Multiple linear regression
S6

Sander Roberti
s.roberti@nki.nl

November 21, 2018

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Introduction

Two main motivations for doing multiple linear regression:

1. There are often numerous variables that might be associated with the dependent variable of interest
   - For example, blood pressure might be related to factors such as body weight, level of physical activity, gender, socioeconomic status, alcohol consumption and tobacco use

2. Adding more variables into the model leads to an increase in R^2 and thus to more accurate predictions of the dependent variable

Multiple linear regression

Multiple linear regression postulates that in the population

$$y = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \ldots + \beta_k \cdot x_k + \epsilon,$$

where:

- y is the dependent variable
- $x_1, x_2, \ldots, x_k$ are the independent variables
- $\alpha, \beta_1, \beta_2, \ldots, \beta_k$ are the population regression coefficients
  - α is called the intercept
  - $\beta_1, \beta_2, \ldots, \beta_k$ are called the partial regression coefficients
  - β_1 is the parameter associated with x_1, β_2 is the parameter associated with x_2, and so on
- ϵ is the random error term, which allows the value of y to vary for any given set of values for the explanatory variables $x_1, x_2, \ldots, x_k$

Multiple linear regression

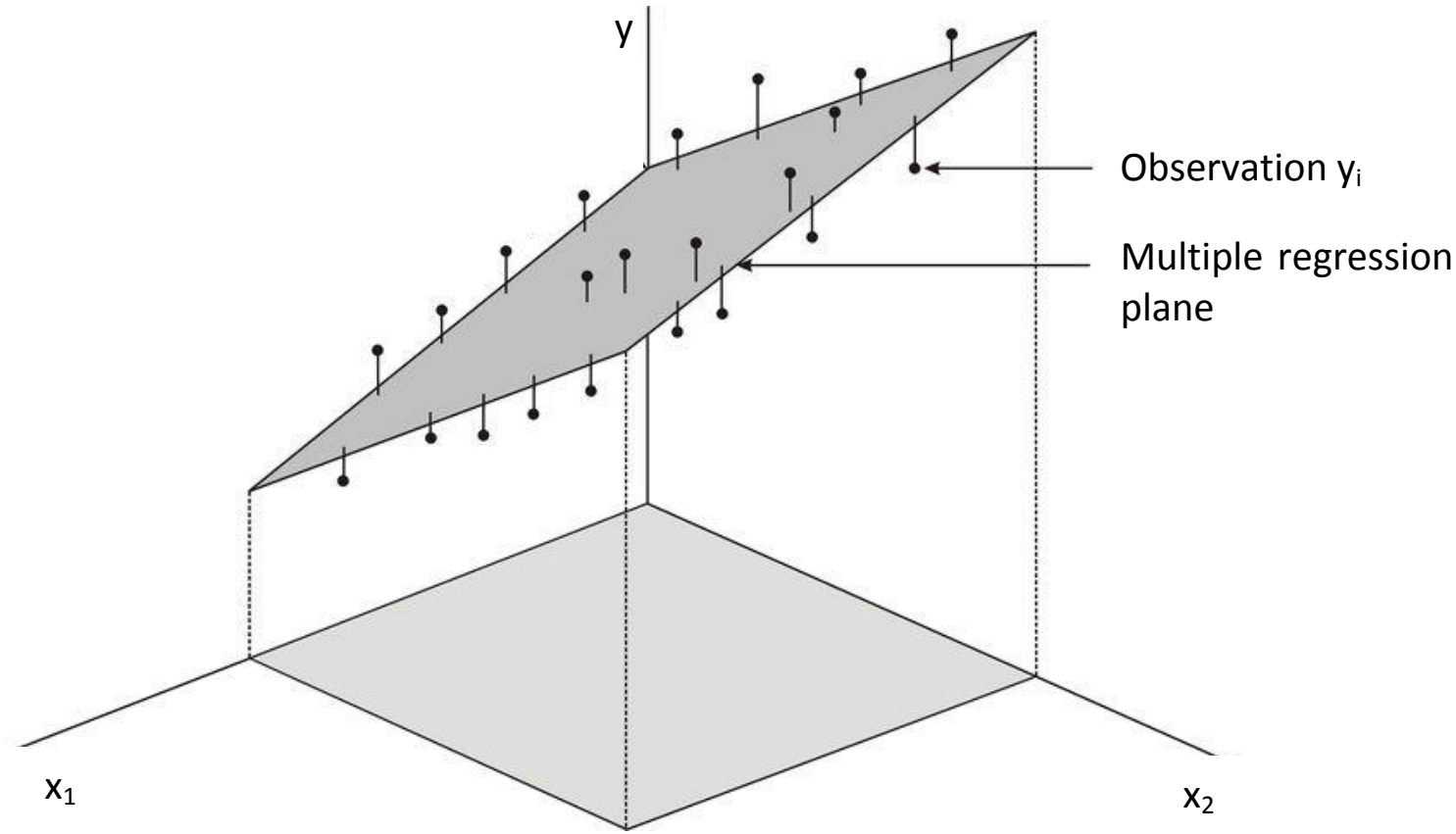
The population regression function now becomes:

$$E(y|x_1, x_2, \ldots, x_k) = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \ldots + \beta_k \cdot x_k,$$

where $\mathrm{E}(y|x_1, x_2, \ldots, x_k)$ is the mean value of y for a given set of values of $x_1, x_2, \ldots, x_k$.

- The population regression function is not a straight line anymore
- If there are only two explanatory variables, the population regression function is a plane in three-dimensional space
- If there are more than two explanatory variables, the population regression function is a hyperplane

Multiple linear regression



NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Multiple linear regression

- α is the mean value of y when all explanatory variables equal zero, i.e. when $x_1 = x_2 = \ldots = x_k = 0$
- β_i is the mean change in y due to one unit increase in the value of x_i when all other variables are held constant. This is seen by looking at the difference in the mean values:

$$E(y|x_1, ..., x_i + 1, ..., x_k) - E(y|x_1, ..., x_i, ..., x_k) =$$

$$[\alpha+\beta_1 \cdot x_1+...+\beta_i \cdot (x_i+1)+...+\beta_k \cdot x_k]-[\alpha+\beta_1 \cdot x_1+...+\beta_i \cdot x_i+...+\beta_k \cdot x_k] = \beta_i$$

  - The magnitude of β_i does not depend on the values at which the other x's than x_i are fixed
  - The value of β_i is not generally the same as the slope when you fit a line with x_i alone

Multiple linear regression

The least squares method chooses a, b_1, b_2, ..., b_k (estimates for $\alpha, \beta_1, \beta_2, \ldots, \beta_k$) to minimize the sum of the squares of the residuals

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2,$$

where

$$\hat{y}_i = a + b_1 \cdot x_{1i} + b_2 \cdot x_{2i} + \ldots + b_k \cdot x_{ki}$$

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Multiple linear regression

- When the model contains two explanatory variables x_1 and x_2, the least squares estimates for β_1, β_2 and α are:

$$b_1 = \frac{s_y}{s_{x_1}} \cdot \frac{r(y, x_1) - r(y, x_2) \cdot r(x_1, x_2)}{1 - [r(x_1, x_2)]^2},$$

$$b_2 = \frac{s_y}{s_{x_2}} \cdot \frac{r(y, x_2) - r(y, x_1) \cdot r(x_1, x_2)}{1 - [r(x_1, x_2)]^2},$$

$$a = \bar{y} - b_1 \cdot \bar{x}_1 - b_2 \cdot \bar{x}_2,$$

where $\bar{x}_1$, $\bar{x}_2$ and $\bar{y}$ are the sample means of x_1, x_2 and y; s_{x_1}, s_{x_2} and s_y are the sample standard deviations of x_1, x_2 and y; $r(y, x_1)$ is the sample correlation between y and x_1; $r(y, x_2)$ is the sample correlation between y and x_2; $r(x_1, x_2)$ is the sample correlation between x_1 and x_2.

Multiple linear regression

Test of overall model significance (overall F-test):

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$
(there is no linear relationship between y and x variables)

$$H_1 : \text{ not all } \beta_i = 0$$
(at least one of the explanatory variables is linearly related to y)

Under H_0, the test statistic

$$F = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2/k}{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2/(n-k-1)}$$

follows an F-distribution with k and $n-k-1$ degrees of freedom.

Multiple linear regression

Test of significance of a specific explanatory variable x_i (t-test):

$$H_0 : \beta_i = 0$$
(there is no linear relationship between y and x_i)

$$H_1 : \beta_i \neq 0$$
(there is a linear relationship between y and x_i)

Under H_0, the test statistic

$$T = \frac{b_i}{SE(b_i)}$$

follows a Student-t distribution with $n - k - 1$ degrees of freedom. Here, $SE(b_i)$ is the standard error of b_i calculated from the data.

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Multiple linear regression

To assess goodness of fit of a regression model (i.e. how well the model predicts the observed values of the dependent variable) we can:

1. Calculate the correlation coefficient R between the predicted and observed values of y
   - The closer the correlation to either 1 or -1, the better the model fit
2. Calculate R-square:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

   - R^2 is the proportion of total variability in y explained by a set of explanatory variables $x_1, x_2, \ldots, x_k$
   - The closer R^2 to 1, the better fit of the model

Multiple linear regression

Example: Blood pressure (mmHg), body weight (kg) and pulse (beats/min) in 20 patients with hypertension[1]

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	514.412	2	257.206	95.913	.000[b]
	Residual	45.588	17	2.682		
	Total	560.000	19			

a. Dependent Variable: BP

b. Predictors: (Constant), Pulse, Weight

[1] Daniel, W.W. and Cross, C.L.(2013). *Biostatistics: a foundation for analysis in the health sciences, 10th edition*.

Multiple linear regression

Example: Blood pressure (mmHg), body weight (kg) and pulse (beats/min) in 20 patients with hypertension

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-1.453	8.394		-.173	.865
	Weight	1.061	.116	.839	9.118	.000
	Pulse	.240	.131	.168	1.826	.085

a. Dependent Variable: BP

$$BP = -1.45 + 1.06 \cdot \text{Weight} + 0.24 \cdot \text{Pulse}$$

- After adjusting for pulse, every 1 kg increase in body weight leads to an average increase in blood pressure of 1.06 mmHg
- After adjusting for body weight, every 1 beat/min increase in pulse rate results in an average increase in blood pressure of 0.24 mmHg
- Weight contributes more to the prediction of blood pressure than pulse ($Beta_{Weight} = 0.839$, $Beta_{Pulse} = 0.168$)

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Multiple linear regression

Example: Blood pressure (mmHg), body weight (kg) and pulse (beats/min) in 20 patients with hypertension

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.958[a]	.919	.909	1.63757

a. Predictors: (Constant), Pulse, Weight

Multiple linear regression

Assumptions of multiple linear regression:

1. Independence: the observations are independent, i.e. there is only one set of observations $(y, x_1, \ldots, x_k)$ per subject
2. Linearity: y is a linear function of $x_1, x_2, \ldots, x_k$
3. Constant variance: the variance of y is constant for each possible combination of the values of $x_1, x_2, \ldots, x_k$
4. Normality: Given $x_1, \ldots x_k$, the dependent variable y has a normal distribution
5. No multicollinearity: there is no exact linear association between two or more explanatory variables

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Multiple linear regression

Checking linearity assumption:

- We cannot use scatter plot of y versus x to evaluate the linearity assumption
- Instead, we plot the residuals versus each explanatory variable separately
  - each graph should show a random scatter of points around the horizontal line at zero and no systematic pattern

Multiple linear regression

Multicollinearity:

- Occurs when any of the explanatory variables has a perfect or nearly perfect linear relationship with at least one other variable in the model

  Examples:

  - Inclusion of a variable that is computed from other variables in the model (e.g. BMI is a function of body weight and height, and regression model includes all 3 variables)
  - Inclusion of the same variable twice (e.g. height in centimeters and in meters)
  - Improper use of dummy variables (i.e. failure to remove a dummy for the reference category)
  - Inclusion of truly highly correlated variables (e.g. BMI and waist circumference)
- Results in imprecise and unreliable estimates of partial regression coefficients or even no estimates at all

Multiple linear regression

Typical signals of multicollinearity:

- The estimated partial regression coefficients change drastically when an explanatory variable is added or removed
- The signs of the estimated partial regression coefficients do not conform to theoretical considerations or prior experience
  - Example: the estimated partial regression coefficient of x is negative when theoretically y should increase with increasing values of x
- The overall F-test rejects the null hypothesis, but none of the partial regression coefficients is significant on the basis of t-test
- Large correlation coefficients between pairs of explanatory variables

Multiple linear regression

Detection of multicollinearity:

- Pairwise correlations between explanatory variables
  - Large correlations ($\geq$0.7 or $\leq$-0.7) are a sign of multicollinearity
- *Tolerance* associated with each explanatory variable x_i equal to 1-R_i^2, where R_i^2 is the R^2 of a model with x_i as the dependent variable and the remaining x variables as the explanatory variables
  - If all *tolerance* values are 1 then none of the variables is linearly related to others $\rightarrow$ no multicollinearity
  - If some *tolerance* values are smaller than 1 multicollinearity might be present
  - General rule of thumb: *tolerance* values < 0.2 are a cause of concern while *tolerance* values < 0.1 are a sign of serious multicollinearity
- *Variance inflation factor* (VIF) associated with each variable (=1/*tolerance*)
  - General rule of thumb: VIF values > 5 are a cause of concern while VIF values > 10 are a sign of serious multicollinearity

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Multiple linear regression

Remedial measures for multicollinearity:

- Make sure there are no flagrant errors, e.g. improper use of computed or dummy variables
- Collect additional data that break the pattern of multicollinearity
- Remove one or more explanatory variables in order to lessen multicollinearity

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Multiple linear regression

Example: BMI in 100 women with diabetes

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	18.430	3.281		5.618	.000		
	Children	1.218	1.001	.250	1.218	.226	.188	5.317
	Pregnancies	-1.380	.778	-.373	-1.773	.079	.180	5.563
	Age	.026	.100	.030	.259	.796	.581	1.722
	BloodPressure	.170	.043	.390	3.996	.000	.832	1.202
	Diabetes	5.218	1.905	.263	2.739	.007	.858	1.166

a. Dependent Variable: BMI

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Multiple linear regression

Example: BMI in 100 women with diabetes

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	18.430	3.281		5.618	.000		
	Children	1.218	1.001	.250	1.218	.226	.188	5.317
	Pregnancies	-1.380	.778	-.373	-1.773	.079	.180	5.563
	Age	.026	.100	.030	.259	.796	.581	1.722
	BloodPressure	.170	.043	.390	3.996	.000	.832	1.202
	Diabetes	5.218	1.905	.263	2.739	.007	.858	1.166

a. Dependent Variable: BMI

Correlations

		Children	Pregnancies
Children	Pearson Correlation	1	.901**
	Sig. (2-tailed)		.000
	N	100	100
Pregnancies	Pearson Correlation	.901**	1
	Sig. (2-tailed)	.000	
	N	100	100

**. Correlation is significant at the 0.01 level (2-tailed).

Multiple linear regression

Example: BMI in 100 women with diabetes

Does removal of "Children" reduce multicollinearity?

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	18.430	3.281		5.618	.000		
	Children	1.218	1.001	.250	1.218	.226	.188	5.317
	Pregnancies	-1.380	.778	-.373	-1.773	.079	.180	5.563
	Age	.026	.100	.030	.259	.796	.581	1.722
	BloodPressure	.170	.043	.390	3.996	.000	.832	1.202
	Diabetes	5.218	1.905	.263	2.739	.007	.858	1.166

a. Dependent Variable: BMI

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	18.696	3.282		5.697	.000		
	Pregnancies	-.554	.382	-.150	-1.450	.150	.750	1.334
	Age	.032	.100	.037	.319	.750	.582	1.718
	BloodPressure	.168	.043	.384	3.928	.000	.834	1.199
	Diabetes	5.213	1.910	.263	2.730	.008	.858	1.166

a. Dependent Variable: BMI

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Multiple linear regression

Example: BMI in 100 women with diabetes

Does removal of "Pregnancies" reduce multicollinearity?

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	18.430	3.281		5.618	.000		
	Children	1.218	1.001	.250	1.218	.226	.188	5.317
	Pregnancies	-1.380	.778	-.373	-1.773	.079	.180	5.563
	Age	.026	.100	.030	.259	.796	.581	1.722
	BloodPressure	.170	.043	.390	3.996	.000	.832	1.202
	Diabetes	5.218	1.905	.263	2.739	.007	.858	1.166

a. Dependent Variable: BMI

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	18.826	3.310		5.688	.000		
	Children	-.329	.495	-.067	-.664	.509	.784	1.275
	Age	-.005	.100	-.006	-.053	.958	.599	1.669
	BloodPressure	.167	.043	.382	3.869	.000	.834	1.199
	Diabetes	5.275	1.926	.266	2.739	.007	.858	1.165

a. Dependent Variable: BMI

Multiple linear regression

Methods of variables selection based on mathematical criteria

- Forced entry: All predictors are forced into the model simultaneously
- Hierarchical blockwise entry (called "Remove" in SPSS): all variables in a block are removed from the model simultaneously
- Stepwise selection: gives regression model containing only significant predictors of the dependent variable based on a set of candidate predictor variables
  - Bidirectional elimination (called "Stepwise" in SPSS)
  - Forward elimination
  - Backward elimination

Multiple linear regression

Bidirectional elimination:

- The model is built by successfully adding or removing variables based on *t*-tests for their partial regression coefficients
  - At each step, a variable is added whose *t*-test p-value is the smallest below some threshold (Probability-of-F-to-enter in SPSS; usually 0.05)
  - At each step, a variable is removed whose *t*-test p-value is the highest above some threshold (Probability-of-F-to-remove in SPSS; usually 0.1)

- Step 1:
  - Fit k simple linear regression models, one for each candidate predictor variable x_i ($i = 1, \ldots, k$)
  - Find a variable with the smallest individual coefficient *t*-test p-value
    - If the p-value is below the entry threshold, add the variable to the null model and go to Step 2.
    - If not, stop. No variable is significant predictor of the dependent variable.

Multiple linear regression

- Step 2: Suppose that x_1 entered the model at Step 1.
  - Fit $k-1$ two-predictor regression models with x_1 as one of the explanatory variables
  - Find a variable (other than x_1) with the smallest individual coefficient t-test p-value
    - If the p-value is below the entry threshold, add the variable to the model.
    - If not, stop. Variable x_1 is the only significant predictor of the dependent variable.
  - Suppose x_2 entered the model at Step 2. Step back to check p-value for β_1 in the model involving x_1 and x_2.
    - If the p-value is above the removal threshold, remove x_1 from the model and repeat Step 2 to find the second important predictor after x_2.
    - If not, keep x_1 in the model and go to Step 3 to find the third significant predictor.
- ... This procedure is continued until no more variables can be added.

Multiple linear regression

Example: Clinical data of 20 patients with hypertension

Excluded Variables[a]

Model		Beta In	t	Sig.
1	Age	,387[b]	4,225	,001
	Dur	,183[b]	1,604	,127
	Pulse	,407[b]	4,279	,001
	Stress	,148[b]	1,277	,219
2	Age	,249[c]	2,699	,016
	Dur	,049[c]	,523	,608
	Stress	-,080[c]	-,775	,450
3	Dur	,019[d]	,235	,818
	Stress	-,111[d]	-1,289	,217

a. Dependent Variable: BP

b. Predictors in the Model: (Constant), BSA

c. Predictors in the Model: (Constant), BSA, Pulse

d. Predictors in the Model: (Constant), BSA, Pulse, Age

Coefficients[a]

Model		Unstandardized Coefficients		t	Sig.
		B	Std. Error		
1	(Constant)	45,183	9,392	4,811	,000
	BSA	34,443	4,690	7,343	,000
2	(Constant)	19,791	8,955	2,210	,041
	BSA	26,921	3,782	7,117	,000
	Pulse	,581	,136	4,279	,001
3	(Constant)	9,774	8,504	1,149	,267
	BSA	25,771	3,260	7,906	,000
	Pulse	,380	,138	2,755	,014
	Age	,541	,201	2,699	,016

a. Dependent Variable: BP

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Multiple linear regression

Example: Clinical data of 20 patients with hypertension

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	BSA	.	Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100).
2	Pulse	.	Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100).
3	Age	.	Stepwise (Criteria: Probability-of-F-to-enter <= ,050, Probability-of-F-to-remove >= ,100).

a. Dependent Variable: BP

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Multiple linear regression

Forward elimination

- Method the same as bidirectional elimination except that each time a predictor is added to the equation, a removal test is not made to eliminate the least useful predictor

Backward elimination

- Method opposite to the forward elimination method
  - All predictors are placed in the model and contribution of each predictor is obtained based on *t*-test
  - Predictor with the smallest contribution is removed if it meets the removal criterion and the model is reevaluated
  - Procedure is continued until no more variables can be removed

Multiple linear regression

Approaches of variable selection decisions:

- There are different approaches for selecting model variables
- You should always choose the one that meets the aim of your study
  - The most common study aims:
    1. Identification of predictors of the dependent variable of interest
    2. Evaluation of association between the dependent variable and one primary explanatory variable
    3. Prediction of the dependent variable

Multiple linear regression

Aim 1: Identification of important predictors

- Goal: find out if any of the potential predictor variables are significant predictors of the dependent variable and if so, which one(s)
- To achieve this goal we can use:
  - The overall *F*-test in combination with the individual coefficient *t*-tests
  - Automated methods of variable selection can be applied

Multiple linear regression

- Overall *F*-test in combination with *t*-tests:
  - Step 1: Fit linear regression model with all potential predictors as explanatory variables
  - Step 2: Perform overall *F*-test
    - If the null hypothesis is not rejected, conclude that none of the explanatory variables are significant predictors of the dependent variable
    - If the null hypothesis is rejected, go to Step 3
  - Step 3: Conduct individual *t*-test on each partial regression coefficient
    - Variables for which p-value of the test statistic is less than 0.05 are deemed to be significant predictors of the dependent variable

Multiple linear regression

Example: Clinical data of 20 patients with hypertension

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	518,672	5	103,734	35,141	,000[b]
	Residual	41,328	14	2,952		
	Total	560,000	19			

a. Dependent Variable: BP

b. Predictors: (Constant), Stress, BSA, Dur, Age, Pulse

BSA- body surface area; Dur - duration of hypertension

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Multiple linear regression

Example: Clinical data of 20 patients with hypertension

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	6,212	9,429		,659	,521
	Age	,563	,206	,259	2,737	,016
	BSA	24,554	3,452	,617	7,114	,000
	Pulse	,456	,159	,320	2,866	,012
	Dur	,077	,204	,030	,376	,713
	Stress	-,017	,013	-,114	-1,284	,220

a. Dependent Variable: BP

BSA- body surface area; Dur - duration of hypertension

Multiple linear regression

Example: Clinical data of 20 patients with hypertension

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,962[a]	,926	,900	1,71813

a. Predictors: (Constant), Stress, BSA, Dur, Age, Pulse

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,958[a]	,917	,902	1,70232

a. Predictors: (Constant), Pulse, BSA, Age

BSA- body surface area; Dur - duration of hypertension

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Multiple linear regression

Aim 2: Evaluation of association between the dependent variable and one primary explanatory variable

- Goals: obtain unbiased estimate of association between the explanatory variable of interest and the dependent variable in observational studies; increase precision and power of the treatment effect estimate in randomized studies
- To achieve these goals: we include the variable of primary interest as well as confounding variables in the model for observational data; we include treatment variable as well as covariates which are imbalanced and strongly predictive of the outcome in the model for randomized study data

Multiple linear regression

- To identify confounders we may use '10% change-in-estimate' approach
  1. Identify variables that could potentially affect the association under study
  2. Fit simple linear regression model for the explanatory variable of primary interest
  3. Fit the model with the variable of interest and separately each potential confounding variable x_i
     - If the estimate of the regression coefficient from the simple linear regression model changes by 10% or more, then x_i is considered a confounder and is added to the model

Multiple linear regression

Example: Clinical data of 20 patients with hypertension

Dependent variable: blood pressure (BP)
Variable of primary interest: body surface area (BSA)

Simple linear regression function: $BP = 45.18 + \underline{34.44 \cdot \text{BSA}}$

Potential confounder	Estimate of BSA coefficient in two-predictor model	Percentage change in estimate (%)
Age	28.62	– 16.91
Pulse	26.92	– 21.84
Dur	33.49	– 2.76
Stress	34.33	– 0.32

$$BP = 9.77 + 0.54 \cdot \text{Age} + \underline{25.77 \cdot \text{BSA}} + 0.38 \cdot \text{Pulse}$$

NKI-AVL
The Netherlands Cancer Institute
Antoni van Leeuwenhoek Hospital

Multiple linear regression

Aim 3: Prediction of the dependent variable

- Goal: build a model to help predict the dependent variable
- To achieve this goal:
  - Identify variables that may explain the variation in the dependent variable
  - Fit a regression model with many variables to maximize R^2
  - Evaluate models with different subsets of predictors
  - Model generalizability based on prediction error obtained with different independent set of observations (validation)

Multiple linear regression

General remarks about selection of model variables:

- Stepwise regressions does not take into account a researcher's knowledge about the predictors. That is, it is possible that some unimportant variables will end up in the model and some important variable will not be included. Thus, this approach should only be used to help guide your decisions.
- General rules of thumb:
  - The number of partial regression coefficients in the model must be smaller than the number of observations
  - There should be at least 10 observations per partial regression coefficient
    (i.e. # observations/$k \geq 10$)