158 - Homework 4

2023-02-11

1.

We have the model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. We can compute the residuals $e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ which approximate the ϵ_i if the model fits (otherwise it's not clear what we're computing). Recalling that the variance of the errors is the "average squared deviation from the mean", we would love to compute

$$\frac{1}{n}\sum(\epsilon_i - \mu_\epsilon)^2 = \frac{1}{n}\sum \epsilon_i^2$$

But since the ϵ are unknowable, we instead compute

$$\frac{1}{n-2}\sum(e_i - \bar{e})^2 = \frac{1}{n-2}\sum e_i^2 = \frac{1}{n-2}SSE = MSE$$

The $n-2$ is to fix the fact that SSE is smaller than $\sum \epsilon_i^2$ by virtue of doing optimization.

If the model fits, MSE is estimating σ_ϵ^2 (unbiasedly).

A more general model is that $y_i = \mu_{x_i} + \epsilon_i$, and this model certainly fits (though requires estimating a lot more parameters than 2, one for each unique value of x in the data set. Also, it is unable to predict observations at x variables you haven't seen yet). We can once again compute residuals in this model, $e_i^* = y_i - \bar{y}_{x_i}$. We are going to need replications at some of the x values to keep these from all being 0. Suppose that there are c unique values of x in the data set, with $c < n$.

From these, we can compute

$$\frac{1}{n-c}\sum(e_i^* - \bar{e^*})^2 = \frac{1}{n-c}\sum e_i^{*2} = \frac{1}{n-c}SSPE = MSPE$$

which is certainly a (less stable) estimate of σ^2. (Less stable because it's based on less information due to having to estimate more parameters, but we know the model fits). PE stands for "pure error".

What shows up in the formal test of linearity is $MSLF$ (lack of fit) $= \frac{1}{c-2}(SSE - SSPE)$. Aruge using the above facts and some simple algebra that $MSLF$ is also estimating σ_ϵ^2.

As a consequence, argue that the F statistic used should be about 1 if the null (linear model is appropriate) is true.

2.

The sakai site contains a data set consisting of brain mass to body mass for 27(ish) species. Fit a model predicting brain mass based on body mass, and additionally include a confidence interval and prediction interval (to be introduced on Tuesday, you'll be able to steal code from those lecture notes). Fully justify and interpret your model, writing in complete sentences.