STAC67H: Regression Analysis

Fall, 2014

Instructor: Jabed Tomal

Department of Computer and Mathematical Sciences
University of Toronto Scarborough
Toronto, ON
Canada

September 17, 2014

Simple Linear Regression Model:

Estimation of Error Terms Variance σ^2

Recall that the variance of each observation Y_i for the regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \;\; ; \;\; i = 1, 2, \cdots, n.$$

is σ^2, the same as that of each error term ϵ_i.

Recall that the variance of each observation Y_i for the regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \;\; ; \;\; i = 1, 2, \cdots, n.$$

is σ^2, the same as that of each error term ϵ_i.

1. The deviation of an observation Y_i from its estimated mean $\hat{Y}_i$ is called the residual

$$e_i = Y_i - \hat{Y}_i.$$

Recall that the variance of each observation Y_i for the regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \;\; ; \; i = 1, 2, \cdots, n.$$

is σ^2, the same as that of each error term ϵ_i.

1. The deviation of an observation Y_i from its estimated mean $\hat{Y}_i$ is called the residual

$$e_i = Y_i - \hat{Y}_i.$$

2. The sum of squares of the residuals is denoted by *SSE*

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (e_i)^2$$

where *SSE* stands for *error sum of squares* or *residual sum of squares*.

Simple Linear Regression Model:

Estimation of Error Terms Variance σ^2

The sum of squares *SSE* has $n-2$ degrees of freedom associated with it. Two degrees of freedom are lost because both β_0 and β_1 had to be estimated in obtaining the estimated means $\hat{Y}_i$. Hence, the mean square, denoted by *MSE* or s^2, is:

$$s^2 = MSE = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n}(e_i)^2}{n-2},$$

where *MSE* stands for *error mean squares* or *residual mean square*.

1. It can be shown that MSE is an unbiased estimator of σ^2 for the above first-order regression model

$$E(MSE) = \sigma^2.$$

1. It can be shown that MSE is an unbiased estimator of σ^2 for the above first-order regression model

$$E(MSE) = \sigma^2.$$

2. An estimator of the standard deviation of σ is simply

$$s = \sqrt{MSE},$$

the positive square room of MSE.

In our *airfreight breakage* exercise (1.21), the *residual* values are in the fourth row of the following table

i:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
X_i:	1	0	2	0	3	1	0	1	2	0
Y_i:	16	9	17	12	22	13	8	15	19	11
$\hat{Y}_i$:	14.2	10.2	18.2	10.2	22.2	14.2	10.2	14.2	18.2	10.2
e_i:	1.8	-1.2	-1.2	1.8	-0.2	-1.2	-2.2	0.8	0.8	0.8

1. For this exercise, $n = 10$ and

$$SSE = \sum_{i=1}^{10} e_i^2 = (1.8)^2 + (-1.2)^2 + \cdots + (0.8)^2 = 17.6.$$

1. For this exercise, $n = 10$ and

$$SSE = \sum_{i=1}^{10} e_i^2 = (1.8)^2 + (-1.2)^2 + \cdots + (0.8)^2 = 17.6.$$

2. Hence, the mean square MSE or s^2 is:

$$s^2 = MSE = \frac{SSE}{n-2} = \frac{17.6}{10-2} = 2.2.$$

1. For this exercise, $n = 10$ and

$$SSE = \sum_{i=1}^{10} e_i^2 = (1.8)^2 + (-1.2)^2 + \cdots + (0.8)^2 = 17.6.$$

2. Hence, the mean square *MSE* or s^2 is:

$$s^2 = MSE = \frac{SSE}{n-2} = \frac{17.6}{10-2} = 2.2.$$

3. An estimator of the standard deviation of σ is

$$s = \sqrt{MSE} = \sqrt{2.2} = 1.48324.$$

1. No matter what may be the form of the distribution of the error terms ϵ_i (and hence of the Y_i), the least squares method provides unbiased point estimators of β_0 and β_1 that have minimum variance among all unbiased linear estimators.

Simple Linear Regression Model:

Normal Error Regression Model

1. No matter what may be the form of the distribution of the error terms ϵ_i (and hence of the Y_i), the least squares method provides unbiased point estimators of β_0 and β_1 that have minimum variance among all unbiased linear estimators.
2. To set up interval estimates and make tests, we need to make an assumption about the form of the distribution of the ϵ_i.

The normal error regression model is as follows

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \;\; ; \;\; i = 1, 2, \cdots, n.$$

The normal error regression model is as follows

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \;\; ; \;\; i = 1, 2, \cdots, n.$$

Where,

1. Y_i is the observed response in the ith trial,

The normal error regression model is as follows

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \ \ ; \ i = 1, 2, \cdots, n.$$

Where,

1. Y_i is the observed response in the *i*th trial,
2. X_i is a known constant, the level of the predictor variable in the *i*th trial,

The normal error regression model is as follows

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \;\; ; \;\; i = 1, 2, \cdots, n.$$

Where,

1. Y_i is the observed response in the ith trial,
2. X_i is a known constant, the level of the predictor variable in the ith trial,
3. β_0 and β_1 are the parameters,

The normal error regression model is as follows

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \;\; ; \; i = 1, 2, \cdots, n.$$

Where,

1. Y_i is the observed response in the *i*th trial,
2. X_i is a known constant, the level of the predictor variable in the *i*th trial,
3. β_0 and β_1 are the parameters,
4. ϵ_i are independent $N(0, \sigma^2)$.

Comments:

1. The symbol $N(0, \sigma^2)$ stands for normally distributed, with mean 0 and variance σ^2.

Comments:

1. The symbol $N(0, \sigma^2)$ stands for normally distributed, with mean 0 and variance σ^2.
2. The normal error model is the same as the regression model with unspecified error distribution, except that the former assumes that the errors ϵ_i are normally distributed.

Comments:

1. The symbol $N(0, \sigma^2)$ stands for normally distributed, with mean 0 and variance σ^2.
2. The normal error model is the same as the regression model with unspecified error distribution, except that the former assumes that the errors ϵ_i are normally distributed.
3. Normal error regression model implies that the Y_i are independent normal random variables, with mean $E\{Y_i\} = \beta_0 + \beta_1 X_i$ and variance σ^2.

In our normal error regression model $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$. That is

$$f(Y_i|\beta_0, \beta_1, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}\left(Y_i - \beta_0 - \beta_1 X_i\right)^2\right\}$$

1. The likelihood function is the joint density function of $Y_1, Y_2, \cdots, Y_n$

$$L(\beta_0, \beta_1, \sigma^2 | Y_1, Y_2, \cdots, Y_n) = f(Y_1, Y_2, \cdots, Y_n | \beta_0, \beta_1, \sigma^2).$$

1. The likelihood function is the joint density function of $Y_1, Y_2, \cdots, Y_n$

$$L(\beta_0, \beta_1, \sigma^2 | Y_1, Y_2, \cdots, Y_n) = f(Y_1, Y_2, \cdots, Y_n | \beta_0, \beta_1, \sigma^2).$$

2. As the responses $Y_1, Y_2, \cdots, Y_n$ are independent, we get

$$f(Y_1, Y_2, \cdots, Y_n | \beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} f(Y_i | \beta_0, \beta_1, \sigma^2).$$

Hence, our likelihood function becomes

$$L(\beta_0, \beta_1, \sigma^2 | Y_1, \cdots, Y_n) = (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \right\}$$

Taking $\log_e$ on both sides, we get

$$\log_e L = -\frac{n}{2}\log_e(2\pi) - \frac{n}{2}\log_e(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2$$

By maximizing $\log_e L(\beta_0, \beta_1, \sigma^2)$ with respect to β_0, β_1 and σ^2, we get

By maximizing $\log_e L(\beta_0, \beta_1, \sigma^2)$ with respect to β_0, β_1 and σ^2, we get

$$\hat{\beta}_1 = b_1 \qquad \text{(the least square estimator of } \beta_1\text{)},$$

By maximizing $\log_e L(\beta_0, \beta_1, \sigma^2)$ with respect to β_0, β_1 and σ^2, we get

$$\hat{\beta}_1 = b_1 \qquad \text{(the least square estimator of } \beta_1\text{)},$$

$$\hat{\beta}_0 = \bar{Y} - b_1\bar{X} \qquad \text{(the least square estimator of } \beta_0\text{)},$$

By maximizing $\log_e L(\beta_0, \beta_1, \sigma^2)$ with respect to β_0, β_1 and σ^2, we get

$$\hat{\beta}_1 = b_1 \qquad \text{(the least square estimator of } \beta_1\text{)},$$

$$\hat{\beta}_0 = \bar{Y} - b_1\bar{X} \qquad \text{(the least square estimator of } \beta_0\text{)},$$

and

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n} \qquad \text{(a biased estimator of } \sigma^2\text{)}.$$

By maximizing $\log_e L(\beta_0, \beta_1, \sigma^2)$ with respect to β_0, β_1 and σ^2, we get

$$\hat{\beta}_1 = b_1 \qquad \text{(the least square estimator of } \beta_1\text{)},$$

$$\hat{\beta}_0 = \bar{Y} - b_1 \bar{X} \qquad \text{(the least square estimator of } \beta_0\text{)},$$

and

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n} \qquad \text{(a biased estimator of } \sigma^2\text{)}.$$

Exercise: show the step-by-step works in obtaining the maximum likelihood estimators of β_0, β_1, and σ^2.

Simple Linear Regression Model:
Normal Error Regression Model

In our *airfreight breakage* exercise, the maximum likelihood estimates of β_1, β_0 and σ^2 are

1. $$\hat{\beta}_1 = 4,$$

In our *airfreight breakage* exercise, the maximum likelihood estimates of β_1, β_0 and σ^2 are

1. $$\hat{\beta}_1 = 4,$$
2. $$\hat{\beta}_0 = 10.2,$$

In our *airfreight breakage* exercise, the maximum likelihood estimates of β_1, β_0 and σ^2 are

1. $$\hat{\beta}_1 = 4,$$
2. $$\hat{\beta}_0 = 10.2,$$
3. and
$$\hat{\sigma}^2 = \frac{SSE}{n} = \frac{17.6}{10} = 1.76,$$
respectively.

Throughout we assume the normal error regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

Throughout we assume the normal error regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

where:

1. β_0 and β_1 are parameters

Throughout we assume the normal error regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

where:

1. β_0 and β_1 are parameters
2. X_i are known constants

Throughout we assume the normal error regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

where:

1. β_0 and β_1 are parameters
2. X_i are known constants
3. ϵ_i are independent $N(0, \sigma^2)$.

At times, tests concerning β_1 are of interests, particularly one of the form:

$$H_0 : \beta_1 = 0,$$

At times, tests concerning β_1 are of interests, particularly one of the form:

$$H_0 : \beta_1 = 0,$$

against

$$H_a : \beta_1 \neq 0.$$

The point estimator b_1 is:

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}.$$

For normal error regression model, the sampling distribution of b_1 is normal, with mean

$$E\{b_1\} = \beta_1,$$

and variance

$$Var\{b_1\} = \frac{\sigma^2}{\sum(X_i - \bar{X})^2}.$$

The estimated variance of b_1 is

$$s^2\{b_1\} = \frac{MSE}{\sum(X_i - \bar{X})^2}$$

an unbiased estimator of $Var\{b_1\}$. Hence, $s\{b_1\}$ is an unbiased estimator of $\sqrt{Var\{b_1\}}$

Here,

$$\frac{b_1 - \beta_1}{\sqrt{\frac{\sigma^2}{\sum(X_i - \bar{X})^2}}} \sim N(0,1),$$

Here,

$$\frac{b_1 - \beta_1}{\sqrt{\frac{\sigma^2}{\sum(X_i - \bar{X})^2}}} \sim N(0, 1),$$

and

$$\frac{(n-2)MSE}{\sigma^2} \sim \chi^2_{(n-2)},$$

and are independent of each other.

Hence,

$$\frac{b_1 - \beta_1}{\sqrt{\frac{MSE}{\sum(X_i - \bar{X})^2}}} = \frac{b_1 - \beta_1}{s\{b_1\}} \sim t_{(n-2)}$$

Let $t_{(\alpha;n-2)}$ be such that $P(t_{(n-2)} \leq t_{(\alpha;n-2)}) = \alpha$
Then, reject the null hypothesis with level of significance α if

$$|t_{est}| = \left| \frac{b_1}{\sqrt{\frac{MSE}{\sum(X_i - \bar{X})^2}}} \right| \geq t_{(1-\alpha/2,n-2)}.$$

The confidence of interval of β_1 is

$$P\left\{b_1 - t_{(1-\alpha/2;n-2)}s\{b_1\} \leq \beta_1 \leq b_1 + t_{(1-\alpha/2;n-2)}s\{b_1\}\right\} = 1-\alpha.$$

For the airfreight breakage example,

For the airfreight breakage example,

1. Residual mean square $MSE = SSE/(n-2) = 2.2$,

For the airfreight breakage example,

1. Residual mean square $MSE = SSE/(n-2) = 2.2$,
2. $\sum(X_i - \bar{X})^2 = 10$,

For the airfreight breakage example,

1. Residual mean square $MSE = SSE/(n-2) = 2.2$,
2. $\sum(X_i - \bar{X})^2 = 10$,
3. $Var(b_1) = 2.2/10 = 0.22$,
4. $s(b_1) = \sqrt{Var(b_1)} = \sqrt{2.2} = 0.4690416$,

For the airfreight breakage example,

1. Residual mean square $MSE = SSE/(n-2) = 2.2$,
2. $\sum(X_i - \bar{X})^2 = 10$,
3. $Var(b_1) = 2.2/10 = 0.22$,
4. $s(b_1) = \sqrt{Var(b_1)} = \sqrt{2.2} = 0.4690416$,
5. $t_{est} = \frac{b_1}{s(b_1)} = 8.528029$,

For the airfreight breakage example,

1. Residual mean square $MSE = SSE/(n-2) = 2.2$,
2. $\sum(X_i - \bar{X})^2 = 10$,
3. $Var(b_1) = 2.2/10 = 0.22$,
4. $s(b_1) = \sqrt{Var(b_1)} = \sqrt{2.2} = 0.4690416$,
5. $t_{est} = \frac{b_1}{s(b_1)} = 8.528029$,
6. $t_{0.975,10-2} = 2.306004$,

For the airfreight breakage example,

1. Residual mean square $MSE = SSE/(n-2) = 2.2$,
2. $\sum(X_i - \bar{X})^2 = 10$,
3. $Var(b_1) = 2.2/10 = 0.22$,
4. $s(b_1) = \sqrt{Var(b_1)} = \sqrt{2.2} = 0.4690416$,
5. $t_{est} = \frac{b_1}{s(b_1)} = 8.528029$,
6. $t_{0.975,10-2} = 2.306004$,
7. pvalue $= 2.748669e - 05$

The confidence interval of β_1 is

The confidence interval of β_1 is

$$4 - 2.306004 \times 0.4690416 \leq \beta_1 \leq 4 + 2.306004 \times 0.4690416$$

The confidence interval of β_1 is

$$4 - 2.306004 \times 0.4690416 \leq \beta_1 \leq 4 + 2.306004 \times 0.4690416$$

$$2.918388 \leq \beta_1 \leq 5.081612.$$

The confidence interval of β_1 is

$$4 - 2.306004 \times 0.4690416 \leq \beta_1 \leq 4 + 2.306004 \times 0.4690416$$

$$2.918388 \leq \beta_1 \leq 5.081612.$$

If you draw many many samples of size 10, almost 95% of such samples will provide b_1 between 2.918388 and 5.081612

The point estimator of β_0 is

The point estimator of β_0 is

$$b_0 = \bar{Y} - b_1 \bar{X}.$$

For the normal error regression model, the sampling distribution of b_0 is normal with mean

For the normal error regression model, the sampling distribution of b_0 is normal with mean

$$E\{b_0\} = \beta_0,$$

For the normal error regression model, the sampling distribution of b_0 is normal with mean

$$E\{b_0\} = \beta_0,$$

and variance

$$Var\{b_0\} = \sigma^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}\right].$$

An unbiased estimator of $Var\{b_0\}$ is

$$s^2\{b_0\} = MSE\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}\right].$$

An unbiased estimator of $Var\{b_0\}$ is

$$s^2\{b_0\} = MSE\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}\right].$$

The positive square root $s\{b_0\}$ is an unbiased estimator of $\sqrt{Var\{b_0\}}$.

The sampling distribution concerning b_0 is

$$\frac{b_0 - \beta_0}{s\{b_0\}} \sim t_{(n-2)}.$$

The sampling distribution concerning b_0 is

$$\frac{b_0 - \beta_0}{s\{b_0\}} \sim t_{(n-2)}.$$

The $100(1-\alpha)\%$ confidence interval of β_0 is

$$b_0 \pm t_{(1-\alpha/2;n-2)} s\{b_0\}.$$

For the airfreight breakage example,

For the airfreight breakage example,

1. Residual mean square $MSE = SSE/(n-2) = 2.2$,

For the airfreight breakage example,

1. Residual mean square $MSE = SSE/(n-2) = 2.2$,
2. $\sum(X_i - \bar{X})^2 = 10$,

For the airfreight breakage example,

1. Residual mean square $MSE = SSE/(n-2) = 2.2$,
2. $\sum(X_i - \bar{X})^2 = 10$,
3. $Var(b_0) = 0.44$,
4. $s(b_0) = \sqrt{Var(b_0)} = \sqrt{0.44} = 0.663325$,

For the airfreight breakage example,

1. Residual mean square $MSE = SSE/(n-2) = 2.2$,
2. $\sum(X_i - \bar{X})^2 = 10$,
3. $Var(b_0) = 0.44$,
4. $s(b_0) = \sqrt{Var(b_0)} = \sqrt{0.44} = 0.663325$,
5. $t_{est} = \frac{b_0}{s(b_0)} = 15.37708$,

For the airfreight breakage example,

1. Residual mean square $MSE = SSE/(n-2) = 2.2$,
2. $\sum(X_i - \bar{X})^2 = 10$,
3. $Var(b_0) = 0.44$,
4. $s(b_0) = \sqrt{Var(b_0)} = \sqrt{0.44} = 0.663325$,
5. $t_{est} = \frac{b_0}{s(b_0)} = 15.37708$,
6. $t_{0.975,10-2} = 2.306004$,

For the airfreight breakage example,

1. Residual mean square $MSE = SSE/(n-2) = 2.2$,
2. $\sum(X_i - \bar{X})^2 = 10$,
3. $Var(b_0) = 0.44$,
4. $s(b_0) = \sqrt{Var(b_0)} = \sqrt{0.44} = 0.663325$,
5. $t_{est} = \frac{b_0}{s(b_0)} = 15.37708$,
6. $t_{0.975,10-2} = 2.306004$,
7. pvalue $= 3.178273e - 07$

The confidence interval of β_0 is

The confidence interval of β_0 is

$$10.2 - 2.306004 \times 0.663325 \leq \beta_0 \leq 10.2 + 2.306004 \times 0.663325$$

The confidence interval of β_0 is

$$10.2 - 2.306004 \times 0.663325 \leq \beta_0 \leq 10.2 + 2.306004 \times 0.663325$$

$$8.67037 \leq \beta_1 \leq 11.72963.$$

The confidence interval of β_0 is

$$10.2 - 2.306004 \times 0.663325 \leq \beta_0 \leq 10.2 + 2.306004 \times 0.663325$$

$$8.67037 \leq \beta_1 \leq 11.72963.$$

If you draw many many samples of size 10, almost 95% of such samples will provide b_0 between 8.67037 and 11.72963.

Prove that the least squares estimator b_1 of β_1 has the minimum variance than any other linear unbiased estimator $\hat{\beta}_1^*$.

Prove that the least squares estimator b_1 of β_1 has the minimum variance than any other linear unbiased estimator $\hat{\beta}_1^*$.

Proof: Consider $\hat{\beta}_1^* = \sum_{i=1}^{n} c_i Y_i$ a linear combination of the observations $Y_1, Y_2, \cdots, Y_n$. We know that $\hat{\beta}_1^*$ is unbiased estimator of β_1, so $E(\hat{\beta}_1^*) = \beta_1$.

Prove that the least squares estimator b_1 of β_1 has the minimum variance than any other linear unbiased estimator $\hat{\beta}_1^*$.

Proof: Consider $\hat{\beta}_1^* = \sum_{i=1}^{n} c_i Y_i$ a linear combination of the observations $Y_1, Y_2, \cdots, Y_n$. We know that $\hat{\beta}_1^*$ is unbiased estimator of β_1, so $E(\hat{\beta}_1^*) = \beta_1$.
Hence,

$$E(\hat{\beta}_1^*) = \sum_{i=1}^{n} c_i E(Y_i) = \sum_{i=1}^{n} c_i(\beta_0 + \beta_1 X_i) = \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i X_i = \beta_1$$

which implies

$$\sum_{i=1}^{n} c_i = 0 \text{ and } \sum_{i=1}^{n} c_i X_i = 1.$$

Inferences in Regression

Let $c_i = k_i + d_i$ where $k_i = (X_i - \bar{X})/\sum_{i=1}^{n}(X_i - \bar{X})^2$ and d_i are any arbitrary constants. We write

Inferences in Regression

Let $c_i = k_i + d_i$ where $k_i = (X_i - \bar{X}) / \sum_{i=1}^{n} (X_i - \bar{X})^2$ and d_i are any arbitrary constants. We write

$$Var(\hat{\beta}_1^*) = Var(\sum_{i=1}^{n} c_i Y_i) = Var\{\sum_{i=1}^{n} (k_i + d_i) Y_i\} = \sum_{i=1}^{n} (k_i + d_i)^2 \sigma^2.$$

Inferences in Regression

Let $c_i = k_i + d_i$ where $k_i = (X_i - \bar{X}) / \sum_{i=1}^n (X_i - \bar{X})^2$ and d_i are any arbitrary constants. We write

$$Var(\hat{\beta}_1^*) = Var(\sum_{i=1}^n c_i Y_i) = Var\{\sum_{i=1}^n (k_i + d_i) Y_i\} = \sum_{i=1}^n (k_i + d_i)^2 \sigma^2.$$

Hence,

$$Var(\hat{\beta}_1^*) = \left(\sum_{i=1}^n k_i^2 + \sum_{i=1}^n d_i^2 + 2 \sum_{i=1}^n k_i d_i \right) \sigma^2 = \left(\sum_{i=1}^n k_i^2 + \sum_{i=1}^n d_i^2 \right) \sigma^2.$$

Inferences in Regression

Let $c_i = k_i + d_i$ where $k_i = (X_i - \bar{X})/\sum_{i=1}^{n}(X_i - \bar{X})^2$ and d_i are any arbitrary constants. We write

$$Var(\hat{\beta}_1^*) = Var(\sum_{i=1}^{n} c_i Y_i) = Var\{\sum_{i=1}^{n}(k_i + d_i)Y_i\} = \sum_{i=1}^{n}(k_i + d_i)^2\sigma^2.$$

Hence,

$$Var(\hat{\beta}_1^*) = \left(\sum_{i=1}^{n} k_i^2 + \sum_{i=1}^{n} d_i^2 + 2\sum_{i=1}^{n} k_i d_i\right)\sigma^2 = \left(\sum_{i=1}^{n} k_i^2 + \sum_{i=1}^{n} d_i^2\right)\sigma^2.$$

Here,

$$\sum_{i=1}^{n} k_i d_i = \sum_{i=1}^{n} k_i(c_i - k_i) = 0.$$

Hence,

$$Var(\hat{\beta}_1^*) = \left(\sum_{i=1}^{n} k_i^2 + \sum_{i=1}^{n} d_i^2 \right) \sigma^2 = Var(b_1) + \sum_{i=1}^{n} d_i^2 \sigma^2.$$

Hence,

$$Var(\hat{\beta}_1^*) = \left(\sum_{i=1}^{n} k_i^2 + \sum_{i=1}^{n} d_i^2\right)\sigma^2 = Var(b_1) + \sum_{i=1}^{n} d_i^2\sigma^2.$$

Which implies

$$Var(\hat{\beta}_1^*) \geq Var(b_1).$$

Hence,

$$Var(\hat{\beta}_1^*) = \left(\sum_{i=1}^{n} k_i^2 + \sum_{i=1}^{n} d_i^2 \right) \sigma^2 = Var(b_1) + \sum_{i=1}^{n} d_i^2 \sigma^2.$$

Which implies

$$Var(\hat{\beta}_1^*) \geq Var(b_1).$$

This proves that the least squares estimator b_1 of β_1 has the minimum variance, i.e., most efficient among all other linear unbiased estimator.