MASM22/FMSN30: Linear and Logistic Regression, 7.5 hp
FMSN40: ... with Data Gathering, 9 hp

Lecture 4, spring 2019

Global and partial F-test

Mathematical Statistics / Centre for Mathematical Sciences
Lund University

3/4-19

Testing the significance of several variables at once

- When we have categorical variables with more than two categories, we replace one variable by two, or more, dummy-variables.
- How to test the significance of the original variable?
- Not t-test. That just tests one of the categories against the reference.
- Solution: Divide the variability in Y into different parts, depending on (groups of) different variables. ANOVA (ANalysis Of VAriance)

ANOVA-tables in regression

We now **decompose the variability** of Y in various ways $\rightarrow$ fundamental to build advanced tools to answer statistical questions.
Idea: the variation in Y_i can be split into two parts: one that is due to the relationship with X ($\mathbf{X}\beta$), and one that is due to the "noise" (ϵ_i).

$$
\begin{aligned}
SS(\text{Total}_{\text{uncorrected}}) &= \sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} (\hat{Y}_i + e_i)^2 \\
&= \underbrace{\sum_{i=1}^{n} \hat{Y}_i^2}_{SS(\text{Model})} + \underbrace{\sum_{i=1}^{n} e_i^2}_{SS(\text{Error})} + \underbrace{2\sum_{i=1}^{n} \hat{Y}_i \cdot e_i}_{=0} \\
&= n\bar{Y}^2 + \underbrace{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}_{SS(\text{Regr})} + \underbrace{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}_{SS(\text{Error})}
\end{aligned}
$$

Actually $SS(Model)$ is **not very interesting**, includes contribution from the constant $n\bar{Y}^2$ whereas we are solely interested in contribution from $X_1, ..., X_p$

$$SS(\text{Total}_{\text{uncorrected}}) = \underbrace{\underbrace{n\bar{Y}^2}_{\text{due to } \beta_0} + \underbrace{SS(\text{Regr})}_{\text{due to } \beta_1 X_1 + \beta_2 X_2 + \ldots}}_{\text{due to } \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots} + \underbrace{SS(\text{Error})}_{\text{due to noise}}$$

$$SS(\text{Total}_{\text{corrected}}) = SS(\text{Total}_{\text{uncorr}}) - n\bar{Y}^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

$$= SS(\text{Regr}) + SS(\text{Error})$$

Thus the sum of squared deviations from $\bar{Y}$ can be split into a part due to the regression and a part due to noise.

SS(Regr) is way more interesting: it gives the contribution to the response variation solely due to the covariates!

Degrees of freedom

Degrees of freedom: "the number of values in the final calculation of a statistic that are free to vary." or "The number of independent ways by which a dynamic system can move, without violating any constraint imposed on it".

Example:

- $\sum_{i=1}^{n} Y_i^2$: a sum of n unconstrained random variables (the Y_i can in principle take any value)
- $SS(Model)$: depends on the $\hat{\beta}_0, ..., \hat{\beta}_p$ these being estimated, without constraints

$$SS(\text{Total}_{\text{uncorr}}) : df = n \quad \text{(observations)}$$
$$SS(\text{Model}) : df = p + 1 \quad \text{(regression params, including } \beta_0\text{)}$$
$$SS(\text{Regr}) : df = p \quad \text{(slope params, excluding } \beta_0\text{)}$$
$$SS(\text{Error}) : df = n - (p + 1)$$

Mean Squares

$$MS(\text{Model}) = \frac{SS(\text{Model})}{p+1} \sim \chi^2_{p+1}$$

$$MS(\text{Regr}) = \frac{SS(\text{Regr})}{p} \sim \chi^2_{p}$$

$$MS(\text{Error}) = \frac{SS(\text{Error})}{n-(p+1)} = s^2 \sim \chi^2_{n-(p+1)}$$

$MS(\text{Error})$ is an unbiased estimate of σ^2, i.e. $E(MS(\text{Error})) = \sigma^2$

Question: does the regression with predictors $X_1, ..., X_p$ contribute significantly towards the total variance in the response?

Global F-test

How to test if the structural part of our model $\boldsymbol{X}\boldsymbol{\beta}$ is "good" at explaining $\boldsymbol{Y}$?

Test $H_0 : \beta_1 = ... = \beta_p = 0$ vs H_1 : at least a $\beta_j \neq 0$ $(j = 1, ..., p)$

Idea

- If H_0: $\beta_1 = \cdots = \beta_p = 0 \rightarrow$ model $\boldsymbol{X}\boldsymbol{\beta}$ is **worthless**
  $SS(Tot_{corr})$ mostly "explained" by the error
  $SS(Error)$ "large" then $\frac{MS(\text{Regr})}{MS(\text{Error})} \lessapprox 1$
- If H_1 plausible a consistent part of the response variation is due to $SS(Regr)$ then $\frac{MS(\text{Regr})}{MS(\text{Error})} \gg 1$

Global F-test

Then it make sense to use the following test, as the distribution of the ratio is known: If H_0 is true then:

$$F = \frac{MS(\text{Regr})}{MS(\text{Error})} \sim F(p,\, n-(p+1)) \qquad \text{(one-sided test)}$$

and we can reject H_0, in favour of H_1: "at least one of $\beta_1, \ldots, \beta_p$ is $\neq 0$", at significance level α if $F > F_\alpha(p,\, n-(p+1))$.

Difference with t-test

- t-test is for the significance of a single covariate, given all others in the model;
- F-test test globally the general aptness of the model (it does not say *which* β_j's are significantly $\neq 0$)
- a global F always is "one-sided" because it is not interesting find that we can explain even *less* than the noice does.

Example: cabbage weight and soil type

$$E(\mathrm{Y}) = \beta_0 + \beta_1 X_{\text{clay}} + \beta_2 X_{\text{loam}}$$

In R: run `anova()` for your model:

$$F = \frac{MS(\text{Regr})}{MS(\text{Error})} = \frac{SS(\text{Regr})/p}{SS(\text{Error})/(n-(p+1))} = \frac{10.022/2}{20.533/(30-(2+1))}$$
$$= \frac{5.011}{0.7605} = 6.5892 > F_{0.05}(2,27) = 3.35$$
$$p - \text{value} = 0.004672 < 0.05$$

Reject H_0 at $\alpha = 0.05$. Our data suggest that the model does contain at least one relevant covariate (it doesn't say which one!).

In R: run the `summary()` for your model
Look for the line: `F-statistic: 6.589 on 2 and 27 DF, p-value: 0.004672`

ANOVA table

A handy way to represent the introduced concepts is the Analysis of Variance table:

variation	SS	df	MS	F
Regression	SS(Regr)	p	MS(Regr)	MS(Regr)/MS(Err)
Error	SS(Err)	$n-(p+1)$	MS(Err)	
Total	SS(Tot_{corr})	$n-1$		

Partial F-test: testing a subset of parameter

Model: $Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_p X_{pi} + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$.
Question: are k specific β-parameters $= 0$
(e.g. the k last: $\beta_{p-k+1} = \ldots = \beta_p = 0$).

Procedure:

- Estimate the full model with all $p+1$ parameters:
  $SS(\text{Error}_{\text{full}})$ with $df = n - (p+1)$.
- Estimate the reduced model with $p+1-k$ parameters:
  $SS(\text{Error}_{\text{reduced}})$ with $df = n - (p+1-k)$.
- Calculate the increase in $SS(\text{Error})$:
  $Q = SS(\text{Error}_{\text{reduced}}) - SS(\text{Error}_{\text{full}})$ with
  $df = n - (p+1-k) - (n - (p+1)) = k$.
- Is this increase too large? Reject H_0 at $\alpha = 5\%$ if

$$F = \frac{Q/k}{s^2_{\text{full}}} > F_{1-\alpha}(k, n-(p+1))$$

Example: do we need the interactions terms?

Compare the full model:

$$E(Y) = \beta_0 + \beta_1 X_{\text{clay}} + \beta_2 X_{\text{loam}} + \beta_3 X_3 + \beta_4 X_{\text{clay}} X_3 + \beta_5 X_{\text{loam}} X_3$$

aginst the reduced model

$$E(Y) = \beta_0 + \beta_1 X_{\text{clay}} + \beta_2 X_{\text{loam}} + \beta_3 X_3.$$

Test H_0: $\beta_4 = \beta_5 = 0$ against H_1: "$\beta_4 \neq 0$ and/or $\beta_5 \neq 0$"

$$\begin{aligned} F &= \frac{Q/k}{s^2_{\text{full}}} = \frac{(18.405 - 17.143)/2}{0.8451779^2} = \frac{1.2611/2}{0.7143} \\ &= 0.8828 \not> F_{0.05}(2, 24) = 3.40 \\ p - \text{value} &= 0.4266 \not< 0.05 \end{aligned}$$

No. We do not need the interaction terms.

Partial F-test

Important!

- Partial F-test for comparison between a large and reduced model only makes sense if models are **nested**.
- Data used must be the same for both models.
- Partial F is **sensitive to ordering** unless tested covariates are uncorrelated (zero covariances).

The R `anova(model)` function builds the ANOVA table, by decomposing the variability explained by the several regression terms. So SS(Regr) is splitted in different lines: sum them to obtain SS(Regr).
This is a *sequential* construction of the ANOVA table. As such, ordering matters.