Tests

Specification testing

Michel Bierlaire

Mathematical Modeling of Behavior

EPFL

Outline

Difference with classical hypothesis testing

Informal tests and t tests

Likelihood ratio test

Non nested hypotheses

Prediction tests

Motivation

Modeling

- Impossible to determine the most appropriate model specification.
- A good fit does not mean a good model.
- Formal testing is necessary, but not sufficient.
- No clear-cut rules can be given.
- Subjective judgments of the analyst.
- Good modeling = good judgment + good analysis.

Wilkinson (1999) "The grammar of graphics". Springer

... some researchers who use statistical methods pay more attention to goodness of fit than to the meaning of the model... Statisticians must think about what the models mean, regardless of fit, or they will promulgate nonsense.

Classical hypothesis testing

Null hypothesis (H_0)

A simple hypothesis contradicting a theoretical assumption.

Analogy: court trial

- Theoretical assumption: an individual has committed a felony.
- Null hypothesis: she is innocent.
- Main principle: the defendant is presumed innocent until proved guilty.
- Similarly, H_0 is considered correct, until the data provide sufficient evidences that it is not.

Classical hypothesis testing: example

Lady testing tea

- Theory: a lady is able to tell if the milk has been poured before of after the tea in a cup.
- H_0: the outcome of the taste is purely random.

[Fisher, 1956]

Classical hypothesis testing: errors

	Accept H_0	Reject H_0
H_0 is true		Type I error (prob. α)
H_0 is false	Type II error (prob. β)	

Specification testing: example

Null hypothesis (H_0)

A simple hypothesis contradicting a theoretical assumption.

Explanatory variable

- Theory: a variable explains the choice behavior.
- H_0: the coefficient of the variable is zero.

Errors in hypothesis testing

Type I error

- H_0 rejected and H_0 true.
- Include an irrelevant variable.
- Loss of efficiency.
- Cost: C_I.

Type II error

- H_0 accepted and H_0 false.
- Omit a relevant variable.
- Specification error.
- Cost: $C_{II} \gg C_I$.

Note

In classical hypothesis testing, $C_I \approx C_{II}$

Impact of an error

Probability of an error

$$\begin{aligned}
\text{P(Type I)} = \quad & \underbrace{\text{P}(H_0 \text{ rejected} | H_0 \text{ true})}_{\alpha} \quad \underbrace{\text{P}(H_0 \text{ true})}_{\lambda} \\
\text{P(Type II)} = \quad & \underbrace{\text{P}(H_0 \text{ accepted} | H_0 \text{ false})}_{\beta} \quad \underbrace{\text{P}(H_0 \text{ false})}_{(1-\lambda)}
\end{aligned}$$

Expected cost

$$\begin{aligned}
\text{Expected cost} &= \text{P(Type I)} \quad C_I \quad + \quad \text{P(Type II)} \quad C_{II} \\
&= \quad \alpha\lambda \quad\quad C_I \quad + \quad \beta(1-\lambda) \quad\quad C_{II}
\end{aligned}$$

Classical hypothesis testing

$\lambda \approx 1$, $C_I \approx C_{II}$: prefer small α.

Impact of an error

Probability of an error

$$\begin{aligned}
\text{P(Type I)} = &\quad \underbrace{\text{P}(H_0 \text{ rejected} | H_0 \text{ true})}_{\alpha} \quad \underbrace{\text{P}(H_0 \text{ true})}_{\lambda} \\
\text{P(Type II)} = &\quad \underbrace{\text{P}(H_0 \text{ accepted} | H_0 \text{ false})}_{\beta} \quad \underbrace{\text{P}(H_0 \text{ false})}_{(1-\lambda)}
\end{aligned}$$

Expected cost

$$\begin{aligned}
\text{Expected cost} &= \text{P(Type I)} \quad C_I \quad + \quad \text{P(Type II)} \quad C_{II} \\
&= \quad \alpha\lambda \quad\quad C_I \quad + \quad \beta(1-\lambda) \quad\quad C_{II}
\end{aligned}$$

Specification testing

$\lambda \approx 0.5$, $C_{II} \gg C_I$: larger α can be used.

Outline

Difference with classical hypothesis testing

Informal tests and t tests

Likelihood ratio test

Non nested hypotheses

Prediction tests

Informal tests

Objective

Identify early inconsistencies between the model and a priori expectations.

Examples

- Sign of the coefficients (e.g. cost, travel time).
- Coefficients in monetary units (e.g. value of time).

t-test

Question

Is the parameter θ equal to a given value θ^*?

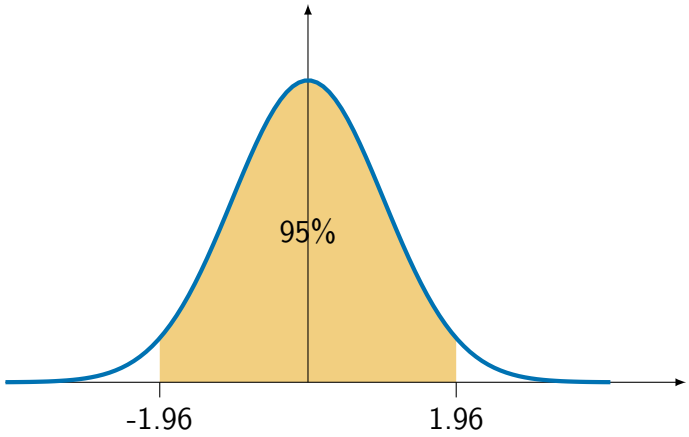
- $H_0 : \theta = \theta^*$.
- $H_1 : \theta \neq \theta^*$.

Statistic (assuming maximum likelihood estimator)

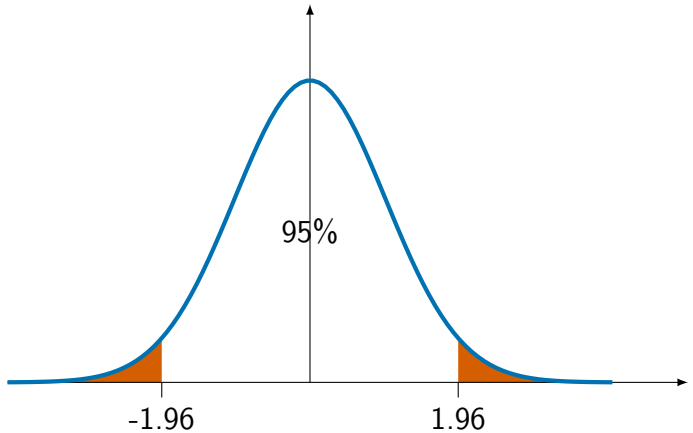
Under H_0, if $\hat{\theta}$ is normally distributed with known variance σ^2:

$$\frac{\hat{\theta} - \theta^*}{\sigma} \sim N(0, 1).$$

t-test: under H_0



t-test: if the statistic lies outside



H_0 is rejected at the 5% level.

Applying the test

Statistic

$$P(-1.96 \leq \frac{\hat{\theta} - \theta^*}{\sigma} \leq 1.96) = 0.95 = 1 - 0.05.$$

Decision

H_0 can be rejected at the 5% level ($\alpha = 0.05$) if

$$\left| \frac{\hat{\theta} - \theta^*}{\sigma} \right| \geq 1.96.$$

Comments

- If $\hat{\theta}$ is asymptotically normal,
- if variance is unknown,
- a t test should be used with N degrees of freedom.
- When $N \geq 30$, the Student t distribution is well approximated by a $N(0, 1)$.

p value

- Probability to get a t statistic at least as large (in absolute value) as the one reported, under the null hypothesis.
- It is calculated as
$$p = 2(1 - \Phi(t)),$$
where $\Phi(\cdot)$ is the CDF of the standard normal.
- The null hypothesis is rejected when the p-value is lower than the significance level α.

Comparing two coefficients

Hypothesis

$$H_0 : \beta_1 = \beta_2.$$

Statistic

$$\frac{\widehat{\beta}_1 - \widehat{\beta}_2}{\sqrt{\mathrm{Var}(\widehat{\beta}_1 - \widehat{\beta}_2)}},$$

where

$$\mathrm{Var}(\widehat{\beta}_1 - \widehat{\beta}_2) = \mathrm{Var}(\widehat{\beta}_1) + \mathrm{Var}(\widehat{\beta}_2) - 2\,\mathrm{Cov}(\widehat{\beta}_1, \widehat{\beta}_2).$$

Distribution

Under H_0, distributed as $N(0, 1)$.

Outline

Difference with classical hypothesis testing

Informal tests and t tests

Likelihood ratio test

Non nested hypotheses

Prediction tests

Likelihood ratio test

Objective

Investigate parsimonious versions of a given specification, by introducing linear restrictions on the parameters.

Null hypothesis

The parsimonious, or restricted, model is the true model

Likelihood ratio test

Test

Under H_0,

$$-2(\mathcal{L}(\hat{\beta}_R) - \mathcal{L}(\hat{\beta}_U)) \sim \chi^2_{(K_U - K_R)},$$

where

- $\mathcal{L}(\hat{\beta}_R)$ is the log likelihood of the restricted model,
- $\mathcal{L}(\hat{\beta}_U)$ is the log likelihood of the unrestricted model,
- K_R is the number of parameters in the restricted model, and,
- K_U is the number of parameters in the unrestricted model.

Benchmarking

Unrestricted model

$$V_{in} = \beta_1 x_{ink} + \cdots$$
$$V_{jn} = \beta_2 x_{jnk} + \cdots$$
$$\vdots$$

Restricted model

Equal probability model

$$V_{in} = 0$$
$$V_{jn} = 0$$
$$\vdots$$

Restrictions

$$\beta_k = 0, \ \forall k$$

Benchmarking

Log likelihood of the unrestricted model

$$\mathcal{L}(\widehat{\beta})$$

Log likelihood of the restricted model

$$P_{in} = 1/J_n, \ \forall i \in \mathcal{C}_n, \forall n$$

$$\mathcal{L}(0) = -\sum_{n=1}^{N} \log(J_n)$$

Statistic

$$-2(\mathcal{L}(0) - \mathcal{L}(\widehat{\beta})) \sim \chi^2_K$$

Benchmarking revisited

Unrestricted model

$$V_{in} = \beta_1 x_{ink} + \cdots$$
$$V_{jn} = \beta_2 x_{jnk} + \cdots$$
$$\vdots$$

Restricted model

Only alternative specific constants

$$V_{in} = \beta_i,$$
$$V_{jn} = \beta_j,$$
$$\vdots$$

Restrictions

All coefficients but the constants are constrained to zero.

Benchmarking revisited

Log likelihood of the unrestricted model

$$\mathcal{L}(\widehat{\beta})$$

Log likelihood of the restricted model

$$P_{in} = N_i/N \; \forall i \in \mathcal{C}, \forall n.$$

$$\mathcal{L}(c) = \sum_{i=1}^{J} N_i \log(N_i/N).$$

Statistic

$$-2(\mathcal{L}(c) - \mathcal{L}(\widehat{\beta})) \sim \chi^2_d \text{ with } d = K - J + 1.$$

Benchmarking

Classical output of estimation software

Summary statistics

Number of observations = 2544

$$\begin{aligned} \mathcal{L}(0) &= -2794.870 \\ \mathcal{L}(c) &= -2203.160 \\ \mathcal{L}(\hat{\beta}) &= -1640.525 \\ -2[\mathcal{L}(0) - \mathcal{L}(\hat{\beta})] &= 2308.689 \end{aligned}$$

Test of generic attributes

Unrestricted model

Alternative specific

$$V_{in} = \beta_{1i} x_{ink} + \cdots$$
$$V_{jn} = \beta_{1j} x_{jnk} + \cdots$$
$$\vdots$$

Restricted model

Generic

$$V_{in} = \beta_{1} x_{ink} + \cdots$$
$$V_{jn} = \beta_{1} x_{jnk} + \cdots$$
$$\vdots$$

Restriction

$$\beta_{1i} = \beta_{1j} = \cdots$$

Test of generic attributes

Log likelihood of the unrestricted model

$$\mathcal{L}(\widehat{\beta}_{AS})$$

Log likelihood of the restricted model

$$\mathcal{L}(\widehat{\beta}_{G})$$

Statistic

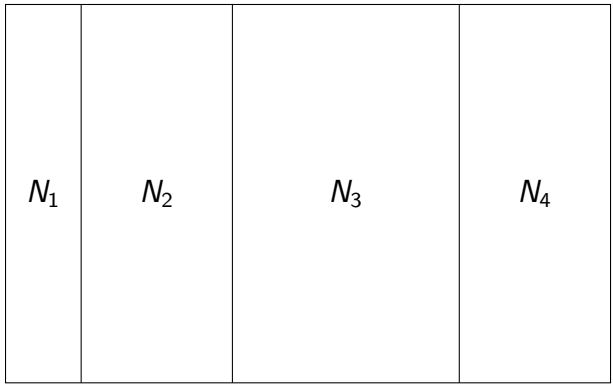
$$-2(\mathcal{L}(\widehat{\beta}_{G}) - \mathcal{L}(\widehat{\beta}_{AS})) \sim \chi^2_d \text{ with } d = K_{AS} - K_G.$$

Test of taste variations

Segmentation

- Classify the data into G groups. Size of group g: N_g.
- The same specification is considered for each group.
- A different set of parameters is estimated for each group.

Test of taste variations



$\mathcal{L}_{N_1}(\widehat{\beta}^1)$ $\mathcal{L}_{N_2}(\widehat{\beta}^2)$ $\mathcal{L}_{N_3}(\widehat{\beta}^3)$ $\mathcal{L}_{N_4}(\widehat{\beta}^4)$ $\sum_{g=1}^{G} \mathcal{L}_{N_g}(\widehat{\beta}^g)$

Test of taste variations

Unrestricted model

Group specific coefficients

$$V_{in} = \sum_{g=1}^{G} (\delta_{ng}\beta_{1g}) x_{ink} + \cdots$$

$$V_{jn} = \sum_{g=1}^{G} (\delta_{ng}\beta_{2g}) x_{jnk} + \cdots$$

$$\vdots$$

Restricted model

Generic coefficients

$$V_{in} = \beta_1 x_{ink} + \cdots$$

$$V_{jn} = \beta_2 x_{jnk} + \cdots$$

$$\vdots$$

Restrictions

$$\beta_{k1} = \beta_{k2} = \cdots = \beta_{kG}, \ \forall k.$$

Test of taste variations

Log likelihood of the unrestricted model

$$\sum_{g=1}^{G} \mathcal{L}_{N_g}(\widehat{\beta}^g)$$

Log likelihood of the restricted model

$$\mathcal{L}_N(\widehat{\beta})$$

Statistic

$$-2\left[\mathcal{L}_N(\widehat{\beta}) - \sum_{g=1}^{G} \mathcal{L}_{N_g}(\widehat{\beta}^g)\right] \sim \chi^2_d \text{ with } d = \sum_{g=1}^{G} K - K = (G-1)K.$$

Tests of nonlinear specifications

Unrestricted model

Power series

$$V_{in} = \sum_{\ell=1}^{L} \beta_{1\ell} \frac{x_{ink}{}^{\ell}}{x_{\text{ref}}} + \cdots$$
$$V_{jn} = \beta_2 x_{jnk} + \cdots$$
$$\vdots$$

Restricted model

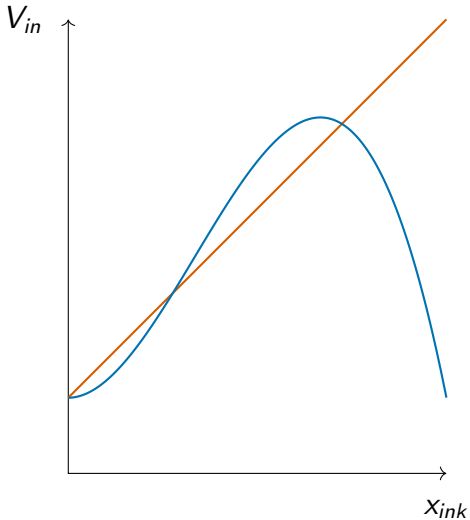
Linear specification

$$V_{in} = \beta_1 x_{ink} + \cdots$$
$$V_{jn} = \beta_2 x_{jnk} + \cdots$$
$$\vdots$$

Restrictions

$$\beta_{12} = \beta_{13} = \cdots = \beta_{1L} = 0.$$

Power series



Test of nonlinear specifications

Log likelihood of the unrestricted model

$$\mathcal{L}(\widehat{\beta}_U)$$

Log likelihood of the restricted model

$$\mathcal{L}(\widehat{\beta}_R)$$

Statistic

$$-2\left[\mathcal{L}(\widehat{\beta}_R) - \mathcal{L}(\widehat{\beta}_U)\right] \sim \chi^2_d \text{ with } d = L - 1.$$

Notes

- Usually not behaviorally meaningful
- Danger of overfitting
- Polynomials are most of the time inappropriate for extrapolation due to oscillation.
- Other nonlinear specifications can be used for testing:
  - piecewise linear,
  - Box-Cox.

Outline

Difference with classical hypothesis testing

Informal tests and t tests

Likelihood ratio test

Non nested hypotheses

Prediction tests

Non nested hypotheses

Nested hypotheses

- Restricted and unrestricted models.
- Linear restrictions.
- H_0: restricted model is correct.
- Test: likelihood ratio test.

Non nested hypotheses

- Need to compare two models.
- None of them is a restriction of the other.
- Likelihood ratio test cannot be used.

Example

Model 1

$$V_{in} = \beta_1 x_{ink} + \cdots$$
$$V_{jn} = \beta_2 x_{jnk} + \cdots$$
$$\vdots$$

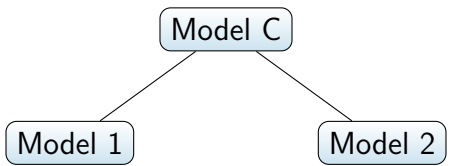
Model 2

$$V_{in} = \beta_1 \log(x_{ink}) + \cdots$$
$$V_{jn} = \beta_2 \log(x_{jnk}) + \cdots$$
$$\vdots$$

Cox test

Back to nested hypotheses

- We want to test model 1 against model 2.
- We generate a composite model C such that both models 1 and 2 are restricted cases of model C.



Example

Model 1

$$
\begin{aligned}
V_{in} &= \beta_1 x_{ink} + \cdots \\
V_{jn} &= \beta_2 x_{jnk} + \cdots \\
&\vdots
\end{aligned}
$$

Model 2

$$
\begin{aligned}
V_{in} &= \beta_1 \log(x_{ink}) + \cdots \\
V_{jn} &= \beta_2 \log(x_{jnk}) + \cdots \\
&\vdots
\end{aligned}
$$

Model C

$$
\begin{aligned}
V_{in} &= \beta_{11} x_{ink} + \beta_{12} \log(x_{ink}) + \cdots \\
V_{jn} &= \beta_{21} x_{jnk} + \beta_{22} \log(x_{jnk}) + \cdots \\
&\vdots
\end{aligned}
$$

Cox test

Testing

- We test 1 against C using the likelihood ratio test.
- We test 2 against C using the likelihood ratio test.

Conclusions

C against 1	C against 2	Conclusion
1 is not rejected	2 is rejected	Prefer 1
1 is rejected	2 is not rejected	Prefer 2
1 is rejected	2 is rejected	Develop better models
1 is not rejected	2 is not rejected	Use another test

Davidson and McKinnon J test

Motivation

Cox test may require to estimate a model with a potentially very large number of parameters.

Consider two specifications

$$M_1 : U_{in} = V_{in}^{(1)}(x_{in}; \beta) + \varepsilon_{in}^{(1)},$$
$$M_2 : U_{in} = V_{in}^{(2)}(x_{in}; \gamma) + \varepsilon_{in}^{(2)}.$$

Null hypothesis

M_1 is correct.

Davidson and McKinnon J test

First step

Estimate the parameters γ of M_2.

Composite specification

$$M_C : U_{in} = (1-\alpha)V_{in}^{(1)}(x_{in};\beta) + \alpha V_{in}^{(2)}(x_{in};\widehat{\gamma}) + \varepsilon_{in},$$

where $\widehat{\gamma}$ are the estimated parameters of M_2.

Estimation

Estimate β and α.

Test

Under H_0, the true value of α is 0. A t-test can be used.

Adjusted likelihood ratio index

Likelihood ratio index

$$\rho^2 = 1 - \frac{\mathcal{L}(\widehat{\beta})}{\mathcal{L}(0)}, \tag{1}$$

Adjusted likelihood ratio index

$$\bar{\rho}^2 = 1 - \frac{\mathcal{L}(\widehat{\beta}) - K}{\mathcal{L}(0)}, \tag{2}$$

where K is the number of unknown parameters in the model.

Model selection

Select the model with the highest value of $\bar{\rho}^2$.

Other criteria

Akaike Information Criterion

$$2K - 2\mathcal{L}(\widehat{\beta}).$$

The smallest, the better.

Bayesian Information Criterion

$$K \ln(N) - 2\mathcal{L}(\widehat{\beta}).$$

The smallest, the better.

Outline

Difference with classical hypothesis testing

Informal tests and t tests

Likelihood ratio test

Non nested hypotheses

Prediction tests

Prediction tests

Motivation

Check if the model is able to predict.

Outlier analysis

Procedure

- Apply the model on the sample.
- Examine observations where the predicted probability is the smallest for the observed choice.
- Test model sensitivity to outliers, as a small probability has a significant impact on the log likelihood.
- Potential causes of low probability:
  - coding or measurement error in the data,
  - model misspecification,
  - Inexplicable variation in choice behavior.

Coding or measurement error in the data

Look for signs of data errors

- Travel time is negative.
- Number is coded as a string.
- etc.

Correct or remove the observation

- Go back to the original survey.
- Correct only if you are certain.

Model misspecification

Improve the specification

- Seek clues of missing variables from the observation.
- Why is the model associating such a low probability for this choice?
- Did we forget to account for age, income, or any other variable?
- Should a nonlinear specification be investigated?
- Use behavioral intuition.

Inexplicable variation in choice behavior

Keep the observation

- If no acceptable explanation is found, keep the observation.
- Avoid overfitting of the model to the data.
- The model should reflect how people behave, not how they should behave.

Cross-validation

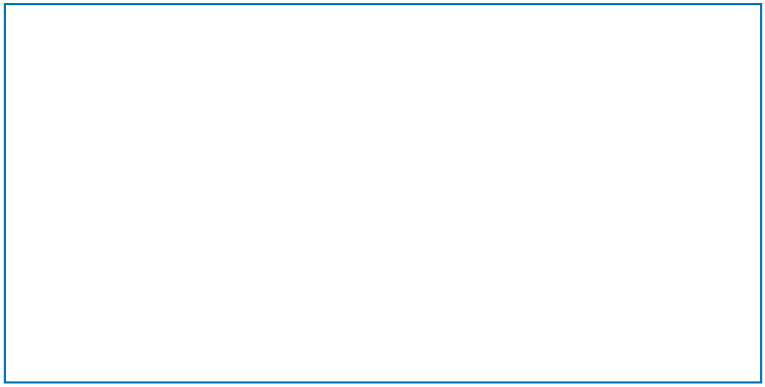
Motivation

- Purpose of the model: prediction.
- Is the model able to predict?

Cross-validation

Motivation

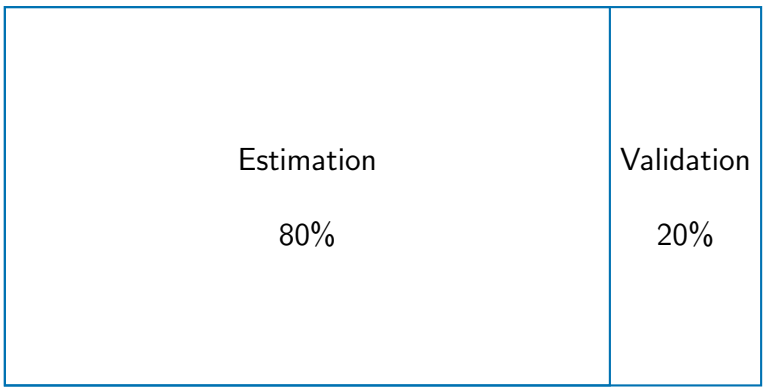
- Purpose of the model: prediction.
- Is the model able to predict?



Cross-validation

Motivation

- Purpose of the model: prediction.
- Is the model able to predict?



Methodology

Split the sample

- Decide the size of the validation set (e.g. 20%)
- Draw randomly an estimation set and a validation set.
- Repeat R times.

Evaluate

- For each pair of estimation/validation set...
- Estimate the parameters of the model with the estimation set.
- Calculate a measure of fit of the estimated model on the validation set.
- Typically, the log-likelihood $\sum_{n=1}^{N_V} \log P(i_n|x_n)$, or the expected number of correctly predicted observations $\sum_{n=1}^{N_V} P(i_n|x_n)$.
- Calculate the average measure of fit.
- Select the model with the highest average fit on the validation sets.

Practical recommendations

- Tests are designed to check meaningful hypotheses.
- Do not test hypotheses that do not make sense.
- Do not apply the tests blindly.
- Always use your judgment.

Summary

- Specification testing different from classical hypothesis testing.
- Informal tests.
- t-tests.
- Likelihood ratio test.
- Non nested hypotheses.
- Prediction tests.

90%, 95% and 99% of the χ^2 distribution with K degrees of freedom

K	90%	95%	99%	K	90%	95%	99%
1	2.706	3.841	6.635	21	29.615	32.671	38.932
2	4.605	5.991	9.210	22	30.813	33.924	40.289
3	6.251	7.815	11.345	23	32.007	35.172	41.638
4	7.779	9.488	13.277	24	33.196	36.415	42.980
5	9.236	11.070	15.086	25	34.382	37.652	44.314
6	10.645	12.592	16.812	26	35.563	38.885	45.642
7	12.017	14.067	18.475	27	36.741	40.113	46.963
8	13.362	15.507	20.090	28	37.916	41.337	48.278
9	14.684	16.919	21.666	29	39.087	42.557	49.588
10	15.987	18.307	23.209	30	40.256	43.773	50.892
11	17.275	19.675	24.725	31	41.422	44.985	52.191
12	18.549	21.026	26.217	32	42.585	46.194	53.486
13	19.812	22.362	27.688	33	43.745	47.400	54.776
14	21.064	23.685	29.141	34	44.903	48.602	56.061
15	22.307	24.996	30.578	35	46.059	49.802	57.342
16	23.542	26.296	32.000	36	47.212	50.998	58.619
17	24.769	27.587	33.409	37	48.363	52.192	59.893
18	25.989	28.869	34.805	38	49.513	53.384	61.162
19	27.204	30.144	36.191	39	50.660	54.572	62.428
20	28.412	31.410	37.566	40	51.805	55.758	63.691

Bibliography I

Fisher, R. A. (1956).
Mathematics of a lady tasting tea.
In Newman, J. R., editor, The world of Mathematics, volume 3, pages 1512–1521, New-York. Simon qnd Schuster.