Estimation and Prediction

Bret Larget

Departments of Botany and of Statistics
University of Wisconsin—Madison

January 30, 2007

The Big Picture

- The least squares regression line is an estimate of the true relationship between the explanatory variable x and the response variable y.
- The accuracy of the estimate is not the same at all x.
- The estimate of the mean $\mu_x = E(y \mid x)$ is less variable than a prediction of y for an individual with a given x.
- Estimation accounts for uncertainty in the regression line.
- Prediction accounts for uncertainty in the regression line *and in the individual observation*.

Standard Error

- The point estimate of response y at explanatory variable x for both estimation ($\hat{\mu}_x$) and prediction ($\hat{y}$) is $\hat{\beta}_0 + \hat{\beta}_1 x$.
- The standard error for estimation of μ_x at x is

$$\mathrm{SE}(\hat{\mu}_x) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

where $\hat{\sigma} = \sqrt{\frac{1}{n-2}\sum_{i=1}^n (y_i - \hat{y}_i)^2}$.

- The standard error for prediction of y at x is

$$\mathrm{SE}(\hat{y}) = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

Remarks

$$\mathrm{SE}(\hat{\mu}_x) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \qquad \mathrm{SE}(\hat{y}) = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

- Notice the only difference is that $\mathrm{SE}(\hat{y})$ has an extra 1.
- As the sample size n goes to infinity, $\mathrm{SE}(\hat{\mu}_x)$ tends to 0 but $\mathrm{SE}(\hat{y})$ tends to σ.
- For fixed n, both standard errors are smaller when the x values are more spread out.
- Estimation/prediction near $\bar{x}$ is more accurate than further away.

Estimation and Prediction Using R

- The `predict` function in R is used for both estimation and prediction.
- The first argument is a linear model object, created with R function `lm`.
- The second argument is a data frame holding the explanatory variable values where estimation/prediction is desired.
- The third argument specifies the type of standard error, `none`, `confidence`, or `prediction`.
- Reconsider the soil phosphorous data.

soilP	1	4	5	9	13	11	23	23	28
plantP	64	71	54	81	93	76	77	95	109

Estimation and Prediction Using R

- Consider estimates at $x = 10$ (near the mean $\bar{x} = 13$) and $x = 25$ (further away).

```
> soilP = c(1, 4, 5, 9, 13, 11, 23, 23, 28)
> plantP = c(64, 71, 54, 81, 93, 76, 77, 95, 109)
> x = data.frame(soilP = c(10, 25))
> phos.lm = lm(plantP ~ soilP)
> predict(phos.lm, x, interval = "none")

       1        2
75.74932 97.00272

> predict(phos.lm, x, interval = "confidence")

       fit      lwr       upr
1 75.74932 66.86786  84.63078
2 97.00272 82.98577 111.01968

> predict(phos.lm, x, interval = "prediction")

       fit      lwr      upr
1 75.74932 48.94919 102.5494
2 97.00272 68.09180 125.9136
```