Lec 12.

Last class

Summaries in ANOVA table

Source	SS	df	Mean Squares → MS	F
Regression	SS(reg)	p	$MS(reg) = SS(reg)/p$	$MS(reg)/MS(res)$
Residual	SS(res)	$n-p-1$	$MS(res) = SS(res)/n-p-1$	
Total	SS(tot)	$n-1$		

$$SS(tot) = SS(res) + SS(reg)$$
$$\sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}(y_i-\hat{\mu}_i)^2 + \sum_{i=1}^{n}(\hat{\mu}_i-\bar{y})^2$$

The ANOVA table allows us to test for overall significance of our model.
($H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$ vs. H_a: at least one of $\beta_1, \dots, \beta_p$ is not 0).

Note: For general linear hypothesis (i.e. $H_0: A\vec{\beta} = 0$)

- Constraint A is a $\ell \times (p+1)$ matrix is ℓ
- Careful not to have redundant constraints : rank$(A) = \ell$
  → make sure rows of A are linearly independent

Multicollinearity

Recall: $\hat{\vec{\beta}} = (X^T X)^{-1} X^T \vec{Y}$

Consider $\vec{Y} = X\vec{\beta} + \vec{\varepsilon}$ where X includes $\{\vec{1}, \vec{x}_1, \vec{x}_2, \vec{x}_3\}$

Suppose that $\vec{x}_3 = \alpha_0 \vec{1} + \alpha_1 \vec{x}_1 + \alpha_2 \vec{x}_2$. That is, $\vec{x}_3$ is a linear combination of other columns of X $\therefore$ Columns of X linearly dependent.

- In this case, we have perfect multicollinearity.
- In LS estimation, we cannot estimate $(X^T X)^{-1}$.
- Intuition: $\vec{x}_3$ cannot explain anything that is not already explained by $\vec{x}_1$ and $\vec{x}_2$.
  ($\vec{x}_3$ does not add any additonal info.).

General Multicollinearity

- Occur when some covariates are highly correlated w/ other covariates.
  e.g. $\vec{x}_3 \approx \alpha_0 \vec{1} + \alpha_1 \vec{x}_1 + \alpha_2 \vec{x}_2$. (columns of X are closely linearly dependent).
- In practice, almost no information is added from including $\vec{x}_3$ given $\vec{x}_1$ and $\vec{x}_2$ are already in model.

* Cause $Var(\hat{\beta}_j)$ to be inflated. This can cause inaccurate inference (e.g. conclusions that we make about hypothesis tests about parameters; CI).
- As a result, $SE(\hat{\beta}_j)$ can change drastically w/ inclusion/omission of some Variable
- Recall $Var(\hat{\vec{\beta}}) = \sigma^2 (X^T X)^{-1}$ $\quad A^{-1} = \frac{1}{\det(A)} (\dots)$
  - $\Rightarrow$ When a matrix A is non-invertible, determinant of A is 0.
  - $\Rightarrow$ When a matrix A is close to being non-invertible, " " close to 0.
- Intuition: Hard to separate variability explained by correlated variables
  ($\Rightarrow$ larger uncertainty w/ parameter estimates)

Examples Suppose we have the following covariates:

1. x_1 = height in cm
2. x_2 = height in inches

$\Rightarrow x_1 = 2.54 x_2$

3. x_3 = Income from 1st half of year
4. x_4 = " 2nd half " "
5. x_5 = total income in a year.

$\Rightarrow x_5 = x_3 + x_4$.

Examples of perfect multicollinearity.

Example : Hospital data. (example of general multicollinearity).

"Beds" and "Census" are highly correlated. The higher the # of patients, the higher # of hospital beds in use.

Q: How do we detect multicollinearity?

1. Scatterplot matrix (all pairwise scatterplots of variables)
2. Calculate correlation matrix (all pairwise correlations b/t variables).
3. In general (>2 predictors that are highly correlated), we use variance inflation factor (VIF).

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

where R_j^2 is the R^2 value from a regression of x_j on other explanatory variables.

VIF in more detail.

Suppose that $Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$, $\varepsilon_i \overset{iid}{\sim} N(0,\sigma^2)$

$$r_{XX} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{12} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1p} & r_{2p} & \cdots & 1 \end{bmatrix}$$

correlation matrix of $\vec{x}_1, \ldots, \vec{x}_p$

s_{x_j} : sample std. deviation of x_j

s_Y : sample std. deviation of Y.

Consider the following transformation :

$$x_{ij}^* = \left(\frac{x_{ij} - \bar{x}_j}{s_{x_j}}\right)\frac{1}{\sqrt{n-1}} \quad ; \quad Y_i^* = \left(\frac{Y_i - \bar{Y}}{s_Y}\right)\frac{1}{\sqrt{n-1}}$$

Instead, fit $Y_i^* = \beta_1^* x_{i1}^* + \cdots + \beta_p^* x_{ip}^* + \varepsilon_i^*$, $\varepsilon_i^* \overset{iid}{\sim} N(0,\sigma^{*2})$

(LS estimation always give an estimate of β_0^* of 0).

$$X^* = \begin{bmatrix} | & | & & | \\ \vec{x}_1^* & \vec{x}_2^* & \cdots & \vec{x}_p^* \\ | & | & & | \end{bmatrix} \quad ; \quad X^{*T}X^* = r_{XX} \text{ (exercise)}$$

Then, $Var(\hat{\beta}^*) = \sigma^{*2}(r_{XX}^{-1})$

When $p=2$, then $Y_i^* = \beta_1^* x_{i1}^* + \beta_2^* x_{i2}^* + \varepsilon_i^*$

$$r_{xx} = \begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix} \quad \text{and} \quad r_{xx}^{-1} = \begin{bmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{bmatrix} \frac{1}{1-r_{12}^2}$$

$\Rightarrow$ If $r_{12} = 0$, then $Var(\hat{\beta}_i^*) = \sigma^{*2}$ since $r_{xx}^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

$\Rightarrow$ If $r_{12} \neq 0$ (say close to 1), the diagonal of r_{xx}^{-1} will be large and inflated by a factor of $\frac{1}{1-r_{12}^2}$.

$\Rightarrow$ Then $Var(\hat{\beta}_1^*) = \sigma^{*2} \underbrace{\frac{1}{1-r_{12}^2}}_{VIF_1}$

More generally, $\underline{VIF_j = \frac{1}{1-R_j^2}}$

- Let's think about R_j^2 by considering the regression of x_j on other explanatory variables
- Consider the correlation b/t x_j and $\hat{x}_j$ (fitted values of x_j)
- Recall in SLR, $r_{xy}^2 = R^2$; r_{xy} is the sample correlation b/t x and y.
- In MLR, $r_{y,\hat{\mu}}^2 = R^2$ (assignment 2 problem); correlation b/t y and fitted values of y

$$r_{y,\hat{\mu}}^2 = \frac{\left[\sum_{i=1}^n (y_i - \bar{y})(\hat{\mu}_i - \bar{\hat{\mu}})\right]^2}{\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (\hat{\mu}_i - \bar{\hat{\mu}})^2}$$

$$\text{(show)} = SS(reg)/SS(tot) = R^2$$

Hint : show that $\sum_{i=1}^n (\hat{\mu}_i - \bar{y})(y_i - \hat{\mu}_i) = 0$

- $\therefore r_{x_j,\hat{x}_j}^2 = R_j^2$ (R_j^2 is R^2 values from regression of x_j on other predictors).
- Intuition: the closer R_j^2 is to 1, this implies x_j may be highly correlated w/ other predictors.

Notes :
- Since R_j^2 is always in $[0,1]$, this implies that $VIF \geq 1$.
- $SE(\hat{\beta}_j)$ is larger when R_j^2 is larger ($\because 1-R_j^2$ is smaller).

☆ Rule of thumb: If $VIF_j \geq 10$, this implies strong multicollinearity. ($R_j^2 \geq 0.9$).

☆ Procedure: remove predictors with large VIF and repeat process until no more strong multicollinearity.