Hw 5

Nihal Karim

1. A student says "If extremely influential outlying cases are detected in a data set, simply delete all those cases from the data set." Would you agree? If not, what would you do?

No, this would require domain knowledge to see / understand exactly what those outliers are and understand the cause of them. There may be a case where they are extremely important to the real world data

2. Express the OLS solution for $\hat{\beta}$ in terms of the singular value decomposition of the design matrix $X = UDV^T$. In the case of extreme multicollinearity, the singular values of the design are very close to zero. Explain how this creates instability in the OLS estimator.

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

$$\hat{B} = ((UDV^T)^T \cdot UDV^T)^{-1} UDV^T \cdot y$$

$$\hat{\beta} = (VD^T U^T \cdot UDV^T)^{-1} UDV^T \cdot y$$

$U^T U = I$ (U is orthogonal)

$$\hat{B} = (VD^T \cdot DV^T)^{-1} VD^T U^T y$$

$V^T = V^{-1}$ (orthogonal) $\quad V^T \cdot V = I$

$$\hat{\beta} = V(D^T \cdot D)^{-1} V^T \cdot V D^T U^T y$$

$$\hat{B} = V(D^T D)^{-1} D^T U^T y$$

$$D^T D = \text{diag}(\sigma_1^2, \sigma_2^2 \dots \sigma_p^2)$$

$$(D^T D)^{-1} = \text{diag}(1/\sigma_1^2, 1/\sigma_2^2 \dots)$$

$$(D^T D)^{-1} D^T = \text{diag}(\sigma_i/\sigma_1^2, \sigma_2/\sigma_2^2, \dots$$

$$= \text{diag}(1/\sigma_1, 1/\sigma_2, \dots) = D^{-1}$$

$$\hat{B} = V D^{-1} U^T \cdot y$$

Since $D^{-1} = \text{diag}(1/\sigma_i, 1/\sigma_2 \dots 1/\sigma_p)$ some values of σ_i will be very low in cases of multicollin $1/\sigma_i$ will "blow up" and so small changes in data lead to large changes in estimations (instability)

1. For the dataset `KelleyBlueBookData.csv`, consider using price as the response and regressing against the following predictors: mileage, type, cylinder, liter, cruise, sound, and leather. Treat Leather (0 for not-leather, 1 for leather), Type and Cylinder as categorical variables.

   (a) Report the estimated coefficient of "leather" and interpret the t test result for testing whether or not there is a leather effect. Interpret in the context of the problem to comment on the impact of "leather" to price.

   T stat = 3.873

   coef = 1677.95 P val = 0.00012

   There is strong evidence to suggest that leather has a statistically significant impact on price. On average, cars with leather seats tend to be more expensive by ~$1678

   (b) Look at the coefficients associated with the "Type" variable. Which type was used as the reference level? Which type seems to have the highest average price?

   reference level was the 'Convertible' Since all others are negative relative to it. Convertible was the highest average price

(c) What conclusion you can make about the price when Cylinder=6 compared to other cylinder levels?

Cyl = 6 , coef = 1360.13 , p-val = 0.206

6 cylinders are suggested to be more expensive by ~1360, but since the p-val is high for common significance values it suggests the difference from price to the reference (4cyl) is not significant. This is different dles sin 8 cyl has coef ~$14K and pval almost 0

(d) Run a partial ANOVA, interpret the F test result for Cylinder. Combine the results of t-test and F-test for Cylinder, should we conclude that it's a significant predictor?

While the individual t-test for cyl=6 showed no statistical significance. The F test for cylinder as a whole indicates there is significance across all levels. So yes, including all levels cylinder is a significant predictor

2. Download the data set `IceCreamConsumption.csv` and regress cons against income, price, and temp.

(a) Obtain the variance inflation factors. What do these suggest about the effects of multicollinearity in this model?

The VIF values are all close to 1

Suggest that multicollinearity is not a significant concern in this model

(b) Explain how the VIF for income is calculated step by step.

$$VIF_{(income)} = \frac{1}{1-R^2}$$

$$= 1/1-0.125$$

$$\approx 1.144$$

$income = B_0 + B_1 \cdot Price + B_2 \cdot temp + \varepsilon$

Get R^2 value of this model and plug in

(c) Draw an influence plot of this model where x-axis is the leverage, y-axis is the (externalaly) studentized residuals, and the size of the points are Cook's distance. Which observation has the highest studentized residual? Which observation has the highest leverage? Which observation has highest Cook's distance?

See notebook for plot.

Observation 0 has highest studentized residual

Observation 29 has the highest leverage

and also highest cook's distance

3. Consider the data set `BrandPreference.csv` and a model in which we regress BrandLiking (scale 0-100, 100 being most preferred) against MoistureContent (scale 1-10, 10 being most moist) and Sweetness (scale 1-5, 5 being sweetest). Treat both predictors as numerical variables.

(a) Perform the regression in python and write down the fitted model.

$\hat{Y} = B_0 + B_1 \cdot$ MoistureContent + B_2 Sweetness

$\hat{Y} =$ Brand Liking

(b) Find the fitted value $\hat{y}_1$ for the first observation of the data. Hint: Don't forget that Python indices start at 0.

$\hat{y}_i \approx 64.1$ see notebook

(c) Calculate the hat matrix H, and show that

$$\hat{y}_1 = \sum_{i=1}^{n} h_{1i} y_i.$$

$H = X(X^T X)^{-1} X^T$

See Python notebook for calcs

64.1 ✓