7. Bias of Estimators

a. Recall definition of expectation of random variable
   i. Recall also that expectation of sum is sum of expectations
   ii. Recall that the expectation of constant time variable is constant times expectation.

b. An estimator of a parameter is called unbiased if the expectation of the estimator is the quantity to be estimated.

c. Our analysis will be conditional on explanatory variable
   i. Effectively, we treat these as fixed.

d. Hence

\[
E \left[ \hat{\beta}_1 \right] = E \left[ \frac{\sum_{i=1}^{n} (X_i - \bar{X}) Y_i}{\sum_{j=1}^{n} (X_j - \bar{X})^2} \right] \\
= \sum_{i=1}^{n} (X_i - \bar{X}) E [Y_i] / \sum_{j=1}^{n} (X_j - \bar{X})^2 \\
= \sum_{i=1}^{n} (X_i - \bar{X})(\beta_0 + \beta_1 X_i) / \sum_{j=1}^{n} (X_j - \bar{X})^2 \\
= \beta_1 \sum_{i=1}^{n} (X_i - \bar{X}) X_i / \sum_{j=1}^{n} (X_j - \bar{X})^2 = \beta_1
\]

i. \( \hat{\beta}_1 \) is unbiased.
e. $\hat{\beta}_0$ is unbiased

i. \[
E \left[ \hat{\beta}_0 \right] = E \left[ \bar{Y} - \hat{\beta}_1 \bar{X} \right] \\
= E \left[ \bar{Y} \right] - E \left[ \hat{\beta}_1 \right] \bar{X} \\
= \sum_{i=1}^{n} E \left[ Y_i \right] / n - \beta_1 \bar{X} \\
= \sum_{i=1}^{n} E \left[ \beta_0 + \beta_1 X_i \right] / n - \beta_1 \bar{X} \\
= \beta_0 + \beta_1 \bar{X} - \beta_1 \bar{X}.
\]

MPV: 2.3-2.3.2

C. Inference on Parameters

1. Approximate Normality of Estimates

a. Both estimates are sums of multipliers times observations

i. If the observations are exactly normal, the estimates are exactly normal.

ii. If the observations are not far from normal,

- Intercept involves the sample average, and central limit theorem implies this is approximately normal.
- Slope parameter is approximately normal, as long as the
multipliers aren’t dominated by a few (Lindeberg-Levy CLT).

2. Variances of Estimates
   a. \( \text{Var} \left[ \hat{\beta}_1 \right] = \text{Var} \left[ \sum_{j=1}^{n} c_j Y_j \right] = \sum_{j=1}^{n} c_j^2 \text{Var} \left[ Y_j \right] = \sum_{j=1}^{n} c_j^2 \sigma^2 = \sigma^2 / S_{xx} . \)
   b. Note that intercept is fitted value at zero.
      i. More generally, calculate general fitted value at \( x \)
      ii. \( \text{Var} \left[ \hat{Y}(x) \right] = \text{Var} \left[ \bar{Y} - \beta_1 \bar{X} + x\beta_1 \right] = \text{Var} \left[ \bar{Y} \right] + \text{Var} \left[ \hat{\beta}_1 \right] (x - \bar{X})^2 + 2(x - \bar{X}) \text{Cov} \left[ \bar{Y}, \hat{\beta}_1 \right] \)
      iii. \( \text{Cov} \left[ \bar{Y}, \hat{\beta}_1 \right] = \sum_{j=1}^{n} \text{Cov} \left[ 1/n Y_j, c_j Y_j \right] + 2 \sum_{j<i} \text{Cov} \left[ 1/n Y_j, c_i Y_i \right] = \sum_{j=1}^{n} 1/nc_j \sigma^2 = 0 . \)
      iv. Hence covariance between \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) is zero if covariate is first centered to have mean zero
      v. \( \text{Var} \left[ \hat{Y}(x) \right] = \sigma^2 (1/n + (x - \bar{X})^2 / S_{xx}) . \)
      vi. \( \text{Var} \left[ \hat{\beta}_0 \right] = \sigma^2 (1/n + \bar{X}^2 / S_{xx}) . \)

3. Gauss-Markov Theorem:
   a. Under assumptions of uncorrelated errors, constant variance, least squares estimates are ones with smallest variance.
   b. Gauss (1777-1855), Markov (1856-1922) never worked together.
Lecture 2

4. Hypothesis Tests

   a. Use $\hat{\beta}_j \sim N(\beta_j, \text{Var} \left[ \hat{\beta}_j \right])$

   b. Note that for both $j = 0, 1$, variance is a known multiple of $\sigma^2$:
      \[ \text{Var} \left[ \hat{\beta}_j \right] = d_j \sigma^2 \]

   c. If you knew $\sigma$, use it and the normal distribution for inference

      i. Ex., $1 - \alpha$ CI for $\beta_j$ is $\hat{\beta}_j \pm z_{\alpha/2} \sigma \sqrt{d_j}$.

   d. If you don’t know $\sigma$,

      i. use unbiased estimator $\hat{\sigma}$,  

      ii. and substitute $t$-distribution for normal.

      • $z_{\alpha/2}$ replaced by $t_{\alpha/2, n-2}$

   iii. Works because

      • $(\hat{\beta}_j - \beta_j)/\left( \sqrt{d_j} \sigma \right) \sim N(0, 1)$,
      • $(n - 2) \hat{\sigma}^2 / \sigma^2 \sim \chi^2_{n-2}$
      • $\hat{\sigma}^2 \perp \hat{\beta}_j$.

   iv. Last two to be proven more generally later.

5. Inference

   a. A quantity that one can divide the difference between a statistic $T$ minus its hypothesized mean to get something approximately normal is called a standard error
Lecture 2

i. Hence the standard error is $\text{SE} \left[ \hat{\beta}_j \right] = \sqrt{d_j \hat{\sigma}}$.

ii. In this case, but not in general, $\text{SE} \left[ \hat{\beta}_j \right]$ approximates $\sqrt{\text{Var} \left[ \hat{\beta}_j \right]}$.

b. Reject $H_0 : \beta_j = \beta_j^\circ$ if $(\hat{\beta}_j - \beta_j^\circ) / \text{SE} \left[ \hat{\beta}_j \right] \geq z_{\alpha/2}$ or $t_{\alpha/2}$

i. $H_0$ represents null hypothesis. See Fig. 3.

Fig. 3: Area under Normal Density

Area is $\alpha/2$: Input

Desired area ends here: Output $z_{\alpha/2}$
ii. In particular, $\beta_j^0 = 0$ coincides to the regressor having no impact.

c. Construct $p$-value by taking the area above the normal or $t$ quantile.

i. As is done in simpler contexts. See Fig. 4.

*Fig. 4: Area under Normal Density giving p-value*
a. Confidence interval for $\beta_j$
   
   i. Choose confidence level $1 - \alpha$.
   
   ii. Confidence Interval is a data-dependent set of parameter values $I$ such that for any parameter value,
   
   $$P_{\beta_j}[\beta_j \in I] \geq 1 - \alpha.$$  
   
   iii. Solution is
   
   - $I = (\hat{\beta}_j - z_{\alpha/2}SE[\hat{\beta}_j], \hat{\beta}_j + z_{\alpha/2}SE[\hat{\beta}_j])$ for $\sigma$ known,
   
   - or with $t$ substituted if $\sigma$ estimated by $\hat{\sigma}$.
   
   MPV: 2.5

7. Fitted Values
   a. Recall $\hat{Y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$ is the fitted value for covariate value $x$.
   
   i. $\text{Var}[\hat{\beta}_0 + \hat{\beta}_1 x] = \sigma^2\left\{1/n + (x - \bar{X})^2/S_{xx}\right\}$
   
   ii. Hence $\text{Var}[\hat{\beta}_0 + \hat{\beta}_1 x] = \sigma^2\left\{1/n + (x - \bar{X})^2/S_{xx}\right\}$

8. Predicted Values
   a. Predict a new response at explanatory level $x$
   
   b. A new value is $\beta_0 + \beta_1 x + \epsilon$
   
   c. Estimate this as $\hat{\beta}_0 + \hat{\beta}_1 x$: Same as fitted value
   
   i. As before $\text{Var}[\hat{\beta}_0 + \hat{\beta}_1 x] = \sigma^2\left\{1/n + (x - \bar{X})^2/S_{xx}\right\}$
Lecture 2

ii. \( \text{Var}[\epsilon] = \sigma^2 \)

iii. New \( \epsilon \perp \hat{\beta}_0, \hat{\beta}_1 \)

iv. Hence \( \text{Var} \left[ \hat{\beta}_0 + \hat{\beta}_1 x + \epsilon \right] = \sigma^2 \left\{ 1/n + (x - \bar{X})^2 / S_{xx} + 1 \right\} \)

9. An Identity for Sums of Squares

a. Definitions

i. Recall \( SS_{Res} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \) sum of squared residuals

ii. Recall \( SS_t = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \) total sum of squares

iii. Define \( SS_R = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 \) total sum of squares

b. Residuals and differences of fitted values from mean are orthogonal: \( \sum_{i=1}^{n} (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) = 0 \).

i. Because

\[
\sum_{i=1}^{n} (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) = \sum_{i=1}^{n} (Y_i - \bar{Y} + \bar{Y} - \hat{Y}_i)\hat{\beta}_1(X_i - \bar{X})
\]

\[
= \sum_{i=1}^{n} (Y_i - \bar{Y})\hat{\beta}_1(X_i - \bar{X}) - \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})\hat{\beta}_1(X_i - \bar{X})
\]

\[
= \hat{\beta}_1 \sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X}) - \hat{\beta}_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2 = 0
\]

c. Total sum of squares is sum of regression and residual sum of
Lecture 2

squares

i. Because

\[ SS_t = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \]

\[ = \sum_{i=1}^{n} ((Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y}))^2 \]

\[ = \sum_{i=1}^{n} [(Y_i - \hat{Y}_i)^2 + (\hat{Y}_i - \bar{Y})^2 + 2(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y})] \]

\[ = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 \]

\[ = SS_R + SS_{Res} \]

ii. Cross term is zero by previous result.

10. Coefficient of Determination

a. Heuristically, \( SS_R / SS_t \) represents the proportion of variability present in the original data

i. measured about the mean

ii. which is explained by the regression model

b. Called the Coefficient of Determination

c. Denoted by \( R^2 \geq 0 \).

d. Properties come from the decomposition identity.

i. \( 1 \geq R^2 \).

ii. Equals 1 \( - SS_{Res} / SS_t \), the proportion of variability left over.
e. Answers a related, but different, question from a statistical test of null hypothesis $H_0 : \beta_1 = 0$

i. $R^2$ reflects quality of model

ii. Test reflects whether null model is adequate.

f. Can get low $R^2$ but reject $H_0$ if large data set reflects a weak relationship.

g. Can get high $R^2$ but not reject $H_0$ if small data set reflects a strong relationship.

MPV: 2.10

11. Regression through the Origin

a. Suppose that one wants to consider only models for which the constant term is zero.

b. Then one minimizes $\sum_i (Y_i - \beta_1 X_i)^2$ with respect to $\beta_1$

i. Set $-2 \sum_i (Y_i - \beta_1 X_i)X_i = 0$

ii. Then $\sum_i Y_i X_i = \beta_1 \sum_i X_i^2$

iii. Then $\hat{\beta}_1 = \sum_i Y_i X_i / \sum_i X_i^2$.

iv. Second derivative is $\sum_i X_i^2 > 0$, and so this is a minimizer.

c. Estimator is unbiased

i. $E\left[\hat{\beta}_1\right] = E\left[\sum_i Y_i X_i / \sum_i X_i^2\right] = \sum_i E[Y_i] X_i / \sum_i X_i^2 = \hat{\beta}_1$
\[ \sum_i E[\beta_1 X_i] X_i / \sum_i X_i^2 = \beta_1: \text{ Unbiased} \]

d. \[ \text{Var} \left[ \hat{\beta}_1 \right] = \text{Var} \left[ \sum_i Y_i X_i / \sum_i X_i^2 \right] = \sum_i \text{Var} \left[ Y_i \right] X_i^2 / (\sum_i X_i^2)^2 = \sigma^2 / \sum_i X_i^2 \]

e. \[ E \left[ SS_{Res} \right] = (n - 1)\hat{\sigma}^2, \text{ similarly with the intercept-present case} \]

i. \[ \hat{\sigma}^2 = SS_{Res} / (n - 1) \]

ii. Divisor is once again the number of parameters estimated before the dispersion parameter.

MPV: 2.11

12. Least squares linear estimates coincide with maximum likelihood

a. Likelihood is the probability of observing data

i. viewed as a function of the unknown parameter

ii. For the simple regression model likelihood is

\[
\prod_{i=1}^{n} \exp\left( -\frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2} \right) \left(2\pi\right)^{-1/2} / \sigma
\]

\[ = \exp\left( -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \right) \left(2\pi\right)^{-n/2} \sigma^{-n} \]

b. Maximum likelihood estimators are those that maximize likelihood

i. Often easier to maximize the log of the likelihood
ii. For simple linear regression model \( \ell(\beta_0, \beta_1, \sigma) = \)

\[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 - (n/2) \log(2\pi) - n \log(\sigma).\]

iii. So \( \beta_0, \beta_1 \) maximizing this minimize the sum of squares.

iv. For \( \sigma \),

- \( \frac{\partial}{\partial \sigma} \ell = \frac{SS_{Res}}{\sigma^3} - \frac{n}{\sigma} \),
- \( \frac{\partial^2}{\partial \sigma^2} \ell = -3 \frac{SS_{Res}}{\sigma^4} + \frac{n}{\sigma^2} \), negative if \( \sigma^2 = \frac{SS_{Res}}{n} \)
- So maximizer is \( \hat{\sigma} = \sqrt{\frac{SS_{Res}}{n}} \).
- This is biased.

c. Can demonstrate (960:583) that maximum likelihood has some optimality properties

i. So by this measure, least squares is optimal for the normal regression. desirable properties