Lecture 03

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

09 September 2015

Taylor B. Arnold
Yale Statistics
STAT 312/612

Yale

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:

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

With independent, identically distributed normal error terms:

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

Last time we calculated the MLE estimator,

$$\widehat{\beta}_{MLE} = \frac{\sum_i y_i x_i}{\sum_i x_i^2}$$

Last time we calculated the MLE estimator,

$$\widehat{\beta}_{MLE} = \frac{\sum_i y_i x_i}{\sum_i x_i^2}$$

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

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

When we know σ^2 that is all we need to do, however outside of simulations we (very) rarely known the true variance of the noise. Otherwise, we first need to estimate it.

T-Test

The residuals from a given prediction of β are given by:

$$\begin{aligned} r_i &= y_i - \widehat{y}_i \\ &= y_i - x_i\widehat{\beta} \end{aligned}$$

These represent an estimate of the error terms ϵ_i.

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

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

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

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:

$$
\begin{aligned}
s^2 &= \frac{1}{n-1} \cdot \sum_i r_i^2 \\
&= \frac{1}{n-1} \cdot \sum_i (y_i - \widehat{y}_i)^2 \\
&= \frac{1}{n-1} \cdot \sum_i (y_i - x_i\beta)^2
\end{aligned}
$$

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

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

The standard error is then given by:

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

The standard error is then given by:

$$\begin{aligned} \text{S.E.}(\widehat{\beta}) &= \sqrt{\frac{s^2}{\sum_i x_i^2}} \\ &= \sqrt{\frac{(y - x_i\widehat{\beta})^2}{(n-1) \cdot \sum_i x_i^2}} \end{aligned}$$

Finally, we can construct a test statistic:

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

Finally, we can construct a test statistic:

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

And under the null hypothesis, we have

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

On a related note, we can similarly calculate a confidence interval for β 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 β 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:

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

F-Test

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

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

F-Test

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

$$
\begin{aligned}
T^2 &= \left(\frac{\widehat{\beta} - b}{\text{S.E.}(\widehat{\beta})} \right)^2 \\
&= \frac{\left(\frac{\widehat{\beta} - b}{\sqrt{\sigma^2 / \sum_i x_i^2}} \right)^2}{s^2 / \sigma^2}
\end{aligned}
$$

F-Test

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

$$
\begin{aligned}
T^2 &= \left(\frac{\widehat{\beta} - b}{\text{S.E.}(\widehat{\beta})} \right)^2 \\
&= \frac{\left(\frac{\widehat{\beta} - b}{\sqrt{\sigma^2 / \sum_i x_i^2}} \right)^2}{s^2 / \sigma^2} \\
&= \frac{U}{V}
\end{aligned}
$$

F-Test

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

$$\begin{aligned} T^2 &= \left(\frac{\widehat{\beta} - b}{\text{S.E.}(\widehat{\beta})} \right)^2 \\ &= \frac{\left(\frac{\widehat{\beta} - b}{\sqrt{\sigma^2 / \sum_i x_i^2}} \right)^2}{s^2 / \sigma^2} \\ &= \frac{U}{V} \end{aligned}$$

Where $U \sim \chi^2_1$ and $(n-1) \cdot V \sim \chi^2_{n-1}$.

F-Test

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

$$\begin{aligned} T^2 &= \left(\frac{\widehat{\beta} - b}{\text{S.E.}(\widehat{\beta})} \right)^2 \\ &= \frac{\left(\frac{\widehat{\beta} - b}{\sqrt{\sigma^2 / \sum_i x_i^2}} \right)^2}{s^2 / \sigma^2} \\ &= \frac{U}{V} \end{aligned}$$

Where $U \sim \chi^2_1$ and $(n-1) \cdot V \sim \chi^2_{n-1}$.

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:

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

$$\text{S.E.}(\beta) = \sqrt{\frac{s^2}{\sum_i (x_i - \bar{x})^2}}$$

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.

$$R^2 = 1 - \frac{\sum_i (y_i - \widehat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

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

$$R^2 = 1 - \frac{\sum_i (y_i - \widehat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

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

$$\begin{aligned} R^2 &= \left(\frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \cdot \sum_i (y_i - \bar{y})^2}} \right)^2 \\ &= cor(x, y)^2 \end{aligned}$$

Applications

Some parting words from Sir Francis Galton

as any other quality. Also, that although it exerts an immense influence for good or ill on domestic happiness, it seems that good temper has not been especially looked for, nor ill temper especially shunned, as it ought to be in marriage-selection.