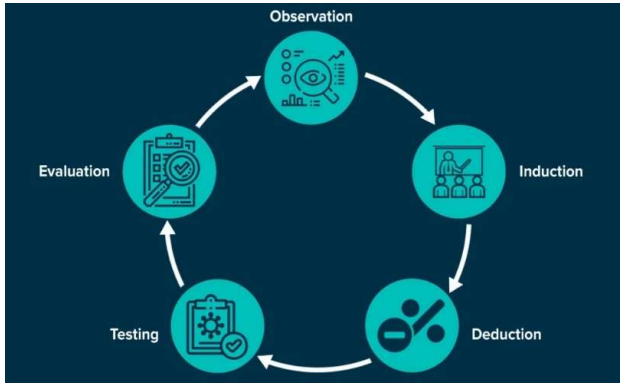
# *Empirical research in management and economics*

## *Multiple regression*




Thorsten Pachur

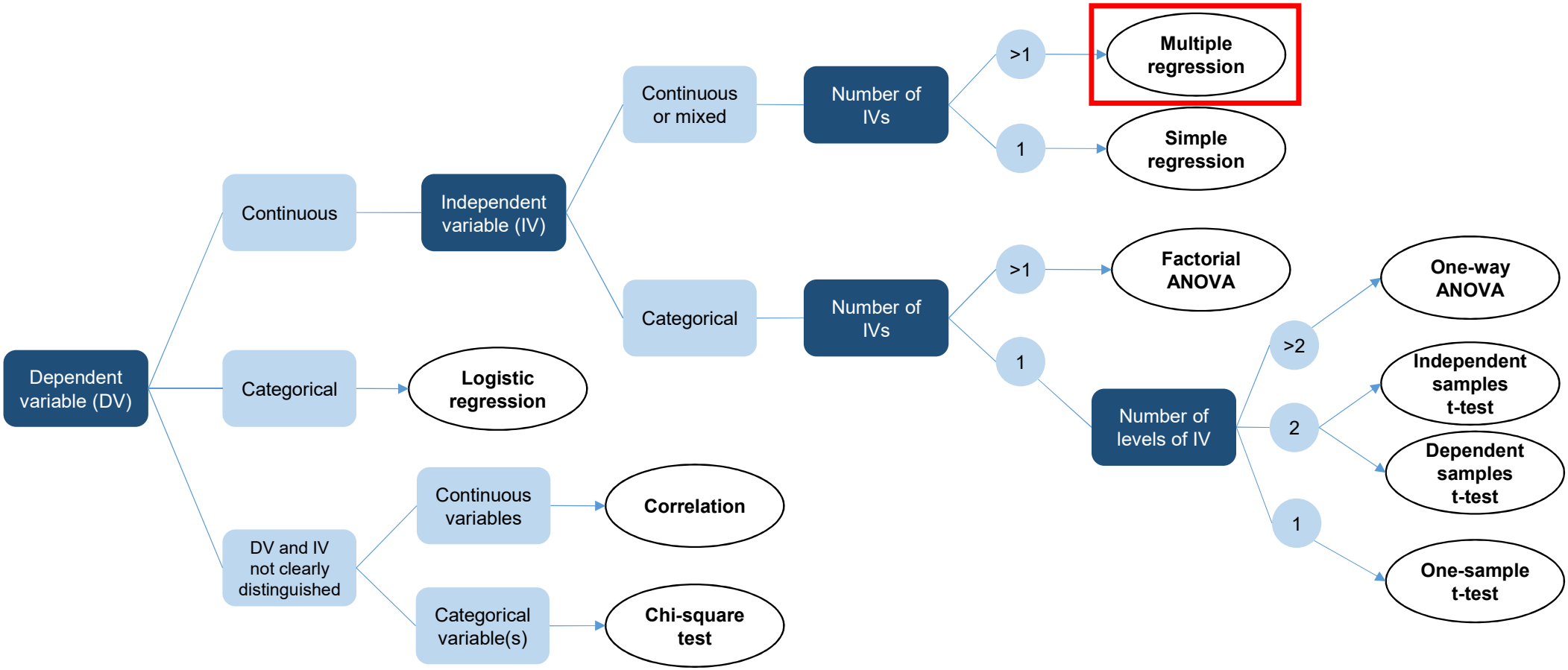
*Technical University of Munich*
*School of Management*
*Chair of Behavioral Research Methods*

# *Recap from last week*

- What are the key purposes of estimating a regression model?
- What are the key parameters of a regression model?
- How are the parameters of a regression model estimated?
- Why can it be helpful to center a predictor?
- How is a regression model evaluated statistically—both in terms of the overall model fit and in terms of the regression coefficients?
- What are key assumptions in simple linear regression—and how can you check whether the assumptions are fulfilled?

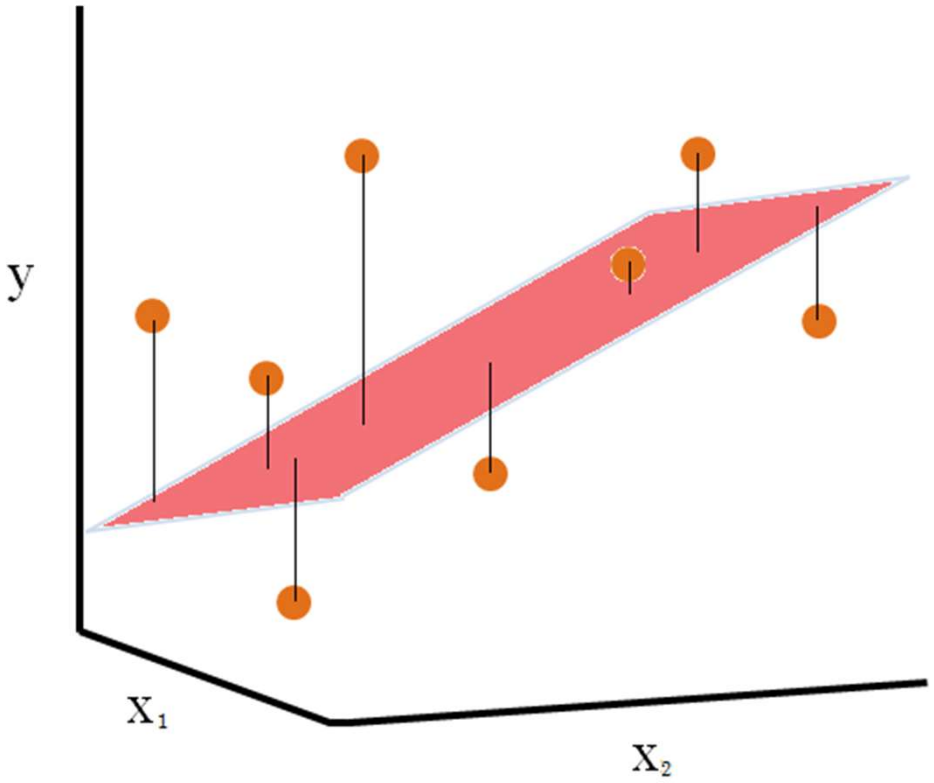
# *Agenda for the semester*

| Session | Date | Topic |
|---|---|---|
| 1 | 13 October | Introduction |
| 2 | 20 October | Descriptive data analysis |
| 3 | 27 October | Hypothesis development and measurement |
| 4 | 3 November | Inferential data analysis I |
| 5 | 10 November | Inferential data analysis II |
| 6 | 17 November | Simple regression |
| **7** | **24 November** | **Multiple regression** |
| 8 | 1 December | Logistic regression |
| 9 | 8 December | Factor analysis |
| 10 | 15 December | Cluster analysis |
| 11 | 12 January | Conjoint analysis |
| 12 | 19 January | The replication crisis and open science |
| 13 | 26 January | Summary and questions |
| | 11 February | Exam |

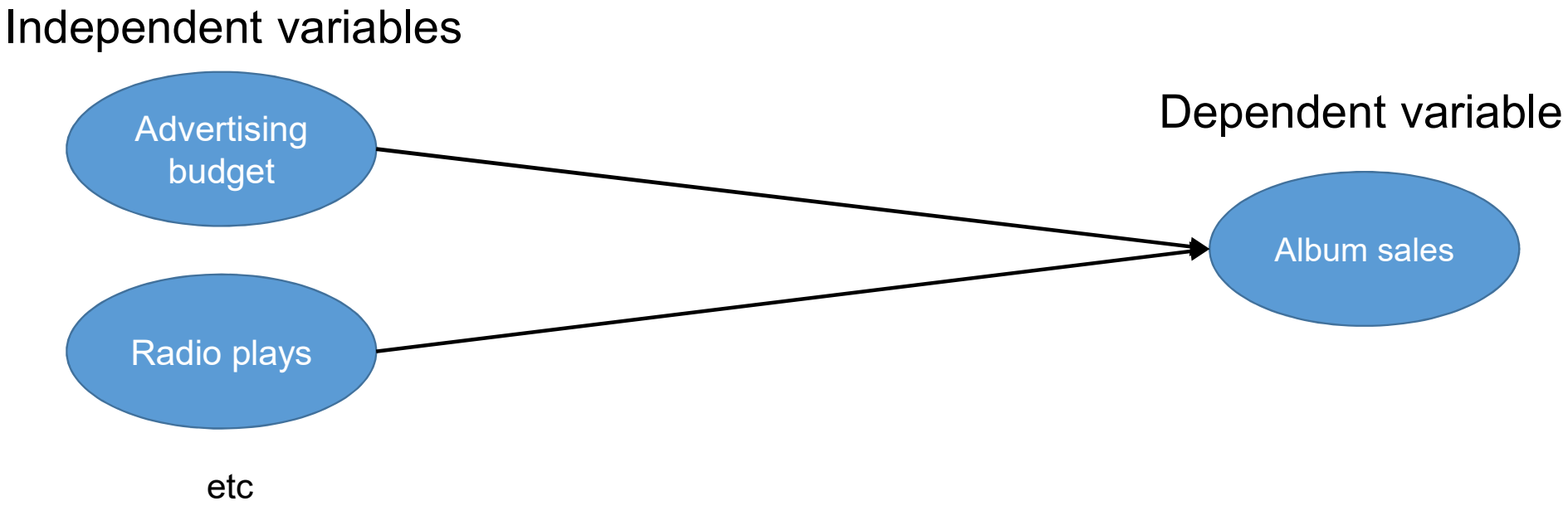

Dependent variable (DV)
Continuous
Independent variable (IV)
Continuous or mixed
Number of IVs
>1
Multiple regression
1
Simple regression
Categorical
Number of IVs
>1
Factorial ANOVA
1
Number of levels of IV
>2
One-way ANOVA
2
Independent samples t-test
Dependent samples t-test
1
One-sample t-test
Categorical
Logistic regression
DV and IV not clearly distinguished
Continuous variables
Correlation
Categorical variable(s)
Chi-square test

# *Goals for this week*

- You know the equation of a multiple regression model
- You know how to evaluate a multiple regression model statistically
- You know what is meant by multicollinearity and how to test for it
- You are familiar with how to do a power analysis for a multiple regression analysis
- You can conduct a moderation analysis and a mediation analysis
- You are familiar with dummy coding and how it is used to include a categorical predictor in a regression model




# *Multiple regression*


Independent variables
Advertising budget
Radio plays
etc
Dependent variable
Album sales

# *Multiple regression equation*

**Slope**

(the amount of change in $\hat{Y}$ associated with a one-unit change in $X_i$, keeping all other predictors constant: $b_i = \frac{\Delta \hat{Y}}{\Delta X_i}$

→ regression coefficient

Predicted value of *Y*

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_n X_n$$

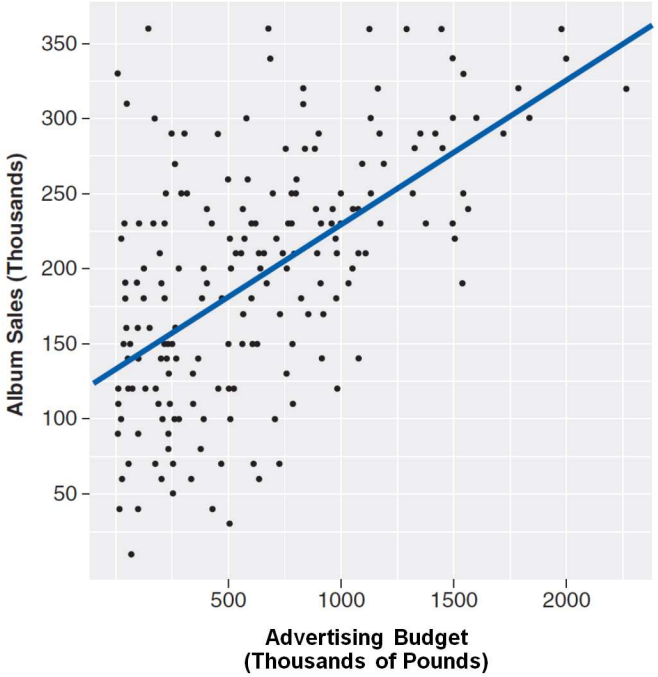
Value of the predictor variables

**Intercept**

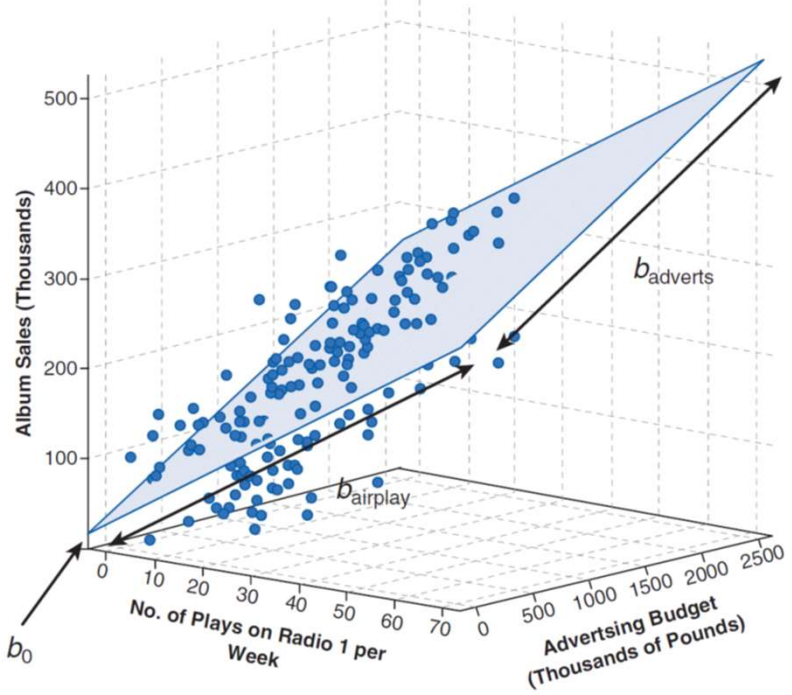
(the value of $\hat{Y}$ when all predictors $X_{1\ldots n} = 0$)

# Geometric representation

$$\widehat{Sales} = b_0 + b_{adverts} \times Adverts$$

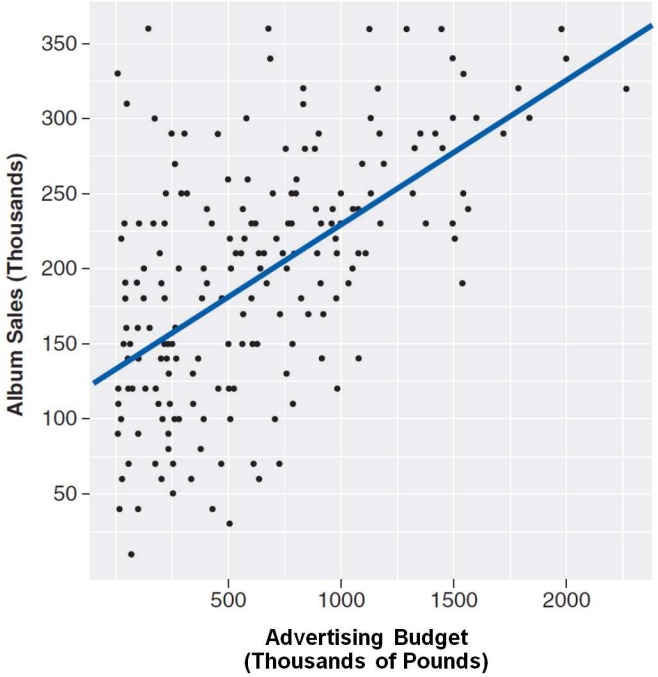



$$\widehat{Sales} = b_0 + b_{adverts} \times Adverts + b_{airplay} \times Airplay$$

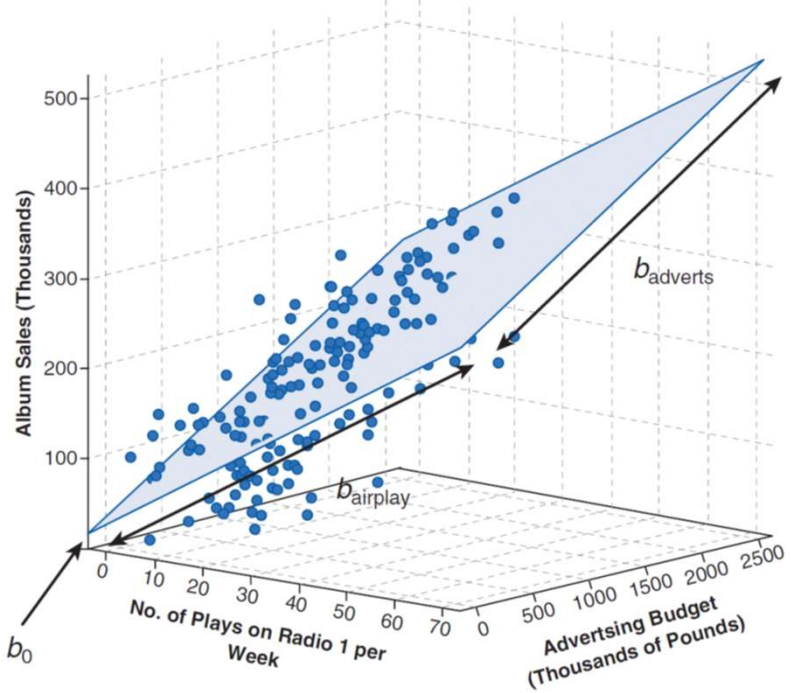


# Geometric representation

$$\widehat{Sales} = 134.1 + 0.096 \times Adverts$$




$$\widehat{Sales} = 41.1 + 0.087 \times Adverts + 3.59 \times Airplay$$



# *Standardized regression coefficients*

→ Allows for a comparison of regression coefficients between predictors

$$\beta_i = b_i \times \frac{SD_{X_i}}{SD_Y}$$

$$\widehat{Sales} = 41.1 + 0.087 \times Adverts + 3.59 \times Airplay$$

$$\beta_{adverts} = 0.087 \times \frac{485.7}{80.7} = .523$$

$$\beta_{airplay} = 3.59 \times \frac{12.3}{80.7} = .546$$

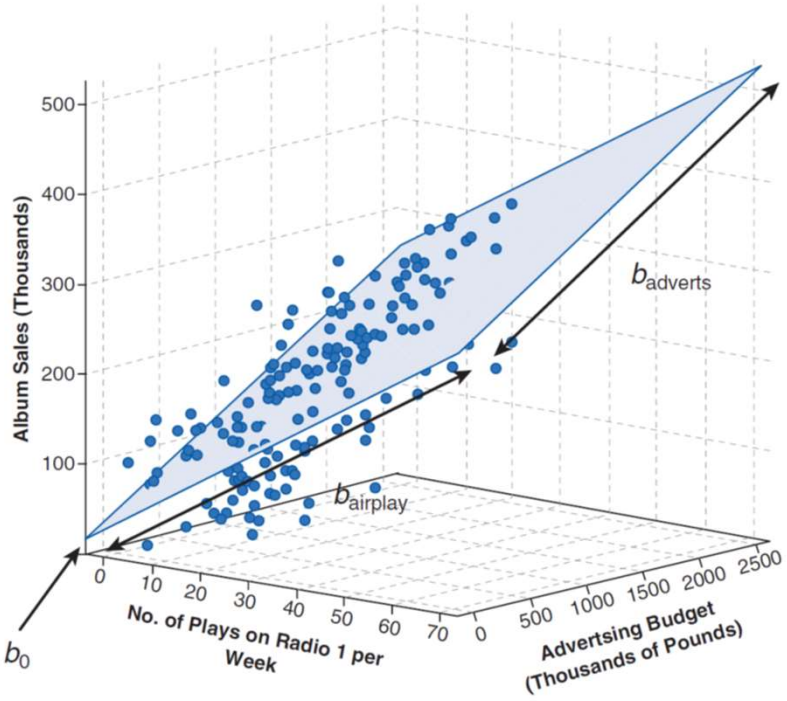
$$z(\widehat{Sales}) = 0 + 0.523 \times z(Adverts) + 0.546 \times z(Airplay)$$

Intercept!

Interpretation of standardized regression coefficients: changes in terms of standard deviations in z-transformed data

$SD_{adverts}$ = 485.7
$SD_{airplay}$ = 12.3
$SD_{sales}$ = 80.7

# *Statistical evaluation of a regression model*



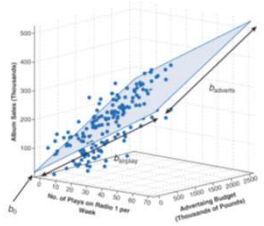

- Evaluating the goodness of fit
  - How much variance in the outcome variable is accounted for by the predictors?
  - Does the model perform better than the baseline model ($\widehat{Y}_i = \bar{Y}$)?
- Do the regression coefficients (*b*s) differ significantly from zero?

# *Statistical evaluation of a regression model*

- Evaluating the fit of the regression model
  - Amount of explained variance

$$R^2 = \frac{\sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2} = \frac{SS_{\hat{Y}}}{SS_Y}$$

$$R^2 = \frac{815{,}524.1}{1{,}295{,}952} = .629$$



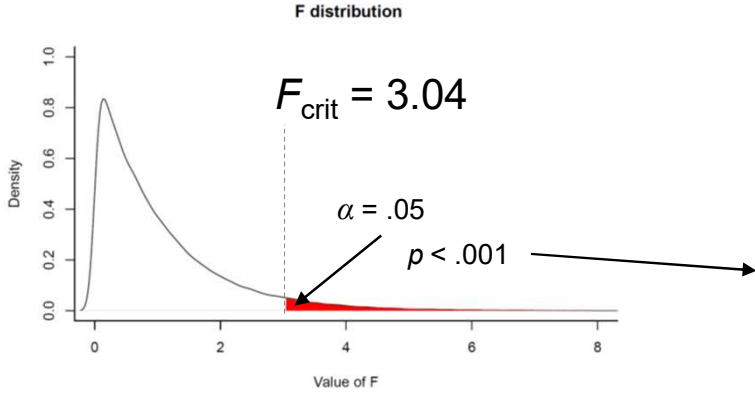
  - *F*-statistic

$$F = \frac{(N - k - 1)R^2}{k(1 - R^2)}$$

$$F = 167.2$$

$df_1 = k$ $df_2 = N\text{-}k\text{-}1$

*k* = number of predictors in the model




- Evaluating the individual regression coefficients *b:* *t*-statistic

$b_{advert} = .087$ $t = 11.087$ $p < .001$ $b_{aiplay} = 3.59$ $t = 12.51$ $p < .001$

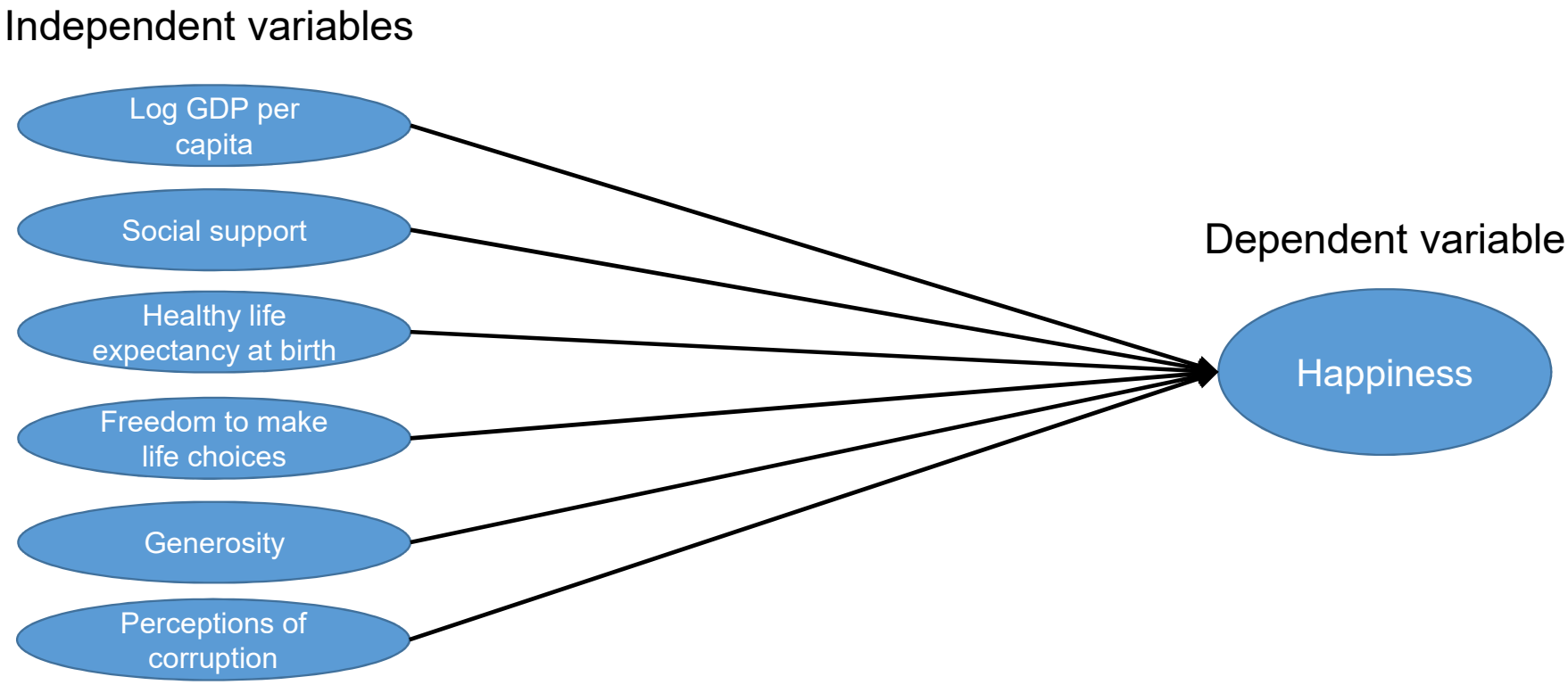
# *Assumptions in multiple regression*

- *Linearity*: The relationship between outcome variable and the predictor variables is linear
- *Homoscedasticity*: At each level of the predictor variable, the variance of the residuals is the same
- The residuals are *normally distributed*
- Absence of (multi)collinearity

  Correlations among predictors should not be too large (~ $r < .8$)

  → If there is multicollinearity, the estimated regression coefficients of the intercorrelated predictors are unstable (i.e., it is difficult to assess the importance of an individual predictor) and their standard errors large

# *Example: World happiness*



# *Checking for multicollinearity*

Intercorrelations (Pearson correlations): Problematic when $r > .8$

| | Log GDP per capita | Social support | Healthy life expectancy at birth | Freedom to make life choices | Generosity | Perceptions of corruption |
|---|---|---|---|---|---|---|
| Log GDP per capita | — | | | | | |
| Social support | 0.753 | — | | | | |
| Healthy life expectancy at birth | 0.857 | 0.720 | — | | | |
| Freedom to make life choices | 0.326 | 0.427 | 0.335 | — | | |
| Generosity | -0.008 | 0.088 | 0.015 | 0.320 | — | |
| Perceptions of corruption | -0.408 | -0.322 | -0.368 | -0.451 | -0.398 | — |

# *Checking for multicollinearity*

- Check intercorrelations among predictors
- Tolerance (= $1 - R^2_{X_i}$; $R^2_{X_i}$ is the correlation of $X_i$ with all other predictors)
  → Degree to which a given predictor can be predicted by the other predictors
  *Guideline*
  - Tolerance below 0.1 indicates that there is a serious problem
  - Tolerance below 0.2 indicates that there is a potential problem
- Variance inflation factor (VIF; = 1/tolerance)
  *Guideline*
  - The largest VIF should not be greater than 10
  - The average VIF should not be substantially greater than 1

If there is multicollinearity
- Consider dropping redundant predictors
- Combine highly correlated predictors with factor analysis

!

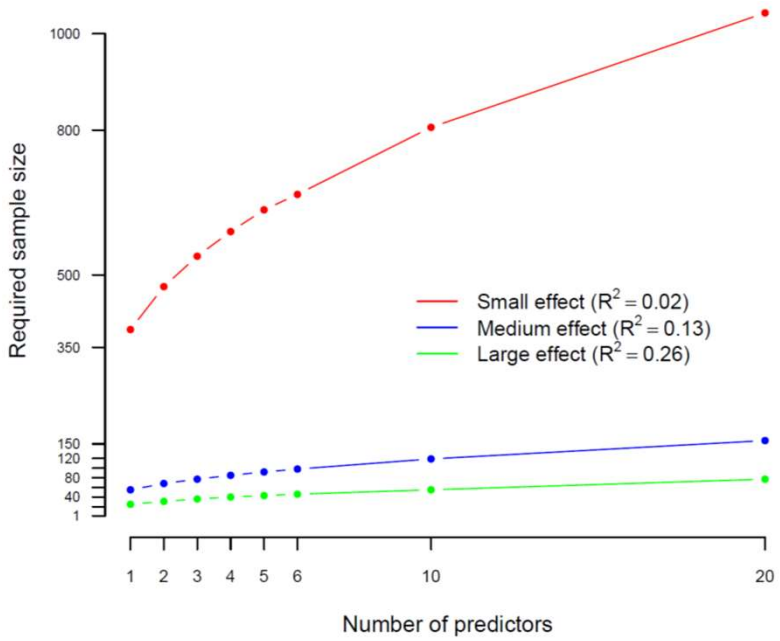
# *Multicollinearity statistics*

| | Tolerance | VIF |
|---|---|---|
| Log GDP per capita | 0.177 | 5.656 |
| Social support | 0.366 | 2.733 |
| Healthy life expectancy at birth | 0.230 | 4.347 |
| Freedom to make life choices | 0.684 | 1.462 |
| Generosity | 0.766 | 1.305 |
| Perceptions of corruption | 0.608 | 1.644 |

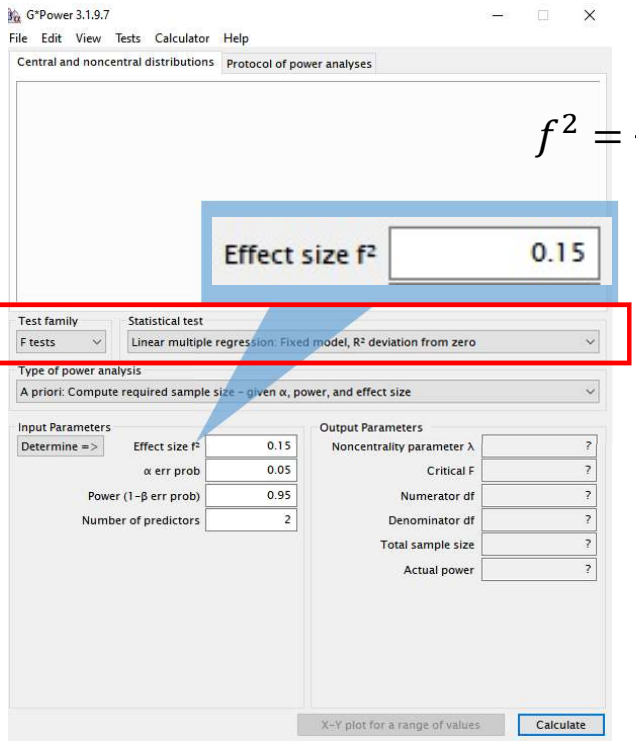
# *Sample-size considerations*

Do power analysis (e.g., with G*Power)

For power = .8

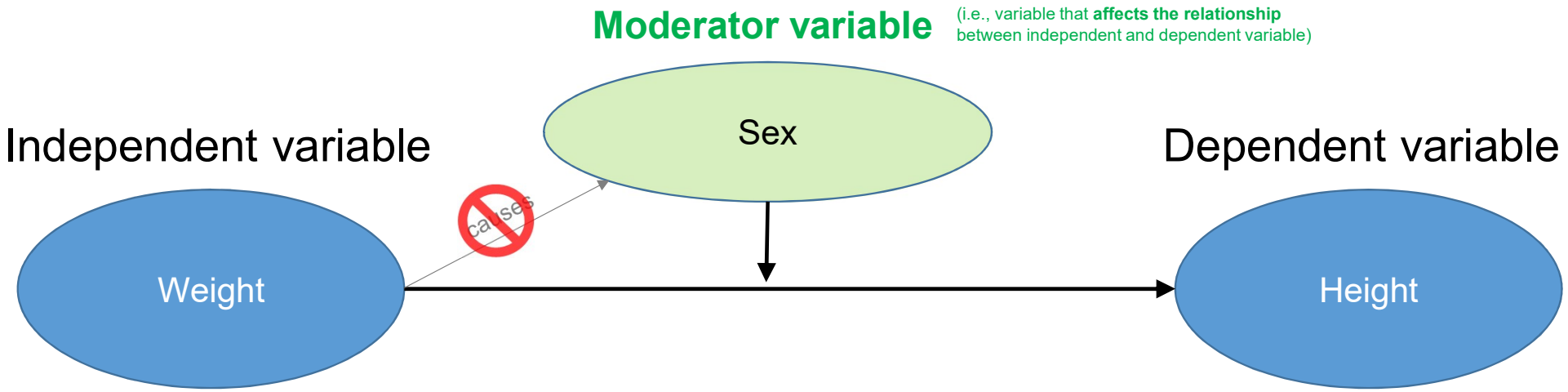



$$f^2 = \frac{R^2}{1 - R^2}$$



# *Applying and Extending Multiple Regression*

# *Moderation analysis*




Does the effect of weight (i.e., the independent variable) on height (i.e., the dependent variable) differ between males and females (i.e., the moderator)?

→ Tested by including the independent variable, the moderator, **and their interaction** as predictors

$$\hat{Y} = b_0 + \underbrace{b_1 \times X_1}_{\text{Independent variable}} + \underbrace{b_2 \times X_2}_{\text{Moderator variable}} + \boxed{b_3 \times X_1 \times X_2}$$

Interaction between independent variable and moderator

# Moderation analysis

$$\widehat{Height} = b_0 + b_{Weight} \times Weight + b_{Sex} \times Sex + b_{Weight \times Sex} \times Weight \times Sex$$

1 = female
0 = male

Interaction term

Coefficients

| Model | Unstandardized | Standard Error | Standardized | t | p |
|---|---|---|---|---|---|
| (Intercept) | 115.987 | 0.490 | | 236.724 | < .001 |
| Weight | 0.700 | 0.006 | 1.042 | 121.786 | < .001 |
| Sex | 4.277 | 0.615 | | 6.954 | < .001 |
| Weight × Sex | -0.025 | 0.008 | | -3.036 | 0.002 |

# *Interpretation of the regression coefficient for the interaction term*

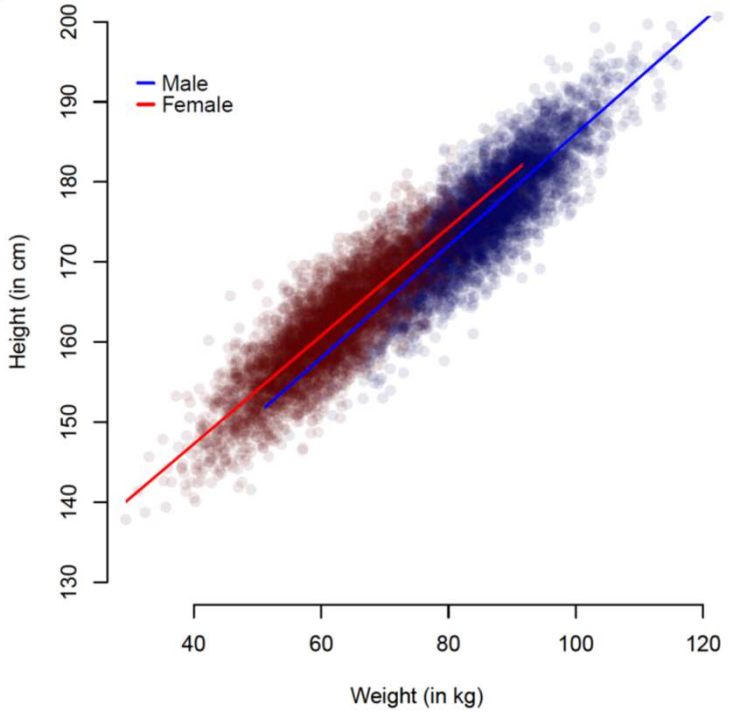
$$\widehat{Height} = 115.987 + 0.700 \times Weight + 4.277 \times Sex - 0.025 \times Weight \times Sex$$

1 = female
0 = male

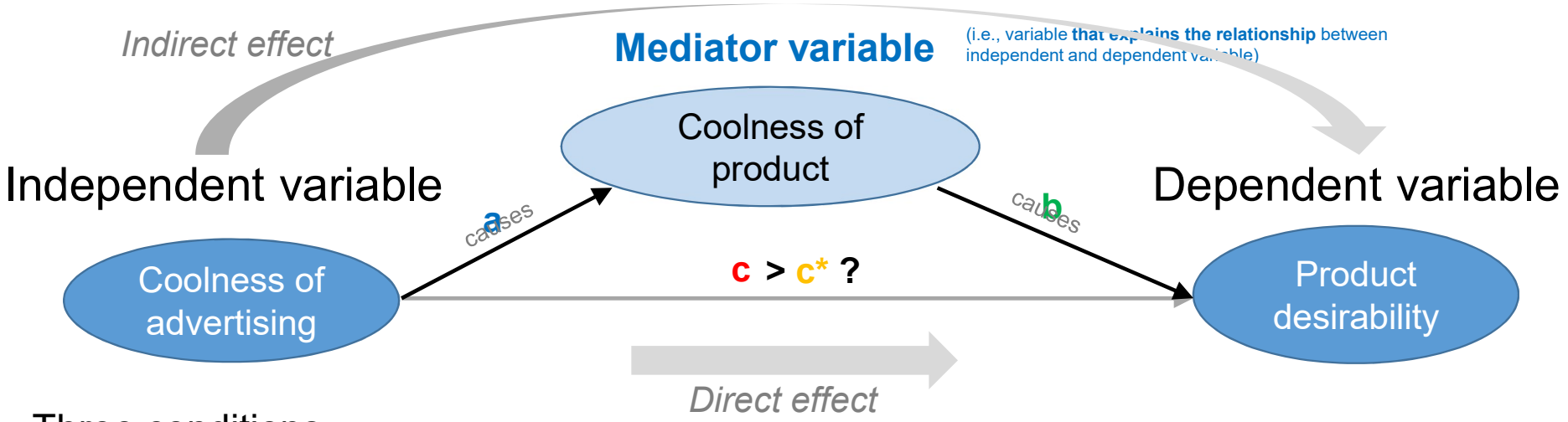
$$\widehat{Height}_{male} = 115.987 + 0.700 \times Weight$$

$$\widehat{Height}_{female} = 120.263 + 0.675 \times Weight$$

0.700−0.675 = 0.025



# Mediation analysis



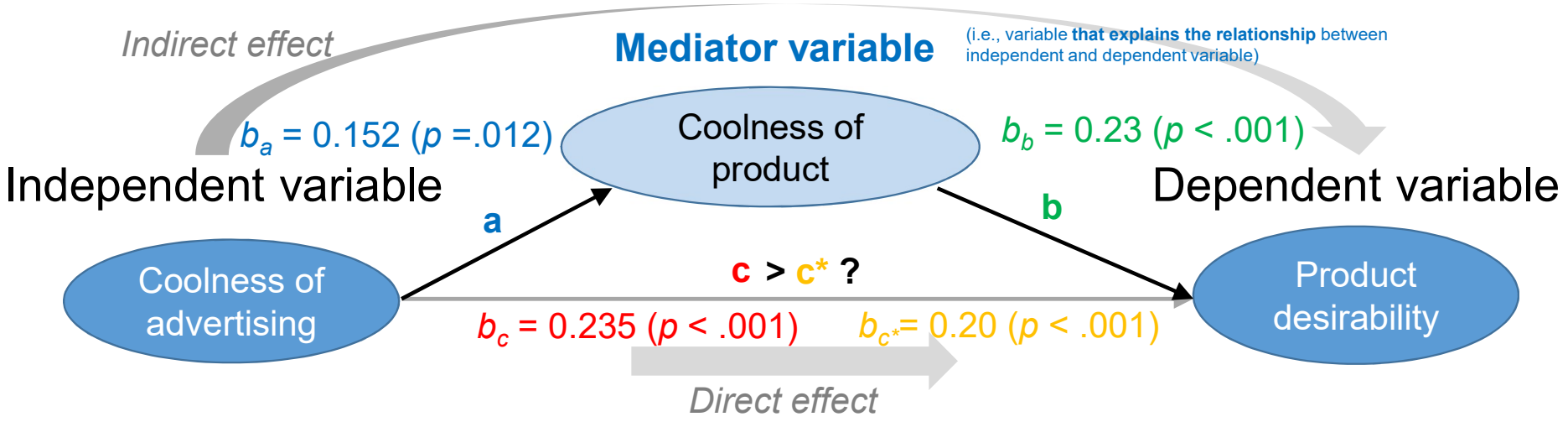

- Three conditions
  1) Is there a significant relationship between independent and dependent variable (**c**)?
  2) Is there a significant relationship between independent variable and mediator (**a**)?
  3) Is there a significant relationship between mediator and dependent variable (**b**), controlling for the independent variable?

→ Is path from IV to DV reduced when mediator and IV are used simultaneously to predict the DV (**c***)?

→ Statistical test: Is the indirect effect (path **a** × **b**) reliably different from zero?

Baron & Kenny (1986)

# Mediation analysis




Three regression models

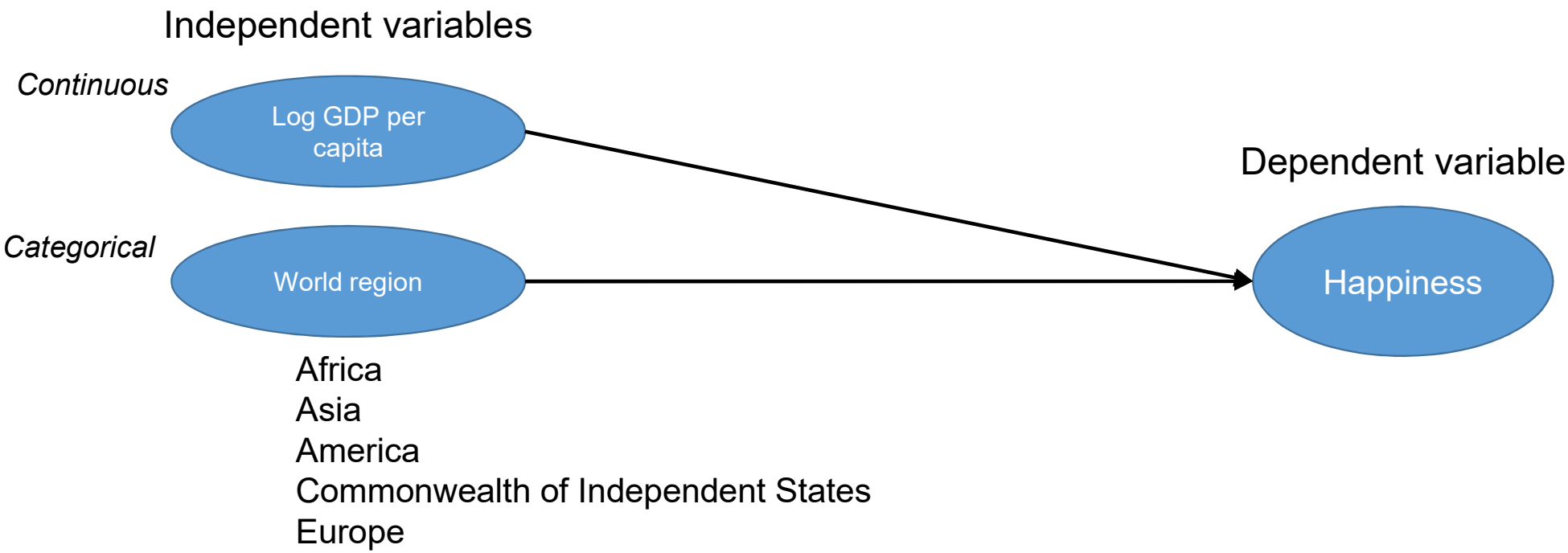
- $\hat{Y} = b_0 + b_c X$
- $\hat{M} = b_0 + b_a X$
- $\hat{Y} = b_0 + b_{c^*} X + b_b M$

Statistical evaluation of the indirect path (a × b)

a × b = 0.035, $CI_{95\%}$ = [0.002, 0.068], $p$ = .038

Baron & Kenny (1986)

# *Multiple regression with a categorical predictor*

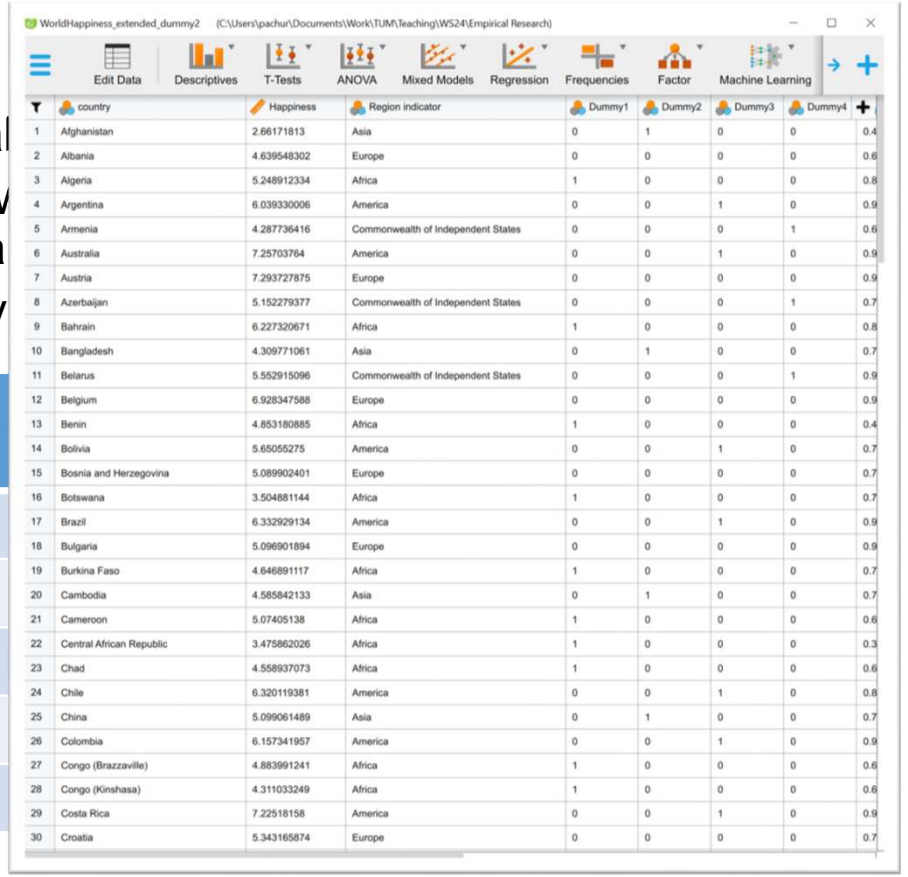


# *Using a categorical predictor: Dummy coding*

Procedure

- If there are *k* categories, create *k*-1 dummy varial
- Choose a reference category (e.g., Europe), to w
  compared. For this category, set all dummy varia
- For each of the other categories, set one dummy

| | Dummy variable 1 | Dummy variable 2 |
|---|---|---|
| Africa | 1 | 0 |
| Asia | 0 | 1 |
| America | 0 | 0 |
| Independent | 0 | 0 |
| Europe | 0 | 0 |




| | country | Happiness | Region indicator | Dummy1 | Dummy2 | Dummy3 | Dummy4 |
|---|---|---|---|---|---|---|---|
| 1 | Afghanistan | 2.66171813 | Asia | 0 | 1 | 0 | 0 |
| 2 | Albania | 4.639548302 | Europe | 0 | 0 | 0 | 0 |
| 3 | Algeria | 5.248912334 | Africa | 1 | 0 | 0 | 0 |
| 4 | Argentina | 6.039330006 | America | 0 | 0 | 1 | 0 |
| 5 | Armenia | 4.287736416 | Commonwealth of Independent States | 0 | 0 | 0 | 1 |
| 6 | Australia | 7.25703764 | America | 0 | 0 | 1 | 0 |
| 7 | Austria | 7.293727875 | Europe | 0 | 0 | 0 | 0 |
| 8 | Azerbaijan | 5.152279377 | Commonwealth of Independent States | 0 | 0 | 0 | 1 |
| 9 | Bahrain | 6.227320671 | Africa | 1 | 0 | 0 | 0 |
| 10 | Bangladesh | 4.309771061 | Asia | 0 | 1 | 0 | 0 |
| 11 | Belarus | 5.552915096 | Commonwealth of Independent States | 0 | 0 | 0 | 1 |
| 12 | Belgium | 6.928347588 | Europe | 0 | 0 | 0 | 0 |
| 13 | Benin | 4.853180885 | Africa | 1 | 0 | 0 | 0 |
| 14 | Bolivia | 5.65055275 | America | 0 | 0 | 1 | 0 |
| 15 | Bosnia and Herzegovina | 5.089902401 | Europe | 0 | 0 | 0 | 0 |
| 16 | Botswana | 3.504881144 | Africa | 1 | 0 | 0 | 0 |
| 17 | Brazil | 6.332929134 | America | 0 | 0 | 1 | 0 |
| 18 | Bulgaria | 5.096901894 | Europe | 0 | 0 | 0 | 0 |
| 19 | Burkina Faso | 4.646891117 | Africa | 1 | 0 | 0 | 0 |
| 20 | Cambodia | 4.585842133 | Asia | 0 | 1 | 0 | 0 |
| 21 | Cameroon | 5.07405138 | Africa | 1 | 0 | 0 | 0 |
| 22 | Central African Republic | 3.475862026 | Africa | 1 | 0 | 0 | 0 |
| 23 | Chad | 4.558937073 | Africa | 1 | 0 | 0 | 0 |
| 24 | Chile | 6.320119381 | America | 0 | 0 | 1 | 0 |
| 25 | China | 5.099061489 | Asia | 0 | 1 | 0 | 0 |
| 26 | Colombia | 6.157341957 | America | 0 | 0 | 1 | 0 |
| 27 | Congo (Brazzaville) | 4.883991241 | Africa | 1 | 0 | 0 | 0 |
| 28 | Congo (Kinshasa) | 4.311033249 | Africa | 1 | 0 | 0 | 0 |
| 29 | Costa Rica | 7.22518158 | America | 0 | 0 | 1 | 0 |
| 30 | Croatia | 5.343165874 | Europe | 0 | 0 | 0 | 0 |

# Multiple regression with a dummy-coded predictor

$$\widehat{Happiness} = b_0 + b_1 Dummy_1 + b_2 Dummy_2 + b_3 Dummy_3 + b_4 Dummy_4$$

| | Dummy Variable 1 | Dummy Variable 2 | Dummy Variable 3 | Dummy Variable 4 |
|---|---|---|---|---|
| Africa | 1 | 0 | 0 | 0 |
| Asia | 0 | 1 | 0 | 0 |
| America | 0 | 0 | 1 | 0 |
| Independent | 0 | 0 | 0 | 1 |
| Europe | 0 | 0 | 0 | 0 |

Average happiness in reference category

Deviations from average in reference category

| Model | | Unstandardized | Standard Error | Standardized[a] | t | p |
|---|---|---|---|---|---|---|
| $H_1$ | (Intercept) | 6.378 | 0.141 | | 45.102 | < .001 |
| | Dummy1 (1) | -1.667 | 0.187 | | -8.935 | < .001 |
| | Dummy2 (1) | -1.308 | 0.239 | | -5.479 | < .001 |
| | Dummy3 (1) | -0.158 | 0.232 | | -0.684 | 0.495 |
| | Dummy4 (1) | -1.074 | 0.286 | | -3.758 | < .001 |

[a] Standardized coefficients can only be computed for continuous predictors.

# *Self-quiz questions*

- Why can the regression coefficient for a predictor in a multiple regression differ from the regression coefficient in a simple regression?
- How can a multiple regression analysis be evaluated statistically (overall model fit, regression coefficients)?
- What is multicollinearity, when is it a problem, and how can one test for it?
- What aspects are relevant for a power analysis for multiple regression?
- How do you test whether a variable acts as a moderator in the relationship between two other variables?
- What are the steps of a mediation analysis?
- What is dummy coding and how it is used to include a categorical predictor in a regression model?

# *Background readings for next week*

Backhaus, K., Erichson, B., Gensler, S., Weiber, R., & Weiber, T. (2021). Logistic regression. In K. Backhaus, B. Erichson, S. Gensler, R. Weiber, & T. Weiber, *Multivariate analysis: An application-oriented introduction* (p. 267-354). Springer.