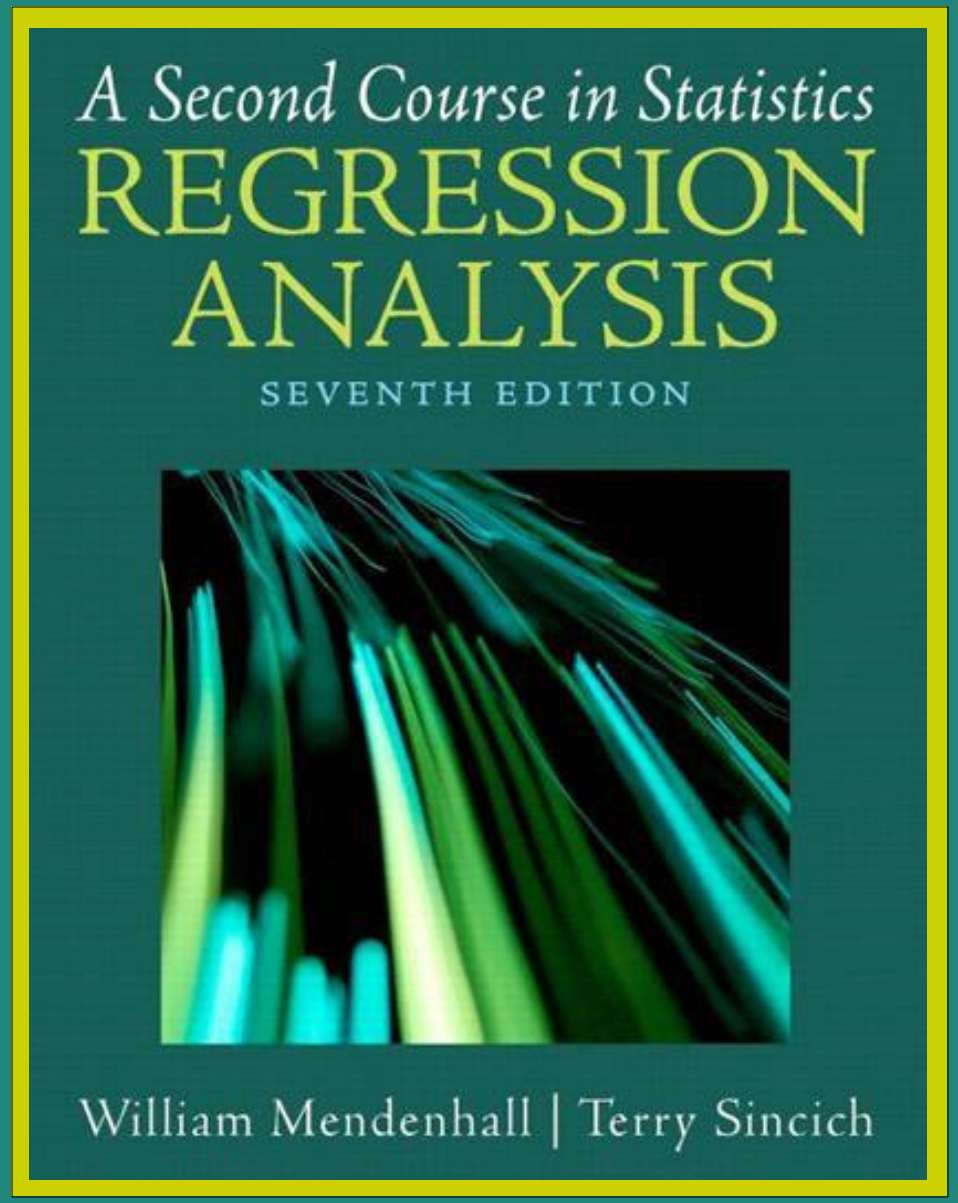
Chapter 6

Variable Screening Methods

(not §6.3)



PEARSON

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 6 Outline

- Parsimony in statistical modelling
- Multicollinearity
- Akaike's Information Criterion
- Stepwise regression only (not §6.3)
  - Forward selection
  - Backward selection

Parsimony in statistical modelling

"All models are wrong but some are useful"

George Box

A model "*should be as simple as possible but no simpler.*"

Albert Einstein

Parsimony in statistical modelling

- Models should have as few parameters as possible
- Linear models should be preferred to non-linear models
- Experiments relying on few assumptions are preferable to those relying on many
- Models should be pared down until they are minimally adequate
- Simple explanations should be preferred to complex ones.

Source: Crawley, M. (2007). The R Book, p. 325

Model fitting

Random SS: how varied the residuals are around the fitted component.
Smaller = more terms, how much smaller depends on usefulness of terms.

- Adding extra terms to the systematic part of a model will increase the systematic SS and decrease the random component
- **Total SS = Systematic SS + Random SS**

- In the extreme, we could include as many terms as there are data points so that the model would have no error (*overfitting)*.
- In other words, the model is too closely aligned with the sample and not representative of the population.

- The model is intended to explain both the systematic effects and the *natural random variation in the population*.

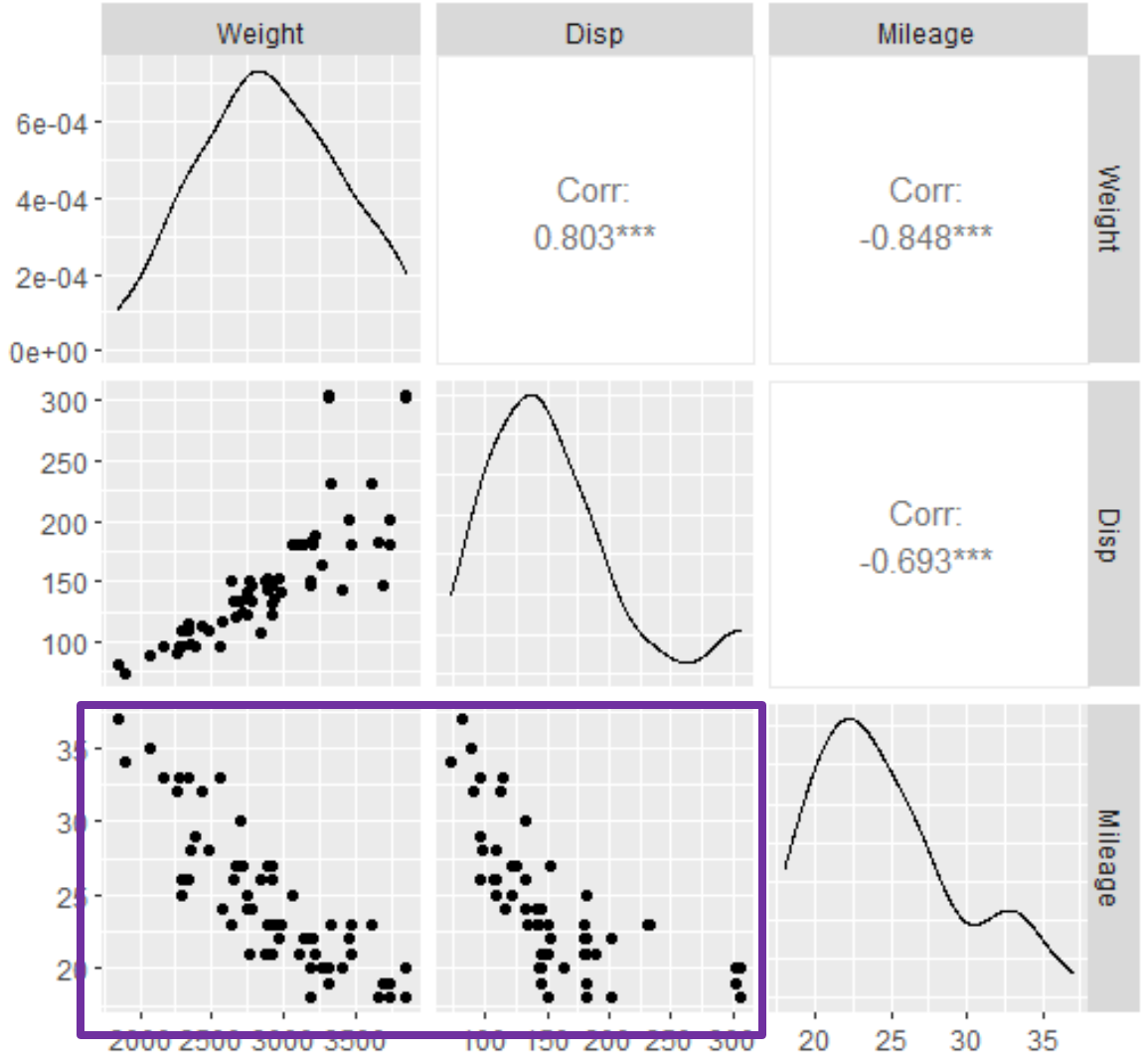
NOTE: We want to fit a model to represent the population *NOT* the data

Example:

Mileage per gallon of 60 cars of different weights and engine size (displacement)

Car	Weight (kg)	Disp	Mileage
1	2560	97	33
2	2345	114	33
3	1845	81	37
4	2260	91	32
:	:	:	:
20	2775	146	24

Exploratory Plot



Linear models

Simple Linear regression

$$\text{Mileage} = \beta_0 + \beta_1 \text{ Weight} + \epsilon$$

$$\text{Mileage} = \beta_0 + \beta_1 \text{ Disp} + \epsilon$$

Multiple regression

$$\text{Mileage} = \beta_0 + \beta_1 \, Weight + \beta_2 \text{ Disp} + \epsilon$$

What other types of models?

What other types of models?

Interaction

$$\text{Mileage} = \beta_0 + \beta_1 \text{Weight} + \beta_2 \text{Disp} + \beta_3 \text{Weight*Disp} + \epsilon$$

Complete second order

$$\text{Mileage} = \beta_0 + \beta_1 \text{Weight} + \beta_2 \text{Disp} + \beta_3 \text{Weight}^2 + \beta_4 \text{Disp}^2 + \beta_5 \text{Weight*Disp} + \epsilon$$

SLR v MLR

SLR1: Mileage ~ Weight
SLR2: Mileage ~ Disp
MLR: Mileage ~ Weight + Disp

	SLR1	SLR2	MLR
Intercept	48.3	34	48
Weight	-0.008 (-0.009,-0.007)		-0.008 (-0.01,-0.005)
Disp		-0.06 (-0.08,-0.04)	-0.003 (-0.02,0.02)
ESS	380	704	380

Error Sums of Squares: Unexplained variability

Q: Compare the CIs for the same coefficient under the different models

Q: What does a comparison of the ESS tell us?

SLR v MLR

SLR1: Mileage ~ Weight
SLR2: Mileage ~ Disp
MLR: Mileage ~ Weight + Disp

	SLR1	SLR2	MLR
Intercept	48.3	34	48
Weight	-0.008 (-0.009,-0.007)		-0.008 (-0.01,-0.005)
Disp		-0.06 (-0.08,-0.04)	-0.003 (-0.02,0.02)
ESS	380	704	380

Error Sums of Squares: Unexplained variability

Q: Compare the CIs for the same coefficient under the different models

Q: What does a comparison of the ESS tell us?

SLR v MLR

SLR1: Mileage ~ Weight
SLR2: Mileage ~ Disp
MLR: Mileage ~ Weight + Disp

	SLR1	SLR2	MLR
Intercept	48.3	34	48
Weight	-0.008 (-0.009,-0.007)		-0.008 (-0.01,-0.005)
Disp		-0.06 (-0.08,-0.04)	-0.003 (-0.02,0.02)
ESS	380	704	380

Error Sums of Squares: Unexplained variability

Q: Compare the CIs for the same coefficient under the different models

Q: What does a comparison of the ESS tell us?

SLR v MLR

SLR1: Mileage ~ Weight
SLR2: Mileage ~ Disp
MLR: Mileage ~ Weight + Disp

	SLR1	SLR2	MLR
Intercept	48.3	34	48
Weight	-0.008 (-0.009,-0.007)		-0.008 (-0.01,-0.005)
Disp		-0.06 (-0.08,-0.04)	-0.003 (-0.02,0.02)
ESS	380	704	380

Error Sums of Squares:
Unexplained variability

Q: Compare the CIs for the same coefficient under the different models

Q: What does a comparison of the ESS tell us?

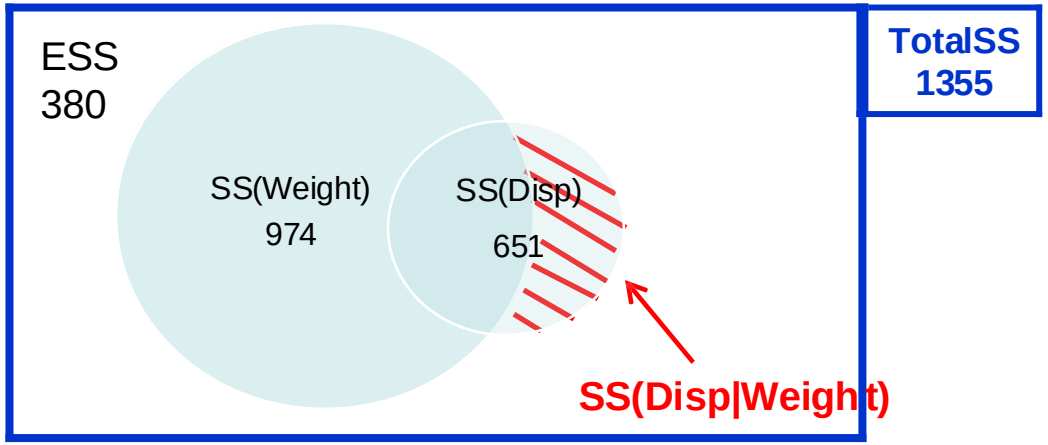
Issues

- In SLR2 Disp is significant in SLR
- In MLR
  - The CI for the coefficient of Disp does **not contain** the estimate for the coefficient of Disp in SLR2
  - Disp is **not** significant in MLR

Q. Why is there discrepancy?

Multicollinearity

- Disp and Weight are correlated – the information they contain about Mileage ***overlaps***



Akaike's Information Criterion

$$AIC = -2 * \log(Likelihood) + 2p\widehat{\sigma^2}$$

Maximise the likelihood (information available from the model)

or

Minimise the AIC

-2*log(likelihood)

can be thought of as the ESS for normal data

Stepwise Regression

1. Fit an initial model.
2. Fit all possible models that can be obtained by dropping (or adding) a term to the current model and compute the AIC for each.
3. Add (or remove) the term which ***reduces*** the AIC the most to form a new model.
4. Repeat the steps above until the AIC cannot be reduced any further at which time the process terminates.

Forward or Backward

Backward selection: starting with all parameters in the model and *removing* them.

Forward selection: starting with no parameters in the model and *adding* them.

Both based off AIC.

Stepwise regression: Step 1

```
# minimal model
formL <- formula(~ 1)
# maximum model
formU <- formula(~ Weight + Disp)

no.model <- lm(Mileage ~ 1,data=fuel.df)

fstep.model  <-  step(no.model,
     direction="forward",
     scope=list(lower=formL,upper=formU))
```

Stepwise regression: Steps 2&3

```
Start:  AIC=189.01
Mileage ~ 1

           Df  Sum of Sq     RSS     AIC
+ Weight    1      973.75   380.8     115
+ Disp      1      650.90   703.7     152
<none>
1355      189
```

Initial AIC =189.01
(no predictor variables)

Minimal AIC (115)
associated with fitting weight

Stepwise regression: Steps 3&4

```
Step:  AIC=115
Mileage ~ Weight
```

AIC =115 (Weight)

```
          Df   Sum of Sq   RSS      AIC
<none>                    381      115
+ Disp    1         0.6   380      117
```

AIC ***increases*** to 117 if we also include Disp

Including Weight ***minimizes*** the AIC, but including Disp does not.

Estimates of the regression coefficients & their confidence intervals:

Intercept: estimate of milage when weight = 0

```
confint(fstep.model)
```

```
                2.5 %     97.5 %
(Intercept)  44.38712  52.31157
Weight       -0.00954  -0.00685
```

Weight: estimate of the change in milage when weight increases by 1kg.

Q: What do the CIs indicate?

Example 2: pig carcass measurements

lean muscle %	LMpc
Fat depths	fdP2, fd34, fdH1, fdH2
muscle depths	mdP2, md34
weight	Wt

Data set is saved as *carc.txt*

	Wt	fdP2	fd34	fdH1	fdH2	mdP2	md34	LMpc
1	52.8	12.8	11.7	10.9	12.6	41.5	44.3	48.0
2	53.4	14.1	11.6	17.2	12.9	43.3	46.9	53.4
3	54.3	9.4	10.9	10.4	11.1	46.0	43.4	53.8
4	55.8	9.2	9.2	10.2	20.0	55.2	54.9	63.0
5	58.0	11.3	10.4	10.1	10.7	47.1	55.5	60.3
6	61.4	26.7	27.2	18.1	18.3	40.7	45.2	37.1
:								

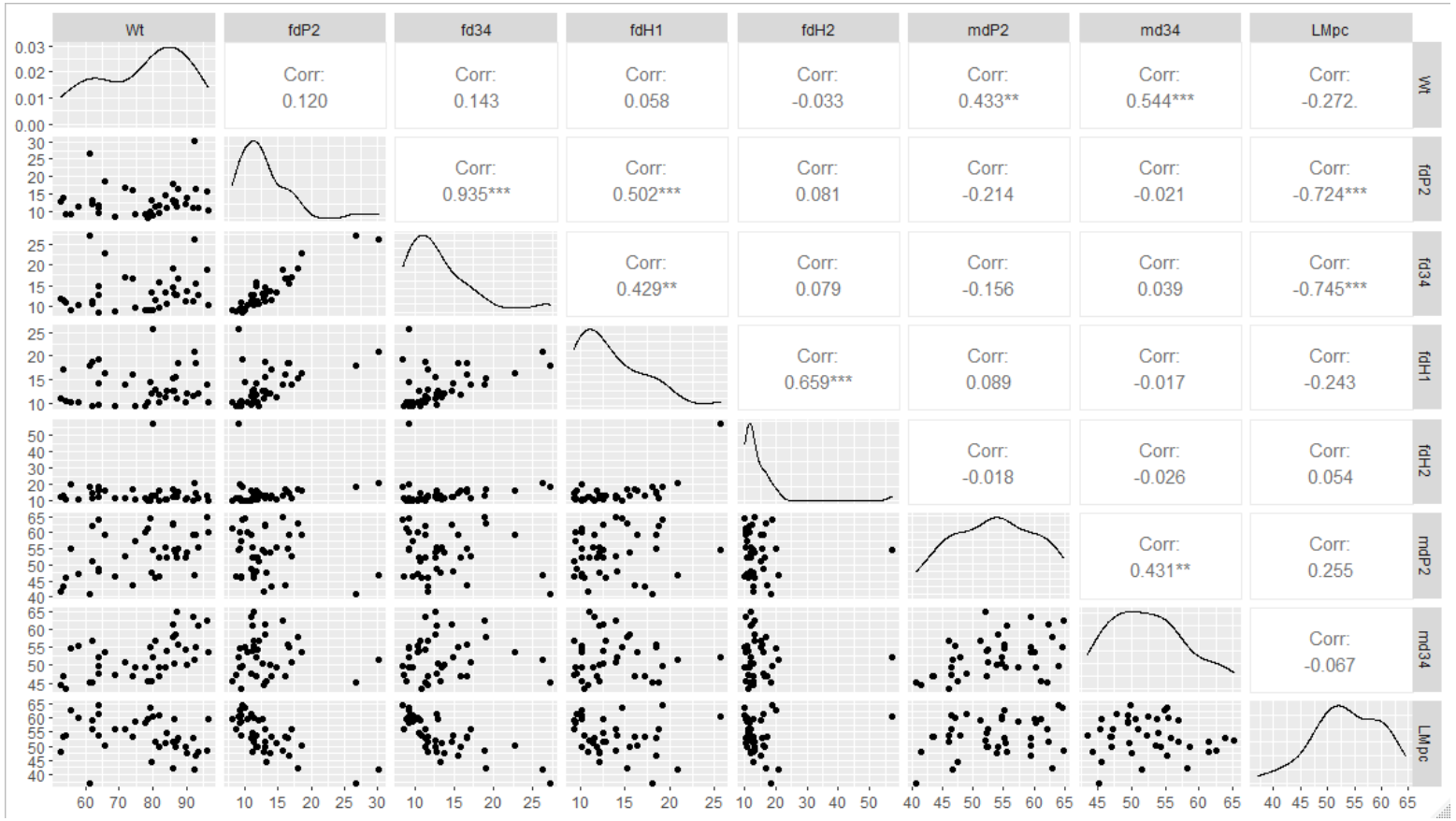
Pairs Plot of pig carcass measurements

Exercise: Interpret the pairs plot, and deduce a likely set of predictors.

- Biological reason for selecting certain variables, subsets of variables?
- Which subsets of predictors are collinear? Is it likely that all predictors in these subsets should be included in the model?
- Which predictors are most highly correlated with the response variable (LMpc)?
- Are there any predictors that should be included?

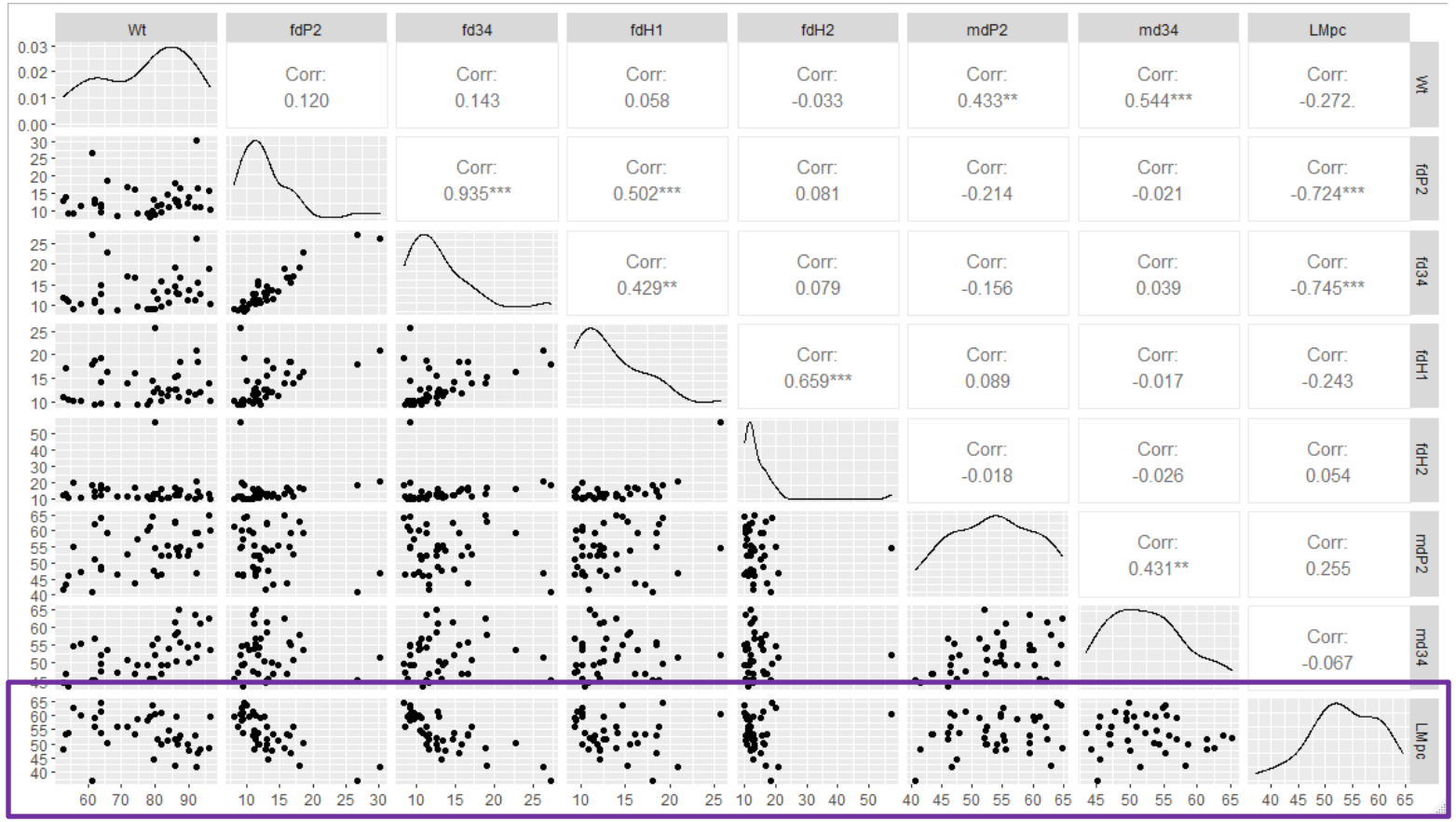
Pairs Plot of pig carcass measurements

```
pigs.df <- read.table("carc.txt", header=TRUE)
library(GGally)
ggpairs(pigs.df)
```



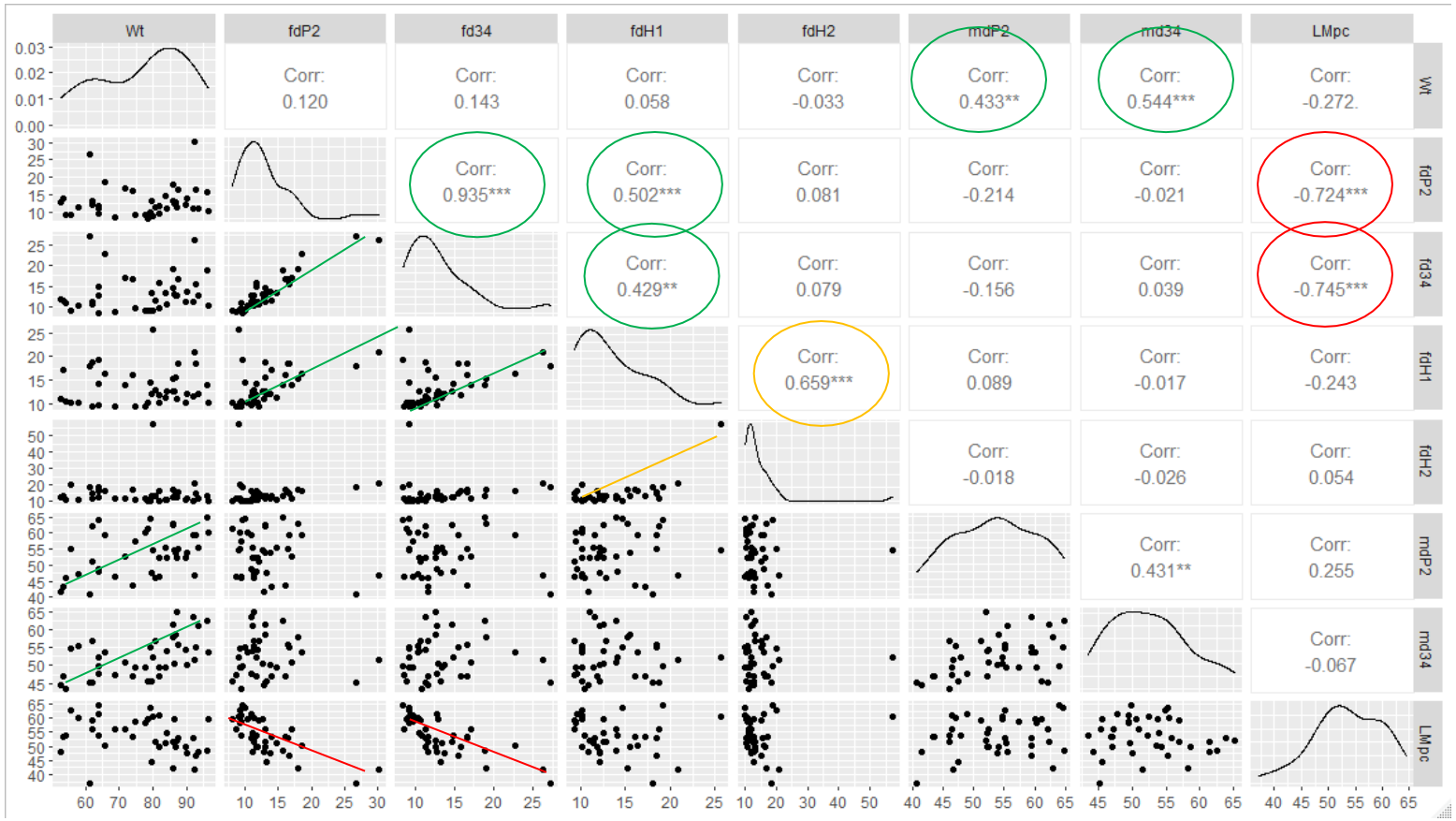
Pairs Plot of pig carcass measurements

```
pigs.df <- read.table("carc.txt", header=TRUE)
library(GGally)
ggpairs(pigs.df)
```



Pairs Plot of pig carcass measurements

```
pigs.df <- read.table("carc.txt", header=TRUE)
library(GGally)
ggpairs(pigs.df)
```



Forward selection

```
# Upper model allows for all variables
# to be included
formU <- formula(~ Wt+fdP2+fd34+fdH1+
                        fdH2 +mdP2 + md34)
formL <- formula(~ 1)

start.model <- lm(LMpc ~ 1,data=lmdf)

step.model <- step(start.model,
           direction="forward”,
         scope=list(upper=formU,lower=formL) )
summary(step.model)
```

First Step

```
Start:  AIC=149
LMpc ~ 1

          Df  Sum of Sq   RSS  AIC
+ fd34     1       879   704  119
+ fdP2     1       830   753  121
+ Wt       1       118  1465  148
+ mdP2     1       103  1480  148
+ fdH1     1        93  1489  149
<none>                  1583  149
+ md34     1         7  1576  151
+ fdH2     1         5  1578  151
```

Last Step

Q: Explain from output why final model only includes fd34+Wt+mdP2

Step: AIC=114
LMpc ~ fd34 + Wt + mdP2

	Df	Sum of Sq	RSS	AIC
<none>			564	114
+ fdH2	1	16.3	547	115
+ fdP2	1	2.0	562	116
+ fdH1	1	2.0	562	116
+ md34	1	0.0	564	116

Estimates & CIs for regression coefficients

```
#CI for regression coefficients
confint(step.model)

             Estimate     Std. Error      2.5 %  97.5 %
(Intercept)   63.1          5.8         (51.30, 74.90)
fd34          -0.93         0.15        (-1.20, -0.64)
Wt            -0.15         0.06        (-0.26, -0.03)
mdP2           0.265        0.11        ( 0.05,  0.48)
```

Backward selection

```
# Upper model allows for all variables
# to be included

 start.model <- lm(LMpc ~ Wt + fdP2 + fd34
    + fdH1 + fdH2 + mdP2 + md34,data=lmdf)

bstep.model <- step(start.model)
summary(bstep.model)
```

For this example the final model is the same.

NB: results will *not* always be the same using forward and backward selection – see Prac 4

Stepwise model selection: Cautionary Note

These procedures are easy to run but you should be aware of the following drawbacks:

- Because of the one-at-a-time nature of adding/ dropping variables, it is possible to miss the “optimal model”.
- Use the AIC as your guide. The p-values should not be taken literally and may not be valid due to the multiple comparisons.
- Keep in mind the purpose of the investigation - prediction &/or explanation. The stepwise procedures are not directly linked to these objectives.
- For a range of reasons you may wish/ need to force retention of certain terms.

Faraway, J. (2005) *Linear Models with R*, Chapman & Hall, p.23

Example: Concrete strength

Concrete.csv

Variables

A study has been undertaken to explore what gives concrete its strength. There are 4 variables in this dataset:

Cement: weight (in kg) contained in a m3 mixture

Water: weight (in kg) contained in a m3 mixture

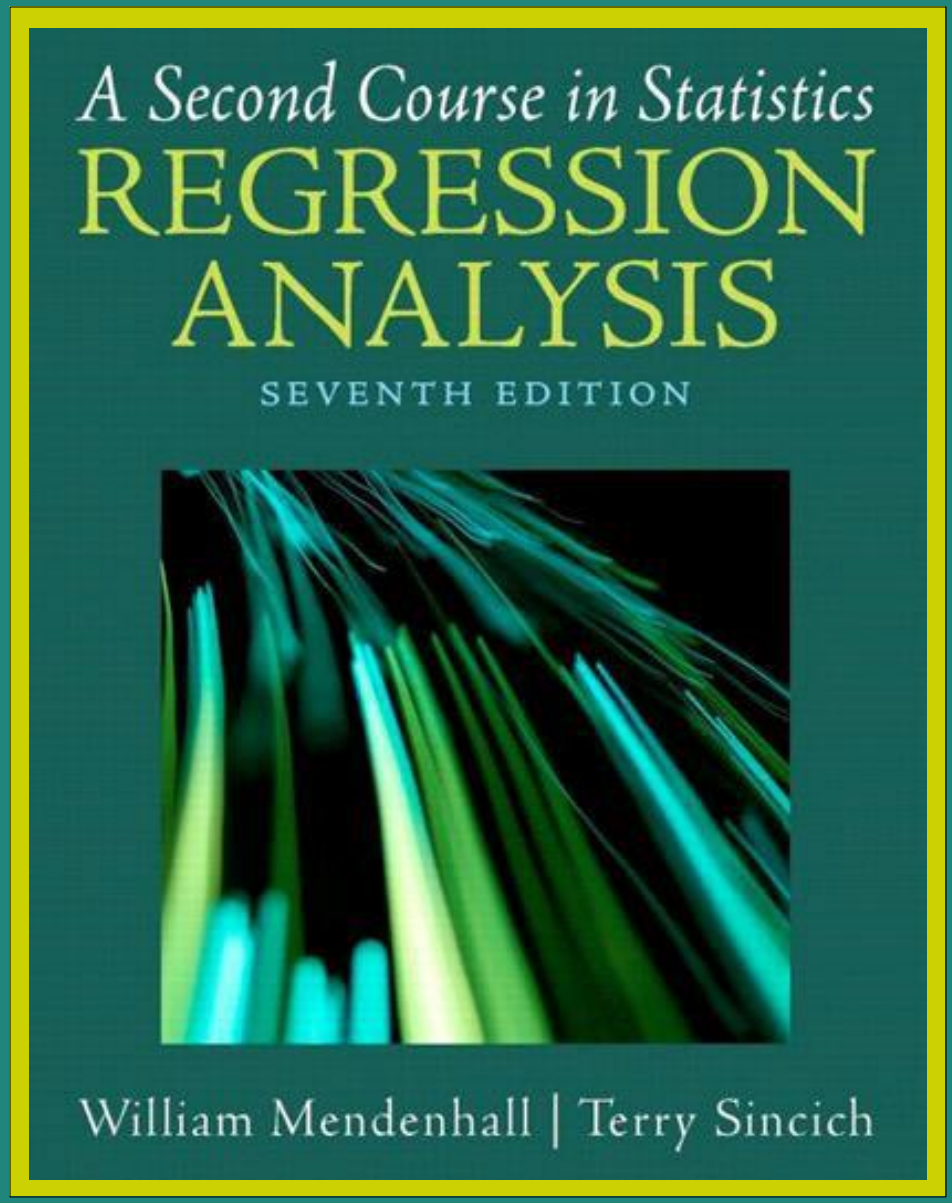
Course.aggregate: amount of course aggregate (in kg) in 1m3 mixture

Strength: compressive strength of concrete, measured in MPa (response variable)

Chapter 7

Some Regression Pitfalls

(not §7.6)



PEARSON

Chapter 7 outline

- Describe the difference between observational studies and designed experiments
- Identify possible limitations/ pitfalls of analysis:
  - Parameter estimability and interpretation
  - Multicollinearity
  - Extrapolation

Choosing types of models

What can you fit with the data you have?

Observational Study vs Designed Experiment

- Values of independent variables controlled?
- Random allocation of treatments?
  - Cause and effect?

Figure 7.1 Creativity and flexibility data for three children

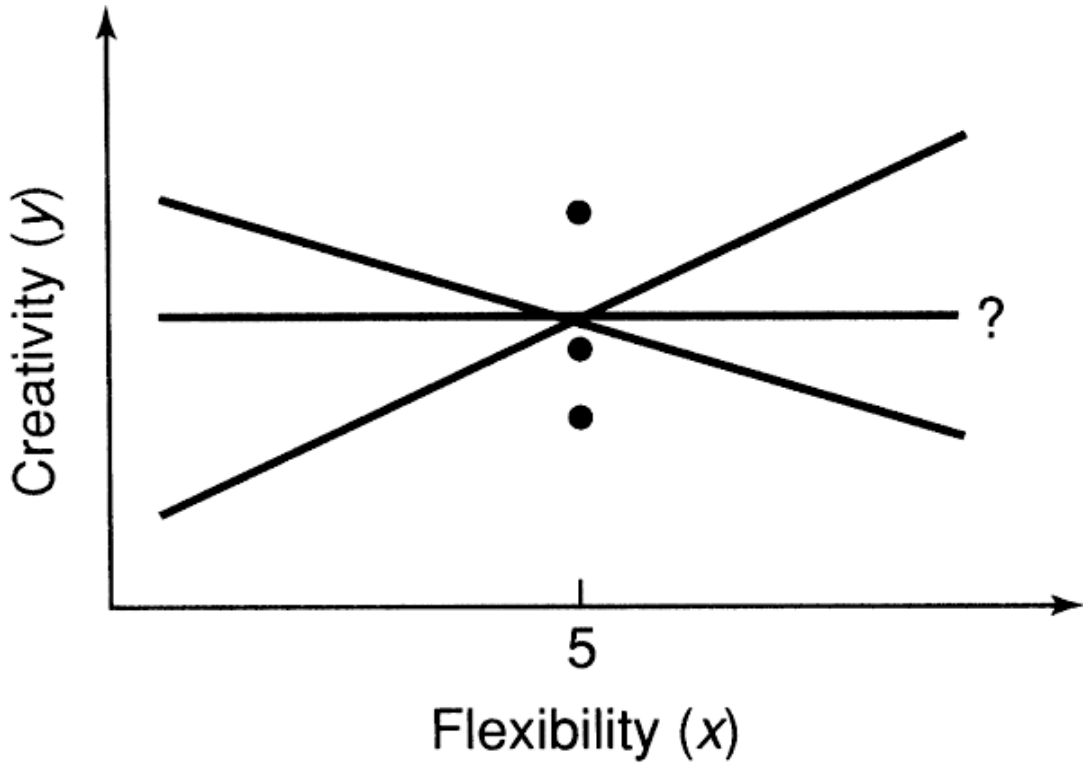
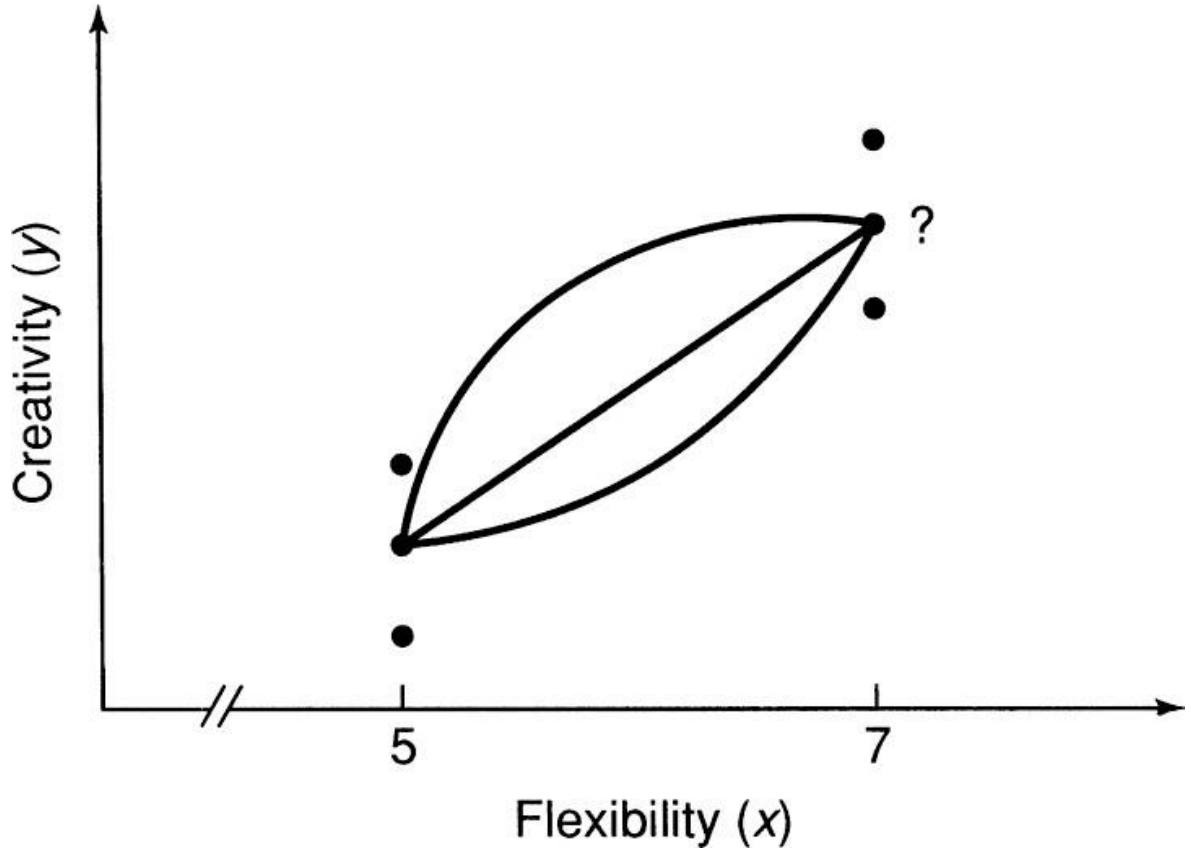


Figure 7.2 Only two different *x*-values observed - the second-order model is not estimable



Requirements for Fitting a *p*th-Order Polynomial Regression Model

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

1. The number of levels of x must be greater than or equal to $(p + 1)$.
2. The sample size n must be greater than $(p + 1)$ to allow sufficient degrees of freedom for estimating σ^2.

p=k-1

Fuel and engine brand revisited

Three levels of FUEL (1, 2, 3) and

Two engine BRANDS (1, 2)

Interaction model:

$$E(y) = \beta_0 + \beta_1 \text{FUEL2} + \beta_2 \text{FUEL3} + \beta_3 \text{BRAND2}$$
$$+ \beta_4 \text{FUEL2:BRAND2} + \beta_5 \text{FUEL3:BRAND2}$$

Table 7.1 Performance data for Example 7.3

		Brand	
		B_1	B_2
Fuel Type	F_1	73, 68	
	F_2	78, 82	50, 43
	F_3		61, 62

Need all combinations

R output for interaction model, Example 7.3

Note interaction term does *not* appear in ANOVA.

```
Response: perf
          Df Sum Sq Mean Sq F value   Pr(>F)
FUEL       2     96      48    4.23 0.10322
BRAND      1   1122    1122   98.66 0.00058
Residuals  4     45      11

Coefficients: (2 not defined because of singularities)
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)         70.50       2.38   29.56  7.8e-06
FUELF2               9.50       3.37    2.82  0.04799
FUELF3              24.50       4.77    5.14  0.00681
BRANDB2            -33.50       3.37   -9.93  0.00058
FUELF2:BRANDB2         NA         NA      NA       NA
FUELF3:BRANDB2         NA         NA      NA       NA
```

Interaction

Can’t infer that there is no interaction, only that we have insufficient data to be able to test for an interaction.

Multicollinearity

Relationships between predictor variables

Detecting Multicollinearity

Detecting Multicollinearity in the Regression Model

ggpairs plot

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

The following are indicators of multicollinearity:

1. Significant correlations between pairs of independent variables in the model
2. Nonsignificant t-tests for all (or nearly all) the individual β parameters when the F-test for overall model adequacy H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ is significant
3. Opposite signs (from what is expected) in the estimated parameters
4. A variance inflation factor (VIF) for a β parameter greater than 10, where

$$(\text{VIF})_i = \frac{1}{1 - R_i^2}, \quad i = 1, 2, \ldots, k$$

and R_i^2 is the multiple coefficient of determination for the model

$$E(x_i) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_{i-1} x_{i-1} + \alpha_{i+1} x_{i+1} + \cdots + \alpha_k x_k$$

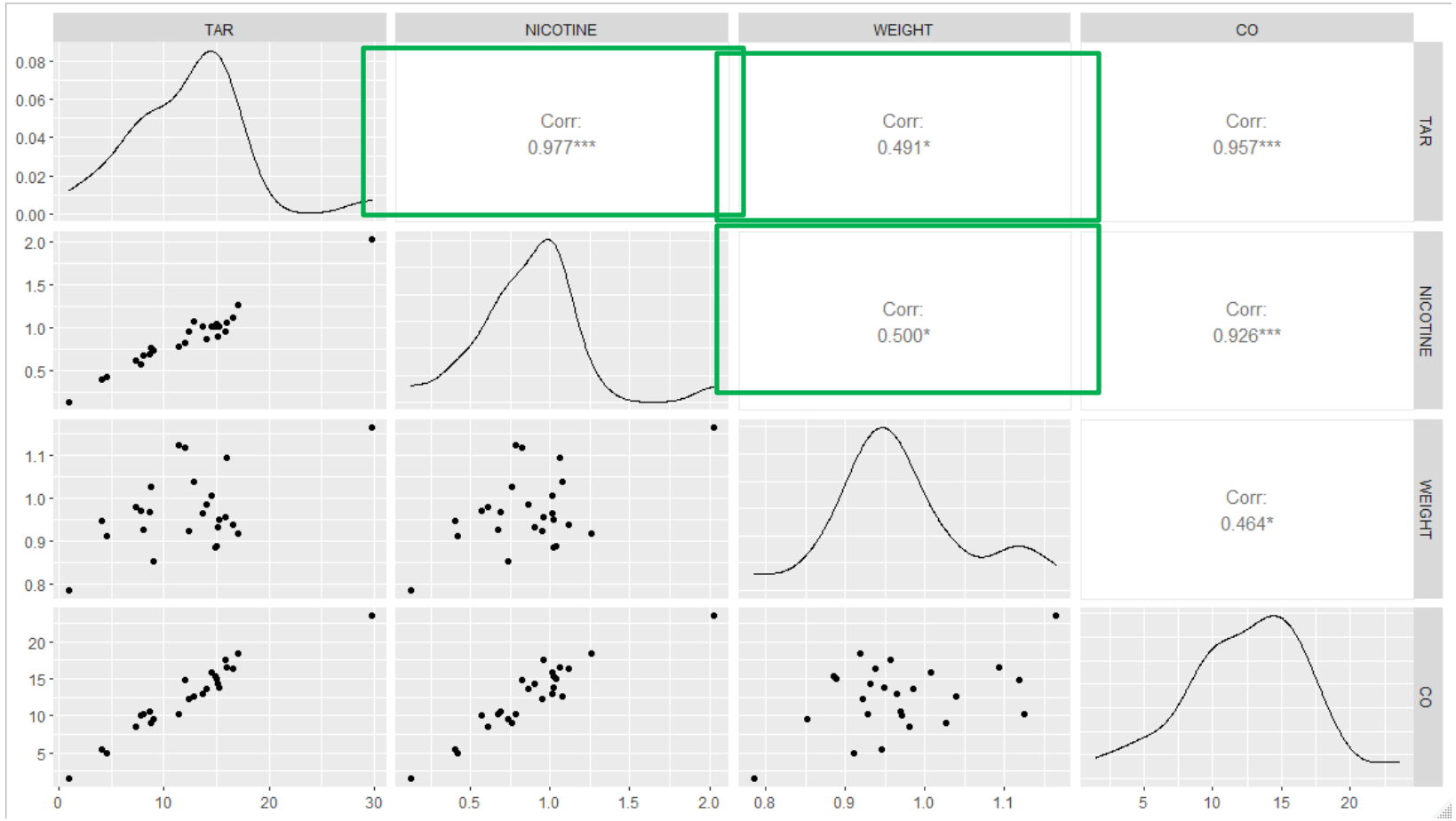
Table 7.2 FTC cigarette data for Example 7.5

Brand	Tar x_1, milligrams	Nicotine x_2, milligrams	Weight x_3, grams	Carbon Monoxide y, milligrams
Alpine	14.1	.86	.9853	13.6
Benson & Hedges	16.0	1.06	1.0938	16.6
Bull Durham	29.8	2.03	1.1650	23.5
Camel Lights	8.0	.67	.9280	10.2
Carlton	4.1	.40	.9462	5.4
Chesterfield	15.0	1.04	.8885	15.0
Golden Lights	8.8	.76	1.0267	9.0
Kent	12.4	.95	.9225	12.3
Kool	16.6	1.12	.9372	16.3
L&M	14.9	1.02	.8858	15.4
Lark Lights	13.7	1.01	.9643	13.0
Marlboro	15.1	.90	.9316	14.4
Merit	7.8	.57	.9705	10.0
Multifilter	11.4	.78	1.1240	10.2
Newport Lights	9.0	.74	.8517	9.5
Now	1.0	.13	.7851	1.5
Old Gold	17.0	1.26	.9186	18.5
Pall Mall Light	12.8	1.08	1.0395	12.6
Raleigh	15.8	.96	.9573	17.5
Salem Ultra	4.5	.42	.9106	4.9
Tareyton	14.5	1.01	1.0070	15.9
True	7.3	.61	.9806	8.5
Viceroy Rich Lights	8.6	.69	.9693	10.6
Virginia Slims	15.2	1.02	.9496	13.9
Winston Lights	12.0	.82	1.1184	14.9

Source: Federal Trade Commission.

Pairs plot and correlation

```
cigar.df <- read.table("FTCCIGAR.txt", header=TRUE)
ggpairs(cigar.df)
```



2. Checking for non-significant regression coefficients

```
mod<-lm(CO~TAR+NICOTINE+WEIGHT, data=ftc)
summary(mod)
```

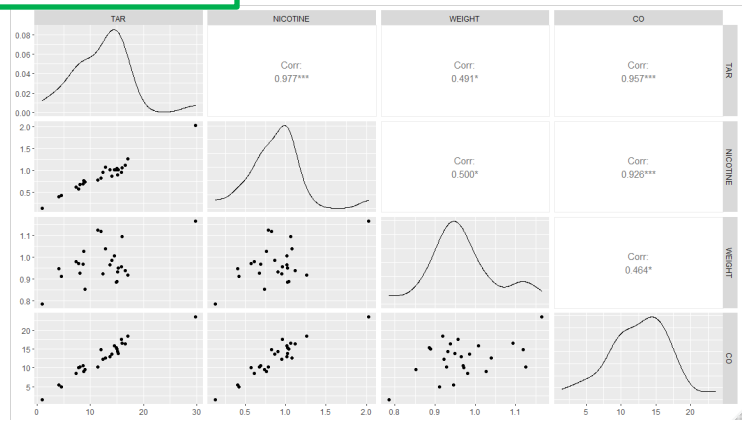
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.202      3.462    0.93  0.36546
TAR            0.963      0.242    3.97  0.00069
NICOTINE      -2.632      3.901   -0.67  0.50723
WEIGHT        -0.130      3.885   -0.03  0.97353

Residual standard error: 1.45 on 21 degrees of freedom
Multiple R-squared:  0.919,    Adjusted R-squared:  0.907
F-statistic:   79 on 3 and 21 DF,  p-value: 1.33e-11
```

3. Checking for opposite signs than expected for regression parameters

```
mod<-lm(CO~TAR+NICOTINE+WEIGHT, data=ftc)
summary(mod)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.202      3.462    0.93  0.36546
TAR             0.963      0.242    3.97  0.00069
NICOTINE       -2.632      3.901   -0.67  0.50723
WEIGHT         -0.130      3.885   -0.03  0.97353
```



4. VIF (Variance Inflation factor)

$VIF_i = 1/(1-R_i^2) > 10?$

```
library(car)
vif(mod)
```

```
     TAR   NICOTINE     WEIGHT
   21.63      21.90       1.33
```

What does this mean?

Problems when interpreting the coefficients:

- High correlation means we can't keep variables constant
- hard to determine how the independent variables influence the dependent variable individually
- Less reliable statistical inferences
- Potentially large standard errors

Multicollinearity – what to do?

- Model selection with only a subset of predictors
- If, for some reason, you decide to keep all predictors in the model, avoid making inference about individual coefficients
- Avoid extrapolation

Multicollinearity – what to do?

- Model selection with only a subset of predictors

TAR	NICOTINE	WEIGHT
21.63	21.90	1.33

Multicollinearity – what to do?

```
Coefficients:
             Estimate   Std. Error    t value   Pr(>|t|)
(Intercept)   3.11433    3.41620       0.912     0.372
Tar           0.80419    0.05904      13.622    3.36e-12
Weight       -0.42287    3.81299      -0.111     0.913

Residual standard error: 1.428 on 22 degrees of freedom
Multiple R-squared:  0.9168,  Adjusted R-squared:  0.9093
F-statistic: 121.3 on 2 and 22 DF,  p-value: 1.318e-12
```

Tar	Weight
1.317264	1.317264

Extrapolation

Predicting outside the observed data

Figure 7.7 Using a regression model outside the experimental region

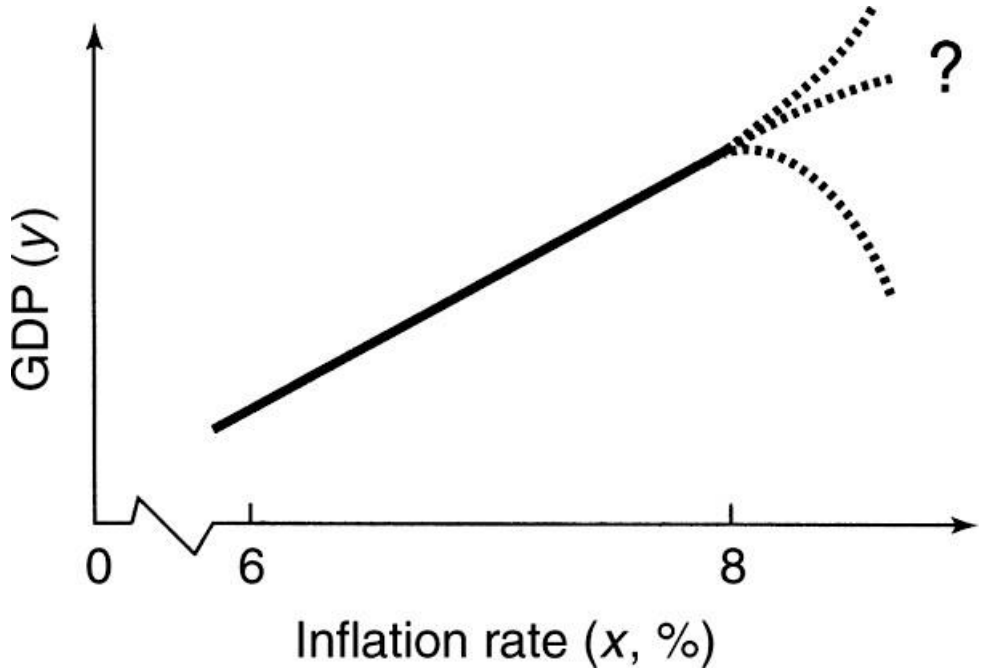


Figure 7.8 Experimental region for modeling GDP (y) as a function of inflation rate (x_1) and prime interest rate (x_2)

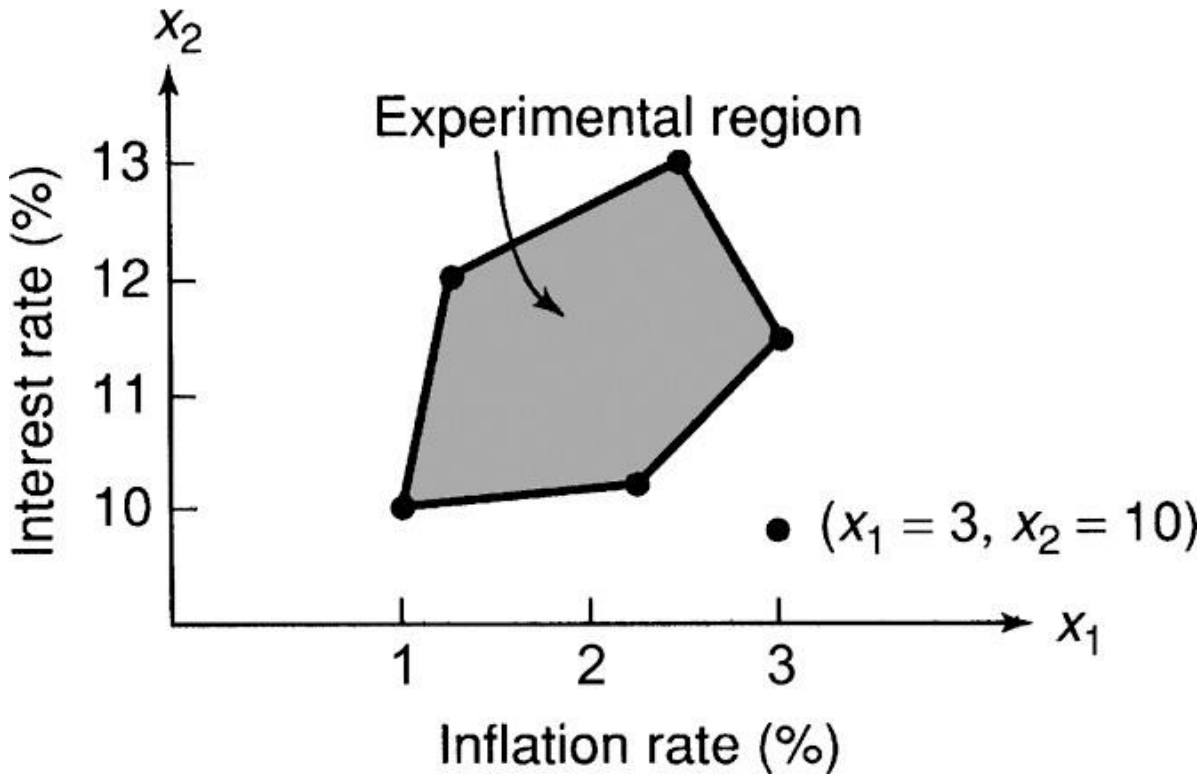


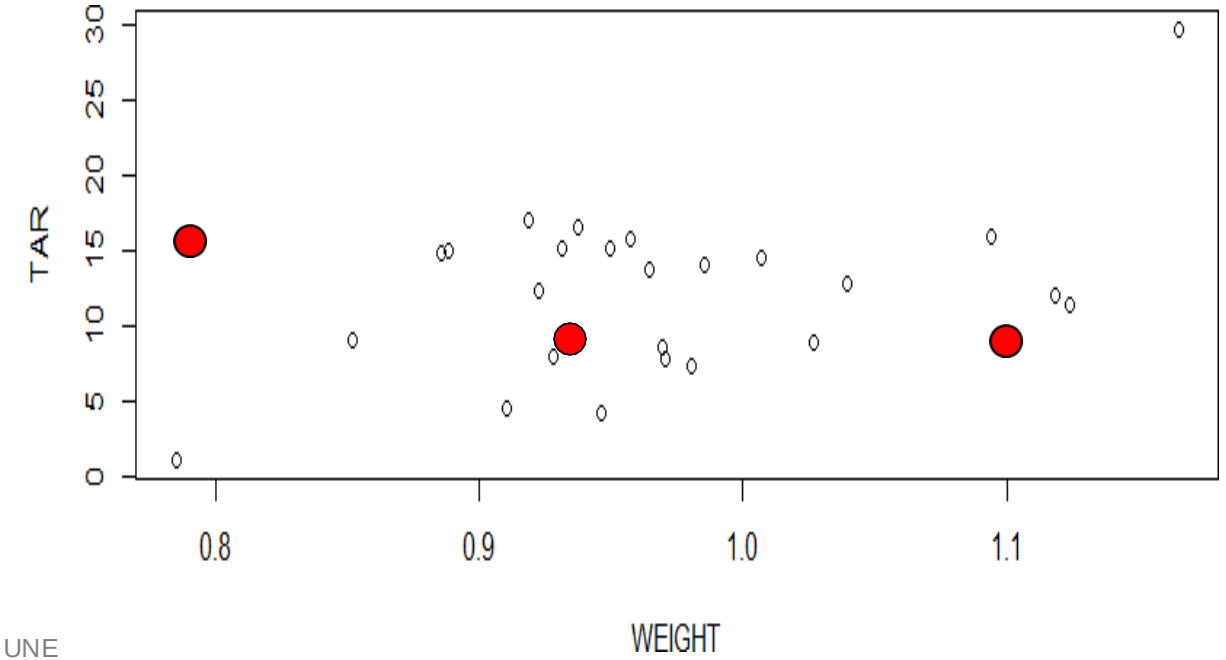
Figure 7.9 Descriptive statistics for independent variables

```
summary(ftc[1:3])
      TAR              NICOTINE          WEIGHT      
 Min.   : 1.0   Min.   :0.130   Min.   :0.785  
 1st Qu.: 8.6   1st Qu.:0.690   1st Qu.:0.922  
 Median :12.8   Median :0.900   Median :0.957  
 Mean   :12.2   Mean   :0.876   Mean   :0.970  
 3rd Qu.:15.1   3rd Qu.:1.020   3rd Qu.:1.007  
 Max.   :29.8   Max.   :2.030   Max.   :1.165  
```

Extrapolation

Which of the following would constitute extrapolation when predicting from a model with TAR and WEIGHT as predictors?

A. WEIGHT= 0.78, TAR = 15
B. WEIGHT= 0.93, TAR = 5
C. WEIGHT= 1.1, TAR = 5



Example: mussel weight

Mussel.csv

Variables

Estimating mussel flesh weight based on shell morphology. The data set has six variables:

- **WT:** weight (g) (response)
- **UW:** Umbo width (mm)
- **HL:** Hinge length (mm)
- **TL:** total length (mm)
- **TD:** total depth (mm)
- **TW:** total width (mm)