2.2.2 Model Checking

Using the General Linear Model approach to regression, we can fit models with different numbers of predictors, and

- assess whether any individual covariate is influential in the model (look at $\widehat{\beta}, s_{\widehat{\beta}}$ and t-statistics
- assess whether there is any explanatory power in the variables combined (look at ANOVA statistics)

For the multiple regression model, the ANOVA table takes the form

SOURCE	DF	SS	MS	F
REGRESSION	k	SSR	MSR	$F = \dfrac{MSR}{MSE}$
ERROR	$n-k-1$	SSE	MSE	
TOTAL	$n-1$	SS		

where

$$MSR = \frac{SSR}{k} \qquad MSE = \frac{SSE}{n-k-1}$$

the F statistic is

$$F = \frac{MSR}{MSE}$$

and if H_0 is true

$$F \sim \text{Fisher-F}(k, n-k-1)$$

Here

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ H_a &: \text{At least one } \beta_j \neq 0 \end{aligned}$$

The model for H_0 has one parameter β_0.
The model for H_a has $k+1$ parameters

$$\beta_0, \beta_1, \beta_2, \ldots, \beta_k$$

Therefore the number of extra parameters for model H_a is

$$(k+1) - 1 = k$$

i.e. to obtain model H_0 from model H_a we constrain k parameters to be zero.

Because we can constrain model H_a by setting some parameters equal to zero to obtain model H_0, we say that

Model H_0 is **nested** inside Model H_a

The number, k, of constraints $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ explains why the ANOVA table Regression degrees of freedom is k

- the multiple regression brings in k extra parameters.

In addition, we can use the R^2 or Adjusted R^2 statistic to check overall model adequacy

$$R^2 = 1 - \frac{SSE}{SS_{yy}} = \frac{SS_{yy} - SSE}{SS_{yy}} = \frac{SSR}{SS}$$

which is equal to

$$\frac{\text{VARIATION EXPLAINED BY THE REGRESSION}}{\text{TOTAL VARIATION}}$$

Also

$$\text{Adj. } R^2 = 1 - \frac{SSE/(n-k-1)}{SS/(n-1)}$$

$R^2 > 0.7$ implies that the model is a good fit, that is, most of the variation observed is explained by the regression model.

We can now fit completely general models in the form of the General Linear Model; if y is the response, and $x_1, \ldots, x_k$ are the covariates or factor predictors, we can include combinations of

- Polynomial Main Effects : $x_j, x_j^2, x_j^3, \ldots$
- Two-way Interactions: $x_{j_1} . x_{j_2}$
- Three-way Interactions: $x_{j_1} . x_{j_2} . x_{j_3}$

etc.

In SPSS, we can use the

General Linear Model $\longrightarrow$ *Univariate*

pulldown menus to build and fit the model.

- To fit factor predictors, we used the *Fixed Factor* option
- To build models, we use the

  Model $\longrightarrow$ *Custom*

  selections on the *Univariate* dialog

SEE SCREENS ON THE COURSE WEBSITE

Dummy Variables

Note: We can fit the factor predictor using the Linear Regression pulldown if we create **dummy variables**.

For example, if factor predictor X has L levels, we create L **new** binary predictors $X_1, \ldots, X_L$, where, for $l = 1, \ldots, L$

$$X_l = \begin{cases} 1 & \text{whenever } X = l \\ 0 & \text{otherwise} \end{cases}$$

We can then include $X_1, \ldots, X_L$ in the regression model.

Example: $L = 4$.

X	X_1	X_2	X_3	X_4
3	0	0	1	0
1	1	0	0	0
3	0	0	1	0
4	0	0	0	1
2	0	1	0	0
2	0	1	0	0

See McClave and Sincich 10, Section 12.7.

2.2.3 Stepwise Model Selection

We seek a method that allows us to compare nested models.

Suppose we want to compare

$$\begin{aligned} \text{MODEL 1} \quad &: \quad y = \beta_0 + \beta_1 x + \beta_2 x^2 \\ \text{MODEL 2} \quad &: \quad y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 \end{aligned}$$

Model 1 is nested inside Model 2 as if we set $\beta_3 = 0$ in Model 2, we get Model 1.

If

$$\begin{aligned} \text{MODEL 1} &: \quad y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \\ \text{MODEL 2} &: \quad y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12}(x_1 . x_2) \end{aligned}$$

we can set $\beta_{12} = 0$ in Model 2 to obtain Model 1, so again the models are nested.

We can set up a hypothesis test to assess whether the simplification of Model 2 to Model 1 (by setting one or more parameters equal to zero) is justified by the data.

ANOVA tests for Comparing Nested Models

Terminology

- *Complete Model*

$$E[Y] = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

- *Reduced Model*

$$E[Y] = \beta_0 + \beta_1 x_1 + \cdots + \beta_g x_g$$

where $g < k$. The reduced model is obtained from the complete model by setting

$$\beta_{g+1} = \beta_{g+2} = \cdots = \beta_k = 0$$

The **reduced** model is **nested** inside the **complete** model.

We wish to test the hypothesis

$$\begin{aligned} H_0 &: \beta_{g+1} = \beta_{g+2} = \cdots = \beta_k = 0 \\ H_a &: \text{At least one of these} \beta_j \neq 0 \end{aligned}$$

We can test this hypothesis by fitting both models, and combining the results; we focus on the sums of squares quantities.