Simple Linear Regression: Confidence and Prediction Intervals

BIOS 6611

CU Anschutz

Week 7

Prediction and Estimation

Prediction and Estimation

Reference Range (Population Confidence Interval): description of the variability in the underlying population (usually the central 95% of the population), generally estimated from large samples of individuals representative of the population.

Confidence Interval: description of the variability in our sample estimate of the true underlying mean (or any other population parameter).

Example

In our FEV example we know $\bar{Y} = 2.64$, $\hat{\sigma}_Y = 0.867$, and $n = 654$.

What is the expected FEV for a single child ($\hat{Y}$)?

What is the 95% reference range? (Note: need to assume FEV is normally distributed.)

What is our estimate of the true underlying mean FEV in children ($\hat{\mu}$)?

What is the 95% CI? (Note: FEV doesn't need to be normal if CLT applies.)

Prediction and Estimation in SLR

Estimation and Prediction

In regression, we assume that the underlying mean changes according to the level of an explanatory variable(s).

Estimation: The expected mean (average) μ for a given value of X, say X_0, in the underlying population is:

$$\hat{\mu}_{Y|X_0} = \hat{\beta}_0 + \hat{\beta}_1 X_0$$

The standard error of the estimate is given by:

$$SE(\hat{\mu}_{Y|X_0}) = \sqrt{\frac{\hat{\sigma}^2_{Y|X}}{n} + \frac{\hat{\sigma}^2_{Y|X}}{n-1}\left(\frac{(X_0 - \bar{X})^2}{\hat{\sigma}^2_X}\right)}$$

Estimation and Prediction

Prediction: The predicted value of Y for a given value of X, say X_0, for a randomly selected individual from the underlying population is:

$$\hat{Y}|X_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0$$

Its standard error is given by:

$$SE(\hat{Y}|X_0) = \sqrt{\hat{\sigma}^2_{Y|X} + \frac{\hat{\sigma}^2_{Y|X}}{n} + \frac{\hat{\sigma}^2_{Y|X}}{n-1}\left(\frac{(X_0 - \bar{X})^2}{\hat{\sigma}^2_X}\right)}$$

This is broken down as the variance of the individual around $\mu_{Y|X}$ and the variance in estimating $\mu_{Y|X}$.

Confidence Intervals and Prediction Intervals

Confidence Interval: A 95% confidence interval describes the variability in the estimate of the underlying mean. It can be calculated (without assuming normality of the errors) as:

$$\hat{\mu}_{Y|X} \pm t_{n-2,0.975} SE(\hat{\mu}_{Y|X})$$

Prediction Interval: A 95% prediction interval (like a reference range) describes the variability in the underlying population. It can be calculated (assuming normality of the errors) as:

$$(\hat{Y}|X) \pm t_{n-2,0.975} SE(\hat{Y}|X)$$

Prediction intervals will be wider than confidence intervals since there is more variability around estimating an individual point as compared to the mean (i.e., prediction intervals take the true error term into account).

FEV in Children Example

```
fev <- read.csv('FEV_rosner.csv', header=T)
reg_out <- lm(fev ~ age, data=fev) #save regression results
sum( reg_out$residuals^2 ) / (nrow(fev)-2) # calculate MSE
```

```
## [1] 0.3220869
```

```
sum( fev$age ) # calculate sum(x_i)
```

```
## [1] 6495
```

```
sum( fev$age^2 ) # calculate sum(x_i^2)
```

```
## [1] 70201
```

```
mean( fev$age ) # calculate mean of X
```

```
## [1] 9.931193
```

```
var( fev$age ) # calculate variance of X
```

```
## [1] 8.725733
```

FEV in Children Example - CI

Calculate the 95% confidence interval for the underlying mean FEV among all children aged 16.

First we calculate our mean:

$$\hat{\mu}_{Y|X=16} = 0.43165 + 0.22204\,(16) = 3.98 \text{ liters}$$

Then we calculate our SE:

$$\begin{aligned} SE\left(\hat{\mu}_{Y|X=16}\right) =& \sqrt{\frac{\hat{\sigma}^2_{Y|X}}{n} + \frac{\hat{\sigma}^2_{Y|X}}{n-1}\left(\frac{(X_0 - \bar{X})^2}{\hat{\sigma}^2_X}\right)} \\ =& \sqrt{\frac{0.32209}{654} + \frac{0.32209}{653}\left(\frac{(16 - 9.931)^2}{8.726}\right)} \\ \approx& 0.05074 \end{aligned}$$

Considering $t_{654-2,0.975} = 1.963609$, our 95% CI is

$$3.98 \pm 1.96(0.05074) = (3.885, 4.989)$$

FEV in Children Example - PI

Calculate the 95% prediction interval in a randomly selected child aged 16.

First we calculate our mean:

$$(\hat{Y}|X = 16) = 0.43165 + 0.22204\,(16) = 3.98 \text{ liters}$$

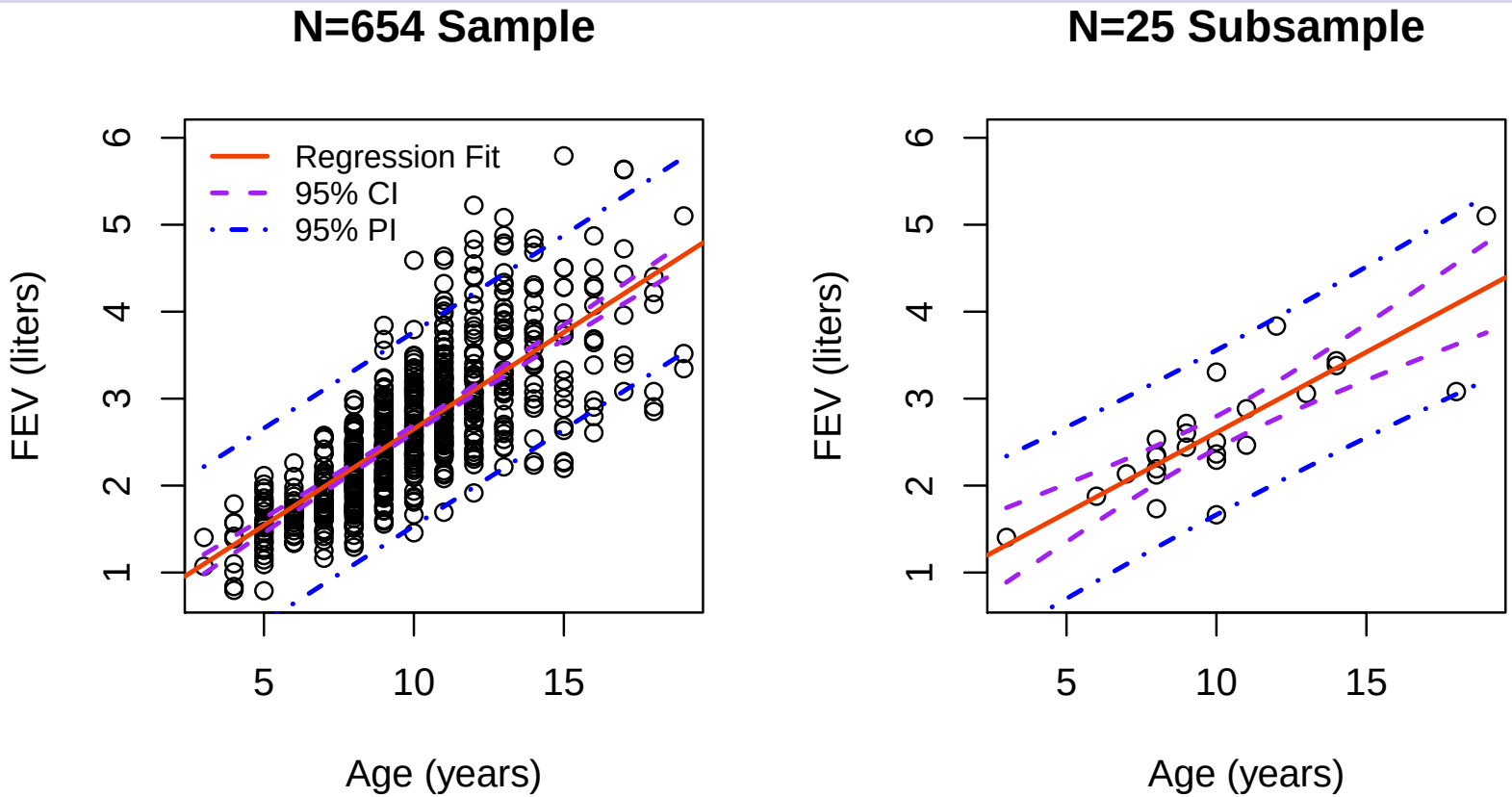
Then we calculate our SE:

$$\begin{aligned} SE\left(\hat{Y}|X = 16\right) =& \sqrt{\hat{\sigma}^2_{Y|X} + \frac{\hat{\sigma}^2_{Y|X}}{n} + \frac{\hat{\sigma}^2_{Y|X}}{n-1}\left(\frac{(X_0 - \bar{X})^2}{\hat{\sigma}^2_X}\right)} \\ =& \sqrt{0.32209 + \frac{0.32209}{654} + \frac{0.32209}{653}\left(\frac{(16 - 9.931)^2}{8.726}\right)} \\ \approx& 0.569793 \end{aligned}$$

Considering $t_{654-2,0.975} = 1.963609$, our 95% PI is

$$3.98 \pm 1.96(0.569793) = (2.867, 5.101)$$

FEV in Children Example



From these two figures of the entire data set and a small subsample, we can note a few interesting observations:

1. The change in slope has a larger impact on the intervals, especially CI, at the extremes. This is because when $X_0 = \bar{X}$ the standard error is minimized.
2. A smaller sample size more greatly impacts the confidence interval width than the prediction interval width.

Properties of Prediction Intervals and Confidence Intervals

Prediction intervals are wider than confidence intervals.

Both prediction intervals and confidence intervals are wider at the ends of the data (the SEs are at a minimum at $\bar{X}$).

Confidence intervals shrink considerably as the sample size grows.
Prediction intervals stay about the same as the sample size grows.

Large samples are generally required if prediction intervals are to be used as reference ranges.