Lecture 8
Multiple Linear Regression II
Reading: Chapter 12

STAT 8020 Statistical Methods II
September 6, 2019

Agenda

1. Coefficient of Determination
2. General Linear Test
3. Multicollinearity

Coefficient of Determination

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

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

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

```r
> summary(fit1)
Call: lm(formula = y ~ x1)
Residuals:
     Min      1Q  Median      3Q     Max
-1.6015 -0.5056 -0.2152  0.6932  2.0118
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.1728    0.1534   33.71  < 2e-16 ***
x1          1.8668    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: 2.47e-12
```

Model 2 Fit

```r
> summary(fit2)
Call: lm(formula = y ~ x1 + x2)
Residuals:
     Min      1Q  Median      3Q     Max
-1.3906 -0.5775 -0.1383  0.5229  1.8385
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.1792    0.1518   34.189  < 2e-16 ***
x1          1.8994    0.1593   11.923  2.88e-12 ***
x2          -0.2189    0.1797   -1.217    0.233
---
Signif. codes:  
  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.8381 on 27 degrees of freedom
Multiple R-squared: 0.8448,  Adjusted R-squared: 0.8291
F-statistic: 71.32 on 2 and 27 DF,  p-value: 1.677e-11
Multiple Linear Regression II

Coefficient of Determination
General Linear Test
Multicollinearity

$R^2$: Model 1 vs. Model 2

![Graph showing $R^2$ values for Model 1 vs. Model 2]

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

![Graph showing $R^2_{adj}$ values for 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 ($l < k$)
- Test statistic: $F^* = \frac{\text{SSE}(R) - \text{SSE}(F)/(k-l)}{\text{SSE}(F)/(n-k-1)} = \frac{\text{SSE}(R) - \text{SSE}(F)/(k-l)}{\frac{\text{SSE}(F)}{n-k-1}}$  

Testing $H_0$ that the regression coefficients for the extra variables are all zero

Notes
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  -14.600  4.634 259.180

Coefficients:    Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.33511   19.28029    0.58      0.56
Elevation   0.08079    0.03465     2.34    0.0291 ***
---
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 27 DF, p-value: 3.17e-06

Species Diversity on the Galapagos Islands Revisited: Full Model

> summary(gala_fit2)

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

Residuals:     Min      1Q  Median      3Q     Max
-392.629  -33.534  -19.199    7.541 261.514

Coefficients:    Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.10519  28.94211    0.61      0.54
Elevation    0.17174   0.05317     3.23    0.00325 **
Area         0.01880   0.02594     0.72    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: 0.594,    Adjusted R-squared: 0.521
F-statistic: 16.77 on 2 and 27 DF,  p-value: 1.845e-05

Perform a General Linear Test

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

> 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

Notes
### P-value Calculation

![F test statistic and density curve]

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

### Multicollinearity

**Another 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 $\varepsilon \sim N(0,1)$ and $X_1$ and $X_2$ are positively correlated with $\rho = 0.9$. Let’s fit the following model:

$$Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \varepsilon$$

Call:

\texttt{lm(formula = Y ~ X1 + X2)}

Residuals:

<table>
<thead>
<tr>
<th>Min</th>
<th>3Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.63956</td>
<td>-0.59972</td>
<td>0.06897</td>
<td>0.39601</td>
<td>1.74513</td>
</tr>
</tbody>
</table>

Coefficients:

\begin{tabular}{cccc}
Estimate & Std. Error & t value & Pr(>|t|) \\
(Intercept) & 4.0154 & 0.1646 & 24.298 < 2e-16 *** \\
X1 & -0.8052 & 0.3450 & -2.350 & 0.021 & 0.760 \\
X2 & 1.7471 & 0.3504 & 4.871 & 5.46e-05 *** \\
\end{tabular}

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8658 on 27 degrees of freedom
Multiple R-squared: 0.8366, Adjusted R-squared: 0.825
F-statistic: 58.12 on 2 and 27 DF, p-value: 1.186e-10

### Multicollinearity cont’d

- **Numerical issue** ⇒ the matrix $X^TX$ is nearly singular
- **Statistical issue**
  - $\beta$’s are not well estimated
  - $\beta$’s may be meaningless
  - $R^2$ and predicted values are usually OK