Inference in regression: F-test

ST552 Lecture 8

Charlotte Wickham

2019-01-25

Homework #1

Solutions on canvas

I graded the initial data analysis.

- Everyone was looking at the right things!
- But, the writeups could use some improvement
- HW #3 gets you to repeat this process on a different data set

Homework #3

I've posted an example with some guidelines as `01-initial-data-analysis-report`, but I started from `01-initial-data-analysis-draft`.

Key things I'll be looking for in HW #3:

- $<$ 2 pages (notice my draft is 10 pages, but report is only 1.5 pages)
- you control what output/code is in the final version
- plots are labelled and sized appropriately
- narrative leads the reader through important findings

Today

- The F-test
- Practice with F-tests

Motivation

t-tests on individual parameters only allow us to ask a limited number of questions.

To ask questions about more than one coefficient we need something more complicted.

F-tests do this by comparing nested models. In practice, the hard part is translating a scientific question in a comparison of two models.

F-test

Let Ω denote a larger model of interest with p parameters and ω a smaller model that represents some simplification of Ω with q parameters.

X_Ω $n \times p$

$q < p$

X_ω $n \times q$

Intuition: If both models "fit" as well as each other, we should prefer the simpler model, ω. If Ω shows substantially better fit than ω, that suggests the simplification is not justified.

How do we measure fit? What is substantially better fit?

We require ω to be "nested" within Ω

Mathematical: the colspace$(X_\omega) \subset$ colspace(X_Ω)

In practice (the small model...):

- some β in Ω are set to a value e.g. $\beta_1 = 0$
- put linear constraints on β in Ω $\beta_1 = \beta_2$

F-statistic

RSS = residual sum of squares

$$F = \frac{\overbrace{(\mathrm{RSS}_\omega - \mathrm{RSS}_\Omega)}^{\text{how much better does the full model fit?}}/\overbrace{(p-q)}^{\text{\# extra parameters in full model.}}}{\text{compare to variation due to error} \longrightarrow \boxed{\underbrace{\mathrm{RSS}_\Omega/(n-p)}_{\hat{\sigma}^2}}}$$

Null hypothesis: the simplification to Ω implied by the simpler model, ω.

Under the null hypothesis, the F-statistic has an F-distribution with $\underbrace{\boxed{p-q}}_{\text{numerator d.f.}}$ and $\underbrace{\boxed{n-p}}_{\text{denominator d.f.}}$ degrees of freedom.

Leads to tests of the form: reject H_0 for $F > F^{(\alpha)}_{p-q,n-p}$.

Deriving this fact is beyond this class (take Linear Models).

Example: Overall regression F-test

The overall regression F-test asks if any predictors are related to the response.

Full model: $Y = X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I)$
Reduced model: $Y = \beta_0 + \epsilon$

Null hypothesis: $H_0 : \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$
All the parameters (other than the intercept) are zero.

Alternative hypothesis: At least one parameter is non-zero.

Exercise: question #1 on handout

If there is **evidence against the null hypothesis**:

- The null is not true, or
- the null is true but we got unlucky, or significance level
- the full model isn't true and the F-test is meaningless.

If there is **no evidence against the null hypothesis**:

- The null is true, or
- the null is false but we didn't gather enough evidence to reject it, or
- the full model isn't true and the F-test is meaningless.

→ checking diagnostics

Example: One predictor

Null hypothesis: $\beta_j = 0$

Equivalent to the t-test, reject null if

$$|t_j| = \left| \frac{\hat{\beta}_j}{\mathrm{SE}\left(\hat{\beta}_j\right)} \right| > t_{n-p}^{\alpha/2}$$

In fact, in this case, $F = t_j^2$.

Exercise: questions #2 & #3 on handout

Other examples

- More than one parameter
- A subspace of the parameter space

can not get from t-tests in output of summary.lm( )

Exercise: questions #4 & #5 on handout

do at home

We can't do F-tests when

- we want to test non-linear hypotheses, e.g. $H_0 : \beta_j \beta_k = 1$ (we might be able to make use of the Delta method, though)
- we want to compare non-nested models (find an example on the handout)
- the models fit use different data (most often comes up when a variable of interest has some missing values)