Basic Medical Statistics Course

Multiple linear regression

S6

Patrycja Gradowska
p.gradowska@nki.nl

November 23, 2016

NETHERLANDS CANCER INSTITUTE
ANTONI VAN LEEUWENHOEK

Introduction

- In the previous lecture, we learned about simple linear regression which involves a single continuous dependent variable y and a single, continuous or categorical, independent variable x
- We are now going to discuss **multiple linear regression**, i.e. an extension of simple linear regression that accommodates two or more independent variables

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

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

where:

- y is the dependent variable
- $x_1, x_2, \ldots, x_k$ are the explanatory variables, predictors or covariates
- $\alpha, \beta_1, \beta_2, \ldots, \beta_k$ are the unknown population regression coefficients
  - α is the intercept or constant term
  - $\beta_1, \beta_2, \ldots, \beta_k$ 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:

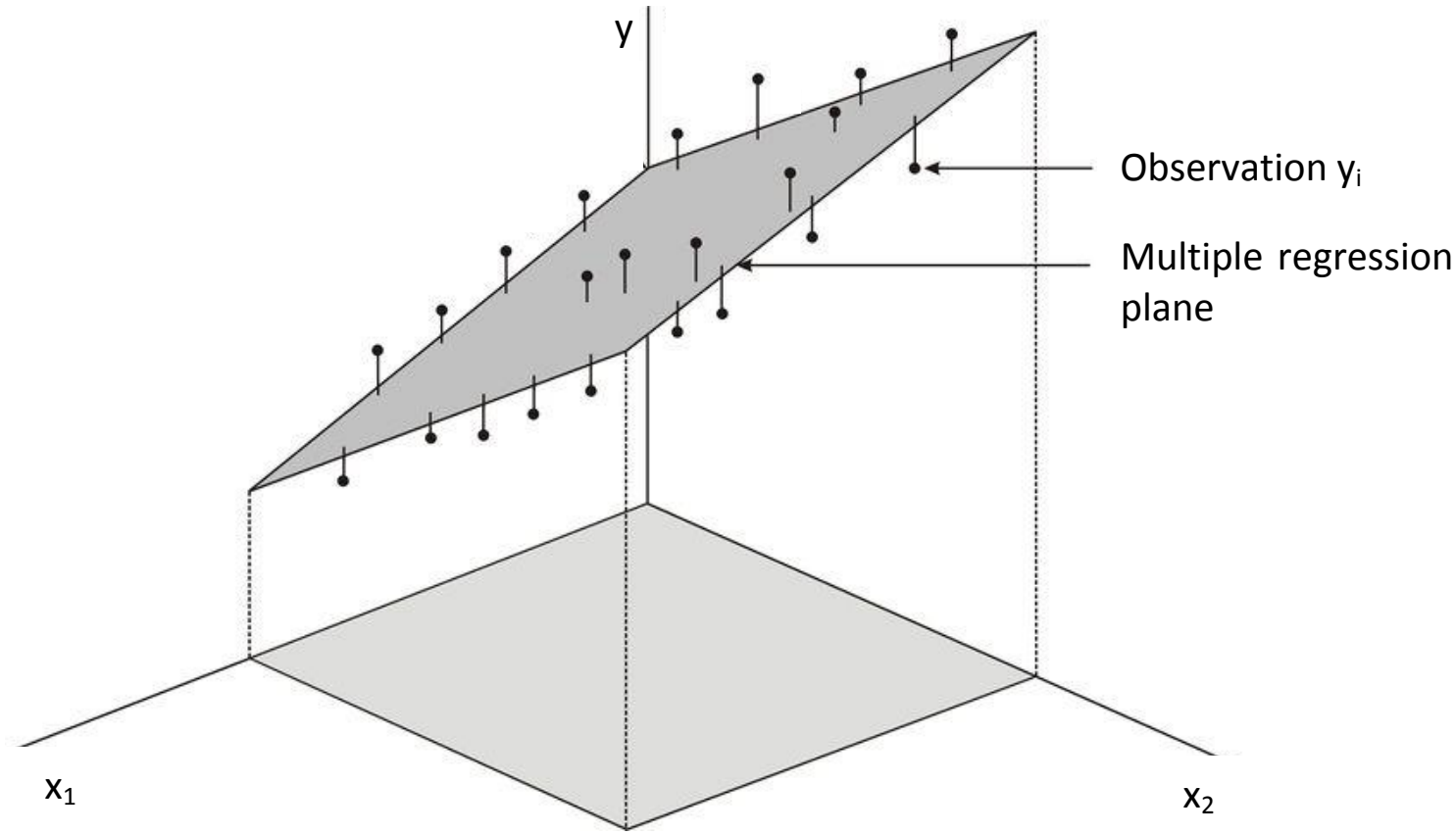
$$E(y|x_1, x_2, \ldots, x_k) = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \ldots + \beta_k \cdot x_k,$$

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

Note:

- 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



Multiple linear regression

Interpretation of model parameters:

- α is the mean value of y when all explanatory variables equal to 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:

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

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

Note:

- The magnitude of β_i does not depend on the values at which the other x's 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

Estimation of model parameters:

- Based on available data on y and $x_1, x_2, \ldots, x_k$, we wish to estimate $\alpha, \beta_1, \beta_2, \ldots, \beta_k$
- As in the case of simple linear regression, we can use the **least squares method** that minimizes 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

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

and $a, b_1, b_2, \ldots, b_k$ are sample estimates of $\alpha, \beta_1, \beta_2, \ldots, \beta_k$, respectively.

Multiple linear regression

- Unlike the simple linear regression, the estimates of coefficients in the multiple linear regression model have somewhat complicated forms
- In case when the model contains two explanatory variables x_1 and x_2, the coefficients are calculated as

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

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

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

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

Testing partial regression coefficients:

- The significance tests used for simple linear regression model were the *t*-test and the *F*-test, which always generated the same conclusion
- In multiple linear regression, these tests have different purposes:
  1. The *F*-test is used to determine the significance of the overall model, i.e. whether there is a linear relationship between the dependent variable *y* and the set of all explanatory variables
  2. The *t*-test is used to determine the significance of each independent variable individually, i.e. whether there is a linear association between the dependent variable *y* and each individual explanatory variable

Multiple linear regression

***F*-test of overall significance (*overall F-test*)**:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$
(there is no linear relationship between y and the 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.

If the null hypothesis can be rejected, then there is enough evidence to conclude that at least one explanatory variable is linearly associated with y.

Multiple linear regression

***t*-test of significance of a specific explanatory variable x_i:**

$$H_0 : \beta_i = 0$$
(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.

If the null hypothesis can be rejected, then there is enough evidence to conclude that the explanatory variable x_i is linearly related to y.

Multiple linear regression

Goodness of fit:

As in simple linear regression, we assess goodness of fit (i.e. how well the model predicts the observed values of the dependent variable) by looking at:

1. The product moment correlation between the observed and predicted values of the dependent variable
   - The closer the correlation is to either 1 or -1, the better the model fit
2. R-square calculated as

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

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

Multiple linear regression

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

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

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

Multiple linear regression

Assumptions of multiple linear regression:

1. **Independence**: the observations are independent, i.e. there is only one set of observations 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**: y has a Normal distribution
5. **No multicollinearity:** there is no exact linear association between two or more explanatory variables

Multiple linear regression

Comment on checking the 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
  - If the assumption of linearity is met, then each graph should show a random starter of points around the horizontal line at zero and no systematic pattern

Multiple linear regression

Multicollinearity:

- It is the situation when any of the explanatory variables has a perfect or nearly perfect linear relationship with some or all other variables 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. body weight and height)
- 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. For example, the estimated partial regression coefficient is negative when theoretically y should increase with increasing values of that x variable
- 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

Detect multicollinearity by examining:

- Pairwise correlations between explanatory variables
  - Large correlations (e.g. 0.7 or above in abs. value) 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 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

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
- It may be that the best thing to do is simply to do nothing - realize that multicollinearity is present and be aware of its consequences

Multiple linear regression

Example 2: U.S. Census Bureau data on 12 states

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	27,713	11,104		2,496	,055		
	White	-,040	,031	-,231	-1,283	,256	,560	1,784
	Crime	-,004	,003	-,221	-1,274	,259	,608	1,645
	Traf Deaths	-,271	3,403	-,030	-,080	,940	,132	7,603
	University	,030	,240	,051	,125	,906	,108	9,257
	Unemployed	1,255	,552	,408	2,274	,072	,566	1,765
	Income	-,324	,076	-1,142	-4,277	,008	,256	3,914

a. Dependent Variable: Poverty

Multiple linear regression

Example 2: U.S. Census Bureau data on 12 states

Does removal of University reduce multicollinearity?

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	28,763	6,614		4,349	,005		
	White	-,039	,027	-,226	-1,412	,208	,595	1,682
	Crime	-,004	,003	-,225	-1,453	,196	,635	1,576
	Traf Deaths	-,609	1,877	-,067	-,324	,757	,361	2,768
	Unemployed	1,260	,503	,409	2,503	,046	,569	1,757
	Income	-,321	,064	-1,129	-5,009	,002	,300	3,338

a. Dependent Variable: Poverty

Multiple linear regression

Selection of model variables:

- There are different approaches to selecting model variables
- But you should always choose the one that meets the aim of your study
- The most common study aims are:
  - Identification of predictors of the dependent variable of interest
  - Evaluation of association between the dependent variable and one primary explanatory variable
  - 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
  - Stepwise regression

Multiple linear regression

Overall *F*-test in combination with *t*-tests:

Step 1: Fit the linear regression model with all potential predictors as explanatory variables

Step 2: Perform the 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 the 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 3: Clinical data of 20 patients with hypertension

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	557.844	6	92.974	560.641	.000[b]
	Residual	2.156	13	.166		
	Total	560.000	19			

a. Dependent Variable: BP

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

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-12.870	2.557		-5.034	.000
	Weight	.970	.063	.767	15.369	.000
	Age	.703	.050	.324	14.177	.000
	BSA	3.776	1.580	.095	2.390	.033
	Stress	.006	.003	.038	1.633	.126
	Pulse	-.084	.052	-.059	-1.637	.126
	Dur	.068	.048	.027	1.412	.182

a. Dependent Variable: BP

BSA- body surface area; Dur - duration of hypertension

Multiple linear regression

Stepwise regression:

- One of the most commonly used variable selection methods
- Goal: develop regression model containing only significant predictors of the dependent variable based on a set of candidate predictor variables
- 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)

Multiple linear regression

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 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.

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 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 3 revisited: Clinical data of 20 patients with hypertension

Variables Entered/Removed[a]

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

a. Dependent Variable: BP

NETHERLANDS CANCER INSTITUTE
ANTONI VAN LEEUWENHOEK

Multiple linear regression

Example 3 revisited: Clinical data of 20 patients with hypertension

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2.205	8.663		.255	.802
	Weight	1.201	.093	.950	12.917	.000
2	(Constant)	-16.579	3.007		-5.513	.000
	Weight	1.033	.031	.817	33.154	.000
	Age	.708	.054	.326	13.235	.000
3	(Constant)	-13.667	2.647		-5.164	.000
	Weight	.906	.049	.717	18.490	.000
	Age	.702	.044	.323	15.961	.000
	BSA	4.627	1.521	.116	3.042	.008

a. Dependent Variable: BP

Multiple linear regression

Example 3 revisited: Clinical data of 20 patients with hypertension

Excluded Variables[a]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics: Tolerance
1	Age	.326[b]	13.235	.000	.955	.834
	BSA	.147[b]	.962	.350	.227	.234
	Stress	.131[b]	1.913	.073	.421	.999
	Pulse	.168[b]	1.826	.085	.405	.565
	Dur	.106[b]	1.463	.162	.334	.960
2	BSA	.116[c]	3.042	.008	.605	.233
	Stress	.019[c]	.748	.465	.184	.848
	Pulse	-.046[c]	-1.353	.195	-.321	.418
	Dur	.019[c]	.784	.445	.192	.877
3	Stress	.021[d]	1.073	.300	.267	.847
	Pulse	-.014[d]	-.452	.658	-.116	.355
	Dur	.026[d]	1.359	.194	.331	.866

a. Dependent Variable: BP

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

c. Predictors in the Model: (Constant), Weight, Age

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

Multiple linear regression

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

- Goal: obtain unbiased estimate of association between the explanatory variable of interest and the dependent variable
- We can achieve this goal by including in the model the variable of primary interest as well as confounding variables
- 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 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 3 revisited: Clinical data of 20 patients with hypertension

Dependent variable: blood pressure (BP)
Variable of primary interest: body weight

Simple linear regression function: $BP = 2.20 + \underline{1.20 \cdot \text{Weight}}$

Potential confounder	Estimate of BW coefficient in two-predictor model	Percentage change in estimate (%)
Age	1.03	– 13.99
BSA	1.04	– 13.49
Pulse	1.06	– 11.66
Dur	1.17	– 2.25
Stress	1.19	– 0.50

$$BP = -13.89 + 0.72 \cdot \text{Age} + \underline{0.92 \cdot \text{Weight}} + 4.33 \cdot \text{BSA} - 0.02 \cdot \text{Pulse}$$

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 all variables to maximize R^2
  - Due to practical reasons or due to high cost of obtaining information on a large number of variables, we may want to fit a model with fewer number of variables (selected using e.g. stepwise regression)

Multiple linear regression

General remarks about selection of model variables:

- Stepwise regression 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.
- Stepwise regression may include only a subset of the dummy variables representing a single categorical variable, leading to problems in interpretation
- General rules of thumb:
  - The number of variables in the model must be smaller than the number of observations
  - There should be at least 10 observations per variable (i.e. # observations/$k \geq 10$)