Multiple Linear Regression (MLR) Handouts

Yibi Huang

- Data and Models
- Least Squares Estimate, Fitted Values, Residuals
- Sum of Squares
- Do Regression in R
- Interpretation of Regression Coefficients
- t-Tests on Individual Regression Coefficients
- F-Tests on Multiple Regression Coefficients/Goodness-of-Fit

Data for Multiple Linear Regression

Multiple linear regression is a generalized form of simple linear regression, in which the data contains multiple explanatory variables.

	SLR		MLR				
	$\mathbf{x}$	$\mathbf{y}$	$\mathbf{x}_1$	$\mathbf{x}_2$	$\ldots$	$\mathbf{x}_p$	$\mathbf{y}$
case 1:	x_1	y_1	x_{11}	x_{12}	$\ldots$	x_{1p}	y_1
case 2:	x_2	y_2	x_{21}	x_{22}	$\ldots$	x_{2p}	y_2
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
case n:	x_n	y_n	x_{n1}	x_{n2}	$\ldots$	x_{np}	y_n

- For SLR, we observe pairs of variables.
  For MLR, we observe rows of variables.
  Each row (or pair) is called a *case*, a *record*, or a *data point*
- y_i is the *response* (or *dependent variable*) of the ith observation
- There are p *explanatory variables* (or *covariates*, *predictors*, *independent variables*), and x_{ik} is the value of the explanatory variable $\mathbf{x}_k$ of the ith case

Multiple Linear Regression Models

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i \quad \text{where } \varepsilon_i\text{'s are i.i.d. } N(0, \sigma^2)$$

In the model above,

- ε_i's (errors, or noise) are i.i.d. $N(0, \sigma^2)$
- Parameters include:

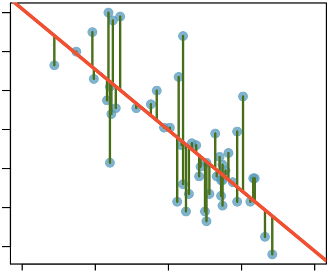
$$\begin{aligned} \beta_0 &= \text{ intercept;} \\ \beta_k &= \text{ regression coefficients (slope) for the } k\text{th} \\ &\qquad \text{explanatory variable,} \quad k = 1, \ldots, p \\ \sigma^2 &= \mathrm{Var}(\varepsilon_i) \text{ is the variance of errors} \end{aligned}$$

- Observed (known): $y_i, x_{i1}, x_{i2}, \ldots, x_{ip}$
  Unknown: β_0, $\beta_1, \ldots, \beta_p$, σ^2, ε_i's
- Random variables: ε_i's, y_i's
  Constants (nonrandom): β_k's, σ^2, x_{ik}'s

Fitting the Model — Least Squares Method

Recall for SLR, the least squares estimate $(\widehat{\beta}_0, \widehat{\beta}_1)$ for (β_0, β_1) is the intercept and slope of the straight line with the minimum sum of squared vertical distance to the data points

$$\sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2.$$



MLR is just like SLR. The least squares estimate $(\widehat{\beta}_0, \ldots, \widehat{\beta}_p)$ for $(\beta_0, \ldots, \beta_p)$ is the intercept and slopes of the (hyper)plane with the minimum sum of squared vertical distance to the data points

$$\sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \ldots - \widehat{\beta}_p x_{ip})^2$$

Solving the Least Squares Problem (1)

From now on, we use the "hat" symbol to differentiate the estimated coefficient $\widehat{\beta}_j$ from the actual unknown coefficient β_j.

To find the $(\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p)$ that minimize

$$L(\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p) = \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \ldots - \widehat{\beta}_p x_{ip})^2$$

one can set the derivatives of L with respect to $\widehat{\beta}_j$ to 0

$$\frac{\partial L}{\partial \widehat{\beta}_0} = -2 \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \ldots - \widehat{\beta}_p x_{ip})$$

$$\frac{\partial L}{\partial \widehat{\beta}_k} = -2 \sum_{i=1}^{n} x_{ik} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \ldots - \widehat{\beta}_p x_{ip}), \ k = 1, 2, \ldots, p$$

and then equate them to 0. This results in a system of $(p+1)$ equations in $(p+1)$ unknowns.

Solving the Least Squares Problem (2)

The least squares estimate $(\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p)$ is the solution to the following system of equations, called *normal equations*.

$$
\begin{array}{lllll}
n\widehat{\beta}_0 & + \widehat{\beta}_1 \sum_{i=1}^n x_{i1} & + \cdots + \widehat{\beta}_p \sum_{i=1}^n x_{ip} & = \sum_{i=1}^n y_i \\
\widehat{\beta}_0 \sum_{i=1}^n x_{i1} & + \widehat{\beta}_1 \sum_{i=1}^n x_{i1}^2 & + \cdots + \widehat{\beta}_p \sum_{i=1}^n x_{i1}x_{ip} & = \sum_{i=1}^n x_{i1}y_i \\
 & & \vdots & \\
\widehat{\beta}_0 \sum_{i=1}^n x_{ik} & + \widehat{\beta}_1 \sum_{i=1}^n x_{ik}x_{i1} & + \cdots + \widehat{\beta}_p \sum_{i=1}^n x_{ik}x_{ip} & = \sum_{i=1}^n x_{ik}y_i \\
 & & \vdots & \\
\widehat{\beta}_0 \sum_{i=1}^n x_{ip} & + \widehat{\beta}_1 \sum_{i=1}^n x_{ip}x_{i1} & + \cdots + \widehat{\beta}_p \sum_{i=1}^n x_{ip}^2 & = \sum_{i=1}^n x_{ip}y_i
\end{array}
$$

- Don't worry about solving the equations.
  R and many other softwares can do the computation for us.
- In general, $\widehat{\beta}_j \neq \beta_j$, but they will be close under some conditions

Fitted Values

The fitted value or predicted value:

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \ldots + \widehat{\beta}_p x_{ip}$$

- Again, the "hat" symbol is used to differentiate the fitted value $\widehat{y}_i$ from the actual observed value y_i.

Residuals

- One cannot directly compute the errors

$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip}$$

  since the coefficients $\beta_0, \beta_1, \ldots, \beta_p$ are *unknown*.
- The errors ε_i can be estimated by the **residuals** e_i defined as:

$$\begin{aligned}
\text{residual } e_i &= \text{observed } y_i - \text{predicted } y_i \\
&= y_i - \widehat{y}_i \\
&= y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \ldots + \widehat{\beta}_p x_{ip}) \\
&= \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i \\
&\quad - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \ldots - \widehat{\beta}_p x_{ip}
\end{aligned}$$

- $e_i \neq \varepsilon_i$ in general since $\widehat{\beta}_j \neq \beta_j$
- Graphical explanation

Properties of Residuals

Recall the least squares estimate $(\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p)$ satisfies the equations

$$\sum_{i=1}^{n} (\underbrace{y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \ldots - \widehat{\beta}_p x_{ip}}_{= y_i - \widehat{y}_i = e_i = \text{residual}}) = 0 \text{ and}$$

$$\sum_{i=1}^{n} x_{ik} (\overbrace{y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \ldots - \widehat{\beta}_p x_{ip}}) = 0, \ k = 1, 2, \ldots, p.$$

Thus the residuals e_i have the properties

$$\underbrace{\sum\nolimits_{i=1}^{n} e_i = 0}_{\text{Residuals add up to 0.}} \quad , \quad \underbrace{\sum\nolimits_{i=1}^{n} x_{ik} e_i = 0, \ k = 1, 2, \ldots, p.}_{\text{Residuals are orthogonal to covariates.}}$$

Sum of Squares

Observe that

$$y_i - \overline{y} = (\widehat{y}_i - \overline{y}) + (y_i - \widehat{y}_i)$$

Squaring up both sides we get

$$(y_i - \overline{y})^2 = (\widehat{y}_i - \overline{y})^2 + (y_i - \widehat{y}_i)^2 + 2(\widehat{y}_i - \overline{y})(y_i - \widehat{y}_i)$$

Summing up over all the cases $i = 1, 2, \ldots, n$, we get

$$\overbrace{\sum_{i=1}^{n}(y_i - \overline{y})^2}^{\text{SST}} = \overbrace{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}^{\text{SSR}} + \overbrace{\sum_{i=1}^{n}(\underbrace{y_i - \widehat{y}_i}_{=e_i})^2}^{\text{SSE}}$$

$$+ \underbrace{2\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})(y_i - \widehat{y}_i)}_{=\,0,\text{ see next page.}}$$

Why $\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})(y_i - \widehat{y}_i) = 0$?

$$
\begin{aligned}
&\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})(\underbrace{y_i - \widehat{y}_i}_{=e_i}) \\
&= \sum_{i=1}^{n}\widehat{y}_i e_i - \sum_{i=1}^{n}\overline{y} e_i \\
&= \sum_{i=1}^{n}(\widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \ldots + \widehat{\beta}_p x_{ip}) e_i - \sum_{i=1}^{n}\overline{y} e_i \\
&= \widehat{\beta}_0 \underbrace{\sum_{i=1}^{n} e_i}_{=0} + \widehat{\beta}_1 \underbrace{\sum_{i=1}^{n} x_{i1} e_i}_{=0} + \ldots + \widehat{\beta}_p \underbrace{\sum_{i=1}^{n} x_{ip} e_i}_{=0} - \overline{y} \underbrace{\sum_{i=1}^{n} e_i}_{=0} \\
&= 0
\end{aligned}
$$

in which we used the properties of residuals that $\sum_{i=1}^{n} e_i = 0$ and $\sum_{i=1}^{n} x_{ik} e_i = 0$ for all $k = 1, \ldots, p$.

Interpretation of Sum of Squares

$$\underbrace{\sum_{i=1}^{n}(y_i - \overline{y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}_{\text{SSR}} + \underbrace{\sum_{i=1}^{n}(\overbrace{y_i - \widehat{y}_i}^{=e_i})^2}_{\text{SSE}}$$

- SST = **total sum of squares**
  - total variability of $\mathbf{y}$
  - depends on the response $\mathbf{y}$ only, not on the form of the model
- SSR = **regression sum of squares**
  - variability of $\mathbf{y}$ explained by $\mathbf{x}_1, \ldots, \mathbf{x}_p$
- SSE = **error (residual) sum of squares**
  - $= \min_{\beta_0, \beta_1, \ldots, \beta_p} \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2$
  - variability of $\mathbf{y}$ not explained by $\mathbf{x}$'s

Degrees of Freedom

If the MLR model $y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i$, ε_i's i.i.d. $\sim N(0, \sigma^2)$ is true, it can be shown that

$$\frac{\text{SSE}}{\sigma^2} \sim \chi^2_{n-p-1},$$

If we further assume that $\beta_1 = \beta_2 = \cdots = \beta_p = 0$, then

$$\frac{\text{SST}}{\sigma^2} \sim \chi^2_{n-1}, \quad \frac{\text{SSR}}{\sigma^2} \sim \chi^2_p$$

and SSR is independent of SSE.

Note the **degrees of freedom** of the 3 chi-square distributions

$$dfT = n - 1, \quad dfR = p, \quad dfE = n - p - 1$$

break down similarly

$$dfT = dfR + dfE$$

just like $SST = SSR + SSE$.

Why SSE Has $n - p - 1$ Degrees of Freedom?

The n residuals $e_1, \ldots, e_n$ cannot all vary freely.

There are $p + 1$ constraints:

$$\sum_{i=1}^{n} e_i = 0 \quad \text{and} \quad \sum_{i=1}^{n} x_{ki} e_i = 0 \text{ for } k = 1, \ldots, p.$$

So only $n - (p + 1)$ of them can be *freely varying*.

The $p + 1$ constraints comes from the $p + 1$ coefficients $\beta_0, \ldots, \beta_p$ in the model, and each contributes one constraint $\frac{\partial}{\partial \beta_k} = 0$.

Mean Square Error (MSE) — Estimate of σ^2

The **mean squares** is the sum of squares divided by its degrees of freedom:

$$MST = \frac{\text{SST}}{dfT} = \frac{\text{SST}}{n-1} = \text{sample variance of } Y,$$
$$MSR = \frac{\text{SSR}}{dfR} = \frac{\text{SSR}}{p},$$
$$MSE = \frac{\text{SSE}}{dfE} = \frac{\text{SSE}}{n-p-1} = \widehat{\sigma}^2$$

- From the fact $\frac{\text{SSE}}{\sigma^2} \sim \chi^2_{n-p-1}$ and that the mean of a χ^2_k distribution is k, we know that **MSE is an unbiased estimator for σ^2**.
- Though SSE always decreases as we add terms to the model, adding unimportant terms may increases MSE.

Example: Housing Price

Price	BDR	FLR	FP	RMS	ST	LOT	BTH	CON	GAR	LOC
53	2	967	0	5	0	39	1.5	1	0.0	0
55	2	815	1	5	0	33	1.0	1	2.0	0
56	3	900	0	5	1	35	1.5	1	1.0	0
58	3	1007	0	6	1	24	1.5	0	2.0	0
64	3	1100	1	7	0	50	1.5	1	1.5	0
44	4	897	0	7	0	25	2.0	0	1.0	0
49	5	1400	0	8	0	30	1.0	0	1.0	0
70	3	2261	0	6	0	29	1.0	0	2.0	0
72	4	1290	0	8	1	33	1.5	1	1.5	0
82	4	2104	0	9	0	40	2.5	1	1.0	0
85	8	2240	1	12	1	50	3.0	0	2.0	0
45	2	641	0	5	0	25	1.0	0	0.0	1
47	3	862	0	6	0	25	1.0	1	0.0	1
49	4	1043	0	7	0	30	1.5	0	0.0	1
56	4	1325	0	8	0	50	1.5	0	0.0	1
60	2	782	0	5	1	25	1.0	0	0.0	1
62	3	1126	0	7	1	30	2.0	1	0.0	1
64	4	1226	0	8	0	37	2.0	0	2.0	1
.										
.										
.										
50	2	691	0	6	0	30	1.0	0	2.0	0
65	3	1023	0	7	1	30	2.0	1	1.0	0

Price = Selling price in $1000
BDR = Number of bedrooms
FLR = Floor space in sq. ft.
FP = Number of fireplaces
RMS = Number of rooms
ST = Storm windows (1 if present, 0 if absent)

LOT = Front footage of lot in feet
BTH = Number of bathrooms
CON = Construction (1 if frame, 0 if brick)

GAR = Garage size (0 = no garage, 1 = one-car garage, etc.)

LOC = Location (1 if property is in zone A, 0 otherwise)

How to Do Regression Using R?

```
> housing = read.table("housing.txt",h=TRUE)    # to load the data
> lm(Price ~ FLR+LOT+BDR+GAR+ST, data=housing)

Call:
lm(formula = Price ~ FLR + LOT + BDR + GAR + ST, data = housing)

Coefficients:
(Intercept)          FLR          LOT          BDR          GAR           ST
   24.63232      0.02009      0.44216     -3.44509      3.35274     11.64033
```

The `lm()` command above asks R to fit the model

$$\text{Price} = \beta_0 + \beta_1\text{FLR} + \beta_2\text{LOT} + \beta_3\text{BDR} + \beta_4\text{GAR} + \beta_5\text{ST} + \varepsilon$$

and R gives us the regression equation

$$\widehat{\text{Price}} = 24.63 + 0.02\text{FLR} + 0.44\text{LOT} - 3.45\text{BDR} + 3.35\text{GAR} + 11.64\text{ST}$$

$$\widehat{\text{Price}} = 24.63 + 0.02\text{FLR} + 0.44\text{LOT} - 3.45\text{BDR} + 3.35\text{GAR} + 11.64\text{ST}$$

The regression equation tells us:

- an extra square foot in floor area increases the price by \$20,
- an extra foot in front footage by \$440,
- an additional bedroom by .... −\$3450,
- an additional space in the garage by \$3350,
- using storm windows by .... \$11640.

Question:

Why an additional bedroom makes a house less valuable?

Interpretation of Regression Coefficients

- β_0 = intercept = the mean value of y when all x_j' are 0.
  - may not have practical meaning
    e.g., β_0 is meaningless in the housing price model as no housing unit has 0 floor space.

- β_j: regression coefficient for x_j, is the mean change in the response y when x_j is increased by one unit *holding all other x_j's constant*
  - Interpretation of β_j depends on the presence of other covariates in the model
    e.g., the meaning of the 2 β_1's in the following 2 models are different

$$\begin{aligned} \text{Model 1}: &\quad y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i \\ \text{Model 2}: &\quad y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i. \end{aligned}$$

What's Wrong?

```
# Model 1
> lm(Price ~ BDR, data=housing)

(Intercept)          BDR
     43.487        3.921
```

The regression coefficient for BDR is 3.921 in the Model 1 above but -3.445 in the Model 2 below.

```
# Model 2
> lm(Price ~ FLR+LOT+BDR+GAR+ST, data=housing)

(Intercept)         FLR         LOT         BDR         GAR          ST
   24.63232     0.02009     0.44216    -3.44509     3.35274    11.64033
```

Considering `BDR` alone, house prices *increase* with `BDR`.

However, an extra bedroom makes a housing unit less valuable when when other covariates (`FLR`, `LOT`, etc) are fixed.

Does this make sense?

More R Commands

```
> lm1 = lm(Price ~ FLR+RMS+BDR+GAR+LOT+ST+CON+LOC, data=housing)
> summary(lm1)        # Regression output with more details
                      # including multiple R-squared,
                      # and the estimate of sigma

> lm1$coef            # show the estimated beta's
> lm1$fitted          # show the fitted values
> lm1$res             # show the residuals
```

```
> lm1 = lm(Price ~ FLR+LOT+BDR+GAR+ST, data=housing)
> summary(lm1)

Call:
lm(formula = Price ~ FLR + LOT + BDR + GAR + ST, data = housing)

Residuals:
    Min      1Q  Median      3Q     Max
-9.7530 -2.9535  0.1779  3.7183 12.9728

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 24.632318   4.836743   5.093 5.56e-05 ***
FLR          0.020094   0.003668   5.478 2.31e-05 ***
LOT          0.442164   0.150023   2.947 0.007965 **
BDR         -3.445086   1.279347  -2.693 0.013995 *
GAR          3.352739   1.560239   2.149 0.044071 *
ST          11.640334   2.688867   4.329 0.000326 ***
---
Residual standard error: 5.79 on 20 degrees of freedom
Multiple R-squared:  0.8306,Adjusted R-squared:  0.7882
F-statistic: 19.61 on 5 and 20 DF,  p-value: 4.306e-07
```

t-Tests on Individual Regression Coefficients

For a MLR model $Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \varepsilon_i$, to test the hypotheses,

$$H_0 : \beta_j = c \quad \text{v.s.} \quad H_a : \beta_j \neq c$$

the t-statistic is

$$t = \frac{\widehat{\beta}_j - c}{\mathrm{SE}(\widehat{\beta}_j)}$$

in which $\mathrm{SE}(\widehat{\beta}_j)$ is the standard error for $\widehat{\beta}_j$.

- General formula for $\mathrm{SE}(\widehat{\beta}_j)$ is a bit complicate but unimportant in STAT222 and hence is omitted
- R can compute $\mathrm{SE}(\widehat{\beta}_j)$ for us
- Formula for $\mathrm{SE}(\widehat{\beta}_j)$ for a few special cases will be given later

This t-statistic also has a t-distribution with $n - p - 1$ degrees of freedom

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 24.632318   4.836743   5.093 5.56e-05 ***
FLR          0.020094   0.003668   5.478 2.31e-05 ***
LOT          0.442164   0.150023   2.947 0.007965 **
......(some rows are omitted)
ST          11.640334   2.688867   4.329 0.000326 ***
```

- the first column gives variable names
- the column `Estimate` gives the LS estimate $\widehat{\beta}_j$'s for β_j's
- the column `Std. Error` gives $\mathrm{SE}(\widehat{\beta}_j)$, the standard error of $\widehat{\beta}_j$
- the column `t value` gives t-value $= \frac{\widehat{\beta}_j}{SE(\widehat{\beta}_j)}$
- column `Pr(>|t|)` gives the P-value for testing H_0: $\beta_j = 0$ v.s. H_a: $\beta_j \neq 0$.

E.g., for `LOT`, we see

$$\widehat{\beta}_{LOT} \approx 0.442,\ \mathrm{SE}(\widehat{\beta}_{LOT}) \approx 0.150,\ t = \frac{\widehat{\beta}_{LOT}}{\mathrm{SE}(\widehat{\beta}_{LOT})} \approx \frac{0.442}{0.150} \approx 2.947.$$

The P-value 0.007965 is the 2-sided P-value for testing H_0: $\beta_{LOT} = 0$

Nested Models

We say Model 1 is **nested in** Model 2 if Model 1 is a special case of Model 2 (and hence Model 2 is an extension of Model 1). E.g., for the 4 models below,

$$
\begin{aligned}
\text{Model A}: \; Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \\
\text{Model B}: \; Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \\
\text{Model C}: \; Y &= \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \varepsilon \\
\text{Model D}: \; Y &= \beta_0 + \beta_1 (X_1 + X_2) + \varepsilon
\end{aligned}
$$

- B is nested in A . . . . since A reduces to B when $\beta_3 = 0$
- C is also nested in A . . . . since A reduces to C when $\beta_2 = 0$
- D is nested in B . . . . since B reduces to D when $\beta_1 = \beta_2$
- B and C are NOT nested in either way
- D is NOT nested in C

Nested Relationship is Transitive

If Model 1 is nested in Model 2, and Model 2 is nested in Model 3, then Model 1 is also nested in Model 3.

For example, for models in the previous slide,

D is nested in B, and B is nested in A,

implies D is also nested in A, which is clearly true because Model A reduces to Model D when

$$\beta_1 = \beta_2, \text{ and } \beta_3 = 0.$$

When two models are nested (Model 1 is nested in Model 2),

- the smaller model (Model 1) is called the **reduced model**, and
- the more general model (Model 2) is called the **full model**.

SST of Nested Models

Question: Compare the SST for Model A and the SST for Model B. Which one is larger? Or are they equal?

What about the SST for Model C? For Model D?

SSE of Nested Models

When a reduced model is nested in a full model, then

$$\text{(i) } SSE_{reduced} \geq SSE_{full}, \text{ and (ii) } SSR_{reduced} \leq SSR_{full}.$$

Proof. We will prove (i) for

- the full model $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$ and
- the reduced model $y_i = \beta_0 + \beta_1 x_{i1} + \beta_3 x_{i3} + \varepsilon_i$.

The proofs for other nested models are similar.

$$\begin{aligned} SSE_{full} &= \min_{\beta_0,\beta_1,\beta_2,\beta_3} \sum_{i=1}^{n} (y_1 - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \beta_3 x_{i3})^2 \\ &\leq \min_{\beta_0,\beta_1,\beta_3} \sum_{i=1}^{n} (y_1 - \beta_0 - \beta_1 x_{i1} - \beta_3 x_{i3})^2 \\ &= SSE_{reduced} \end{aligned}$$

Part (ii) follows directly from (i), the identity $SST = SSR + SSE$, and the fact that all MLR models of the same data set have a common SST

General Framework for Testing Nested Models

$\boxed{\text{H}_0\text{: reduced model is true}}$ v.s. $\boxed{\text{H}_a\text{ : full model is true}}$

- As the reduced model is nested in the full model,

$$SSE_{reduced} \geq SSE_{full}$$

- Simplicity or Accuracy?
  - The full model fits the data better (with smaller SSE) but is more complicate
  - The reduced model doesn't fit as well but is simpler.
  - If $\text{SSE}_{reduced} \approx \text{SSE}_{full}$, one can sacrifice a bit of accuracy in exchange for simplicity
  - If $\text{SSE}_{reduced} \gg \text{SSE}_{full}$, it would cost to much in accuracy in exchange for simplicity. The full model is preferred.

The F-Statistic

$$F = \frac{(SSE_{reduced} - SSE_{full})/(df_{reduced} - df_{full})}{MSE_{full}}$$

- $SSE_{reduced} - SSE_{full}$ is the reduction in SSE from replacing the reduced model with the full model.
- $df_{reduced}$ is the df for error for the reduced model.
- df_{full} is the df for error for the full model.
- $F \geq 0$ since $SSE_{reduced} \geq SSE_{full} \geq 0$
- The smaller the F-statistic, the more we favor the reduced model
- Under H_0, the F-statistic has an F-distribution with $df_{reduced} - df_{full}$ and df_{full} degrees of freedom.

Testing All Coefficients Equal Zero

Testing the hypotheses

$$H_0\text{: } \beta_1 = \cdots = \beta_p = 0 \text{ v.s. } H_a\text{: not all } \beta_1 \ldots, \beta_p = 0$$

is a test to evaluate the **overall significance** of a model.

$$\begin{aligned} \text{Full :} y_i &= \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i \\ \text{Reduced :} y_i &= \beta_0 + \varepsilon_i \quad \text{(all covariates are unnecessary)} \end{aligned}$$

- The LS estimate for β_0 in the reduced model is $\widehat{\beta}_0 = \overline{y}$, so

$$SSE_{reduced} = \sum_{i=1}^{n} (y_i - \widehat{\beta}_0)^2 = \sum_i (y_i - \overline{y})^2 = SST_{full}$$

- $df_{reduced} = dfE_{reduced} = n - 1$,
  because the reduced model has only one coefficient β_0
- $df_{full} = dfE_{full} = n - p - 1$.

Testing All Coefficients Equal Zero

Hence

$$
\begin{aligned}
F &= \frac{(SSE_{reduced} - SSE_{full})/(df_{reduced} - df_{full})}{MSE_{full}} \\
&= \frac{(SST_{full} - SSE_{full})/[n-1-(n-p-1)]}{SSE_{full}/(n-p-1)} \\
&= \frac{SSR_{full}/p}{SSE_{full}/(n-p-1)}.
\end{aligned}
$$

Moreover, $F \sim F_{p,n-p-1}$ under H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$.

In R, the F statistic and p-value are displayed in the last line of the output of the `summary()` command.

```
> lm1 = lm(Price ~ FLR+LOT+BDR+GAR+ST, data=housing)
> summary(lm1)
... (output omitted)

Residual standard error: 5.79 on 20 degrees of freedom
Multiple R-squared:  0.8306,Adjusted R-squared:  0.7882
F-statistic: 19.61 on 5 and 20 DF,  p-value: 4.306e-07
```

ANOVA and the F-Test

The test of all coefficients equal zero is often summarized in an ANOVA table.

Source	df	Sum of Squares	Mean Squares	F
Regression	$dfR = p$	SSR	$MSR = \dfrac{SSR}{dfR}$	$F = \dfrac{MSR}{MSE}$
Error	$dfE = n - p - 1$	SSE	$MSE = \dfrac{SSE}{dfE}$	
Total	$dfT = n - 1$	SST		

Testing Some Coefficients Equal to Zero

E.g., for the housing price data, we may want to test if we can eliminate `BDR` and `GAR` from the model,
i.e., H_0: $\beta_{BDR} = \beta_{GAR} = 0$.

```
> lmfull = lm(Price ~ FLR+LOT+BDR+GAR+ST, data=housing)
> lmreduced = lm(Price ~ FLR+LOT+ST, data=housing)
> anova(lmreduced, lmfull)
Analysis of Variance Table

Model 1: Price ~ FLR + LOT + ST
Model 2: Price ~ FLR + LOT + BDR + GAR + ST
  Res.Df     RSS Df Sum of Sq      F   Pr(>F)
1     22 1105.01
2     20  670.55  2    434.46 6.4792 0.006771 **
```

Note SSE is called RSS (residual sum of square) in R.

Testing Equality of Coefficients

Example. To test H_0: $\beta_1 = \beta_2 = \beta_3$, the reduced model is

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 X_1 + \beta_1 X_2 + \beta_1 X_3 + \beta_4 X_4 + \varepsilon \\
&= \beta_0 + \beta_1 (X_1 + X_2 + X_3) + \beta_4 X_4 + \varepsilon
\end{aligned}
$$

1. Make a new variable $W = X_1 + X_2 + X_3$
2. Fit the reduced model by regressing Y on W and X_4
3. Find $\text{SSE}_{reduced}$ and $\text{df}_{reduced} - \text{df}_{full} =$ 2
4. In R

```
> lmfull = lm(Y ~ X1 + X2 + X3 + X4)
> lmreduced = lm(Y ~ I(X1 + X2 + X3) + X4)
> anova(lmreduced, lmfull)
```

The line `lmreduced = lm(Y ~ I(X1 + X2 + X3) + X4)` is equivalent to

```
> W = X1 + X2 + X3
> lmreduced = lm(Y ~ W + X4)
```