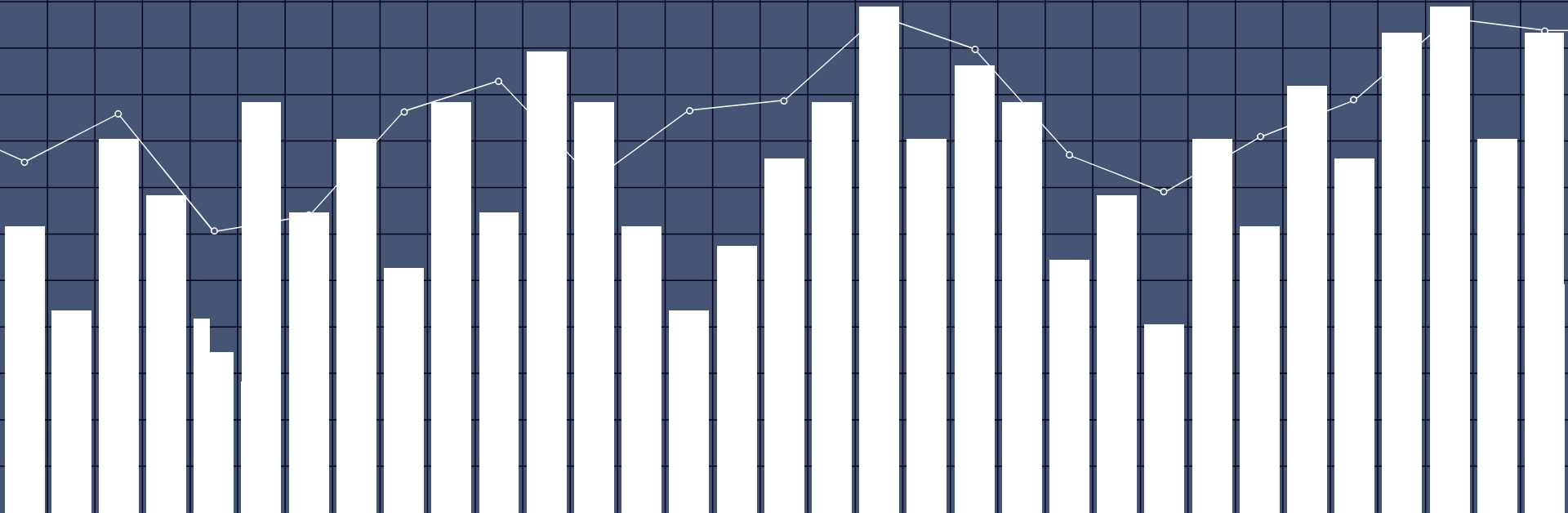
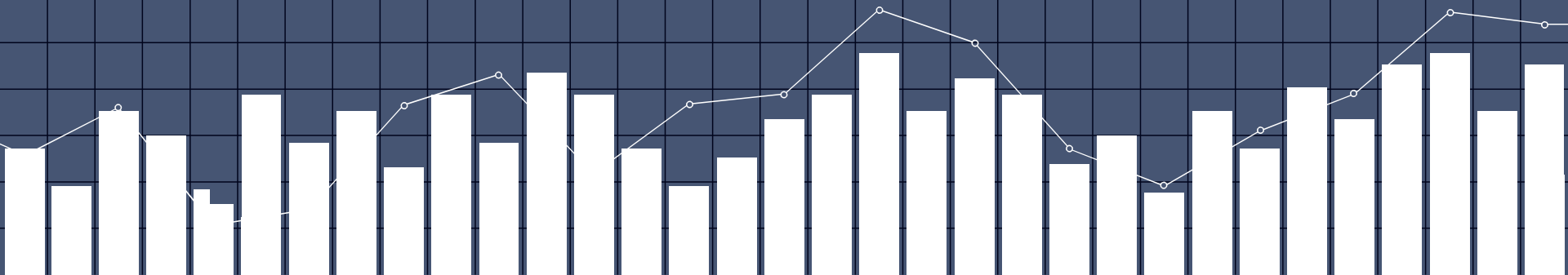
10. Vícenásobná lineární regrese



Harmonogram

- 01. Vícenásobná lineární regrese
- 02. Mediace
- 03. Moderace



Data

JAN. 23, 2017, AT 12:18 PM

Higher Rates Of Hate Crimes Are Tied To Income Inequality

By Maimuna Majumder

Filed under Hate Crimes

Get the data on GitHub

In the 10 days after the 2016 election, nearly 900 hate incidents were reported to the Southern Poverty Law Center, averaging out to 90 per day. By comparison, about 36,000 hate crimes were reported to the FBI from 2010 through 2015 — an average of 16 per day.

RECOMMENDED

Election Update: Some Competitive Races Have Little To No Polling. That's A Problem.

House Update: Keeping An Eye On Democratic 'Reach' Districts

Our Final Forecast In The Senate, House And Gubernatorial Races

Společnost

NENÁVIST NA SÍTÍCH JE TŘEBA TRESTAT. PŘESTUPKY NIC NEŘEŠÍ

S Klárovu Kalibovou, šéfkou organizace In IUSTITIA, o přelomovém rozsudku

Klára Kalibová • Autor: DVTV

Nejčtenější články

1. Babiš se topí a my s ním
   Erik Tabery
2. Zdraví dcery Andreje Babiše už znalec přezkoumal
   Jaroslav Spurný, Ivana Svobodová
3. Jak jsem se mýlil v Nohavicovi
   Petr Třešňák
4. České tajné služby prověřují ukrajinskou přítelkyni Babišova syna
   Ondřej Kundra
5. Policie nepředala žalobcům informace o podezření z Babišova únosu
   Ondřej Kundra, Jaroslav Spurný, Ivana Svobodová

Lineární regrese

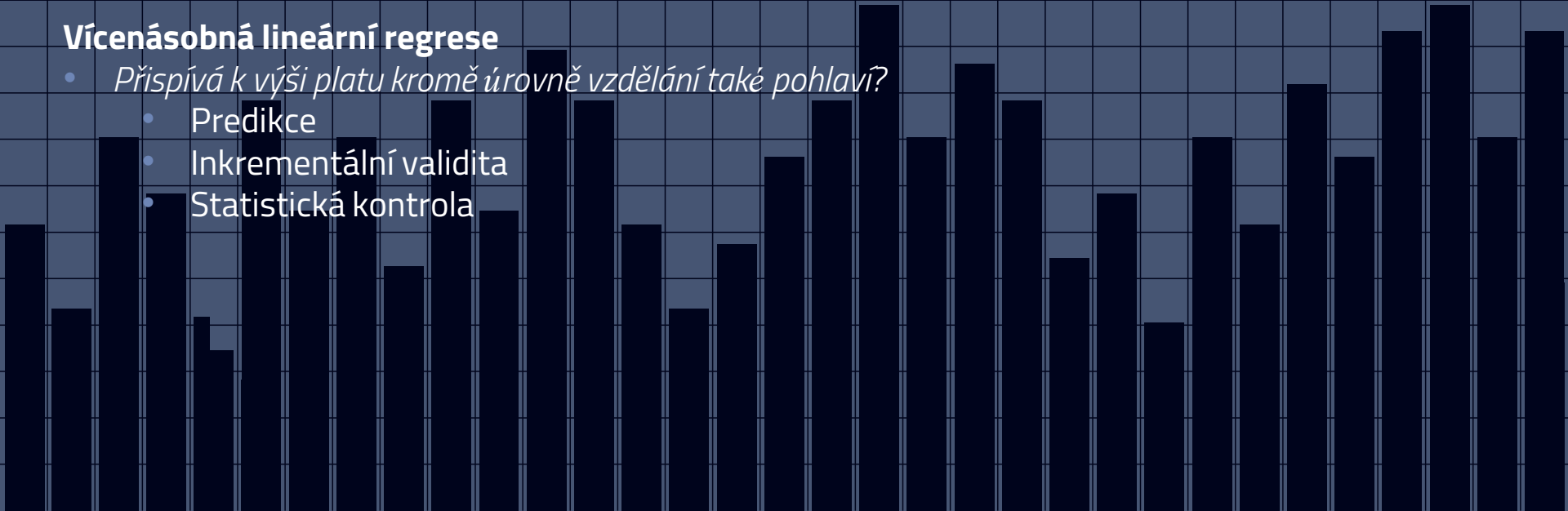
K čemu slouží?

Lineární regrese

- *Nakolik lze z IQ skóru usuzovat o výkonu v matematice?*
  - Predikce

Vícenásobná lineární regrese

- *Přispívá k výši platu kromě úrovně vzdělání také pohlaví?*
  - Predikce
  - Inkrementální validita
  - Statistická kontrola



Lineární regrese

Notace

$Y = Y' + e$

Lineární regrese

$Y' = a + bX$

$Y' = b_0 + b_1X_1$

Vícenásobná lineární regrese

$Y' = a + b_nX_n$

$Y' = b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n + e$

Y = Predikovaná (= závislá; outcome) proměnná
Y' = Náš model
e = Chyba měření
a nebo b_0 = průsečík (= intercept)
b nebo $b_{1\ldots n}$ = směrnice (= slope)
$X_{1\ldots n}$ = Prediktor (= nezávislá proměnná; predictor)

Lineární regrese

Grafické znázornění

$$Y = Y' + e$$
$$Y' = a + bX$$
$$Y' = b_0 + b_1X_1$$

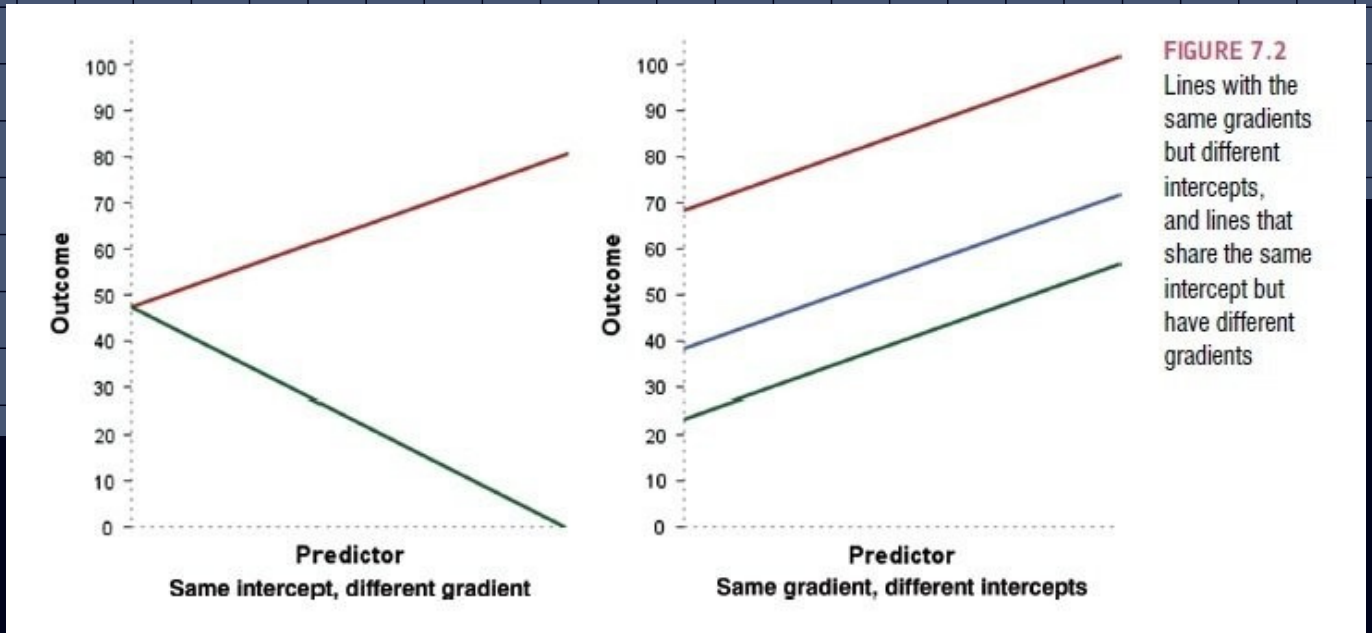


FIGURE 7.2
Lines with the same gradients but different intercepts, and lines that share the same intercept but have different gradients

Lineární regrese

Přímka (model) je proložena daty tak, aby jim co nejlépe odpovídala.

Metoda odhadu nejmenších čtverců (Least Squares Estimation)
Suma (druhých mocnin) vzdáleností modelu od dat je nejmenší možná

$$ss_M = \frac{\sum(m_y - Y')^2}{n-1}$$

$$ss_R = \frac{\sum(Y - Y')^2}{n-1}$$

$$ss_T = \frac{\sum(Y - m_y)^2}{n-1}$$

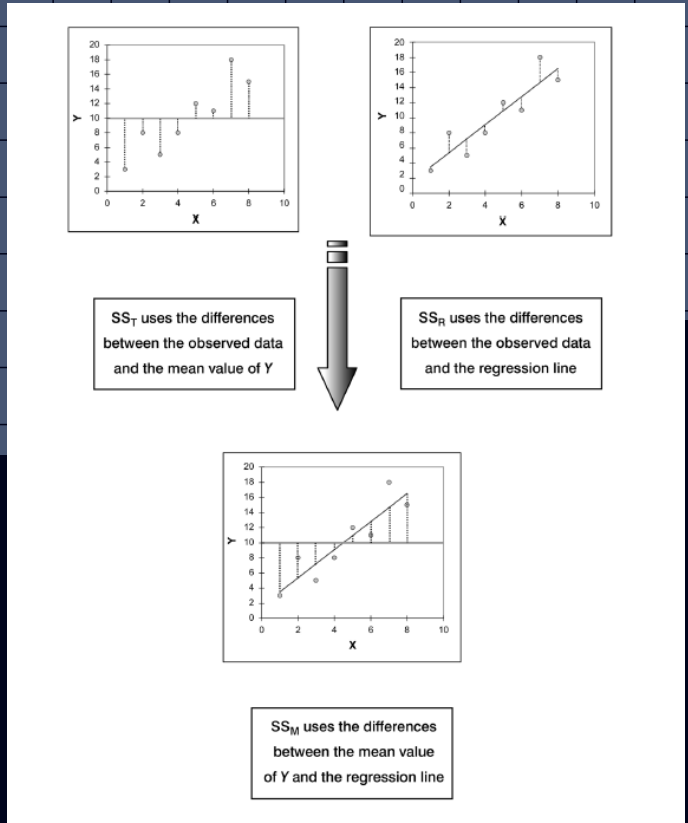
$s_T^2 = s_M^2 + s_R^2$

$R^2 = ss_M^2 / ss_T^2$

SS_M = Rozdíl mezi **nulovým modelem** (průměr Y) a **námi stanoveným modelem** (přímkou)
SS_R = Rozdíl mezi **daty** a **námi stanoveným modelem** (přímkou)
SS_T = Rozdíl mezi **daty** a **nulovým modelem** (průměr Y)
R^2 = Podíl rozptylu závislé (outcome) proměnné **vysvětlené modelem** (= *koeficient determinance*)

Lineární regrese

Metoda nejmenších čtverců graficky



$$ss_M = \frac{\sum(m_y - Y')^2}{n-1}$$

$$ss_R = \frac{\sum(Y - Y')^2}{n-1}$$

$$ss_T = \frac{\sum(Y - m_y)^2}{n-1}$$

Lineární regrese

Příklad modelu

```
ModelHateCrime <- lm(formula = share_voters_voted_trump ~ share_white_poverty +
                     share_non_citizen, data = Hate_Crimes)

# Compute the summary statistics for model
# Generic functions (summary) change their behaviour based on an object's class.
summary(ModelHateCrime)

# Perform an analysis of variance on model
anova(ModelHateCrime)

# Predict based on the fitted function model_erc
predict(ModelHateCrime)
```

Lineární regrese

Koeficienty

b_i

Vyjadřuje nárůst Y' při nárůstu Xi o jednu jednotku v jednotkách Y, při kontrole všech ostatních prediktorů (tj. semiparciální korelace); jedinečný přínos

- K porovnání síly prediktoru v různých skupinách, modelech, vzorcích

β_i ; Beta

Vyjadřuje nárůst Y' při nárůstu X_i o 1; jsou-li X_i i Y standardizovány, při kontrole všech ostatních prediktorů (tj. semiparciální korelace), jedinečný přínos

- K porovnání prediktorů mezi sebou v rámci jednoho modelu
- K porovnání různě operacionalizovaného prediktoru v různých modelech
- Ukazatel velikosti účinku

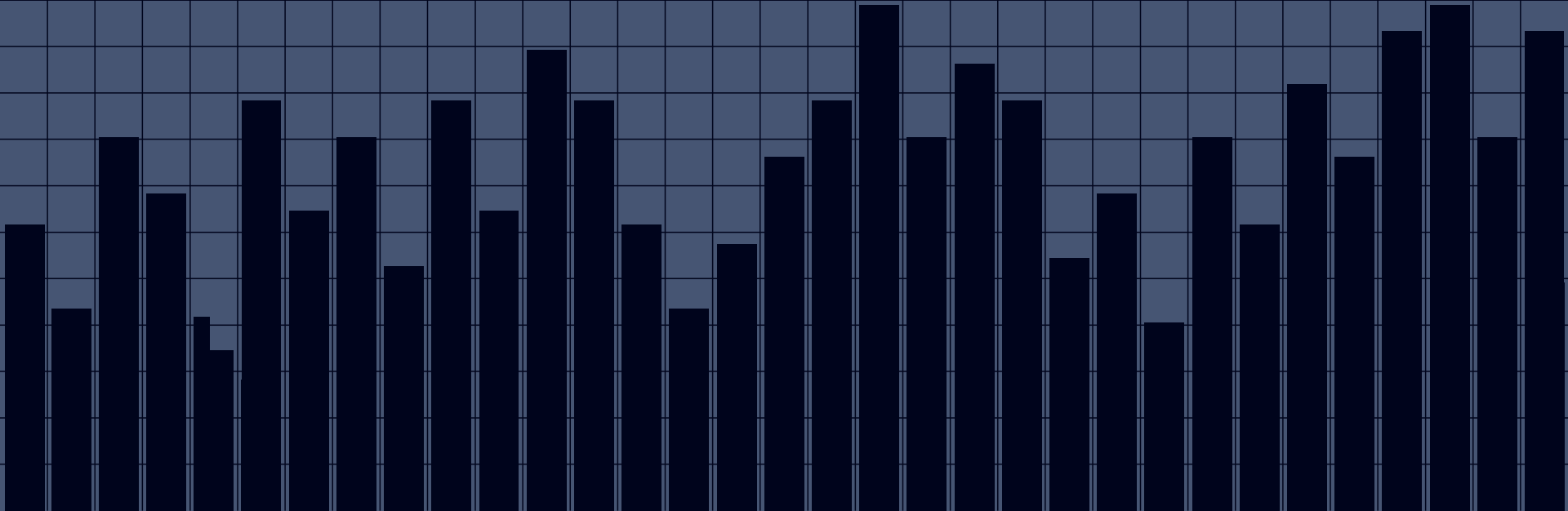
b_0

Po vycentrování (odečtení průměru od všech hodnot X_1) odpovídá průměru Y.

Lineární regrese

Koeficienty

```
install.packages("QuantPsyc")
library(QuantPsyc)
lm.beta(ModelHateCrime)
```



Lineární regrese

Předpoklady použití I.

"*To draw conclusions about a population based on a regression analysis done on a sample, several assumptions must be true.*" (Field, 2009 , s. 220)

Proměnné

1. Povaha proměnných - spojité, kvantitativní a kardinální nebo dummy (jen v případě prediktorů).
2. Nenulová variabilita prediktorů (tj. nejde o konstantu).

Prediktory

3. Absence (dokonalé) multikolinearity - prediktory by spolu neměly vysoce korelovat.
4. Prediktory nekorelují s vnějšími proměnnými – absence třetí (intervenující, vnější) proměnné.

Lineární regrese

Předpoklady použití I. – příklad

Povaha proměnných a nenulová variabilita

```
Hate_Crimes_Select <- Hate_Crimes%>%
  select(c("share_voters_voted_trump", "share_white_poverty", "share_non_citizen"))
lapply(Hate_Crimes_Select, class)
summary(Hate_Crimes_Select)
describe(Hate_Crimes_Select) # library("psych")
```

Multikolinearita

```
# Ověření skrze funkce (např.):
library("car")
vif(ModelHateCrime) # variance inflation factors
sqrt(vif(ModelHateCrime)) > 2 # problem?
rcorr.adjust(Hate_Crimes_Select)
```

Lineární regrese

Předpoklady použití I.

Rezidua

5. Homoskedascita - rozptyl reziduí by měl být konstantní napříč různými úrovněmi prediktoru.
6. Nezávislost reziduí - Reziduální hodnoty kterýchkoliv dvou případů by spolu neměly souviset.
7. Normálně rozložená rezidua - jejich rozložení by mělo být náhodné.

Outcome

8. Nezávislost kterýchkoliv dvou hodnot závislé proměnné (každá hodnota v rámci ní pochází z unikátního zdroje).
9. Linearita - přímka jako vhodný model popisu dat.

Lineární regrese

Homoskedascita a linearita

dle Field, 2009, s. 248

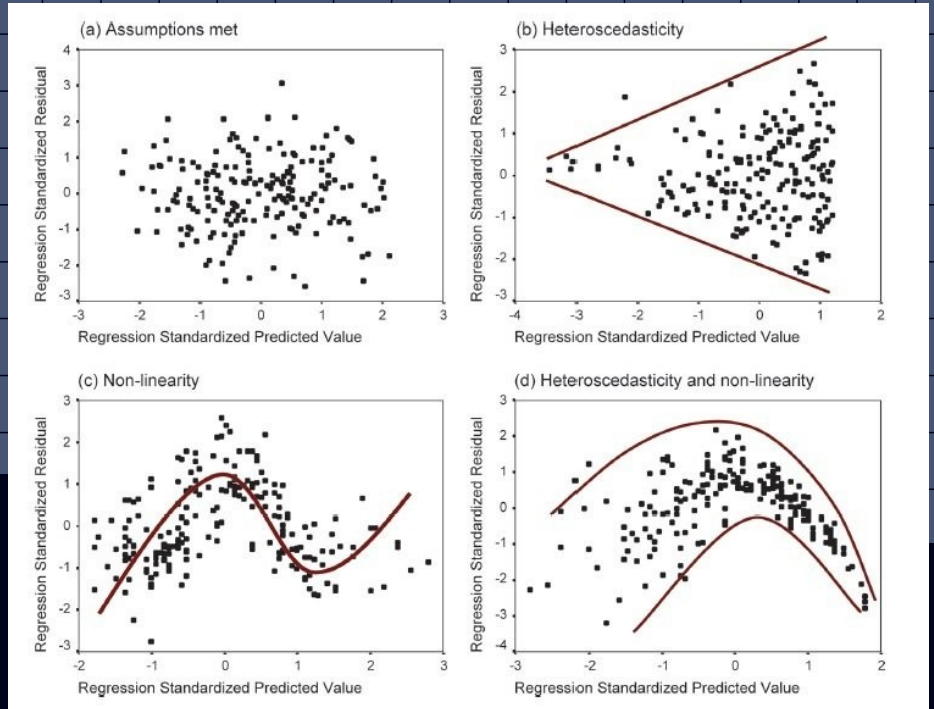


FIGURE 7.19 Plots of *ZRESID against *ZPRED

Lineární regrese

Předpoklady použití II. – příklad

Homoskedascita

```
# Evaluate homoscedasticity
# non-constant error variance test
ncvTest(ModelHateCrime)

# plot studentized residuals vs.
# fitted values
spreadLevelPlot(ModelHateCrime)
```

Nezávislost reziduí

```
# Test for Autocorrelated Errors
durbinWatsonTest(ModelHateCrime)
```

Normálně rozložená rezidua

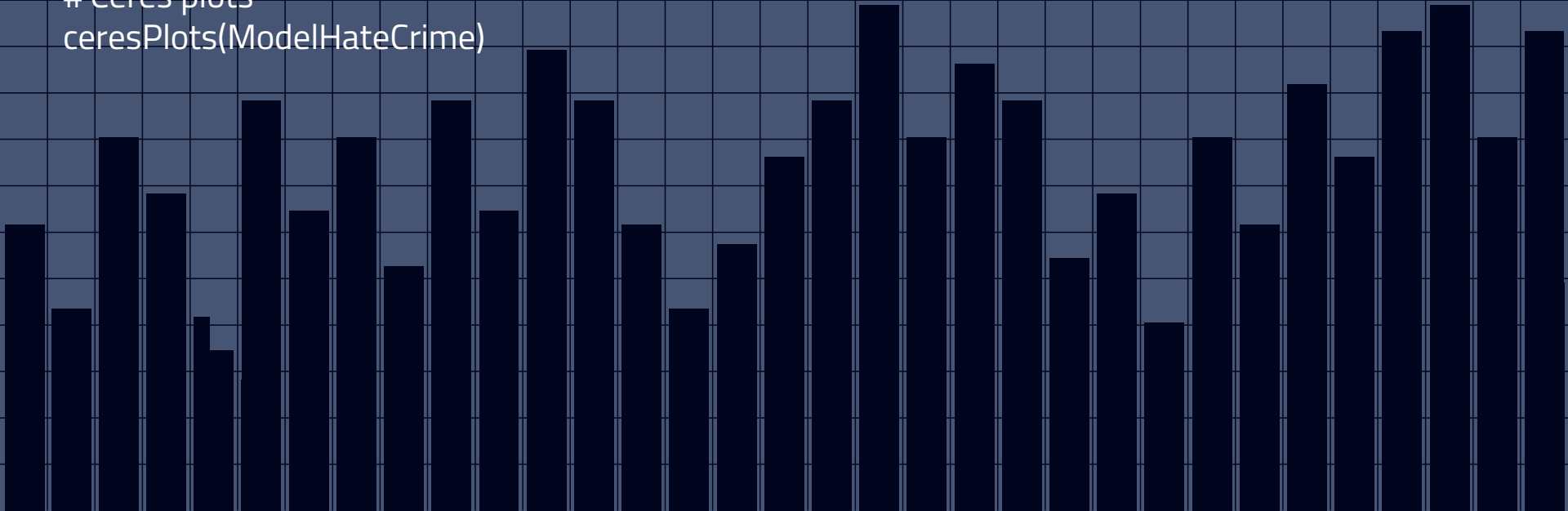
```
# distribution of studentized residuals
library(MASS)
sresid <- studres(ModelHateCrime)
hist(sresid, freq=FALSE,
    main="Distribution of Studentized
    Residuals")
```

Lineární regrese

Předpoklady použití II. – příklad

Linearita

```
# component + residual plot
crPlots(ModelHateCrime)
# Ceres plots
ceresPlots(ModelHateCrime)
```



Lineární regrese

Diagnostika I. – Outliers a Influentials

Nemají některé případy příliš velký vliv na výsledky regrese?

- **Outliery** – mohou zvyšovat i snižovat b
  - **Rezidua** – případy s vysokými rezidui regrese predikuje nejhůř, standardizovaná, ± 3
  - **Vlivné případy** – případy, které nejvíc ovlivňují parametry modelu
    - Co se stane s parametry regrese, když případ odstraníme?
    - *DFBeta* – rozdíl mezi parametrem s a bez, standardizované > 1
    - *DFFit* – rozdíl mezi predikovanou hodnotou a predikovanou hodnotou bez případu (adjustovanou)
    - *Cookova vzdálenost* > 1
    - *Leverage* > 2(k+1)/ n , kde k = počet prediktorů, n= velikost vzorku

Případy s vysokými rezidui či vlivné případy **NEODSTRAŇUJEME**
...leda by šlo o zjevnou chybu v datech či vzorku
...leda by nám šlo výhradně o zpřesnění predikce (nikoli o testy hypotéz**)**

Lineární regrese

Diagnostika II. – Kolinearita

- Když dva prediktory vysvětlují **tutéž část variability** závislé proměnné, jeden z nich je téměř zbytečný
- **Komplikuje porovnávání** síly prediktorů
- **Snižuje stabilitu** odhadu parametrů
- V extrému (*když lze jeden prediktor přesně vypočítat z ostatních*) regresi úplně **znemožňuje**
- "**Rules of Thumb**"
  - Korelace nad 0,9
  - Tolerance (= 1/VIF) cca pod 0,1
  - VIF (= 1/tolerance) cca nad 10)

Lineární regrese

Outliers and Influentials – příklad

Outliers

```
# Bonferonni p-value for most extreme
observations
outlierTest(ModelHateCrime)

#qq plot for studentized resid
qqPlot(ModelHateCrime, main="QQ
Plot")
```

Influentials

```
# Cook's D plot
# identify D values > 4/(n-k-1)
cutoff <- 4/((nrow(Hate_Crimes)-
length(ModelHateCrime$coefficients)-2))
plot(ModelHateCrime, which=4,
cook.levels=cutoff)

# leverage plots
leveragePlots(ModelHateCrime)
```

Lineární regrese

Dummy coding I. – obecně a postup

Dummy proměnné - kategorické proměnné **upravené** tak, aby mohly vstoupit do (vícenásobné) lineární regrese

Postup (dle Field, 2009, s. 254)

1. Count the number of groups you want to recode and subtract 1.
2. Create as many new variables as the value you calculated in step 1. These are your dummy variables.
3. Choose one of your groups as a baseline (i.e. a group against which all other groups should be compared). This should usually be a control group, or, if you don't have a specific hypothesis, it should be the group that represents the majority of people (because it might be interesting to compare other groups against the majority).
4. Having chosen a baseline group, assign that group values of 0 for all of your dummy variables.
5. For your first dummy variable, assign the value 1 to the first group that you want to compare against the baseline group. Assign all other groups 0 for this variable.
6. For the second dummy variable assign the value 1 to the second group that you want to compare against the baseline group. Assign all other groups 0 for this variable.
7. Repeat this until you run out of dummy variables.
8. Place all of your dummy variables into the regression analysis!

Lineární regrese

Dummy coding I. – obecně a postup

Indikátorové kódování *(Indicator coding)*
Referenční kategorie = 0

Efektové kódování *(Effect coding)*
Referenční kategorie = -1

Úroveň vzdělání	Původní hodnota	Indikátorové kódování		Efektové kódování	
		Vysokoškolské	*Středoškolské*	*Vysokoškolské*	*Středoškolské*
Vysokoškolské	1	1	0	1	0
Středoškolské	2	0	1	0	1
Základní	3	0	0	-1	-1

Lineární regrese

Dummy coding II. – kódování

$$Y = b_0 + b_{A1}X_{A1} + b_{A2}X_{A2} + \ldots + b_mX_m + e$$

- Po dosazení do regresní rovnice predikujeme případu průměr jeho skupiny (pokud nejsou žádné další prediktory).
- **Indikátorové kódování**
  - b_{Ai} udává rozdíl průměrných hodnot Y mezi indikovanou skupinou a referenční skupinou; signifikanci b Ai referenční skupinou; signifikance b_{Ai} znamená signifikanci rozdílu
  - b_{Ai} udává o kolik nám členství ve skupině zvyšuje/snižuje predikovanou hodnotu oproti referenční skupině
  - b_0 udává (při absenci jiných prediktorů) průměr Y v referenční skupině
- **Efektové kódování**
  - b_{Ai} udává rozdíl průměrných hodnot Y mezi indikovanou skupinou a celkovým průměrem
  - b_0 udává (při absenci jiných prediktorů) celkový průměr

Lineární regrese

Dummy coding III. – příklad

```
Hate_Crimes$UrbanRural = cut(Hate_Crimes$share_population_in_metro_areas,breaks=c(-Inf, 0.5, Inf), labels=c("Rural","Urban"))
```

The factor function

```
class(Hate_Crimes$UrbanRural)
contrasts(Hate_Crimes$UrbanRural)
summary(lm(share_voters_voted_trump ~ share_white_poverty +
        share_non_citizen + UrbanRural, data = Hate_Crimes))
```

Lineární regrese

Mediace

A mediation analysis is typically conducted to better understand an observed effect of an IV on a DV or a correlation between X and Y

- Why, and how, does X influence/correlates with Y?

If X and Y are correlated BECAUSE of the mediator M, then (X -> M -> Y)

$$Y = B_0 + B_1M + e$$

&

$$M = B_0 + B_1X + e$$

&

$$Y = B_0 + B_1M + B_2X + e$$

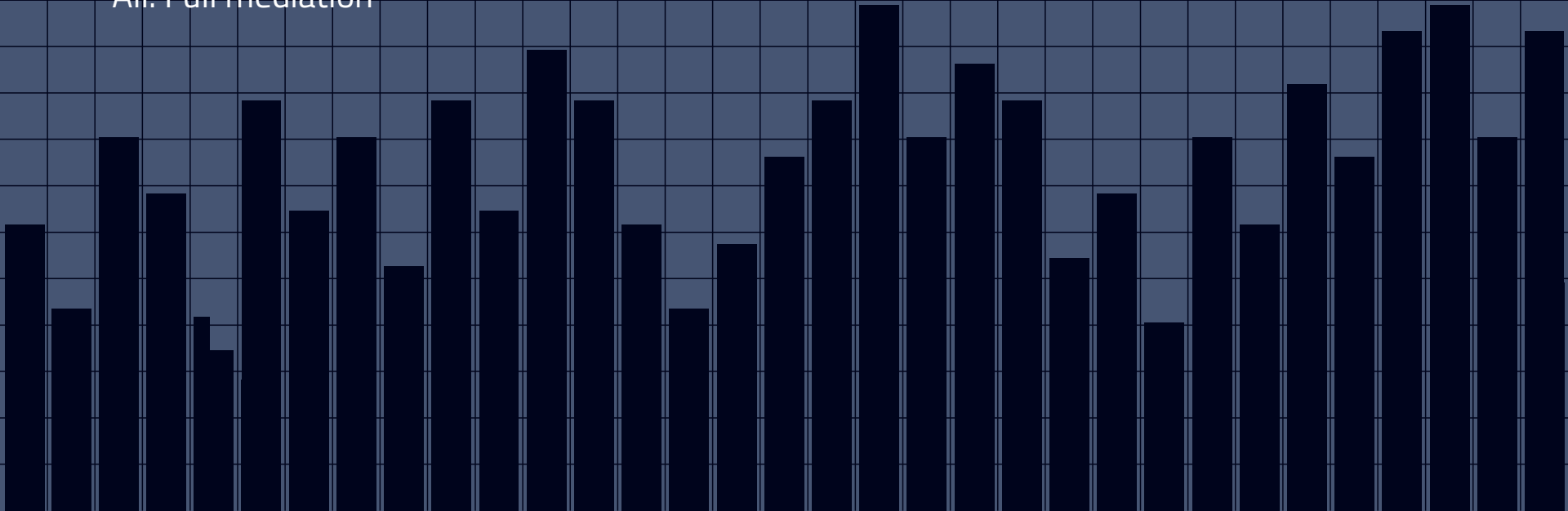
What will happen to the predictive value of X?
In other words, will B_2 be significant?

Lineární regrese

Mediace

A mediator variable (M) accounts for some or all of the relationship between X and Y:

- Some: Partial mediation
- All: Full mediation



Lineární regrese

Mediace – příklad

```
# Run the three regression models
model_yx <- lm(Hate_Crimes$share_voters_voted_trump ~
Hate_Crimes$median_household_income)
model_mx <- lm(Hate_Crimes$share_population_in_metro_areas ~
Hate_Crimes$median_household_income)
model_yxm <- lm(Hate_Crimes$share_voters_voted_trump ~
Hate_Crimes$median_household_income + Hate_Crimes$share_population_in_metro_areas)

# Make a summary of the three models
summary(model_yx)
summary(model_mx)
summary(model_yxm)
```

Lineární regrese

Mediace – příklad – Sobelův test

```
library("multilevel") # install.packages("multilevel")

# Compare the previous results to the output of the sobel function
model_all <- sobel(Hate_Crimes$share_population_in_metro_areas,
Hate_Crimes$median_household_income, Hate_Crimes$share_voters_voted_trump)

# Print out model_all
model_all
```

Lineární regrese

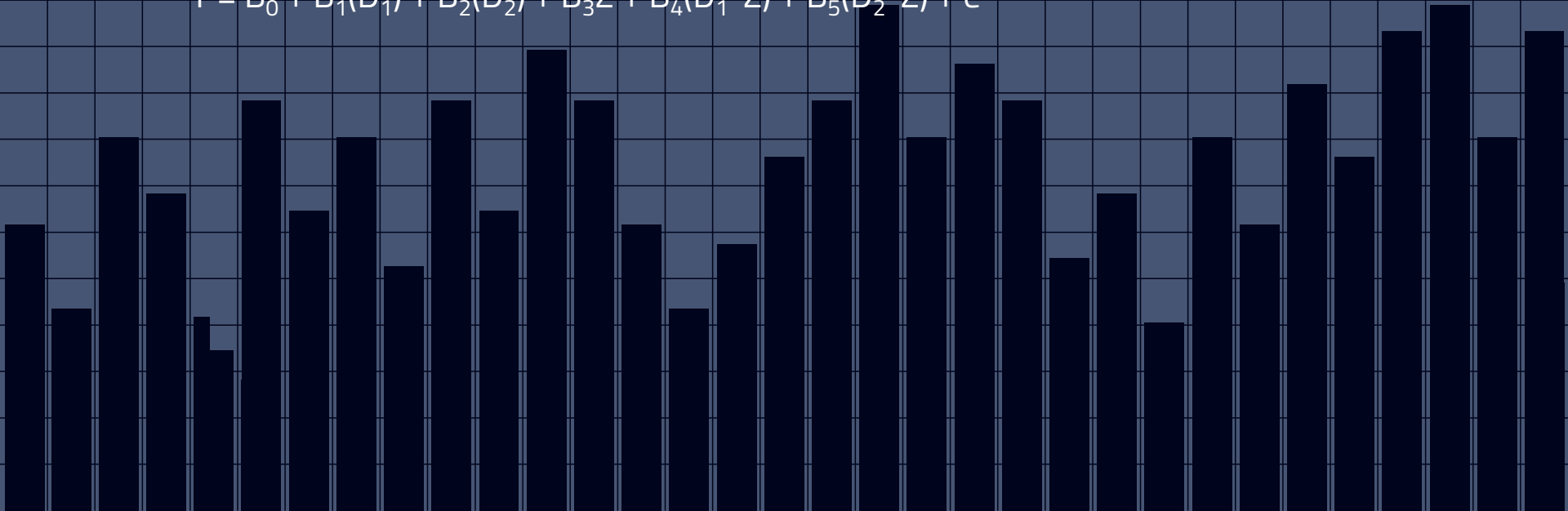
Moderace – představení

- Experimentální design
  - Manipulace s nezávislou proměnnou (X) vede ke změně v závislé proměnné (Y)
  - Moderátor (Z) zavádíme z toho kvůli předpokladu, že vliv (účinek) X na Y **NENÍ** konzistentní napříč rozložením (různými úrovněmi) Z
- Korelační design
  - Předpokládáme souvislost mezi proměnnými X a Y
  - Moderátor (Z) zavádíme kvůli předpokladu, že korelace mezi X a Y **NENÍ** konzistentní napříč rozložením (různými úrovněmi) Z

Lineární regrese

Moderace – představení

- Pokud jsou oboje X a Z spojité (resp. intervalové úrovně měření)
  - $Y = B_0 + B_1X + B_2Z + B_3(X*Z) + e$
- Pokud je X kategorická a Z spojitá (3 úrovně X)
  - $Y = B_0 + B_1(D_1) + B_2(D_2) + B_3Z + B_4(D_1*Z) + B_5(D_2*Z) + e$



Lineární regrese

Moderace – příklad

```
mod <- Hate_Crimes %>%
                select(share_voters_voted_trump, median_household_income, UrbanRural)

# Summary statistics
describeBy(mod, mod$UrbanRural)
# Create a boxplot of the data
boxplot(formula = mod$share_voters_voted_trump ~ mod$UrbanRural, main = "Boxplot", xlab = "Group
UrbanRural", ylab = "share_voters_voted_trump")

# Create subsets of the two groups
# Make the subset for the group UrbanRural = "Rural"
mod_Rural <- subset(mod, mod$UrbanRural == "Rural")
# Make the subset for the group UrbanRural = "Urban"
mod_Urban <- subset(mod, mod$UrbanRural == "Urban")
# Calculate the correlations
cor(mod_Rural$share_voters_voted_trump, mod_Rural$median_household_income)
cor(mod_Urban$share_voters_voted_trump, mod_Urban$median_household_income)
```

Lineární regrese

Moderace – příklad

```
# Model without moderation (tests for "first-order effects")
model_1 <- lm(mod$share_voters_voted_trump ~ mod$median_household_income + mod$UrbanRural)
# Make a summary of model_1
summary(model_1)
# Create new predictor variables
moderator <- mod$median_household_income * as.numeric(mod$UrbanRural)

# Model with moderation
model_2 <- lm(mod$share_voters_voted_trump ~ mod$median_household_income + mod$UrbanRural +
moderator)
# Make a summary of model_2
summary(model_2)

# Compare model_1 and model_2
anova(model_1, model_2)
```

Lineární regrese

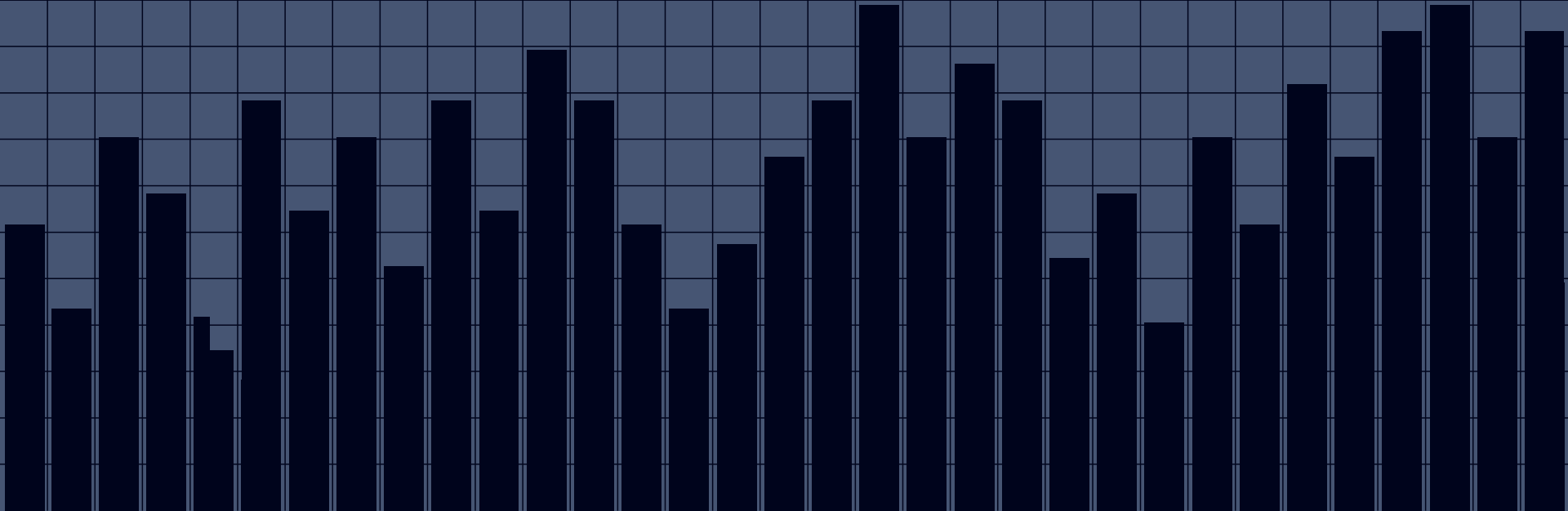
Moderace – příklad

```
# Choose colors to represent the points by group
color <- c("red","green","blue")

# Illustration of the first-order effects of working memory on share_voters_voted_trump
ggplot(mod, aes(x = median_household_income, y = share_voters_voted_trump)) +
geom_smooth(method = "lm", color = "black") +
  geom_point(aes(color = UrbanRural))

# Illustration of the moderation effect of working memory on share_voters_voted_trump
ggplot(mod, aes(x = median_household_income, y = share_voters_voted_trump)) +
  geom_smooth(aes(group = UrbanRural), method = "lm", se = T, color = "black", fullrange = T) +
  geom_point(aes(color = UrbanRural))
```

Lineární regrese

Moderace a mediace

A moderator has influence over other effects or relationships, whereas the mediator explains a relationship.



Zdroje

Field, A. (2009). Discovering statistics using SPSS, 3th Ed. Los Angeles: Sage.

Fox, J. (2016). Applied Regression Analysis and Generalized Linear Models, 3th Ed. Los Angeles: Sage.

Robotková, A., & Ježek, S. (2012). Vícenásobná lineární regrese. Prezentace ke kurzu PSY252.