MSCI152: Introduction to Business Intelligence and Analytics

Lecture 9: Simple linear regression

Dr Anna Sroginis

Lancaster University Management School

Agenda

1. Introduction to the simple linear regression
2. Estimating the regression: Method of Least Squares
3. Assessing the Fit of the Simple Linear Regression Model
4. Confidence Intervals
5. Extrapolation

More details and explanations can be found in Camm et al., Sections 7.1, 7.2, 7.3, 7.5 and 6.4

Regression Modelling

Where are we?

- We have our data
  - probably a sample representing a much larger population
- We've drawn a scatter diagram
- We've calculated the correlation coefficient

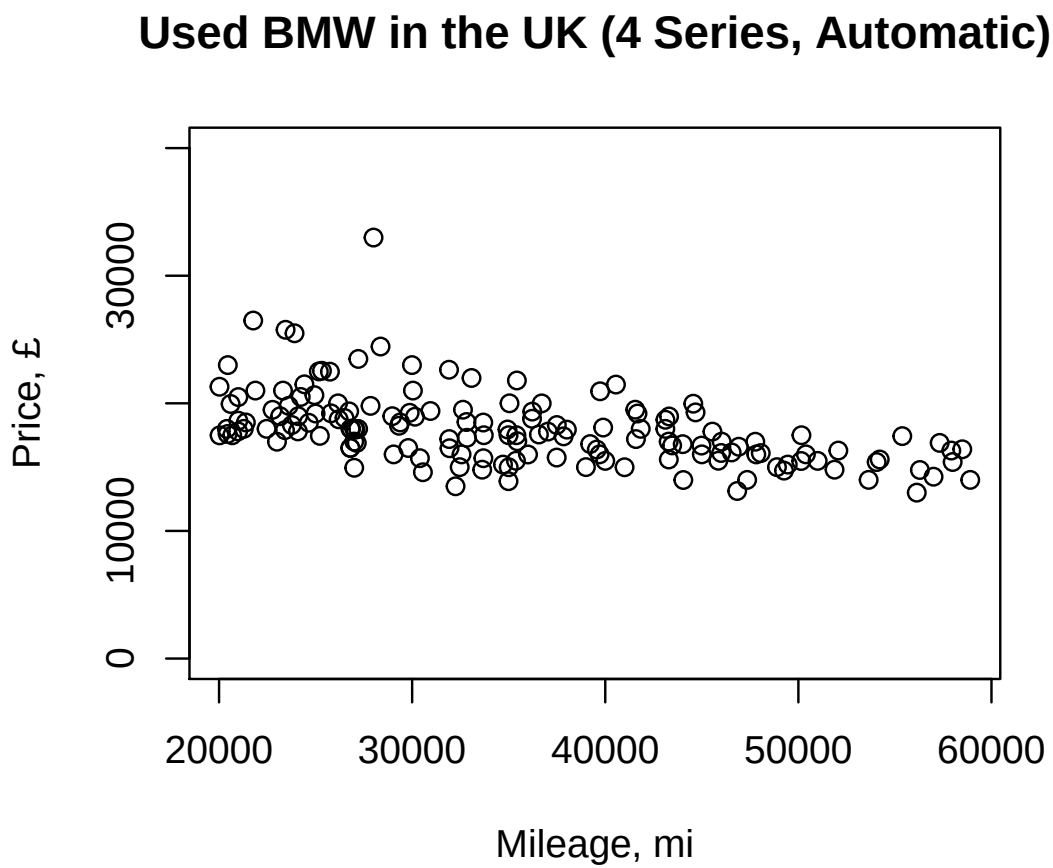
What do we do now?

We want a **model** that

- Provides an estimate of what happens more generally
- Can be applied to other hypothetical situations

Assuming we believe there is a causal relationship

BMW cars: mileage vs price

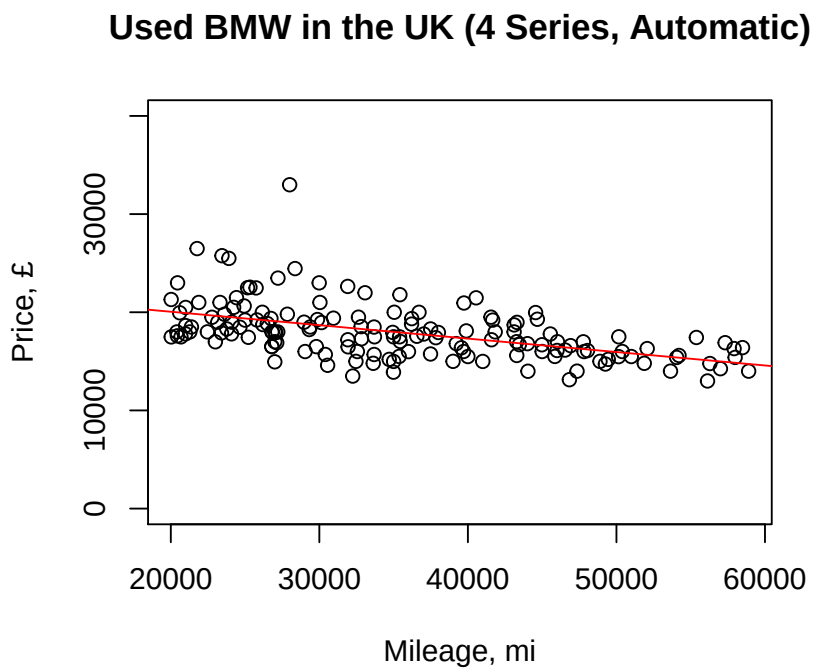


- We concluded that there is a linear relationship between price and mileage so that with the growth of mileage, the price tends to reduce.
- **What is the expected price of a car with a specific mileage?**

Simple linear regression model

A simple linear regression model is a mathematical equation of a straight line involving y and x

$$y_j = \beta_0 + \beta_1 x_j + \epsilon_j, \qquad (1)$$

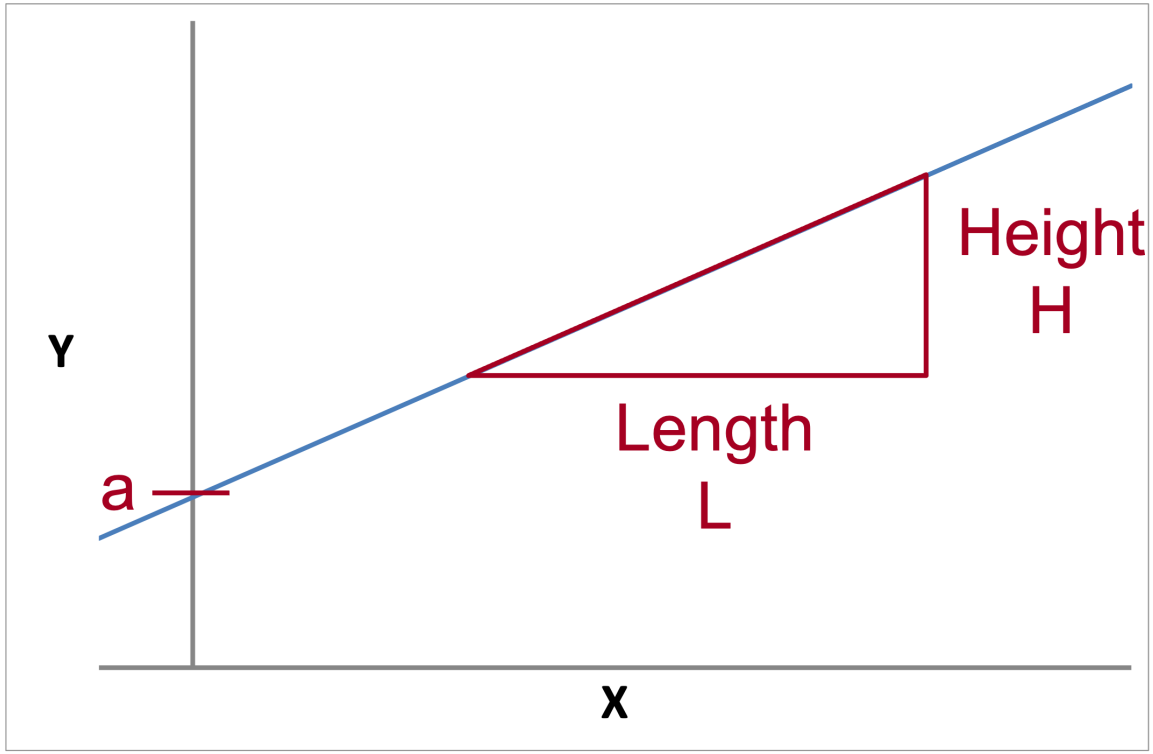


- where y_j is the price,
- x_j is the mileage,
- ϵ_j is an error term (some randomness),
- β_0 is constant in the population
- and β_1 is the coefficient for the slope in the population

Not exact: usually points won't all lie precisely on a straight line, so we want to choose **the line of best fit**

Equation of a straight line

$y = a + bx$ (where a and b are constants)



a is the intercept on the y axis
b is the gradient or slope, $b = H/L$

- negative if the line slopes downwards

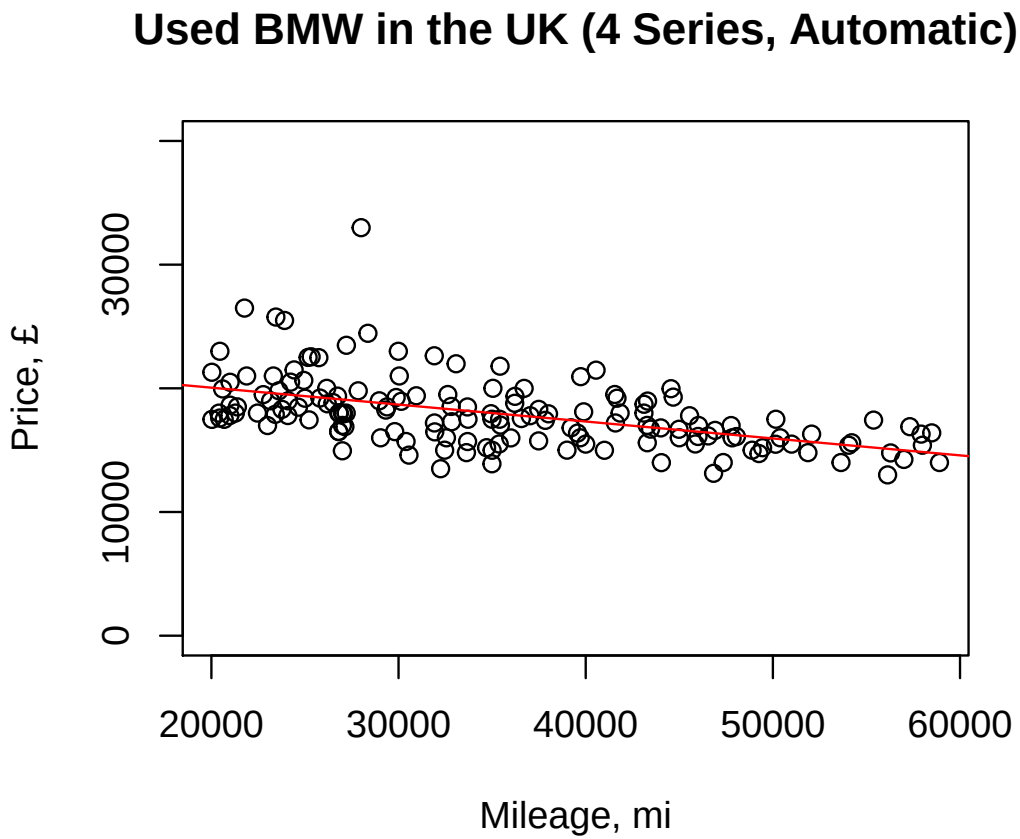
β_0 and β_1

While we might know price y_j and mileage x_j, we don't know the values of parameters.

- We don't even know if this should be a linear function.

How can we estimate the parameters?

By eye?
$\beta_0 \approx$
$\beta_1 \approx$
'likely size of $\varepsilon \approx$



Simple linear regression equation

We deal with a sample, so the parameters are just estimates of the truth:

$$\hat{y}_j = b_0 + b_1 x_j, \tag{2}$$

- b_0 is only an estimate of the "true" β_0,
- b_1 is only an estimate of the "true" β_1,
- $\hat{y}_j$ is the predicted value.

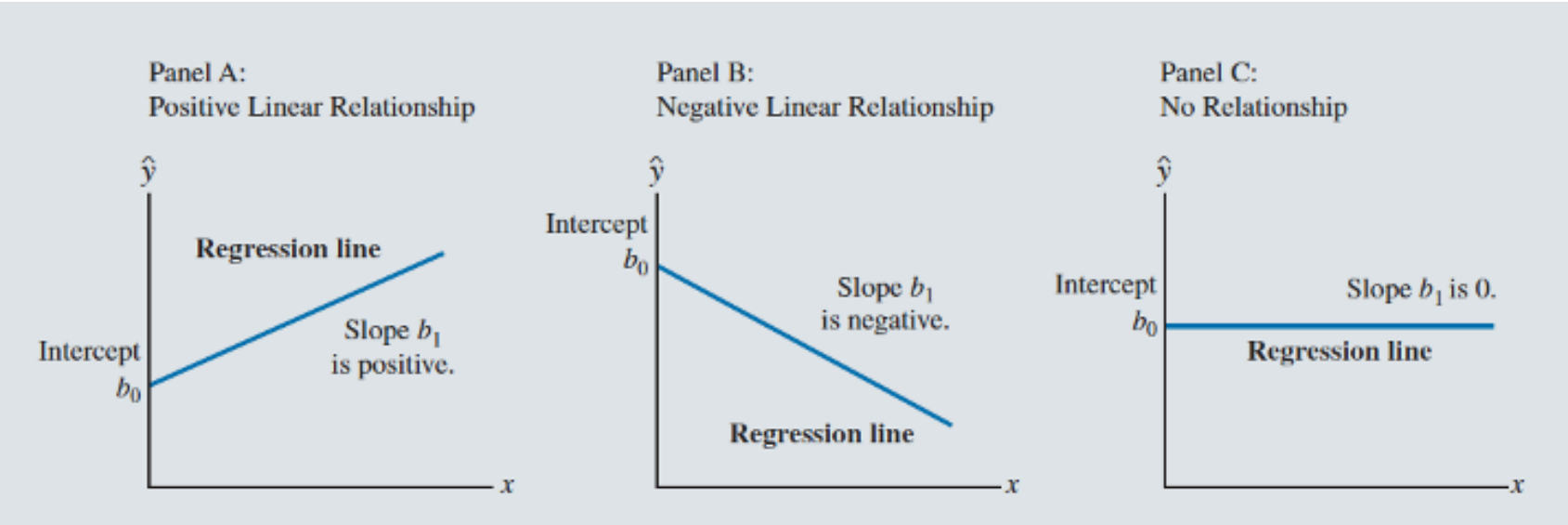
And the respective estimated regression model is:

$$y_j = \hat{y}_j + e_j = b_0 + b_1 x_j + e_j \tag{3}$$

Note that $e_j \neq \epsilon_j$

- *Why?*

Examples of possible regression lines



- The estimated mean value of y is related positively/negatively to x, with larger/smaller values of y associated with larger/smaller values of x;
- But do not forget about correlation $\neq$ causation!

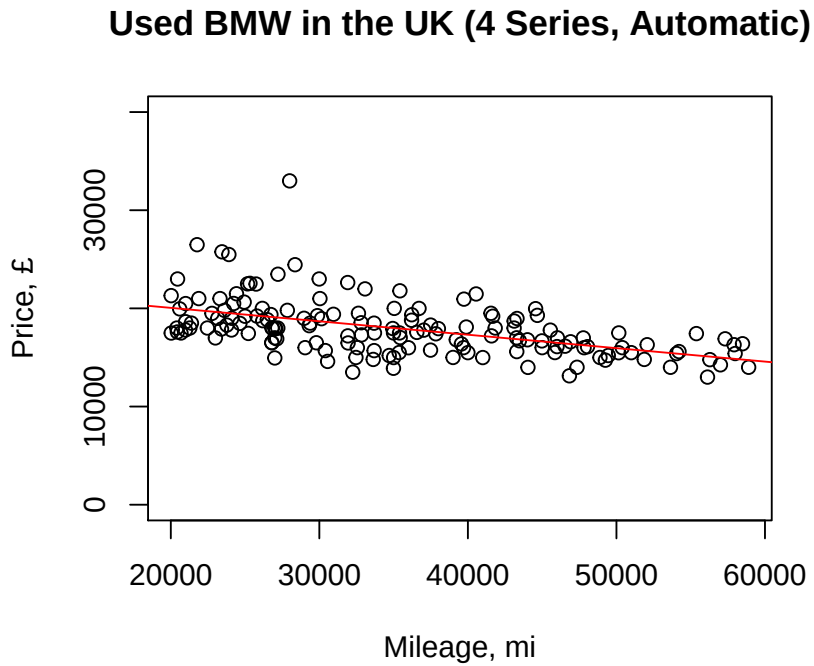
Introduction to the simple linear regression

While the population regression line is:

$$E(y_j|x_j) = \beta_0 + \beta_1 x_j, \qquad (4)$$

the estimated is:

$$\hat{y}_j = b_0 + b_1 x_j, \qquad (5)$$



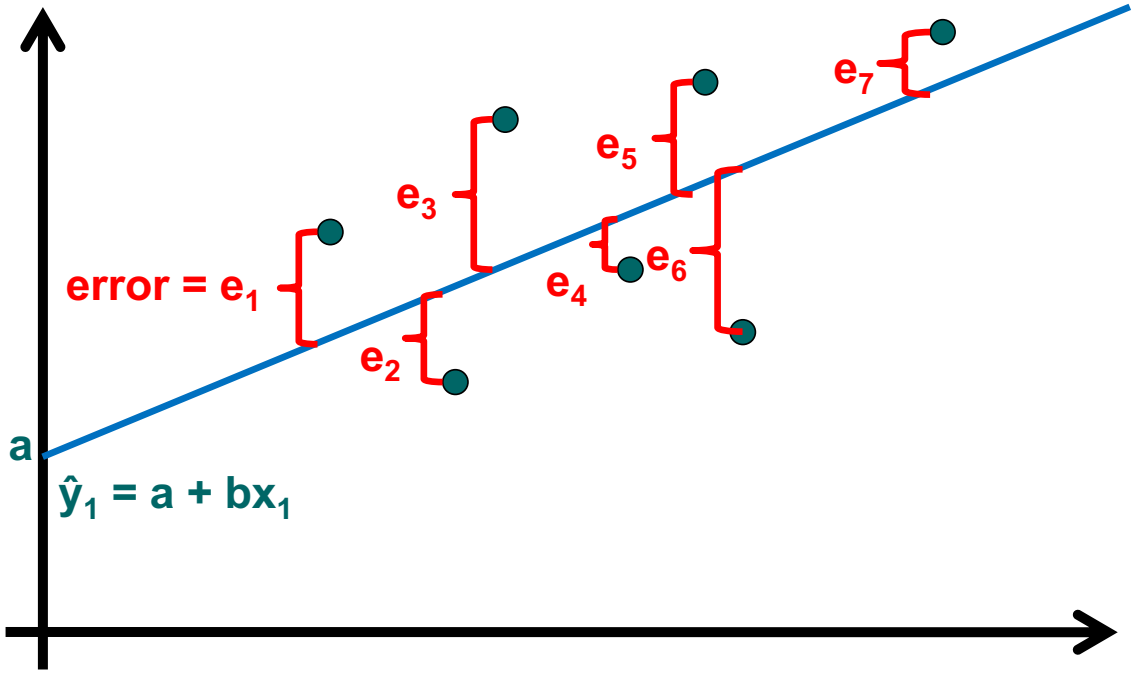
Why are they different?
So, how do we estimate the parameters?

Agenda

1. Introduction to the simple linear regression
2. Estimating the regression: Method of Least Squares
3. Assessing the Fit of the Simple Linear Regression Model
4. Confidence Intervals
5. Extrapolation

Line of best fit

It aims to predict all the y values



Measure how well the line fits using the error distance in the y direction – how far and above the line each point is

- Estimate by the sum of the squared differences
- Best fit will be the smallest value

Regression Line

The **Least squares regression line:**

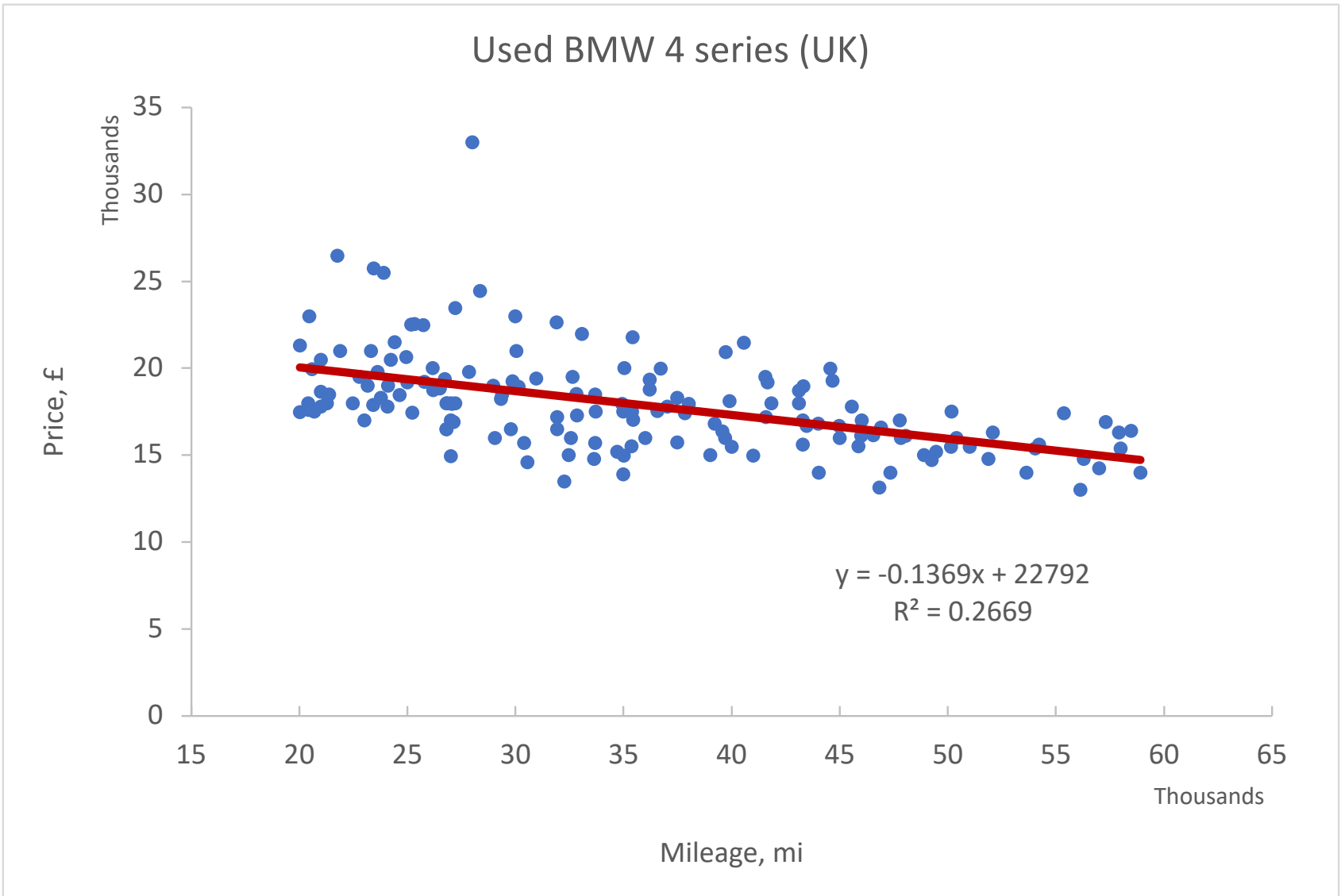
- Minimises **the sum of squares due to error**

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

  - y_i is the **observed value** of the data point
  - $\hat{y}_i$ is the **estimated value** of the line
- squaring the differences stops positive and negative values from cancelling each other out
- squaring the distances penalise large distances from the line
- otherwise tends to follow the greatest concentration of points rather than going through the middle

Important: switching x and y alters the equation of the regression line

Regression Line in Excel



Comments

Line looks sensible

- always plot to check it looks okay

Interpretation of constants (coefficients):

- **Intercept**, $a = 22,792$: **if** the line applies down to x = 0, this is the average price of mileage is close to zero (new?)
- **Gradient**, $b = -0.1369$: the price reduction for each **unit** of mileage

Note: **We are using Excel**, but the coefficients can be calculated using

$$b = r \times \frac{s_Y}{s_X}$$

$$a = \bar{y} - (b \times \bar{x})$$

Agenda

1. Introduction to the simple linear regression
2. Estimating the regression: Method of Least Squares
3. Assessing the Fit of the Simple Linear Regression Model
4. Confidence Intervals
5. Extrapolation

Coefficient of determination R^2

We have two models:
Price (£) = 22,792 - 0.137 Mileage (mi)
Price (£) = 26,084 - 144 Mpg (mi per gallon)
Which one to use? Which one explains more the available data?

R^2

Coefficient of determination R^2 measures the % of total variation explained

$$R^2 = \frac{\textit{Explained Variation}}{\textit{Total Variation}} = \frac{\sum(y_i - \bar{y})^2 - \sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2} \tag{6}$$

$$0 \leq R^2 \leq 1$$

What is a good R^2?

R^2 interpretation

How "**good**" is the model?

- R^2 (square of correlation coefficient, r) gives a good indication of how closely the straight line fits the points
- R^2 is **always between 0 and 1**
  - closer to 1 means a better fit
- R^2 measures the proportion of variation in the data that the regression equation can explain
- but it's not the only possible measure of the model
  - adjusted R^2 (see multiple regression)
  - hypothesis tests
  - Information Criteria (AIC or BIC)
- Our examples:
  - Price (£) = 22,792 - 0.137 Mileage (mi): $R^2 = 0.267$
  - Price (£) = 26,084 - 144 Mpg (mi per gallon): $R^2 = 0.232$

So which model is better?

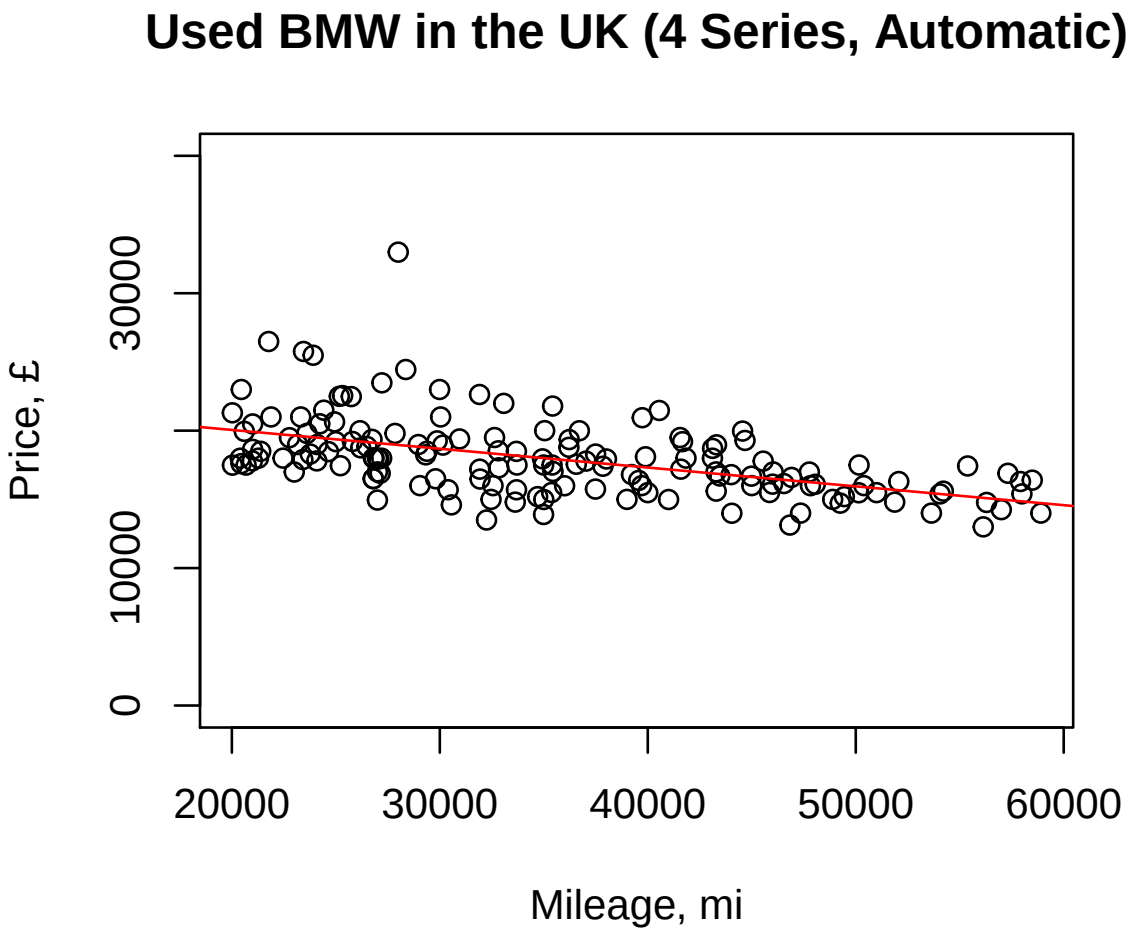
Our examples:

- Price (£) = 22,792 - 0.137 Mileage (mi): $R^2 = 0.267$
- Price (£) = 26,084 - 144 Mpg (mi per gallon): $R^2 = 0.232$

Agenda

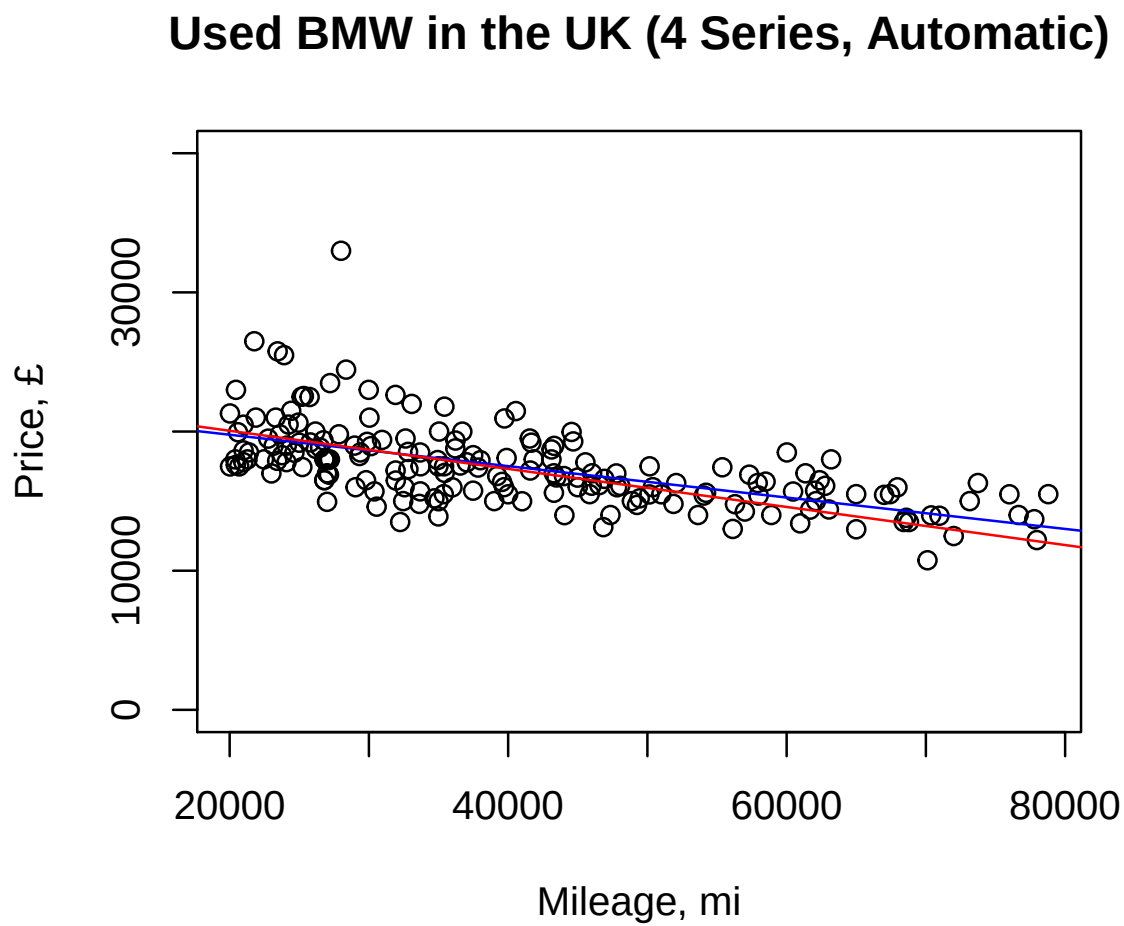
Confidence Intervals

Let's say that we have the following data and the regression line:
Price (£) = 22,792 - 0.137 Mileage (mi)



Confidence Intervals

Then we add more observations (a different sample from the same population) and get another line:



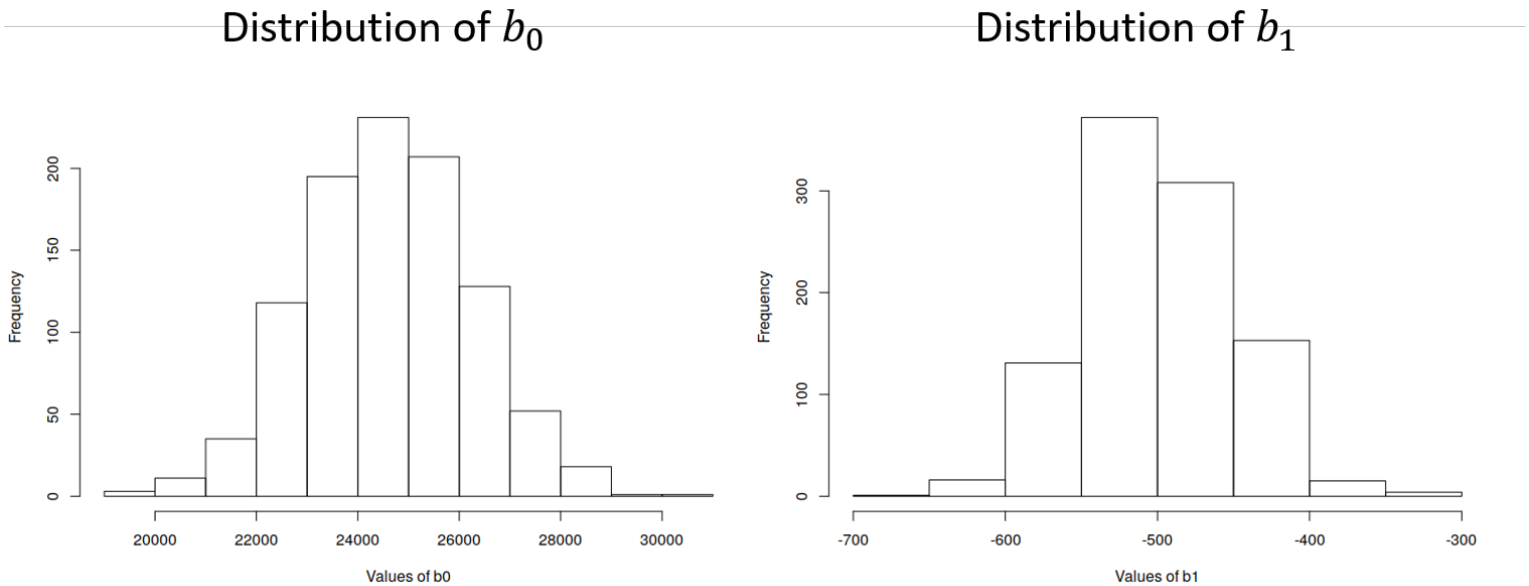
So, which one is correct?

Confidence intervals

- We **cannot** make conclusions based only on the regression line;
- We need to make sure that the parameters are meaningful (**not zeroes**);
- We deal with a **sample**, so we deal with **uncertainty**:
  - The parameters represent the values for a specific sample;
  - This also implies uncertainty in the parameters estimation;
- We need to be able to measure the uncertainty.

Confidence intervals

By adding or removing data, we will have a distribution of parameters (a collection of parameters obtained using different samples):



We can calculate standard errors of the parameters (something like standard deviation).

Confidence intervals

Definition

A **confidence interval** for a regression parameter β_i is an estimated interval believed to contain the true value of β_i at some level of confidence.

We can construct confidence intervals for parameters:

$$b_j \pm t_{\alpha/2} s_{b_j}, \tag{7}$$

- b_j is the point estimate of the regression parameter β_j;
- s_{b_j} is the estimated standard deviation of b_j and
- $t_{\alpha/2}$ is the t value based on the sample size and specified $100(1-\alpha)\%$ confidence level of the interval.

Regression in Excel: raw output

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.516617					
R Square	0.2668931					
Adjusted R Square	0.2621634					
Standard Error	2413.852					
Observations	157					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	1	328793295.6	328793296	56.428913	4.3259E-12	
Residual	155	903135615.3	5826681.4			
Total	156	1231928911				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	22792.242	671.829	33.926	0.000	21465.120	24119.364
Mileage	-0.137	0.018	-7.512	0.000	-0.173	-0.101

Regression in Excel: comments

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.516617					
R Square	0.2668931					
Adjusted R Square	0.2621634					
Standard Error	2413.852					
Observations	157					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	1	328793295.6	328793296	56.428913	4.326E-12	
Residual	155	903135615.3	5826681.4			
Total	156	1231928911				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	22792.242	671.829	33.926	0.000	21465.120	24119.364
Mileage	-0.137	0.018	-7.512	0.000	-0.173	-0.101

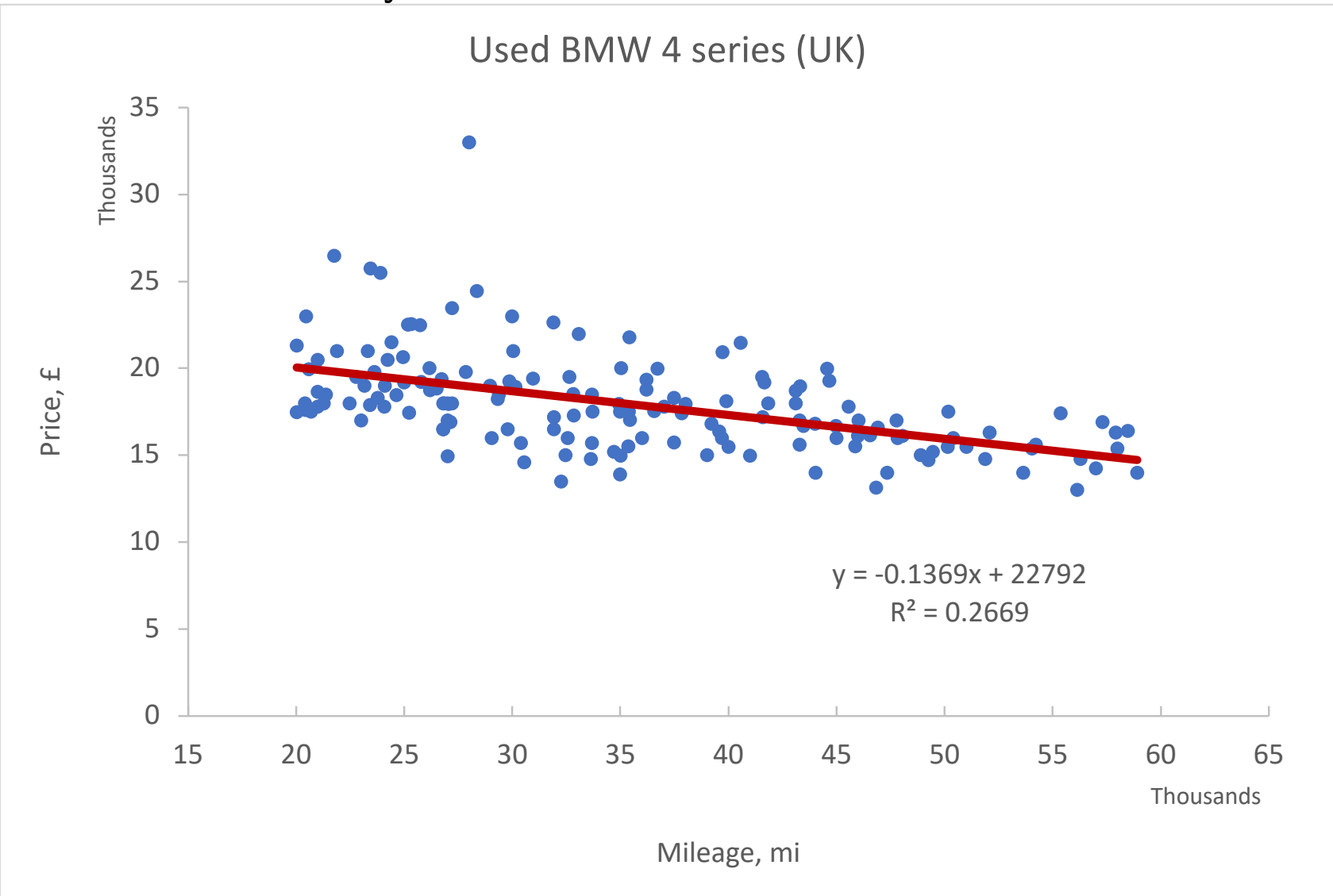
So how do we know if a variable is meaningful?

Agenda

1. Introduction to the simple linear regression
2. Estimating the regression: Method of Least Squares
3. Assessing the Fit of the Simple Linear Regression Model
4. Confidence Intervals
5. Extrapolation

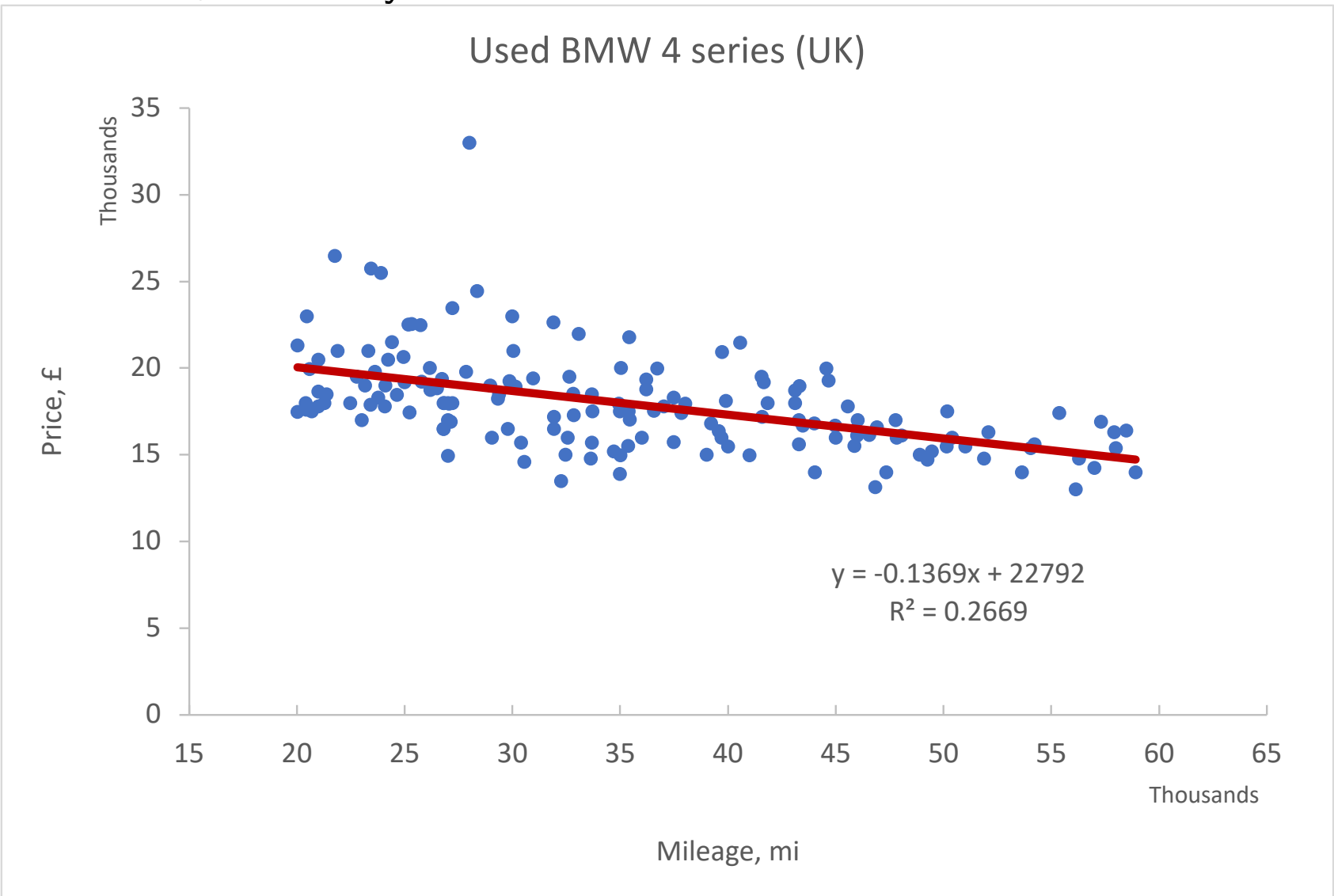
Predictions: Assume a causal relationship!

Predict the price if mileage is around 50,000 miles,
so x = 50,000 and y = ?



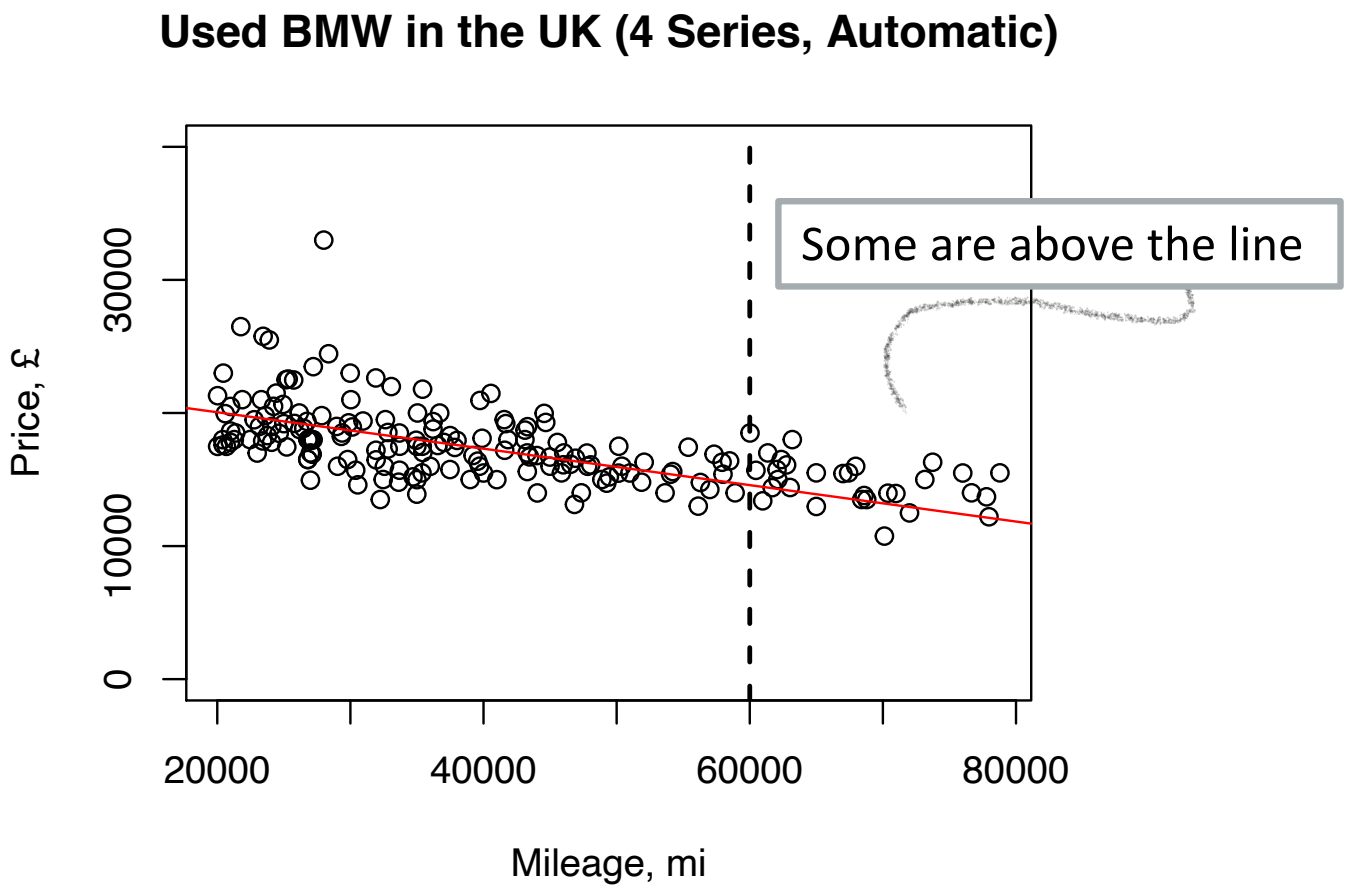
Predictions: Assume a causal relationship!

Predict the price if mileage is around 65,000 miles,
so x = 65,000 and y = ?

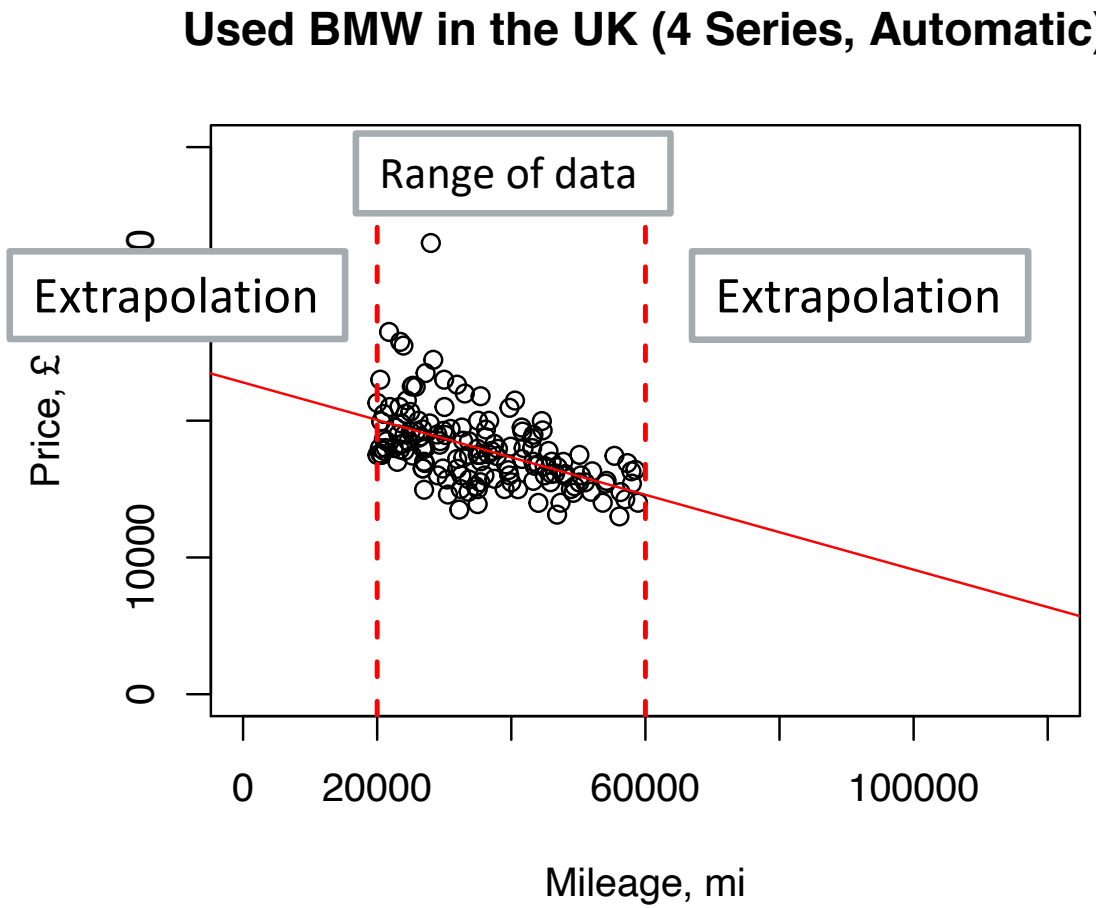


Extrapolation

Data is often approximately linear within a certain range but curves over a wider range

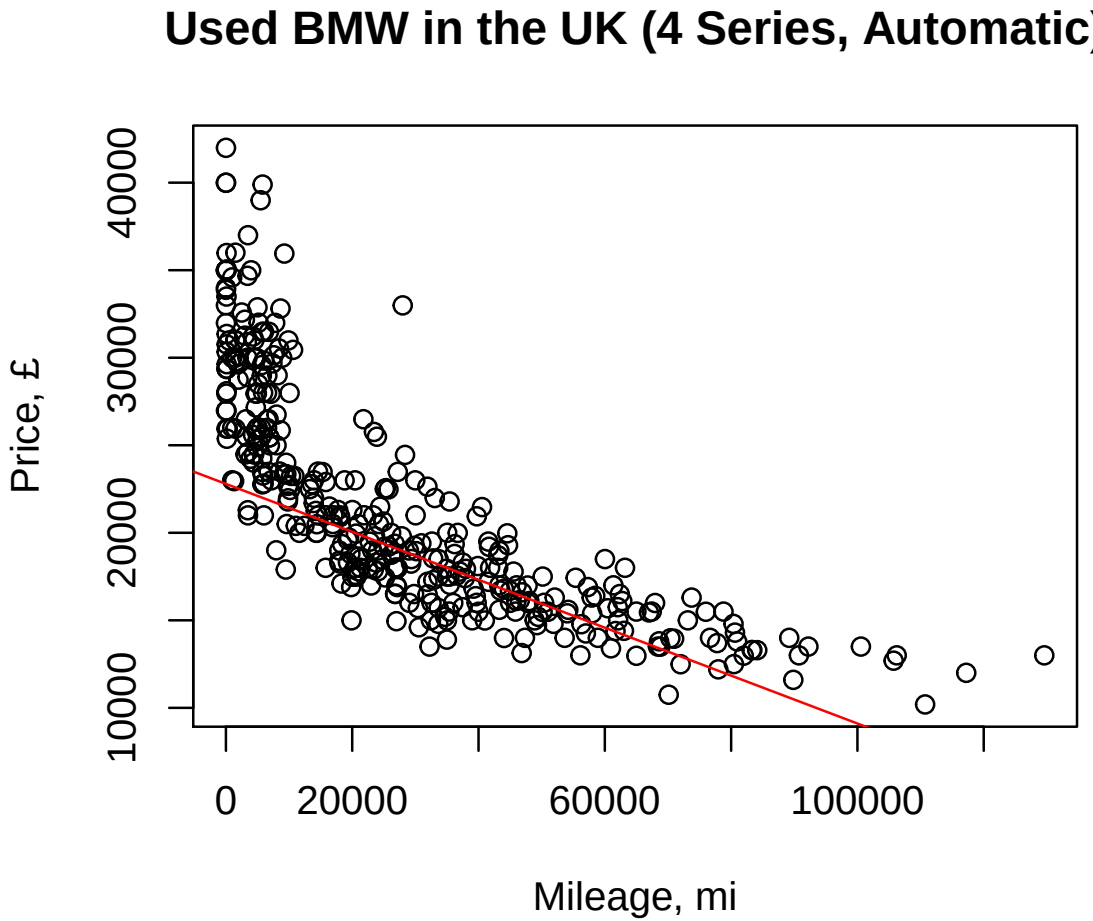


Extrapolation



Be less confident about predictions outside the range of the data

Real data



What do you think about this regression line now?

Summary

1. Draw a scatter chart of data
   - Look for patterns, curves, outliers
2. Calculate correlation
3. Use knowledge of situation to interpret
   - Correlation **does not necessarily imply** cause and effect
4. Use regression to fit best line
5. Regression line for prediction and decision making
   - Much less confidence when extrapolating beyond data

Wrap up

Here we:

- Modelling relationships between two variables: **Simple linear regression**

Next time:

- **Multiple Regression**