Lecture 6

Multiple Linear Regression II

Reading: Chapter 12

STAT 8020 Statistical Methods II
September 8, 2020

Whitney Huang
Clemson University

Agenda

1. General Linear Test
2. Multicollinearity
3. Variable Selection Criteria

Review: Coefficient of Determination

- Coefficient of Determination R^2 describes proportional of the variance in the response variable that is predictable from the predictors

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}, \quad 0 \le R^2 \le 1$$

- R^2 usually increases with the increasing p, the number of the predictors
  - Adjusted R^2, denoted by $R^2_{\text{adj}} = 1 - \frac{\text{SSE}/(n-p)}{\text{SST}/(n-1)}$ attempts to account for p

R^2 vs. R^2_{adj} Example

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

$$Y = 5 + 2X_1 + \varepsilon,$$

where $\varepsilon \sim$ N(0, 1) and X_1 and X_2 are independent to each other. Let's fit the following two models to the "data"

Model 1: $Y = \beta_0 + \beta_1 X_1 + \varepsilon^1$

Model 2: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon^2$

Question: Which model will "win" in terms of R^2?

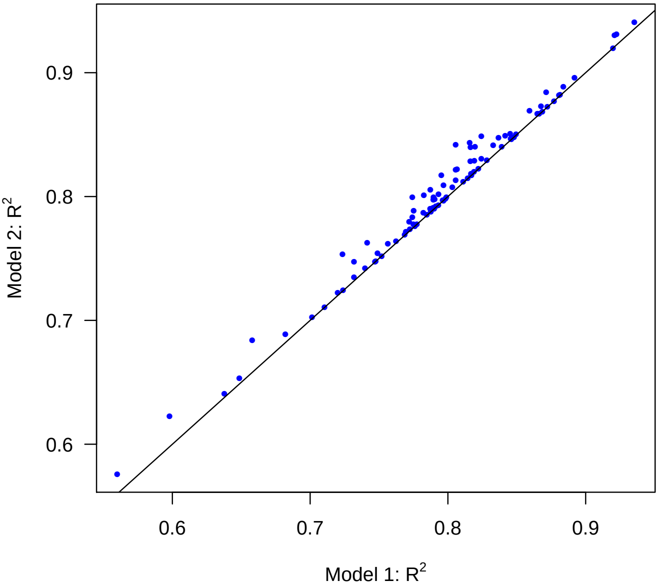
Model 1 Fit

```
> summary(fit1)

Call:
lm(formula = y ~ x1)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.6085 -0.5056 -0.2152  0.6932  2.0118 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   5.1720     0.1534   33.71  < 2e-16 ***
x1            1.8660     0.1589   11.74 2.47e-12 ***
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8393 on 28 degrees of freedom
Multiple R-squared:  0.8313,	Adjusted R-squared:  0.8253 
F-statistic:   138 on 1 and 28 DF,  p-value: 2.467e-12
```

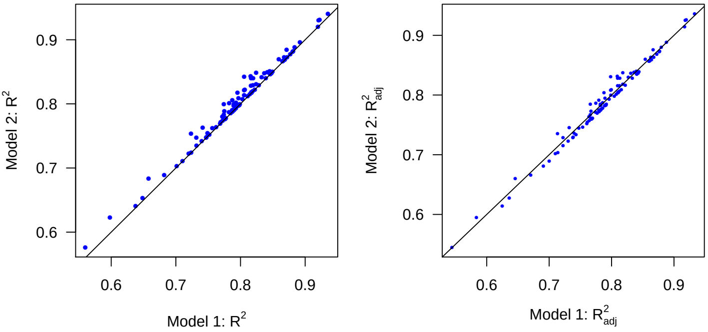
Model 2 Fit

```
> summary(fit2)

Call:
lm(formula = y ~ x1 + x2)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.3926 -0.5775 -0.1383  0.5229  1.8385 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   5.1792     0.1518  34.109  < 2e-16 ***
x1            1.8994     0.1593  11.923 2.88e-12 ***
x2           -0.2289     0.1797  -1.274    0.213    
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8301 on 27 degrees of freedom
Multiple R-squared:  0.8408,	Adjusted R-squared:  0.8291 
F-statistic: 71.32 on 2 and 27 DF,  p-value: 1.677e-11
```

R^2: Model 1 vs. Model 2



R^2_{adj}: Model 1 vs. Model 2



General Linear Test

- Comparison of a "full model" and "reduced model" that involves a subset of full model predictors
- Consider a full model with k predictors and reduced model with ℓ predictors ($\ell < k$)
- Test statistic: $F^* = \frac{\text{SSE(R)} - \text{SSE}(F)/(k-\ell)}{\text{SSE}(F)/(n-k-1)} \Rightarrow$ Testing H_0 that the regression coefficients for the extra variables are all zero
  - Example 1: $X_1, X_2, \cdots, X_{p-1}$ vs. intercept only $\Rightarrow$ Overall F test
  - Example 2: $X_j, 1 \leq j \leq p-1$ vs. intercept only $\Rightarrow$ t test for β_j
  - Example 3: X_1, X_2, X_3, X_4 vs. $X_1, X_3 \Rightarrow H_0 : \beta_2 = \beta_4 = 0$

Species Diversity on the Galapagos Islands Revisited: Full Model

```
> summary(gala_fit2)

Call:
lm(formula = Species ~ Elevation + Area)

Residuals:
     Min       1Q   Median       3Q      Max
-192.619  -33.534  -19.199    7.541  261.514

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.10519   20.94211   0.817  0.42120
Elevation    0.17174    0.05317   3.230  0.00325 **
Area         0.01880    0.02594   0.725  0.47478
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 79.34 on 27 degrees of freedom
Multiple R-squared:  0.554,     Adjusted R-squared:  0.521
F-statistic: 16.77 on 2 and 27 DF,  p-value: 1.843e-05
```

Species Diversity on the Galapagos Islands Revisited: Reduce Model

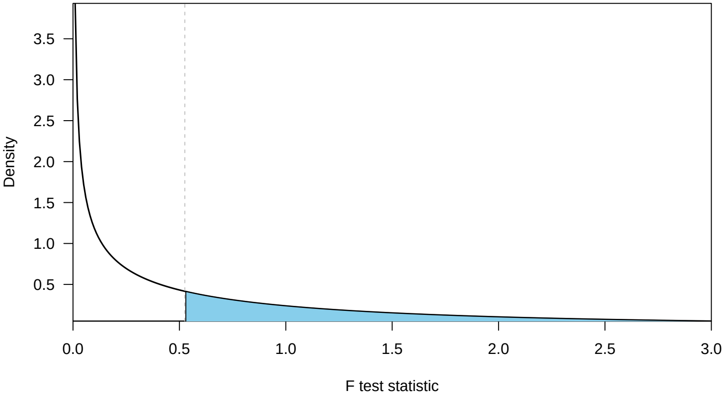
```
> summary(gala_fit1)

Call:
lm(formula = Species ~ Elevation)

Residuals:
     Min       1Q   Median       3Q      Max 
-218.319  -30.721  -14.690    4.634  259.180 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 11.33511   19.20529   0.590     0.56    
Elevation    0.20079    0.03465   5.795 3.18e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 78.66 on 28 degrees of freedom
Multiple R-squared:  0.5454,    Adjusted R-squared:  0.5291 
F-statistic: 33.59 on 1 and 28 DF,  p-value: 3.177e-06
```

Perform a General Linear Test

- $H_0 : \beta_{\texttt{Area}} = 0$ vs. $H_a : \beta_{\texttt{Area}} \neq 0$
- $F^* = \frac{(173254-169947)/(2-1)}{169947/(30-2-1)} = 0.5254$
- P-value: $\mathrm{P}[F > 0.5254] = 0.4748$, where $F \sim \mathsf{F}(1, 27)$

```
> anova(gala_fit1, gala_fit2)
Analysis of Variance Table

Model 1: Species ~ Elevation
Model 2: Species ~ Elevation + Area
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     28 173254
2     27 169947  1      3307 0.5254 0.4748
```

P-value Calculation



P-value is the shaped area under the under the density curve

Another Example of General Linear Test: Full Model

```
> full <- lm(Species ~ Area + Elevation + Nearest + Scruz + Adjacent,
  data = gala)
> anova(full)
Analysis of Variance Table

Response: Species
          Df Sum Sq Mean Sq F value    Pr(>F)
Area       1 145470  145470 39.1262 1.826e-06 ***
Elevation  1  65664   65664 17.6613 0.0003155 ***
Nearest    1     29      29  0.0079 0.9300674
Scruz      1  14280   14280  3.8408 0.0617324 .
Adjacent   1  66406   66406 17.8609 0.0002971 ***
Residuals 24  89231    3718
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Another Example of General Linear Test: Reduced Model

```
> reduced <- lm(Species ~ Elevation + Adjacent)
> anova(reduced)
Analysis of Variance Table

Response: Species
          Df Sum Sq Mean Sq F value    Pr(>F)
Elevation  1 207828  207828  56.112 4.662e-08 ***
Adjacent   1  73251   73251  19.777 0.0001344 ***
Residuals 27 100003    3704
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Perform a General Linear Test

- $H_0 : \beta_{\texttt{Area}} = \beta_{\texttt{Nearest}} = \beta_{\texttt{Scruz}}$ vs.
  $H_a :$ at least one of the three coefficients $\neq 0$
- $F^* = \frac{(100003-89231)/(5-2)}{89231/(30-5-1)} = 0.9657$
- P-value: $\mathrm{P}[F > 0.9657] = 0.425$, where $F \sim$ F$(3, 24)$

```
> anova(reduced, full)
Analysis of Variance Table

Model 1: Species ~ Elevation + Adjacent
Model 2: Species ~ Area + Elevation + Nearest + Scruz + Adjacent
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     27 100003
2     24  89231  3     10772 0.9657  0.425
```

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

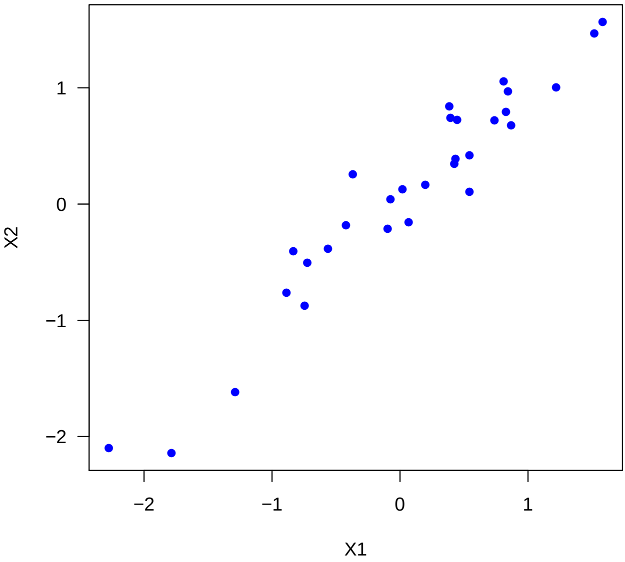
Suppose the true relationship between response Y and predictors (X_1, X_2) is

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

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
```

Variable Selection

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

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

- Akaike's information criterion (AIC)

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

- Bayesian information criterion (BIC)

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

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