Linear models

Gabriela Ciołek

March 8, 2017

Student's t-test

1. We consider linear model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_j x_j + u$$

2. We use t-test in order to test hypotheses about a particular β_k
3. Remark: β_k are unknown features of the population and we will never know them with certainty. Nevertheless, we can hypothesize about the value of β_k and then use statistical inference to test our hypothesis.

Testing against one-sided alternatives

- We consider null hypothesis
$$H_0 : \beta_k = 0$$
- Intuition: since β_k measures the partial effect of x_k on y, H_0 means that once $x_1, x_2, \cdots, x_{k-1}, x_{k+1}, \cdots, x_j$ have been accounted for, $x_k$has no effect on the expected value of y
- We test H_0 against $H_1 : \beta_k > 0$.

t statistic

- The statistic we use to test H_0 is called the t statistic or the t ratio of $\hat{\beta}_k$ and is defined as
$$t_{\hat{\beta}_k} := \frac{\hat{\beta}_k}{se(\hat{\beta}_k)}.$$
- It is reasonable to use $t_{\hat{\beta}_k}$ to detect $\beta_j \neq 0$ since
  - $se(\hat{\beta}_k)$ is always positive
  - $t_{\hat{\beta}_k}$ has the same sign as $\hat{\beta}_k$
  - for a given value of $se(\hat{\beta}_k)$ a larger value of $\hat{\beta}_k$ leads to larger values of $t_{\hat{\beta}_k}$.

t-test

Few remarks:

- Since we are testing $H_0 : \beta_k = 0$ it is only natural to look at our unbiased estimator of β_k.
- In practice the point estimate $\hat{\beta}_k$ will be never exactly zero
- A sample value of $\hat{\beta}_k$ very far from zero provides evidence against H_0
- $t_{\hat{\beta}_k}$ measures how many estimated standard deviations $\hat{\beta}_k$ is away from zero
- Values of $t_{\hat{\beta}_k}$ sufficiently far from zero will result in rejection of H_0.
- Determining a rule for rejecting H_0 at a given significance level, that is a probability of rejecting H_0 when it is true, requires knowing the sample distribution of $t_{\hat{\beta}_k}$ which is t_{n-k-1}, where $k+1$ is a number of unknown parameters.

Choice of rejection rule

- Firstly, decide on a significance level or the probability of rejecting H_0 when it is in fact true
- For example: suppose we have decided on a 5% significance level. It means that we are willing to mistakenly reject H_0 when it is true 5%of time
- We are looking at suffiently large positive value of $t_{\hat{\beta}_k}$ in order to reject H_0.
- The definition sufficiently large with a 5% significance level is the 95th percentile in a t distribution with $n-k-1$ degrees of freedom, denote this by c.
- The rejection rule is that H_0 is rejected in favor of H_1 at the 5% significance level if

$$t_{\hat{\beta}_k} > c.$$

- By our choice of the critical value c, rejection of H_0 will occur for 5% of all random samples when H_0 is true.

Two-Sided alternatives

- In applications, it is common to test the null hypothesis $H_0 : \beta_k = 0$ against a two-sided alternative that is
$$H_1 : \beta_k \neq 0.$$
- When the alternative is two-sided, we are interested in the absolute value of the t statistic. The rejection rule for H_0 is
$$|t_{\hat{\beta}_k}| > c.$$
- In order to find c, we again specify a significance level, let say 5%. For a two-tailed test, c is chosen to make an area in each tail of the t distribution equal to 2.5%.
- p is the 97.5%th percentile in the t distribution with $n - k - 1$ degrees of freedom.
- Check, if $n - k - 1 = 25$, the critical value for a two-sided test is $c = 2.060$.

Two-sided Alternatives

- If H_0 is rejected in favor of H_1 at the 5% level, we say that x_k is statistically significant or statistically different from zero, at the 5% level.
- If H_0 is not rejected, we say that x_k is statistically insignificant at the 5% level.

Testing other hypotheses about β_j

- $H_0 : \beta_j = a_j$.
- the appropriate t statistic is

$$t = (\hat{\beta}_j - a_j)/se(\hat{\beta}_j).$$

- As before, t measures how many estimated standard deviations $\hat{\beta}_j$ is from the hypothesized value of β_j.
- The general statistic t is usefully written as

$$t = \frac{estimate - hypothesized\ value}{standard\ error}.$$

Computing p-values for t tests

- Rather than testing a different significance levels, it is more informative to answer the following question: Given the observed value of the t statistic, what is the smallest level at which the null hypothesis would be rejected?
- this level is known as p-value for the test.
- The p value for testing the null hypothesis $H_0 : \beta_j = 0$ against two-sided alternative is given by

$$\mathbb{P}(|T| > |t|),$$

  where for clarity we let T denote a t distributed random variable with $n - j - 1$ degrees of freedom and t is numerical value of the test statistic.
- the p-value is the probability of observing a t statistic as extreme as we did if the null hypothesis is true. That means that small p-values are evidence against null, large p-values provide little evidence against H_0.

Student's t-test with Matlab

- Explain $wage|educ,\ exper,\ tenure$
- load WAGE1.raw

$$y = wage1(:,1)$$

$$[n,k] = size(wage1)$$

$$X = [ones(n,1), wage1(:,[2,3,4])]$$

$$[n,k] = size(X)$$

Standard deviation

- $\beta = (X' \times X)^{-1} \times X' \times y$
- $u = y - X \times \beta$
- $sig2 = u' \times u/(n-4)$ because we have 3 variables and intercept
- $std = sqrt(diag(sig2 \times inv(X' \times X)))$

- $\beta =$
$$\begin{matrix} -2.8727 \\ 0.5990 \\ 0.0223 \\ 0.1693 \end{matrix}$$
- $std =$
$$\begin{matrix} 0.7290 \\ 0.0513 \\ 0.0121 \\ 0.0216 \end{matrix}$$

Exercise

- Download modul *jplv6*, decompress it in your working directory
- Test $H_0 : \beta_{exper} = 0$ (one-sided and two-sided test)
- Calculate the test statistic

$$t = \frac{\beta}{std}.$$

T-statistic

- Calculate: $t = \beta ./ std$
- $t =$

$$\begin{matrix} -3.9408 \\ 11.6795 \\ 1.8528 \\ 7.8204 \end{matrix}$$

- Formulate H_0 : 'One year of studies brings additional 60 centimes of hourly wage'
- Test H_0
- H_1 ? Positive or...? Remember of 95% percentile.

Critical values and p-values

- Command *tdis_inv*
- Calculate p-value for two tests
- Validation of H_0?
- Command *tdis_prb*.

Exercise

- Trace the histogram of u.
- What are the properties of distribution? It is normal distribution?

Exercise

- Perform a regression for the logarithm of the salary and calculate the parameter beta. Draw a new histogram of residuals.
- Detect 10 observations which outlie the most (5 the smallest, and 5 the largest), discard them, and recalculate beta.
- How to interpret wage changes for an additional year of education? Test $H_0 : \beta_{educ} = 0.1$.
- Reject or accept?
- Calculate the p value

- $\beta =$

$$\begin{matrix} 0.2844 \\ 0.0920 \\ 0.0041 \\ 0.0221 \end{matrix}$$

- $std =$

$$\begin{matrix} 0.1042 \\ 0.0073 \\ 0.0017 \\ 0.0031 \end{matrix}$$

- $t =$

$$\begin{matrix} 2.7292 \\ 12.5552 \\ 2.3914 \\ 7.1331 \end{matrix}$$

Testing hypotheses about a single linear combination of the parameters

- Testing hypotheses concerning two parameters $H_0 : \beta_i = \beta_k$
- For the most part, the alternative is one-sided $H_1 : \beta_i < \beta_k$
- t-statistic is of the following form

$$t = \frac{\hat{\beta}_k - \hat{\beta}_i}{se(\hat{\beta}_k - \hat{\beta}_i)}$$

- Once we have the t statistic, testing proceeds as before. We choose significance level for the test, based on df obtain the critical value. Because of the form of H_1, the rejection rule is of the form $t < -c$. Or, we compute t statistic, and then the p-value.

Testing multiple linear restrictions: the F test

- We wish to test multiple hypotheses about underlying parameters $\beta_1, \cdots, \beta_k$. We want to test whether a set of independent variables has no partial effect on a dependent variable.
- Unresticted model with k independent variables

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

  Number of parameteres in unrestiricted model is $k + 1$.
- The null hypothesis is stated as

$$H_0 : \beta_{k-q+1} = 0, \cdots, \beta_k = 0,$$

  which puts q restrictions on the model.
- $H_1 : H_0$ is not true.
- Restricted model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{k-q} x_{k-q} + u.$$

F test

- $$F := \frac{(SSS\_r - SSR\_ur)/q}{SSR\_ur/(n-k-1)},$$
  where SSR_r is the sum of squared residuals from the resticted model and SSR_ur is the sum of squared residuals from the unrestricted model.
- $$q = \quad \textit{numerator degrees of freedom} \quad = df\_r - df\_ur$$
- $$n-k-1 = \quad \textit{denominator degrees of freedom} \quad = df\_ur$$
- One can show that under H_0, F is distributed as a F random variable with $(q, n-k-1)$ degrees of freedom
- We will reject H_0 in favor of H_1 when F is 'sufficiently' large (how large it depends of chosen significance level). The critical value depends on q and $n-k-1$

F test

- Once c has been obtained, we reject H_0 in favor of H_1 at the chosen significance level if
$$F > c.$$
- If H_0 is rejected, than we say that $x_{k-q+1}, \cdots, x_k$ are jointly statistically significant at the appropriate significance level.
- Remark:The test alone does not allow us to say which of the variables has a partial effect on y; they may all affect y or maybe just only one affects y. If H_0 is not rejected, then the variables are jointly insignificant

The R-squared Form of the F Statistic

- In most applications it is convenient to use a form of the F statistic that can be computed using the R-squareds from the restricted and unrestricted models. The reason is that R-squared is always between zero and one
- R-squared form of the F statistic

$$F := \frac{(R_{ur}^2 - R_r^2)/q}{(1 - R_{ur}^2)/(n - k - 1)}$$

Computing p-values for F Tests

- For reporting the outcomes of F tests, p-values are especially useful. Since the F distribution depends on the numerator and denominator df, it is difficult to get a feel for how strong or weak the evidence is against the null hypothesis simply by looking at the value of the F statistic and one or two critical values. In the F testing context, the p-value is defined as

$$p - value \quad = \mathbb{P}(\mathcal{F} > F),$$

  where $\mathcal{F}$ is an F random variable with $(q, n-k-1)$ degrees of freedom, and F is actual value of the test statistic.
- The same interpretation as it did for t statistics: it is the probability of observing a value of the F at least as large as we did, given that the null hypothesis is true. A small p-value is evidence against H_0.

Joint hypothesis testing with matlab

- Effect (on a wage) of education = effect of professional experience
- Test $H_0 : \beta_{educ} = \beta_{exper}$
- Create a variable capitaltot = educ + exper
- $Log(wage)|educ,\ capitaltot,\ tenure$
- Test the nullity of the coefficient associated with capitaltot

Construction of the test variable

- Calculate educ+exper
- $$test = X(:,2) + X(:,3);$$
  $$X = [X(:,[1,2,4]), test];$$

- $\beta =$

$$\begin{matrix} 0.2844 \\ 0.0879 \\ 0.0221 \\ 0.0041 \end{matrix}$$

- $std =$

$$\begin{matrix} 0.1042 \\ 0.0070 \\ 0.0031 \\ 0.0017 \end{matrix}$$

- $t =$

$$\begin{matrix} 2.7292 \\ 12.5880 \\ 7.1331 \\ 2.3914 \end{matrix}$$

F test

- Test $H_0 : \beta_{educ} = 0, \beta_{exper} = 0$
- 2 restrictions
- Estimate the unrestricted model

$$log(wage)|educ, \; exper, \; tenure,$$

  calculate SSR0
- Estimate the restricted model

$$log(wage)|tenure,$$

  calculer SSR1
- Calculate F

Unconstrained model

- Sums of squarres of errors $SSR0 = u'u$
- $SSR0 = 101.4556$

Constrained model

- Remove variables educ and exper
- $X = X(:, [1, 4]);$
- $SSR1 = 132.6105$

F test

- Compare sum of squares of deviations of 2 models (constrained and unconstrained)
- $$F = ((SSR1 - SSR0)/SSR0)(n-k)/2 = 80.1478 > F_{2,n-4}$$
- We reject H_0.
- $fdis\_prb(F, 2, n-k)$

Restricted model

- We test $\beta(educ) = 0$
- Using the restricted model $SSR2 = 132.09$

F test

- Compare sum of squares of deviations of two models (restricted and unrecsticted)
- $$F = ((SSR2 - SSR0)/SSR0)(n-k)/1 => F_{1,522}$$
- We reject H_0
- Remark $F_{1,522} = t^2_{522}$

Binary observations

Qualitative factors often come in the form of binary information: a person is female or male; a person does or does not own a personal computer; a firm offers a certain kind of employee pension plan or it does not; a state administers capital punishment or it does not. In all of these examples, the relevant information can be captured by defining a binary variable or a zero-one variable. In econometrics, binary variables are most commonly called dummy variables, although this name is not especially descriptive.

Example

Consider the following simple model of hourly wage determination:

$$wage = \beta_0 + \delta_0 female0 + \beta_1 educ + u.$$

We use δ_0 as the parameter on female in order to highlight the interpretation of the parameters multiplying dummy variables; later, we will use whatever notation is most convenient. In above model, only two observed factors affect wage: gender and education. Since $female = 1$ when the person is female, and $female = 0$ when the person is male, the parameter δ_0 has the following interpretation: δ_0 is the difference in hourly wage between females and males, given the same amount of education (and the same error term u). Thus, the coefficient δ_0 determines whether there is discrimination against women: if $\delta_0 < 0$, then, for the same level of other factors, women earn less than men on average.

Binary observations

- load wage1.raw
- Carry out regression:

$$Log(wage)|female,\ married,\ educ,\ exper,\ tenure$$

- Test $\beta_{female} = 0$

- $y = wage1(:, 1);$
- $[n, k] = size(wage1);$
- $X = [ones(n, 1), wage1(:, [2, 3, 4, 6, 7])];$
- $[n, k] = size(X)$
- $y = log(y);$
- $beta = inv(X' * X) * X' * y$
- $u = y - X * beta;$
- $sig2 = u' * u/(n - k)$
- $std = sqrt(diag(sig2 * inv(X' * X)))$
- $t = beta./std$

Results

- $\beta =$

$$\begin{matrix} 0.4901 \\ 0.0839 \\ 0.0031 \\ 0.0169 \\ -0.2855 \\ 0.1257 \end{matrix}$$

- std $=$

$$\begin{matrix} 0.1011 \\ 0.0070 \\ 0.0017 \\ 0.0030 \\ 0.0373 \end{matrix}$$

Interactions

- Sometimes it is natural for the partial effect, elasticity, or semi-elasticity of the dependent variable with respect to an explanatory variable to depend on the magnitude of yet another explanatory variable.
- Consider

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_j x_i x_j + \cdots + \beta_i x_i + \cdots + \beta_k x_k$$

- the partial effect of x_i on y is

$$\frac{\Delta(y)}{\Delta(x_i)} = \beta_i + \beta_j x_j$$

- If $\beta_j > 0$, then, there is an interaction effect between x_i and x_j.
- We want to test if $\beta_j = 0$.

Interactions

- Create :

$$marrmale = (1 - female) * married$$

$$marrfem = female * married$$

$$-singfem = female * (1 - married)$$

- Carry out regression:

$$log(wage)|marrmale, marrfem, singfem, educ, exper, tenure$$

- Test the effect of being a women

Creation of variables

$$educ = X(:,2);$$

$$exper = X(:,3);$$

$$tenure = X(:,4);$$

$$female = X(:,5);$$

$$married = X(:,6);$$

$$marrmale = (1 - female). * married;$$

$$marrfem = female. * married;$$

$$singfem = female. * (1 - married);$$

Regression

$$X = [ones(n,1)|educ,\ exper,\ tenure,\ marrmale,\ marrfem,\ singfem];$$

$$[n,k] = size(X)$$

Then we do the regression with the new matrix.

Results

- $\beta =$

$$\begin{matrix} 0.3878 \\ 0.0835 \\ 0.0032 \\ 0.0157 \\ 0.2921 \\ -0.1202 \\ -0.0967 \end{matrix}$$

- $std =$

$$\begin{matrix} 0.1022 \\ 0.0069 \\ 0.0017 \\ 0.0029 \\ 0.0553 \\ 0.0579 \\ 0.0574 \end{matrix}$$

- $t =$

$$\begin{array}{c} 3.7938 \\ 12.1870 \\ 1.9136 \\ 5.3747 \\ 5.2785 \\ -2.0756 \\ -1.6853 \end{array}$$

Restricted model

$$SSR0 = u' * u$$

$$SSR0 = 85.4648$$

$$X = [ones(n,1)|educ,\ exper,\ tenure,\ marrmale];$$

$$SSR1 = u' * u$$

$$SSR1 = 98.46$$

F test

$$F = ((SSR1 - SSR0)/SSR0) * (n - k)/2$$

$$F = 39.57$$

$$fdist_{prob}(F, n - k, 2)$$

Interactions

- Interaction between female and educ
- Calculate $femeduc = female * educ$
- Perform the regression:

$$log(wage)|female, educ, femeduc, exper, tenure$$

- Test the significance of effect of being a woman

Unrestricted model

$$femeduc = female. * educ;$$

$$X = [ones(n, 1)|educ,\ exper,\ tenure,\ female,\ femeduc];$$

- $\beta =$

$$\begin{matrix} 0.4647 \\ 0.0903 \\ 0.0046 \\ 0.0174 \\ -0.2104 \\ -0.0072 \end{matrix}$$

- $std =$

$$\begin{matrix} 0.1228 \\ 0.0087 \\ 0.0016 \\ 0.0030 \\ 0.1738 \\ 0.0135 \end{matrix}$$

- $t =$

$$\begin{matrix} 3.7851 \\ 10.3685 \\ 2.8530 \\ 5.8545 \\ -1.2104 \\ -0.5347 \end{matrix}$$

$$SSR0 = 90.0950$$

$$SSR1 = 101.4556$$

$$F = 32.7848$$

Logarithmic transformation

- Use WAGE1.RAW
- Perform regression

$$Log(wage)|educ, exper, tenure, educ * fem, exper * fem, tenure * fem, fem$$

- Test the coefficients associated with $women = 0$

Logarithmic transformation

- Use WAGE1.RAW
- $Log(wage)|educ, exper, tenure$
- Test the stability of coefficients men/women
- unrestricted model: $SSR0 = 88.6$
- restricted model: $SSR1 = 101.3$

$$F = (101.3 - 88.6)/88.6/(4/(526 - 8) = 18.8$$

$$Fdis_{prb}(18.8, 4, 516) < .01$$

2 separate regressions

- Create sample consisting with men and women separately (see: TP Matlab 1)
- Do two regressions and add sum of squares of deviations : sum of squares of deviations of unrestricted model.

$$Log(wage)|educ, exper, tenure$$

  for men, SSR01

$$Log(wage)|educ, exper, tenure,$$

  for women, SSR02
- $SSR0 = SSR01 + SSR02 = 88.6$