Tutorial 2 Business Analytics: Statistics

"Test Manual" – Overview

1. i) 1 sample or 2 samples

   ii) If 1 sample: σ_x known or unknown
   If 2 samples: dependent or independent

2. State H_0 and H_1 (given)

3. Select and calculate the test statistic

4. Select α (given)

5. Find the critical value in the table

6. Result

Tutorial 2 Business Analytics: Statistics

"Test Manual" – 3rd Step

When to use which test? We want to make a statement about the mean of a population, μ_x, based on a sample with size n_x and mean $\bar{x}$

1 Sample

- σ_x known → Gauss/z-test $z_0 = \frac{\bar{x}-\mu_0}{\sigma_X}\sqrt{n} \sim N(0,1)$
- σ_x unknown → t-test $t_0 = \frac{\bar{x}-\mu_0}{s_X}\sqrt{n} \sim t_{n-1}$ with $s_x^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n}(x_i - \bar{x})^2$

2 Samples

- independent → Welch-test $t_0 = \frac{\bar{x}-\bar{w}-\mu_0}{s_{\bar{x}-\bar{w}}} \sim_{\text{approx}} t_{\text{df}}$ with $s_{\bar{x}-\bar{w}}^2 = \frac{s_x^2}{n_x} + \frac{s_W^2}{n_W}$ and

  $s_x^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n}(x_i - \bar{x})^2$ $\left(\text{df} = \frac{\left(s_{\bar{x}-\bar{w}}^2\right)^2}{\frac{s_x^4}{n_x^2(n_x-1)} + \frac{s_W^4}{n_W^2(n_W-1)}}\right.$ rounded to nearest integer number)

- dependent → Paired t-test $t_0 = \frac{\bar{d}-\mu_0}{s_d}\sqrt{n} \sim t_{n-1}$ with $s_d^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n}(d_i - \bar{d})^2$ and

  $\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i = \bar{x} - \bar{w}$, $d_i = x_i - w_i$, $\mu_D = \mu_X - \mu_W$

Tutorial 2 Business Analytics: Statistics

“Test Manual” – 5th Step

How to find the critical value in the table? For

- Gauss/z-Test → use normal distribution
- t-Test, Welch-Test and Paired t-Test → use t-distribution

H_1	t^c range	t^c value
$\mu_x \neq \mu_0$	can be any, $\mathbb{R}$	$\left\|t^c_{1-\frac{\alpha}{2};\,\mathrm{df}}\right\| = \left\|t^c_{\frac{\alpha}{2};\,\mathrm{df}}\right\|$
$\mu_x > \mu_0$	must be positive, $\mathbb{R}_{>0}$	$t^c_{1-\alpha;\,\mathrm{df}}$
$\mu_x < \mu_0$	must be negative, $\mathbb{R}_{<0}$	$t^c_{\alpha;\,\mathrm{df}}$

Tutorial 2 Business Analytics: Statistics

“Test Manual” – 6th Step

Reject H_0:

H_1	p-value criterion	test statistic criterion
$\mu_x \neq \mu_0$	$p < \alpha$	$\lvert t_0 \rvert > \left\lvert t^c_{1-\frac{\alpha}{2};\,\mathrm{df}} \right\rvert$
$\mu_x > \mu_0$	$p < \alpha$	$t_0 > t^c_{1-\alpha;\,\mathrm{df}}$
$\mu_x < \mu_0$	$p < \alpha$	$t_0 < t^c_{\alpha;\,\mathrm{df}}$

Tutorial 2 Business Analytics: Statistics

Example: Learning Method Comparison

In order to compare two learning methods, results have been measured for a group of students. Test if the students got better (higher) results using method 2. Assume the difference follows a normal distribution, (significance level of 5%, i.e., $\alpha = 0.05$).

student	1	2	3	4	5
method 1 (x)	8	6	8	8	4
method 2 (w)	10	9	7	12	7

1.) i) 2 samples ii) dependent

2.) $H_0\colon\ \mu_D = \mu_X - \mu_W \geq \mu_0 = 0$ $H_1\colon\ \mu_D = \mu_X - \mu_W < \mu_0 = 0$

3.) → Paired t-Test: $t_0 = \frac{\bar{d}-\mu_0}{s_d}\sqrt{n} \sim t_{n-1}$ with unbiased sample variance $s_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2$

sample means: $\bar{x} = 6.8$, $\bar{w} = 9.0$, difference $\bar{d} = -2.2$,

$s_d^2 = 3.7, \quad s_d = 1.9235 \quad \Rightarrow \quad t_0 = -2.5574$

4.) $\alpha = 0.05$

5.) $\rightarrow\ t^c_{\alpha;n-1} = -t^c_{1-\alpha;n-1}\ (\text{sym.}) \ \Rightarrow\ t^c_{0.05;4} = -t^c_{0.95;4} \overset{\text{table}}{=} -2.132$

6.) $t_0 = -2.557 < -2.132 = t^c_{0.05;4} \ \Rightarrow$ Reject H_0: Learning method 2 is significantly better.

Tutorial 2 Business Analytics: Statistics

Example: Learning Method Comparison – step 3 details

In order to compare two learning methods, results have been measured for a group of students. Test if the students got better (higher) results using method 2. Assume the difference follows a normal distribution, (significance level of 5%, i.e., $\alpha = 0.05$).

student	1	2	3	4	5
method 1 (x)	8	6	8	8	4
method 2 (w)	10	9	7	12	7

3.)

sample means: $\bar{x} = \frac{1}{5}(8+6+8+8+4) = 6.8$, $\bar{w} = \frac{1}{5}(10+9+8+12+7) = 9.0$

difference: $\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i = \bar{x} - \bar{w} = -2.2$

sample variance: $s_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2$, $d_i = x_i - w_i$,

$$s_d^2 = \frac{1}{4}\left((8-10+2.2)^2 + (6-9+2.2)^2 + (8-7+2.2)^2 + (8-12+2.2)^2 + (4-7+2.2)^2\right) = 3.7$$

$$s_d = 1.9235$$

Tutorial 2 Business Analytics: Statistics

Confidence Intervals

Find confidence intervals for μ_x, which—under H_0—contain the true value μ_x with a probability of at least $1-\alpha$ (confidence level). We differentiate two cases:

- σ_x known:

confidence interval: $$\boxed{[I_u(x),\ I_o(x)] = \left[\bar{x} - z^c_{1-\frac{\alpha}{2}} \frac{\sigma_x}{\sqrt{n}},\ \bar{x} + z^c_{1-\alpha/2} \frac{\sigma_x}{\sqrt{n}}\right]}$$

- σ_x unknown, use s_x as estimate instead:

confidence interval: $$\boxed{[I_u(x),\ I_o(x)] = \left[\bar{x} - t^c_{1-\frac{\alpha}{2};\, n-1} \frac{s_x}{\sqrt{n}},\ \bar{x} + t^c_{1-\frac{\alpha}{2};\, n-1} \frac{s_x}{\sqrt{n}}\right]}$$

- Values of μ_0 within the confidence interval cannot be rejected regarding a significance level of α

  → Reject H_0 if μ_0 is not in the confidence interval

Tutorial Business Analytics

Finding the estimators

- Squared error of a point (residual): $e_i^2 = \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)^2$
- Residual Sum Squares: $\text{RSS} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}\left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)^2$

$$\min_{\hat{\beta}_0,\hat{\beta}_1} \left\{RSS = \sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2\right\}$$

… (set partial derivatives equal to zero)

$$\Rightarrow \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\Rightarrow \quad \hat{\beta}_1 = \frac{Cov(x,y)}{Var(x)} = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_i^n (x_i - \bar{x})^2} = \frac{\frac{1}{n}\sum_i^n x_i y_i - \bar{x}\bar{y}}{\frac{1}{n}\sum_i^n x_i^2 - \bar{x}^2} = \frac{S_{xy}}{S_{xx}}$$

Tutorial Business Analytics

Finding the estimators

- Squared error of a point (residual): $e_i^2 = \left(y_i - (\hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{ij})\right)^2$

- Residual Sum Squares: RSS $= e^T e = (y - \mathbf{X}\hat{\beta})^T (y - \mathbf{X}\hat{\beta})$

$$\min_{\hat{\beta}} \left\{RSS = (y - \mathbf{X}\hat{\beta})^T (y - \mathbf{X}\hat{\beta})\right\}$$

… (take derivative and use FOC and SOC)

$$\Rightarrow \qquad \hat{\beta} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T y$$

Tutorial Business Analytics

Testing the significance of regression coefficients

- Follow “test manual ” from Tutorial 2 to do the Hypothesis testing

- The **test statistic** is calculated as follows:

$$t_0 = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \sim t_{n-2}$$

$$SE(\hat{\beta}_1) = \sqrt{\frac{RSS}{\sum_{i=1}^{n}(x_i - \bar{x})^2} * \frac{1}{n-2}}$$

Tutorial Business Analytics

When to **reject H_0**?

$\boldsymbol{H_1}$	**using p-value**	**using test statistic**
$\hat{\beta}_j \neq 0$	p < α	$\lvert t_0 \rvert \geq \left\lvert t^c_{1-\frac{\alpha}{2};df} \right\rvert$
$\hat{\beta}_j > 0$	p < α	$t_0 \geq t^c_{1-\alpha;df}$
$\hat{\beta}_j < 0$	p < α	$t_0 \leq t^c_{\alpha;df}$

Tutorial Business Analytics

Evaluation of model

Measure the difference between true observations and the regression line

- Residual Sum of Squares (RSS):

$$\text{RSS} = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)^2$$

- Mean Squared Error (MSE):

$$\text{MSE} = \frac{RSS}{n}$$

- Root Mean Squared Error (RMSE):

$$\text{RMSE} = \sqrt{MSE}$$

- Coefficient of Determination (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{RSS}{TSS}$$

Tutorial Business Analytics

Gauss-Markov Theorem

Property	What does it mean?	Why do we need that?	How can we test that?
Linearity	Regression linear in the coefficients β	Core assumption of **linear** regression	Do not transform β, only the covariates
No Multicollinearity	• $rank(\mathbf{X}) = p$ • No high correlation between covariates	• Impossible to estimate coefficients • Non-significant coefficients	Variance Inflation Factor
Homoskedasticity	$Var(\varepsilon_i \| \boldsymbol{X}) = \sigma^2 \ \forall i$	• Some observations have more „weight" • Biased standard errors	• White Test • Breusch-Pagan Test
No Autocorrelation	$\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0 \ \ \forall i, j$	• Omitted variables • Functional misfit • Measurement errors	Durbin-Watson Statistic
Exogeneity	$\mathrm{E}(\varepsilon_i \| \boldsymbol{X}) = 0 \ \ \forall i$	• Omitted variables • Measurement errors	Instrument Variables

Under these assumptions, the OLS estimator is BLUE

Tutorial Business Analytics

Panel regression

- **Fixed Effects Model:**

$$y_{it} = (\beta_0 + \lambda_i) + \beta_1 x_{1it} + \beta_2 x_{2it} + \ldots + \beta_p x_{pit} + \varepsilon_{it}$$

- **Random Effects Model:**

$$y_{it} = \beta_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \ldots + \beta_p x_{pit} + \lambda_i + u_{it}$$

- **Lagrange Multiplier Test:** Test of individual effects for panel models

  H_0: No individual effects

- **Hausman Test:** Test of fixed effects vs. random effects

  H_0: Random effects estimator is consistent and efficient

DSS
Decision Sciences & Systems

Tutorial Business Analytics

Generalized Linear Models

- GLMs are a general class of linear models
- Consist of three components:

- **Random:** Identifies dependent variable μ and probability distribution
- **Systematic:** Identifies the set of explanatory variables $(X_1, \dots, X_k)$
- **Link function:** Identifies function of μ that is linear

$$g(\mu) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

Example: Linear regression uses identity link ($g(\mu) = \mu$)

Question: Which link function could be useful for a binary dependent variable?

Tutorial Business Analytics

From Logistic Function to Logit

Logistic Function: $p(x_i) = \frac{e^{x_i{}'\beta}}{1+e^{x_i{}'\beta}}$

transform …

Logit: $\ln(\frac{p(x_i)}{1-p(x_i)}) = x_i{}'\beta$

$\Leftrightarrow$ $\boldsymbol{\frac{p(x_i)}{1-p(x_i)}} = e^{x_i{}'\beta}$ **odds**

Logistic Regression: $\ln(\frac{p(x_i)}{1-p(x_i)}) = x_i{}'\beta + \varepsilon_i$

Tutorial Business Analytics

Interpreting the coefficient of logistic regression

$x_{ij} \epsilon x_i$: $\ln(\frac{p(x_i)}{1-p(x_i)}) = x_i{}'\beta$

$(x_{ij}+1) \epsilon \tilde{x}_i$: $\ln(\frac{p(\tilde{x}_i)}{1-p(\tilde{x}_i)}) = \tilde{x}_i{}'\beta$

$$\ln(\frac{p(\tilde{x}_i)}{1-p(\tilde{x}_i)}) - \ln(\frac{p(x_i)}{1-p(x_i)}) = \tilde{x}_i{}'\beta - x_i{}'\beta = \beta_j$$

$$\Leftrightarrow \qquad \beta_j = \ln(\frac{\frac{p(\tilde{x}_i)}{1-p(\tilde{x}_i)}}{\frac{p(x)}{1-p(x)}})$$

$$\Leftrightarrow \qquad e^{\beta_j} = \frac{\frac{p(\tilde{x}_i)}{1-p(\tilde{x}_i)}}{\frac{p(x_i)}{1-p(x_i)}}$$

odds ratio

Tutorial Business Analytics

Summary: Interpreting the coefficient of logistic regression

Effect of change in x_{ij}: on **log-odds (A)**, **odds (B)** and **probability (C)**

$$\Delta x_{ij} = 1 > 0$$

$$\Rightarrow \qquad \Delta\, ln\left(\frac{p(x_i)}{1-p(x_i)}\right) = \ln\left(\frac{p(\tilde{x}_i)}{1-p(\tilde{x}_i)}\right) - \ln\left(\frac{p(x_i)}{1-p(x_i)}\right) = \beta_j \qquad \textbf{(A)}$$

$$\Leftrightarrow \qquad e^{\beta_j} = \frac{\frac{p(\tilde{x}_i)}{1-p(\tilde{x}_i)}}{\frac{p(x_i)}{1-p(x_i)}} \qquad \textbf{(B), (C)}$$

$\boldsymbol{\beta_j}$	$\boldsymbol{ln(\frac{p}{1-p})}$ **(A)**	$\boldsymbol{\frac{p}{1-p}}$ **(B)**	$\boldsymbol{p}$ **(C)**
$\beta_j > 0$	increases by β_j	increases by a factor of e^{β_j}	Magnitude of increase unknown
$\beta_j < 0$	decreases by β_j	decreases by a factor of e^{β_j}	Magnitude of decrease unknown

Tutorial Business Analytics

From Incidence Rate to Link Function

Incidence Rate: $\mu(x) = e^{x_i{}'\beta}$

transform …

Link Function: $\ln\big(\mu(x)\big) = x_i{}'\beta$

Poisson Regression: $\boxed{\ln\big(\mu(x)\big) = x_i{}'\beta + \varepsilon_i}$

DSS
Decision Sciences & Systems

Tutorial Business Analytics

Interpreting the coefficient of poisson regression

$x_{ij} \epsilon\, x_i$: $\quad \ln\big(\mu(x_i)\big) = {x_i}'\beta$

$\big(x_{ij} + 1\big) \epsilon\, \tilde{x}_i$: $\quad \ln\big(\mu(\tilde{x}_i)\big) = {\tilde{x}_i}'\beta$

$$\ln\big(\mu(\tilde{x}_i)\big) - \ln\big(\mu(x_i)\big) = {\tilde{x}_i}'\beta - {x_i}'\beta = \beta_j$$

$$\Leftrightarrow \qquad \beta_j = \ln\left(\frac{\mu(\tilde{x}_i)}{\mu(x_i)}\right)$$

$$\Leftrightarrow \qquad e^{\beta_j} = \frac{\mu(\tilde{x}_i)}{\mu(x_i)}$$

incidence rate ratio

Tutorial Business Analytics

Summary: Interpreting the coefficient of poisson regression

Effect of change in x_{ij}: on **log-incidence rate (A)**, **incidence rate (B)**

$$\Delta x_{ij} = 1 > 0$$

$$\Rightarrow \qquad \Delta \ln\big(\mu(x_i)\big) = \ln\big(\mu(\tilde{x}_i)\big) - \ln\big(\mu(x_i)\big) = \beta_j \qquad \textbf{(A)}$$

$$\Leftrightarrow \qquad e^{\beta_j} = \frac{\mu(\tilde{x}_i)}{\mu(x_i)} \qquad \textbf{(B)}$$

$\boldsymbol{\beta_j}$	$\boldsymbol{ln\big(\mu(x_i)\big)}$ **(A)**	$\boldsymbol{\mu(x_i)}$ **(B)**
$\beta_j > 0$	increases by β_j	**increases by a factor of** e^{β_j}
$\beta_j < 0$	decreases by β_j	**decreases by a factor of** e^{β_j}

DSS
Decision Sciences & Systems

Tutorial Business Analytics

Maximum Likelihood Estimation

Goal: Maximize the joint probability of observing the set of dependent variables of the random sample

- Logistic regression: $L = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i}$ with $p = \frac{e^{X\beta}}{1+e^{X\beta}}$
- Poisson regression: $L = \prod_{i=1}^{n} p$ with $p = \frac{e^{X\beta y}}{y!} e^{-e^{X\beta}}$

Use numerical algorithm to find the maximum → gradient ascent

$k\ =\ 1$, feasible start point $\beta^{(1)} \in \mathbb{R}^n$, parameter $\varepsilon > 0$

While ($\|\nabla L(\beta^{(k)})\| \geq \varepsilon$) {

- Choose step size $\alpha > 0$
- Set $\beta^{(k+1)} = \beta^{(k)} + \alpha^* \nabla L\big(\beta^{(k)}\big)$
- $k++$

}

Tutorial Business Analytics

Evaluation and Goodness-of-Fit

- Null deviance: $-2\ln\left(\mathrm{L(null)}\right)$
- Residual deviance: $-2\ \ln\left(\mathrm{L(fitted)}\right)$
- McFadden R²:

$$R^2_{McFadden} = 1 - \frac{LL(fitted)}{LL(null)}$$

- Likelihood ratio test: Does fitted model explain significantly more variance than null model?

$$\mathrm{D} = -2\ln\left(\frac{\mathrm{L(null)}}{\mathrm{L(fitted)}}\right) = -2\left(\mathrm{LL(null)} - \mathrm{LL(fitted)}\right)$$

- Wald test: Is a particular coefficient significant?

$$\mathrm{H_0}: \beta_i = 0$$