HW 4

Nihal Karim

1. Suppose we have fit a MLR model between response variable Y and predictors $X_1, ..., X_{p-1}$. using a data of size n. The global F-test aka omnibus test considers the hypotheses:

$$H_0: \ y_i = \beta_0 + \epsilon_i \quad \text{vs.} \quad H_1: y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_{p-1} X_{(p-1)i} + \epsilon_i$$

using the statistic $F = \dfrac{MSR}{MSE}$ where MSR and MSE are calculated under the full model. Show that this definition is equivalent to the alternative formulation of the F statistic as:

$$F_{alt} = \frac{\dfrac{SSE_{H_0} - SSE_{H_1}}{df_{SSE_{H_0}} - df_{SSE_{H_1}}}}{\dfrac{SSE_{H_1}}{df_{SSE_{H_1}}}}.$$

$H_0: \hat{Y}_i = \bar{Y}$

$SSE_{H_0} = \Sigma(Y_i - \hat{Y}_i)^2 = \Sigma(Y_i - \bar{Y})^2 = SST_{H_0}$

$SSR_{H_0} = \Sigma(\hat{Y}_i - \bar{Y})^2 = 0$

$H_1: \hat{Y}_i = \beta_0 + \beta_1 X_{1i} + \ldots$

$SSE_{H_1} = \Sigma(Y_i - \hat{Y}_i)^2$

$SSR_{H_1} = \Sigma(\hat{Y}_i - \bar{Y})^2$

$SST_{H_1} = SST_{H_0} = \Sigma(Y_i - \bar{Y})^2$ SST is property of data not predictors

$MSR = \dfrac{SSR}{p-1}$ $\quad SSR = SST - SSE$

$= SSE_{H_0} - SSE_{H_1}$

$df_{SSE_{H_0}} = n-1$ only intercept (B_0) as predictor

$df_{SSE_{H_1}} = n-p$ lose d.f for each predictor

$$p-1 = (n-1)-(n-p) = df_{SSE_{H_0}} - df_{SSE_{H_1}}$$

$$MSR = \frac{SSE_{H_0} - SSE_{H_1}}{df_{SSE_{H_0}} - df_{SSE_{H_1}}}$$

$$MSE = \frac{SSE}{n-p} = \frac{SSE_{H_1}}{df_{SSE_{H_1}}}$$

$$F = \frac{MSR}{MSE} = F_{alt}$$ ✓

2. If a predictor variable is categorical with six states and we want to include it in a regression model, how many dummy variables do we need to use?

$k - 1 \rightarrow 6 - 1 = 5$

with all 6 we would have multicollinearity

instead use 5 and have a 6th state represented

as all 5 are null

3. Suppose a predictor variable is categorical with three states "C1", "C2", "C3". When we include it in a regression model and the individual t tests used "C1" as reference level, and showed "C2" is significant and "C3" is not. Would you conclude that "we should drop C3 and fit a new model"? Why or why not?

No, C3 must be considered in context as

it could have some affect on C2 or the

model overall. Unless there is strong theoretical

justification or other data-driven reason (eg.

more diagnostic tests or analysis on data collection, etc.)

then C3 should not be dropped

8.14. In a regression study of factors affecting learning time for a certain task (measured in minutes), gender of learner was included as a predictor variable (X_2) that was coded $X_2 = 1$ if male and 0 if female. It was found that $b_2 = 22.3$ and $s\{b_2\} = 3.8$. An observer questioned whether the coding scheme for gender is fair because it results in a positive coefficient, leading to longer learning times for males than females. Comment.

The coding scheme itself is not inherently unfair based on the data. The positive coefficient is statistically significant and implies an underlying relationship between gender and learning time.

$$t = \frac{22.3}{3.8} = 5.87$$

t test implies statistical significance

5. This question will help you to understand the calculation of ANOVA in MLR using an example. For the dataset KelleyBlueBookData.csv, response= Price against the following predictors: Mileage, Liter, Cylinder (in this order).

(a) Run the sequential ANOVA for the fitted model. Report null and alternative hypothesis, the F stat, and the p-value for the F-test for dropping or including the 'Cylinder' predictor. What is the conclusion of this test?

H_0: Cylinder predictor $= 0$

H_1: cylinder predictor $\neq 0$

Fstat $= 15.997$ Pval : 6.93×10^{-5}

Since p value is low reject null hypothesis

so cylinder is a significant predictor

on price statistically

(b) Manually run the test in part (a) yourself: 1. fit the null model in python and extract SSE and degrees of freedom of this SSE; then 2. fit the alternative model in python and extract SSE and degrees of freedom of this SSE. Plug in the numbers to

$$F_{alt} = \frac{\frac{SSE_{H_0} - SSE_{H_1}}{df_{SSE_{H_0}} - df_{SSE_{H_1}}}}{\frac{SSE_{H_1}}{df_{SSE_{H_1}}}}.$$

Does the value you calculated match the F-statistic from part (a)?

$SSE_{H_0} = 52,637,552,166$ $SSE_{H_1} = 51,605,604,120$

$df_{SSE_{H_0}} = 801$ $df_{SSE_{H_1}} = 800$

$F_{alt} \simeq 15.99745$ ✓

Matches yes

(c) Run the partial ANOVA (typ=2) for the fitted model. Does the F-test for 'Cylinder' match the F-test from part (a)? Why or why not?

$F = 15.997$ pval $= 6.9 \times 10^{-5}$

Yes it matches because cylinder was the last variable added to the model so the effect that is measured is the same as having order not matter as all other variables are included even in type 1 since since its last.

(d) From the partial ANOVA (typ=2) table in (c), report the null and alternative hypothesis, the F stat, and the p-value for the F-test for dropping or including the 'Mileage' predictor. Interpret the result of this test.

H_0 : Mileage predictor $= 0$

H_1 : Mileage predictor $\neq 0$

Fstat $= 19.9$ pval $= 9 \times 10^{-6}$

Again, p value is very small indicating there is statistically significant predictor Mileage to Price and we reject the null hypothesis.

(e) Manually run the test in part (d) yourself: 1. fit the null model in python and extract SSE and degrees of freedom of this SSE; then 2. fit the alternative model in python and extract SSE and degrees of freedom of this SSE. Plug in the numbers to

$$F_{alt} = \frac{\frac{SSE_{H_0} - SSE_{H_1}}{df_{SSE_{H_0}} - df_{SSE_{H_1}}}}{\frac{SSE_{H_1}}{df_{SSE_{H_1}}}}.$$

Does the value you calculated match the F-statistic from part (d)?

$SSE_{H0} = 52{,}889{,}600{,}780$

$df\ SSE_{H0} = 801$

$SSE_{H1} = 5{,}165{,}604{,}120$

$df\ SSE_{H1} = 800$

$F_{alt} = 19.985$ ✓

Matches

6. This question will help you to understand the calculation of R^2 and R^2_{adj} in MLR using an example. For the dataset KelleyBlueBookData.csv:

(a) Fit Model 1: a model which considers Price as the response and regresses it against the predictors Mileage and Cylinder. Report the R^2 and R^2_{adj} values from the summary table.

$R^2 = 0.340$ $R^2adj = 0.338$

(b) Calculate R^2 and R^2_{adj} for the model in part (a) yourself. Obtain the SSE and SST of the model in part (a), then plug in the formulas: $R^2 = 1 - \frac{SSE}{SST}$ and $R^2_{adj} = 1 - \dfrac{SSE/n-p}{SST/n-1}$ Do the values match with the python output in (a)?

SSE = 51,799,580,167 n = 804

SST = 78,461,382,864 p = 3

$R^2 = 0.3398$ ✓

$R^2adj = 0.3382$ ✓

They do match

(c) Fit Model 2: a model which considers Price as the response and regresses it against the predictors Mileage, Liter and Cylinder. Report the $R^2$and R^2_{adj} values from the summary table. Which model is preferable according to R^2_{adj} between Model 1 and Model 2? Why?

$R^2 = 0.342$ $R^2adj = 0.340$

The 2nd model is slightly more preferable in fit but not by much with the small increase in R^2 and R^2adj. So the liter variable only improves model fit by a small amount.

(d) Open question: Consider simultaneously the t-test results, ANOVA, R^2_{adj} and any other concepts we have covered so far (e.g. diagnostics). Which model would you choose, Model 1 or Model 2? Argue for your model in terms of these statistics and also the real life meaning of the problem.

t test shows pval of litar = 0.084
which means the predictor is not
necessarily significant at 5% level
With the marginal increase in R^2 and R^2adj
Model 1 actually seems to be the
better choice since it simplifies
the model which would make it more
straightforward and reliable since
we have less predictors when
considering future data.