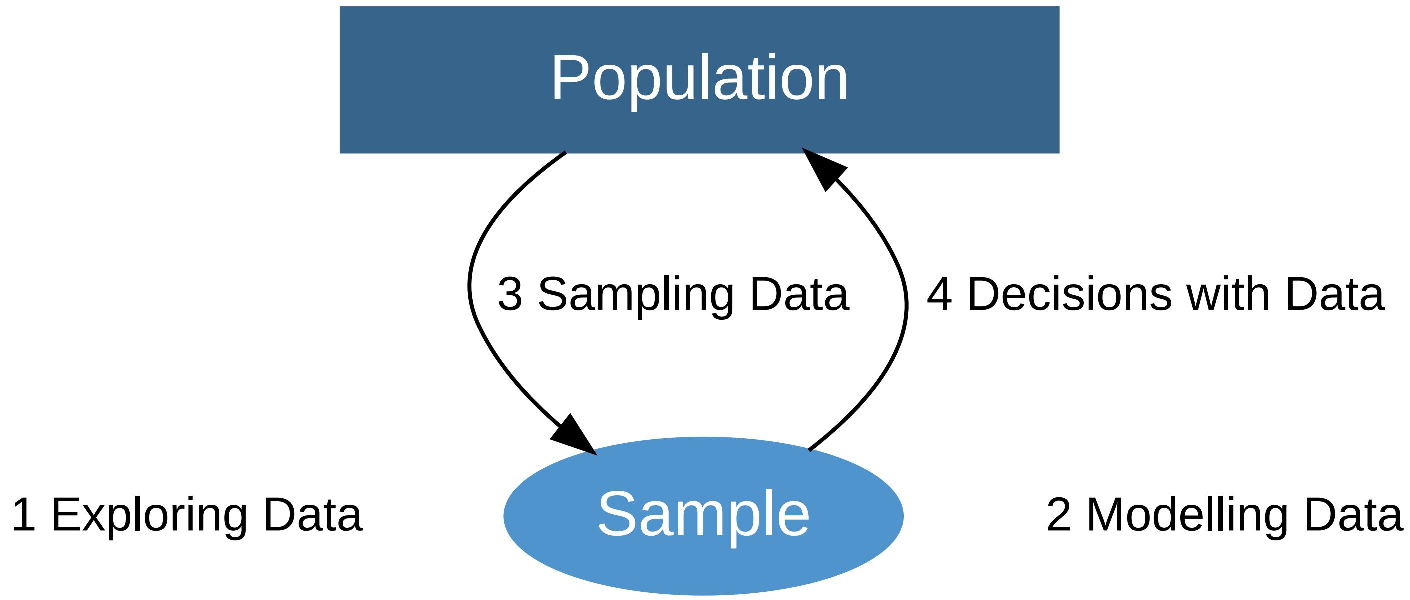
Linear Regression Summary

Modelling Data | Linear Model

Unit Overview



Module2 Modelling Data

Normal Model

What is the Normal Curve? How can we use it to model data?

Linear Model

How can we describe the relationship between 2 variables? When is a linear model appropriate?

Linear Regression Summary

Data Story

What affects the price of a Newtown property?

Data on Newtown Property Sales

- Data is taken from domain.com.au:
  - All properties sold in Newtown (NSW 2042) between April-June 2017.
  - The variable `Sold` has price in $1000s.

```
data <- read.csv("data/NewtownJune2017.csv", header = T)
head(data, n = 2)
```

```
##                ï..Property  Type    Agent Bedrooms Bathrooms Carspots Sold
## 1 19 Watkin Street Newtown House RayWhite        4         1        1 1975
## 2  30 Pearl Street Newtown House RayWhite        2         1        0 1250
##      Date
## 1 23/6/17
## 2 23/6/17
```

```
attach(data)  # Attaches names, so we don't need to use $
```

Which variables are most helpful in predicting property price?

Summary of linear regression

Summary of linear regression

Given bivariate data (x, y):

1. Produce a scatter plot

Does it look linear?

2. Produce a Regression line

$$\hat{y} = a + bx$$

3. Calculate the correlation coefficient (r)

How strong is the linear assocation?

4. Produce a residual plot

Does it look random? Is the linear model appropriate or would another model be better?

5. Check assumptions

Does the data look homoscedastic?

6. Perform predictions

- Predict y for given x
- Predict y within a vertical strip

Applying to Newtown data

1. Produce a scatter plot

```
plot(Bedrooms, Sold, xlab = "Bedrooms", ylab = "Sold Price (1000s)")
points(mean(Bedrooms), mean(Sold), col = "indianred", pch = 19, cex = 2)  # point of averages (centre)
```

Is this linear?

2. Produce a Regression line

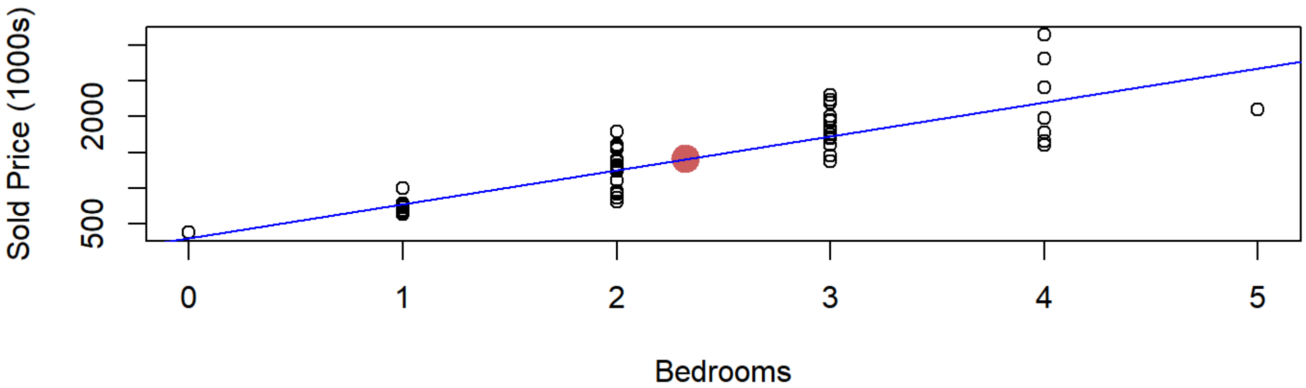
```
lm(Sold ~ Bedrooms)
```

```
##
## Call:
## lm(formula = Sold ~ Bedrooms)
##
## Coefficients:
## (Intercept)     Bedrooms
##       298.9        477.4
```

So

$$Sold = 298.9 + 477.4Bedrooms$$

```
plot(Bedrooms, Sold, xlab = "Bedrooms", ylab = "Sold Price (1000s)")
points(mean(Bedrooms), mean(Sold), col = "indianred", pch = 19, cex = 2)  # point of averages (centre)
abline(lm(Sold ~ Bedrooms), col = "blue")
```



3. Calculate the correlation coefficient

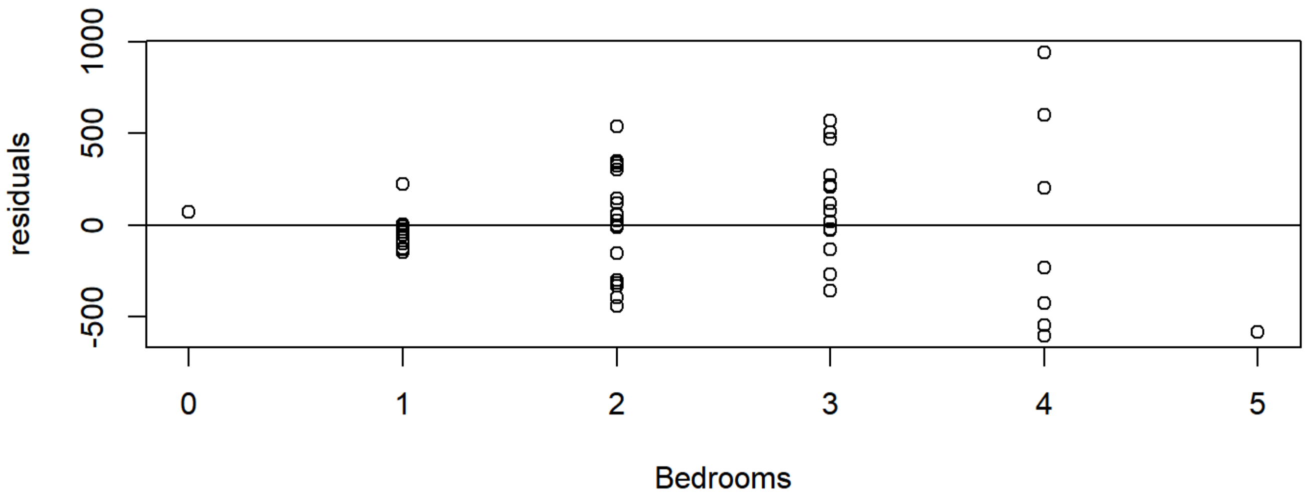
```
cor(Sold, Bedrooms)
```

```
## [1] 0.8475377
```

How strong is the linear association?

4. Produce a residual plot

```
l = lm(Sold ~ Bedrooms)
plot(Bedrooms, l$residuals, ylab = "residuals")
abline(h = 0)
```



This shows “fanning” (rather than randomness) which means more investigation is needed. You might have already identified this problem in the scatterplot.

5. Check assumptions

From the fanning, the data is not homoscedastic.

6. Perform predictions

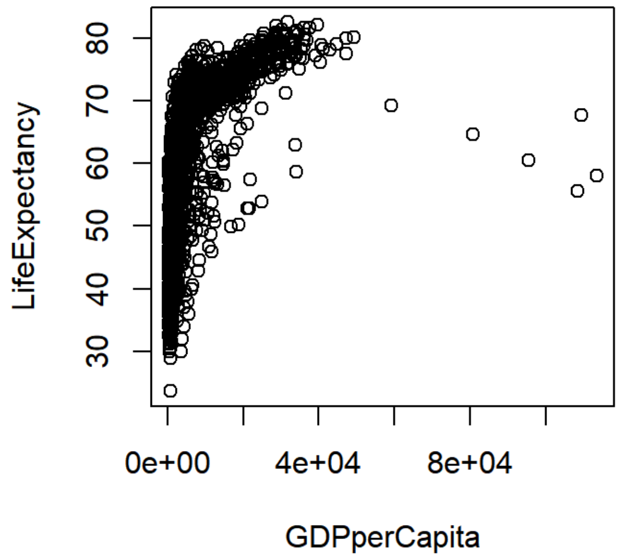
Predictions would not be appropriate.

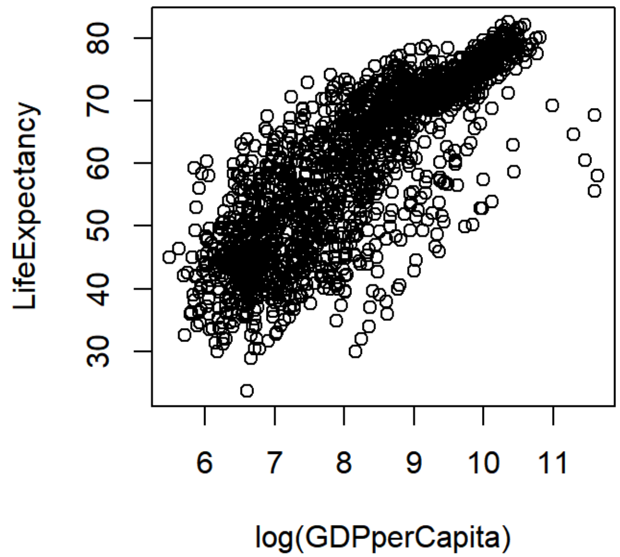
More on regression

Transformation

If the data is very spread out, then we can try transforming 1 or both of the original variables.

For example, we could take `ln(y)` or `ln(x)` as the new variables.





Multiple regression (Quick intro)

- The natural extension to linear regression is multiple regression, in which we look at the connection between y and 2+ x variables.
- The equation becomes

$$\hat{y}_i = a + b_1 x_{i,1} + b_2 x_{i,2} + \ldots + b_n x_{i,n}$$

- The coefficient b_j represent the association between variables $x_{i,j}$ and y_i. The sign of b_j is the direction of the association.
- Changing the set of variables can change the model suprisingly.
- Multicollinearity occurs when 2 variables are highly correlated with each other.
- A binary quantiative variable can be added to a multiple regression by coding a "dummy variable" as 0 and 1.

```r
head(data, n = 2)
```

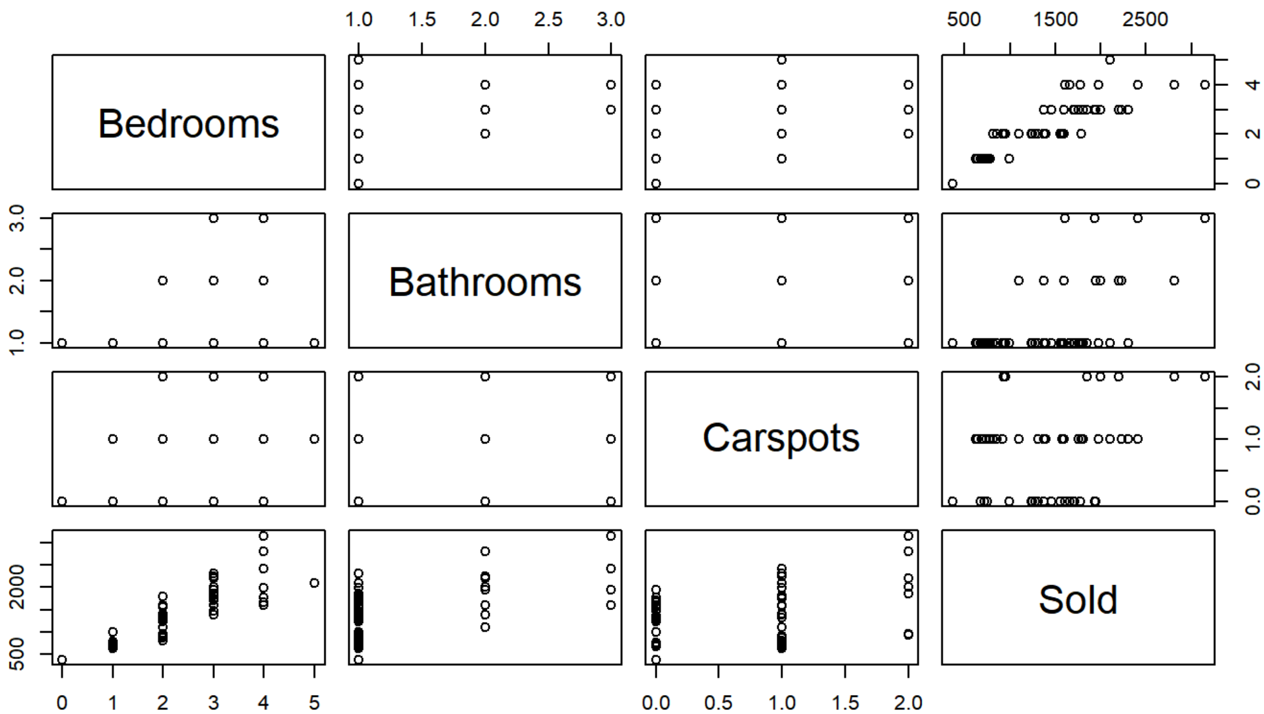
```
##                ï..Property  Type    Agent Bedrooms Bathrooms Carspots Sold
## 1 19 Watkin Street Newtown House RayWhite        4         1        1 1975
## 2  30 Pearl Street Newtown House RayWhite        2         1        0 1250
##      Date
## 1 23/6/17
## 2 23/6/17
```

```r
lm(Sold ~ Bedrooms + Bathrooms + Carspots)
```

```
##
## Call:
## lm(formula = Sold ~ Bedrooms + Bathrooms + Carspots)
##
## Coefficients:
## (Intercept)     Bedrooms    Bathrooms     Carspots
##      103.33       416.23       211.47        81.72
```

$$Sold = 103.33 + 416.23 Bedrooms + 211.47 Bathrooms + 81.72 Carspots$$

```r
pairs(data[, 4:7])
```



Summary

Fitting a linear model is easy in R, but requires careful thought to make sure it is appropriate. Otherwise any predictions are invalid.

Key Words

transformation, multiple regression