ECON 340
Economic Research Methods

Div Bhagia

Lecture 17
Inference in Regression Models

Assumptions for Causal Inference

Assumption 1 (Linearity): The relationship between X and Y is given by:

$$Y = \beta_0 + \beta_1 X + u$$

u is the mean zero error term, $E(u) = 0$.

Assumption 2 (Random Sample): The observed data (Y_i, X_i) for $i = 1, 2, ..., n$ represent a random sample of size n from the above population model.

Assumptions for Causal Inference

Assumption 3 (No large outliers): Fourth moments (or Kurtosis) of X and Y are finite.

Assumption 4 (Mean Independence/Exogeneity): The expected value of the error term is the same conditional on any value of the explanatory variable.

$$E(u|X) = E(u) = 0$$

When the exogeneity assumption fails

$$Y = \beta_0 + \beta_1 X + u$$

- Y: test scores, X: class-size, u : teacher quality
- If schools with higher student-teacher ratios have worse teachers ($\uparrow X, \downarrow u$)
- Then, if we see test scores decline with class size ($\uparrow X, \downarrow Y$), hard to say if it's due to teacher quality or class size.

Sampling Distribution for OLS Estimators

Under Assumptions 1-4, in large samples ($n > 100$),

$$\hat{\beta}_0 \sim N(\beta_0, \sigma^2_{\hat{\beta}_0}), \qquad \hat{\beta}_1 \sim N(\beta_1, \sigma^2_{\hat{\beta}_1})$$

where

$$\sigma^2_{\hat{\beta}_1} = \frac{1}{n} \frac{Var[(X_i - \mu_X)u_i]}{Var(X_i)}$$

Test Scores and Class Size

We estimated the following model:

$$TestScore_i = \beta_0 + \beta_1 \cdot STR_i + u$$

And found:

$$\hat{\beta}_0 = 698.93 \quad \text{and} \quad \hat{\beta}_1 = -2.28$$

Even if $E(u|STR) = 0$, some uncertainty around estimates due to sampling variation. Do we really know whether -2.28 is statistically significantly different from 0?

We want to rule out having found a negative impact due to sampling variation when there was no impact.

Hypothesis Testing

Since $\hat{\beta}_1 \sim N(\beta_1, \sigma^2_{\hat{\beta}_1})$ in large samples,

$$T = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim t_{n-k}$$

Remember, the t-distribution has fatter tails but is similar to the standard normal in large samples.

Hypothesis Testing

Null and alternative hypothesis:

$$H_0 : \beta_1 = 0 \qquad H_1 : \beta_1 \neq 0$$

The test statistic under the null:

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$

If $|t| > z_{\alpha/2}$ we reject the null at α% level of significance and say that β_1 is statistically significant at α% level of significance.

Remember: $z_{\alpha/2}$ is the value of z that leaves $\alpha/2$ area in the upper tail of the standard normal distribution.

Output from R

```
summary(lm(testscr ~ str, data))
```

```
##
## Call:
## lm(formula = testscr ~ str, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.727 -14.251   0.483  12.822  48.540
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 698.9330     9.4675  73.825  < 2e-16 ***
## str          -2.2798     0.4798  -4.751 2.78e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Hypothesis Testing

From the output we can see that,

$$\hat{\beta}_1 = -2.28 \quad \text{and} \quad SE(\hat{\beta}_1) = 0.48$$

In which case, the t-statistic:

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{-2.28}{0.48} = -4.75$$

Since $|-4.75| > 2.58$, we can say that $\hat{\beta}_1$ is statistically significant at 1% level of significance.

Hypothesis Testing

From the output we can see that,

$$\hat{\beta}_1 = -2.28 \quad \text{and} \quad SE(\hat{\beta}_1) = 0.48$$

In which case, the t-statistic:

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{-2.28}{0.48} = -4.75$$

Since $|-4.75| > 2.58$, we can say that $\hat{\beta}_1$ is statistically significant at 1% level of significance.

Is it also significant at 5% level of significance?

p-Value

The p-value is the probability of drawing an outcome as or more extreme given the null hypothesis.

$$\text{p-value} = 2P(Z > |t|)$$

In our example,

$$\text{p-value} = 2P(Z > 4.75) = 0.00$$

Remember if $p < \alpha$, reject the null with α% level of significance.

Output from R using Stargazer

	Dependent variable:
	testscr
str	-2.280***
	(0.480)
Constant	698.933***
	(9.467)
Observations	420
R2	0.051
Adjusted R2	0.049
Note:	*p<0.1; **p<0.05; ***p<0.01

Confidence Intervals

As before, we can also create confidence intervals to summarize the uncertainty associated with our estimates.

A $(1 - \alpha)\%$ confidence interval for β_1:

$$\hat{\beta}_1 \pm z_{\alpha/2} \cdot SE(\hat{\beta}_1)$$

If 0 is not in the 95% confidence interval, then once again we can say that β_1 is statistically significant at 5% level of significance.

Confidence Intervals

```
> model <- lm(testscr ~ str, data)
> confint(model, level = 0.95)
                2.5 %     97.5 %
(Intercept) 680.32313 717.542779
str          -3.22298  -1.336637
```