TMA4267 Linear Statistical Models V2017 (L14)

Part 3: Hypothesis testing and analysis of variance
The universal F-test [F:3.3]
One-way ANOVA [H:8.1.1]

Mette Langaas

Department of Mathematical Sciences, NTNU

To be lectured: March 7, 2017

Today

- Linear hypotheses in regression vs. nested models.
- The universal F-test for linear hypotheses: two formulas.
- The two formulas: one easy to use, one easy for proving F-distribution.
- Special cases of the universal F-test.
- New problem: categorical covariate with effect coding (for interpretation)

Happiness ($n = 39$)

Are love and work the important factors determining happiness?

- y, `happiness`. 10-point scale, with 1 representing a suicidal state, 5 representing a feeling of «just muddling along», and 10 representing a euphoric state.
- x_1, `money`. Annual family income in thousands of dollars.
- x_2, `sex`. Sex was measured as the values 0 or 1, with 1 indicating a satisfactory level of sexual activity.
- x_3, `love`. 3-point scale, with 1 representing loneliness and isolation, 2 representing a set of secure relationships, and 3 representing a deep feeling of belonging and caring in the context of some family or community.
- x_4, `work`. 5-point scale, with 1 indicating that an individual is seeking other employment, 3 indicating the job is OK, and 5 indicating that the job is enjoyable.

Data taken from library faraway, data set happy.

3.13 Testing Linear Hypotheses

Hypotheses

1. General linear hypothesis:

$$H_0 : \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{d} \qquad \text{against} \qquad H_0 : \boldsymbol{C}\boldsymbol{\beta} \neq \boldsymbol{d}$$

   where $\boldsymbol{C}$ is a $r \times p$-matrix with $\text{rk}(\boldsymbol{C}) = r \leq p$ (r linear independent restrictions).
2. Test of significance (t-test):

$$H_0 : \beta_j = 0 \qquad \text{against} \qquad H_1 : \beta_j \neq 0$$

3. Composite test of a subvector:

$$H_0 : \boldsymbol{\beta}_1 = \boldsymbol{0} \qquad \text{against} \qquad H_1 : \boldsymbol{\beta}_1 \neq \boldsymbol{0}$$

4. Test for significance of regression:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \text{ against}$$
$$H_1 : \beta_j \neq 0 \text{ for at least one } j \in \{1, \ldots, k\}$$

Box from our text book: Fahrmeir et al (2013): Regression. Springer. (p.135)

Constrained and unconstrained estimate

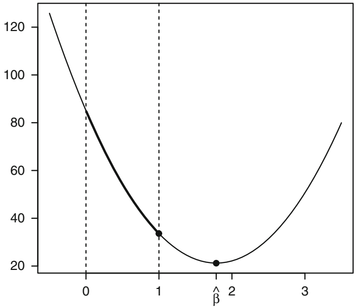

Fig. 3.15 Illustration of the difference in goodness of fit between the unconstrained least squares estimator and the estimator under the constraint $0 \leq \beta \leq 1$. The (unconstrained) least squares estimator is labeled as $\hat{\beta}$. For the constrained solution, we have $\hat{\beta} = 1$

Figure 3.15 from our text book: Fahrmeir et al (2013): Regression. Springer. (p.1329)

3.13 Testing Linear Hypotheses

Hypotheses

1. General linear hypothesis:

$$H_0 : \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{d} \qquad \text{against} \qquad H_0 : \boldsymbol{C}\boldsymbol{\beta} \neq \boldsymbol{d}$$

   where $\boldsymbol{C}$ is a $r \times p$-matrix with $\text{rk}(\boldsymbol{C}) = r \leq p$ (r linear independent restrictions).
2. Test of significance (t-test):

$$H_0 : \beta_j = 0 \qquad \text{against} \qquad H_1 : \beta_j \neq 0$$

3. Composite test of a subvector:

$$H_0 : \boldsymbol{\beta}_1 = \boldsymbol{0} \qquad \text{against} \qquad H_1 : \boldsymbol{\beta}_1 \neq \boldsymbol{0}$$

4. Test for significance of regression:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \text{ against}$$
$$H_1 : \beta_j \neq 0 \text{ for at least one } j \in \{1, \ldots, k\}$$

Box from our text book: Fahrmeir et al (2013): Regression. Springer. (p.135)

Test Statistics

Assuming normal errors we obtain under H_0:

1. $F = 1/r\,(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d})' \left(\hat{\sigma}^2 \boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{C}'\right)^{-1} (\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{d}) \sim \mathrm{F}_{r,n-p}$
2. $t_j = \dfrac{\hat{\beta}_j}{\mathrm{se}_j} \sim \mathrm{t}_{n-p}$
3. $F = \frac{1}{r}(\hat{\boldsymbol{\beta}}_1)'\widehat{\mathrm{Cov}(\hat{\boldsymbol{\beta}}_1)}^{-1}(\hat{\boldsymbol{\beta}}_1) \sim \mathrm{F}_{r,n-p}$
4. $F = \dfrac{n-p}{k}\dfrac{R^2}{1-R^2} \sim \mathrm{F}_{k,n-p}$

Critical Values

Reject H_0 in the case of:

1. $F > F_{r,n-p}(1-\alpha)$
2. $|t| > t_{n-p}(1-\alpha/2)$
3. $F > F_{r,n-p}(1-\alpha)$
4. $F > F_{k,n-p}(1-\alpha)$

The tests are relatively robust against moderate departures from normality. In addition, the tests can be applied for large sample size, even with non-normal errors.

Box from our text book: Fahrmeir et al (2013): Regression. Springer. (p.135)

3.14 Confidence Regions and Prediction Intervals

Provided that we have (at least approximately) normally distributed errors or a large sample size, we obtain the following confidence intervals or regions and prediction intervals:

Confidence Interval for β_j

A confidence interval for β_j with level $1-\alpha$ is given by

$$[\hat{\beta}_j - t_{n-p}(1-\alpha/2)\cdot \mathrm{se}_j, \hat{\beta}_j + t_{n-p}(1-\alpha/2)\cdot \mathrm{se}_j].$$

Confidence Ellipsoid for Subvector $\boldsymbol{\beta}_1$

A confidence ellipsoid for $\boldsymbol{\beta}_1 = (\beta_1, \ldots, \beta_r)'$ with level $1-\alpha$ is given by

$$\left\{ \boldsymbol{\beta}_1 : \frac{1}{r}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1)' \widehat{\mathrm{Cov}(\hat{\boldsymbol{\beta}}_1)}^{-1} (\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) \leq F_{r,n-p}(1-\alpha) \right\}.$$

Confidence Interval for μ_0

A confidence interval for $\mu_0 = \mathrm{E}(y_0)$ of a future observation y_0 at location $\boldsymbol{x}_0$ with level $1-\alpha$ is given by

$$\boldsymbol{x}_0'\hat{\boldsymbol{\beta}} \pm t_{n-p}(1-\alpha/2)\hat{\sigma}(\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0)^{1/2}.$$

Prediction Interval

A prediction interval for a future observation y_0 at location $\boldsymbol{x}_0$ with level $1-\alpha$ is given by

$$\boldsymbol{x}_0'\hat{\boldsymbol{\beta}} \pm t_{n-p}(1-\alpha/2)\hat{\sigma}(1 + \boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0)^{1/2}.$$

Box from our text book: Fahrmeir et al (2013): Regression. Springer. (p.137)

Concrete aggregates data

Aggregate:	1	2	3	4	5	
	551	595	639	417	563	
	457	580	615	449	631	
	450	508	511	517	522	
	731	583	573	438	613	
	499	633	648	415	656	
	632	517	677	555	679	
Total	3320	3416	3663	2791	3664	16,854
Mean	553.33	569.33	610.50	465.17	610.67	561.80

Table 13.1 of Walepole, Myers, Myers, Ye: Statistics for Engineers and Scientists – our textbook from the introductory TMA4240/TMA4245 Statistics course.

Today

- Linear hypotheses in regression vs. nested models.
- The universal F-test for linear hypotheses: two formulas.
- The two formulas: one easy to use, one easy for proving F-distribution.
- Special cases of the universal F-test.
- Next time: categorical covariate with effect coding (for interpretation)