MSCI152: Introduction to Business Intelligence and Analytics

Lecture 10: Multiple Linear Regression

Dr Anna Sroginis

Lancaster University Management School

Agenda

"All models are wrong, but some are useful" George Box

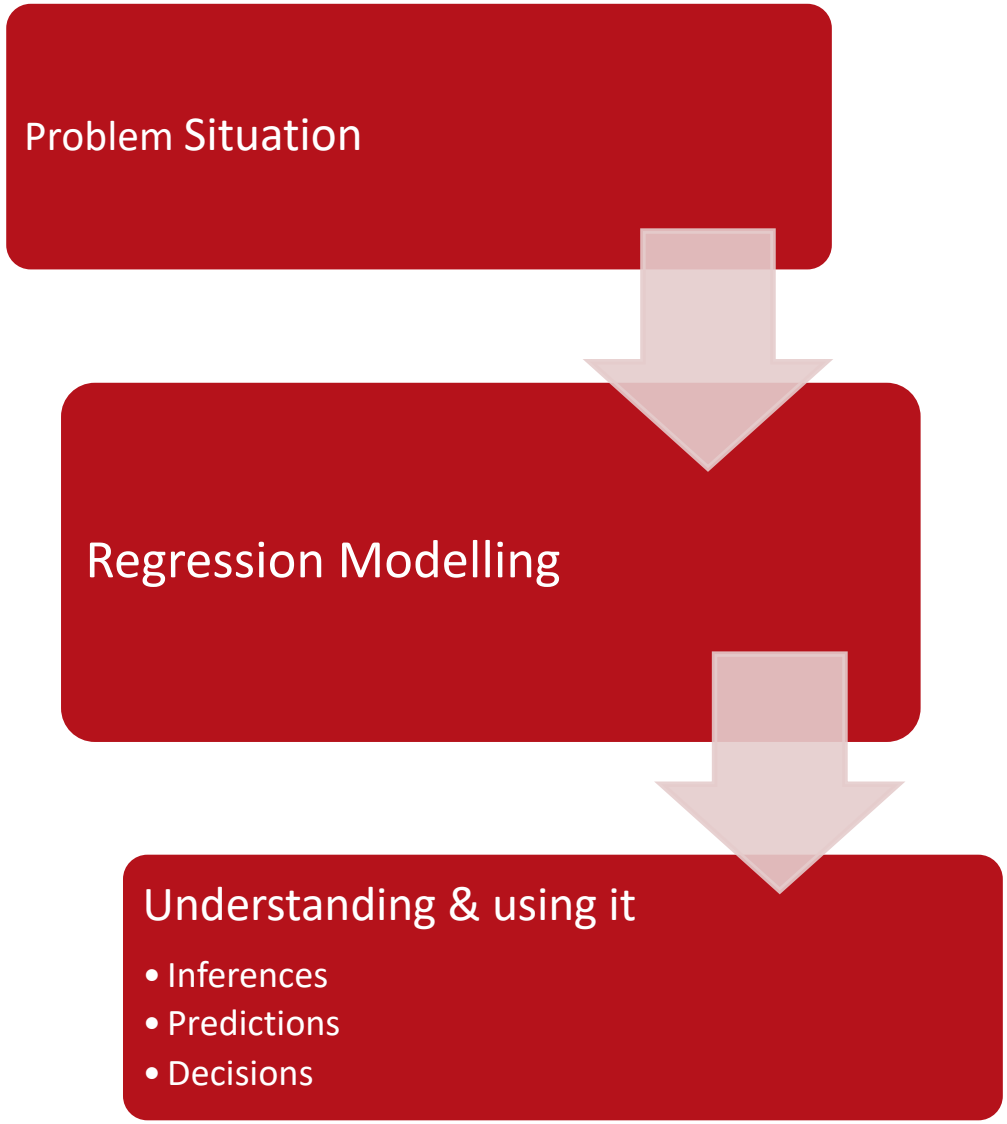
1. Recap
2. Statistical model building
3. Multiple linear regression
4. Building a regression model
   - Plot your variables
   - Fit a model
   - Validate your model
   - Test and interpret your coefficients
   - Consider alternatives
5. Example

More details can be found in Camm et al., Section 7.4 & 7.5

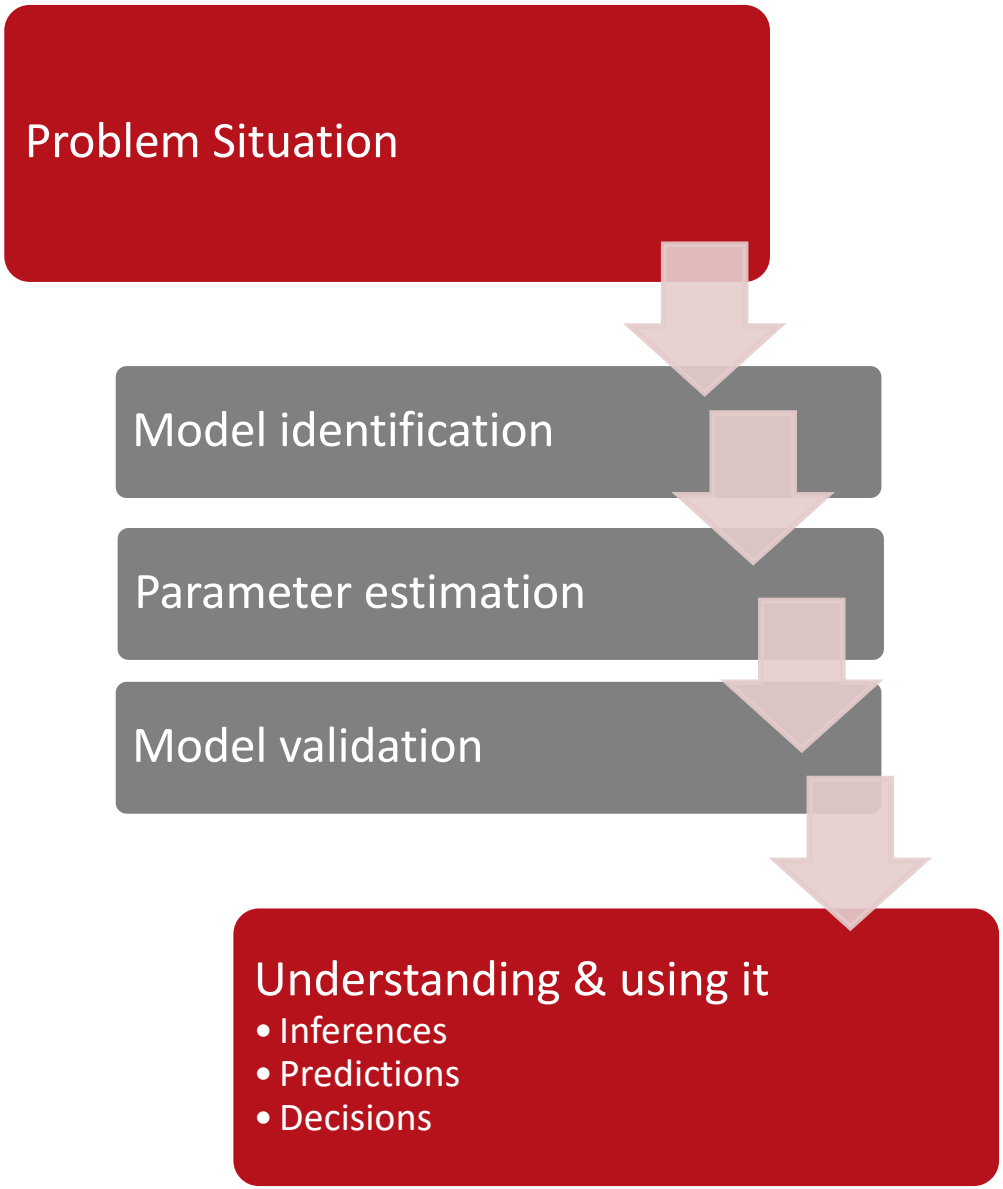
Recap of the previous lecture

- Simple linear regression gives an average estimate of linear relation:
  - $\hat{y}_j = b_0 + b_1 x_j + e_j$, where
    - y_j is a dependent (response) variable (the one that we want to model/predict);
    - x_j is an independent variable (explanatory);
- There is an uncertainty around the model
- Confidence intervals allow reducing this uncertainty
- We can measure the quality of a fit of a model: R^2
- But we should be careful with extrapolation (over/under fitting)

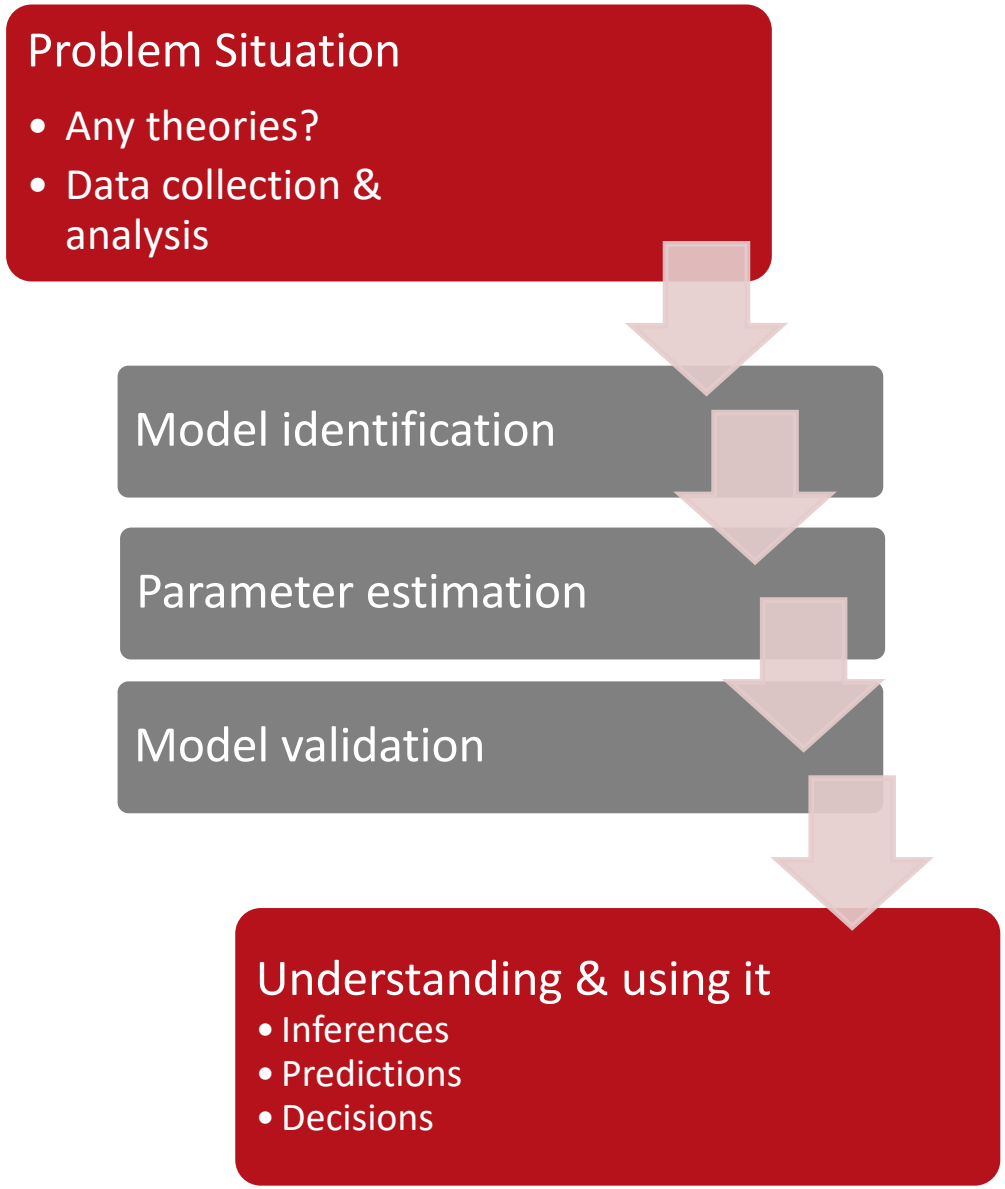
Statistical Model Building



Statistical Model Building



Statistical Model Building



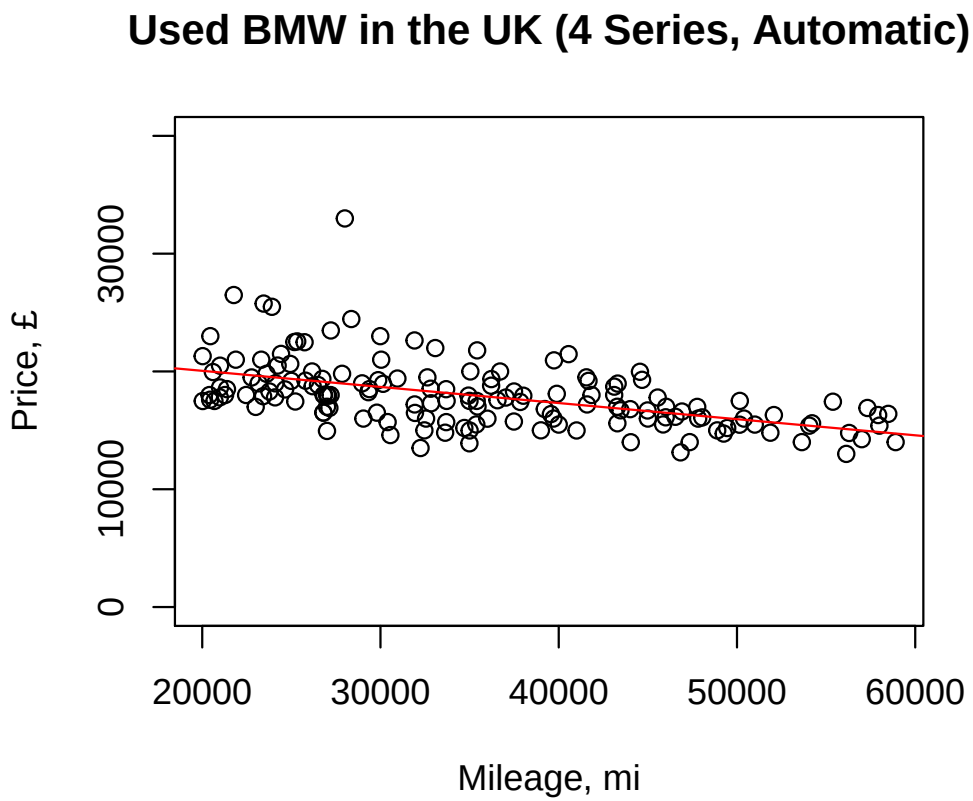
You might need to revise and repeat it several times!

Agenda

1. Recap
2. Statistical model building
3. Multiple linear regression
4. Building a regression model
    - Plot your variables
    - Fit a model
    - Validate your model
    - Test and interpret your coefficients
    - Consider alternatives
5. Example

Additional variables

- Remember the model of **Price and Mileage** (used BMW example)?
- What if we also have information about the *number of cylinders, year, tax, mpg and fuel type*?
  - Any other potential variables?



Multiple linear regression

- We can include those variables in the model to obtain the multiple linear regression.
- In the population it will be:

$$y_j = \beta_0 + \beta_1 x_{1,j} + \beta_2 x_{2,j} + ... + \beta_{k-1} x_{k-1,j} + \epsilon_j, \qquad (1)$$

where

- y_j is is a dependent (response) variable (the one that we want to model/predict);
- $x_{1,j}$, $x_{2,j}$, ..., $x_{k-1,j}$ are independent variables (explanatory);
- k is the number of estimated parameters (k is smaller than sample size);
- j is the observation number;
- every parameter $\beta_1, \beta_2, ..., \beta_{k-1}$ are slopes for the respective variables.

Multiple linear regression

- Now the regression equation is:
  - Population:

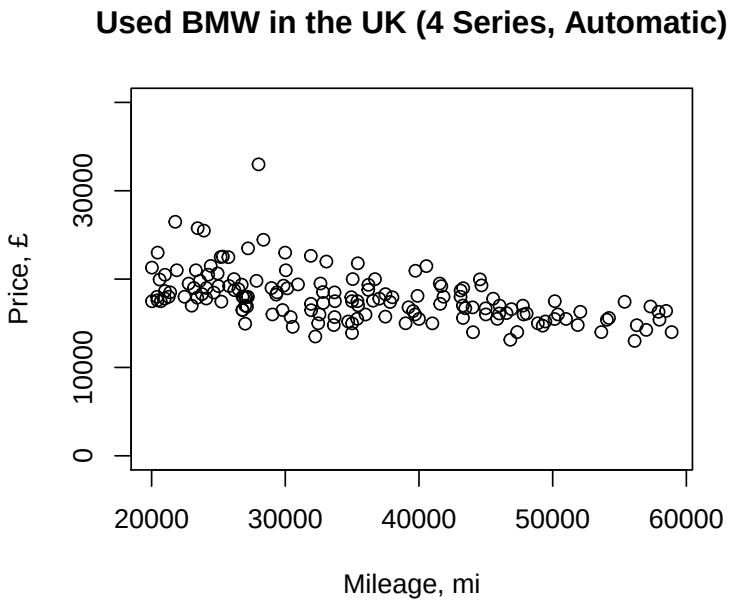
$$E(y_j|x_j) = \beta_0 + \beta_1 x_{1,j} + \beta_2 x_{2,j} + ... + \beta_{k-1} x_{k-1,j} \quad (2)$$

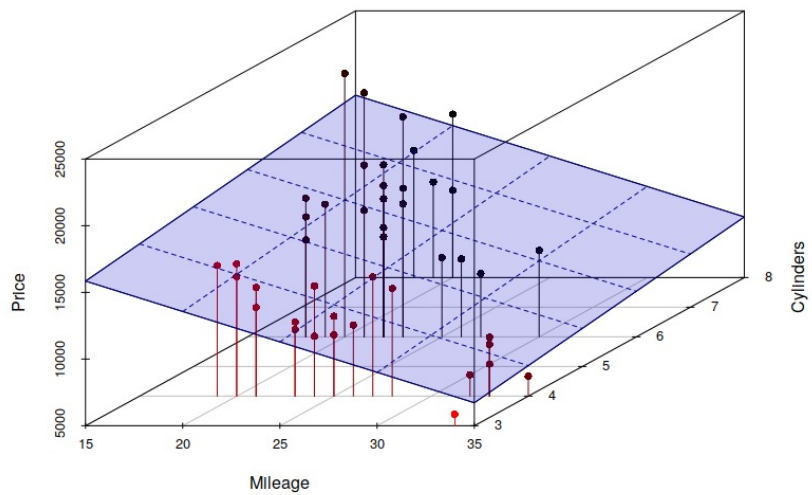
  - Sample:

$$\hat{y}_j = b_0 + b_1 x_{1,j} + b_2 x_{2,j} + ... + b_{k-1} x_{k-1,j} \quad (3)$$

Multiple linear regression

- Very similar to regression with 1 explanatory variable
- Predict y using k independent variables: $x_1, x_2, \ldots, x_k$
- Fitting a k dimensional flat surface:
  - 2D vs 3D





- Still minimises the sum of squared differences between y and $\hat{y}$
- Just as easy to do in Excel or stats package as with 1 explanatory variable

Agenda

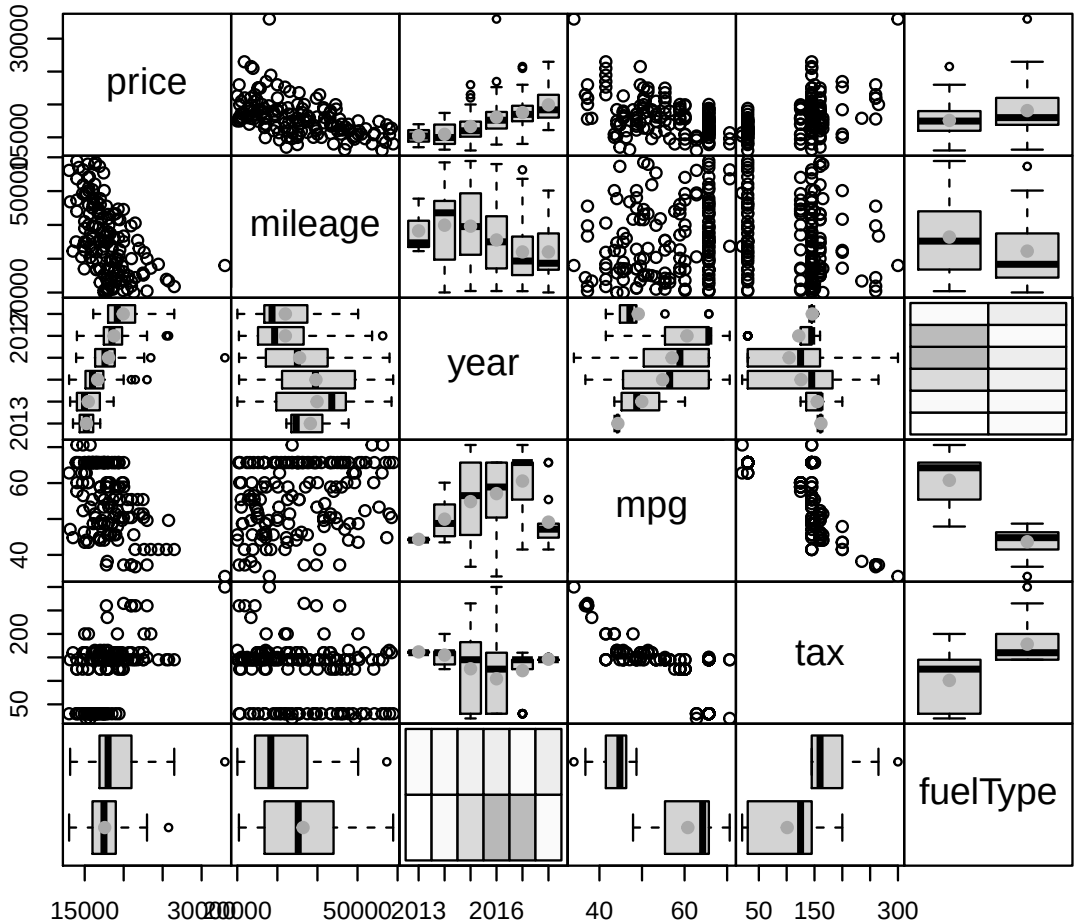
1. Recap
2. Statistical model building
3. Multiple linear regression
4. Building a regression model
   - Plot your variables
   - Fit a model
   - Validate your model
   - Test and interpret your coefficients
   - Consider alternatives
5. Example

General Approach

1. Plot charts for each variable
   - As before, look for the shape of the relationship and outliers
   - But, shape may be obscured by the effect of other variables
2. Think about what variables to include and how
3. Use Excel or a stats package to fit a regression equation
4. Validate your model
5. Use Excel output to assess the strength of the relationship overall and for each variable (parameter estimation)
   - Any statistically insignificant or missing variables? Wrong specification?
6. Consider alternative models
   - We have to decide which variables to include, so there are lots of choices

Multiple regression for used BMW prices

1. Plot charts for each variable (multiple scatter plots)



2. What independent variables are we going to include? What variables might be omitted here?

Note: this plot is done using R, Excel doesn't plot it automatically

Building regression models

❸ Model 1

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.6967544
R Square	0.4854667
Adjusted R Sc	0.4719263
Standard Erre	2042.1037
Observations	157

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	598060408	149515102	35.853328	4.426E-21
Residual	152	633868503	4170187.5		
Total	156	1.232E+09			

	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	28554.638	2441.019	11.698	0.000	23731.931	33377.344
mileage	-0.113	0.016	-7.133	0.000	-0.144	-0.082
tax	7.795	4.240	1.838	0.068	-0.583	16.173
mpg	-165.807	41.114	-4.033	0.000	-247.036	-84.578
fuelType	2406.159	644.505	3.733	0.000	1132.815	3679.504

Before looking at parameters

Regression models should not violate their assumptions. How can we check this?

The first assumptions are satisfied by construction, as long as the relationship between inputs and target is linear. The others can be checked either visual or using statistical tests.

4. We need to make sure that our model is statistically valid:
   1. The errors have zero means $\rightarrow$ The model is unbiased
   2. The error terms have equal variances ($Var(e) = \sigma_e^2$)
   3. Errors are independent of each other (and of everything else!)

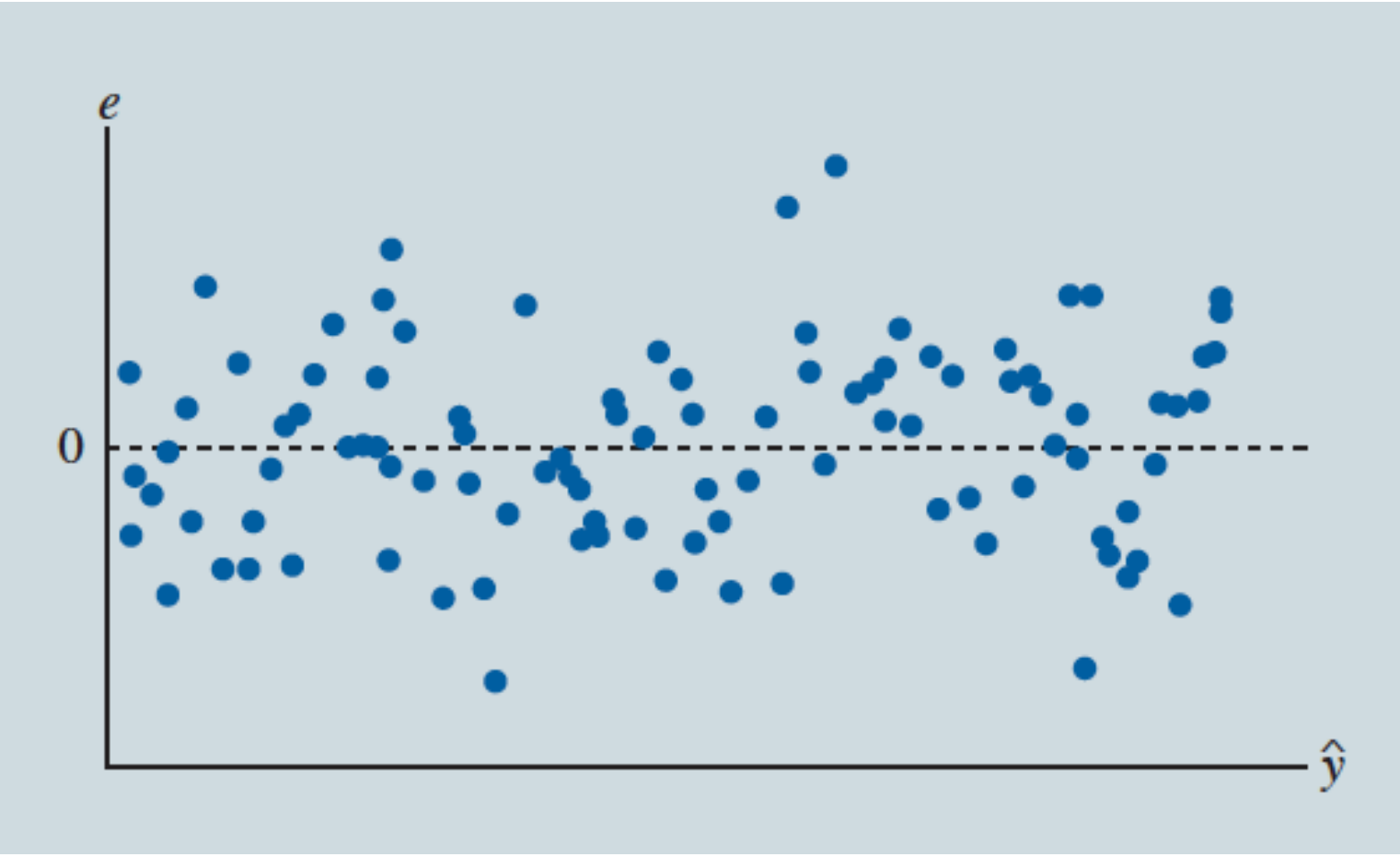
Check

e $\approx$ N(0, σ_e) and statistically independent
Note: This does NOT mean that y_j is normally distributed

Analysis of residuals is important since if some of the assumptions do not hold, then all estimates might be misleading!

Analysis of residuals

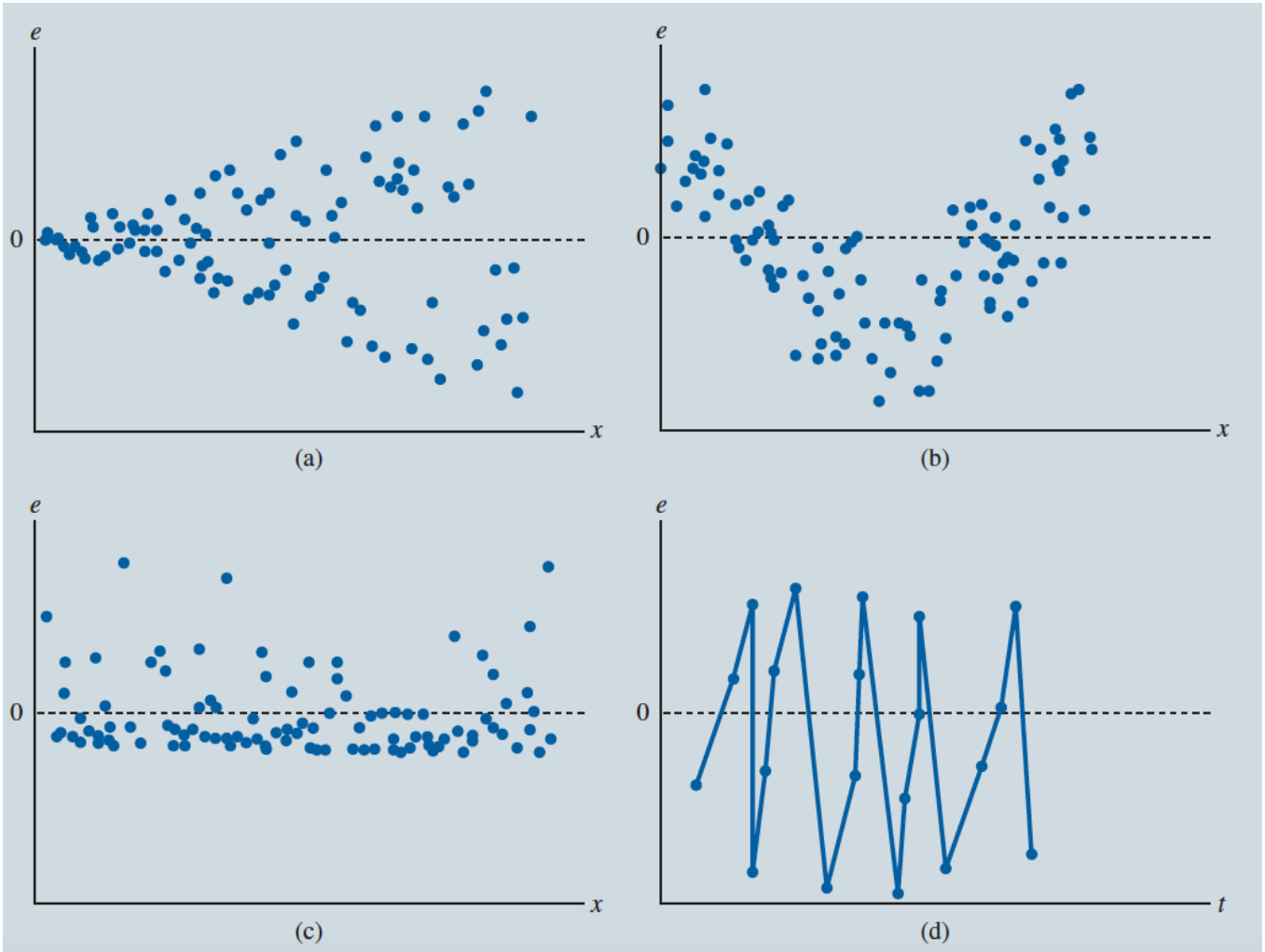
How random errors should look like:



No discernible pattern!

Analysis of residuals

How random errors should **NOT** look like:

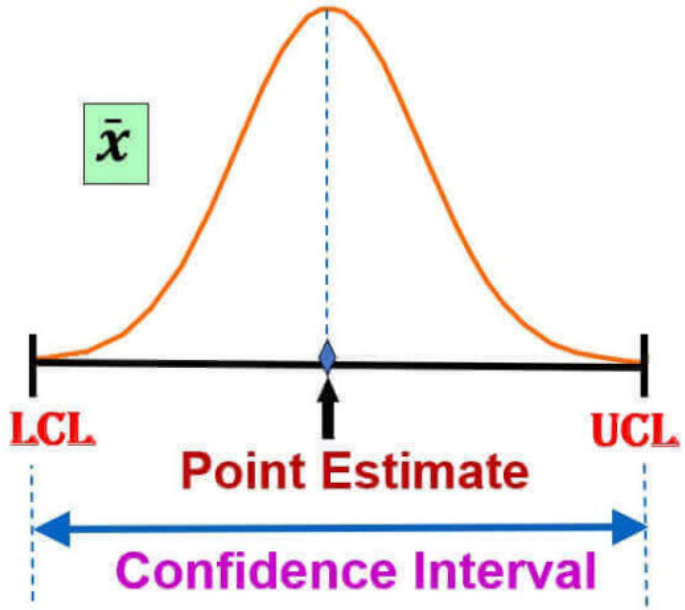


Observe distinct patterns, each suggesting a violation of at least one of the regression model conditions.

Testing and interpreting coefficients

5. If the residuals provide little evidence that our regression model violates the error assumptions necessary for reliable inference, then **we can test the parameters:**

    - There is hypothesis testing for each parameter, but it is **NOT** covered in this module
    - We use confidence intervals to assess our parameters: if zero not included in your confidence intervals, your b is significantly different from zero.



Testing and interpreting coefficients

Interpreting the slope coefficients:

- For the specification $y_j = \beta_0 + \beta_1 x_{1,j} + \beta_2 x_{2,j} + \epsilon_j$:
  - β_0 is estimated value of y_j when all independent variables are zeroes
  - β_j is the **average** effect on y_j of a one unit increase in x_j, **holding all other predictors fixed**
  - An increase/decrease of one unit in the value of x_1 produces an increase/decrease of β_1 units in the *expected value of* y_j - but **always think about units**

Non-significant independent variables

- **What do we do when some confidence intervals include zero?**
- Do we use the model as originally formulated with the non-significant independent variables, or
- Do we rerun the regression without it?

  To answer, consider:
  - a theoretical basis for your regression (practical experience);
  - does your model sufficiently explain the dependent variable without the insignificant ones?

  Note: the estimates of the other regression coefficients may change considerably when we remove the non-significant independent variables.

Building regression models

6 Model 2

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.68849654
R Square	0.47402748
Adjusted R S	0.46371429
Standard Err	2057.9207
Observations	157

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	583968156	1.95E+08	45.9632	3.0978E-21
Residual	153	647960755	4235038		
Total	156	1231928911			

	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	32282.796	1369.023	23.581	0.000	29578.167	34987.425
mileage	-0.114	0.016	-7.170	0.000	-0.146	-0.083
mpg	-218.392	29.760	-7.339	0.000	-277.185	-159.600
fuelType	2709.616	627.831	4.316	0.000	1469.278	3949.954

Comparing regression models

Model 1

	Coefficients	tandard Erro	t Stat	P-value	Lower 95%	Upper 95%
Intercept	28554.638	2441.019	11.698	0.000	23731.931	33377.344
mileage	-0.113	0.016	-7.133	0.000	-0.144	-0.082
tax	7.795	4.240	1.838	0.068	-0.583	16.173
mpg	-165.807	41.114	-4.033	0.000	-247.036	-84.578
fuelType	2406.159	644.505	3.733	0.000	1132.815	3679.504

According to CI (95%), a parameter for "tax" is likely to be zero, so reject this variable from the model. Remove it and re-estimate.

Model 2

	Coefficients	tandard Erro	t Stat	P-value	Lower 95%	Upper 95%
Intercept	32282.796	1369.023	23.581	0.000	29578.167	34987.425
mileage	-0.114	0.016	-7.170	0.000	-0.146	-0.083
mpg	-218.392	29.760	-7.339	0.000	-277.185	-159.600
fuelType	2709.616	627.831	4.316	0.000	1469.278	3949.954

Observe how the coefficients have changed.

- Is this model reasonable? Better than Model 1?

Comparing models

1. **Adjusted R^2**
   - Due to its structure, R^2 increases as the number of variables increases - **Always!**
     - I can even add random variables and get the same effect!
   - So for multiple linear regression, we can adjust R^2 to penalise additional variables
     - Each additional variable must introduce useful information
     - Useful to compare models with different number of variables

$$AdjR^2 = \frac{n-1}{n-k-1}\left(R^2 - \frac{k}{n-1}\right), \tag{4}$$

   where n is a number of observations and
   k is a number of parameters
2. **Standard error S**
   - This is the unbiased estimator of the in-sample 1-step ahead MSE (the smaller, the better).
3. **Information criteria (AIC, BIC)**
   - These balance *the goodness of fit* (Mean Squared Error) with the model complexity (k) - this is a modern way to compare models, but not covered in this module!

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. Validate your regression model
6. Check your coefficients, can you interpret them?
7. Consider alternatives
8. Regression line for prediction and decision making - next lecture!
   - Much less confidence if extrapolate beyond data

Some uses of regression

Identify **relationship** between the variables

- Regression equation

Identify **strength of relationship** between variables

- R^2 value, chart
- With multiple regression we can compare variables

Make **predictions**

- Predict what will happen for a specific x value
- Can feed into a decision-making model

Use as a **benchmark**

- Compare actual values with regression line to identify those performing well or badly

Agenda

1. Recap
2. Statistical model building
3. Multiple linear regression
4. Building a regression model
    - Plot your variables
    - Fit a model
    - Validate your model
    - Test and interpret your coefficients
    - Consider alternatives
5. Example

Example: Advertising spending vs Sales

TV, $000	Radio, $000	Newspaper, $000	Sales, 000 units
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	9.3
151.5	41.3	58.5	18.5
180.8	10.8	58.4	12.9
8.7	48.9	75	7.2
57.5	32.8	23.5	11.8
120.2	19.6	11.6	13.2
8.6	2.1	1	4.8
199.8	2.6	21.2	10.6

Based on this data, you are asked to suggest a marketing plan for next year that will result in high product sales. How are you going to do so? Discuss it in your groups.

Initial questions

- Is there a relationship between advertising budget and sales?
- How strong is the relationship between advertising budget and sales?
- Which media are associated with sales?
- How large is the association between each medium and sales?
- Is the relationship linear?

One possible model is:

$$sales = \beta_0 + \beta_1 TV + \beta_2 radio + \beta_3 newspaper + \epsilon. \qquad (5)$$

Correlation

	TV	radio	newspaper	sales
TV	1.000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

Regression model

What would you think about this model?

$$sales = 2.939 + 0.046\,TV + 0.189 radio - 0.001 newspaper \quad (6)$$

Useful questions:

- Is at least one of the predictors X1,X2, . . . ,Xp useful in predicting the response?
- Do all the predictors help to explain Y, or is only a subset of the predictors useful?
- How well does the model fit the data?
- Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

Wrap up

Here we:

- Modelling relationships between two and more variables: **Multiple linear regression**

Next time:

- **More multiple linear regression**