ANOVA F-test for SLR model usefulness

Use case

We have a single simple linear regression model (SLR) and want to assess its usefulness. Does the model provide useful predictions about our variable of interest?

The population regression line is defined by
$\mu_y = \beta_0 + \beta_1 x$, and the least-squares regression line is
$\hat{y} = b_0 + b_1 x$.

The ANOVA approach uses the following quantities to assess how much of the data's variation is explained by our model.

$$\text{Total SS (SST)} = \sum (y_i - \bar{y})^2 \qquad \text{df} = n - 1$$

$$\text{Regression SS (SSR)} = \sum (\hat{y}_i - \bar{y})^2 \qquad \text{df} = 1$$

$$\text{Residual SS (SSE)} = \sum (y_i - \hat{y}_i)^2 \qquad \text{df} = n - 2$$

$$\text{Note: SST} = \text{SSR} + \text{SSE}$$

Using these values, we can readily compute the *coefficient of determination*,
r^2.

$$r^2 = \frac{\text{SSR}}{\text{SST}}$$

Finally, we compute the *mean squares* values before we can calculate the test statistic.

$$\text{Total MS (MST)} = \frac{\text{SST}}{\text{df}_{\text{SST}}} = \frac{\text{SST}}{n-1}$$

$$\text{Regression MS (MSR)} = \frac{\text{SSR}}{\text{df}_{\text{SSR}}} = \frac{\text{SSR}}{1}$$

$$\text{Residual MS (MSE)} = \frac{\text{SSE}}{\text{df}_{\text{SSE}}} = \frac{\text{SSE}}{n-2}$$

$$\text{Note: In general, SST} \neq \text{SSR} + \text{SSE (mean squares are not additive)}$$

Preconditions

For an ANOVA F-test to be valid, we should verify that we already meet the preconditions for the validity of SLR. (If SLR is inappropriate, an F test on your model would be pointless.)

In our case, this means that we should verify that residuals are scattered randomly around zero with uniform variation. This satisfies the *linearity*, *independence*, and *constant variance* assumptions that underpin SLR.

Hypotheses

Null hypothesis,
H_0**:** the model is not useful (
$\beta_1 = 0$)

Alternative hypothesis,
H_a**:** the model is useful (
$\beta_1 \neq 0$)

Test statistic

$$F = \frac{\text{MSR}}{\text{MSE}}$$

p-value

We compute our p value as follows, where
$F(k_1, k_2)$ indicates a random variable distributed according to the
$F(k_1, k_2)$ distribution.

$$p = P(F(k_1, k_2) \geq F), \text{ where } k_1 = 1 \text{ and } k_2 = n - 2$$

p-value (Python)

```
p = 1 - stats.f.cdf(F, 1, n-2)
```