BIOS 7345: Advanced Regression Analysis I

Andrew J. Spieker, Ph.D.

Assistant Professor of Biostatistics
Vanderbilt University Medical Center

Set 9: Confidence and prediction intervals

Version: 09/28/2023

Table of Contents

Confidence intervals for regression parameters

Recall:

- We spent a meaningful amount of time developing the theory for inference regarding regression parameters.
- For these notes, we'll focus on "inverting" the tests in order to create confidence regions/intervals.
- Model: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{X}_{N\times K}$ is of full rank (for simplicity of presentation), $\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}$, and $\mathrm{Cov}[\boldsymbol{\epsilon}] = \sigma^2\mathbf{I}$.
- I will make the formal assumption that $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, so that the results presented are exact. However, the results surrounding confidence intervals are otherwise approximate for sufficiently large samples.
- Throughout, let $\widehat{\boldsymbol{\beta}}$ denote the OLS estimator.

Special case of the *F*-statistic:

- Throughout, we've written the F-statistic as:

$$F = \frac{(\text{RSS}_H - \text{RSS})/Q}{\text{RSS}/(N-K)} = \frac{(\mathbf{C}\widehat{\boldsymbol{\beta}})^T(\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T)^{-1}\mathbf{C}\widehat{\boldsymbol{\beta}}/Q}{S^2}$$

- In these notes, we will generally be choosing $\mathbf{C} = \mathbf{c}^T$ (one row; $Q = 1$), in which case we can freely write the F-statistic as:

$$F = \frac{(\mathbf{c}^T\widehat{\boldsymbol{\beta}})^2}{S^2\mathbf{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{c}}.$$

- Under $H_0 : \mathbf{c}^T\boldsymbol{\beta} = 0$, this will follow an $F_{1,N-K}$ distribution.

Lemma 9.1: Relationship between the t- and F-distribution

Let $Z \sim \mathcal{N}(0,1)$ and let $U \sim \chi^2_K$, with $U \perp\!\!\!\perp Z$. Further, let

$$T = \frac{Z}{\sqrt{U/K}}.$$

Then, $T \sim t_K$ and $T^2 \sim F(1,K)$.

Justification for the *t*-statistic:

- Rather than using the F-statistic:

$$F = \frac{(\mathbf{c}^T\widehat{\boldsymbol{\beta}})^2}{S^2\mathbf{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{c}},$$

  note that Lemma 9.1 gives justification instead for the t-statistic:

$$t = \frac{\mathbf{c}^T\widehat{\boldsymbol{\beta}}}{S\sqrt{\mathbf{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{c}}}.$$

- Under $H_0 : \mathbf{c}^T\boldsymbol{\beta} = 0$, this will follow an t_{N-K} distribution.

Confidence intervals for regression parameters

Important note:

- Suppose that the value of $\mathbf{c}^T\boldsymbol{\beta}$ is not zero. Then,

$$\frac{\mathbf{c}^T\widehat{\boldsymbol{\beta}} - \mathbf{c}^T\boldsymbol{\beta}}{S\sqrt{\mathbf{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{c}}} \sim t_{N-K}$$

- Note that the left-hand side is not a statistic, but a just a pivotal quantity (one for which the distribution does not depend upon unknown parameters).

Confidence intervals for regression parameters

A confidence interval for β_k:

- Suppose we seek to develop a confidence interval for β_k.
- Letting $V_k = [\text{diag}((\mathbf{X}^T\mathbf{X})^{-1})]_{k,k}$, we have that

$$\frac{\widehat{\beta}_k - \beta_k}{S\sqrt{V_k}} \sim t_{N-K}.$$

- This suggests the following $100(1-\alpha)\%$ confidence interval for β_k:

$$\widehat{\beta}_k \pm t_{1-\alpha/2,N-K} S\sqrt{V_k}.$$

- Two interpretations:
  - $100(1-\alpha)\%$ of intervals generated in this way contain β_j.
  - This interval contains the set of all null hypotheses β_j^* such that $H_0 : \beta_j = \beta_j^*$ cannot be rejected at the nominal level α (two-sided).

Confidence intervals for regression parameters

A confidence interval for $\mathbf{c}^T\boldsymbol{\beta}$:

- Suppose, more generally, that we seek to develop a confidence interval for $\mathbf{c}^T\boldsymbol{\beta}$. Then,

$$\frac{\mathbf{c}^T\widehat{\boldsymbol{\beta}} - \mathbf{c}^T\boldsymbol{\beta}}{\sqrt{S^2\mathbf{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{c}}} \sim t_{N-K}.$$

- This suggests the following $100(1-\alpha)\%$ confidence interval for $\mathbf{c}^T\boldsymbol{\beta}$:

$$\mathbf{c}^T\widehat{\boldsymbol{\beta}} \pm t_{1-\alpha/2,N-K}S\sqrt{\mathbf{c}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{c}}.$$

- Quantities of the form $\mathbf{c}^T\boldsymbol{\beta}$ include comparisons of non-reference categories, subgroup-specific effects, etc.

Table of Contents

Confidence intervals for means

A confidence interval for $\mathbf{x}_0^T\boldsymbol{\beta}$:

- Parameters of the form $\mathbf{c}^T\boldsymbol{\beta}$ include subgroup means: $\mathrm{E}[\mathbf{y}|\mathbf{x}_0] = \mathbf{x}_0^T\boldsymbol{\beta}$.
- Indeed, $\mathbf{x}_0$ does not need to mark a subgroup that is specifically represented in the data (though it should be "in the range" of the covariate space to avoid extrapolation).
- Letting $\widehat{y}(\mathbf{x}_0) = \mathbf{x}_0^T\widehat{\boldsymbol{\beta}}$, we have that:

$$\frac{\widehat{y}(\mathbf{x}_0) - \mathbf{x}_0^T\boldsymbol{\beta}}{\sqrt{S^2\mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0}} \sim t_{N-K}.$$

- This suggests the following $100(1-\alpha)\%$ confidence interval for $\mathbf{c}^T\boldsymbol{\beta}$:

$$\widehat{y}(\mathbf{x}_0) \pm t_{1-\alpha/2,N-K}S\sqrt{\mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0}.$$

Confidence intervals for means

A confidence interval for $\mathbf{x}_0^T\boldsymbol{\beta}$:

- In the special case of simple linear regression, $\mathbf{x}_0 = \begin{bmatrix} 1 & x_0 \end{bmatrix}^T$, and the following $100(1-\alpha)\%$ confidence interval can be derived:

$$\widehat{y}(x_0) \pm t_{1-\alpha/2,N-2} S \sqrt{\left(\frac{1}{N} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{N} (x_i - \overline{x})^2} \right)}.$$

Table of Contents

A confidence interval for the error variance

A confidence interval for σ^2:

- Recall: $\text{RSS}/\sigma^2 = (N-K)S^2/\sigma^2 \sim \chi^2_{N-K}$.
- This suggests the following $100(1-\alpha)\%$ confidence interval for $1/\sigma^2$:

$$\left[\frac{\chi_{\alpha/2,N-K}}{(N-K)S^2}, \frac{\chi_{1-\alpha/2,N-K}}{(N-K)S^2}\right]$$

- This suggests the following $100(1-\alpha)\%$ confidence interval for σ^2:

$$\left[\frac{N-K}{\chi_{\alpha/2,N-K}}S^2, \frac{N-K}{\chi_{1-\alpha/2,N-K}}S^2\right]$$

- This suggests the following $100(1-\alpha)\%$ confidence interval for σ:

$$\left[\sqrt{\frac{N-K}{\chi_{\alpha/2,N-K}}}S, \sqrt{\frac{N-K}{\chi_{1-\alpha/2,N-K}}}S\right]$$

Table of Contents

Prediction intervals

Future/external observations:

- Imagine we want to use a model to establish a range of typical values for outcomes corresponding to a covariate profile $\mathbf{x}_0$.
  - Also termed a prediction interval.
  - Characterizes a "reference range."
- Similar to a confidence interval, we want a certain percentage of intervals we create in this fashion to capture the value of Y_0, a random variable marking a future observation with covariate vector $\mathbf{x}_0$.

Prediction intervals

Future/external observations:

- Let $\widehat{Y}(\mathbf{x}_0)$ denote the fitted value (estimated mean value of Y in the subgroup $\mathbf{x}_0$.
- By assumption, the future observation Y_0 is taken to be independent of all observations that contributed to our fitted value, $\widehat{Y}(\mathbf{x}_0)$.
- Therefore, we find that:

$$\begin{aligned}
\text{Var}[Y_0 - \widehat{Y}(\mathbf{x}_0)] &= \text{Var}[Y_0] + \text{Var}[\widehat{Y}(\mathbf{x}_0)] \\
&= \sigma^2 + \text{Var}[\mathbf{x}_0^T \widehat{\boldsymbol{\beta}}] \\
&= \sigma^2 + \mathbf{x}_0^T[\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}]\mathbf{x}_0 \\
&= \sigma^2(1 + \mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0),
\end{aligned}$$

which can be estimated as:

$$\widehat{\text{Var}}[Y_0 - \widehat{Y}(\mathbf{x}_0)] = S^2(1 + \mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0),$$

Future/external observations:

- Since $Y_0 \perp\!\!\!\perp \widehat{Y}(\mathbf{x}_0)$ and $S^2 \perp\!\!\!\perp \widehat{Y}(\mathbf{x}_0)$, we have:

$$\frac{Y_0 - \widehat{Y}(\mathbf{x}_0)}{S\sqrt{(1 + \mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0)}} \sim t_{N-K}.$$

- This suggests the following $100(1-\alpha)\%$ prediction interval for Y_0:

$$\widehat{y}(\mathbf{x}_0) \pm t_{1-\alpha/2,N-K} S\sqrt{1 + \mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0}.$$

Prediction intervals

Future/external observations:

- In the special case of simple linear regression, $\mathbf{x}_0 = \begin{bmatrix} 1 & x_0 \end{bmatrix}^T$, and the following $100(1-\alpha)\%$ prediction interval can be derived:

$$\widehat{y}(x_0) \pm t_{1-\alpha/2,N-2} S \sqrt{\left(1 + \frac{1}{N} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{N}(x_i - \overline{x})^2}\right)}.$$

Future/external observations: Simulation (setup)

```
## Sample size
n <- 100

## Number of simulations
nsim <- 5000

## Simulation parameters
beta0 <- 1
beta1 <- 2
sigma <- 2

## Predictor of interest (fixed)
x <- runif(n, 1, 4)
xbar <- mean(x)

## Subgroup of interest
x0 <- 4

## Place to store results
coverage <- matrix(1, nrow = nsim, ncol = 1)
```

Future/external observations: Simulation (data generation/model fit)

```
for (j in 1:nsim)
{
  ## External observation
  y0 <- rnorm(1, beta0 + beta1 * x0, sigma)

  ## Generate outcome
  y <- beta0 + beta1*x + rnorm(n, 0, sigma)

  ## OLS (coefficients and error variance)
  zz <- lm(y ~ x)
  bhat <- coef(zz)
  S <- sqrt(sum(zz$residuals^2)/(n - 2))

  ## Critical value
  cv <- qt(0.975, df = n - 2)
```

Future/external observations: Simulation (interval and results)

```
  ## Fitted value
  yhat <- as.numeric(bhat[1] + bhat[2]*x0)

  ## Prediction interval
  pilo <- yhat - cv*S*sqrt(1 + 1/n + (x0 - xbar)/((n - 1)*var(x)))
  pihi <- yhat + cv*S*sqrt(1 + 1/n + (x0 - xbar)/((n - 1)*var(x)))

  ## Decide whether coverage was achieved
  if (pilo > y0 | pihi < y0) {coverage[j] <- 0}
}

## > mean(coverage)
## [1] 0.9572
```

Summary

So far:

- Confidence and prediction.

Summary

Up next:

- Diagnostics.