Linear models

Gabriela Ciołek

February 9, 2018

Transformations

We consider linear model

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

Some non-linear re-expression of the dependent variable is indicated when any of the following apply:

- The residuals have a skewed distribution. The purpose of a transformation is to obtain residuals that are approximately symmetrically distributed (about zero, of course).
- The spread of the residuals changes systematically with the values of the dependent variable ("heteroscedasticity"). The purpose of the transformation is to remove that systematic change in spread, achieving approximate "homoscedasticity."
- To linearize a relationship.

- When scientific theory indicates. For example, chemistry often suggests expressing concentrations as logarithms (giving activities or even the well-known pH).
- When a more nebulous statistical theory suggests the residuals reflect "random errors" that do not accumulate additively.
- To simplify a model. For example, sometimes a logarithm can simplify the number and complexity of "interaction" terms.

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_j.
- In practice the point estimate $\hat{\beta}_k$ will be very rarely equal to 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-j-1}, where $j+1$ 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 $H_0 : \beta_j = 0$ against a two-sided alternative that is
$$H_1 : \beta_j \neq 0.$$
- When the alternative is two-sided, we are interested in the absolute value of the t statistic. The rejection rule for H_0 is
$$|t_{\hat{\beta}_j}| > c.$$
- 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_j is statistically significant or statistically different from zero, at the 5% level.
- If H_0 is not rejected, we say that x_j is statistically insignificant at the 5% level.

Testing other hypotheses about β_j

- $H_0 : \beta_j = a_j$.
- the apprioriate 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

$$t = \frac{estimate - hypothesizedvalue}{standarderror}.$$

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 $n - k - 1$ 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.

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

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

- 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.

Student's t-test with Matlab

- Explain wage — educ, exper, tenure
- load WAGE1.raw

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

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

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

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

Calculate:

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

Draw histograms:

- for u in a model $y_{wage} = \beta_{educ} x_{educ} + \beta_{exper} x_{exper} + \beta_{tenur} x_{tenur} + u$.
- for u in a model $log(y)$.

Compare the histograms. Which model do you choose?

Code for the model with *log* transformation

$$logy = log(y)$$

$$beta = inv(X' \times X)X' \times logy$$

$$u = logy - X \times beta$$

$$f = figure;$$

$$hist(u, 20)$$

title('Distribution of residuals of the transformed model')

xlabel('Residuals')

ylabel ('Frequency')

- Download modul *jplv*6, decompress it in your working directory
- Test $H_0 : \beta_{exper} = 0$. (one-sided and two-sided test)
- Calculate the test statistic
$$t = \frac{\beta}{std}.$$
- take 5% significance level

Attention: degree of freedom in the model is $n - 4$, thus when computing variance and standard variation we take:

$$sig2 = u' \times u/(n-4)$$

$$std_dv = sqrt(diag(sig2 \times inv(X' \times X)))$$

Calculate: $t = \beta./std_dv$

$$t = t(3)$$

Calculate a critical value c which is a 95% percentile of t_{n-k-1} distribution. Degree of freedom is 4 (we have 3 variables + intercept). Thus:

$$k_0 = 4$$

$$c1 = tdis\_inv(0.95, n - k)$$

Do we reject or accept H_0?

Perform a two-sided test. Do we reject or accept H_0 in this case?
Reminder: the rejection rule for H_0 is

$$|t_{\hat{\beta}_j}| > c.$$

- Formulate H_0 : 'One year of studies brings additional 60 centimes of hourly wage'
- Test $H_0 : \beta_{educ} = 0.6$
- Calculate t statistitic, and critical value. Take the 5% significance level.

Do we reject or accept H_0?

- Test $H_0 : \beta_{exper} = 0.6$
- Perform one-sided and two-sided test:
- Calculate the value of the test statistic
- Using the known distribution of the test statistic, calculate the p-value: "If the null hypothesis is true, what is the probability that we'd observe a more extreme test statistic in the direction of the alternative hypothesis than we did?"
  Code to calculate p-value $tdis\_prb(t, n - k)$
- Set the significance level of $\alpha = 5\%$

Compare *p*-value with α. Do we reject or accept H_0? Consider one-sided and two-sided tests.
Reminder: if the *p*-value is less than (or equal to) α, then the null hypothesis is rejected in favor of the alternative hypothesis. And, if the *p*-value is greater than α, then the null hypothesis is not rejected.

- Test:

$$\beta_{educ} = \beta_{exper}$$

  Under this hypothesis, one extra year of studies has the same effect on the salary as one more year of experience

- Define a new parameter

$$\theta = \beta_{educ} - \beta_{exper}$$

- Consider $H_0 : \theta = 0$ versus $H_1 = \theta < 0$.
- Create a variable $capitaltot = educ + exper$
- Do a regression

$$log(wage) = \beta_0 + \theta educ + \beta_2(capitaltot) + \beta_3 tenure + u$$

- Test the coeffcient associated with variable $educ$.