10 oct - LEC 4

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:

$$\begin{cases} H_0: \text{the model does not help to explain the variability of } Y \\ H_1: \text{the model helps to explain the variability of } Y \end{cases}$$

- simple linear model: $Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$ (only one covariate x)

  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:
  If H_0 is true, the correct model is the null model $Y_i = \beta_1 + \varepsilon_i$
  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 x
- $R^2 \approx 1$: strong linear relation between y and the covariate x

We can do a formal statistical test:

TEST ON R^2

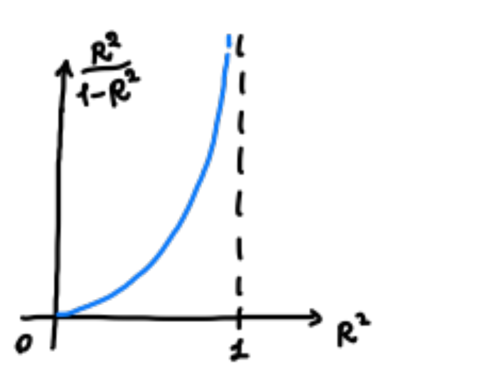
$$\begin{cases} H_0: R^2 = 0 \\ H_1: R^2 \neq 0 \quad (\text{i.e. } R^2 > 0) \end{cases}$$

→ 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: $\frac{R^2}{1-R^2}$

this is a monotone increasing function of R^2.

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

$$= \frac{SST - SSE}{SSE} = \frac{SST}{SSE} - 1 =$$

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

what are the two quantities A and B?

(A) is the SUM OF THE SQUARED RESIDUALS of the NULL MODEL: model with only the intercept
Since the null model is the model assumed under H_0, (A) is the SUM of SQUARED RESIDUALS UNDER H_0.

Recall that if $Y_i = \beta_1 + \varepsilon_i$ $\Rightarrow$ the estimate is $\tilde{\beta}_1 = \bar{y}$
$\Rightarrow$ the predicted values are $\tilde{y}_i = \bar{y}$ for all i

Let's define the residuals $y_i - \bar{y} = e_i^*$

sum of squared residuals is $\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} e_i^{*2}$

(B) is the SUM OF SQUARED RESIDUALS OF THE FULL MODEL.
the model is the unconstrained model (i.e., the model under H_1).
(B) is the SUM of SQUARED RESIDUALS UNDER H_1.

Model $Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$

Let's define the residuals $y_i - \hat{y}_i = e_i$

sum of squared residuals is $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$

Returning now to the TEST STATISTIC

$$\frac{R^2}{1-R^2} = \frac{SST}{SSE} - 1 = \frac{SSE_{H_0}}{SSE_{H_1}} - 1 = \frac{\sum_{i=1}^{n} e_i^{*2}}{\sum_{i=1}^{n} e_i^2} - 1$$

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 $\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$ is the estimate of the variance of the error under the model with only the intercept (H_0).

The denominator is $\sum_{i=1}^{n} e_i^2 = n\hat{\sigma}^2$ where $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the estimate of the variance of the error under the full model (H_1).

Hence,
$$\frac{R^2}{1-R^2} = \frac{\sum_{i=1}^{n} e_i^{*2}}{\sum_{i=1}^{n} e_i^2} - 1 = \frac{n\tilde{\sigma}^2}{n\hat{\sigma}^2} - 1 = \frac{\tilde{\sigma}^2}{\hat{\sigma}^2} - 1 = \frac{\tilde{\sigma}^2 - \hat{\sigma}^2}{\hat{\sigma}^2}$$

→ we are comparing the estimated variance of the error under the two models.
Now, we need to study what values the test statistic can assume.
First of all, notice that the quantity is always positive

What values of the test statistic do we expect under H_0 and H_1?
i.e., how are the REJECT and ACCEPTANCE REGIONS defined?

- IF H_0 IS TRUE, x is not useful in explaining y
  - → hence the models under H_0 and H_1 will have similar performances at predicting y.
    (the full model can not be worse in terms of prediction, at most is the same as the null model)
  - → 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
  - → the quantities $\sum_{i=1}^{n} e_i^{*2}$ and $\sum_{i=1}^{n} e_i^2$ will be similar (hence also $\tilde{\sigma}^2$ and $\hat{\sigma}^2$).

$$\overbrace{\frac{\sum_{i=1}^{n} e_i^{*2}}{\sum_{i=1}^{n} e_i^2}}^{\approx 1} - 1 = \frac{\overbrace{\tilde{\sigma}^2 - \hat{\sigma}^2}^{\approx 0}}{\hat{\sigma}^2}$$

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)
  - → the predictions under H_1 will be more accurate
  - → the total amount of error of the full model will be smaller
  - → $\sum_{i=1}^{n} e_i^{*2} \gg \sum_{i=1}^{n} e_i^2$
  - → $\tilde{\sigma}^2 \gg \hat{\sigma}^2$

$$\overbrace{\frac{\sum_{i=1}^{n} e_i^{*2}}{\sum_{i=1}^{n} e_i^2}}^{\gg 1} - 1 = \frac{\overbrace{\tilde{\sigma}^2 - \hat{\sigma}^2}^{\gg 0}}{\hat{\sigma}^2} > 0$$

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.

> Preliminary result
>
> If $X \sim \chi^2_{\nu_1}$ and $W \sim \chi^2_{\nu_2}$ independent, $\frac{X/\nu_1}{W/\nu_2} \sim F_{\nu_1,\nu_2}$ F distribution with (ν_1, ν_2) degrees of freedom

It is possible to show that:

$$\frac{SSR}{\sigma^2} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sigma^2} \overset{H_0}{\sim} \chi^2_1$$

$$\frac{SSE}{\sigma^2} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sigma^2} \overset{H_0}{\sim} \chi^2_{n-2}$$

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

The test statistic is $\frac{R^2}{1-R^2} = \frac{SSR}{SSE}$

Hence it holds
$$F = \frac{SSR/1}{SSE/(n-2)} = \frac{\left(\frac{SSR}{\sigma^2}\right)/1}{\left(\frac{SSE}{\sigma^2}\right)/n-2} \overset{H_0}{\sim} F_{1,n-2}$$

Hence to perform the test we can use this quantity (known distribution under H_0)

$$\left.\begin{aligned} F &= \frac{R^2}{1-R^2} \cdot (n-2) = \frac{SSR}{SSE}(n-2) = \\ &= \left(\frac{SST}{SSE} - 1\right)(n-2) = \\ &= \left(\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} - 1\right)(n-2) = \\ &= \frac{\tilde{\Sigma}^2 - \hat{\Sigma}^2}{\hat{\Sigma}^2}(n-2) \overset{H_0}{\sim} F_{1,n-2} \end{aligned}\right\} \text{equivalent formulations}$$

To finish the test

1) FIXED SIGNIFICANCE LEVEL α

$\alpha = \mathbb{P}(\text{reject } H_0 \mid H_0 \text{ true})$

the reject region is on the right tail $\Rightarrow \alpha = \mathbb{P}_{H_0}(F \in (k, +\infty))$
$= \mathbb{P}_{H_0}(F > k)$

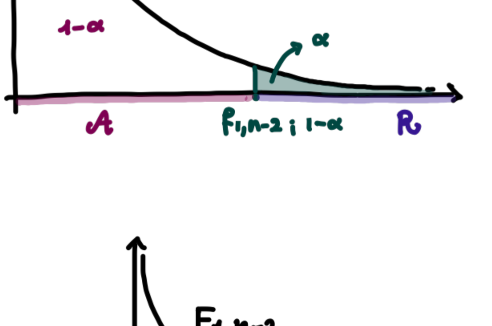
what is the value k that guarantees that the probability that F will assume values larger than k is exactly α?
(i.e., the value that guarantees that the probability that F assumes values smaller than k is $1-\alpha$)

$k = f_{1,n-2;1-\alpha}$ quantile of level $(1-\alpha)$ of a $F_{1,n-2}$ distribution

$\mathbb{P}_{H_0}(F > f_{1,n-2;1-\alpha}) = \alpha$

acceptance region $\mathcal{A} = (0, f_{1,n-2;1-\alpha})$
reject region $\mathcal{R} = (f_{1,n-2;1-\alpha}, +\infty)$

if $f^{obs} < f_{1,n-2;1-\alpha} \Rightarrow$ we do not reject H_0
if $f^{obs} > f_{1,n-2;1-\alpha} \Rightarrow$ we reject H_0



2) P.VALUE $\alpha^{obs} = \mathbb{P}_{H_0}(F \geq f^{obs})$ where $F \overset{H_0}{\sim} F_{1,n-2}$

Remark: To see what values lead to rejecting H_0, we could also do a reasoning about the values of $(n-2)\cdot\frac{R^2}{1-R^2}$ 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, large values of R^2 correspond to large values of F
$\Rightarrow$ reject region: $(k; +\infty)$