Inference and prediction in SLR

(STATS544.1: Applied simple linear regression)

Anna Helga Jonsdottir and Gunnar Stefansson

September 22, 2012

Introduction

The following slides will focus on inference in SLR.

- Expected values, variances and distribution of estimators in SLR
- Confidence intervals for β_0 and β_1
- Hypothesis test for β_1
- Prediction interval for new measurements

The linear regression model

Recall that if we have n paired measurements $(x_1, y_1), \ldots, (x_n, y_n)$, the regression model can be written as

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

- β_0 is the true intercept (population intercept) that we do not know the value of.
- β_1 is the true slope (population slope)
- ϵ_i are the errors where $\epsilon \sim n(0, \sigma^2)$ and

$$\hat{\sigma}^2 = \frac{\sum (y - \hat{y})^2}{n - 2}$$

β_0 and β_1 are therefore statistics, that we both want to estimate and make inference on.

Expected value of the slope estimator *

When assumption hold, the estimator for the slope is unbiased.
Notation: $E[\hat{\beta}_1] = \beta_1$.

Variance of the slope estimator *

The variance of the estimator can be derived:

$$V\left[\hat{\beta}_1\right] = \ldots = \frac{\sigma^2}{\sum(x-\bar{x})^2}$$

Given an estimate of the unknown σ^2, this variance can also be estimated.

Note: This result only depends on the mean and variance structure of Y_i, not the p.d.f.

A proof based on simple rewrite is not difficult, but may be omitted.

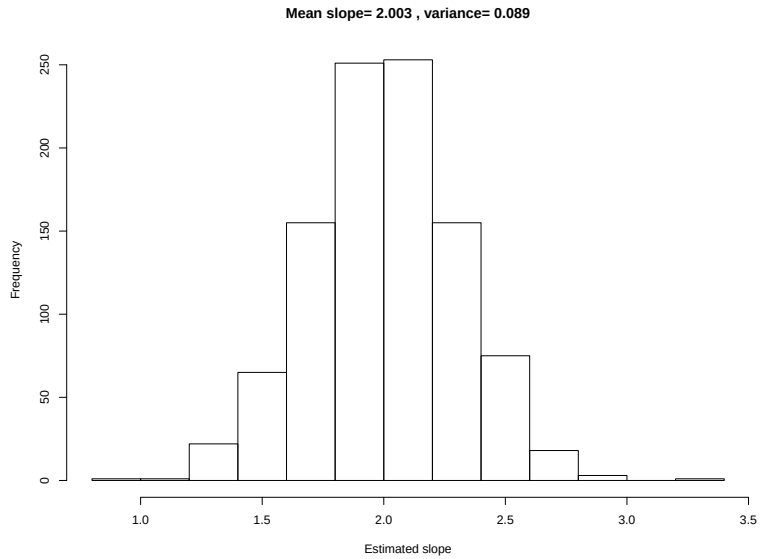

Image is done with

```
set.seed(19)
x<-c(2,3,1,0,4)
out<-NULL
```

Expected value of the intercept estimator *

The estimate of the intercept is unbiased:

$$\begin{aligned} E[\hat{\beta}_0] &= E\left[\bar{Y} - \hat{\beta}_1 \bar{x}\right] \\ &= E\left[\bar{Y}\right] - \beta_1 \bar{x} \\ &= (\beta_0 + \beta_1 \bar{x}) - \beta_1 \bar{x} \\ &= \beta_0. \end{aligned}$$

Note: This result only depends on the mean and variance structure of Y_i, not the p.d.f.

Variance of intercept estimator *

The variance of the estimator can be derived:

$$V\left[\hat{\beta}_0\right] = \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\right)\sigma^2$$

Note: This result only depends on the mean and variance structure of Y_i, not the p.d.f.

Marginal distribution of estimator of slope

Recall that

$$E[\hat{\beta}_1] = \beta_1$$

and

$$V\left[\hat{\beta}_1\right] = \frac{\sigma^2}{\sum(x - \bar{x})^2}$$

Under normality, the estimator also has a Gaussian (normal) distribution:

$$\hat{\beta}_1 \sim n\left(\beta_1, \frac{\sigma^2}{\sum(x - \bar{x})^2}\right)$$

Marginal distribution of estimator of intercept

Recall that

$$E[\hat{\beta}_0] = \beta_0$$

and

$$V\left[\hat{\beta}_0\right] = \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)\sigma^2$$

Under normality, the estimator also has a Gaussian (normal) distribution:

$$\hat{\beta}_0 \sim n\left(\beta_0, \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)\sigma^2\right)$$

Estimating slope and intercept accuracy

The standard error of the slope is:

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{\hat{\sigma}^2}{\sum(x - \bar{x})^2}$$

and the standard error of the intercept is:

$$\hat{\sigma}^2_{\hat{\beta}_0} = \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) \hat{\sigma}^2$$

where

$$\hat{\sigma}^2 = \frac{\sum(y - \hat{y})^2}{n - 2}$$

Elements of inference in simple linear regression

Basic inference: Test hypotheses and generate confidence intervals for slope and intercept.

Testing hypotheses concerning the slope

Hypothesis test for β_1

The null hypothesis is:

$$H_0 : \beta_1 = \beta_{10}$$

The test statistic is:

$$t = \frac{b_1 - \beta_{10}}{\hat{\sigma}_{\hat{\beta}_1}}$$

If the null hypothesis is true the test statistic follows the t distribution with n-2 degrees of freedom or $t \sim t(n-2)$.

Alternative hypothesis	Reject H_0 if:
$H_1 : \beta_1 < \beta_{10}$	$t < -t_{1-\alpha}$
$H_1 : \beta_1 > \beta_{10}$	$t > t_{1-\alpha}$
$H_1 : \beta_1 \neq \beta_{10}$	$t < -t_{1-\alpha/2}$ or $t > t_{\alpha/2}$

Testing hypotheses concerning the slope

Hypothesis test for β_0

The null hypothesis is:

$$H_0 : \beta_0 = \beta_{00}$$

The test statistic is:

$$t = \frac{b_0 - \beta_{00}}{\hat{\sigma}_{\hat{\beta}_0}}$$

If the null hypothesis is true the test statistic follows the t distribution with n-2 degrees of freedom or $t \sim t(n-2)$.

Alternative hypothesis	Reject H_0 if:
$H_1 : \beta_0 < \beta_{00}$	$t < -t_{1-\alpha}$
$H_1 : \beta_0 > \beta_{00}$	$t > t_{1-\alpha}$
$H_1 : \beta_0 \neq \beta_{00}$	$t < -t_{1-\alpha/2}$ or $t > t_{\alpha/2}$

Confidence interval for β_1

Confidence interval for β_1

The lower bound of$1 - \alpha$ confidence interval for β_1 is:

$$b_1 - t_{1-\alpha/2,(n-2)} \cdot \hat{\sigma}_{\hat{\beta}_1}$$

The upper bound of $1 - \alpha$ confidence interval is:

$$b_1 + t_{1-\alpha/2,(n-2)} \cdot \hat{\sigma}_{\hat{\beta}_1}$$

where b_1 is calculated the same way as usual, n is the number of paired measurements and $t_{1-\alpha/2,(n-2)}$ is found in the t-distribution table.

Confidence interval for β_0

Confidence interval for β_0

The lower bound of a $1-\alpha$ confidence interval for β_0 is:

$$b_0 - t_{1-\alpha/2,(n-2)} \cdot \hat{\sigma}_{\hat{\beta}_0}$$

The upper bound of $1-\alpha$ confidence interval is:

$$b_0 + t_{1-\alpha/2,(n-2)} \cdot \hat{\sigma}_{\hat{\beta}_0}$$

where b_0 is calculated the same way as usual, n is the number of paired measurements and $t_{1-\alpha/2,(n-2)}$ is in the table for the t-distribution.

Estimating a point on the regression line

Estimate mean response at x_h:

$$\widehat{E[Y_h]} = \hat{\beta}_0 + \hat{\beta}_1 x_h$$

Then

$$E\left[\widehat{E[Y_h]}\right] = \beta_0 + \beta_1 x_h$$

$$Var[\widehat{E[Y_h]}] = \sigma^2 \left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum (x_j - \bar{x})^2} \right)$$

We use $s^2_{y_h}$ to denote the numerical outcome or

$$s^2_{y_h} = \hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum (x_j - \bar{x})^2} \right)$$

Confidence interval for a point on the regression line

Confidence interval for a point on the regression line

The lower bound of$1 - \alpha$ confidence interval for $\hat{Y}_h$ is:

$$(b_0 + b_1 x_0) - t_{1-\alpha/2,(n-2)} \cdot s_{y_h}$$

The upper bound of $1 - \alpha$ confidence interval is:

$$(b_0 + b_1 x_0) + t_{1-\alpha/2,(n-2)} \cdot s_{y_h}$$

where b_0 and b_1 are calculated the same way as usual, n is the number of paired measurements and $t_{1-\alpha/2,(n-2)}$ is found in the t-distribution table.

Predicting a new observation

We would like to predict a new observation Y_h, at x_h
Use $\hat{Y}_h = \hat{\beta}_0 + \hat{\beta}_1 x_h$
The variance is:

$$V\left[\hat{Y}_h - Y_h\right] = \sigma^2\left(1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum(x_j - \bar{x})^2}\right)$$

We use s^2_{pred} to denote the numerical outcome or:

$$s^2_{pred} = \hat{\sigma}^2\left(1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum(x_j - \bar{x})^2}\right)$$

Prediction interval for a new observation

Prediction interval for a new observation

The lower bound of$1 - \alpha$ prediction interval for $\hat{Y}_h$ is:

$$(b_0 + b_1 x_0) - t_{1-\alpha/2,(n-2)} \cdot s_{pred}$$

The upper bound of $1 - \alpha$ prediction interval is:

$$(b_0 + b_1 x_0) + t_{1-\alpha/2,(n-2)} \cdot s_{pred}$$

where b_0 and b_1 are calculated the same way as usual, n is the number of paired measurements and $t_{1-\alpha/2,(n-2)}$ is found in the t-distribution table.