MSCI152: Introduction to Business Intelligence and Analytics

Lecture 12: Multiple Linear Regression

Dr Anna Sroginis

Lancaster University Management School

Agenda

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

1. Recap

2. Example: Wage data

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

Recap of the previous lectures

- Multiple linear regression gives an **average estimate of linear relation**:
  - $\hat{y}_j = b_0 + b_1 x_{1,j} + b_2 x_{2,j} + ... + b_{k-1} x_{k-1,j}$, where
    - y_j 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);
- As soon as you fit any model, you need to **validate this model**
- Confidence intervals help you to **test and interpret the coefficients**
- We can **measure the quality of a fit of a model**: Adjusted R^2 is better for multiple regression
- But we should be careful with **over/under fitting**

Where can we use regression analysis?

- **Economics**: analysing the relationships between different economic factors such as GDP, inflation, unemployment, and consumer spending
- **Finance**: asset pricing models, risk assessment, and portfolio management to understand the relationships between various financial variables
- **Marketing**: consumer behaviour, market trends, and the impact of marketing/advertising strategies on sales
- **Human Resources**: predicting employee turnover, understand factors influencing performance, and optimise workforce management strategies
- **Engineering and Science**: analysing experimental data, quality control, and predicting physical phenomena in various fields such as physics, chemistry, and engineering

Where can we use regression analysis?

- **Healthcare**: it's applied in epidemiology to study the relationships between risk factors and disease incidence, as well as in predicting patient outcomes based on different medical parameters
- **Predictive Analysis**: in business, regression analysis is utilised for forecasting, and predicting sales, demand, and trends to make strategic decisions
- **Sports Analytics**: analysing player performance, team strategies, and predicting game outcomes
- **Urban Planning**: predicting population distribution, traffic patterns, and infrastructure development
- **Criminal Justice**: analysing factors related to crime rates, recidivism, and the effectiveness of various interventions or policies.
- many more...

General Approach

1. Plot charts for each variable
   - As before, look for the shape of relationship and outliers
   - But, shape may be obscured by effect of other variables
2. Think what variables to include and how
3. Use Excel or stats package to fit regression equation
4. Validate your model
5. Use Excel output to assess the strength of 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

Example: Pay Equality

How can we make sure that there is pay equality in the company?
Imagine that you have this dataset:

- **wage** – wage in GBP, daily
- **education** – number of years of education (from primary school)
- **experience** – number of years of work experience
- **age** – age in years
- **ethnicity** – variable, indicating whether the respondent is Caucasian or of another ethnicity
- **region** – variable, showing, whether the respondent works in the south of England or elsewhere in the UK
- **gender** – gender of the respondent
- **occupation** – the occupation of a person. This can be:
  - worker – tradesperson or assembly line worker
  - technical – technical or professional worker
  - services – service worker
  - office – office and clerical worker
  - sales – sales worker
  - management – management and administration

Example: Pay Equality

- What is your response variable?
- Are all explanatory variables numerical?
  - Quantitative variables: ...
  - Qualitative variables: ...
- How would you plot each variable?
  - wage
  - education
  - experience
  - age
  - ethnicity
  - region
  - gender
  - occupation

Example: Pay Equality

- What relationships would you expect between your response and explanatory variables?
  - wage & education
  - wage & experience
  - wage & age
  - wage & ethnicity
  - wage & region
  - wage & gender
  - wage & occupation

Example: Pay Equality

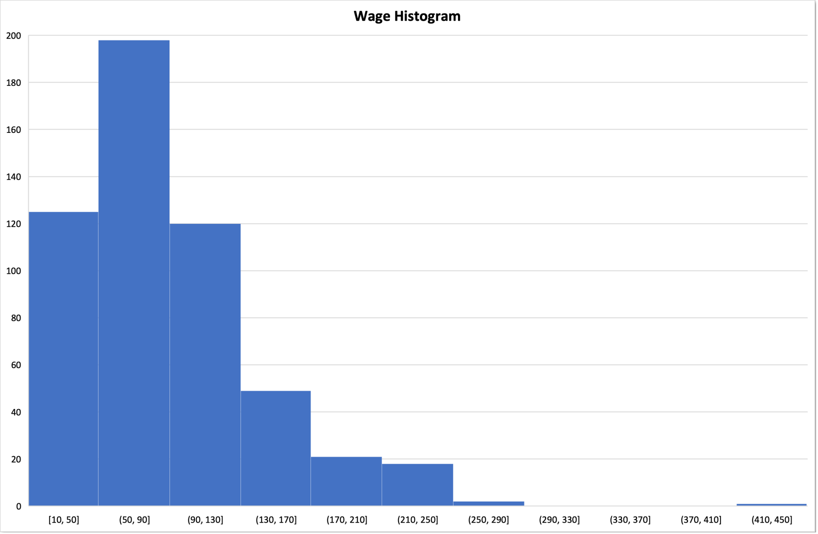
- Would you expect any relationships between any independent variables?
  - education
  - experience
  - age
  - ethnicity
  - region
  - gender
  - occupation

Example: Pay Equality

- Any outliers?
  - education
  - experience
  - age
  - ethnicity
  - region
  - gender
  - occupation
- Any unexpected visual patterns?

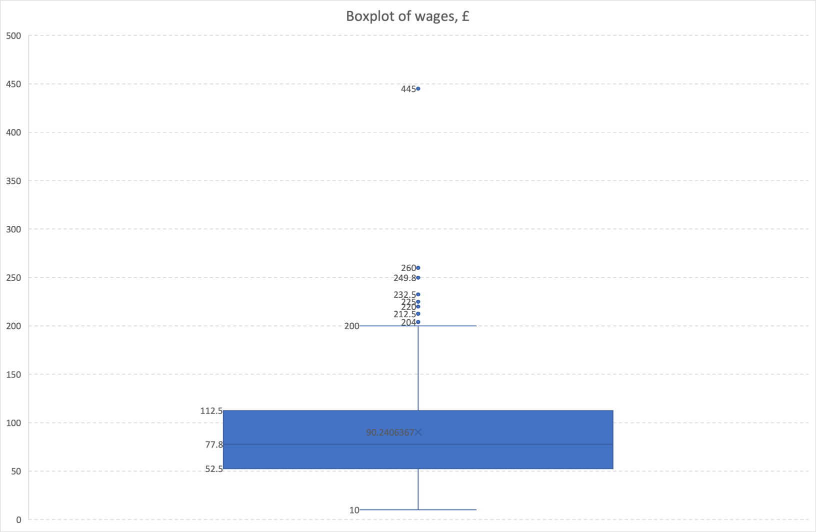
Don't forget to use summary statistics wherever possible!

Wage histogram



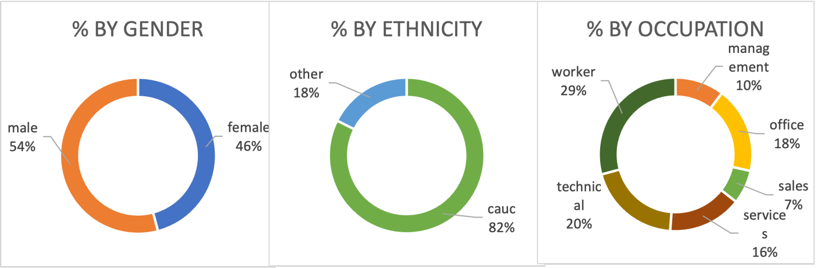
Can you spot any problems with this chart?

Wage boxplot



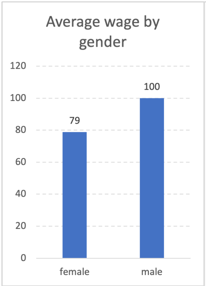
Percentages

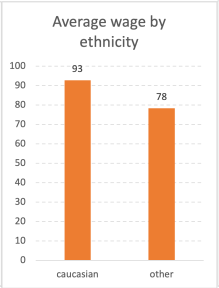
Head count: 534

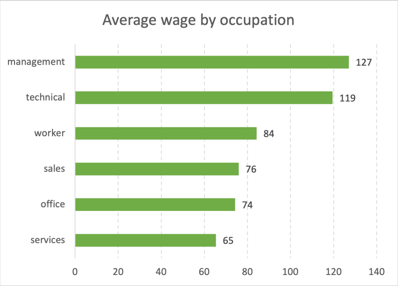


Any insights? Any problems?

Average wages by category

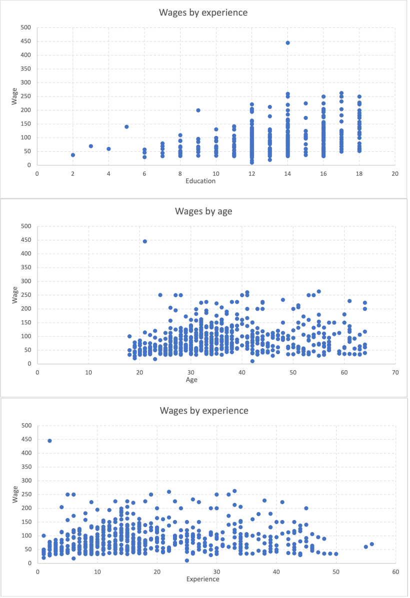






Any insights? Any problems?

Wages by years



Example: Pay Equality

- Correlation Analysis
- Any strong linear associations?

	wage	*education*	*experience*	*age*	*cauc*	*south*	*female*	*managemen*	*office*	*sales*	*services*	*technical*
wage	1											
education	0.38	1										
experience	0.09	-0.35	1									
age	0.18	-0.15	0.98	1								
cauc	0.11	0.12	-0.02	0.01	1							
south	-0.14	-0.14	-0.01	-0.04	-0.12	1						
female	-0.21	0.00	0.08	0.08	0.02	-0.02	1					
management	0.24	0.20	0.01	0.05	0.01	-0.06	-0.05	1				
office	-0.15	-0.01	-0.01	-0.01	-0.04	0.05	0.31	-0.16	1			
sales	-0.08	0.02	0.01	0.02	0.05	0.03	-0.01	-0.09	-0.13	1		
services	-0.21	-0.23	0.08	0.04	-0.11	0.01	0.11	-0.15	-0.20	-0.12	1	
technical	0.28	0.50	-0.09	0.01	0.08	-0.09	0.04	-0.17	-0.23	-0.14	-0.21	1

Example: Pay Equality

- Correlation Analysis
- Any strong linear associations?

	wage	*education*	*experience*	*age*	*cauc*	*south*	*female*	*managemen*	*office*	*sales*	*services*	*technical*
wage	1											
education	0.38	1										
experience	0.09	-0.35	1									
age	0.18	-0.15	0.98	1								
cauc	0.11	0.12	-0.02	0.01	1							
south	-0.14	-0.14	-0.01	-0.04	-0.12	1						
female	-0.21	0.00	0.08	0.08	0.02	-0.02	1					
management	0.24	0.20	0.01	0.05	0.01	-0.06	-0.05	1				
office	-0.15	-0.01	-0.01	-0.01	-0.04	0.05	0.31	-0.16	1			
sales	-0.08	0.02	0.01	0.02	0.05	0.03	-0.01	-0.09	-0.13	1		
services	-0.21	-0.23	0.08	0.04	-0.11	0.01	0.11	-0.15	-0.20	-0.12	1	
technical	0.28	0.50	-0.09	0.01	0.08	-0.09	0.04	-0.17	-0.23	-0.14	-0.21	1

Note: if you include two explanatory variables that have a strong correlation between each other, it will cause problems (age and experience). It is better to include only one with the strongest association with a response variable.

Model Pay Equality

Model 0: including education, age, cauc, south, female, occupation dummies
Note: we include *experience*!

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.56
R Square	0.31
Adjusted R S	0.29
Standard Err	43.19
Observation:	534

ANOVA

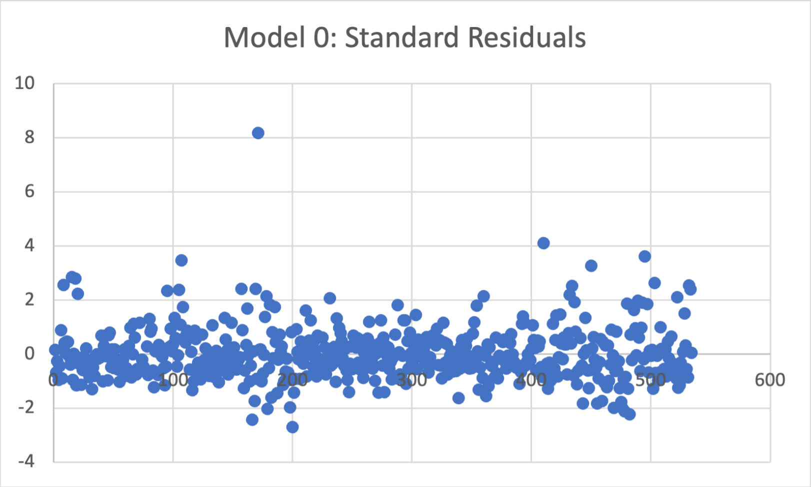
	df	*SS*	*MS*	*F*	*Significance F*
Regression	11	434078.17	39461.65	21.16	0.00
Residual	522	973591.69	1865.12		
Total	533	##########			

	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-10.08	55.92	-0.18	0.86	-119.93	99.78
education	7.24	10.94	0.66	0.51	-14.26	28.74
experience	1.41	10.89	0.13	0.90	-20.00	22.81
age	-0.40	10.88	-0.04	0.97	-21.78	20.98
cauc	6.03	5.02	1.20	0.23	-3.83	15.89
south	-7.36	4.20	-1.75	0.08	-15.60	0.88
female	-20.18	4.17	-4.84	0.00	-28.37	-11.99
managemen	24.43	7.52	3.25	0.00	9.67	39.20
office	-7.38	6.33	-1.17	0.24	-19.82	5.06
sales	-15.80	8.10	-1.95	0.05	-31.71	0.11
services	-13.49	6.15	-2.19	0.03	-25.56	-1.42
technical	14.19	6.94	2.04	0.04	0.55	27.83

- Is it a good model?
- Any insignificant variables?
- Can we validate this model? Residuals analysis

Model 0: residuals

- Residuals Analysis
- Any visual problems?



Model Pay Equality

Model 1: including education, age, cauc, south, female, occupation dummies

Note: we exclude *experience*!

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.56
R Square	0.31
Adjusted R Square	0.30
Standard Error	43.15
Observations	534

ANOVA

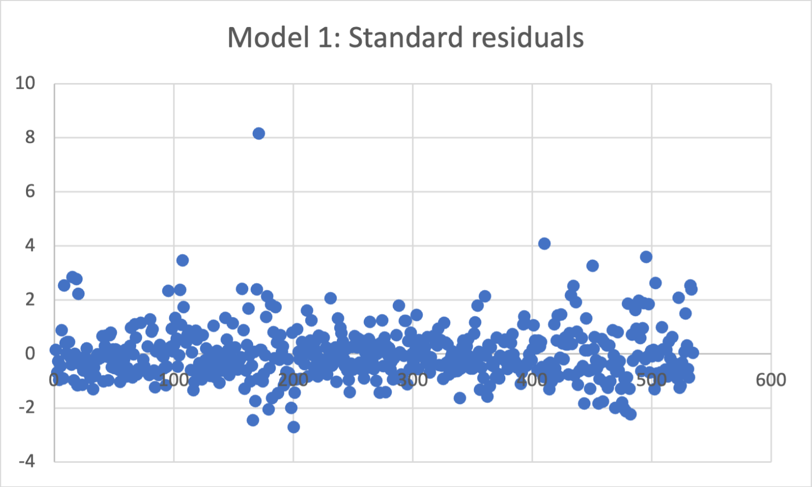
	df	SS	MS	F	Significance F
Regression	10	434047.11	43404.71	23.32	0.00
Residual	523	973622.76	1861.61		
Total	533	1407669.87			

	Coefficients	tandard Erro	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-17.05	14.49	-1.18	0.24	-45.51	11.42
education	5.83	0.95	6.15	0.00	3.97	7.69
age	1.01	0.16	6.11	0.00	0.68	1.33
cauc	6.04	5.01	1.20	0.23	-3.81	15.89
south	-7.37	4.19	-1.76	0.08	-15.60	0.86
female	-20.15	4.16	-4.85	0.00	-28.32	-11.98
management	24.43	7.51	3.25	0.00	9.68	39.18
office	-7.40	6.33	-1.17	0.24	-19.82	5.03
sales	-15.80	8.09	-1.95	0.05	-31.70	0.10
services	-13.50	6.14	-2.20	0.03	-25.56	-1.43
technical	14.24	6.92	2.06	0.04	0.64	27.84

- Is it a good model?
- Any insignificant variables?
- Can we validate this model? Residuals analysis

Model 1: residuals

- Residuals Analysis
- Any visual problems?



Model 2

Model 2: including education, age, female, occupation dummies
Note: we exclude *experience, cauc, south*!

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.55
R Square	0.30
Adjusted R Sq	0.29
Standard Errc	43.27
Observations	534

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	8	424619.684	53077.4605	28.3461284	1.0204E-36
Residual	525	983050.184	1872.47654		
Total	533	1407669.87			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-18.83	13.86	-1.36	0.17	-46.06	8.40
education	6.14	0.94	6.52	0.00	4.29	7.99
age	1.03	0.16	6.24	0.00	0.70	1.35
female	-19.63	4.16	-4.72	0.00	-27.81	-11.45
management	24.13	7.52	3.21	0.00	9.35	38.91
office	-8.51	6.32	-1.35	0.18	-20.93	3.90
sales	-16.26	8.11	-2.00	0.05	-32.19	-0.33
services	-14.28	6.12	-2.33	0.02	-26.30	-2.25
technical	13.95	6.94	2.01	0.04	0.31	27.59

Model 2

Model 2: including education, age, female, occupation dummies
Note: we exclude *experience, cauc, south*!

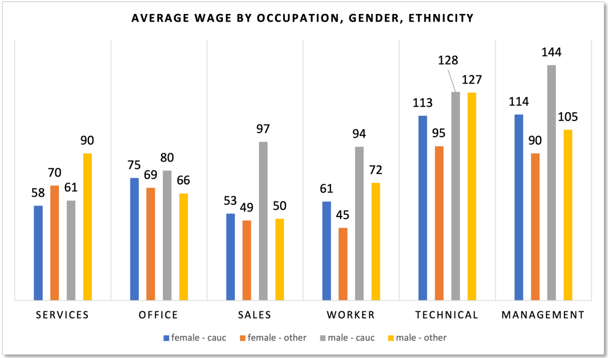
$$\begin{aligned} wage = &\beta_0 + \beta_1 Education + \beta_2 Age + \beta_3 Female + \\ &\beta_4 Management + \beta_5 Office + \beta_6 Sales + \\ &\beta_7 Services + \beta_8 Technical + \epsilon \end{aligned} \tag{1}$$

$$\begin{aligned} wage = &-18.83 + 6.14 Education + 1.03 Age - 19.63 Female + \\ &24.13 Management - 8.51 Office - 16.26 Sales - \\ &14.28 Services + 13.95 Technical + \epsilon \end{aligned} \tag{2}$$

- Any striking insights?
- How would you interpret coefficients?
  - Education
  - Age
  - Female
  - Different occupations

Potential improvements

- Delete an outlier
  - Wage £445 at age 21 and with 2 years of experience! It makes sense to assume that there is some mistake, even though this person is in a management position
- Include an interaction effect for females at different occupations
- Include an interaction effect for ethnicity and occupations



Interpretation of this model

- The model confirms that males earn approximately £20 more than females per day on average.
- Our linear regression doesn't show any effect of ethnicity on wages, even though the initial visualisation claims otherwise. Possibly, it is because of an insufficient sample size (just one-fifth of the workforce).
- We see that employees from "services", "sales", and "office" earn approximately the same. In contrast, "management" and "technical" make significantly more (around £24 or £14 increase on average), compared to workers as the baseline.
- There is no pronounced effect of region on salaries.
- **The average differences in gender pay are alarming and should be carefully reevaluated in this company.**

Wrap up

Here we:

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

Next time:

- **Introduction to forecasting**