Základy lineární regrese

David Hampel

Ústav statistiky a operačního výzkumu,
Mendelova univerzita v Brně

Kurz pokročilých statistických metod
Global Change Research Centre AS CR, 5.–7. 8. 2015

Tato akce se koná v rámci projektu: Vybudování vědeckého týmu environmentální metabolomiky a ekofyziologie a jeho zapojení do mezinárodních sítí (ENVIMET; r.č. **CZ.1.07/2.3.00/20.0246**) realizovaného v rámci Operačního programu Vzdělávání pro konkurenceschopnost

Obsah

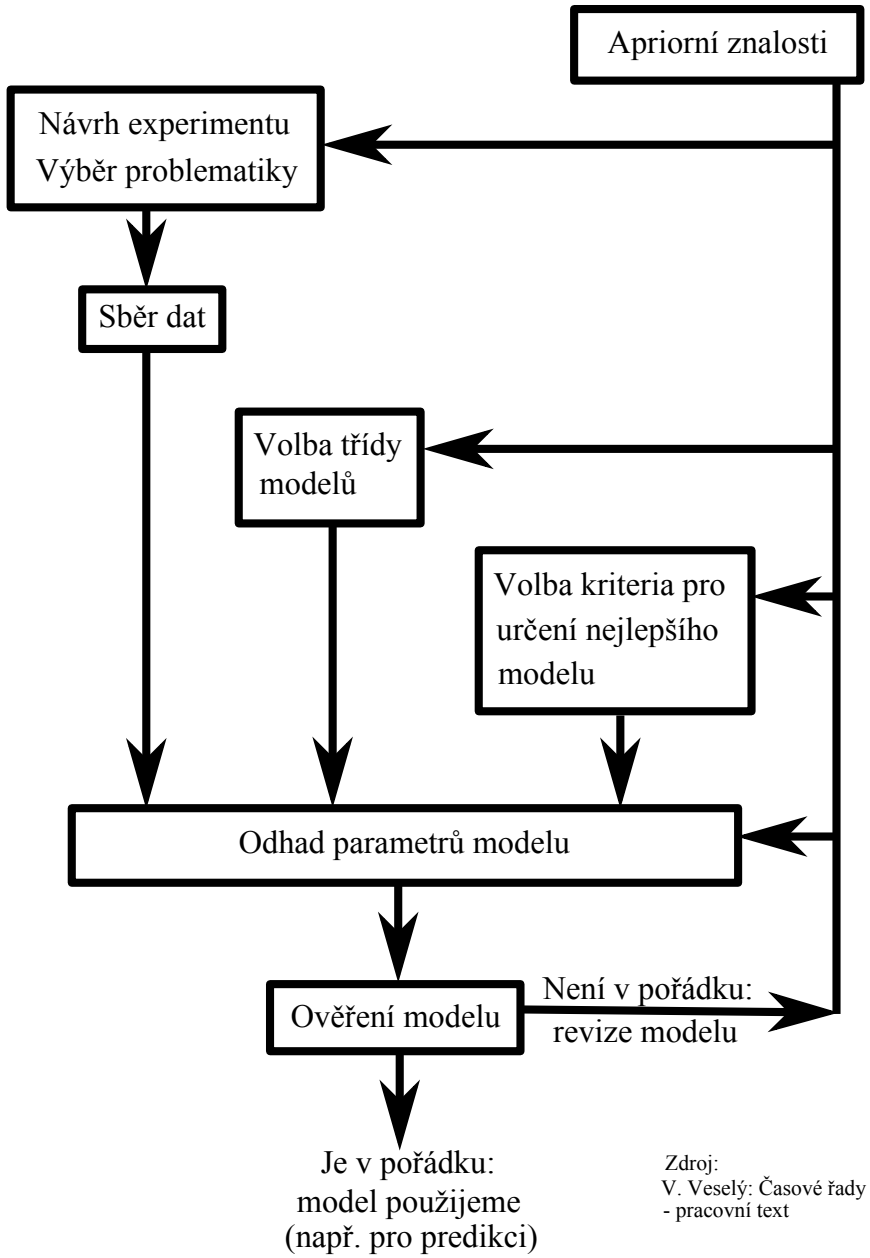
Motivace

- Regrese je jedna z nejčastěji používaných metod.
- Cílem je vysvětlit variabilitu jedné proměnné pomocí jiných proměnných.
- Prakticky můžeme ověřovat zda vybraná proměnná má vliv na hodnoty jiné proměnné, jak silný je tento vliv, můžeme jej i vyčíslit. Dále např. můžeme predikovat proměnnou pomocí odhadnutého modelu.
- Řada praktických úloh lze vyřešit pomocí regrese, ačkoliv tak nemusí být přímo zadána (t-test, faktorová ANOVA, atd.).

Základní kroky v aplikované regresní analýze

1. Přehled literatury a odvození teoretického modelu.
2. Specifikace modelu:
   - výběr nezávisle proměnných;
   - výběr funkční formy.
3. Vyslovení hypotéz o očekávaných znaménćích parametrů.
4. Sběr dat.
5. Odhad modelu a ověření předpokladů.
6. Zdokumentování výsledků.

Logické schéma hledání adekvátního modelu



Zdroj:
V. Veselý: Časové řady - pracovní text

Vysvětlovaná proměnná

- Vysvětlovaná proměnná Y neboli
  - závislá proměnná;
  - regresand.

Vysvětlující proměnná

- Vysvětlující proměnná X neboli
  - nezávislá proměnná;
  - regresor.
- Jako vysvětlující proměnná může vystupovat také
  - zpožděná proměnná;
  - umělá (dummy, indikativní) proměnná.

Zpožděná proměnná

- V některých případech uběhne určitý čas mezi změnou ve vysvětlující proměnné a reakcí ve vysvětlované proměnné – zpoždění.

- Zapisujeme např. $X_{1,t-1}$, kde index $t-1$ značí, že pozorování proměnné X_1 je z předcházejícího období.

Umělá (dummy, indikativní) proměnná

- Proměnná, která obvykle nabývá hodnot nula nebo jedna.
- V některých případech je potřeba do modelu zařadit i slovní znaky, které je nutné převést na kvantitativní znaky.
- Příklad: pohlaví
  - muž $= 1$;
  - žena $= 0$.

Přehled funkčních forem

Volba funkční formy

Rozeznáváme

- modely lineární v parametrech;
- modely nelineární v parametrech, avšak transformovatelné na modely lineární v parametrech (linearizovatelné modely);
- modely nelineární v parametrech.

Polynomiální model s jednou proměnnou

- Polynom stupně 1 (přímka)

$$Y_i = \beta_0 + \beta_1 X_i, \quad i = 1, \ldots, n.$$

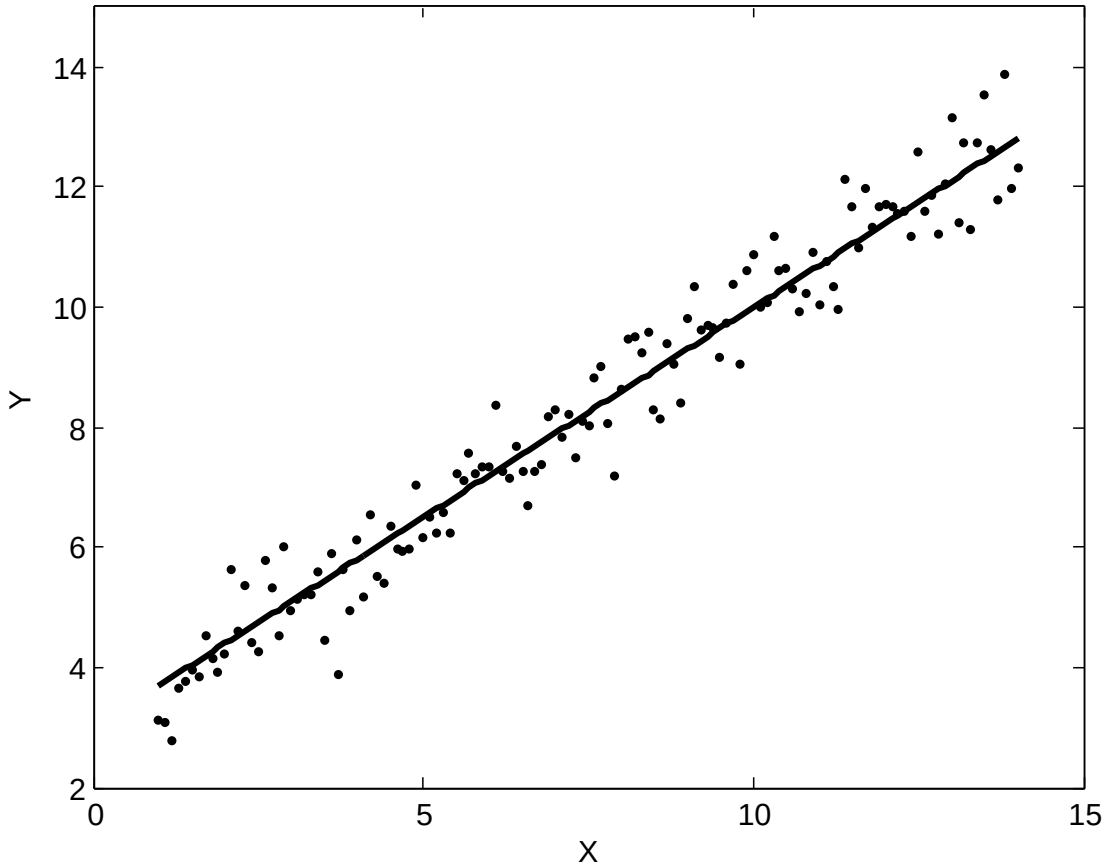
- Polynom stupně 2 (parabola)

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2, \quad i = 1, \ldots, n.$$

- Polynom stupně 3

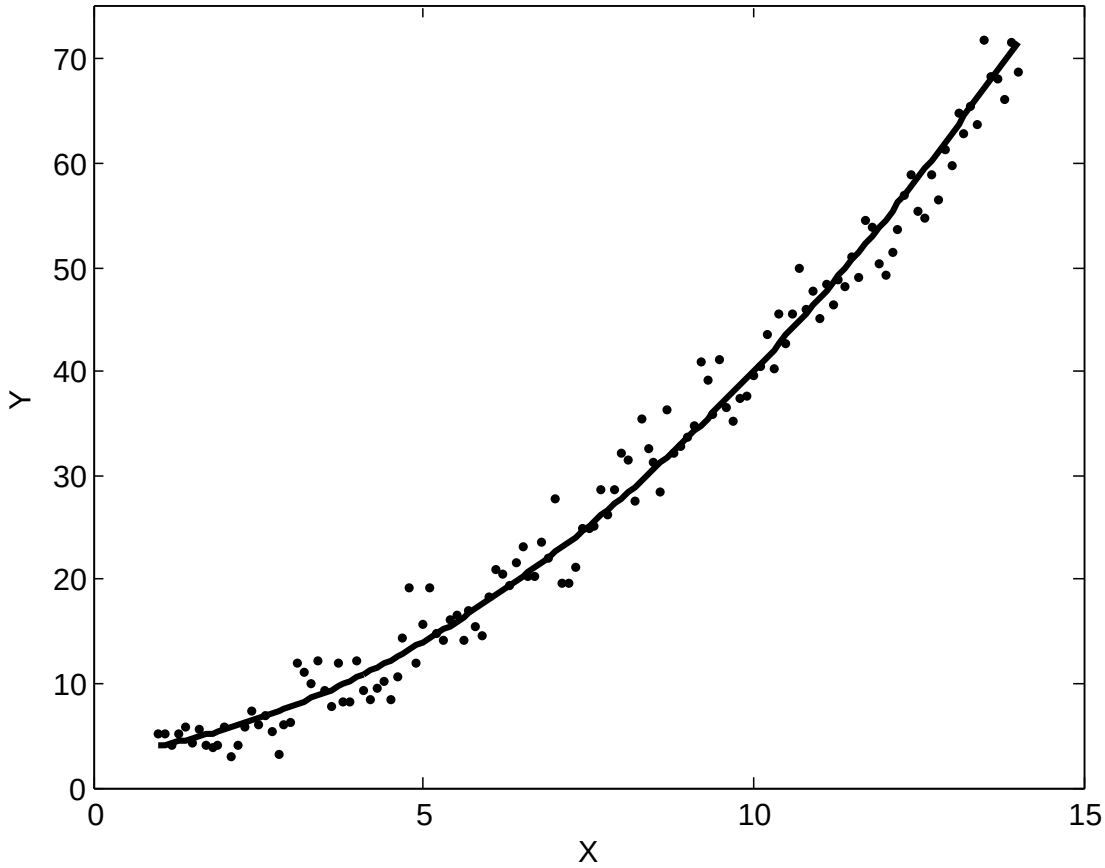
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3, \quad i = 1, \ldots, n.$$

Lineární modely



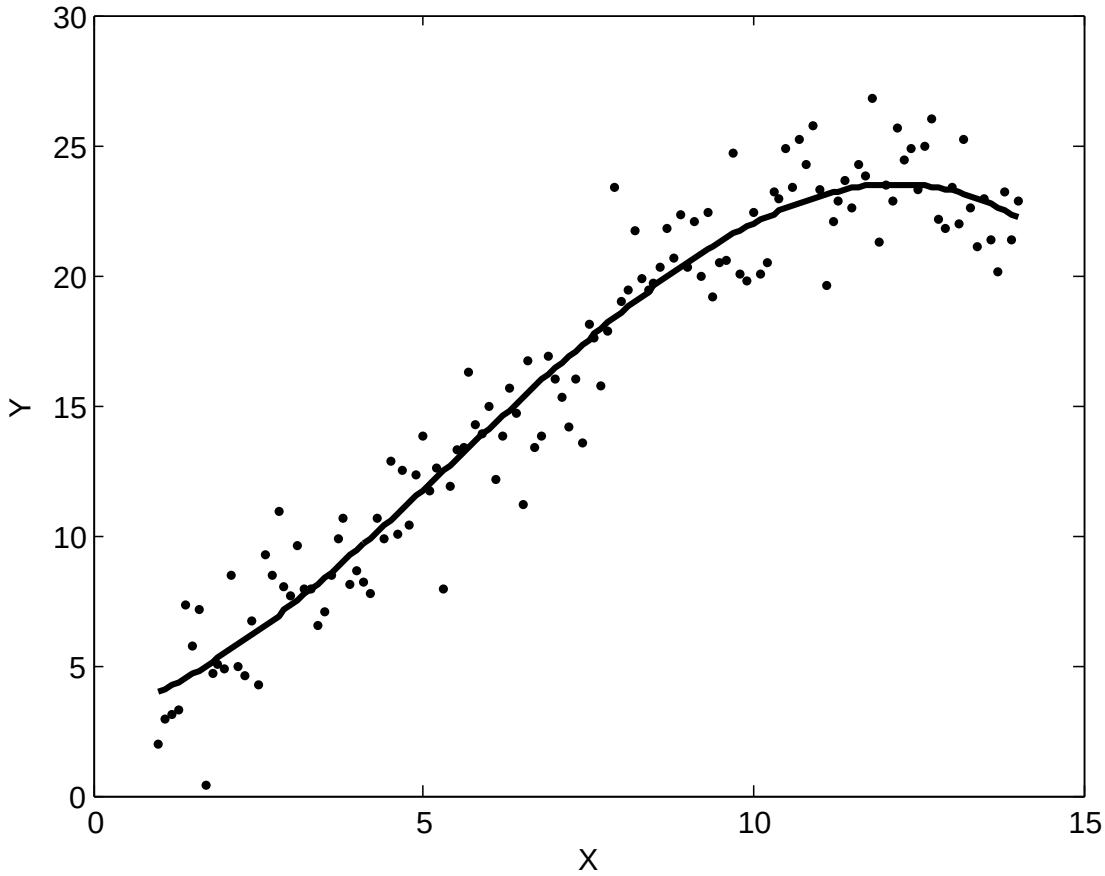
Polynom 1. stupně ($\beta_0 = 3$, $\beta_1 = 0{,}7$).

Lineární modely



Polynom 2. stupně ($\beta_0 = 3$, $\beta_1 = 0{,}7$, $\beta_2 = 0{,}3$).

Lineární modely



Polynom stupně 3 ($\beta_0 = 3$, $\beta_1 = 0{,}7$, $\beta_2 = 0{,}3$, $\beta_3 = -0{,}018$).

Lineární modely

Polynomiální model s více proměnnými

- Vysvětlovaná veličina je modelována pomocí více různých proměnných.
- Obvykle tyto proměnné nejsou umocněny, někdy se používají 2. či 3. mocniny vybraných proměnných.

Polynomiální model s více proměnnými

- Proměnné neumocněny

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{3,i}, \quad i = 1, \ldots, n.$$

- Proměnné neumocněny, jiné značení než výše

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 V_i, \quad i = 1, \ldots, n.$$

- Proměnné v polynomech různých stupňů

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 Z_i + \beta_4 Z_i^2 + \beta_5 Z_i^3 + \beta_6 V_i, \quad i = 1, \ldots, n.$$

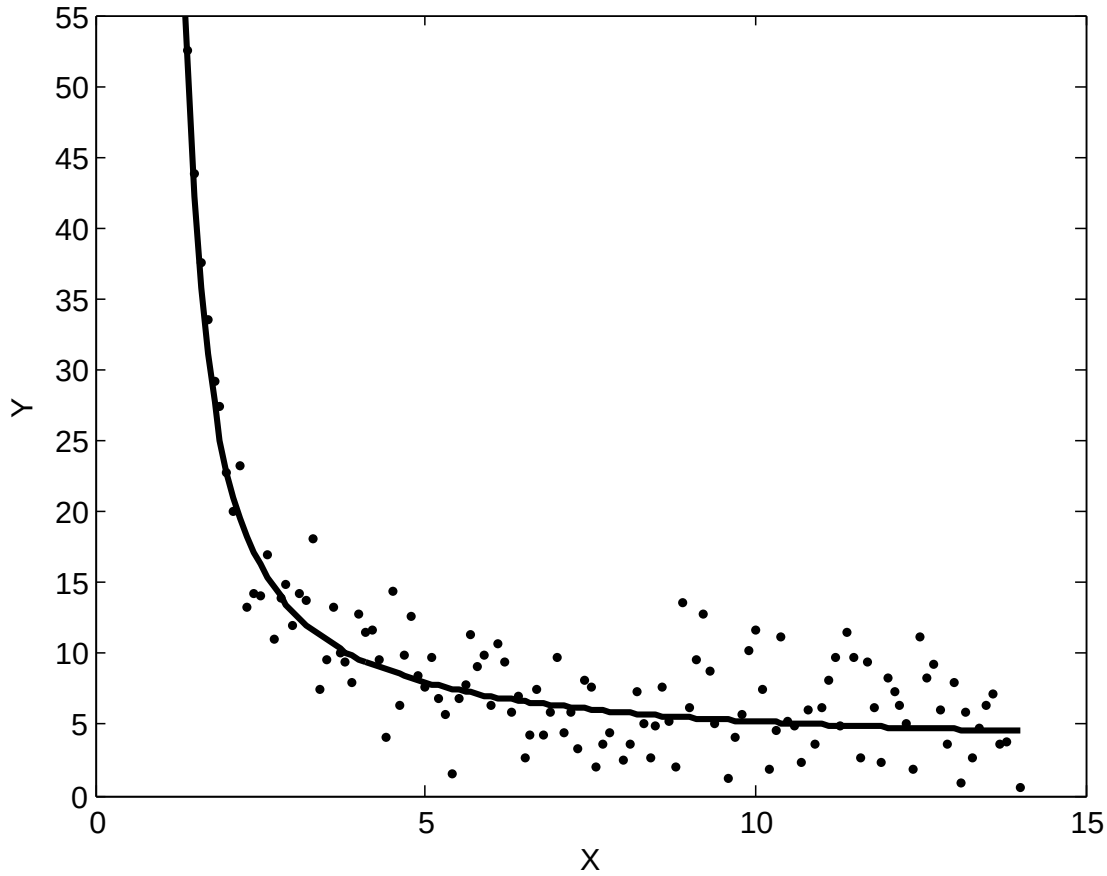
Inverzní (reciproční) model

Jedna či více proměnných modelu je umocněna na -1. Tento model může mimo jiné nabývat následujících forem:

- Právě jedna proměnná, ta umocněna na -1

$$Y_i = \beta_0 + \beta_1 \frac{1}{X_i}, \quad i = 1, \ldots, n.$$

Lineární modely



Inverzní model s jednou proměnnou ($\beta_0 = 3$, $\beta_1 = 20$).

Inverzní (reciproční) model

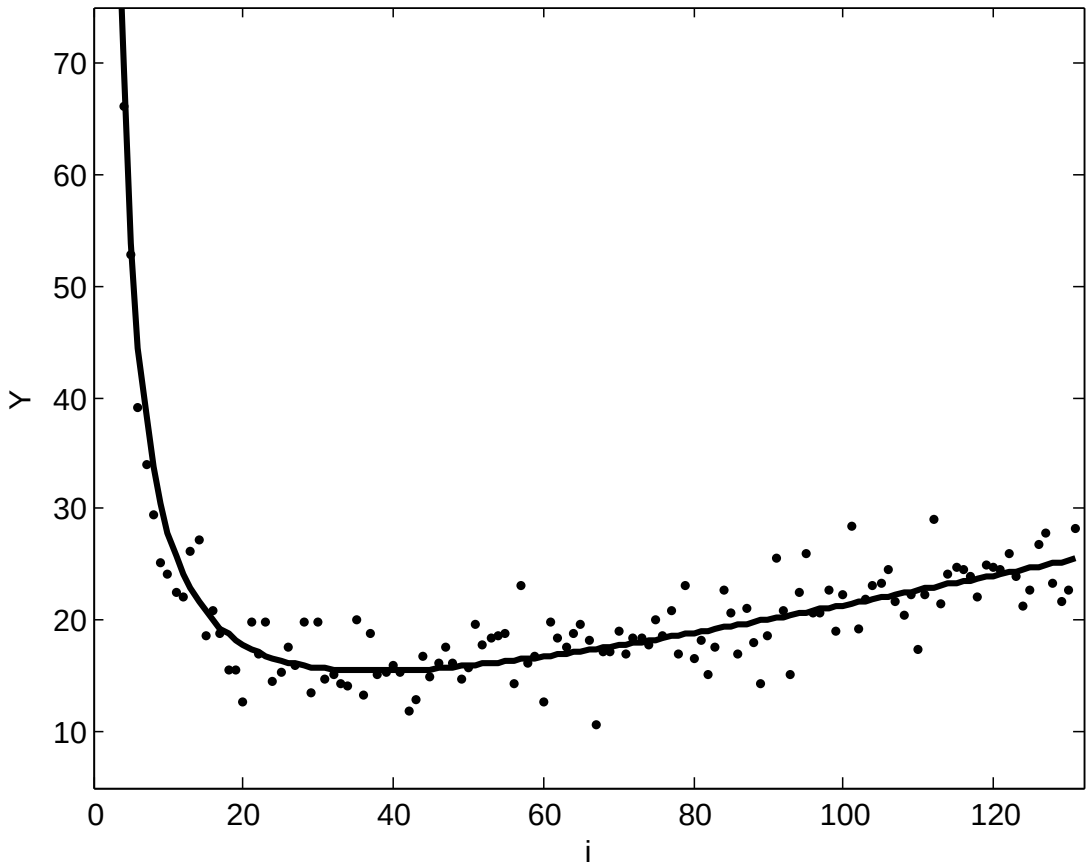
Jedna či více proměnných modelu je umocněna na -1. Tento model může mimo jiné nabývat následujících forem:

- Právě dvě proměnné, z toho jedna proměnná umocněna na -1

$$Y_i = \beta_0 + \beta_1 \frac{1}{X_i} + \beta_2 Z_i, \quad i = 1, \ldots, n,$$

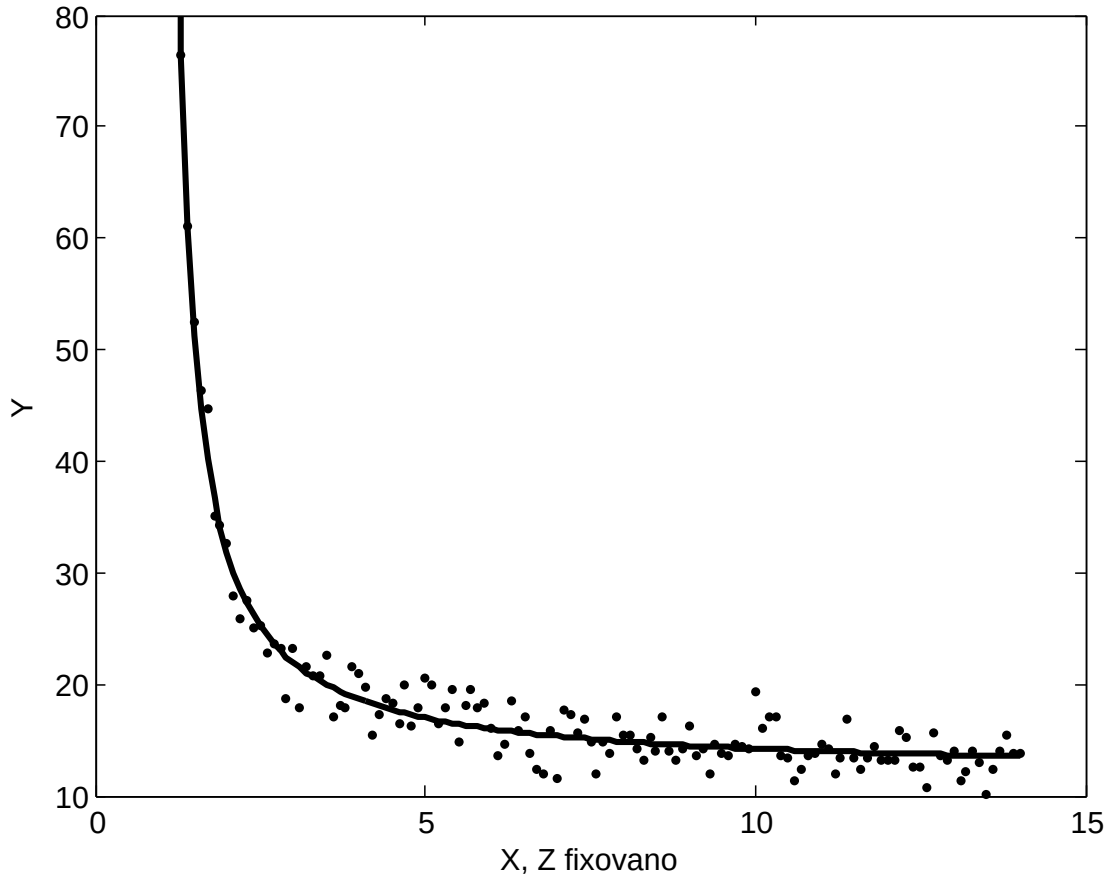
Při znalosti modelu můžeme fixovat jednu z proměnných a vykreslit model pomocí dvojrozměrného bodového grafu.

Lineární modely



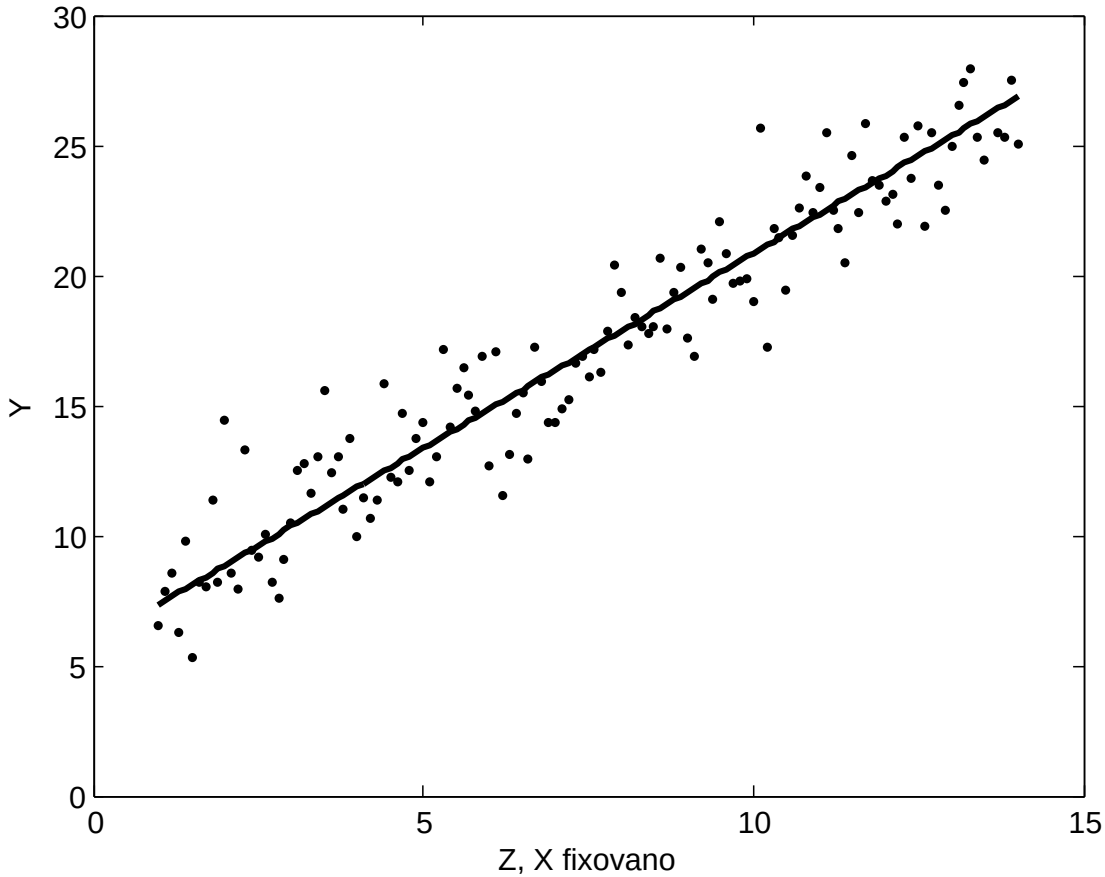
Inverzní model se dvěma proměnnými ($\beta_0 = 3$, $\beta_1 = 20$, $\beta_2 = 0{,}3$).

Lineární modely



Inverzní model se dvěma proměnnými, fixována lineární proměnná ($Z_i = 30$).

Lineární modely



Inverzní model se dvěma proměnnými, fixována inverzní proměnná $(X_i = 7)$.

Inverzní (reciproční) model

Jedna či více proměnných modelu je umocněna na -1. Tento model může mimo jiné nabývat následujících forem:

- Právě tři proměnné, z toho dvě proměnné umocněny na -1

$$Y_i = \beta_0 + \beta_1 \frac{1}{X_i} + \beta_2 \frac{1}{Z_i} + \beta_3 V_i, \quad i = 1, \ldots, n.$$

Linearizovatelné modely

- Některé nelineární modely lze vhodnou transformací převést na modely lineární, linearizovat je. Obvykle se k tomuto účelu používá přirozený logaritmus.
- Odhady parametrů pomocí linearizace jsou většinou horší než pomocí technik pro nelineární modely.

Linearizovatelné modely

Exponenciální (mocninný) model

Jedná se o nelineární model

$$Y_i = \beta_0 X_i^{\beta_1} e^{u_i}, \quad i = 1, \ldots, n,$$

který je možno linearizovat zlogaritmováním:

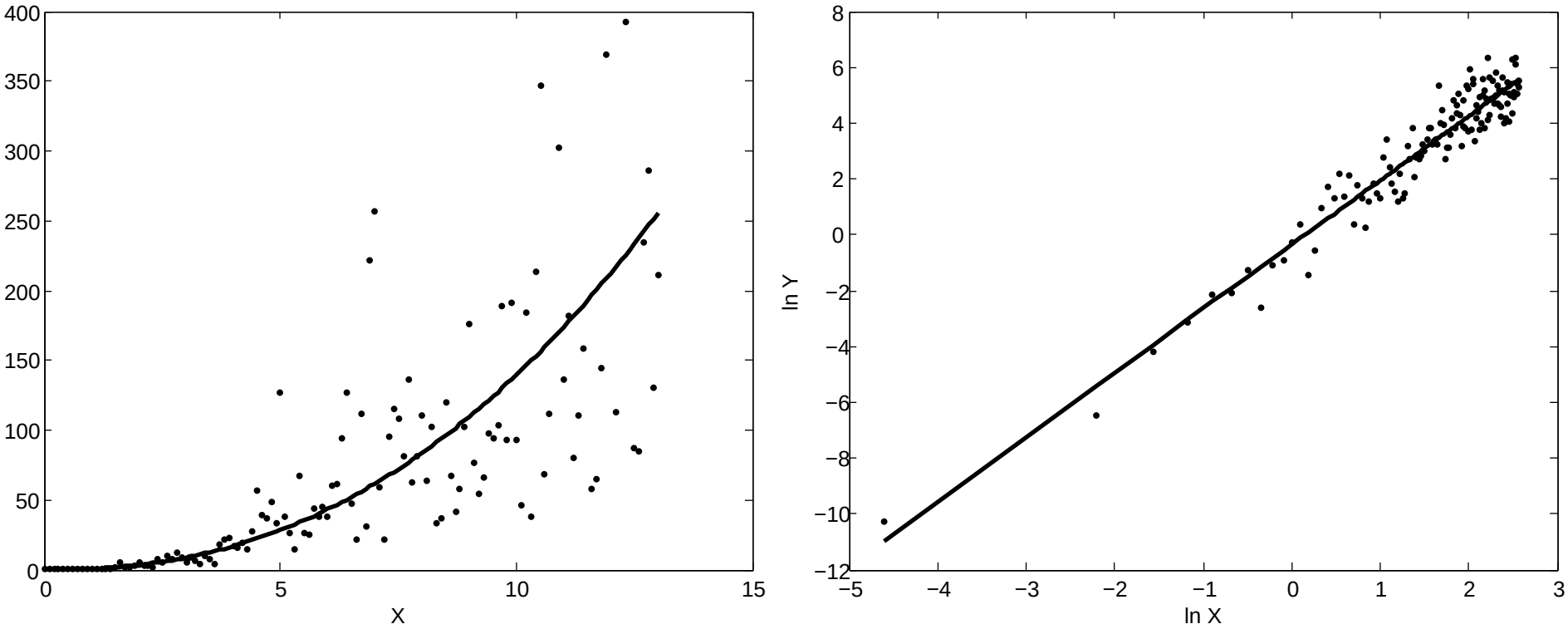
$$\ln Y_i = \ln \beta_0 + \beta_1 \ln X_i + u_i, \quad i = 1, \ldots, n.$$

Po zavedení parametru $\alpha = \ln \beta_0$ dostaneme finální tvar modelu jako

$$\ln Y_i = \alpha + \beta_1 \ln X_i + u_i, \quad i = 1, \ldots, n,$$

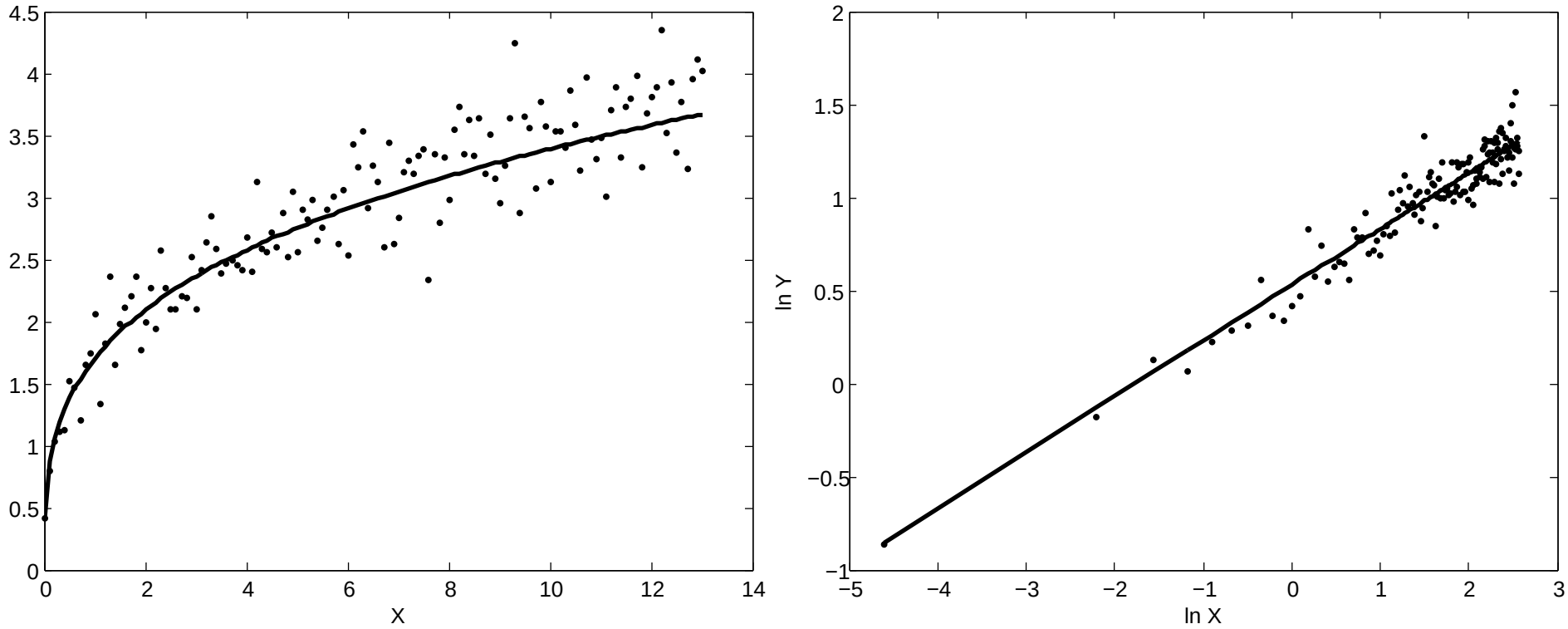
což už je model lineární v parametrech. Zda je tento model vhodný pro analyzovaná data zjistíme vykreslením zlogaritmovaných proměnných pomocí bodového grafu – body by se měly řadit kolem pomyslné přímky.

Linearizovatelné modely



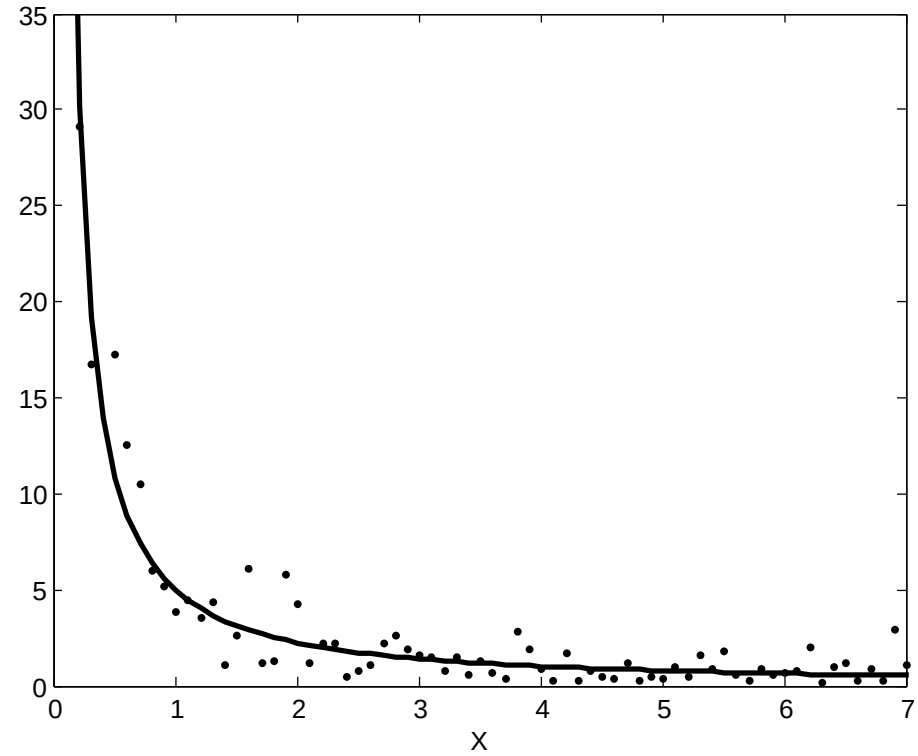
Exponenciální model s parametrem $\beta_1 > 0$ ($\beta_0 = 0{,}7$, $\beta_1 = 2{,}3$). Na levém grafu zobrazen původní model, na pravém linearizovaný model.

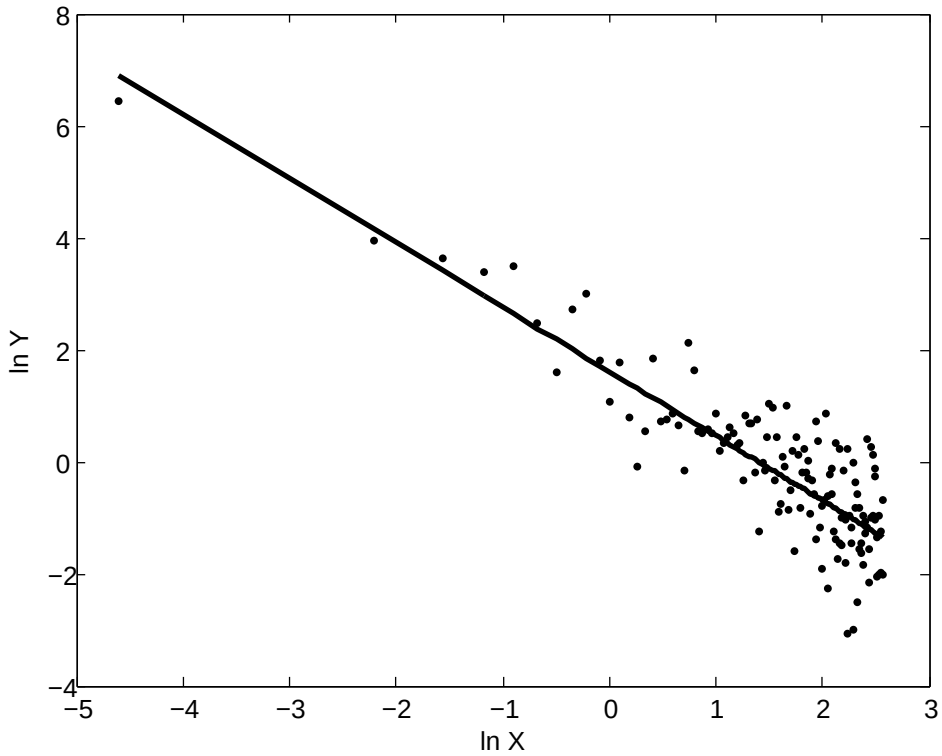
Linearizovatelné modely



Exponenciální model s parametrem $0 < \beta_1 < 1$ ($\beta_0 = 1{,}7$, $\beta_1 = 0{,}3$). Na levém grafu zobrazen původní model, na pravém linearizovaný model.

Linearizovatelné modely





Exponenciální model s parametrem $\beta_1 < 0$ ($\beta_0 = 5$, $\beta_1 = -1{,}15$). Na levém grafu zobrazen původní model, na pravém linearizovaný model.

Linearizovatelné modely

Dvojitě exponenciální (mocninný) model

Jedná se o nelineární model

$$Y_i = e^{\beta_0} X_i^{\beta_1} Z_i^{\beta_2} e^{u_i}, \quad i = 1, \ldots, n,$$

který je opět možno linearizovat zlogaritmováním:

$$\ln Y_i = \beta_0 + \beta_1 \ln X_i + \beta_2 \ln Z_i + u_i, \quad i = 1, \ldots, n,$$

což je model lineární v parametrech. Pokud jednu z proměnných fixujeme, bude průběh analogický k exponenciálnímu modelu.

Linearizovatelné modely

Logaritmický inverzní model

Nelineární model

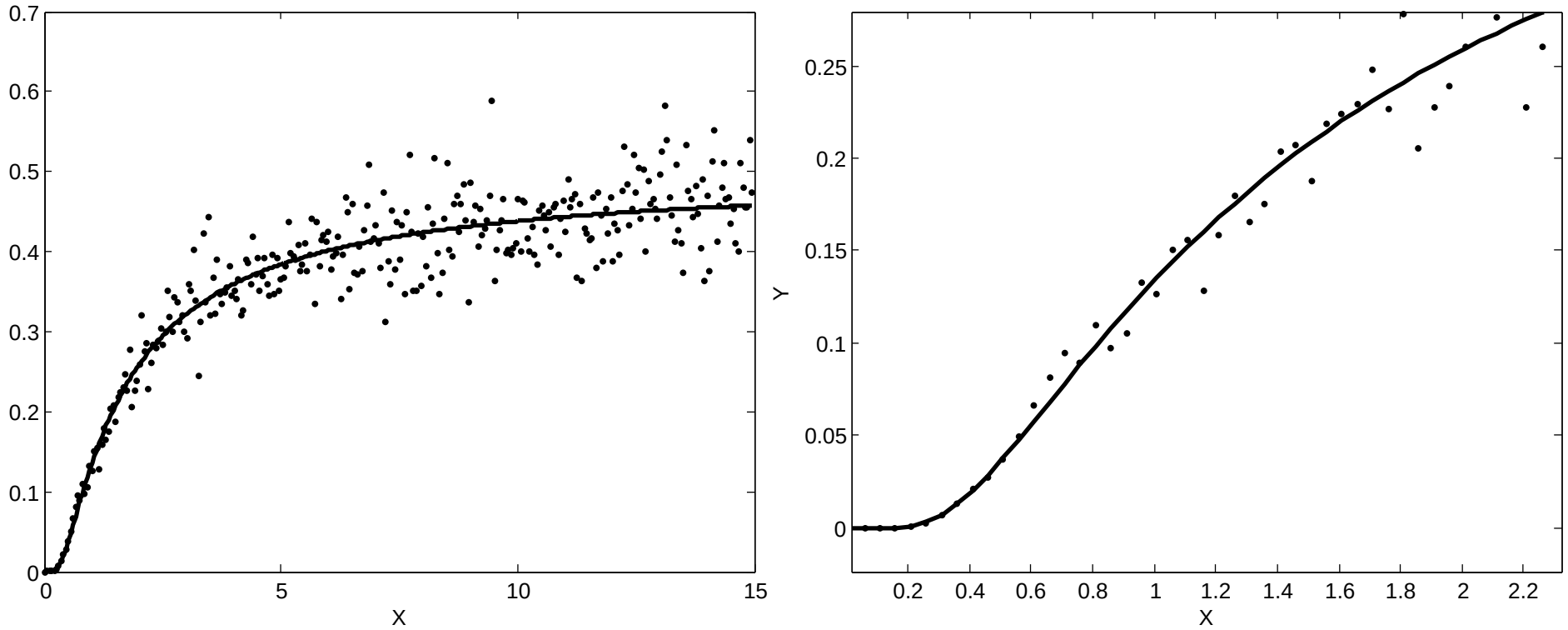
$$Y_i = \alpha \left(e^{\frac{1}{X_i}} \right)^{-\beta_1}, \quad \alpha, \beta_1 > 0, \quad i = 1, \ldots, n,$$

lze logaritmováním převést na model lineární v parametrech

$$\ln Y_i = \beta_0 - \beta_1 \frac{1}{X_i}, \quad i = 1, \ldots, n,$$

kde $\beta_0 = \ln \alpha$.

Linearizovatelné modely



Vlevo logaritmický inverzní model v nelineární formě ($\alpha = 0{,}5$, $\beta_1 = 1{,}32$), vpravo detail.

Nelineární modely

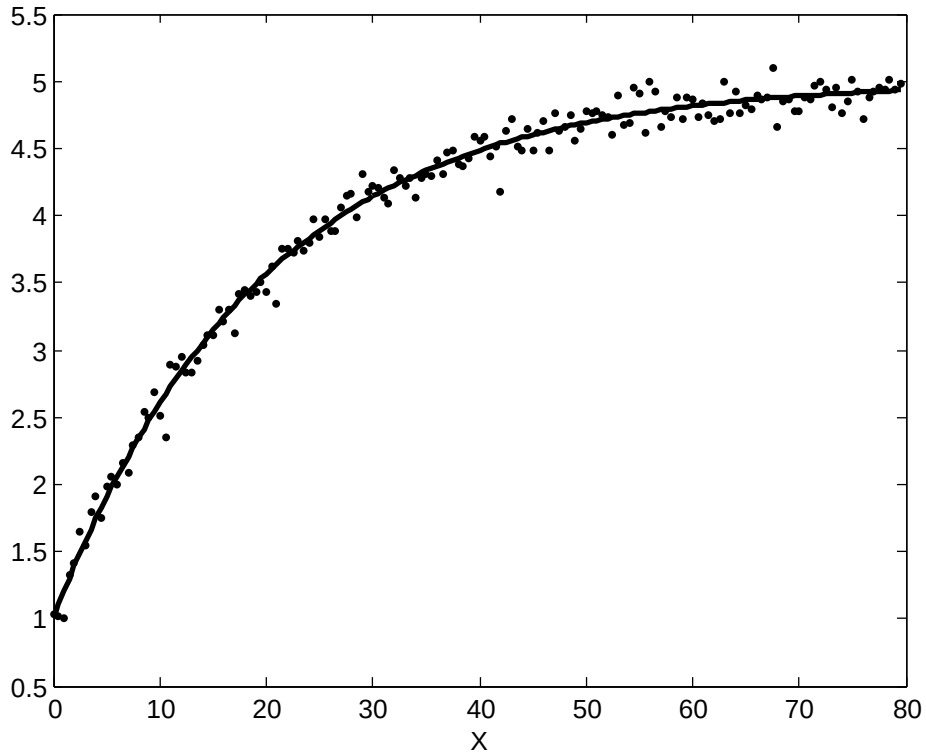
Následující modely nelze linearizovat, pro odhad jejich parametrů je nutno použít speciální postupy či obecné metody odhadů v nelineárních modelech.

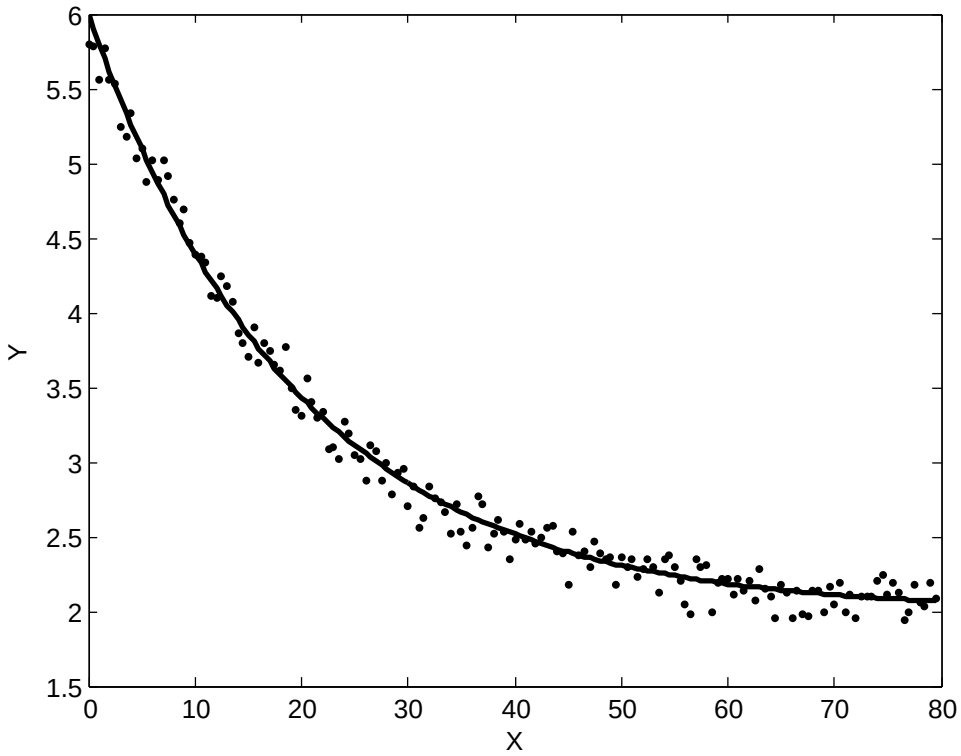
Modifikovaný exponenciální model

$$Y_i = \beta_0 + \beta_1 \beta_2^{x_i}, \quad \beta_2 > 0, \quad i = 1, \ldots, n.$$

Parametr β_0 odpovídá asymptotické úrovni, kam regresní funkce roste ($\beta_1 < 0$), popř. klesá ($\beta_1 > 0$).

Nelineární modely





Modifikovaný exponenciální model. Na levém grafu $\beta_0 = 5$, $\beta_1 = -4$ a $\beta_2 = 0{,}95$; na pravém grafu $\beta_0 = 2$, $\beta_1 = 4$ a $\beta_2 = 0{,}95$.

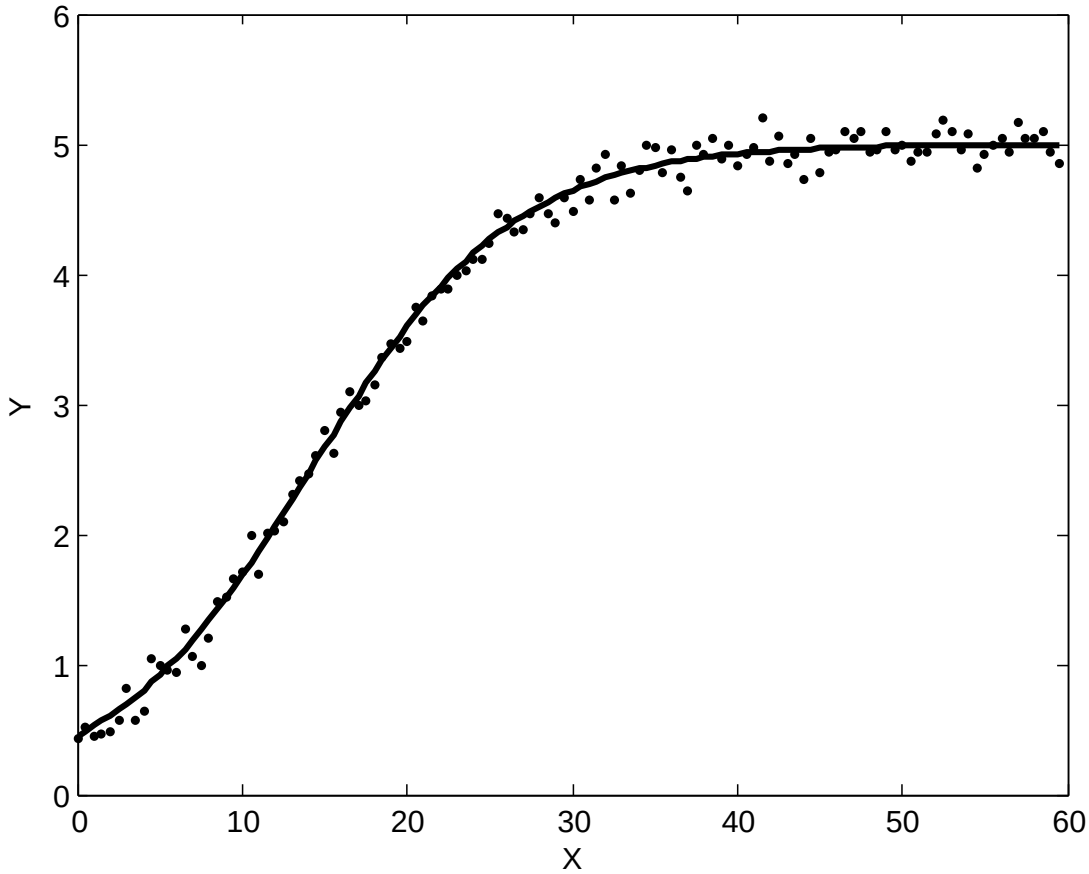
Nelineární modely

Logistický model

Tento model je rozšířením modifikovaného exponenciálního modelu o inflexní bod.

$$Y_i = \frac{\beta_0}{1 + \beta_1 \beta_2^{x_i}}, \quad \beta_0, \beta_2 > 0, \quad i = 1, \ldots, n,$$

Nelineární modely



Logistický model ($\beta_0 = 5$, $\beta_1 = 10$, $\beta_2 = 0{,}85$).

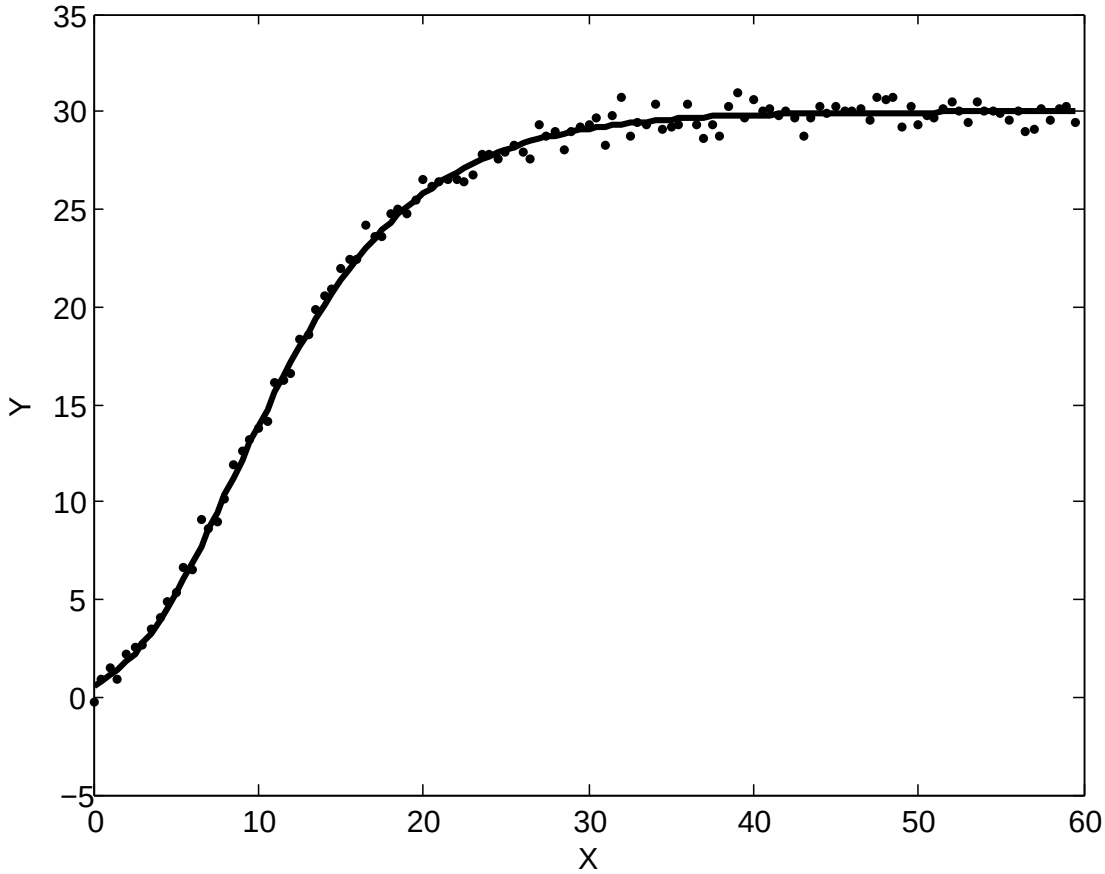
Nelineární modely

Gompertzova křivka

$$Y_i = e^{\beta_0 + \beta_1 \beta_2^{x_i}}, \quad \beta_2 > 0, \quad i = 1, \ldots, n.$$

Podobná jako logistický model, jen inflexní bod je umístěn jiným způsobem.

Nelineární modely



Gompertzova křivka ($\beta_0 = 3{,}4$, $\beta_1 = -3{,}91$, $\beta_2 = 0{,}85$).

Typy datových souborů

1. Průřezová data (viz náhodné výběry)
   - pozorování mnoha jevů v jednom časovém okamžiku.
2. Časové řady
   - pozorování jednoho jevu přes několik časových období.
3. Panelová data
   - pozorování mnoha jevů po několik časových období.

Regresní analýza

Regresní analýza

Regresní analýza je označení statistických metod, pomocí nichž odhadujeme hodnotu jisté náhodné veličiny (tzv. závisle proměnné, vysvětlované proměnné) na základě znalosti jiných veličin (tzv. nezávisle proměnných, vysvětlujících proměnných).

$$Y = f(X_1, X_2, \ldots, X_k)$$

Jednorozměrná regresní analýza

Jednorozměrná regresní analýza je využívána ke studiu vztahu mezi dvěmi proměnnými:

$$Y = f(X).$$

Vícerozměrná regresní analýza

Vícerozměrná regresní analýza je využívána ke studiu vztahu mezi závisle proměnnou a dvěmi či více nezávisle proměnnými:

$$Y = f(X_1, X_2, \ldots, X_k).$$

Odhad parametrů

Odhad parametrů regresního modelu

Snahou je na základě datového souboru a vyrovnávacích kritérií odhadnout parametry $\beta_0, \beta_1, \ldots, \beta_k$ teoretického regresního modelu

$$Y = f(X_1, \ldots, X_k, \beta_0, \beta_1, \ldots, \beta_k) + \epsilon.$$

Odhad parametrů regresního modelu

Pokud do funkčního vztahu dosadíme za parametry $\beta_0, \beta_1, \ldots, \beta_k$ jejich odhady $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$, dostaneme odhad $\hat{Y}$

$$\hat{Y} = f(X_1, \ldots, X_k, \hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)$$

získaný na základě modelu

$$Y = f(X_1, \ldots, X_k, \beta_0, \beta_1, \ldots, \beta_k) + \epsilon.$$

Metoda nejmenších čtverců (OLS)

- OLS je nejrozšířenější metoda odhadu parametrů teoretického modelu.
- Tato metoda je založena na minimalizaci sumy čtverců rozdílů mezi původními hodnotami vysvětlované proměnné a jejími odhadnutými hodnotami

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

Metoda nejmenších čtverců

- Rozdíl mezi empirickou hodnotou a teoretickou hodnotou

$$e_i = Y_i - \hat{Y}_i$$

nazveme **reziduum**.

- OLS je tak založena na minimalizaci sumy čtverců reziduí:

$$\sum_{i=1}^{n} (e_i)^2 \to \min .$$

Odhad jednorozměrného regresního modelu pomocí metody OLS

Naším cílem je minimalizace výrazu Q, tedy

$$Q = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \to \min .$$

Minimalizaci můžeme provést pomocí techniky známé z matematické analýzy – nejedná se o nic jiného než o hledání extrému funkce jedné či více proměnných:

- Vypočteme parciální derivace podle všech proměnných (zde parametrů).
- Parciální derivace položíme rovny 0.
- Z těchto rovnic utvoříme soustavu rovnic, kterou vyřešíme.

Odhad jednorozměrného regresního modelu pomocí metody OLS

Příklad:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, \ldots, n.$$

Odhad jednorozměrného regresního modelu pomocí metody OLS

- Teoretický jednorozměrný regresní model (např. lineární funkční tvar):

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, \ldots, n.$$

- Tento model můžeme odhadnout jako

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, \quad i = 1, \ldots, n.$$

Odhad jednorozměrného regresního modelu pomocí metody OLS

Minimalizujeme sumu čtverců reziduí:

$$Q = \sum_{i=1}^{n} (e_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2.$$

Po dosazení $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$, $i = 1, \ldots, n$ dostaneme

$$Q = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2.$$

Odhad jednorozměrného regresního modelu pomocí metody OLS

Naším cílem je minimalizace výrazu Q, tedy

$$Q = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \to \min .$$

Připomeňme:

- Vypočteme parciální derivace podle všech proměnných (zde parametrů).
- Parciální derivace položíme rovny 0.
- Z těchto rovnic utvoříme soustavu rovnic, kterou vyřešíme.

Odhad jednorozměrného regresního modelu pomocí metody OLS

V našem případě:

$$\frac{\partial Q}{\partial \beta_0} = 2 \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) \cdot (-1)$$

a

$$\frac{\partial Q}{\partial \beta_1} = 2 \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) \cdot (-X_i).$$

Odhad jednorozměrného regresního modelu pomocí metody OLS

V našem případě:

$$2\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)\cdot(-1) = 0$$

$$2\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)\cdot(-X_i) = 0$$

Odhad jednorozměrného regresního modelu pomocí metody OLS

$$\sum_{i=1}^{n} Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i$$

$$\sum_{i=1}^{n} Y_i X_i = \hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2$$

Maticový zápis modelu vícenásobné regrese

Model vícenásobné regrese můžeme vyjádřit jako

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

kde:

$\mathbf{Y}$ – náhodný vektor pozorování závisle proměnné;

$\mathbf{X}$ – matice pozorování nezávisle proměnných typu $(n, k+1)$;

$\boldsymbol{\beta}$ – vektor neznámých parametrů;

$\boldsymbol{\epsilon}$ – vektor náhodných chyb.

Maticový zápis modelu vícenásobné regrese

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{k1} \\ 1 & X_{12} & X_{22} & \cdots & X_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{kn} \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Maticový zápis modelu vícenásobné regrese

- Vektor parametrů regresního modelu lze pak odhadnout pomocí vztahu

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

- Odhad vhodné regresní funkce lze pak zapsat jako

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}.$$

Chybový člen

Chybový člen (náhodná složka)

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, \ldots, n.$$

- Popisuje závislost vysvětlované proměnné na neznámých nebo nepozorovaných proměnných a popisuje i vliv náhody.
- Nelze ji funkčně vyjádřit.

Důvody zapojení chybového členu

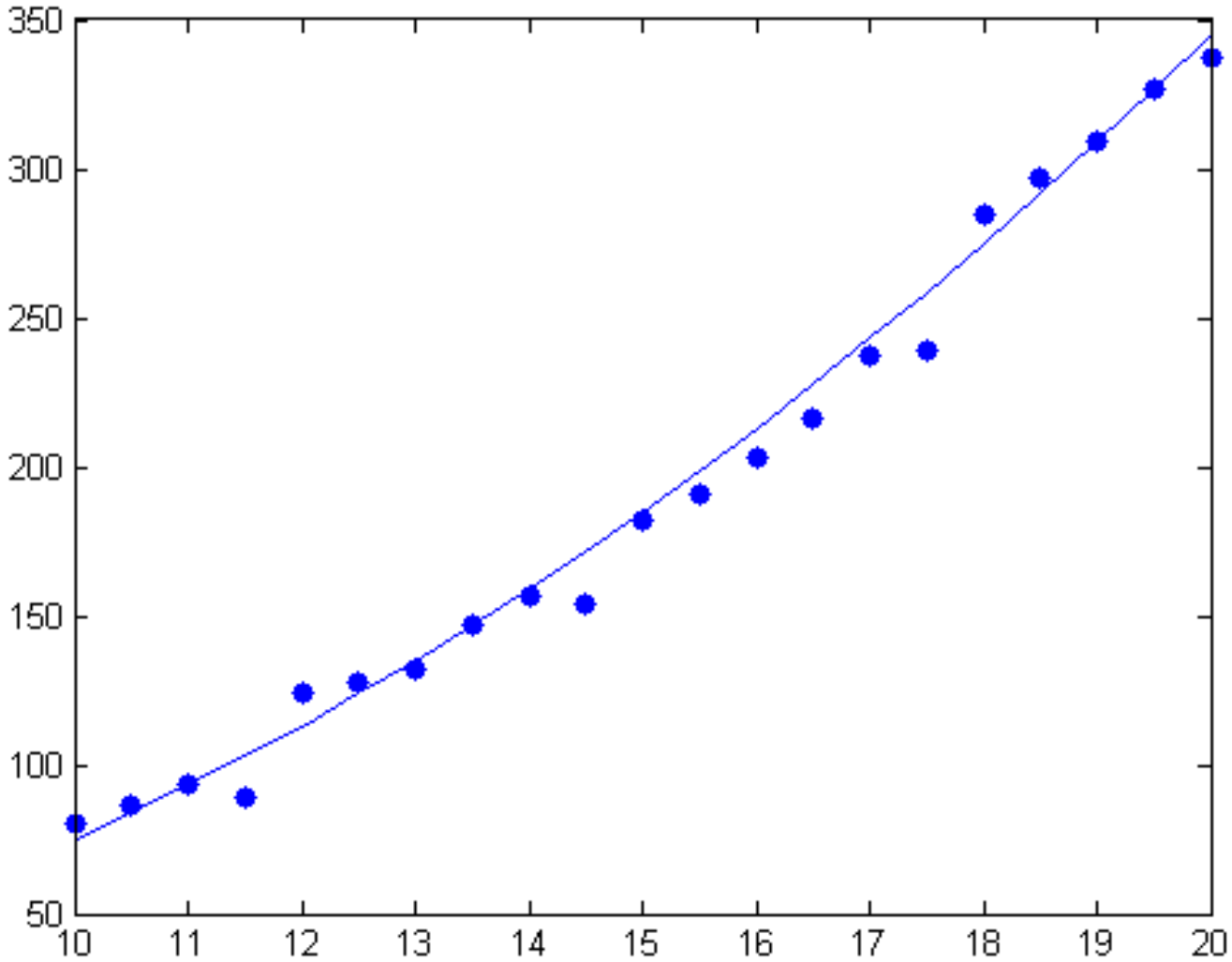
- Opomenuty vlivy nezávisle proměnných na Y v regresní rovnici.
- Chyby měření v modelu.

Využití odhadu chybového členu

- Odhadem náhodné složky ϵ_i, $i = 1, \ldots, n$ konkrétního modelu jsou rezidua e_i, $i = 1, \ldots, n$.
- Analýza reziduí může pomoci ověřit, zda zvolený model je pro data vhodný či ne.
- Graf reziduí může orientačně ukázat zanedbanou funkční formu vysvětlující proměnné.

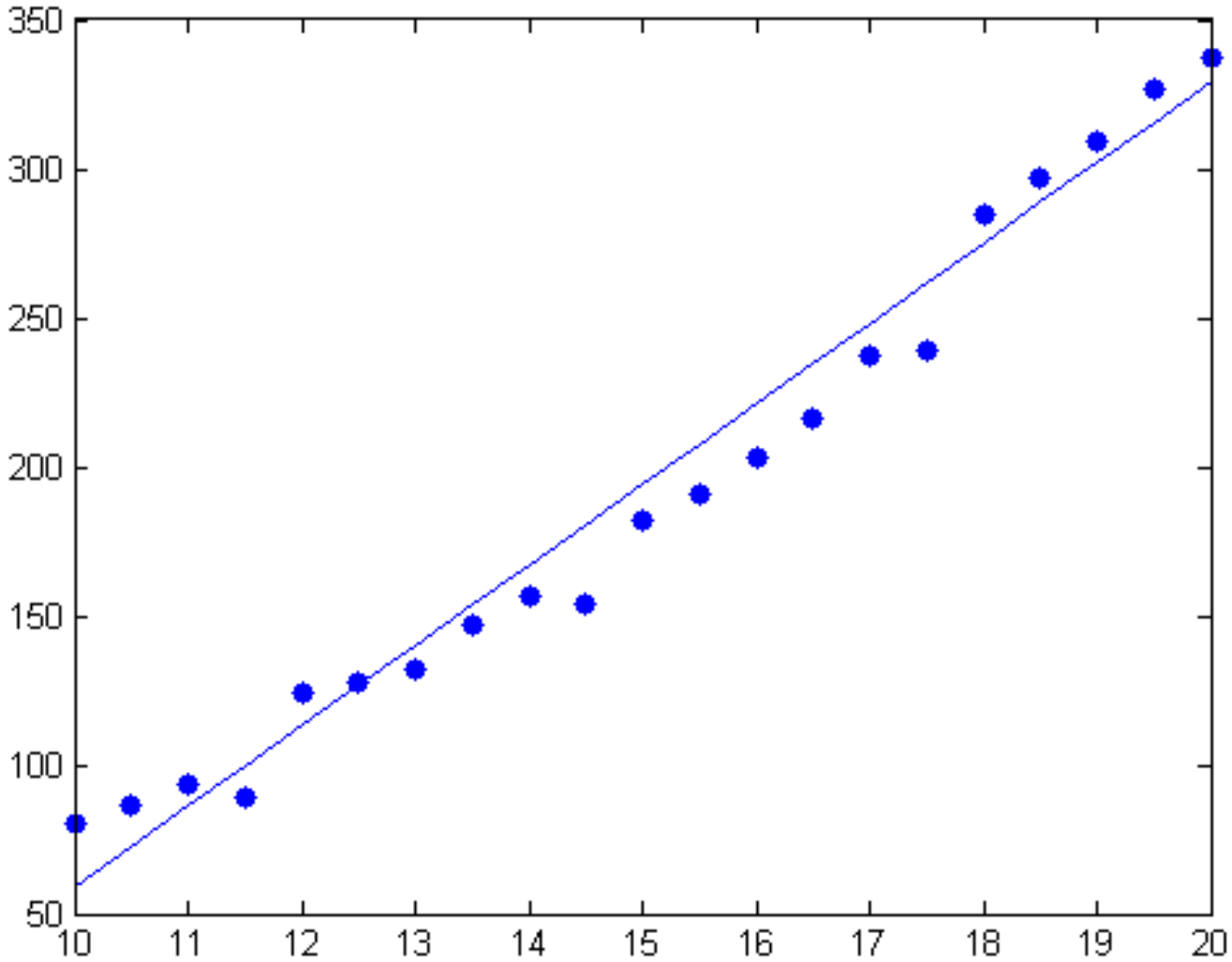
Využití odhadu chybového členu

Data modelovaná adekvátně pomocí polynomu 2. stupně.



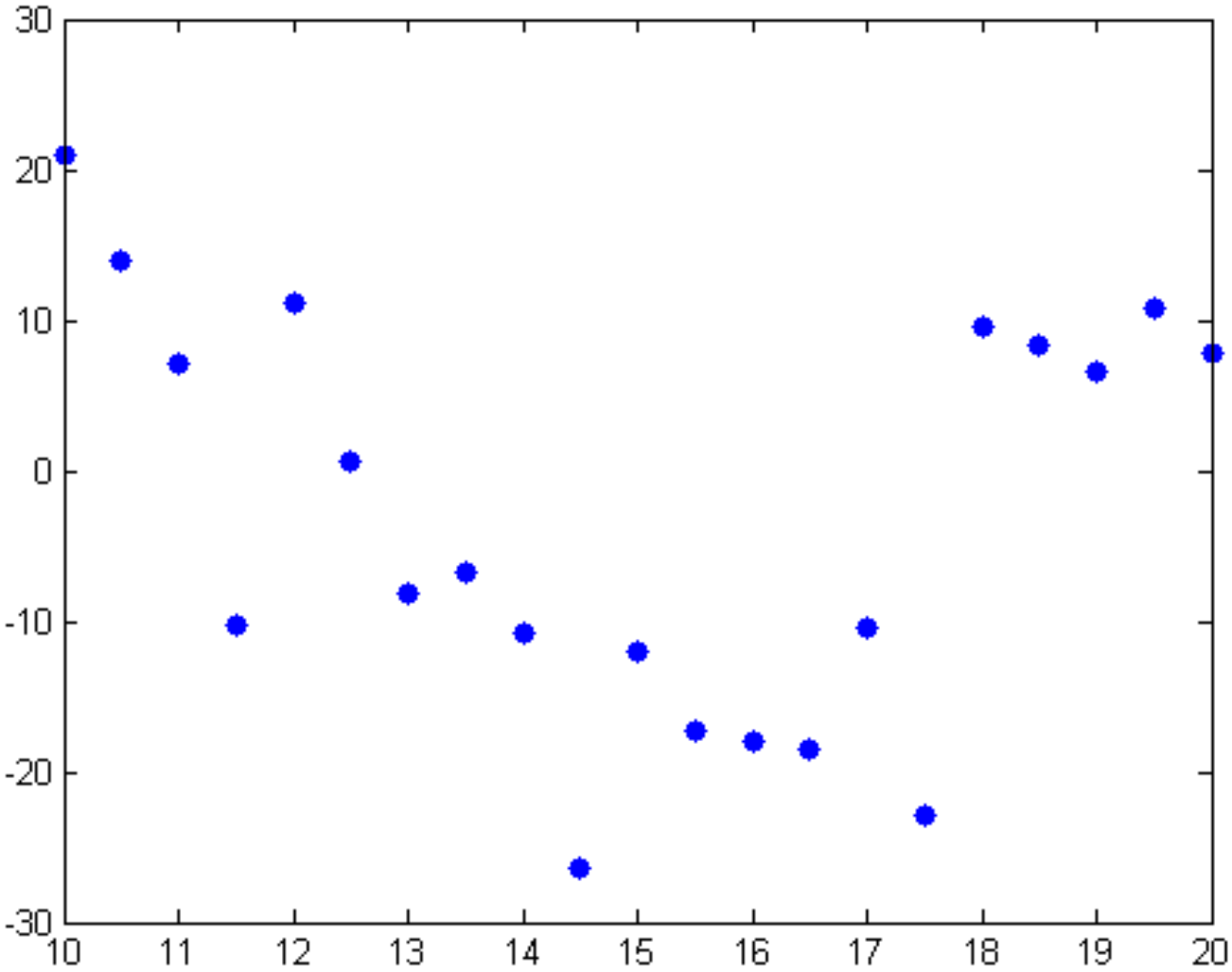
Využití odhadu chybového členu

Data modelovaná pomocí přímky. Rozdíl není na první pohled příliš vidět.



Využití odhadu chybového členu

Graf reziduí modelu s přímkou ukazuje, že rezidua se shlukují do tvaru paraboly, která byla v tomto modelu zanedbána.



Využití odhadu chybového členu

- Graf reziduí testuje vhodnost či dostatečnost modelu pouze orientačně.
- Rezidua by teoreticky měla být náhodně rozložena kolem nulové hodnoty.
- V dalším uvedeme testy parametrů i celého modelu pro exaktnější posouzení.

Dodatečné podmínky

$$E(\boldsymbol{\epsilon}) = \mathbf{0} \qquad \text{var}\,(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$$

$$\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$$

Dodatečné podmínky

Matice $\mathbf{X}$ je plné hodnosti, což znamená,

že matice $\mathbf{X}^T\mathbf{X}$ je regulární.

Testy hypotéz o jednom parametru

Rozdělení pravděpodobnosti v regresním modelu

$$\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$$

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$$

Rozdělení pravděpodobnosti v regresním modelu

- Odhad $\hat{\boldsymbol{\beta}}$ je nestranným odhadem $\boldsymbol{\beta}$.
- Rozptyly odhadu $\hat{\boldsymbol{\beta}}$ jsou na diagonále matice $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$.

Oboustranný interval spolehlivosti pro regresní parametr

$$(\hat{\beta}_i - SE(\hat{\beta}_i)t_{1-\alpha/2}(n-p); \hat{\beta}_i + SE(\hat{\beta}_i)t_{1-\alpha/2}(n-p))$$

Směrodatná chyba odhadu parametru β_i:

$$SE(\hat{\beta}_i) = \sqrt{\frac{RSS}{n-p} h_{i+1,i+1}}$$

Připomeňme: n je počet pozorování, p je počet parametrů modelu.

$$H = (\mathbf{X}^T\mathbf{X})^{-1}$$

Testování hypotéz o jednom regresním parametru

- Tento test se obvykle označuje jako t-test.
- Většinou používán pro testování významnosti regresního parametru v modelu.
- Obecně může být použit k testování hypotézy, že je skutečný regresní parametr roven konkrétnímu reálnému číslu.

Testování hypotéz o jednom regresním parametru

$$T_i = \frac{\hat{\beta}_i - c}{SE(\hat{\beta}_i)} \quad i = 0, \ldots, p$$

$$SE(\hat{\beta}_i) = \sqrt{\frac{RSS}{n-p} h_{i+1,i+1}}$$

$$\mathbf{H} = (\mathbf{X}^T \mathbf{X})^{-1}$$

Testování hypotéz o jednom regresním parametru

- Oboustranná alternativa:

$$H_0 : \beta_i = c \qquad H_1 : \beta_i \neq c$$

$$W = (-\infty; -t_{1-\alpha/2}(n-p)\rangle \cup \langle t_{1-\alpha/2}(n-p); \infty)$$

- Levostranná alternativa:

$$H_0 : \beta_i = c \qquad H_1 : \beta_i < c$$

$$W = (-\infty; -t_{1-\alpha}(n-p)\rangle$$

- Pravostranná alternativa:

$$H_0 : \beta_i = c \qquad H_1 : \beta_i > c$$

$$W = \langle t_{1-\alpha}(n-p); \infty)$$

Testování hypotéz o jednom regresním parametru

- T-test netestuje věcnou správnost zařazení vysvětlující proměnné do modelu.
- T-test neidentifikuje věcnou důležitost vysvětlující proměnné.

Testování hypotéz o jednom regresním parametru

Test významnosti:

- oboustranný test s $c = 0$;
- testuje, zda by odpovídající proměnná v modelu být měla, či nikoliv (nelze ovšem používat samostatně ani automaticky!);
- testová statistika se zjednoduší:

$$T_i = \frac{\hat{\beta}_i - c}{SE(\hat{\beta}_i)} \quad \rightarrow \quad T_i = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}.$$

Analýza rozptylu

Analýza rozptylu

Celkový součet čtverců

$$TSS = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2$$

Regresní součet čtverců

$$ESS = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2$$

Reziduální součet čtverců

$$RSS = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

Analýza rozptylu

	SS	Stupně volnosti	MS	F
Regrese	ESS	$p-1$	$\frac{ESS}{p-1}$	$\frac{ESS/(p-1)}{RSS/(n-p)}$
Rezidua	RSS	$n-p$	$\frac{RSS}{n-p}$	
Celkem	TSS	$n-1$		

SS – suma čtverců

MS – průměrný čtverec

p – počet regresních parametrů

n – počet pozorování

Analýza rozptylu

$$F = \frac{ESS/(p-1)}{RSS/(n-p)} \sim F(p-1, n-p)$$

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

$$H_1 : \mathrm{non} H_0$$

$$W = \langle F_{1-\alpha}(p-1, n-p); \infty)$$

Analýza rozptylu

	SS	Stupně volnosti	MS	p-hodnota
Regrese	ESS	$p-1$	$\frac{ESS}{p-1}$	p-hodnota
Rezidua	RSS	$n-p$	$\frac{RSS}{n-p}$	
Celkem	TSS	$n-1$		

p-hodnota $< \alpha$: H_0 zamítám

p-hodnota $> \alpha$: H_0 nemůžu zamítnout

Analýza rozptylu

Rozklad čtverců:

$$\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2 + \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

$$TSS = ESS + RSS$$

Vztah mezi stupni volnosti:

$$DF_{cel} = DF_{reg} + DF_{rez}$$

$$n - 1 = (p - 1) + (n - p)$$

Příklad

V tabulce jsou uvedeny zisky [tis. Kč] společnosti nabízející venkovní bazény v letech 1991 až 2009, dále kalendářní rok, maximální roční teplota [°C] a celková doba [hod] reklamy v televizi.

Lze z těchto dat usoudit, že zisky jsou ovlivňovány uvedenými prediktory?

(Převzato z Budíková, Králová, Maroš: Průvodce základními statistickými metodami, Grada 2010; upraveno a doplněno.)

Příklad

Zisk	Rok	Teplota	Reklama	Zisk	Rok	Teplota	Reklama
17504	1991	34,9	0,0	18287	2001	35,3	6,5
18971	1992	35,8	3,6	20111	2002	34,8	6,5
20121	1993	35,0	6,8	17890	2003	36,8	7,2
19420	1994	36,0	5,9	19864	2004	32,5	7,0
19563	1995	35,0	6,1	21408	2005	36,4	6,0
19996	1996	34,8	6,7	18244	2006	35,3	6,5
19635	1997	35,1	6,2	19884	2007	37,3	7,5
20219	1998	36,1	7,0	19910	2008	32,5	6,6
20287	1999	34,8	7,0	20588	2009	34,1	6,2
20111	2000	35,6	6,8				

Příklad

Zvolíme $\alpha = 0{,}05$. Počet úrovní je $n = 19$.

Počet regresorů je $k = 3$ a jsou to rok, teplota, reklama (včetně konstanty máme $p = 4$ parametry).

Označíme:

- $Y =$ zisk,
- $x_{1,1} = 1991, \ldots, x_{19,1} = 2009$,
- $x_{1,2} = 34{,}9, \ldots, x_{19,2} = 34{,}1$,
- $x_{1,3} = 0{,}0, \ldots, x_{19,3} = 6{,}2$.

Místo $x_{1,1} = 1991, \ldots, x_{19,1} = 2009$ můžeme také položit $x_{1,1} = 1, \ldots, x_{19,1} = 19$, čímž se pouze změní odhad konstanty β_0.

Příklad

Odhadneme parametry modelu

$$Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \epsilon_i \quad \text{pro } i = 1, \ldots, 19.$$

Obdržíme:

$$\text{zisk} = -24255{,}9 - 1{,}18\ \text{rok} - 115{,}79\ \text{teplota} + 287{,}88\ \text{reklama}$$

	Koeficient	Směr. chyba	t-podíl	p-hodnota
const	24255,9	95845,0	0,2531	0,8036
rok	−1,18190	47,3632	−0,0250	0,9804
teplota	−115,791	185,261	−0,6250	0,5414
reklama	287,875	155,833	1,8473	0,0845

Příklad

	DF	SS	MS	F	p-hodnota
Regrese	3	$4{,}46084 \cdot 10^6$	$1{,}48695 \cdot 10^6$	1,69756	0,2102
Rezidua	15	$1{,}3139 \cdot 10^7$	875935		
Celek	18	$1{,}75999 \cdot 10^7$			

$$W = \langle F_{1-\alpha}(p-1, n-p); \infty) = \langle F_{0,95}(3, 15); \infty) = \langle 3{,}2874; \infty)$$

$F \notin W \Rightarrow$ Nevýznamnost modelu nezamítám.

p-hodnota $> \alpha \Rightarrow$ Nevýznamnost modelu nezamítám.

Příklad – párový koeficient korelace

zisk	rok	teplota	reklama	
1,0000	0,2775	−0,1269	0,4831	zisk
	1,0000	−0,1690	0,5298	rok
		1,0000	0,0304	teplota
			1,0000	reklama

Příklad

$$\text{zisk} = 21869{,}5 \; -114{,}781 \text{ teplota} \; +285{,}763 \text{ reklama}$$

	Koeficient	Směr. chyba	t-podíl	p-hodnota
const	21869,5	6183,98	3,5365	0,0027
teplota	−114,781	175,052	−0,6557	0,5213
reklama	285,763	126,708	2,2553	0,0385

p-hodnota F-testu: 0,096510

Příklad

$$\text{zisk} = 17848{,}9 + 283{,}236 \text{ reklama}$$

	Koeficient	Směr. chyba	t-podíl	p-hodnota
const	17848,9	787,785	22,6571	0,0000
reklama	283,236	124,508	2,2748	0,0362

p-hodnota F-testu: 0,036155

Ověření kvality modelu

Koeficient determinace

Dalším kriteriem pro posouzení kvality modelu může být **koeficient determinace**.

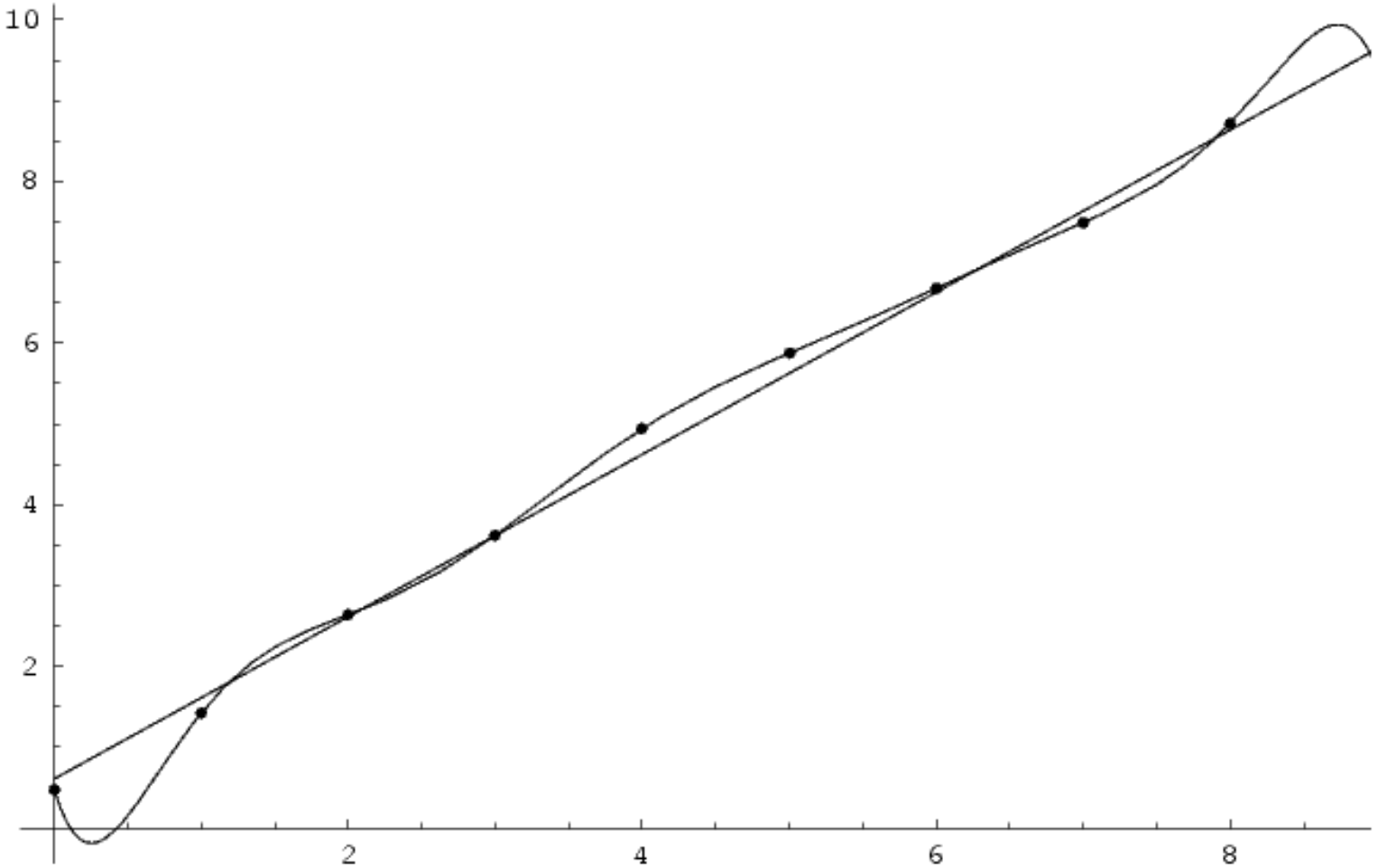
- Informace koncentrovaná do jednoho čísla.
- Může nabývat hodnot od 0 do 1.
  - Hodnoty blízké 1: model vyhovuje.
  - Hodnoty blízké 0: model nevyhovuje.
- Jednoduchý výpočet.
- Nezahrnuje počet vysvětlujících proměnných, s přidáním další vysvětlující proměnné roste – to není vždy žádoucí.

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

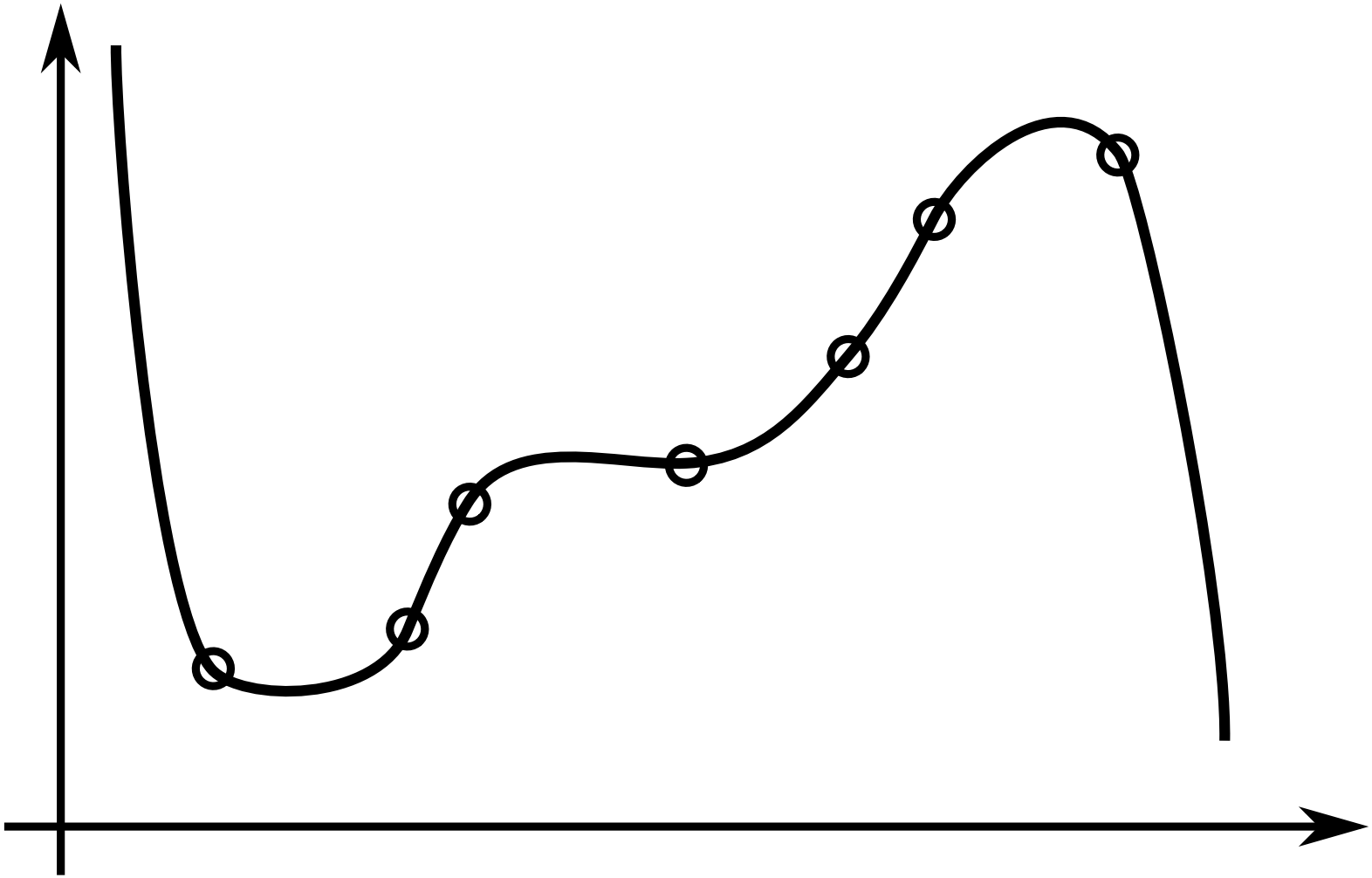
Přeparametrizování modelu

- Přidáním dalších a dalších vysvětlujících proměnných se zlepší některá kriteria hodnotící model (např. index determinace).
- Příliš mnoho vysvětlujících proměnných ale může být na škodu:
  - Numerické problémy.
  - Neinterpretovatelnost.
  - Nereálná znaménka u odhadů parametrů.
  - Nepoužitelnost v praxi (např. pro predikce).

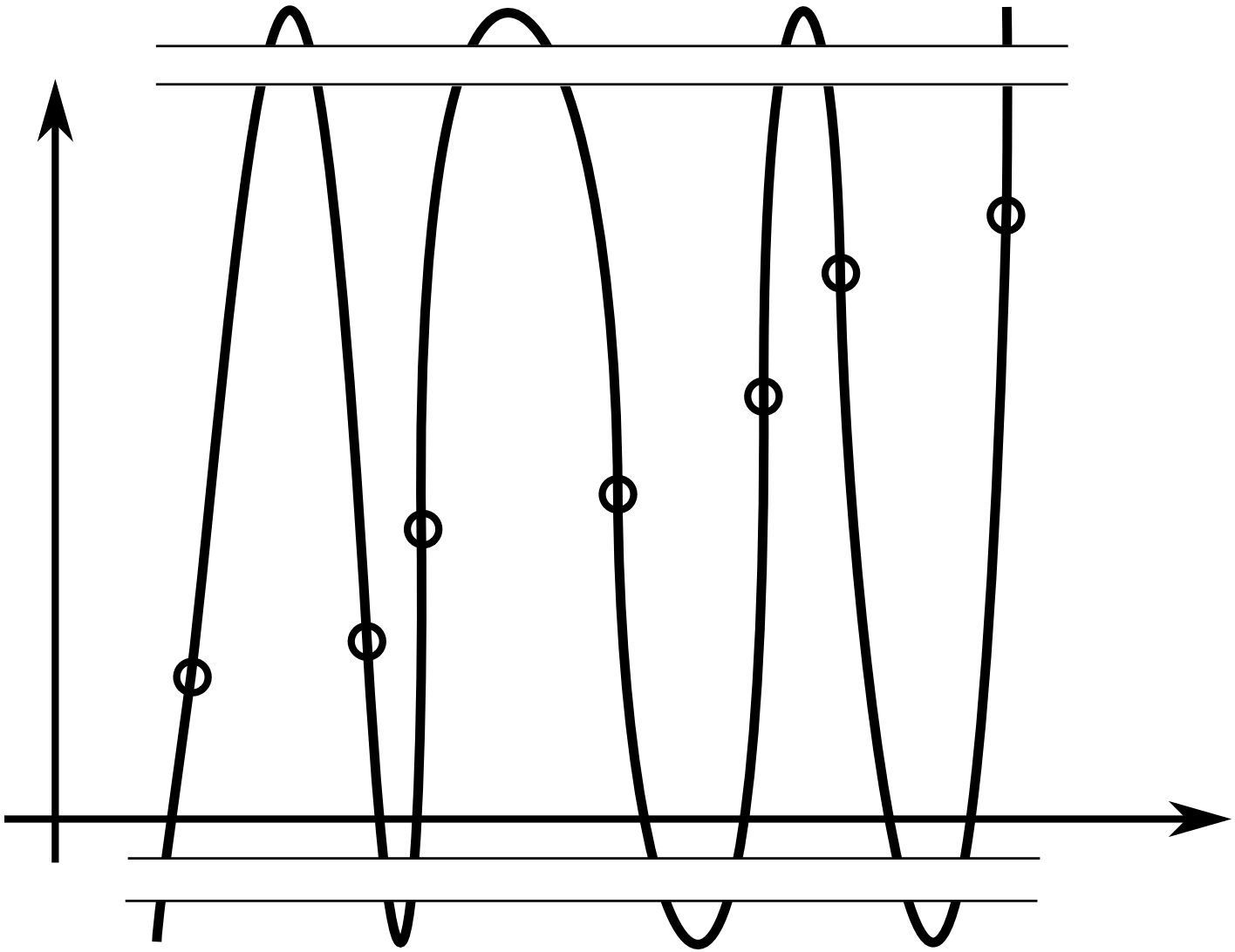
Přeparametrizování modelu



Přeparametrizování modelu



Přeparametrizování modelu



Přeparametrizování modelu

- Model je „trénován“ na dostupných datech, ale pro odhad budoucích hodnot nebo vůbec pro použitelnost modelu s naměřenými budoucími daty je žádoucí, aby model zachycoval hlavní směr vývoje, nikoliv jednotlivá náhodná vychýlení.

- Zásadou je, že počet pozorování n je mnohem větší než počet parametrů (vysvětlujících proměnných) p.

- Proto byla odvozena kriteria, která počet vysvětlujících proměnných zachycují.

Korigovaný koeficient determinace

- Informace koncentrovaná do jednoho čísla.
- Může nabývat hodnot od 0 do 1.
  - Hodnoty blízké 1: model vyhovuje.
  - Hodnoty blízké 0: model nevyhovuje.
- Jednoduchý výpočet.
- Zahrnuje počet vysvětlujících proměnných.

Korigovaný koeficient determinace

Pro model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon,$$

kde $p = k + 1$, dostaneme

$$\bar{R}^2 = 1 - \left(1 - R^2\right) \cdot \frac{n-1}{n-p}.$$

Informační kriteria

- Informační kriteria slouží zejména k porovnání více modelů se stejnou vysvětlovanou proměnnou a různými vysvětlujícími proměnnými.
- Jejich hodnota není omezena.
- Jako nejlepší model vybereme ten, který bude mít nejmenší hodnotu informačního kriteria (platí pro kriteria zde uvedená).
- Různá kriteria mohou doporučit jiné modely, v tom případě je třeba podrobit modely vybrané různými kritérii podrobnější analýze.

Akaikeho informační kritérium

Původní verze:

$$AIC = 2k + n[\ln(2\pi RSS/n) + 1]$$

Po úpravách:

$$AIC = 2k + n\ln(2\pi) + n\ln(RSS) - n\ln(n) + n$$

Vybereme pouze složky závislé na počtu proměnných:

$$AIC = 2k + n\ln(RSS)$$

Schwarzovo (Bayesovské) informační kritérium

$$SIC = BIC = n \ln\left(\frac{RSS}{n}\right) + k \ln(n)$$

Hannanovo-Quinnovo informační kriterium

$$HQC = n \ln\left(\frac{RSS}{n}\right) + 2k \ln\ln(n)$$

Interval a pás spolehlivosti pro model

Interval spolehlivosti pro regresní model

- Též predikční interval spolehlivosti.
- Je to interval, ve kterém se s vysokou pravděpodobností $(1 - \alpha)$ bude vyskytovat regresní model s teoretickými parametry.
- Obecně používaný vzorec:

$$\left(\hat{Y} - SE(\hat{Y}_i)t_{1-\alpha/2}(n-p); \hat{Y} + SE(\hat{Y}_i)t_{1-\alpha/2}(n-p)\right).$$

Interval spolehlivosti pro regresní model

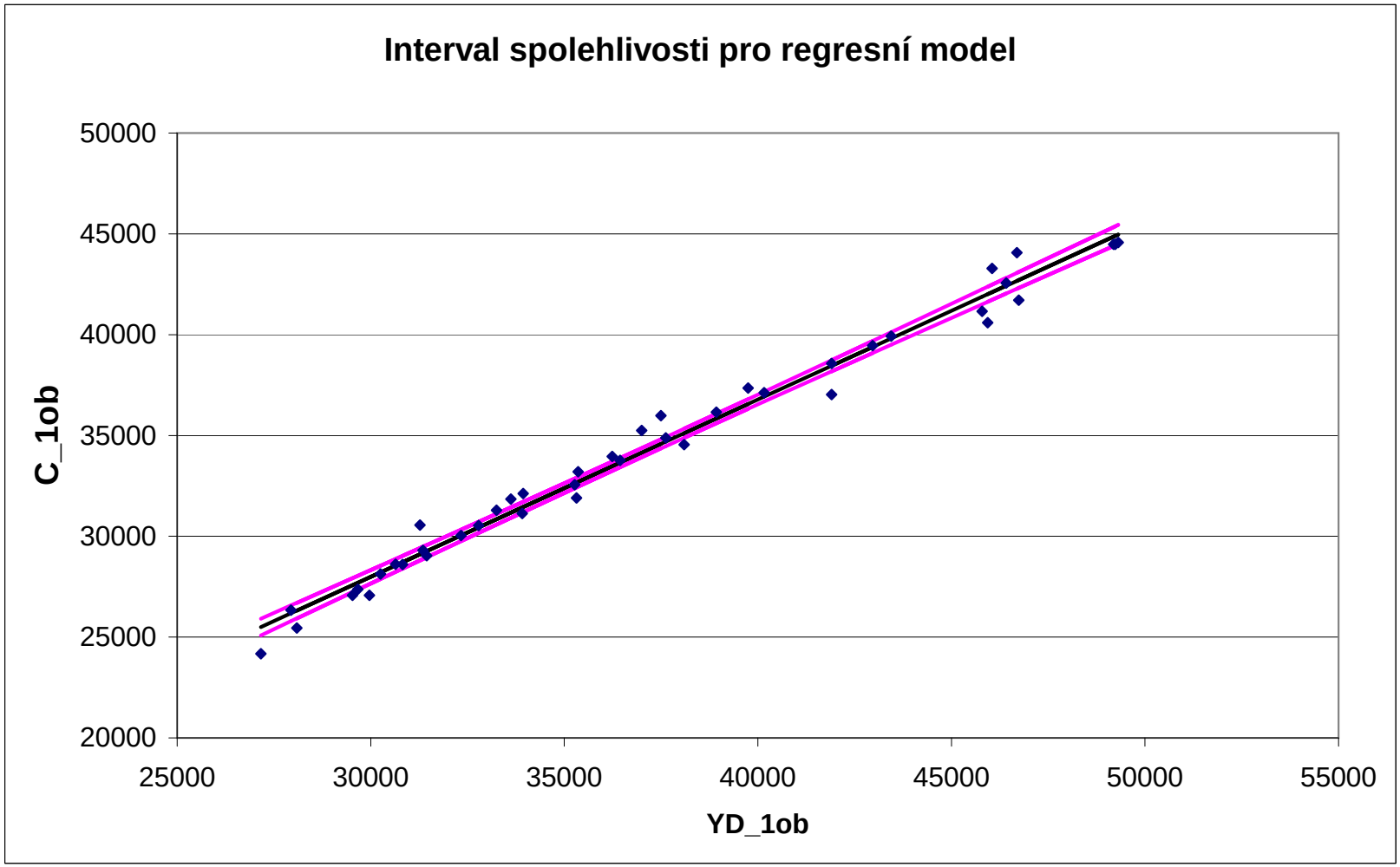
Standardní chyba

$$SE(\hat{Y}_i) = \sqrt{\frac{RSS}{n-p} m_{ii}}$$

m_{ii} je diagonální prvek matice

$$\mathbf{M} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

Interval spolehlivosti pro regresní model



Pás spolehlivosti pro regresní model

- Je to plocha, v níž se za předpokladu vyhovujícího modelu s vysokou pravděpodobností $(1-\alpha)$ budou vyskytovat napozorované hodnoty.
- Přesný vzorec:

$$\left(\hat{Y} - \sqrt{\frac{RSS}{n-p} + (SE(\hat{Y}_i))^2} t_{1-\alpha/2}(n-p); \right.$$

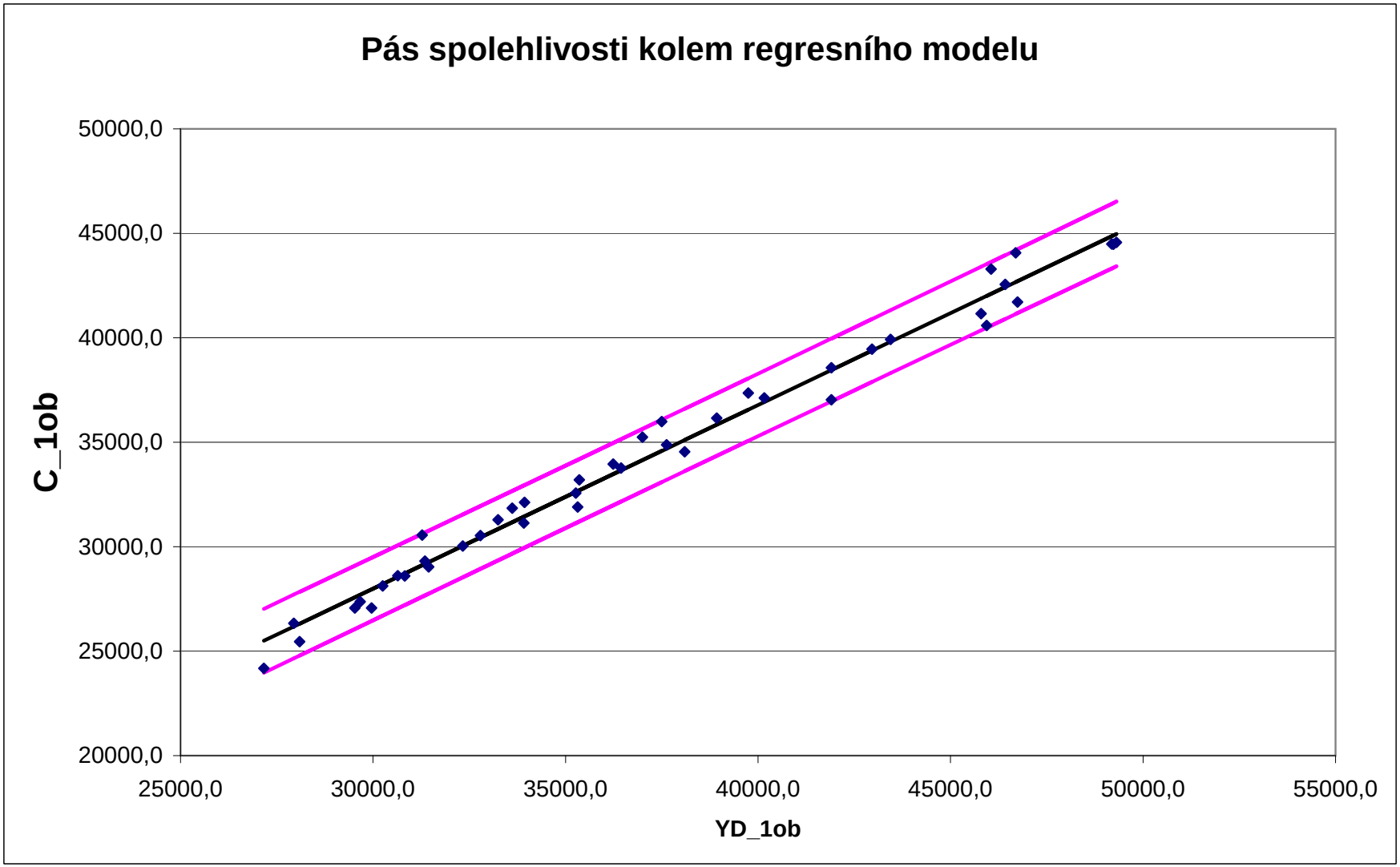
$$\left. \hat{Y} + \sqrt{\frac{RSS}{n-p} + (SE(\hat{Y}_i))^2} t_{1-\alpha/2}(n-p) \right).$$

- Přibližný vzorec:

$$\left(\hat{Y} - \sqrt{\frac{RSS}{n-p}} t_{1-\alpha/2}(n-p); \hat{Y} + \sqrt{\frac{RSS}{n-p}} t_{1-\alpha/2}(n-p) \right).$$

- Pás spolehlivosti, ačkoliv na grafu vypadá podobně jako interval spolehlivosti pro model, má zcela jinou interpretaci.

Pás spolehlivosti pro regresní model



Klasický lineární model

Motivace

- Odhady a testy vztahující se k lineární regresi vykazují za splnění určitých předpokladů velmi dobré vlastnosti.
- Předpoklady kladené na regresi je nutno ověřovat, a to zejména statistickými testy.
- S některými předpoklady jsme se již seznámili, další uvedeme či precizujeme nyní.

Klasický lineární model

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{k1} \\ 1 & X_{12} & X_{22} & \cdots & X_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{kn} \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Předpoklady klasického lineárního regresního modelu

- Jedná se o předpoklady, při jejichž splnění je odhad parametrů metodou nejmenších čtverců považován za nejlepší možný.
- Odhady parametrů dané metodou nejmenších čtverců pak mají „dobré“ vlastnosti:
  - jsou nestranné;
  - jsou maximálně vydatné;
  - jsou konzistentní;
  - jsou normálně rozdělené.

Předpoklady klasického lineárního regresního modelu

I. Regresní model je lineární v parametrech, je správně specifikován a má aditivně připojen chybový člen.

II. Chybový člen má nulovou střední hodnotu.

III. Všechny vysvětlující proměnné jsou nekorelované s chybovým členem.

IV. Pozorování chybového členu nejsou korelována se sebou samými (nedochází k autokorelaci chybového členu).

Předpoklady klasického lineárního regresního modelu

V. Chybový člen má konstantní rozptyl (homoskedasticita chybového členu).

VI. Žádná vysvětlující proměnná není lineární kombinací jiné vysvětlující proměnné (nedochází k perfektní multikolinearitě).

VII. Chybový člen je normálně rozdělen.

Předpoklad VII je volitelný, ale obvykle je zahrnut.

Předpoklady klasického lineárního regresního modelu

- Chybový člen splňující předpoklady I. až V. je nazýván klasický chybový člen.
- V případě přidání a splnění sedmého předpokladu je chybový člen nazýván klasickým normálním chybovým členem.

Klasický předpoklad I.

Regresní model je lineární v parametrech, je správně specifikován a má aditivně připojen chybový člen.

- Regresní model lineární v parametrech:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \epsilon_i$$

- Regresní model s aditivně připojeným chybovým členem:

$$Y = f(X_1, \ldots, X_k, \beta_0, \beta_1, \ldots, \beta_k) + \epsilon.$$

- Pokud není splněno, někdy možno upravit:

$$Y_i = \beta_0 \cdot \beta_1^{X_{1i}} \cdot \ldots \cdot \beta_k^{X_{ki}} \cdot \epsilon_i$$

$$\ln Y_i = \ln \beta_0 + X_{1i} \ln \beta_1 + \cdots + X_{ki} \ln \beta_k + \ln \epsilon_i$$

$$\ln Y_i = \alpha_0 + X_{1i}\alpha_1 + \cdots + X_{ki}\alpha_k + \nu_i$$

Klasický předpoklad I.

Správná specifikace modelu:

- výběr vhodných vysvětlujících proměnných;
- výběr vhodné funkční formy.

Chyby specifikace:

- opomenutá proměnná;
- nadbytečná proměnná;
- chybně zvolená funkční forma.

Klasický předpoklad I. – Opomenutá proměnná

- Opomenutá proměnná je definována jako důležitá vysvětlující proměnná (nezávisle proměnná), která je v regresním modelu vynechána.
- Opomenutá proměnná způsobuje vychýlení odhadnutých parametrů modelu (specifikační vychýlení), tj.

$$E(\hat{\beta}) \neq \beta.$$

Klasický předpoklad I. – Nadbytečná proměnná

- Nadbytečná proměnná je definována jako vysvětlující proměnná v modelu, která však do modelu nepatří.
- Přidání nadbytečné proměnné do modelu nezpůsobuje vychýlení, ale zvyšuje variabilitu parametrů zahrnutých v modelu.

Klasický předpoklad I. – Výběr vysvětlujících proměnných

Umělá (dummy) proměnná

- Proměnná, která nabývá pouze dvou hodnot (obvykle nula nebo jedna).
- V některých případech je potřeba do modelu zařadit i slovní znaky, které je nutné převést na kvantitativní znaky, např. pohlaví:
  - muž $= 1$;
  - žena $= 0$.

Metody přidávání nezávisle proměnných

- Velmi často jsou při hledání optimální podmnožiny vysvětlujících proměnných využívány metody tzv. postupného výběru regresorů (tzv. sekvenční metody), jejichž reprezentanty jsou:
  - vzestupný výběr (forward selection);
  - sestupný výběr (backward selection);
  - kroková (schodovitá) regrese (stepwise regression).

Metody přidávání nezávisle proměnných

Vzestupný výběr (forward selection)

- Metoda probíhá v několika krocích, kdy se k absolutnímu členu postupně přidává taková proměnná, která:
  - má největší příspěvek k vysvětlení variability závisle proměnné;
  - je zároveň statisticky významná.
- Problémy způsobuje multikolinearita vysvětlujících proměnných (viz další přednášky).

Metody přidávání nezávisle proměnných

Sestupný výběr (backward selection)

- Začíná se s modelem, ve kterém jsou zařazeny všechny potenciální vysvětlující proměnné.
- Následně se model redukuje o ty vysvětlující proměnné, které:
  - nejméně přispívají k vysvětlení variability závisle proměnné;
  - jsou statisticky nevýznamné.

Metody přidávání nezávisle proměnných

Kroková regrese (stepwise regression)

- Kombinace dvou předchozích metod.
- Metoda reaguje na skutečnost, že v určitém kroku může klesnout vliv některé vysvětlující proměnné, která již byla do modelu zařazena a taková proměnná se musí z modelu vyřadit.
- Postup končí, pokud do modelu nelze žádnou další proměnnou zařadit ani z modelu žádnou vyloučit.

Volba funkční formy

Rozeznáváme

- modely lineární v parametrech;
- modely nelineární v parametrech, avšak transformovatelné na modely lineární v parametrech (linearizovatelné modely);
- modely nelineární v parametrech.

Klasický předpoklad I. – Výběr vysvětlujících proměnných

Důležitá specifikační kritéria

1. Teorie: Je navrhovaná proměnná relevantní?
2. Statistický test: Je parametr navrhované proměnné statisticky významný v očekávaném směru?
3. Korigovaný koeficient determinace: Zvýší se zařazením navrhované proměnné hodnota $\bar{R}^2$?
4. Vychýlení: Změní se signifikantně ostatní parametry modelu po přidání navrhované proměnné do modelu?

Klasický předpoklad I. – Výběr vysvětlujících proměnných

Další specifikační kritéria:

- Testy specifikace:
  - RESET test,
  - testy založené na Lagrangeových multiplikátorech.
- Informační kritéria:
  - Akaikeho informační kritérium (AIC);
  - Schwarzovo (Bayesovské) informační kritérium (SIC, BIC);
  - Hannan-Quinnovo informační kritérium (HQC).

Klasický předpoklad I. – Výběr vysvětlujících proměnných

Pokud do funkčního vztahu dosadíme za parametry $\beta_0, \beta_1, \ldots, \beta_k$ jejich odhady $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$, dostaneme odhad $\hat{Y}$

$$\hat{Y} = f(X_1, \ldots, X_k, \hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)$$

získaný na základě modelu

$$Y = f(X_1, \ldots, X_k, \beta_0, \beta_1, \ldots, \beta_k) + \epsilon.$$

Klasický předpoklad I. – Výběr vysvětlujících proměnných

- Teoretický regresní model (např. lineární funkční tvar):

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, \ldots, n.$$

- Tento model můžeme odhadnout jako

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, \quad i = 1, \ldots, n.$$

Klasický předpoklad I. – Výběr vysvětlujících proměnných

RESET test

- Test chyb specifikace modelu založený na přidání $\hat{Y}^2$ a $\hat{Y}^3$ do původního modelu.

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \cdots + \beta_k X_{k,i} + \epsilon_i$$

- Pomocný model

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \cdots + \beta_k X_{k,i} + \beta_{k+1,i} \hat{Y}_i^{\,2} + \beta_{k+2,i} \hat{Y}_i^{\,3} + \nu_i$$

Klasický předpoklad I. – Výběr vysvětlujících proměnných

RESET test

- Hypotézy: H_0: model je správně specifikován, H_1: model není správně specifikován.
- Testovací statistika:

$$\frac{\frac{R_1^2 - R_0^2}{2}}{\frac{1 - R_1^2}{n-k-3}},$$

kde R_0^2 je koeficient determinace původního modelu a R_1^2 je koeficient determinace pomocného modelu

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \cdots + \beta_k X_{k,i} + \beta_{k+1,i} \hat{Y}_i^2 + \beta_{k+2,i} \hat{Y}_i^3 + \nu_i.$$

- Kritický obor:

$$W = \langle F_{1-\alpha}(2, n-k-3), \infty)$$

Klasický předpoklad I. – Výběr vysvětlujících proměnných

Test založený na Lagrangeových multiplikátorech

- založen na pomocném regresním modelu, kde vysvětlovanou proměnnou jsou rezidua původního modelu a vysvětlujícími proměnnými vysvětlující proměnné původního modelu plus:

  a) čtverce vysvětlujících proměnných;

  b) logaritmy vysvětlujících proměnných.

Klasický předpoklad I. – Výběr vysvětlujících proměnných

Test založený na Lagrangeových multiplikátorech

- Původní model:

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_k X_{i,k} + \epsilon_i, \quad i = 1, \ldots, n$$

- Odhadneme parametry, pomocí nich $\hat{Y}_i$ a nakonec $e_i = Y_i - \hat{Y}_i$.
- Pomocný model:

  a)

$$e_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_k X_{i,k} + \beta_{k+1} X_{i,1}^2 + \cdots + \beta_{k+k} X_{i,k}^2$$

  b)

$$e_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_k X_{i,k} + \beta_{k+1} \ln(X_{i,1}) + \cdots + \beta_{k+k} \ln(X_{i,k})$$

- Vypočteme koeficient determinace R^2.

Klasický předpoklad I. – Výběr vysvětlujících proměnných

Test založený na Lagrangeových multiplikátorech

- Hypotézy:

  H_0: model je správně specifikován, H_1: model není správně specifikován.

- Testovací statistika:

$$nR^2$$

- Kritický obor:

$$W = \langle \chi^2_{1-\alpha}(k), \infty)$$

Klasický předpoklad I. – Výběr vysvětlujících proměnných

Test založený na Lagrangeových multiplikátorech – jedna proměnná

- Původní model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, \ldots, n$$

- Odhadneme parametry, pomocí nich $\hat{Y}_i$ a nakonec $e_i = Y_i - \hat{Y}_i$.
- Pomocný model:

$$e_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3, \quad i = 1, \ldots, n$$

- Vypočteme koeficient determinace R^2.

Klasický předpoklad I. – Výběr vysvětlujících proměnných

Test založený na Lagrangeových multiplikátorech – jedna proměnná

- Hypotézy: H_0: model je správně specifikován, H_1: model není správně specifikován.
- Testovací statistika:

$$nR^2$$

- Kritický obor:

$$W = \langle \chi^2_{1-\alpha}(2), \infty)$$

Klasický předpoklad I. – Analýza dat „naslepo“

Pro 4 různé sady dat platí:

Popisné charakteristiky

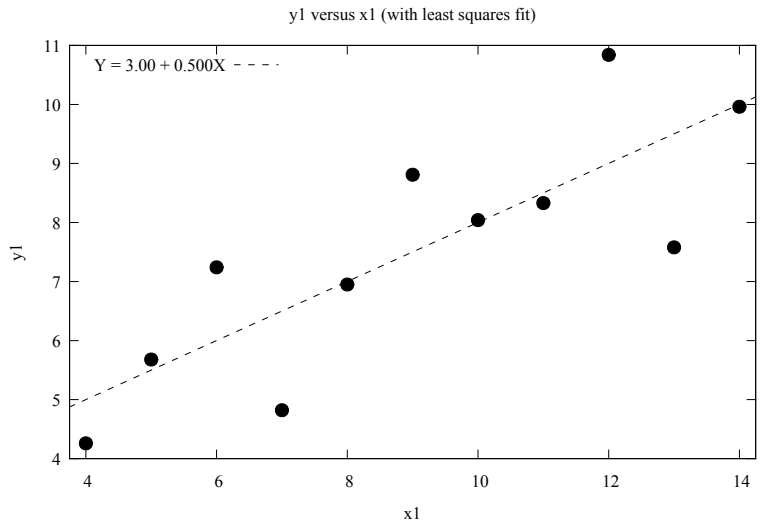
Proměnná	n	Průměr	Sm. odch.
x	11	9,00	3,32
y	11	7,50	2,03

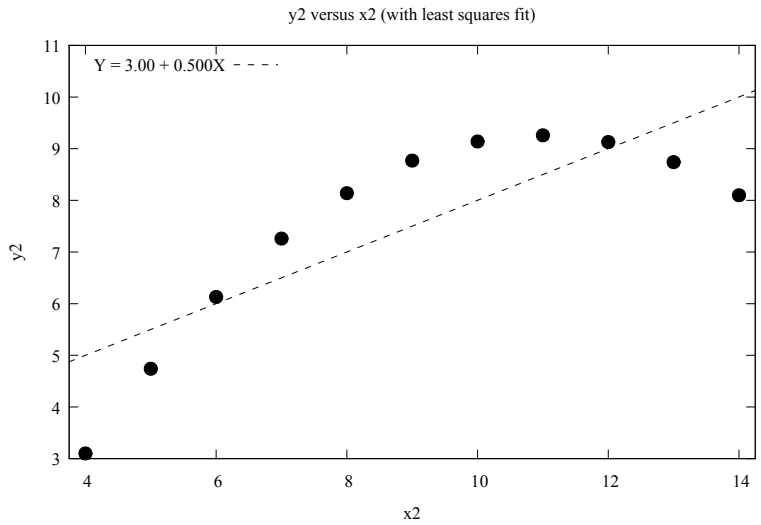
Výsledky lineární regrese – závislost y na x

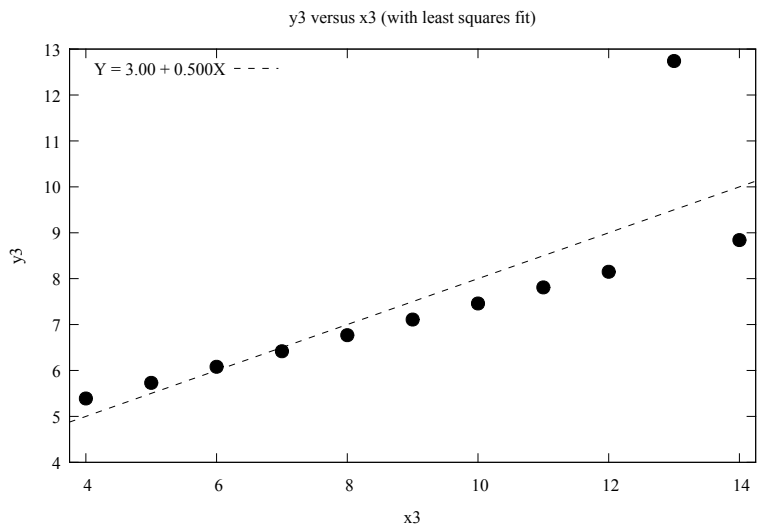
	$\hat{\beta}_i$	$SE(\hat{\beta}_i)$	T	p-hodnota
Konstanta	3,00	1,125	2,67	0,026
x	0,50	0,118	4,24	0,002

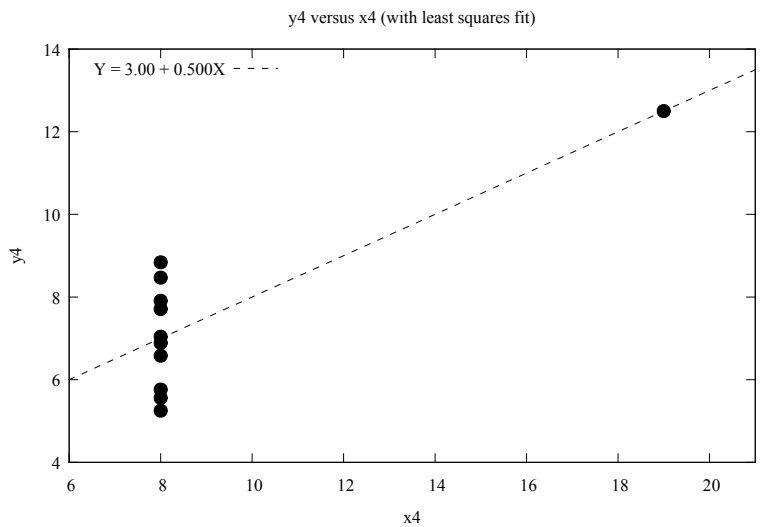
Rovnice regresní přímky:	$y = 3{,}00 + 5{,}00x$
Koeficient determinace R^2:	0,67
Korelační koeficient:	0,82
Směrodatná odchylka reziduí:	1,24

Klasický předpoklad I. – Analýza dat „naslepo“









Klasický předpoklad II.

- Chybový člen má nulovou střední hodnotu.

Klasický předpoklad III.

- Všechny vysvětlující proměnné jsou nekorelované s chybovým členem.
- Jestliže jsou chybový člen a vysvětlující proměnné korelovány, pak metoda nejmenších čtverců chybně přisoudí vysvětlujícím proměnným část variability ve vysvětlované proměnné, která však pochází z chybového členu.

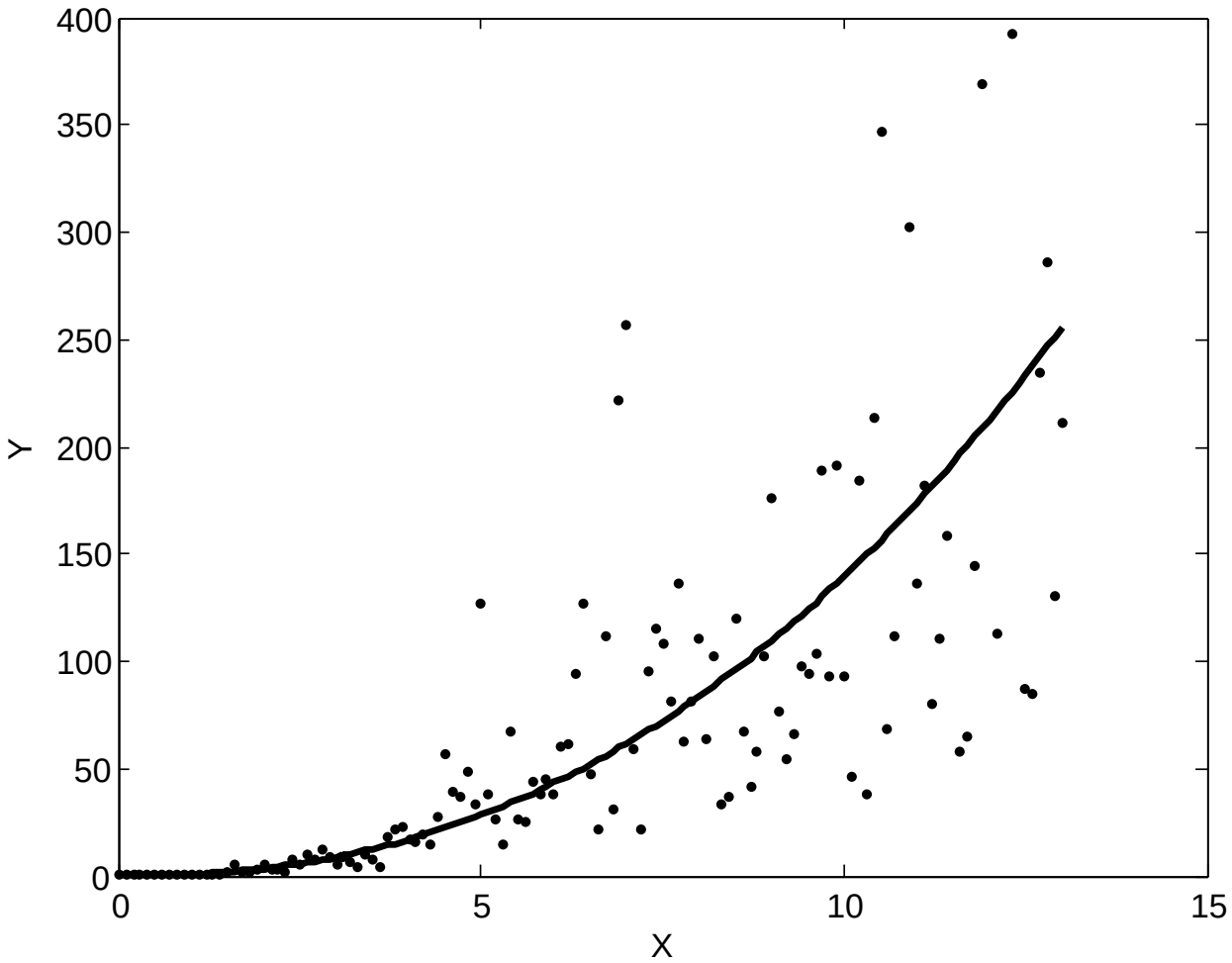
Klasický předpoklad IV.

- Pozorování chybového členu nejsou korelována se sebou samými, tj. nedochází k významné autokorelaci.
- Pokud existuje korelace mezi pozorováními chybového členu, pak dochází ke zvýšení variability odhadu parametrů modelu.

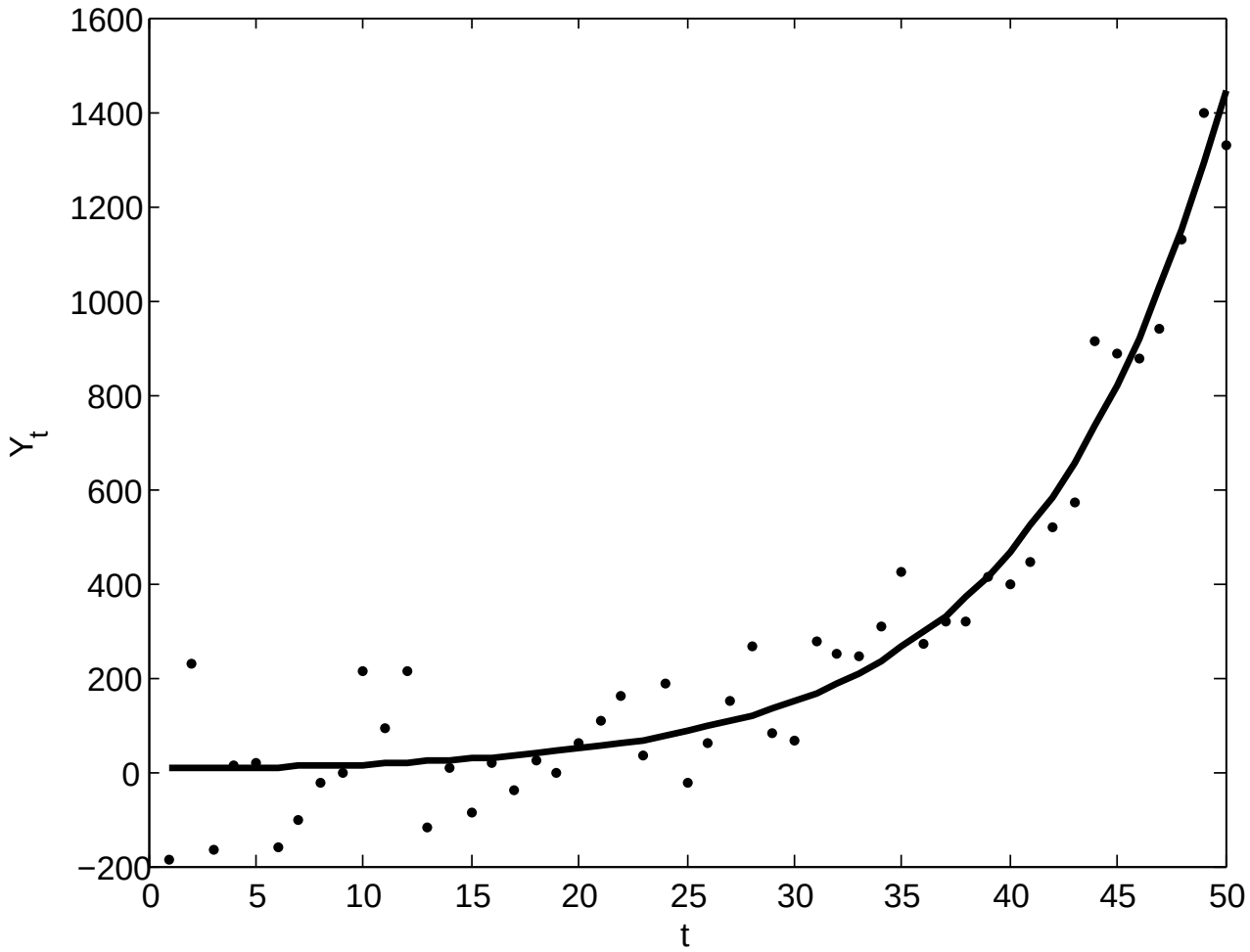
Klasický předpoklad V.

- Chybový člen má konstantní rozptyl, tj. je homoskedastický.
- Heteroskedasticita znamená, že se rozptyl (variabilita) rozdělení chybového členu v jednotlivých úsecích pozorování mění.

Klasický předpoklad V.



Klasický předpoklad V.



Klasický předpoklad VI.

- Žádná vysvětlující proměnná není perfektní lineární kombinací jiné vysvětlující proměnné – neexistuje perfektní (multi)kolinearita.
- Perfektní kolinearita:
  - stejné proměnné;
  - jedna proměnná je lineární transformací jiné proměnné.
- Problém činí i neperfektní kolinearita, případně pokud jedna z proměnných má nulovou variabilitu.

Klasický předpoklad VII.

Chybový člen je normálně rozdělen.

Důvody přidání tohoto předpokladu:

1. Chybový člen zahrnuje mnoho minoritních vlivů (opomenuté proměnné) nebo chyb – v případě rostoucího počtu těchto minoritních vlivů nebo chyb má rozdělení chybového členu tendenci konvergovat k normálnímu rozdělení — důvodem je centrální limitní věta.
2. Jeho hlavní využití je při testování hypotéz, neboť bez splnění tohoto předpokladu bude většina testů (např. t-test, F-test) neplatná.

Gaussova-Markovova věta

Nechť jsou splněny klasické předpoklady I. až VI.

Pak odhad parametru $\hat{\beta}_i$ daný metodou nejmenších čtverců má minimální rozptyl mezi všemi lineárními nevychýlenými odhady parametru $\hat{\beta}_i$, $i = 0, 1, 2, \ldots, k$.

Gaussova-Markovova věta

- Gaussova-Markovova věta je lépe zapamatovatelná jako BLUE (= Best Linear Unbiased Estimator).
- Pokud přidáme předpoklad VII., pak lze ukázat, že odhad parametru daný metodou nejmenších čtverců je tzv. BUE (= Best Unbiased Estimator), tj. nejlepší (maximálně vydatný) nevychýlený odhad parametrů ze všech možných odhadů (tj. nejen ze všech lineárních odhadů).

Příklad

- `data_PCA.sta`