Question

Do a regression for the model

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u$$

Code:

$$load\ \ WAGE1.raw$$

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

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

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

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

Estimation of model's parameters:

$$\beta = inv(X' * X) * X' * y$$

- Variance of residuals:

  $$\hat{\sigma}^2 = \frac{\hat{u}'\hat{u}}{n - k - 1},$$

  where $n - k - 1$ is the number of degrees of freedom of the model

- Variance of the estimator:

  $$Var(\hat{\beta}|x) = \hat{\sigma}^2 (x'x)^{-1}$$

Code:

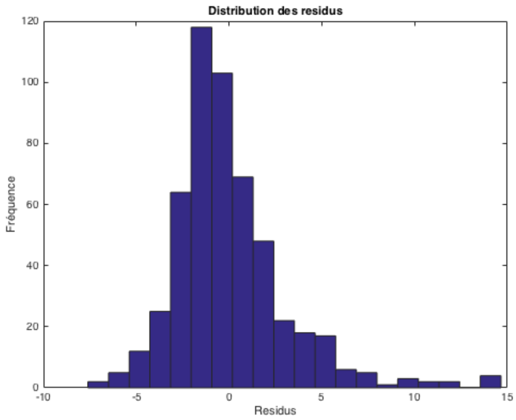
$$u = y - X * \beta$$

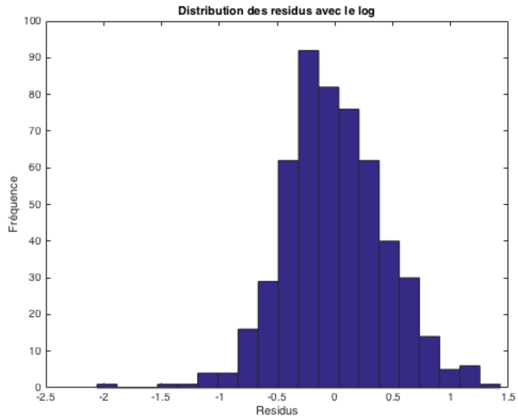
$$sig2 = u' * u/(n - 4)$$

$$std = sqrt(diag(sig2 * inv(X' * X)))$$

Question

Draw histograms for $wage$ model and for $log(wage)$.





The second distribution is less skewed. The logarithimic transformation is recommended when the residuals have a "strongly" positively skewed distribution.
Code:

$$f = figure;$$

$$hist(u, 20)$$

$$title('Distribution\ of\ residuals')$$

$$xlabel('Residals')$$

$$ylabel('Frequency')$$

Regression with logarithmic transformation:

$$logy = log(y)$$

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

$$u = logy - X * beta$$

$$f = figure;$$

$$hist(u, 20)$$

$$title('Distribution\ of\ the\ residuals')$$

$$xlabel('Residuals')$$

$$ylabel('Frequency')$$

Hypothesis testing

Redo regression:

$$log(y) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u$$

Code:

$$y = wage1(:, 1);$$
$$logy = log(y)$$
$$[n, k] = size(wage1)$$
$$X = [ones(n, 1), wage1(:, [2, 3, 4])];$$
$$[n, k] = size(X)$$
$$\beta = inv(X' * X) * X' * logy$$
$$u = logy - X * \beta$$
$$sig2 = u' * u/(n - 4)$$
$$std\_dv = sqrt(diag(sig2 * inv(X' * X)))$$

- **MLR1**We consider linear equation:

  $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

  .

- **MLR2**: We have a sample consisting of n observations $\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i) : i = 1, 2, \ldots, n)$
- **MLR3**: $\mathbb{E}[u|x_1, \ldots, x_k] = 0$.
- **MLR4**: In the sample (and thus in the population), none of the independent variables is constant, and there is no relationship between the independent variables.
- **MLR5** (homoskedascity): $Var(u|x_1, x_2, \ldots, x_k) = \sigma^2$.
- **MLR6**:Error for the population u is independent of the explanatory variables $x_1, \ldots, x_k$ and is distributed according to normal distribution with zero mean and variance σ^2: $u \sim \mathcal{N}(0, \sigma^2)$.

Theorem 0.1 *Under the hypotheses* $1-4$, $\mathbb{E}[\hat{\beta}_j] = \beta_j$ *for* $j = 0, 1, \ldots, k$.

Theorem 0.2 *Under the hypotheses 1-5,* $Var(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1-R_j^2)}$ *for* $j = 0, 1, \ldots, k$, *with* $SST_j = \sum_{i=1}^{n}(x_{ij} - \bar{x_j})$*- the total variation of the variable* j *in the sample, and* R_j^2 *-the R-square of the regression of* x_j *for the other independent variables.*

Theorem 0.3 *Under the hypotheses* $1-6$, $(\hat{\beta}_j - \beta_j)/std(\hat{\beta}_j) \sim t_{n-k-1}$, *where* $n - k - 1$ *is the number of degrees of freedom.*

Question

Test $H_0 : \beta_{exper} = 0$.
In most of applications, we test the null hypothesis $H_0 : \beta_j = 0$. Thus, it is natural to consider unbiased estimator β_j which we denote by $\hat{\beta}_j$. To test H_0 we use t-statistic $\hat{\beta}_j$ given by $t_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{std(\hat{\beta}_j)}$. Values of $t_{\hat{\beta}_j}$ too far from 0 result in rejecting H_0.

Remark 0.1 *Since $std(\hat{\beta}_j)$ is always positive, $t_{\hat{\beta}_j}$ has the same sign as $\hat{\beta}_j$.*

We test $H_0 : \beta_{exper} = 0$. Thus, we want to test the hypothesis that the number of years of professional experience does not affect the hourly wage. We calculate:

$$t = beta./std_{dv}$$

$$t = t(3)$$

We obtain $t_{\hat{\beta}_{exper}} = 2.39$. According to the theorem, if H_0 is true, $t_{\hat{\beta}_j}$ has the t distribution with $n-k-1$ degrees of freedom: $t_{\hat{\beta}_j} \sim t_{n-k-1}$.

One-sided test

In order to decide if we should accept or reject H_0, we consider the alternative hypothesis H_1 of the following form $H_1 : \beta_j > 0$. This means that we are not concerned with alternatives of H_0 of the form $H_1 : \beta_j < 0$, for intuitive reasons, or coming from economic theory, for example. In order to test H_0 we need to choose significance level which is the probability of rejecting H_0 when it is in fact true.

We reject H_0 in favor of H_1 at the significance level 5% if: $t_{\hat{\beta}_j} > c$ where c is called the *critical value*, is the 95% percentile of t_{n-k-1} distribution. In order to obtain c, we need to calculate the number of degrees of freedom, here $n-k-1 = n-4$ because we have 3 variables and intercept.

$$k0 = 4$$

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

We obtain $c = 1.6478$. Thus, $t_{\hat{\beta}_{exper}} > c$ so we reject H_0 with 5% . We reject H_0 also with 1% significance level since in this case $c = 2.3335$ and $t_{\hat{\beta}_{exper}}$ is greater than c.

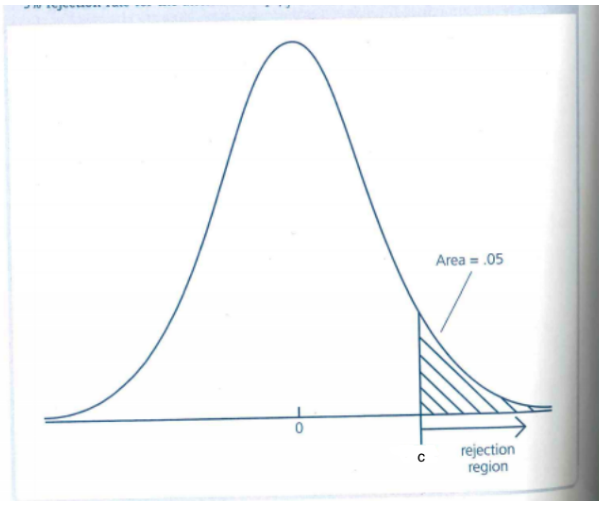


Figure 1: One-sided test - Distribution t_{n-k-1}, 95% percentile c and rejection region.

Two-sided test

We test H_0 against the two-sided alternative $H_1 : \beta_j \neq 0$, if the sign of β_j is not obvious or if we want to be more careful. In this case, the rule of rejecting H_0 is of the form : $|t_{\hat{\beta}_j}| > c$. As before, we choose the significance level, for example for a significance level 5%, c must be chosen such that the area of the two distribution tails is equal to 2.5.

Code:

$$c2 = tdis\_inv(0.975, n - k0)$$

We obtain c= 1.9645 so we have $|t_{\beta_{exper}}| = 2.3914 > c$, we reject H_0 in favor of H_1 with 5%.

Question

Test $H_0 : \beta_{educ} = 0.6$.
We will test H_0 : One year of studies increases by 60 centimes the hourly wage. It is sufficient to modify the t statistic:

$$t = (beta - 0.6)./std_{dv}$$

$$t = t(2)$$

$$c = tdis\_inv(0.95, n - k0)$$

We obtain $t_{\hat{\beta}_{educ}} = -69.3010$ and c=1.6478 , thus we do not reject H_0.

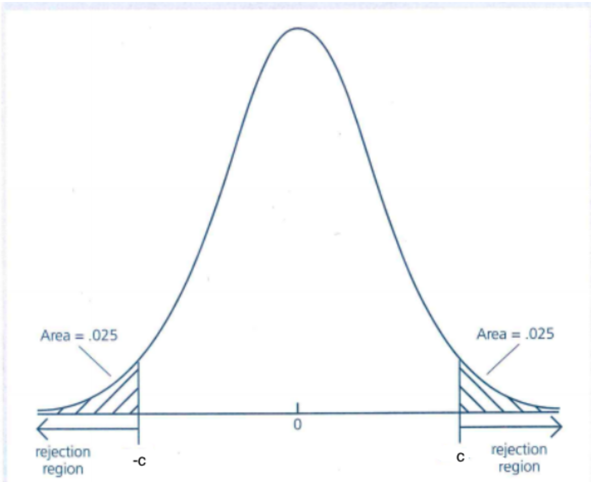


Figure 2: Two-sided test

Question

Test $H_0 : \beta_{educ} = 0.6$. 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? We call this level *p-value* for the test.

In case of two-sided test $H_0 : \beta_j = 0$ against alternative, the p value is given by $\mathbb{P}[|T| > |t|]$, where $\mathbb{P}[T > a]$ is the area under the right curve of the value a. Code to obtain p-value

$$tdis\_prb(t, n - k0)$$

We find that p= 0 so we reject stongly H_0.

Question

Test $H_0 : \beta_{educ} = \beta_{exper}$.
We want to test that a year of additional education has the same effect as a year of additional experience.
We define a new variable $\theta : \beta_{educ} - \beta_{exper}$.
We test $H_0 : \theta = 0$ versus $H_1 : \theta < 0$. Code:

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

$$logy = log(y)$$

$$test = X(:, 2) + X(:, 3)$$

$$X = [X(:,[1,2,4]), test];$$
$$[n,k] = size(X)$$
$$beta = inv(X' * X) * X' * logy$$
$$u = logy - X * beta$$
$$sig2 = u' * u/(n-4)$$
$$std_dv = sqrt(diag(sig2 * inv(X' * X)))$$

$$t = (beta)./std_dv$$
$$t = t(2)$$
$$p = tdis\_prb(t, n-k0).$$

The p-value is 0, thus we reject H_0.