2.3: Simple Model Inference

Dr. Bean - Stat 5100

Recall the simple linear model

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

Inference is the process by which we make a decision about whether an observed difference from an expectation was simply due to chance or not.

In other words, inference is the process of making conclusions given incomplete information.

1 Why Inference?

Hypothetical questions:

- Suppose you found out that there is not significant relationship between study time and final grades in Stat 5100, how would this effect your approach to this course?
- Suppose you have the flu and you find out from a clinical trial of a new flu drug that those who took the drug had slightly shorter flu durations than those who took the placebo, but that the difference was likely due to chance. How likely would you be to purchase this drug?

In the absence of complete information, inference is an efficient way to decide what associations are "real" and which are not.

(Individual) Can you think of an example scenario where a test of significance would be of interest to researchers?

2 Hypothesis Testing

Recall that hypothesis testing is the formal way by which we determine if an observed difference from an expectation was due to chance.

Process

- Define a null and alternative hypothesis.
  - H_0 : "no effect"
  - H_a : "some effect"
- Define a test statistic:
  - Compares what we observed to what we expected if the null hypothesis was true.
- Determine the "sampling distribution"
  - Determines the natural variation in the test statistic that we would expect if we took many different samples from the same population.
  - In practice, we only ever take one sample. Statistical theory is what allows us to determine what the distribution would look like if we could take many samples.
  - The distribution often relies on **model assumptions**.

- Get p-value
  - This is the probability of obtaining an observation as far, or farther, away from what we expected if the null was true.
- Make conclusion in context.
  - If the p-value is small ($< \alpha$), then it is unlikely that we would have obtained our observation if the null hypothesis is true. This provides evidence that the observed difference between our observation and expectation is REAL, and not simply due to chance.

2.1 Toluca Example:

If model assumptions are satisfied, then $b_1 \sim N(0, \sigma^2\{b_1\}$.

$\sim$ means "follows" while $\sigma^2\{b_1\}$ represents the true variance of b_1, as estimated by $s\{b_1\}$.

Our test statistic then becomes

$$t = \frac{b_1 - 0}{s\{b_1\}} \sim t_{df_E} = 10.29$$

with $25 - 2 = 23$ degrees of freedom with a **p-value** $<$ **0.0001**.

where df_E is the degrees of freedom for the residuals, which is $n-2$ in the simple linear model case

draw t-distribution and shade the area that represents the p-value

Since our p-value is lower than our level of significance (which is typically 0.05 and something we set beforehand), we would **reject** the null hypothesis **and conclude** that there is significant evidence that lot size and work hours are linearly related.

Where did $\alpha = 0.05$ come from?

Short answer: Sir Ronald Fisher, a prominent statistician, made it up:

> It is a common practice to judge a result significant, if it is such a magnitude that it would have been produced by chance not more frequently than once in twenty trials. This is an arbitrary, but convenient, level of significance for the practical investigator...[1]

However, $\alpha = 0.05$ has proven to be a good level of significance that balances the probability of Type I (claiming a difference when there isn't one) and Type II (claiming *no* difference when there *is* one).

[1] As on p99 of "The Lady Tasting Tea" (2001) by David Salsburg. See `http://jse.amstat.org/v16n2/velleman.pdf` for more discussion about statistical theory.

Consider the following

You wish to determine if Aggie ice cream is more fattening than other ice cream shops in Logan. Suppose your null hypothesis is: "Aggie ice cream has the same number of calories per cup as Charlie's ice cream." You then conduct a test and obtain a p-value of 0.048, indicating that there is evidence that the average caloric counts are significantly different. You then realize that you forgot to include 5 sample in your study. When you include these additional samples, you obtain a p-value of 0.052, indicating no significant difference.

(Groups) What will be your final conclusion based on this information?

It depends. What is the cost of claiming a difference when there isn't one, versus the cost of claiming no difference when there is one?

P-values should inform an analysis, rather than become the analysis.

Confidence Intervals

(Groups) Why are confidence intervals equivalent to hypothesis testing?

If the expected value under the null hypothesis falls outside of a $x\%$ confidence interval, then we know that the p-value of the test statistic is less than $(100 - x)\%$.

- General Form:

$$\text{estimate} \pm (\text{critical value}) \times (\text{SE of estimate})$$

- For β_1:

$$b_1 \pm t^* \times s\{b_1\}$$

- Interpretation:
  - We are 95% confident that the true value of β_1 is contained in this interval.
  - If we were to create 100 confidence intervals from 100 different samples, we would expect about 95 of them to contain the true β_1.

Testing $H_0 : \beta_1$ at level α is the same as checking whether 0 is inside the $(1 - \alpha)100\%$ CI for β_1.

(Individual) Why might the confidence interval approach be preferred to the p-value approach?

Confidence intervals tell us the magnitude of the estimate relative to the standard error of the estimate.

(Individual) Why are we not usually interested in confidence intervals for β_0?

The intercept tells us the expected value of Y when X is 0, which is often nonsensical.

Model Inference

All previous examples test whether an individual X variable has a significant linear relationship with Y. We will now look at some measures of model usefulness that apply when there is more than one X variable.

Ingredients of Model Inference

- Sum of Squares
  - $SS_{total} = \sum_i \left(Y_i - \bar{Y}\right)^2 \propto$ variance of Y
  - $SS_{error} = \sum_i \left(Y_i - \hat{Y}_i\right)^2 = \sum_i e_i^2 \propto$ variance not explained by model
  - $SS_{model} = SS_{total} - SS_{error}$

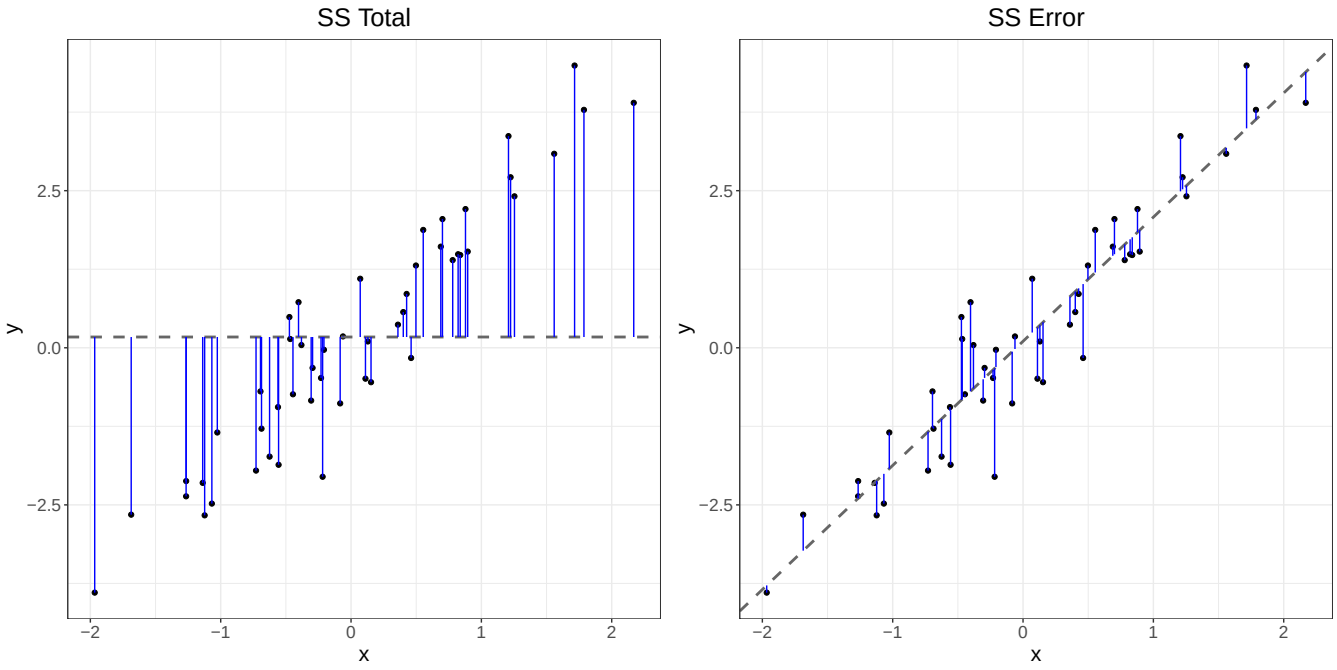

Figure 1: Illustration of SS_{total} and SS_{error}.

- Mean Square: $MS = \frac{SS}{df}$
- $F = \frac{MS_{model}}{MS_{error}}$
- $R^2 = \frac{SS_{model}}{SS_{total}} = 1 - \frac{SS_{error}}{SS_{total}}$
  - Interpretation: the percent of the variation in Y that is explained by the model.
- MSE = Mean Square Error = $\hat{\sigma}^2$ = our best estimate of the error variance ($\epsilon \sim N(0, \sigma^2)$).

Toluca Example:

$R^2 = 0.82$ (from Handout 2.1.1) which means that about 82% of the variation in work hours is explained by lot size.

Two other ways to look at $H_0 : \beta_1 = 0$:

1. How much worse would the model fit be if we dropped the β_1 term?

   Reduced Model (null hypothesis): $Y_i = \beta_0 + \epsilon_i$.

   Full Model: $Y_i = \beta_0 + \beta_1 X_{i,1} + \epsilon_i$.

   F-statistic looks at chance in SS_{error} between these two models.

   Can be extended to consider removal of *subsets* of X variables (more later in the semester).

2. Let $\rho = Corr(X, Y) =$ true, unknown correlation coefficient, which we estimate with the sample correlation (r).

   - When there is only one x-variable in the model it follows that

$$H_0 : \beta_1 = 0 \equiv H_0 : \rho = 0.$$

Inference on the Response Variable Y

We can create interval estimates for the response variable.

$$\hat{Y} \pm t_{df_E}(1 - \frac{\alpha}{2}) * SE\{\hat{Y}\}$$

Two Intervals:

- **Confidence Interval**: Interval estimate of **mean** (or expected) Y for **population** of all $X = X_h$.

$$SE\{\hat{Y}\} = s\{\hat{Y}_h\} = \hat{\sigma}^2 \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_i (X_i - \bar{X})^2}\right]$$

- **Prediction Interval**: Interval estimate of **predicted** Y for a single [new] observation at $X = X_h$

$$SE\{\hat{Y}\} = s\{\hat{Y}_{h(new)}\} = \hat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_i (X_i - \bar{X})^2}\right]$$

(Groups) Why would prediction intervals always be larger than confidence intervals?

The variability of individuals is ALWAYS more than the variability of groups.

Toluca Example (with $X_h = 10$)

- If we were to go to a new (single) lot of 10 acres, we are 95% confident that the work hours would be between

  -13.6 (truncate at 0) and 209.7.

- If we were to consider all possible 10 acre sized lots, we are 95% confident that the mean work hours across all these lots would be between

  50.5 and 145.6.

Note: Most models have more than one predictor variable, we will use the following common notation throughout the remainder of this course:

- $n =$ sample size
- $p =$ number of β_j's in the model (including the intercept)
- $df_E = n - p$