GOODNESS OF FIT

The goodness of fit of a model describes how well it fits the observations.
There are several tools that can be used to evaluate it.

We start with the first "tool": tests to assess whether the model is useful.
In general, these tests evaluate the following system of hypotheses:
$\begin{cases} H_0: \text{the model } \underline{\text{does not}} \text{ help to explain the variability of } Y \\ H_1: \overline{H_0} \end{cases}$

- simple linear model: $Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$
  The question becomes: does the inclusion of x help to explain the variability of y?
  Under H_0 the inclusion of x is not useful
  For this special case, we have already seen that we can answer to this question using a test $H_0: \beta_2 = 0$ vs $H_1: \beta_2 \neq 0$ ($\rightsquigarrow$ test t)

To test the fit of the model we can also use R^2: we have seen that
- $R^2 \approx 0$ no linear relation between y and the covariate
- $R^2 \approx 1$ strong linear relation between y and the covariate

We can do a formal statistical test:

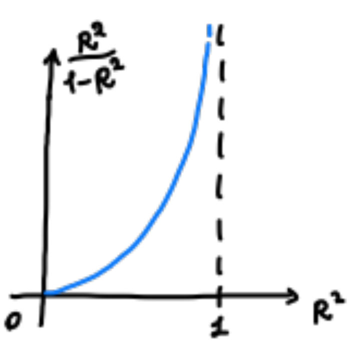
TEST ON R^2

$\begin{cases} H_0: R^2 = 0 \\ H_1: R^2 \neq 0 \quad (\text{i.e. } R^2 > 0) \end{cases}$ → under H_0, including x is not useful

$\rightsquigarrow$ under H_0 I use the "null model" $Y_i = \beta_1 + \varepsilon_i$
$\rightsquigarrow$ under H_1 I use the full model $Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$

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

→ we use a transformation of R^2: $\frac{R^2}{1-R^2} = \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} - 1$



this is a monotone increasing function of R^2.

$\frac{R^2}{1-R^2} = \frac{SSR}{SST} \cdot \left(1 - \frac{SSR}{SST}\right)^{-1} = \frac{SSR}{\cancel{SST}} \cdot \frac{\cancel{SST}}{SST - SSR} = \frac{SSR}{SSE}$

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

what are the two quantities A and B?

Ⓐ it is the sum of squared residuals of a model with only the intercept
Recall that if $y_i = \beta_1 + \varepsilon_i \Rightarrow \tilde{\beta}_1 = \bar{y} \Rightarrow \tilde{y}_i = \bar{y}$ for all i
Let's call $y_i - \bar{y} = e_i^* \Rightarrow \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} e_i^{*2}$
Notice this is also the model we assume if H_0 is true
Ⓐ is the SUM of SQUARED RESIDUALS UNDER H_0

Ⓑ it is the sum of squared residuals of the full model
Let's call $y_i - \hat{y}_i = e_i \Rightarrow \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$
Ⓑ is the SUM of SQUARED RESIDUALS UNDER H_1

Returning now to the test statistic

$\frac{R^2}{1-R^2} = \frac{SSE_{H_0}}{SSE_{H_1}} - 1$

$= \frac{\sum_{i=1}^{n} e_i^{*2}}{\sum_{i=1}^{n} e_i^2} - 1$

we are comparing the residuals of the model we would estimate in the absence of information (ie, x) and the residuals of the model that includes x.

Notice that $\sum_{i=1}^{n} e_i^{*2} = n\tilde{\sigma}^2$ where $\tilde{\sigma}^2 = \frac{1}{n}\sum(y_i - \bar{y})^2$ is the estimate of the variance under the model with only the intercept (H_0).
The denominator $\sum_{i=1}^{n} e_i^2 = n\hat{\sigma}^2$ where $\hat{\sigma}^2 = \frac{1}{n}\sum(y_i - \hat{y}_i)^2$ estimate under the full model (H_1).

$\frac{\sum_{i=1}^{n} e_i^{*2}}{\sum_{i=1}^{n} e_i^2} - 1 = \frac{n\tilde{\sigma}^2}{n\hat{\sigma}^2} - 1 = \frac{\tilde{\sigma}^2}{\hat{\sigma}^2} - 1 = \frac{\tilde{\sigma}^2 - \hat{\sigma}^2}{\hat{\sigma}^2}$ → we are comparing the estimated variance of the error under the two models

How do we interpret this quantity?
What value of the test statistic do we expect under H_0 and H_1? (How is the reject region defined?)

- if H_0 is true, x is not helpful for explaining y
  → hence the models under H_0 and H_1 will have similar performances at predicting y.
  → if the predictions under the two models are similar, also the residuals will be similar
  → the "total amount of error" of the two models will be similar
  → $\sum e_i^{*2}$ and $\sum e_i^2$ will be similar (hence also $\tilde{\sigma}^2$ and $\hat{\sigma}^2$)

  $\underbrace{\frac{\sum_{i=1}^{n} e_i^{*2}}{\sum_{i=1}^{n} e_i^2}}_{\approx 1} - 1 = \frac{\overbrace{\tilde{\sigma}^2 - \hat{\sigma}^2}^{\approx 0}}{\hat{\sigma}^2}$ under H_0 I expect this quantity to be close to zero

  → the ACCEPTANCE REGION will be $(0; k)$

- What happens if H_0 is not true?
  In this case, the full model (H_1) is better than the null model (H_0)
  (it can not be worse in terms of prediction, at most is the same)
  → the predictions under H_1 will be more accurate
  → the total amount of error of the full model will be smaller
  → $\sum e_i^{*2} > \sum e_i^2$
  → $\tilde{\sigma}^2 > \hat{\sigma}^2$

  $\underbrace{\frac{\sum_{i=1}^{n} e_i^{*2}}{\sum_{i=1}^{n} e_i^2}}_{>1} - 1 = \frac{\overbrace{\tilde{\sigma}^2 - \hat{\sigma}^2}^{>0}}{\hat{\sigma}^2} > 0$ under H_1 I expect large positive values!

  → the REJECT REGION will be $(k; +\infty)$

Now we only need a distribution to determine the threshold k.

def: if $X \sim \chi^2_{v_1}$ and $W \sim \chi^2_{v_2}$ independent, $\frac{X/v_1}{W/v_2} \sim F_{v_1, v_2}$ F distribution with (v_1, v_2) degrees of freedom

It is possible to show that:

$\frac{SSR}{\sigma^2} \overset{H_0}{\sim} \chi^2_1$

$\frac{SSE}{\sigma^2} \overset{H_0}{\sim} \chi^2_{n-2}$

$SSR \perp\!\!\!\perp SSE$

Recall that the test statistic is $\frac{R^2}{1-R^2} = \frac{\tilde{\sigma}^2 - \hat{\sigma}^2}{\hat{\sigma}^2} = \frac{SSR}{SSE}$

Hence it holds $F = \frac{SSR}{SSE/(n-2)} = \frac{\left(\frac{SSR}{\sigma^2}\right)/1}{\left(\frac{SSE}{\sigma^2}\right)/n-2}$

$\Rightarrow F \overset{H_0}{\sim} F_{1, n-2}$

$F = \frac{R^2}{1-R^2} \cdot (n-2) = \frac{\tilde{\sigma}^2(Y) - \hat{\sigma}^2(Y)}{\hat{\sigma}^2(Y)} \cdot (n-2) = \frac{SSR}{SSE}(n-2) \overset{H_0}{\sim} F_{1, n-2}$

PROOF FOR THE CASE of SIMPLE LM

We can also demonstrate the distribution of F (this proof is only for the simple lm).

Let's start from $\frac{SSA}{SSE} = \frac{\sum_{i=1}^{n} e_i^{*2}}{\sum_{i=1}^{n} e_i^2}$

Now, notice that we can write

$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)^2 = \sum_{i=1}^{n}(y_i - \bar{y} + \hat{\beta}_2 \bar{x} - \hat{\beta}_2 x_i)^2 =$

$= \sum_{i=1}^{n}[(y_i - \bar{y}) - \hat{\beta}_2(x_i - \bar{x})]^2 = \sum_{i=1}^{n}(y_i-\bar{y})^2 + \hat{\beta}_2^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 \underbrace{- 2\hat{\beta}_2 \sum_{i=1}^{n}(x_i-\bar{x})(y_i - \bar{y})}$

$-2\hat{\beta}_2 \cdot \underbrace{\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}}_{=\hat{\beta}_2} \cdot \sum(x_i - \bar{x})^2$

$-2\hat{\beta}_2^2 \sum(x_i-\bar{x})^2$

$= \sum_{i=1}^{n}\underbrace{(y_i - \bar{y})^2}_{(e_i^*)^2} - \hat{\beta}_2^2 \sum_{i=1}^{n}(x_i - \bar{x})^2$

$\Rightarrow \sum_{i=1}^{n} e_i^{*2} = \sum_{i=1}^{n} e_i^2 + \hat{\beta}_2^2 \sum_{i=1}^{n}(x_i - \bar{x})^2$

Moreover, recall that

$V(\hat{\beta}_2) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$; $\hat{V}(\hat{\beta}_2) = \frac{S^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$; $\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}$

$\frac{R^2}{1-R^2} = \frac{\sum_{i=1}^{n}(e_i^*)^2}{\sum_{i=1}^{n} e_i^2} - 1 = \frac{\sum_{i=1}^{n} e_i^2 + \hat{\beta}_2^2 \sum_{i=1}^{n}(x_i-\bar{x})^2}{\sum_{i=1}^{n} e_i^2} - 1 = \cancel{1} + \frac{\hat{\beta}_2^2}{\sum_{i=1}^{n} e_i^2 / \sum_{i=1}^{n}(x_i-\bar{x})^2} \cancel{-1} =$

$\sum e_i^2 = \sum(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)^2 = n\hat{\sigma}^2 = (n-2)S^2$

$= \frac{\hat{\beta}_2^2}{(n-2)\frac{S^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \cdot \frac{\sigma^2}{\sigma^2} = \frac{\frac{\hat{\beta}_2^2}{\frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}}{\frac{(n-2)S^2}{\sigma^2}}$

$\Rightarrow \frac{R^2}{1-R^2} = \frac{\left(\frac{\hat{\beta}_2}{\sqrt{V(\hat{\beta}_2)}}\right)^2 \overset{H_0}{\sim} N(0,1)^2}{\frac{(n-2)S^2}{\sigma^2} \cdot \frac{1}{(n-2)} \overset{H_0}{\sim} \frac{\chi^2_{n-2}}{n-2}} \cdot \frac{1}{(n-2)} = \frac{1}{(n-2)} T^2$

where $T = \frac{\hat{\beta}_2}{\sqrt{\frac{S^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}} = \frac{\hat{\beta}_2}{\sqrt{\hat{V}(\hat{\beta}_2)}} \overset{H_0}{\sim} t_{n-2}$

def: if $V \sim t_m$, then $V^2 \sim F_{1,m}$.

Hence $F = (n-2)\frac{R^2}{1-R^2} = (n-2)\frac{SSR}{SSE} \overset{\text{only in the simple LM}}{=} T^2 \overset{H_0}{\sim} F_{1, n-2}$

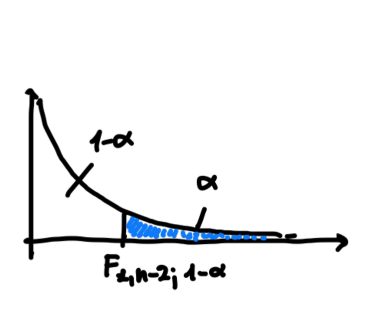
So, we have derived the null distribution of $F = \frac{SSR}{SSE/(n-2)} = (n-2)\frac{R^2}{1-R^2}$
(recall: monotone increasing fun of R^2)

To see what values lead to rejecting H_0, we can also do a reasoning about the values of $\left((n-2)\cdot\frac{R^2}{1-R^2}\right)$ directly.
If I am testing $H_0: R^2 = 0$ vs $H_1: R^2 > 0$, I would reject for large values of R^2
Since F is a monotone increasing transformation
$\Rightarrow$ corresponds to large values of F $\Rightarrow$ reject region: $(k; +\infty)$
With the data, we can compute the observed value of the test, f^{obs}
$f^{obs} = \frac{R^2}{1-R^2}(n-2)$ or $f^{obs} = \frac{\tilde{\sigma}^2 - \hat{\sigma}^2}{\hat{\sigma}^2}(n-2)$

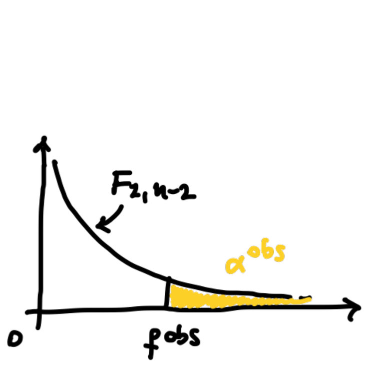
or, in the case of a simple lm $f^{obs} = \left(\frac{\hat{\beta}_2}{\sqrt{S^2/\sum_{i=1}^{n}(x_i-\bar{x})^2}}\right)^2 = (t^{obs})^2$

TEST

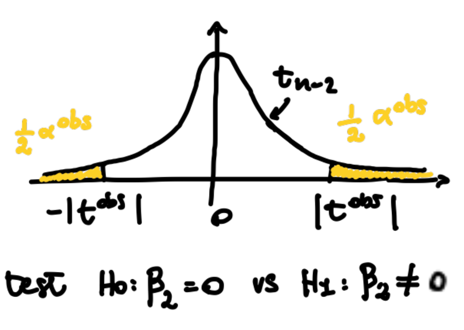
- fixed significance level α: $\alpha = \mathbb{P}(\text{reject } H_0 \mid H_0 \text{ true})$
  $\mathbb{P}_{H_0}(F > F_{1,n-2;1-\alpha}) = \alpha$
  the acceptance region is
  $A = (0; F_{1,n-2;1-\alpha})$
  if $f^{obs} < F_{1,n-2;1-\alpha} \Rightarrow$ we do not reject H_0
  if $f^{obs} \geq F_{1,n-2;1-\alpha} \Rightarrow$ we reject H_0



- pvalue: $\underset{\alpha^{obs}}{\mathbb{P}_{H_0}(F > f^{obs})}$ where $F \overset{H_0}{\sim} F_{1,n-2}$



However, notice that
$\mathbb{P}_{H_0}(F > f^{obs}) = \mathbb{P}_{H_0}(T^2 > (t^{obs})^2)$
$= \mathbb{P}_{H_0}(|T| > |t^{obs}|) =$
$= 2\,\mathbb{P}_{H_0}(T > |t^{obs}|)$ $T \sim t_{n-2}$



where T is exactly the test statistic we derived to test $H_0: \beta_2 = 0$ vs $H_1: \beta_2 \neq 0$