- <u>EXACT DISTRIBUTION</u> of $\hat{B}_1$ and $\hat{B}_2$

> **Preliminary result**
> Given $Y_1, \dots, Y_n$ independent with distribution $Y_i \sim N(\mu_i, \sigma^2)$ $i=1,\dots,n$
> and a sequence of known constants a_i, $i=1,\dots,n$,
> $$\sum_{i=1}^{n} a_i Y_i \sim N\left(\sum_{i=1}^{n} a_i \mu_i ,\ \sigma^2 \sum_{i=1}^{n} a_i^2 \right)$$

We have seen that $\hat{B}_1$ and $\hat{B}_2$ are linear combinations of $Y_1, \dots, Y_n$ of the form

$$\hat{B}_1 = \sum_{i=1}^{n} v_i Y_i \qquad \hat{B}_2 = \sum_{i=1}^{n} w_i Y_i$$

hence $\hat{B}_1$ and $\hat{B}_2$ are <u>exactly</u> Gaussian-distributed r.v. (see res. 1)

Moreover, the expression of the two estimators are the same we obtained with OLS.
In fact, the Gaussian linear model is a special case. Hence the properties we computed still hold.
In particular, we computed

$$\mathbb{E}[\hat{B}_1] = \beta_1 \qquad var(\hat{B}_1) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \right)$$

$$\mathbb{E}[\hat{B}_2] = \beta_2 \qquad var(\hat{B}_2) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

The exact distributions are then easily obtained as

$$\hat{B}_1 \sim N\left(\beta_1 ;\ \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \right) \right)$$

$$\hat{B}_2 \sim N\left(\beta_2 ;\ \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \right)$$

- <u>EXACT DISTRIBUTION</u> of $\hat{\Sigma}^2$

$$\hat{\Sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{B}_1 - \hat{B}_2 x_i)^2$$

it is possible to show that $\frac{n\hat{\Sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$ Chi-squared with n-2 degrees of freedom

In general, for a χ^2_v r.v., the expected value is v

$$\mathbb{E}\left[\frac{n\hat{\Sigma}^2}{\sigma^2}\right] = (n-2) \Rightarrow \mathbb{E}[\hat{\Sigma}^2] = \frac{(n-2)}{n}\sigma^2 \quad \text{biased}$$

hence again we obtain an unbiesed estimator as

$$S^2 = \frac{n}{n-2}\hat{\Sigma}^2 \qquad \mathbb{E}[S^2] = \frac{n}{n-2}\mathbb{E}[\hat{\Sigma}^2] = \frac{n}{n-2}\cdot\frac{n-2}{n}\sigma^2 = \sigma^2$$

and $\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}$.

Moreover, it is possible to show that $\hat{\Sigma}^2 \perp\!\!\!\perp (\hat{B}_1, \hat{B}_2)$
(hence also $S^2 \perp\!\!\!\perp (\hat{B}_1, \hat{B}_2)$)

<u>INFERENCE ABOUT β</u>

We have derived the exact distributions of the estimators.
With these distributions we can test statistical hypotheses, compute confidence intervals.

Examples

Test: $\begin{cases} H_0: \beta_j = b \\ H_1: \beta_j \neq b \end{cases}$ $\begin{cases} H_0: \beta_j = 0 \\ H_1: \beta_j > 0 \end{cases}$ $j = 1,2$

Confidence interval, of level $1-\alpha$: $\hat{C}_j$ such that $\mathbb{P}(\hat{C}_j \ni \beta_j) = 1-\alpha \quad \forall \beta_j \in \mathbb{R}$

Recall that: $\hat{B}_1 \sim N(\beta_1, V(\hat{B}_1))$
$\hat{B}_2 \sim N(\beta_2, V(\hat{B}_2))$
$\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}$

where $V(\hat{B}_1) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)$
$V(\hat{B}_2) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$

We need to find a pivotal quantity.
PIVOTAL QUANTITY : a transformation of the data (and of the parameter) whose distribution does not depend on the parameter (hence is completely known).

> **Preliminary result:**
> If $Z \sim N(0,1)$ and $W \sim \chi^2_v$ independent, then $\frac{Z}{\sqrt{W/v}} \sim t_v$.
> (Student's t with v degrees of freedom)
> ↳ · symmetric distrib.
> · heavier tails than a normal
> · for large v it is very close to a normal

Since $\hat{B}_j \sim N(\beta_j, V(\hat{B}_j))$, the simplest (and most intuitive) transformation is

$$\frac{\hat{B}_j - \beta_j}{\sqrt{V(\hat{B}_j)}} \sim N(0,1)$$

however, $V(\hat{B}_j)$ includes σ^2 which is unknown

In place of $V(\hat{B}_j)$ we use an estimate, $\hat{V}(\hat{B}_j) = \frac{S^2}{\sigma^2} V(\hat{B}_j)$ (eg. $\hat{V}(\hat{B}_2) = \frac{S^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$)

$$T_j = \frac{\hat{B}_j - \beta_j}{\sqrt{\hat{V}(\hat{B}_j)}}$$

what is its distribution? Notice that $\hat{V}(\hat{B}_j)$ includes $\hat{S}^2$ (transformation of Y)

$$T_j = \frac{\hat{B}_j - \beta_j}{\sqrt{\hat{V}(\hat{B}_j)}} = \frac{\dfrac{\hat{B}_j - \beta_j}{\sqrt{V(\hat{B}_j)}}}{\sqrt{\dfrac{\hat{V}(\hat{B}_j)}{V(\hat{B}_j)}}}$$

moreover, $\hat{B}_j \perp\!\!\!\perp S^2$

$$= \frac{\dfrac{\hat{B}_j - \beta_j}{\sqrt{V(\hat{B}_j)}}}{\sqrt{\dfrac{S^2}{\sigma^2}}} \qquad \sim N(0,1)$$

$$\sqrt{\frac{S^2}{\sigma^2}} = \sqrt{\frac{(n-2)S^2}{\sigma^2}\cdot\frac{1}{(n-2)}} \sim \sqrt{\frac{\chi^2_{n-2}}{n-2}}$$

$$\Rightarrow T_j \sim t_{n-2}$$

- <u>CONFIDENCE INTERVAL for β_j</u>

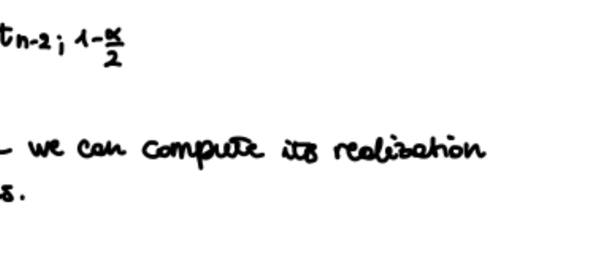
We want to find an interval (u,v) such that

$\mathbb{P}(u < T_j < v) = 1-\alpha$

and then "isolate" the parameter to find an interval for β_j

$T_j \sim t_{n-2}$ hence

$$\mathbb{P}\left(t_{n-2;\frac{\alpha}{2}} < T_j < \underbrace{t_{n-2;1-\frac{\alpha}{2}}}_{\text{quantile } 1-\frac{\alpha}{2} \text{ of a } t_{n-2} \text{ distrib.}} \right) = 1-\alpha$$



$t_{n-2;\frac{\alpha}{2}} = -t_{n-2;1-\frac{\alpha}{2}}$ for symmetry

$$\mathbb{P}\left(-t_{n-2;1-\frac{\alpha}{2}} < \frac{\hat{B}_j - \beta_j}{\sqrt{\hat{V}(\hat{B}_j)}} < t_{n-2;1-\frac{\alpha}{2}} \right) = 1-\alpha$$

$$\mathbb{P}\left(\hat{B}_j - \sqrt{\hat{V}(\hat{B}_j)}\cdot t_{n-2;1-\frac{\alpha}{2}} < \beta_j < \hat{B}_j + \sqrt{\hat{V}(\hat{B}_j)}\, t_{n-2;1-\frac{\alpha}{2}} \right) = 1-\alpha$$

$$\mathbb{P}(\beta_j \in \hat{C}) = 1-\alpha \qquad \text{where } \hat{C} = \hat{B}_j \pm \sqrt{\hat{V}(\hat{B}_j)}\, t_{n-2;1-\frac{\alpha}{2}}$$

$\hat{C}$ is a random interval. After observing the data we can compute its realisation by substituting the estimators with their estimates.

We obtain $\beta_j \in \hat{\beta}_j \pm t_{n-2,1-\frac{\alpha}{2}} \sqrt{\hat{V}(\hat{B}_j)}$.

That is

$$\beta_1 \in \hat{\beta}_1 \pm t_{n-2;1-\frac{\alpha}{2}} \sqrt{s^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)}$$

$$\beta_2 \in \hat{\beta}_2 \pm t_{n-2;1-\frac{\alpha}{2}} \sqrt{\frac{s^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

- <u>HYPOTHESIS TEST on β_j</u>

$$\begin{cases} H_0: \beta_j = b \\ H_1: \beta_j \neq b \end{cases}$$

Following the same reasoning as before, we use the TEST STATISTIC

$$T_j = \frac{\hat{B}_j - b}{\sqrt{\hat{V}(\hat{B}_j)}} \overset{H_0}{\sim} t_{n-2}$$

(under H_0 b is the true value of the parameter, so we are subtracting the true mean of B_j)

T_j is a random variable. After observing $y_1, \dots, y_n$ we can compute the OBSERVED VALUE OF THE TEST t^{obs}.

How do we define the acceptance and reject regions?

(1) what values of the test do we expect when H_0 is true? And what values do we expect instead when H_0 is not true (under H_1)?

If H_0 true: $\beta_j = b$.
If the data support this hypothesis, then the estimate $\hat{\beta}_j$ will be close to b ($\mathbb{E}[\hat{B}_j] = \beta_j = b$).
$\Rightarrow \hat{\beta}_j - b \approx 0 \Rightarrow t_j^{obs} \approx 0$
Hence we expect that, if H_0 is true, t^{obs} will be small (in absolute value)

If H_0 is not true, then $\beta_j \neq b$. The estimate $\hat{\beta}_j$ will be different from b
$\Rightarrow |\hat{\beta}_j - b|$ large $\Rightarrow |t_j^{obs}|$ large
Hence we expect that, under H_1, t_j^{obs} will be large (in absolute value)

(2) The acceptance region thus will contain the values around 0 $(-a,+a) = A$
The reject region will contain values far from 0 $(-\infty,-a) \cup (a,+\infty) = R$
We need to define the thresholds $-a, a$

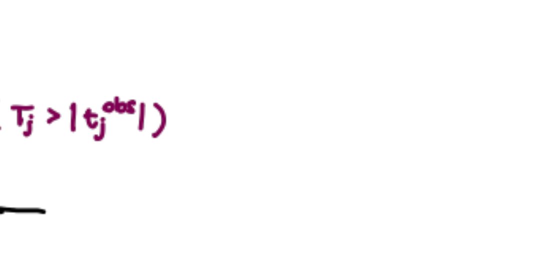
(3a) fixed significance level α: $\alpha = \mathbb{P}(\text{reject } H_0 \mid H_0 \text{ true})$

$\mathbb{P}_{H_0}(|T_j| > t_{n-2;1-\frac{\alpha}{2}}) = \alpha$

the acceptance region is $A = (t_{n-2;\frac{\alpha}{2}},\ t_{n-2,1-\frac{\alpha}{2}})$

the reject region is $R = R_1 \cup R_2 = (-\infty; t_{n-2;\frac{\alpha}{2}}) \cup (t_{n-2;1-\frac{\alpha}{2}}; +\infty)$

if $t_j^{obs} \in A \Rightarrow$ we do not reject H_0
if $t_j^{obs} \notin A \Rightarrow$ we reject H_0



(3b) p-value

it is the probability of observing "more extreme" values than t_j^{obs}

$$\alpha^{obs} = 2\min\{ \mathbb{P}_{H_0}(T \geq t_j^{obs}) ; \mathbb{P}_{H_0}(T \leq t_j^{obs}) \}$$

the t distribution is symmetric, so

$$\alpha^{obs} = \mathbb{P}_{H_0}(|T_j| > |t_j^{obs}|) = 2\cdot\mathbb{P}_{H_0}(T_j > |t_j^{obs}|)$$

connection between the two types of test
- if $\alpha^{obs} < \alpha \Rightarrow$ reject H_0 at level α
- if $\alpha^{obs} > \alpha \Rightarrow$ do not reject H_0 at a level α

In practical applications, these methods are useful tools to investigate relevant applicative questions. For example:

· does the covariate x have a significant effect on Y?

The effect of x on Y is summarised by the coefficient β_2.
Hence this question can be formalized by the statistical test

$$\begin{cases} H_0: \beta_2 = 0 & \rightarrow \text{no effect} \\ H_1: \beta_2 \neq 0 \end{cases}$$

Indeed the model $Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$
under H_0 becomes $Y_i = \beta_1 + \varepsilon_i$ (x has no impact on Y)
↳ this is called the "NULL MODEL"