STAT021 Statistical Methods II

Lecture 15 MLR Model Assessment

Lu Chen
Swarthmore College
11/1/2018

Outline

Multiple linear regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + \epsilon, \text{ where } \epsilon \overset{iid}{\sim} N(0, \sigma)$$

- MLR analysis of variance (ANOVA)
  - F test and R^2
- Three tests
  - t test for the slopes
  - F test for the MLR model
  - t test for the correlations
- Nested F test for a subset of predictors
- Adjusted R^2 for model comparison

Analysis of variance (ANOVA)

	Data		Model		Error
	Data	=	Model	+	Error
Population:	Y	=	$\beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K$	+	ϵ
Sample:	y	=	$b_0 + b_1 x_1 + \cdots + b_K x_K$	+	e
	y	=	$\hat{y}$	+	$y - \hat{y}$
	$y - \bar{y}$	=	$\hat{y} - \bar{y}$	+	$y - \hat{y}$
SS:	$SSTotal$	=	$SSModel$	+	SSE
	$\sum(y - \bar{y})^2$	=	$\sum(\hat{y} - \bar{y})^2$	+	$\sum(y - \hat{y})^2$
	Total variability in response Y	=	Variability explained by the MLR model	+	Variability in residuals
DF:	df_{Total}	=	df_{Model}	+	df_{Error}
	$n - 1$	=	K	+	$n - K - 1$

Analysis of variance (ANOVA)

$$MSModel = \frac{SSModel}{df_{Model}} = \frac{\sum(\hat{y}-\bar{y})^2}{K}, \quad MSE = \frac{SSE}{df_{Error}} = \frac{\sum(y-\hat{y})^2}{n-K-1}$$

$$F = \frac{MSModel}{MSE} = \frac{\frac{\sum(\hat{y}-\bar{y})^2}{K}}{\frac{\sum(y-\hat{y})^2}{n-K-1}} \sim F(K, n-K-1)$$

$$R^2 = \frac{SSModel}{SSTotal} = \frac{\sum(\hat{y}-\bar{y})^2}{\sum(y-\hat{y})^2}$$

- MSE is the estimate to the variance of error. The residual standard error is

$$\hat{\sigma} = \sqrt{MSE} = \sqrt{\frac{\sum(y-\hat{y})^2}{n-K-1}}$$

- F test indicates the significance of the MLR model. R^2 measures the strength (the fraction of variability explained by) of the MLR model.

ANOVA Table

To test the effectiveness of the multiple linear model, the hypotheses are

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_K = 0$ and H_a: at least one $\beta_k \neq 0$.

The **ANOVA table** is

	Degree of Freedom	Sum of Squares	Mean Square	F statistic	P-value
Model	K	$SSModel$	$MSModel$	$F = \frac{MSModel}{MSE}$	$P(F_{K,n-K-1} > F)$
Error	$n - K - 1$	SSE	MSE		
Total	$n - 1$	SST			

If the conditions for the multiple linear regression model hold, the P-value is obtained from the upper tail of an F-distribution with K and $n - K - 1$ degrees of freedom.

ANOVA F test and R squared

Understand MLR ANOVA F test

- $H_0 : \beta_1 = \beta_2 = \cdots = \beta_K = 0$
  - None of the predictors has any linear relationship with the response variable - ineffective model.
- H_a: at least one $\beta_k \neq 0$
  - At least one of the predictors is effective in the model - effective model.
  - But the ANOVA F test does not identify which predictors are significant. This is the role for the individual t tests.

Understand *coefficient of multiple determination* R^2

- It measures the strength of the model as a whole.
- It is interpreted as the fraction of variability explained by the MLR model that includes the multiple predictors.

ANOVA *F* test and *R* squared

```r
summary(m3 <- lm(Happiness ~ log(GDPpc) + LifeExp, data=HappyPlanet))
```

```
## Multiple R-squared:  0.4355, Adjusted R-squared:  0.4262
## F-statistic: 46.68 on 2 and 121 DF,  p-value: 9.436e-16
```

- $F = 46.68$ on
- $K = 2$ and $n - K - 1 = 121$ degrees of freedom
- $P = 9.4 \times 10^{-16}$
  - At least one of *log(GDPpc)* and *LifeExp* is significant in explaining *Happiness*.
  - The model with both *log(GDPpc)* and *LifeExp* is significant in explaining *Happiness*.
- $R^2 = 0.4355$
  - 43.55% of the variability in *Happiness* is explained by the model including *log(GDPpc)* and *LifeExp*.

Three tests

1. t **tests for the slopes**
2. F **test for the model**

```r
summary(m3) # Happiness ~ log(GDPpc) + LifeExp
```

```
## Call: lm(formula = Happiness ~ log(GDPpc) + LifeExp, data = HappyPlanet)
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.64252    4.33521   5.915 3.16e-08 ***
## log(GDPpc)  -5.68509    0.76664  -7.416 1.82e-11 ***
## LifeExp      0.96482    0.09991   9.657  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.371 on 121 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.4355, Adjusted R-squared:  0.4262
## F-statistic: 46.68 on 2 and 121 DF,  p-value: 9.436e-16
```

Three tests

3. ***t* tests for the correlations** (equivalent to the *t* tests in SLR models)

```
cor.test(~ Happiness + log(GDPpc), data=HappyPlanet)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Happiness and log(GDPpc)
## t = 0.24027, df = 122, p-value = 0.8105
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1551637  0.1973079
## sample estimates:
##        cor
## 0.02174784
```

- Relationship between *Happiness* and *log(GDPpc)* without considering *LifeExp*.

Three tests

3. ***t* tests for the correlations** (equivalent to the *t* tests in SLR models)

```
cor.test(~ Happiness + LifeExp, data=HappyPlanet)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Happiness and LifeExp
## t = 5.1155, df = 126, p-value = 1.135e-06
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2598608 0.5487341
## sample estimates:
##       cor
## 0.4146913
```

- Relationship between *Happiness* and *LifeExp* without considering *log(GDPpc)*.

Three tests

	t test for slope	F test for model	t test for correlation
H_0	$\beta_k = 0$	$\beta_1 = \cdots = \beta_K = 0$	$\rho_k = 0$
Test statistic	$t = -7.42$ $t = 9.66$	$F = 46.68$	$t = 0.24$ $t = 5.12$
Distribution	$t(n-K-1) = t(121)$	$F(K, n-K-1) = F(2, 121)$	$t(n-2) = t(122)$
P-value	1.82×10^{-11} $< 2.2 \times 10^{-16}$	9.44×10^{-16}	0.811 1.14×10^{-6}

- **t test for the slope**: significance of a predictor given other predictors are held constant.
- **F test for the model**: significance of the whole model including all the predictors.
- **t test for the correlation**: significance of a predictor without considering other predictors.

Three tests

	t test for slope	F test for model	t test for correlation
H_0	$\beta_k = 0$	$\beta_1 = \cdots = \beta_K = 0$	$\rho_k = 0$
Test statistic	$t = -7.42$ $t = 9.66$	$F = 46.68$	$t = 0.24$ $t = 5.12$
Distribution	$t(n-K-1) = t(121)$	$F(K, n-K-1) = F(2, 121)$	$t(n-2) = t(122)$
P**-value**	1.82×10^{-11} $< 2.2 \times 10^{-16}$	9.44×10^{-16}	0.811 1.14×10^{-6}

Note:

- In MLR, F is NOT the square of any of the t statistics.
- P-value of the F test is NOT equal to any of the P-values of the t tests, either.

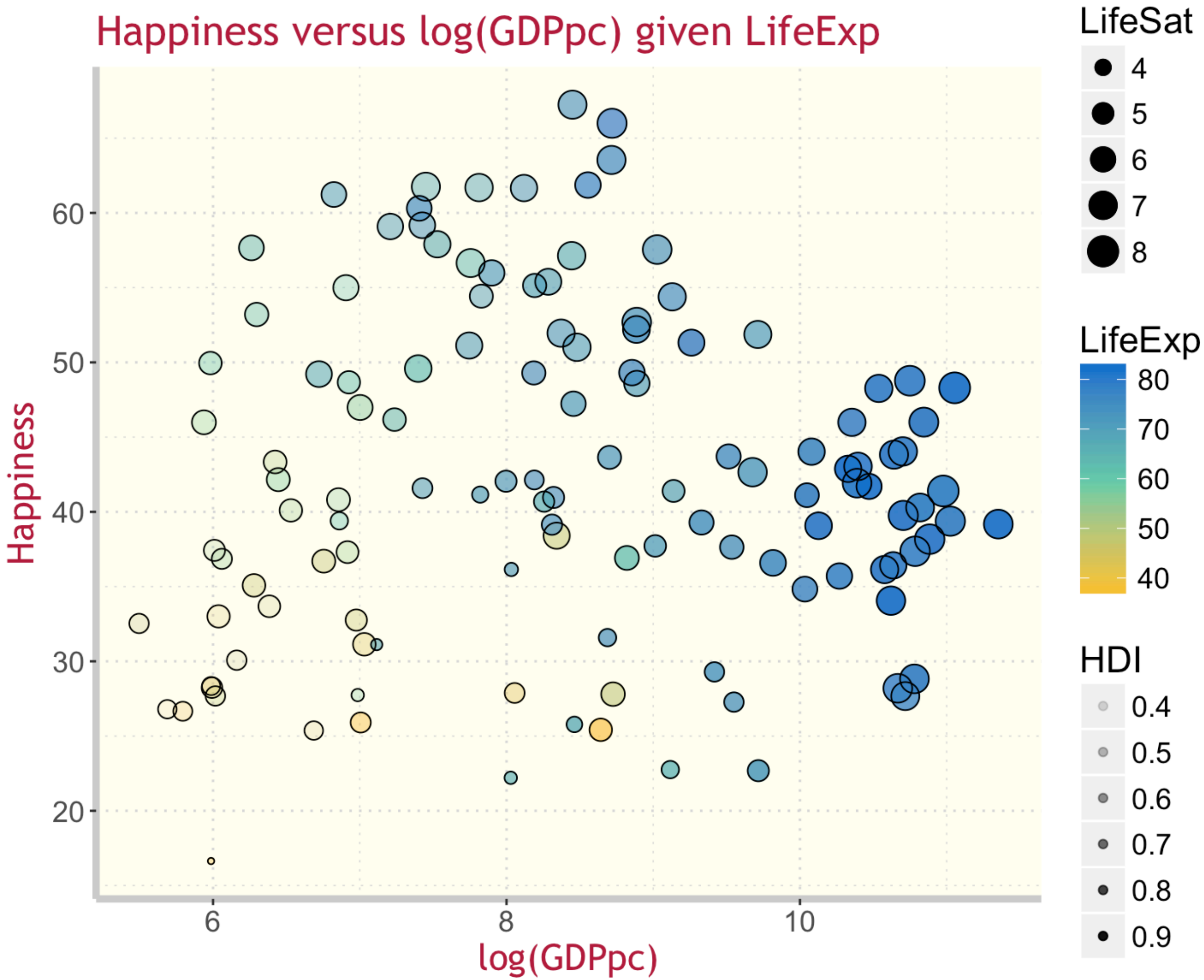
A new model

- Current predictors: *log(GDPpc)*, *LifeExp*
- Two new predictors:
  - *LifeSat*: Life satisfaction, level of experienced well-being.
  - *HDI*: Human development index, computed based on life expectancy, education index and gross national income of each country.
- Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon, \text{ where } \epsilon \overset{iid}{\sim} N(0, \sigma)$$

  - Y : *Happiness*; X_1 : *log(GDPpc)*; X_2 : *LifeExp*; X_3 : *LifeSat*; X_4 : *HDI*

A new model



- Although we can display multiple dimensions of data by color, size and transparency, it becomes difficult to tell the actual relationship between the multiple variables visually.
- Therefore, the three tests (t test for slope, F test for model and t tests for correlations) are of great help.

A new model

```r
summary(m5 <- lm(Happiness ~ log(GDPpc)+LifeExp+LifeSat+HDI, data=HappyPlanet))
```

```
## Call: lm(formula = Happiness ~ log(GDPpc) + LifeExp + LifeSat + HDI,
##     data = HappyPlanet)
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.67493    3.26129   3.886 0.000168 ***
## log(GDPpc)  -8.91513    0.94204  -9.464 3.49e-16 ***
## LifeExp      0.73834    0.08806   8.384 1.20e-13 ***
## LifeSat      7.55823    0.58596  12.899  < 2e-16 ***
## HDI         14.09056   11.77288   1.197 0.233738
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.412 on 119 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.7679, Adjusted R-squared:  0.7601
## F-statistic: 98.46 on 4 and 119 DF,  p-value: < 2.2e-16
```

$$\widehat{Happiness} = 12.7 - 8.9 \times log(GDPpc) + 0.7 \times LifeExp + 7.6 \times LifeSat + 14.1 \times HDI$$

A new model

```r
summary(m5)$coefficients
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.67493    3.26129   3.886 0.000168 ***
## log(GDPpc)  -8.91513    0.94204  -9.464 3.49e-16 ***
## LifeExp      0.73834    0.08806   8.384 1.20e-13 ***
## LifeSat      7.55823    0.58596  12.899  < 2e-16 ***
## HDI         14.09056   11.77288   1.197 0.233738
```

- *Happiness* and *log(GDPpc)* are significantly negatively associated given that *LifeExp*, *LifeSat* and *HDI* are held constant.
- Adjusted for *log(GDPpc)*, *LifeSat* and *HDI*, *Happiness* and *LifeExp* are significantly positively associated.
- *Happiness* and *LifeSat* are significantly positively associated after adjusting for *log(GDPpc)*, *LifeExp* and *HDI*.
- *Happiness* and *HDI* are NOT significantly associated after adjusting for *log(GDPpc)*, *LifeExp* and *LifeSat*.

A new model

```
summary(m5) # Happiness ~ log(GDPpc)+LifeExp+LifeSat+HDI
```

```
## Multiple R-squared:  0.7679, Adjusted R-squared:  0.7601
## F-statistic: 98.46 on 4 and 119 DF,  p-value: < 2.2e-16
```

- $F = 98.46$ on $K = 4$ and $n - K - 1 = 119$ degrees of freedom; $P < 2.2 \times 10^{-16}$.
  - The model with all the four predictors is highly significant in explaining *Happiness*.
- $R^2 = 0.7679$. 76.79% of variability is explained by this new model including four predictors.

How to tell whether this model with two new predictors *LifeSat* and *HDI* is better than the previous one ($F = 46.68$ with $P = 9.55 \times 10^{-16}$ and $R^2 = 0.4355$)?

- Nested F test and adjusted R^2.

Nested F test

To test a subset of predictors (J out of K predictors) in a MLR model,
$H_0 : \beta_j = 0$ for all J predictors in the subset versus
$H_a : \beta_j \neq 0$ for at least one predictor in the subset.

Let the **full model** denote one with all K predictors and the **reduced model** be the nested model with $K - J$ predictors obtained by dropping the J predictors that are being tested. The test statistic is

$$F = \frac{(SSModel_{full} - SSModel_{reduced})/J}{SSE_{full}/(n-K-1)} \sim F(J, n-K-1)$$

The P-value is computed from an F distribution with J and $n - K - 1$ degrees of freedom. Note that since $SST_{full} = SST_{reduced}$,

$$SSModel_{full} - SSModel_{reduced} = SSE_{reduced} - SSE_{full}$$

Nested F test

- Reduced model: *Happiness* ~ *log(GDPpc)* + *LifeExp*
- Full model: *Happiness* ~ *log(GDPpc)* + *LifeExp* + *LifeSat* + *HDI*
- Subset of predictors that are of interest: *LifeSat* and *HDI*

$$F = \frac{(SSModel_{full} - SSModel_{reduced})/J}{SSE_{full}/(n-K-1)} = \frac{(SSE_{reduced} - SSE_{full})/J}{SSE_{full}/(n-K-1)} \sim F(J, n-K-1)$$

- $n = 124, K = 4, J = 2$
- $n - K - 1 = 119$
- $SSE = MSE \times df_{Error} = \hat{\sigma}^2 \times df_{Error}$

Nested F test in R

```r
anova(m3, m5) # compare the reduced model m3 to the full model m5
```

```
## Analysis of Variance Table
##
## Model 1: Happiness ~ log(GDPpc) + LifeExp
## Model 2: Happiness ~ log(GDPpc) + LifeExp + LifeSat + HDI
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    121 8478.9
## 2    119 3485.6  2    4993.3 85.239 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- `Res.Df`: df_{Error}
- `RSS`: SSE, residual sum of squares.
- $F = \frac{(8478.9-3485.6)/2}{3485.6/119} = 85.2$ on 2 and 119 degrees of freedom.
- $P < 2.2 \times 10^{-16}$.

Nested F test in R

```r
anova(m3, m5) # compare the reduced model m3 to the full model m5
```

```
## Analysis of Variance Table
##
## Model 1: Happiness ~ log(GDPpc) + LifeExp
## Model 2: Happiness ~ log(GDPpc) + LifeExp + LifeSat + HDI
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    121 8478.9
## 2    119 3485.6  2    4993.3 85.239 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion:

- (About the association) We reject the null hypothesis that $\beta_3 = \beta_4 = 0$. At least one of the two predictors *LifeSat* and *HDI* is sigficant in explaining *Happiness* given *log(GDPpc)* and *LifeExp* are held constant.
- (About model comparison) The full model including all the 4 predictors is significantly better than the reduced model with only 2 predictors.

Nested F test for a single predictor

```r
anova(m3, m4 <- lm(Happiness ~ log(GDPpc)+LifeExp+LifeSat, data=HappyPlanet))
```

```
## Analysis of Variance Table
##
## Model 1: Happiness ~ log(GDPpc) + LifeExp
## Model 2: Happiness ~ log(GDPpc) + LifeExp + LifeSat
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    121 8478.9
## 2    120 3527.5  1    4951.4 168.44 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- For testing the significance of *LifeSat*, $F = 168.44$, $P < 2.2 \times 10^{-16}$.

```r
summary(m4)$coefficients
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.18764    3.02069   3.704 0.000323 ***
## log(GDPpc)  -7.97988    0.52709 -15.140  < 2e-16 ***
## LifeExp      0.80851    0.06582  12.283  < 2e-16 ***
## LifeSat      7.38646    0.56914  12.978  < 2e-16 ***
```

- For testing the significance of *LifeSat*, $t = 12.978$, $P < 2.2 \times 10^{-16}$.

- $F = 168.44 = t^2 = 12.978^2$, P-values are equal.

Nested F test for a single predictor

```r
anova(m4, m5)
```

```
## Analysis of Variance Table
##
## Model 1: Happiness ~ log(GDPpc) + LifeExp + LifeSat
## Model 2: Happiness ~ log(GDPpc) + LifeExp + LifeSat + HDI
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    120 3527.5
## 2    119 3485.6  1    41.958 1.4325 0.2337
```

- $F = 1.4325 = t^2 = 1.197^2$, both $P = 0.2337$.

```r
summary(m5)$coefficients
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.67493    3.26129   3.886 0.000168 ***
## log(GDPpc)  -8.91513    0.94204  -9.464 3.49e-16 ***
## LifeExp      0.73834    0.08806   8.384 1.20e-13 ***
## LifeSat      7.55823    0.58596  12.899  < 2e-16 ***
## HDI         14.09056   11.77288   1.197 0.233738
```

- Nested F test for a single predictor is equivalent to the t test for the slope of that predictor in the full model.

Adjusted R squared

```r
summary(m3) # Happiness ~ log(GDPpc)+LifeExp
```

```
## Multiple R-squared:  0.4355, Adjusted R-squared:  0.4262
## F-statistic: 46.68 on 2 and 121 DF,  p-value: 9.436e-16
```

```r
summary(m5) # Happiness ~ log(GDPpc)+LifeExp+LifeSat+HDI
```

```
## Multiple R-squared:  0.7679, Adjusted R-squared:  0.7601
## F-statistic: 98.46 on 4 and 119 DF,  p-value: < 2.2e-16
```

- For the reduced model, $R^2 = 0.4355$, $R^2_{adj} = 0.4262$
- For the full model, $R^2 = 0.7679$, $R^2_{adj} = 0.7601$

$$
\begin{aligned}
R^2 &= \frac{SSModel}{SST} = 1 - \frac{SSE}{SST} \\
R^2_{adj} &= 1 - \frac{SSE/(n-K-1)}{SST/(n-1)}
\end{aligned}
$$

Ajudsted R squared

$$R^2_{adj} = 1 - \frac{SSE/(n-K-1)}{SST/(n-1)} < R^2$$

- R^2_{adj} is always smaller than R^2 because $n-1 > n-K-1$.
- R^2_{adj} accounts for both the variability explained by the model as well as the sample size n and number of predictors K used in the model. It penalizes the R^2 values of models with more predictors.
- Although adding new predictors to a model will always increase the R^2 value, it will not necessarily increase the R^2_{adj} value to the same amount or even decrease it.

Adjusted R squared

```r
summary(m3) # Happiness ~ log(GDPpc)+LifeExp
```

```
## Multiple R-squared:  0.4355, Adjusted R-squared:  0.4262
## F-statistic: 46.68 on 2 and 121 DF,  p-value: 9.436e-16
```

```r
summary(m4) # Happiness ~ log(GDPpc)+LifeExp+LifeSat
```

```
## Multiple R-squared:  0.7652, Adjusted R-squared:  0.7593
## F-statistic: 130.3 on 3 and 120 DF,  p-value: < 2.2e-16
```

```r
summary(m5) # Happiness ~ log(GDPpc)+LifeExp+LifeSat+HDI
```

```
## Multiple R-squared:  0.7679, Adjusted R-squared:  0.7601
## F-statistic: 98.46 on 4 and 119 DF,  p-value: < 2.2e-16
```

- Adding *LifeSat*, R^2 increased 0.3297 and R^2_{adj} increased 0.3331.
- However, adding *HDI*, R^2 increased 0.0027 and R^2_{adj} increased 0.0008.
- For model comparisons, we usually use R_{adj} because it measures both the variablity explained by the model and the complexity of the model.

Summary

Multiple linear regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + \epsilon, \text{ where } \epsilon \overset{iid}{\sim} N(0, \sigma)$$

- MLR analysis of variance (ANOVA)
  - F test and R^2
- Three tests
  - t test for the slopes
  - F test for the MLR model
  - t test for the correlations
- Nested F test for a subset of predictors
- Adjusted R^2 for model comparison