - \hat{\beta}_1 S_{xy}\right] = 371.91$$

$$H_0 : \beta_1 = 0, \qquad \beta_1 \neq 0.$$

$$T = \frac{\hat{\beta}_1 - 0}{S\sqrt{1/S_{xx}}} = 2.023$$

$$P\text{-value} = 2P(t(13) > |T|) = 0.064$$

Fitting nonlinear relations

1. The exponential growth curve. $E[Y] = AB^x$.

$$\ln Y = \ln A + \ln B \cdot x + \varepsilon.$$

or

$$W = \beta_0 + \beta_1 x + \varepsilon.$$

Example: Dry weights (1 unit $= 10^{-3}$ grams) of chick embryos from ages 6 to 16 days.

age X	6	7	8	9	10	11	12	13	14	15	16
weight Y	29	52	79	125	181	261	425	738	1130	1882	2812

$$\ln Y = 0.7156 + 0.4510x$$

$$Y = 2.046 \cdot 1.57^{x}.$$

Discussion on the assumptions and other regression method.

2. A general method for fitting nonlinear regressions.

$$Y = f(\beta, x) + \varepsilon.$$

f: a nonlinear function of known form.

β: the unknown parameters to be estimated from data.

The method of least square:

$$\min_{\beta} \sum_{i=1}^{n} [Y_i - f(\beta, x_i)]^2 .$$