Module 2: Inference with Simple Linear Regression

Jeffrey Woo

MSDS, University of Virginia

Assumptions for Linear Regression Model

(a)

$$\mathrm{E}(\epsilon_i) = 0, \quad 1 \leq i \leq n;$$
$$\mathrm{Var}(\epsilon_i) = \sigma^2, \quad 1 \leq i \leq n;$$
$$\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0, \quad i \neq j.$$

(b) $\epsilon_1, \ldots, \epsilon_n$ i.i.d. $\sim \mathrm{N}(0, \sigma^2)$ (i.i.d. means independent and identically distributed)

Assumptions (a) and (b) allow us to proceed with statistical inference. We will cover how we assess if these are met in Module 3.

Confidence Intervals

Goal: Provide a range of plausible values for the unknown parameter of interest, e.g. β_1.
General form for CI:

$$\text{statistic} \pm (\text{multiplier} \times \text{s.e of statistic}).$$

- Statistic: numerical quantity that describes a sample
- Multiplier: determined by confidence level and relevant probability distribution
- Standard error of statistic: measure of precision of the statistic

CI for β_1, σ unknown

A $100(1-\alpha)\%$ CI for β_1 when σ is unknown is

$$\hat{\beta}_1 \pm t_{1-\alpha/2;df}\, se(\hat{\beta}_1) \tag{1}$$

where $se(\hat{\beta}_1) = \sqrt{\frac{MS_{res}}{S_{xx}}}$. The $df = n-2$, or more generally $n-p$.

Question: Do you recall how to find $t_{1-\alpha/2;df}$ in R?

Hypothesis Testing

Investigate if a population parameter is equal to a specific value. In the context of simple linear regression, we usually want to test if β_1 is 0 or not. If $\beta_1 = 0$, there is no linear relationship between the variables. We usually state the hypotheses statements as null and alternative hypotheses.

- Step 1: State the null and alternative hypotheses.
- Step 2: A test statistic is calculated using the sample, **assuming null is true**. The value of the test statistic is affected by the degree to which the sample deviates from the null.
- Step 3: Make conclusion. Two equivalent approaches: critical region and p-value approach.

t Statistic

General framework of a t statistic is

$$t = \frac{\text{statistic} - \text{value of parameter under null}}{\text{s.e of statistic}}.$$

t Test

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

In this setting for a linear regression, the t statistic is

$$t = \frac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)}.$$

Two-Sided Hypothesis Testing

Critical region approach: Reject H_0 at level of significance α if $|t| > t_{1-\alpha/2;n-2}$.

P-value approach: The two-sided p-value is $2 \times P\left\{t_{n-2} > |t|\right\}$. If p-value is less than α, reject H_0.

NOTE: Rejecting H_0 indicates our data support H_a. Not rejecting H_0 indicates our data do not support H_a.

Using R

Critical region: For two-sided test, type qt$(1-\alpha/2, df)$, where α is the significance level and $df = n-2$ in SLR.

Question: How does this change if we have a one-sided test?

NOTE 1: Conclusions from 2-sided tests and CI at the same significance level are always consistent.

NOTE 2: P-value approach and critical value always lead to same conclusions at the same significance level.

ANOVA F Test

In SLR, the ANOVA F test tests

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

The ANOVA F statistic is

$$F = \frac{MS_R}{MS_{res}}.$$

Results from the ANOVA F test and t test are the same in SLR when $H_0 : \beta_1 = 0$ and the alternative is two sided.

R Output

```
> result<-lm(purity~hydro)
> summary(result)

Call:
lm(formula = purity ~ hydro)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6724 -3.2113 -0.0626  2.5783  7.3037

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   77.863      4.199  18.544 3.54e-13 ***
hydro         11.801      3.485   3.386  0.00329 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.597 on 18 degrees of freedom
Multiple R-squared:  0.3891,    Adjusted R-squared:  0.3552
F-statistic: 11.47 on 1 and 18 DF,  p-value: 0.003291
```

Two Interpretations of $\hat{y}$

Regression equation in population: $\mu_y = \beta_0 + \beta_1 x$
Regression equation in sample: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.
Two interpretations of $\hat{y}$:

1. $\hat{y}$ estimates μ_y, the mean of the response for specific value of x.
2. $\hat{y}$ predicts the value of the response for a specific observation of x.

Depending on which interpretation you are interested in, we have 2 CIs, one for the mean of the response for specific x, another for the prediction of the response for a specific observation of x.

Outline

Distribution of $\hat{\mu}_{y|x_0}$

Sometimes, we want to find a CI for the **mean response**, $\mathrm{E}(y|x_0)$, at $x = x_0$. The point estimator of $\mathrm{E}(y|x_0)$ is $\hat{\mu}_{y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

If σ^2 is unknown, a $100(1-\alpha)\%$ CI for μ_0 is

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

Outline

Prediction for a New Observation

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 for a future observation y_0.

Prediction for New Observations

Prediction interval for $\hat{y}_0$ takes into account

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

By comparison, the CI for $\mathrm{E}(y|x_0)$ only deals with the first element.

Prediction for New Observations

For unknown σ^2, the prediction interval is

$$\hat{y}_0 \pm t_{1-\alpha/2,n-2} s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}. \tag{3}$$

Comparing (3) with (2), PI is always wider.

Pointwise Confidence and Prediction Intervals