L7. Modeling Problems

During Model Evaluation, we should try to identity modeling problems that could potentially hurt the model performance and inferences. The model problems can be summarized as

- Data Structural Problems:
  - Multicollinearity
  - Influential Points
- Model Assumption violatings:
  - Heteroscedasiticity
  - Non-normality residuals
  - Y and X are not linearly related but modeled that way: False assumption of linearity between y and predictors.

For all the modeling problems, we will break down to

a. Damage it might cause (less on math, more on symptoms)

b. Detection

c. Solution.

In L7, we will start with

I. Multicollinearity

II. Influential Points.

I. Multicollinearity.

Problem: <u>two or more of the predictors are highly correlated</u>, or in a broader description, one of the columns of $\vec{X}^T\vec{X}$ is a linear combination of the others, i.e. $\vec{X}^T\vec{X}$ is close to non-invertible.

(a) Damage it might cause:

Recall: $\vec{b} = (\vec{X}^T\vec{X})^{-1}\vec{X}^T\vec{y}$ ← becomes very "large"

sensitive to the change in $\vec{X}$.

$E(\vec{b}) = \vec{\beta}$

$Var(\vec{b}) = \sigma^2(\vec{X}^T\vec{X})^{-1}$ ← becomes very "large"

- The wefficient estimates:

can swing wildly based on which other independent variables are in the model. The coefficient $\vec{b}$ beom very sensitive to small changes in the model.

$\boxed{\text{symptom}}$: when you add and drop variables. there are huge changes to the value of coefficients, sometimes even the sign of the fitted coefficients

$\boxed{\text{Ex}}$:

A perfect data not showing Multicollineanity at all:

Model	$\hat{\beta}_1$
$y \sim X_1$	-1
$y \sim X_1 + X_2$	-1

Model	$\hat{\beta}_2$
$y \sim X_2$	-1.75
$y \sim X_1 + X_2$	-1.75

Adding X_2 or not, doesn't change the estimate of $\hat{\beta}_1$

A data with serious multicollineanity.

	$\hat{\beta}_1$
$y \sim X_1$	1.201
$y \sim X_1 + X_2$	1.039

$y \sim X_2$	34.44
$y \sim X_1 + X_2$	5.83

adding X_1 changed $\hat{\beta}_2$ a lot.

- Test result unreliable due to the inflation in $se(\hat{\beta}_K)$

(1) In t test: $t_{stat} = \dfrac{\hat{\beta}_K}{se(\hat{\beta}_K)\uparrow}$ $\downarrow$

the inflation in $se(\hat{\beta}_K)$ caused smaller t statistics, harder to have a rejection / significant predictor when having a lot of predictors in the model.

$\boxed{\text{symptom}}$: There are very few coefficients show significance from t test.

(2) In anova (typ=1), by changing the order of variables, the significance might change dramatically.

$\boxed{\text{Ex}}$

Data without Multicollinearity.

Model	$se(\hat{\beta}_1)$	t_{stat} for $\hat{\beta}_1$	ANOVA (typ=1)
$y \sim X_1$	1.47	-0.68	$SS(X_1 \mid Null) = 8$
$y \sim X_2 + X_1$	1.41	-0.71	$SS(X_1 \mid X_2) = 8.$

Data with serious Multicolinearity

Model	$se(\hat{\beta}_1)$	tstat	ANOVA (typ=1)
$y \sim X_1$	0.093	12.92	$SS(X_1 \mid Null) = 505.5$
$y \sim X_2 + X_1$	0.193	5.39	$SS(X_1 \mid X_2) = 88.43$

- Note about Multicollinearity:

(1) In reality, it always exists. When it's not too "serious" we don't need to worry about it too much.

(2) When the problem is serious, we can still use the model to do the prediction, since all estimation are unbiased. But the std. err of prediction would increase;

It affects more in the model inference: testing. It becomes difficult for us to examine the relationship between each predictor and response variable independently.

(b) Detection:

To detect Multicollinearity in the data.

(1) Plot - heatmap of pair-wise correlation between predictors (Naive!! but pretty)

Ex: Credit.csv.

leave out the response variable "Balance"

some correlations that might be tricky:

"Income" v.s. "Limit" (0.79)

"Income" v.s. "Rating" (0.79)

Normally we say correlation ≥ 0.8 is strong.

those are on the boundaries.

"Limit" v.s. "Rating" (1!)

(2) Variance Inflation Factor (VIF) - most common.

VIF measures how much the variance is inflated in the coefficient estimates.

- VIF_j for $\hat{\beta}_j$ is

$$VIF_j = \frac{1}{1-R_j^2}$$

where R_j^2 is the R^2-value obtained by fitting the model without X_j.

Match with textbook: R_j^2 is the R^2 value obtained by regressing X_j again the rest p-2 predictors

i.e. $y \sim X_1 + X_2 + X_3$

$VIF_1 = \frac{1}{1-R_1^2}$ where R_1^2 is

the R^2 value from $y \sim X_2 + X_3$

Recall $R^2 = \frac{SSR}{SST}$ ← without X_1, how much variation in the model has lost.

SST ← stay the same

- When $VIF_j = 1$: we don't lose any variation without X_j, A.K.A. it's not correlated with any other predictors. / the variance in X_j is not inflated at all.
- When $1 < VIF_j \leq 4$: light.

  $4 < VIF_j \leq 10$: moderate

  } fine.
- When $VIF_j > 10$: X_j is highly ~~correlated with other predictors~~ impacted by the multicolinearity.

Ex: Credit.csv.

"Limit" "Rating" and "Eduction"

are highly inflated by the multicolinearity.

Recall from the heatmap; "Limit" and "Rating" are

highly correlated.

(c) Solutions.

Fast solutions: • drop some of the highly correlated predictors

based on practical understanding;

i.e. I dropped "Limit" since I feel

"Rating" is more important to stay.

• linearly combine some independent variables

together when it make sense: like when

they are on the same scale. (dimension

depth + length + width)

or define a new one:

use BMI ~~fat~~ instead of height and weight.

More complex solutions (~~More~~ Not include here):

- Use Regularization (Machine learning) to standarize the data

- Use other regression methods:

  Principle Components Analysis / Partial Least squares.

Note: (1) Don't drop intercept easily. It's the constant that adjust the scale of your response.

(2) If dropping one or two variables don't help, you don't have to drop more.

It's a balance between inflation and information!

(battle between overfitting and underfitting again)

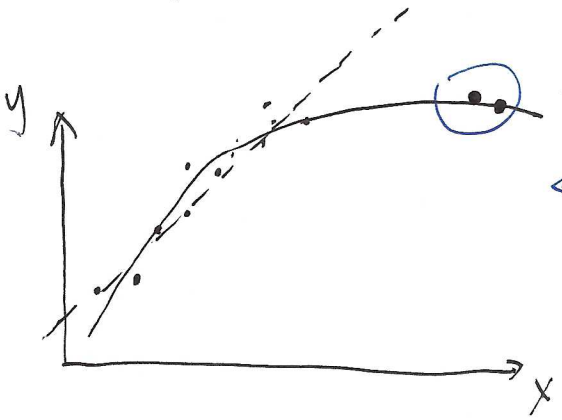
II. Influential Points

Data points can potentially be influential to your model in different ways:

(i) Outlier: an observation has a $\boxed{\text{response value}}$ y_i that is very far from the fitted value $\hat{y}_i$.

(ii) High leverage point: an observation has a particular unusual combination of $\boxed{\text{predictor values}}$ is said to have high leverage, that could influence $\hat{y}_i$

(ii) Observations ~~wh~~ that's an outlier with high leverage.



← outlier

← high leverage

Influential or not needed to be measured for the impact on your model.

(a) Prolem: The influential points significantly affect the model estimation: with or without the observations lead to significantly different models.

(b) Detection

(i) Just identifying outliers: obs with y_i "far away" from $\hat{y}_i$

- Ordinary residuals? restricted by the units of measurements

  i.e. on weight, $e_i = 20$ (lb)

  on length, $e_i = 1$ (inch)

  which one is "bigger"?

- Standarize the residuals:

  — semi-studentized residuals: $e_i^* = \dfrac{e_i}{\sqrt{MSE}}$

  "rule of thumb": $|e_i^*| > 3$ can be identified as outlier.

(ii) Just identify high leverage: obs with X values that impact the $\hat{y}_i$ a lot.

Recall:

$$\vec{y} = \vec{X}\vec{\beta} + \vec{\varepsilon}$$

$$\vec{b} = (\vec{X}^T\vec{X})^{-1}\vec{X}^T\vec{y}$$

$$\hat{\vec{y}} = \vec{X}\vec{b} = \underbrace{\vec{X}(\vec{X}^T\vec{X})^{-1}\vec{X}^T}\vec{y} = \vec{H}\vec{y}$$

$\vec{H}$: "hat" matrix" calculated by $\vec{X}$

and measure when $\vec{y}$ changes, how much it changes $\hat{\vec{y}}$.

look closer:

$$\begin{pmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{pmatrix} = \begin{pmatrix} h_{11} & h_{12} & \cdots & h_{n} \\ h_{21} & h_{22} & \cdots & h_{2n} \\ \vdots & \vdots & & \vdots \\ h_{n1} & h_{n2} & \cdots & h_{nn} \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

$$\hat{\vec{y}} = \vec{H} \quad * \quad \vec{y}$$

Therefore: $\hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + \cdots + h_{in}y_n$

- h_{ii} quantifies the influence that observed y_i has on its predicted value $\hat{y}_i$.
- h_{ii} completed decided by $\vec{X}$
- If h_{ii} is large, y_i plays a big role in estimating $\hat{y}_i$

h_{ii} is defined as the leverage of the ith observation.

- Properties of leverage (No math proof required)

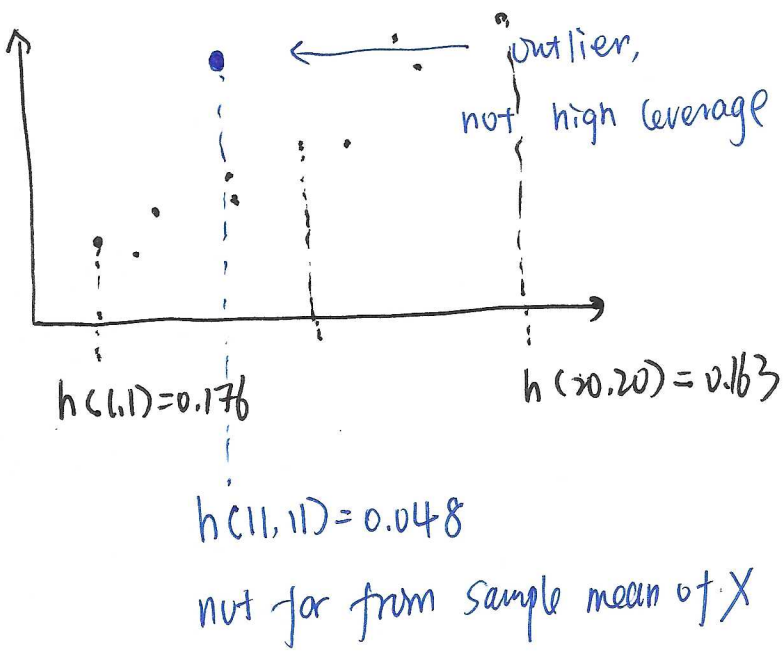
$$h_{ii} = \vec{H}_{i,i} = \left(\vec{X}(\vec{X}^T\vec{X})^{-1}\vec{X}^T \right)_{i,i}$$

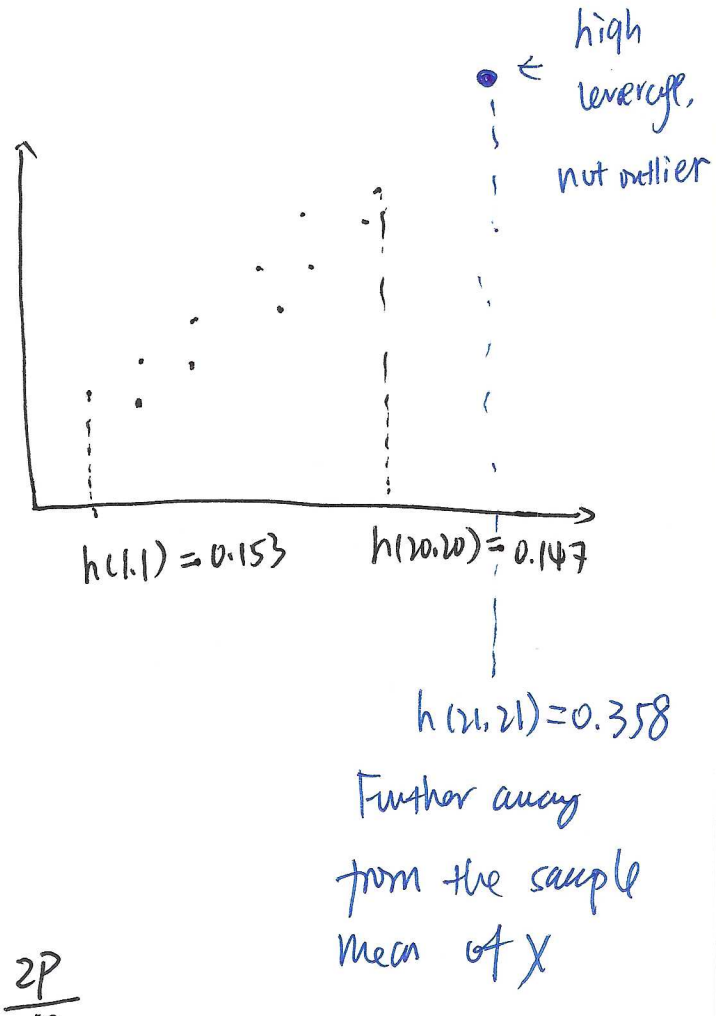
(i) $0 \le h_{ii} \le 1$

(ii) $\sum h_{ii} = p$

(iii) $\bar{h} = \frac{p}{n}$

(iv) h_{ii} measure the "distance" between ~~[illegible]~~ $\vec{x}_i$ and $\bar{\vec{X}}$.





- Rule of Thumb: Obs with $h_{ii} > \frac{2p}{n}$

flag as high leverage points.

(iii) Just being an outlier or high leverage points doesn't necessarily mean it's disturbing the model. We care more about how the observation influence the regression model — Influence Points

- Externaly Studentized Residuals

Recall semi-studentized residuals : $e_i^* = \frac{e_i}{\sqrt{MSE}}$

"semi" because MSE is not the estimate of $Var(e_i)$, i.e., $Var(e_i) \neq \sigma^2$

Recall $\vec{e} = \vec{y} - \hat{\vec{y}} = (I_n - \vec{H})\vec{y}$.

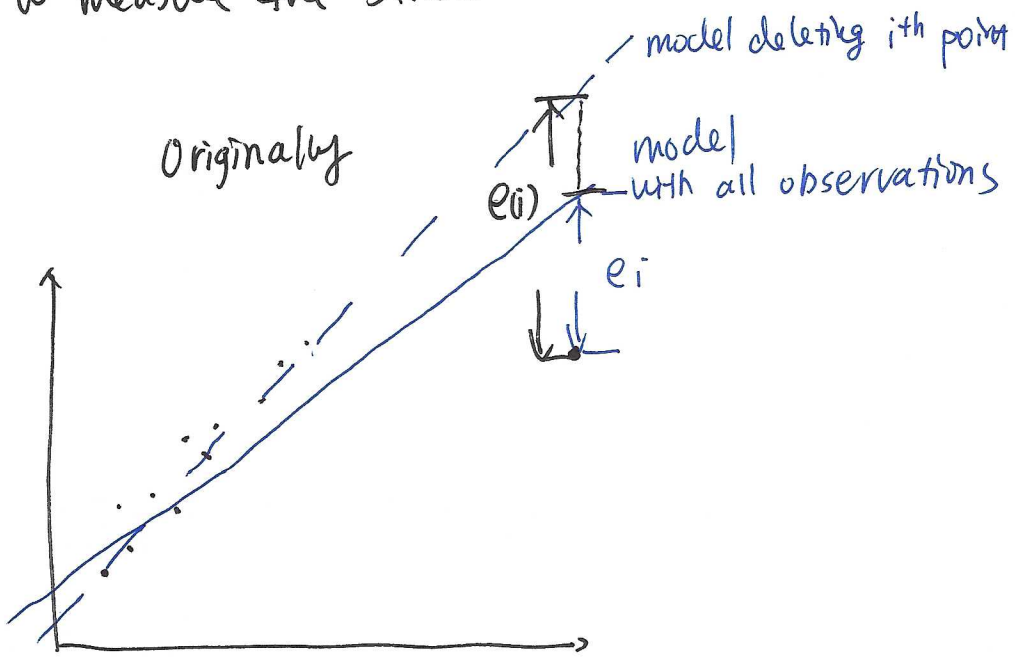
then $Var(\vec{e}) = Var[(I_n - \vec{H})\vec{y}] = (I_n - \vec{H}) \underbrace{Var(\vec{y})}_{= \sigma^2 I_n}$

$= \sigma^2 (I_n - \vec{H})$

- so we can estimate $Var(\vec{e})$ by $MSE(I_n - \vec{H})$
- $\boxed{\widehat{Var}(e_i) = MSE(1 - h_{ii})}$
- The "read" Studentized Residual $= \frac{e_i}{\sqrt{MSE(1-h_{ii})}}$
  (Internal)

Now we are trying to measure how much an obs impacts the regression line. When the influence is substantial, the line is "pulled" towards the observation. So we can measure this by deleting one obs at a time and fitting the model on the remaining (n-1) observations. then use $e_{(i)} = y_i - \hat{y}_{(i)}$ to measure the studentized residual.

from the model with i^{th} obs. deleted.



The $\boxed{\text{Externally Studentized Residual}}$ is defined as

$$studs(e_i) = \frac{e_{(i)}}{S(e_{(i)})}$$

Text book p395~396

$$S(e_{(i)}) = \sqrt{\frac{MSE_{(i)}}{1-h_{ii}}}$$

$$e_{(i)} = \frac{e_i}{1-h_{ii}}$$

Note: index i indicates the model using all data
(i) indicates the model using data deleting obs. i.

Textbook p 396.

$$(n-p)MSE = (n-p-1)MSE_{(i)} + \frac{e_i^2}{1-h_{ii}}$$

$$\text{studs}(e_i) = \frac{e_{(i)}}{\sqrt{\frac{MSE_{(i)}}{1-h_{ii}}}}$$

(model with data deleting obs i)

or

$$= \frac{e_i}{\sqrt{\frac{SSE(1-h_{ii}) - e_i^2}{n-p-1}}}$$

(model with full data)

- Rule of using studs(e_i) to identify outliers with high influence: $\text{studs}(e_i) \sim t(\underbrace{n}_{\text{sample size}}-1-p)$

When $|\text{studs}(e_i)| \geq t_{\alpha/2, df = n-1-p}$, we can detect the ith obs. as an outlier with high influence.

Another might be more common one to detect influential points:

$\boxed{\text{Cook's distance}}$.

It again consider refitting the data without observation i, and it's a combination of e_i and h_{ii}:

The COOK's distance D_i for i^{th} obs:

$$D_i = \frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}_{j(i)})^2}{p\,MSE} = \frac{e_i^2}{p\,MSE} \cdot \frac{h_{ii}}{(1-h_{ii})^2}$$

- D_i measures directly how much ALL the fitted values changes after deleting i^{th} observation.

- Rule of thumbs: flag the observation when $D_i > 1$ (practical)

  Correction: Di >4/n

  the exact comparison can be decided by a F distribution, not discussed here.

Summary :

- Multicolinearity

Problem	Detection	Solution (so far)
• unstable estimation of $\vec{b}$	• Heatmap plot for correlations in X	• Delete one of the highly correlated variables
• unreliable t test : harder to have rejections/ significant predictors	• VIF > 10	• Combine some of the variables together if applies.
• unreliable anova test.		

- Influential Points

Problem	Detection	Solution
• Observations that are outliers or with high leverage might influence the fitted model	• $\lvert studs(e_i) \rvert \geq t_{\alpha/2}, df = n-1-p$	• Report the observations
	• cook's distance $D_i > 1$	• Report the model result with and without the influential points.

Correction: Di >4/n