36-401, Chapter 3: Inference for β Parameters in Linear Regression

Zach Branson, Fall 2025

Basic Properties of the $\widehat{\beta}$

In this chapter, we'll still consider the **simple linear regression model**, which involves parameters β_0, β_1, and σ^2. We'll focus on the **least squares estimators** $\hat{\beta}_0$ and $\hat{\beta}_1$, and study properties about these estimators. We'll also consider consequences of having to estimate the variance σ^2.

First, let's recall the three basic assumptions of simple linear regression:

1. **Mean-Zero Noise**: $\mathbb{E}(\epsilon_i \mid X_i) = 0$ for all i
2. **Constant Variance** ("homoskedasticity") $\text{Var}(\epsilon_i \mid X_i) = \sigma^2$ for all i
3. **Uncorrelated Noise**: The ϵ_i are uncorrelated, i.e., $\text{Cov}(\epsilon_i, \epsilon_j \mid X_i) = 0$ for all $i \neq j$.

As we've seen, the least squares estimators are:

$$\begin{aligned}
\widehat{\beta}_1 &= \frac{\sum_i \left(Y_i - \overline{Y}\right)\left(X_i - \overline{X}\right)}{\sum_i \left(X_i - \overline{X}\right)^2} = \frac{s_{xy}}{s_{xx}} \\
\widehat{\beta}_0 &= \overline{Y} - \widehat{\beta}_1 \overline{X}.
\end{aligned}$$

Under the above assumptions, we can find that the following properties for $\widehat{\beta}_0$ and $\widehat{\beta}_1$ hold:

To prove the above results, it is useful to note the following fact: $\widehat{\beta}_1$ can be written as a **linear combination** of the Y_i in the following way.

$$\widehat{\beta}_1 = \sum_{i=1}^{n} k_i Y_i, \quad \text{where } k_i = \frac{X_i - \overline{X}}{\sum_{j=1}^{n} \left(X_j - \overline{X}\right)^2}$$

Exercise 0.1. Prove that the variance of $\widehat{\beta}_1$ takes the stated form.

An important quantity for inference will be the **standard error**. The standard error of an estimator is the square root of its variance. As we saw above, the variance may have to be estimated, such that the standard error will also have to be estimated. Thus, we have:

$$\text{SE}(\hat{\beta}_1) = \sqrt{\text{Var}(\hat{\beta}_1)}, \quad \text{and} \quad \widehat{\text{SE}}(\hat{\beta}_1) = \sqrt{\widehat{\text{Var}}(\hat{\beta}_1)}$$

The *estimated* SEs are displayed in `summary()` output in R.

Exercise 0.2. Let's consider the BEA example from previous chapters; `summary()` output is shown below. State and interpret the standard errors.

```
bea <- read.csv("data/bea-2006.csv")
bea_fit <- lm(pcgmp ~ log(pop), data = bea)
summary(bea_fit)
```

```
## 
## Call:
## lm(formula = pcgmp ~ log(pop), data = bea)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -21572  -4765  -1016   3686  40207 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -23306.2     4957.1  -4.702 3.67e-06 ***
## log(pop)      4449.8      390.9  11.383  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 7929 on 364 degrees of freedom
## Multiple R-squared:  0.2625,Adjusted R-squared:  0.2605 
## F-statistic: 129.6 on 1 and 364 DF,  p-value: < 2.2e-16
```

Adding on the Normality Assumption

So far, we have not made any distributional assumptions. Given additional distributional assumptions, we will have additional properties that will be useful for inference. We'll focus our discussion on inference for β_1.

Exercise 0.3. Assume $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$, where $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. Because $\widehat{\beta}_1 = \sum_{i=1}^{n} k_i Y_i$, what can we say about the distribution of $\widehat{\beta}_1$ in this case?

From now on, when we assume $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, we will call the resulting model the **normal simple linear regression model**.

A key result: Under the normal simple linear regression model,

$$\left(\widehat{\beta}_1 - \beta_1\right) \Big/ \widehat{\mathrm{SE}}(\widehat{\beta}_1) = \left(\widehat{\beta}_1 - \beta_1\right) \Big/ \left(\frac{\widehat{\sigma}^2}{\sum_i \left(X_i - \overline{X}\right)^2}\right)^{1/2} \sim t_{n-2}$$

Note that the ratio $\widehat{\beta}_1/\widehat{\mathrm{SE}}(\widehat{\beta}_1)$ is named the "`t value`" in the R output.

Exercise 0.4. Comment on the practical interpretation of the "t value."

When the degrees of freedom is small, the t distribution has heavier tails than a standard Normal. Meanwhile, as the degrees of freedom becomes large, the t distribution looks more and more like a standard Normal.

In what follows, we'll consider **confidence intervals** based on t-distribution quantiles. These will be very similar to Normal quantiles for large samples.

Exercise 0.5. Use the above “key result” to show that

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

is a $100(1-\alpha)\%$ confidence interval for β_1.

(Here $t_{1-\alpha/2,n-2}$ refers to the $1-\alpha$ quantile of the t_{n-2} distribution. It can be calculated in R with `qt(1 - alpha/2, df = n - 2)`.)

Example 0.1. Here we construct 95% confidence intervals for the linear regression used in the BEA example.

First, we can look up the appropriate t quantile for $\alpha = 0.05$, and find $\hat{\beta}_1$ and $\widehat{\mathrm{SE}}(\hat{\beta}_1)$:

```
#compute t quantile
t_quant <- qt(1 - 0.05/2, df = nrow(bea) - 2)
#obtain beta-hat
betaHat = coef(bea_fit)["log(pop)"]
#obtain estimated SE
betaHat.se = coef(summary(bea_fit))["log(pop)", "Std. Error"]
#manual calculation of CI:
c(betaHat - betaHat.se*t_quant, betaHat + betaHat.se*t_quant)


## log(pop) log(pop)
## 3681.029 5218.486


#using confint() for CI:
confint(bea_fit)


##                  2.5 %     97.5 %
## (Intercept) -33054.300 -13558.099
## log(pop)      3681.029   5218.486
```

Reporting and Interpreting Estimates

With confidence intervals, we now have what we need to report our regression estimates *with uncertainty*. We must communicate the uncertainty clearly to readers so they can understand our results.

Interpreting $\hat{\beta}_0$ and $\hat{\beta}_1$

To report estimates, we must first be able to interpret what they mean. Consider simple linear regression with the model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i.$$

We use least squares to obtain $\hat{\beta}_0$ and $\hat{\beta}_1$ for a particular sample of data.

Exercise 0.6. Interpret $\hat{\beta}_0$ and $\hat{\beta}_1$.

Reporting Estimates with Uncertainty

Uncertainty is important, because simply reporting "$\hat{\beta}_1 = 0.42$" may conceal a great deal. Results should *always* be reported with confidence intervals, or at least SEs, so readers can see the scale of uncertainty.

Example 0.2. Suppose we surveyed CMU students to ask (a) how many hours they sleep per night and (b) their GPA. We fit a regression using hours of sleep as the covariate and GPA as the outcome, hoping to understand how sleep habits relate to grades.

Exercise 0.7. Suppose we obtain $\hat{\beta}_1 = 0.4$. Interpret this result in context, and show why the size of the confidence interval matters a lot.

Exercise 0.8. Suppose instead that we obtained $\hat{\beta}_1 = 0.002$. Show why the size of the confidence interval matters a lot.

Hypothesis Testing for β_1

Review: There are five key components to a **statistical hypothesis test**:

1. *Null hypothesis* H_0: Tentative assumption that an effect or parameter is "null" (or equal to zero). We'll assess to what extent the data are consistent with H_0. For example: $H_0 : \beta_1 = 0$.
2. *Alternative hypothesis* H_A: Characteristic about an effect or parameter we assume if we reject the null hypothesis. For example: $H_A : \beta_1 \neq 0$.
3. *Test statistic*: Measures how consistent the data are with H_0. Ideally, (1) the more "false" H_0 becomes, the more the test statistic changes; and (2) we know its distribution when H_0 is true.
4. *Rejection region*: Range of test statistic values for which we reject H_0.
5. *Significance level* α: Determines size of rejection region and frequency we're willing to falsely reject when H_0 is true.

Exercise 0.9. To test $H_0 : \beta_1 = 0$, a natural test statistic is $T = \widehat{\beta}_1 \Big/ \widehat{\text{SE}}(\widehat{\beta}_1)$. Why? Furthermore, what's the practical value of testing $H_0 : \beta_1 = 0$?

Recall that we need to consider separately the **one-sided** and **two-sided** hypothesis tests. In this case,

If testing. . .	**reject** H_0 **if** . . .
$H_0\colon \beta_1 = 0$ versus $H_A\colon \beta_1 > 0$	$T > t_{1-\alpha,n-2}$
$H_0\colon \beta_1 = 0$ versus $H_A\colon \beta_1 < 0$	$T < -t_{1-\alpha,n-2}$
$H_0\colon \beta_1 = 0$ versus $H_A\colon \beta_1 \neq 0$	$T > t_{1-\alpha/2,n-2}$ **or** $T < -t_{1-\alpha/2,n-2}$

Importantly, the **p-value** can be calculated using the appropriate **tail probability**. We reject H_0 if $p < \alpha$.

Exercise 0.10. Depict the appropriate tail probability to be calculated in each of the three possibilities given above. Furthermore, clarify what distribution you use to compute the tail probability.

R provides p-values for **two-sided** alternatives in its `Coefficients` table.

Exercise 0.11. Below is the BEA example again. Make a statistical and practical interpretation of hypothesis testing regarding β_1.

```
  summary(bea_fit)


##
## Call:
## lm(formula = pcgmp ~ log(pop), data = bea)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -21572  -4765  -1016   3686  40207
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -23306.2     4957.1  -4.702 3.67e-06 ***
## log(pop)      4449.8      390.9  11.383  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7929 on 364 degrees of freedom
## Multiple R-squared:  0.2625,Adjusted R-squared:  0.2605
## F-statistic: 129.6 on 1 and 364 DF,  p-value: < 2.2e-16
```

Testing & p-values: Uses & Abuses

Statistical vs. practical significance

If we test $H_0 : \beta_1 = b$ and reject it, it is common to say that the difference between β_1 and b is **statistically significant**. Because many professions have an overwhelming urge to test $H_0 : \beta_1 = 0$, it's common for people to say "β_1 is statistically significant" when they really mean "β_1 is statistically significantly different from zero."

However, despite the word "significance," within a given application, rejecting a null hypothesis may not be practically meaningful in and of itself. For example, we may reject the null hypothesis that $H_0 : \beta_1 = 0$, thereby thinking that there is a "significant association." However, after looking at our confidence interval, we may be confident that the association, while non-zero, is very small and "practically insignificant." For example, perhaps we were 95% confident that the true β_1 for a linear regression that uses GPA as the outcome and hours of sleep as the covariate is [-0.0002, -0.0001]. Would this motivate you to sleep less in order to increase your GPA, because there is a statistically significant effect?

Relatedly, there is a notable difference between being quite confident in a very small but non-zero effect, and having no certainty about the effect (whether it's zero, very negative, or very positive). In the former situation, we are confident that a large effect does not exist (at least given the data at hand, and assuming our modeling assumptions are correct); but in the second situation, we cannot declare any magnitude about the effect. This is why we stated earlier that it is good practice to report confidence intervals.

It is also always tempting to "accept" the null when we fail to reject it, but failing to reject is not synonymous with accepting. There are at least two reasons why we might fail to reject the null $\beta_1 = 0$:

1. β_1 is, in fact, zero,
2. $\beta_1 \neq 0$, but $\text{SE}(\hat{\beta}_1)$ is so large that we can't tell anything about β_1 with any confidence.

Even a huge $\widehat{\beta}_1$ can be statistically insignificant if the standard error is large enough. Conversely, a very small $\widehat{\beta}_1$ will become statistically significant once its standard error is small enough. Since the standard error goes to zero as $n \to \infty$, the t value $\frac{\hat{\beta}_1}{\mathrm{SE}(\beta_1)} \to \pm\infty$, unless β_1 is exactly 0.

Statistical significance involves a mixture of effect size, sample size, and variation. Unfortunately, this has little to do with "significance" in the typical, ordinary language sense. Perhaps a better phrase is "statistically detectable" or "statistically distinguishable from 0". However, for the time being, "statistical significance" is a commonly used phrase, and thus we should be aware of its implicit nuances.

Misinterpretations of the p-value

The p-value has become somewhat controversial due to its widespread misuse. Here we try to clarify some common misconceptions:

- a large p-value is NOT strong evidence in favor of H_0
- the p-value is NOT equal to $P(H_0$ true given the data$)$
- $p = 0.048$ is not very different from $p = 0.053$, even though only the former leads to rejecting H_0 at level $\alpha = 0.05$.

However, we should not abandon the p-value simply because people misuse it. We should improve statistical literacy, and be honest about the limitations/meanings of p-values. It's also important to keep in mind that confidence intervals have similar limitations, in the sense that we can use them to make fail-to-reject and reject hypothesis testing conclusions by checking if zero is contained in the confidence interval. But, confidence intervals at least give some quantification of uncertainty, which p-values do not.