12.3: Inferences About the Regression Coefficient β_1

Taylor

University of Virginia

Introduction

When we estimate the βs, we usually are much more interested in β_1. If this number is non-zero, then there is a relationship between the inputs and the outputs.

Our estimate $\hat{\beta}_1$ is based on random data, so it is random itself. $\hat{\beta}_1$ is a point estimator for β_1. We can do a lot of the same stuff we did when we were estimating μ with $\bar{X}$.

Notation

Assuming all the Xs are known/nonrandom, then we can write the estimated slope coefficient like this

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(Y_i - \bar{Y})}{\sum(x_i - \bar{x})^2}$$

The book is capitalizing the Ys to emphasize that this is the source if randomness.

A Little Trick

Call $\sum(x_i - \bar{x})^2 = S_{xx}$

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{\sum(x_i - \bar{x})(Y_i - \bar{Y})}{S_{xx}} \\
&= \frac{\sum(x_i - \bar{x})Y_i - \sum(x_i - \bar{x})\bar{Y}}{S_{xx}} \\
&= \frac{\sum(x_i - \bar{x})Y_i}{S_{xx}} \\
&= \sum_i c_i Y_i
\end{aligned}
$$

because $\sum(x_i - \bar{x}) = 0$. This is a linear combination of independent (but not identical) normal rvs.

A Distribution for $\hat{\beta}_1$

So

1. $\hat{\beta}_1 \sim \mathcal{N}(\beta_1, \frac{\sigma^2}{S_{xx}})$
2. $\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$
3. $\hat{\beta}_1$ is independent from $\frac{(n-2)\hat{\sigma}^2}{\sigma^2}$

where $\hat{\sigma}^2 = \sum_i \frac{(y_i - \hat{y}_i)^2}{n-2}$

so that means...

Our distribution

$$T = \left[\frac{\hat{\beta}_1 - \beta_1}{\frac{\sigma}{\sqrt{S_{xx}}}} \right] \div \left[\sqrt{\frac{(n-2)\hat{\sigma}^2}{\sigma^2(n-2)}} \right] = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{S_{xx}}}$$

follows a t_{n-2} distribution.

A CI for β_1

Based on

$$P\left(-t_{\alpha/2,n-2} \leq \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{S_{xx}}} \leq t_{\alpha/2,n-2}\right) = 1 - \alpha$$

we can do a bit of arithmetic and

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2}\frac{\hat{\sigma}}{\sqrt{S_{xx}}}$$

A Hypothesis Test for β_1

1. $H_0 : \beta_1 = \beta_{1,0}$
2. $t = \frac{\hat{\beta}_1 - \beta_{1,0}}{\hat{\sigma}/\sqrt{S_{xx}}}$
3. if $H_a : \beta_1 > \beta_{1,0}$, reject if $t > t_{\alpha,n-2}$
4. if $H_a : \beta_1 < \beta_{1,0}$, reject if $t < -t_{\alpha,n-2}$
5. if $H_a : \beta_1 \neq \beta_{1,0}$, reject if $t > t_{\alpha/2,n-2}$ or if $t < -t_{\alpha/2,n-2}$

The most common situation is when $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$. This is basically testing if X is associated (linearly) with Y.

A Note About Fitting Logistic Regression Models

$Y_i \sim \text{Bernoulli}[p(x_i)], i = 1, \ldots, n$. The likelihood is then

$$\prod_i f(y_i) = \prod_i p(x_i)^{y_i}[1 - p(x_i)]^{1-y_i}$$

We still fit this using maximum likelihood estimation, but there's no closed-form solution for the coefficients. Also notice how each $p(x_i)$ is different possibly, so we can't combine/simplify stuff.

Many software packages fit these. In R it would be something like

```
my_mod <- glm(y ~ x, family = "binomial")
```