INFERENCE in the MULTIPLE LINEAR MODEL

we will work under the assumption that the model always includes the intercept $\underline{x}_1 = \underline{1}_n$ with $\beta_1$ the associated coefficient.

## 1. TEST about an individual coefficient $\beta_j$ $(j = 2, \dots, p)$

assume that we want to test a single coefficient:

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

In particular, we are often interested in testing the statistical significance of an individual coefficient

$$\begin{cases} H_0: \beta_j = 0 \\ H_1: \beta_j \neq 0 \end{cases}$$

Recall:
- $\hat{\underline{\beta}}(\underline{Y}) \sim N_p(\underline{\beta}, (X^TX)^{-1}\sigma^2)$
- the $j$-th element $\hat{\beta}_j(\underline{Y}) \sim N(\beta_j, \underbrace{\sigma^2[(X^TX)^{-1}]_{j,j}}_{V(\hat{\beta}_j)})$
- $\frac{n\,\hat{\sigma}^2(Y)}{\sigma^2} \sim \chi^2_{n-p}$
- $\frac{(n-p)S^2}{\sigma^2} \sim \chi^2_{n-p}$
- $\hat{\underline{\beta}}(\underline{Y}) \perp\!\!\!\perp \hat{\sigma}^2(\underline{Y})$ and $\hat{\underline{\beta}}(Y) \perp\!\!\!\perp S^2$

We need to define a pivotal quantity

$$\frac{\hat{\beta}_j(\underline{Y}) - b_j}{\sqrt{V(\hat{\beta}_j)}} \overset{H_0}{\sim} N(0,1)$$ but it depends on the unknown $\sigma^2$ (hence we can't use it)

we consider instead

$$T_j = \frac{\hat{\beta}_j - b_j}{\sqrt{\hat{V}(\hat{\beta}_j)}} = \frac{\hat{\beta}_j - b_j}{\sqrt{\frac{S^2}{\sigma^2}V(\hat{\beta}_j)}} = \frac{\dfrac{\hat{\beta}_j - b_j}{\sqrt{V(\hat{\beta}_j)}} \sim N(0,1)}{\sqrt{\frac{S^2}{\sigma^2}} \sim \sqrt{\frac{\chi^2_{n-p}}{(n-p)}}}$$

$$\hat{V}(\hat{\beta}_j) = S^2[(X^TX)^{-1}]_{jj} \cdot \frac{\sigma^2}{\sigma^2}$$
$$= \left(\sigma^2[(X^TX)^{-1}]_{jj}\right) \cdot \frac{S^2}{\sigma^2} = V(\hat{\beta}_j) \cdot \frac{S^2}{\sigma^2}$$
general expression

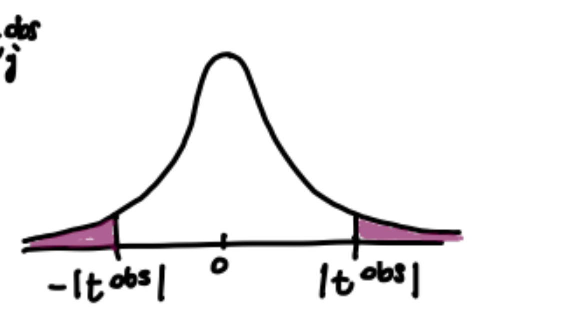
$$\Rightarrow \quad T_j = \frac{\hat{\beta}_j - b_j}{\sqrt{\hat{V}(\hat{\beta}_j)}} \overset{H_0}{\sim} t_{n-p}$$

with the data, I compute the observed value of the test $t_j^{obs}$

p-value $= \mathbb{P}_{H_0}(|T_j| \geq |t_j^{obs}|) =$

$= 2\,\mathbb{P}_{H_0}(T_j \geq |t_j^{obs}|)$ with $T_j \sim t_{n-p}$

in the simple lm we had $(t-2)$ degrees of freedom. Indeed $p=2$ for the simple lm $X = [\underline{1}\ \underline{x}]$




## 2. TEST about the OVERALL SIGNIFICANCE

We want to test if the model is useful to explain the variability of $y$.

$$\begin{cases} H_0: \beta_2 = \beta_3 = \dots = \beta_p = 0 \\ H_1: \overline{H_0} \quad (\text{at least one } \beta_j \neq 0 \quad j = 2, \dots, p) \end{cases}$$

Under $H_0$, all coefficients but $\beta_1$ (intercept) are zero: none of the covariates is useful to predict $y$.

If $H_0$ is true, the model is

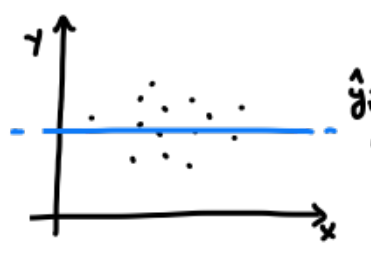
$Y_i = \beta_1 + \varepsilon_i$

$\rightarrow$ I estimate $\tilde{\beta}_1 = \bar{y} \Rightarrow$ predicted values $\tilde{y}_i = \bar{y}\ \forall i$

the residuals are $\tilde{e}_i = (y_i - \bar{y})$. The estimate of $\sigma^2$ is $\tilde{\sigma}^2 = \frac{1}{n}\tilde{\underline{e}}^T\tilde{\underline{e}}$

This model corresponds to the case of "no linear relationship between $y$ and the covariates". We have seen that the coefficient $R^2$ in this case is close to zero. Similarly to what we have seen for the simple linear model, we can reformulate this hypothesis as a test on the value of the coefficient $R^2$ associated with the model:

$$\begin{cases} H_0: R^2 = 0 \\ H_1: R^2 \neq 0 \end{cases}$$




$\hat{y}_i = \tilde{\beta}_1 = \bar{y}\ \forall i$ and $R^2 = 0$

We used a transformation of $\frac{R^2}{1-R^2}$

where, similarly to simple lm,

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

$$= \frac{\tilde{\underline{e}}^T\tilde{\underline{e}}}{\underline{e}^T\underline{e}} - 1 = \frac{\tilde{\sigma}^2}{\hat{\sigma}^2} - 1$$

Notice that also in this case we are comparing the estimated variances $\tilde{\sigma}^2$ and $\hat{\sigma}^2$

$\tilde{\sigma}^2$: estimate under $H_0$ (model with only intercept: 1 covariate) "restricted model"

$\hat{\sigma}^2$: estimate under the full model ($H_1$) $p$ covariates

Distribution:

it is possible to show that

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-p}{p-1} = \frac{\tilde{\sigma}^2 - \hat{\sigma}^2}{\hat{\sigma}^2} \cdot \frac{n-p}{p-1} = \frac{\dfrac{\tilde{\sigma}^2 - \hat{\sigma}^2}{p-1}}{\dfrac{\hat{\sigma}^2}{n-p}} \overset{H_0}{\sim} F_{p-1,n-p}$$

NOTE: to remember the degrees of freedom (and the constants in the test)

$\tilde{\sigma}^2 \sim \chi^2_{n-1}$

$\hat{\sigma}^2 \sim \chi^2_{n-p}$

$$F = \frac{\dfrac{\tilde{\sigma}^2 - \hat{\sigma}^2}{p-1}}{\dfrac{\hat{\sigma}^2}{n-p}} \overset{H_0}{\sim} F_{p-1,\,n-p}$$

$\frac{\text{difference of the estimators}}{\text{difference of their d.o.f.}} = \frac{\tilde{\sigma}^2 - \hat{\sigma}^2}{(n-1)-(n-p)}$ ; $\cancel{n}-1-\cancel{n}+p = p-1$

$\frac{\hat{\sigma}^2}{\text{its d.o.f.}}$

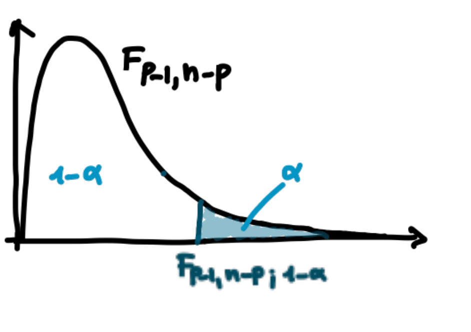
How do we define the rejection and acceptance regions?
- acceptance region (values that suggest that the data support $H_0$)
  under $H_0$: $\tilde{\sigma}^2 \approx \hat{\sigma}^2 \Rightarrow F \approx 0$
- rejection region (values that suggest that the data are against $H_0$)
  under $H_1$: $\tilde{\sigma}^2 \gg \hat{\sigma}^2 \Rightarrow F \gg 0$ I reject $H_0$ for large values

hence $\mathcal{A} = (0, k)$ and $\mathcal{R} = (k, +\infty)$

If we fix the significance $\alpha$, $k$ will be the quantile of level $(1-\alpha)$ of an $F_{p-1,n-p}$ distribution

$$\mathcal{R} = (F_{p-1,n-p;\,1-\alpha}\ ;\ +\infty)$$

with the data: I can compute $f^{obs}$
- reject $H_0$ if $f^{obs} > F_{p-1,n-p;\,1-\alpha}$




Alternatively, the p-value is $\alpha^{obs} = \mathbb{P}_{H_0}(F \geq f^{obs})$