Single and multiple linear regression analysis

Marike Cockeran
2017

It all starts here

NORTH-WEST UNIVERSITY
YUNIBESITI YA BOKONE-BOPHIRIMA
NOORDWES-UNIVERSITEIT
POTCHEFSTROOM CAMPUS

Outline of the session

- Introduction
- Simple linear regression analysis
- SPSS example of simple linear regression analysis
- Additional topics in multiple linear regression analysis
  - Adjusted R-squared
  - Standardised regression coefficients
  - Multicollinearity
  - A note on categorical predictors (dummy variables)

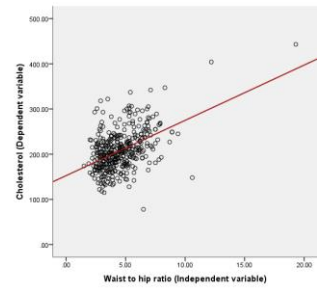
Introduction: Simple linear regression (1)

- Suppose we collect data on two variables:
  - Waist to hip ratio (X).
  - Cholesterol (Y).
- For each participant we now have a pair of observations (X_i, Y_i).

ID	Cholesterol (Y)	Ratio (X)
1	203	3.60
2	165	6.90
3	228	6.20
4	78	6.50
5	249	8.90
⋮	⋮	⋮

Introduction: Simple linear regression (2)

- We want to fit a straight line that describes the linear relationship between X and Y.

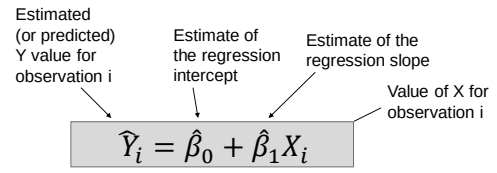


Introduction: Simple linear regression (3)

- Simple linear regression is a technique that is used to explore the nature of the relationship between **two** variables.
- Regression analysis enables us to investigate the change in one variable, called the response (dependent variable), which corresponds to a given change in the other, known as the explanatory variable (independent variable).
- The ultimate objective of regression analysis is to predict or estimate the value of the response that is associated with a fixed value of the explanatory variable.

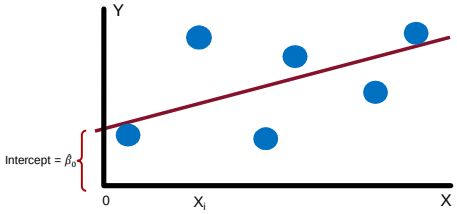
Simple Linear Regression Equation (Prediction Line)

The simple linear regression equation provides an estimate of the population regression line



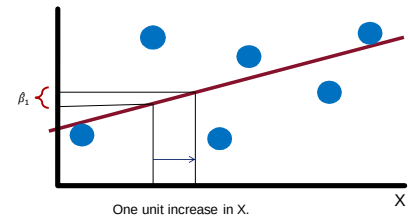
Interpretation of intercept term

- $\hat{\beta}_0$ is the estimated average values of Y when the value of X is zero.
- $\hat{\beta}_0$ has only practical application if X=0 is in the range of observed X values.



Interpretation of slope term

- $\hat{\beta}_1$ estimates the change in the average value of Y as a result of a one-unit increase in X.



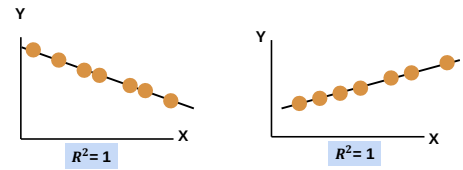
Coefficient of Determination (R^2)

- The **coefficient of determination** is the portion of the total variation in the dependent variable that is explained by variation in the independent variable
- The coefficient of determination is also called **R-squared** and is denoted as R^2.

$$R^2 = \frac{SSR}{SST} = \frac{\text{regression sum of squares}}{\text{total sum of squares}}$$

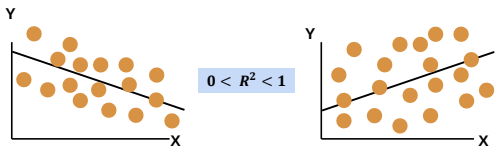
$$0 \le R^2 \le 1$$

Examples of approximate R^2 values (1)



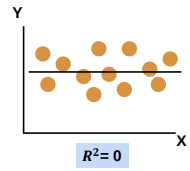
Perfect linear relationship between X and Y.
100% of the variation in Y is explained by variation in X.

Examples of approximate R^2 values (2)



Weak linear relationships between X and Y.
Some but not all of the variation in Y is explained by variation in X.

Examples of Approximate R^2 values (3)



No linear relationship between X and Y.
The value of Y does not depend on X.
None of the variation in Y is explained by the variation in X.

Assumption of simple linear regression models

- Linearity
  - The relationship between X and Y is linear.
- Independence of errors
  - Error values are statistically independent.
- Normality of error
  - Error values are normally distributed for any given value of X.
- Equal variance (homoscedasticity)
  - The probability distribution of the errors has constant variance.

Test assumptions after fitting the model by making use of the residuals.

Residual analysis

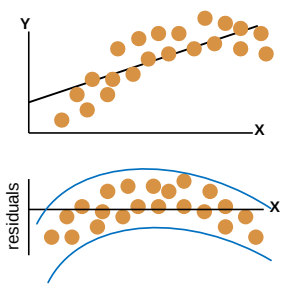
- The residual for observation i, e_i, is the difference between the observed and predicted values.

$$e_i = Y_i - \hat{Y}_i$$

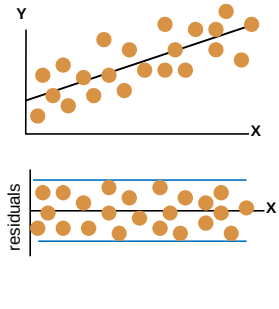
- Check the assumptions of regression by examining the residuals.
  - Linearity assumption
  - Independence assumption
  - Normal distribution assumption
  - Constant variance for all levels of X (homoscedasticity).
- Graphical analysis of residuals
  - Plot the residuals against X.

Residual analysis for linearity

✘ Not linear

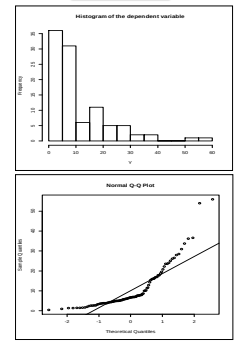


✓ Linear

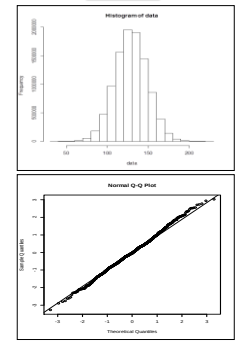


Residual analysis for normality

✘ Not normal

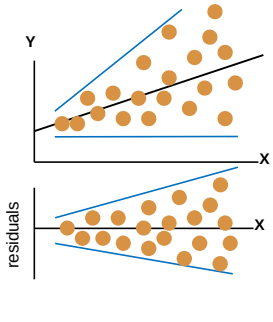


✓ Normal

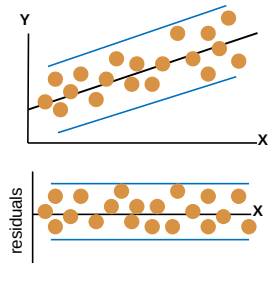


Residual analysis for equal variance

✗ Non-constant variance



✓ Constant variance



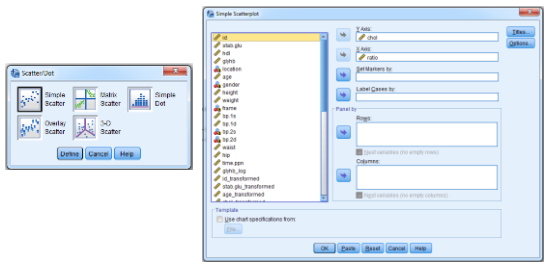
SPSS example: The data

- Suppose we collect data on two variables:
  - Waist to hip ratio (X).
  - Cholesterol (Y).
- For each participant we now have a pair of observations (X_i, Y_i).

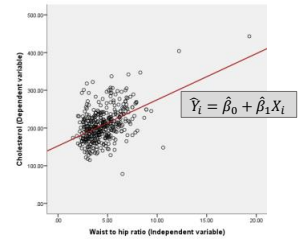
ID	Cholesterol (Y)	Ratio (X)
1	203	3.60
2	165	6.90
3	228	6.20
4	78	6.50
5	249	8.90
⋮	⋮	⋮

Steps in SPSS: Scatter plot and prediction line

Graphs → Legacy dialogs → Scatter/Dot



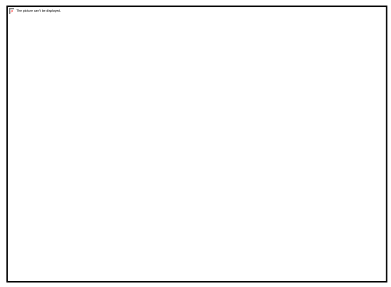
SPSS output: Scatter plot and prediction line



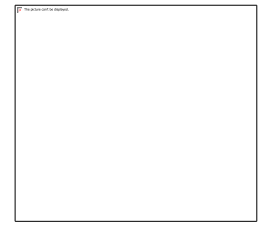
Predicted cholesterol $= 0.970 + 0.387 \times Waist$

Steps in SPSS: Simple linear regression model

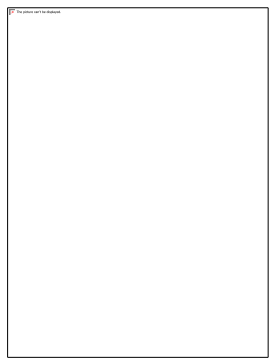
Analyze → Regression → Linear



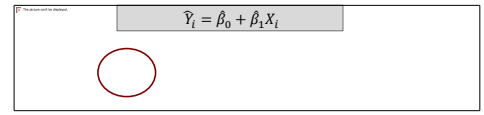
Steps in SPSS: Statistics tab



Steps in SPSS: Save tab

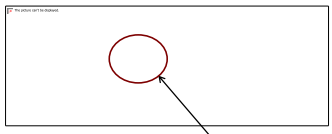


SPSS output: Regression equation



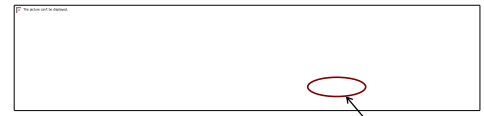
Predicted cholesterol = $171.233 + 0.970 \times Waist$

SPSS output: R^2 values



1.5% of the variance of cholesterol could be explained by waist to hip ratio.

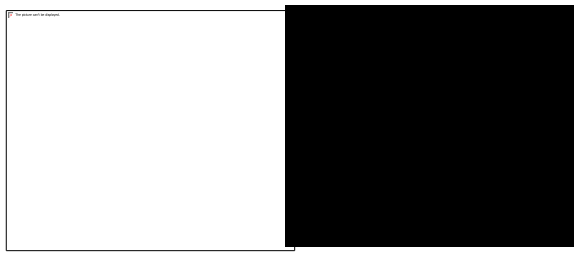
SPSS output: Inferences about the slope



H_0: $\beta_1 = 0$ H_A: $\beta_1 \neq 0$

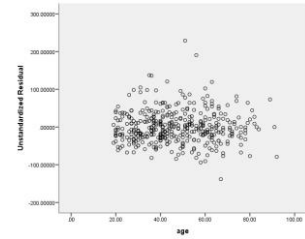
Waist to hip ratio is a significant predictor of the dependent variable cholesterol, p=0.013.

SPSS output: Residual analysis for normality



The residuals are normally distributed.

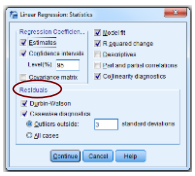
SPSS output: Residual analysis for equal variance



The residuals have constant variance.

SPSS output: Outliers

- If the absolute standardised residual value is larger than 3 then the observation is considered as an outlier.



Casewise Diagnostics[a]

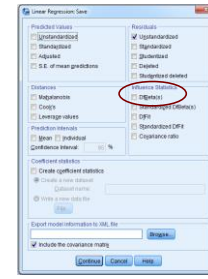
Case Number	Std. Residual	chol	Predicted Value	Residual
4	-3.935	78.00	232.0443	-154.04427
47	-3.428	148.00	282.1938	-134.19385
134	3.320	318.00	188.0105	129.98951
378	3.025	337.00	218.5895	118.41050
381	3.235	322.00	195.3495	126.65055

a. Dependent Variable: chol

SPSS output: Influential cases

Analyze → Regression → Linear → Save tab

- $DFBETAS$: It is the difference between the estimated regression coefficient $\hat{\beta}_k$ based on all n cases and the regression coefficient obtained when the i^{th} case is omitted.
- If the absolute value of the $DFBETAS$ value exceeds $\frac{2}{\sqrt{n}}$ the value can be viewed as an influential value.



Introduction: Multiple linear regression analysis

- In the preceding chapter, we saw how simple linear regression can be used to explore the nature of the relationship between two variables.
- If knowing the value of a single explanatory variable improves our ability to predict the response, we might expect that additional explanatory variables could be used to our advantage.
- To investigate the more complicated relationship among a number of different variables, we use a natural extension of simple linear regression analysis known as multiple linear regression analysis.

Introduction: Multiple linear regression analysis

- Multiple linear regression equation:

$$\hat{Y} = \hat{B}_0 + \hat{B}_1 X_1 + \hat{B}_2 X_2 + \cdots + \hat{B}_p X_p$$

- The regression coefficients are still estimated by using the method of least squares.
- The independent variables can be continuous or categorical variables.
- In the case of categorical variables we need to use dummy variables.

Adjusted R-squared

- The inclusion of an additional variable in a model can never cause R^2 to decrease.
- To get around this problem, we can use a second measure, called the $adjusted\ R^2$, that compensated for the added complexity of a model.
- The $adjusted\ R^2$ increases when the inclusion of a variable improves our ability to predict the response and decreases when it does not.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.506[a]	.256	.251	38.30893	.256	45.624	3	397	.000	1.985

a. Predictors: (Constant), weight, age, ratio
b. Dependent Variable: chol

Standardised coefficients

- **Unstandardised coefficients**
  - The value of the unstandardised coefficient is dependent on the units of measurement of the variables.
  - It is not possible to compare the relative magnitude of coefficients.
- **Standardised coefficients**
  - The value of the unstandardised coefficient now does not depend on the units of measurement of the variables.
  - It is now possible to compare the relative magnitude of coefficients.
  - How to standardise:

$$\frac{Y_i - \bar{Y}}{sd(Y_i)} \quad \frac{X_i - \bar{X}}{sd(X_i)}$$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	143.981	10.679		13.482	.000	122.986	164.976		
	ratio	11.908	1.172	.465	10.157	.000	9.603	14.212	.893	1.120
	age	.448	.119	.165	3.756	.000	.213	.682	.967	1.034
	weight	-.060	.050	-.055	-1.210	.227	-.159	.038	.911	1.098

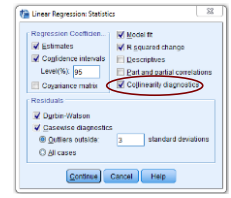
a. Dependent Variable: chol

Multicollinearity

- Multicollinearity exists when there is a strong correlation between two or more predictors (independent variables) in a regression model.
- High levels of collinearity increase the probability that a good predictor of the outcome variable will be found non-significant and rejected from the model (Type II error).
- VIF (variance inflation factor) > 10 indicates a potential problem.
- Tolerance below 0.2 indicates a potential problem.

SPSS output: Multicollinearity

Analyze → Regression → Linear → Statistics



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	143.981	10.679		13.482	.000	122.986	164.976		
	ratio	11.908	1.172	.465	10.157	.000	9.603	14.212	.893	1.120
	age	.448	.119	.165	3.756	.000	.213	.682	.967	1.034
	weight	-.060	.050	-.055	-1.210	.227	-.159	.038	.911	1.098

a. Dependent Variable: chol