Reading Assignment 5

Andrew Shaw

09-14-25

Linear Regression

1. Define an equation that relates e-cigarette use to tax rates similar to 4.4 from the text.

$$eCigUse = \beta_0 + \beta_1 \times TaxRate + u_i \tag{1}$$

$$where\ \beta_1 = \frac{\Delta eCigUse}{\Delta TaxRate} \tag{2}$$

2. What is the interpretation of the value of β_0?

It is the level of e-cig use when the tax rate is set to 0.

3. Plug in a value for the tax rate in 1, what is the e-cigarette use rate?

$$eCigUse = 20 - 0.5X \tag{3}$$

$$19.5 = 20 - 0.5 * (1) \tag{4}$$

4. Define the error term for your hypothetical plugged in value for 3.

u_i represents other factors which can affect the population's e-cig usage beyond the tax rate. It's hypothetical until the regression analysis is run and is the difference between the observed value (Y_i) and the predicted value ($\hat{Y}_i$).

5. Define the slope of a regression as the ratio of variances.

The slope of the regression is the covariance of X and Y divided by the variance of X.

6. Define the slope.

On average, it's the change in Y resulting from a 1 unit increase of X.

7. What is the measure for best fit?

The R^2 is the measure for best fit which is a ratio of explained squared residuals and total residuals which relays how much of Y can be explained by X.

8. What is the first assumption of OLS to be unbiased and consistent?

It's the conditional distribution of the error term u_i given X_i has a mean of 0. This states on average over many samples the error terms average to 0 giving the 'best fit line'.

9. Why might not this assumption hold?

Assumption might not and probably does not hold because there is likely some correlation among X and factors that are considered in the error term.

10. The second assumption what is it? Describe it in your own words.

This restates that for OLS to work, observation and samples must be i.i.d so the observations are identically distributed and independent meaning observations do not influence each other. This assumption doesn't hold for data such as time-series as a prior period has bearing on the current.

11. Third assumption, what is it? Describe in your own words.

The last assumption is that large outliers are unlikely. This is due to the nature of what is tested, often real-world measures have some finite bounds and therefore do not have infinite variance.

Regression and Hypothesis Testing

1. How can we define the variance of β_1?

The variance is a ratio of the product of the sum of square deviations of the independent variable and the squared error term and the average deviation of X which according to the appendix is the heteroskedastic-robust standard error.

2. Why do we usually use a two-sided hypothesis for β_1?

Because unless it's quite clear that the alternative is one-sided, not doing so could mask downsides of a trial because it's designed to look only in one direction potentially missing other alternatives.

3. How do you estimate a t-statistic for your hypothesis?

Take the difference of the estimated beta coefficient and the hypothesized beta divided by the standard error of the $\hat{\beta}_i$.

4. What would be the reason to test whether β_0 is zero?

Could be useful to find out if there is expected to be no effect on the dependent variable if the independent variable is zero or test estimates for a predicted baseline if it means something significant when X is actually zero.

5. What is the formula for the 90% confidence interval for β_1? What is your best interpretation of this confidence interval?

$$[\beta_1 - 1.645 * SE(\hat{\beta}_1), \beta_1 + 1.645 * SE(\hat{\beta_1}))] \tag{5}$$

The confidence interval is the range of values the true mean of β_1 falls within and with repeated sampling the true value would fall within that interval 90% of the time.

6. What is an indicator variable?

An indicator variable, or dummy variable, is a binary variable so takes the value of 0 or 1.

7. What does the indicator variable tell us?

The indicator can tell us whether some subgroup in the sample data affirmatively has some characteristic or not.

8. What is heteroskedasticity? What is it important for?

Well OLS has nice properties if homoskedastic, like being best linear unbiased estimator but data is rarely homoskedastic. It's important because it heteroskedasticity is ignored hypothesis testing and the tests will return inaccurate results.

9. When should we use the z-distribution and when should we use the t-distribution?

T-distributions are used when the population variance isn't known or when the same size is small. This leads to more variance and the t-distribution has fatter tails that Z-distribution to account for the additional variance.