M6120 – 6. cvičení : **M6120cv06** (*Porovnávání několika výběrů – ANOVA, vícenásobná srovnávání*)

A. Úvod

Předpokládejme, že $Y_{11}, \ldots, Y_{1n_1}$ je výběr z $\mathcal{L}(\mu_1, \sigma_1^2)$

$$\vdots$$

$Y_{p1}, \ldots, Y_{pn_p}$ je výběr z $\mathcal{L}(\mu_p, \sigma_p^2)$.

Nechť všechny výběry jsou stochasticky nezávislé.

Zaveďme tečkovou notaci $Y_{j\cdot} = \sum_{k=1}^{n_j} Y_{jk}$ pro $j = 1, \ldots, p$

$y_{j\cdot} = \frac{Y_{j\cdot}}{n_j}$

$n = n_1 + \cdots + n_p$

$Y_{\cdot\cdot} = Y_{1\cdot} + \cdots + Y_{p\cdot}$

$\bar{Y} = \frac{Y_{\cdot\cdot}}{n}$,

Analýza rozptylu je vlastně porovnání <u>středních hodnot</u> **více nezávislých náhodných výběrů z normálního rozdělení**.

Předpoklady platné pro každý náhodný výběr:

NORMALITA: pro $\forall j$ $\mathcal{L}(\mu_j, \sigma_j^2) = N(\mu_j, \sigma_j^2)$.

HOMOGENITA ROZPTYLU: pro $\forall j$ $\sigma_j^2 = \sigma^2$.

Ověření homogenity rozptylů se většinou provádí pomocí tzv. **Bartletova testu**, popř. pomocí **Levenova testu**.

<u>Bartletův test</u>:

Označme: $s_j^2 = \frac{1}{n_j - 1}\left(\sum_{k=1}^{n_j} Y_{jk}^2 - n_j y_{j\cdot}^2\right)$ $j = 1, \ldots, p$

$s^2 = \frac{1}{n_j - 1}\sum_{j=1}^{p}(n_j - 1)s_j^2$

$C = 1 + \frac{1}{3(p-1)}\left(\sum_{j=1}^{p}\frac{1}{n_j - 1} - \frac{1}{n-p}\right)$

Pak $B = \frac{1}{C}\left[(n-p)\ln s^2 - \sum_{j=1}^{p}\frac{1}{n_j - 1}\ln s_j^2\right] \overset{A}{\sim} \chi^2(n-p)$ $(\forall\, n_j > 6)$

má za platnosti nulové hypotézy přibližně χ^2-rozdělení o $n - p$ stupních volnosti, je-li $n_j > 6$ pro $j = 1, \ldots, p$.

Dostaneme-li

$$B \geq \chi^2_{1-\alpha}(p-1)$$

zamítneme hypotézu na hladině, která je přibližně rovna α.

Levenův test:

Označme: $Z_{jk} = |Y_{jk} - \bar{Y}^*_{j.}|$ kde $\bar{Y}^*_{j.}$ je výběrový průměr
výběrový medián
10% trimmed mean (ořezaný průměr)

Pak testová statistika má tvar

$$W = \frac{N-p}{p-1} \frac{\sum_{j=1}^{p} n_j (\bar{Z}_j - \bar{Z}_{..})^2}{\sum_{j=1}^{p} \sum_{k=1}^{n_p} (Z_{jk} - \bar{Z}_j)^2}$$

a má za platnosti nulové hypotézy přibližně F-rozdělení o $p-1$ a $n-p$ stupních volnosti.

Dostaneme-li

$$W \geq F_{1-\alpha}(p-1, n-p)$$

zamítneme hypotézu na hladině, která je přibližně rovna α.

B. Analýza rozptylu jako speciální případ lineárního regresního modelu

Základní situace je obdobná jako u dvouvýběrového t-testu, ale nyní je problém komplikován tím, že výběrů může být víc.

Je třeba testovat hypotézu

$$H_0 : \mu_1 = \cdots = \mu_p \quad \text{vs} \quad H_1 : \ \exists\, j \neq k : \mu_j \neq \mu_k$$

V případě zamítnutí této hypotézy bývá téměř vždy nutné najít všechny ty dvojice μ_j, μ_k, které toto zamítnutí způsobily (např. pomocí Tukeovy metody).

V analýze rozptylu vetšinou střední hodnotu j-tého výběru rozkládáme na součet společné střední hodnoty plus odchylky

$$\mu_j = \mu + a_j$$

a pak testujeme hypotézu, ze všechny odchylky jsou nulové.

Regresní model můžeme napsat ve tvaru

$$Y_{jk} = \mu + a_j + \varepsilon_{jk} \qquad \varepsilon_{jk} \sim \ iid\ N(0, \sigma^2), \quad j = 1, \cdots, p, \quad k = 1, \cdots, n_j.$$

Maticově, lze napsat $\boxed{\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}}$ a rozepíšeme-li matice podrobně, dostaneme model, který **není plné hodnosti**, neboť první sloupec je součtem všech ostatních. Říkáme, že model je **přeparametrizován** (tzv. *"overparameterized model"*).

$$\underbrace{\begin{pmatrix} Y_{11} \\ \vdots \\ \vdots \\ \underline{Y_{1n_1}} \\ Y_{21} \\ \vdots \\ \vdots \\ \underline{Y_{2n_2}} \\ \vdots \\ \vdots \\ \underline{Y_{p1}} \\ \vdots \\ \vdots \\ Y_{pn_p} \end{pmatrix}}_{\mathbf{Y}} = \underbrace{\begin{pmatrix} 1 & 1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 1 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \underline{1} & \underline{1} & \underline{0} & \underline{\cdots} & \underline{\cdots} & \underline{\cdots} & \underline{\cdots} & \underline{0} \\ 1 & 0 & 1 & 0 & \cdots & \cdots & \cdots & 0 \\ 1 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \underline{1} & \underline{0} & \underline{1} & \underline{\cdots} & \underline{\cdots} & \underline{\cdots} & \underline{\cdots} & \underline{0} \\ 1 & 0 & 0 & \cdots & \cdots & 0 & 1 & 0 \\ 1 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \underline{1} & \underline{0} & \underline{0} & \underline{\cdots} & \underline{\cdots} & \underline{0} & \underline{1} & \underline{0} \\ 1 & 0 & 0 & \cdots & \cdots & \cdots & 0 & 1 \\ 1 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & \cdots & \cdots & 0 & 1 \end{pmatrix}}_{\mathbf{X}} \underbrace{\begin{pmatrix} \mu \\ a_1 \\ \vdots \\ a_p \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \vdots \\ \underline{\varepsilon_{1n_1}} \\ \varepsilon_{21} \\ \vdots \\ \vdots \\ \underline{\varepsilon_{2n_2}} \\ \vdots \\ \vdots \\ \underline{\varepsilon_{p1}} \\ \vdots \\ \vdots \\ \varepsilon_{pn_p} \end{pmatrix}}_{\boldsymbol{\varepsilon}} \tag{1}$$

Protože matice plánu $\mathbf{X}$ není plné hodnosti, je možné vypustit některý ze sloupců, popř. provést reparametrizaci.

Často se některá z a_j zvolí jako **referenční** a položí

$$a_j = 0$$

(nejčastěji hned první), tím jeden sloupec vypadne a matice plánu je plné hodnosti.

Koeficient determinace

Předpokládejme, že v regresním modelu

$$\boxed{M}: \quad \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \text{kde} \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$$

matice plánu $\mathbf{X}$ (typu $n \times (p+1)$) má v prvním sloupci vektor jedniček. Pak velmi důležitou roli v regresní analýza hraje tzv. **nulový (minimální) model**, což je model ve tvaru

$$\boxed{M_0}: \quad Y_i = \beta_0 + \epsilon_i = \mu + \epsilon_i, \quad \text{kde} \quad \epsilon_i \sim \; iid \; N(0, \sigma_\epsilon^2), \quad \text{tj.} \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma_\epsilon^2\mathbf{I}_n).$$

Označíme–li matici plánu nulového modelu symbolem $\mathbf{X}_0 = \mathbf{1}_n$, kde $\mathbf{1}_n$ je jednotkový vektor, pak řešením normálních rovnic dostaneme

$$\mathbf{X}_0'\mathbf{X}_0\beta_0 = \mathbf{X}_0\mathbf{Y} \quad \Rightarrow \quad \mathbf{1}_n'\mathbf{1}_n\beta_0 = \mathbf{1}_n\mathbf{Y} \quad \Rightarrow \quad n\beta_0 = n\overline{Y} \quad \Rightarrow \quad \widehat{\beta}_0 = \widehat{\mu} = \overline{Y}.$$

Bývá zvykem v regresní analýze označovat

$$SSE=(\mathbf{Y}-\widehat{\mathbf{Y}})'(\mathbf{Y}-\widehat{\mathbf{Y}})=(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})=\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2 \qquad \text{sum of squares, error}$$

$$SST=(\mathbf{Y}-\widehat{\mathbf{Y}}_0)'(\mathbf{Y}-\widehat{\mathbf{Y}}_0)=(\mathbf{Y}-\overline{Y}\mathbf{1}_n)'(\mathbf{Y}-\overline{Y}\mathbf{1}_n)=\sum_{i=1}^{n}(Y_i-\overline{Y})^2 \qquad \text{sum of squares, total}$$

$$SSR=(\widehat{\mathbf{Y}}-\overline{Y}\mathbf{1}_n)'(\widehat{\mathbf{Y}}-\overline{Y}\mathbf{1}_n)=\sum_{i=1}^{n}(\hat{Y}_i-\overline{Y})^2 \qquad \text{sum of squares, regression}$$

Pak nestrannými odhady rozptylu σ^2_ϵ v minimálním modelu M_0 a σ^2 ve výchozím modelu M jsou v tomto značení

$$\widehat{\sigma}^2_\epsilon=\tfrac{SST}{n-1} \quad \text{a} \quad \widehat{\sigma}^2=\tfrac{SSE}{n-p-1}.$$

Protože minimální model $\boxed{M_0}$ je podmodelem výchozího modelu $\boxed{M}$, tak lze dokázat, že platí

$$SSR=SST-SSE \quad \Rightarrow \quad \boxed{SST=SSR+SSE}.$$

Koeficient determinace R^2 je vlastně výběrový korelační koeficient mezi $\mathbf{Y}$ a $\widehat{\mathbf{Y}}$ a ukazuje, jak velký díl výchozí variability hodnot závisle proměnné (charakterizované výrazem SST) se podařilo vysvětlit uvažovanou regresní závislostí. Nevysvětlená variabilita je dána reziduálním součtem čtverců SSE.

S využitím vztahu $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i$ se dá ukázat, že

$$R^2(\mathbf{Y},\widehat{\mathbf{Y}})=\frac{\left[\sum_{i=1}^n(Y_i-\overline{Y})(\widehat{Y}_i-\overline{Y})\right]^2}{\sum_{i=1}^n(Y_i-\overline{Y})^2\sum_{i=1}^n(\widehat{Y}_i-\overline{Y})^2}=\frac{\left[(\mathbf{Y}-\overline{Y}\mathbf{1}_n)'(\widehat{\mathbf{Y}}-\overline{Y}\mathbf{1}_n)\right]^2}{(\mathbf{Y}-\overline{Y}\mathbf{1}_n)'(\mathbf{Y}-\overline{Y}\mathbf{1}_n)(\widehat{\mathbf{Y}}-\overline{Y}\mathbf{1}_n)'(\widehat{\mathbf{Y}}-\overline{Y}\mathbf{1}_n)}$$

$$=\frac{\{[(\mathbf{Y}-\widehat{\mathbf{Y}})+(\widehat{\mathbf{Y}}-\overline{Y}\mathbf{1}_n)]'(\widehat{\mathbf{Y}}-\overline{Y}\mathbf{1}_n)\}^2}{(\mathbf{Y}-\overline{Y}\mathbf{1}_n)'(\mathbf{Y}-\overline{Y}\mathbf{1}_n)(\widehat{\mathbf{Y}}-\overline{Y}\mathbf{1}_n)'(\widehat{\mathbf{Y}}-\overline{Y}\mathbf{1}_n)}=\frac{[(\widehat{\mathbf{Y}}-\overline{Y}\mathbf{1}_n)'(\widehat{\mathbf{Y}}-\overline{Y}\mathbf{1}_n)]^2}{(\mathbf{Y}-\overline{Y}\mathbf{1}_n)'(\mathbf{Y}-\overline{Y}\mathbf{1}_n)(\widehat{\mathbf{Y}}-\overline{Y}\mathbf{1}_n)'(\widehat{\mathbf{Y}}-\overline{Y}\mathbf{1}_n)}$$

$$=\frac{(\widehat{\mathbf{Y}}-\overline{Y}\mathbf{1}_n)'(\widehat{\mathbf{Y}}-\overline{Y}\mathbf{1}_n)}{(\mathbf{Y}-\overline{Y}\mathbf{1}_n)'(\mathbf{Y}-\overline{Y}\mathbf{1}_n)}=\frac{SSR}{SST}=1-\frac{SSE}{SST}=R^2$$

Označíme–li vychýlené varianty odhadů příslušných rozptylů symboly

$$\widetilde{\sigma}^2_\epsilon=\tfrac{SST}{n} \quad \text{a} \quad \widetilde{\sigma}^2=\tfrac{SSE}{n},$$

pak můžeme psát

$$R^2=1-\frac{SSE/n}{SST/n}=1-\frac{\widetilde{\sigma}^2}{\widetilde{\sigma}^2_\epsilon}.$$

Nahradíme–li v tomto vzorci vychýlené odhady rozptylů nevychýlenými, dostaneme tzv. **upravený (adjustovaný) koeficient determinace**

$$R^2_{adj}=1-\frac{\widehat{\sigma}^2}{\widehat{\sigma}^2_\epsilon}=1-\tfrac{n-1}{n-p-1}(1-R^2).$$

S ohledem na rozklad celkové sumy SST na součet dvou složek SSR a SSE bývá zvykem jako výstup regresní analýzy nabízet tzv. ANOVA tabulku ve formě

Source	*df*	*SS*	*MS*	*F*	*p-valule*
Total	$n-1$	SST			
Regression	p	SSR	$MSR=\frac{SSR}{p}$	$\frac{MSR}{MSE}$	$P\left(F>\frac{MSR}{MSE}\right)$
Residual	$n-p-1$	SSE	$MSE=\frac{SSE}{n-p-1}$		

Statistika F má za platnosti nulové hypotézy $(\beta_1,\ldots,\beta_p)'=(0,\ldots,0)'$ F– rozdělení o p a $n-p-1$ stupních volnosti.

PŘÍKLAD 1: *Blood coagulation times*

Klinickou studií byla sledována závislost mezi dietou a dobou, za kterou dojde ke koagulaci krve.

Data jsou uložena ve dvou souborech, v prvním je popis, ve druhém samotná data. Nejprve načteme popisný soubor

```
> fileTxt <- paste(data.library, "blood.inf", sep = "")
> con <- file(fileTxt)
> (popis <- readLines(con))
```

```
[1] "Blood coagulation times"
[2] "The data consist of blood coagulation times for 24 animals"
[3] "fed one of 4 different diets."
```

```
> close(con)
```

Protože data ke každé dietě jsou na samostatném řádku, řádky načteme po jednom postupně.

```
> fileDat <- paste(data.library, "blood.txt", sep = "")
> diet1 <- scan(fileDat, nlines = 1)
> diet2 <- scan(fileDat, skip = 1, nlines = 1)
> diet3 <- scan(fileDat, skip = 2, nlines = 1)
> diet4 <- scan(fileDat, skip = 3, nlines = 1)
```

Pro identifikaci diety nejprve vytvoříme proměnnou `diet` typu faktor, následně datový rámec. Abychom viděl vstupní data, vypíšeme je pomocí funkce `cat`.

```
> gr <- factor(c(rep(1, length(diet1)), rep(2, length(diet2)), rep(3,
    length(diet3)), rep(4, length(diet4))))
> TxtGr <- "diet"
> ngr <- 4
> y <- c(diet1, diet2, diet3, diet4)
> TxtY <- "Blood coagulation times"
> data <- data.frame(y = y, gr = gr)
> str(data)
```

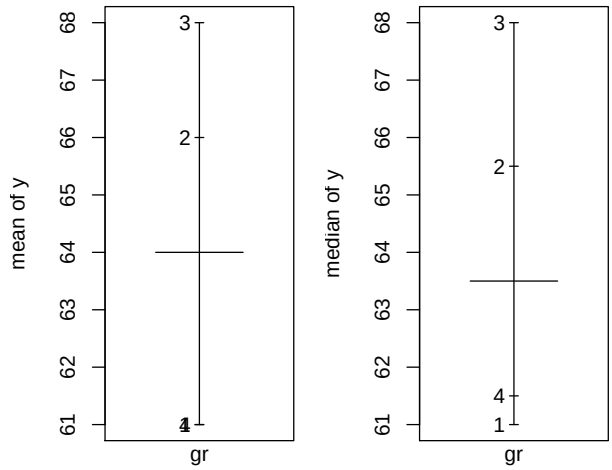
```
,data.frame,:   24 obs. of  2 variables:
 $ y : num  62 60 63 59 63 67 71 64 65 66 ...
 $ gr: Factor w/ 4 levels "1","2","3","4": 1 1 1 1 2 2 2 2 2 2 ...
```

```
> for (igr in as.character(1:ngr)) cat(paste(TxtGr, igr, ":", paste(y[gr ==
    igr], collapse = ", "), "\n"))
```

```
diet 1 : 62, 60, 63, 59
diet 2 : 63, 67, 71, 64, 65, 66
diet 3 : 68, 66, 71, 67, 68, 68
diet 4 : 56, 62, 60, 61, 63, 64, 63, 59
```

Zajímavým grafem je `plot.design()`, který pro jednotlivé skupiny (v našem případě diety) podle zadání vykreslí buď polohy výběrových středních hodnot (implicitně nastaveno) či výběrových mediánů.

```
> par(mfrow = c(1, 2), mar = c(2.25, 3.5, 0, 1) + 0.25)
> plot.design(data)
> plot.design(data, fun = median)
> par(mfrow = c(1, 1))
```



Obrázek 1: `plot.design` grafy pro data *Blood coagulation times*

Vizuální představu o hodnotách výběrových průměrů a mediánů doplníme o numerické hodnoty pomocí příkazů `tapply()`.

```
> with(data, tapply(y, gr, mean))
```

```
 1  2  3  4
61 66 68 61
```

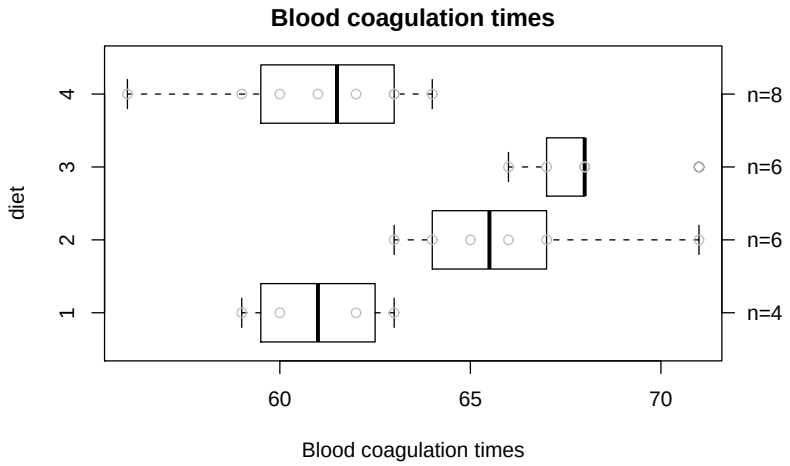
```
> with(data, tapply(y, gr, median))
```

```
   1    2    3    4
61.0 65.5 68.0 61.5
```

Ověřování homogenity rozptylu

Abychom graficky posoudily homogenitu rozptylu, použijeme příkaz `boxplot()`.

```
> par(mar = c(5, 5, 2, 5) + 0.25)
> boxplot(y ~ gr, data = data, horizontal = TRUE, main = popis[1], xlab = TxtY,
      ylab = TxtGr)
> points(y, gr, col = "gray")
> axis(4, at = 1:ngr, labels = paste("n=", table(data$gr), sep = ""),
      las = 2)
```



Obrázek 2: Krabicové grafy pro data *Blood coagulation times*

Na základě krabicových grafů nevypadá variabilita všech diet stejně. Máme však k dispozici příliš malé výběry.

Vedle grafického posouzení provedeme ještě Bartlettův a Levenův test.

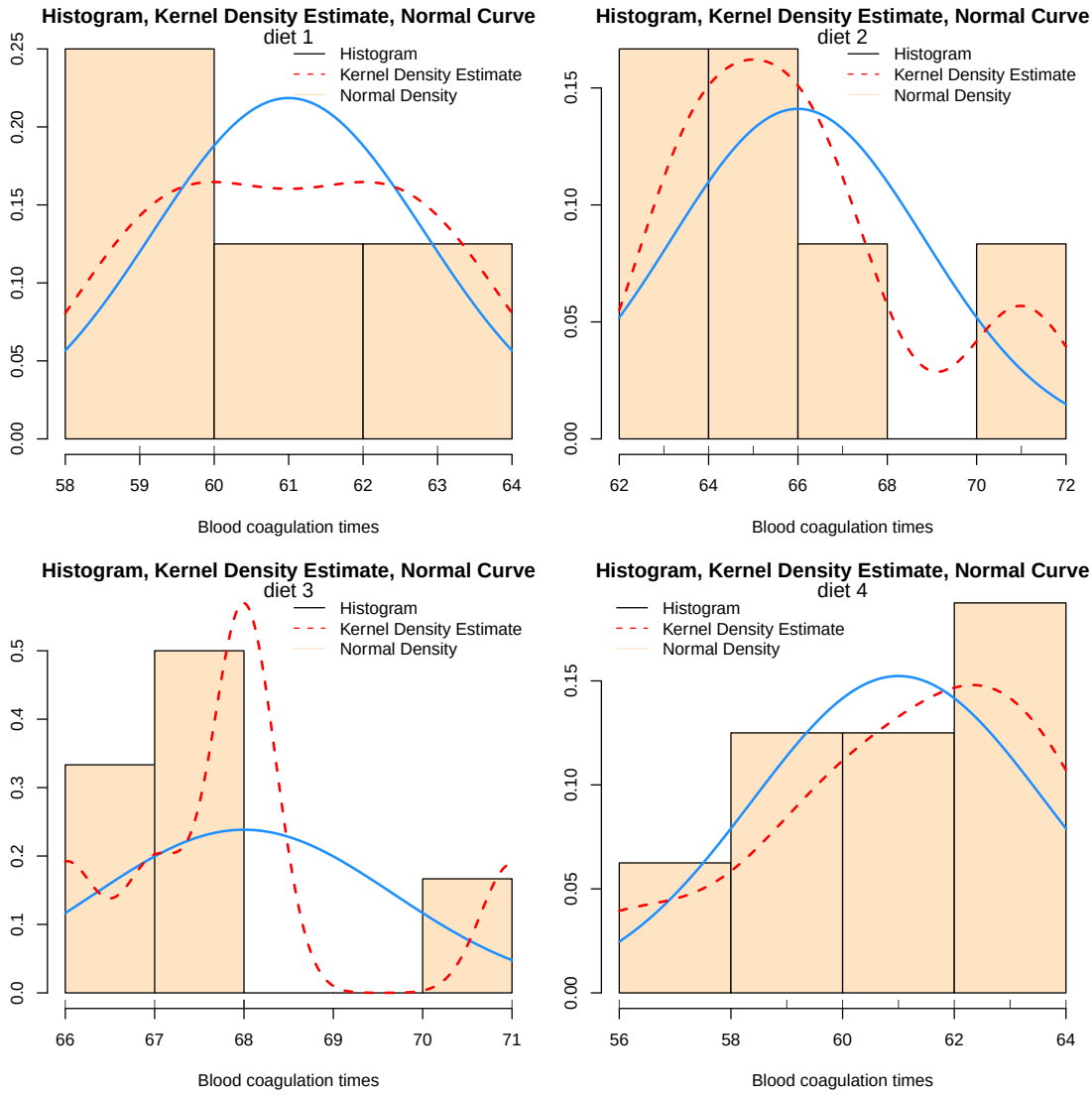
```
> bartlett.test(y ~ gr, data = data)


    Bartlett test of homogeneity of variances

data:  y by gr
Bartlett,s K-squared = 1.668, df = 3, p-value = 0.6441


> library(car)
> leveneTest(y ~ gr, data = data)


Levene,s Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  3  0.6492 0.5926
      20
```

Protože p–hodnota není ani v jednom případě menší než 0.05, hypotézu o shodnosti rozptylů **nezamítáme**.

Ověřování normality

Pomocí funkce `HistFit()` (skript `FunkceM6120.R`) vykreslíme histogram, jádrový odhad hustoty a odhad hustoty normálního rozdělení.

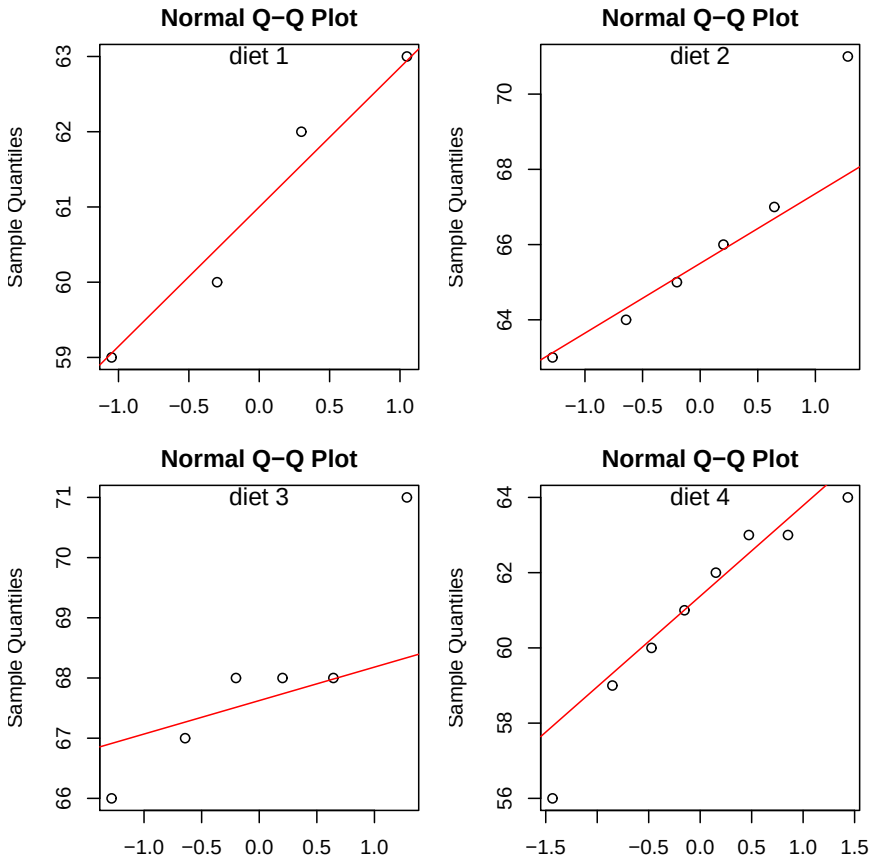
```
> par(mfrow = c(2, 2))
> for (igr in as.character(1:ngr)) {
     HistFit(y[gr == igr], xlab = TxtY)
     mtext(paste(TxtGr, igr), line = -0.7)
 }
> par(mfrow = c(1, 1))
```



Obrázek 3: `HistFit` grafy pro data *Blood coagulation times*

Dále ověříme normalitu pomocí Q-Q grafů.

```
> par(mfrow = c(2, 2), mar = c(2.25, 3.5, 2, 1) + 0.25)
> for (igr in as.character(1:ngr)) {
      qqnorm(y[gr == igr])
      qqline(y[gr == igr], col = "red")
      mtext(paste(TxtGr, igr), line = -1)
  }
> par(mfrow = c(1, 1))
```



Obrázek 4: `qqnorm` grafy pro data *Blood coagulation times*

Normalitu nakonec otestujeme pomocí Shapiro–Wilkova testu.

```
> for (igr in as.character(1:ngr)) {
     TXT <- paste(TxtY, ":", TxtGr, igr, "\n", sep = "")
     oddel <- paste(paste(rep("=", nchar(TXT) - 1), collapse = ""), "\n",
         sep = "")
     cat(paste(oddel, TXT, oddel, sep = ""))
     print(shapiro.test(y[gr == igr]))
 }


==============================
Blood coagulation times:diet1
==============================

    Shapiro-Wilk normality test

data:  y[gr == igr]
W = 0.9497, p-value = 0.7143

==============================
Blood coagulation times:diet2
==============================
```

```
    Shapiro-Wilk normality test

data:  y[gr == igr]
W = 0.9224, p-value = 0.5227

============================
Blood coagulation times:diet3
============================

    Shapiro-Wilk normality test

data:  y[gr == igr]
W = 0.8728, p-value = 0.2375

============================
Blood coagulation times:diet4
============================

    Shapiro-Wilk normality test

data:  y[gr == igr]
W = 0.9317, p-value = 0.5319
```

Protože p–hodnoty Shapiro–Wilkova testu u všech diet nejsou menší než 0.05, **nezamítáme** normalitu. Obdobně dopadla homogenita rozptylu, takže můžeme provádět analýzu rozptylu pomocí klasického regresního modelu.

Analýza rozptylu pomocí funkce `lm()`

Pro analýzu rozptylu jednoduchého třídění můžeme využít funkci `lm()`.

```
> model.lm <- lm(y ~ gr, data = data)
```

Výsledek je uložen do objektu, který jsme nazvali `model.lm`. Objekt je tvořen položkami, jejichž názvy můžeme získat příkazem `names()`.

```
> names(model.lm)
```

```
 [1] "coefficients"  "residuals"     "effects"       "rank"          "fitted.values"
 [6] "assign"        "qr"            "df.residual"   "contrasts"     "xlevels"
[11] "call"          "terms"         "model"
```

Pomocí příkazu `coefficients()` (lze použít i kratší verzi `coef`) vypíšeme hodnoty ohadnadnutých neznámých parametrů.

```
> coef(model.lm)
```

```
  (Intercept)           gr2           gr3           gr4
 6.100000e+01  5.000000e+00  7.000000e+00 -1.071287e-14
```

Abychom je však mohli vůbec interpretovat, potřebujeme znát kódování jednotlivých úrovní. Proto si pomocí příkazu `model.matrix()` vypíšeme matici plánu.

```
> model.matrix(model.lm)
```

```
   (Intercept) gr2 gr3 gr4
1            1   0   0   0
2            1   0   0   0
3            1   0   0   0
4            1   0   0   0
5            1   1   0   0
6            1   1   0   0
7            1   1   0   0
8            1   1   0   0
9            1   1   0   0
10           1   1   0   0
11           1   0   1   0
12           1   0   1   0
13           1   0   1   0
14           1   0   1   0
15           1   0   1   0
16           1   0   1   0
17           1   0   0   1
18           1   0   0   1
19           1   0   0   1
20           1   0   0   1
21           1   0   0   1
22           1   0   0   1
23           1   0   0   1
24           1   0   0   1
attr(,"assign")
[1] 0 1 1 1
attr(,"contrasts")
attr(,"contrasts")$gr
[1] "contr.treatment"
```

Vidíme, že standardní (implicitní, default) nastavení **kontrastu** v prostředí R je kontrast typu `contr.treatment`, takže funkce `lm()` uvažuje lineární regresní model ve tvaru

$$Y_{jk} = \mu + a_j + \varepsilon_{jk} \qquad \text{kde} \qquad a_1 = 0.$$

Podrovnější informace o odhadnutém lineárním modelu získáme pomocí příkazu `summary()`.

```
> summary(model.lm)
```

```
Call:
lm(formula = y ~ gr, data = data)

Residuals:
       Min         1Q     Median         3Q        Max
-5.000e+00 -1.250e+00  1.488e-16  1.250e+00  5.000e+00

Coefficients:
             Estimate Std. Error   t value Pr(>|t|)
(Intercept)  6.100e+01  1.183e+00    51.554  < 2e-16 ***
gr2          5.000e+00  1.528e+00     3.273 0.003803 **
gr3          7.000e+00  1.528e+00     4.583 0.000181 ***
gr4         -1.071e-14  1.449e+00 -7.39e-15 1.000000
```

```
---
Signif. codes:  0 ,***, 0.001 ,**, 0.01 ,*, 0.05 ,., 0.1 , , 1

Residual standard error: 2.366 on 20 degrees of freedom
Multiple R-squared: 0.6706, Adjusted R-squared: 0.6212
F-statistic: 13.57 on 3 and 20 DF,  p-value: 4.658e-05
```

V odstavci `Coefficients` je vždy vedle bodového odhadu j-tého koeficientu také střední chyba tohoto odhadu $s\sqrt{v_{jj}}$ ($s^2 = \frac{SSE}{n-p-1}$, v_{jj} je diagonální prvek matice $(\mathbf{X}'\mathbf{X})^{-1}$), dále testová statistika $T_j = \frac{\hat{\beta}_j}{s\sqrt{v_{jj}}}$ pro test nulové hybotézy, že j-tý koeficient je nulový, a nakonec dosažená p–hodnota při oboustranné alternativě.

Odhad uvedený v řádku `(Intercept)` je odhadem střední hodnoty referenční skupiny. Ostatní parametry značí odchylky od referenční skupiny. Protože p–hodnoty u proměnných `diet2` a `diet3` jsou velmi malé, obě dvě se významně liší od první diety. U první a poslední diety se rozdílnost neprokázala.

Pod označením `Residual standard error` je statistika s, dále následuje koeficient determinace R^2 a upravený (adjustovaný) koeficient determinace R^2_{adj}. Koeficient determinace ukazuje, jak velký díl výchozí variability hodnot závisle proměnné se podařilo vysvětlit uvažovanou regresní závislostí.

Abychom nemuseli sami počítat odhady středních hodnot pro jednotlivé diety (na základě hodnot coeficientů označených ve výstupu `(Intercept)`, `diet2`, `diet3` a `diet4`), použijeme k tomu příkaz `predict()`.

```
> predict(model.lm, newdata = data.frame(gr = as.character(1:ngr)))


 1  2  3  4
61 66 68 61
```

Podívejme se, jakou formu ANOVA tabulky nabízí prostředí R pomocí funkce `anova()`.

```
> anova(model.lm)


Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
gr         3    228    76.0  13.571 4.658e-05 ***
Residuals 20    112     5.6
---
Signif. codes:  0 ,***, 0.001 ,**, 0.01 ,*, 0.05 ,., 0.1 , , 1
```

Protože p-hodnota u statistiky F v ANOVA tabulce je menší než 0.05, zamítáme nulovou hypotézu, že vektor $(a_2, a_3, a_4)' = (0, 0, 0)'$, a lze říci, **že doba koagulace krve se liší v závislosti na podávané dietě**.

Analýza rozptylu pomocí funkce `aov()`

Kromě obecné funkce `lm()` pro klasickou lineární regresi lze použít speciální funkci jen pro analýzu rozptylu, a to funkci `aov()`.

```
> model.aov <- aov(y ~ gr, data = data)
```

Výsledek je uložen do objektu, který jsme nazvali `model.aov`. Objekt je tvořen položkami, jejichž názvy můžeme získat příkazem `names()`.

```
> names(model.aov)
```

```
 [1] "coefficients"  "residuals"     "effects"       "rank"          "fitted.values"
 [6] "assign"        "qr"            "df.residual"   "contrasts"     "xlevels"
[11] "call"          "terms"         "model"
```

Opět ke zjištní kódování jednotlivých úrovní použijeme příkaz

```
> model.matrix(model.aov)
```

```
   (Intercept) gr2 gr3 gr4
1            1   0   0   0
2            1   0   0   0
3            1   0   0   0
4            1   0   0   0
5            1   1   0   0
6            1   1   0   0
7            1   1   0   0
8            1   1   0   0
9            1   1   0   0
10           1   1   0   0
11           1   0   1   0
12           1   0   1   0
13           1   0   1   0
14           1   0   1   0
15           1   0   1   0
16           1   0   1   0
17           1   0   0   1
18           1   0   0   1
19           1   0   0   1
20           1   0   0   1
21           1   0   0   1
22           1   0   0   1
23           1   0   0   1
24           1   0   0   1
attr(,"assign")
[1] 0 1 1 1
attr(,"contrasts")
attr(,"contrasts")$gr
[1] "contr.treatment"
```

Vidíme, že kódování je úplně stejné jako u funkce `lm()`, neboť implicitně jsou kontrasty nastaveny na `contr.treatment`. Požijeme–li příkaz `summary` v souvislosti s výstupy z funkce `aov()`, dostaneme výstupní ANOVA tabulku.

```
> summary(model.aov)
```

```
            Df Sum Sq Mean Sq F value    Pr(>F)
gr           3    228    76.0  13.571 4.658e-05 ***
Residuals   20    112     5.6
---
Signif. codes:  0 ,***, 0.001 ,**, 0.01 ,*, 0.05 ,., 0.1 , , 1
```

Chceme–li znát odhady středních hodnot jednotlivých diet, použijeme pro výstup z funkce `aov()` příkaz `model.tables()`

```
> model.tables(model.aov, type = "means")
```

```
Tables of means
Grand mean

64

 gr
    1  2  3  4
   61 66 68 61
rep  4  6  6  8
```

Mnohonásobná srovnávání

Pokud zamítneme nulovou hypotézu

$$H_0 : \mu_1 = \cdots = \mu_p \quad \text{vs} \quad H_1 : \ \exists\, j \neq k : \mu_j \neq \mu_k$$

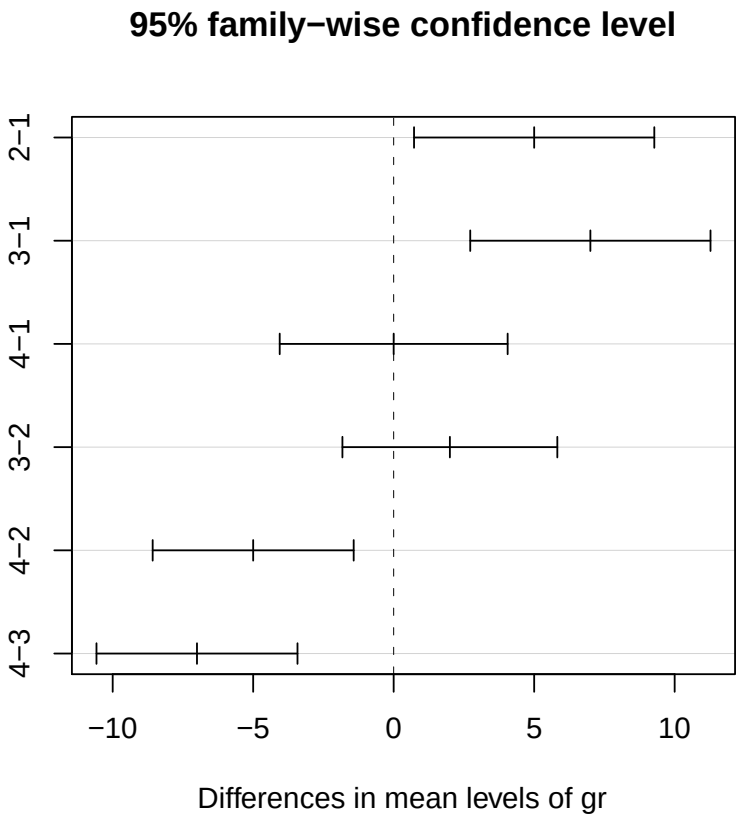
ve prospěch alternativní hypotézy, zajímá nás, které dvojice se významně liší. K tomu použijeme testy mnohonásobného srovnávání.

```
> print(TukeyTest <- TukeyHSD(model.aov))
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = y ~ gr, data = data)

$gr
             diff         lwr       upr     p adj
2-1  5.000000e+00   0.7245544  9.275446 0.0183283
3-1  7.000000e+00   2.7245544 11.275446 0.0009577
4-1 -1.421085e-14  -4.0560438  4.056044 1.0000000
3-2  2.000000e+00  -1.8240748  5.824075 0.4766005
4-2 -5.000000e+00  -8.5770944 -1.422906 0.0044114
4-3 -7.000000e+00 -10.5770944 -3.422906 0.0001268
```

Na závěr vysledky Tukeova testu mnohonásobného srovnávání ještě vykreslíme.

```
> plot(TukeyTest)
```



Obrázek 5: Tukeova metoda mnohonásobného srovnávání pro data *Blood coagulation times*

Z výsledků je patrné, že se

(a) **prokázal** významný rozdíl

- mezi 1. a 2. dietou
- mezi 1. a 3. dietou
- mezi 2. a 4. dietou
- mezi 3. a 4. dietou

nebot' intervaly spolehlivosti neobsahují nulu,

(b) **neprokázal** významný rozdíl

- mezi 1. a 4. dietou
- mezi 2. a 3. dietou

nebot' intervaly spolehlivosti **obsahují nulu**.

C. Úkol 1:

(a) Načtěte soubor informací `brambory.inf` a soubor dat `brambory.txt`.

(b) Před samotnou analýzou rozptylu proveďte ověření homogenity rozptylu a normality.

(c) Testujte hypotézu, že výnosy všech odrůd brambor jsou stejné.

(d) Pokud se prokáže významný rozdíl, nalezněze ty dvojice odrůd, které se významně liší.

D. Úkol 2:

(a) Načtěte soubor informací `b515.inf` (s kódováním utf8), resp. `b515w.inf` (s kódováním cp 1250) a soubor dat `b515.txt`.

(b) Pro tato data proveďte úkoly (b) až (d) z předchozího úkolu.

Poznámka:
V prvním řádku souboru `b515.txt` je uveden pro jednotlivé druhy myší počet jedinců. Chybějící hodnoty (*missing values*) jsou označeny jako nuly.