Ch 4: An introduction to regression

Why regression is **EXTREMELY** important

- The subject underpins nearly all mathematical applied statistics
- Subject also underpins the subject of econometrics and financial econometrics
- Rigorous statistical techniques are often an integral part of MSc dissertations
- **(High-level) financial research is inherently quantitative**
  - This is not to deny that finance is inherently subjective
  - But even calculating subjective benchmarks is inherently quantitative
  - Finance is also about much more than just applied mathematics
- **High-level research in the social sciences is increasingly quantitative in nature**

Mathematical good advice

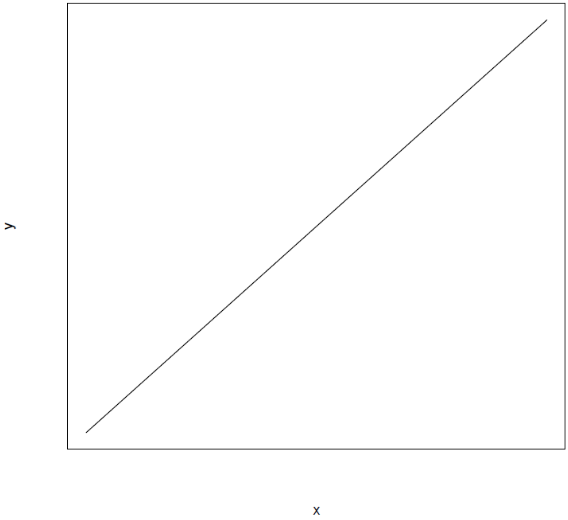
- **Math is not easy...**
  - often easier than it first looks
  - often the questions asked are deceptively simple

Basic question in regression

- What happens to Y as X increases?
  - increases?
  - decreases?
  - nothing?
- **In this way regression can be seen as a more advanced version of high-school maths**

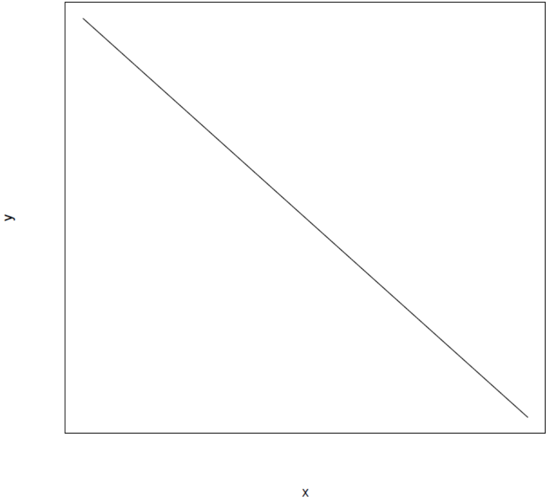
Positive gradient

- As X increases Y increases



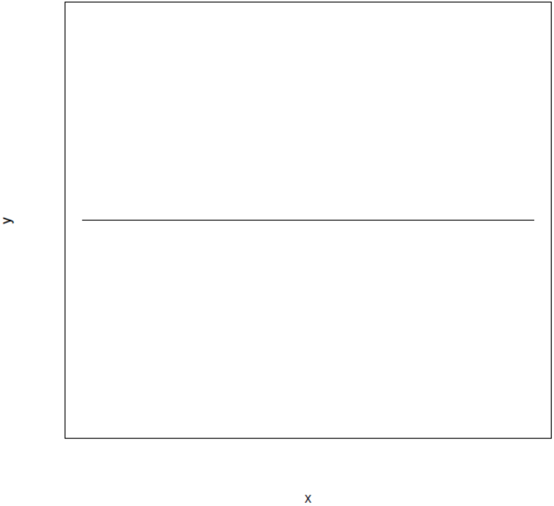
Negative gradient

- As X increases Y decreases



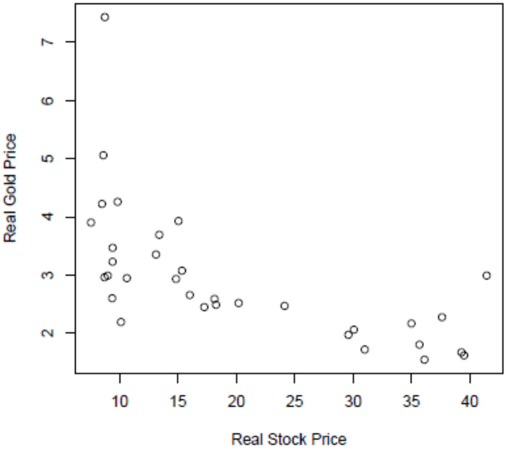
Zero gradient

- Changes in X do not affect Y



Real data example – more imperfect

- Need to remember that a real data set will be more imperfect
- But the same basic idea applies



Golden rule – don't panic!

- Regression problems can look a lot harder than they really are
  - Basic question remains what happens to Y as X increases?
- Beware of jargon. Gujarati and Porter (2009) distinguish between
  - Two variable regression model
  - Multiple regression model
- **Despite this apparent difference the mathematical methodology and the regression-fitting commands in R for both models are essentially the same**

Beware of cosmetic differences in notation and jargon

- Some authors use different terminology and notation for essentially the same thing
- **Remember that math is usually a lot simpler than you first imagine...**
- Two-variable regression model (Gujarati and Porter, 2009)

$$Y_i = \beta_1 + \beta_2 X_{2,i} + u_i$$

- Three-variable regression model (Gujarati and Porter, 2009)

$$Y_i = \beta_1 + \beta_2 X_{2,i} + \beta_3 X_{3,i} + u_i$$

- Multiple regression model (Gujarati and Porter, 2009)

$$Y_i = \beta_1 + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \ldots + \beta_p X_{p,i} + u_i$$

- **In terms of mathematical methodology and R commands etc. all these are special cases of the multiple linear regression model**

Outline of the lecture: Solving four basic regression problems

1. Plotting variables in R
   - cross-check formal statistical results with graphical analyses
   - deceptively important in practical research work
2. R^2 measures the proportion of variability in the data explained by the model
   - the higher the better
   - anything 0.3 or higher is potentially worthwhile
3. t-test
   - test the significance of individual parameters
4. F-test
   - test the significance of multiple parameters
5. Additional multiple regression example

1.1 Plotting variables

• Interested in the relationship between real (inflation-adjusted) stock prices and gold prices

• Want to plot the two variables together

- Cross-check the results of a formal statistical analysis
- **Very important in real project work**
- We do the statistical analysis so we are not restricted to simply looking at the graph and guessing

Financial context

• Some suggestion that as the stock price falls there is a flight to quality and people buy gold, the increased demand may, in turn, increase gold prices

• The reverse may also be true – people leave gold to play the market when the stock price rises.

• **Suggests that (inflation adjusted) stock prices and gold prices may be negatively correlated**

1.2 Plotting variables in R – reading in the data

- Data is in the file `L3eg1data.txt` available on Canvas
- Save to your USB stick then read in the data using the `read.table` command

```
data1<-read.table(‘‘E:L3eg1data.txt")
```

- The dataset contains two columns containing the real gold price (left column) and the real stock price (right column)
- In R assign variables linking the real gold price to the first column and the real stock price to the second column. (Note that this has to match the name `data1` given in the above sequence of commands)

```
realgoldprice<-data1[,1]
realstockprice<-data1[,2]
```

1.3 Actually plotting variables in R

• The basic plotting command in R is `plot`. But you have to variously specify

- axes titles, plotting style (line or dots; dots often simplest so is the default option), scaling (using the commands `xlim=c(lower, upper)`, `ylim=c(lower, upper)`)

- Usually best to stick with the simplest default options and then change these only if you need to.

• The plot command is applied to the X-variable first and then the Y-variable. For our simple lecture example

```
plot(realstockprice, realgoldprice, xlab=‘‘Real
Stock Price", ylab=‘‘Real Gold Price")
```

1.4 The graph – revisited

- Suggests that the two variables are indeed negatively correlated
- Still need to cross-check with the results of a formal statistical regression analysis
- But the same basic idea applies

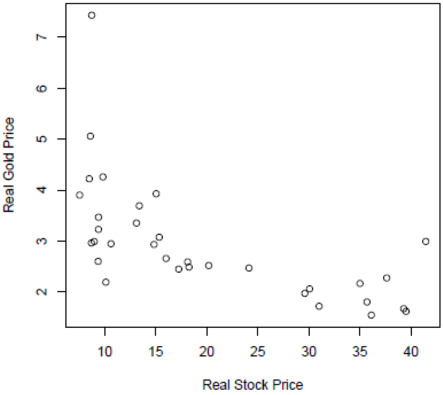


Figure:

2.1 R^2

- As part of this module you will need to demonstrate an ability to understand and interpret computer-generated model output
- R^2 is often one of the quickest and easiest things to make sense of
- **When running a regression R generates R^2 values automatically**
- The R^2 statistic gives you the proportion of the variability in the data explained by the regression model – the higher the better!
- **Important caveat.** R^2 automatically increases as additional X variables are added to a regression model. An **Adjusted** R^2 can be constructed that tries to take account of this although this statistic does not appear to be widely used.

2.2 Some observations about R^2

- R^2 lies between 0 and 1
  - $R^2 = 0$ models explains nothing
  - $R^2 = 1$ model explains everything
  - Generally the higher the value of R^2 the better the model
  - Textbook examples often have high R^2 values e.g. 0.7 or higher
- **There is no hard and fast rule about the interpretation of R^2. Usually an R^2 value of say 0.3 or higher is enough to say that there is a nontrivial amount of variation in the data explained by the model. In our example there is an R^2 value of 0.395325 showing us that the stock price clearly affects the price of gold. However, it is clear that other also factors affect the price of gold**

2.3 Where the R^2 statistic comes from

- Consider the following ANOVA table for a regression model (we will return to the ANOVA table later!)
- The ANOVA table shows that

$$R^2 = 1 - \frac{SSE}{SST}$$

Source	df	S. S.	M.S.	F
Regression	$p-1$	SSR	$MSR = \frac{SSR}{p-1}$	$F = \frac{MSR}{MSE}$
Error	$n-p$	SSE	$MSE = \frac{SSE}{n-p}$	
Total	$n-1$	SST	$MST = \frac{SST}{n-1}$	

2.4 Construction of the R^2 statistic...

$$
\begin{aligned}
R^2 &= \frac{\text{Variation explained by the model}}{\text{Total variation in the data}} \\
&= \frac{SSR}{SST} = \frac{SST - SSE}{SST} \\
&= 1 - \frac{SSE}{SST}
\end{aligned}
$$

2.5 Running the regression in R

- The basic command used is `lm` for linear model
- You specify the Y variable and then the X variables with a $\sim$ sign between the X and Y variables (mathematically this means "related to") + sign between the different X variables
- The best way to do this is to
  1. Run the regression analysis and store the results
  2. Get R to summarise the results for you in a second command
- For our simple lecture example

```
a.lm<-lm(realgoldprice~realstockprice)
summary(a.lm)
```

2.6 Interpreting the results of the regression

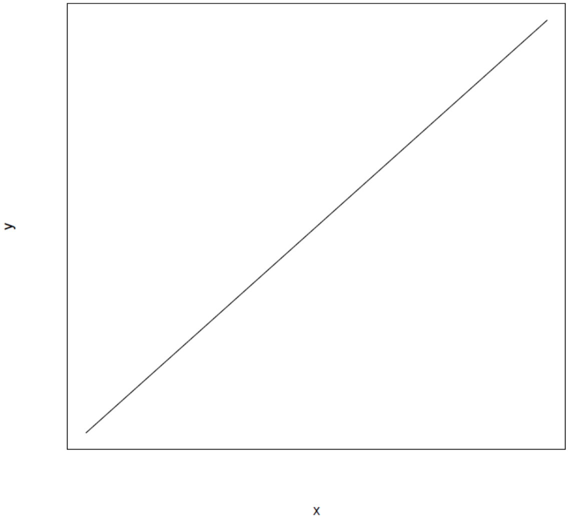
- Running the regression in either R produces a wealth of results – we only need a small portion of the results actually generated
- Interesting and useful bits of the results produced by R
  1. `R-squared` 0.395325
  2. `t-Statistic` for the variable `REALSTOCKPRICE` -4.502
  3. `F-statistic` 20.27
- **The rest of the lecture discusses what these t and F statistics really mean**

3.1 Regression and the t-statistic

- Basic question is always what happens to Y as X increases?
  - Increases?
  - Decreases?
  - Nothing?
- **As promised these are all very simple concepts and easy to visualise pictorally**

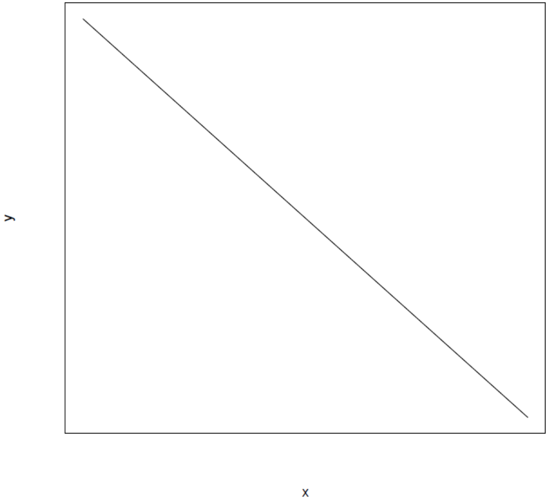
3.2 Positive gradient

- As X increases Y increases



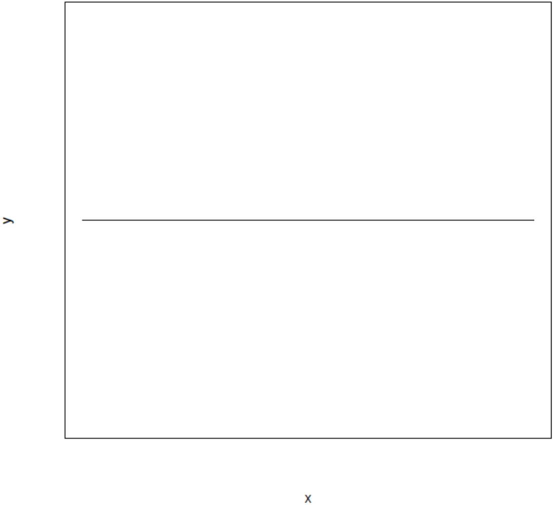
3.3 Negative gradient

- As X increases Y decreases



3.4 Zero gradient

- Changes in X do not affect Y



3.5 t-test

• I am afraid that some mathematics and some equations are unavoidable ...

• Consider the two-variable linear regression model

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

• **Want to see if X affects Y**

• It is a slightly strange way of thinking but the easiest way to do this is by testing the hypothesis

$$\begin{aligned} H_0 : \quad & \beta_2 = 0 \\ H_1 \quad & \beta_2 \neq 0 \end{aligned}$$

3.6 t-test: Gold price example re-visited

Variable	Coefficient	Std. Error	t-value	$Pr(> \|t\|)$
(Intercept)	4.21285	0.32351	13.022	4.14e-14***
realstockprice	-0.06409	0.01424	-4.502	8.90e-05***

• Usually always fit a constant term so the first row of the table is not really informative

• **The second row of the table (and downwards if a larger model) is THE INFORMATIVE part of the table**

• The asterisks denote statistical significance. 8.90e-05 may look weird but means 8.90×10^{-5}

3.7 Construction and interpretation of the t-test in regression

- Construction and interpretation of the t-test follows the example in Lecture 2 but this time with $n - p$ degrees of freedom

$$t = \frac{\text{Estimate} - \text{Hypothesised Value}}{\text{e.s.e}}$$

- Because it is extremely common to test the hypothesis $\beta_2 = 0$ the usual form of the t-statistic becomes

$$t = \frac{\text{Estimate} - 0}{\text{e.s.e}}$$

3.8 Computation in R

- The t-statistic computed in R can be constructed as

$$t = \frac{\text{Estimate} - 0}{\text{e.s.e}} = \frac{-0.064086}{0.014235} = -4.502 \text{ 3 d.p.}$$

- R calculates the p-value to be 8.90×10^{-5} (Slide 3.6).
- We can't calculate the exact p-value by hand but we can produce a bound for the p-value using tables.
- **The increased accuracy hints at how worthwhile computers are!**

3.9 Reconstructing what R does ...

- R calculates the p-value to be 8.90×10^{-5} (Slide 3.6).

$$\begin{aligned} n &= \text{No. of data points} = 33 \\ p &= \text{No. of variables in the model} = 2 \\ df &= n - p = 33 - 2 = 31 \end{aligned}$$

- $t_{31}(0.025) = 2.040$

$$|t| = 4.502 > t_{31}(0.025) = 2.040, \text{ therefore } p < 0.05$$

Interpretation

- Some evidence ($p < 0.05$) that stock prices affect gold prices
- As the coefficient is negative (and statistically significant) as stock prices increase gold prices decrease and vice versa.

4.1 F-test: Testing the significance of multiple parameters simultaneously

- **We want some way of systematically testing the overall fit of the model**
- It is possible to perform a sequence of t-tests in order to do this although for statistical reasons this is not really desirable
- The F-test performed automatically by R is only one possibility amongst many and may only have limited value in itself
- We will see in the next lecture that F-tests and the extra sum of squares principle can be applied much more generally

4.2 F-test for the overall fit of the model

- **The F-test produced automatically by R tests the overall fit of the model**

  - "Does at least one of the X-variables in the model have a statistically significant affect on Y?"

Formal hypothesis testing

- Multiple linear regression model

$$Y_i = \beta_1 + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \ldots + \beta_p X_{p,i} + u_i$$

$$\begin{aligned} H_0 &: \beta_2 = \beta_3 = \ldots = \beta_p = 0 \\ H_1 &: \text{At least one of the } \beta\text{s is non-zero} \end{aligned}$$

- Two-variable regression model

$$Y_i = \beta_1 + \beta_2 X_{2,i} + u_i$$

$$\begin{aligned} H_0 &: \beta_2 = 0 \\ H_1 &: \beta_2 \neq 0 \end{aligned}$$

- The output produced by R states

```
F-statistic:  20.27 on 1 and 31 DF, p-value:
8.904e-05
```

- This would be best interpreted as

"We have strong evidence ($p = 0.000$) that the real stock price affects the real gold price"

- We will see in the next lecture example that the interpretation of the F-statistic changes slightly when we have more than one X-variable in the regression model (in addition to the constant term).

4.4 Reconstructing calculation of the F-statistic in R

- Want to show where the numbers produced by R come from and give some additional practice of using the F-tables
- In general terms for the multiple linear regression model

$$Y_i = \beta_1 + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \ldots + \beta_p X_{p,i} + u_i$$

- Want to test the hypothesis

$$\begin{aligned} H_0 &: \beta_2 = \beta_3 = \ldots = \beta_p = 0 \\ H_1 &: \text{At least one of the } \beta\text{s is non-zero} \end{aligned}$$

- Construct the F-statistic as

$$\begin{aligned} F &= \frac{\frac{\text{Difference in SS}}{\text{Difference in d.f.}}}{\frac{\text{Residual SS (big model)}}{\text{Residual d.f.}}} = \frac{\frac{(R^2)TSS}{p-1}}{\frac{(1-R^2)TSS}{n-p}} \\ F &= \frac{\frac{R^2}{p-1}}{\frac{1-R^2}{n-p}} \sim F_{p-1,n-p} \end{aligned}$$

4.5 Two-variable regression example revisited ...

- The R output states `Multiple R-squared:  0.3953`
- Construct the F-statistic as

$$F = \frac{\frac{R^2}{p-1}}{\frac{1-R^2}{n-p}} = \frac{(n-p)R^2}{(p-1)(1-R^2)} = \frac{31(0.395325)}{1(0.604675)} = 20.267 \text{ (3 d.p.)}$$

- This needs to be compared to the value for $F_{1,31}$. From tables $F_{1,30} = 4.17$, $F_{1,40} = 4.08$

$$\begin{aligned} 31 &= 0.9(30) + 0.1(40), \\ F_{1,31} &= 0.9F_{1,30} + 0.1F_{1,40}, \\ F_{1,31} &= 0.9(4.17) + 0.1(4.08) = 4.161 \end{aligned}$$

- $F > F_{1,31}$ **so evidence (**$p < 0.05$**) that the real stock price affects the real gold price**

5.1 Multiple linear regression example

• To show you how to interpret the results from a multiple linear regression model use an example from the classical Longley dataset
• Overall aim is to explain the number of employed people in the US in terms of

1. X_2, GNP
2. X_3 the number of unemployed
3. X_4 the unemployment rate
4. X_5 the "non-institutionalised" population over the age of 14
5. $X_6$the yearly trend

5.2 R commands for reading in the data

- Data in the file `longley.txt`

```
longley<-read.table(``E:longley.txt")
x2<-longley[,1]
x3<-longley[,2]
x4<-longley[,3]
x5<-longley[,4]
x6<-longley[,5]
y<-longley[,6]
```

5.3 R multiple regression example

- Fit the model in the usual way using

```
a.lm<-lm(y~x2+x3+x4+x5+x6)
summary(a.lm)
```

- **R will produce a lot of irrelevant information. The obvious things to look at are**

1. The R^2 statistic
2. The individual t-statistics
3. The F-statistic to assess overall fit

5.4 Interpreting R output

- R produces a lot of information
  - **Not all of it will be relevant**

1. The R^2 statistic

R states `Multiple R-squared:  0.9955`

- R^2 is very high which suggests we might have quite a good model
- $R^2 = 0.9955$ which means that the model explains around 99.6% of the variability in the data
- Whilst this R^2 value is very high there is a chance that this is potentially too high to be true (see later)

5.5 Interpreting R output

- R produces a lot of information
  - **Not all of it will be relevant**

2. The t statistic

- For this course we need to look at the variables for which $p < 0.05$
- In project work, like dissertations, sometimes the interpretation might be different and a p-value satisfying $0.1 < p < 0.05$ might give weak evidence of an effect

- Need to analyse the results carefully
- Results given by R suggest that not all of the variables are statistically significant

5.6 Shortened version of the results given by R

```
Coefficients:
Estimate Std.  Error t value Pr(>|t|)
(Intercept) -3.450e+03 8.282e+02 -4.165 0.001932 **
x2 -3.196e-02 2.420e-02 -1.321 0.216073
x3 -1.972e-02 3.861e-03 -5.108 0.000459 ***
x4 -1.020e-02 1.908e-03 -5.345 0.000326 ***
x5 -7.754e-02 1.616e-01 -0.480 0.641607
x6 1.814e+00 4.253e-01 4.266 0.001648 **
```

5.7 Interpreting *t*-statistics

1. *t*-statistics show that not all the variables are statistically significant
2. Since the convention is to usually include a constant term in the model anyway the *t*-statistic for the constant term is not usually very informative
3. *p*-values suggest that the variables X2 and X5 are not statistically significant ($p > 0.05$)
   - the sign is irrelevant there is no formal statistical evidence of an effect
4. The coefficient of X3 is negative and statistically significant ($p < 0.05$)
   - As the number of unemployed people increases the number of employed people decreases

5. The coefficient of X4 is negative and statistically significant ($p < 0.05$)

- As the unemployment rate increases the number employed decreases decrease

6. The coefficient of X_6 is positive and statistically significant ($p < 0.05$)

- As X_6 is the time trend, this suggests that the number employed is generally increasing every year over the period in question.

5.9 F-statistic for assessing overall fit

• According to R

```
F-statistic:  438.8 on 5 and 10 DF, p-value:
2.242e-11
```

• **This presents evidence ($p = 2.242 \times 10^{-11} < 0.05$) that at least one of the X-variables in the study affects Y**

• **However, for example, do we need to include both the unemployment rate and the number of unemployed people in the same model?**