INTRODUCTION TO STATISTICS

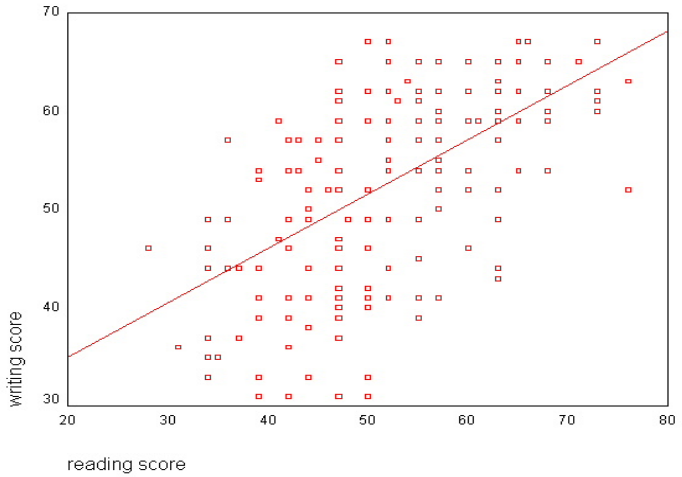
Session 6

May 22rd 2017

Madoka Takeuchi

Review: Regression

- Regression investigates how the change or lack of change of one variable corresponds and relates to the change of another variable.
- Linear regression consists of finding the best-fitting straight line through the data.
  The best-fitting line is called a regression line.



Review:
Least Square Regression Line

- Multiple lines can be drawn through the data
- Least squares regression line minimizes the squared distances between the line and the data points, thus this line is most likely the line that best fits the data.
- The slope of the least square line is related to the correlation coefficient of the data.

  Slope (a)= r (S_y/S_x) where the basic equation is $y= ax+ b$

- Every least squares line passes through the mean x coordinate and mean y coordinate thus can calculate the intercept (b)

Example

Recall the study of ten students actual study time and test score

Student	Hours spent	Test Score
1	1	81
2	5	75
3	7	89
4	10	88
5	3	65
6	5	92
7	6	71
8	2	78
9	1	54
10	0	68

Least square regression line

- What is the dependent/independent variable?

- What do we need to know/ calculate to find the least square regression line?

Prediction/residuals

- The equation for the regression line can be used for prediction- given a value for x, can predict a value for y
  - Using the equation, what is the predicted test score for a student who studies for 10 hours?
- The residuals measure the error of each observation away from what the model predicts them to be
  - e= actual data value- predicted data value
  - The equation predicted a score of 90 for a student studying 10 hours, what is the e ?

Residual Properties

- Residuals are assumed to have the following properties
  - Each residual is independent of each other
  - Normally distributed
  - Centered at zero
  - Constant variance

This can be written as $e \sim N(0,\sigma)$

Extending Regression…

- One could study more details of how single independent variables change the dependent variable but in certain situations it is more instructive, to look at more complex models- models with several independent variables-
  - this is called ***multiple regression***
  - Multiple regression is an extension of simple linear regression
  - Multiple regression can also determine the overall fit of the model and the relative contribution of each of the independent variables to the model

Example

- In the example of hours spent studying and test score, only evaluated if the hours spent studying predicts test score.
- However there may other variable such as lecture attendance, and gender that may or even better predict test score

Equations

- In simple linear regression
  - $Y=ax+b$
- In multiple regression
  - $y= a_1x_1 + a_2x_2 + a_3x_3 + \ldots\ldots b$
  - $x_1, x_2, x_3, \ldots\ldots x_n$ are the independent variables
  - $a_1, a_2, a_3, \ldots a_n$ are called coefficients
- In the previous example
  - Test score = a_1(hours) + a_2(attendance)+ a_3(gender) +intercept

Multiple regression: Assumptions

- Prior to analyzing data with multiple regression, must check that the data can actually be analyzed using multiple regression.
- There are eight assumptions that the data must pass to obtain valid results from multiple regression
- In real world data, it is often the case that the data does not "pass" all the assumptions and this does not mean that multiple regression can not be used. But it is necessary to know how the data "failed to pass"

Assumptions (1)

Assumption 1:

- The dependent variable must be measured on a continuous scale, can not be categorical

Assumption 2:

- Must have 2 or more independent variables- the variables can be continuous or categorical
  - If you only have categorical independent variables (i.e. no continuous independent variables), it is more common to approach the analysis from the perspective of a two-way ANOVA- covered in later lectures

Assumptions (2)

Assumption 3:

- The observations must be independent of each other

Assumption 4:

- There must be a linear relationship between the each of the independent variables and the dependent variable
  - Check by creating scatterplots for each of the independent variables or calculate *r*

Assumptions (3)

Assumption 5:

- The data needs to show homoscedasticity-variance around the regression line is the same for all values of the predictor variables

Assumption 6:

- The data must not show multicollinearity- meaning that the independent variable are highly correlated.
  - Mulitcollinearity will cause problems when trying the understand the relation of one independent variable to the dependent variable
  - i.e. can`t have weight and BMI as two variables

Assumptions (4)

Assumption 7:

- There should be no significant outliers

Assumption 8:

- Check that the residuals (errors) are approximately normally distributed- histogram of the residuals

Important concepts

- Interpretation of regression coefficients: A coefficient for a independent variable estimates how much the dependent variable changes if
  – the independent variable changes by one unit
  – All other independent variables are held constant- do not change
- The 'effect' of each independent variable on the dependent variable when other variables are held constant is linear (assumption is linear regression).

Multiple regression in SPSS

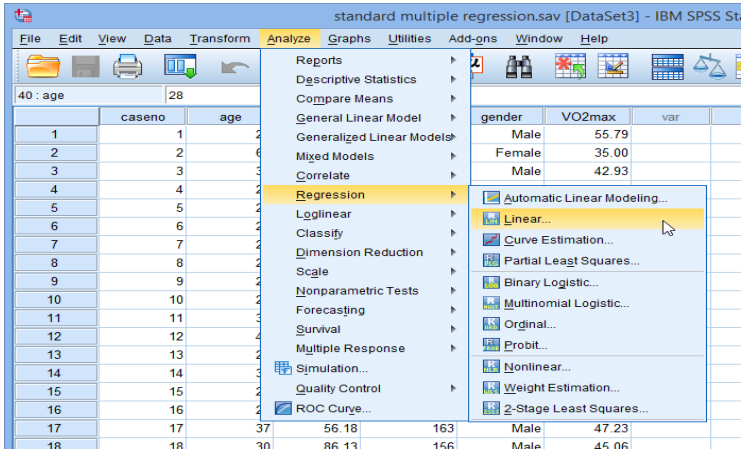
A health researcher wants to be able to predict "cholesterol" using gender, age and weight(kg)

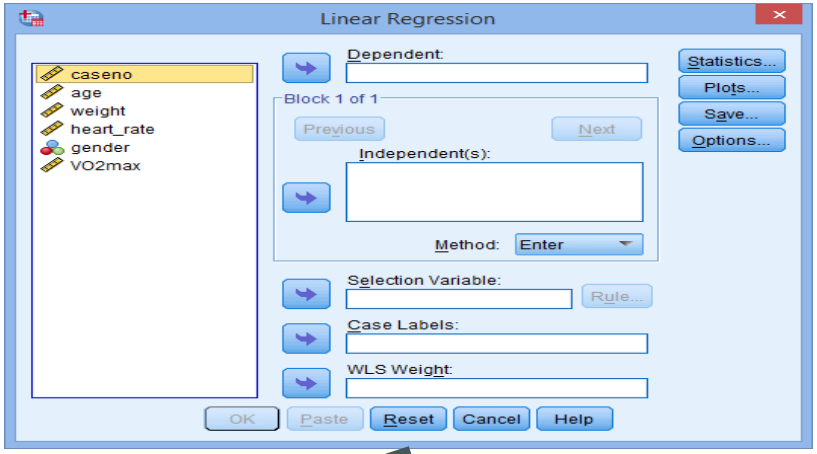
ID	Gender	age	weight	chol
1	Male	81	70	201
2	Female	75	43	150
3	Male	18	59	140
4	Male	65	83	230
5	Male	73	60	165
6	Female	66	51	180
7	Female	85	44	174
8	Male	70	76	216
9	Male	79	85	221
10	Male	80	79	203

Data assumption check

- Are all the assumptions cleared??

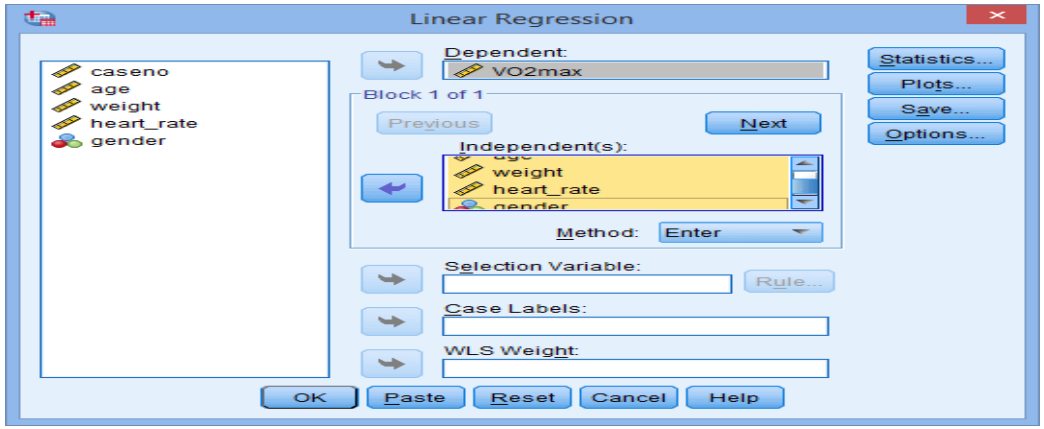
SPSS





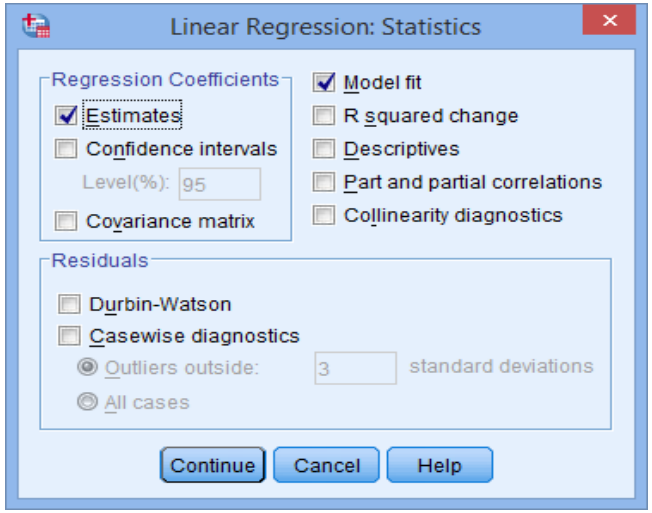
Analyze→ Regression→ Linear

Transfer/drag the "chol" variable to the dependent box and drag the age, weigh and gender variables into the independent box



* ignore the previous and next buttons

Click the statistics button



Click continue→ OK

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.760[a]	.577	.559	5.69097

a. Predictors: (Constant), gender, age, heart_rate, weight

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	**87.830**	6.385		13.756	.000	75.155	100.506
	age	**-.165**	.063	-.176	-2.633	.010	-.290	-.041
	weight	**-.385**	.043	-.677	-8.877	.000	-.471	-.299
	heart_rate	**-.118**	.032	-.252	-3.667	.000	-.182	-.054
	gender	**13.208**	1.344	.748	9.824	.000	10.539	15.877

a. Dependent Variable: VO2max

RESULTS

- To determine the model fit→

look at the model summary table which provides R and R^2

R^2- proportion of variance in the dependent variable that can be explained by the independent variables

-if R^2= 0.8- independent variables explain 80% of the variability of our dependent variable

- To determine the model coefficient estimates→ look at the coefficients table
  - Constant= slope

What is the multiple regression equation?

- Coeffcients
  - Age=-.638
  - Weight= 1.561

- Constant= 42.387

- R^2= 0.845

Equation

Chol= .638(Age) + 1.561(Weight) + 42.387

This means that for each 1 year increase in age, there is an increase in chol of .638

Age and weight (the independent variables) explain approx 85% of the variability of cholesterol (the dependent variable)

Statistically significant?

- One can test for the statistical significance of each of the independent variables.
- Tests whether the unstandardized coefficients are equal to zero in the population- meaning the is not effect of the independent variable on the dependent variable.
- If $p < .05$, you can conclude that the coefficients are statistically significantly different to 0 (zero).
  - The definition and interpretation of the p-value will be discussed in future lectures
- In SPSS, statistical significance of a variable is evaluated in the "Sig." column

Best model?

- If an independent variable is found not to be statistically significant, may consider removing the variable from the regression equation.
  - Re-run the regression analysis without the variable an look at the R and R^2 value to determine which model is better.
- In our previous example, R and R^2 are very high so the model is a good fit with to the data with the age, gender, weight variables
- If the "Sig." for age is not significant- re-run the model without age
  - What is the new R, R^2- should we keep or drop the age variable for the model to predict cholesterol?

Model Selection

- Since simple linear regression and multiple regression are used for prediction, it is often the case that an one wants to come up with a model that will best predict future observations.
- A simple model with the fewest necessary variables is favorable- easy to understand and interpret.
- Choose the model with the smallest variance- Want a model with a large R^2

In class exercise

One would like to develop a model to predict the number of goals by U23 soccer players. It is thought that age, years of play and hours of practice are important factors

What is the best model?
What independent variables are in the model?

id	age	Years of play	Hours practice	goals
1	18	10	40	16
2	22	15	35	15
3	20	9	22	7
4	21	10	36	15
5	22	8	45	5
6	19	13	30	17
7	25	15	20	25
8	17	2	50	3
9	18	12	35	15