EFR summary

Applied Statistics 2, FEB12005X

2022-2023

Lectures 1 to 7

Weeks 1 to 7

Deloitte.

DeNederlandscheBank
EUROSYSTEEM

financials
LAUDAME

Details

Subject: Applied Statistics 2 IBEB 2022-2023

Teacher: dr. C. Cavicchia

Date of publication: 07.10.2022

Applied statistics 2 – IBEB – Lecture 1, week 1

Introduction

The purpose of Statistics is represented by four main activities:

- Asking a question about a population
- Observing data from a sample smaller than the population (gathering evidence)
- Making a decision rule
- Drawing conclusions regarding the population based on the information provided by the data from the observed sample

Observation: While analysing the data, you need to keep in mind what the population is and whether the selected sample is representative for it.

Hypothesis testing

In order to do the testing, you need to respect the following four steps:

1. **Question** (formulation of the hypotheses)
2. **Evidence/data** (calculation of the test statistic)
3. **Decision rule** (Implementation of a decision rule that reflects when you can reject the formulated hypothesis)
4. **Conclusion** (accept/reject hypothesis)

One-sample z test

A one-sample z-test is used for a population with unknown mean (μ) and known standard deviation (σ).

Distribution of sample mean

We can say the sample mean is (approximately) normally distributed with mean μ and standard deviation σ if n is sufficiently large.

- The distribution of the sample mean X is approximately $N(\mu, \frac{\sigma}{\sqrt{n}})$
- Standardised sample mean $Z \sim N(0,1) = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$

If the population distribution is normal, then any n is sufficient.
The more non-normal the population is, the larger n is needed.

Significance testing

A **decision rule** is a procedure, which we use in order to decide whether we accept or reject the null hypothesis. For example, when H0 is true, there is still a small probability of error that results in rejecting H0. The probability for such an error is called the significance level. It should be small.

The **significance level** (denoted also as alpha (α)) is the probability of rejecting the null hypothesis when it is true. For example, a significance level of 0.05 indicates a 5% risk of concluding that a difference exists when there is no actual difference.

Ways of performing hypothesis testing

testing based on critical value

In correspondence with the significance level we have z_α * and $-z_\alpha$ *, which represent the **critical values** on the test distribution that are compared to the test statistic to determine whether or not the null hypothesis is rejected. H0 is rejected if the test statistic is in the rejection region on the test distribution.

testing based on P value

P-value: If H0 is true, the probability that the test statistic would be as extreme or more extreme than the observed value.("Extreme" means deviating from H0 in favor of Ha.)
The smaller the P-value, the stronger the evidence against H0. If P-value < significance level α, then H0 is rejected.
Note: $\alpha = 0$ should not be used. $\alpha = P(rejecting\ H_0\ when\ H_0 is\ true)$ resulting in critical values $z_{\frac{\alpha}{2}}$ *= ∞, hence H0 will never be rejected (even if it is wrong).

Two types of error

		Truth about population	
		H_0 is true	H_0 is not true
Decision based on sample	Reject H_0	Type I error	Correct decision
	Do not reject H_0	Correct decision	Type II error

Slide 27, Lecture 1, dr. Carlo Cavicchia (2022)

The probability of getting a Type I error is given by significance level and is rejecting null hypothesis that is actually true.
On the other hand, Type II error is failing to reject null hypothesis that is actually false.

One-sided vs two-sided z test

Two-sided z test example: Test H0 : $\mu = x$ against Ha :$\mu \neq x$
One-sided z test: Test H0 : $\mu = x$ against Ha : $\mu > x$
OR Test H0 : $\mu = x$ against Ha : $\mu < x$

Confidence interval

The sample mean $\overline{X}$ can be used to construct a confidence interval for the unknown mean μ. A **confidence interval** represents a range of plausible values for the population parameter.

Confidence interval of $100(1 - \alpha)\%$ for the mean μ is:
$C = \overline{x} \pm z^*_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
where n-sample size, σ- standard deviation, C- area between critical values -z* and z* under the standard Normal curve.

One-sample t-test

A one-sample t test is used for a population with unknown mean μ as well as standard deviation σ.

For sufficiently large n, sample mean $\overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$. σ can be approximated by sample standard error denoted s.

Then, $t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$,

following t distribution with (n-1) degrees of freedom.

One-sample test for the mean summarised

$H_0 : \mu = \mu_0$, significance level α		σ known	σ unknown
Test statistic		$z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$	$t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$
$H_a : \mu > \mu_0$	Rejection region	$z > z^*_\alpha$	$t > t^*_\alpha(n-1)$
	P value	$P(Z \geq z)$	$P(T \geq t)$
$H_a : \mu < \mu_0$	Rejection region	$z < -z^*_\alpha$	$t < -t^*_\alpha(n-1)$
	P value	$P(Z \leq z)$	$P(T \leq t)$
$H_a : \mu \neq \mu_0$	Rejection region	$\lvert z\rvert > z^*_{\alpha/2}$	$\lvert t\rvert > t^*_{\alpha/2}(n-1)$
	P value	$2P(Z \geq \lvert z\rvert)$	$2P(T \geq \lvert t\rvert)$
Confidence interval		$\bar{x} \pm z^*_{\alpha/2}\frac{\sigma}{\sqrt{n}}$	$\bar{x} \pm t^*_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}$

Slide 44, Lecture 1, dr. Carlo Cavicchia (2022)

Matched pairs t test

Matched pairs t test refers to the situation when there are two measurements for each individual in the population. The respective measurements correspond to a random sample described by the μ_1 and μ_2 respectively.

Use difference $d = x_1 - x_2$ creating a one-sample t test of the difference.

We want to test whether $\mu_1 = \mu_2$, by using difference we can formulate hypothesis as follows:
H0: $\mu_1 - \mu_2 = 0$ against Ha: $\mu_1 - \mu_2 > 0$ (becomes *one-sample t-test* of the difference)

Note: t test in SPSS always gives the P value for two-sided test (needs to be divided by two for one-sided test)

T-test under non-normality

- t test requires sample data to follow normal distribution
- Mean and standard deviation are sensitive to outliers, hence so is the t statistic

t test can still be used if there are no outliers and…

- n sufficiently large (e.g., n > 100)
- n moderate (e.g., 20 ≤ n ≤ 100) and little skewness
- n small (e.g., n < 20) and data approximately normally distributed

Applied statistics 2 – IBEB – Lecture 2, week 2

Sign test

Sign test is a test for matched pairs, when there are two measurements for each individual in the population.
Unlike matched pairs t test, the sign test evaluates the medians instead of the means. Additionally, the sign test is more robust (i.e. "capable of performing without failure under a wide range of conditions") to skewness of the distribution and outliers.

When performing a sign test we want to compare the number of positive differences and negative differences. A **tie** is a pair of measurements with difference 0, which is neither positive nor negative, these types of measurements are removed from the sample. Let p denote proportion of positive differences.

1. The formulated hypothesis are as follows:
   H0: p=0.5 against Ha: p>0.5
2. Test statistic x= the number of positive differences
   $X \sim Binomial(n, p = 0.5)$ (under the null hypothesis)
3. P-value: $P(X \geq x) = P(X = x) + \ldots + P(X = n)$
4. Reject H0 if $p\ value < \alpha$

Sign test with normal approximation

Binomial distribution can be approximated by normal distribution $N(np, \sqrt{np(1-p)})$, where μ=np and σ=$\sqrt{np(1-p)}$ if:

- np ≥ 10
- n(1-p) ≥ 10

continuity correction

When using normal approximation, continuity correction is required as binomial is an integer-valued distribution whereas normal is a continuous distribution.
Example: $P(X \geq 25)^{Binomial} = P(X > 24.5)^{Normal}$

Sign test vs. t-test

Matched pairs t test is applicable if:

- No outliers
- Sample mean approximately normal

When both matched pairs t test and sign test can be applied, sign test is less powerful. For the same P(Type I error), sign test will have a higher P(Type II error).

Wilcoxon signed rank test

Wilcoxon signed rank test is a nonparametric test for matched pairs. Compared to the sign test it is also robust to outliers, but not skewness, however it is more powerful than sign test. For the same P(Type I error), sign test will have a higher P(Type II error) than a Wilcoxon test.

When performing Wilcoxon signed rank test we use the sum of positive/negative ranks.

1. Test statistic W^+: the sum of (+) ranks

Total rank sum = n(n+1)/2
Expected positive (negative) rank sum = n(n+1)/4
Under H0: $W^+ \sim N(\mu_{W+}, \sigma_{W+})$
$\mu_{W+} = \frac{n(n+1)}{4}$; $\sigma_{W+} = \sqrt{n(n+1)(2n+1)/24}$

2. P-value: $P(W^+ \geq x) = P(Z \geq \frac{x-\mu_{W+}}{\sigma_{W+}})$, x-test statistic
3. Reject H0 if $p\ value < \alpha$

two types of ties:

1. Observations with difference 0, which are removed from the sample; v
2. Several observations with the same absolute difference for which we need to assign the average rank (0.5 rank) (e.g., if we had two similar observations then we take the average of the two ranks).

Overview of matched pairs tests

Most powerful to Least powerful:

1. Matched pairs t test: no outliers + approximately normal distribution
2. Wilcoxon signed rank test: symmetrical distribution, BUT allows for outliers
3. Sign test: allows for outliers and skewness of distribution

Applied statistics 2 – IBEB – Lecture 3, week 2

Two-sample t test

matched pairs t test vs. two-sample t test

For both tests H_0: $\mu_1 = \mu_2$

Matched pairs t test: two measurements for each individual

Two-sample t test: two independent samples

- Individuals do not have to be paired
- Group sizes do not have to be the same

Two types of two-sample t tests

Perform a statistical test on H0: σ1 = σ2 vs. Ha: σ1 ≠ σ2

o F test
o Levene's test
If the variances are equal, two-sample t test with equal variance is preferred.

1. Two-sample t test with equal variance assumed. Assume that σ1 = σ2
2. Two-sample t test with equal variance not assumed. Do not assume that σ1 = σ2

F-test on equality of variances

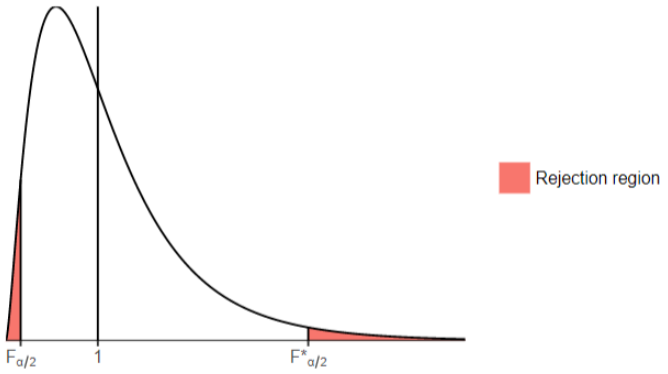
To determine which t test we should proceed with, we perform a test for equality of variances. Below is the procedure for an F-test:

Sample n1 observations from Population 1 with a normal distribution $N(\mu_1, \sigma_1^2)$.

Sample n2 observations from Population 2 with a normal distribution $N(\mu_2, \sigma_2^2)$

1. Formulate hypothesis: H0: $\sigma_1^2 = \sigma_2^2$ against Ha: $\sigma_1^2 \neq \sigma_2^2$.
2. Calculate sample variances: s_1^2, s_2^2.
3. Reject H0 when $s_1^2 >> s_2^2$ or $s_1^2 << s_2^2 \Rightarrow s_1^2/s_2^2 >> 1$ or $s_1^2/s_2^2 << 1$

Under H0, $F = s_1^2/s_2^2$ follows an F distribution with degrees of freedom $(n_1 - 1, n_2 - 1)$.



$\longrightarrow$ Reject H_0 when
$s_1^2/s_2^2 > F^*_{\alpha/2}(n_1 - 1, n_2 - 1)$ or $s_1^2/s_2^2 < F_{\alpha/2}(n_1 - 1, n_2 - 1)$

Slide 7, Lecture 3, dr. Carlo Cavicchia (2022)

Rejection region: $s_L^2/s_s^2 > F^*_{\alpha/2}(n_L - 1, n_s - 1)$

(s_L denotes the larger value of s_1 and s_2, s_s the smaller one.)

Levene's test

Levene's test is:

- used by SPSS instead of F test
- compares two or more variances
- not equivalent to the F test when two variances are compared

Two-sample t-test for means

Equal Variances $\sigma_1^2 = \sigma_2^2$

Test statistic: $t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$,

where $s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$ (s_p is pooled standard deviation)

Degrees of freedom: $n_1 + n_2 - 2$ (t distribution)

Unequal variances $\sigma_1^2 \neq \sigma_2^2$

Test statistic: $t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

Degrees of freedom: *k* either approximate by a software or $k = min(n_1 - 1, n_2 - 1)$ (t distribution)

Wilcoxon rank sum test

Parametric tests as F test and t test require (approximately) normal distributions of both samples. A non-parametric alternative is the **Wilcoxon Rank sum test** (robust to outliers). The test is performed as follows:

1. Formulate hypothesis:
   H0: no difference in [measurement] between groups
   Ha: systematically higher/lower [measurement] in one group
2. Rank results lowest to highest irrespective of group, in the event of a tie, assign the average rank
3. Test statistic W: Sum ranks per group to find the rank sum of one group W
   Under H0, W can be approximated by Normal distribution with mean and standard deviation:
   $\mu_W = n_1(N + 1)/2$, $\sigma_W = \sqrt{n_1 n_2 (N + 1)/12}$
4. Find p-value or critical region.
5. Reject H0 if $p\ value < \alpha$

One-way ANOVA (analysis of variances)

A **one-way ANOVA** is a more general test that allows us to compare the means from two or more groups. Test based on the ratio of

- between-group variation: differences between the group averages
- within-group variation: overall population variance

Assumptions:

1. Independent random samples from groups
2. Data is normally distributed in each group
3. Within-group variances are equal for all groups

Variation between groups is:

- small if the sample means are close
- large if the sample means differ much

Notation:

x_{ij} observation j in group i

$\overline{x_i}$ sample mean in group i

$\overline{x}$ overall sample mean

n_i number of observations in group i

N total number of observations

Measure of variation:

Total: $SST = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (x_{ij} - \overline{x})^2$

Between group: $SSB = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (\overline{x_i} - \overline{x})^2$

Within group: $SSW = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (x_{ij} - \overline{x_i})^2$

SST =SSB+SSW

To perform the test:

1. Formulate hypothesis:

   H0: $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_i$

Ha: not all of the μ are equal

2. Test statistic: $F = \frac{MSB}{MSW} = \frac{SSB/(I-1)}{SSW/(N-I)}$

   Under H0, F follows F-distribution with $I - 1$ and $N - I$ degrees of freedom
3. Reject H0 if F is too large (test statistic > critical value)

Limitations to one-way ANOVA:

- Outliers
- Skewed distributions (for small samples)
- Unequal variances within the groups
  - Rule of thumb for standard deviations: $s_{largest} < 2s_{smallest}$
  - More precision => Levene's test

Kruskal-Wallis test

A **Kruskal-Wallis** test is a non-parametric, rank-based alternative to the one-way ANOVA test.

The test is performed as follows:

1. Hypothesis:

   H0: [measurement] has the same distribution in all groups

   Ha: distributions of [measurement] are different for some groups
2. Rank lowest to highest (assign average rank for ties).
3. Rank sum R_i for each group.
4. Average rank sum $\frac{R_i}{n_i}$ for each group.

   Overall average rank sum = (N+1)/2, should be close to each Ri/ni under H0.
5. Test statistic: $H = \frac{12}{N(N+1)} \sum_{i=1}^{I} \sum_{j=1}^{n_i} \left(\frac{R_i}{n_i} - \frac{N+1}{2}\right)^2 = \frac{12}{N(N+1)} \sum_{i=1}^{I} \frac{R_i^2}{n_i} - 3(N+1)$

   Under H0, H approximately follows a χ^2-distribution (Chi-squared distribution) with I-1 degrees of freedom.

Reject H0 if test statistic > critical value (if H is too large).

Applied statistics 2 – IBEB – lecture 4, week 3

One-way ANOVA continued

one-way ANOVA vs two-sample t test

With equal variance and only two groups, one-way ANOVA and two-sample t test are equivalent: $F = t^2$

One-way ANOVA

Assumptions:

1. Independent random samples from groups
2. Data is normally distributed in each group
3. Within-group variances are equal for all groups → $x_{ik} \sim N(\mu_i, \sigma)$

Observations can be modelled as:

$x_{ik} = \mu_i + \epsilon_{ik}$, $(x_{ik} - \mu_i) \sim N(0, \sigma)$

where: x_{ik} observation k in group i

μ_i mean in group i

ϵ_{ik} error term of observation k in group i, (independent draws from N(0, σ))

$H_0 : \mu_1 = \ldots = \mu_i$

$$x_{ik} = \mu_i + \epsilon_{ik}, \text{ where } \mu_i = \mu + \tau_i$$

μ overall mean

τ_i group effect (centres around 0)

$$\text{then, } x_{ik} = \mu + \tau_i + \epsilon_{ik}$$

$\Rightarrow H_0 : \tau_1 = \ldots = \tau_i = 0$

Between-group variation (MSB): how much does τ_i vary around 0

Within-group variation (MSW): how much does ϵ_{ik} vary around 0

Two-way ANOVA

One-way ANOVA tests if a categorical variable (or factor) influences the mean of the continuous response variable. **Two-way ANOVA** tests whether there is an influence of two categorical variables on the means of the continuous response variable.

Two-way ANOVA considers:
• Response variable: dependent variable, we test if response variable is affected by factors (2 factors for two-way ANOVA)

Testing three effects and three sets of hypothesis:

1. Test for the main effect of Factor A:
   H0: Factor A has no effect on the mean
   Ha: Factor A has an effect on the mean
2. Test for the main effect of Factor B:
   H0: Factor B has no effect on the mean
   Ha: Factor B has an effect on the mean
3. Test for the interaction effect:
   H0: There is no interaction effect on the mean
   Ha: There is an interaction effect on the mean

Profile Plot

- No effect if the lines are overlapping and horizontal
- Effect of Factor A (x-axis) if lines are non-horizontal
- Effect of Factor B (y-axis) if lines are not overlapping
- Interaction effect if lines are not parallel (vice versa)

Two-way ANOVA model

$$x_{ijk} = \mu_{ij} + \epsilon_{ijk}$$
$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

Where:

x_{ijk} observation k in group (i, j)
μ overall mean
α_i effect of level i of factor A

β_j effect of level i of factor B

γ_{ij} interaction effect of level i of factor A and level j of factor B

(The above 3 terms measure deviations from the overall mean that can be attributed to factor A/B/interaction)

ϵ_{ijk} error term

$$x_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$$

SST: Total sum of squares
SSA: Sum of squares for main effect of factor A
SSB: Sum of squares for main effect of factor B
SSAB: Sum of squares for interaction effect
SSW: Within-group sum of squares (error sum of squares)

If the group sample sizes n_{ij} are equal, the variation can be decomposed into:
SST = SSA + SSB + SSAB + SSW

Two-way ANOVA model and test distribution

Source of variation	Sum of squares	Degrees of freedom	Mean Square	F	P-value
Factor A	SSA	$I-1$	$MSA=\frac{SSA}{I-1}$	$F_A=\frac{MSA}{MSW}$	
Factor B	SSB	$J-1$	$MSB=\frac{SSB}{J-1}$	$F_B=\frac{MSB}{MSW}$	
Interaction	$SSAB$	$(I-1)(J-1)$	$MSAB=\frac{SSAB}{(I-1)(J-1)}$	$F_{AB}=\frac{MSAB}{MSW}$	
Within groups	SSW	$N-IJ$	$MSW=\frac{SSW}{N-IJ}$		
Total	SST	$N-1$			

Slide 34, Lecture 4, dr. Carlo Cavicchia (2022)

Under H0: $F_A/F_B/F_{AB}$ follows F distribution with degrees of freedom of SSA/SSB/SSAB in the numerator and degrees of freedom of SSW in the denominator

P-value and rejection region:

Effect is significant if corresponding variation is too large compared to within-group variation.

Critical value is given by F^*_α with corresponding degrees of freedom

P-value: P(F > observed)

Reject H0 (no effect) if F statistic is too large:

Test statistic > critical value

P-value < significance level

χ^2 test on independence

One-way ANOVA: tests whether a continuous response variable is independent of a categorical variable

χ^2 **test on independence:** tests whether a categorical response variable is independent of a categorical variable

Example (Lecture 4, dr. C. Cavicchia): "Colour of packaging & product rating
Two categorical variables:

1. packaging colours (blue, pink)
2. rating options (poor, normal, good, excellent)

Free samples are given to customers and collects product ratings

500 customers receive blue
1000 customers receive pink

Question: does product rating depend on the colour of the packaging? Or more generally, is there a relationship between two categorical variables? "

Two-way tables

Hypothesis:

H0: no relationship between [one categorical variable] and [another categorical variable] (independence)
Ha: a relationship between [one categorical variable] and [another categorical variable]

Observed counts O_{ij}

- i: the color
- j: the rating

O_{ij}	Poor	Normal	Good	Excellent	Total
Blue	100	150	150	100	500
Pink	125	225	375	275	1000
Total	225	375	525	375	1500

Slide 44, Lecture 4, dr. Carlo Cavicchia (2022)

expected counts

Under H0, expected counts E_{ij} keeps the proportions among $\{ R_i \}$ and $\{ C_j \}$

R_i : row total; C_j: column total

$$E_{ij} = \frac{R_i * C_j}{n}$$

E_{ij}	Poor	Normal	Good	Excellent	Total
Blue	75	125	175	125	$R_1 = 500$
Pink	150	250	350	250	$R_2 = 1000$
Total	$C_1 = 225$	$C_2 = 375$	$C_3 = 525$	$C_4 = 375$	$n = 1500$

Slide 47, Lecture 4, dr. Carlo Cavicchia (2022)

χ^2 test on independence

Test statistic: $\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$,

where r is number of rows and c is the number of columns

Degrees of freedom: test distribution can be approximated by a χ^2-distribution with (r-1)*(c-1) degrees of freedom

(the approximation is reasonable if all $E_{ij} \geq 5$ (not O_{ij}))

Reject H0 if χ^2 is large

Critical value: Reject H0 if $\chi^2 > (\chi^2)^*_\alpha((r-1)*(c-1))$

P-value: P(test stat > observed)

χ^2 goodness-of-fit test

χ^2 **goodness-of-fit test** is to test whether the data fit a certain distribution.
Data: The observed count O_i for each category

To perform the test:

1. Hypothesis:

   H0: data fit a multinomial distribution with parameters…(p_1 =…, p_2 =…)

   Ha: data do not fit this distribution
2. Calculate the expected count E_i for each category, under H0. The expected counts are $E_i = np_i$ where n is the total number of observations.
3. Test statistic:

   $$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

   where k is the number of categories
4. Degrees of freedom: test distribution can be approximated by χ^2- distribution with (k-1) degrees of freedom.

   (Approximation is reasonable if all $E_i \geq 5$ (otherwise combine adjacent categories))
5. Reject H0:

   Critical value: $\chi^2 > (\chi^2)^*_\alpha(k-1)$

continuous distributions

For continuous distributions, construct intervals and compute expected and observed counts.
If we have to estimate parameters (e.g. for Normal distribution estimate μ by $\bar{x}$ and σ by s), we must adjust the degrees of freedom of the χ^2 -distribution: degrees of freedom = (k – 1 – # estimated parameters).

Applied statistics 2 – IBEB – Lecture 5, week 4

Linear regression

Simple linear regression model

Linear relationship between a response variable and an explanatory variable(s).
Mean of the response variable depends on the value of the explanatory variable

Notations:

$$y = \beta_0 + \beta_1 x + \epsilon$$

y	response variable
x	explanatory variable, predictors variable
β_0, β_1	regression coefficients
ϵ	error term

Response = Model + Error

- Model: the part of y explained by x
- Error: the part of y that is not explained by x
- $\epsilon \sim N(0, \sigma)$
- Interpretation: one unit increase in x is associated with a change of β_1 in y on average
- Prediction: the model can be used to make predictions about y by substituting x into the model

Estimating the line

ordinary least squares (OLS)

Assume the model is known as:

$$y = \beta_0 + \beta_1 x + \epsilon \text{ (data=model+error)}$$

To estimate the line:

1. β_0, β_1 estimated by b_0, b_1

2. fitted value $\hat{y}_i = b_0 + b_1 x_i$
3. residual $e_i = y_i - \hat{y}_i$

$$y_i = \hat{y}_i + e_i \text{ (data=fit+residual)}$$

The regression line of best fit is such that:

- $\hat{y}_i$ are the closest to y_i
- e_i are the closest to 0

The way to achieve that is to find b_0, b_1 that minimises the residual sum of squares:

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 = \sum_{i=1}^{n} \left(y_i - (b_0 + b_1 x_1)\right)^2$$

Note:

- Since $\epsilon \sim N(0, \sigma)$ presence of outliers can influence the fitted result
- It is sometimes not advisable to extrapolate too much outside of the range of observed data
- Regression results do not indicate the causality only correlation

Linear regression model

Linear relationship between a response variable and one or more explanatory variables:

- Simple regression (only one explanatory variable): p=1

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- Multiple regression (multiple explanatory variables): p>1

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i$$

Notation:

y	response variable
$x_{i1} \ldots x_{ip}$	explanatory variables, predictors variables
$\beta_1 \ldots \beta_p$	regression coefficients
ϵ_i	error term

In practice, $\beta_0, \ldots, \beta_p$ and error terms ϵ_i are unknown and estimated from the data.

$\beta_0, \ldots, \beta_p$ can be estimated by $b_0, \ldots, b_p$

Fitted values: $\widehat{y}_i = b_0 + b_1 x_{i1} + \ldots + b_p x_{ip}$

Residuals: $e_i = y_i - \widehat{y}_i = y_i - (b_0 + b_1 x_{i1} + \ldots + b_p x_{ip})$

ordinary least squares (OLS)

To minimise the residual sum of squares:

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \widehat{y}_i\right)^2 = \sum_{i=1}^{n} \left(y_i - (b_0 + b_1 x_{i1} + \ldots + b_p x_{ip})\right)^2$$

Significance of coefficients

assumption:

- $\epsilon_i \sim N(0, \sigma)$ and are independent
- Constant variance σ^2 for all error terms ϵ_i (homoskedasticity)

Then, coefficient estimators b_j follow normal distribution $b_j \sim N(\beta_j, \sigma_{b_j})$

t test: H0: β_j=0 against Ha: $\beta_j \neq 0$

standard errors

$$\epsilon_i \sim N(0, \sigma)$$

σ is estimated by the regression standard error s, where:

$$s = \sqrt{\frac{1}{n-p-1} \sum_{i=1}^{n} e_i^2}$$

In simple linear regression:

Standard error of intercept b_0: $SE_{b0} = s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$

Standard error of slope b_1: $SE_{b1} = s \sqrt{\frac{1}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$

significance of coefficients

Hypothesis:

H0: $\beta_j = 0$

Ha: $\beta_j \neq 0$

Observe b_j where $b_j \sim N(\beta_j, \sigma_{b_j})$

Test statistic:

$$t_{b_j} = \frac{b_j - 0}{SE_{bj}}$$

Degrees of freedom:

Under H0, test distribution follows t distribution with n-p-1 degrees of freedom

Reject H0 if $\left|t_{b_j}\right|$ is too large:

Critical value: $t^*_{\alpha/2}(n - p - 1)$

P-value: $P(|T| \geq \left|t_{b_j}\right|)$

confidence interval

Confidence interval for β_j: $b_j \pm t^*_{\alpha/2}(n - p - 1)SE_{b_j}$

interpretation

Each coefficient describes a **partial effect**: the effect of a change in the respective variable given that <u>all other variables remain constant</u>

Interpretation of constant term: Constant term β_0 can only be interpreted if value 0 is realistic for all explanatory variables x_i

Note: Too much correlation in explanatory variables is not desirable

Correctness of the model

Assumptions:

1. Linear relationship
   - Residual plot: e_i against $\hat{y}_i$
     - Residuals are randomly scattered around 0 with no pattern
2. Errors are independent
   - Fulfilled in case of random sample
3. Errors are normally distributed
   - Check histogram of residuals
4. Constant variance
   - Residual plot: e_i against $\hat{y}_i$
     - Residuals are equally scattered (no funnel shape)

Residual plots

residuals

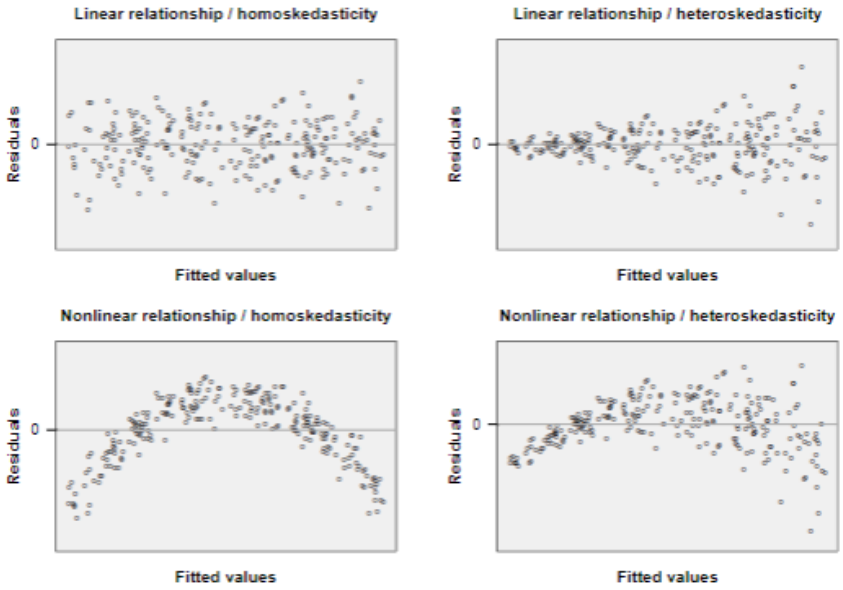
It is useful to check whether the linear regression model is valid by looking at:

- Residual plot
- Histogram of residuals

Residuals are fitted such that: $\sum_{i=1}^{n} e_i = 0$

$\chi 2$ goodness-of-fit test for normality for general data:

degree of freedom = n − 1 - #parameters



→ Mean of residuals always equal to 0

Slide 46, Lecture 5, dr. Carlo Cavicchia (2022)

Usefulness of the model

We need a model to

- explain the patterns/variations in y using $x_1 \dots x_p$
- predict values in y using $x_1 \dots x_p$

A model is **useful** if model explains much variation in y

Different $x_{i1}...x_{ip}$ give different $\hat{y}_i$

A model is not so useful if it does not explain the variation in y

Different $x_{i1}...x_{ip}$ give similar $\hat{y}_i$

The variation is shown in e_i

The measure of the usefulness of the model

The variation can be measured in terms of sums of squares:

Regression: $SSR = \sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}_i\right)^2$ (variation explained by model)

Residual: $SSE = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$ (unexplained variation)

Total: $SST = \sum_{i=1}^{n} \left(y_i - \bar{y}_i\right)^2$

SST=SSR+SSE

R^2 and sums of squares

R^2 is a squared correlation of observed values y and fitted values $\hat{y}$

$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$

Note:Adding more variables to the model increases R^2 (and vice versa)

ANOVA for regression

Use ANOVA to compare explained variation with unexplained variation

Hypothesis:

H0: $\beta_1 = ... = \beta_p = 0$

Ha: at least one $\beta_j \neq 0$

Test statistic:

$F = \frac{MSR}{MSE} = \frac{SSR/p}{SSE/(n-p-1)}$

Degrees of freedom:

Under H0, test distribution follows F distribution with p and n-p-1 degrees of freedom

Reject H0 if p-value < significance level

Using the model for prediction

two prediction problems

- What is the prediction of a new (unknown) response? (predict y)
- What is the prediction of the average of responses? (predict μ_y)

The best point predictions for y and μ_y are the same, but there's uncertainty about the point predictions, and the confidence intervals for y and μ_y are different.

Point prediction for μ_y : $\hat{\mu}_y = b_0 + b_1 x_1 + \ldots + b_p x_p$

Confidence interval $\mu_y : \hat{\mu}_y \pm t^*_{\alpha/2}(n - p - 1)SE_{\hat{\mu}_y}$

Point prediction for y : $\hat{y} = b_0 + b_1 x_1 + \ldots + b_p x_p$

Confidence interval $y: \hat{y} \pm t^*_{\alpha/2}(n - p - 1)SE_{\hat{y}}$

Although $\hat{y}$=$\hat{\mu}_y$, confidence interval for y is wider than confidence interval for μ_y as $SE_{\hat{y}} > SE_{\hat{\mu}_y}$

$$SE_{\hat{\mu}} = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

$$SE_{\hat{y}} = s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Slide 72, Lecture 5, dr. Carlo Cavicchia (2022)

Note that for multiple regressions, the formulas are more complicated

approximate intervals

With a large number of observations n:

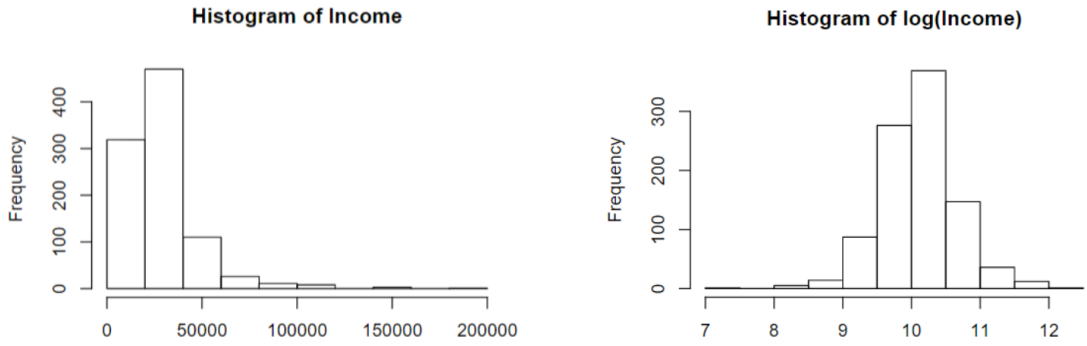
$SE_{\hat{\mu}_y}$ is close to 0

$SE_{\hat{y}}$ can be approximated by s (regression standard error)

Applied statistics 2 – IBEB – lecture 6, week 5

logarithms

Distribution of economic variables is often right-skewed
(e.g. income, sales, prices, returns, etc)
The use of **logarithm** may yield more symmetric distributions, however the interpretation of the models is different when logarithms are used.



Slide 9 and 11, Lecture 6, dr. Carlo Cavicchia (2022)

interpretation of original variables

Regression model: $y = \beta_0 + \beta_1 x + \epsilon$

- Mathematically, change is measured by differentiation: $\frac{dy}{dx} = \beta_1$
  1 unit change in x means β_1 units change in y
- Coefficient β_1 indicates the absolute change in y for an absolute change in x

logarithm of response variable

Regression model: $z = ln(y) = \beta_0 + \beta_1 x + \epsilon$
Mathematically: $\beta_1 = \frac{dz}{dx} = \frac{dz}{dy}\frac{dy}{dx} = \frac{1}{y}\frac{dy}{dx} = \frac{dy/y}{dx}$
1 unit change in x means (100 * β_1)% change in y

Coefficient β_1 indicates the relative change in y for an absolute change in x

Point prediction:

For the logarithm of response log (y): $\widehat{log(y)} = \hat{\mu}_{log(y)} = b_0 + b_1 x_1 + \ldots + b_p x_p$

For the response y: $\hat{y} = \hat{\mu}_y = e^{\hat{\mu}_{log(y)} + s^2/2}$, s = regression standard error

logarithm of explanatory variable

Regression model: $y = \beta_0 + \beta_1 ln(x) + \epsilon$

Mathematically: $\beta_1 = \frac{dy}{d\, ln(x)} = \frac{dy}{dx/x}$

1% increase in x means $\beta_1/100$ units change in y

Coefficient β_1 indicates the absolute change in y for a relative change in x

logarithm of response and explanatory variable

Regression model: $ln(y) = \beta_0 + \beta_1 ln(x) + \epsilon$

Mathematically: $\beta_1 = \frac{d\, ln(y)}{d\, ln(x)} = \frac{dy/y}{dx/x}$

1% increase in x means β_1% change in y

Coefficient β_1 indicates the relative change in y for a relative change in x (elasticity of y with respect to x)

squared effects

linear and squared effects

Sometimes data does not follow the linear relationship e.g. average income tends to increase with age at persons' early stage of career, but then decreases with age in later stages of career

- **parabolic relationship** can be used to model effect rather than a linear relationship
- we do that by add squared age as explanatory variable to the model as follows:

The model: $ln(income) = \beta_0 + \beta_1 * age + \beta_2 * age^2 + error$

Test the significance of effect of age on logarithm of income:
t tests assess the significance of the two variables individually
ANOVA test for joint significance of effect of age on logarithm of income (joint significance of linear and squared effects)

dummy variables

Dummy variable- is a variable that can take on only the value of either 1 or 0 and represents the presence/absence of some categorical effect

For example, by adding a dummy variable for gender (value 1 for women and value 0 for men) we can see if gender has an effect on income

Regression coefficient indicates change in logarithm of income for women compared to men

The model: $ln(income) = \beta_0 + \beta_1 * age + \beta_2 * age^2 + \beta_3 * 1_{female} + error$

1_{female} takes on value 1 for females and 0 for males

$\beta_3 < 0$ means for a given age, income is on average β_3*100 % lower for females than males

Note: If changing the reference category, constant coefficient β_0 will also change

categorical variables

If we want to incorporate categorical variable (with more than 2 outcomes) into the model we can consider a dummy variable for each outcome.

Example: To see if economics status has an effect on income we use a categorical variable with three outcomes:

- Working full time
- Working part time
- Retired or gave up business

One of the dummy variables is unnecessary as if the values in two dummy variables are known, the value in the third dummy variable is known as well.

For categorical variable with k outcomes, k - 1 dummy variables are added to the model

Reference category is the left out category

Regression coefficients indicate change in the response variable with respect to the reference category
Example: in this case of economic status effect on income, full-time work seems reasonable reference category as it is the majority

Model:
$ln(income) = \beta_0 + \beta_1 * age + \beta_2 * age^2 + \beta_3 * 1_{female} + \beta_4 * 1_{part\ time} + \beta_5 * 1_{retired} + error$
Interpretation:
β_4<0: For a given age and gender income is on average β_4*100% lower for part-time workers than full-time workers
β_5<0: For a given age and gender income is on average β_5*100% lower for retired than full-time workers

comparing linear regression models

Model: $y_i = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \epsilon$
t-test: H_0: $\beta_p = 0$ (each coefficient)
ANOVA test: H_0: $\beta_1 = \beta_2 = ... = \beta_p = 0$ (all the coefficients)
F-test: H_0: $\beta_4 = \beta_5 = ... = \beta_p = 0$ (a subset)

F-test

When adding a variable to the model it has an effect of increasing R^2 whether or not the variable has an explanatory power
F test aims to test if there is a significant increase in R^2 by comparing the full model with a restricted model

- Full Model: $y_i = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \epsilon$
- Restricted model: $y_i = \beta_0 + \beta_1 x_1 + ... + \beta_{p-q} x_{p-q} + \epsilon$ (the model without the last q variables)

To test if the q variables $x_{p-q+1}, \dots, x_p$ add explanatory power:
Hypothesis:
H0: $\beta_{p-q+1} = ... = \beta_p = 0$
Ha: at least one of $\beta_{p-q+1}, \dots, \beta_p$ is not equal to 0

Test statistic: measure the change in R^2

$$F = \frac{(R_F^2 - R_R^2)/q}{(1-R_F^2)/(n-p-1)}, R_F^2 \geq R_R^2$$

R_F^2= R^2 of full model

R_R^2= R^2 of restricted model

Degrees of freedom: F distribution with degrees of freedom q and (n-p-1)

Reject H0 if F is too large (change in R^2 is too large)

Critical value: $F_\alpha^*(q,\ n - p - 1)$

interaction effects

Example: does age have a different effect on income for men and women?
Is the interaction effect significant?

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
5	.470[a]	.221	.217	.486	.221	53.424	5	942	.000
6	.479[b]	.229	.223	.484	.008	4.874	2	940	.008

a. Predictors: (Constant), Age, AgeSq, Female, PartTime, Retired

b. Predictors: (Constant), Age, AgeSq, Female, AgeFemale, AgeSqFemale, PartTime Retired

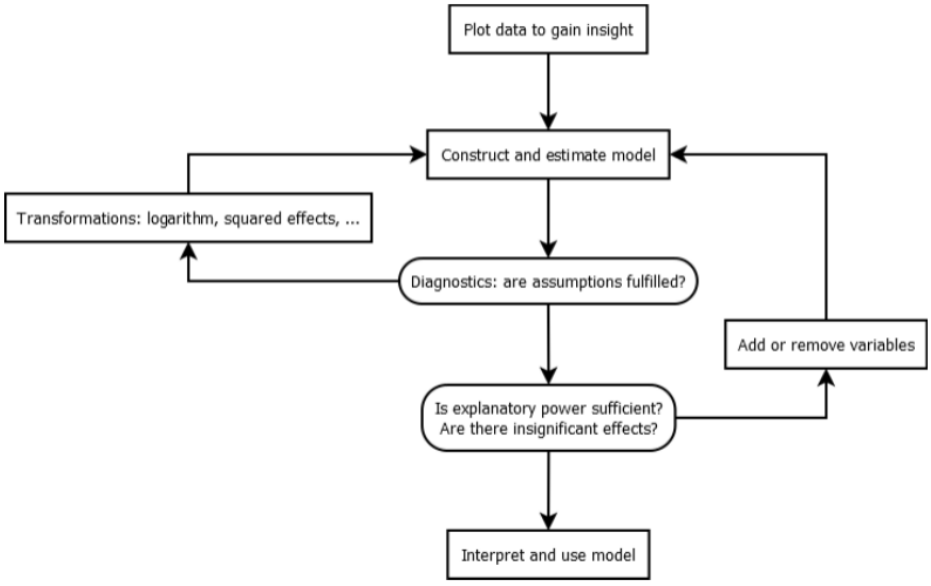
$\longrightarrow$ P-value $= 0.008 < 0.05$, hence interaction effect is significant

$\longrightarrow$ Effect of age on logarithm of income is significantly different for men and women

Slide 48, Lecture 6, dr. Carlo Cavicchia (2022)

model building

Model Building in practice:



Slide 65, Lecture 6, dr. Carlo Cavicchia (2022)

model building and standard errors

Standard errors are derived under the assumption that:

1. Variables are selected based on (economic) theory
2. Only afterwards data are collected and analysis is performed

Data snooping - deciding which procedure to use after looking at the data.
Standard errors become invalid

If you intend to modify your analysis according to the results from the data (if you intend to snoop the data): Randomly separate the data into two parts:

- One part of the data for finding out the best model
- One part for applying the model and calculate the standard error

Applied statistics 2 – IBEB – Lecture 7, week 6

Time series

- The study of the variable over time.
- Often used in economics (e.g. GDP, inflation, sales).
- Measurements are taken at (regular) intervals over time

trend

Time series plot shows an overall behaviour over time (e.g. general increase of GDP with time)
Linear trend can be modelled by regression model, with time as an explanatory variable:
$y_t = \beta_0 + \beta_1 t + \epsilon_t$

seasonality

Time series plot shows seasonal pattern (seasonality)
To add seasonal effects into the model:

1. Add dummy variables for different months
2. Use one month as reference category (for example, December):

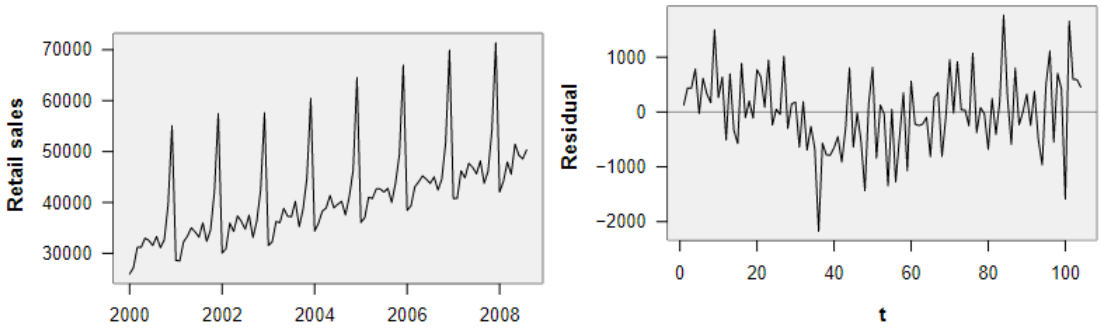
$y_t = \beta_0 + \beta_1 t + \beta_2 Jan + \ldots + \beta_{12} Nov + \epsilon_t$

Removing trend and seasonality

In practice, trend and seasonality are often estimated to remove them from the time series to obtain the dataset that is stable over time (a time series whose expected value does not change overtime)
We can do so by plotting residuals against time:

- Estimated model: $y_t = b_0 + b_1 t + b_2 Jan + \ldots + b_{12} Nov + e_t$
- Define new time series: $z_t = y_t - (b_0 + b_1 t + b_2 Jan + \ldots + b_{12} Nov)$

  where $z_t = e_t$



Slide 14, Lecture 7, dr. Carlo Cavicchia (2022)

Further analysis is performed on stationary time series z_t with another regression model (**autoregressive model**): $z_t = \gamma_0 + \gamma_1 z_{t-1} + \gamma_2 z_{t-2} + \gamma_3 z_{t-3} + \epsilon_t$

Autoregressive models

Shows the linear relationship between successive values of a time series (predicting current value with past values)
AR(1): $y_t = \beta_0 + \beta_1 y_{t-1} + \epsilon_t$ (where $y_t =$ current value and $y_{t-1} =$ past value)
AR(L): $y_t = \beta_0 + \beta_1 y_{t-1} + \ldots + \beta_L y_{t-L} + \epsilon_t$

Notes on the model:

- Additional explanatory variables can be added to the model
- Can be estimated by ordinary least squares (OLS) regression
- Can be used for predictions similarly to other linear models

Differences

Differences

Taking differences $y_t - y_{t-1}$ often removes trend

Differences of logarithms (log returns)

Difference of logarithms is a good approximation of growth rate (percentage change from t-1 to t)
differences of logarithms: $ln(y_t) - ln(y_{t-1}) \approx \frac{(y_t - y_{t-1})}{y_{t-1}}$

Note: for small values of x ln(1+x)≈x

Goldfeld - Quandt test

Similar to the F-test for equality of variances. Goldfeld-quandt test tests if variance of error terms is constant over time.
Linear regression model:
$y = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \epsilon$
Test is performed as follows:

1. Entire time period is split in three parts and the middle part is removed (first part becomes period 1 and third part becomes period 2)
   Size of deleted middle part:
   n/5 if n is small ($\frac{2}{5}n$ is the size of period 1 and period 2)
   n/3 otherwise
2. Estimate model and error variance σ_1^2 in first period
3. Estimate model and error variance σ_2^2 in second period

Perform F-test to test if $\sigma_1^2 = \sigma_2^2$

1. Hypothesis:
   H0: $\sigma_1^2 = \sigma_2^2$
   Ha: $\sigma_1^2 \neq \sigma_2^2$
2. Test statistic:

$F = \frac{larger\ s^2}{smaller\ s^2} = \frac{s_L^2}{s_S^2}$

3. Degrees of freedom: test statistic follows F distribution with degrees of freedom $(n_L - p - 1)$ in numerator and $(n_S - p - 1)$ in denominator
   n_L= number of observations in the sample of period with larger s
   p= number of explanatory variables in the linear regression
4. Reject H0 if $F > F^*_{\alpha/2}(n_L - p - 1, n_S - p - 1)$ *(Note: two sided test => F^*_{α})*

Chow break test

Chow break test is to test whether the linear relationship is constant over time.

- Consider an assumed break point in time period
- Test if regression parameters are different before and after break

Before break: $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_m x_m + \epsilon$

After break: $y = (\beta_0 + \gamma_0) + (\beta_1 + \gamma_1)x_1 + \ldots + (\beta_m + \gamma_m)x_m + \epsilon$

There is no change if $\gamma_0 = \ldots = \gamma_m = 0$

A dummy variable d with value 0 before the break and value 1 after is introduced.

The regression model can be written as:

$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_m x_m + \gamma_0 d + \gamma_1 (d \times x_1) + \ldots + \gamma_m (d \times x_m) + \epsilon$

Change in relationship if any of $\gamma_0, \ldots, \gamma_m$ are not 0

F test on increase R^2

Full model (with dummy variables):

$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_m x_m + \gamma_0 d + \gamma_1 (d \times x_1) + \ldots + \gamma_m (d \times x_m) + \epsilon$

Restricted model (before the break):

$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_m x_m + \epsilon$

Test is performed as follows:

1. Hypothesis:
   H0: $\gamma_0 = \ldots = \gamma_m = 0$
   Ha: at least one of $\gamma_0, \ldots, \gamma_m$ not 0
2. The test statistic:
   $F = \frac{(R_F^2 - R_R^2)/q}{(1 - R_F^2)/(n-p-1)}$
   p = number of variables in full model = 2m+1
   q = number of restrictions = m+1
3. Degrees of freedom: F distribution with q and (n-p-1) degrees of freedom
4. Reject H0 if $F > F_\alpha^*(q, n-p-1)$ (p-value computed only in upper tail of distribution)

Granger causality test

Recall, the results of regression models do not indicate causality, only correlation.

With time series, it is possible to test if past values of one variable are useful to predict current values of another variable.

Note: granger causality does not guarantee causal relationship

Granger causality test

F test on increase R^2

Full model: $y = \beta_0 + \beta_1 y_{t-1} + \ldots + \beta_L y_{t-L} + \beta_{L+1} x_{t-1} + \beta_p x_{t-L} + \epsilon_t$

Restricted model: $y = \beta_0 + \beta_1 y_{t-1} + \ldots + \beta_L y_{t-L} + \epsilon_t$

Perform the test as follows:

1. Hypothesis:

   H0: $\beta_{L+1} = \ldots = \beta_p = 0$

   Ha: at least one $\beta_{L+1}, \ldots, \beta_p$ not 0

2. The test statistic:

   $$F = \frac{(R_F^2 - R_R^2)/q}{(1 - R_F^2)/(n-p-1)}$$

   q = number of restrictions = L

   p = number of variables in full model = 2L

3. Degrees of freedom: F distribution with q and (n-p-1) degrees of freedom
4. Reject H0 if $F > F_\alpha^*(q, n-p-1)$ (p-value computed only in upper tail of distribution)

Reference list

Cavicchia, C. (2022). *Lecture 1: Hypothesis Testing Review* [PowerPoint slides]. Retrieved from: https://canvas.eur.nl/courses/39842/files/69266574

Cavicchia, C. (2022). *Lecture 2: Matched Pairs Tests* [PowerPoint slides]. Retrieved from: https://canvas.eur.nl/courses/39842/files/69266572

Cavicchia, C. (2022). *Lecture 3: Two-sample and multi-sample tests* [PowerPoint slides]. Retrieved from: https://canvas.eur.nl/courses/39842/files/69266579

Cavicchia, C. (2022). *Lecture 4: Two-way ANOVA, χ^2 tests* [PowerPoint slides]. Retrieved from: https://canvas.eur.nl/courses/39842/files/70133983

Cavicchia, C. (2022). *Lecture 5: Linear regression: estimation and diagnostics* [PowerPoint slides]. Retrieved from: https://canvas.eur.nl/courses/39842/files/70414212

Cavicchia, C. (2022). *Lecture 6: Transformations and model selection* [PowerPoint slides]. Retrieved from: https://canvas.eur.nl/courses/39842/files/69266604

Cavicchia, C. (2022). *Lecture 7: Linear regression with time series data* [PowerPoint slides]. Retrieved from: https://canvas.eur.nl/courses/39842/files/69266606