Lecture 03 Simple Linear Models: Leverage, Hypothesis Tests, Goodness of Fit

09 September 2015

Taylor B. Arnold Yale Statistics STAT 312/612



### **Notes**

- Problem Set #1 Online: Due Next Wednesday, 2015-09-16
- R code; online
- Course Pace
- Classroom

# Goals for today

- 1. simulation of leverage
- 2. hypothesis tests for simple linear regression
- 3. goodness of fit,  $R^2$
- 4. Galton's heights data

# LEVERAGE SIMULATION

# Hypothesis Tests

#### **Z-Test**

Take the simple linear regression model:

$$y_i = x_i \beta + \epsilon_i, \quad i = 1, \dots n.$$

With independent, identically distributed normal error terms:

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Last time we calculated the MLE estimator,

$$\widehat{eta}_{MLE} = rac{\sum_i y_i x_i}{\sum_i x_i^2}$$

Last time we calculated the MLE estimator.

$$\widehat{eta}_{ ext{MLE}} = rac{\sum_{i} y_{i} x_{i}}{\sum_{i} x_{i}^{2}}$$

And showed that it has a normal distribution with the following mean and variance:

$$\widehat{\beta} \sim \mathcal{N}(\beta, \frac{\sigma^2}{\sum_i x_i^2})$$

If we want to test the hypothesis  $H_0: \beta = b$ , we could construct a test statistic as follows:

If we want to test the hypothesis  $H_0: \beta = b$ , we could construct a test statistic as follows:

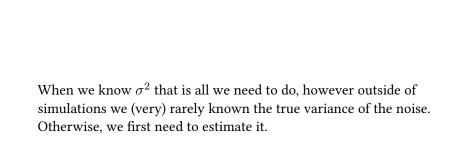
$$z = \frac{\widehat{\beta} - b}{\sqrt{\frac{\sigma^2}{\sum_i x_i^2}}}$$

If we want to test the hypothesis  $H_0: \beta = b$ , we could construct a test statistic as follows:

$$z = \frac{\widehat{\beta} - b}{\sqrt{\frac{\sigma^2}{\sum_i x_i^2}}}$$

Under the null hypothesis, we have

$$z|H_0 \sim \mathcal{N}(0,1)$$



#### T-Test

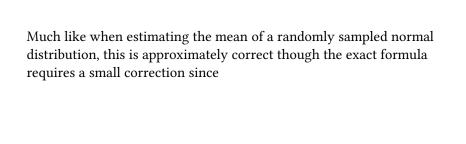
The residuals from a given prediction of  $\beta$  are given by:

$$r_i = y_i - \widehat{y}_i$$
$$= y_i - x_i \widehat{\beta}$$

These represent an estimate of the error terms  $\epsilon_i$ .

If  $r_i$  is the sampled and estimated version of  $\epsilon_i$ , it would seem reasonable to have:

$$\frac{1}{n} \sum_{i=1}^{n} r_i^2 \approx \mathbb{E}\epsilon^2$$
$$= \sigma^2$$



Much like when estimating the mean of a randomly sampled normal distribution, this is approximately correct though the exact formula requires a small correction since

$$\mathbb{E}\left(\sum_{i} r_i^2\right) = (n-1) \cdot \sigma^2$$

Much like when estimating the mean of a randomly sampled normal distribution, this is approximately correct though the exact formula requires a small correction since

$$\mathbb{E}\left(\sum_{i} r_i^2\right) = (n-1) \cdot \sigma^2$$

I will delay a formal derivation of this until the multivariate case; conceptually seems reasonable that the estimate will be slightly smaller due to the estimation of  $r_i$  by the same data.

So, we instead use a corrected form to estimate the error variance,

an estimator that we will call 
$$s^2$$
: 
$$s^2 = \frac{1}{n-1} \cdot \sum_i r_i^2$$

 $=\frac{1}{n-1}\cdot\sum_{i}(y_i-\widehat{y}_i)^2$ 

 $=\frac{1}{n-1}\cdot\sum_{i}(y_i-x_i\beta)^2$ 

The ratio of our estimator to the true variance has a  $\chi^2$  distribution with n-1 degrees of freedom.

$$(n-1)\cdot\frac{s^2}{\sigma^2}\sim\chi_{n-1}^2$$

The standard error is then given by:

$$S.E.(\widehat{\beta}) = \sqrt{\frac{s^2}{\sum_i x_i^2}}$$

The standard error is then given by:

$$S.E.(\widehat{\beta}) = \sqrt{\frac{s^2}{\sum_i x_i^2}}$$

$$\sqrt{s^2}$$

 $= \sqrt{\frac{(y-x_i\widehat{\beta})^2}{(n-1)\cdot\sum_i x_i^2}}$ 

Finally, we can construct a test statistic:

$$t = \frac{\widehat{\beta} - b}{\text{S.E.}(\widehat{\beta})}$$

Finally, we can construct a test statistic:

$$t = \frac{\widehat{\beta} - b}{\text{S.E.}(\widehat{\beta})}$$

 $t|H_0 \sim t_{n-1}$ 

And under the null hypothesis, we have

On a related note, we can similarly calculate a confidence interval for  $\beta$  using the standard error. A  $100(1-\alpha)\%$ . confidence interval is given by:

$$\widehat{\beta} \pm t_{n-1,1-\alpha/2} \cdot \text{S.E.}(\widehat{\beta})$$

On a related note, we can similarly calculate a confidence interval for  $\beta$  using the standard error. A  $100(1-\alpha)\%$ . confidence interval is given by:

$$\widehat{\beta} \pm t_{n-1,1-\alpha/2} \cdot \text{S.E.}(\widehat{\beta})$$

For a reasonably large sample size n, we can approximate this by a normal distribution:

$$\widehat{\beta} \pm z_{1-\alpha/2} \cdot \text{S.E.}(\widehat{\beta})$$

As an alternative to the T-test, consider squaring the test statistic

$$T^2 = \left(\frac{\widehat{\beta} - b}{\text{S.E.}(\widehat{\beta})}\right)^2$$

As an alternative to the T-test, consider squaring the test statistic

$$T^{2} = \left(\frac{\widehat{\beta} - b}{\text{S.E.}(\widehat{\beta})}\right)^{2}$$
$$= \frac{\left(\frac{\widehat{\beta} - b}{\sqrt{\sigma^{2}/\sum_{i}x_{i}^{2}}}\right)^{2}}{s^{2}/\sigma^{2}}$$

As an alternative to the T-test, consider squaring the test statistic

$$T^{2} = \left(\frac{\widehat{\beta} - b}{\text{S.E.}(\widehat{\beta})}\right)^{2}$$
$$= \frac{\left(\frac{\widehat{\beta} - b}{\sqrt{\sigma^{2}/\sum_{i} x_{i}^{2}}}\right)^{2}}{s^{2}/\sigma^{2}}$$
$$= \frac{U}{V}$$

As an alternative to the T-test, consider squaring the test statistic

$$T^{2} = \left(\frac{\widehat{\beta} - b}{\text{S.E.}(\widehat{\beta})}\right)^{2}$$

$$= \frac{\left(\frac{\widehat{\beta} - b}{\sqrt{\sigma^{2}/\sum_{i} x_{i}^{2}}}\right)^{2}}{s^{2}/\sigma^{2}}$$

$$= \frac{U}{V}$$

Where  $U \sim \chi_1^2$  and  $(n-1) \cdot V \sim \chi_{n-1}^2$ .

As an alternative to the T-test, consider squaring the test statistic

$$T^{2} = \left(\frac{\widehat{\beta} - b}{\text{S.E.}(\widehat{\beta})}\right)^{2}$$

$$= \frac{\left(\frac{\widehat{\beta} - b}{\sqrt{\sigma^{2} / \sum_{i} x_{i}^{2}}}\right)^{2}}{s^{2} / \sigma^{2}}$$

$$= \frac{U}{V}$$

Where  $U \sim \chi_1^2$  and  $(n-1) \cdot V \sim \chi_{n-1}^2$ .

And therefore  $T^2 \sim F_{1,n-1}$ .

# Intercept Model

When we have the model  $y = \alpha + x\beta + \epsilon$ , the form of  $s^2$  changes slightly:

$$s^2 = \frac{1}{n-2} \cdot \sum_{i} (y_i - \widehat{y}_i)^2$$

# Intercept Model

When we have the model  $y = \alpha + x\beta + \epsilon$ , the form of  $s^2$  changes slightly:

$$s^2 = \frac{1}{n-2} \cdot \sum_{i} (y_i - \widehat{y}_i)^2$$

as well as the standard errors:

S.E.
$$(\alpha) = \sqrt{s^2 \cdot \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}\right)}$$
  
S.E. $(\beta) = \sqrt{\frac{s^2}{\sum_i (x_i - \bar{x})^2}}$ 

# GOODNESS OF FIT

## $R^2$

A common measurement of how well a linear model explains the data is the  $R^2$ . For the non-intercept version, it can be written as:

$$R^{2} = 1 - \frac{\sum_{i} (y_{i} - \widehat{y}_{i})^{2}}{\sum_{i} (y_{i})^{2}}$$

## $R^2$

A common measurement of how well a linear model explains the data is the  $R^2$ . For the non-intercept version, it can be written as:

$$R^{2} = 1 - \frac{\sum_{i} (y_{i} - \widehat{y}_{i})^{2}}{\sum_{i} (y_{i})^{2}}$$

We can re-write this as:

$$R^2 = \left(\frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2 \cdot \sum_i y_i^2}}\right)^2$$

The more typically seen version compares the estimated residuals with the centered values of y.

$$R^{2} = 1 - \frac{\sum_{i} (y_{i} - \widehat{y}_{i})^{2}}{\sum_{i} (y_{i} - \overline{y})^{2}}$$

The more typically seen version compares the estimated residuals with the centered values of y.

$$R^{2} = 1 - \frac{\sum_{i} (y_{i} - \widehat{y}_{i})^{2}}{\sum_{i} (y_{i} - \overline{y})^{2}}$$

With a bit of algebraic manipulation, we see that this is equal to the squared sample correlation of x and y:

$$R^{2} = \left(\frac{\sum_{i}(x_{i} - \bar{x})(y_{i} - \bar{y})}{\sqrt{\sum_{i}(x_{i} - \bar{x})^{2} \cdot \sum_{i}(y_{i} - \bar{y})^{2}}}\right)^{2}$$
$$= cor(x, y)^{2}$$