DISCUSSION 12

- Review on linear regression

1. Model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

   (a) fitting $\hat{\beta} = (X^T X)^{-1} X^T Y$,where $X = [1_n, X_1, X_2, ..., X_k], Y = [y_1, y_2, ..., y_n]^T$
   $\hat{\sigma}^2 \sim \frac{MeanRSS}{n-k-1}$

   (b) testing:

   (c) model selection:

   (d) model checking:

2. Hierarchy rule, parsimony rule:

   Hierarchy rule: when you include interaction of two factors, say A, B, then you should want to keep A, B in the model. Help with the interpretation of the model.

   Parsimony rule: when two models perform similarly, we prefer the simple models to the complicated one.

3. Model comparison of two nested models(**Lack-of-fit** test)

   full model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$

   reduced model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_l + \epsilon, l < k$

Source	Df	RSS	MRSS	F
reduced model	$n-l-1$	RSS(reduced)	MRSS(reduced)	
full model	$n-k-1$	RSS(full)	MRSS(full)	
Difference	$k-l$	RSS1-RSS2	MSS(diff)	MSS(diff)/MSS(full)

   – If the F-test is not significant, this indicates that there is no significant difference between the two tested models. Based on **rule of parsimony**, we keep the simple(reduced) model instead of the full model. Otherwise, we keep the full model.

   – Using **anova(model1,model2)** in R

4. **Departure from underlying assumptions:**

   (a) Effects of outliers, Influential points, Non-normality etc.

   (b) collinearity: measure how close x_j is to being a linear combination of the other explanatory variables.
   – The Variance Inflation Factor is defined as $\text{VIF}_j = \frac{1}{1-R_j^2}$.
   – effects : (1) inflate variance of $\hat{\beta}_j$, make the estimation of $\hat{\beta}_j$ un stable.
   (2) prediction at $\mathbf{x}$ will have a big standard error, if the $\mathbf{x}$ is far from center $\bar{\mathbf{x}}$ of observed data.

-Exercises

1. go through this example to learn about effects of collinearity in regression. Adding a variable correlated with current variable will

   – affect the estimation of regression coefficients

   – affect the precision of the regression coefficients (inflate the variance)

   – affect the anova table

   – affect the testing on $\beta_j = 0$.

   https://onlinecourses.science.psu.edu/stat501/node/346

2. Data centerization: Chap 10, Question 6

   6. Using the method of least squares, an experimenter fitted the model

   $$\eta = \beta_0 + \beta_1 x \qquad \text{(I)}$$

   to the data below. It was known that σ was about 0.2. A friend suggested it would be better to fit the following model instead:

   $$\eta = \alpha + \beta(x - \bar{x}) \qquad \text{(II)}$$

   where $\bar{x}$ is the average value of the x's.

x	10	12	14	16	18	20	22	24	26	28	30	32
y	80.0	83.5	84.5	84.8	84.2	83.3	82.8	82.8	83.3	84.2	85.3	86.0

   (a) Is $\hat{\alpha} = \hat{\beta}_0$? Explain your answer. (The caret indicates the least squares estimate of the parameter.)
   (b) Is $\hat{\beta} = \hat{\beta}_1$? Explain your answer.
   (c) Are $\hat{y}_{\text{I}}$ and $\hat{y}_{\text{II}}$, the predicted values of the responses for the two models, identical at $x = 40$? Explain your answer.
   (d) Considering the two models above, which would you recommend the experimenter use: I or II or both or neither? Explain your answer.

3. go over the project 3.