# Lecture 3

## Multiple Linear Regression: Inference and Prediction

Reading: Faraway, 2014 Chapters 3, 4

*DSA 8020 Statistical Methods II*

January 24-28, 2021

Whitney Huang

Clemson University

# Agenda

1. **General Linear F-Test**
2. **Prediction**
3. **Multicollinearity**

# Review: T-Test and F-Test in Linear Regression

- T-Test: Testing one predictor
  1. **Null/Alternative Hypotheses**: $H_0 : \beta_j = 0$ vs. $H_a : \beta_j \neq 0$
  2. **Test Statistic**: $t^* = \frac{\hat{\beta}_j - 0}{SE(\hat{\beta}_j)}$
  3. Reject $H_0$ if $|t^*| > t_{1-\alpha/2, n-p}$

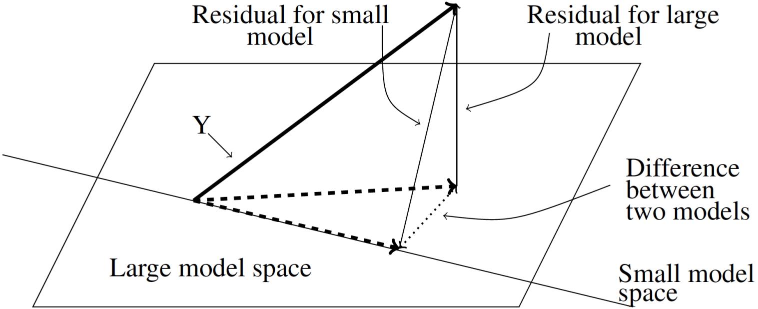
- Overall F-Test: Test of all the predictors
  1. $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$
  2. $H_a :$ at least one $\beta_j \neq 0, 1 \leq j \leq p-1$
  3. **Test Statistic**: $F^* = \frac{\text{MSR}}{\text{MSE}}$
  4. Reject $H_0$ if $F^* > F_{1-\alpha, p-1, n-p}$

Both tests are special cases of General Linear F-Test

## General Linear F-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 $\ell$ 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 $\beta_j$
  - Example 3: $x_1, x_2, x_3, x_4$ vs. $x_1, x_3 \Rightarrow H_0 : \beta_2 = \beta_4 = 0$

# Geometric Illustration of General Linear F-Test




**Source**: Faraway, *Linear Models with* R, 2014, p.34

# Species Diversity on the Galapagos Islands: 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: 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
```

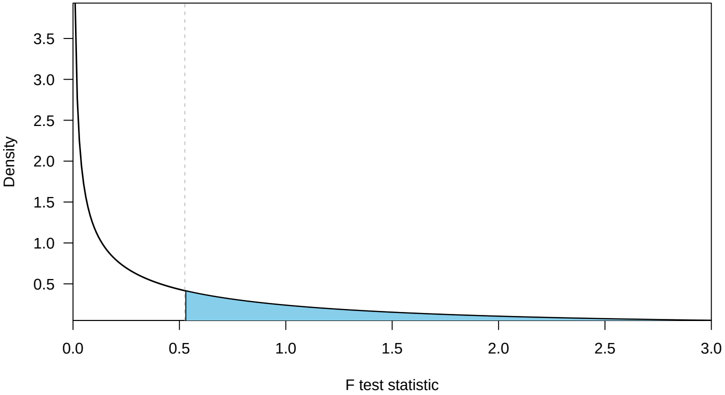
# Performing a General Linear Test

- $H_0 : \beta_{\texttt{Area}} = 0$ 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 $F \sim \mathrm{F}_{\underbrace{1}_{k-\ell}, \underbrace{27}_{n-k-1}}$

```
> 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
```

# Visualizing P-value




P-value is the shaped area under the density curve of the null distribution

# 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
```

# Performing 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 $F \sim \mathrm{F}_{3,24}$

```
> 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
```

# Multiple Linear Regression Prediction



Given a new set of predictors, $\boldsymbol{x}_0 = (1, x_{0,1}, x_{0,2}, \cdots, x_{0,p-1})^{\mathrm{T}}$, the predicted response is

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{0,1} + \cdots + \hat{\beta}_{p-1} x_{0,p-1}.$$

Again, we can use matrix representation to simplify the notation

$$\hat{y}_0 = \boldsymbol{x}_0^{\mathrm{T}} \hat{\boldsymbol{\beta}}.$$

We will use this to carry out two different kinds of predictions

# Two Kinds of Predictions

There are two kinds of predictions can be made for a given $\boldsymbol{x}_0$:

- **Predicting a future response:**

  Based on MLR, we have $y_0 = \boldsymbol{x}_0^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon$. Since $E(\varepsilon) = 0$, therefore the predicted value is

  $$\hat{y}_0 = \boldsymbol{x}_0^{\mathrm{T}}\hat{\boldsymbol{\beta}}$$

- **Predicting the mean response:**

  Since $E(y_0) = \boldsymbol{x}_0^{\mathrm{T}}\boldsymbol{\beta}$, there we have the predicted mean response

  $$\widehat{E(y_0)} = \boldsymbol{x}_0^{\mathrm{T}}\hat{\boldsymbol{\beta}},$$

  the same predicted value as predicting a future response

Next, we need to assess the prediction uncertainty

# Prediction Uncertainty

From slides 2 page 21, we have $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2 \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}$. Therefore we have

$$\mathrm{Var}(\hat{y}_0) = \mathrm{Var}(\boldsymbol{x}_0^{\mathrm{T}}\hat{\boldsymbol{\beta}}) = \sigma^2 \boldsymbol{x}_0^{\mathrm{T}} \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1} \boldsymbol{x}_0$$

We can now construct $100(1-\alpha)\%$ CI for the two kinds of predictions:

- **Predicting a future response:**

$$\boldsymbol{x}_0^{\mathrm{T}}\hat{\boldsymbol{\beta}} \pm t_{n-p,\alpha/2} \times \hat{\sigma}\sqrt{\underbrace{1}_{\text{accounting for } \varepsilon} + \boldsymbol{x}_0^{\mathrm{T}} \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1} \boldsymbol{x}_0}$$

- **Predicting the mean response:**

$$\boldsymbol{x}_0^{\mathrm{T}}\hat{\boldsymbol{\beta}} \pm t_{n-p,\alpha/2} \times \hat{\sigma}\sqrt{\boldsymbol{x}_0^{\mathrm{T}} \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1} \boldsymbol{x}_0}$$

# Example: Predicting Body Fat (Taken from Faraway 2014, Linear Models with R)

```
lm(formula = brozek ~ age + weight + height + neck + chest +
    abdom + hip + thigh + knee + ankle + biceps + forearm + wrist,
    data = fat)

Residuals:
    Min      1Q  Median      3Q     Max
-10.264  -2.572  -0.097   2.898   9.327

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -15.29255   16.06992  -0.952  0.34225
age           0.05679    0.02996   1.895  0.05929 .
weight       -0.08031    0.04958  -1.620  0.10660
height       -0.06460    0.08893  -0.726  0.46830
neck         -0.43754    0.21533  -2.032  0.04327 *
chest        -0.02360    0.09184  -0.257  0.79740
abdom         0.88543    0.08008  11.057  < 2e-16 ***
hip          -0.19842    0.13516  -1.468  0.14341
thigh         0.23190    0.13372   1.734  0.08418 .
knee         -0.01168    0.22414  -0.052  0.95850
ankle         0.16354    0.20514   0.797  0.42614
biceps        0.15280    0.15851   0.964  0.33605
forearm       0.43049    0.18445   2.334  0.02044 *
wrist        -1.47654    0.49552  -2.980  0.00318 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.988 on 238 degrees of freedom
Multiple R-squared:  0.749,     Adjusted R-squared:  0.7353
F-statistic: 54.63 on 13 and 238 DF,  p-value: < 2.2e-16
```

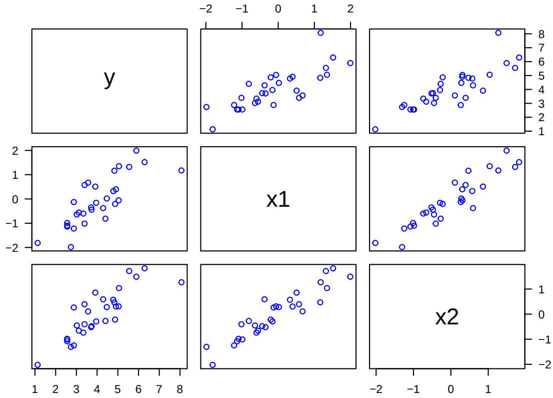
What is our prediction for the future response of a "typical" (e.g., each predictor take its median value) man?

# Example: Predicting Body Fat Cont'd

1. Calculate the median for each predictor to get $\boldsymbol{x}_0$
2. Compute the predicted value $\hat{y}_0 = \boldsymbol{x}_0^{\mathrm{T}} \hat{\boldsymbol{\beta}}$
3. Quantify the prediction uncertainty

```
> X <- model.matrix(lmod)
> (x0 <- apply(x, 2, median))
(Intercept)         age      weight      height        neck       chest       abdom
       1.00       43.00      176.50       70.00       38.00       99.65       90.95
        hip       thigh        knee       ankle      biceps     forearm       wrist
      99.30       59.00       38.50       22.80       32.05       28.70       18.30
> (y0 <- sum(x0 * coef(lmod)))
[1] 17.49322
> predict(lmod, new = data.frame(t(x0)))
       1
17.49322
> predict(lmod, new = data.frame(t(x0)), interval = "prediction")
       fit     lwr      upr
1 17.49322 9.61783 25.36861
> predict(lmod, new = data.frame(t(x0)), interval = "confidence")
       fit      lwr      upr
1 17.49322 16.94426 18.04219
```

# Multicollinearity




```
> cor(sim1)
           y        x1        x2
y  1.0000000 0.7987777 0.8481084
x1 0.7987777 1.0000000 0.9281514
x2 0.8481084 0.9281514 1.0000000
```

# Multicollinearity Cont'd





**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
  - $\beta$'s are not well estimated
  - Spurious regression coefficient estimates
  - $R^2$ and predicted values are usually OK

## An Simulated Example



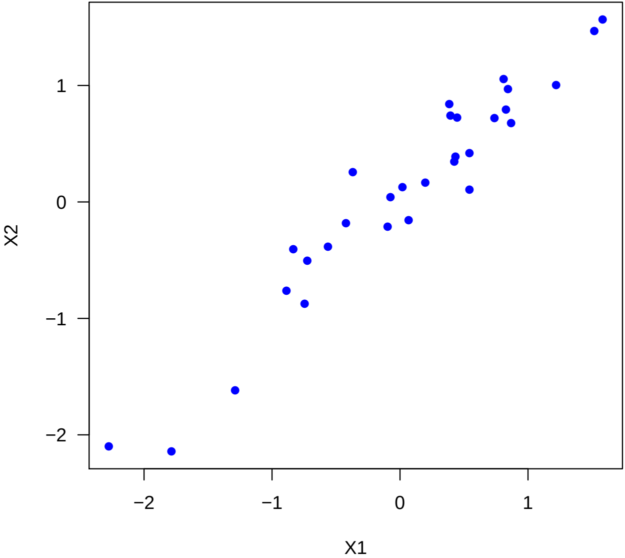
Suppose the true relationship between response $Y$ and predictors $(x_1, x_2)$ is

$$Y = 4 + 0.8x_1 + 0.6x_2 + \varepsilon,$$

where $\varepsilon \sim \mathrm{N}(0, 1)$ and $x_1$ and $x_2$ are positively correlated with $\rho = 0.9$. Let's fit the following models:

- Model 1: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_1$
- Model 2: $Y = \beta_0 + \beta_1 x_1 + \varepsilon_2$

# 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
```

# Variance Inflation Factor (VIF)

We can use the variance inflation factor (VIF)

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

to quantifies the severity of multicollinearity in MLR, where $\text{R}_i^2$ is the **coefficient of determination** when $X_i$ is regressed on the remaining predictors

**R example code**

```
> library(faraway)
> vif(sim1[, 2:3])
      x1       x2
7.218394 7.218394
```

$\sqrt{\text{VIF}}$ indicates how much larger the standard error increases compared to if that variable had 0 correlation to other predictor variables in the model.