Inference with Simple Linear Regression (SLR)

1 Introduction

Oftentimes, the data we collect come from a random sample that is representative of the population of interest. A common example is an election poll before a presidential election. Random sampling allows the sample to be representative of the population. However, if we obtain another random sample, the characteristics of the new sample are unlikely to be exactly the same as the first sample. For example, the sample proportion who will vote for a certain party is unlikely to be the same for both random samples. What this tells us is that even with representative samples, sample proportions are unlikely to be equal to the population proportion, and sample proportions vary from sample to sample.

Dr. W. Edwards Deming's Red Bead experiment illustrates this concept. A video of this experiment can be found here.

In this video, the number of red beads, which represent bad products, varies each time the worker obtains a random sample of 50 beads. The fact that the number of red beads increases in his second sample does not indicate that he performed his task any worse, as this increase is due to the random variation associated with samples.

Note: Deming's Red Bead experiment was developed to illustrate concepts associated with management. He is best known for his work in developing the Japanese economy after World War II. You will be able to find many blogs/articles discussing the experiment on the World Wide Web. Although many of the articles discuss how this experiment applies in management, it can be used to illustrate concepts of variation.

The same idea extends to the slope and intercept of a regression line. The estimated slope and intercept will vary from sample to sample and are unlikely to be equal to the population slope and intercept. In inferential statistics, we use hypothesis tests and confidence intervals to aid us in accounting for this random variation. In this module, you will learn how to account for and quantify the random variation associated with the estimated regression model, and how to interpret the estimated regression model while accounting for random variation.

1.1 Review from previous module

The **simple linear regression model** is written as

$$y = \beta_0 + \beta_1 x + \epsilon. \tag{1}$$

We make some assumptions for the error term ϵ. They are:

1. The errors have mean 0.
2. The **errors have variance denoted by** σ^2. Notice this variance is constant.
3. The errors are independent.
4. The errors are normally distributed.

These assumptions allow us to derive the distributional properties associated with our least squares estimators $\hat{\beta}_0, \hat{\beta}_1$, which then enables us to compute reliable confidence intervals and perform hypothesis tests on our SLR reliably.

$\hat{\beta}_1, \hat{\beta}_0$ are the estimators for β_1, β_0 respectively. These estimators can be interpreted in the following manner:

- $\hat{\beta}_1$ **denotes the change in the predicted** y **when** x **increases by 1 unit. Alternatively, it denotes the change in** y**, on average, when** x **increases by 1 unit.**
- $\hat{\beta}_0$ **denotes the predicted** y **when** $x = 0$**. Alternatively, it denotes the average of** y **when** $x = 0$**.**

How do the values of these estimators vary from sample to sample?

2 Hypothesis Testing in SLR

2.1 Distribution of least squares estimators

Gauss Markov Theorem: Under assumptions for a regression model, the least squares estimators $\hat{\beta}_1$ and $\hat{\beta}_0$ are unbiased and have minimum variance among all unbiased linear estimators.

Thus, the least squares estimators have the following properties:

1. $\mathrm{E}(\hat{\beta}_1) = \beta_1$, $\mathrm{E}(\hat{\beta}_0) = \beta_0$

Note: An estimator is **unbiased** if its expected value is exactly equal to the parameter it is estimating.

2. The variance of $\hat{\beta}_1$ is

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \tag{2}$$

3. The variance of $\hat{\beta}_0$ is

$$\mathrm{Var}(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right] \tag{3}$$

4. $\hat{\beta}_1$ and $\hat{\beta}_0$ both follow a normal distribution.

Note that in (2) and (3), we use $s^2 = MS_{res}$ to estimate σ^2 since σ^2 is a unknown value.

What these imply is that if we standardize $\hat{\beta}_1$ and $\hat{\beta}_0$, these standardized quantities will follow a t_{n-2} distribution, i.e.

$$\frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} \sim t_{n-2} \tag{4}$$

and

$$\frac{\hat{\beta}_0 - \beta_0}{se(\hat{\beta}_0)} \sim t_{n-2}, \tag{5}$$

where

$$se(\hat{\beta}_1) = \sqrt{\frac{MS_{res}}{\sum (x_i - \bar{x})^2}} = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}} \tag{6}$$

and

$$se(\hat{\beta}_0) = \sqrt{MS_{res} \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right]} = s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}} \tag{7}$$

Note:

- $se(\hat{\beta}_1)$ is read as the **standard error of** $\hat{\beta}_1$. The standard error of any estimator is essentially the sample standard deviation of that estimator, and measures the spread of that estimator.
- A t_{n-2} distribution is read as a t **distribution with** $n-2$ **degrees of freedom**.

2.2 Testing regression coefficients

Hypothesis testing is used to investigate if a population parameter is **different from a specific value**. In the context of SLR, we usually want to test if β_1 is 0 or not. If $\beta_1 = 0$, there is no linear relationship between the variables.

The general steps in hypothesis testing are:

- Step 1: State the null and alternative hypotheses.
- Step 2: A test statistic is calculated using the sample, assuming the null is true. The value of the test statistic measures how the **sample deviates from the null**.
- Step 3: Make conclusion, using either critical values or p-values.

In the previous module, we introduced the ANOVA F test. In SLR, this tests if the slope of the SLR equation is 0 or not. It turns out that we can also perform a t test for the slope. In the t test for the slope, the null and alternative hypotheses are:

$$H_0 : \beta_1 = 0, H_a : \beta_1 \neq 0.$$

The test statistic is

$$t = \frac{\hat{\beta}_1 - \text{ value in null}}{se(\hat{\beta}_1)} \tag{8}$$

which is compared with a t_{n-2} distribution. Notice that (8) comes from (4).

Let us go back to our simulated example that we saw in the last module. We have data from 6000 UVa undergraduate students on the amount of time they spend studying in a week (in minutes), and how many courses they are taking in the semester (3 or 4 credit courses).

```
##create dataframe
df<-data.frame(study,courses)

##fit regression
result<-lm(study~courses, data=df)
##look at regression coefficients
summary(result)$coefficients
```

```
##              Estimate Std. Error   t value      Pr(>|t|)
## (Intercept)  58.44829  1.9218752  30.41211 4.652442e-189
## courses     120.39310  0.4707614 255.74125  0.000000e+00
```

The t statistic for testing $H_0 : \beta_1 = 0, H_a : \beta_1 \neq 0$ is reported to be 255.7412482, which can be calculated using (8): $t = \frac{120.39310-0}{0.4707614}$. The reported p-value is virtually 0, so we reject the null hypothesis. The data support the claim that there is a linear association between study time and the number of courses taken.

3 Confidence Intervals for Regression Coefficients

Confidence intervals (CIs) are similar to hypothesis testing in the sense that they are also based on the distributional properties of an estimator. CIs may differ in their use in the following ways:

1. We are not assessing if the parameter is different from a specific value.

2. We are more interested in exploring a plausible **range of values for an unknown parameter**.

Because CIs and hypothesis tests are based on the distributional properties of an estimator, their conclusions will be consistent (as long as the significance level is the same).

Recall the general form for CIs:

$$\text{estimator} \pm (\text{multiplier} \times \text{s.e of estimator}). \tag{9}$$

We have the following components of a CI

- **estimator (or statistic)**: numerical quantity that describes a sample
- **multiplier**: determined by confidence level and relevant probability distribution
- **standard error of estimator**: measure of variance of estimator (basically the square root of the variance of estimator)

Following (9) and (4), the $100(1-\alpha)\%$ CI for β_1 is

$$\hat{\beta}_1 \pm t_{1-\alpha/2;n-2} se(\hat{\beta}_1) = \hat{\beta}_1 \pm t_{1-\alpha/2;n-2} \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}}. \tag{10}$$

Going back to our study time example, the 95% CI for β_1 is (119.470237, 121.3159601).

```
##CI for coefficients
confint(result,level = 0.95)[2,]
```

```
##    2.5 %   97.5 %
## 119.4702 121.3160
```

An interpretation of this CI is that we have 95% confidence that the true slope β_1 lies between (119.470237, 121.3159601). In other words, for each additional course taken, the predicted study time increases between 119.470237 and 121.3159601 minutes.

3.1 Thought questions

- Is the conclusion from this 95% CI consistent with the hypothesis test for $H_0 : \beta_1 = 0$ in the previous section at 0.05 significance level?
- I have presented hypothesis tests and CIs for the slope, β_1.
  - How would you calculate the t statistic if you wanted to test $H_0 : \beta_0 = 0, H_0 : \beta_0 \neq 0$?
  - How would you calculate the 95% CI for the intercept β_0?

Generally, we are usually more interested in the slope than the intercept.

4 CI of the Mean Response

We have established that the least squares estimators $\hat{\beta}_1, \hat{\beta}_0$ have their associated variances. Since the estimated SLR equation is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x, \tag{11}$$

it stands to reason that $\hat{y}$ has an associated variance as well, since it is a function of $\hat{\beta}_1, \hat{\beta}_0$.

There are two interpretations of $\hat{y}$:

1. it **estimates the mean of y when $x = x_0$**;
2. it **predicts the value of y for a new observation when $x = x_0$**.

Note: x_0 denotes a specific numerical value for the predictor variable.

Depending on which interpretation we want, there are two different intervals based on $\hat{y}$. The first interpretation is associated with the **confidence interval for the mean response, $\hat{\mu}_{y|x_0}$, given the predictor**. This is used when we are interested in the average value of the response variable, when the predictor is equal to a specific value. This CI is

$$\hat{\mu}_{y|x_0} \pm t_{1-\alpha/2,n-2} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}. \tag{12}$$

Going back to our study time example, suppose we want the average study time for students who take 5 courses, the 95% CI is

```
##CI for mean y when x=5
newdata<-data.frame(courses=5)
predict(result, newdata, level=0.95, interval="confidence")
```

```
##        fit      lwr      upr
## 1 660.4138 659.2224 661.6052
```

We have 95% confidence that the average study time for students who take 5 courses is between 659.2223688 and 661.605187 minutes.

5 PI of a New Response

Previously, we found a CI for the mean of y given a specific value of x, (12). This CI gives us an idea about the location of the regression line at a specific of x.

Instead, we may have interest in finding an interval for a new value of $\hat{y}_0$, when we have a new observation $x = x_0$. This is called a **prediction interval (PI) for a future observation y_0 when the predictor is a specific value**. This interval follows from the second interpretation of $\hat{y}$.

The PI for $\hat{y}_0$ takes into account:

1. Variation in location for the distribution of y (i.e. where is the center of the distribution of y?).
2. Variation **within the probability distribution of** y.

By comparison, the confidence interval for the mean response (12) only takes into account the first element. The PI is

$$\hat{y}_0 \pm t_{1-\alpha/2,n-2} s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}. \tag{13}$$

Going back to our study time example, suppose we have a newly enrolled student who wishes to take 5 courses, and the student wants to predict his study time

```
##PI for y when x=5
predict(result, newdata, level=0.95, interval="prediction")
```

```
##        fit      lwr      upr
## 1 660.4138 602.0347 718.7928
```

We have 95% confidence that the study time for this student is between 602.0347305 and 718.7928253 minutes.

5.1 Thought questions

- In the following two scenarios, decide if we are more interested in the CI for the mean response given the predictor (12), or the PI for a future response given the predictor (13).
  - We wish to estimate the waiting time, on average, of DMV customers if there are 10 people in line at the DMV.
  - I enter the DMV and notice 10 people in line. I want to estimate my waiting time.
- Look at the standard errors associated with the intervals given in (12) and (13). How are they related to each other?