Examining several factors together (multivariable)

- In previous sessions, we looked at examining:
  - one factor and comparing it between two different groups of people (or things)
  - the association between two factors, both measured on a single group of people (or things)
- These can be referred to as '**univariate**' or '**univariable**' analyses

- E.g. examining the relationship between a single response variable (blood pressure) and only one other variable (age)
- A **linear regression** (previous session) of blood pressure and age was:

$$BP = 13.3+1.7(Age)$$

- But what if blood pressure is also affected by gender?
- How can we allow for this?

Multivariable Analysis

- Here we look at the same relationships as we may do with a univariable analysis, but we want to <u>simultaneously</u> consider several other factors

- Reasons for doing this could include:
    - Adjusting for confounders when looking at a single risk factor and its effect on the risk of a disease/event
    - Finding a set of prognostic factors that could be used to predict the risk of disease/event
    - To correct for imbalances in subject characteristics in a clinical trial or laboratory experiment
    - To examine interactions between factors

- Multivariable regressions are just an extension of the regression techniques already seen to examine a single factor

Outcome measure		**Method**	**Effect size produced in terms of:**
Taking measurements on people	Continuous data	Multiple linear regression	Mean difference for categorical data or slope for continuous data
Counting people	Binary or categorical data	Multiple logistic regression	Odds ratio
Time-to-event	(not everyone has had event of interest)	Multiple Cox regression	Hazard ratio

Regression models are of the form:

$$Y = a + bX_1 + cX_2 + dX_3$$

Outcome measure (Y)

Can be one of the following:
• Continuous (taking measurements)
• Binary (counting people)
• Ordered categorical (counting people) [but not easy!]
• Time-to-event

Factors to examine (X_1, X_2, X_3) – called covariates

Can be any mixture of:
• Continuous (taking measurements), including time-to-event, but only if everyone has had the event
• Binary (counting people)
• Categorical (counting people)

This is what determines which method to use

Example

- Blood pressure (mmHg) of 50 patients is measured

- We want to know blood pressure is associated with other factors:
  - Age (in years)
  - Gender (male, female)
  - Social Class (low, lower middle, upper middle, high)

- For binary factor, it is best to code as 0 and 1
- For categorical factor, code as 0,1, 2 and 3

Linear Regression Output - **Unadjusted**

Parameter Estimates

Dependent Variable: Systolic_BP

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	13.300	7.331	1.814	.076	-1.440	28.041
Age	1.749	.128	13.695	.000	1.492	2.005

Age is a significant predictor / important

Blood pressure increases by 1.7 mmHg as age increases by a year

Linear Regression Output - **Unadjusted**

Tests of Between-Subjects Effects

Dependent Variable: Systolic_BP

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	6203.902[a]	1	6203.902	187.550	.000
Intercept	108.873	1	108.873	3.291	.076
Age	6203.902	1	6203.902	187.550	.000
Error	1587.778	48	33.079		
Total	647146.000	50			
Corrected Total	7791.680	49			

a. R Squared = .796 (Adjusted R Squared = .792)

Age is a significant predictor / important (it is the only predictor in the model)

~80% of the variability in blood pressure is explained by age alone used in the model. This does **not** tell us how well the model fits!

Linear Regression Output - **Unadjusted**

Parameter Estimates

Dependent Variable: Systolic_BP

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	119.880	2.137	56.100	.000	115.583	124.177
[Gender=female]	-13.600	3.022	-4.500	.000	-19.676	-7.524
[Gender=male]	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

Gender is a significant predictor / important

Blood pressure is 13.6 mmHg lower in females than males

Linear Regression Output - **Unadjusted**

Tests of Between-Subjects Effects

Dependent Variable: Systolic_BP

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	2312.000[a]	1	2312.000	20.252	.000
Intercept	639354.320	1	639354.320	5600.511	.000
Gender	2312.000	1	2312.000	20.252	.000
Error	5479.680	48	114.160		
Total	647146.000	50			
Corrected Total	7791.680	49			

a. R Squared = .297 (Adjusted R Squared = .282)

Gender is a significant predictor / important (it is the only predictor in the model)

~30% of the variability in blood pressure is explained by gender alone used in the model

Linear Regression Output - **Unadjusted**

Don't use for factors with multiple groups!

Parameter Estimates

Dependent Variable: Systolic_BP

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	111.923	3.598	31.104	.000	104.680	119.166
[SE_class=High]	.462	5.089	.091	.928	-9.782	10.705
[SE_class=Upper middle	1.994	5.194	.384	.703	-8.461	12.448
[SE_class=Lower middle	2.327	5.194	.448	.656	-8.128	12.782
[SE_class=Low]	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

There is not much difference in blood pressure between the social classes (e.g. mean difference of 0.462 between low and high)

Linear Regression Output - **Unadjusted**

Tests of Between-Subjects Effects

Dependent Variable: Systolic_BP

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	48.513[a]	3	16.171	.096	.962
Intercept	638767.102	1	638767.102	3794.738	.000
SE_class	48.513	3	16.171	.096	.962
Error	7743.167	46	168.330		
Total	647146.000	50			
Corrected Total	7791.680	49			

a. R Squared = .006 (Adjusted R Squared = -.059)

<1% of the variability in blood pressure is explained by social class alone used in the model

Social class is not a significant predictor / important (it is the only predictor in the model)

Linear Regression Output – **After Adjusting** for Age, Gender, and Social Class

Blood pressure increases by 1.7 mmHg as age increases by 1 year, after adjusting for the other factors
Blood pressure 3.3 mmHg lower in females than males, after adjusting for the other factors

Parameter Estimates

Dependent Variable: Systolic_BP

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	17.344	8.852	1.959	.056	-.496	35.183
[Gender=Female]	-3.278	1.793	-1.828	.074	-6.892	.336
[Gender=Male]	0	.	.	.	.	.
[SE_class=High]	-4.583	2.186	2.096	.042	-8.989	-.177
[SE_class=Upper middle]	-2.194	2.183	1.005	.320	-6.595	2.206
[SE_class=Lower middle]	-1.006	2.190	.459	.648	-5.420	3.409
[SE_class=Low]	0	.	.	.	.	.
Age	1.672	.142	11.749	.000	1.385	1.959

Don't use for factors with multiple groups!

e.g. Blood pressure is 4.6 mmHg lower in high compared to low category, after adjusting for the other factors

Linear Regression Output – **After Adjusting** for Age, Gender, and Social Class

At least one of the factors is a significant predictor / important

Tests of Between-Subjects Effects

Dependent Variable: Systolic_BP

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	6484.529[a]	5	1296.906	43.655	.000
Intercept	138.928	1	138.928	4.676	.036
Age	4101.015	1	4101.015	138.044	.000
Gender	99.272	1	99.272	3.342	.074
SE_class	144.147	3	48.049	1.617	.199
Error	1307.151	44	29.708		
Total	647146.000	50			
Corrected Total	7791.680	49			

a. R Squared = .832 (Adjusted R Squared = .813)

Social class is not a significant predictor / important

>80% of the variability in blood pressure is explained by the multivariable regression

Interpretation

- For **continuous variables** (i.e. age), the ‘parameter estimate’ represents the increase in blood pressure for an increase in age of 1 unit (i.e. as age increases by 1 year, blood pressure increases by 1.7 mmHg. This is <u>adjusted for all the other variables</u>. The 95% CI is the range of possible **true** effect sizes

- For **binary variables** (i.e. gender), the ‘parameter estimate’ represents the difference in the mean blood pressure <u>adjusted for all the other variables</u>. E.g., the estimated difference in blood pressure between males and females is 3.3 mmHg

- These are all effect sizes (they involve making comparisons), they are **mean differences**, and the no effect value is 0

- For both of these types of variables, the p-value given alongside is the one to use to determine whether each variable is an important factor or not, i.e. whether the observed effect size could be a chance finding in this particular study (there is only 1 p-value for each factor)

Interpretation

- For **categorical variables** with ≥3 levels (i.e. social class), you need to specify which level becomes the reference group (check coding - usually first or last group)

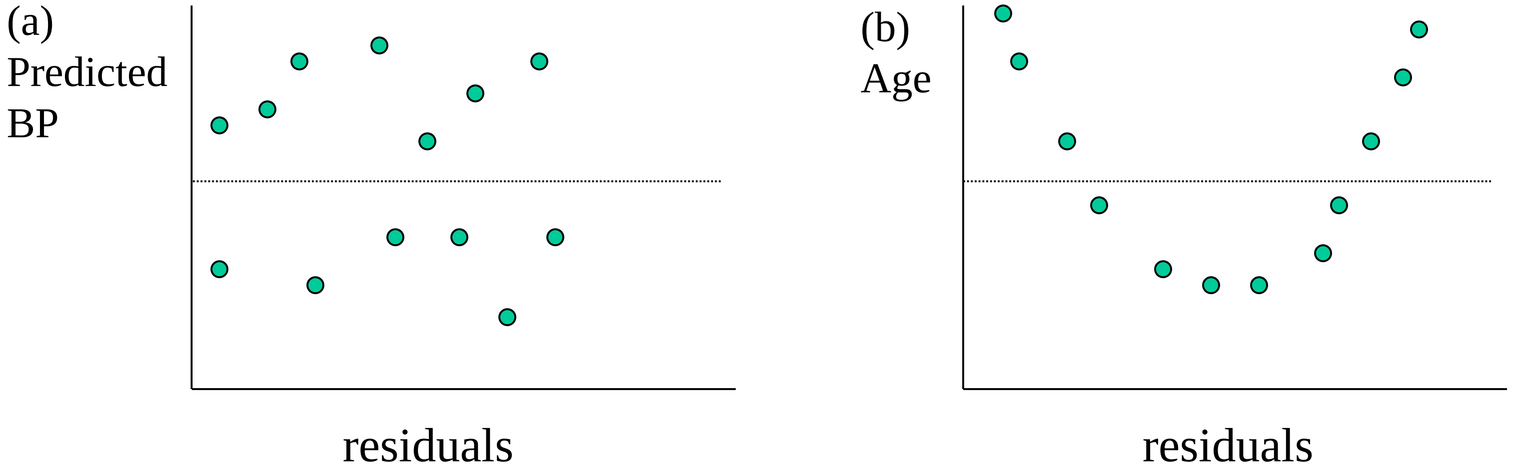
- The ‘parameter estimate’ is then the **mean difference** in blood pressure between each level and the reference group, <u>adjusted for all the other variables</u>
  - The difference between High and Low categories = -4.6 mmHg
  - The difference between Upper middle and Low categories = -2.2 mmHg
  - The difference between Lower middle and Low categories = -1.0 mmHg

- However, do not use the p-value alongside each level. You now have 3 p-values for the factor ‘social class’ (if it had 5 levels, you would have 4 p-values) – this can be difficult to interpret

- Use p-value from an **F-test** to determine whether ‘social class’ is important or not. It tells us overall whether social class is an important predictor of blood pressure (we now have only 1 p-value to consider for each factor)

Model Checks

- Plot of **residuals versus predicted** blood pressure should be a random scatter around zero (a)
- Residual = observed value minus predicted value from model
- Plot of residuals versus all other variables should be a random scatter around zero. For example age (b)



Is a Linear Model suitable for Age? Probably not

Multiple Logistic Regression

- We can extend logistic regression to adjust for multiple factors when the outcome has two levels (i.e. binary), such as in the hospital admission example seen earlier
- Similar principles as (multiple) linear regression, except the effect sizes are now ‘Odds Ratios’ and the no effect value is 1
- It has some useful mathematical properties that allow easier modelling (compared to relative risk)
- If there are many cells with no responses, the model could be unreliable (the estimates of effect size and standard errors could be extremely small or big). Therefore, consider combining cells with small numbers

Logistic Regression Output – **After Adjusting** for Age, Gender, and Social Class

Age is a significant predictor (p=0.003). The odds of hospital admission increases by 4.1 % as age increases by 1 year, after adjusting for the other factors

The odds of admission increases may be 80% lower or more than six-times higher in females than males, after adjusting for the other factors

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	Age	.340	.113	9.078	1	.003	1.041	1.013	1.075
	Gender	.145	.899	.026	1	.872	1.156	.199	6.729
	SE_class			3.509	3	.320			
	SE_class(1)	-1.794	1.263	2.018	1	.155	.166	.014	1.976
	SE_class(2)	.232	1.376	.029	1	.866	1.262	.085	18.722
	SE_class(3)	-1.447	1.227	1.391	1	.238	.235	.021	2.607
	Constant	-17.333	5.819	8.873	1	.003	.000		

a. Variable(s) entered on step 1: SE_class.

Don't use for factors with multiple groups!

Each 'estimate' is the log-odds, so we take exponentials (the **effect size** is the **odds ratio**)

Logistic Regression Output – **After Adjusting** for Age, Gender, and Social Class

- As in the linear regression analysis, if the factor is categorical with ≥3 levels we use a different test to see whether the factor is important or not (called the 'change in deviance')
- This avoids having to interpret several p-values for a single factor

Social class is not a significant predictor / important (p=0.25)

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1[a]	Step	4.129	3	.248
	Block	4.129	3	.248
	Model	25.832	5	.000

a. Variable(s) entered on step 1: SE_class.

At least one of the factors is a significant predictor / important (p<.001)

- Goodness of Fit: How well does the data fit the model?
- In Multiple Linear Regression check residuals versus all of the variables by plotting them (non constant variance)
- In Multiple Logistic regression – the most common method is called the **Hosmer & Lemeshow Test**. If significant, this suggests the model does not fit the data well

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	10.941	8	.205

The model fit is fine

Time-to-Event Outcomes – Cox regression

- The approach is analogous to multiple linear or multiple logistic regression, but the outcome measure is time until an event has occurred
- One main difference is that this method produces the **hazard ratio** as the effect size
- This is the risk of having an event in one group, compared to the risk in the reference group, <u>at the same point in time</u>
- Like other multivariate methods, the hazard ratio (effect size) can be adjusted for any other variables

Cox Regression Output – **After Adjusting** for Age, Gender, and Social Class

Age is a significant predictor (p<0.001). The risk of death increases by 15.7% as age increases by 1 year, after adjusting for the other factors

The risk of death may be 25% lower or four-times higher in females than males, after adjusting for the other factors

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Age	.146	.036	16.268	1	.000	1.157	1.078	1.242
Gender	.564	.439	1.650	1	.199	1.757	.744	4.153
SE_class			5.369	3	.147			
SE_class(1)	-.578	.511	1.284	1	.257	.561	.206	1.525
SE_class(2)	-.508	.451	1.267	1	.260	.602	.249	1.457
SE_class(3)	-1.244	.540	5.316	1	.021	.288	.100	.830

Don't use for factors with multiple groups!

Each 'estimate' is the log-odds, so we take exponentials (the **effect size** is the **hazard ratio**)

Cox Regression Output – **After Adjusting** for Age, Gender, and Social Class

- Again, as in the other types of regression, if the factor is categorical with ≥3 levels we use a different test to see whether the factor is important or not
- This avoids having to interpret several p-values for a single factor

Social class is not a significant predictor / important (p=0.13)

Omnibus Tests of Model Coefficients

-2 Log Likelihood	Overall (score)			Change From Previous Step		
	Chi-square	df	Sig.	Chi-square	df	Sig.
220.808	26.964	5	.000	5.607	3	.132

At least one of the factors is a significant predictor / important (p<.001)

All of the regression models covered above can be of the same form, and the factors that can be examined together can be any mixture of continuous, binary or categorical. You just need to know what the coefficients mean

	Regression model	Continuous (taking measurements)	Binary (2 levels) (counting)	Categorical (≥3 levels) (counting)
Type of outcome measure, Y	Y = a ('a' is intercept)	+ B x Age	+ C x Gender	+ D x Social class (low, low-mid, upper-mid, high)
Taking measurements (continuous) E.g. Y=blood pressure	Linear	As age increases by 1 unit, Y increases by B (same interpretation as simple linear **regression slope**)	C is the **mean difference** in Y between males and females (e.g. mean blood pressure in males minus mean blood pressure in females)	There will be three values for D. Each one is the **mean difference** in Y between the reference group which you choose (e.g. low) and each of the other categories
Counting people (binary) E.g. Y=hospital admission/none	Logistic	B is the **odds ratio** (log scale) for Y. As age increases by 1 unit. E.g. if OR=1.25, as age increases by 1 year, the chance of admission increases by 25%	C is the **odds ratio** (log scale) for Y (e.g. admission) for males compared to females. E.g. if OR=0.75, then the risk of admission in females is 25% lower than the risk in males	There will be three values for D. Each one is the **odds ratio** (log scale) of Y (e.g. admission) between the reference group which you choose (e.g. low) and each of the other categories. Same interpretation as C.
Time-to-event E.g. Y=survival time (the time it takes to die, but also some people haven't died yet)	Cox	B is the **hazard ratio** (log scale); but the same interpretation as odds ratio above. E.g. if HR=1.25, as age increases by 1 year, the chance of dying increases by 25%	C is the **hazard ratio** (log scale); but the same interpretation as odds ratio above. It is the risk of having an event (for whatever you have defined as an event)	D is the **hazard ratio** (there will be three values, log scale); but the same interpretation as odds ratio above.