STAT 224 Lecture 4
Multiple Linear Regression, Part 3

Yibi Huang
Department of Statistics
University of Chicago

Outline

- Accuracy of Predictions
    - Confidence Intervals for Predictions
    - Prediction Intervals for Predictions
- Sum of Squares
- Model Comparison

Accuracy of Predictions for SLR

Two Kinds of Predictions

There are *TWO kinds of predictions* for the response Y given $X = x_0$ based on a SLR model $Y = \beta_0 + \beta_1 X + \varepsilon$:

- given $X = x_0$, estimation of the **mean response**

$$\mathrm{E}[Y|X = x_0] = \beta_0 + \beta_1 x_0$$

- given $X = x_0$, prediction of the response for **one specific observation**

$$Y = \beta_0 + \beta_1 x_0 + \varepsilon$$

For the Fire Damage example in L03, one may want to

- estimate the **average** fire damage for **all** houses located 2 miles away from the nearest fire station, which is $\beta_0 + 2\beta_1$
- predict the fire damage for **a specific house** located 2 miles away from the nearest fire station which is $\beta_0 + 2\beta_1 + \varepsilon$

Estimation v.s. Prediction

The first one is an **estimation** problem as $\beta_0 + \beta_1 x_0$ only involve fixed parameters β_0, β_1, and a known number x_0.

The second one is a **prediction** problem as $\beta_0 + \beta_1 x_0 + \varepsilon$ involve a random number ε

Estimated Value and Predicted Value

Both

$$\mathrm{E}[Y|X_0] = \beta_0 + \beta_1 x_0 \quad \text{and} \quad Y = \beta_0 + \beta_1 x_0 + \varepsilon$$

are estimated/predicted by

$$\widehat{\beta}_0 + \widehat{\beta}_1 x_0$$

The noise ε for a future observation is predicted to be its mean 0. We cannot make a better prediction for ε from the observed (x_i, y_i)'s since ε independent of all observed (x_i, y_i)'s.

The Two Prediction Problems Differ in Uncertainty!

For estimating $\mathrm{E}[Y|X = x_0] = \beta_0 + \beta_1 x_0$, the variance for the estimate $\widehat{\beta}_0 + \widehat{\beta}_1 x_0$ can be shown to be

$$\mathrm{Var}\left(\widehat{\beta}_0 + \widehat{\beta}_1 x_0\right) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)$$

To predict $Y = \beta_0 + \beta_1 x_0 + \varepsilon$, we need to include the extra variability from the noise ε.

$$\begin{aligned}\mathrm{Var}\left(\widehat{\beta}_0 + \widehat{\beta}_1 x_0 + \varepsilon\right) &= \mathrm{Var}\left(\widehat{\beta}_0 + \widehat{\beta}_1 x_0\right) + \mathrm{Var}(\varepsilon) \\ &= \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) + \sigma^2\end{aligned}$$

As n gets large,

- $\mathrm{Var}(\widehat{\beta}_0 + \widehat{\beta}_1 x_0)$ would go down to 0, but
- $\mathrm{Var}(\widehat{\beta}_0 + \widehat{\beta}_1 x_0 + \varepsilon)$ just goes down to σ^2.

What Affects the Accuracy of Prediction?

Recall the variances for the two prediction problems are

$$\begin{cases} \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) & \text{for estimating } \mathrm{E}[Y|X = x_0] = \beta_0 + \beta_1 x_0 \\ \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) & \text{to predict } Y \text{ when } X = x_0 \end{cases}$$

An accurate prediction (less variance) comes from

- small σ^2 (i.e., small noise ε's)
- large sample size n
- large $\sum_{i=1}^n (x_i - \bar{x})^2$ (more spread in predictors)
- small $(x_0 - \bar{x})^2$

Confidence Intervals and Prediction Intervals

The $100(1-\alpha)\%$ confidence interval for $\beta_0 + \beta_1 x_0$ is

$$\widehat{\beta}_0 + \widehat{\beta}_1 x_0 \pm t_{(n-2,\alpha/2)}\,\widehat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

The $100(1-\alpha)\%$ prediction interval for $Y = \beta_0 + \beta_1 x_0 + \varepsilon$ is

$$\widehat{\beta}_0 + \widehat{\beta}_1 x_0 \pm t_{(n-2,\alpha/2)}\,\widehat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

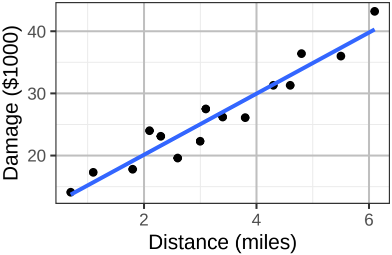
where $\widehat{\sigma} = \sqrt{\text{MSE}}$.

Example: Fire Damage Data

Recall the fire damage data in L03. The variables are

- `dist`: distance to the nearest fire station in miles
- `damage`: amount of fire damage in $1000

```
fire = data.frame(
  dist=c(0.7,1.1,1.8,2.1,2.3,2.6,3.0,3.1,3.4,3.8,4.3,4.6,4.8,5.5,6.1),
  damage=c(14.1,17.3,17.8,24.0,23.1,19.6,22.3,27.5,26.2,26.1,31.3,
           31.3,36.4,36.0,43.2)
  )
```

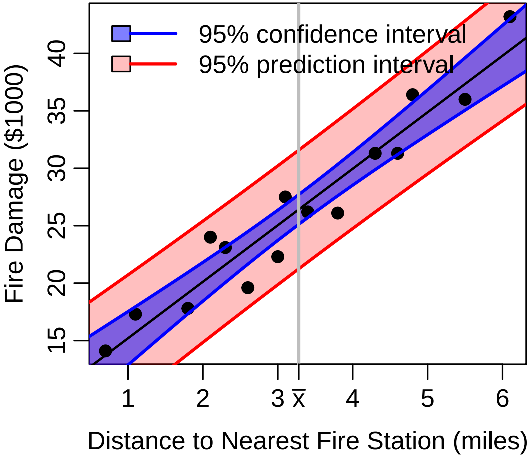


Confidence Intervals and Prediction Intervals in R

```
lmfire = lm(damage ~ dist, data = fire)
predict(lmfire, data.frame(dist=2), interval="confidence")
    fit   lwr  upr
1 20.12 18.43 21.8
predict(lmfire, data.frame(dist=2), interval="prediction")
    fit   lwr  upr
1 20.12 14.84 25.4
```

- For houses located 2 miles away from the nearest fire station, the <u>average</u> fire damage is estimated to be \$20,120 with a 95% confidence interval from \$18,430 to \$21.800.
- When a house located 2 miles away from the nearest fire station, the fire damage is between \$14,840 to \$25,400 with 95% confidence.
- The prediction interval for a **single** house is wider.

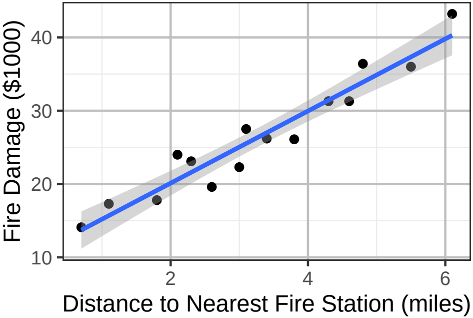
The plot below shows the 95% confidence intervals and the 95% prediction intervals at different values of x_0.



Both the confidence intervals and the prediction intervals are **narrowest when** $x_0 = \bar{x}$.

`geom_smooth(method='lm')` in `ggplot()` by default includes the 95% confidence intervals for estimating $\mathrm{E}(y|X = x_0)$.

```
library(ggplot2)
ggplot(fire, aes(x=dist, y=damage)) + geom_point() +
  geom_smooth(method='lm', formula='y~x') +
  xlab("Distance to Nearest Fire Station (miles)") +
  ylab("Fire Damage ($1000)")
```



Accuracy of Predictions for MLR

Accuracy of Predictions for MLR

An MLR model $Y = \beta_0 + \beta_1 X_1 + \cdots \beta_p X_p + \varepsilon$ also has **two** kinds of conditional prediction problems of the response Y given the values of the predictors:

$$X_1 = x_{01}, \ldots, X_p = x_{0p}.$$

- estimation of the **mean response** given $X_1 = x_{01}, \ldots, X_p = x_{0p}$

$$\mathrm{E}[Y|X_0] = \beta_0 + \beta_1 x_{01} + \cdots + \beta_p x_{0p}$$

- prediction of the response for **one specific observation** given $X_1 = x_{01}, \ldots X_p = x_{0p}$

$$Y = \beta_0 + \beta_1 x_{01} + \cdots + \beta_p x_{0p} + \varepsilon$$

Just like SLR, two problems have identical estimated/predicted values

$$\widehat{\beta}_0 + \widehat{\beta}_1 x_{01} + \cdots + \widehat{\beta}_p x_{0p}$$

but their standard errors are different

$$s.e.(\widehat{\mathrm{E}(Y|X_0)}) = \widehat{\sigma}\sqrt{\mathbf{x}_0^T (X^T X)^{-1} \mathbf{x}_0}$$

$$s.e.(\widehat{Y}|X_0) = \widehat{\sigma}\sqrt{1 + \mathbf{x}_0^T (X^T X)^{-1} \mathbf{x}_0}$$

where $\mathbf{x}_0^T = (1, x_{01}, \ldots, x_{0p})^T$.

Confidence Intervals and Prediction Intervals

The $100(1-\alpha)\%$ confidence interval for
$\mathrm{E}[Y|X_1 = x_{01}, \ldots X_p = x_{0p}] = \beta_0 + \beta_1 x_{01} + \cdots + \beta_p x_{0p}$ is

$$\widehat{\beta}_0 + \widehat{\beta}_1 x_{01} + \cdots + \widehat{\beta}_p x_{0p} \pm t_{(n-p-1,\alpha/2)}\, s.e.(\widehat{\mathrm{E}(Y|X_0})) $$

The $100(1-\alpha)\%$ prediction interval for
$Y = \beta_0 + \beta_1 x_{01} + \cdots + \beta_p x_{0p} + \varepsilon$ is

$$\widehat{\beta}_0 + \widehat{\beta}_1 x_{01} + \cdots + \widehat{\beta}_p x_{0p} \pm t_{(n-p-1,\alpha/2)}\, s.e.(\widehat{Y}|X_0)$$

For the `trees` data in L03

```
data(trees)
trees$Diameter = trees$Girth
lmtrees = lm(log(Volume) ~ log(Diameter) + log(Height), data=trees)
predict(lmtrees, data.frame(Diameter=10, Height = 70),
        interval = "confidence")
   fit   lwr   upr
1 2.68 2.633 2.726
predict(lmtrees, data.frame(Diameter=10, Height = 70),
        interval = "prediction")
   fit   lwr   upr
1 2.68 2.507 2.853
```

- The mean log(Volume) for all 70-ft-tall, 10 ft in diameter, cherry trees is estimated to between 2.633 to 2.726, at 95% confidence level
- The log(Volume) for a randomly selected 70-ft-tall cherry tree with a diameter of 10 ft is predicted to be between 2.507 to 2.853.

One can exponentiate the intervals to get intervals for `Volume` rather than for `log(Volume)`.

```
predict(lmtrees, data.frame(Diameter=10, Height = 70),
        interval = "confidence")
   fit   lwr   upr
1 2.68 2.633 2.726
predict(lmtrees, data.frame(Diameter=10, Height = 70),
        interval = "prediction")
   fit   lwr   upr
1 2.68 2.507 2.853
```

- The mean Volume for all 70-ft-tall, 10 ft in diameter, cherry trees is estimated to between $e^{2.633} \approx 13.92$ to $e^{2.726} \approx 15.27$ cubic ft, at 95% confidence level
- The Volume for a randomly selected 70-ft-tall cherry tree with a diameter of 10 ft is predicted to be between $e^{2.507} \approx 12.26$ to $e^{2.853} \approx 17.34$ cubic ft.

Sum of Squares, Degrees of Freedom, Mean Squares

Sum of Squares

Observe that

$$y_i - \bar{y} = \underbrace{(\widehat{y_i} - \bar{y})}_{a} + \underbrace{(y_i - \widehat{y_i})}_{b}$$

Squaring up both sides using the identity $(a+b)^2 = a^2+b^2+2ab$, we get

$$(y_i - \bar{y})^2 = \underbrace{(\widehat{y_i} - \bar{y})^2}_{a^2} + \underbrace{(y_i - \widehat{y_i})^2}_{b^2} + \underbrace{2(\widehat{y_i} - \bar{y})(y_i - \widehat{y_i})}_{2ab}$$

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

$$\overbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}^{\text{SST}} = \overbrace{\sum_{i=1}^{n}(\widehat{y_i} - \bar{y})^2}^{\text{SSR}} + \overbrace{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}^{\text{SSE}} + 2\underbrace{\sum_{i=1}^{n}(\widehat{y_i} - \bar{y})(y_i - \widehat{y_i})}_{\text{= 0, see next page.}}$$

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

$$
\begin{aligned}
&\sum_{i=1}^{n}(\widehat{y_i}-\overline{y})(\underbrace{y_i-\widehat{y_i}}_{=e_i})\\
&=\sum_{i=1}^{n}\widehat{y_i}e_i-\sum_{i=1}^{n}\overline{y}e_i\\
&=\sum_{i=1}^{n}\left(\widehat{\beta}_0+\widehat{\beta}_1x_{i1}+\ldots+\widehat{\beta}_px_{ip}\right)e_i-\sum_{i=1}^{n}\overline{y}e_i\\
&=\widehat{\beta}_0\underbrace{\sum_{i=1}^{n}e_i}_{=0}+\widehat{\beta}_1\underbrace{\sum_{i=1}^{n}x_{i1}e_i}_{=0}+\ldots+\widehat{\beta}_p\underbrace{\sum_{i=1}^{n}x_{ip}e_i}_{=0}-\overline{y}\underbrace{\sum_{i=1}^{n}e_i}_{=0}\\
&=0
\end{aligned}
$$

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 Y
  - depends on the response Y only, not on the form of the model
- SSR = **regression sum of squares**
  - variability of Y explained by $X_1, \ldots, X_p$
- 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 Y not explained by the 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.

Multiple R^2 and Adjusted R^2

Multiple R-Squared

Multiple R^2, also called the **coefficient of determination**, is defined as

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$
$$= \text{proportion of variability in } Y \text{ explained by } X_1, \ldots, X_p$$

- $0 \le R^2 \le 1$

Multiple R-Squared

Multiple R^2, also called the **coefficient of determination**, is defined as

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$
$$= \text{proportion of variability in } Y \text{ explained by } X_1, \ldots, X_p$$

- $0 \leq R^2 \leq 1$
- For SLR, $R^2 = r_{xy}^2$ is the square of the correlation between X and Y. So multiple R^2 is a generalization of the correlation

Multiple R-Squared

Multiple R^2, also called the **coefficient of determination**, is defined as

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$
$$= \text{proportion of variability in } Y \text{ explained by } X_1, \ldots, X_p$$

- $0 \leq R^2 \leq 1$
- For SLR, $R^2 = r_{xy}^2$ is the square of the correlation between X and Y. So multiple R^2 is a generalization of the correlation
- For MLR, R^2 is the square of the correlation between Y and $\widehat{Y}$

Multiple R-Squared

Multiple R^2, also called the **coefficient of determination**, is defined as

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$
$$= \text{proportion of variability in } Y \text{ explained by } X_1, \ldots, X_p$$

- $0 \leq R^2 \leq 1$
- For SLR, $R^2 = r_{xy}^2$ is the square of the correlation between X and Y. So multiple R^2 is a generalization of the correlation
- For MLR, R^2 is the square of the correlation between Y and $\widehat{Y}$
- When more terms are added into a model, R^2 may increase or stay the same but never decrease

Multiple R-Squared

Multiple R^2, also called the **coefficient of determination**, is defined as

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$
$$= \text{proportion of variability in } Y \text{ explained by } X_1, \ldots, X_p$$

- $0 \leq R^2 \leq 1$
- For SLR, $R^2 = r_{xy}^2$ is the square of the correlation between X and Y. So multiple R^2 is a generalization of the correlation
- For MLR, R^2 is the square of the correlation between Y and $\widehat{Y}$
- When more terms are added into a model, R^2 may increase or stay the same but never decrease
- Is large R^2 always preferable?

Adjusted R-Squared

Since R^2 always increases as we add terms to the model, some people prefer to use an **adjusted** R^2 defined as

$$R^2_{adj} = 1 - \frac{\text{SSE/dfE}}{\text{SST/dfT}} = 1 - \frac{\text{SSE}/(n-p-1)}{\text{SST}/(n-1)}$$
$$= 1 - \frac{n-1}{n-p-1}(1-R^2).$$

- $-\frac{p}{n-p-1} \le R^2_{adj} \le R^2 \le 1$

Adjusted R-Squared

Since R^2 always increases as we add terms to the model, some people prefer to use an **adjusted** R^2 defined as

$$R^2_{adj} = 1 - \frac{\text{SSE/dfE}}{\text{SST/dfT}} = 1 - \frac{\text{SSE}/(n-p-1)}{\text{SST}/(n-1)}$$
$$= 1 - \frac{n-1}{n-p-1}(1-R^2).$$

- $-\frac{p}{n-p-1} \le R^2_{adj} \le R^2 \le 1$
- Unlike R^2, R^2_{adj} can be negative

Adjusted R-Squared

Since R^2 always increases as we add terms to the model, some people prefer to use an **adjusted** R^2 defined as

$$R^2_{adj} = 1 - \frac{\text{SSE/dfE}}{\text{SST/dfT}} = 1 - \frac{\text{SSE}/(n-p-1)}{\text{SST}/(n-1)}$$
$$= 1 - \frac{n-1}{n-p-1}(1-R^2).$$

- $-\frac{p}{n-p-1} \leq R^2_{adj} \leq R^2 \leq 1$
- Unlike R^2, R^2_{adj} can be negative
- R^2_{adj} does not always increase as more variables are added. In fact, if unnecessary terms are added, R^2_{adj} may decrease.

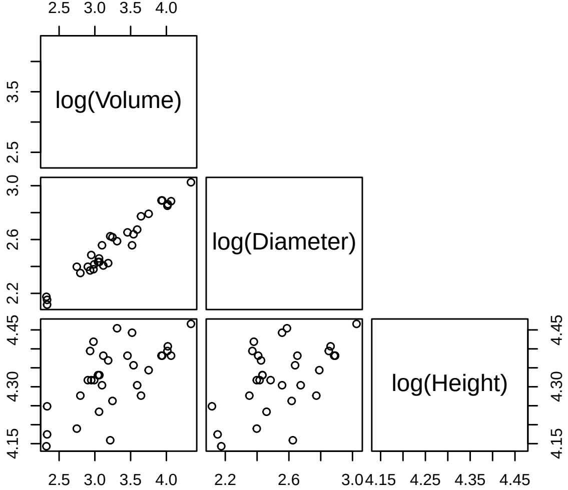
R^2 and R^2_{adj} in R

```
> lmtrees = lm(log(Volume) ~ log(Diameter) + log(Height), data = trees)
> summary(lmtrees)
... (output omitted)
Residual standard error: 0.08139 on 28 degrees of freedom
Multiple R-squared:  0.9777,    Adjusted R-squared:  0.9761
F-statistic: 613.2 on 2 and 28 DF,  p-value: < 2.2e-16
```

The R output above shows that $R^2 = 0.9777$ and $R^2_{adj} = 0.9761$.

The predictors `log(Diameter)` and `log(Height)` can explain 97.77% of the variation in `log(Volume)`.

Model	R^2	R^2_{adj}
`log(Volume) ~ log(Height)`	0.4207	0.4008
`log(Volume) ~ log(Diameter)`	0.9539	0.9523
`log(Volume) ~ log(Diameter) + log(Height)`	0.9777	0.9761



F-Tests on Multiple Regression Coefficients

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,

$$\text{Model A} : Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

$$\text{Model B} : Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

$$\text{Model C} : Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \varepsilon$$

$$\text{Model D} : Y = \beta_0 + \beta_1 (X_1 + X_2) + \varepsilon$$

- B is nested in A . . . . since A reduces to B when $\beta_3 = 0$

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 $\beta_2 = 0$

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,

$$\text{Model A}: Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

$$\text{Model B}: Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

$$\text{Model C}: Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \varepsilon$$

$$\text{Model D}: Y = \beta_0 + \beta_1 (X_1 + X_2) + \varepsilon$$

- 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 $\beta_1 = \beta_2$

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,

$$\text{Model A}: Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

$$\text{Model B}: Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

$$\text{Model C}: Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \varepsilon$$

$$\text{Model D}: Y = \beta_0 + \beta_1 (X_1 + X_2) + \varepsilon$$

- 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

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,

$$\text{Model A}: Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

$$\text{Model B}: Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

$$\text{Model C}: Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \varepsilon$$

$$\text{Model D}: Y = \beta_0 + \beta_1 (X_1 + X_2) + \varepsilon$$

- 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

Nesting 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

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

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

SST of Nested Models

Question: Compare the SST's for Model A, B, C, and D. Which one is the largest? Or are they equal?

SST of Nested Models

Question: Compare the SST's for Model A, B, C, and D. Which one is the largest? Or are they equal?

The 4 models have an identical SST.
SST $= \sum_{i=1}^{n}(y_i - \overline{y})^2$ only depends on the response y but not on which predictors are included in the model.

SSE of Nested Models

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

$$\text{(i) } \text{SSE}_{reduced} \geq \text{SSE}_{full}, \text{ and (ii) } \text{SSR}_{reduced} \leq \text{SSR}_{full}.$$

Proof.

- Observe that $\min\{a, b, c, d\} \leq \min\{a, b, c\}$ is always true for any numbers a, b, c, and d
- In general, $\min S_1 \leq \min S_2$ if S_2 is a subset of S_1 where S_1 and S_2 are two sets of numbers
- We will prove (i) for

$$\text{full model } y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$
$$\text{reduced model } y_i = \beta_0 + \beta_1 x_{i1} + \beta_3 x_{i3} + \varepsilon_i$$

The proofs for other nested models are similar.

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

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

General Framework for Testing Nested Models

$\boxed{\text{H}_0\text{: reduced model is true}}$ v.s. $\boxed{\text{H}_1\text{ : full model is true}}$

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

$$\text{SSE}_{reduced} \geq \text{SSE}_{full}$$

- Simplicity or Accuracy?
    - The full model fits the data better (with a 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 sacrifice too much in accuracy in exchange for simplicity. The full model is preferred.

- Hence, a larger difference $\text{SSE}_{reduced} - \text{SSE}_{full}$ is stronger evidence against the reduced model
- How large $\text{SSE}_{reduced} - \text{SSE}_{full}$ is considered large?
  - It depends on the difference in the **complexity** of the two models, which can be reflected by the **difference in the number of parameters** of the two models,

$$dfE_{reduced} - dfE_{full}$$

  - The larger the magnitude of the noise, σ^2, the larger $\text{SSE}_{reduced} - \text{SSE}_{full}$ is even if H_0 is true
- Hence a reasonable test statistic is

$$\frac{(\text{SSE}_{reduced} - \text{SSE}_{full})/(\text{dfE}_{reduced} - \text{dfE}_{full})}{\sigma^2}$$

- Need to estimate the unknown σ^2 with the MSE.
- Should estimate σ^2 using MSE_{full} rather than $\text{MSE}_{reduced}$ as the full model is always true since the reduced model is a special case of the full model

The F-Statistic

$$F = \frac{(\text{SSE}_{reduced} - \text{SSE}_{full})/(\text{dfE}_{reduced} - \text{dfE}_{full})}{\text{MSE}_{full}}$$

- $\text{dfE}_{reduced}$ is the df for SSE for the reduced model.
- dfE_{full} is the df for SSE for the full model.
- $F \geq 0$ since $\text{SSE}_{reduced} \geq \text{SSE}_{full}$
- The smaller the F-statistic, the more the reduced model is favored
- Under H_0, the F-statistic has an F-distribution with $\text{dfE}_{reduced}$-dfE_{full} and dfE_{full} degrees of freedom.

Testing All Coefficients Equal Zero

Testing the hypotheses

$$\text{H}_0\text{: } \beta_1 = \cdots = \beta_p = 0 \text{ v.s. } \text{H}_a\text{: not all } \beta_1 \ldots, \beta_p = 0$$

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

$$\text{Full :} y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

$$\text{Reduced :} y_i = \beta_0 + \varepsilon_i \quad \text{(all predictors are unnecessary)}$$

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

$$\text{SSE}_{reduced} = \sum_{i=1}^{n} (y_i - \widehat{\beta}_0)^2 = \sum_i (y_i - \overline{y})^2 = \text{SST}_{full}$$

Testing All Coefficients Equal Zero

Testing the hypotheses

$$\text{H}_0\text{: } \beta_1 = \cdots = \beta_p = 0 \text{ v.s. } \text{H}_a\text{: not all } \beta_1 \ldots, \beta_p = 0$$

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

$$\text{Full : } y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

$$\text{Reduced : } y_i = \beta_0 + \varepsilon_i \quad \text{(all predictors are unnecessary)}$$

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

$$\text{SSE}_{reduced} = \sum_{i=1}^{n}(y_i - \widehat{\beta}_0)^2 = \sum_i (y_i - \overline{y})^2 = \text{SST}_{full}$$

- $\text{dfE}_{full} = n - p - 1$.

Testing All Coefficients Equal Zero

Testing the hypotheses

$$\text{H}_0\text{: } \beta_1 = \cdots = \beta_p = 0 \text{ v.s. } \text{H}_a\text{: not all } \beta_1 \ldots, \beta_p = 0$$

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

$$\text{Full :} y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

$$\text{Reduced :} y_i = \beta_0 + \varepsilon_i \quad \text{(all predictors are unnecessary)}$$

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

$$\text{SSE}_{reduced} = \sum_{i=1}^{n} (y_i - \widehat{\beta}_0)^2 = \sum_i (y_i - \bar{y})^2 = \text{SST}_{full}$$

- $\text{dfE}_{full} = n - p - 1$.
- $\text{dfE}_{reduced} = n - 1$ since the reduced model has 0 predictors.

Testing All Coefficients Equal Zero

Hence

$$
\begin{aligned}
F &= \frac{(\text{SSE}_{reduced} - \text{SSE}_{full})/(\text{dfE}_{reduced} - \text{dfE}_{full})}{\text{MSE}_{full}} \\
&= \frac{(\text{SST}_{full} - \text{SSE}_{full})/[n-1-(n-p-1)]}{\text{SSE}_{full}/(n-p-1)} \\
&= \frac{\text{SSR}_{full}/p}{\text{SSE}_{full}/(n-p-1)} = \frac{\text{MSR}_{full}}{\text{MSE}_{full}}.
\end{aligned}
$$

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.

```
> lmtrees = lm(log(Volume) ~ log(Diameter) + log(Height), data = trees)
> summary(lmtrees)
... (part of output omitted)...
Residual standard error: 0.08139 on 28 degrees of freedom
Multiple R-squared:  0.9777,    Adjusted R-squared:  0.9761
F-statistic: 613.2 on 2 and 28 DF,  p-value: < 2.2e-16
```

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= $\frac{\text{SSR}}{\text{dfR}}$	$F = \frac{\text{MSR}}{\text{MSE}}$
Error	$dfE = n - p - 1$	SSE	MSE= $\frac{\text{SSE}}{\text{dfE}}$	
Total	$dfT = n - 1$	SST		

ANOVA is the shorthand for **an**alysis **o**f **va**riance.
It decomposes the total variation in the response (SST) into separate pieces that correspond to different sources of variation, like SST = SSR + SSE in the regression setting.

Example Tree Data

```
lmfull = lm(log(Volume) ~ log(Diameter) + log(Height), data=trees)
lmreduced = lm(log(Volume) ~ 1, data=trees)
anova(lmreduced, lmfull)
Analysis of Variance Table

Model 1: log(Volume) ~ 1
Model 2: log(Volume) ~ log(Diameter) + log(Height)
  Res.Df  RSS Df Sum of Sq   F Pr(>F)
1     30 8.31
2     28 0.19  2      8.12 613 <2e-16
```

Testing Some Coefficients Equal to Zero

Ex. Testing H_0: $\beta_2 = \beta_3 = 0$ under the model
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$.

- full model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$
- reduced model: $Y = \beta_0 + \beta_1 X_1 + \varepsilon$

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

Testing Some Coefficients Equal to Non-Zero Values

In the model for the `trees` data,

$$\log(\text{Volume}) = \beta_0 + \beta_1 \log(\text{Diameter}) + \beta_2 \log(\text{Height}) + \varepsilon$$

recall we think that $\beta_1 = 2$ and $\beta_2 = 1$.

We can test both coefficients in one test. Under H_0: $\beta_1 = 2$ and $\beta_2 = 1$, the full model becomes the reduced model below

$$\log(\text{Volume}) = \beta_0 + \mathit{2} \log(\text{Diameter}) + \mathit{1} \log(\text{Height}) + \varepsilon$$

- Note in the reduced model, the coefficients of $\log(\text{Diameter})$ and $\log(\text{Height})$ are both *known*

```
lmreduced = lm(log(Volume) ~ 1, offset=2*log(Diameter)+log(Height),
               data=trees)
```

Testing Some Coefficients Equal to Non-Zero Values

In the model for the `trees` data,

$$\log(\text{Volume}) = \beta_0 + \beta_1 \log(\text{Diameter}) + \beta_2 \log(\text{Height}) + \varepsilon$$

recall we think that $\beta_1 = 2$ and $\beta_2 = 1$.

We can test both coefficients in one test. Under H_0: $\beta_1 = 2$ and $\beta_2 = 1$, the full model becomes the reduced model below

$$\log(\text{Volume}) = \beta_0 + 2 \log(\text{Diameter}) + 1 \log(\text{Height}) + \varepsilon$$

- Note in the reduced model, the coefficients of $\log(\text{Diameter})$ and $\log(\text{Height})$ are both *known*
- Terms with known coefficients in an MLR model are called *offsets*. One can add an `offset` term in an `lm()` model like

```
lmreduced = lm(log(Volume) ~ 1, offset=2*log(Diameter)+log(Height),
               data=trees)
```

One can then test H_0: $\beta_1 = 2$ and $\beta_2 = 1$ simultaneously in one test as follows

```
lmfull = lm(log(Volume) ~ log(Diameter) + log(Height), data=trees)
lmreduced = lm(log(Volume) ~ 1, offset=2*log(Diameter)+log(Height),
               data=trees)
anova(lmreduced, lmfull)
Analysis of Variance Table

Model 1: log(Volume) ~ 1
Model 2: log(Volume) ~ log(Diameter) + log(Height)
  Res.Df   RSS Df Sum of Sq    F Pr(>F)
1     30 0.188
2     28 0.185  2   0.00222 0.17   0.85
```

Testing Equality of Coefficients

Ex1. Testing H_0: $\beta_1 = \beta_2 = \beta_3$ under the model

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$, the reduced model is

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

- Make a new variable $W = X_1 + X_2 + X_3$
- Fit the reduced model by regressing Y on W and X_4
- Find $\text{SSE}_{reduced}$ and $\text{dfE}_{reduced} - \text{dfE}_{full} =$ ___

```
lmfull = lm(Y ~ X1 + X2 + X3 + X4)

W = X1 + X2 + X3
lmreduced = lm(Y ~ W + X4)
# or simply
lmreduced = lm(Y ~ I(X1 + X2 + X3) + X4)

anova(lmreduced, lmfull)
```

Testing Equality of Coefficients

Ex1. Testing H_0: $\beta_1 = \beta_2 = \beta_3$ under the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$, the reduced model is

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

- Make a new variable $W = X_1 + X_2 + X_3$
- Fit the reduced model by regressing Y on W and X_4
- Find SSE$_{reduced}$ and dfE$_{reduced}$ − dfE$_{full}$ = 2

```r
lmfull = lm(Y ~ X1 + X2 + X3 + X4)

W = X1 + X2 + X3
lmreduced = lm(Y ~ W + X4)
# or simply
lmreduced = lm(Y ~ I(X1 + X2 + X3) + X4)

anova(lmreduced, lmfull)
```

Testing Equality of Coefficients (2)

Ex2. Testing H_0: $\beta_1 = \beta_2$ and $\beta_3 = \beta_4$ under the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$, the reduced model is

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_1 X_2 + \beta_3 X_3 + \beta_3 X_4 + \varepsilon \\ &= \beta_0 + \beta_1 (X_1 + X_2) + \beta_3 (X_3 + X_4) + \varepsilon \end{aligned}$$

- Make new variables $W_1 = X_1 + X_2$, $W_2 = X_3 + X_4$
- Fit the reduced model by regressing Y on W_1 and W_2
- Find $\text{SSE}_{reduced}$ and $\text{dfE}_{reduced} - \text{dfE}_{full} =$ ___
- In R

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

Testing Equality of Coefficients (2)

Ex2. Testing H_0: $\beta_1 = \beta_2$ and $\beta_3 = \beta_4$ under the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$, the reduced model is

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_1 X_2 + \beta_3 X_3 + \beta_3 X_4 + \varepsilon \\ &= \beta_0 + \beta_1 (X_1 + X_2) + \beta_3 (X_3 + X_4) + \varepsilon \end{aligned}$$

- Make new variables $W_1 = X_1 + X_2$, $W_2 = X_3 + X_4$
- Fit the reduced model by regressing Y on W_1 and W_2
- Find SSE$_{reduced}$ and dfE$_{reduced}$ − dfE$_{full}$ = __2__
- In R

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

Testing Coefficients under Constraints (1)

Ex1 say the full model is

$$\text{Full model : } Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$$

If H_0: $\beta_2 = \beta_3 + \beta_4$, then the reduced model is

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + (\beta_3 + \beta_4) X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon \\ &= \beta_0 + \beta_1 X_1 + \beta_3 (X_2 + X_3) + \beta_4 (X_2 + X_4) + \varepsilon \end{aligned}$$

- Make new variables $W_1 = X_2 + X_3$, $W_2 = X_2 + X_4$
- Fit the reduced model by regressing Y on X_1, W_1 and W_2
- Find $\text{SSE}_{reduced}$ and $\text{dfE}_{reduced} - \text{dfE}_{full}$ = ___

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

Testing Coefficients under Constraints (1)

Ex1 say the full model is

$$\text{Full model}: \; Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$$

If H_0: $\beta_2 = \beta_3 + \beta_4$, then the reduced model is

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + (\beta_3 + \beta_4) X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon \\ &= \beta_0 + \beta_1 X_1 + \beta_3 (X_2 + X_3) + \beta_4 (X_2 + X_4) + \varepsilon \end{aligned}$$

- Make new variables $W_1 = X_2 + X_3$, $W_2 = X_2 + X_4$
- Fit the reduced model by regressing Y on X_1, W_1 and W_2
- Find $\text{SSE}_{reduced}$ and $\text{dfE}_{reduced} - \text{dfE}_{full}$ = 1

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

Testing Coefficients under Constraints (2)

Ex2: If we suspect $\beta_2 = 2\beta_1$, then the reduced model is

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon \\
&= \beta_0 + \beta_1 X_1 + 2\beta_1 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon \\
&= \beta_0 + \beta_1 (X_1 + 2X_2) + \beta_3 X_3 + \beta_4 X_4 + \varepsilon
\end{aligned}
$$

- Make a new variable $W = X_1 + 2X_2$
- Fit the reduced model by regressing Y on W, X_3 and X_4
- Find $\text{SSE}_{reduced}$ and $\text{dfE}_{reduced} - \text{dfE}_{full} =$ ___
- Can be done in R as follows

```
lmfull = lm(Y ~ X1 + X2 + X3 + X4)
lmreduced = lm(Y ~ I(X1 + 2*X2) + X3 + X4)
anova(lmreduced, lmfull)
```

Testing Coefficients under Constraints (2)

Ex2: If we suspect $\beta_2 = 2\beta_1$, then the reduced model is

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon \\ &= \beta_0 + \beta_1 X_1 + 2\beta_1 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon \\ &= \beta_0 + \beta_1 (X_1 + 2X_2) + \beta_3 X_3 + \beta_4 X_4 + \varepsilon \end{aligned}$$

- Make a new variable $W = X_1 + 2X_2$
- Fit the reduced model by regressing Y on W, X_3 and X_4
- Find SSE$_{reduced}$ and dfE$_{reduced}$ − dfE$_{full}$ = 1
- Can be done in R as follows

```
lmfull = lm(Y ~ X1 + X2 + X3 + X4)
lmreduced = lm(Y ~ I(X1 + 2*X2) + X3 + X4)
anova(lmreduced, lmfull)
```

Example (Tree Data)

In the model for the `trees` data,

$$\log(\text{Volume}) = \beta_0 + \beta_1 \log(\text{Diameter}) + \beta_2 \log(\text{Height}) + \varepsilon$$

to test whether H_0: $\beta_1 = 2\beta_2$ is true, we can conduct the test below

```
lmfull = lm(log(Volume) ~ log(Diameter) + log(Height), data=trees)
lmreduced = lm(log(Volume) ~ I(2*log(Diameter) + log(Height)), data=tre
anova(lmreduced, lmfull)
Analysis of Variance Table

Model 1: log(Volume) ~ I(2 * log(Diameter) + log(Height))
Model 2: log(Volume) ~ log(Diameter) + log(Height)
  Res.Df   RSS Df Sum of Sq    F Pr(>F)
1     29 0.188
2     28 0.185  1   0.00204 0.31   0.58
```

Testing Coefficients under Constraints (3)

Ex3: To test H_0: $\beta_1 + \beta_2 = 1$ against H_1: $\beta_1 + \beta_2 \neq 1$ for the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, then say, $\beta_1 + \beta_2 = 1 + \delta$.

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \\
Y &= \beta_0 + \beta_1 X_1 + (1 - \beta_1 + \delta) X_2 + \varepsilon \\
= &= \beta_0 + X_2 + \beta_1 (X_1 - X_2) + \delta X_2 + \varepsilon
\end{aligned}
$$

- Testing whether $\beta_1 + \beta_2 = 1$ is equivalent to testing $\delta = 0$.
- Note the term $+X_2$ has a *known* coefficient $+1$ and hence is an *offset*

```
lmfull = lm(Y ~ X1 + X2)
lmreduced = lm(Y ~ I(X1 -X2), offset = X2)
anova(lmreduced, lmfull)
```

Testing Coefficients under Constraints (3)

Ex3: To test H$_0$: $\beta_1 + \beta_2 = 1$ against H$_1$: $\beta_1 + \beta_2 \neq 1$ for the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, then say, $\beta_1 + \beta_2 = 1 + \delta$.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

$$Y = \beta_0 + \beta_1 X_1 + (1 - \beta_1 + \delta) X_2 + \varepsilon$$

$$= = \beta_0 + X_2 + \beta_1 (X_1 - X_2) + \delta X_2 + \varepsilon$$

- Testing whether $\beta_1 + \beta_2 = 1$ is equivalent to testing $\delta = 0$.
- Note the term $+X_2$ has a *known* coefficient $+1$ and hence is an *offset*

```
lmfull = lm(Y ~ X1 + X2)
lmreduced = lm(Y ~ I(X1 -X2), offset = X2)
anova(lmreduced, lmfull)
```

F-Test on a Single β_j is Equivalent to t-Test

F-Test on a Single β_j is Equivalent to t-Test

Say one wants to test a single $\beta_3 = 0$ in the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

- one can do a t-test by reading the t-statistic and P-value for X_3 from the output for `summary(lm(Y ~ X1 + X2 + X3))`
- alternatively, one can conduct an F-test comparing the models

$$\text{Full model}: Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$
$$\text{Reduced model}: Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

```
anova(lm(Y ~ X1 + X2 + X3), lm(Y ~ X1 + X2))
```

One can show that the F-statistic = (t-statistic)^2 and the P-values are the same, and thus the two tests are equivalent.

The proof involves complicate matrix algebra and is hence omitted.

E.g., for the `trees` data, one might test the β_j for `log(Height)` using an F-test,

```
lm1 = lm(log(Volume) ~ log(Diameter) + log(Height), data = trees)
lmreduced = lm(log(Volume) ~ log(Diameter), data = trees)
anova(lmreduced,lm1)
Analysis of Variance Table

Model 1: log(Volume) ~ log(Diameter)
Model 2: log(Volume) ~ log(Diameter) + log(Height)
  Res.Df   RSS Df Sum of Sq    F    Pr(>F)
1     29 0.383
2     28 0.185  1     0.198 29.9 0.0000078
```

```
summary(lm1)$coef
                 Estimate Std. Error t value  Pr(>|t|)
(Intercept)        -6.632    0.79979  -8.292 5.057e-09
log(Diameter)       1.983    0.07501  26.432 2.423e-21
log(Height)         1.117    0.20444   5.464 7.805e-06
```

Observe

- (t-statistics)^2 $= (5.4644)^2 \approx 29.86 = F$-statistic.
- The P-values are both 0.0000078.

The slight difference is due to rounding.