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 → *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 → *Custom*

  selections on the *Univariate* dialog

Dummy Variables

Recall that 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

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

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.

Method

1. Fit the **complete model** and obtain the sum of squared errors, SSE_C, available from the ANOVA table.
2. Fit the **reduced model** and obtain the sum of squared errors, SSE_R, available from the ANOVA table.
3. Form the test statistic

$$F = \frac{(SSE_R - SSE_C)/(k-g)}{SSE_C/(n-k-1)}$$

If H_0 is **true**, then $F \sim$ Fisher-F$(k-g, n-k-1)$

Note: $k-g$ is the number of parameters we set equal to zero when moving from complete to reduced model.

Using this F statistic, we can assess whether there is evidence to support the reduced model over the complete model.

Complete Model ANOVA table:

SOURCE	DF	SS	MS	F
COMPLETE MODEL	k	SSR_C	MSR_C	F_C
ERROR$_C$	$n-k-1$	SSE_C	MSE_C	
TOTAL	$n-1$	SS		

Reduced Model ANOVA table:

SOURCE	DF	SS	MS	F
REDUCED MODEL	g	SSR_R	MSR_R	F_R
ERROR$_R$	$n-g-1$	SSE_R	MSE_R	
TOTAL	$n-1$	SS		

The result holds for comparing any two nested models where the standard model assumptions hold:

- ϵ uncorrelated
- ϵ independent of $x_1, \ldots, x_k$
- ϵ has constant variance
- $\epsilon \sim N(0, \sigma^2)$

Note: It does not allow us to compare non-nested models; for example

$$\begin{aligned} \text{MODEL 1} &: \quad y = \beta_0 + \beta_1 x_1 + \epsilon \\ \text{MODEL 2} &: \quad y = \beta_0 + \beta_2 x_2 + \epsilon \end{aligned}$$

- **NOT NESTED !**

$$F = \frac{(SSE_R - SSE_C)/(k-g)}{SSE_C/(n-k-1)} = \frac{\textcircled{1}/\textcircled{2}}{\textcircled{3}/\textcircled{4}}$$

① - $SSE_R - SSE_C$: this is the improvement in fit when the reduced model is extended to the complete model

② - $k - g$: this is the number of extra parameters needed to extend the reduced model to the complete model

③ - SSE_C
④ - $n - k - 1$

③/④ - this is the best guess we have at the true value of σ^2, that is, the estimate of σ^2 constructed using as much information as possible, once the effects of

$$x_1, \ldots, x_k$$

have been accounted for.

Example (Hooker's Data)

We consider the two models:

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

Here

- MODEL 1: Reduced Model
- MODEL 2: Complete Model

$k = 2$, $g = 1$.

IS THE QUADRATIC TERM NEEDED ?

Example (Hooker's Data)

COMPLETE MODEL	SSR_C	2286.933
	SSE_C	4.382
REDUCED MODEL	SSR_R	2272.474
	SSE_R	18.840

with $n = 31, k = 2, g = 1$

$$\Longrightarrow k - g = 1, n - k - 1 = 28$$

So

$$F = \frac{(SSE_R - SSE_C)/(k-g)}{SSE_C/(n-k-1)} = \frac{(18.840 - 4.382)/1}{4.382/28} = 92.383$$

Example (Hooker's Data)

We compare F with the

$$\text{Fisher-F}(k-g, n-k-1) \equiv \text{Fisher-F}(1, 28)$$

distribution.

$$F_{0.05}(1, 28) = 4.20$$

Thus

$$92.383 = F > F_{0.05}(1, 28) = 4.20$$

and $H_0 : E[Y] = \beta_0 + \beta_1 x$ is **REJECTED** in favour of $H_a : E[Y] = \beta_0 + \beta_1 x + \beta_2 x^2$.

i.e. the **quadratic model** fits better than the straight-line model.

NOTE: From the original ANOVA tables, we already know that Model 1 and Model 2 both fit better than the null model

$$\begin{array}{lrcl} \text{MODEL 0} \qquad & E[Y] & = & \beta_0 \\ & y & = & \beta_0 + \epsilon \end{array}$$

where there is no dependence on x.

We have now confirmed that Model 2 fits better than Model 1.

Example (Diabetes Data)

Factor Predictor: **group** (X_2)
Continuous Covariate: **loggluf** (X_1)
Response: **logglut** (Y)

We have five models to confirm:

$$
\begin{aligned}
\text{MODEL 0} &: 1 \\
\text{MODEL 1} &: X_2 \\
\text{MODEL 2} &: X_1 \\
\text{MODEL 3} &: X_1 + X_2 \\
\text{MODEL 4} &: X_1 + X_2 + X_1.X_2
\end{aligned}
$$

Example (Diabetes Data)

MODEL 4 us the most complex model with 6 parameters

$$\beta_{10}, \beta_{11}, \beta_{20}, \beta_{21}, \beta_{30}, \beta_{31}$$

MODEL 4:

$$E[Y] = \begin{cases} \beta_{10} + \beta_{11} x_1 & \text{GROUP 1} \\ \beta_{20} + \beta_{21} x_1 & \text{GROUP 2} \\ \beta_{30} + \beta_{31} x_1 & \text{GROUP 3} \end{cases}$$

All of the other models are nested inside Model 4; we can obtain them all by setting parameters equal to zero.

Example (Diabetes Data)

In the SPSS parameterization:

β_{30}, β_{31}	Group 3 Intercept and Slope
$\beta_{10} = \beta_{30} + \delta_{10}$	Changes in the Intercepts in
$\beta_{20} = \beta_{30} + \delta_{20}$	Groups 1 and 2 are δ_{10} and δ_{20}
$\beta_{11} = \beta_{31} + \delta_{11}$	Changes in the Slopes in
$\beta_{21} = \beta_{31} + \delta_{21}$	Groups 1 and 2 are δ_{11} and δ_{21}

Thus the six new parameters are

$$\beta_{30}, \beta_{31}, \delta_{10}, \delta_{20}, \delta_{11}, \delta_{21}$$

MODEL 0 $\beta_{31} = 0$
$\delta_{10} = \delta_{20} = \delta_{11} = \delta_{21} = 0$

MODEL 1 $\beta_{31} = \delta_{11} = \delta_{21} = 0$

MODEL 2 $\delta_{10} = \delta_{20} = \delta_{11} = \delta_{21} = 0$

MODEL 3 $\delta_{11} = \delta_{21} = 0$

Note: $\beta_{31} = 0 \Longrightarrow \delta_{11} = \delta_{21} = 0$, as X_1 is not included in the model.

Counting Parameters

- Whenever we remove a **continuous covariate**, from a model, we set **one** parameter to zero.
- Whenever we remove a **factor predictor** with L levels from a model, we set $L - 1$ parameters to zero.
- Whenever we remove a two-way interaction between these variables from a model, we set $1.(L - 1) = L - 1$ parameters to zero.

Models 0,1,2,3 are nested inside Model 4.

Two approaches to finding the best model are used:

1. Start with Model 0 and try to add terms that improve the model fit (**Forward Selection**)
2. Start with Model 4 and try to remove terms that improve the model fit (**Backward Selection**)

Note:

- Models 0,1 and 2 are nested inside Model 3.
- Model 0 is nested inside Models 1 and 2.

Therefore we can begin with Model 4, or Model 3 or Model 1 or 2, and simplify to a nested model.

Example (Diabetes Data)

Here $n = 144$. From SPSS output handouts:

Model	Description	SSE	p
0	1	28.504	1
1	X_2	4.160	3
2	X_1	3.738	2
3	$X_1 + X_2$	1.472	4
4	$X_1 + X_2 + X_1.X_2$	1.318	6

p is the number of non-zero parameters; k or g is always $p - 1$ in the following analysis.

Backward Selection:

Complete Model : Model 4
Reduced Model : Model 3

Here $k = 5, g = 3$ so $k - g = 2$, and

$$n - k - 1 = 144 - 5 - 1 = 138.$$

We have

$$F = \frac{(SSE_R - SSE_C)/(k - g)}{SSE_C/(n - k - 1)} = \frac{(1.472 - 1.318)/2}{1.318/138} = 8.062$$

We compare this with the

$$\text{Fisher-F}(k - g, n - k - 1) = \text{Fisher-F}(2, 138)$$

distribution; we have $F_\alpha(2, 138) = 3.061$, so we

$$\text{Reject } H_0 \text{ at } \alpha = 0.05$$

i.e. Model 4

$$X_1 + X_2 + X_1.X_2$$

fits **significantly better** than Model 3

$$X_1 + X_2.$$

- we cannot simplify the complete model to the reduced model without the loss of significant explanatory power.

The Interaction is Necessary in the Model

Backward selection stops here; we cannot simplify further from the complete model.

Forward Selection: we start with Model 0 and build up.

Model 1 vs Model 0 $\quad F = 412.568$

Model 2 vs Model 0 $\quad F = 940.846$

It seems that Model 2 is the better improvement, so we try the selection path

$$\text{Model } 0 \longrightarrow \text{Model } 2 \longrightarrow \text{Model } 3 \longrightarrow \text{Model } 4$$

Model	SSE	$SSE_R - SSE_C$
0	28.504	-
2	3.738	24.766
3	1.472	2.266
4	1.318	0.154

ie we work down the table, $28.504 - 3.738 = 24.766$ etc.

Comparison	k	g	SSE_C	$SSE_R - SSE_C$	F
2 vs 0	1	0	3.738	24.766	940.82
3 vs 2	3	1	1.472	2.266	107.76
4 vs 3	5	3	1.318	0.154	8.06

Recall that $n = 144$, and

$$F = \frac{(SSE_R - SSE_C)/(k-g)}{SSE_C/(n-k-1)}$$

Under each H_0,

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

- $F_{0.05}(1, 142) \doteqdot 3.92 < 940.82$
  Therefore Model 0 is **NOT** an adequate simplification of Model 2
- $F_{0.05}(2, 140) \doteqdot 3.07 < 107.76$
  Therefore Model 2 is **NOT** an adequate simplification of Model 3
- $F_{0.05}(2, 138) \doteqdot 3.07 < 8.06$
  Therefore Model 3 is **NOT** an adequate simplification of Model 4

All of the null hypotheses are **rejected**.

Therefore by both forward and backward selection, we select Model 4

$$X_1 + X_2 + X_1.X_2$$

as the most appropriate model.

Note: In this sequence of hypothesis tests, the convention is **not** to correct for multiple testing (we don't know how many tests we are going to do), although a correction could be used.

F-tests for Unbalanced Designs

Example (Potato Damage Data)

The damage to potato plants caused by cold temperatures is to be studied.

In this experimental study, three binary factor predictors were used: we label them A, B and C rather than X_1, X_2, X_3 to recall the link with Factorial Designs in ANOVA. Each factor takes two levels:

	Factor	Levels
A	Potato Variety	0- Variety 1, 1- Variety 2
B	Acclimatization Routine	0- Room Temp, 1- Cold Room
C	Preparation Treatment	0- -4C, 1- -8C

Thus we have a $2 \times 2 \times 2$ three-way factorial design.

F-tests for Unbalanced Designs

However, the design is **unbalanced**; we have different numbers of replicates in each of the 8 factor-level combinations.

This means we cannot use conventional 3-way ANOVA; the lack of balance means that the stated *p*-values **may be misleading if we perform a standard ANOVA**.

Thus we are forced to use the General Linear Model F-test approach.

We begin with the most complex model and do backward selection.

Here the most complex model can be written

$$A + B + C + A.B + A.C + B.C + A.B.C$$

that is,

- all main effects (terms 1,2 and 3)
- all two-way interactions (terms 4,5 and 6)
- all three-way interactions (term 7)

We may write this model

$$A * B * C$$

which is termed the full factorial model.

Counting the numbers of parameters

Term	Parameters	
A	$(a-1)$	1
B	$(b-1)$	1
C	$(c-1)$	1
$A.B$	$(a-1)(b-1)$	1
$A.C$	$(a-1)(c-1)$	1
$B.C$	$(b-1)(c-1)$	1
$A.B.C$	$(a-1)(b-1)(c-1)$	1
Total		7

where $a = b = c = 2$.

We have 7 parameters in total (excluding the baseline mean) when all terms are considered, so

$$k = 7$$

In the following tables columns are:

Complete Model
Reduced Model
SSE_C
SSE_R
k
g
F (test statistic)
$F_{0.05}(k-g, n-k-1)$

We denote the critical value by F_α and check whether $F > F_\alpha$.

Potato Damage Data: ANOVA-F Tests

We compare four models: M_{R_1}, M_{R_2} and M_{R_3} are nested within the complete model M_C.

$$\begin{aligned}
M_C &: \quad A + B + C + A.B + A.C + B.C + A.B.C \\
M_{R_1} &: \quad A + B + C + A.B \\
M_{R_2} &: \quad A + B + C \\
M_{R_3} &: \quad A + B + A.B
\end{aligned}$$

COMP.	RED.	SSE_C	SSE_R	k	g	F	F_α
M_C	M_{R_1}	4968.876	5093.746	7	4	0.561	2.76
M_{R_1}	M_{R_2}	5093.746	7183.674	4	3	28.721	3.92
M_{R_1}	M_{R_3}	5093.746	6319.640	4	3	16.846	3.92

Note: The quoted F_α values are approximate as the textbook does not tabulate all Fisher-F distributions. We take $\alpha = 0.05$

Conclusions

Taking the comparisons in order:

1. M_C *vs* M_{R_1} : $F < F_\alpha$. Therefore the result is **not significant**: Model M_{R_1} **is an adequate simplification** of Model M_C, and we choose M_{R_1} over M_C.

   The model M_{R_1} now becomes the complete model.

2. M_{R_1} *vs* M_{R_2} : $F > F_\alpha$. Therefore the result **is significant**: Model M_{R_2} **is not an adequate simplification** of Model M_{R_1}

3. M_{R_1} *vs* M_{R_3} : $F > F_\alpha$. Therefore the result **is significant**: Model M_{R_3} **is not an adequate simplification** of Model M_{R_1}

Thus the final model is

$$A + B + C + A.B$$

i.e. all main effects, plus the interaction between potato variety and acclimatization routine.

We cannot simplify this model further without significant loss in terms of goodness of fit.

Note: $R^2 = 0.631$ and Adjusted $R^2 = 0.610$, so we have a reasonable fit.

Task Distraction Data

Example (Task Distraction Data)

In an experimental study, the number of errors made in performing a specified task was recorded. The experiment investigated the influence of various predictors on the numbers of errors made.

There are two factor predictors (A, B) and one continuous covariate (X).

We have a balanced design with 15 people (replicates) in each factor-level subgroup.

Example (Task Distraction Data)

A	Group	1 : Non-smoker 2 : Delayed smoker 3 : Active smoker
B	Task	1 : Pattern Recognition 2 : Cognitive Task 3 : Driving Simulation
X	Distraction Level	

We compare four models with the **complete** model

Complete Model : $A * B * X$

$$A + B + X + A.B + A.X + B.X + A.B.X$$

Number of parameters

Term	Parameters		Tot.
A	$(a-1)$	$= 3-1$	2
B	$(b-1)$	$= 3-1$	2
X	(1)		1
$A.B$	$(a-1)(b-1)$	$= 2 \times 2$	4
$A.X$	$(a-1)(1)$	$= 2 \times 1$	2
$B.X$	$(b-1)(1)$	$= 2 \times 1$	2
$A.B.X$	$(a-1)(b-1)(c-1)$	$= 2 \times 2 \times 1$	4
Total			17

For illustration we consider the following sequence of models:

- Reduced Model 1: M_{R_1}

$$A + B + X + A.X + B.X$$

- Reduced Model 2: M_{R_2}

$$A + B + X + B.X$$

- Reduced Model 3: M_{R_3}

$$B + X + B.X$$

- Reduced Model 4: M_{R_4}

$$B + X$$

Task Distraction Data: ANOVA-F Tests

$$\begin{aligned}
M_C &: A + B + X + A.B + A.X + B.X + A.B.X \\
M_{R_1} &: A + B + X + A.X + B.X \\
M_{R_2} &: A + B + X + B.X \\
M_{R_3} &: B + X + B.X \\
M_{R_4} &: B + X
\end{aligned}$$

COMP.	RED.	SSE_C	SSE_R	k	g	F	F_α
M_C	M_{R_1}	5660.010	7627.479	17	9	5.084	2.02
M_{R_1}	M_{R_2}	7627.479	7971.274	9	7	2.817	3.07
M_{R_2}	M_{R_3}	7971.274	8404.654	7	5	3.452	3.07
M_{R_3}	M_{R_4}	8404.654	11154.715	5	3	21.105	3.07

Conclusions

Taking the comparisons in order:

1. M_C *vs* M_{R_1} : $F > F_\alpha$. Therefore the result is **significant**: Model M_{R_1} **is not an adequate simplification** of Model M_C

2. M_{R_1} *vs* M_{R_2} : $F < F_\alpha$. Therefore the result **is not significant**: Model M_{R_2} **is an adequate simplification** of Model M_{R_1}

3. M_{R_2} *vs* M_{R_3} : $F > F_\alpha$. Therefore the result **is significant**: Model M_{R_3} **is not an adequate simplification** of Model M_{R_2}

4. M_{R_3} *vs* M_{R_4} : $F > F_\alpha$. Therefore the result **is significant**: Model M_{R_4} **is not an adequate simplification** of Model M_{R_3}

Follow-up Analysis

In a follow up analysis (see Handout), it transpires that the model

$$A + B + X + A.B + A.X + B.X$$

ie selected.

Note: $R^2 = 0.863$ and Adjusted $R^2 = 0.831$, so we have a good fit.

Note: we must take great care with the sequence of models.

Stepwise Selection in SPSS: Options

It is possible to carry out stepwise selection in SPSS using the *Linear Regression* pulldown menu, and the *Method* pulldown list.

- **Enter :** All variables in a *block* are entered in a single step.
- **Stepwise :** At each step, the independent variable not in the equation that has the **smallest** p-value in the F-test is entered, if that probability is sufficiently small. Variables already in the regression equation are **removed** if their p-value becomes sufficiently large. The method terminates when no more variables are eligible for inclusion or removal.
- **Remove :** All variables in a block are removed in a single step.

Stepwise Selection in SPSS: Options

- **Backward :** Variables are entered into the equation and then sequentially removed. The variable with the smallest *partial correlation* with the dependent variable is considered first for removal. After the first variable is considered, the variable remaining in the equation with the smallest partial correlation is considered next. The procedure stops when there are no variables in the equation that satisfy the removal criteria.
- **Forward :** Variables are sequentially entered into the model starting from the null model. The first variable considered for entry into the equation is the one with the largest positive or negative correlation with the dependent variable. This variable is entered into the equation only if it satisfies the criterion for entry. If the first variable is entered, the independent variable not in the equation that has the largest partial correlation is considered next. The procedure stops when there are no variables that meet the entry criterion.

2.2.5 Pitfalls of Regression Modelling

Five issues to bear in mind in ANOVA, Regression and General Linear Modelling.

1. Model assumptions
2. Data transformations
3. Model selection
4. Multicollinearity
5. Predicting beyond the range of the covariates

See Handout.