INTRODUCTION TO STATISTICS

Session 5

May 8th 2017

Madoka Takeuchi

Review: Response and Explanatory variables

- Response Variable/dependent variable
  - measures the outcome
  - denoted as *y*
- Explanatory Variable/ independent variable
  - explains the change in the response variable
  - *may be* the cause of change in the response variable-not necessarily the cause
  - denoted as *x*

$$Y=ax+b$$

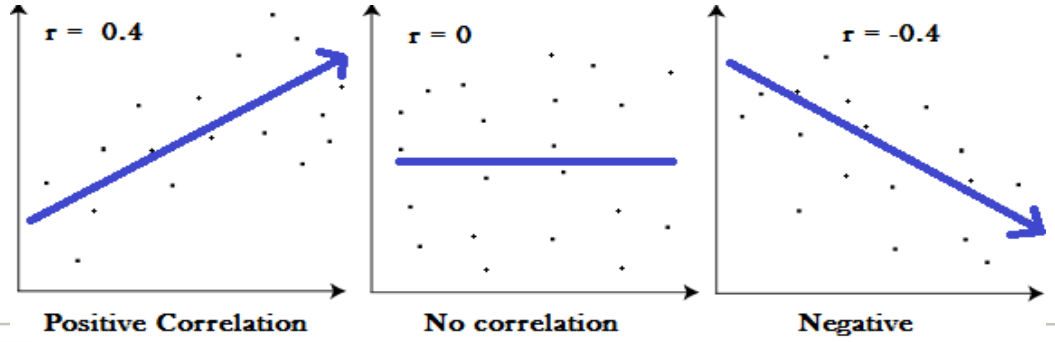
Review: Causation vs Association

- Causation- changes in one variable *directly (100%)* causes changes in the other variable
  - i.e when you exercise the amount of calories you burn increases→ exercise directly causes the increase in calorie burn
- Association- relationship between two variables.
  - i.e. cloudy weather is associated with rainfall, but cloudy weather does not 100% result in rain,cloudy weather does not cause rain
- Association (correlation) does not imply causation

Review: Correlation

- Association refers to the general relationship between two variables. Correlation refers a **linear relationship** between the variables.

- Correlation is a measure of association.
  - Correlation Coefficient (*r*), measures the **strength** and the **direction** of a linear relationship between two variables
    - tells us how closely data in a scatterplot fall along a straight line.
  - Ranges from 1 to -1 , 0 meaning there is no linear relationship between the variables.
    - r=1,indicates that the variables are positively linearly related and the scatterplot falls almost along a straight line with positive slope
    - r=-1,indicates that the variables are negatively linearly related and the scatter plot almost falls along a straight line with negative slope



Correlation Coefficient

(Pearson Correlation Coefficient)

- Although SPSS can easily calculate the correlation coefficient, must understand the equation

$$r = \frac{1}{n-1} \sum \left(\frac{x - \bar{x}}{s_x} \right) \left(\frac{y - \bar{y}}{s_y} \right)$$

- n= number of paired data
- S_x=sample standard deviation of the x values
- S_y= sample standard deviation of the y values

Study- Want to know if aging and weight gain is associated

Data from 5 random people-

What is the explanatory variable (x)?
What is the response variable (y)?

	Age	Weight
1	23	50
2	38	51
3	60	63
4	30	49
5	45	54

Correlation coefficient calculation

1. Need to calculate the sample mean of the independent and dependent variables

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(xi - \overline{x}\right)^2}{n-1}}$$

1. After the sample means are calculated, calculate the sample standard deviation of the independent and dependent variables.

2. Plug the sample means and standard deviations into the correlation coefficient equation

$$r = \frac{1}{n-1}\sum\left(\frac{x-\overline{x}}{s_x}\right)\left(\frac{y-\overline{y}}{s_y}\right)$$

n=5

Mean of the x values= (23+38+60+30+45)/5
= 39.2

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(xi - \overline{x}\right)^2}{n-1}}$$

SD of x= $\sqrt{((23-39.2)^2+(38-39.2)^2+(60-39.2)^2+(30-39.2)^2+(45-39.2)^2/4)}$
= $\sqrt{((-16.2)^2+ (-1.2)^2+ (20.8)^2+ (-9.2)^2+(5.8)^2/4))}$
=$\sqrt{203.6}$
=14.27

Mean of the y values= (50+51+63+49+54)/5
= 53.4

SD of y= 5.68

$$r = \frac{1}{n-1}\sum\left(\frac{x-\overline{x}}{s_x}\right)\left(\frac{y-\overline{y}}{s_y}\right)$$

r= (1/4) {[(23-39.2)/14.26]*[(50-53.4)/5.68)}+{[(38-39.2)/14.26]*[(51-53.4)/5.68]}
+ {[(60-39.2)/14.26]*[(63-53.4)/5.68]}+{[(30-39.2)/14.26]*[(49-53.4)/5.68]}
+ {[(45-39.2)/14.26]*[(54-53.4)/5.68]}
= 0.9296

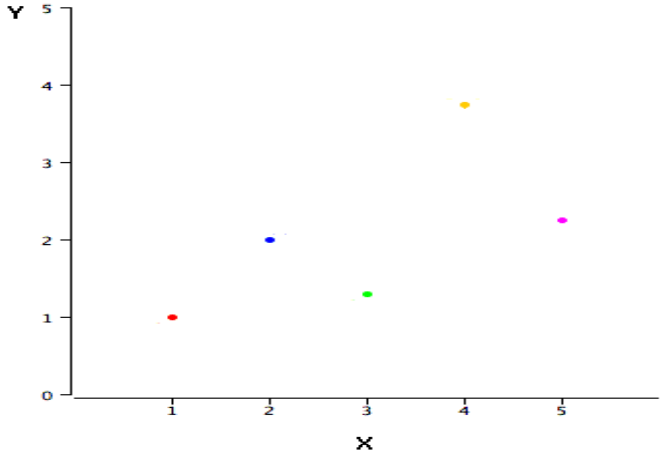
Please check in SPSS

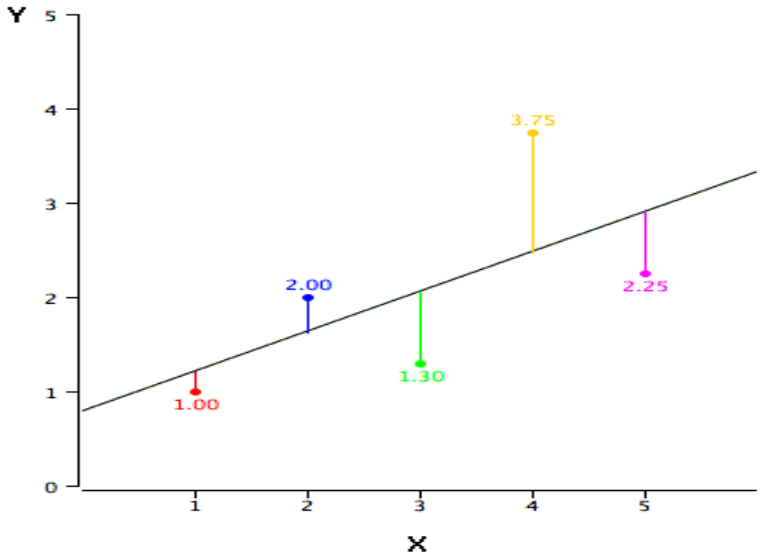
Regression Analysis

- regression analysis= The study of the analysis of data aimed at discovering how one or more variables (*independent variables, explanatory variables*) are associated with the values of other variables (called *dependent or response variables*)
- **Linear regression** - finding the best-fitting straight line through the data.
- The best-fitting line is called a *regression line*.

Example data

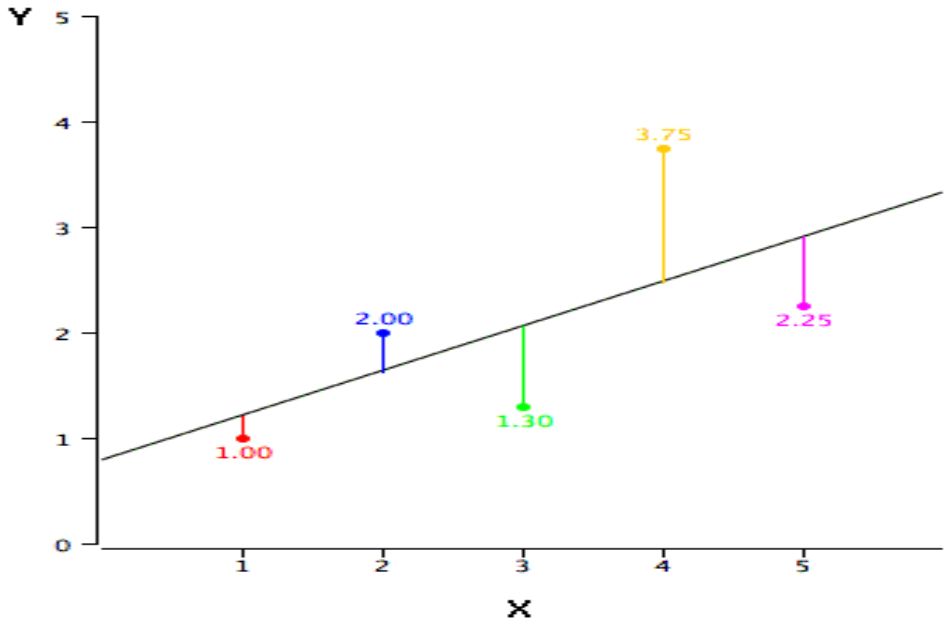
X	Y
1	1
2	2
3	1.3
4	3.75
5	2.25





Simple linear regression- predict scores on one variable from the scores on a second variable

- Points are the actual data
- Black line is the regression line and consists of the predicted score on Y for each possible value of X



-The **error of prediction** for a point is the value of the point minus the predicted value

-The vertical lines from the points to the regression line represent the errors of prediction.

-The red point is very near the regression line

-error of prediction is *small*

-The yellow point is much higher than the regression line

-error of prediction is *large*

Errors of Prediction

X	Y	Y'	Y-Y'	$(Y-Y')^2$
1	1	1.2	-.2	0.040
2	2	1.62	.38	0.144
3	1.3	2.04	-.74	0.548
4	3.75	2.46	1.29	1.664
5	2.25	2.9	-.65	0.423

Y'- predicted value
Y-Y'- errors of prediction

Simple linear regression can over-predict or under-predict the response variables.
The aim is to find the best fitting line to the data

- best fitting line is the line that minimizes the sum of the squared errors of prediction

Least Square Regression Line

- Least squares 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 connected to the correlation coefficient of the data.

  Slope= r (S_y/S_x) where S_y= SD of Y

  S_x= SD of X
- **When slope > 0** – there is a positive trend — that is, as x increases, y tends to increase
- **When slope < 0**- – there is a negative trend — that is, as x increases, y tends to decrease

Intercept

- Every least squares line passes through the mean x coordinate and mean y coordinate thus can calculate the intercept
  - After solving for the slope(a), plug in mean y coordinate and mean x coordinate into the equation y= ax+ b and solve for a to get the intercept

$$b = y - ax$$

Regression line equation

From the study of aging and weight gain, we calculated the mean of x, mean of y, SD of x, SD of Y and the correlation coefficient

Mean x=39.2 Mean y=53.4
SD x=14.27 SD y=5.68
r= 0.9296

Slope= r (S_y/S_x)
= 0.9296 (5.68/14.27)
=.367

Intercept = y-slope(x)
=53.4- (.367)(39.2)
=39.014

Equation for the regression line y= .367x+ 39.014

Prediction model

- Once the regression equation is calculated, we can use this equation as a prediction model
  - given a particular value of *x*, we can predict what the value of *y* would be (*y* along the line at *x*)
    - Predicted y value is denoted as $\hat{y}$
  - Example- from the age and weight equation we can predict the weight of a 30 yr old

$\hat{y}$ = .367(30)+39.014

= 50.024

Residuals

- The errors of prediction are also referred to as residuals and are denoted as *e*
- Residuals= measure the error of each observation away from what the model predicts them to be
  - *e*= actual data value- predicted data value
- In the age and weight data, a 30 yr weighed 49kg, however using the calculated regression equation, we predicted that the weight of a 30yr would be around 50kg
- The residual *e*= 49-50

  =-1

r^2

- The square of the correlation coefficient, r^2, is a useful value in linear regression
- r^2=fraction of the variation in one variable that may be explained by the other variable.
  - If a correlation of 0.8 is observed between two variables (i.e, age and weight), then a linear regression model attempting to explain *either* variable in terms of the other variable will account for 64% of the variability in the data.

Example: the correlation coefficient for crime rate and unemployment rate was found to be 0.42 thus r^2 is approximately 0.18

Interpretation: the regression line explains 18% of the variability in violent crime rate, or leaves 82% unexplained.

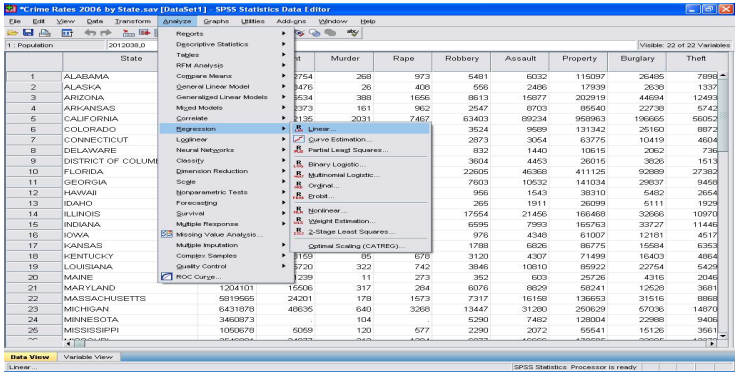
Going back to the previous study of aging and weight gain…

	Age	Weight
1	23	50
2	38	51
3	60	63
4	30	49
5	45	54

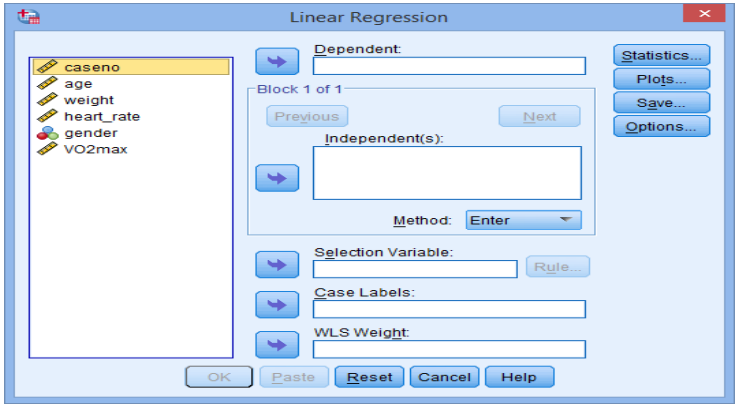
Lets check the linear regression line we calculated using SPSS

Linear Regression in SPSS

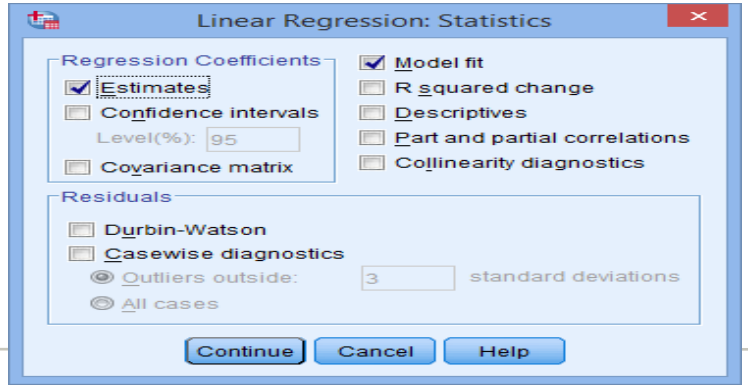
Step 1
Analyze→ Regression→ Linear



Step 2
Select dependent variable and independent variable



Step 3
Click statistics button and select estimates and model fit



Understanding the output

First table of interest is the model summary- provides *r* and r^2

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1				

a. Predictors: (Constant), Months since Hire, Educational Level (years)

To calculate the regression equation, look at the **Coefficients** table

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)					
	x					

The constant is the intercept and the value for x is the coefficient for the slope- the equation for the above data would be

y= _____x + constant

The equation we calculated by hand was

$$y = .367x + 39.014$$

What is the regression equation from SPSS???

Are the equations the same?