Simple Linear Regression: A Simple Application and How We Make Inference

BIOS 6611

CU Anschutz

Week 7

Refresher

Refresher

Recall, we are learning about the Simple Linear Regression (SLR) model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \epsilon_i$$

where $i = 1, \ldots, n$ indexes the data pairs (Y_i, X_i).

A *simple* linear regression model has a single explanatory variable (X_1). (Linear regression models that are not simple can have more than one explanatory variable.)

The assumptions are Existence, Linearity, Independent, Homoscedasticity, and Normality of the error term.

Inference for Least Squares Estimators

Inference for Least Squares Estimators

Under the assumptions, we have

$$\begin{aligned} \epsilon_i &\sim N(0, \sigma_e^2) \\ Y_i|X_i &\sim N(\mu_{Y|X}, \sigma_{Y|X}^2) \end{aligned}$$

where $\mu_{Y|X}$ is allowed to change (linearly) with the explanatory variable. That is,

$$\mu_{Y|X} = \beta_0 + \beta_1 X$$

Therefore, if we assume normality of errors, then the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are normally distributed, since $\hat{\beta}_0$ and $\hat{\beta}_1$ will be functions of independent normally distributed random variables (See Corollary 4.6.10 in Casella & Berger).

Inference for Least Squares Estimators (cont.)

Alternatively, if we cannot assume normality of the error terms:

1. If we have a large sample size, asymptotic normality may be assumed for the estimators (CLT!)
2. If we don't have a large sample size and errors are not normally distributed, bootstrap or Monte Carlo methods may be appropriate.

Testing for Significant Associations

Testing for Significant Associations

Say we want to test if there is a *linear* association between the explanatory (X) and response (Y) variables. This would be equivalent to testing if the slope is zero in the SLR model. Thus, we test the hypothesis:

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_A : \beta_1 \neq 0$$

To perform this test, we use the fact that the ratio of the estimate to its standard errors, called the t-statistic, follows a t-distribution with $n - 2$ degrees of freedom:

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \sim t_{n-2}$$

($n - 2$ degrees of freedom because we estimate both the intercept and the predictor beta coefficients)

Testing for Significant Associations (cont.)

95% CI for the slope coefficient:

$$\hat{\beta}_1 \pm t_{n-2,1-\alpha/2} SE(\hat{\beta}_1)$$

If we fail to reject H_0, it generally means one of three things:

1. There is no association
2. There is no *linear* association
3. We've made a Type II error

Example Regression Code and Output for FEV Data

R code

```
# Load in FEV dataset
fev <- read.csv("FEV_rosner.csv")

# Fit SLR FEV = B0 + B1*Age + E
fev_slr <- lm(fev ~ age, data=fev)
summary(fev_slr)
```

R code (cont.)

```
##
## Call:
## lm(formula = fev ~ age, data = fev)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.57539 -0.34567 -0.04989  0.32124  2.12786
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.431648   0.077895   5.541 4.36e-08 ***
## age         0.222041   0.007518  29.533  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5675 on 652 degrees of freedom
## Multiple R-squared:  0.5722, Adjusted R-squared:  0.5716
## F-statistic: 872.2 on 1 and 652 DF,  p-value: < 2.2e-16
```

SAS code

SAS Code:

```
proc reg data=fev;
    model fev = age;
run;
```

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	280.91916	280.91916	872.18	<.0001
Error	652	210.00068	0.32209		
Corrected Total	653	490.91984			

Root MSE	0.56753	**R-Square**	0.5722
Dependent Mean	2.63678	**Adj R-Sq**	0.5716
Coeff Var	21.52349		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	**1**	$\hat{\beta}_0 = 0.43165$	0.07790	5.54	<.0001
age	**1**	$\hat{\beta}_1 = 0.22204$	0.00752	29.53	<.0001

Interpreting and Utilizing the Regression Output

Interpreting and Utilizing the Regression Output

Q: What is the regression equation?
A:

$$\hat{FEV} = 0.432 + 0.222 \times Age$$

Q: What is the interpretation of the slope parameter?
A: For every 1 year increase in age between the ages of 3 and 19, the FEV increases on average by 0.222 liters. *(Note: restricted to observed range of age values, no extrapolation!)*

Q: What is the interpretation of the intercept?
A: When age is 0 years, average FEV is 0.432 liters. *(Note: not scientifically meaningful and also extrapolating outside the range of age values)*

Interpreting and Utilizing the Regression Output

Q: Is there a significant linear relationship between age and FEV?
A: Yes.

$$t = \frac{0.22204}{0.00752} = 29.53 \sim t_{654-2}$$
$$\Rightarrow p < 0.0001$$

Thus, we reject H_0: $\beta_1 = 0$, and conclude there is a significant linear relationship between age and FEV.

Interpreting and Utilizing the Regression Output

Q: Calculate the 95% CI for age. Interpret.
A:

$$\begin{aligned} 0.22204 &\pm t_{652,0.975} \times 0.007518 \\ &= 1.963609 \times 0.007518 \\ &= (0.207, 0.237) \end{aligned}$$

We are 95% confident that FEV increases between 0.207 and 0.237 liters on average for every 1-year increase in age (between the ages of 3 and 19).

Interpreting and Utilizing the Regression Output

Q: What is the predicted FEV for an 11-year old child?
A:

$$\hat{FEV} = 0.432 + 0.222(11) = 2.874 \text{ liters}$$

Q: What is the predicted difference in FEV for 16-year old children versus 11-year old children?
A:

$$\begin{aligned} E(FEV|Age = 16) - E(FEV|Age = 11) &= [\beta_0 + \beta_1 16] - [\beta_0 + \beta_1 11] \\ &= \beta_1(16 - 11) \\ &= 5\beta_1 \end{aligned}$$

Based on the estimate from our regression model we predict the difference between a 16- and 11-year old child to be $5\hat{\beta}_1 = 5(0.22204) = 1.1102$ liters.

Interpreting and Utilizing the Regression Output

Q: What is the 95% CI around this predicted difference?
A:

$$\begin{aligned} 5\hat{\beta}_1 &\pm t_{652,0.975} SE(5\hat{\beta}_1) \\ 5(0.22204) &\pm 1.96 \times 5 \times 0.00752 \\ &= (1.037, 1.184) \text{ liters} \end{aligned}$$

We are 95% confident that FEV increases between 1.037 and 1.184 liters on average from ages 11 to 16 (or any other 5 year age difference between 3 and 19 years).