

## VĚTA 3.10

Nechť v modelu  $(**)$  platí, že  $\mathbf{e} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$  a  $h(\mathbf{X}) = m + 1$ . Označme  $SSE_F$  reziduální s.č. plného modelu a  $SSE_R$  reziduální s.č. modelu, kde platí  $H_0: \mathbf{C}\boldsymbol{\beta} = \mathbf{b}$ . Potom je za platnosti  $H_0$  splněno

$$F = \frac{\frac{\Delta SSE}{r}}{s_n^2} \sim F(r, n - m - 1), \quad \text{kde } \Delta SSE = SSE_R - SSE_F.$$

## Lemma 3.2

Označme  $\hat{\boldsymbol{\beta}}_F$  a  $\hat{\boldsymbol{\beta}}_R$  LSE parametru  $\boldsymbol{\beta}$  v plném a redukovaném modelu. Potom platí

- 1)  $\hat{\boldsymbol{\beta}}_R = \hat{\boldsymbol{\beta}}_F - (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{C}^T \mathbf{A} (\mathbf{C} \hat{\boldsymbol{\beta}}_F - \mathbf{b})$ , kde  $\mathbf{A} = \left( \mathbf{C} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{C}^T \right)^{-1}$ ,
- 2)  $\Delta SSE = SSE_R - SSE_F = \left( \mathbf{C} \hat{\boldsymbol{\beta}}_F - \mathbf{b} \right)^T \mathbf{A} \left( \mathbf{C} \hat{\boldsymbol{\beta}}_F - \mathbf{b} \right).$

Důkaz.

Důkaz Věty 3.10.



### POZNÁMKA 3.10

- použitím rozkladu  $SST = SSE + SSR$  dostaneme

$$\Delta SSE = SSR_F - SSR_R.$$

**interpretace:** nárůst regresního součtu čtverců díky neplatnosti  $H_0$

- dále

$$SSR_F = SSR_R + \Delta SSE,$$

kde  $\Delta SSE$  se nazývá **extra sum of squares** (přidaný k  $SSR$  díky neplatnosti  $H_0$ )

- např. pokud  $\beta_R = (\beta_0, \beta_1, \dots, \beta_{m-1}, 0)$ , tzn.  $\beta_m = 0$  a skutečný model má  $\beta = \beta_F$ , potom  $\Delta SSE$  je extra regresní součet čtverců získaný díky přidání  $\beta_m$  do modelu
- umožňuje rozklad  $SSR$  plného modelu na jednotlivé části  $(x_1, x_2 | x_1, x_3 | x_2 x_1, \dots)$

### PŘÍKLAD 3.7 (Porodní váha)

```
mod <- lm(Weight ~ Age + Sex + Sex:Age)
anova(mod)
```

```
## Analysis of Variance Table
## Response: Weight
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Age         1 1013799 1013799 31.0779 1.862e-05 ***
## Sex         1  157304  157304  4.8221  0.04006 *
## Age:Sex      1    6346    6346  0.1945  0.66389
## Residuals  20  652425   32621
```

```
mod0 <- lm(Weight ~ Age + Sex)
anova(mod0,mod)
```

```
## Analysis of Variance Table
## Model 1: Weight ~ Age + Sex
## Model 2: Weight ~ Age + Sex + Sex:Age
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1       21 658771
## 2       20 652425  1    6346.2 0.1945 0.6639
```

```
SSR <- sum((predict(mod)-mean(y))^2)
SSR0 <- sum((predict(mod0)-mean(y))^2)
D.SSE <- SSR-SSR0; D.SSE
## 6346.225
MSE<- sum((predict(mod)-Weight)^2)
F<- D.SSE/MSE*(24-3-1); F
## 0.1945428
```

### PŘÍKLAD 3.8 (Data cement)

```
mod <- lm(y ~ x1 + x2 + x3 + x4)
anova(mod)
```

```
## Analysis of Variance Table
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x1         1 1450.08 1450.08 242.3679 2.888e-07 ***
## x2         1 1207.78 1207.78 201.8705 5.863e-07 ***
## x3         1    9.79    9.79   1.6370  0.2366
## x4         1    0.25    0.25   0.0413  0.8441
## Residuals  8   47.86    5.98
## ---
```

```
y.hat <- predict(mod)
SSR <- sum((y.hat-mean(y))^2); SSR
## 2667.899
```

## 3.6 Predikce

### a) Intervaly spolehlivosti pro $E(Y_{\mathbf{x}})$

- necht'  $\mathbf{x}_0 = (1, x_{0,1}, \dots, x_{0,m})^T$  je nový bod proměnné  $\mathbf{x}$
- bodový odhad  $E(Y_{\mathbf{x}_0})$  je roven

$$\hat{Y}_{\mathbf{x}_0} = \hat{\beta}_0 + \sum_{j=1}^m x_{0,j} \hat{\beta}_j = \mathbf{x}_0^T \hat{\beta}$$

- tzn.  $\text{Var}(\hat{Y}_{\mathbf{x}_0}) = \mathbf{x}_0^T \cdot \text{Var}(\hat{\beta}) \cdot \mathbf{x}_0 = \sigma^2 \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0$
- a může být odhadnut pomocí

$$\hat{\sigma}^2(\hat{Y}_{\mathbf{x}_0}) = s_n^2 [\mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0] \quad (\text{rozptyl predikce})$$

- speciálně pokud  $\mathbf{x}_0^T = \mathbf{x}_i^T$  ( $i$ -tý řádek matice  $\mathbf{X}$ )

$$\hat{\sigma}^2(\hat{Y}_{\mathbf{x}_i}) = s_n^2 [\mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i] = s_n^2 h_{ii} \quad \text{kde} \quad h_{ii} = (\mathbf{H})_{ii} \quad \text{a} \quad \mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$$

- pro normální chyby lze odvodit interval spolehlivosti pro  $E(Y_{\mathbf{x}_0}) = \mu_{\mathbf{x}_0}$

- $\hat{Y}_{\mathbf{x}_0}$  má totiž normální rozdělení (je to LK složek vektoru  $\hat{\beta}$ ) a platí

$$E(\hat{Y}_{\mathbf{x}_0}) = \mu_{\mathbf{x}_0} = \mathbf{x}_0^T \beta \quad \text{a} \quad \text{Var}(\hat{Y}_{\mathbf{x}_0}) = \sigma^2 \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0$$

- tzn.

$$\frac{\hat{Y}_{\mathbf{x}_0} - \mu_{\mathbf{x}_0}}{\sigma \sqrt{\mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0}} \sim N(0, 1)$$

a díky nezávislosti  $\hat{\beta}$  a  $s_n^2$

$$\frac{\hat{Y}_{\mathbf{x}_0} - \mu_{\mathbf{x}_0}}{s_n \sqrt{\mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0}} \sim t(n - m - 1)$$

- 100(1 -  $\alpha$ )% IS pro  $\mu_{\mathbf{x}_0}$  tedy je:

$$\left( \hat{Y}_{\mathbf{x}_0} \pm t_{1-\frac{\alpha}{2}}(n - m - 1) \cdot s_n \sqrt{\mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0} \right)$$

POZNÁMKA: Intervaly spolehlivosti v 

# 95% intervaly spolehlivosti

```
CI<-predict(mod.lin, new, interval = "confidence")
```

## b) Intervaly predikce pro $Y_{\mathbf{x}_0}$

- bodový odhad je opět  $\hat{Y}_{\mathbf{x}_0}$
- pokud  $Y_{\mathbf{x}_0}$  je skutečná hodnota  $Y_{\mathbf{x}}$  v bodě  $\mathbf{x} = \mathbf{x}_0$ , potom  $Y_{\mathbf{x}_0}$  a  $\hat{Y}_{\mathbf{x}_0}$  budou nezávislé za předpokladu, že pozorování  $Y_{\mathbf{x}_0}, Y_1, \dots, Y_n$  jsou nezávislá (což předpokládáme)
- potom

$$\text{Var}(\hat{Y}_{\mathbf{x}_0} - Y_{\mathbf{x}_0}) = \text{Var}(\hat{Y}_{\mathbf{x}_0}) + \text{Var}(Y_{\mathbf{x}_0}) = \sigma^2(1 + \mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0),$$

- takže

$$\frac{\hat{Y}_{\mathbf{x}_0} - Y_{\mathbf{x}_0}}{\sigma\sqrt{1 + \mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0}} \sim N(0, 1) \quad \text{a} \quad \frac{\hat{Y}_{\mathbf{x}_0} - Y_{\mathbf{x}_0}}{s_n\sqrt{1 + \mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0}} \sim t(n - m - 1)$$

(za předpokladu normality chyb)

- $100(1 - \alpha)\%$  IP pro  $Y_{\mathbf{x}_0}$  tedy je

$$\left( \hat{Y}_{\mathbf{x}_0} \pm t_{1-\frac{\alpha}{2}}(n - m - 1) \cdot s_n \sqrt{1 + \mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0} \right)$$

**POZNÁMKA:** 95% Intervaly predikce v :

```
PI<-predict(mod.lin, new, interval = "prediction")
```

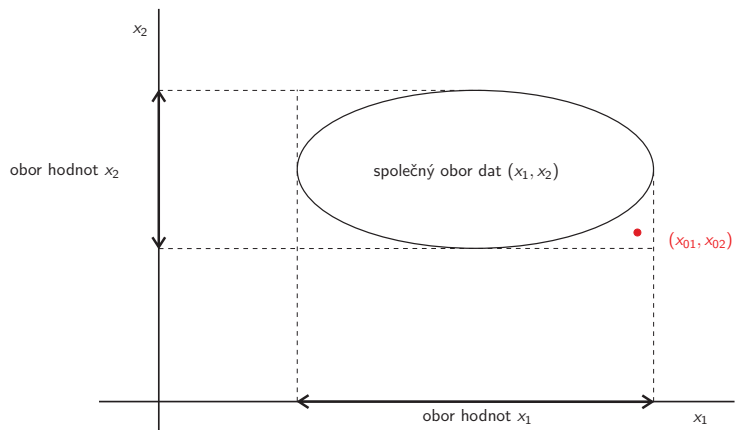
### POZNÁMKA 3.11 (Extrapolace)

- u jednoduché LR kvalita predikce závisela na vzdálenosti  $x_0$  od  $\bar{x}$
- je třeba si dát pozor na predikce mimo  $\langle x_{min}, x_{max} \rangle$
- podobné závěry platí i pro vícerozměrnou LR
- protože rozptyl predikce je úměrný  $\mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0$ , v bodech s velkými hodnotami této veličiny nebude predikce spolehlivá
- speciálně pokud  $\mathbf{x}_i$  jsou pozorovaná data, můžeme očekávat, že body s nejvyššími hodnotami  $\mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i = h_{ii}$  budou na hranici množiny, kde je predikce spolehlivá
- tzn., že vnitřek elipsoidu

$$\mathbf{x}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x} \leq \max_{1 \leq i \leq n} h_{ii}$$

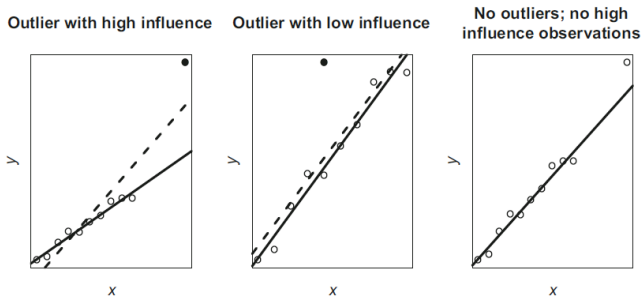
může být považován za **přípustný obor predikce**





## 4. Rezidua, diagnostika a transformace

- je třeba ověřit adekvátnost modelu  
máme  $R^2$ ,  $t$ ,  $F$  statistiky, byly odvozeny za předpokladu linearity modelu a dalších podmínek  
pro ověření je důležitý nástroj **analýza reziduí**
- je také třeba ověřit vliv jednotlivých pozorování na model  
**analýza odlehlých** (outliers) a **vlivných pozorování** (influential observations)



- při detekci problémů s modelem  
mohou pomoci **transformace proměnných** nebo **metoda na korekci nekonstantního rozptylu**

## 4.1 Rezidua

Připomenutí:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{H}\mathbf{y}, \quad \text{kde} \quad \mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T, \quad \hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I}_n - \mathbf{H})\mathbf{y} = (\mathbf{I}_n - \mathbf{H})\mathbf{e}$$

Dále jsme ukázali:

$$E(\hat{\mathbf{e}}) = \mathbf{0}, \quad \text{Cov}(\hat{\mathbf{e}}) = \sigma^2(\mathbf{I}_n - \mathbf{H}), \quad \text{pokud } \mathbf{e} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n) \text{ potom } \hat{\mathbf{e}} \sim N_n(\mathbf{0}, \sigma^2(\mathbf{I}_n - \mathbf{H}))$$


Při označení  $h_{ij} = \mathbf{H}_{ij}$  tedy platí  $\text{Var}(\hat{e}_i) = \sigma^2(1 - h_{ii})$  a  $\text{Cov}(\hat{e}_i, \hat{e}_j) = -\sigma^2 h_{ij}$

Obecně bývá výhodnější pracovat s normovanými rezidui

$$r_i = \frac{\hat{e}_i}{\sigma\sqrt{1 - h_{ii}}}, \quad \text{pro která platí } \text{Var}(r_i) = 1$$


- pokud  $\sigma^2$  odhadneme pomocí  $s_n^2 = \frac{1}{(n-m-1)} SSE$ , dostaneme

$$\hat{r}_i = \frac{\hat{e}_i}{s_n\sqrt{1 - h_{ii}}}, \quad i = 1, \dots, n, \quad \text{tzv. interně studentizovaná rezidua}$$

někdy také **standardizovaná rezidua**, funkce v : `rstandard(mod)`

- pokud  $\sigma^2$  odhadneme na základě modelu, ve kterém bylo vynecháno  $i$ -té pozorování, ozn.  $\hat{\sigma}_{(-i)}^2$ , dostaneme

$$\hat{t}_i = \frac{\hat{e}_i}{\hat{\sigma}_{(-i)} \sqrt{1 - h_{ii}}}, \quad i = 1, \dots, n, \quad \text{tzv. externě studentizovaná rezidua}$$

někdy také **studentizovaná rezidua**, funkce v : `rstudent(mod)`

např.  $\hat{\sigma}_{(-i)}^2 = \frac{1}{n-m-2} SSE_{(-i)}$  je nestranný odhad  $\sigma^2$  v modelu bez  $i$ -tého pozorování

#### POZNÁMKA 4.1

- pokud je  $h_{ii}$  malé, pro velká  $n$  by se měla  $\hat{e}_i, \hat{r}_i, \hat{t}_i$  chovat přibližně stejně a  $\hat{r}_i, \hat{t}_i \approx N(0, 1)$
- pro malá  $n$  ( $n < 20$ ) a/nebo  $h_{ii} \approx 1$  je preferováno použití  $\hat{r}_i$  nebo  $\hat{t}_i$
- častěji bývá doporučováno  $\hat{t}_i$
- $h_{ii}$  hraje zásadní roli v diagnostice modelu, probereme jeho základní vlastnosti

**Leverage**  $h_{ii}$  - potenciál  $i$ -tého pozorování

**leverage point** - pákový bod, vzdálený bod

# Vlatnosti potenciálu $h_{ii}$

- $\text{Var}(\hat{e}_i) = \sigma^2(1 - h_{ii}) \geq 0 \Rightarrow h_{ii} \leq 1$
- $\mathbf{H}^2 = \mathbf{H} \Rightarrow h_{ii} = \sum_{j=1}^n h_{ij}h_{ji} = \sum_{j=1}^n (h_{ij})^2$  tedy  $h_{ii} > 0$  (dá se ukázat:  $h_{ii} \geq \frac{1}{n}$ )
- $\mathbf{HX} = \mathbf{X} \Rightarrow \sum_{j=1}^n h_{ij}x_{j1} = \sum_{j=1}^n h_{ij} = x_{i1} = 1$  tedy  $\sum_{j=1}^n h_{ij} = 1, \forall i \in \hat{n}$  (v mod. s interceptem)
- význam  $h_{ii}$  vyplýne z následujících úvah:  $\hat{\mathbf{y}} = \mathbf{Hy} \Rightarrow \hat{y}_i = \sum_{j=1}^n h_{ij}y_j = h_{ii}y_i + \sum_{j=1, j \neq i}^n h_{ij}y_j$ 
  - pokud  $h_{ii} \approx 1$ , potom  $\hat{y}_i \approx y_i$  a model je nucen proložit přímkou bodem  $(\mathbf{x}_i, y_i)$  i když když tam neplatí
  - body s „velkým  $h_{ii}$ “ - **body s velkým potenciálem** (high leverage points)
  - tyto body by měly být detekovány pro další zkoumání
- otázka je, jaká hodnota  $h_{ii}$  je „velká“  
**Heuristické pravidlo:**  $\sum_{i=1}^n h_{ii} = \text{tr}(\mathbf{H}) = m + 1$ , tzn.  $\frac{m+1}{n}$  je průměrná hodnota  $h_{ii}$   
 $i$ -té pozorování má velký potenciál, jestliže  $h_{ii} > \frac{3(m+1)}{n}$

## 4.2 Grafy reziduí

A) **ověření normality** - histogramy, Q-Q plots

tyto obrázky nezávisí na počtu nezávislých proměnných  $x$ , vše stejné jako v jednorozměrné LR  
*testy normality*: Shapiro-Wilk, Lilliefors, Anderson-Darling

B) pro **ověření funkční formy** pro  $E(Y_x)$  a / nebo **konstantního rozptylu** se nejčastěji používají:

- 1) grafy  $\hat{e}_i, \hat{r}_i$  nebo  $\hat{t}_i$  oproti  $\mathbf{x}_j^c$ ,  $j = 1, \dots, m$ , kde  $\mathbf{x}_j^c$  je  $j$ -tý sloupec  $\mathbf{X}$
- 2) grafy  $\hat{e}_i, \hat{r}_i$  nebo  $\hat{t}_i$  oproti  $\hat{y}_i$
- 3) partial residual plots

*testy konstantního rozptylu*: Breusch-Pagan, Levene

### POZNÁMKA 4.2

Zdůvodnění 1): normální rovnice  $\mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\beta}) = 0$  implikují  $\mathbf{X}^T(\mathbf{y} - \hat{\mathbf{y}}) = \mathbf{X}^T\hat{\mathbf{e}} = 0$

připomenutí:  $\mathbf{Y}_i = \beta_1 x_i + e_i$ , 
$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} = \frac{\mathbf{x}^T \mathbf{y}}{\|\mathbf{x}\|^2}$$

- pro LR model bez interceptu pro  $\hat{\mathbf{e}}$  v závislosti na  $\mathbf{x}_j^c$  bude odhad směrnice

$$\hat{\beta}_j^* = \frac{(\mathbf{x}_j^c)^T \hat{\mathbf{e}}}{\|\mathbf{x}_j^c\|^2} = 0$$

- graf  $\hat{e}_i, \hat{r}_i, \hat{t}_i$  oproti  $\mathbf{x}_j^c$  by měl dávat náhodně rozptýlené body kolem osy  $x$  (bez trendů,  $\hat{r}_i, \hat{t}_i$  uvnitř  $\approx \pm 2$  )
- pokud tomu tak není, může to naznačovat nelinearitu v  $\mathbf{x}_j$  nebo nekonstantní rozptyl

Zdůvodnění 2): ukázali jsme  $\sum_{i=1}^n \hat{y}_i \hat{e}_i = 0$

- pro LM bez interceptu pro  $\hat{e}_i$  oproti  $\hat{y}_i$  tedy platí

$$\hat{\beta} = \frac{\hat{\mathbf{e}}^T \hat{\mathbf{y}}}{\|\hat{\mathbf{y}}\|^2} = 0$$

- body by opět měly být náhodně rozptýlené kolem osy  $x$
- trychtýřovitý tvar indikuje nekonstantní rozptyl
- trendy indikují nelinearitu

## Ad 3) Partial residual plots

- i když grafy  $\hat{\mathbf{e}}$  oproti  $\mathbf{x}_j^c$  a  $\hat{\mathbf{y}}$  mohou indikovat nedostatky modelu, nemusí být zřejmé, jaké ty nedostatky jsou
- V SLR graf  $\hat{e}_i$  oproti  $x_i$  lze použít pro detekci nelinearity
- ale v MLR tyto grafy mohou být zavádějící, protože  $\hat{\mathbf{e}}$  závisí na všech prediktorech, nemusí být tedy izolován efekt dané proměnné při odstranění efektů ostatních
- pro zkoumání efektů  $j$ -té proměnné lze použít **partial residual plots**
  - lze je chápat jako ekvivalent scatterplotů v SLR
- definujme  $\hat{\mathbf{e}}_j^* = \hat{\mathbf{e}} + \hat{\beta}_j \mathbf{x}_j^c$ ,  
kde  $\hat{\mathbf{e}}$  je vektor reziduí modelu,  $\hat{\beta}_j$  je LSE parametru  $\beta_j$ ,  $\mathbf{x}_j^c$  je  $j$ -tý sloupec  $\mathbf{X}$
- **partial residual plot (PRP)**: graf  $\hat{\mathbf{e}}_j^*$  oproti  $\mathbf{x}_j^c$ ,  $j = 1, \dots, m$
- pokud je model správný, měly by být body náhodně rozmístěné kolem přímky se směrnici  $\hat{\beta}_j$



## Zdůvodnění:

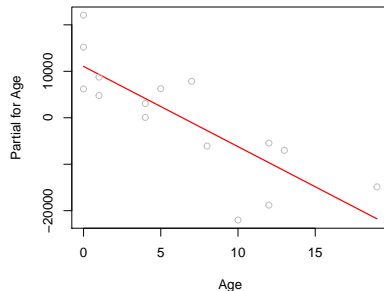
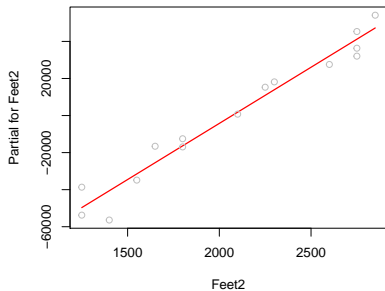
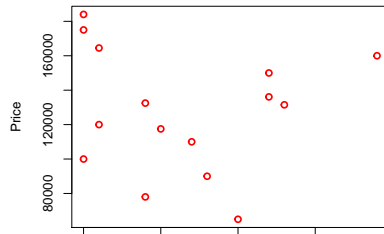
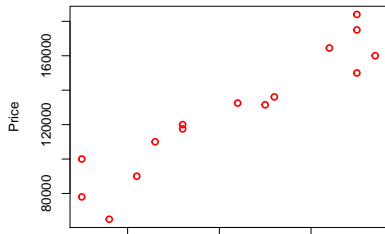
- vztah mezi  $\hat{\mathbf{e}}_j^*$  a  $\mathbf{x}_j^c$  má formu SLR bez interceptu
- pokud je model správný, platí  $E(\hat{\mathbf{e}}_i) = 0$  a  $\text{Var}(\hat{\mathbf{e}}_i) = \sigma^2(1 - h_{ii})$
- má tedy smysl uvažovat regresní model pro  $\hat{\mathbf{e}}_j^*$  oproti  $\mathbf{x}_j^c$  ( $\hat{\mathbf{e}}_j^* = \gamma_j \mathbf{x}_j^c + \mathbf{e}$ )
- pro odhad koeficientu platí

$$\hat{\gamma}_j = \frac{(\hat{\mathbf{e}}_j^*)^T \mathbf{x}_j^c}{\|\mathbf{x}_j^c\|^2} = \frac{(\hat{\mathbf{e}} + \hat{\beta}_j \mathbf{x}_j^c)^T \mathbf{x}_j^c}{\|\mathbf{x}_j^c\|^2} = \frac{\hat{\mathbf{e}}^T \mathbf{x}_j^c + \hat{\beta}_j \|\mathbf{x}_j^c\|^2}{\|\mathbf{x}_j^c\|^2} = \hat{\beta}_j$$

## PŘÍKLAD 4.1 (Housing Price Data)

	Feet2	Age	Price
1	1800	1	120000
2	1650	7	110000
3	2750	12	150000
4	1550	8	90000
⋮	⋮	⋮	⋮

```
mod <- lm(Price ~ Feet2 + Age)
summary(mod)
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13238.600   7677.183   1.724 0.110273
## Feet2       60.589     3.644   16.625 1.19e-09 ***
## Age       -1726.762    364.172  -4.742 0.000479 ***
## ---
## Residual standard error: 7763 on 12 degrees of freedom
## Multiple R-squared:  0.9586,    Adjusted R-squared:  0.9517
## F-statistic: 139 on 2 and 12 DF,  p-value: 5.021e-09
```

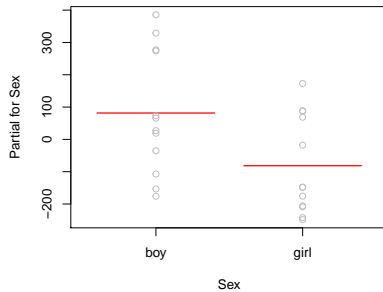
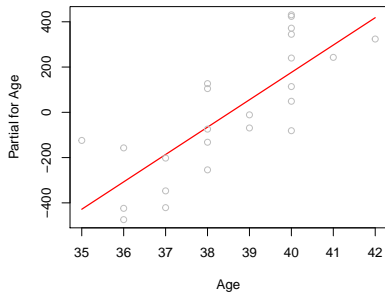


```
termplot(mod,partial.resid=TRUE, terms="Feet2")
```

```
termplot(mod, partial.resid=TRUE, terms="Age")
```

## PŘÍKLAD 4.2 (Porodní váha)

```
mod <- lm(Weight ~ Age + Sex)
termplot(mod, partial.resid=TRUE, terms="Age")
termplot(mod, partial.resid=TRUE, terms="SEX")
```



- PRPs jsou někdy kritizovány za nadhodnocování efektu  $\mathbf{x}_j^c$
- alternativa: **partial regression plot (added variable plot)**
- motivace: ptáme se, zda přidat novou proměnnou do modelu a chceme odhadnout její efekt

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \gamma\mathbf{w} + \mathbf{e}, \quad \text{kde } \mathbf{w} \text{ je nový vektor regresorů}$$

lze zapsat jako

$$\mathbf{Y} = [\mathbf{X} \ \mathbf{w}] \begin{bmatrix} \boldsymbol{\beta} \\ \gamma \end{bmatrix} + \mathbf{e} = \mathbf{X}_w \boldsymbol{\beta}_w + \mathbf{e}$$

požitím normálních rovnic lze odvodit

$$\hat{\gamma} = \frac{\hat{\mathbf{e}}^T (\mathbf{I} - \mathbf{H}) \mathbf{w}}{\|(\mathbf{I} - \mathbf{H}) \mathbf{w}\|^2} \quad (4.1)$$

$\hat{\gamma}$  je směrnice RM pro  $\hat{\mathbf{e}}$  v závislosti na  $\mathbf{w}_{res} = (\mathbf{I} - \mathbf{H}) \mathbf{w}$

- uvažujme teď naopak, že  $\mathbf{w}$  je sloupec původní matice, např.  $\mathbf{x}_j^c$  a ozn.  $\mathbf{X}_{(-j)}$  matici  $\mathbf{X}$  bez sl.  $j$
- v předchozím modelu položíme  $\mathbf{X} = \mathbf{X}_{(-j)}$  a  $\mathbf{w} = \mathbf{x}_j^c$ , potom LSE parametru  $\beta_j$  je

$$\hat{\beta}_j = \frac{\hat{\mathbf{e}}_{(-j)}^T \mathbf{x}_{j,res}^c}{\|\mathbf{x}_{j,res}^c\|^2}, \quad \text{kde } \hat{\mathbf{e}}_{(-j)} \text{ jsou rezidua modelu bez } \mathbf{x}_j^c$$

a  $\mathbf{x}_{j,res}^c = (\mathbf{I} - \mathbf{H}_{(-j)}) \mathbf{x}_j^c$ , tedy jsou to rezidua modelu pro  $\mathbf{x}_j^c$  v závislosti na ostatních proměnných

**Added variable plot:** graf  $\hat{e}_{(-j)}$  proti  $x_{j,res}^c$ ,  $j = 1, \dots, m$

- pokud je model správný, body by měly být náhodně rozptýlené kolem přímky se směrnici  $\hat{\beta}_j$  procházející počátkem
- pokud závislost na  $x_j^c$  není lineární, projeví se to odklonem bodů od přímky
- funkce v **R**: `avPlots()` knihovna `car`

### PŘÍKLAD 4.3 (Housing Price Data)

```
mod <- lm(Price ~ Feet2 + Age)
avPlots(mod)
```

