Testing and confidence intervals

Emilija Perković

Dept. of Statistics
University of Washington

Distributions of estimators

We assume:

$$\underline{y} = X\underline{\beta} + \underline{\epsilon}$$

with $\underline{\epsilon} \sim \mathcal{N}(\underline{0}, \sigma^2 I_n)$.
Then

$$\underline{y} \sim \mathcal{N}(X\underline{\beta}, \sigma^2 I_n).$$

Since $\underline{\hat{\beta}} = (X'X)^{-1}X'\underline{y}$,

$$\underline{\hat{\beta}} \sim \mathcal{N}(\underline{\beta}, \sigma^2 (X'X)^{-1}).$$

Distributions of estimators

Since

$$\underline{\hat{y}} = X\underline{\hat{\beta}} = X(X'X)^{-1}X'\underline{y},$$

$$\underline{\hat{y}} \sim \mathcal{N}(X\underline{\beta}, \sigma^2 X(X'X)^{-1}X').$$

And for the residuals, we have that $\underline{\hat{\epsilon}} = \underline{y} - X\underline{\hat{\beta}}$, so

$$\underline{\hat{\epsilon}} \sim \mathcal{N}(0, \sigma^2(I_n - X(X'X)^{-1}X')).$$

Then for each $i = 0, \ldots, p$, $\hat{\beta}_i \sim \mathcal{N}(\beta_i, \sigma^2((X'X)^{-1})_{ii})$,

and for each $j = 1, \ldots, n$, $\hat{\epsilon}_j \sim \mathcal{N}(0, \sigma^2(I_n - X(X'X)^{-1}X')_{ii})$, where A_{ii} denotes the i-th row and i-th column element of the matrix A.

Recall: Hypothesis testing

Recall that performing a α level hypothesis test means:

$$P(\text{reject } H_0 \mid H_0 \text{ true}) \leq \alpha.$$

Then

	H_0 true	H_A true
Do not reject H_0	$1-\alpha$	β
Reject H_0	α	$1-\beta$

- α - Type I error,
- β - Type II error.

Ideally, we want both α and β to be small.
However, there is a trade-off to consider. Performing an $\alpha_1 < \alpha$ hypothesis test, implies that the type II error for this test β_1 is larger than β, that is, $\beta_1 > \beta$.

Hypothesis test for single $\hat{\beta}_i$

Under the assumption $\underline{\epsilon} \sim \mathcal{N}(\underline{0}, \sigma^2 I_n)$
- which we assume throughout this lecture -
$\hat{\beta}_i \sim \mathcal{N}(\beta_i, \sigma^2((X'X)^{-1})_{ii})$, so

$$\frac{\hat{\beta}_i - \beta_i}{\sigma\sqrt{((X'X)^{-1})_{ii})}} \sim \mathcal{N}(0, 1).$$

If σ is known, we can test the null hypothesis $H_0 : \beta_i = \beta_i^*$ using the test statistic

$$Z = \frac{\hat{\beta}_i - \beta_i^*}{\sigma\sqrt{((X'X)^{-1})_{ii})}},$$

where $Z \sim \mathcal{N}(0, 1)$ if H_0 is true.

Distributions of estimators | Testing one coefficient | Testing multiple coefficients | Multiple testing

Hypothesis test for single $\hat{\beta}_i$

Performing an α level hypothesis test means that we reject the null hypothesis H_0 for p-values smaller than α.

For testing $H_0 : \beta_i = \beta_i^*$ against $H_1 : \beta_i \neq \beta_i^*$, use the two-sided p-value: $P(|Z| \geq |z| \big| H_0 \text{ true})$.

For testing $H_0 : \beta_i < \beta_i^*$ against $H_1 : \beta_i > \beta_i^*$, use the following one-sided p-value: $P(Z \geq z \big| H_0 \text{ true})$.

Alternatively, for testing $H_0 : \beta_i > \beta_i^*$ against $H_1 : \beta_i < \beta_i^*$, use the following one-sided p-value $P(Z \leq z \big| H_0 \text{ true})$.

$1-\alpha$ confidence interval for $\hat{\beta}_i$

If $\beta_i = \beta_i^*$, $Z \sim \mathcal{N}(0,1)$, we can choose $z_{1-\alpha/2}$ such that

$$P(-z_{1-\alpha/2} \leq Z \leq z_{1-\alpha/2}) = 1-\alpha,$$

- $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of $\mathcal{N}(0,1)$.

Then

$$P(-z_{1-\alpha/2} \leq \frac{\hat{\beta}_i - \beta_i^*}{\sigma\sqrt{((X'X)^{-1})_{ii})}} \leq z_{1-\alpha/2}) = 1-\alpha,$$

so a two-sided $(1-\alpha)$ confidence interval for β_i is:

$$\left(\hat{\beta}_i - \sigma\sqrt{((X'X)^{-1})_{ii})} \cdot z_{1-\alpha/2}, \hat{\beta} + \sigma\sqrt{((X'X)^{-1})_{ii})} \cdot z_{1-\alpha/2}\right).$$

Testing $\hat{\beta}_i$ when σ is unknown

Often σ is not known a priori. We estimate σ^2 as $\hat{\sigma}^2 = RSS/(n-p-1)$.
To test the null hypothesis $H_0 : \beta_i = \beta_i^*$, we use the following T statistic:

$$T = \frac{\hat{\beta}_i - \beta_i^*}{\hat{\sigma}\sqrt{((X'X)^{-1})_{ii})}},$$

under H_0, $T \sim t_{n-p-1}$ (see Linear models handout for details).
Note that $\hat{\sigma}\sqrt{((X'X)^{-1})_{ii})}$ is the standard error of $\hat{\beta}_i$.

$$SE(\hat{\beta}_i) = \hat{\sigma}\sqrt{((X'X)^{-1})_{ii})}.$$

Exercise: `fuel2001`.
Suppose $\underline{fuel} = \beta_0 + \beta_{Tax}\underline{Tax} + \beta_{Dlic}\underline{Dlic} + \underline{\epsilon}$, with $\underline{\epsilon} \sim \mathcal{N}(\underline{0}, \sigma^2 I_n)$.
Conduct a .05 level hypothesis test of $H_0 : \beta_{Tax} = 0$ (against $H_1 : \beta_{Tax} \neq 0$) and obtain a 95% confidence interval for β_{Tax}.

Testing multiple regression coefficients

Testing whether multiple regression coefficients are zero is usually done by comparing the fits of two regression models.For example comparing

$$\underline{y} = \beta_0 + \beta_1 \underline{x}_1 + \beta_2 \underline{x}_2 + \cdots + \beta_p \underline{x}_p + \underline{\epsilon} \tag{1}$$

and

$$\underline{y} = \beta_0 + \beta_1 \underline{x}_1 + \beta_2 \underline{x}_2 + \cdots + \beta_q \underline{x}_q + \underline{\epsilon} \tag{2}$$

for $0 \leq q \leq p$. Note that $\underline{x}_1, \ldots \underline{x}_q$ are used in both (1) and (2).
To test $H_0 : (\beta_{q+1}, \ldots \beta_p)' = (0, \ldots 0)$ against $H_1 : \beta_i \neq 0$ for at least one $i = q+1, \ldots, p$, we then use the F-statistic

$$F = \frac{(RSS_q - RSS_p)(n-p-1)}{(p-q)RSS_p},$$

where RSS_q is the residual sum of squares after fitting model (1) and RSS_p is the the residual sum of squares after fitting model (2).
If H_0 is true, $F \sim F_{p-q,n-p-1}$ (see Linear model handout for details).

Example: Fuel data

In R this test can be performed by using the `anova(.)` function.

```
> anova(fit3,fit1)
Analysis of Variance Table

Model 1: Fuel ~ Tax + Dlic
Model 2: Fuel ~ Tax + Dlic + Income + logMiles
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     48 289681
2     46 193700  2     95981 11.397 9.546e-05
```

Decision: Reject $H_0 : \beta_{Income} = \beta_{logMiles} = 0$. We prefer the bigger model.

Example: Fuel data

We can also use this test when $q = 0$. This way we compare the full model with the model containing only the intercept ("empty model").

```
> anova(fit.empty,fit1)
Analysis of Variance Table

Model 1: Fuel ~ 1
Model 2: Fuel ~ Tax + Dlic + Income + logMiles
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     50 395694
2     46 193700  4     201994 11.992 9.331e-07
```

Decision: Reject $H_0 : \beta_{Tax} = \beta_{Dlic} = \beta_{Income} = \beta_{logMiles} = 0$. We prefer the bigger model. See also

```
> summary(fit1)
....
F-statistic: 11.99 on 4 and 46 DF,  p-value: 9.331e-07
```


Multiple testing

In hypothesis testing, we control the type I error α:

$$P(\text{ reject } H_0 | H_0 \text{ is true}) = \alpha$$

- α - probability of a false positive.

Suppose you are conducting m independent hypothesis tests: $H_0^1, \dots H_0^m$ each at level α.
The probability of having at least one false positive is:

$$\begin{aligned}
&P(\text{reject } H_0^i, \text{ for at least one } i \,\big|\, \text{all } H_0^1, \dots, H_0^m \text{ true })\\
&= 1 - P(\text{do not reject any } H_0^i \big| \text{ all } H_0^1, \dots, H_0^m \text{ true })\\
&= 1 - \Pi_{i=1}^m P(\text{do not reject } H_0^i \,\big| H_0^i \text{ true }) = 1 - (1-\alpha)^m
\end{aligned}$$

If $m = 20$ and $\alpha = .05$, the probability of at least one false positive is:

$$1 - (1-\alpha)^m = 1 - .95^{20} = 0.6415141$$

Multiple testing

The probability of making at least one false positive discovery when conducting m hypothesis tests is called the Family Wise Error Rate (FWER).
The FWER for m hypothesis tests where each test is conducted at level α is: $\alpha \le FWER \le 1-(1-\alpha)^m$.
$FWER$ is controlled at level α_1, for some pre-chosen α_1, $0 < \alpha_1 < 1$ if

$$FWER < \alpha_1.$$

How to control FWER:

- Bonferroni correction
- Holm correction
- etc.

Bonferroni correction

In order to ensure that $FWER \le \alpha$,

- Perform each of the m hypothesis test at level α/m. That is, reject H_0^i if the p-value for this test p_i: $p_i < \alpha/m$.

This procedure ensures that $FWER \le \alpha$:

$$\begin{aligned}
&P(\text{ reject } H_0^i, \text{ for at least one } i \mid \text{all } H_0^1, \dots, H_0^m \text{ true })\\
&\le \sum_{i=1}^{m} P(\text{ reject } H_0^i \mid \text{all } H_0^1, \dots, H_0^m \text{ true })\\
&= m\frac{\alpha}{m} = \alpha,
\end{aligned}$$

where the second line follows using Boole's inequality.
For example, if $\alpha = .05$ and $m = 20$, then in the worst case:

$$FWER = 1-(1-\alpha/m)^m = 1-(1-.05/20)^{20} = 0.04883012.$$

Bonferroni correction

Alternative view of the Bonferroni procedure:

- Instead of rejecting hypothesis tests with p-values $p_i < \alpha/m$.
- "Adjust" the p-values p_i to obtain adjusted p-values p_i^*:

$$p_i^* = p_i \cdot m.$$

- Reject only the null hypothesis corresponding to hypothesis tests for which $p_i^* < \alpha$.

In R this is implemented in function `p.adjust(.)`.

Holm correction

Less conservative than the Bonferroni correction, but still ensures that $FWER \leq \alpha$.

- Order the p-values form the m hypothesis tests from smallest to largest: $p_{(1)}, \dots p_{(m)}$.
- Calculate $\alpha/(m-i+1)$ for $i \in \{1, \dots, m\}$.
- Let i_0 be the smallest index such that $p_{(i_0)} \geq \frac{\alpha}{m-i_0+1}$.
- Reject only the null hypothesis corresponding to p-values $p_{(1)}, \dots p_{(i_0-1)}$ (if $i_0 = 1$ do not reject any H_0^i and if $p_{(i)} < \frac{\alpha}{m-i+1}$ for all i, reject all H_0^i).

Holm correction

Suppose that you are conducting three hypothesis tests and want to control *FWER* at a .05 level. Suppose additionally that you obtain p-values: $0.02, 0.01, 0.035$.
Using the Holm procedure:

1. Sort the p-values $0.01 < 0.02 < 0.035$.
2. Calculate $\alpha/(m-i+1)$ for all i: $.05/3 \approx 0.017$, $.05/2 = 0.025$, and $0.05/1 = 0.05$.
3. Find the smallest p-value $p_{(i_0)}$ that is larger than its corresponding adjusted significance level.
4. Reject the null hypothesis corresponding to p-values $p_{(1)}, \ldots p_{(i_0-1)}$.

The Holm procedure would advise rejecting all three null hypothesis. By contrast, since $\alpha/m = .017$, the Bonferroni procedure would only advise rejecting the hypothesis corresponding to p-value .01.

Holm correction

What would be an alternative way to phrase the Holm correction procedure using "adjusted" p-values?

1. Order the p-values form the m hypothesis tests from smallest to largest: $p_{(1)}, \ldots p_{(m)}$.
2. Obtain the adjusted p-values $p^*_{(i)}$ as

$$p^*_{(i)} = p_{(i)} \cdot (m - i + 1).$$

3. Find the smallest index i_0 such that $p^*_{(i_0)} \geq \alpha$.
4. Reject the null hypothesis corresponding to the adjusted p-values $p^*_{(1)}, \ldots p^*_{(i_0-1)}$.

In R this is implemented in function `p.adjust(.)`.

Controlling False Discoveries

The FWER criterion aims to control the probability of making even one false rejection among m simultaneous hypothesis tests.

Number of decisions

Decision/Truth	H_0 true	H_A true	Total
Do not reject H_0	U	T	$m-R$
Reject H_0	V	S	R
Total	m_0	$m-m_0$	m

- U - true negatives, S - true positives,
- T - false negatives, V - false positives
- R - number of null hypothesis rejected.

In terms above, $FWER = P(V \geq 1)$. Controlling FWER can prove too conservative when conducting many tests (e.g., if $m > 20$).
It may make more sense to control the proportion of false discoveries made: V/R.

Distributions of estimators Testing one coefficient Testing multiple coefficients Multiple testing

Benjamini-Hochberg, FDR

We define False Discovery Rate (FDR) as:

$$FDR = \mathrm{E}[V/R].$$

(see Benjamini & Hochberg, 1995.)
FDR is controlled at level q, for a pre-chosen q, $0 < q < 1$ if $FDR < q$.
In order to control FDR at level q:

- Order the p-values form the m hypothesis tests from smallest to largest: $p_{(1)}, \dots p_{(m)}$.
- Calculate $\frac{i}{m}q$ for all $i \in \{1, \dots, m\}$.
- Let i_0 be the largest index such that $p_{(i_0)} < \frac{i_0}{m}q$.
- Reject only the null hypothesis corresponding to p-values $p_{(1)}, \dots p_{(i_0)}$.

In practice, $q = .1$ is typically chosen.
In general, FDR control at level q will be less conservative than the Holm correction which controls FWER at level q.

Benjamini-Hochberg, FDR

Exercise: What would be an alternative way to phrase the Benjamini-Hochberg FDR procedure using "adjusted" p-values?