

PLSC 503 – Spring 2017

Bivariate Regression II: Inference

January 24, 2017

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

Estimators:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

and

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^N (X_i - \bar{X})^2} \\ &= \frac{\sum (X_i - \bar{X}) Y_i}{\sum (X_i - \bar{X})^2}\end{aligned}$$

$$\text{Var}(\hat{\beta}_1)$$

$$u_i \sim \text{i.i.d. } N(0, \sigma^2)$$

meaning:

$$\text{Var}(Y|X, \beta) = \sigma^2$$

so:

$$\begin{aligned} \text{Var}(\hat{\beta}_1) &= \text{Var} \left[\frac{\sum_{i=1}^N (X_i - \bar{X}) Y_i}{\sum_{i=1}^N (X_i - \bar{X})^2} \right] \\ &= \left[\frac{1}{\sum (X_i - \bar{X})^2} \right]^2 \sum (X_i - \bar{X})^2 \text{Var}(Y) \\ &= \left[\frac{1}{\sum (X_i - \bar{X})^2} \right]^2 \sum (X_i - \bar{X})^2 \sigma^2 \\ &= \frac{\sigma^2}{\sum (X_i - \bar{X})^2}. \end{aligned}$$

$$\text{Var}(\hat{\beta}_0) \text{ and } \text{Cov}(\hat{\beta}_0, \hat{\beta}_1)$$

Similarly:

$$\text{Var}(\hat{\beta}_0) = \frac{\sum X_i^2}{N \sum (X_i - \bar{X})^2} \sigma^2$$

and :

$$\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\bar{X}}{\sum (X_i - \bar{X})^2} \sigma^2$$

Important Things

- $\text{Var}(\hat{\beta}_0)$ and $\text{Var}(\hat{\beta}_1) \propto \sigma^2$
- $\text{Var}(\hat{\beta}_0)$ and $\text{Var}(\hat{\beta}_1) \propto -\sum (X_i - \bar{X})$
- $\text{Var}(\hat{\beta}_0)$ and $\text{Var}(\hat{\beta}_1) \propto -N$
- $\text{sign}[\text{Cov}(\hat{\beta}_0, \hat{\beta}_1)] = \text{sign}(\bar{X})$

Gauss-Markov Theorem

Imagine:

$$\hat{\beta}_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

Rewrite:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^N (X_i - \bar{X}) Y_i}{\sum_{i=1}^N (X_i - \bar{X})^2}.$$

k are “weights”:

$$\hat{\beta}_1 = \sum k_i Y_i$$

with $k_i = \frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2}$.

Gauss-Markov (continued)

Alternative (non-LS) estimator:

$$\tilde{\beta}_1 = \sum w_i Y_i$$

Unbiasedness requires:

$$\begin{aligned} E(\tilde{\beta}_1) &= \sum w_i E(Y_i) \\ &= \sum w_i (\beta_0 + \beta_1 X_i) \\ &= \beta_0 \sum w_i + \beta_1 \sum w_i X_i \end{aligned}$$

Gauss-Markov (continued)

Variance:

$$\begin{aligned}\text{Var}(\tilde{\beta}_1) &= \text{Var}\left(\sum w_i Y_i\right) \\ &= \sigma^2 \sum w_i^2 \\ &= \sigma^2 \sum \left[w_i - \frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2} + \frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2} \right]^2 \\ &= \sigma^2 \sum \left[w_i - \frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2} \right]^2 + \sigma^2 \left[\frac{1}{\sum (X_i - \bar{X})^2} \right]\end{aligned}$$

Gauss-Markov (continued)

Because $\sigma^2 \left[\frac{1}{\sum (X_i - \bar{X})^2} \right]$ is a constant, $\min[\text{Var}(\tilde{\beta}_1)]$ minimizes

$$\sum \left[w_i - \frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2} \right]^2.$$

Minimized at:

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

implying:

$$\begin{aligned} \text{Var}(\tilde{\beta}_1) &= \frac{\sigma^2}{\sum (X_i - \bar{X})^2} \\ &= \text{Var}(\hat{\beta}_1) \end{aligned}$$

If $u_i \sim N(0, \sigma^2)$, then:

$$\hat{\beta}_0 \sim N[\beta_0, \text{Var}(\hat{\beta}_0)]$$

and

$$\hat{\beta}_1 \sim N[\beta_1, \text{Var}(\hat{\beta}_1)]$$

Means:

$$\begin{aligned} z_{\hat{\beta}_1} &= \frac{(\hat{\beta}_1 - \beta_1)}{\sqrt{\text{Var}(\hat{\beta}_1)}} \\ &= \frac{(\hat{\beta}_1 - \beta_1)}{\text{s.e.}(\hat{\beta}_1)} \\ &= \sim N(0, 1) \end{aligned}$$

A Small Problem...

$$\sigma^2 = ???$$

Solution: use

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{N - k}$$

Gives:

$$\widehat{\text{Var}}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\sum (X_i - \bar{X})^2},$$

and

$$\widehat{\text{Var}}(\hat{\beta}_0) = \frac{\sum X_i^2}{N \sum (X_i - \bar{X})^2} \hat{\sigma}^2$$

Inference (continued)

$$\begin{aligned}\widehat{\text{s.e.}}(\hat{\beta}_1) &= \sqrt{\widehat{\text{Var}}(\hat{\beta}_1)} \\ &= \sqrt{\frac{\hat{\sigma}^2}{\sum (X_i - \bar{X})^2}} \\ &= \frac{\hat{\sigma}}{\sqrt{\sum (X_i - \bar{X})^2}}\end{aligned}$$

implies:

$$\begin{aligned}t_{\hat{\beta}_1} &\equiv \frac{(\hat{\beta}_1 - \beta_1)}{\widehat{\text{s.e.}}(\hat{\beta}_1)} = \frac{(\hat{\beta}_1 - \beta_1)}{\frac{\hat{\sigma}}{\sqrt{\sum (X_i - \bar{X})^2}}} \\ &= \frac{(\hat{\beta}_1 - \beta_1) \sqrt{\sum (X_i - \bar{X})^2}}{\hat{\sigma}} \\ &\sim t_{N-k}\end{aligned}$$

Predictions and Variance

Point prediction:

$$\hat{Y}_k = \hat{\beta}_0 + \hat{\beta}_1 X_k$$

Y_k is unbiased:

$$\begin{aligned} E(\hat{Y}_k) &= E(\hat{\beta}_0 + \hat{\beta}_1 X_k) \\ &= E(\hat{\beta}_0) + X_k E(\hat{\beta}_1) \\ &= \beta_0 + \beta_1 X_k \\ &= E(Y_k) \end{aligned}$$

Variability:

$$\begin{aligned} \text{Var}(\hat{Y}_k) &= \text{Var}(\hat{\beta}_0 + \hat{\beta}_1 X_k) \\ &= \frac{\sum X_i^2}{N \sum (X_i - \bar{X})^2} \sigma^2 + \left[\frac{\sigma^2}{\sum (X_i - \bar{X})^2} \right] X_k^2 + 2 \left[\frac{-\bar{X}}{\sum (X_i - \bar{X})^2} \sigma^2 \right] X_k \\ &= \sigma^2 \left[\frac{1}{N} + \frac{(X_k - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right] \end{aligned}$$

Variability of Predictions

$$\text{Var}(\hat{Y}_k) = \sigma^2 \left[\frac{1}{N} + \frac{(X_k - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right]$$

means that $\text{Var}(\hat{Y}_k)$:

- Decreases in N
- Decreases in $\text{Var}(X)$
- Increases in $|X - \bar{X}|$

Predictions and Inference

Standard error of the prediction:

$$\widehat{\text{s.e.}}(\hat{Y}_k) = \sqrt{\sigma^2 \left[\frac{1}{N} + \frac{(X_k - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right]}$$

→ (e.g.) confidence intervals:

$$95\% \text{ c.i.}(\hat{Y}_k) = \hat{Y}_k \pm [1.96 \times \widehat{\text{s.e.}}(\hat{Y}_k)]$$

Back to the Example

```
> IMdata<-na.omit(Data[c("infantmortalityperK","DPTpct")])
> IMDPT<-with(Data,lm(infantmortalityperK~DPTpct,na.action=na.exclude))
> summary(IMDPT)
```

Call:

```
lm(formula = infantmortalityperK ~ DPTpct, data = IMdata)
```

Residuals:

Min	1Q	Median	3Q	Max
-56.8	-16.3	-5.1	11.8	86.6

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	173.277	8.489	20.4	<2e-16 ***
DPTpct	-1.576	0.101	-15.6	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26.2 on 175 degrees of freedom

Multiple R-Squared: 0.582, Adjusted R-squared: 0.58

F-statistic: 244 on 1 and 175 DF, p-value: <2e-16

$\text{Var}(\hat{\beta})$:

```
> vcov(IMDPT)
```

	(Intercept)	DPTpct
(Intercept)	72.0677	-0.83317
DPTpct	-0.8332	0.01018

95 percent c.i.s:

```
> confint(IMDPT)
```

	2.5 %	97.5 %
(Intercept)	156.523	190.032
DPTpct	-1.775	-1.377

Predictions

```
> SEs<-predict(IMDPT,interval="confidence")
> SEs
```

	fit	lwr	upr
1	25.10	20.53	29.68
3	17.22	12.05	22.40
4	23.53	18.84	28.21
.			
.			
<rows omitted>			
.			
.			
189	21.95	17.15	26.75
190	39.29	35.36	43.23
191	17.22	12.05	22.40

A Plot, With CIs

Scatterplot of Infant Mortality and DPT Immunizations, along with Least-Squares Line and 95% Prediction Confidence Intervals

