Lecture 7

Multiple Linear Regression III

Reading: Chapter 13

STAT 8020 Statistical Methods II
September 10, 2020

Whitney Huang
Clemson University

Agenda

Multicollinearity

Multicollinearity is a phenomenon of high inter-correlations among the predictor variables

- Numerical issue $\Rightarrow$ the matrix $\boldsymbol{X}^T\boldsymbol{X}$ is nearly singular
- Statistical issue
  - β's are not well estimated
  - Spurious regression coefficient estimates
  - R^2 and predicted values are usually OK

Example

- Consider a two predictor model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

- We can show

$$\hat{\beta}_{1|2} = \frac{\hat{\beta}_1 - \sqrt{\frac{\hat{\sigma}_Y^2}{\hat{\sigma}_{X_1}^2}} r_{X_1,X_2} r_{Y,X_2}}{1 - r_{X_1,X_2}^2},$$

where $\hat{\beta}_{1|2}$ is the estimated partial regression coefficient for X_1 and $\hat{\beta}_1$ is the estimate for β_1 when fitting a simple linear regression model $Y \sim X_1$

An Simulated Example

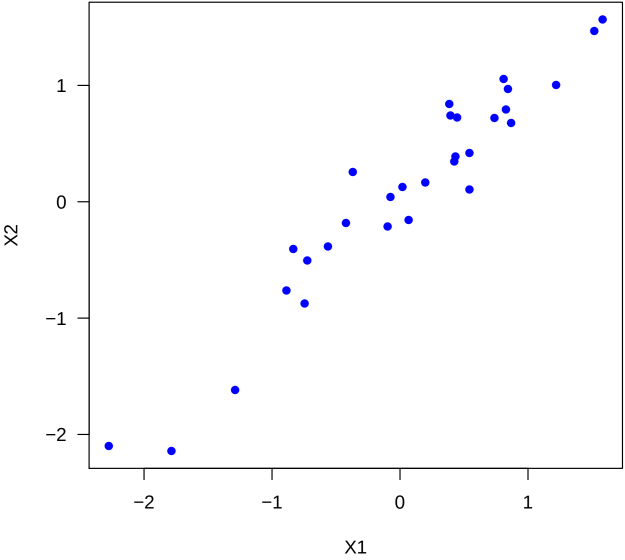
Suppose the true relationship between response Y and predictors (X_1, X_2) is

$$Y = 4 + 0.8X_1 + 0.6X_2 + \varepsilon,$$

where $\varepsilon \sim \mathrm{N}(0, 1)$ and X_1 and X_2 are positively correlated with $\rho = 0.95$. Let's fit the following models:

- Model 1: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$
- Model 2: $Y = \beta_0 + \beta_1 X_1 + \varepsilon_1$
- Model 3: $Y = \beta_0 + \beta_2 X_2 + \varepsilon_2$

Scatter Plot: X_1 vs. X_2



Model 1 Fit

```
Call:
lm(formula = Y ~ X1 + X2)

Residuals:
     Min       1Q   Median       3Q      Max
-1.91369 -0.73658  0.05475  0.87080  1.55150

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.0710     0.1778  22.898  < 2e-16 ***
X1            2.2429     0.7187   3.121  0.00426 **
X2           -0.8339     0.7093  -1.176  0.24997
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9569 on 27 degrees of freedom
Multiple R-squared:  0.673,     Adjusted R-squared:  0.6488
F-statistic: 27.78 on 2 and 27 DF,  p-value: 2.798e-07
```

Model 2 Fit


```
Call:
lm(formula = Y ~ X1)

Residuals:
     Min       1Q   Median       3Q      Max
-2.09663 -0.67031 -0.07229  0.87881  1.49739

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.0347     0.1763  22.888  < 2e-16 ***
X1            1.4293     0.1955   7.311 5.84e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9634 on 28 degrees of freedom
Multiple R-squared:  0.6562,    Adjusted R-squared:  0.644
F-statistic: 53.45 on 1 and 28 DF,  p-value: 5.839e-08
```

Model 3 Fit


```
Call:
lm(formula = Y ~ X2)

Residuals:
    Min      1Q  Median      3Q     Max
-2.2584 -0.7398 -0.3568  0.8795  2.0826

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.9882     0.2014   19.80  < 2e-16 ***
X2            1.2973     0.2195    5.91 2.33e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.096 on 28 degrees of freedom
Multiple R-squared:  0.555,	Adjusted R-squared:  0.5391
F-statistic: 34.92 on 1 and 28 DF,  p-value: 2.335e-06
```

Variance Inflation Factor (VIF)

We can use the variance inflation factor (VIF)

$$\mathsf{VIF}_i = \frac{1}{1 - \mathsf{R}_i^2}$$

to quantifies the severity of multicollinearity in MLR, where R_i^2 is the **coefficient of determination** when X_i is regressed on the remaining predictors

Variable Selection

Multiple Linear Regression Model:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1} + \varepsilon_i, \quad \varepsilon_i \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma^2)$$

- What is the appropriate subset size?
- What is the best model for a fixed size?

In the next few slides we will discuss some commonly used model selection criteria

Mallows' C_p Criterion

$$
\begin{aligned}
(\hat{Y}_i - \mu_i)^2 &= (\hat{Y}_i - \mathrm{E}(\hat{Y}_i) + \mathrm{E}(\hat{Y}_i) - \mu_i)^2 \\
&= \underbrace{(\hat{Y}_i - \mathrm{E}(\hat{Y}_i))^2}_{\text{Variance}} + \underbrace{(\mathrm{E}(\hat{Y}_i) - \mu_i)^2}_{\text{Bias}^2},
\end{aligned}
$$

where $\mu_i = \mathrm{E}(Y_i | X_i = x_i)$

- Mean squared prediction error (MSPE): $\sum_{i=1}^n \sigma^2_{\hat{Y}_i} + \sum_{i=1}^n (\mathrm{E}(\hat{Y}_i) - \mu_i)^2$
- C_p criterion measure:

$$
\begin{aligned}
\Gamma_p &= \frac{\sum_{i=1}^n \sigma^2_{\hat{Y}_i} + \sum_{i=1}^n (\mathrm{E}(\hat{Y}_i) - \mu_i)^2}{\sigma^2} \\
&= \frac{\sum \text{Var}_{\text{pred}} + \sum \text{Bias}^2}{\text{Var}_{\text{error}}}
\end{aligned}
$$

C_p Criterion

- Do not know σ^2 nor numerator
- Use $\text{MSE}_{X_1,\cdots,X_{p-1}} = \text{MSE}_\text{F}$ as the estimate for σ
- For numerator:
  - Can show $\sum_{i=1}^n \sigma^2_{\hat{Y}_i} = p\sigma^2$
  - Can also show $\sum_{i=1}^n (\text{E}(\hat{Y}_i) - \mu_i)^2 = \text{E}(\text{SSE}_\text{F}) - (n-p)\sigma^2$

$$\Rightarrow C_p = \frac{\text{SSE} - (n-p)\text{MSE}_\text{F} + p\text{MSE}_\text{F}}{\text{MSE}_\text{F}}$$

C_p Criterion Cont'd

Recall

$$\Gamma_p = \frac{\sum_{i=1}^n \sigma^2_{\hat{Y}_i} + \sum_{i=1}^n (\mathrm{E}(\hat{Y}_i) - \mu_i)^2}{\sigma^2}$$

- When model is correct $\mathrm{E}(C_p) \approx p$
- When plotting models against p
  - Biased models will fall above $C_p = p$
  - Unbiased models will fall around line $C_p = p$
  - By definition: C_p for full model equals p

Adjusted R^2 Criterion

Adjusted R^2, denoted by R^2_{adj}, attempts to take account of the phenomenon of the R^2 automatically and spuriously increasing when extra explanatory variables are added to the model.

$$R^2_{\text{adj}} = 1 - \frac{\text{SSE}/(n-p-1)}{\text{SST}/(n-1)}$$

- Choose model which maximizes R^2_{adj}
- Same approach as choosing model with smallest MSE

Predicted Residual Sum of Squares $PRESS$ Criterion

- For each observation i, predict Y_i using model generated from other $n-1$ observations
- $PRESS = \sum_{i=1}^{n}(Y_i - \hat{Y}_{i(i)})^2$
- Want to select model with small $PRESS$

Other Approaches: Information criteria

- Akaike's information criterion (AIC)

$$n \log\left(\frac{\text{SSE}_k}{n}\right) + 2k$$

- Bayesian information criterion (BIC)

$$n \log\left(\frac{\text{SSE}_k}{n}\right) + k \log(n)$$

- Can be used to compare **non-nested** models

Automatic Search Procedures

- Forward Selection
- Backward Elimination
- Stepwise Search
- All Subset Selection

Model Assumptions

Model:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1} + \varepsilon_i, \quad \varepsilon_i \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma^2)$$

We make the following assumptions:

- Linearity:

$$\mathrm{E}(Y|X_1, X_2, \cdots, X_{p-1}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1}$$

Model Assumptions

Model:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1} + \varepsilon_i, \quad \varepsilon_i \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma^2)$$

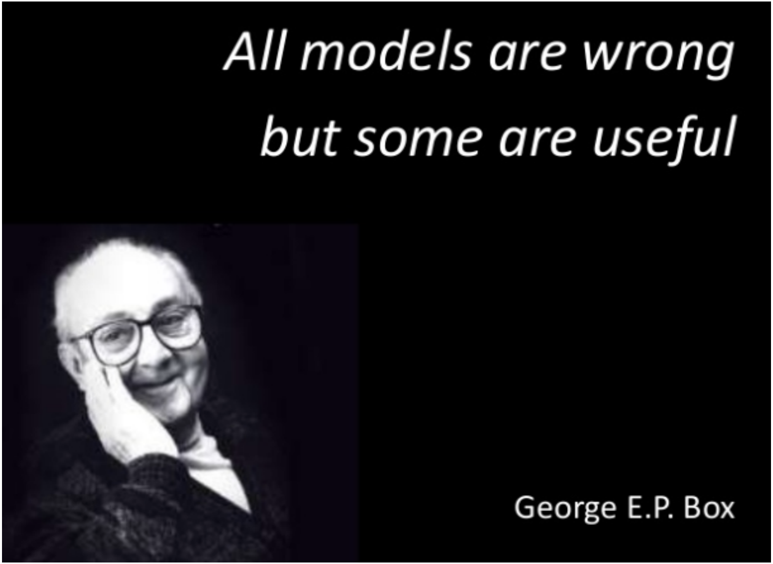
We make the following assumptions:

- Linearity:

$$\mathrm{E}(Y|X_1, X_2, \cdots, X_{p-1}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1}$$

- Errors have constant variance, are independent, and normally distributed

$$\varepsilon_i \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma^2)$$


All models are wrong
but some are useful
George E.P. Box