Sum of Squares Decomposition & & the F-test

An alternative way to test whether a predictor X is a "significant predictor" of Y is through something called the "sum of squares decomposition"?

The decomposition provides a breakdown of the total variation in Y into 2 pts:

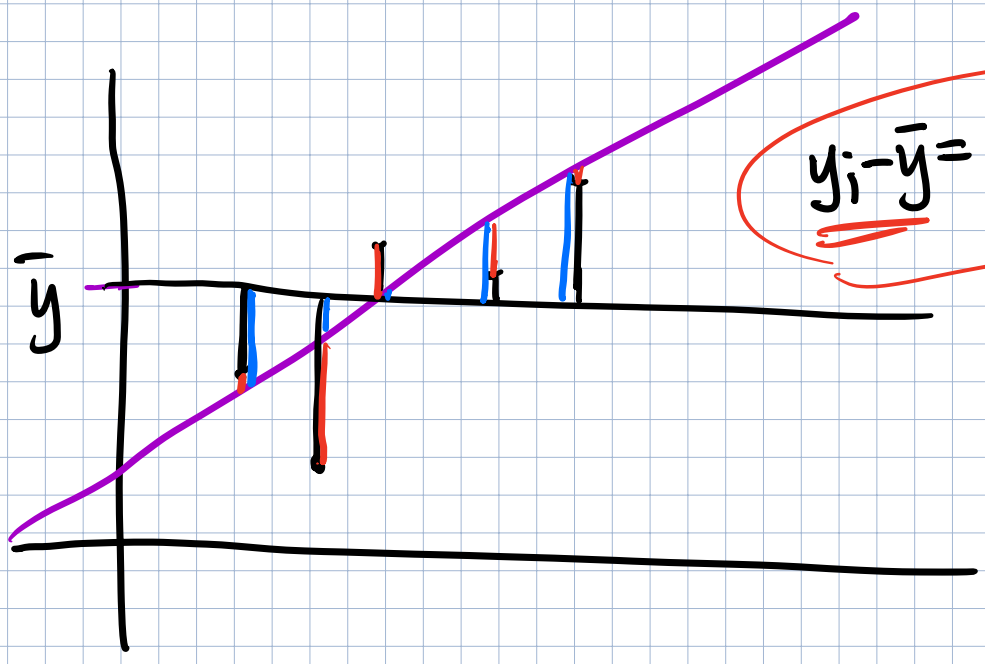
1. the sum of sq. errors (SSE)
2. the regression sum of squares (SSR)

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

SST: "total sum of squares"

SSE

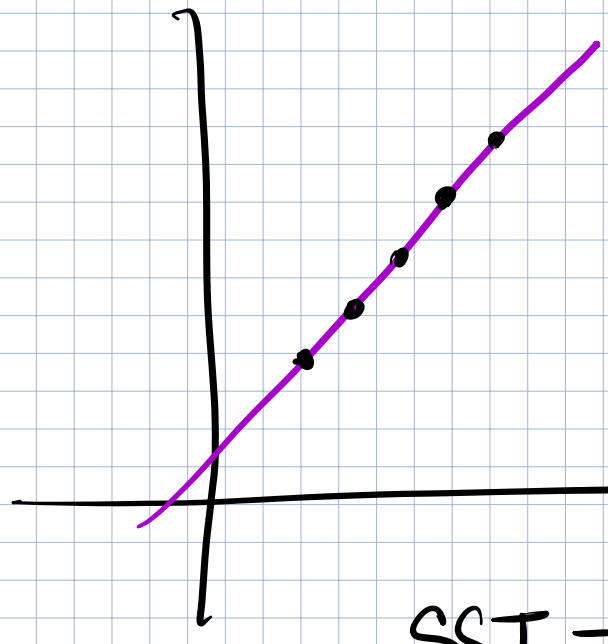
SSR



$$y_i - \bar{y} = y_i - \hat{y}_i + \hat{y}_i - \bar{y}$$

To build our intuition let's look @ some extreme cases:

① our predictions are perfect: $y_i = \hat{y}_i \quad \forall i$



$$e_i = y_i - \hat{y}_i = 0 \quad \forall i$$

$$\Rightarrow \Sigma_i e_i^2 = 0$$

$$SSE = 0$$

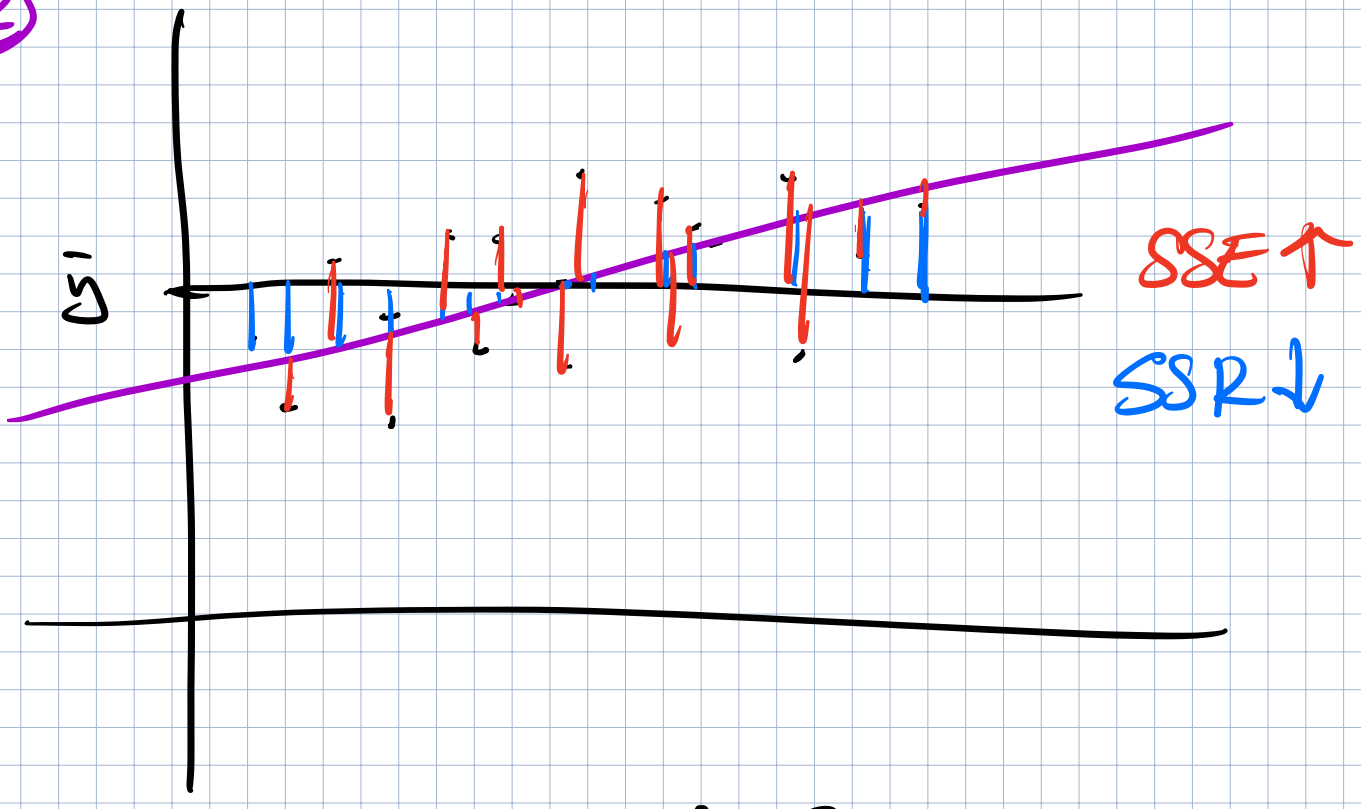
$$SST = \overset{0}{SSE} + SSR$$

$$SST = SSR$$

Intuition ⇒ reject H_0 ⇒ the slope is not 0

⇒ X is a useful predictor of y.

②



If SSR is low & SSE is high

⇒ fail to reject H_0

⇒ X is not useful in predicting y.

$$e_i = y_i - \hat{y}_i$$

$$y_i = \hat{y}_i + e$$

theoretical version

$$Var(y) = Var(\hat{y}) + Var(e)$$

$$\sum_i (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} e_i^2$$

$$SST = SSR + SSE$$

need to prove...

Now let's explicitly define a statistic:

$$F = \frac{SSR/1}{SSE/n-2} \sim F_{1,n-2}$$

Define an F RV:

Suppose ① $U \sim \chi^2_n$

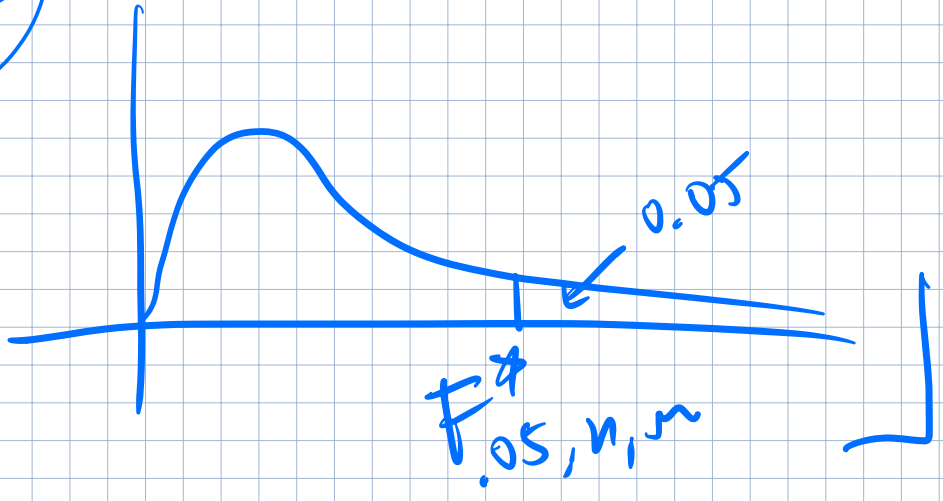
② $V \sim \chi^2_m$

③ $U \perp\!\!\!\perp V$

$\perp\!\!\!\perp$ = "independent"

then:

$$F = \frac{U/n}{V/m} \sim F_{n,m}$$



To show that $F = \frac{MSR}{MSE}$

$$F = \frac{SSR/1}{SSE/n-2} \sim F_{1,n-2}$$ we

need to know:

p = # of betas

$$\frac{SSR}{\sigma^2} \sim \chi^2_1 = p - 1$$

$$\frac{SSE}{\sigma^2} \sim \chi^2_{n-2} = n - p$$

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

& SSR ⫫ SSE

$$SSE = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

2 params est

$n-2$ df

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \sim \chi^2_n$$

0

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

← one parameter est

↓

$n-1$ df

$$\frac{SST}{\sigma^2} = \frac{SSE}{\sigma^2} + \frac{SSR}{\sigma^2}$$

$$\chi^2_{n-1} = \chi^2_{n-2} + \chi^2_{(1)}$$

Decision Rule:

If $F > F^*_{1-\alpha,\, 1,\, n-2}$

$\Rightarrow$ reject H_0

$\Rightarrow$ X is a sig. pred. of Y.

$$\sum_i (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2$$

$$= \sum_i (y_i - \hat{y}_i)^2 + 2\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$

$$+ \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

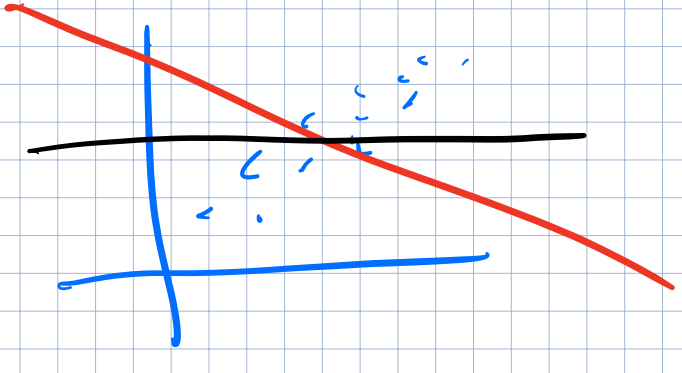
SSE

SSR

Coef. of Determination

$$R^2 := 1 - \frac{SSE}{SST}$$

① For the LSRL, $0 \le R^2 \le 1$



② For SLR

it turns out that

$R^2 = r^2$ where r is the sample correlation

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$