Chapter 4

Multiple Linear Regression

Lecture 3

University of New England

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 4 Outline

Lecture 1

- Intro to MLR
- Fitting the model, testing the overall utility of a model
- Interpreting regression coefficients

Lecture 2

- Inferences about the individual β_i
- Multiple Coefficients of determination, R^2 and R^2_{adj}
- Using the model for estimation and prediction

Lecture 3

- An interaction model with quantitative predictors

Lecture 4

- Models with qualitative predictors

NB: Sections 4.11, 4.13 and 4.14 of the text will ***not*** be covered

Reminder: Multiple Linear Regression Equation

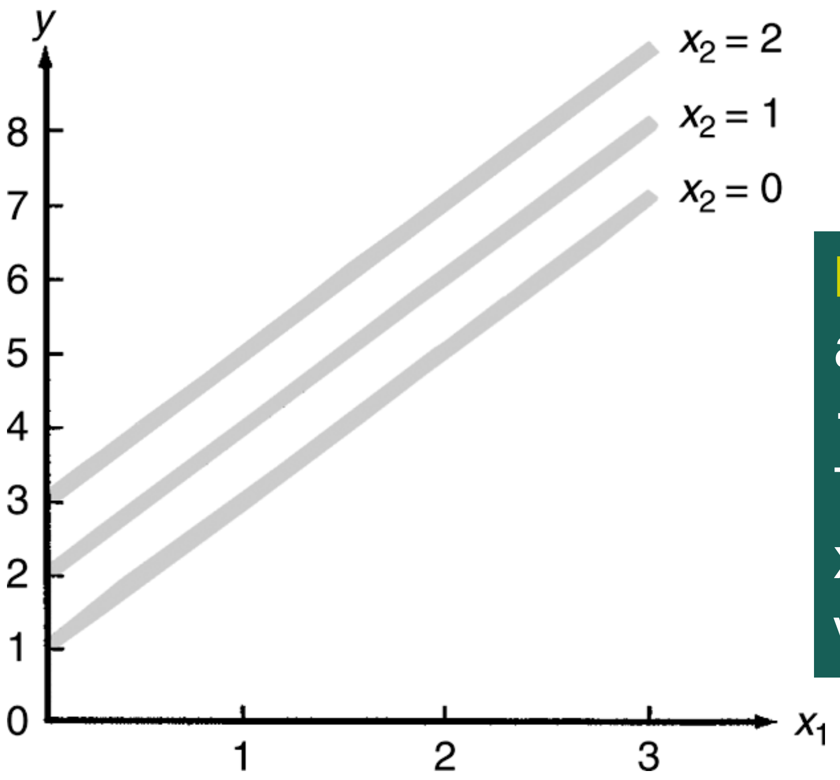
$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

When there is no interaction:

- A positive β coefficient indicates that the predictor variable increases the response, while the other predictors are constant.

- A negative β coefficient indicates that the predictor variable decreases the response, while the other predictors are constant.

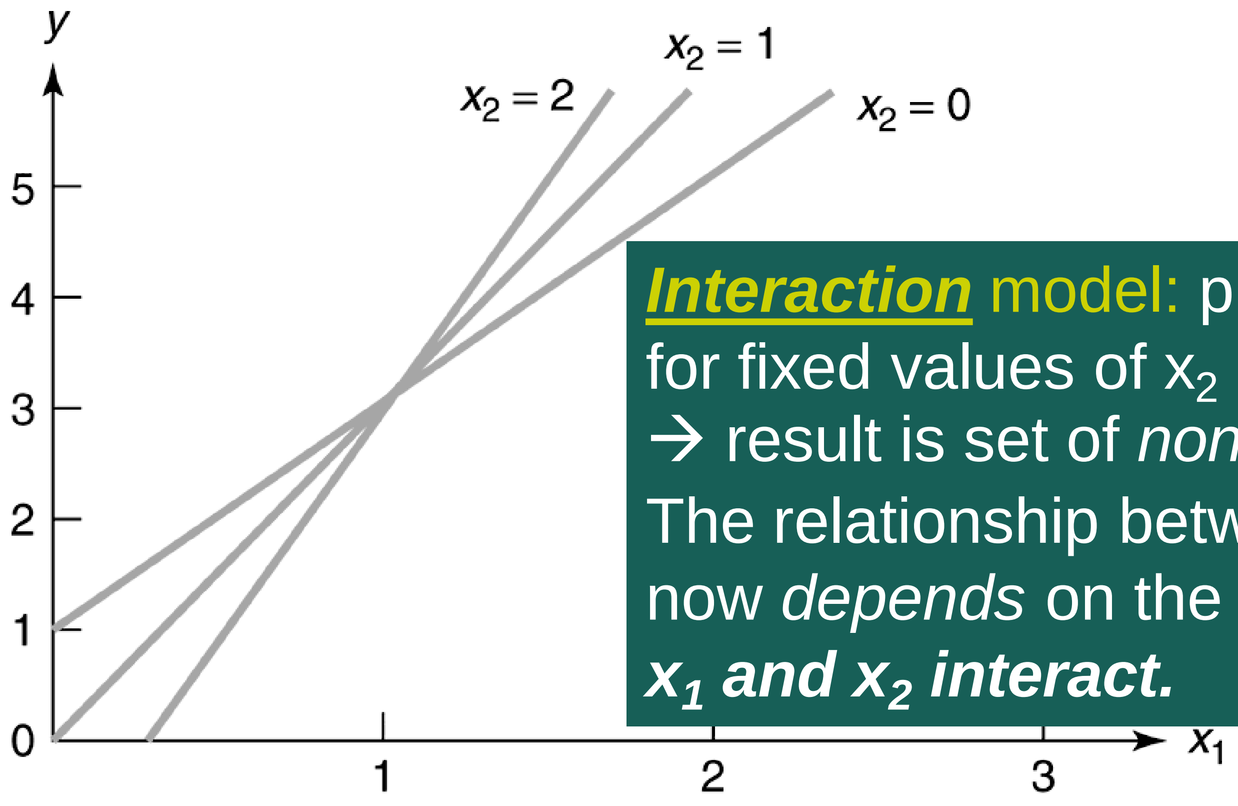
MLR model with no interaction



First-order model: plot y against x_1 for fixed values of x_2 .
→ result is set of *parallel* lines.
The relationship between y and x_1 does ***not*** depend on the values of x_2

Figure 4.1 Graphs of $E(y) = 1 + 2x_1 + x_2$ for $x_2 = 0,1,2$

MLR model with interaction



Interaction model: plot y against x_1 for fixed values of x_2
→ result is set of *non-parallel* lines.
The relationship between y and x_1 now *depends* on the value of x_2.
x_1 and x_2 interact.

Figure 4.9: MLR equation: $\mathbf{y = 1 + 2x_1 - 1x_2 + x_1x_2}$ for $x_2 = 0, 1, 2$

An Interaction Model Relating $E(y)$ to Two Quantitative Independent Variables

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

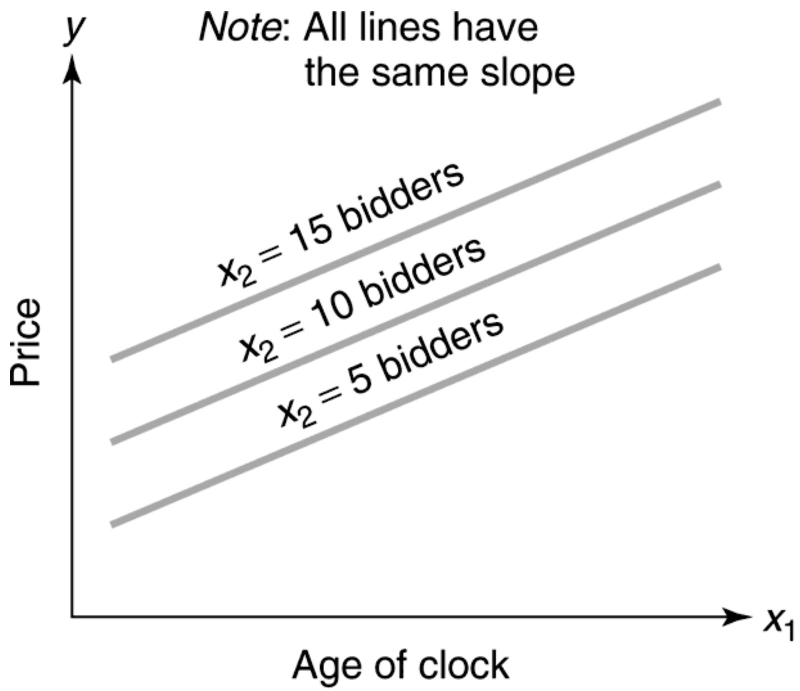
where

$(\beta_1 + \beta_3 x_2)$ represents the change in $E(y)$ for every 1-unit increase in x_1, holding x_2 fixed

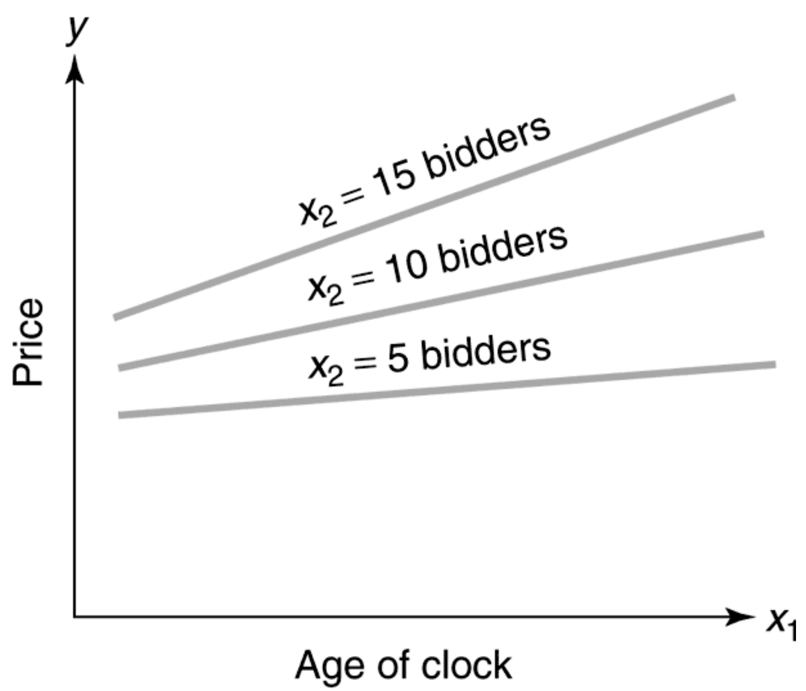
$(\beta_2 + \beta_3 x_1)$ represents the change in $E(y)$ for every 1-unit increase in x_2, holding x_1 fixed

Figure 4.10 Examples of no-interaction and interaction models

Price of antique clocks sold at auction depends on the age of the clock (108-194 yrs) and the number of bidders at the auction (5-15).



(a) No interaction between x_1 and x_2



(b) Interaction between x_1 and x_2

Exercise: Give a practical interpretation of the two scenarios.

Multiple Linear Regression Equation for an interaction

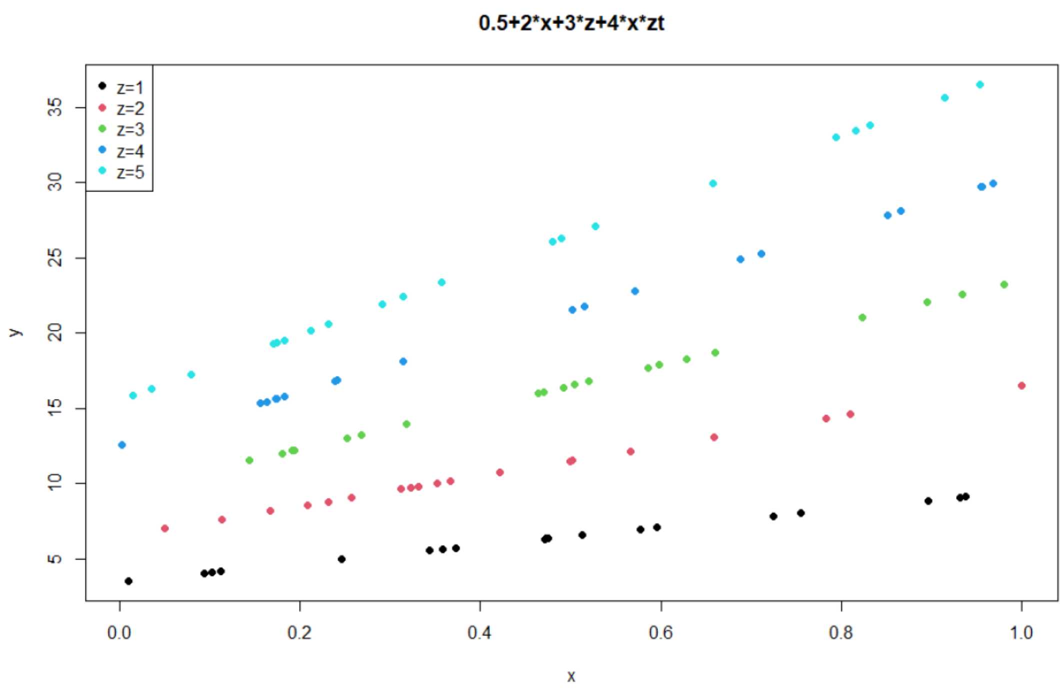
$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

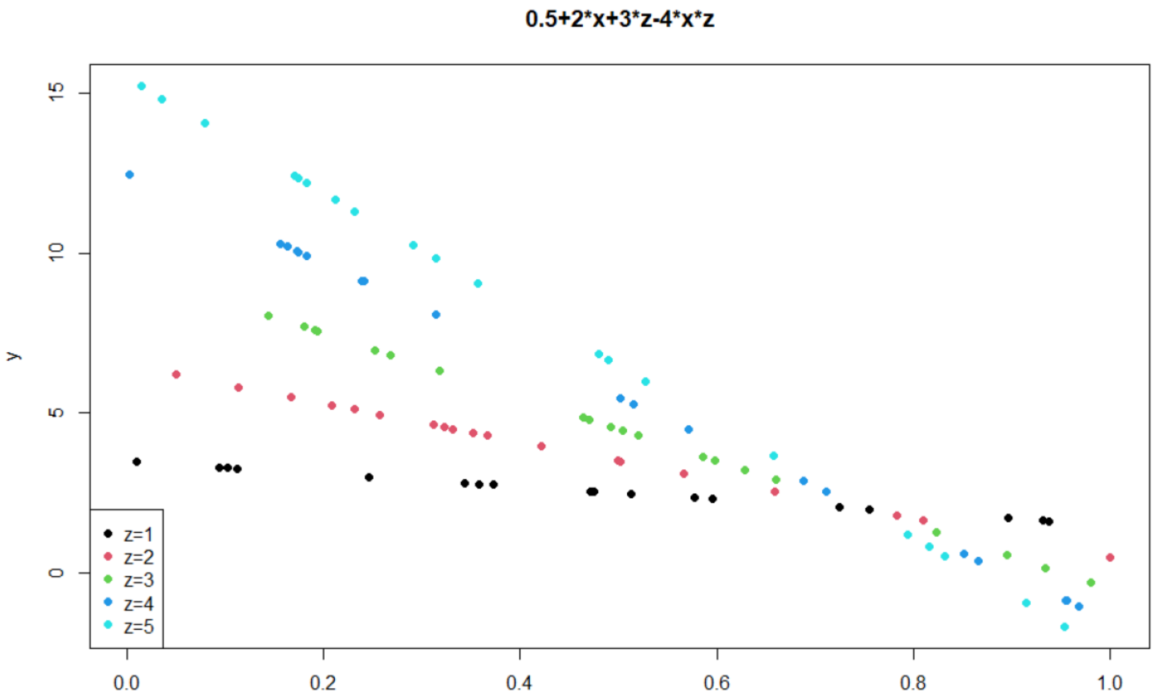
When there is an interaction:

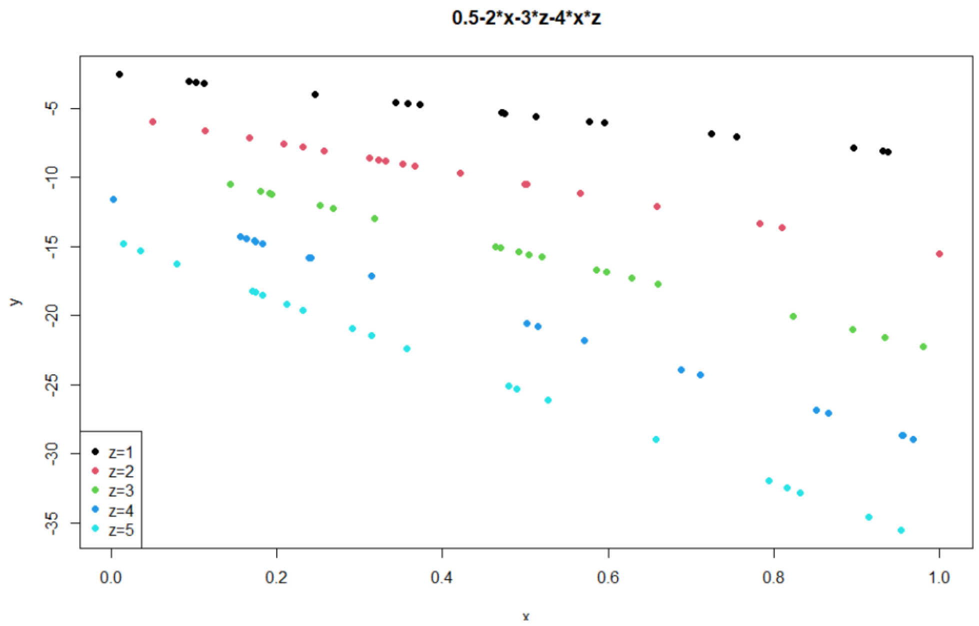
- The change in the two variables is now different depending on the interaction with the other variable.

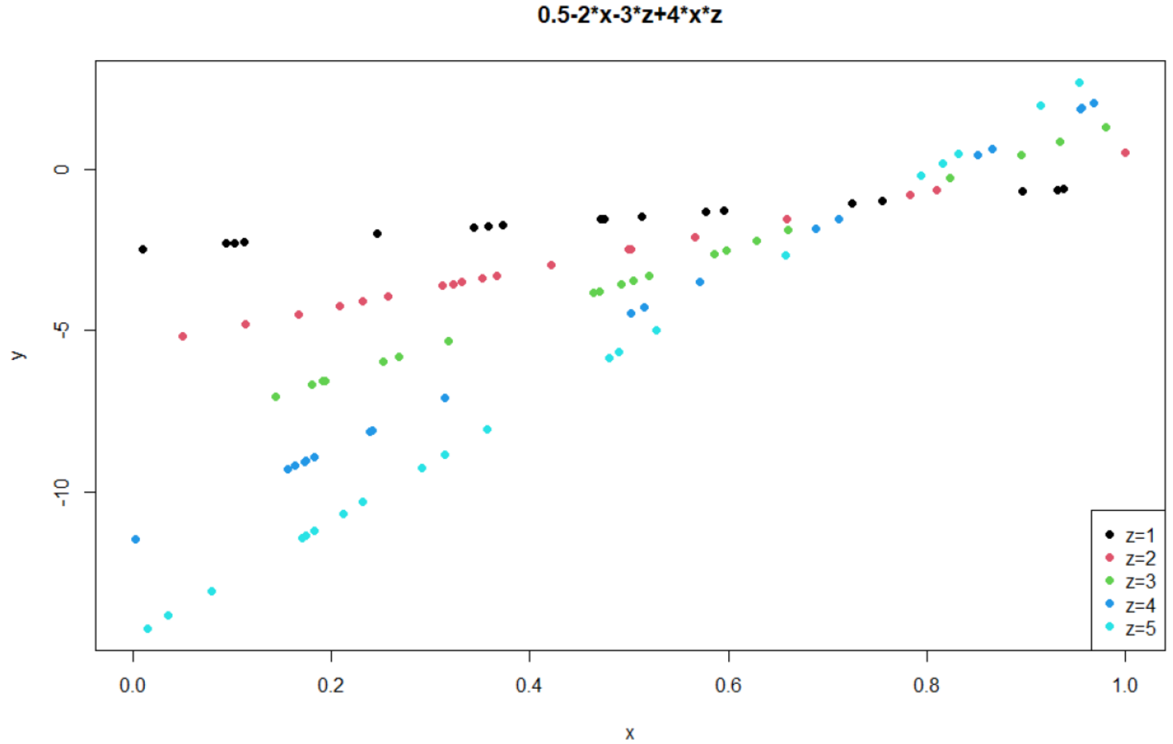
- **Positive interaction:** influence of the first variable increases with the second OR slope becomes more positive with higher values of the second variable.

- **Negative interaction:** influence of the first variable decreases with the second OR slope becomes more negative with higher values of the second variable.


0.5+2*x+3*z+4*x*zt
z=1
z=2
z=3
z=4
z=5
y
x


0.5+2*x+3*z-4*x*z
z=1
z=2
z=3
z=4
z=5
y

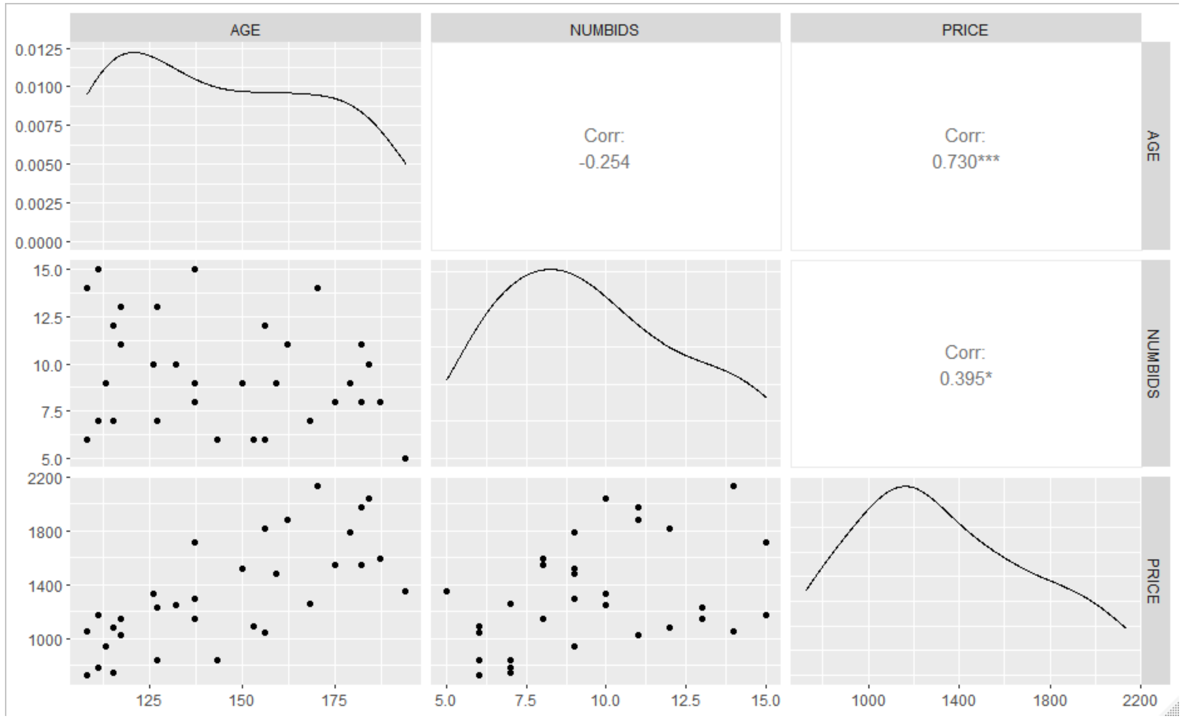

0.5-2*x-3*z-4*x*z
y
x
-5
-10
-15
-20
-25
-30
-35
0.0
0.2
0.4
0.6
0.8
1.0
z=1
z=2
z=3
z=4
z=5


0.5-2*x-3*z+4*x*z
y
x
0
-5
-10
0.0
0.2
0.4
0.6
0.8
1.0
z=1
z=2
z=3
z=4
z=5

Grandfather clocks example

Price of antique clocks sold at auction depends on the age of the clock (108-194 yrs) and the number of bidders at the auction (5-15).

```
gfclocks.df <- read.table("GFCLOCKS.txt", header=T)
ggpairs(gfclocks.df, columns = 1:3)
```



Grandfather clocks example

Fitting a MLR without interaction

```
mod1<-lm(PRICE ~ AGE + NUMBIDS, data = gfclocks.df)
summary(mod1)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1338.951    173.809   -7.70  1.7e-08
AGE            12.741      0.905   14.08  1.7e-14
NUMBIDS        85.953      8.729    9.85  9.3e-11

Residual standard error: 133 on 29 degrees of freedom
Multiple R-squared:  0.892,       Adjusted R-squared:  0.885
F-statistic:  120 on 2 and 29 DF,  p-value: 9.22e-15
```

Grandfather clocks example

Fitting a MLR with an interaction term

```
mod2<-lm(PRICE ~ AGE*NUMBIDS, data = gfclocks.df)
summary(mod2)
```

Price = 320 + 0.88*Age- 93.3*Numbids+ 1.3*Age*Numbids

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   320.458    295.141    1.09   0.2868
AGE             0.878      2.032    0.43   0.6690
NUMBIDS       -93.265     29.892   -3.12   0.0042
AGE:NUMBIDS     1.298      0.212    6.11  1.4e-06

Residual standard error: 88.9 on 28 degrees of freedom
Multiple R-squared:  0.954,      Adjusted R-squared:  0.949
F-statistic:  193 on 3 and 28 DF,  p-value: <2e-16
```

AGE:NUMBIDs is the interaction between AGE and NUMBIDS

If an interaction term is significant then do ***not*** test the “main effects” (AGE and NUMBIDS).

However, these *main effect terms must be kept in the model* (even if the p-value is not significant).

Predicting from the model

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x_1 + \widehat{\beta_2} x_2 + \widehat{\beta_3} x_1 x_2$$

?

```
Price = 320 + 0.88*Age - 93.3*Numbids
            + 1.3*Age*Numbids
```

Exercise: Predict mean price when age of clock is 100 yrs and no. of bidders is 10

Positive β interaction coefficient = steeper slope for larger Age or Numbids values.

Coding an interaction model in R

Model

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

In R:

```
mod1<-lm(y~x1*x2, data = data.df)
```

where

`x1*x2` = $x_1 + x_2 + x_1 : x_2$

That is, `x1*x2` gives the *main* effect of x_1, the *main* effect of x_2 <u>*and*</u> the *interaction* of the two ($x_1 : x_2$)

Coding an interaction model in R

Model

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

Other options:

```
mod1<-lm(y~(x1+x2)^2, data = data.df)
```

Where `^2` says include second order interaction term for variables in the brackets.

```
mod1<-lm(y~x1+x2+x1:x2, data = data.df)
```

Coding an interaction model in R

Model

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

```
mod1<-lm(y~x1*x2,data=data.df)
```

Always test the significance of the *interaction first*.

- If it is significant then do ***not*** test the main effects.
- If the interaction is ***not significant*** then *remove* the term and fit the main effects model only:

```
mod2<-lm(y~x1+x2, data = data.df)
```

View an interaction in R

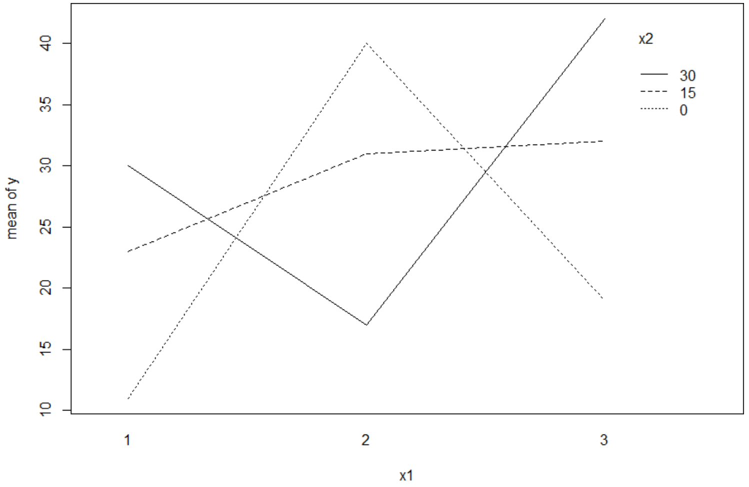
Variable plotted on x-axis is listed first (x_1)

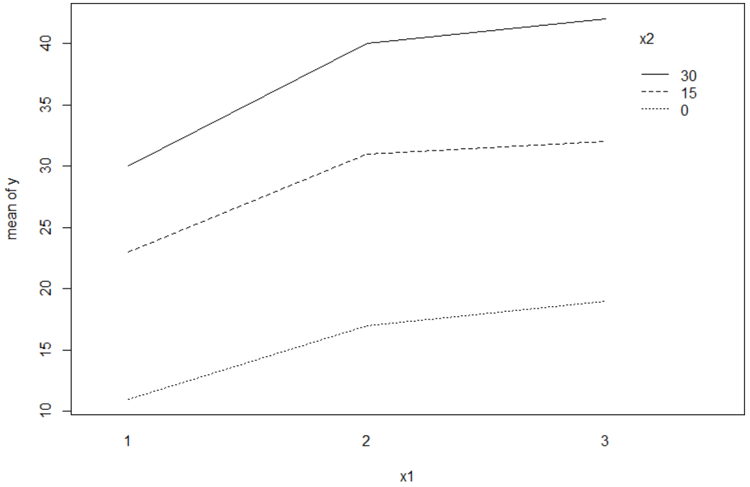
“Trace” variable (x_2) is listed second

with(data.df, interaction.plot(x_1, x_2, y))

data.df contains the variables x1 and x2

Variable plotted on y-axis (response, y) is listed last





Non-parallel lines suggest interaction

Parallel lines suggest NO interaction

Why fit an interaction?

- Residuals not following constant variance with a mean of zero and/or not normally distributed.
- Want to investigate an interaction
  - In the literature
  - Common sense
  - You're curious
- Assignment question asks you to…

Steps to Fitting Multiple Linear Regression

1. Exploratory analysis:
   - Look at correlations between variables
   - Later this will also include looking at multicollinearity as well.
2. Fit main effects model and look at output
   - summary and ANOVA (include equation)
   - Check Global usefulness (F stat and associated p-value)
   - Check if each predictor is useful (t-value and associated p-value for each variable)
3. Refit a model with only useful predictors
   - Include interaction term?
4. Check residuals to make sure none of the conditions for the residuals are violated
5. Interpret final model output
   - adjusted R^2
   - what predictors are fitted and if they are all still significant
   - include final equation
   - Influence of predictors on response

Example: Cereal

Variables:

- **Energy:** the kilojoules contained in a recommended serving
- **Protein:** measured in grams
- **Fat:** measured in grams
- **Fibre:** dietary fibre, measured in grams
- **Carbs:** carbohydrates, measured in grams

Questions?

Next lecture

Lecture 1

- Intro to MLR
- Fitting the model, testing the overall utility of a model
- Interpreting regression coefficients

Lecture 2

- Inferences about the individual β_i
- Multiple Coefficients of determination, R^2 and R^2_{adj}
- Using the model for estimation and prediction

Lecture 3

- An interaction model with quantitative predictors

Lecture 4

- Models with qualitative predictors

NB: Sections 4.11, 4.13 and 4.14 of the text will ***not*** be covered

Lecture 4

Multiple Linear Regression

Reminder: Multiple Linear Regression Equation

General Form of the Multiple Regression Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

where y is the dependent variable

$x_1, x_2, \ldots, x_k$ are the independent variables

$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$ is the deterministic portion of the model

β_i determines the contribution of the independent variable x_i

Note: The symbols $x_1, x_2, \ldots, x_k$ may represent higher-order terms for quantitative predictors (e.g., $x_2 = x_1^2$) or terms for qualitative predictors.

A Model Relating $E(y)$ to a Qualitative Independent Variable with Two Levels

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

where

$$x = \begin{cases} 1 & \text{if level A} \\ 0 & \text{if level B} \end{cases}$$

Interpretation of β's:

$$\beta_0 = \mu_B \text{ (Mean for base level)}$$

$$\beta_1 = \mu_A - \mu_B$$

A Model Relating $E(y)$ to a Qualitative Independent Variable with Three Levels

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

where

$$x_1 = \begin{cases} 1 & \text{if level A} \\ 0 & \text{if not} \end{cases} \qquad x_2 = \begin{cases} 1 & \text{if level B} \\ 0 & \text{if not} \end{cases} \qquad \text{Base level} = \text{Level C}$$

Interpretation of β's:

$$\beta_0 = \mu_C \text{ (Mean for base level)}$$

$$\beta_1 = \mu_A - \mu_C$$

$$\beta_2 = \mu_B - \mu_C$$

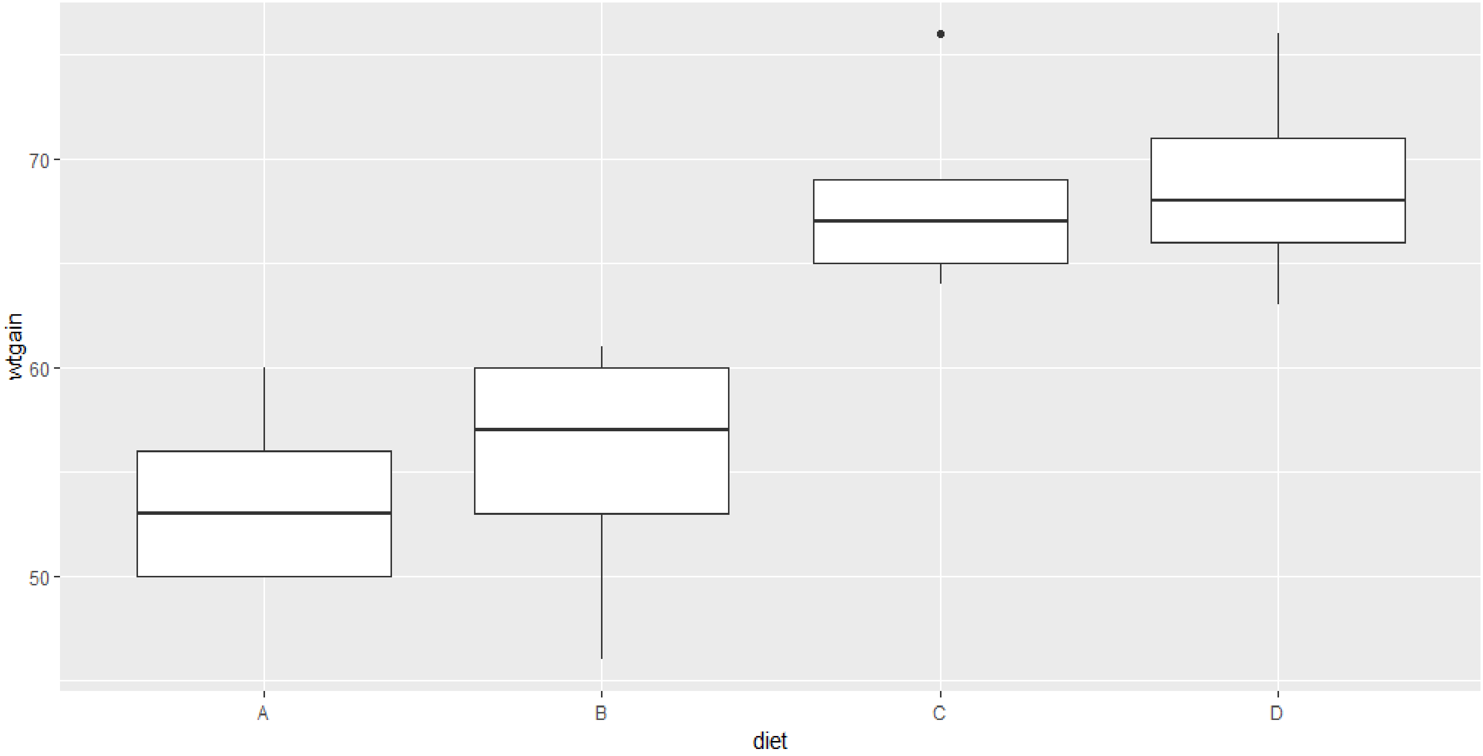
Chick weight gain

The following data are weight gains of chicks fed on different diets, and the mean weight gain of each diet.

Diet						Mean
A	60	50	50	53	56	53.8
B	46	60	61	53	57	55.4
C	64	67	76	69	65	68.2
D	66	63	71	68	76	68.8

Exploratory Plot

```
diet.df <- read.table("diet.txt",header=T)
# declare variable as qualitative/ factor
# optional if levels coded alphabetically
diet.df$diet <- factor(dietdf$diet)
library(ggplot2)
ggplot(diet.df, aes(x=diet, y=wtgain)) +
              geom_boxplot()
```



Hypotheses

- H_o: $\mu_A = \mu_B = \mu_C = \mu_{D,}$ i.e., mean weights are the same
- H_a: not ALL the treatment means are equal.

- The test of the null hypothesis H_o: $\mu_A = \mu_B = \mu_C = \mu_D$ $(= \mu)$
- asks the question:
- Is the model with a ***common mean*** for all diets adequate? That is,
    - $Y_{ij} \sim N(\mu, \sigma^2)$
- *or*

- Is more than one mean needed to represent the data?

Indicator (dummy variables)

$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

where

Base level = Level A (diet A)

x_1 = 1 if level B, 0 otherwise

x_2 = 1 if level C, 0 otherwise

x_3 = 1 if level D, 0 otherwise

Diet	x_1	x_2	x_3
A	0	0	0
B	1	0	0
C	0	1	0
D	0	0	1

If regression model contains a constant (intercept), then for a factor with k *levels*, k-1 *indicator variables* will uniquely define the k levels.

The way the indictor variables can be defined is **not** unique, but the 0-1 coding above is default in R

$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

For reference (baseline) level, Diet A, $(x_1 = x_2 = x_3 = 0)$

$$E(y) = \mu_A = \beta_0 + \beta_1 * 0 + \beta_2 * 0 + \beta_3 * 0$$

$$\mu_A = \beta_0$$

$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

For reference (baseline) level, Diet A, $(x_1=x_2=x_3=0)$

$$E(y) = \mu_A = \beta_0 + \beta_1{*}0 + \beta_2{*}0 + \beta_3{*}0$$

$$\mu_A = \beta_0$$

Diet B, $(x_1=1,\ x_2=x_3=0)$

$$E(y) = \mu_B = \beta_0 + \beta_1{*}1 + \beta_2{*}0 + \beta_3{*}0$$

$$\mu_B = \beta_0 + \beta_1$$

UNE

$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

For reference (baseline) level, Diet A, $(x_1=x_2=x_3=0)$

$$E(y) = \mu_A = \beta_0 + \beta_1 {*} 0 + \beta_2 {*} 0 + \beta_3 {*} 0$$

$$\mu_A = \beta_0$$

Diet B, $(x_1=1,\ x_2=x_3=0)$

$$E(y) = \mu_B = \beta_0 + \beta_1 {*} 1 + \beta_2 {*} 0 + \beta_3 {*} 0$$

$$\mu_B = \beta_0 + \beta_1$$

Diet C, $(x_1=0,\ x_2=1,\ x_3=0)$

$$E(y) = \mu_C = \beta_0 + \beta_1 {*} 0 + \beta_2 {*} 1 + \beta_3 {*} 0$$

$$\mu_C = \beta_0 + \beta_2$$

$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

For reference (baseline) level, Diet A, $(x_1=x_2=x_3=0)$

$$E(y) = \mu_A = \beta_0 + \beta_1{*}0 + \beta_2{*}0 + \beta_3{*}0$$

$$\mu_A = \beta_0$$

Diet B, $(x_1=1,\ x_2=x_3=0)$

$$E(y) = \mu_B = \beta_0 + \beta_1{*}1 + \beta_2{*}0 + \beta_3{*}0$$

$$\mu_B = \beta_0 + \beta_1$$

Diet C, $(x_1=0,\ x_2=1,\ x_3=0)$

$$E(y) = \mu_C = \beta_0 + \beta_1{*}0 + \beta_2{*}1 + \beta_3{*}0$$

$$\mu_C = \beta_0 + \beta_2$$

Diet D, $(x_1=x_2=0,\ x_3=1)$

$$E(y) = \mu_D = \beta_0 + \beta_1{*}0 + \beta_2{*}0 + \beta_3{*}1$$

$$\mu_D = \beta_0 + \beta_3$$

Interpreting the coefficients

$$\beta_0 = \mu_A$$
$$\beta_1 = \mu_B - \mu_A$$

β_1 is the ***difference*** *in mean weight gain* between chicks fed diet B and Diet A

Similarly:

$$\beta_2 = \mu_C - \mu_A$$
$$\beta_3 = \mu_D - \mu_A$$

Regression coefficients

β_0 = mean wtgn for diet A = 53.8

```
mod1<-lm(wtgain~diet, data= diet.df);
summary(mod1)
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)  β0  53.80        2.27    23.72  6.8e-14
dietB        β1   1.60        3.21     0.50  0.62472
dietC        β2  14.40        3.21     4.49  0.00037
dietD        β3  15.00        3.21     4.68  0.00025
```

Exercise:

Interpret the coefficient of dietB (β_1) dietC (β_2) and dietD (β_3), and calculate the estimated mean wt gain for chicks fed Diets C and D

Regression coefficients

β_0=mean wtgn for diet A = 53.8

```
mod1<-lm(wtgain~diet, data= diet.df);
summary(mod1)
Coefficients:
```

		Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	β_0	53.80	2.27	23.72	6.8e-14
dietB	β_1	1.60	3.21	0.50	0.62472
dietC	β_2	14.40	3.21	4.49	0.00037
dietD	β_3	15.00	3.21	4.68	0.00025

β_1=difference in mean wtgn between diet B and diet A
Hence mean wt gn for diet B = 53.8+1.6= 55.4

Regression coefficients

β_0 = mean wtgn for diet A = 53.8

```
mod1<-lm(wtgain~diet, data= diet.df);
summary(mod1)
Coefficients:
```

		Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	β_0	53.80	2.27	23.72	6.8e-14
dietB	β_1	1.60	3.21	0.50	0.62472
dietC	β_2	14.40	3.21	4.49	0.00037
dietD	β_3	15.00	3.21	4.68	0.00025

β_2 = difference in mean wtgn between diet C and diet A
Hence mean wt gn for diet B = 53.8+14.4= 68.2

Regression coefficients

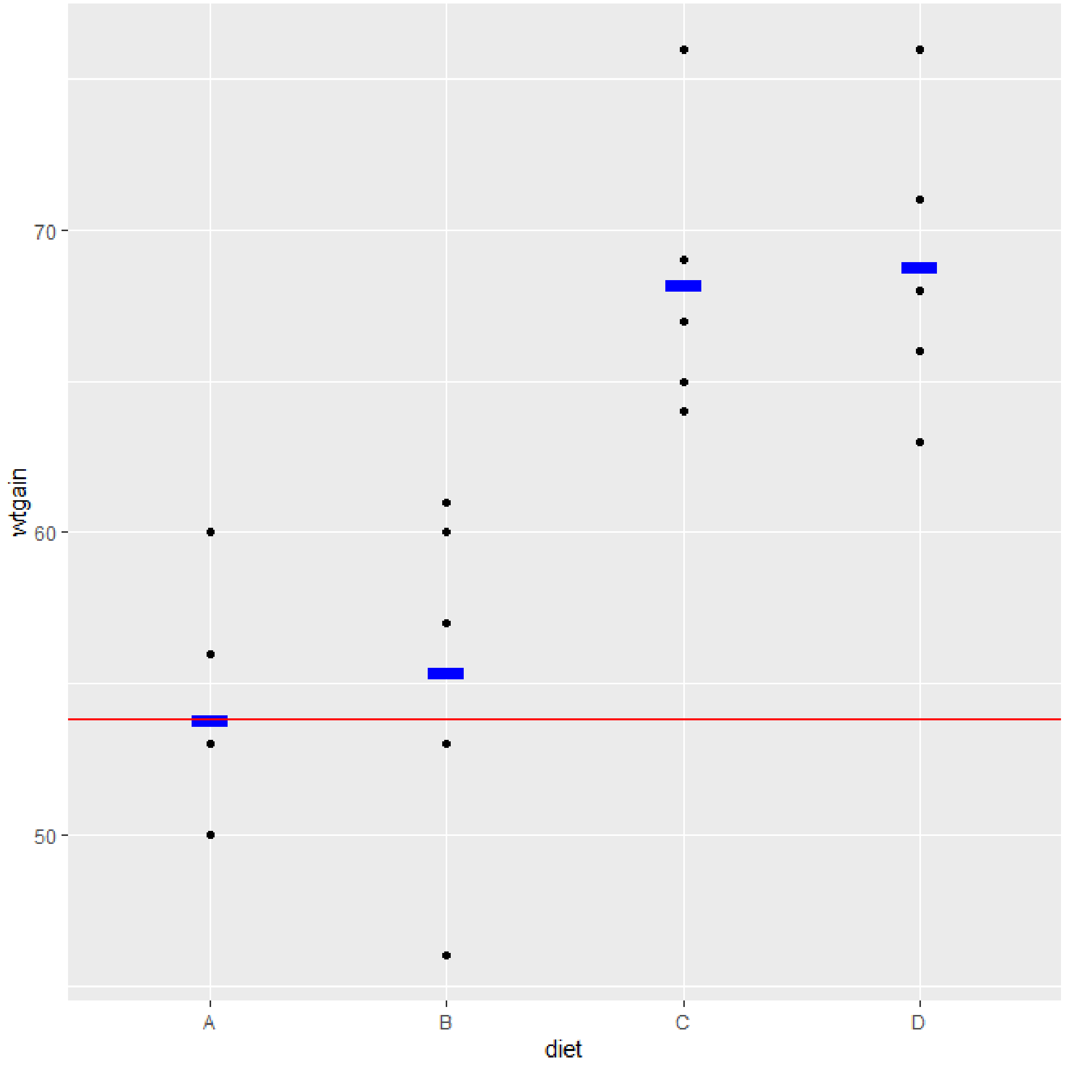
β_0=mean wtgn for diet A = 53.8

```
mod1<-lm(wtgain~diet, data= diet.df);
summary(mod1)
Coefficients:
```

		Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	β_0	53.80	2.27	23.72	6.8e-14
dietB	β_1	1.60	3.21	0.50	0.62472
dietC	β_2	14.40	3.21	4.49	0.00037
dietD	β_3	15.00	3.21	4.68	0.00025

β_3=difference in mean wtgn between diet D and diet A
Hence mean wt gn for diet B = 53.8+15= 68.8

Thinking about our βs graphically



Interpreting Confidence Intervals

```
mod1<-lm(wtgain~diet, data= diet.df)
confint(mod1)
```

	2.5 %	97.5 %
(Intercept)	49.0	58.6
dietB	-5.2	8.4
dietC	7.6	21.2
dietD	8.2	21.8

The confidence interval for the intercept in the model is the mean weight gain of diet A.

With 95% confidence mean weight gain for chicks fed diet A is between 49.0 and 58.6 g.

Interpreting Confidence Intervals

```
mod1<-lm(wtgain~diet, data= diet.df)
confint(mod1)
```

	2.5 %	97.5 %
(Intercept)	49.0	58.6
dietB	-5.2	8.4
dietC	7.6	21.2
dietD	8.2	21.8

The confidence interval for diet B in the model is the difference in mean weight gain between diet A and diet B.

With 95% confidence mean weight gain for chicks fed diet B is 5.2 g lower to 8.4g higher than chicks fed diet A.

Interpreting Confidence Intervals

```
mod1<-lm(wtgain~diet, data= diet.df)
confint(mod1)
```

	2.5 %	97.5 %
(Intercept)	49.0	58.6
dietB	-5.2	8.4
dietC	7.6	21.2
dietD	8.2	21.8

The confidence interval for diet C in the model is the difference in mean weight gain between diet A and diet C.

With 95% confidence mean weight gain for chicks fed diet C is 7.6 g and 21.2g higher than chicks fed diet A.

Interpreting Confidence Intervals

```
mod1<-lm(wtgain~diet, data= diet.df)
confint(mod1)
```

	2.5 %	97.5 %
(Intercept)	49.0	58.6
dietB	-5.2	8.4
dietC	7.6	21.2
dietD	8.2	21.8

The confidence interval for diet D in the model is the difference in mean weight gain between diet A and diet D.

Q. Interpret confidence interval for diet D?

Interpreting Confidence Intervals

```
mod1<-lm(wtgain~diet, data= diet.df)
confint(mod1)
```

	2.5 %	97.5 %
(Intercept)	49.0	58.6
dietB	-5.2	8.4
dietC	7.6	21.2
dietD	8.2	21.8

```
mod1<-lm(wtgain~diet-1, data=diet.df)
confint(mod1)
```

	2.5 %	97.5 %
dietA	49.0	58.6
dietB	50.6	60.2
dietC	63.4	73.0
dietD	64.0	73.6

Questions?

Chapter 4 Recap

- MLR: fitting multiple qualitative, quantitative variables and interactions
- General Equation for MLR:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

- Fitting a MLR model, testing the overall utility of a model using F-test
- Interpreting regression coefficients β_i
- Inferences about the individual β_i using t-test and p-value
- Interpreting R^2_{adj}
- Using the model for estimation and prediction