

Lecture 9

Multiple Linear Regression III

Reading: Chapter 12

STAT 8020 Statistical Methods II
September 9, 2019

Whitney Huang
Clemson University

Review: General
Linear Test

Review:
Multicollinearity

1 Review: General Linear Test

2 Review: Multicollinearity

- 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: $F^* = \frac{SSE(R) - SSE(F)/(k - \ell)}{SSE(F)/(n - k - 1)} \Rightarrow$ Testing H_0 that the regression coefficients for the extra variables are all zero
 - Example 1: X_1, X_2, \dots, 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

```
> full <- lm(Species ~ Area + Elevation + Nearest + Scrutz + 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
Scrutz	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

Species Diversity on the Galapagos Islands Revisited: 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

- $H_0 : \beta_{\text{Area}} = \beta_{\text{Nearest}} = \beta_{\text{Scruz}} \text{ vs.}$
 $H_a : \text{at least one of the three coefficients} \neq 0$
- $F^* = \frac{(100003 - 89231)/(5-2)}{89231/(30-5-1)} = 0.9657$
- P-value: $P[F > 0.9657] = 0.425$, 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 is a phenomenon of high inter-correlations or among the predictor variables

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

- 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 $Y \sim X_1$

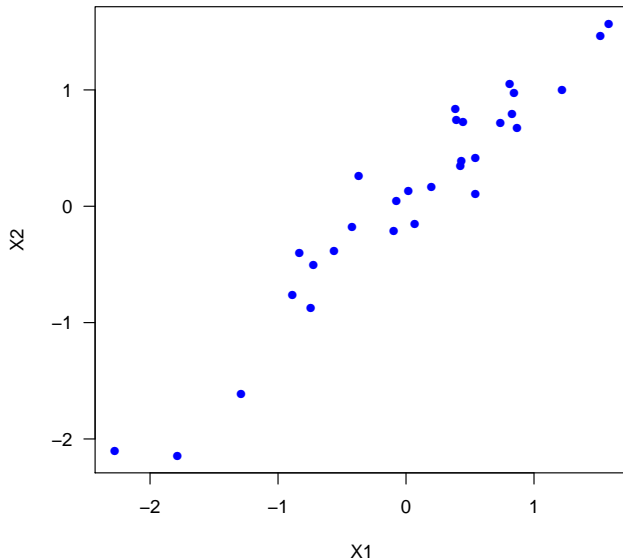
Suppose the true relationship between response Y and predictors (X_1, X_2) is

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

where $\varepsilon \sim N(0, 1)$ and X_1 and X_2 are positively correlated with $\rho = 0.9$. Let's fit the following models:

- Model 1: $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \varepsilon$
- Model 2: $Y = \beta_0 + \beta_1X_1 + \varepsilon^1$
- Model 3: $Y = \beta_0 + \beta_2X_2 + \varepsilon^2$

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

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

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