Lecture 2

Multiple Linear Regression: Estimation and Inference

Reading: Faraway 2014 Chapters 2.1 - 2.6, 3.1 - 3.2; 3.5; ISLR 2021 Chapter 3.2

DSA 8020 Statistical Methods II

Whitney Huang
Clemson University

Agenda

1. **Multiple Linear Regression**
2. **Estimation & Inference**
3. **Assessing Model Fit**

Multiple Linear Regression (MLR)

Goal: To model the relationship between two or more predictors (x's) and a response (y) by fitting a **linear equation** to observed data $\{y_i, x_{1,i}, x_{2,i}, \cdots, x_{p-1,i}\}_{i=1}^n$:

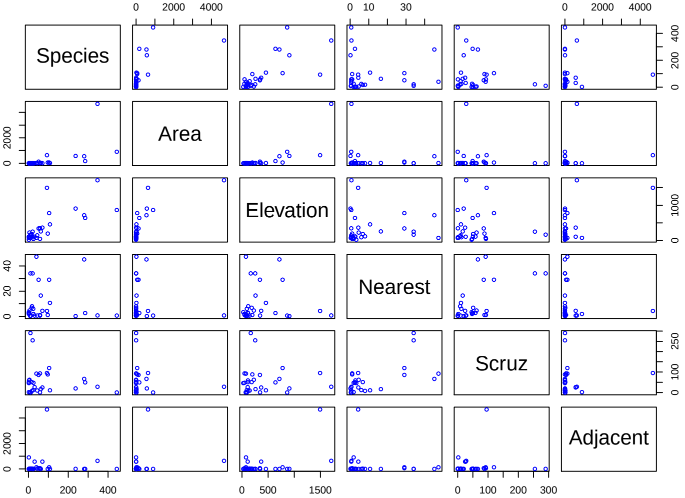
$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_{p-1} x_{p-1,i} + \varepsilon_i, \quad \varepsilon_i \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma^2)$$

Example: Species diversity on the Galapagos Islands. We are interested in studying the relationship between the number of plant species (`Species`) and the following geographic variables: `Area`, `Elevation`, `Nearest`, `Scruz`, `Adjacent`.

Data: Species Diversity on the Galapagos Islands

	Species	Endemics	Area	Elevation	Nearest	Scruz	Adjacent
Baltra	58	23	25.09	346	0.6	0.6	1.84
Bartolome	31	21	1.24	109	0.6	26.3	572.33
Caldwell	3	3	0.21	114	2.8	58.7	0.78
Champion	25	9	0.10	46	1.9	47.4	0.18
Coamano	2	1	0.05	77	1.9	1.9	903.82
Daphne.Major	18	11	0.34	119	8.0	8.0	1.84
Daphne.Minor	24	0	0.08	93	6.0	12.0	0.34
Darwin	10	7	2.33	168	34.1	290.2	2.85
Eden	8	4	0.03	71	0.4	0.4	17.95
Enderby	2	2	0.18	112	2.6	50.2	0.10
Espanola	97	26	58.27	198	1.1	88.3	0.57
Fernandina	93	35	634.49	1494	4.3	95.3	4669.32
Gardner1	58	17	0.57	49	1.1	93.1	58.27
Gardner2	5	4	0.78	227	4.6	62.2	0.21
Genovesa	40	19	17.35	76	47.4	92.2	129.49
Isabela	347	89	4669.32	1707	0.7	28.1	634.49
Marchena	51	23	129.49	343	29.1	85.9	59.56
Onslow	2	2	0.01	25	3.3	45.9	0.10
Pinta	104	37	59.56	777	29.1	119.6	129.49
Pinzon	108	33	17.95	458	10.7	10.7	0.03
Las.Plazas	12	9	0.23	94	0.5	0.6	25.09
Rabida	70	30	4.89	367	4.4	24.4	572.33
SanCristobal	280	65	551.62	716	45.2	66.6	0.57
SanSalvador	237	81	572.33	906	0.2	19.8	4.89
SantaCruz	444	95	903.82	864	0.6	0.0	0.52
SantaFe	62	28	24.08	259	16.5	16.5	0.52
SantaMaria	285	73	170.92	640	2.6	49.2	0.10
Seymour	44	16	1.84	147	0.6	9.6	25.09
Tortuga	16	8	1.24	186	6.8	50.9	17.95
Wolf	21	12	2.85	253	34.1	254.7	2.33

How Do Geographic Variables Affect Species Diversity?

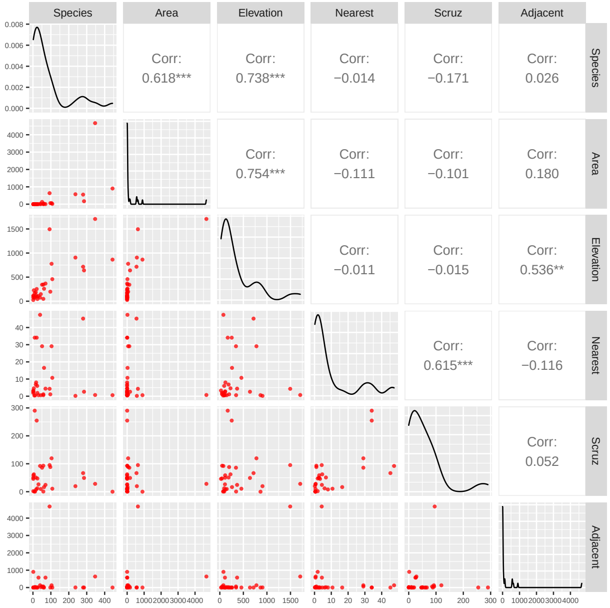


Let’s Take a Look at the Correlation Matrix

Here we compute the correlation coefficients between the response (`Species`) and predictors (all the geographic variables)

```
> round(cor(gala[, -2]), 3)
          Species   Area Elevation Nearest  Scruz Adjacent
Species     1.000  0.618     0.738  -0.014 -0.171    0.026
Area        0.618  1.000     0.754  -0.111 -0.101    0.180
Elevation   0.738  0.754     1.000  -0.011 -0.015    0.536
Nearest    -0.014 -0.111    -0.011   1.000  0.615   -0.116
Scruz      -0.171 -0.101    -0.015   0.615  1.000    0.052
Adjacent    0.026  0.180     0.536  -0.116  0.052    1.000
```

Combining Two Pieces of Information in One Plot

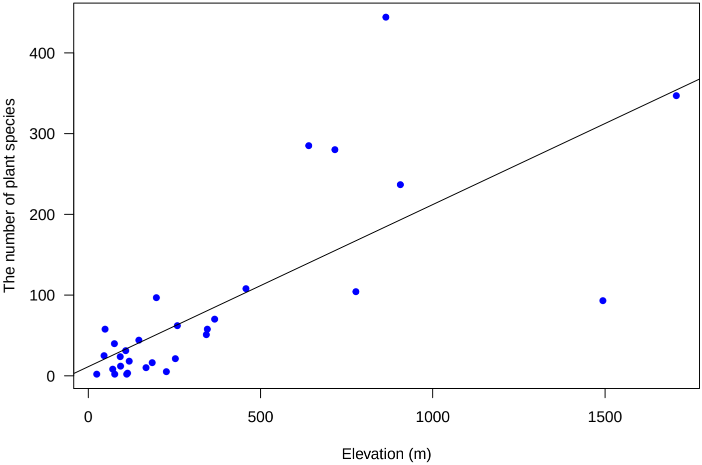


Model 1: `Species ~ Elevation`

```
Call:
lm(formula = Species ~ Elevation, data = gala)

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
```

Model 1 Fit

$$\hat{\text{Species}} = 11.33511 + 0.20079 \times \text{Elevation},$$
$$\hat{\sigma} = 78.66,\ \mathrm{R}^2 = 0.5454$$



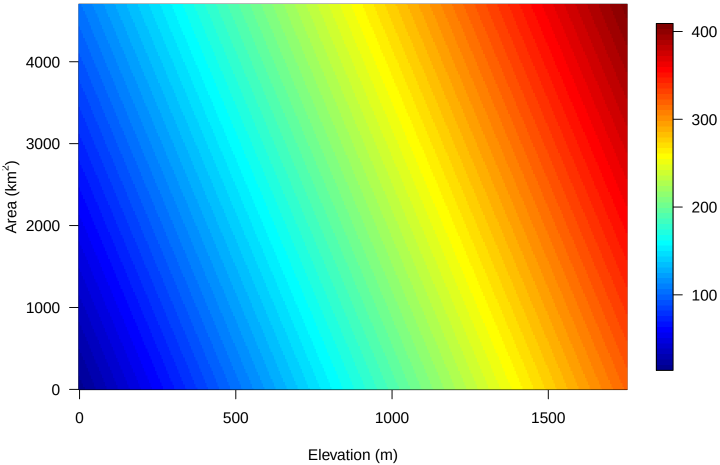
Model 2: Species ~ Elevation + Area

```
Call:
lm(formula = Species ~ Elevation + Area, data = gala)

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
```

Model 2 Fit

$$\widehat{\text{Species}} = 17.10519 + 0.17174 \times \text{Elevation} + 0.01880 \times \text{Area},$$

$$\hat{\sigma} = 79.34, \; \text{R}^2 = 0.554$$



Model 3: Species ~ Elevation + Area + Adjacent

```
Call:
lm(formula = Species ~ Elevation + Area + Adjacent, data = gala)

Residuals:
     Min       1Q   Median       3Q      Max
-124.064  -34.283   -8.733   27.972  195.973

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.71893   16.90706  -0.338  0.73789
Elevation    0.31498    0.05211   6.044  2.2e-06 ***
Area        -0.02031    0.02181  -0.931  0.36034
Adjacent    -0.07528    0.01698  -4.434  0.00015 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 61.01 on 26 degrees of freedom
Multiple R-squared:  0.746,     Adjusted R-squared:  0.7167
F-statistic: 25.46 on 3 and 26 DF,  p-value: 6.683e-08
```

“Full Model”

```
lm(formula = Species ~ Area + Elevation + Nearest + Scruz + Adjacent,
    data = gala)

Residuals:
     Min       1Q   Median       3Q      Max 
-111.679  -34.898   -7.862   33.460  182.584 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.068221  19.154198   0.369 0.715351
Area        -0.023938   0.022422  -1.068 0.296318
Elevation    0.319465   0.053663   5.953 3.82e-06
Nearest      0.009144   1.054136   0.009 0.993151
Scruz       -0.240524   0.215402  -1.117 0.275208
Adjacent    -0.074805   0.017700  -4.226 0.000297

(Intercept)
Area
Elevation   ***
Nearest
Scruz
Adjacent    ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 60.98 on 24 degrees of freedom
Multiple R-squared:  0.7658,    Adjusted R-squared:  0.7171
F-statistic:  15.7 on 5 and 24 DF,  p-value: 6.838e-07
```

MLR Topics

Similar to SLR, we will discuss

- Estimation
- Inference
- Diagnostics and Remedies

We will also discuss some new topics

- Model Selection
- Multicollinearity

Multiple Linear Regression in Matrix Notation

Given the actual data, we can write MLR model as:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & x_{2,1} & \cdots & x_{p-1,1} \\ 1 & x_{1,2} & x_{2,2} & \cdots & x_{p-1,2} \\ \vdots & \cdots & \ddots & \vdots & \vdots \\ 1 & x_{1,n} & x_{2,n} & \cdots & x_{p-1,n} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

It will be more convenient to put this in a matrix representation as:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Error Sum of Squares (SSE) = $\sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \sum_{j=1}^{p-1} \beta_j x_{j,i}\right)\right)^2$ can be expressed as:

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

Next, we are going to find $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_{p-1})$ to minimize SSE as our estimate for $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_{p-1})$

Estimating Regression Coefficients

We apply method of least squares to minimize $(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$ to obtain $\hat{\boldsymbol{\beta}}$

- The resulting least squares estimate is

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

  (see `LS_MLR.pdf` for the derivation)

- Fitted values:

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}$$

- Residuals:

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$$

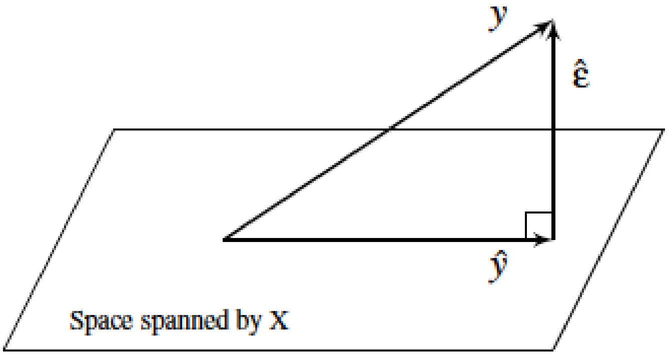
Estimation of σ^2

- Similar as we did in SLR

$$\begin{aligned}\hat{\sigma}^2 &= \frac{\boldsymbol{e}^T\boldsymbol{e}}{n-p} \\ &= \frac{(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})}{n-p} \\ &= \frac{\text{SSE}}{n-p} \\ &= \text{MSE}\end{aligned}$$

Geometrical Representation of the Estimation β

Projecting the observed response $\boldsymbol{y}$ into a space spanned by $\boldsymbol{X}$



Source: Linear Model with R 2nd Ed, Faraway, p. 15

Analysis of Variance (ANOVA) Approach to Regression

Partitioning Sums of Squares

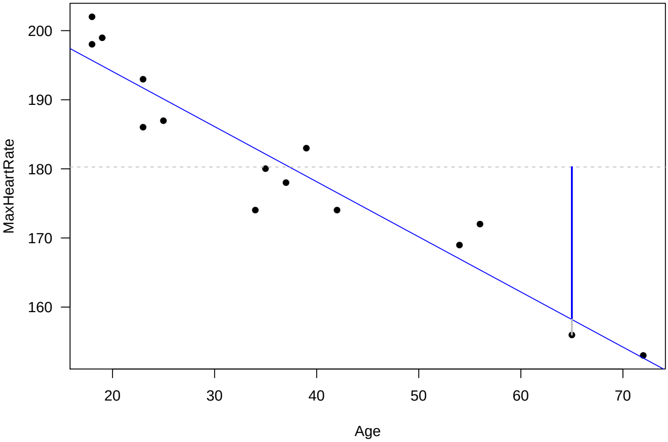
- Total sums of squares in response

$$\text{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

- We can rewrite SST as

$$\begin{aligned}\sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\ &= \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\text{"Error": SSE}} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{\text{Model: SSR}}\end{aligned}$$

Partitioning Total Sums of Squares: A Graphical Illustration



ANOVA Table & F-Test

To answer the question: **Is at least** one of the predictors $x_1, \cdots, x_{p-1}$ useful in predicting the response y?

Source	df	SS	MS	F Value
Model	$p-1$	SSR	MSR = SSR/$(p-1)$	MSR/MSE
Error	$n-p$	SSE	MSE = SSE/$(n-p)$	
Total	$n-1$	SST		

- F-Test: Tests if the predictors $\{x_1, \cdots, x_{p-1}\}$ collectively help explain the variation in y
  - $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$
  - $H_a :$ at least one $\beta_k \neq 0, \quad 1 \leq k \leq p-1$
  - $F^* = \frac{\text{MSR}}{\text{MSE}} = \frac{\text{SSR}/(p-1)}{\text{SSE}/(n-p)} \overset{H_0}{\sim} F_{p-1,n-p}$
  - Reject H_0 if $F^* > F_{1-\alpha,p-1,n-p}$

Testing Individual Predictor

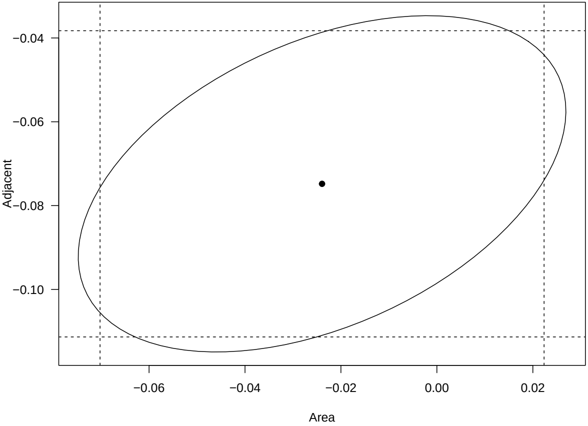
- We can show that $\hat{\boldsymbol{\beta}} \sim \mathrm{N}_p\left(\boldsymbol{\beta}, \sigma^2\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\right) \Rightarrow$ $\hat{\beta}_k \sim \mathrm{N}(\beta_k, \sigma^2_{\hat{\beta}_k})$
- Perform t-Test:
  - $H_0 : \beta_k = 0$ vs. $H_a : \beta_k \neq 0$
  - $\frac{\hat{\beta}_k - \beta_k}{\hat{SE}(\hat{\beta}_k)} \sim t_{n-p} \Rightarrow t^* = \frac{\hat{\beta}_k}{\hat{SE}(\hat{\beta}_k)} \overset{H_0}{\sim} t_{n-p}$
  - Reject H_0 if $|t^*| > t_{1-\alpha/2,n-p}$
- Confidence interval for β_k:

$$\hat{\beta}_k \pm t_{1-\alpha/2,n-p}\hat{SE}(\hat{\beta}_k)$$

Confidence Intervals and Confidence Ellipsoids

Comparing with individual confidence interval, confidence ellipsoids can provide additional information when inference with multiple parameters is of interest. A $100(1-\alpha)\%$ confidence ellipsoid for $\boldsymbol{\beta}$ can be constructed using:

$$(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T \boldsymbol{X}^T \boldsymbol{X} (\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \le p\hat{\sigma}^2 F^{\alpha}_{p,n-p}.$$



Quantifying Model Fit using Coefficient of Determination R^2

- 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 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 $\varepsilon \sim \mathrm{N}(0,1)$ 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?

Let's conduct a Monte Carlo simulation to study this

Outline of Monte Carlo Simulation

1. Generating a large number (e.g., $M = 500$) of "data sets", where each has exactly the same $\{x_{1,i}, x_{2,i}\}_{i=1}^n$ but different values of response $\{y_i = 5 + 2x_{1,i} + \varepsilon_i\}_{i=1}^n$
2. Fitting model 1: $y = \beta_0 + \beta_1 x_1 + \varepsilon^1$ (true model) and model 2: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon^2$, respectively for each simulating data set and calculating their R^2 and R^2_{adj}
3. Summarizing $\{R_j^2\}_{j=1}^M$ and $\{R^2_{adj,j}\}_{j=1}^M$ for model 1 and model 2

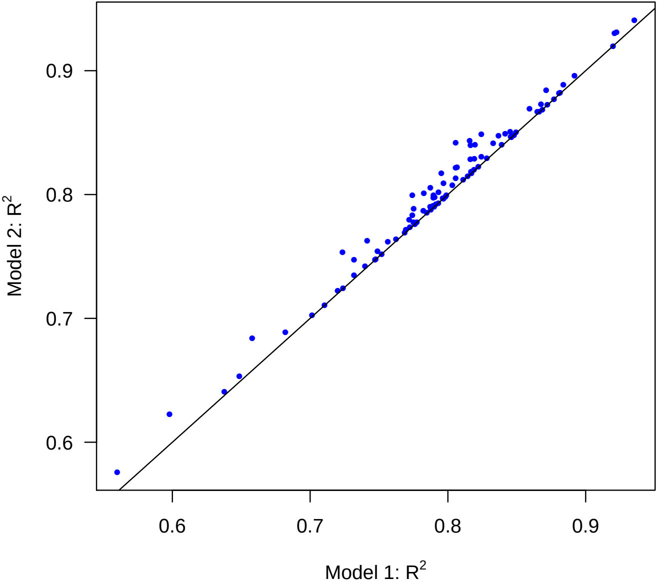
An Example of Model 1 Fit

```
> 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
```

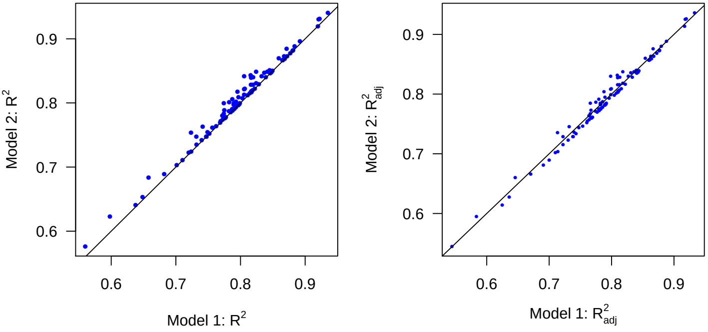
An Example of Model 2 Fit

```
> 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



Takeaways:

- R^2 always pick the more "complex" model (i.e., with more predictors), even the simpler model is the true model
- R^2_{adj} has a better chance to pick the "right" model

Summary

These slides cover:

- Parameter Estimation of MLR
- Inference: F-test and t-test; Confidence intervals/ellipsoids
- Assessing Model Fit: R^2 and R^2_{adj}
- Monte Carlo Simulation

`R` functions to know:

- `image.plot` in the `fields` library and `scatter3D` in the `plot3D` library for visualization
- `anova` for computing the ANOVA table