Okruh 3 - Základy ekonometrie

Hana Drdlová,
Jakub Chalmovianský,
Jakub Moučka

17. prosince 2020

Kapitola 3

Základy ekonometrie v Matlabu

V této kapitole se blíže seznámíme se základy ekonometrie v rámci softwaru Matlab. Podíváme se na možnosti odhadu metodou nejmenších čtverců (OLS), ověření základních předpokladů metody OLS, problém multikolinearity, testování parametrů a modelové specifikace. Prostudujeme možnosti predikce a robustních odhadů, logistické regrese a další možnosti regrese. V závěru jednotlivých částí kapitoly jsou přidány funkce rozšiřující dané téma. Kapitola je včetně řešených příkladů doplněna o neřešené příklady, které procvičí získané vědomosti z oblasti ekonometrie z pohledu Matlabu, pro kontrolu jsou uvedeny výsledky.

Nejdříve budeme pracovat s datovým souborem *fev.csv*, který obsahuje 654 pozorování. Soubor je tvořen 5 proměnnými, *AGE* (věk v letech), *FEV* (maximální vydechnutý objem vzduchu z plic v litrech), *HEIGHT* (výška měřená v palcích), *GENDER* (pohlaví 0 = žena, 1 = muž) a *SMOKE* (kuřák 0 = ne, 1 = ano). Načteme datový soubor pomocí *readmatrix*, kde si jednotlivé sloupce uložíme do proměnných, se kterými budeme dále pracovat.

```
fev_data = readmatrix("fev.csv");
AGE = fev_data(:,2);
FEV = fev_data(:,3);
HEIGHT = fev_data(:,4);
GENDER = fev_data(:,5);
SMOKE = fev_data(:,6);
```

3.1 Odhad LRM metodou nejmenších čtverců

Začneme s odhadem lineárního regresního modelu (LRM) s jednou vysvětlující proměnnou metodou OLS, teoretické vysvětlení použitých pojmů naleznete v části 2.2 Základů ekonometrie (Němec, 2019). Nejdříve sestavíme jednoduchý model, ve kterém prozkoumáme vliv pohlaví na maximální vydechnutý objem vzduchu z plic *FEV*, jinými slovy chceme zjistit, zda existuje rozdíl mezi muži a ženami.

$$FEV_i = \beta_0 + \beta_1 \cdot GENDER_i + \varepsilon_i$$

```
>> OLSregreseS = fitlm(GENDER,FEV,'VarNames',{'Gender',
'FEV'});
```

V možnostech výpisu, který popisuje charakteristiky modelu, je v Matlabu celá řada informací, proto se zmíníme pouze o některých. Základní informace o modelu získáme příkazem *_.disp* či *_.compact*. Na místo podtržítka zadáváme název modelu.

```
>> OLSregreseS.compact;
```

Ekvivalentně:

```
>> compact(OLSregreseS);
```

Linear regression model:
FEV 1 + Gender
Estimated Coefficients:

	Estimate	**SE**	**tStat**	**pValue**
(Intercept)	2.4512	0.047591	51.505	5.591e-232
Gender	0.36128	0.066396	5.4412	7.4958e-08

Number of observations: 654, Error degrees of freedom: 652
Root Mean Squared Error: 0.849
R-squared: 0.0434, Adjusted R-Squared: 0.042
F-statistic vs. constant model: 29.6, p-value = 7.5e-08

Tabulka 3.1: Odhad modelu s regresorem pohlaví

Dle p-hodnot u proměnné a úrovňové konstanty v tabulce 3.1 lze říct, že jsou obě statisticky významné. Pro případ bližšího zkoumání nahlédněte do části 2.2.4 Základů ekonometrie (Němec, 2019). Jaká je interpretace bodových odhadů? Pokud uvážíme ženu, která je kódovaná na 0, odhad vydechnutého objemu vzduchu z plic bude v průměru $2,4512$ l. V případě muže bude odhad vydechnutého objemu vzduchu z plic v průměru $2,81248$ l. V případě potřeby

nahlédněte do části 2.2.2 Základů ekonometrie (Němec, 2019). Počet stupňů volnosti (652) získáme jako rozdíl počtu pozorování a odhadovaných parametrů. Koeficient determinace dostáváme velmi nízký ($R^2 = 4,34\,\%$). F-test, jehož nulová hypotéza hovoří o sdružené hypotéze statistické nevýznamnosti všech parametrů, říká, že alespoň jeden regresor (vysvětlující proměnná) je statisticky významný. Jinými slovy nízká p-hodnota hovoří o statistické významnosti modelu jako celku. Více podrobností lze najít v části 2.2.5 Základů ekonometrie (Němec, 2019).

Výpis koeficientů modelu obdržíme pomocí příkazu _*.Coefficients*. Pro kovarianční matici lze použít příkaz _*.CoefficientCovariance*, pro logaritmus věrohodnosti _*.LogLikelihood*, střední kvadratickou chybu (MSE) _*.MSE* a její odmocninu získáme prostřednictvím příkazu _*.RMSE*. Hodnoty koeficientů, kovarianční matici a MSE lze zároveň získat pomocí funkce *lscov*, výstup nalezneme v tabulce 3.2.

```
>> [LSCOV_X,LSCOV_STDX,LSCOV_MSE,S] = lscov([ones(N,1),
GENDER],FEV)
```

LSCOV_X	LSCOV_STDX	LSCOV_MSE	S	
2.4512	0.0476	0.7202	0.0023	-0.0023
0.3613	0.0664		-0.0023	0.0044

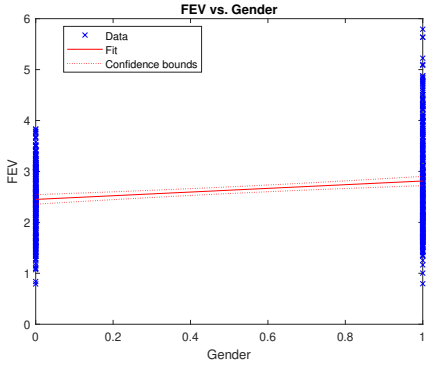
Tabulka 3.2: Odhady koeficientů a směr. chyb, MSE a kovarianční matice

Další detaily použitého modelu získáme následujícími příkazy - rovnici modelu příkazem _*.Formula*, koeficient determinace _*.Rsquared*, součet čtverců chyb (SSE) _*.SSE*, součet čtverců regrese (SSR) _*.SSR* a celkový součet čtverců (SST) _*.SST*. Vysvětlení zmíněných termínů naleznete v části 2.2.3 Základů ekonometrie (Němec, 2019). Počet stupňů volnosti dané specifikace modelu získáme příkazem _*.DFE*.

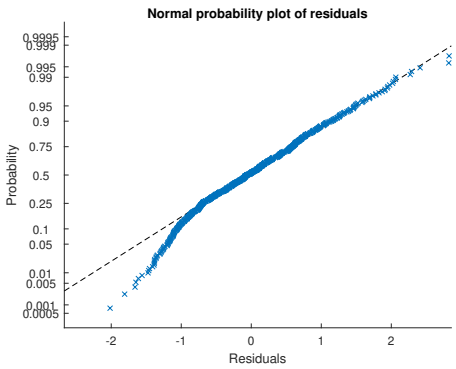
```
>> OLSregreseS.Rsquared
ans =
    struct with fields:
Ordinary: 0.0434
Adjusted: 0.0420
```

Modelem proložená data získáme příkazem _*.Fitted* a rezidua _*.Residuals*. Grafické znázornění provedené regrese získáme funkcí _*.plot* a normální pravděpodobnostní graf reziduí pomocí funkce _*.plotResiduals* s volbou *probability*, což je znázorněno na obrázku 3.1.

```
>> OLSregreseS.plot
```




```
>> OLSregreseS.plotResiduals
('probability')
```



Obrázek 3.1: Grafické znázornění regrese

Výpis informačních kritérií, pro nás zejména Akaikeho a Schwarzovo informační kritérium, lze získat funkcí *_.ModelCriterion*, nebo také pomocí funkce *aicbic*. Funkce vrací hodnotu Akaikeho informačního kritéria po zadání logaritmu věrohodnosti a počtu odhadovaných parametrů. Pro rozšíření výstupního vektoru o Schwarzovo informační kritérium musíme přidat informaci o velikosti vzorku *numObs* (normalizovanou hodnotu IC získáme pomocí volby *'Normalize', true*).

```
>> OLSregreseS.ModelCriterion;
ans =
    struct with fields:
AIC: 1.6433e+03
AICc: 1.6434e+03
BIC: 1.6523e+03
CAIC: 1.6543e+03
```

```
>> [numObs,~] = size(fev_data);
>> [aic,bic] = aicbic
(OLSregreseS.LogLikelihood,
2,numObs);
aic =
    1.6433e+03
bic =
    1.6523e+03
```

Pokud odhadujeme modelovou specifikaci a nepotřebujeme příliš detailní informace o provedeném odhadu, může se nám v takovém případě hodit odhad modelu pomocí funkce *regress*. Získáme koeficienty odhadu B, intervaly spolehlivosti CI, hodnoty reziduí R, intervaly spolehlivosti $RINT$, které slouží při hledání odlehlých hodnot a testové statistiky $STATS$.

```
>> [B,BINT,R,RINT,STATS] = regress(FEV,[ones(N,1),
GENDER],0.05);
```

3.2 Ověřování základních předpokladů OLS

Podívejme se nyní na ověřování základních předpokladů metody OLS, kterými jsou:

- $E(\varepsilon_i) = 0$
  Nulovou střední hodnotu náhodných složek zajistíme přítomností úrovňové konstanty β_0 v modelu.

- $Var(\varepsilon_i) = \sigma^2$
  Zde se ptáme, zda náhodné složky mají konstantní rozptyl, čili jestli není porušen předpoklad homoskedasticity. Této problematice se věnujeme v samostatné části kapitoly 3.6.

- $Cov(\varepsilon_i, \varepsilon_j) = 0$, pro $i \neq j$
  Ověříme, zda nejsou náhodné složky vzájemně korelovány.

- Náhodná složka ε_i má normální rozdělení $N(0, \sigma^2)$.

- Vysvětlující proměnné jsou nenáhodné veličiny.

- Počet vysvětlujících proměnných je menší než počet pozorování.

O nekorelovanosti náhodných složek lze rozhodnout pomocí Durbinova-Watsonova testu, jehož nulová hypotéza zní: H_0: V datovém vzorku není přítomna autokorelace 1. řádu. K tomu využijeme příkaz *_.dwtest* či funkci *dwtest*. Díky velmi nízké p-hodnotě provedeného testu můžeme říct, že rezidua datového vzorku jsou korelovaná.

```
>> OLSregreseS.dwtest;
```

Ekvivalentně:

```
>> dwtest(OLSregreseS);
ans =
    1.3087e-41
```

V předešlé kapitole v části 2.1. jsme se věnovali testování normálního rozdělení. Proto si připomeňme Jarqueův-Beryho test a podívejme se, zda náhodné složky pocházejí z normálního rozdělení. Výpis reziduí pro Jarqueův-Beryho test musíme provést pomocí *OLSregreseS.Residuals.Raw*, neboť vstupem musí být vektor. Hodnota $h = 1$ i nízká p-hodnota hovoří o zamítnutí nulové hypotézy. Díky velmi rozsáhlému datovému souboru lze využít platnost centrálně limitní věty *(CLV)* a pokračovat dál v analýze, více informací lze najít v části B.2 Základů ekonometrie (Němec, 2019).

```
>> [h,p,~,~] = jbtest(OLSregreseS.Residuals.Raw)
h =
    1
p =
    1.0000e-03
```

Doplnění

Část 3.2 doplníme o další funkce, které se mohou při ověřování předpokladů hodit.

- **h = adtest(X)**
  [h,p,adstat,cv] = adtest(X) - Jedná se funkci, která provede Andersonův-Darlingův test, pomocí něhož je testována nulová hypotéza, že data X pocházejí z normálního rozdělení. Při $h = 1$ je zamítnuta H_0. Po zadefinování střední hodnoty a směrodatné odchylky lze testovat normální rozdělení s daným rozdělením pravděpodobnosti. Volbou *Distribuce* lze zvolit také jiné než normální rozdělení, hladinu významnosti volíme pomocí parametru *Alpha*. Kromě h a p hodnoty také lze získat hodnotu statistiky *adstat* a kritickou hodnotu testu *cv*.

- **[h,p,ci,zval] = ztest(X,M,sigma)** - Funkce provede Z-test o normalitě vstupního vektoru dat X s pomocí testovaného průměru na základě nulové hypotézy M a populační směrodatné odchylky *sigma*. Výstupem funkce jsou: rozhodnutí o (ne)zamítnutí nulové hypotézy h, p-hodnota testu p, intervaly spolehlivosti skutečného populačního průměru *ci* a testovací statistika *zval*. Funkce má volitelné vstupní argumenty, které specifikují hladinu významnosti *Alpha*, dimenzi *Dim* a alternativní hypotézu *Tail*, např. pravostranná.

- **[h,p,kstat,critval] = lillietest(X)** - Funkce provádí Lillieforsův test normality dat a vrací rozhodnutí o (ne)zamítnutí nulové hypotézy h, p-hodnotu p, testovací statistiku *kstat* a kritické hodnoty *critval*. Vstupním argumentem je vektor dat X. Volitelný argument *Alpha* slouží k nastavení hladiny významnosti., *Distribution* dovoluje určit rozdělení, defaultně je normální, ale je možné zvolit exponenciální nebo rozdělení extrémních hodnot. Volitelný argument *MCTol* slouží k nastavení maximální Monte Carlo standardní chyby.

3.3 Multikolinearita

Krátce se v této části kapitoly dotkneme problematiky multikolinearity, která nastává v rámci metody OLS, pokud jsou některé či všechny vysvětlující proměnné vzájemně silně korelovány. V takové situaci má model problém rozlišit mezi vlivy vysvětlujících proměnných na vysvětlovanou proměnnou. Tento problém lze poznat podle nízkých hodnot testových statistik (s vysokými p-hodnotami), přičemž koeficient determinace modelu je vysoký. Více informací lze nalézt v části 2.3.5 Základů ekonometrie (Němec, 2019). Pro ověření multikolinearity v datech, si vytvoříme v řešeném příkladu rozšířený model.

Cvičení 3.1

Nejdříve si zopakujeme syntaxi funkce *fitlm* a vytvoříme rozšířený model, pro který bude dříve vytvořený model vnořeným. Úkolem bude vypsat a interpretovat odhady koeficientů, vypsat rovněž pro srovnání obou modelů hodnotu adjustovaného R^2.

$$FEV_i = \beta_0 + \beta_1 \cdot GENDER_i + \beta_2 \cdot HEIGHT_i + \beta_3 \cdot AGE_i + \varepsilon_i$$

Řešení:

```
Y = FEV;
X = [GENDER,HEIGHT,AGE];
OLSregreseE = fitlm(X,Y,'VarNames',{'Gender','Height',
    'Age','FEV'});
OLSregreseE.Coefficients
```

	Estimate	**SE**	**tStat**	**pValue**
(Intercept)	-4.4486	0.22297	-19.952	1.8802e-69
Gender	0.16111	0.033125	4.8638	1.4463e-06
Height	0.10456	0.0047557	21.986	1.5941e-80
Age	0.061364	0.0090694	6.766	2.9567e-11

Tabulka 3.3: Výpis odhadů koeficientů rozšířeného modelu

```
>> OLSregreseE.Rsquared
ans =
    struct with fields:
Ordinary: 0.7746
Adjusted: 0.7736
```

Interpretujme odhady koeficientů z tabulky 3.3. Pokud uvážíme ženu, nulové výšky a věku, která nekouří, odhad vydechnutého objemu vzduchu z plic bude $-4,4486$ l. Úrovňová konstanta v tomto případě nemá příliš rozumnou interpretaci, to nám ale nevadí a úrovňovou konstantu v modelu ponecháme. Pokud budeme mít muže a ženu stejné výšky, věku a vztahu ke kouření, vydechnutý objem vzduchu z plic muže (kódovaného na 1) bude v průměru o $0,16111$ l vyšší. Podobně bychom interpretovali ostatní koeficienty. Podívejme se také na sloupec s p-hodnotami, všechny regresory bychom na 5% hladině významnosti mohli považovat za statisticky významné. Koeficient determinace je roven 77,46%, adjustovaný $R^2 = 77,36\%$, tudíž vysvětlujeme větší část celkové variability v datech.

Abychom prozkoumali multikolinearitu mezi vysvětlujícími proměnnými, nejdříve si vypíšeme korelační matici, ve které najdeme hodnoty korelací vždy dvou proměnných. Symetrickou korelační matici nalezneme v tabulce 3.4. Všimněme si vyšší hodnoty korelace *HEIGHT* a *AGE*, a to 0,7919.

```
>> corrcoef(fev_data(:,[2,4:6]))
ans =
```

	AGE	**HEIGHT**	**GENDER**
AGE	1.0000	0.7919	0.0291
HEIGHT	0.7919	1.0000	0.1590
GENDER	0.0291	0.1590	1.0000

Tabulka 3.4: Korelační matice regresorů

Podívejme se na multikolinearitu z pohledu testování prostřednictvím Belsleyho diagnostiky kolinearity v datové matici nebo datové tabulce. K tomu využijeme funkci *collintest*. Po zadání vstupních hodnot dostáváme výstup uvedený v tabulce 3.5. Nejdříve se díváme na hodnoty ve sloupci *condIdx*, zda některá nepřesáhne defaultně nastavenou hodnotu 30. Pokud ano, hledáme v řádku hodnotu větší než 0,5 (opět defaultní), takové proměnné by způsobovaly multikolinearitu. V našem případě nebyla multikolinearita potvrzena.

```
>> collintest(fev_data(:,[2,4:5]))
```

sValue	condIdx	var1	var2	var3
1.6131	1	0.0066	0.0062	0.0510
0.6119	2.6361	0.0219	0.0134	0.9037
0.1536	10.5019	0.9715	0.9804	0.0452

Tabulka 3.5: Belsleyho diagnostika kolinearity

3.4 Testování parametrů a modelové specifikace

Při testování parametrů využijeme výpis intervalů spolehlivosti odhadů koeficientů *_.coefCI*, které získáme taktéž pomocí funkce *coefCI*. Testování statistické významnosti odhadů parametrů lze pomocí *_.coefTest* či funkcí *coefTest*. Jedná se o F-test, čili testujeme nulovou hypotézu $H_0 : \beta_0 = \beta_1 = \beta_2 = \beta_3 = 0$ oproti alternativě, že alespoň jeden koeficient je různý od nuly. P-hodnota je výrazně menší než obvyklé hladiny významnosti, tudíž alespoň jeden koeficient je statisticky významný.

```
>> coefCI(OLSregreseE,0.05)
ans =
   -4.8864   -4.0107
    0.0961    0.2262
    0.0952    0.1139
    0.0436    0.0792
>> coefTest(OLSregreseE)
ans =
  9.0530e-210
```

Podívejme se nyní na test, pomocí něhož můžeme testovat hypotézy o odhadech parametrů. K tomu použijeme funkci *linhyptest*, její specifikace vypadá takto:

$$[p, t, r] = linhyptest(B, COVB, C, H, DFE)$$

Výstupem funkce je p-hodnota *(p)*, testová statistika *(t)* a hodnost matice H *(r)*. Vstupní argument B je vektor parametrů odhadu, volitelným argumentem *COVB* lze nastavit kovarianční matici. Volitelné argumenty C a H slouží ke specifikaci testované hypotézy. Volitelný argument *DFE* určuje stupně volnosti pro odhad kovarianční matice. Zkusíme si pomocí funkce *linhyptest* otestovat nulovou hypotézu $H_0 : \beta_3 = 0$ neboli $H \cdot b = C$ pro $H = [0, 0, 0, 1]$ a $C = 0$. Čili je naším cílem otestovat statistickou významnost vlivu věku *AGE* na maximální vydechnutý objem vzduchu z plic *FEV*. Z provedeného testu, díky velice nízké p-hodnotě, lze říct, že parametr β_3 má vliv na *FEV*.

```
>> [p,t,r] = linhyptest(OLSregreseE.Coefficients.Estimate,
OLSregreseE.CoefficientCovariance,0,[0,0,0,1],
OLSregreseE.DFE);
p =
    2.9567e-11
t =
    45.7794
r =
    1
```

Zaměříme se nyní na porovnávání modelů, nejdříve pomocí testu Lagrangeových multiplikátorů (LM). Provedeme pomocnou regresi reziduí jednoduššího modelu *OLSregreseS* na proměnné rozšířeného modelu *OLSregreseE*. Spočítáme si hodnotu statistiky $LM = N \cdot R^2$, která se řídí χ^2-rozdělením se stupni volnosti rovnu počtu restrikcí v parametrech *LM_dof*. Na základě malé p-hodnoty lze rozšířený model považovat za lepší, viz *LM_pval*. Pokud bychom test chtěli provádět pomocí funkce *lmtest*, řešení je doplněno v příloze 3.10.

```
OLSregresePOM = fitlm(X,OLSregreseS.Residuals.Raw,
    'VarNames',{'Gender','Height','Age','FEV'})
LM_stat = N * OLSregresePOM.Rsquared.Ordinary;
LM_dof = 2;
LM_pval = chi2cdf(LM_stat,LM_dof,'upper')
LM_pval =
    2.8032e-109
```

Další možnost porovnání modelů je pomocí testu poměru věrohodností *(LR)*, kdy využijeme funkci *lratiotest*. Výstupem je rozhodnutí o (ne)zamítnutí nulové hypotézy (LR_h), p-hodnota *(LR_pValue)*, testovací statistika *(LR_stat)* a kritická hodnota *(LR_cValue)*. Vstupem jsou věrohodnosti neomezeného *OLSregreseE* a omezeného *OLSregreseS* modelu a stupně volnosti *(LR_dof)*. Volitelně lze nastavit hladinu významnosti, nastavíme ji na 5 %. Výsledky provedeného testu z tabulky 3.6 upřednostňují neomezený model.

```
uLL = OLSregreseE.LogLikelihood;
rLL = OLSregreseS.LogLikelihood;
LR_dof = OLSregreseE.DFE;
[LR_h,LR_pValue,LR_stat,LR_cValue] = lratiotest(uLL,rLL,
    LR_dof,0.05)
```

LR_h	LR_pValue	LR_stat	LR_cValue
1	2.5902e-13	945.3693	710.4211

Tabulka 3.6: Test poměru věrohodností - použití funkce *lratiotest*

Podívejme se nyní na Waldův test (W), pomocí něhož porovnáme opět neomezený *OLSregreseE* a omezený *OLSregreseS* model. K tomu slouží funkce *waldtest*, která vrátí rozhodnutí o (ne)zamítnutí nulové hypotézy *(WALD_h)*, p-hodnotu testu *(WALD_pValue)*, testovací statistiku *(WALD_stat)* a kritickou hodnotu *(WALD_cValue)*. Jejím vstupem jsou restriktivní funkce *(Wald_r)*, restriktivní funkce Jakobiho matice *(Wald_R)*, odhady kovariance parametrů neomezeného modelu *(Wald_EstCov)* a hladina významnosti. Na základě Waldova testu z tabulky 3.7 bychom zamítli nulovou hypotézu omezeného modelu $\beta_2 = \beta_3 = 0$ ve prospěch alternativní hypotézy neomezeného modelu. Neuvažujeme omezené modely $\beta_2 = 0$, nebo $\beta_3 = 0$.

```
Wald_r = [OLSregreseE.Coefficients.Estimate(3),
    OLSregreseE.Coefficients.Estimate(4)]
Wald_R = [0,0,1,0;0,0,0,1]
Wald_EstCov = OLSregreseE.CoefficientCovariance
[WALD_h,WALD_pValue,WALD_stat,WALD_cValue] =
    waldtest(Wald_r,Wald_R,Wald_EstCov,0.05)
```

WALD_h	WALD_pValue	WALD_stat	WALD_cValue
1	0	2.1086e+03	5.9915

Tabulka 3.7: Waldův test - použití funkce *waldtest*

3.5 Predikce

Velmi krátce se také zmíníme o predikci pomocí funkce *predict*. Jako nejlepší model jsme po testování zvolili tento (můžete otestovat, že regresor vztahu ke kouření *SMOKE* vykazuje statistickou nevýznamnost):

$$FEV_i = \beta_0 + \beta_1 \cdot GENDER_i + \beta_2 \cdot HEIGHT_i + \beta_3 \cdot AGE_i + \varepsilon_i$$

Pokusíme se nyní predikovat maximální vydechnutý objem vzduchu z plic (v litrech) pro ženu (kódovaná na 0), vysokou 68 palců, ve věku 30 let.

```
>> predict(OLSregreseE,[0,68,30])
ans =
    4.5024
```

3.6 Heteroskedasticita a robustní odhad

V této části kapitoly se zaměříme na problematiku (ne)konstantního rozptylu. Budeme pracovat s datovým souborem *electric*, jedná se o data sledující produkci výroby elektřiny v roce 1970 v USA. Datový soubor obsahuje proměnné: náklady produkce v mil. dolarů (cost), výstup produkce v kWh za rok (output), cena práce v dolarech (price_l), cena kapitálu v dolarech (price_k) a cena paliv v dolarech (price_f). Nejdříve si datový soubor načteme pomocí *readmatrix* a uložíme si do proměnných ty, se kterými budeme dále pracovat (pozn.: nezapomeňte si před novou regresí vyčistit pracovní prostředí.).

```
clear
clc
electric = readmatrix('electric.csv');
cost = electric(:,1);
output = electric(:,2);
price_l = electric(:,3);
price_k = electric(:,4);
price_f = electric(:,5);
```

Sestavíme matici plánu a vektor vysvětlované proměnné. Následně provedeme regresi pomocí funkce *fitlm*.

$$cost_i = \beta_0 + \beta_1 \cdot output_i + \beta_2 \cdot price\_l_i + \beta_3 \cdot price\_k_i + \beta_4 \cdot price\_f_i + \varepsilon_i$$

```
Y = cost;
X = [output,price_l,price_k,price_f];
electricityOLS = fitlm(X,Y);
```

Pokud chceme ověřit předpoklad homoskedasticity, musíme test na ověření předpokladu provést ručně, neboť v Matlabu není žádný test přímo implementován. Provedeme Whiteův test, pro který si nejdříve uložíme predikované hodnoty *y_predicted* a provedeme následující regresi (detailnější popis testů ověřující konstantní rozptyl náhodné složky modelu je k dipozici v části 5.3.3 Základů ekonometrie, Němec, 2019):

$$\hat{\varepsilon}_i^2 = \delta_0 + \delta_1 \cdot \hat{Y}_i + \delta_2 \cdot \hat{Y}_i^2$$

```
y_predicted = predict(electricityOLS, [ ,output,price_l,
    price_k,price_f]);
white_pom_regression = fitlm([y_predicted,
    y_predicted.^2],electricityOLS.Residuals.Raw.^2);
```

Testová statistika Whiteova testu je rovna součinu počtu pozorování a R^2 předešlé regrese, stupně volnosti jsou 3, testová statistika se řídí χ^2-rozdělením. Na základě velice nízké p-hodnoty nulovou hypotézu, H_0 : Odhadnutý model splňuje předpoklad homoskedasticity, zamítáme na všech obvyklých hladinách významnosti.

```
[N,~] = size(X)
whites_test_statistic = N *
    white_pom_regression.Rsquared.Ordinary;
whites_df = N - white_pom_regression.DFE;
white_pval = chi2cdf(whites_test_statistic,whites_df,
    'upper');
white_pval =
    6.6940e-08
```

Ukážeme si také ručně naprogramovaný Goldfeldův-Quandtův test, který předpokládá, že pouze jedna vysvětlující proměnná může způsobovat nekonstantní rozptyl, díky tomuto předpokladu lze vzorek podle proměnné způsobující heteroskedasticitu rozdělit na 2 části. Ze zapříčinění heteroskedasticity budeme podezírat proměnnou *price_f*, protože ceny paliv bývají velmi volatilní a při výrobě nemají žádný substitut. Podle této proměnné seřadíme datový vzorek vzestupně, v rámci obou polovin provedeme regresi a porovnáme součty čtverců reziduí (SSE). Také si uložíme velikost matice plánu a po seřazení jednotlivých prvků v datové matici si do proměnné *Is* vložíme index pozice před přeřazením.

```
[n, m] = size(X);
[Xsort,Is] = sort(X(:,4));
```

Dále je sestrojen cyklus prohledávající pozorování, kde se data rozdělí na dva vzorky, které se porovnávají, jestli se mění rozptyl na začátku a na konci vzorku.

```
for i = 1:size(Y)
       Ysort(i,1) = Y(Is(i),1); % prerazuje podle Xsort
       Xsort(i,1:3) = X(Is(i), 1:3);
   end
Dat = [Xsort Ysort];
c = fix(4*n/15);
k = fix((n - c)/2);
if floor(k) > 0.4
    k = k+1;
end
```

Díváme se nejdříve na první vzorek. Je odhadnut model pomocí metody OLS, jsou uloženy odhady koeficientů *b1*, odchylka vyrovnání *dev1* a struktura statistik *stats1*.

```
Dat1 = Dat(1:k,:);
[b1,dev1,stats1] = glmfit(Dat1(:,1),Dat1(:,2));
S1 = sum(stats1.resid.^2);
```

To stejné provedeme pro 2. vzorek.

```
Dat2 = Dat(n-k+1:n,:);
[b2,dev2,stats2] = glmfit(Dat2(:,1),Dat2(:,2));
S2 = sum(stats2.resid.^2);
```

Nyní přichází samotné testování hypotézy o případném nekonstantním rozptylu, a to pomocí porovnání součtů čtverců reziduí (SSE).

```
if S1 > S2
    Fp = S1/S2;
   else
    Fp = S2/S1;
end
```

Dále je použita funkce *finv*, která vyhodnotí inverzní kumulativní distribuční funkci pro hladinu významnosti 5 % a $k - m - 1$ stupni volnosti.

```
Ft = finv(0.95,k-m-1,k-m-1);
if Fp > Ft
    fprintf('\nHeteroscedasticity is present\n\n')
   else
    fprintf('\nHeteroscedasticity is absent\n\n')
end
ans =
    Heteroscedasticity is present
```

Jelikož je v datovém vzorku přítomna heteroskedasticita, OLS estimátor není BLUE, tj. nejedná se o nejlepší, lineární, nestranný (neboli nevychýlený) odhad. Proto je nutné provést nový robustní odhad zobecněnou metodou nejmenších čtverců (GLS). Lze využít funkci *fgls* s nastavením odpovídající problému s heteroskedasticitou.

```
>> [GLS_coeff,GLS_se,GLS_EstCoeffCov] = fgls(X,Y,'innovMdl',
'HC1')
```

Ekvivalentně lze také provést regresi pomocí funkce *fitlm* doplněnou o vektor vah např. $w = 1/rezidua^2$. Výstup nalezneme v tabulce 3.8.

```
w = electricityOLS.Residuals.Raw.^(-2);
electricityWLS = fitlm(X,Y,'weights',w);
ans =
```

Linear regression model:
y ~ 1 + x1 + x2 + x3 + x4
Estimated Coefficients:

	Estimate	**SE**	**tStat**	**pValue**
(Intercept)	-65.765	1.9238	-34.186	4.658e-63
x1	0.0047369	1.6334e-05	290	3.2193e-170
x2	0.0032739	0.00012119	27.015	2.3318e-52
x3	0.26124	0.010195	25.625	4.8302e-50
x4	0.76336	0.020171	37.845	7.7567e-68

Number of observations: 123, Error degrees of freedom: 118
Root Mean Squared Error: 0.98
R-squared: 0.999, Adjusted R-Squared: 0.999
F-statistic vs. constant model: 2.51e+04, p-value = 7.48e-172

Tabulka 3.8: Robustní odhad modelu - použití funkce *fitlm*

V případě, že kombinace více proměnných způsobuje heteroskedasticitu, neumíme transformovat model a využít GLS estimátor. Pro takovou situaci je vhodné použít heteroskedasticitě a autokorelaci konzistentní estimátor (HAC). Provedeme tedy nový heteroskedasticitě konzistentní odhad pomocí funkce *hac*. Výstup v podobě odhadů koeficientů *HAC_coeff*, směrodatných chyb *HAD_se* a kovarianční matice *HAC_EstCov* nalezneme v tabulce 3.9.

```
>> [HAC_EstCov,HAC_se,HAC_coeff] = hac(X,Y,'type','HC',
'weights','HC1');
```

HAC_se	HAC_coeff	HAC_EstCov				
16.7669	-70.4951	281.1304	0.0005	-0.0141	-1.7605	-1.6736
0.0003	0.0047	0.0005	0.0000	0.0000	-0.0000	0.0000
0.0010	0.0036	-0.0141	0.0000	0.0000	0.0001	0.0001
0.1289	0.2801	-1.7605	-0.0000	0.0001	0.0166	0.0077
0.1409	0.7835	-1.6736	0.0000	0.0001	0.0077	0.0199

Tabulka 3.9: Výstup HAC estimátoru - použití funkce *hac*

Doplnění

- **[b,stats] = robustfit(X,Y,'wfun',tune,'const')** - Funkce provede robustní regresi pro vektor vysvětlovaných hodnot Y na základě matice prediktorů X a vrátí odhady koeficientů b a modelové statistiky *stats*. Volitelnými argumenty jsou funkce robustního váženého vyrovnání *wfun*, ladící konstanta *tune* a indikátor pro úrovňovou konstantu *const*.

3.7 Modely diskrétní volby

Pro studium modelů diskrétní volby se vrátíme zpět k datovému souboru *fev.csv* o maximálním vydechnutém objemu vzduchu z plic (v litrech). Více informací o datovém souboru najdete na začátku této kapitoly 3. Nejdříve si vyčistíme pracovní plochu a načteme si data pomocí *readmatrix*.

```
clear
clc
fev_data = readmatrix("fev.csv");
AGE = fev_data(:,2);
FEV = fev_data(:,3);
GENDER = fev_data(:,5);
SMOKE = fev_data(:,6);
```

V této části kapitoly věnované modelům diskrétní volby budeme vysvětlovat vztah ke kouření (SMOKE - kuřák = 1, nekuřák = 0) pomocí proměnných věk (AGE) a pohlaví (GENDER - žena = 0, muž = 1). Budeme využívat funkci *mnrfit*, pomocí které lze sestavit mnoho modelů diskrétní volby, my si zde ukážeme logit model. Nejdříve je nutné kategorizovat proměnnou SMOKE.

```
Y = categorical(SMOKE);
[LOGIT_B,LOGIT_dev,LOGIT_stats] = mnrfit([AGE,GENDER],Y);
```

LOGIT_B	LOGIT_dev	LOGIT_stats.t	LOGIT_stats.p
7.5861	311.6799	10.5282	0.0000
-0.5009		-8.7759	0.0000
0.7929		2.5729	0.0101

Tabulka 3.10: Výstupy logit estimátoru

Podívejme se na interpretaci parametrů modelu, v prvním sloupci tabulky 3.10 je odhad interceptu a koeficientů pro modelování logistické pravděpodobnosti nekuřáka a kuřáka. Máme-li ženu s nulovým věkem, mezní vliv na podíl šancí, že se bude jednat o kuřačku, je roven $7,5861$. Pokud uvážíme 2 lidi stejného pohlaví, kdy jeden z nich bude o 1 rok starší, jeho mezní vliv na podíl šancí, že se bude jednat o kouřícího člověka, je roven $-0,5009$, jinými slovy pravděpodobnost je $e^{-0,5009} \doteq 0,61x$ vyšší. Na závěr pokud uvážíme muže a ženu stejného věku, mezní vliv na podíl šancí, že se bude jednat o kuřáka, je roven $0,7929$, čili pravděpodobnost je $e^{0,7929} \doteq 2,21x$ vyšší.

Opět si můžeme nechat vypsat detaily provedeného odhadu, např. vektor t-statistik koeficientů získáme pomocí *LOGIT_stats.t* a p-hodno-ty koeficientů pomocí *LOGIT_stats.p*. Obě zmíněné charakteristiky najdeme společně s výstupy logit estimátoru v tabulce 3.10. Vidíme, že odhady jsou na 5% hladině významnosti statisticky významné. Kovarianční matici obdržíme příkazem *LOGIT_stats.covb*, korelační matici pomocí *LOGIT_stats.coeffcorr*. Stupně volnosti provedeného odhadu získáme analogicky *LOGIT_stats.dfe*, pozorovaná rezidua příkazem *LOGIT_stats.resid*, pearsonova rezidua pomocí *LOGIT_stats.residp* a ochylky reziduí prostřednictvím příkazu *LOGIT_stats.residd*.

```
>> LOGIT_stats.t;
>> LOGIT_stats.p;
>> LOGIT_stats.covb;
>> LOGIT_stats.coeffcorr;
>> LOGIT_stats.dfe;
>> LOGIT_stats.resid;
>> LOGIT_stats.residp;
>> LOGIT_stats.residd;
```

Velmi užitečná funkce je také *mnrval*, která slouží pro odhad pravděpodobnosti multinomiální logistické regrese. Pokud uvážíme ženu, kódovanou na 0, ve věku 21 let, poté pravděpodobnost, že se bude jednat o kuřačku je 94,94 %.

```
>> mnrval(LOGIT_B,[21,0])
ans =
    0.0506    0.9494
```

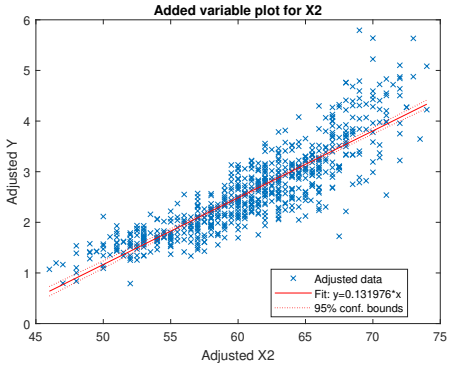
3.8 Jiné typy regrese a zobecněné funkce

V této doplňkové části se podíváme na postupnou regresi, jedná se o interaktivní prostředí odhadu pomocí funkce *stepwise*. Jde o systematickou metodu

pro přidávání a odebírání proměnných z vícenásobné regrese na základě jejich statistické významnosti. Vstupní argumenty jsou stejné jako u běžné regrese, vektor vysvětlovaných hodnot a matice prediktorů, další volitelné parametry jsou počáteční nastavení modelu, vstupní a výstupní p-hodnoty pro F-test. Opět jsou pro ukázku použita data ze souboru *fev.csv*, která byla představena na začátku kapitoly 3. Pro připomenutí uvádíme vektor vysvětlovaných hodnot a matici plánu.

```
Y = FEV;
X = [GENDER,HEIGHT,SMOKE];
stepwise(X,Y);
```

Pomocí *Add Variable Plot* lze vykreslit příslušné grafy, pomocí *next step* lze přidávat nebo ubírat proměnné, pomocí *all steps* přeskočíme na konec. Na obrázku 3.2 je demonstrován graf s provedenou regresí s vysvětlující proměnnou *HEIGHT*.



Obrázek 3.2: Provedený odhad s regresorem výšky - použití funkce *stepwise*

Abychom získali model nejlépe popisující závislost mezi vektorem vysvětlovaných hodnot a maticí plánu datového souboru *fev.csv* pomocí postupné regrese, použijeme funkci *stepwiselm*. Výstup nalezneme v tabulce 3.11. Můžeme si všimnout, že model obsahuje interakci proměnných *GENDER* a *HEIGHT*, jaká by byla interpretace modelu s interakcí? Uvažujeme-li dvě ženy, z nichž jedna je o 1 palec vyšší, její maximální vydechnutý objem vzduchu z plic bude o $0,11243$ l větší. Uvažujeme-li dva muže, z nichž jeden je o 1 palec vyšší, maximální vydechnutý objem plic bude o $0,11243 + 0,027457 \doteq 0,14$ l větší. Tudíž vliv výšky má výraznější dopad u mužů.

```
>> stepwiselm(X,Y,'VarNames',{'Gender','Height','Smoke',
'FEV'});
```

Linear regression model:
FEV 1 + Gender*Height
Estimated Coefficients:

	Estimate	**SE**	**tStat**	**pValue**
(Intercept)	-4.3182	0.29764	-14.508	1.5968e-41
Gender	-1.5456	0.37384	-4.1344	4.0246e-05
Height	0.11243	0.0049276	22.815	4.3784e-85
Gender:Height	0.027457	0.0061193	4.487	8.543e-06

Number of observations: 654, Error degrees of freedom: 650
Root Mean Squared Error: 0.42
R-squared: 0.766, Adjusted R-Squared: 0.765
F-statistic vs. constant model: 709, p-value = 1.8e-204

Tabulka 3.11: Optimální model pomocí postupné regrese - funkce *stepwiselm*

Doplnění

- **[b,se,pval,finalmodel,stats,nextstep,history] = stepwisefit(X, Y, inmodel,penter,premove)** - Pro detailnější informace lze použít funkci *stepwisefit*. Vstupní argumenty jsou stejné jako ve funkci *stepwise*, Volitelně lze specifikovat počáteční model *inmodel*, vstupní *penter* a výstupní *premove* toleranci pro p-hodnoty F-statistiky. Mezi výstupními hodnotami dostaneme odhady koeficientů *(b)*, standardní chyby *(se)*, p-hodnoty *(pval)*, specifikaci proměnných v konečném regresním modelu *(finalmodel)*, statistiku konečného modelu *(stats)*, doporučený další krok *(nextstep)* a informace o historii všech provedených kroků *(history)*.

- **stats = regstats(responses, data, model)** - Funkce provádí regresní diagnózu *stats* vektoru pozorovaných hodnot *responses* na základě matice dat/prediktorů *data* za použití vícenásobného lineárního modelu. Funkce má mnoho volitelných argumentu týkající nastavení modelu, např. *linear, quadratic, purequadratic*.

- **h = leverage(data, model)** - Funkce, která najde vliv (*h*, čili vzdálenost bodu od regresní přímky) pro všechny řádky (body) matice dat. Je možné pomocí volitelného argumentu *model* upřesnit, jaký model chceme využít. (Leverage points jsou hodnoty na diagonále matice vyrovnání H.)

- **[b,stats] = lasso(X,Y)** - Funkce, která provede lasso (least absolute shrinkage and selection operator) a vrátí vyrovnané koeficienty a informace o regresi modelu. Vstupní parametry této funkce jsou numerická matice regresorů X a numerický vektor vysvětlované proměnné Y. Volitelné vstupní argumenty: *NumLambda* - počet použitých hodnot lambd, *LambdaRatio* - poměr nejmenší a největší lambdy, *Lambda* - hodnoty lambd, *CV* - specifikace cross-validation. Funkce obsahuje mnoho dalších volitelných argumentů. Některé z nich jsou totožné s běžnou lineární regresí.

- **[beta,sigma,E,CovB,logL] = mvregress(X,Y)** - Funkce provede vícerozměrnou lineární regresi pro matici vícerozměrných pozorování Y na základě matice prediktorů X. Výstupy funkce jsou: odhadnuté regresní koeficienty *beta*, odhadnutá kovarianční matice *sigma*, rezidua modelu *E*, parametr odhadnuté kovarianční matice *CovB* a věrohodnost modelu *logL*. Funkce má mnoho volitelných argumentů. Mezi ně například se řadí *algorithm*, který umožňuje specifikovat algoritmus odhadu nebo *beta0* a *covar0*, které umožňují specifikovat počáteční nastavení koeficientů regrese a kovarianční matice.

- **[Xl,Yl,Xs,Ys,beta,pctVAR,MSE] = plsregress(X,Y,ncomp)** - Funkce provádí regresi metodou částečných nejmenších čtverců pro vektor vysvětlovaných hodnot Y na základě matice prediktorů X za použití komponent *ncomp*. Funkce vrátí matici zatížení pro regresor *Xl* a vysvětlovaná pozorování *Yl*, rovněž také matice jejich skórů *Xs, Ys*. Dále vrací regresní koeficienty *beta*, matici obsahující procentuální rozptyl vysvětlený modelem $pctVAR$ a průměrnou standardní chybu *MSE*.

- **[Coeff,se,coeffPlots] = recreg(X,Y)** - Funkce provede rekurzivní regresi pro vektor vysvětlovaných hodnot Y na základě matice prediktorů X. Funkce vrací odhadnuté koeficienty *Coeff* a standardní chyby *se*. Funkce také umí vykreslit získané výsledky pomocí parametru *coeffPlots*. Obsahuje volitelné argumenty pro samotnou regresi, např. zahrnutí úrovňové konstanty *Intercept*, ale i pro grafický výstup, např. jména proměnných *VarNames*.

- **[X,resnorm,residual,exitflag,output,lambda,jacobian] = lsqnonlin(FUN,X0)** - Funkce vrací řešení nejmenších čtverců X s počátkem v bodě *X0* pro zadanou funkci *FUN*. Dalším výstupem jsou: čtvercová norma reziduí *resnorm*, rezidua *residual*, důvod kvůli kterému byla funkce ukončena *exitflag*, informace o optimalizačním procesu *output*, Lagrangeovy multiplikátory *lambda* a Jacobiho matici *jacobian*.

Volitelně lze zvolit, kromě různých nastavení, spodní a horní hranici pro řešení *lb,ub*.

- **[beta,R,J,CovB,MSE,ErrorModelInfo] = nlinfit(X,Y, modelfun,beta0)** - Funkce provede nelineární regresi pro vektor vysvětlovaných hodnot Y na základě matice prediktorů X podle specifikované funkce nelinearního regresního modelu *modelfun* a počátečních hodnot koeficientů *beta0*. Výstupy funkce jsou: odhadnuté regresní koeficienty *beta*, rezidua modelu R, Jacobiho matice J, odhad kovarianční matice *CovB*, průměrná standardní chyba *MSE* a informace o chybě vyrovnání modelu *ErrorModelInfo*. Funkce má argument pro různá nastavení *options* pro potřeby uživatele. Funkce má také volitelný argument vah *Weights*.

- **x = fminimax(fun,x0,A,b)**
  [x,fval,maxfval,exitflag,output,lambda] = fminimax(fun,x0, A,b) - Funkce, která řeší soustavu rovnic *fun* za různých podmínek, které definujeme vstupními parametry. Pokud máme soustavu lineárních nerovnic $A \cdot x \leq b$, postačí zadání 4 vstupních parametrů, pro další omezení (lineární rovnosti, mezí atd.) nahlédněte do helpu. Jsou velice široké možnosti v nastavení výstupu při výpočtu, např. jeho zobrazení. Rozšířením lze získat včetně řešení i funkční hodnoty *fval*, maximální funkční hodnotu *maxfval*, podmínku ukončení výpočtu *exitflag*, informace o průběhu optimalizace *output* a hodnoty Lagrangeových multiplikátorů *lambda*.

- **x = fsolve(fun,x0)**
  [x,fval,exitflag,output,jacobian] = fsolve() - Pomocí této funkce lze řešit soustavu nelineárních rovnic, která vychází z počátečního bodu *X0*. Lze pomocí široké škály možností nastavit průběh výpočtu. Rozšířeným výstupním vektorem obdržíme navíc funkční hodnoty *fval*, výstupní podmínku *exitflag*, detailní informace o průběhu optimalizace *output* a jakobián *jacobian*.

- **x = fminunc(fun,x0)**
  [x,fval,exitflag,output,grad,hessian] = fminunc() - Funkce, která hledá minimum neomezené funkce více proměnných. Vychází ze zadaného počátečního bodu *x0*, který může být zadán jako skalár, vektor či matice. Opět je zde široká škála možností pro nastavení optimalizace. Širším výstupním vektorem obdržíme navíc funkční hodnoty *fval*, výstupní podmínku *exitflag*, detailní informace o průběhu optimalizace *output*, gradient *grad* a hesián *hessian*.

3.9 Neřešené příklady

1. Budeme zpracovávat datový soubor *music.csv* zabývající se stanovování cen kompaktních disků, který obsahuje následující proměnné: *PRICE* (cena disku v dolarech), *AGE* (stáří nahrávky jako rozdíl mezi rokem 1999 a rokem vzniku copyrightu k dané nahrávce), *OLD* (1 - nahrávka není nová, 0 jinak), *NET* (1 - internetová cena, 0 jinak). Mixon a Ressler (2000) si všimli si, že vydavatelé a obchodníci často stanovují nižší cenu pro nová vydání disků vzhledem k ceně u vydání starších nahrávek. Cílem je toto tvrzení prozkoumat.

   (a) Načtěte si datový soubor, uložte si proměnné, odhadněte uvedený model a interpretujte jej. Díky tomu, že proměnná *NET* je kategorická, do příslušné funkce je nutné přidat parametr *Categorical-Vars*, pomocí něhož upozorníme, která z vysvětlujících proměnných je kategoriální. *Model 1*:

$$PRICE_i = \beta_0 + \beta_1 \cdot AGE_i + \beta_2 \cdot NET_i + \varepsilon_i$$

   (b) Nyní odhadněte upravený model.

$$PRICE_i = \beta_0 + \beta_1 \cdot OLD_i + \beta_2 \cdot NET_i + \varepsilon_i$$

   (c) Odhadněte rovněž model se všemi vysvětlujícími proměnnými a model nazvěte *model_3*.

   (d) Otestujte hypotézu u modelu 3, že každý rok navíc dané nahrávky zvyšuje cenu kompaktního disku o 30 centů.

   (e) Upravte *model_1* tak, abyste byli schopni testovat hypotézu, že internetová cena kompaktního disku roste mnohem rychleji se stářím nahrávky než neinternetová.

   (f) Nyní modifikujte *model_1*, abyste dokázali otestovat, že vliv stáří nahrávky se projevuje až pro nahrávky staré 3 roky a více.

2. Jakub V. chytil 3 různé ryby ve svém oblíbeném rybníku a rozhodl se zhodnotit své úlovky na místním tržišti s rybami. Chce zjistit, s jak velkými rybami se na tržišti může setkat, aby věděl, jak si svými úlovky stojí. Tabule s cenami ryb v EUR/kg vyvěšená na tržišti je k dispozici v tabulce 3.12. Vaším cílem je odhadnout pomocí lineární regrese váhu ryb na tržišti a predikovat jeho výdělek z jeho chycených ryb. Pracujte s datovým souborem *Fish.csv*, který zaznamenává prodej sedmi běžných druhů ryb na tržišti. Datový soubor obsahuje proměnné: druh

ryby *(Species)*, váhu v gramech *(Weight)*, šířku *(Width)*, vertikální délku *(Length1)*, diagonální délku *(Length2)* a křížovou délku *(Length3)*. V příkladu pracujte s proměnnou délky *(Length1)*.

(a) Vytvořte lineární regresní model pro odhad vah ryb na tržišti podle jejich druhu a délky. V případě, že některé koeficienty jsou záporné, pokuste se odůvodnit, proč tomu tak je?

$$Weight_i = \beta_0 + \beta_1 \cdot Species_i + \beta_2 \cdot Length1_i + \varepsilon_i$$

Pro kategorizaci proměnné *Species* můžete vycházet z tabulky 3.12.

Kódovaní	Druh	Cena v EUR
1	Bream	9
2	Roach	8
3	Whitefish	2.3
4	Parkki	26.7
5	Perch	14.5
6	Pike	7.5
7	Smelt	36

Tabulka 3.12: Popis ryb nabízené na tržišti

(b) Graficky znázorněte provedenou regresi a vykreslete graf reziduí.

(c) Pokuste odhadnout ceny Jakubových ryb. Můžete použít přibližný kurz 26,5 Kč = 1 Euro. Jakub V. chytil tyto ryby: Bream (38,4 cm), Roach (16,1 cm) a Perch (50,2 cm).

3. World Health Organization každým rokem vydává zprávy týkající se zdraví lidí napříč celým světem. Pokuste se na základě datového vzorku z WHO odhadnout vliv na délku lidského života. Pracujte s datovým souborem *life_ expectancy.csv*, který obsahuje: průměrnou délku života *(Life Expectancy)*, mortalitu lidí ve věku 15 až 60 let na 1000 obyvatel *(Adult Mortality)*, binární proměnnou označující, zda se jedná o vyspělou nebo rozvíjející se ekonomiku *(Status) (Developing = 1, Developed = 0)*, počet dětských úmrtí na 1000 obyvatel *(Infant Deaths)*, HDP *(GDP)*, průměrné BMI v populaci *(BMI)*, očkování dětí ve věku 1 roku proti žloutence (v %) *(Hepatitis B)*, očkování proti poliu u dětí ve věku 1 roku (v %) *(Polio)*, očkování proti tetanu u dětí ve věku 1 roku (v %) *(Diphtheria)*, počet případů spalniček na 1000 obyvatel *(Measles)*, počet úmrtí dětí do 5 let na 1000 obyvatel *(Polio)*, vládní výdaje na

healthcare *(Total Expenditure)*, průměrný počet let studia *(Schooling)* a počet dětských úmrtí na AIDS ve věku 0-4 let *(HIV/AIDS)*.

(a) Odhadněte pomocí linearní regrese vliv vybraných proměnných na průměrnou délku života ve světě a jak se jejich vliv změnil za 15 let. Tudíž proveďte dvě regrese, v roce 2000 a 2015. Nicméně jde o real world data, takže očekávejte, že budou zaneřáděná, a je tedy třeba je očistit před začátkem odhadu. Dávejte si pozor, abyste měli pro oba roky vzorky stejných zemí.

(b) Porovnejte výsledky a rozhodněte, co by mohlo vysvětlovat změny ve vlivu vysvětlujících proměnných.

(c) Vyzkoušejte alespoň 1 další model z datového souboru, který by vysvětloval očekávanou délku života *LifeExpect*.

4. V tomto neřešeném příkladu budeme pracovat s datovým souborem *golf.csv*, který obsahuje údaje o skóre *SCORE* nejlepšího golfisty Liona Foresta ze 150 turnajů spolu s údaji o jeho věku *AGE*. Ve věku 45 let však jeho hra přestala být tou, kterou bývala dříve. Svou profesionální kariéru začal, když mu bylo 20 a po dovršení svých 45. narozenin ho náhle začala zajímat analýza historického průběhu jeho výsledků v závislosti na tom, jak postupně stárnul. Prostudujte jeho výsledky.

(a) Testujte, který model je vhodnější, zda-li kvadratický nebo kubický.

$$SCORE_i = \beta_0 + \beta_1 \cdot AGE_i + \beta_2 \cdot AGE_i^2 + \beta_3 \cdot AGE_i^3 + \varepsilon_i$$

(b) Vytvořte si modelové predikce pro kubický model a odpovězte na následující otázky:

i. V jakém věku byl Lion na vrcholu své kariéry?

ii. V jakém období jeho věku docházelo ke zlepšování jeho hry, a to rostoucím tempem růstu

iii. V jakém období docházelo ke zlepšování Lionovy hry, a to snižujícím se tempem růstu?

iv. Ve kterém věku začal hrát Lion hůře než na začátku své kariéry (což bylo ve věku 20 let)?

v. Od kterého věku Lion už nebyl schopen hrát pod par?

vi. Když bude Lionovi 70, bude (podle našeho modelu) schopen zahrát turnaj na 100 úderů? Předpokládáme, že par turnaje je 72.

5. Model ocenění kapitálových aktiv (capital asset pricing model – CAPM) je důležitý model v oblasti financí (spadající do obecné skupiny tzv. stochastických diskontních faktorových modelů). Vysvětluje nám variabilitu v mírách výnosnosti cenných papírů jako funkci míry výnosnosti portfolia skládajícího se ze všech veřejně obchodovatelných akcií, což se nazývá tržním portfóliem. Obecně je míra výnosnosti investice měřena relativně ke svým nákladům obětované příležitosti, které jsou často chápány jako výnosnost bezrizikového aktiva. Výsledný rozdíl je riziková prémie (risk premium), nebot' se jedná o odměnu za rizikovou investici. CAPM říká, že riziková prémie cenného papíru je proporcionální rizikové prémii tržního portfolia. To znamená

$$r_j - r_f = \beta_j(r_m - r_f),$$

kde r_j a r_f jsou postupně výnosy z j-tého cenného papíru a bezriziková úroková míra, r_m, je výnos tržního portfolia a β_j je tzv„beta“ hodnota j-tého cenného papíru. Tato beta je důležitým indikátorem pro investory, nebot' se v ní objevuje volatilita dané akcie. Měří se jí citlivost výnosů j-tého cenného papíru vzhledem k variabilitě celého akciového trhu. Pro hodnoty beta menší než 1 se jedná o „defenzivní“ tituly, protože jejich variabilita je menší než variabilita celého trhu. Hodnoty beta větší než 1 hovoří o „agresivní akcii“. Investor při konstrukci svého portfolia obvykle chce znát betu dané akcie, a to před tím, než se rozhodne k jejímu nákupu. CAPM uvedený výše je tedy „ekonomickým modelem“. „Ekonometrický“ model získáme doplněním úrovňové konstanty (ačkoli teorie říká, že by měla být její hodnota nulová, což můžeme následně otestovat) a náhodné složky:

$$r_j - r_f = \alpha + \beta_j(r_m - r_f) + \epsilon.$$

Pracujte s datovým souborem *CAPM_ data.csv* s výnosností akcií společnosti Apple (2. sloupec), Microsoft (3. sloupec), Tesla (4. sloupec) a Intel (5. sloupec). Datový soubor v 1. sloupci obsahuje úrok bezrizikového aktiva (výnosy tržního portfolia - bezriziková úroková míra).

   (a) Odhadněte povahu akcií každé společnosti. Doporučujeme odhadovat v rámci cyklu pro ušetření práce. Po odhadnutí modelů otestujte u každého modelu normální rozložení reziduí, vykreslete si graf provedené regrese a reziduí.

   (b) V případě porušení normality otestuje, zda se vyskytují v modelu odlehlá pozorování. Pokud se v modelu vyskytují, nahraďte je mediánem a odhadněte modely znovu a opět je otestujte na normalitu.

(c) V případě, že ani po nahrazení odlehlých pozorování nesplňují modely předpoklady normality, spočítejte statistiky pomocí studentova t-rozdělení.

(d) (Dobrovolný) Pokročilí uživatelé mohou vyzkoušet odhadnout model pomocí funkce fminunc a věrohodnostní funkce studentova t-rozdělení.

6. Každé ráno mezi 6:30–8:00 odchází Bill do práce, prozkoumejte proměnné, které ovlivňují délku jeho cestování. K tomu budete potřebovat datový soubor *commute.csv*, který sleduje faktory ovlivňující čas strávený na cestě do práce *time*, čas odjezdu *depart*, počet červených světel na semaforech *reds*, počet čekání na vlakových přejezdech z důvodu projíždějícího vlaku *trains*.

(a) Sestrojte následující model:

$$time_i = \beta_0 + \beta_1 \cdot depart_i + \beta_2 \cdot reds_i + \beta_3 \cdot trains_i + \varepsilon_i$$

(b) Získejte 95 % intervalový odhad pro odhady koeficientů.

(c) Nyní testujte hypotézu, že každé červené světlo zpozdí Billa nejméně o 2 minuty, a to oproti alternativě, že zpozdění je menší než 2 minuty.

(d) Testujte hypotézu, že každé čekání na železničním přejezdu zpozdí Billa o 3 minuty.

(e) Testujte hypotézu, že minimální čas Billovy cesty do práce je menší nebo roven 20 minutám.

(f) Testujte hypotézu, že zpozdění příjezdu do práce způsobené vlakem je stejné jako trojnásobek zpozdění způsobeného červeným světlem na semaforu.

(g) Předpokládejte, že Bill vyjede z domu v 7 hodin a narazí na 6 červených světel a jeden vlak. Odhadněte očekávanou dobu příjezdu do práce.

(h) Ověřte nesplnění předpokladů konstantního rozptylu a normality reziduí.

7. Pojišťovací společnosti se v rámci své obchodní činnosti rozhodují, zda a za jakou výši pojistného pojistí lidi žádající o pojištění. V tomto příkladu se ocitnete v roli pojistného analytika pojišťovny a budete odhadovat vhodnou výši pojistného v závislosti na vybraných faktorech. V datovém souboru *insurance.csv* naleznete potřebná data pro odhad.

Data obsahují: věk *Age*, pohlaví *Sex* (0 - žena, 1 - muž), *BMI*, počet dětí *children*, kouření *Smoker* (0 - ne, 1 - ano), oblast *Region* (1 - severovýchod, 2 - severozápad, 3 - jihovýchod, 4 - jihozápad) a výše pojistného *Charges*.

(a) Sestavte lineární regresní model se všemi proměnnými, dříve než z modelu odstraníte statisticky nevýznamné proměnné proveďte test na ověření homoskedasticity. Vykreslete si provedenou regresi a graf reziduí.

(b) V případě, že je v datech přítomna heteroskedasticita, pokuste se problém vyřešit pomocí metody vážených nejmenších čtverců (WLS).

(c) Pokud se nepodařilo vyřešit problém s nekonstantním rozptylem reziduí, sestavte nový model bez statisticky nevýznamných proměnných z modelu s využitím metody OLS a znovu otestujte model na přítomnost heteroskedasticity. Pokud v novém modelu problém přetrvá, vyzkoušejte opět metodu WLS.

(d) Jestli problém s nekonstantním rozptylem přetrval, vyzkoušejte jiné možnosti vah. V případě, že žádná varianta vah nepovede k odstranění problému heteroskedasticity, odhadněte HAC estimátor.

8. Nejmenovaná americká společnost si zaplatila reklamu na sociálních sítích na podporu prodeje SUV. Na základě nasbíraných dat touto společností prozkoumejte, zda lidé, kteří byli vystaveni reklamě nakoupili v závislosti na několika faktorech SUV. K tomu budete potřebovat datový soubor *Social_Network_Ads.csv*, který obsahuje následující proměnné: ID uživatelů sociálních sítí *User ID*, pohlaví *Gender* (1 - muž, 0 - žena), věk *Age*, odhadovaný roční plat v dolarech *EstimatedSalary*, zakoupil/a SUV *Purchased* (1 - ano, 0 - ne).

(a) Vytvořte lineární pravděpodobností model a logit model pomocí vhodných funkcí.

$$Purchased_i = \beta_0 + \beta_1 \cdot Age_i + \beta_2 \cdot EstimatedSalary_i + \varepsilon_i$$

(b) Spočítejte pravděpodobnost koupě SUV pro následující 3 lidi:

i. Anna, která má 21 let a její roční příjem je 23 000$.

ii. Catherine, která má 42 let, příjem ve výši 130 000$.

iii. Bob ve věku 55 let s příjmem 65 000$.

(c) Na závěr zkuste roztřídit lidi podle věku a pohlaví do příjmové skupiny dle tabulky 3.13 pomocí multinominalního logitu a probitu. Interpretujte získané výsledky.

$$SalaryIdx_i = \beta_0 + \beta_1 \cdot Age_i + \beta_2 \cdot Gender_i + \varepsilon_i$$

Kódování	Zařazení	Roční plat
1	Nízkopříjmová skupina	méně než 40 000$
2	Středněpříjmová skupina	40 000 – 100 000$
3	Vysokopříjmová skupina	více jak 100 000$

Tabulka 3.13: Rozdělení do skupiny dle výše ročního příjmu

9. Existuje nějaký vzorec na základě, kterého by bylo možno určit vítěze ceny Oscara za nejlepší film? Pracujte s datovým souborem *oscar.csv* o nominovaných filmech v kategorii „Nejlepší film“ od roku 1984. V souboru najdete proměnné: rok nominace *year*, název filmu *title*, dummy proměnná indikující Oscara *winner*, celkový počet nominací *nominations*, počet získaných cen Zlatého glóbu *gglobes* a dummy proměnná indikující, zda jde o komedii *comedy*.

   (a) S využitím všech pozorování 1984–2003 odhadněte logit model k predikci vítěze s využitím proměnných *nominations* a *gglobes*. Jaký je mezní efekt dodatečné nominace na pravděpodobnost získání Oscara. Jaký je mezní vliv dodatečné ceny Zlatý globus na pravděpodobnost získání Oscara? Jsou koeficienty statisticky významné?

$$winner_i = \beta_0 + \beta_1 \cdot nominations_i + \beta_2 \cdot gglobes_i + \varepsilon_i$$

   (b) Vypočtěte pravděpodobnosti získání Oscara pro každý rok až do roku 2003. S využitím pravidla, že nejvyšší predikovaná pravděpodobnost je vítěz, spočítejte procento správných predikcí z tohoto modelu.

   (c) S využitím odhadů modelu predikujte vítěze pro rok 2004 (vhodně vyberte vzorek). Skutečným vítězem byl film „Million Dollar Baby“.

10. V tomto neřešeném příkladu prozkoumejte vliv ekonomických zdrojů jednotlivých zemí na počet získaných medailí v OH, pracujte s datovým souborem *olympics.csv*. Přítomné proměnné jsou: rok *year*, HDP

GDP, velikost populace *pop*, počet zlatých/stříbrných/bronzových medailí *gold/silver/bronze*, počet získaných medailí dohromady *medaltot*, dummy proměnná, zda se země účastnila OH jako host *host*, dummy proměnná, jestli se jedná o centrálně plánovanou ekonomiku *planned*, umělá proměnnná, zda byla země součástí bývalého Sovětského svazu *soviet*. Celkový počet medailí v roce 1988 byl 738, v roce 1992 to bylo 815 a v roce 1996 celkem 842 medailí. Proměnná *share* specifikuje podíl získaných medailí. Jedná se o panelová data, ve kterých se nachází mnoho hodnot NA, které nahraďte 0.

(a) S využitím Poissonova regresního modelu zkuste vysvětlit počet získaných medailí *medaltot* jako funkci logaritmů počtu obyvatel a HDP (v dolarech roku 1995). Interpretujte dosažené výsledky. Pro nastavení Poissonova modelu použijte parametr *Distribution*.

$$medaltot_i = \beta_0 + \beta_1 \cdot ln(GDP_i) + \beta_2 \cdot ln(pop_i) + \varepsilon_i$$

(b) V roce 1988 měla Austrálie $HDP = 3 \cdot 10^{11}$ a velikost populace byla 16,5 miliónů obyvatel. Kolik medailí ji predikuje sestavený model? (Ve skutečnosti vyhrála 14 medailí.) Spočítejte pravděpodobnost, že by Austrálie vyhrála 10 medailí a více.

(c) V roce 1988 měla Kanada $HDP = 5,19 \cdot 10^{11}$ a populaci 26,9 miliónů. Kolik medailí by ji predikoval náš model? (Ve skutečnosti získala 10 medailí.) Spočítejte pravděpodobnost, že by Kanada vyhrála 15 medailí a méně.

(d) Využijte data z období 1992–1996 k odhadu modelu vysvětlujícího počet získaných medailí jako funkci logaritmů počtu obyvatel a HDP, porovnejte tyto výsledky s výsledky původního modelu.

(e) Pro období 1992–1996 odhadněte nový Poissonův regresní model s přidáním proměnných *Soviet* a *Host*. Diskutujte získané výsledky. Jsou přidané proměnné statisticky významné?

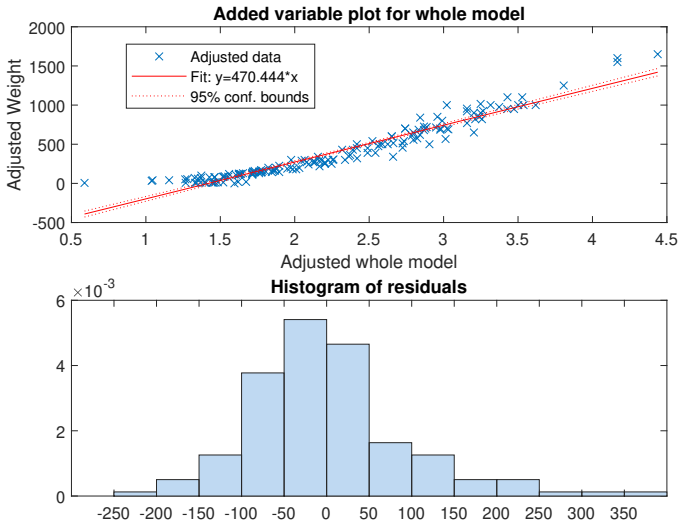
(f) Odhadněte nový rozšířený model, kdy místo proměnné *soviet* použijte *plannes*.

(g) V roce 2000 bylo HDP (v amerických dolarech roku 1995) pro Austrálii vy výši $3,22224 \cdot 10^{11}$ a pro Kanadu $HDP = 6,41256 \cdot 10^{11}$. Počet obyvatel Austrálie byl v roce 2000 celkem 19,071 miliónů a Kanady 30,689 miliónů. S využitím těchto dat predikujte počet medailí pro Kanadu a Austrálii založených na odhadu z modelu předchozí části. Nezapomeňte, že v roce 2000 se olympijské hry konaly v Sydney. (Ve skutečnosti vyhrála Austrálie 58 medailí a Kanada získala 14 medailí.) Jak dobré byly vaše predikce?

3.10 Výsledky neřešených příkladů

1. $d = 0.3576$
   $e = 6.8167 \cdot 10^{-09}$

2. $b =$



Obrázek 3.3: Grafické znázornění regrese a grafu reziduí

$c =$

Kódovaní	Druh	Cena/EURA	Délka	Odhad výdělku
1	Bream	9	38.4	229.2666
2	Roach	8	16.1	45.3232
5	Perch	14.5	50.2	545.7536

Tabulka 3.14: Odhad výdělku Jakuba za chycené ryby

3. $a =$

Linear regression model:
LifeExpect 1 + Status + GDP + BMI + Polio
Estimated Coefficients:

	Estimate	**SE**	**tStat**	**pValue**
(Intercept)	52.694	2.5414	20.735	2.3586e-45
Status_1	-3.9844	1.734	-2.2978	0.022995
GDP	0.00020891	7.2672e-05	2.8746	0.0046502
BMI	0.25661	0.0338	7.5921	3.4407e-12
Polio	0.097723	0.022341	4.3741	2.3062e-05

Tabulka 3.15: Model odhadu délky života v roce 2000

Linear regression model:
LifeExpect 1 + Status + GDP + BMI + Polio
Estimated Coefficients:

	Estimate	**SE**	**tStat**	**pValue**
(Intercept)	61.626	2.2424	27.482	8.6286e-60
Status_1	-6.5163	1.2577	-5.1813	7.1475e-07
GDP	0.00010392	4.2584e-05	2.4403	0.015863
BMI	0.12056	0.023027	5.2359	5.5774e-07
Polio	0.1164	0.019254	6.0454	1.1731e-08

Tabulka 3.16: Model odhadu délky života v roce 2015

4. $a =$

Specifikace	Adjustovaný R^2
Lineární	35,94 %
Kvadratický	62,63 %
Kubický	64,52 %

Tabulka 3.17: Hodnoty adjustovaného koeficientu determinace

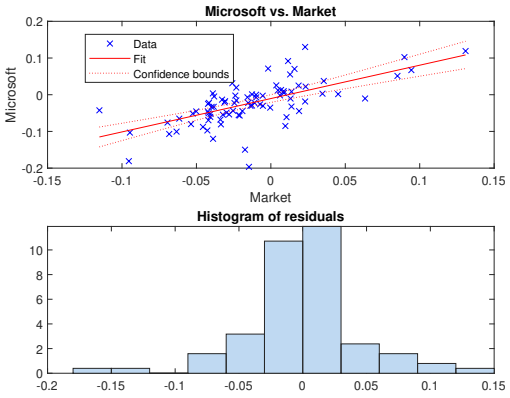
Specifikace	LR_h	LR_pValue	LR_stat	LR_cValue
Kvadratický vs. lineární	0	1	81.8499	176.2938
Kubický vs. lineární	0	0.9999	90.6593	175.1976
Kubický vs. kvadratický	0	1	8.8094	175.1976

Tabulka 3.18: Testování poměrem věrohodností

$b =$

```
  i. 4.4
 ii. 3.1, 3.2, 3.3, 3.4, 3.5, 3.6, 3.7, 3.8, 3.9,
     4.0, 4.1, 4.2, 4.3, 4.4
iii. 2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 2.7, 2.8,
     2.9, 3
 iv. 3
 vi. 8.4712e+05
```

5. $a =$



Obrázek 3.4: Znázornění regrese a reziduí pro společnost Microsoft

6. $b =$

```
17.4439   22.3892
 0.3386    0.3998
 1.0615    1.6091
 2.1562    3.3535
```

$c = 1.4639 \cdot 10^{-18}$
$d = 0.4205$
$e = 1.1428 \cdot 10^{-04}$
$f = 5.4303 \cdot 10^{-30}$

$g =$

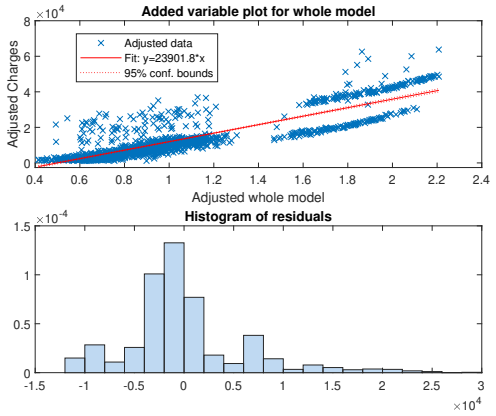
```
predikce =
    41.7602
Ci =
    41.0305   42.4898
```

h =

```
normalita =
    0
rozptyl =
    Heteroscedasticity is absent.
```

7. $a =$

```
rozptyl =
    Heteroscedasticity is present
```



Obrázek 3.5: Znázornění regrese a reziduí plného modelu

8. $a =$

Linear regression model:
Charges 1 + Age + BMI + Smoker + Southeast
Estimated Coefficients:

	Estimate	**SE**	**tStat**	**pValue**
(Intercept)	-11855	960.5	-12.342	3.2305e-33
Age	259.41	11.937	21.732	6.9234e-90
BMI	325.86	27.75	11.742	2.2819e-30
Smoker_1	23837	413.18	57.69	0
Southeast_1	335.65	392.37	0.85545	0.39246

Tabulka 3.19: Výstup modelu bez statisticky nevýznamných proměnných

$b =$

```
  i. 0.0012
 ii. 0.6550
iii. 0.8849
```

$c =$

Logit		**Probit**	
-1.0816	3.5839	-0.6697	2.1422
-0.0047	-0.0667	-0.0026	-0.0396
-0.0442	0.4774	-0.0258	0.2672

Tabulka 3.20: Odhady koeficientů logit a probit modelu

9. $a =$

Odhady logitu	**Statistická významnost**
9.3768	0.000005
-0.6756	0.0004158
-1.1252	0.0005689

Tabulka 3.21: Odhady koeficientů logit a probit modelu

$c =$

Název filmu	**Pravděpodobnost nominace na Oskara**
'Sideways'	0.0230
'Finding Neverla'	0.0095
'Ray'	0.0148
'Million Dollar '	0.0834
'The Aviator'	0.8069

Tabulka 3.22: Predikce vítěze pro rok 2004

10. $a =$

Generalized linear regression model:
log(Medal_Total) 1 + GDP + POP
Distribution = Poisson
Estimated Coefficients:

	Estimate	**SE**	**tStat**	**pValue**
(Intercept)	-16.059	0.16385	-98.008	0
GDP	0.67559	0.0080588	83.832	0
POP	0.042968	0.0096728	4.4421	8.9081e-06

Tabulka 3.23: Poissonův regresní model

$b =$

```
predikce =
    12.2947
pravdepodobnost =
    0.7824
```

$c =$

```
predikce =
    18.2577
pravdepodobnost =
    0.8083
```

$d =$

Generalized linear regression model:
log(Medal_Total) 1 + GDP + POP
Distribution = Poisson
Estimated Coefficients:

	Estimate	**SE**	**tStat**	**pValue**
(Intercept)	-15.573	0.32141	-48.453	0
GDP	0.5467	0.015476	35.327	2.2952e-273
POP	0.20624	0.020835	9.8989	4.2078e-23

395 observations, 392 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 4.06e+03, p-value = 0

Tabulka 3.24: Poissonův regresní model

$e =$

	Estimate	**SE**	**tStat**	**pValue**
(Intercept)	-15.476	0.34482	-44.882	0
GDP	0.57551	0.017109	33.638	4.6457e-248
POP	0.15167	0.022372	6.7797	1.2041e-11
Soviet	2.1042	0.083948	25.065	1.202e-138
Host	0.14788	0.1005	1.4713	0.1412

Tabulka 3.25: Rozšířený Poissonův regresní model

$f =$

	Estimate	**SE**	**tStat**	**pValue**
(Intercept)	-15.161	0.34097	-44.464	0
GDP	0.57499	0.017461	32.931	7.927e-238
POP	0.13815	0.024564	5.624	1.866e-08
Planned	0.6418	0.11881	5.4018	6.5989e-08
Host	0.10596	0.10045	1.0548	0.2915

Tabulka 3.26: Rozšířený Poissonův regresní model

$g =$

```
Australie2000 =
    12.1520
Canada2000 =
    17.3390
```

Příloha

Na závěr přidáváme porovnání modelů prostřednictvím testu Lagrangeových multiplikátorů pomocí funkce *lmtest*. Bohužel neakceptuje vstupy z lineárních regresních modelů (LRM), se kterými jsme doposud v této kapitole pracovali. Musíme k funkci *lmtest* přistoupit cestou *estimate* a *arima*. Funkce *arima* vytvoří model pro časové řady, poté pomocí funkce *estimate* získáme odhad. Více informací o funkcích *arima* a *estimate* najdete ve 4. kapitole.

```
ar0 = arima(0,0,0);
ar0_est = estimate(ar0,Y,'X',X);
b0 = ar0_est.Beta(1);
b1 = ar0_est.Beta(2);
b2 = 0;
b3 = 0;
```

Uložíme si parametry s restrikcemi pro omezený model do proměnné *beta0*. Také si uložíme směrodatnou odchylku odhadu do proměnné *v0* a počet pozorování do proměnné *N*.

```
beta0 = [b0,b1,b2,b3]';
v0 = ar0_est.Variance;
[N,~] = size(fev_data);
```

V dalším kroku vypočítáme gradient, což je vektor 1. parciálních derivací funkce podle jednotlivých proměnných, obecně ve tvaru $(Y - X * beta)/\sigma^2$.

```
G0 = (Y - [ones(N,1),X]*beta0)./v0;
G1 = ((Y - [ones(N,1),X]*beta0).*HEIGHT)./v0;
G2 = ((Y - [ones(N,1),X]*beta0).*GENDER)./v0;
G3 = ((Y - [ones(N,1),X]*beta0).*AGE)./v0;
Gv = -1/(2*v0) + ((Y - [ones(N,1),X]*beta0).^2)./(2*v0^2);
Grad = [G0,G1,G2,G3,Gv];
```

Uložíme si skóre gradientu, což je součet provedených derivací. Do proměnné *LM_EstCov* si vložíme kovarianční matici gradientu. Poté si upřesníme počet restrikcí na parametry a provedeme *lmtest*.

```
LM_score = sum(Grad)';
LM_EstCov = inv(Grad'*Grad);
LM_dof = 2;
[LM_h,LM_pValue,LM_stat,LM_cValue] = lmtest(LM_score,
    LM_EstCov,LM_dof);
```

Výstup v tabulce 3.27 se shoduje s výstupem ručně provedeného testu v části 3.4. Obdrželi jsme $LM\_h = 1$, tudíž byla zamítnuta nulová hypotéza o omezeném modelu *OLSregreseS* ve prospěch neomezeného modelu *OLSregreseE*.

LM_h	LM_pValue	LM_stat	LM_cValue
1	0	653.5420	5.9915

Tabulka 3.27: Test Lagrangeových multiplikátorů - použití funkce *lmtest*