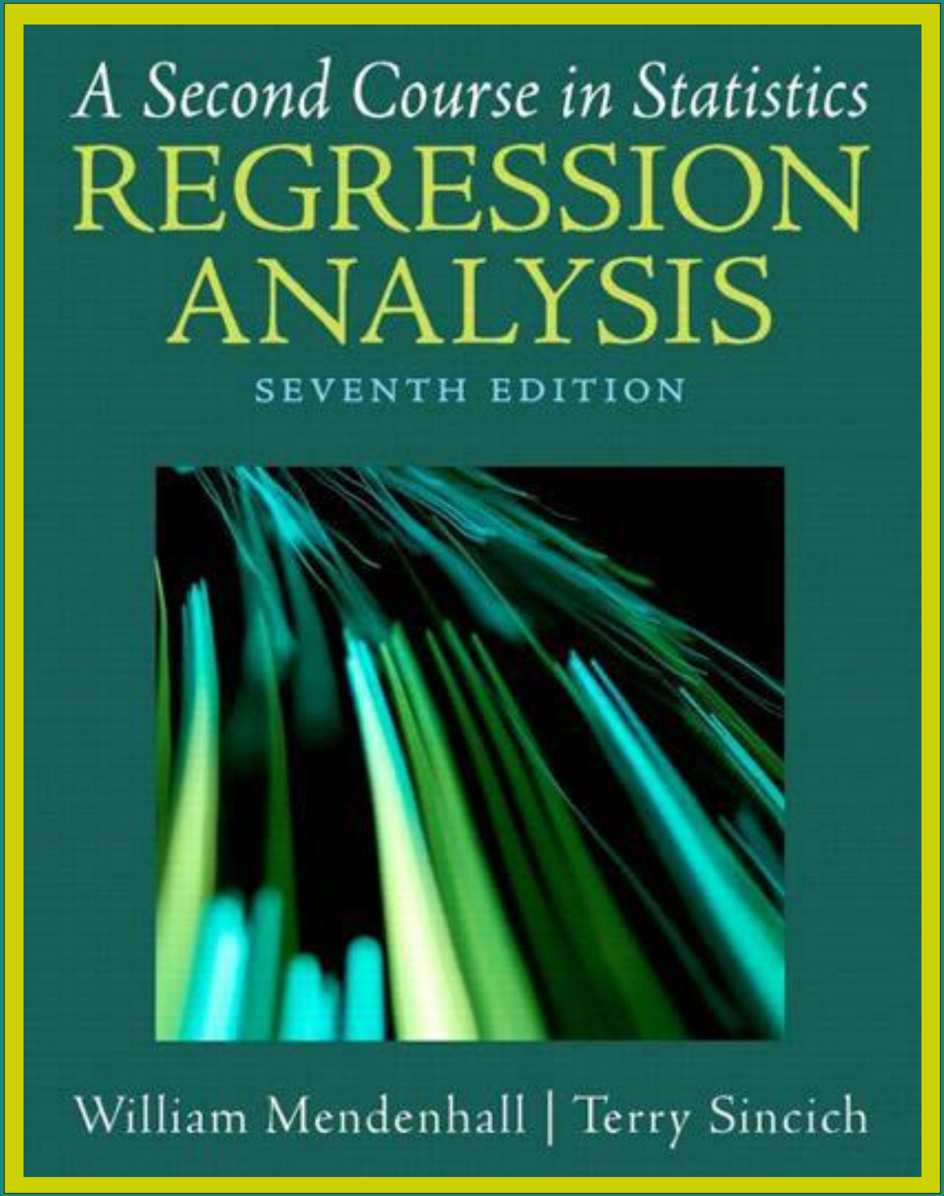
Chapter 5

Principles of Model Building



STAT210/410 Study Plan

Topic	Weeks covered	Readings	Assessment
Topic 1: Simple Linear regression (SLR)	Wk 1	Chapter 3	Online Quiz due 9th March
Topic 2: Multiple Linear Regression (MLR)	Wk2 & 3	Chapter 4	Written Assessment A2 due 23rd March
Topic 3: Model building	Wk 4	Chapter 5	
Topic 4: Variable Screening and regression pitfalls	Wk 5	Chapters 6, 7	
Topic 5: Residual Analysis	Wk 6	Chapter 8	Written Assessment A3 due 13th April
Topic 6 Generalised Linear Models (GLMs)	Wk 9 & 10	Chapter 9	
Topic 7: Principles of Experimental Design	Wk 11	Chapter 11	Written Assessment A4 due 11th May
Topic 8: ANOVA, contrasts	Wk 12 & 13	Chapter 12	
STAT410 ONLY			
ART: Nonparametric Regression		Section 9.9	Written Assessment ART due 18th May

Chapter 5 Outline

Lecture 1

- ❖ Introduction
- ❖ Models with 1 quantitative predictor
- ❖ First - order models with ≥ 2 quantitative predictors
- ❖ Second - order models with ≥ 2 quantitative predictors

Lecture 2

- ❖ Model with 1 qualitative predictor
- ❖ Model with 2 qualitative predictors
- ❖ Model with ≥ 3 qualitative predictors
- ❖ Models with both qualitative & quantitative predictors

§5.6 is *not* covered in this unit

Introduction

Data = systematic* + random component

- ❖ Model building is the key to the success of the regression analysis
- ❖ Use exploratory data plots to help suggest an appropriate model
- ❖ Hypothesize the form of the *systematic/ deterministic* portion of the probabilistic model.
- ❖ An appropriate model should provide
    - a good fit to the observed data
    - reliable estimate of the mean value of y
    - reliable predictions of future values of y for given values of the predictors

Revision: type of variables

- ❖ Quantitative – measurements (e.g. length, blood pressure) or counts (e.g. no. of plants surviving)
- ❖ Qualitative – categorical, non-numerical
  - gender (m/f);
  - eye colour (blue, green, brown, hazel)
  - Age group (<18, 18-30, 30-45, 46-65, >65)

Models with only 1 quantitative predictor

Models with 1 quantitative predictor

A *p*th-Order Polynomial with One Independent Variable

$$E(y) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_p x^p$$

where p is an integer and $\beta_0, \beta_1, \ldots, \beta_p$ are unknown parameters that must be estimated.

Systematic or deterministic component

Models with 1 quantitative predictor

p=1: First-order model

First-Order (Straight-Line) Model with One Independent Variable

$$E(y) = \beta_0 + \beta_1 x$$

Interpretation of model parameters

β_0: y-intercept; the value of $E(y)$ when $x = 0$
β_1: Slope of the line; the change in $E(y)$ for a 1-unit increase in x

p=2: Second - order model

A Second-Order (Quadratic) Model with One Independent Variable

$$E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$$

where β_0, β_1, and β_2 are unknown parameters that must be estimated.

Interpretation of model parameters

β_0: y-intercept; the value of $E(y)$ when $x = 0$
β_1: Shift parameter; changing the value of β_1 shifts the parabola to the right or left (increasing the value of β_1 causes the parabola to shift to the right)
β_2: Rate of curvature

Models with 1 quantitative predictor

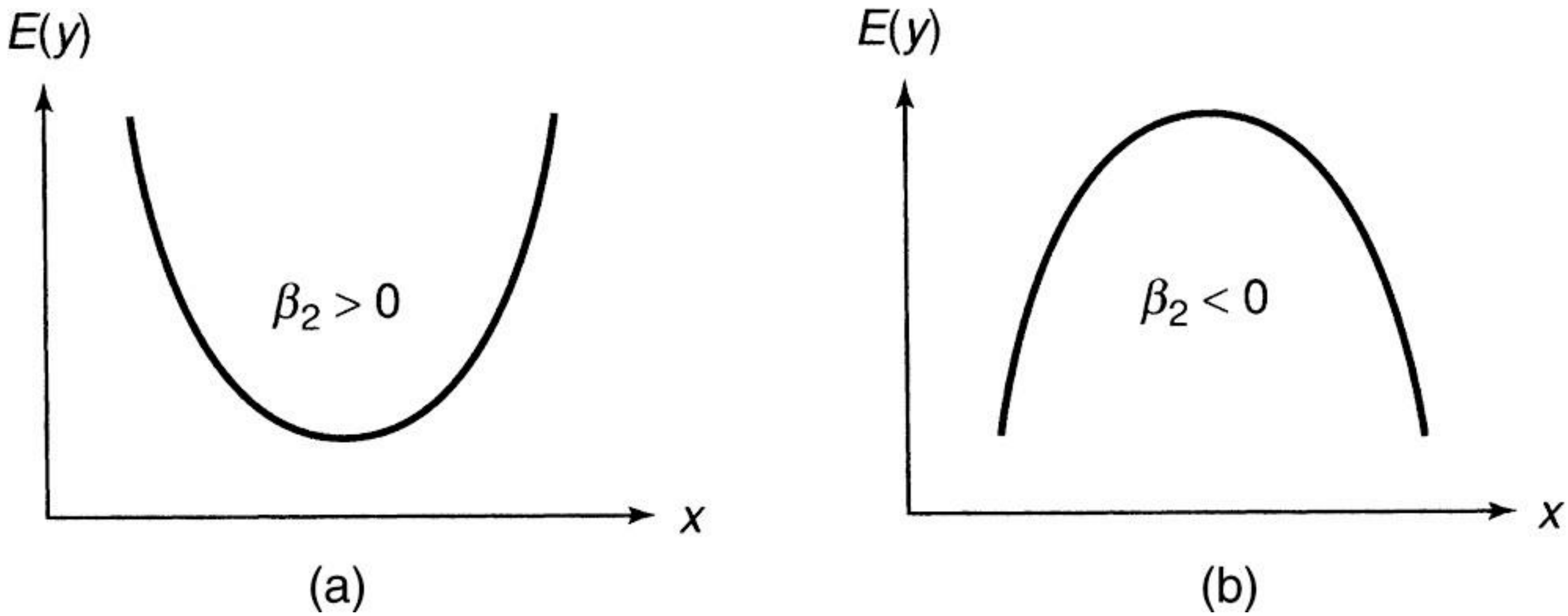


Figure 5.2 Graphs for two second-order polynomial models

Models with 1 quantitative predictor

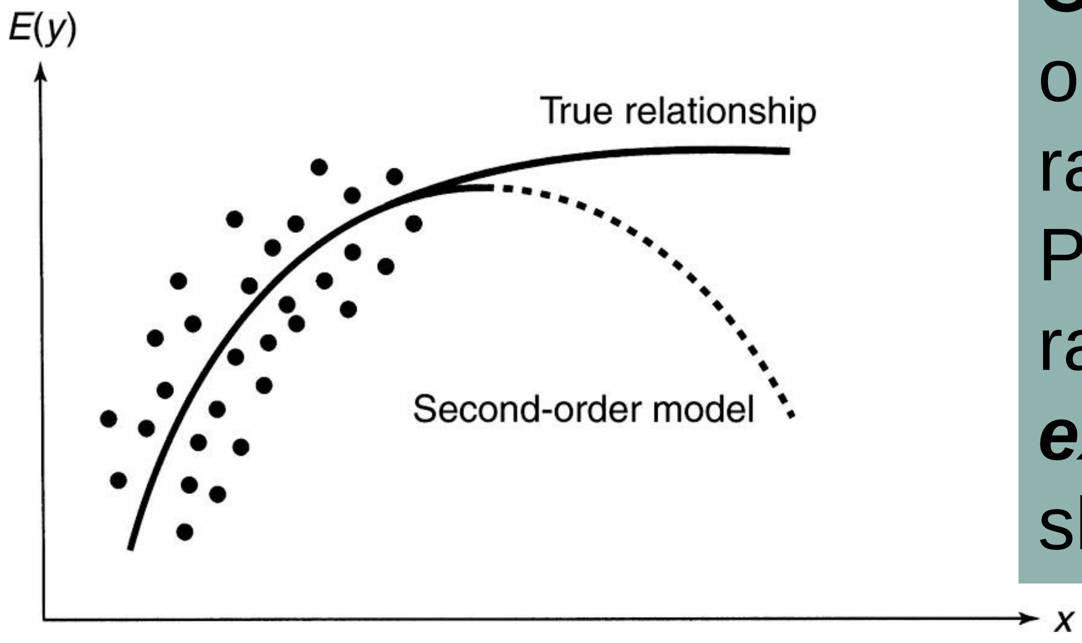


Figure 5.3 Example of the use of a quadratic model

Caution: Model is only valid for the range of observed x. Predicting outside this range is ***extrapolation*** and should be avoided.

Models with 1 quantitative predictor

p=3: Third - order model

Third-Order Model with One Independent Variable

$$E(y) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$$

Interpretation of model parameters

β_0: y-intercept; the value of $E(y)$ when $x = 0$
β_1: Shift parameter (shifts the polynomial right or left on the x-axis)
β_2: Rate of curvature
β_3: The magnitude of β_3 controls the rate of reversal of curvature for the polynomial

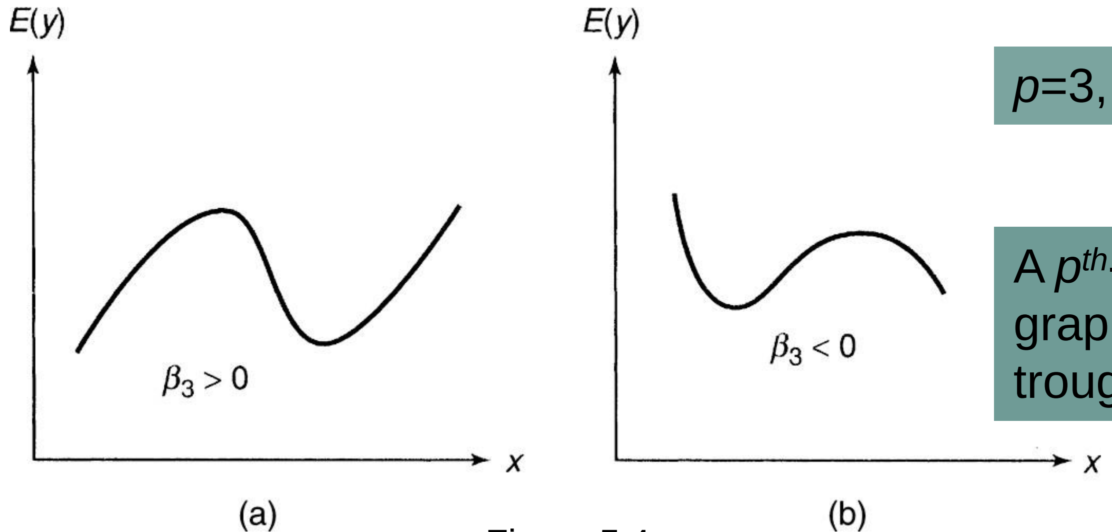


Figure 5.4

p=3, p-1 = 2 peaks/ troughs

A p^{th}-order polynomial when graphed will have (p-1) peaks, troughs, reversals in direction

Example: powerloads p.260-261

Table 5.1 Power load data

Temperature °F	Peak Load megawatts	Temperature °F	Peak Load megawatts	Temperature °F	Peak Load megawatts
94	136.0	106	178.2	76	100.9
96	131.7	67	101.6	68	96.3
95	140.7	71	92.5	92	135.1
108	189.3	100	151.9	100	143.6
67	96.5	79	106.2	85	111.4
88	116.4	97	153.2	89	116.5
89	118.5	98	150.1	74	103.9
84	113.4	87	114.7	86	105.1
90	132.0				

Model **power load** (response variable) against **daily maximum temperature** (predictor) using p^{th} – order polynomial with p = 1, 2, 3.

Example: powerloads p.260-261

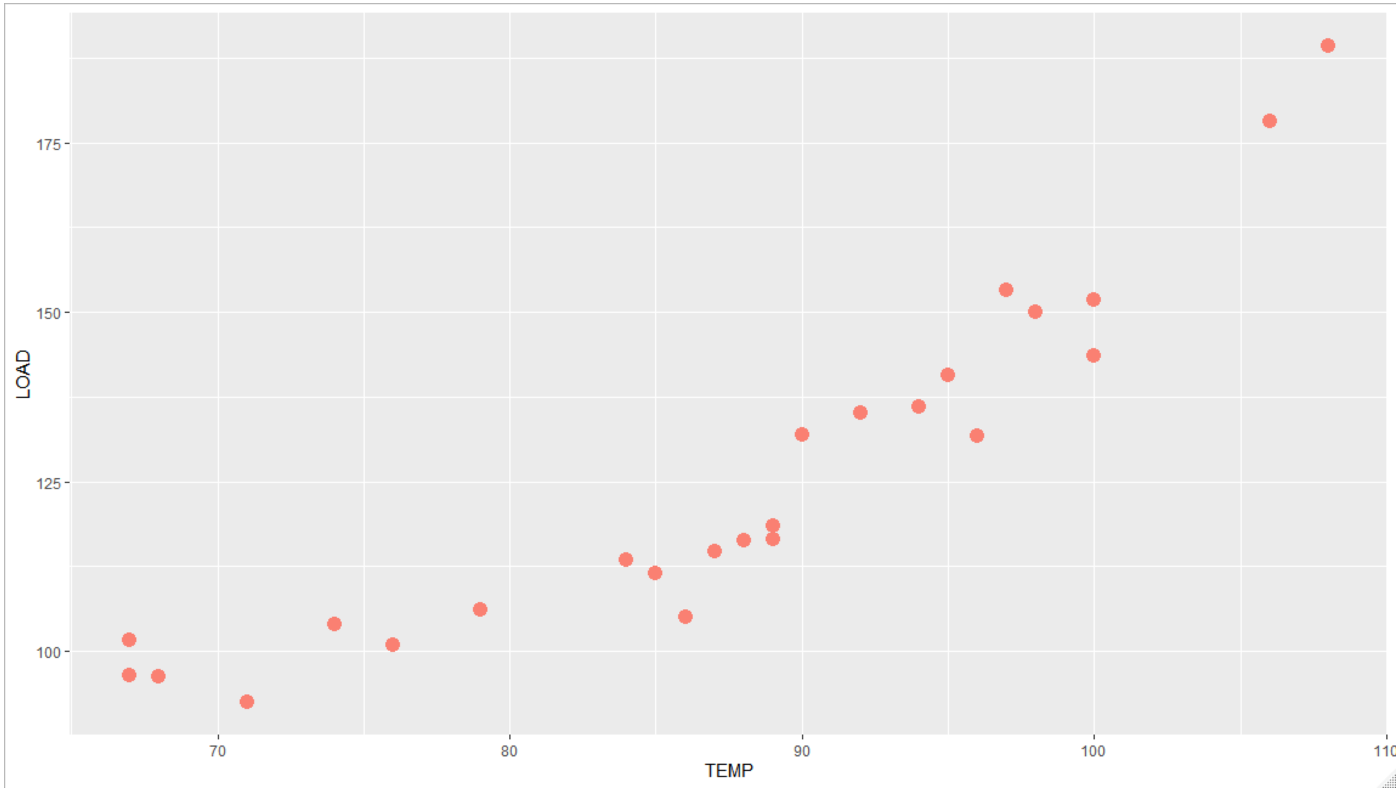


Figure 5.5 : Scatterplot for power load data

Q: Do you think a straight-line model (SLR) is appropriate?   Why and why not?

Example: powerloads p.260-261

First-order model (SLR): $y = \beta_0 + \beta_1 x + \epsilon$

```
pow.df <- read.table("POWERLOADS.txt",header=T)
mod1<-lm(LOAD~TEMP, data=pow.df)
summary(mod1)
```

Power load = - 47.4 + 1.98 * Temp

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -47.394     15.668   -3.02    0.006
TEMP           1.976      0.178   11.13  9.8e-11

Residual standard error: 10.3 on 23 degrees of freedom
Multiple R-squared:  0.843,   Adjusted R-squared:  0.837
F-statistic:  124 on 1 and 23 DF,  p-value: 9.82e-11
```

Q: What can you infer from the output?
What is the relevant hypothesis?

Example: powerloads p.260-261

First-order model (SLR): $y = \beta_0 + \beta_1 x + \epsilon$

```
pow.df <- read.table("POWERLOADS.txt",header=T)
mod1<-lm(LOAD~TEMP, data=pow.df)
summary(mod1)
```

Power load = - 47.4 + 1.98 * Temp

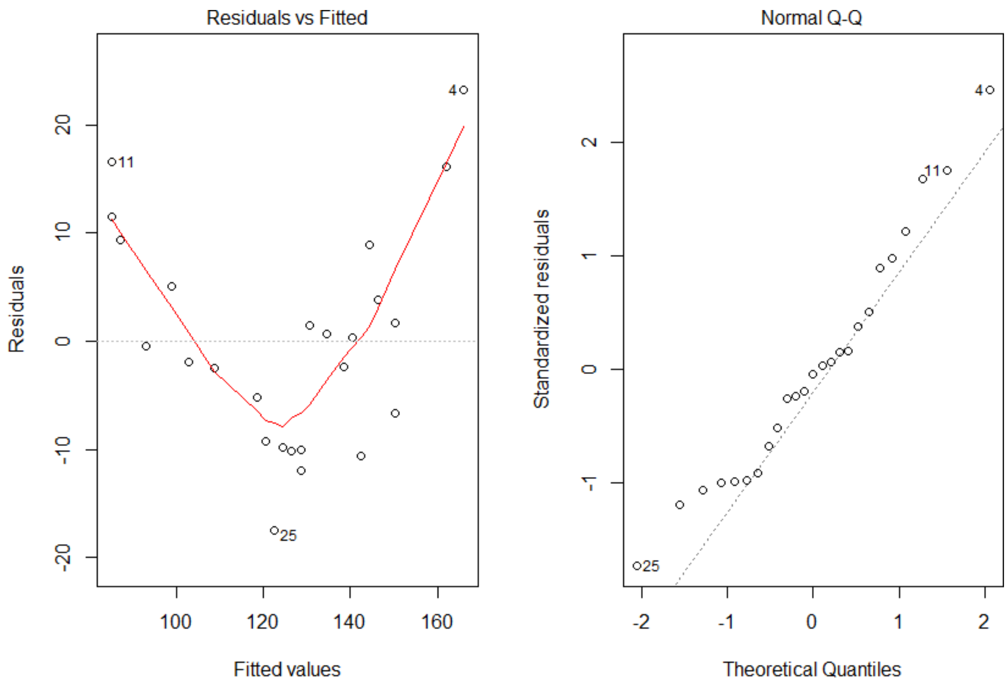
```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -47.394     15.668   -3.02    0.006
TEMP            1.976      0.178   11.13  9.8e-11

Residual standard error: 10.3 on 23 degrees of freedom
Multiple R-squared:  0.843,   Adjusted R-squared:  0.837
F-statistic:  124 on 1 and 23 DF,  p-value: 9.82e-11
```

Q: What can you infer from the output?
What is the relevant hypothesis?

Example: powerloads p.260-261

First-order model (SLR): model assumptions



Q: Interpret the residuals plots?

- Residuals vs fitted: a curved/pattern in the residuals

Example: powerloads p.260-261

Second-order model: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$

```
mod2<-lm(LOAD~TEMP + I(TEMP^2), data=pow.df)
summary(mod2)
```

Power load = 385.05 – 8.29 * Temp + 0.06*Temp^2

```
Coefficients:
             Estimate  Std. Error t value Pr(>|t|)
(Intercept) 385.04809   55.17244    6.98  5.3e-07
TEMP         -8.29253    1.29905   -6.38  2.0e-06
I(TEMP^2)     0.05982    0.00755    7.93  6.9e-08

Residual standard error: 5.38 on 22 degrees of freedom
Multiple R-squared:  0.959,   Adjusted R-squared:  0.956
F-statistic:  260 on 2 and 22 DF,  p-value: 4.99e-16
```

Q: What can you infer from the output?
State the relevant hypothesis.

Example: powerloads p.260-261

Second-order model: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$

```
mod2<-lm(LOAD~TEMP + I(TEMP^2), data=pow.df)
summary(mod2)
```

Power load = 385.05 – 8.29 * Temp + 0.06*Temp^2

```
Coefficients:
             Estimate  Std. Error t value  Pr(>|t|)
(Intercept) 385.04809   55.17244     6.98   5.3e-07
TEMP         -8.29253    1.29905    -6.38   2.0e-06
I(TEMP^2)     0.05982    0.00755     7.93   6.9e-08

Residual standard error: 5.38 on 22 degrees of freedom
Multiple R-squared:  0.959,    Adjusted R-squared:  0.956
F-statistic:  260 on 2 and 22 DF,   p-value: 4.99e-16
```

Q: What can you infer from the output?
State the relevant hypothesis.

Example: powerloads p.260-261

F-tests (ANOVA) vs t-tests

Power load = 385.05 – 8.29 * Temp + 0.06*Temp^2

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	385.04809	55.17244	6.98	5.3e-07
TEMP	-8.29253	1.29905	-6.38	**2.0e-06**
I(TEMP^2)	0.05982	0.00755	7.93	**6.9e-08**

t-test
Tests H_0: $\beta_i=0$, *given that the other predictors have been fitted*

###############

ANOVA

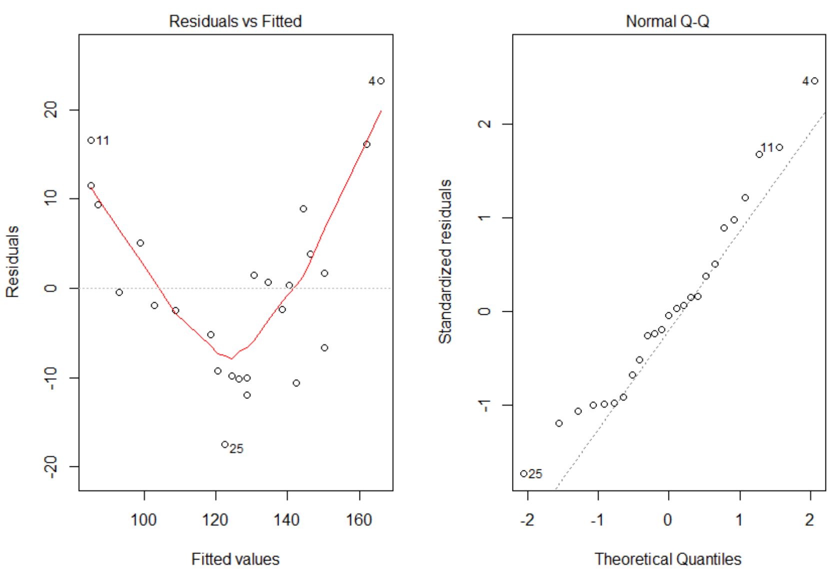
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
TEMP	1	13196	13196	456.6	**3.3e-16**
I(TEMP^2)	1	1815	1815	62.8	**6.9e-08**
Residuals	22	636	29		

F-test
Order of fit is important

Example: powerloads p.260-261

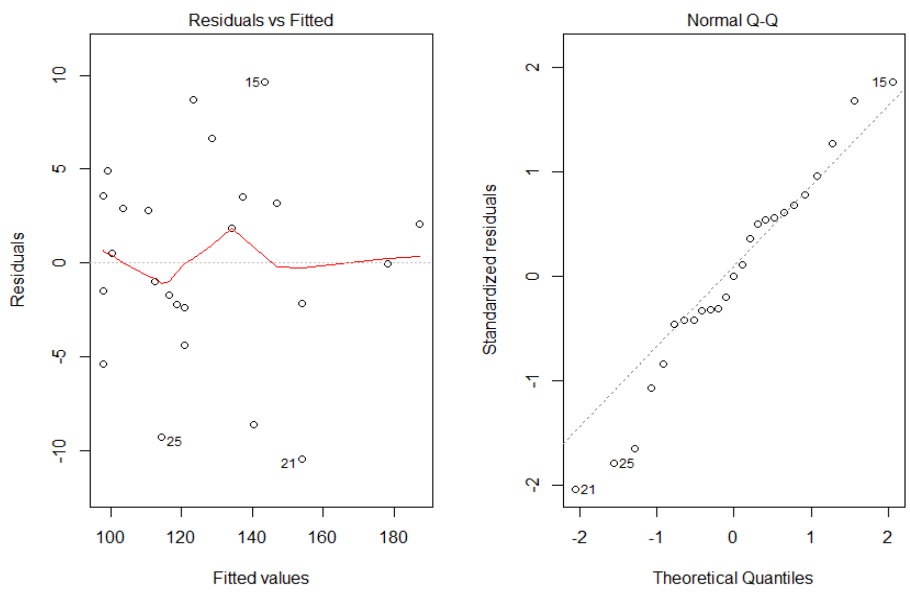
First-order model (SLR):

$$y = \beta_0 + \beta_1 x + \epsilon$$



Second-order model:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$



Q: Interpret the residuals plots?

Example: powerloads p.260-261

Third-order model: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$

```
mod3<-lm(LOAD~TEMP + I(TEMP^2) + I(TEMP^3),data=pow.df)
summary(mod3)
```

Power load = 331 – 6.39 * Temp + 0.0378*Temp^2 + 8.43*10^{-5}* Temp^3

```
Coefficients:
              Estimate  Std. Error t value  Pr(>|t|)
(Intercept)   3.31e+02   4.77e+02    0.69      0.50
TEMP         -6.39e+00   1.68e+01   -0.38      0.71
I(TEMP^2)     3.78e-02   1.95e-01    0.19      0.85
I(TEMP^3)     8.43e-05   7.43e-04    0.11      0.91

Residual standard error: 5.5 on 21 degrees of freedom
Multiple R-squared:  0.959,   Adjusted R-squared:  0.954
F-statistic:  165 on 3 and 21 DF,  p-value: 9.14e-15
```

Q: What can you infer from the output?
State the relevant hypothesis.

Example: powerloads p.260-261

Third-order model: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$

```
mod3<-lm(LOAD~TEMP + I(TEMP^2) + I(TEMP^3),data=pow.df)
summary(mod3)
```

Power load = 331 – 6.39 * Temp + 0.0378*Temp^2 + 8.43*10^{-5}* Temp^3

```
Coefficients:
              Estimate  Std. Error t value  Pr(>|t|)
(Intercept)   3.31e+02   4.77e+02     0.69      0.50
TEMP         -6.39e+00   1.68e+01    -0.38      0.71
I(TEMP^2)     3.78e-02   1.95e-01     0.19      0.85
I(TEMP^3)     8.43e-05   7.43e-04     0.11      0.91

Residual standard error: 5.5 on 21 degrees of freedom
Multiple R-squared:  0.959,   Adjusted R-squared:  0.954
F-statistic:  165 on 3 and 21 DF,  p-value: 9.14e-15
```

?

Q: What can you infer from the output?
State the relevant hypothesis.

Example: powerloads p.260-261

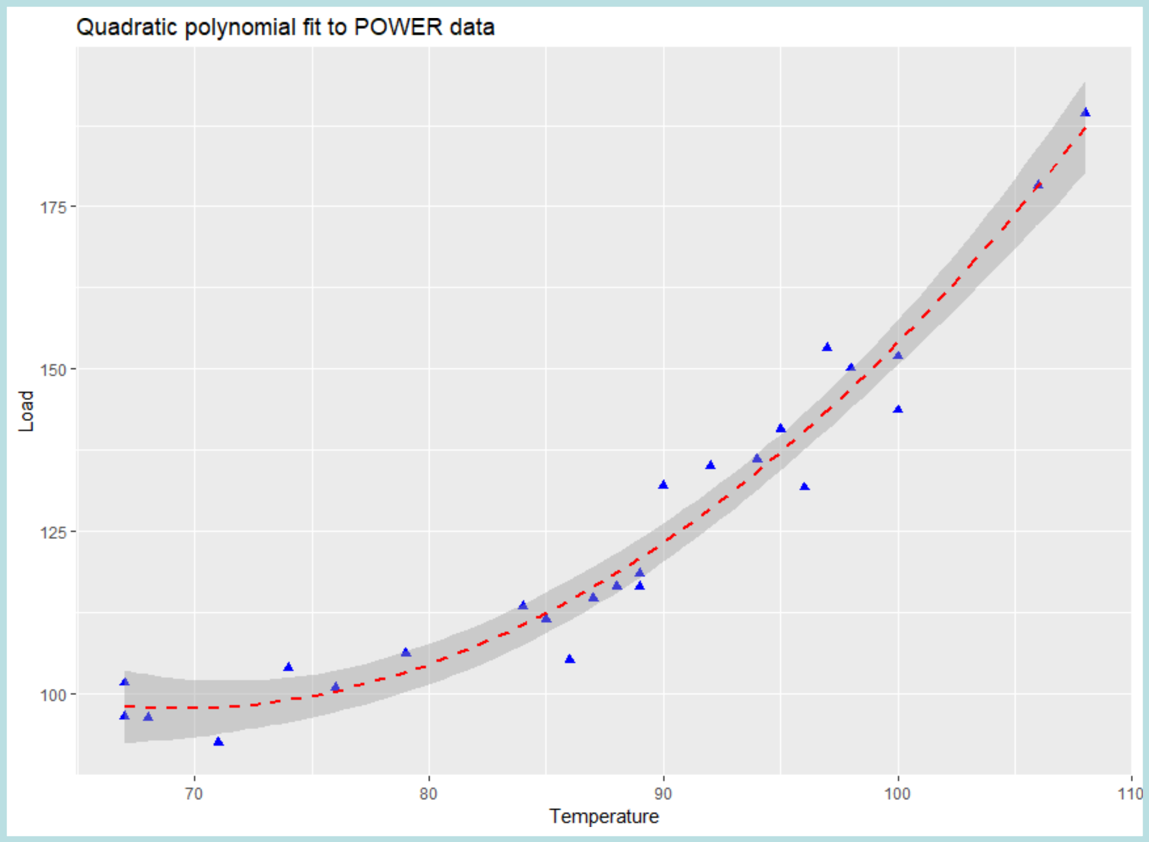
Third-order model (F- test vs t-tests)

```
mod3<-lm(LOAD~TEMP + I(TEMP^2) + I(TEMP^3), data=pow.df)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.31e+02   4.77e+02    0.69     0.50
TEMP         -6.39e+00   1.68e+01   -0.38     0.71
I(TEMP^2)     3.78e-02   1.95e-01    0.19     0.85
I(TEMP^3)     8.43e-05   7.43e-04    0.11     0.91

Response: LOAD
          Df Sum Sq Mean Sq F value  Pr(>F)
TEMP       1  13196   13196  436.08 1.6e-15
I(TEMP^2)  1   1815    1815   59.99 1.4e-07
I(TEMP^3)  1      0       0    0.01    0.91
Residuals 21    635      30
```

```
library(ggplot2)
ggplot(data=pow.df, aes(x=TEMP, y=LOAD))+
  geom_point(pch=17, color="blue", size=2)+
  geom_smooth( method="lm", formula = y ~ poly(x, 2),
  color="red", linetype=2)+
  labs(title="Quadratic polynomial fit to POWER data",
       x="Temperature", y="Load")
```



Models with more than 1 quantitative predictor

Revisit MLR: first-order model

First-Order Model in k Quantitative Independent Variables

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

where $\beta_0, \beta_1, \ldots, \beta_k$ are unknown parameters that must be estimated.

Interpretation of model parameters

β_0: y-intercept of $(k+1)$-dimensional surface; the value of $E(y)$ when $x_1 = x_2 = \cdots = x_k = 0$

β_1: Change in $E(y)$ for a 1-unit increase in x_1, when $x_2, x_3, \ldots, x_k$ are held fixed

β_2: Change in $E(y)$ for a 1-unit increase in x_2, when $x_1, x_3, \ldots, x_k$ are held fixed

$\vdots$

β_k: Change in $E(y)$ for a 1-unit increase in x_k, when $x_1, x_2, \ldots, x_{k-1}$ are held fixed

Example: EXECSAL p.220

A sample of 100 executives is selected. We are interested whether the **salary (y)** of an executive is depend on:

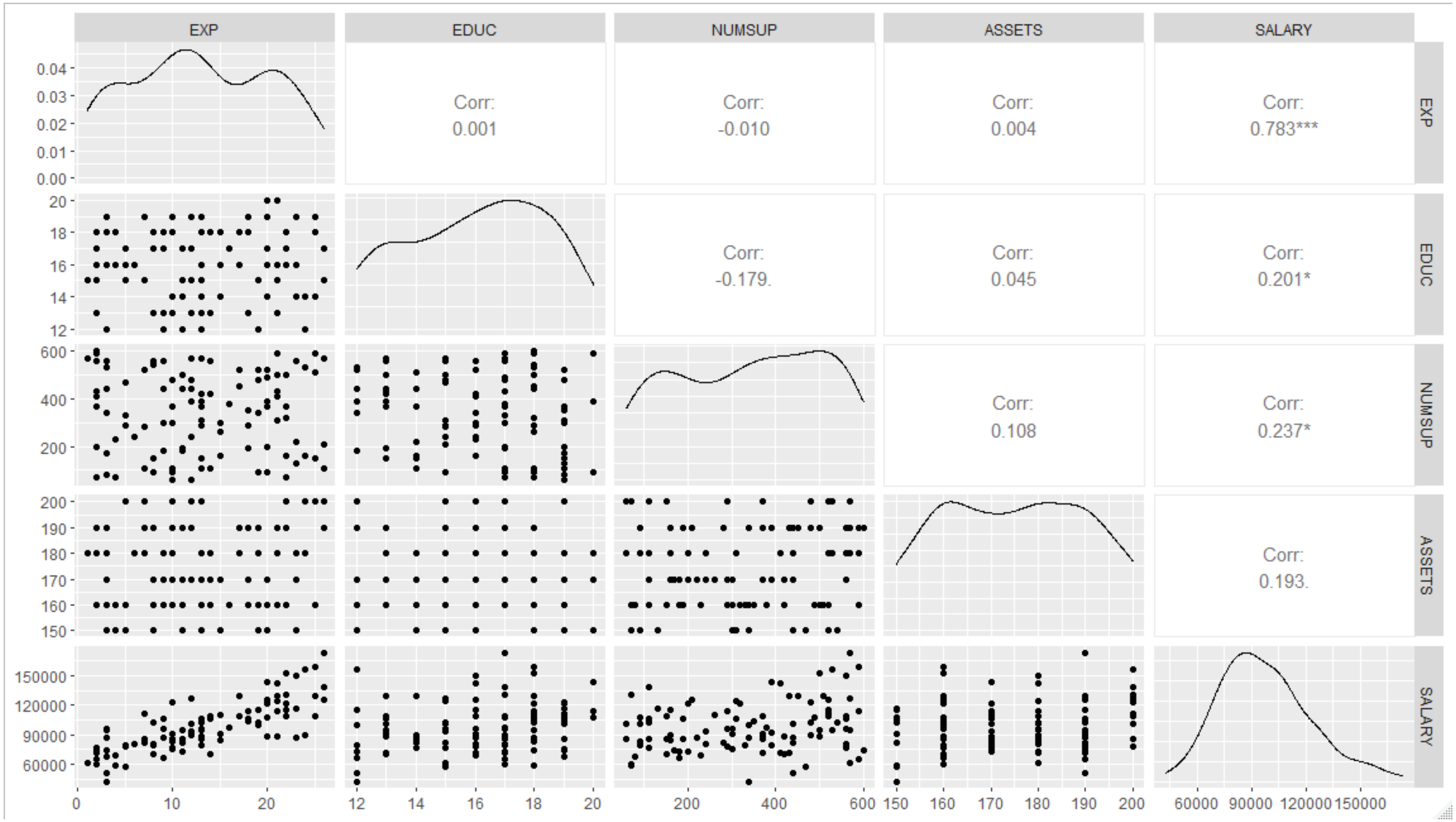
- Years of experience (EXP)
- Years of education (EDUC)
- Number of employees supervised (NUMSUP)
- Corporate assets (millions of dollars) (ASSETS)

The data is saved as *EXECSAL.txt* in the folder

Data sets and R scripts files used in lectures and workshops

Example: EXECSAL p.220

```
exec.df <- read.table("EXECSAL.txt", header=TRUE)
library(GGally)
ggpairs(exec.df[,c(2,3,5,6,1)])
```



Example: EXECSAL p.220

```
exec.df <- read.table("EXECSAL.txt", header=TRUE)
library(GGally)
ggpairs(exec.df[,c(2,3,5,6,1)])
```

Example: EXECSAL p.220

```
mod1<-lm(SALARY ~ EXP + EDUC + NUMSUP + ASSETS,
      data=exec.df)
summary(mod1)
Coefficients:
             Estimate Std. Error t value  Pr(>|t|)
(Intercept) -37082.15   17052.09   -2.17  0.0321
EXP           2696.36     173.65   15.53  < 2e-16
EDUC          2656.02     563.48    4.71  8.3e-06
NUMSUP          41.09       7.81    5.26  8.7e-07
ASSETS         244.57      83.42    2.93  0.0042

Residual standard error: 12700 on 95 degrees of freedom
Multiple R-squared:  0.757,      Adjusted R-squared:  0.747
F-statistic:   74 on 4 and 95 DF,  p-value: <2e-16
```

Example: EXECSAL p.220

```
mod1<-lm(SALARY ~ EXP + EDUC + NUMSUP + ASSETS,
      data=exec.df)
summary(mod1)
Coefficients:
             Estimate Std. Error t value  Pr(>|t|)
(Intercept) -37082.15   17052.09   -2.17  0.0321
EXP           2696.36     173.65   15.53  < 2e-16
EDUC          2656.02     563.48    4.71  8.3e-06
NUMSUP          41.09       7.81    5.26  8.7e-07
ASSETS         244.57      83.42    2.93  0.0042

Residual standard error: 12700 on 95 degrees of freedom
Multiple R-squared:  0.757,     Adjusted R-squared:  0.747
F-statistic:   74 on 4 and 95 DF,  p-value: <2e-16
```

Second-order models

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

interaction

(Chapter 4)

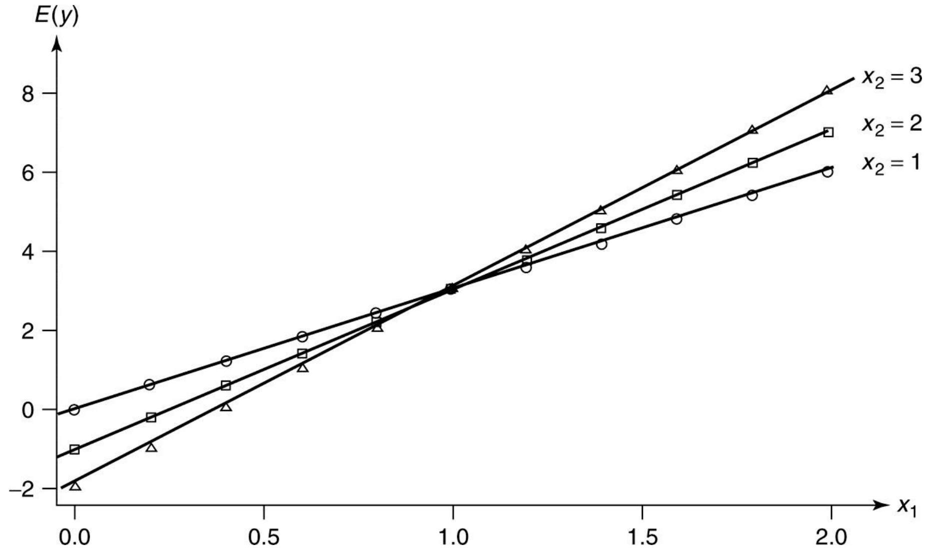


Figure 5.11 Contour lines of E(y) for x_2 = 1,2,3 (first-order model plus interaction)

$$E(y) = 1 + 2x_1 - x_2 + x_1 x_2$$

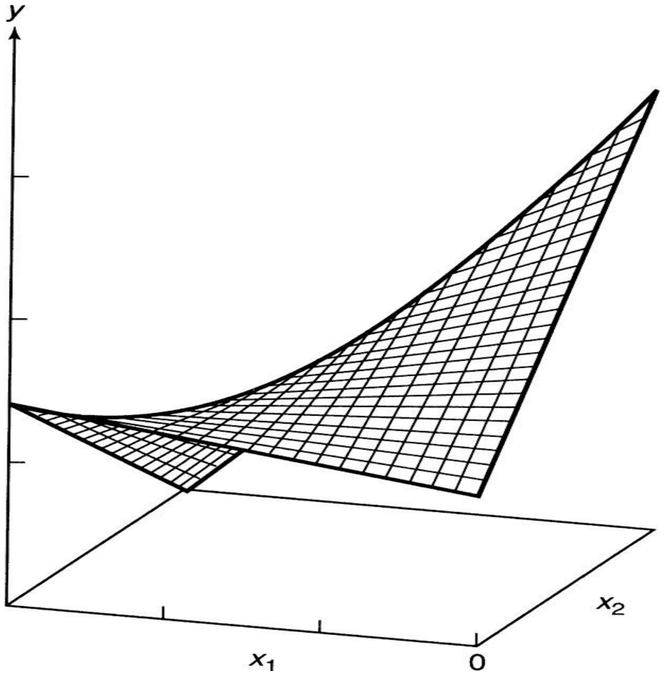


Figure 5.10 Response surface for an interaction model (second-order)

Second-order models

Interaction (Second-Order) Model with Two Independent Variables

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

Interpretation of Model Parameters

β_0: y-intercept; the value of $E(y)$ when $x_1 = x_2 = 0$

β_1 and β_2: Changing β_1 and β_2 causes the surface to shift along the x_1 and x_2 axes

β_3: Controls the rate of twist in the ruled surface (see Figure 5.10)

When one independent variable is held fixed, the model produces straight lines with the following slopes:

$\beta_1 + \beta_3 x_2$: Change in $E(y)$ for a 1-unit increase in x_1, when x_2 is held fixed

$\beta_2 + \beta_3 x_1$: Change in $E(y)$ for a 1-unit increase in x_2, when x_1 is held fixed

Second-order models

An interaction model relating E(y) to two quantitative x's

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

A complete second-order model with two quantitative x's

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2$$

A complete second-order model with three quantitative x's

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3 + \beta_7 x_1^2 + \beta_8 x_2^2 + \beta_9 x_3^2$$

Second-order models

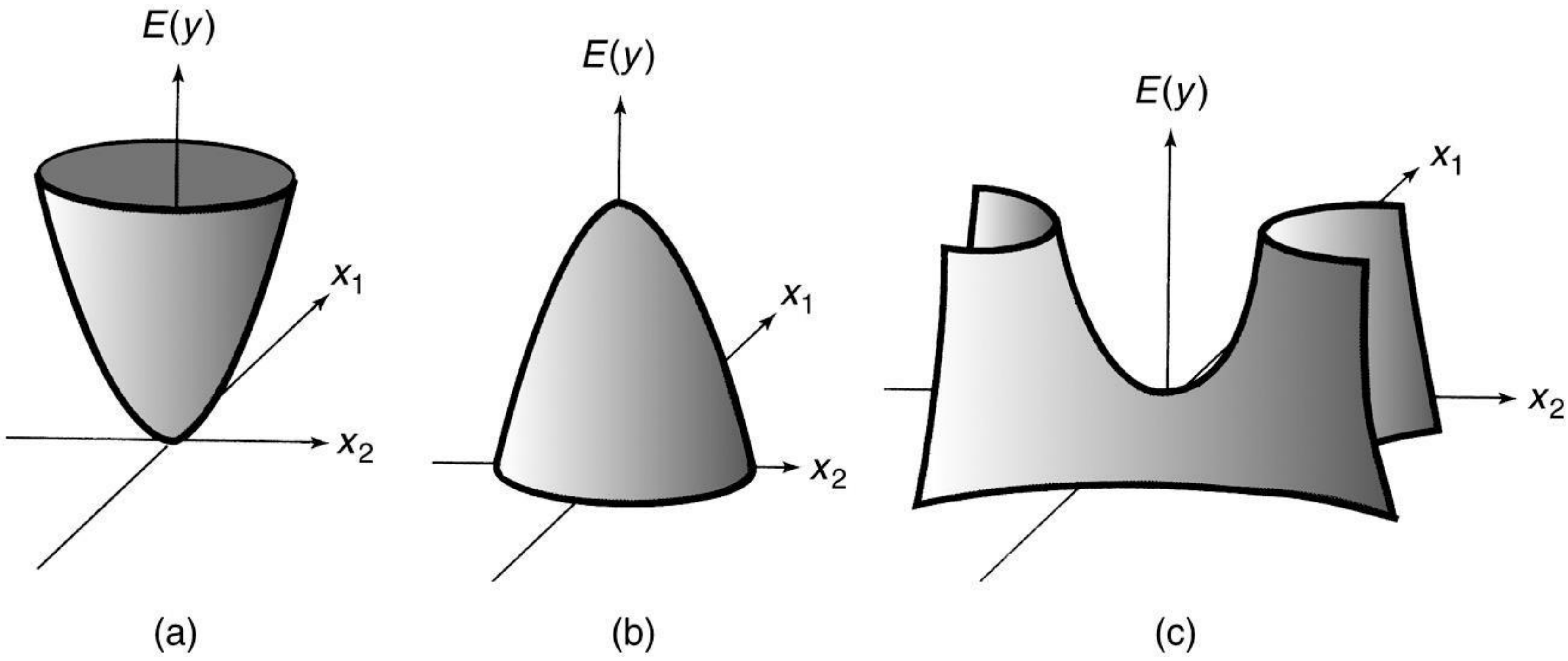
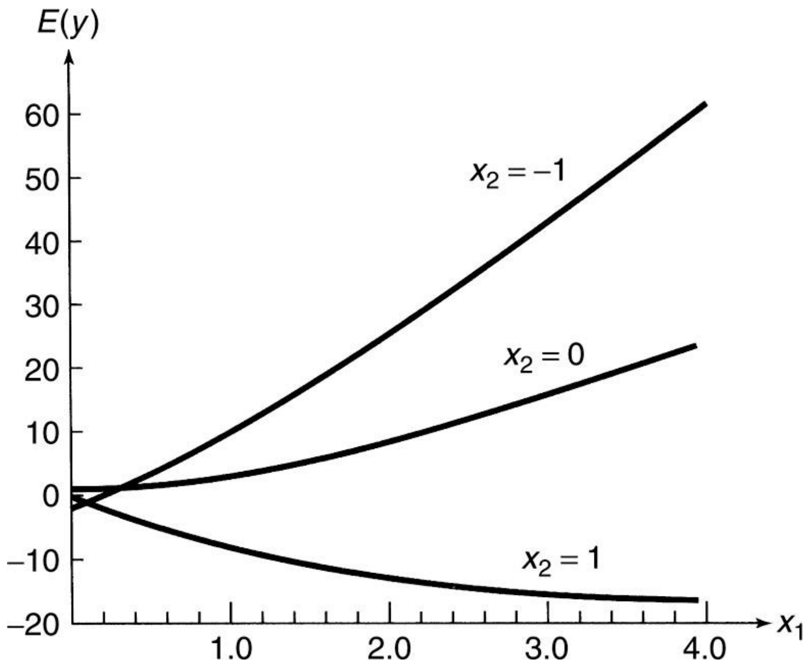


Figure 5.12 Graphs of three second-order surfaces

Second-order models



Q: For x_2 = -1, show how E(y) becomes a quadratic in x_1:
$E(y) = ax_1^2 + bx_1 + c$

Figure 5.13 Contours of *E*(*y*) for complete second-order model when x_2 = -1,0,1

$$E(y) = 1 + 2x_1 + x_2 - 10x_1x_2 + x_1^2 - 2x_2^2$$

When $x_2 = -1$ ➔ $E(y) = 1 + 2x_1 - 1 + 10x_1 + x_1^2 - 2*(1) = x_1^2 + 12x_1 - 2$

Similar $x_2 = 0$ ➔ $E(y) = x_1^2 + 2x_1 + 1$

$x_2 = 1$ ➔ $E(y) = x_1^2 - 8x_1$

Example: PROQUAL p.270

Table 5.2 Temperature, pressure, and quality of the finished product

x_1, °F	x_2, psi	y	x_1, °F	x_2, psi	y	x_1, °F	x_2, psi	y
80	50	50.8	90	50	63.4	100	50	46.6
80	50	50.7	90	50	61.6	100	50	49.1
80	50	49.4	90	50	63.4	100	50	46.4
80	55	93.7	90	55	93.8	100	55	69.8
80	55	90.9	90	55	92.1	100	55	72.5
80	55	90.9	90	55	97.4	100	55	73.2
80	60	74.5	90	60	70.9	100	60	38.7
80	60	73.0	90	60	68.8	100	60	42.5
80	60	71.2	90	60	71.3	100	60	41.4

Model the **quality (y)** of a product as a function of the **temperature (x_1)** and the **pressure (x_2)** at which it's produced.

Fit a complete second order model with 2 quantitative variables

$$\widehat{QUALITY} = \boxed{\beta_0} + \boxed{\beta_1 TEMP + \beta_2 PRESSURE} + \boxed{\beta_3 TEMP^2 + \beta_4 PRESSURE^2} + \boxed{\beta_5 TEMP * PRESSURE}$$

Intercept

Main effects

Polynomial terms

Interaction

Figure 5.14 Output for complete second-order model of quality

Remember the asterisk means the model includes the main effects and the interaction.

```
rm(list=ls())
prod.df <- read.table("PRODQUAL.txt",header=T)
attach(prod.df)
names(proqual.df) ## to see the name of the variables
mod1<-lm(QUALITY ~ TEMP*PRESSURE + I(TEMP^2) + I(PRESSURE^2))
summary(mod1)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -5.13e+03   1.10e+02   -46.5  < 2e-16
TEMP            3.11e+01   1.34e+00    23.1  < 2e-16
PRESSURE        1.40e+02   3.14e+00    44.5  < 2e-16
I(TEMP^2)      -1.33e-01   6.85e-03   -19.5  6.5e-15
I(PRESSURE^2) -1.14e+00   2.74e-02   -41.7  < 2e-16
TEMP:PRESSURE -1.45e-01   9.69e-03   -15.0  1.1e-12

Residual standard error: 1.68 on 21 degrees of freedom
Multiple R-squared:  0.993,   Adjusted R-squared:  0.991
F-statistic:  596 on 5 and 21 DF,  p-value: <2e-16
```

Figure 5.14 Output for complete second-order model of quality

```
rm(list=ls())
prod.df <- read.table("PRODQUAL.txt",header=T)
attach(prod.df)
names(proqual.df) ## to see the name of the variables
mod1<-lm(QUALITY ~ TEMP*PRESSURE + I(TEMP^2) + I(PRESSURE^2))
summary(mod1)
```

Remember the asterisk means the model includes the main effects and the interaction.

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -5.13e+03   1.10e+02   -46.5  < 2e-16
TEMP            3.11e+01   1.34e+00    23.1  < 2e-16
PRESSURE        1.40e+02   3.14e+00    44.5  < 2e-16
I(TEMP^2)      -1.33e-01   6.85e-03   -19.5  6.5e-15
I(PRESSURE^2)  -1.14e+00   2.74e-02   -41.7  < 2e-16
TEMP:PRESSURE  -1.45e-01   9.69e-03   -15.0  1.1e-12

Residual standard error: 1.68 on 21 degrees of freedom
Multiple R-squared:  0.993,  Adjusted R-squared:  0.991
F-statistic:  596 on 5 and 21 DF,  p-value: <2e-16
```

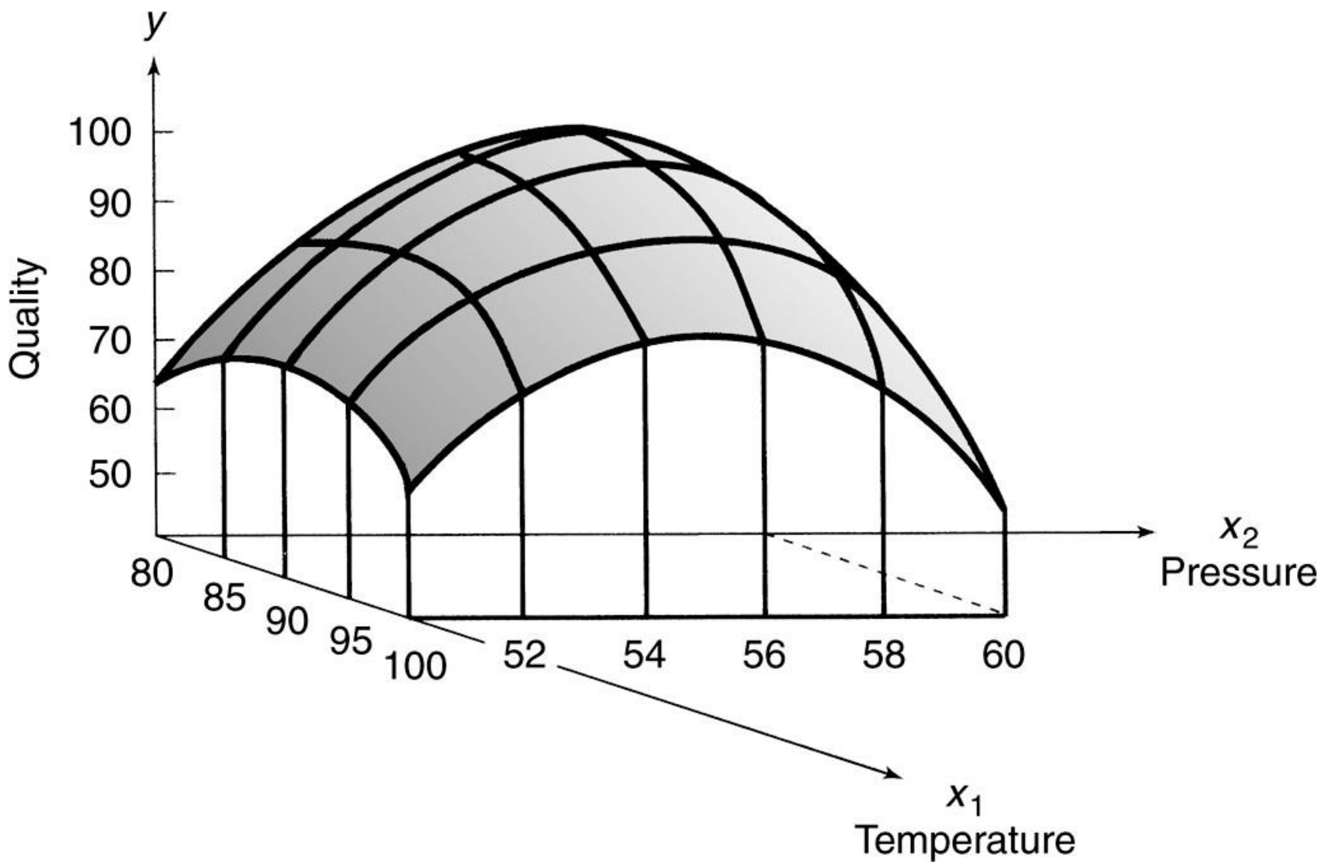
Figure 5.14 Output for complete second-order model of quality

```
rm(list=ls())
prod.df <- read.table("PRODQUAL.txt",header=T)
attach(prod.df)
names(proqual.df) ## to see the name of the variables
mod1<-lm(QUALITY ~ TEMP*PRESSURE + I(TEMP^2) + I(PRESSURE^2))
> anova(mod1)
Analysis of Variance Table

Response: QUALITY
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
TEMP	1	1511	1511	536.1	**< 2e-16**
PRESSURE	1	279	279	99.1	**2.1e-09**
I(TEMP^2)	1	1068	1068	378.8	**6.5e-15**
I(PRESSURE^2)	1	4910	4910	1742.2	**< 2e-16**
TEMP:PRESSURE	1	635	635	225.4	**1.1e-12**
Residuals	21	59	3		

Figure 5.15 Graph of second-order least squares model

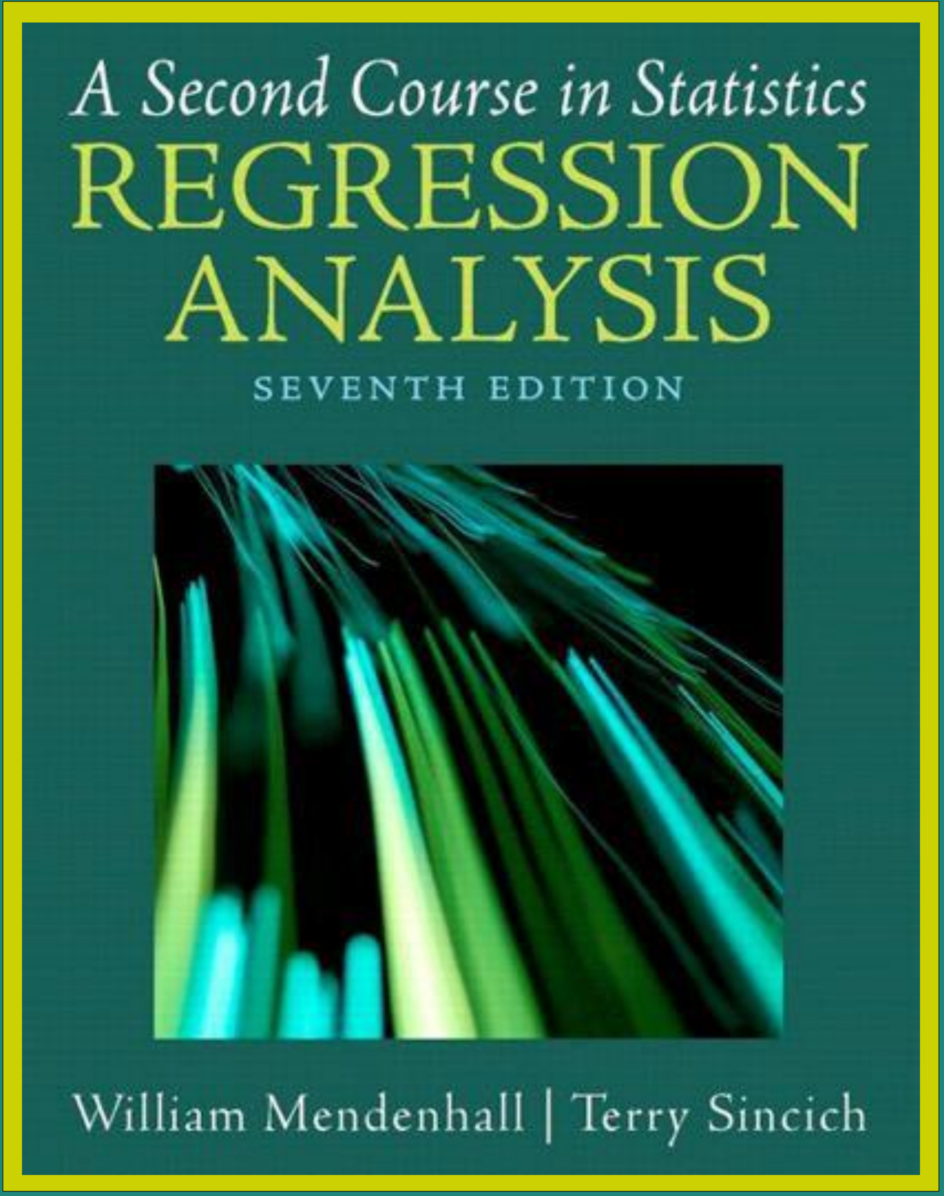


Chapter 5

Principles of Model Building

(Lecture 2)

Dr Brenda Vo



Chapter 5 Outline

Lecture 1

- Introduction
- Models with 1 quantitative predictor
- First - order models with ≥ 2 quantitative predictors
- Second - order models with ≥ 2 quantitative predictors

Lecture 2

- Model with 1 qualitative predictor
- Model with 2 qualitative predictors
- Model with ≥ 3 qualitative predictors
- Models with both qualitative & quantitative predictors

§5.6 is *not* covered in this unit

Models with 1 qualitative predictor

Model with 1 qualitative predictor

Procedure for Writing a Model with One Qualitative Independent Variable at *k* Levels (A, B, C, D, . . .)

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{k-1} x_{k-1}$$

where

$$x_i = \begin{cases} 1 & \text{if qualitative variable at level } i+1 \\ 0 & \text{otherwise} \end{cases}$$

The number of dummy variables for a single qualitative variable is always 1 less than the number of levels for the variable. Then, assuming the base level is A, the mean for each level is

$$\mu_A = \beta_0$$
$$\mu_B = \beta_0 + \beta_1$$
$$\mu_C = \beta_0 + \beta_2$$
$$\mu_D = \beta_0 + \beta_3$$
$$\vdots$$

β *Interpretations*:

$$\beta_0 = \mu_A$$
$$\beta_1 = \mu_B - \mu_A$$
$$\beta_2 = \mu_C - \mu_A$$
$$\beta_3 = \mu_D - \mu_A$$
$$\vdots$$

Model with 1 qualitative predictor

Example 5.5, p. 280

Compare annual maintenance costs of a computerized system for monitoring road construction bids. Mean annual cost is recorded for ten users sampled from three different states.

The dataset is saved as *BIDMAINT.txt*

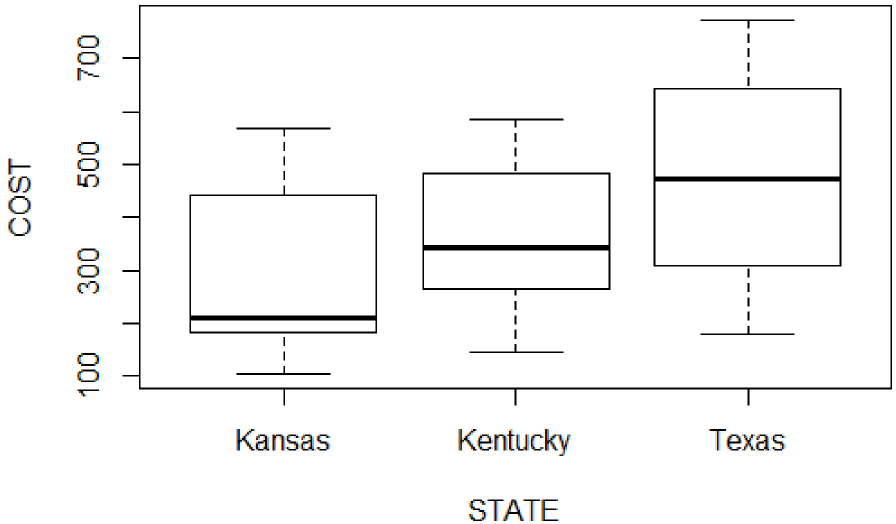
Data sets and R scripts files used in lectures and workshops

Model with 1 qualitative predictor

```
bid.df<-read.table("BIDMAINT.txt",header=T)
bid.df$STATE <- factor(bid.df$STATE)
boxplot(COST~STATE, data=bid.df)
```

Table 5.6 Annual maintenance costs

	State Installation		
	Kansas	Kentucky	Texas
	$ 198	$ 563	$ 385
	126	314	693
	443	483	266
	570	144	586
	286	585	178
	184	377	773
	105	264	308
	216	185	430
	465	330	644
	203	354	515
Totals	$2,796	$3,599	$4,778



What model are we fitting?

When building a model:

- Choose (or know) the baseline: Kansas
- Number of dummy variables: k-1

$$\widehat{COST} = \boxed{\beta_0} + \boxed{\beta_1} STATE_{Kentucky} + \boxed{\beta_2} STATE_{Texas}$$

Mean for baseline (EG Kansas)

Difference between mean cost of baseline (EG Kansas) and mean cost of Kentucky

Difference between mean cost of baseline (EG Kansas) and mean cost of Texas

Annual maintenance costs

```
mod<-lm(COST~STATE, data=bid.df)
summary(mod)
```

Q: Give an informative interpretation of the output, and estimate mean annual cost per state.

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)       279.6       53.4    5.23  1.6e-05
STATEKentucky      80.3       75.6    1.06    0.297
STATETexas        198.2       75.6    2.62    0.014

Residual standard error: 169 on 27 degrees of freedom
Multiple R-squared:  0.205,      Adjusted R-squared:  0.146
F-statistic: 3.48 on 2 and 27 DF,  p-value: 0.0452
```

- The global F-test indicates that ***not all mean costs are the same*** (F=3.48 on 2,27df, p-value =0.045)

Annual maintenance costs

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)      279.6       53.4    5.23  1.6e-05
StateKentucky     80.3       75.6    1.06    0.297
StateTexas       198.2       75.6    2.62    0.014
```

The t-tests indicate that

- There is ***no significant difference in mean annual maintenance costs*** between Kansas and Kentucky (p=0.297)
- The mean cost for Texas is significantly greater than that in Kansas (p=0.014).

Q: Estimate mean annual maintenance cost for each state

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)       279.6       53.4    5.23  1.6e-05
StateKentucky      80.3       75.6    1.06    0.297
StateTexas        198.2       75.6    2.62    0.014
```

$$\widehat{COST} = \boxed{\beta_0} + \boxed{\beta_1} STATE_{Kentucky} + \boxed{\beta_2} STATE_{Texas}$$

$$\widehat{COST_{Kansas}} = \mu_{Kansas} = \beta_0 = 279.6$$

$$\widehat{COST_{Kentucky}} = \mu_{Kentucky} = \beta_0 + \beta_1 = 279.6 + 80.3 = 359.9$$

$$\widehat{COST_{Texas}} = \mu_{Texas} = \beta_0 + \beta_2 = 279.6 + 198.2 = 477.8$$

Q: Estimate mean annual maintenance cost for each state

$$\widehat{COST} = \beta_0 + \beta_1 STATE_{Kentucky} + \beta_2 STATE_{Texas}$$

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)          279.6        53.4    5.23  1.6e-05
StateKentucky         80.3        75.6    1.06    0.297
StateTexas           198.2        75.6    2.62    0.014
```

These are CIs for differences in means

```
> confint(mod)
                 Estimate Std. Error    2.5 %  97.5 %
(Intercept)         279.6      53.43   169.98   389.2
STATEKentucky        80.3      75.56   -74.73   235.3
STATETexas          198.2      75.56    43.17   353.2
```

The mean maintenance cost in Kansas is between $169.98 and $389.20

Q: Estimate mean annual maintenance cost for each state

$$\widehat{COST} = \beta_0 + \beta_1 STATE_{Kentucky} + \beta_2 STATE_{Texas}$$

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	279.6	53.4	5.23	1.6e-05
StateKentucky	80.3	75.6	1.06	$\boxed{0.297}$
StateTexas	198.2	75.6	2.62	0.014

These are CIs for differences in means

```
> confint(mod)
```

	Estimate	Std. Error	2.5 %	97.5 %
(Intercept)	279.6	53.43	169.98	389.2
STATEKentucky	80.3	75.56	-74.73	235.3
STATETexas	**198.2**	75.56	**43.17**	**353.2**

The *difference* in mean maintenance cost between Kansas and Kentucky is between \$74.73 less and \$235.3 more.

NOTE: because the CI include 0, here we can say there is no difference in maintenance cost between Kansas and Kentucky.

Q: Estimate mean annual maintenance cost for each state

$$\widehat{COST} = \beta_0 + \beta_1 STATE_{Kentucky} + \beta_2 STATE_{Texas}$$

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)      279.6       53.4    5.23  1.6e-05
StateKentucky     80.3       75.6    1.06    0.297
StateTexas       198.2       75.6    2.62    0.014

> confint(mod)
              Estimate Std. Error    2.5 % 97.5 %
(Intercept)      279.6      53.43   169.98  389.2
STATEKentucky     80.3      75.56   -74.73  235.3
STATETexas       198.2      75.56    43.17  353.2
```

These are CIs for differences in means

The *difference* in mean maintenance cost between Kansas and Texas is between \$43.17 and \$353.20 more.

Alternatively: maintenance coast in Texas is between \$43.17 and \$353.20 more than maintenance costs in Kansas.

NOTE: because the CI does not include 0, here we can say there is a significant difference in maintenance cost between Kansas and Texas.

Estimate mean annual maintenance cost for each state and 95% CI

$$\widehat{COST}_{Kansas} = \mu_{Kansas} = \beta_0 = 279.6$$

$$\widehat{COST}_{Kentucky} = \mu_{Kentucky} = \beta_0 + \beta_1 = 279.6 + 80.3 = 359.9$$

$$\widehat{COST}_{Texas} = \mu_{Texas} = \beta_0 + \beta_2 = 279.6 + 198.2 = 477.8$$

```
mod2<-lm(COST~STATE -1,data=bid.df)
confint(mod2)
```

	Estimate	Std. Error	2.5 %	97.5 %
STATEKansas	**279.6**	53.43	**170.0**	**389.2**
STATEKentucky	**359.9**	53.43	**250.3**	**469.5**
STATETexas	**477.8**	53.43	**368.2**	**587.4**

Q: What information is still missing?

Annual maintenance costs

Changed the baseline to Texas, to compare maintenance costs between Texas and Kentucky.

```
bid.df$STATE <- factor(bid.df$STATE)
bid.df$STATE <- relevel(bid.df$STATE, ref="Texas")
mod3<-lm(COST ~ STATE, data=bid.df)
summary(mod3)
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)      477.8       53.4    8.94  1.5e-09
StateKansas     -198.2       75.6   -2.62    0.014
StateKentucky   -117.9       75.6   -1.56    0.130
```

Q: Interpret this output

Models with 2 qualitative predictor

Models with 2 qualitative predictors

Table 5.7 The six combinations of fuel type and diesel engine brand

		Brand	
		B_1	B_2
FUEL TYPE	F_1	μ_{11}	μ_{12}
	F_2	μ_{21}	μ_{22}
	F_3	μ_{31}	μ_{32}

Table 5.8 Performance data for combinations of fuel type and diesel engine brand

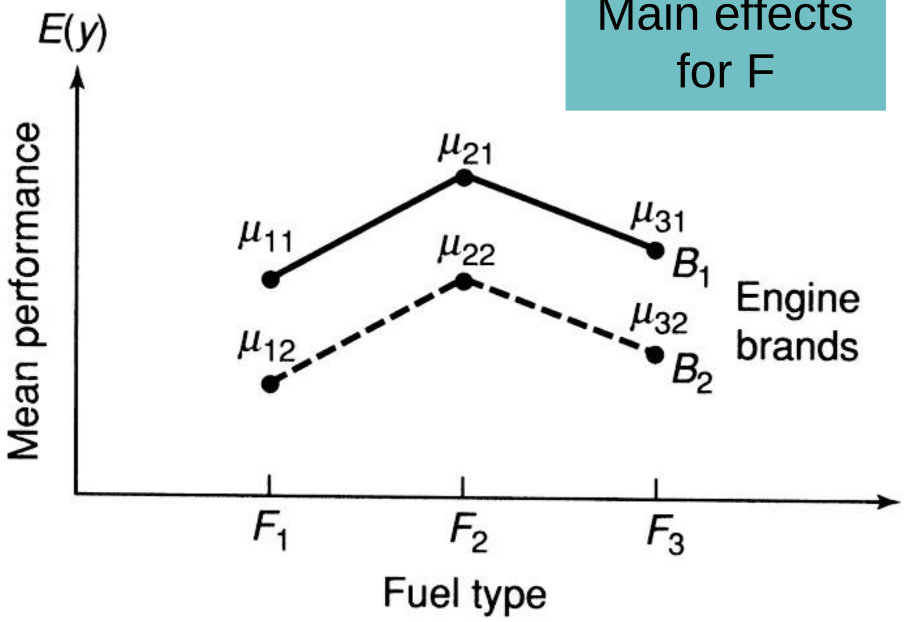
		Brand	
		B_1	B_2
FUEL TYPE	F_1	65 73 68	36
	F_2	78 82	50 43
	F_3	48 46	61 62

We want to model the mean performance, E(y), of a diesel engine as a function of both qualitative predictors: *Fuel type and Brand*.

The data is saved as *DIESEL.txt* file.

Models with 2 qualitative predictors

$$E(y) = \beta_0 + \underbrace{\beta_1 x_1 + \beta_2 x_2}_{\text{Main effects for F}} + \underbrace{\beta_3 x_3}_{\text{Main effect for B}}$$



Without interaction

Figure 5.20 Hypothetical main effects model:
Mean response as a function of *F* and *B* when *F* and *B* affect *E*(*y*) ***independently***

Models with 2 qualitative predictors

Without interaction

Main Effects Model with Two Qualitative Independent Variables, One at Three Levels (F_1, F_2, F_3) and the Other at Two Levels (B_1, B_2)

$$E(y) = \beta_0 + \overbrace{\beta_1 x_1 + \beta_2 x_2}^{\text{Main effect terms for } F} + \overbrace{\beta_3 x_3}^{\text{Main effect term for } B}$$

where

$$x_1 = \begin{cases} 1 & \text{if } F_2 \\ 0 & \text{if not} \end{cases} \qquad x_2 = \begin{cases} 1 & \text{if } F_3 \\ 0 & \text{if not} \end{cases} \qquad (F_1 \text{ is base level})$$

$$x_3 = \begin{cases} 1 & \text{if } B_2 \\ 0 & \text{if } B_1 \quad \text{(base level)} \end{cases}$$

β_1 : Mean differences ($F_2 – F_1$) for brand 1

Interpreting intercept and main effects

Without interaction

Mean for baseline (EG Fuel 1 & Brand 1)

Difference between fuel baseline (EG Fuel 1) and Fuel 2 at Brand 1

Difference between fuel baseline (EG Fuel 1) and Fuel 3 at Brand 1

$$E(Performance) = \boxed{\beta_0} + \boxed{\beta_1} fuel_2 + \boxed{\beta_2} fuel_3 + \boxed{\beta_3} brand_2$$

Difference between brand baseline (EG Brand 1) and Brand 2 at Fuel 1

Models with 2 qualitative predictors

Without interaction

Table 5.7 The six combinations of fuel type and diesel engine brand

		Brand	
		B_1	B_2
FUEL TYPE	F_1	μ_{11}	μ_{12}
	F_2	μ_{21}	μ_{22}
	F_3	μ_{31}	μ_{32}

Main Effects Model with Two Qualitative Independent Variables, One at Three Levels (F_1, F_2, F_3) and the Other at Two Levels (B_1, B_2)

$$E(y) = \beta_0 + \overbrace{\beta_1 x_1 + \beta_2 x_2}^{\text{Main effect terms for } F} + \overbrace{\beta_3 x_3}^{\text{Main effect term for } B}$$

where

$$x_1 = \begin{cases} 1 & \text{if } F_2 \\ 0 & \text{if not} \end{cases} \quad x_2 = \begin{cases} 1 & \text{if } F_3 \\ 0 & \text{if not} \end{cases} \quad (F_1 \text{ is base level})$$

$$x_3 = \begin{cases} 1 & \text{if } B_2 \\ 0 & \text{if } B_1 \quad \text{(base level)} \end{cases}$$

F1 and B1 occur when $x_1 = x_2 = x_3 = 0$

Then:

$E(y) = \beta_0 + \beta_1(0) + \beta_2(0) + \beta_3(0) = \beta_0$ ➔ $\mu_{11} = \beta_0$

Similar F2 and B1 occur when $x_1 = 1, x_2 = x_3 = 0$

$E(y) = \beta_0 + \beta_1(1) + \beta_2(0) + \beta_3(0) = \beta_0 + \beta_1$ ➔ $\mu_{21} = \beta_0 + \beta_1$

Therefore, difference between F1 and F2 for Brand 1: ➔ $\beta_1 = \mu_{21} - \mu_{11}$

Interpreting intercept and main effects

Without interaction

Mean for baseline (EG Fuel 1 & Brand 1)

Difference between fuel baseline (EG Fuel 1) and Fuel 2 at Brand 1

Difference between fuel baseline (EG Fuel 1) and Fuel 3 at Brand 1

$$E(Performance) = \boxed{\beta_0} + \boxed{\beta_1} fuel_2 + \boxed{\beta_2} fuel_3 + \boxed{\beta_3} brand_2$$

Difference between brand baseline (EG Brand 1) and Brand 2 at Fuel 1

Models with 2 qualitative predictors

With interaction

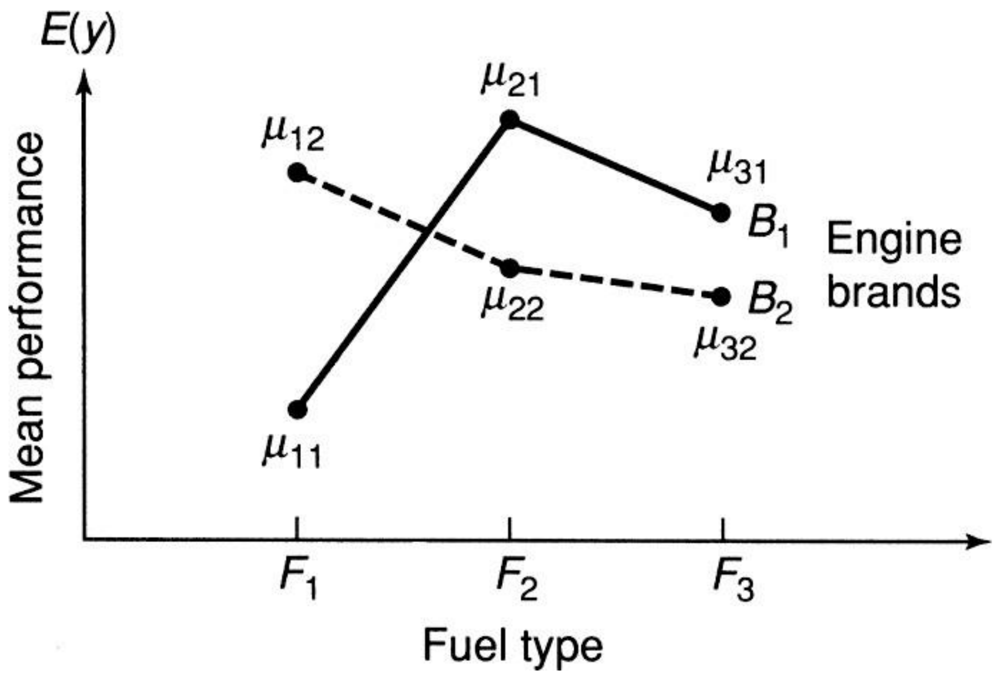


Figure 5.21 Hypothetical Interaction model:
Mean response as a function of *F* and *B* when *F* and *B* ***interact*** to affect *E*(*y*)

Models with 2 qualitative predictors

With interaction

Interaction Model with Two Qualitative Independent Variables, One at Three Levels (F_1, F_2, F_3) and the Other at Two Levels (B_1, B_2)

$$E(y) = \beta_0 + \overbrace{\beta_1 x_1 + \beta_2 x_2}^{\substack{\text{Main effect} \\ \text{terms for } F}} + \overbrace{\beta_3 x_3}^{\substack{\text{Main effect} \\ \text{term for } B}} + \overbrace{\beta_4 x_1 x_3 + \beta_5 x_2 x_3}^{\substack{\text{Interaction} \\ \text{terms}}}$$

where the dummy variables x_1, x_2, and x_3 are defined in the same way as for the main effects model.

Interpretation of Model Parameters

$$\beta_0 = \mu_{11} \text{ (Mean of the combination of base levels)}$$

$$\beta_1 = \mu_{21} - \mu_{11} \text{ (i.e., for base level } B_1 \text{ only)}$$

$$\beta_2 = \mu_{31} - \mu_{11} \text{ (i.e., for base level } B_1 \text{ only)}$$

$$\beta_3 = \mu_{12} - \mu_{11} \text{ (i.e., for base level } F_1 \text{ only)}$$

$$\beta_4 = (\mu_{22} - \mu_{12}) - (\mu_{21} - \mu_{11})$$

$$\beta_5 = (\mu_{32} - \mu_{12}) - (\mu_{31} - \mu_{11})$$

Models with 2 qualitative predictors

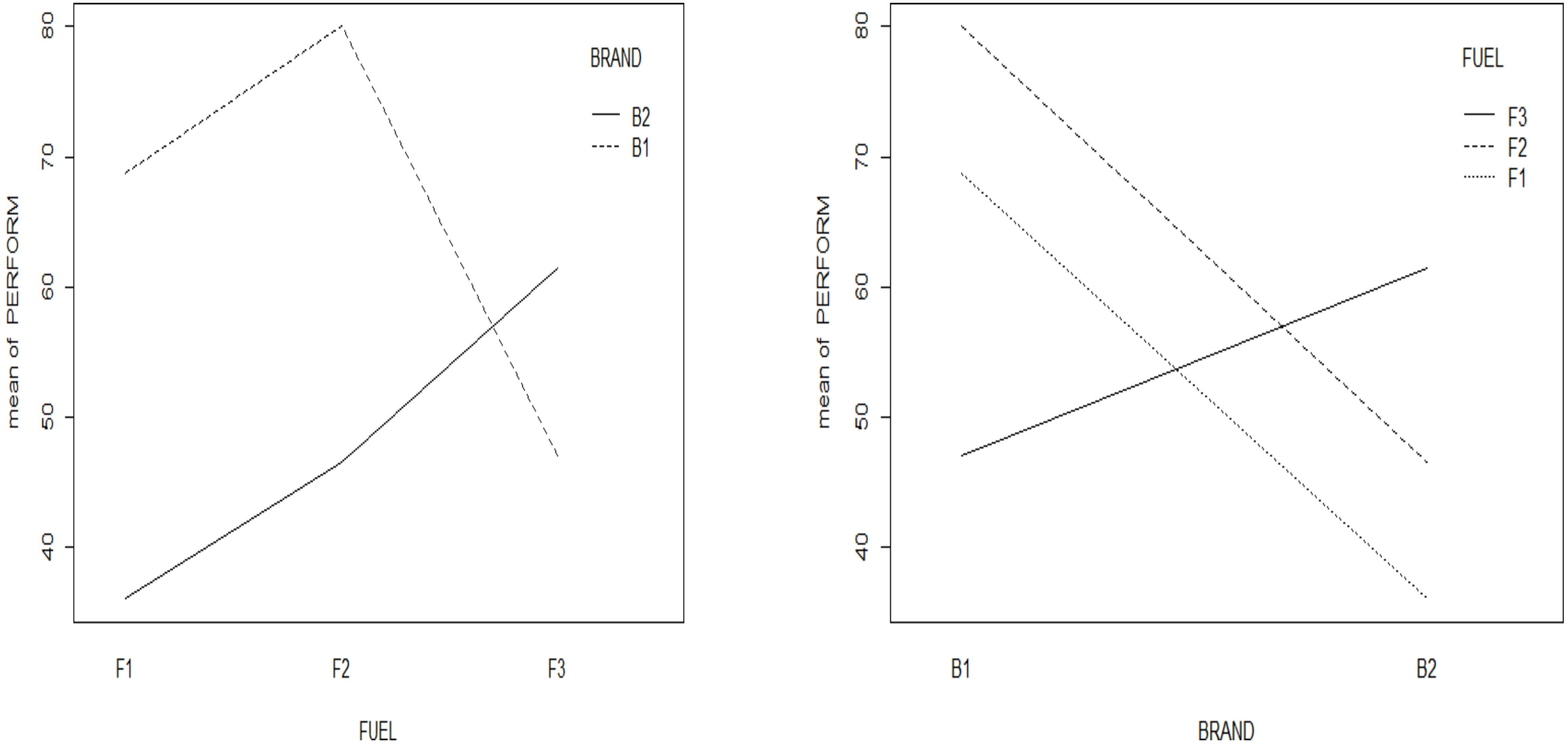
FUELBRAND	PERFORM	FUEL	BRAND
F1B1	65	F1	B1
F1B1	73	F1	B1
F1B1	68	F1	B1
F1B2	36	F1	B2
F2B1	78	F2	B1
F2B1	82	F2	B1
F2B2	50	F2	B2
F2B2	43	F2	B2
F3B1	48	F3	B1
F3B1	46	F3	B1

```
rm(list = ls()) ## remove all of the variables in the
working environment
diesel.df<-read.table("DIESEL.txt",header=T)
```

Models with 2 qualitative predictors

Interaction plots

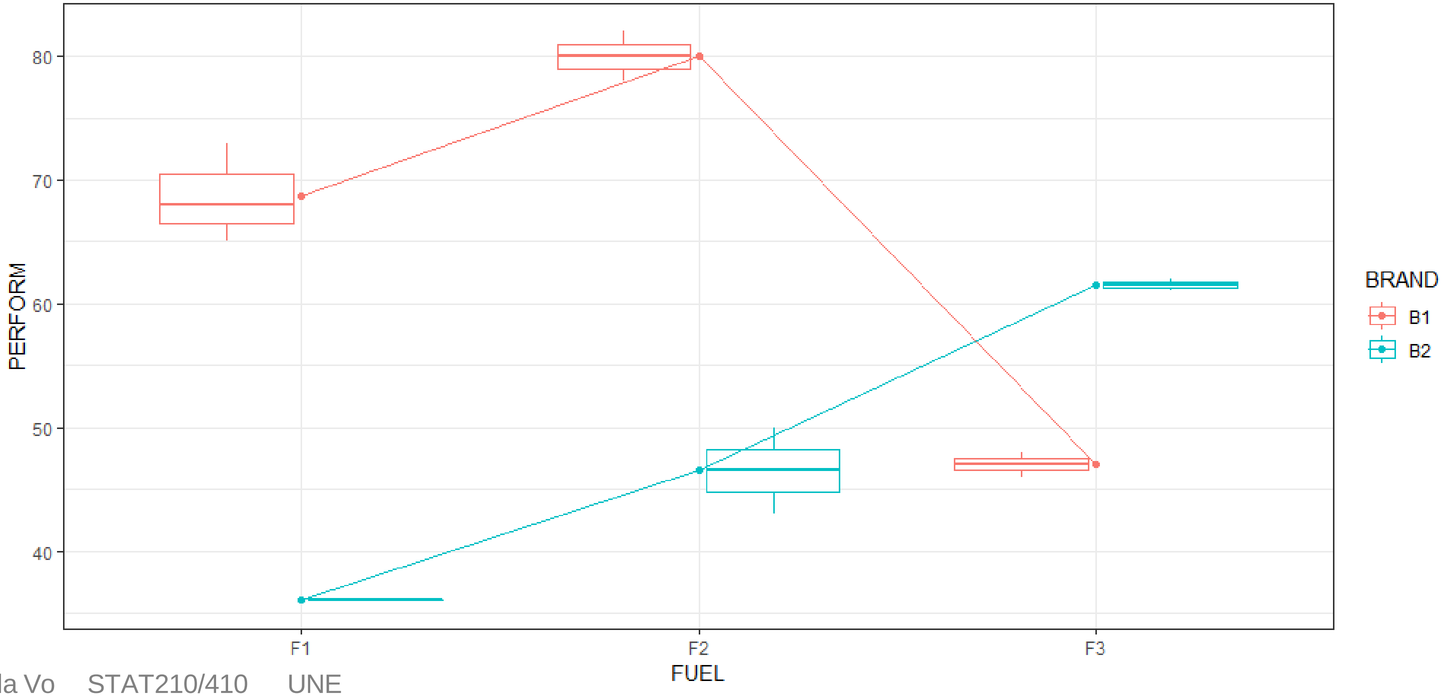
```
with(diesel.df, interaction.plot(FUEL,BRAND,PERFORM))
```



Q: What R code will produce the plot on the right?

```
library(ggplot2)
library(plyr)

# calculate interaction means
dieInt <- ddply(die.df,.(FUEL,BRAND),summarise, val =
mean(PERFORM))
# Interaction plot of means, with corresponding boxplots
ggplot(die.df, aes(x = FUEL, y = PERFORM, colour = BRAND)) +
  geom_boxplot() +
  geom_point(data = dieInt, aes(y = val)) +
  geom_line(data = dieInt, aes(y = val, group = BRAND)) +
  theme_bw()
```



Example: Fuel type - Brand

interaction model

$$E(y) = \beta_0 + \overbrace{\beta_1 x_1 + \beta_2 x_2}^{\substack{\text{Main effect} \\ \text{terms for } F}} + \overbrace{\beta_3 x_3}^{\substack{\text{Main effect} \\ \text{term for } B}} + \overbrace{\beta_4 x_1 x_3 + \beta_5 x_2 x_3}^{\substack{\text{Interaction} \\ \text{terms}}}$$

```
mod <- lm(PERFORM ~ FUEL*BRAND, data = diesel.df)
summary(mod)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       68.667      1.939   35.42  3.4e-08
FUELF2            11.333      3.066    3.70  0.01013
FUELF3           -21.667      3.066   -7.07  0.00040
BRANDB2          -32.667      3.878   -8.42  0.00015
FUELF2:BRANDB2    -0.833      5.130   -0.16  0.87628
FUELF3:BRANDB2    47.167      5.130    9.19  9.3e-05

Residual standard error: 3.36 on 6 degrees of freedom
Multiple R-squared:  0.971,     Adjusted R-squared:  0.948
F-statistic: 40.8 on 5 and 6 DF,  p-value: 0.000148
```

Example: Fuel type - Brand

interaction model

$$E(y) = \beta_0 + \overbrace{\beta_1 x_1 + \beta_2 x_2}^{\substack{\text{Main effect} \\ \text{terms for } F}} + \overbrace{\beta_3 x_3}^{\substack{\text{Main effect} \\ \text{term for } B}} + \overbrace{\beta_4 x_1 x_3 + \beta_5 x_2 x_3}^{\substack{\text{Interaction} \\ \text{terms}}}$$

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      68.667      1.939   35.42  3.4e-08
FUELF2           11.333      3.066    3.70  0.01013
FUELF3          -21.667      3.066   -7.07  0.00040
BRANDB2         -32.667      3.878   -8.42  0.00015
FUELF2:BRANDB2   -0.833      5.130   -0.16  0.87628
FUELF3:BRANDB2   47.167      5.130    9.19  9.3e-05

The regression equation is:
```

E(Perform)= 68.67 + 11.33*Fuel2 -21.67*Fuel3 – 32.67*Brand2 – 0.83*Fuel2*Brand2 + 47.17*Fuel3*Brand2

Models with 2 qualitative predictors

interaction model

```
anova(mod)
Analysis of Variance Table

Response: PERFORM
           Df Sum Sq Mean Sq F value   Pr(>F)
FUEL        2    170      85    7.54 0.02303
BRAND       1    688     688   61.01 0.00023
FUEL:BRAND  2   1445     722   64.05    9e-05
Residuals   6     68      11
```

Interpreting the regression coefficients

$$E(y) = \beta_0 + \overbrace{\beta_1 x_1 + \beta_2 x_2}^{\text{Main effect terms for } F} + \overbrace{\beta_3 x_3}^{\text{Main effect term for } B} + \overbrace{\beta_4 x_1 x_3 + \beta_5 x_2 x_3}^{\text{Interaction terms}}$$

```
E(Perform)= 68.67 + 11.33*Fuel2
     - 21.67*Fuel3 - 32.67*Brand2
          - 0.83*Fuel2*Brand2
        - + 47.17*Fuel3*Brand2
```

Table 5.7 The six combinations of fuel type and diesel engine brand

		Brand	
		B_1	B_2
FUEL TYPE	F_1	μ_{11}	μ_{12}
	F_2	μ_{21}	μ_{22}
	F_3	μ_{31}	μ_{32}

$\boxed{\beta_0 = \mu_{F1B1} = \mu_{11} = 68.67}$ Mean of combined base levels (F1, B1)

$\beta_1 = \boxed{\mu_{F2B1}} - \boxed{\beta_0} = \mu_{21} - \mu_{11} = 11.33$ Difference in means (F2-F1) at base level B1

$\beta_2 = \mu_{F3B1} - \beta_0 = \mu_{31} - \mu_{11} = -21.67$ Difference in means (F3-F1) at base level B1

$\beta_3 = \mu_{F1B2} - \beta_0 = \mu_{12} - \mu_{11} = -32.67$ Difference in means (B2-B1) at base level F1

Interpreting the regression coefficients

$$E(y) = \beta_0 + \overbrace{\beta_1 x_1 + \beta_2 x_2}^{\text{Main effect terms for } F} + \overbrace{\beta_3 x_3}^{\text{Main effect term for } B} + \overbrace{\beta_4 x_1 x_3 + \beta_5 x_2 x_3}^{\text{Interaction terms}}$$

Table 5.7 The six combinations of fuel type and diesel engine brand

		Brand	
		B_1	B_2
FUEL TYPE	F_1	$\boxed{\mu_{11}}$	μ_{12}
	F_2	μ_{21}	μ_{22}
	F_3	$\boxed{\mu_{31}}$	μ_{32}

```
E(Perform)= 68.67 + 11.33*Fuel2
    - 21.67*Fuel3 - 32.67*Brand2
            - 0.83*Fuel2*Brand2
          - + 47.17*Fuel3*Brand2
```

$$\boxed{\beta_0 = \mu_{F1B1} = \mu_{11} = 68.67}$$

Mean of combined base levels (F1, B1)

$$\beta_1 = \mu_{F2B1} - \beta_0 = \mu_{21} - \mu_{11} = 11.33$$

Difference in means (F2-F1) at base level B1

$$\beta_2 = \boxed{\mu_{F3B1}} - \boxed{\beta_0} = \mu_{31} - \mu_{11} = -21.67$$

Difference in means (F3-F1) at base level B1

$$\beta_3 = \mu_{F1B2} - \beta_0 = \mu_{12} - \mu_{11} = -32.67$$

Difference in means (B2-B1) at base level F1

Interpreting the regression coefficients

$$E(y) = \beta_0 + \overbrace{\beta_1 x_1 + \beta_2 x_2}^{\substack{\text{Main effect} \\ \text{terms for } F}} + \overbrace{\beta_3 x_3}^{\substack{\text{Main effect} \\ \text{term for } B}} + \overbrace{\beta_4 x_1 x_3 + \beta_5 x_2 x_3}^{\substack{\text{Interaction} \\ \text{terms}}}$$

Table 5.7 The six combinations of fuel type and diesel engine brand

		Brand	
		B_1	B_2
FUEL TYPE	F_1	μ_{11}	μ_{12}
	F_2	μ_{21}	μ_{22}
	F_3	μ_{31}	μ_{32}

```
E(Perform)= 68.67 + 11.33*Fuel2
      - 21.67*Fuel3 - 32.67*Brand2
              - 0.83*Fuel2*Brand2
           - + 47.17*Fuel3*Brand2
```

$\boxed{\beta_0 = \mu_{F1B1} = \mu_{11} = 68.67}$ Mean of combined base levels (F1, B1)

$\beta_1 = \mu_{F2B1} - \beta_0 = \mu_{21} - \mu_{11} = 11.33$ Difference in means (F2-F1) at base level B1

$\beta_2 = \mu_{F3B1} - \beta_0 = \mu_{31} - \mu_{11} = -21.67$ Difference in means (F3-F1) at base level B1

$\beta_3 = \boxed{\mu_{F1B2}} - \boxed{\beta_0} = \mu_{12} - \mu_{11} = -32.67$ Difference in means (B2-B1) at base level F1

Interpreting intercept and main effects

Mean for baseline (EG Fuel 1 & Brand 1)

Difference between fuel baseline (EG Fuel 1) and Fuel 2 at Brand 1

Difference between fuel baseline (EG Fuel 1) and Fuel 3 at Brand 1

$$E(Perform) = \boxed{68.67} + \boxed{11.33}*Fuel2 \boxed{-21.67}*Fuel3 \boxed{-\ 32.67}*Brand2 - 0.83*Fuel2*Brand2 + 47.17*Fuel3*Brand2$$

Difference between brand baseline (EG Brand 1) and Brand 2 at Fuel 1

Interpreting the regression coefficients

$$E(y) = \beta_0 + \overbrace{\beta_1 x_1 + \beta_2 x_2}^{\text{Main effect terms for } F} + \overbrace{\beta_3 x_3}^{\text{Main effect term for } B} + \overbrace{\beta_4 x_1 x_3 + \beta_5 x_2 x_3}^{\text{Interaction terms}}$$

Table 5.7 The six combinations of fuel type and diesel engine brand

		Brand	
		B_1	B_2
	F_1	μ_{11}	μ_{12}
FUEL TYPE	F_2	μ_{21}	μ_{22}
	F_3	μ_{31}	μ_{32}

Interaction $\beta_4 x_1 x_3$

$x_1 = 1$ (Fuel 2), $x_2 = 0$, $x_3 = 1$ (Brand2)

$$\mu_{22} = E(y) = \beta_0 + \beta_1(1) + \beta_2(0) + \beta_3(1) + \beta_4(1)(1) + \beta_5(0)(1)$$

$$\mu_{22} = \beta_0 + \beta_1 + \beta_3 + \beta_4$$

$$\Rightarrow \beta_4 = \mu_{22} - \beta_0 - \beta_1 - \beta_3$$

$$\Rightarrow \beta_4 = \mu_{22} - \mu_{11} - (\mu_{21} - \mu_{11}) - (\mu_{12} - \mu_{11})$$

$$= \mu_{22} - \cancel{\mu_{11}} - \mu_{21} + \cancel{\mu_{11}} - \mu_{12} + \mu_{11}$$

$$= \mu_{22} - \mu_{21} - \mu_{12} + \mu_{11} = \boxed{(\mu_{22}} - \boxed{\mu_{12})} - \boxed{(\mu_{21}} - \boxed{\mu_{11})}$$

Compares the change in mean performance between Brand 1 and 2, as we move from Fuel 1 to Fuel 2

Interpreting the regression coefficients

$$E(y) = \beta_0 + \overbrace{\beta_1 x_1 + \beta_2 x_2}^{\text{Main effect terms for } F} + \overbrace{\beta_3 x_3}^{\text{Main effect term for } B} + \overbrace{\beta_4 x_1 x_3 + \beta_5 x_2 x_3}^{\text{Interaction terms}}$$

Table 5.7 The six combinations of fuel type and diesel engine brand

		Brand	
		B_1	B_2
FUEL TYPE	F_1	$\boxed{\mu_{11}}$	$\boxed{\mu_{12}}$
	F_2	μ_{21}	μ_{22}
	F_3	$\boxed{\mu_{31}}$	$\boxed{\mu_{32}}$

Interaction $\beta_5 x_2 x_3$

$x_1 = 0$, $x_2 = 1$ (Fuel 3), $x_3 = 1$ (Brand2)

$$\mu_{32} = E(y) = \beta_0 + \beta_1(0) + \beta_2(1) + \beta_3(1) + \beta_4(0)(1) + \beta_5(1)(1)$$

$$\mu_{32} = \beta_0 + \beta_2 + \beta_3 + \beta_5$$

$$\Rightarrow \beta_5 = \mu_{32} - \beta_0 - \beta_1 - \beta_3$$

$$\Rightarrow \beta_5 = \mu_{32} - \mu_{11} - (\mu_{31} - \mu_{11}) - (\mu_{12} - \mu_{11})$$

$$= \mu_{32} - \cancel{\mu_{11}} - \mu_{31} + \cancel{\mu_{11}} - \mu_{12} + \mu_{11}$$

$$= \mu_{32} - \mu_{31} - \mu_{12} + \mu_{11} = \boxed{(\mu_{32}} - \boxed{\mu_{12})} - \boxed{(\mu_{31} -} \boxed{\mu_{11})}$$

Compares the change in mean performance between Brand 1 and 2, as we move from Fuel 1 to Fuel 3

Interpreting the regression coefficients

$$E(y) = \beta_0 + \overbrace{\beta_1 x_1 + \beta_2 x_2}^{\substack{\text{Main effect}\\ \text{terms for } F}} + \overbrace{\beta_3 x_3}^{\substack{\text{Main effect}\\ \text{term for } B}} + \overbrace{\beta_4 x_1 x_3 + \beta_5 x_2 x_3}^{\substack{\text{Interaction}\\ \text{terms}}}$$

```
E(Perform)= 68.67 + 11.33*Fuel2 - 21.67*Fuel3
            - 32.67*Brand2 - 0.83*Fuel2*Brand2
            + 47.17*Fuel3*Brand2
```

Interactions

$$\beta_4 = (\mu_{22} - \mu_{12}) - (\mu_{21} - \mu_{11}) = -0.83$$

Compares the change in mean performance between Brand 1 and 2, as we move from Fuel 1 to Fuel 2

$$\beta_5 = (\mu_{32} - \mu_{12}) - (\mu_{31} - \mu_{11}) = 47.17$$

Compares the change in mean performance between Brand 1 and 2, as we move from Fuel 1 to Fuel 3

Interpreting intercept and main effects

Mean for baseline (EG Fuel 1 & Brand 1)

Difference between fuel baseline (EG Fuel 1) and Fuel 2 at Brand 1

Difference between fuel baseline (EG Fuel 1) and Fuel 3 at Brand 1

$$E(Perform) = \boxed{68.67} + \boxed{11.33}*Fuel2 \boxed{-21.67}*Fuel3 \boxed{-32.67}*Brand2 \boxed{-0.83}*Fuel2*Brand2 \boxed{+47.17}*Fuel3*Brand2$$

Difference between brand baseline (EG Brand 1) and Brand 2 at Fuel 1

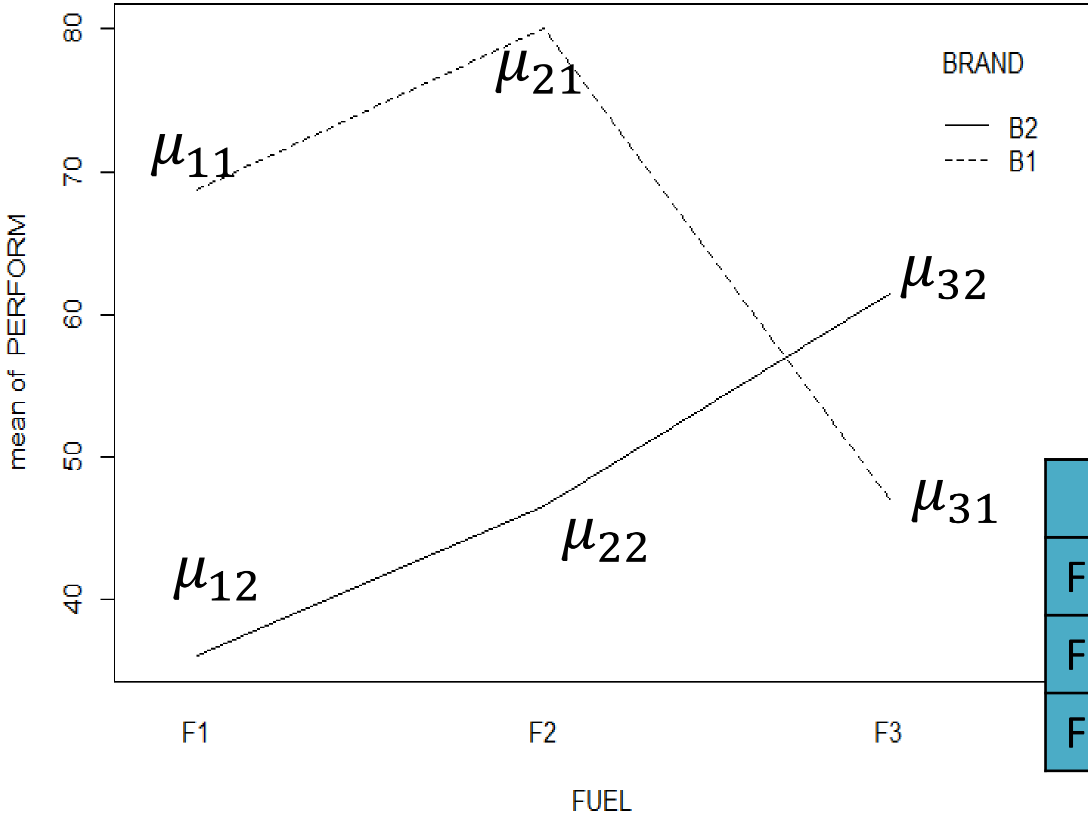
Change in mean performance between Brand baseline (EG Brand 1) and Brand 2, as we move from Fuel baseline (EG Fuel 1) to Fuel 2

Change in mean performance between Brand baseline (EG Brand 1) and Brand 2, as we move from Fuel baseline (EG Fuel 1) to Fuel 3

Interpreting the regression coefficients for the interaction

Table 5.7 The six combinations of fuel type and diesel engine brand

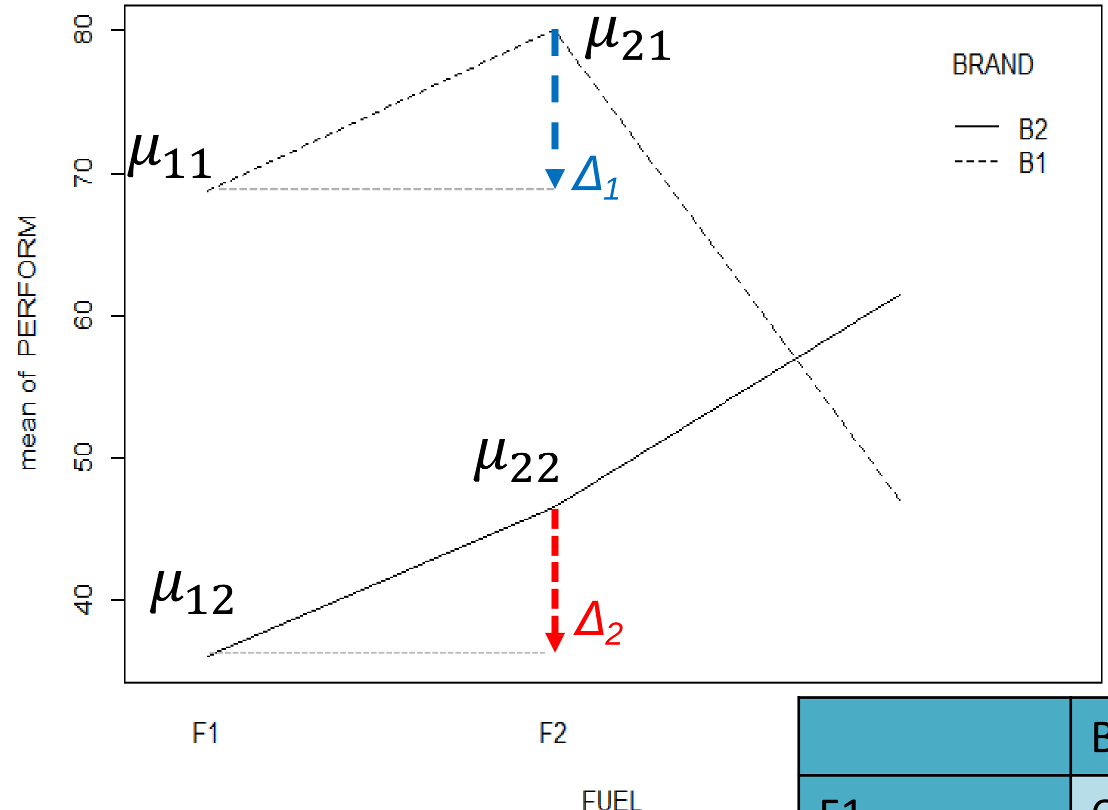
		Brand	
		B_1	B_2
FUEL TYPE	F_1	μ_{11}	μ_{12}
	F_2	μ_{21}	μ_{22}
	F_3	μ_{31}	μ_{32}



	B1	B2
F1	68.67	36
F2	80	46.5
F3	47	61.5

Interaction regression coefficients

$$\beta_4 = (\mu_{F2B2} - \mu_{F1B2}) - (\mu_{F2B1} - \mu_{F1B1})$$
$$= (\mu_{22} - \mu_{12}) - (\mu_{21} - \mu_{11}) = \Delta_2 - \Delta_1$$

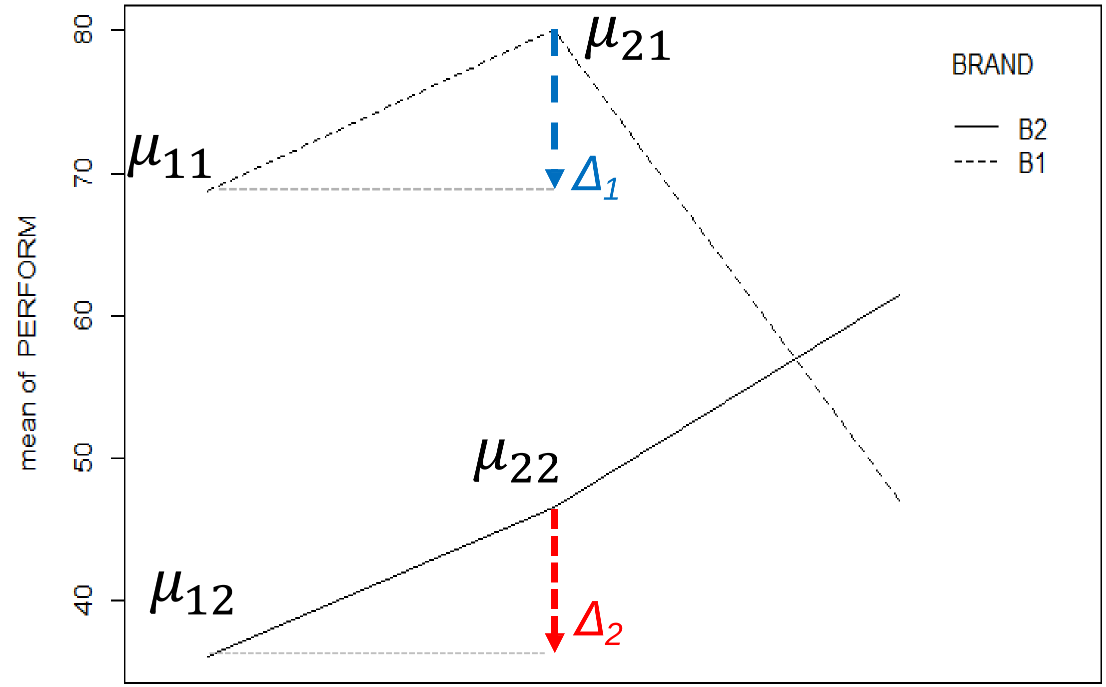


$$\Delta_2 \sim \Delta_1$$

	B1	B2
F1	68.67	36
F2	80	46.5
F3	47	61.5

Interaction regression coefficients

$$\beta_4 = (\mu_{F2B2} - \mu_{F1B2}) - (\mu_{F2B1} - \mu_{F1B1})$$
$$= (\mu_{22} - \mu_{12}) - (\mu_{21} - \mu_{11}) = \Delta_2 - \Delta_1$$



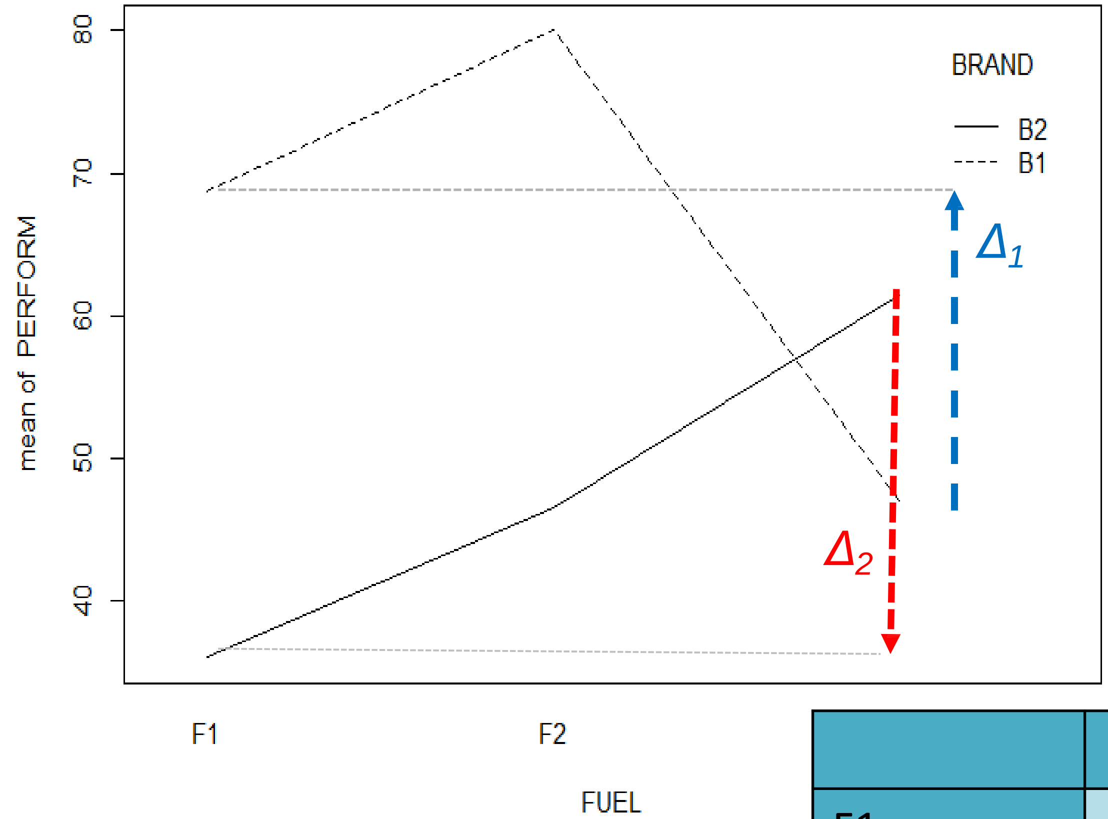
$\Delta_2 \sim \Delta_1$

The interpretation of β_4 (not sig., p = 0.88) is that the change in mean performance as we move from Fuel 1 to Fuel 2 is the _same_ for both brands - but that does *not* explain all of the interaction.

	$\beta_4 = \Delta_2 - \Delta_1$			p-value
FUELF2:BRANDB2	**-0.833**	**5.130**	**-0.16**	**0.87628**

Interaction regression coefficients

$$\beta_5 = (\mu_{F3B2} - \mu_{F1B2}) - (\mu_{F3B1} - \mu_{F1B1})$$
$$= (\mu_{32} - \mu_{12}) - (\mu_{31} - \mu_{11}) = \Delta_2 - \Delta_1$$

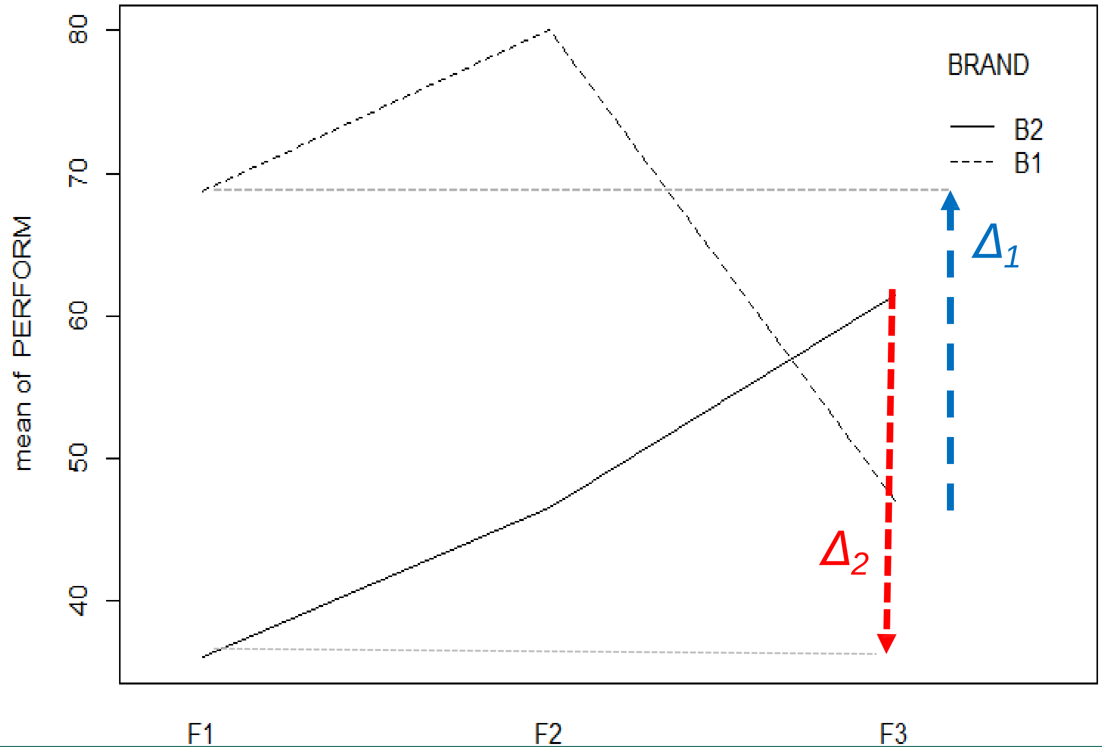


$$\Delta_2 \sim -\Delta_1$$

	B1	B2
F1	68.67	36
F2	80	46.5
F3	47	61.5

Interaction regression coefficients

$$\beta_5 = (\mu_{F3B2} - \mu_{F1B2}) - (\mu_{F3B1} - \mu_{F1B1})$$
$$= (\mu_{32} - \mu_{12}) - (\mu_{31} - \mu_{11}) = \Delta_2 - \Delta_1$$



$$\Delta_2 \sim -\Delta_1$$

The interpretation of β_5 (sig. p = 9.3×10^{-5}) is that the change in mean performance as we move from Fuel 1 to Fuel 3 is *not* the same for both brands. This results in the significant overall interaction in the anova table.

	$\beta_5 = \Delta_2 - \Delta_1$			p-value
FUELF3:BRANDB2	47.167	5.130	9.19	9.3e-05

Regression coefficients

Q: using the regression coefficients verify that the estimated mean performance for a **brand 2 engine, using fuel 2** will be 46.5

$$E(y) = \beta_0 + \overbrace{\beta_1 x_1 + \beta_2 x_2}^{\text{Main effect terms for } F} + \overbrace{\beta_3 x_3}^{\text{Main effect term for } B} + \overbrace{\beta_4 x_1 x_3 + \beta_5 x_2 x_3}^{\text{Interaction terms}}$$

Coefficients	
(Intercept)	68.667
FUELF2	11.333
FUELF3	-21.667
BRANDB2	-32.667
FUELF2:BRANDB2	-0.833
FUELF3:BRANDB2	47.167

$x_1 = 1$ (Fuel 2), $x_2 = 0$, $x_3 = 1$ (Brand2)

$$\mu_{22} = E(y)$$

$$= \beta_0 + \beta_1(1) + \beta_2(0) + \beta_3(1)$$

$$+\beta_4(1)(1) + \beta_5(0)(1)$$

$$\mu_{22} = \beta_0 + \beta_1 + \beta_3 + \beta_4$$

$$\mu_{22} = 68.67 + 11.33 - 32.67 - 0.83$$

$$= 46.5$$

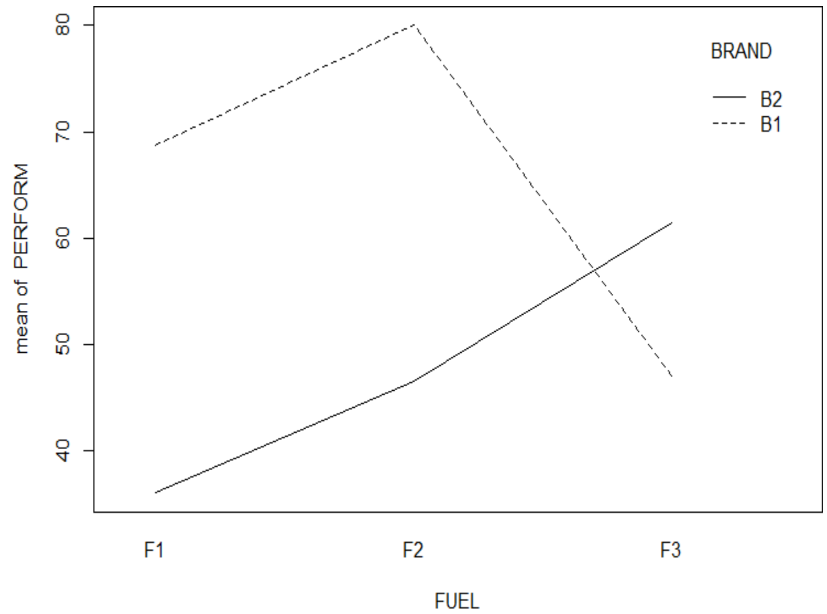
Figure 5.23 R printout for interaction model, Example 5.10

```
With(diesel.df, tapply(PERFORM,list(FUEL, BRAND), mean)
```

Table 5.7 The six combinations of fuel type and diesel engine brand

		Brand	
		B_1	B_2
FUEL TYPE	F_1	μ_{11}	μ_{12}
	F_2	μ_{21}	μ_{22}
	F_3	μ_{31}	μ_{32}

```
     B1   B2
F1 68.7 36.0
F2 80.0 46.5
F3 47.0 61.5
```



Models with 3 or more qualitative predictor

Models with ≥ 3 qualitative predictors

Pattern of the Model Relating $E(y)$ to k Qualitative Independent Variables

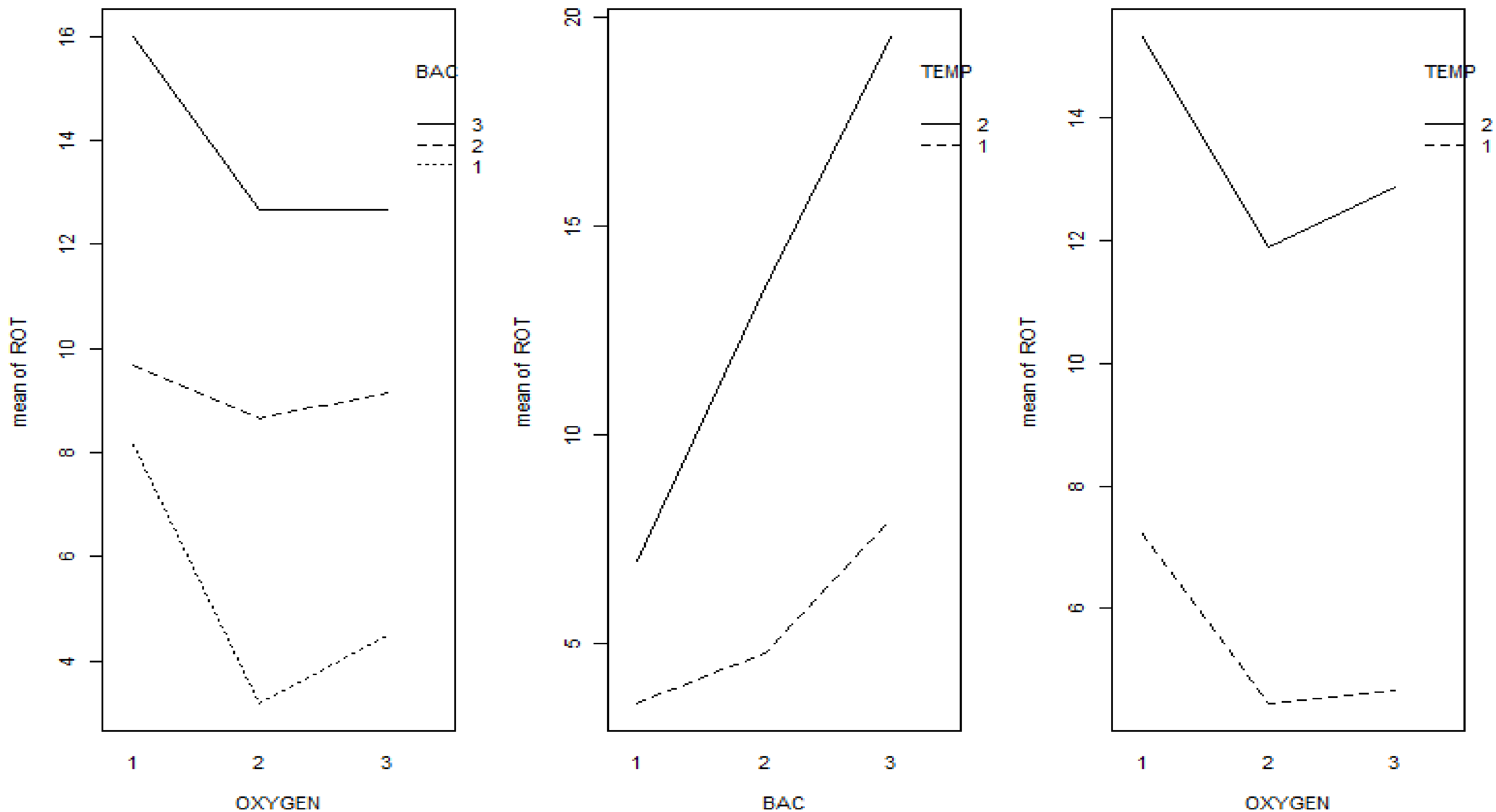
$$
\begin{aligned}
E(y) = \beta_0 &+ \text{Main effect terms for all independent variables} \\
&+ \text{All two-way interaction terms between pairs of independent variables} \\
&+ \text{All three-way interaction terms between different groups of three independent variables} \\
&+ \\
&\vdots \\
&+ \text{All } k\text{-way interaction terms for the } k \text{ independent variables}
\end{aligned}
$$

Example

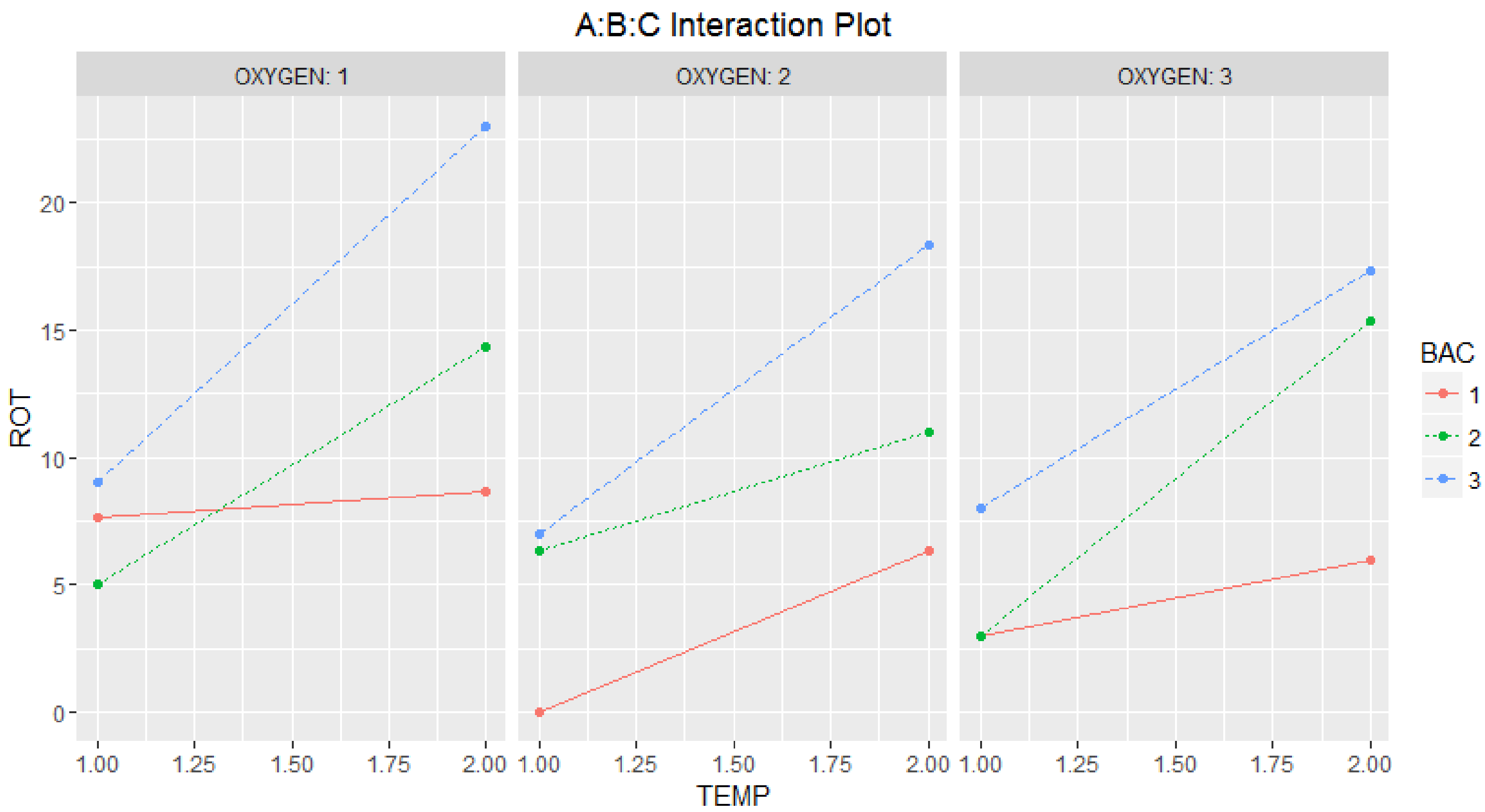
Potato farmers often experience problems with potatoes rotting while in storage. An experiment was conducted to find the conditions under which to keep potatoes to minimise the rate at which rotting occurs.

The variables were oxygen (OXYGEN: 3 levels), temperature (TEMP: 2 levels) and bacterial inoculation (BAC: 3 levels). There were 3 replicates of each treatment combination, completing an orthogonal factorial design.

Two-way Interaction Plots



Three-way Interaction Plots




```
library(dae)
interaction.ABC.plot(ROT,TEMP,BAC,OXYGEN,data=potrot)
```

Three-way model

Analysis of Variance Table

Response: ROT

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
OXYGEN	2	98	49	2.09	$\boxed{0.14}$
BAC	2	652	326	13.91	3.3e-05
TEMP	1	848	848	36.20	6.6e-07
OXYGEN:BAC	4	30	8	0.32	$\boxed{0.86}$
OXYGEN:TEMP	2	2	1	0.03	$\boxed{0.97}$
BAC:TEMP	2	153	76	3.26	0.05
OXYGEN:BAC:TEMP	4	81	20	0.87	0.49
Residuals	36	843	23		

Two-way interactions

```
mod2<-lm(ROT~BAC*TEMP, data=potrot)
anova(mod2)
```

```
Analysis of Variance Table

Response: ROT
          Df Sum Sq Mean Sq F value   Pr(>F)
BAC        2    652     326   14.84 9.6e-06
TEMP       1    848     848   38.61 1.2e-07
BAC:TEMP   2    153      76    3.48   0.039
Residuals 48   1054      22
```

Table of means

```
#Interaction means
tapply(ROT,INDEX = list(TEMP, BAC),  mean)
        1      2     3
1 3.556  4.778  8.00
2 7.000 13.556 19.56
```

The increase in rotting with an increase in bacteria is greater for temperature 2 (hence the 2-way interaction between BAC and TEMP)

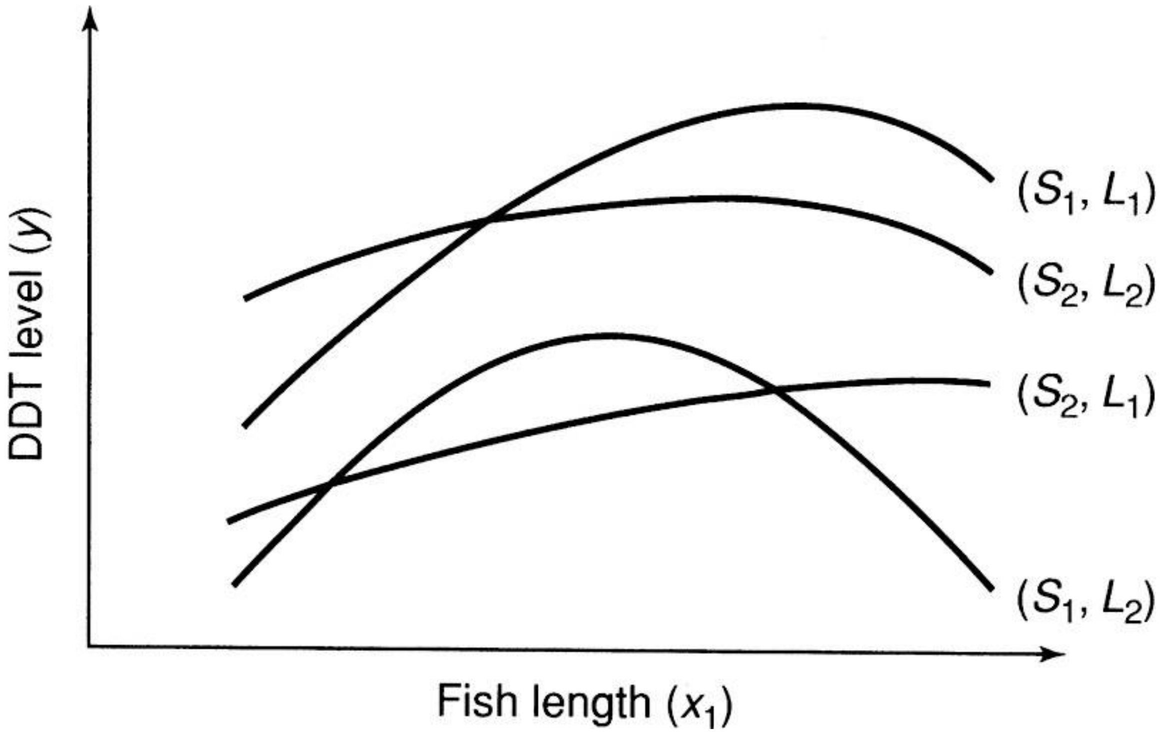
Models with both quantitative and qualitative predictor

Models with Both Quantitative and Qualitative Independent Variables

Example 5.14 p.299

- Response: level of contaminant DDT in fish
- Predictors:
  - – Fish length (Quantitative, cms): x_1
  - – Species (2 levels S_1, S_2) : x_2
  - – Location (2 levels L_1, L_2): x_3

Figure 5.29 Two qualitative (x_2, x_3) and one quantitative (x_1) predictor

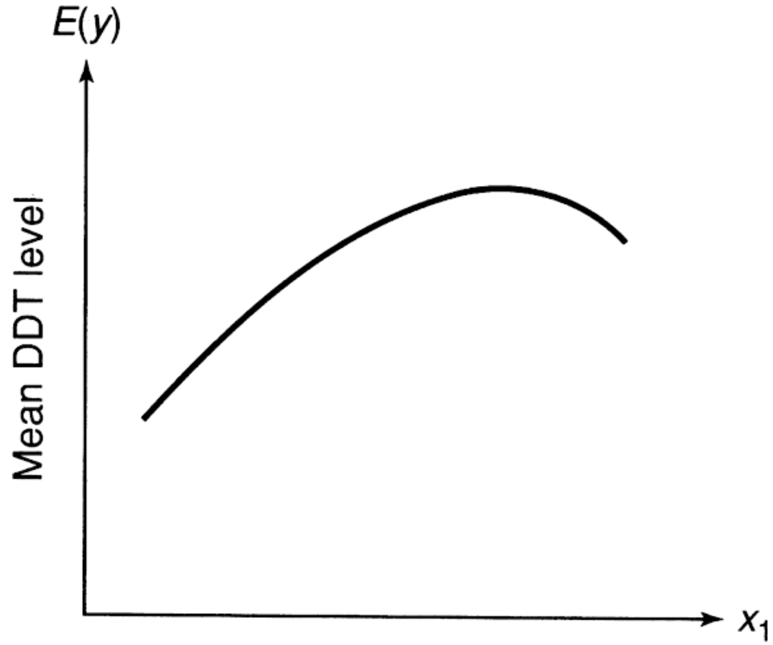


Q: Describe in general terms the association between DDT levels and the predictors

Figure 5.29 Modelling two qualitative (x_2, x_3) and one quantitative (x_1) predictor

$$E(y) = \boxed{\beta_0 + \beta_1 x_1 + \beta_2 x_1^2}$$

Stage 1:
Quantitative variable (x_1) first



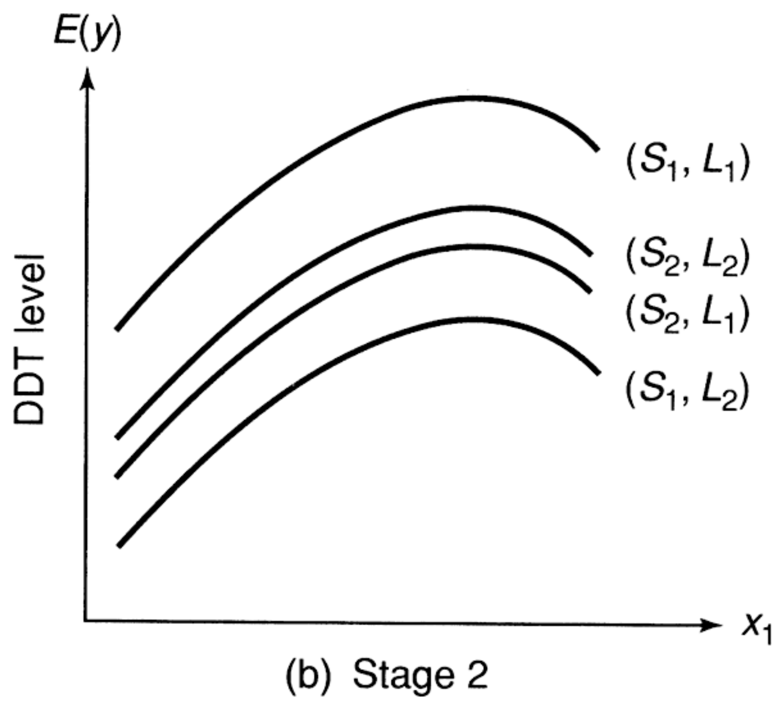
(a) Stage 1

Figure 5.29 Modelling two qualitative (x_2, x_3) and one quantitative (x_1) predictor

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \boxed{\beta_3 x_2 + \beta_4 x_3 + \beta_5 x_2 x_3}$$

Stage 1:
Quantitative variable (x_1) first

Stage 2:
Qualitative variables (x_2, x_3): main effects and interactions



x_2 =1 if species S_1, 0 otherwise
x_3 =1 if location L_1, 0 otherwise

These terms allow for differing intercepts

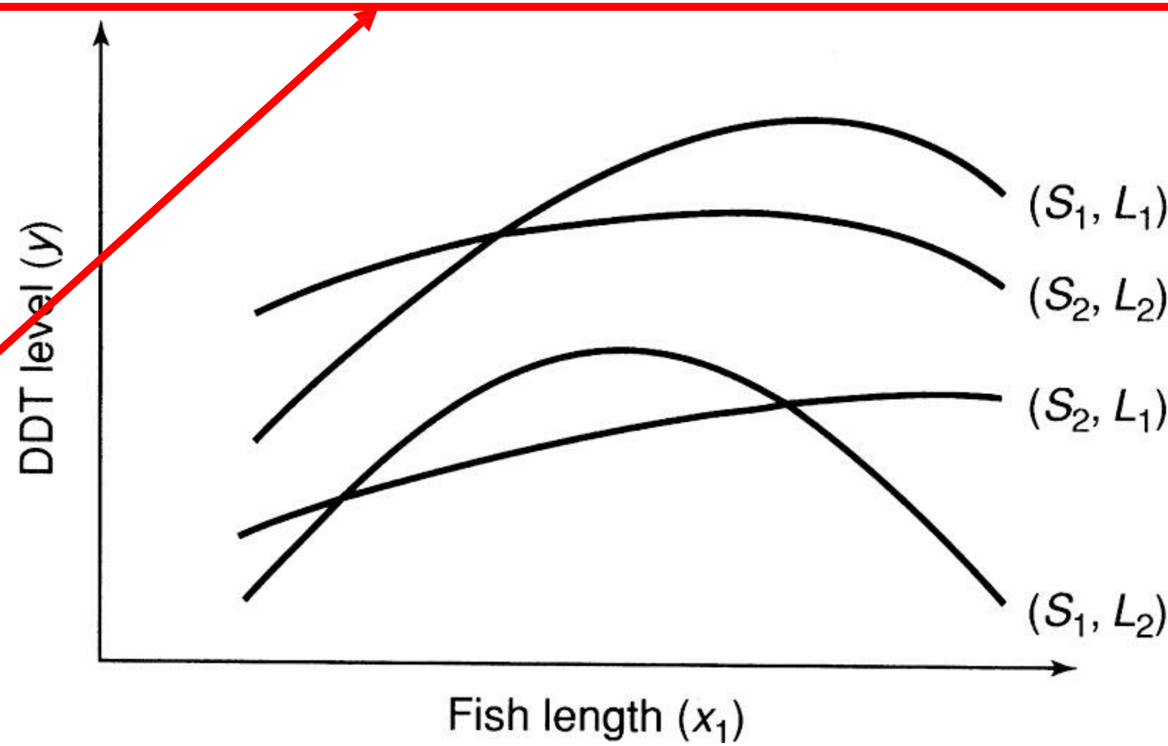
Figure 5.29 Modelling two qualitative (x_2, x_3) and one quantitative (x_1) predictor

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_3 + \beta_5 x_2 x_3$$
$$\boxed{+ \beta_6 x_1 x_2 + \beta_7 x_1 x_3 + \beta_8 x_1 x_2 x_3 + \beta_9 x_1^2 x_2 + \beta_{10} x_1^2 x_3 + \beta_{10} x_1^2 x_2 x_3}$$

Stage 1:
Quantitative variable (x_1) first

Stage 2:
Qualitative variables (x_2, x_3): main effects and interactions

Stage 3:
Interaction between quantitative (x_1, x_1^2) & qualitative variables (x_2, x_3)



These terms allow for shape of response curves to differ

Chapter 5 Recap

- **Models with 1 quantitative predictor**

  ➔ p^th^ – order polynomial : $\mathrm{E}(\boldsymbol{y}) = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 x} + \boldsymbol{\beta_2 x^2} + \cdots + \boldsymbol{\beta_p x^p}$

- **First - order models with ≥ 2 quantitative predictors**

$$\mathrm{E}(\boldsymbol{y}) = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 x_1} + \boldsymbol{\beta_2 x_2} + \cdots + \boldsymbol{\beta_k x_k}$$

- **Second - order models with ≥ 2 quantitative predictors**

Interaction: $\mathrm{E}(\boldsymbol{y}) = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 x_1} + \boldsymbol{\beta_2 x_2} + \boldsymbol{\beta_3 x_1 x_2}$

Complete: $\mathrm{E}(\boldsymbol{y}) = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 x_1} + \boldsymbol{\beta_2 x_2} + \boldsymbol{\beta_3 x_1 x_2} + \boldsymbol{\beta_4 x^2_{\ 1}} + \boldsymbol{\beta_5 x^2_{\ 2}}$

Chapter 5 Recap

❖ **Model with 1 qualitative predictor at k levels**

$$E(\boldsymbol{y}) = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 x_1} + \boldsymbol{\beta_2 x_2} + \cdots + \boldsymbol{\beta_{k-1} x_{k-1}}$$

$$x_i = \begin{cases} 1 \; if \; qualitative \; variable \; at \; level \; i+1 \\ 0 \; otherwise \end{cases}$$

❖ **Model with 2 qualitative predictors**

Without interaction:

$$E(y) = \beta_0 + \overbrace{\beta_1 x_1 + \beta_2 x_2}^{\substack{\text{Main effect} \\ \text{terms for } F}} + \overbrace{\beta_3 x_3}^{\substack{\text{Main effect} \\ \text{term for } B}}$$

With interaction:

$$E(y) = \beta_0 + \overbrace{\beta_1 x_1 + \beta_2 x_2}^{\substack{\text{Main effect} \\ \text{terms for } F}} + \overbrace{\beta_3 x_3}^{\substack{\text{Main effect} \\ \text{term for } B}} \overbrace{+ \beta_4 x_1 x_3 + \beta_5 x_2 x_3}^{\substack{\text{Interaction} \\ \text{terms}}}$$

❖ **Model with $\geq$ 3 qualitative predictors**

❖ **Models with both qualitative & quantitative predictors**