Kraniometrie klokanů

Logistická regrese

Tomáš Spurný

Nahrání dat, popis dat

Byly nahrány knihovny, konstantní proměnné MOTIV, VARIABLES, POHLAVI, a vytvořeny dvě tabulky, jedna dlouhého formátu dfL a druhá tzv. wide formátu dfW, kreré pojímali jen 6 původních proměnných.

Byly vypočítány výběrové průměry a výběrové směrodatné odchylky pro skupiny proměnných dělených dle pohlaví. Výsledek je v následující tabulce

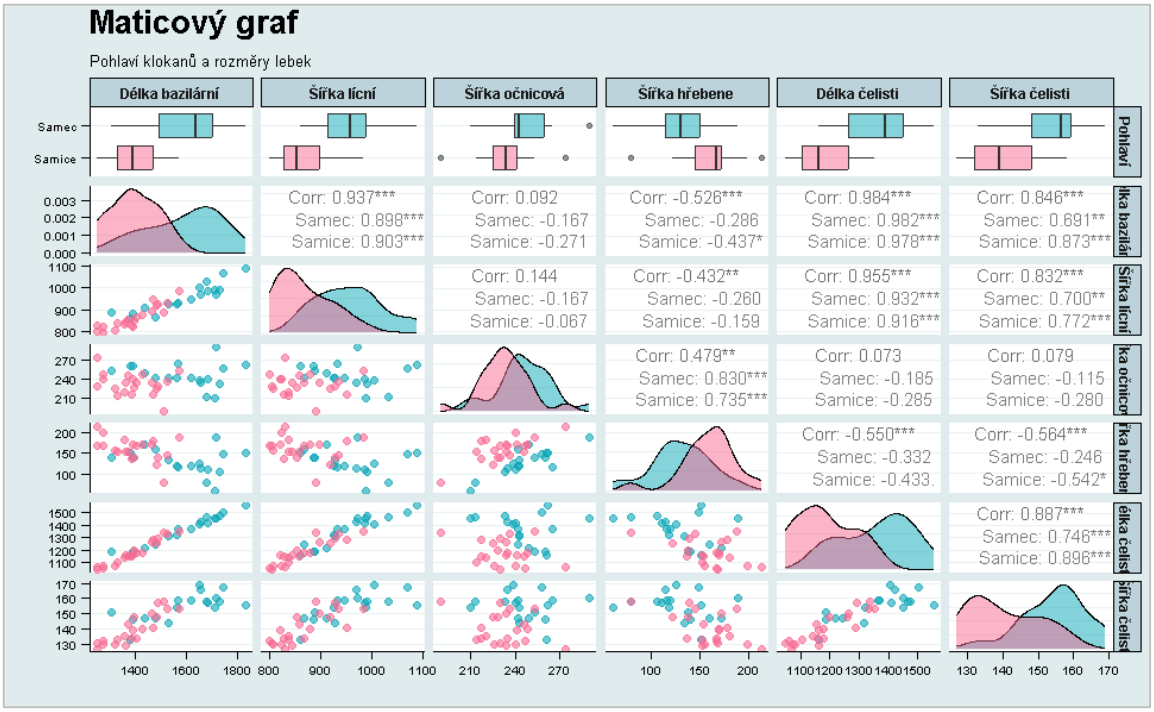
Popisná statistika skupin dle pohlaví

Veličina	Výběrové průměry		Výběrové sm. Odchylky	
	Samci	Samice	Samci	Samice
Délka bazilární	1593,6	1394,6	146,75	93,75
Délka čelisti	1361,8	1184,7	119,66	98,12
Šířka čelisti	154,7	140,2	8,98	9,92
Šířka hřebene	129,5	159,5	31,50	28,73
Šířka lícní	958,3	866,4	65,59	52,32
Šířka očnicová	246,6	232,6	18,76	17,09
Počet ve skupinách:	18	21		

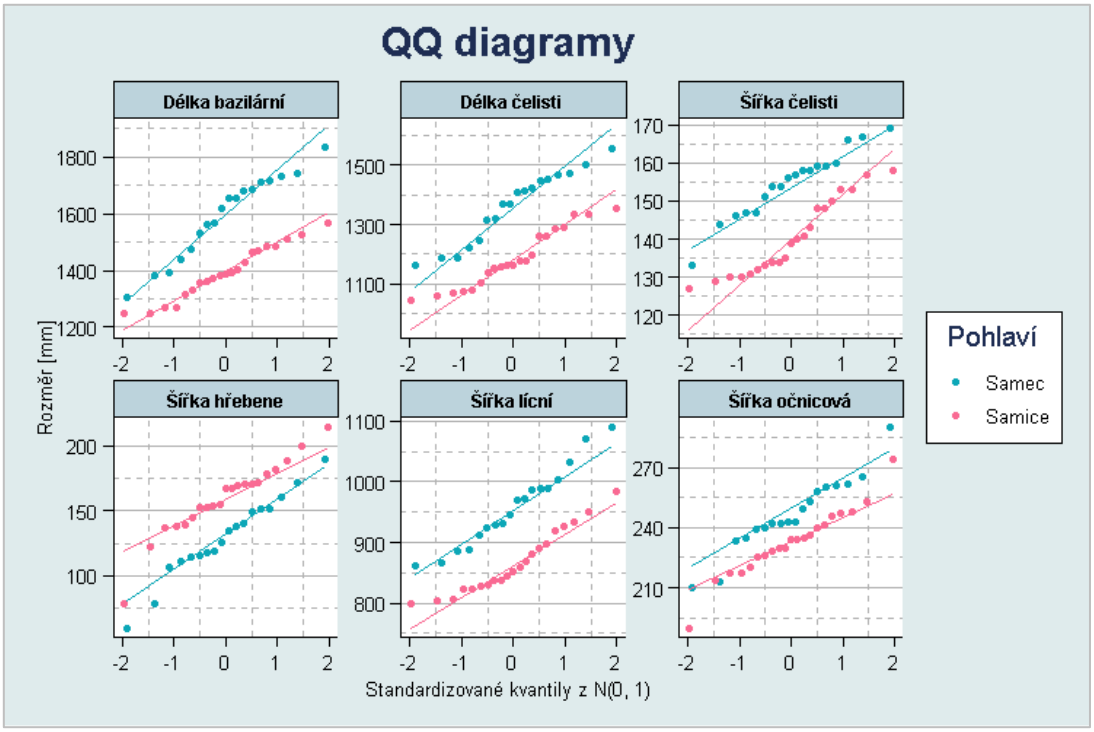
Samci mají vždy větší rozměr lebeční části než samice, krom šířky hřebene. Také mají větší variabilitu v datech, až na šířku čelisti.

Porovnávání skupin samci-samice

Pro zběžný přehled byl vykreslen maticový graf s XY diagramy, boxploty, korelacemi a hustotami rozdělení.



K porovnávání skupin bylo nejprve nutno ověřit předpoklady T-testů. Nejprve normalita graficky na QQ plotech:



Normalita by mohla být splněna, ale existuje podezření, že by některé body mohly být odlehlé, proto byl proveden před T-testy Shapiro-Wilkův test. Stojí za povšimnutí, že proložené přímky u mužů mají větší sklon (-> variabilitu) a jen u šířky hřebene leží křivka samců pod křivkou samic. Snad jen přímky u lícní šířky jsou skoro rovnoběžné.

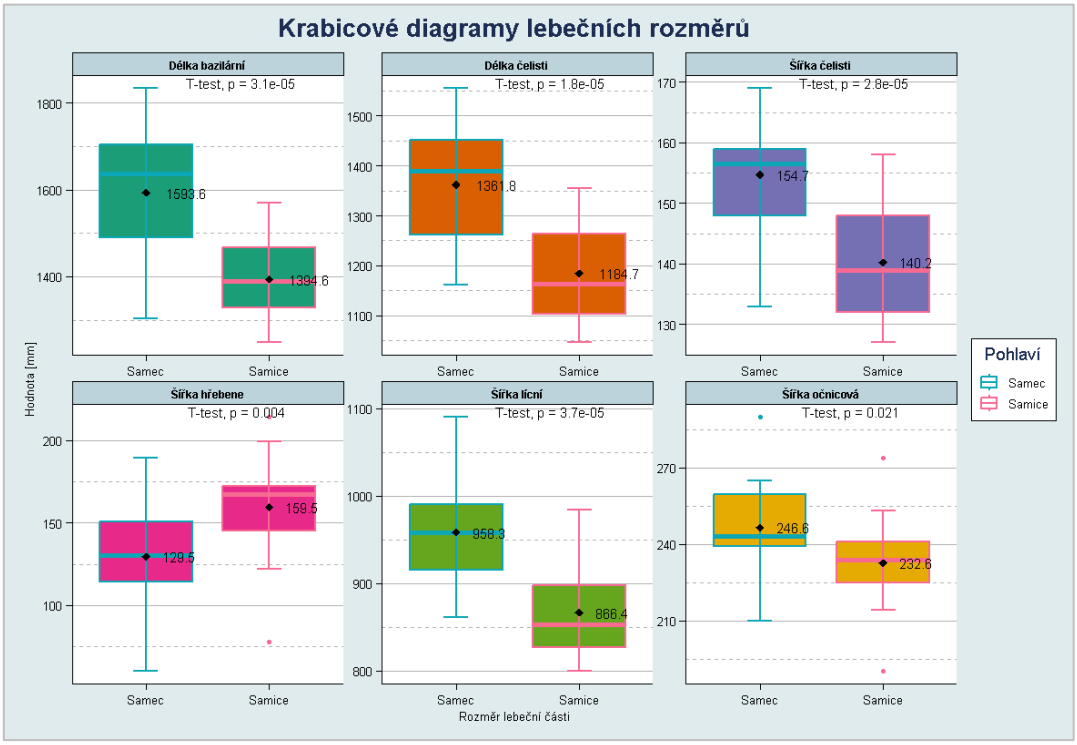
Pro každou skupinu byla následně provedena trojice testů, jejichž p-hodnoty jsou uvedeny v následující tabulce. V algoritmu byl zohledněn výsledek F-testu na homogenitu rozptylů. Hodnoty, které jsou nejblíž hodnotě spolehlivosti (shora i zdola) jsou tučně.

P-hodnoty dvouvýběrových t-testů shody i s p-hodnotami předpokladů

Skupina		Předpoklad		Dvouvýběrový T-test
Proměnná	Pohlaví	SW test normality	F-test stejnosti rozptylů	
Délka bazilární	Samec	0,4990	**0,0575**	0,00001
	Samice	0,6391		
Šířka lícní	Samec	0,6997	0,3315	0,00002
	Samice	0,1564		
Šířka očnicová	Samec	0,4498	0,6845	**0,02020**
	Samice	0,6359		
Šířka hřebene	Samec	0,9136	0,6874	0,00362
	Samice	0,2745		
Délka čelisti	Samec	0,2745	0,3931	0,00001
	Samice	0,1496		
Šířka čelisti	Samec	0,5350	0,6836	0,00003
	Samice	**0,0716**		

Žádný test normalitu nezamítl. Proto nebyl problém přistoupit k F-testům, které taktéž nezamítly nulové hypotézy o shodě rozptylů. Naproti tomu Každý T-test zamítl, že by se skupiny v daném lebečním rozměru shodovali v průměru.

Vyneseno do boxplotů, lze pozorovat rozdíly v polohách mezi klokany a klokanicemi.



Modelování logistické regrese

PLNÝ MODEL

Cílem je předpovědět pohlaví jedince na základě ostatních spojitých proměnných. Nejprve byl sestrojen plný (bez interakcí) a nulový model a navzájem otestovány F-testem.

Plný model

Člen	Odhad	Sm. Chyba	Statistika	P-hodnota	e^{Odhad}
(Intercept)	308,270	254,481	1,2114	0,2258	
Délka bazilární	-0,153	0,150	-1,0261	0,3049	0,86
Šířka lícní	0,062	0,058	1,0683	0,2854	1,06
Šířka očnicová	-0,944	0,660	-1,4297	0,1528	**0,39**
Šířka hřebene	0,348	0,236	1,4749	0,1402	**1,42**
Délka čelisti	0,138	0,171	0,8053	0,4206	1,15
Šířka čelisti	-0,921	0,994	-0,9271	0,3539	0,40

Dle tohoto modelu nejvíce zvyšuje šanci na to, že daná lebka bude samičí, kladný přírustek v šířce hřebene. Nejvíce ji snižuje kladný přírůstek v očnicové šířce.

Nicméně všechny dílčí Waldovy statistiky se realizovali mimo kritický obor, tzn. v absolutní hodnotě do 1,96. Takže žádný člen dle nich není významný.

Plný model byl porovnán s nulovým. Dle p-hodnoty (2,68e-08) je významný oproti modelu konstanty.

STEPWISE PROCEDURY

Jak dopředné, tak zpětné stepwise procedury dospěly k modelu o třech členech bez interakcí: **Bazilární délce, Očnicové šířce a šířce hřebene**. P-hodnoty dílčích Waldových statistik se alespoň přiblížili k hranici 0,05.

Stepwise (pod)model

Člen	Odhad	Sm. Chyba	Statistika	P-hodnota	e^{Odhad}
(Intercept)	85,193	46,639	1,8267	0,0678	
Délka bazilární	-0,015	0,011	-1,4051	0,1600	0,99
Šířka očnicová	-0,356	0,195	-1,8300	0,0673	0,70
Šířka hřebene	0,153	0,084	1,8274	0,0676	1,16

Trendy jsou podobné jako v plném modelu (až na přesnou hodnotu koeficientů).

SROVNÁNÍ MODELŮ

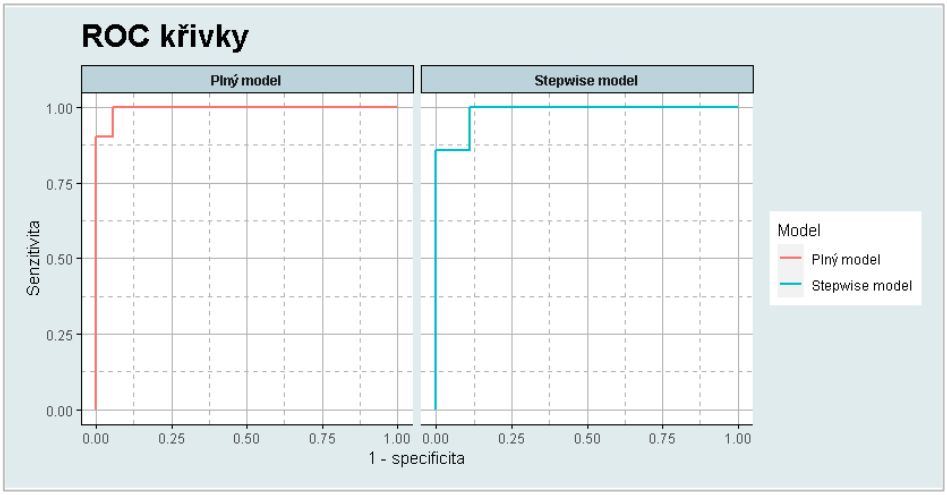
Pro připomenutí, v datech je 18 samců a 21 samic klokanů…. Pomocí for-cyklu byly zjištěny různá kritéria informující o vhodnosti modelů.

Tabulka srovnání modelů

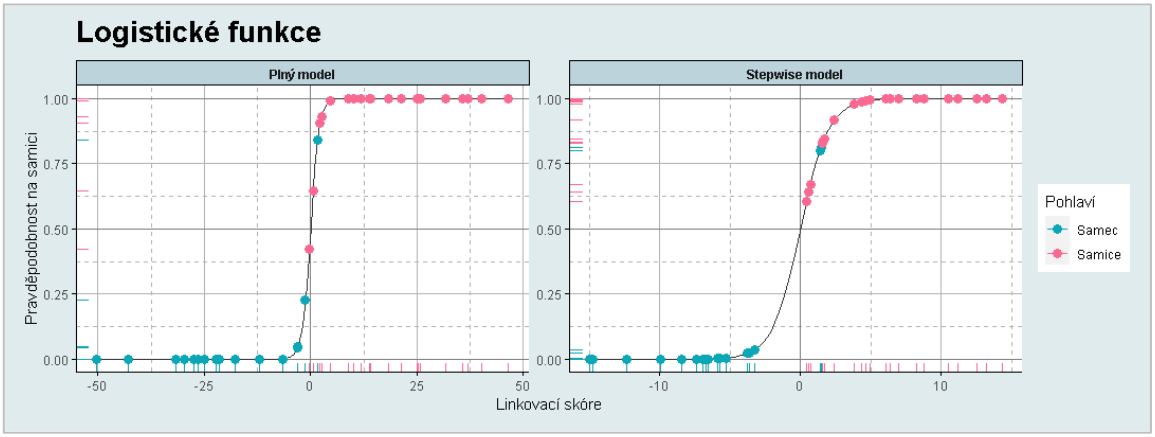
Model	Nagelkerkeho koeficient	AIC	Deviance	Určení dle modelu v 4-polní tabulce					AUC
				Samci správně	Samice správně	Samec jako Samice	Samice jako Samec	Úspěch zařazení	
Stepwise podmodel	0,8925	**18,82**	10,82	16	21	2	0	94,87 %	0,984
Plný model	**0,9275**	21,61	**7,62**	17	20	1	1	94,87 %	**0,995**

Jelikož se v obou případech špatně určili celkem dva klokaní jedinci špatně je úspěšnost rozřazení stejná. Pro plný model mluví Nagelkerkeho koeficient (vysvětluje mírně více variability) a hodnota AUC, která je ale v obou případech výborná. Deviance je z teorie nižší. Pro stepwise podmodel hovoří Akaikovo kritérium.

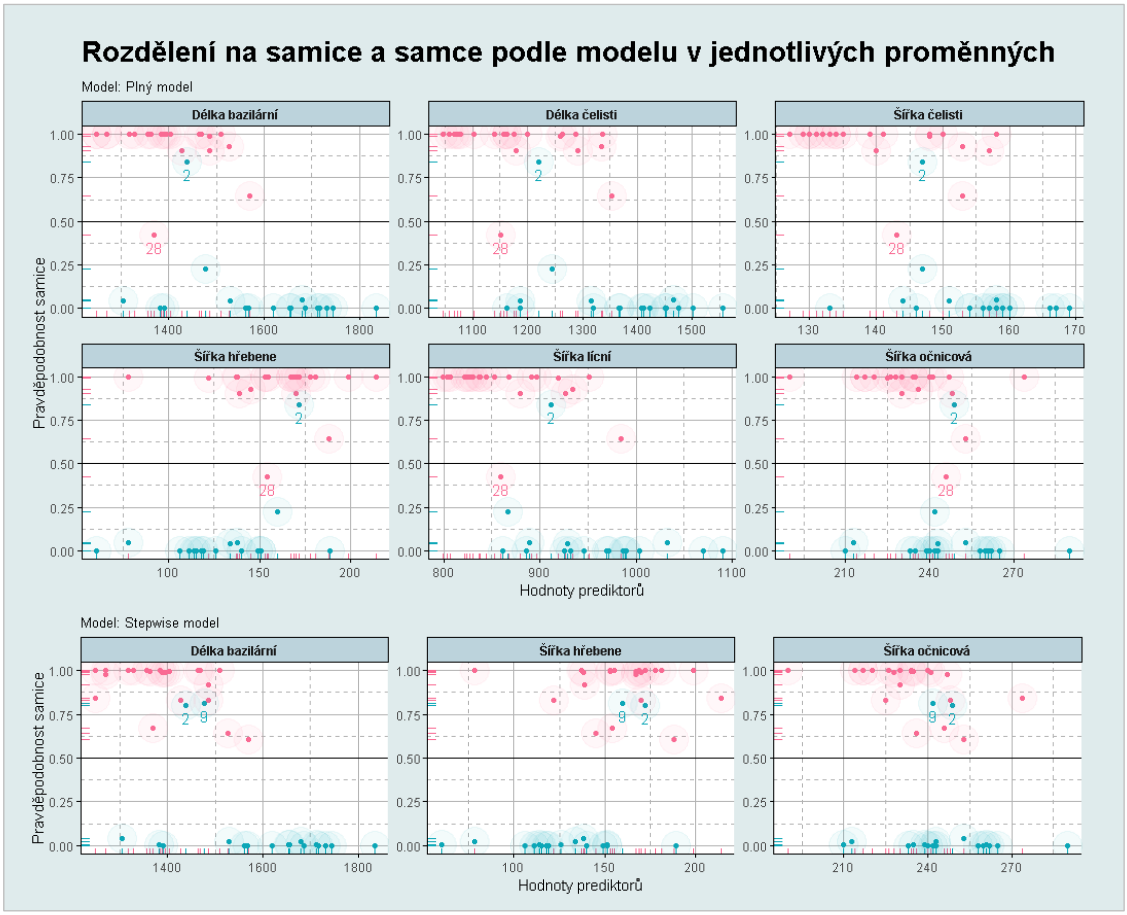
ROC křivky se příliš neliší, ale jednoduší stepwise model má přeci jen menší plochu pod křivkou.



V následujících proložených logistických funkcích lze pozorovat, že plný model prohodil dva jedince a stepwise model určil dva samce jako samice. Také obor linkovacího skóre na ose x se liší – u stepwise modelu je menší.



Dále byly vytvořeny grafy, které vynášejí pravděpodobnost, že jedinec je samice vůči hodnotám jednotlivých lebečních rozměrů u zkoumaných modelů. Označeny jsou špatně určení jedinci (tj. 2. a 28., 2. a 9. pozorování). Za povšimnutí stojí jak jsou vůči sobě postavené mraky množin bodů a srovnat s boxploty.



Dále byl ještě proveden ANOVA F-test těchto dvou modelů. Jeho p-hodnota vyšla 0,3608 a tedy nezamítá hypotézu, že by se modely lišily – z hlediska deviancí není stepwise model lepší jak plný model.

VÝBĚR MODELU

Jelikož mezi modely není drastický rozdíl v kritériích pro vhodnost, zvolil bych radši jednodušší stepwise model (tedy dle Akaikova kritéria). Je to praktičtější, vzhledem k tomu, že nemusím měřit všechno (a tím i chybovat) a myslím, že si to můžu dovolit i vzhledem k tomu, že v modelu neuvažuji interakce.

Pomocí kódu byly spočteny poměry šancí a konfidenční intervaly do následující tabulky:

Tabulky poměrů šancí s CI

		Délka bazilární	Šířka lícní	Šířka očnicová	Šířka hřebene	Délka čelisti	Šířka čelisti
Stepwise m.	Odhad OR	0,99		0,70	1,16		
	CI pro OR (97,5%)	9,96		1,69	4,70		
	CI pro OR (2,5%)	0,10		0,29	0,29		
Plný model	Odhad OR	0,86	1,06	0,39	1,42	1,15	0,40
	CI pro OR (97,5%)	2,46	5,20	0,27	2,96	3,02	0,06
	CI pro OR (2,5%)	0,30	0,22	0,56	0,68	0,44	2,56

Šance na to, že daná lebka bude samičí se zvyšuje 1,16-krát s dalším každým „milimetrem“ šířky hřebene a zmenšuje s koeficientem 0,70 každým „milimetrem“ očnicové šířky a s koeficientem 0,99 s každým „milimetrem“ bazilární délky. Samice mají v průměru širší hřeben a kratší očnicovou šířku a bazilární délku.

Pokud bych použil plný model bude pravděpodobnost na ženu stoupat s každým „milimetrem“:

- šířky hřebene 1,42-krát,
- délky čelisti 1,15-krát,
- lícní šířky 1,06 krát,

a klesat s každým „milimetrem“:

- očnicové šířky 0,39 krát,
- šířky čelisti 0,4-krát,
- a bazilární délky 0,86-krát.

Příloha – kódy

NAHRÁNÍ KNIHOVEN A KONSTANT

```
libs <- c("tidyverse", "GGally", "rsq", "ROCR", "clipr", "broom")

installed_libs <- libs %in% rownames(installed.packages())
if (any(installed_libs == F)){install.packages(libs[!installed_libs])}
invisible(lapply(libs, library, character.only = T)); rm(libs, installed_libs)

MOTIV <- theme(plot.background = element_rect(fill = "#dfebec"),
               strip.background = element_rect(colour = "black", fill = "#bcd3dc"),
               strip.text = element_text(face = "bold", size = 9),
               plot.title = element_text(size = 20, face = "bold", vjust = 2),
               axis.text.y = element_text(angle = 0),
               plot.subtitle = element_text(size = 10),
               axis.line = element_line(colour="black", linewidth = 0.6),
               axis.ticks = element_line(colour="black", linewidth = 0.6),
               panel.grid.major = element_line(colour="grey70", linewidth = 0.2),
               panel.grid.minor = element_line(colour="grey70", linewidth = 0.2,
                                               linetype = "dashed"),
               plot.margin=unit(c(5,5,5,5),"mm"),
               panel.background = element_rect(fill = "white"),
               legend.position = "right",)

VARIABLES <- c("Druh klokana" = "species",
            "Pohlaví" = "sex",
            "Délka bazilární" = "basilar.length",#
            "Délka patra" = "palate.length",
            "Šířka patra" = "palate.width",
            "Délka nosu" = "nasal.length",
            "Šířka nosu" = "nasal.width",
            "Šířka lícní" = "zygomatic.width",    #
            "Šířka očnicová" = "orbital.width",   #
            "Šířka hřebene" = "crest.width",      #
            "Délka otvoru" = "foramina.length",
            "Délka čelisti" ="mandible.length",   #
            "Šířka čelisti" ="mandible.width",    #
            "Hloubka čelisti" ="mandible.depth")

POHLAVI <- c("Samice" = "#F96A93",
             "Samec" = "#0CA7B7")

REV_VARIABLES <- setNames(names(VARIABLES), VARIABLES)
```

NAHRÁNÍ DAT, VYTVOŘENÍ DATOVÝCH DABULEK

```
df <- read.delim('./du-kanga.txt', sep=" ", header=T,
                 stringsAsFactors = T, na.strings = "NA")

setnames(df, old = VARIABLES, new=names(VARIABLES))

dfW <- df |> # wide format tabulka
  as.data.frame(row.names = 1:nrow(df)) |>
  drop_na() |>
  filter(`Druh klokana` == "fuliginosus") |>
  select(`Pohlaví`,
         `Délka bazilární`,
         `Šířka lícní`,
         `Šířka očnicová`,
         `Šířka hřebene`,
         `Délka čelisti`,
         `Šířka čelisti`) |>
  mutate(`Pohlaví` = `Pohlaví` |> recode_factor("Male" = "Samec",
                                                "Female" = "Samice"))
dfL <- dfW |> #long format tabulka
```

```
  as.data.frame(row.names = 1:nrow(dfW)) |>
  rownames_to_column("ID") |>
  pivot_longer(names_to = "Promenna", values_to = "Hodnoty", cols = `Délka
bazilární`:`Šířka čelisti`) |>
  mutate(Promenna = as.factor(Promenna),
         ID = as.factor(ID))
```

POPISNÁ STATISTIKA SKUPIN

```
Popis <- dfL |>
  group_by(`Pohlaví`, Promenna) |>
  dplyr::summarise(`Směrodatná odchylka` = sd(Hodnoty),
                   `Průměr` = mean(Hodnoty),
                   Pocet = length(Hodnoty))

Popisy <- Popis |> split(f = Popis$`Pohlaví`)
Popisy$Samec
```

```
## # A tibble: 6 x 5
## # Groups:   Pohlaví [1]
##   Pohlaví Promenna         `Směrodatná odchylka` Průměr Pocet
##   <fct>   <fct>                            <dbl>  <dbl> <int>
## 1 Samec   Délka bazilární                  147.    1594.    18
## 2 Samec   Délka čelisti                   120.    1362.    18
## 3 Samec   Šířka čelisti                    8.98    155.    18
## 4 Samec   Šířka hřebene                   31.5     130.    18
## 5 Samec   Šířka lícní                     65.6     958.    18
## 6 Samec   Šířka očnicová                  18.8     247.    18
```

```
Popisy$Samice
```

```
## # A tibble: 6 x 5
## # Groups:   Pohlaví [1]
##   Pohlaví Promenna         `Směrodatná odchylka` Průměr Pocet
##   <fct>   <fct>                            <dbl>  <dbl> <int>
## 1 Samice  Délka bazilární                   93.8   1395.    21
## 2 Samice  Délka čelisti                    98.1   1185.    21
## 3 Samice  Šířka čelisti                     9.92   140.    21
## 4 Samice  Šířka hřebene                    28.7    159.    21
## 5 Samice  Šířka lícní                      52.3    866.    21
## 6 Samice  Šířka očnicová                   17.1    233.    21
```

VZTAHY MEZI POHLAVÍMI – MATICOVÝ GRAF

```
# Matice XY grafů ----
matgraf <- ggpairs(data=dfW, switch="both",
        mapping = aes(color=`Pohlaví`),
        lower = list(continuous = wrap("points", alpha = 0.6, size=2)),
        diag = list(discrete="barDiag",
                    continuous = wrap("densityDiag", alpha=0.5 )),
        upper = list(combo = wrap("box_no_facet", alpha=0.5),
                     continuous = wrap("cor", size=4, alignPercent=0.6)),
        legend = 1)+
  theme_stata(base_size = 8) +
  scale_colour_manual(values = POHLAVI)  +
  scale_fill_manual(values = POHLAVI)  +
  labs(x = "", y = "", title = "Maticový graf",
       subtitle = "Pohlaví klokanů a rozměry lebek")

plots = list()
for (i in 1:7){
  plots <- c(plots, lapply(2:matgraf$ncol, function(j) getPlot(matgraf, i=i, j = j)))
}

ggmatrix(plots,
         nrow = 7,
         ncol=matgraf$ncol-1,
         xAxisLabels = matgraf$xAxisLabels[2:matgraf$ncol],
         yAxisLabels = matgraf$yAxisLabels) +
  labs(x = "", y = "", title = "Maticový graf",
```

```
        subtitle = "Pohlaví klokanů a rozměry lebek")+
  theme(plot.background = element_rect(fill = "#dfebec"),
        strip.background = element_rect(colour = "black", fill = "#bcd3dc"),
        strip.text = element_text(face = "bold", size = 9),
        plot.title = element_text(size = 20, face = "bold", vjust = 2),
        legend.position = "bottom",
        axis.text.y = element_text(angle = 0),
        plot.subtitle = element_text(size = 10))
```

VZTAHY MEZI POHLAVÍMI – QQ PLOTY

```
dfL |>
  ggplot(mapping = aes(sample = Hodnoty, color = Pohlaví)) +
  stat_qq()+
  stat_qq_line(show.legend = F)+
  facet_wrap(~Promenna, scales = "free")+
  scale_color_manual(values = POHLAVI) +
  guides(fill = "none") +
  labs(y = "Rozměr [mm]", x = "Standardizované kvantily z N(0, 1)",
       title = "QQ diagramy") +
  theme_stata() +
  MOTIV
```

VZTAHY MEZI POHLAVÍMI- BOXPLOTY

```
fun_mean <- function(x){
  #funkce ke značení průměru v boxplotech
  return(data.frame(y=mean(x),label=round(mean(x,na.rm=T),1)))
}

dfL |>
  ggplot(aes(x = Pohlaví, y = Hodnoty)) +
  geom_boxplot(aes(fill = Promenna, color = Pohlaví), linewidth = 1) +
  stat_boxplot(mapping = aes(color = Pohlaví), geom = 'errorbar',
               width = 0.2, linewidth = 1) +
  stat_summary(fun.y=mean, geom="point", shape=18, size=3, color="black") +
  stat_summary(fun.data = fun_mean, geom="text", hjust=-0.5) +
  facet_wrap(~Promenna, scales = "free", nrow = 3, ncol = 3) +
  scale_fill_brewer(palette="Dark2") +
  scale_color_manual(values = POHLAVI) +
  guides(fill = "none") +
  labs(y = "Hodnota [mm]", x = "Rozměr lebeční části",
       title = "Krabicové diagramy lebečních rozměrů") +
  theme_stata() +
  MOTIV +
  ggpubr::stat_compare_means(method = "t.test", label.x = 1.5)
```

T-TESTY A JEJICH PŘEDPOKLADY

```
Predpoklady <- data.frame(matrix(nrow = 0, ncol = 6))
colnames(Predpoklady) <- c("Promenna",
                           "Pohlaví",
                           "SW",
                           "F-test",
                           "Ttest",
                           "N")

pr <- colnames(dfW)[2:7]
for (i in 1:6){
  promen <- pr[i]
  message(promen)
  for (sex in c("Samec", "Samice")){
    message(sex)
    message("SW test")
    val = dfL |> subset(Promenna == promen & Pohlaví == sex)
    radky <- nrow(val)
```

```
    sw <- shapiro.test(val$Hodnoty); print(sw)

    message("F-test a T-test")
    val = dfL |> subset(Promenna == promen)
    vt <- var.test(val$Hodnoty~val$Pohlaví); print(vt)
    if (vt$p.value < 0.05){b <- F}else{b <- T}

    message("...")
    val = dfL |> subset(Promenna == promen)
    tt <- t.test(val$Hodnoty~val$Pohlaví, var.equal = b); print(tt)

    Predpoklady <- Predpoklady |> rbind(data.frame(Promenna = promen,
                                                   Pohlaví = sex,
                                                   SW = sw$p.value,
                                                   `F-test` = vt$p.value,
                                                   Ttest = tt$p.value,
                                                   N = radky))
  }
  message("####################")
}; rm(sex, i, val, vt, tt, sw, radky, promen, pr, b)
```

PLNÝ MODEL

```
full = formula(`Pohlaví` ~ `Délka bazilární` +
                `Šířka lícní` +
                `Šířka očnicová`+
                `Šířka hřebene` +
                `Délka čelisti` +
                `Šířka čelisti`)
LRfull <- glm(formula = full, family = binomial(logit), data = dfW)
summary(LRfull)
```

```
## Call:
## glm(formula = full, family = binomial(logit), data = dfW)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.91226  -0.00001   0.00000   0.00135   1.31121
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       308.27026  254.48121   1.211    0.226
## `Délka bazilární`  -0.15345    0.14955  -1.026    0.305
## `Šířka lícní`       0.06167    0.05773   1.068    0.285
## `Šířka očnicová`   -0.94356    0.65998  -1.430    0.153
## `Šířka hřebene`     0.34786    0.23586   1.475    0.140
## `Délka čelisti`     0.13797    0.17132   0.805    0.421
## `Šířka čelisti`    -0.92121    0.99361  -0.927    0.354
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 53.8345  on 38  degrees of freedom
## Residual deviance:  7.6172  on 32  degrees of freedom
## AIC: 21.617
##
## Number of Fisher Scoring iterations: 11
```

NULOVÝ MODEL

```
dataLR0 <- glm(formula = `Pohlaví` ~ 1, family = binomial(logit), data = dfW)
summary(dataLR0)
```

```
## Call:
## glm(formula = Pohlaví ~ 1, family = binomial(logit), data = dfW)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.244  -1.244   1.113   1.113   1.113
```

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.1542     0.3212    0.48    0.631
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 53.834  on 38  degrees of freedom
## Residual deviance: 53.834  on 38  degrees of freedom
## AIC: 55.834
##
## Number of Fisher Scoring iterations: 3
```

SROVNÁNÍ PLNÝ MODEL – NULOVÝ MODEL

```
anova(dataLR0, LRfull, test = "Chisq")
## Analysis of Deviance Table
##
## Model 1: Pohlaví ~ 1
## Model 2: Pohlaví ~ `Délka bazilární` + `Šířka lícní` + `Šířka očnicová` +
##     `Šířka hřebene` + `Délka čelisti` + `Šířka čelisti`
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        38     53.834
## 2        32      7.617  6   46.217 2.68e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

STEPWISE METODY MODELOVÁNÍ

```
stepBack <- step(glm(formula = full, family=binomial(logit), data=dfW),
                 direction='backward')
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Start:  AIC=21.62
## Pohlaví ~ `Délka bazilární` + `Šířka lícní` + `Šířka očnicová` +
##     `Šířka hřebene` + `Délka čelisti` + `Šířka čelisti`
##                       Df Deviance    AIC
## - `Délka čelisti`      1    8.760 20.760
## - `Šířka lícní`        1    9.086 21.086
## - `Šířka čelisti`      1    9.254 21.254
## <none>                      7.617 21.617
## - `Délka bazilární`    1   10.265 22.265
## - `Šířka hřebene`      1   23.237 35.237
## - `Šířka očnicová`     1   31.720 43.720
##
## Step:  AIC=20.76
## Pohlaví ~ `Délka bazilární` + `Šířka lícní` + `Šířka očnicová` +
##     `Šířka hřebene` + `Šířka čelisti`
##                       Df Deviance    AIC
## - `Šířka čelisti`      1    9.272 19.272
## <none>                      8.760 20.760
## - `Šířka lícní`        1   10.772 20.772
## - `Délka bazilární`    1   11.239 21.239
## - `Šířka hřebene`      1   23.801 33.801
## - `Šířka očnicová`     1   32.736 42.736
##
## Step:  AIC=19.27
## Pohlaví ~ `Délka bazilární` + `Šířka lícní` + `Šířka očnicová` + `Šířka hřebene`
##                       Df Deviance    AIC
## - `Šířka lícní`        1   10.824 18.824
## <none>                      9.272 19.272
## - `Délka bazilární`    1   12.425 20.425
## - `Šířka hřebene`      1   24.536 32.536
## - `Šířka očnicová`     1   32.783 40.783
##
## Step:  AIC=18.82
## Pohlaví ~ `Délka bazilární` + `Šířka očnicová` + `Šířka hřebene`
##                       Df Deviance    AIC
```

```
## <none>                    10.824 18.824
## - `Délka bazilární`  1    13.354 19.354
## - `Šířka hřebene`    1    24.571 30.571
## - `Šířka očnicová`   1    33.446 39.446
```

```
stepForw <- step(glm(formula = `Pohlaví` ~ 1, family=binomial(logit), data=dfW),
                 scope =  ~ `Délka bazilární` + `Šířka lícní` + `Šířka očnicová`+
                        `Šířka hřebene` + `Délka čelisti` + `Šířka čelisti`,
                 direction='forward')
```

```
## Start:  AIC=55.83
## Pohlaví ~ 1
##                      Df Deviance    AIC
## + `Délka bazilární`  1   34.458 38.458
## + `Délka čelisti`    1   34.920 38.920
## + `Šířka lícní`      1   35.371 39.371
## + `Šířka čelisti`    1   36.604 40.604
## + `Šířka hřebene`    1   44.780 48.780
## + `Šířka očnicová`   1   47.974 51.974
## <none>                    53.834 55.834
##
## Step:  AIC=38.46
## Pohlaví ~ `Délka bazilární`
##                      Df Deviance    AIC
## + `Šířka očnicová`   1   24.571 30.571
## <none>                    34.458 38.458
## + `Šířka hřebene`    1   33.446 39.446
## + `Šířka čelisti`    1   33.871 39.871
## + `Šířka lícní`      1   34.181 40.181
## + `Délka čelisti`    1   34.448 40.448
##
## Step:  AIC=30.57
## Pohlaví ~ `Délka bazilární` + `Šířka očnicová`
##                      Df Deviance    AIC
## + `Šířka hřebene`    1   10.824 18.824
## <none>                    24.571 30.571
## + `Délka čelisti`    1   23.773 31.774
## + `Šířka čelisti`    1   23.824 31.824
## + `Šířka lícní`      1   24.536 32.536
##
## Step:  AIC=18.82
## Pohlaví ~ `Délka bazilární` + `Šířka očnicová` + `Šířka hřebene`
##                      Df Deviance    AIC
## <none>                    10.8244 18.824
## + `Šířka lícní`      1   9.2724 19.272
## + `Délka čelisti`    1  10.3159 20.316
## + `Šířka čelisti`    1  10.7717 20.772
```

```
stepBoth <- step(glm(formula = full, family=binomial(logit), data=dfW),
                 direction='both')
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
## Start:  AIC=21.62
## Pohlaví ~ `Délka bazilární` + `Šířka lícní` + `Šířka očnicová` +
##     `Šířka hřebene` + `Délka čelisti` + `Šířka čelisti`
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
##                      Df Deviance    AIC
## - `Délka čelisti`    1    8.760 20.760
## - `Šířka lícní`      1    9.086 21.086
## - `Šířka čelisti`    1    9.254 21.254
## <none>                    7.617 21.617
## - `Délka bazilární`  1   10.265 22.265
## - `Šířka hřebene`    1   23.237 35.237
## - `Šířka očnicová`   1   31.720 43.720
##
## Step:  AIC=20.76
## Pohlaví ~ `Délka bazilární` + `Šířka lícní` + `Šířka očnicová` +
##     `Šířka hřebene` + `Šířka čelisti`
##                      Df Deviance    AIC
## - `Šířka čelisti`    1    9.272 19.272
## <none>                    8.760 20.760
## - `Šířka lícní`      1   10.772 20.772
```

```
## - `Délka bazilární`  1   11.239 21.239
## + `Délka čelisti`    1    7.617 21.617
## - `Šířka hřebene`    1   23.801 33.801
## - `Šířka očnicová`   1   32.736 42.736
##
## Step:  AIC=19.27
## Pohlaví ~ `Délka bazilární` + `Šířka lícní` + `Šířka očnicová` + `Šířka hřebene`
##                       Df Deviance    AIC
## - `Šířka lícní`        1   10.824 18.824
## <none>                     9.272 19.272
## - `Délka bazilární`    1   12.425 20.425
## + `Šířka čelisti`      1    8.760 20.760
## + `Délka čelisti`      1    9.254 21.254
## - `Šířka hřebene`      1   24.536 32.536
## - `Šířka očnicová`     1   32.783 40.783
##
## Step:  AIC=18.82
## Pohlaví ~ `Délka bazilární` + `Šířka očnicová` + `Šířka hřebene`
##                       Df Deviance    AIC
## <none>                     10.824 18.824
## + `Šířka lícní`        1    9.272 19.272
## - `Délka bazilární`    1   13.354 19.354
## + `Délka čelisti`      1   10.316 20.316
## + `Šířka čelisti`      1   10.772 20.772
## - `Šířka hřebene`      1   24.571 30.571
## - `Šířka očnicová`     1   33.446 39.446
stepBoth$coefficients == stepBack$coefficients &
  stepBack$coefficients == stepForw$coefficients &
  stepBoth$coefficients == stepForw$coefficients
##       (Intercept) `Délka bazilární`   `Šířka očnicová`    `Šířka hřebene`
##              TRUE              TRUE               TRUE               TRUE
summary(stepBack)
## Call:
## glm(formula = Pohlaví ~ `Délka bazilární` + `Šířka očnicová` +
##     `Šířka hřebene`, family = binomial(logit), data = dfW)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.82697  -0.05102   0.00189   0.12860   1.00037
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       85.19341   46.63912   1.827   0.0678 .
## `Délka bazilární` -0.01490    0.01061  -1.405   0.1600
## `Šířka očnicová`  -0.35594    0.19451  -1.830   0.0673 .
## `Šířka hřebene`    0.15261    0.08351   1.827   0.0676 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Dispersion parameter for binomial family taken to be 1)
##     Null deviance: 53.834  on 38  degrees of freedom
## Residual deviance: 10.824  on 35  degrees of freedom
## AIC: 18.824
## Number of Fisher Scoring iterations: 9
stepBack |> tidy() |> write_clip()
```

SROVNÁVÁNÍ MODELŮ

```
Srovnani <- list()
score_data <- list()
rocGrafy <- list()
lm_data <- list()

i = 1
modely = c(Stewise = list(stepBack), LRfull = list(LRfull))
for (each in modely){
  jmeno = c("Stepwise model", "Plný model")[i]

  fitted <- data.frame(Predpoved = predict(each, newdata=dfW, type="response")) |>
```

```
    mutate(Binarne = if_else(Predpoved > 0.5, "SamiceM", "SamecM") |> as.factor())

  KontinTab <- table(fitted$Binarne, dfW$`Pohlaví`)

  auc <- performance(prediction(fitted$Predpoved,
                     as.numeric(dfW$`Pohlaví`)), "auc")@y.values |> as.numeric()
  ## Tabulka pro srovnání ----
  Srovnani[[i]] <- data.frame(
              Model = jmeno,

              Nagelkerke = rsq(each, type='n'),
              McFadden = rsq(each, type='kl'),
              CoxSnell = rsq(each, type='lr'),
              AICval = AIC(each),
              Deviance = each$deviance,

              OKSamec =  c(KontinTab["SamecM", "Samec"]),
              OKSamice = c(KontinTab["SamiceM", "Samice"]),
              RealSamecJakoSamiceM = c(KontinTab["SamiceM","Samec"]),
              RealSamiceJakoSamecM = c(KontinTab["SamecM", "Samice"]),
              SpravneUrcene = c(sum(diag(KontinTab)) / sum(KontinTab)),
              AUC = auc)

  #Data pro prokládané logistické křivky
  #https://blog.revolutionanalytics.com/2016/08/roc-curves-in-two-lines-of-code.html
  score_data[[i]] <- data.frame(link = predict(each, dfW, type="link"),
                          response = fitted$Predpoved, Pohlaví = dfW$Pohlaví,
                          m = c(jmeno), stringsAsFactors=FALSE)

  #Data pro parciální grafy dle jednotlivych promenných
  prediktory <- names(coefficients(each))[-1] |> str_remove_all("`")

  lm_data[[i]] <- data.frame(response = fitted$Predpoved,
                             m = c(jmeno)) |>
    rownames_to_column("ID") |>
    mutate(Binarne = if_else(response > 0.5, "Samice", "Samec") |> as.factor()) |>
    bind_cols(dfW[c("Pohlaví", prediktory)]) |>
    pivot_longer(names_to = "Promenna", values_to = "Hodnoty", cols = -c(1:5)) |>
    mutate(Prohozene = case_when(Pohlaví != Binarne ~ ID, .default = c(""))) |>
    filter(Promenna %in% prediktory)

  #data pro ROC krivky
  roc <- performance(prediction(fitted$Predpoved,
                                as.numeric(dfW$`Pohlaví`)), "tpr", "fpr")
  rocGrafy[[i]] <- data.frame(OsaX = c(roc@x.values[[1]]),
                          OsaY = c(roc@y.values[[1]]),
                          Model = c(jmeno))
  i <- i+1
}
score_data <- do.call(rbind, score_data); rocGrafy <-   do.call(rbind, rocGrafy);
Srovnani <- do.call(rbind, Srovnani)

Srovnani |> write_clip()
Srovnani
```

ROC GRAFY

```
rocGrafy |>
  ggplot(aes(x = OsaX, y = OsaY, color = Model)) +
  geom_line(linewidth = 0.8) +
  facet_wrap(~Model) +
  MOTIV +
  coord_equal() +
  labs(title = "ROC křivky",
       y = "Senzitivita",
       x = "1 - specificita")
```

LOGISTICKÉ PROLOŽENÍ

```
score_data |>
  ggplot(aes(x=link, y=response, col=Pohlaví)) +
  geom_hline(yintercept = c(0, 0.5, 1), color = "grey50")  +
  geom_vline(xintercept = 0.0, color = "grey50")  +
  geom_function(fun = ~1/(1+exp(-.x)), inherit.aes = F, col = "grey20") +
  scale_color_manual(values=POHLAVI) +
  geom_point(size = 3) +
  geom_rug() +
  MOTIV  +
  theme(panel.background = element_rect(fill = "white")) +
  facet_wrap(~m, scale = "free") +
  labs(title = "Logistické funkce",
       # subtitle = paste0("Model: ", jmeno),
       y = "Pravděpodobnost na samici",
       x = "Linkovací skóre")
```

PARCIÁLNÍ GRAFY MODELU

```
p <- list()
for (ind in c(1, 2)){
  jmeno <- lm_data[[ind]]$m[1]
  p[[ind]] <- lm_data[[ind]] |>
    ggplot(aes(x = Hodnoty, y = response, color = Pohlaví, label = Prohozene)) +
    geom_point() + geom_rug(show.legend = F) +
    geom_point(size = 10, alpha = 0.05) +
    geom_text(vjust = 1.5, show.legend = F) +
    geom_hline(yintercept = 0.5, color = "grey0", show.legend = F)  +
    scale_color_manual(values = POHLAVI) +
    labs(x = "Hodnoty prediktorů", y = "Pravděpodobnost samice",
         title = "Rozdělení na samice a samce podle modelu v jednotlivých proměnných",
         subtitle = paste0("Model: ", jmeno)) +
    MOTIV +
    theme(legend.position = "left") +
    facet_wrap(~Promenna, scale = "free", ncol = 3)
  print(p[[ind]])
  Sys.sleep(2)
}
```

ANOVA SROVNÁNÍ MODELŮ

```
anova(stepBack, LRfull, test = "Chisq")
## Analysis of Deviance Table
## Model 1: Pohlaví ~ `Délka bazilární` + `Šířka očnicová` + `Šířka hřebene`
## Model 2: Pohlaví ~ `Délka bazilární` + `Šířka lícní` + `Šířka očnicová` +
##     `Šířka hřebene` + `Délka čelisti` + `Šířka čelisti`
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        35    10.8244
## 2        32     7.6172  3   3.2072   0.3608
```

VÝBĚR MODELU

```
stws <- list()
stws[[1]] <- coef(stepBack) |> exp()
stws[[2]] <- exp(coef(stepBack) - qnorm(0.975 * summary(stepBack)$coefficients[,2]))
stws[[3]] <- exp(coef(stepBack) + qnorm(0.975 * summary(stepBack)$coefficients[,2]))
stws <- do.call(rbind, stws)
stws |> write_clip()

fl <- list()
fl[[1]] <- coef(LRfull) |> exp()
fl[[2]] <- exp(coef(LRfull) - qnorm(0.975 * summary(LRfull)$coefficients[,2]))
fl[[3]] <- exp(coef(LRfull) + qnorm(0.975 * summary(LRfull)$coefficients[,2]))
fl <- do.call(rbind, fl)
fl |> write_clip()
```