Stat 206: Linear Models

Lecture 3

Oct. 2, 2019

ReCap: Properties of LS Estimators

- **LS estimators are unbiased**: For **all** values of β_0, β_1,

$$E(\hat{\beta}_0) = \beta_0, \quad E(\hat{\beta}_1) = \beta_1.$$

- Variances of $\hat{\beta}_0, \hat{\beta}_1$:

$$\begin{aligned} \sigma^2\{\hat{\beta}_0\} &= \sigma^2 \left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^n (X_i - \overline{X})^2} \right] \\ \sigma^2\{\hat{\beta}_1\} &= \frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2}. \end{aligned}$$

Standard errors (SE) of the LS estimators.

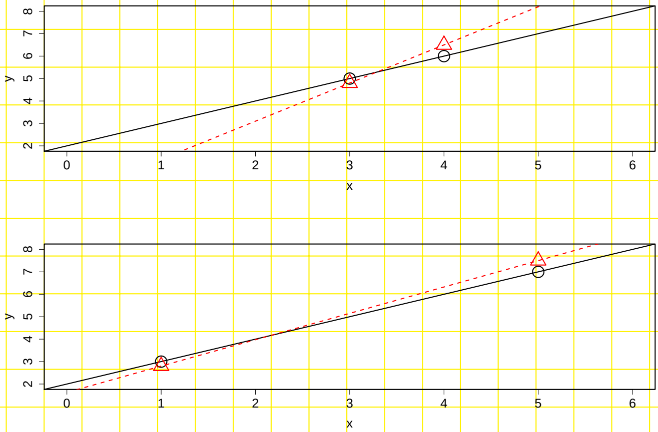
- Replace σ^2 by MSE:

$$
\begin{aligned}
s^2\{\hat{\beta}_0\} &= MSE\left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right], \\
s^2\{\hat{\beta}_1\} &= \frac{MSE}{\sum_{i=1}^n (X_i - \overline{X})^2}.
\end{aligned}
$$

- $s\{\hat{\beta}_0\}$ and $s\{\hat{\beta}_1\}$ are SE of $\hat{\beta}_0$ ad $\hat{\beta}_1$, respectively.
- SEs with the increase of $\sum_{i=1}^n (X_i - \overline{X})^2 = (n-1)s_X^2$, which in turn with the increase of sample size n and sample variance s_X^2 of X.
- SEs tend to with the increase of error variance.

What are the implications?

Figure: Effects of the dispersion of X on the variability of the fitted line



A Simulation Study

Simulate 100 data sets.

- $n = 5$ cases with the X values

$$X_1 = 1.86, \quad X_2 = 0.22, \quad X_3 = 3.55, \quad X_4 = 3.29, \quad X_5 = 1.25,$$

  fixed throughout all data sets.
- For each data set, the response variable is generated by:
  - First generate $\epsilon_1, \cdots, \epsilon_5$ i.i.d. from $N(0,1)$.
  - Then set the response variable as:

$$Y_i = 2 + X_i + \epsilon_i, \quad i = 1, \cdots, 5.$$

- For each data set, derive the LS estimators $\hat{\beta}_0, \hat{\beta}_1$ and MSE.

- Data set 1:

case	X	Y
1	1.86	3.08
2	0.22	2.27
3	3.55	4.38
4	3.29	5.12
5	1.25	1.38

$\hat{\beta}_0 = 1.34,\ \hat{\beta}_1 = 0.94,\ MSE = 0.79.$

- Data set 2:

case	X	Y
1	1.86	2.91
2	0.22	2.13
3	3.55	5.35
4	3.29	5.76
5	1.25	2.01

$\hat{\beta}_0 = 1.19,\ \hat{\beta}_1 = 1.20,\ MSE = 0.52.$

- ..., ...
- Data set 100:

case	X	Y
1	1.86	3.36
2	0.22	2.50
3	3.55	5.93
4	3.29	5.36
5	1.25	2.67

$\hat{\beta}_0 = 1.75,\ \hat{\beta}_1 = 1.09,\ MSE = 0.24.$

Note how the X_is are kept fixed and how the LS estimators vary across these data sets.

Figure: Sampling distributions of $\hat{\beta}_0, \hat{\beta}_1, MSE$. Sample means are 1.99, 1.02, 1.04 respectively. True parameters are 2, 1, 1, respectively.

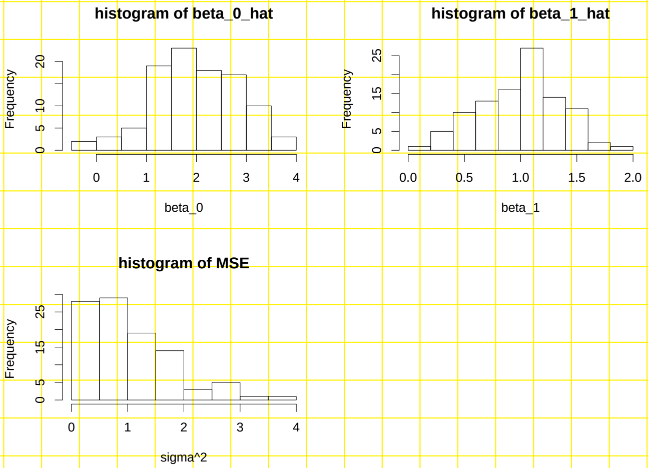
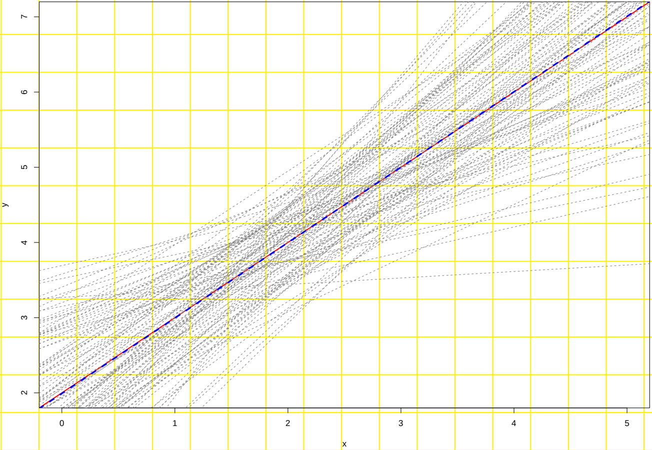


Figure: True: red solid; LS lines: grey broken; mean LS line: blue broken



We calculate the sample mean and sample standard deviation of these 100 realizations of $\hat{\beta}_0, \hat{\beta}_1$, respectively. Then compare them to the respective theoretical values.

- $\hat{\beta}_0$: Theoretical mean and standard deviation:

$$E(\hat{\beta}_0) = \beta_0 = 2, \quad \sigma\{\hat{\beta}_0\} = \sqrt{\sigma^2 \left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^n (X_i - \overline{X})^2} \right]} = 0.854.$$

  Sample mean and sample standard deviation: $1.99, 0.847$.

- $\hat{\beta}_1$: Theoretical mean and standard deviation:

$$E(\hat{\beta}_1) = \beta_1 = 1, \quad \sigma\{\hat{\beta}_0\} = \sqrt{\frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2}} = 0.358.$$

  Sample mean and sample standard deviation: $1.002, 0.36$.

Normal Error Model

Normal error model: Simple regression model $+$ assumption.

- Model equation:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \qquad i = 1, \ldots, n.$$

- Model assumptions: The error terms ϵ_is are

Sampling Distributions of LS Estimators

Under the Normal error model:

- $\hat{\beta}_0, \hat{\beta}_1$ are :

  Notes: Use the facts (i) linear combinations of independent normal random variables are still normal random variables; (ii) $\hat{\beta}_0, \hat{\beta}_1$ are linear combinations of the Y_is.

- SSE/σ^2 follows
- Moreover, SSE is with both $\hat{\beta}_0$ and $\hat{\beta}_1$.

Inference of Regression Coefficients

All inferences are under the Normal error model.

- **Studentized pivotal quantity**:

  where $t_{(n-2)}$ denotes the t-distribution with $n-2$ degrees of freedom.
  - The numerator is the difference between the estimator and the parameter.
  - The denominator is the standard error of the estimator.
  - This quantity follows a known distribution, i.e., the t-distribution.

  Notes: Use the fact that if $Z \sim N(0,1)$, $S^2 \sim \chi^2_{(k)}$ and Z, S^2 are independent, then $\frac{Z}{\sqrt{S^2/k}} \sim t_{(k)}$.

Confidence Interval

$(1-\alpha)$-Confidence interval of β_1:

where $t(1-\alpha/2; n-2)$ is the $(1-\alpha/2)$th percentile of $t_{(n-2)}$.

How to construct confidence intervals for β_0?

Interpretation of Confidence Intervals

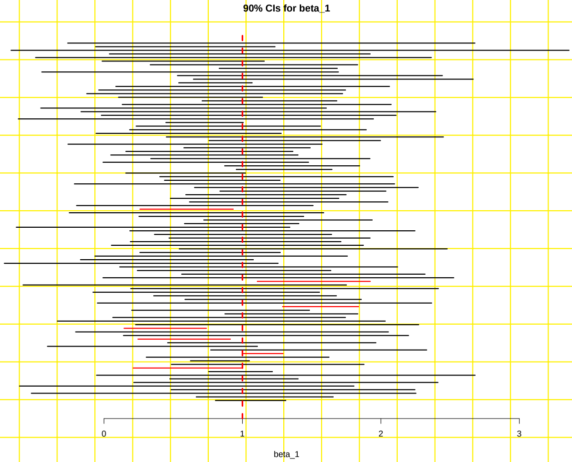


Figure: A Simulation Study

Heights

- Recall $n = 928$, $\overline{X} = 68.316$, $\sum_i X_i^2 = 4334058$, $\sum_{i=1}^n (X_i - \overline{X})^2 = \sum_{i=1}^n X_i^2 - n(\overline{X})^2 = 3038.761$. Also

$$\hat{\beta}_0 = 24.54, \quad \hat{\beta}_1 = 0.637, \quad MSE = 5.031.$$

  So

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

- 95%-confidence interval of β_1:

  We are that the regression slope is in between 0.557 and 0.717.

T-tests

- Null hypothesis: $H_0 : \beta_1 = \beta_1^{(0)}$, where $\beta_1^{(0)}$ is a given constant.
- T-statistic:
- **Null distribution** of the T-statistic:

Can you derive the null distribution?

Decision rule at significance level α.

- **Two-sided alternative** $H_a : \beta_1 \neq \beta_1^{(0)}$: Reject H_0 if and only if $|T^*| > t(1-\alpha/\mathbf{2}; n-2)$, or equivalently, reject H_0 if and only if pvalue:$= P(|t_{(n-2)}| > |T^*|) < \alpha$.
- **Left-sided alternative** $H_a : \beta_1 < \beta_1^{(0)}$: Reject H_0 if and only if $T^* < t(\alpha; n-2)$, or equivalently, reject H_0 if and only if pvalue:$= P(t_{(n-2)} < T^*) < \alpha$.
- **Right-sided alternative** $H_a : \beta_1 > \beta_1^{(0)}$: Reject H_0 if and only if $T^* > t(1-\alpha; n-2)$, or equivalently, reject H_0 if and only if pvalue:$= P(t_{(n-2)} > T^*) < \alpha$.

The decision rule depends on the form of

Why are the critical value approach and the pvalue approach equivalent? How to conduct hypothesis testing with regard to β_0?

Heights

Test whether there is a linear association between parent's height and child's height. Use significance level $\alpha = 0.01$.

- The hypotheses:
- T statistic:
- Critical value:
- Since
- Or the pvalue . Since
- Conclude that