Vilniaus Universitetas

Išgyvenamumo analizė (trukmės modeliai)

Laboratorinis darbas

Darbą atliko:

Vainius Gataveckas, Matas Gaulia, Dovydas Martinkus

Duomenų Mokslas

3 kursas 2 gr.

Vilnius, 2022

Turinys

Naudoti metodai

Šiame darbe naudoti Kaplan-Meier išgyvenamumo funkcijos įverčiai, Cox semiparametrinis ir AFT parametrinis regresijos modeliai. Darbas atliktas naudojant R.

Naudoti R paketai:

tidyverse

survival

survminer

eha

Duomenys ir jų šaltiniai

Telekomunikacijų bendrovės klientų paslaugų atsisakymo (churn) trukmės duomenys pagal bendrovės klientų demografines ir naudojamų paslaugų kovariantes.

Duomenų šaltinis – modeldata R paketas.

Originalus šaltinis IBM Watson Analytics churn data. Prieiga per internetą:

https://ibm.co/2sOvyvy.

Duomenis sudaro šie kintamieji:

“churn” – ar klientas atsisakė bendrovės paslaugų.

„tenure“ – mėnesių skaičius, kurį laiką klientas naudojosi įmonės paslaugomis (cenzūruota iš dešinės).

„female“ – ar klientas yra moteris.

„senior_citien“ – ar klientas yra pensijinio amžiaus.

„partner“ – ar klientas turi partnerį (partnerę).

„dependents“ – ar klientas turi išlaikytinių.

„phone_service“ – ar klientas naudojami bendrovės telefono ryšiu.

„internet_service“ – ar klientas iš bendrovės užsisakęs internetą.

„monthly_charges“ – per mėnesį sumokoma suma.

Kitos kovariantės, esančios originaliame duomenų rinkinyje nėra stipriai susijusios su tyrimo tikslu, todėl iš anksto pasirinkta jų tyrime nenaudoti.

Tikslas ir uždaviniai

Tikslas: Sudaryti trukmės modelius telekomunikacijų bendrovės paslaugų atsisakymo trukmei ir įvertinti demografinių požymių įtaką paslaugų atsisakymui.

Uždaviniai:

Demografinių požymių įtakos įvertinimas naudojant tiriamąja duomenų analizę ir Kaplan-Meier išgyvenamumo funkcijas.

Trukmės regresijos modelių, atsižvelgiančių į kitų kovariančių reikšmes, sudarymas:

- Cox semiparametrinio modelio sudarymas.
- AFT parametrinio modelio sudarymas.

Modelių tinkamumo analizė.

Atliktos analizės aprašymas

Naudojant R

Duomenyse turimas požymis „churn“ parodo ar klientas atsisakė bendrovės paslaugų ar ne. Šį stulpelį galima interpretuoti kaip dešinįjį cenzūravimą laikant tų klientų, kurie paslaugų tyrimo metu dar nebuvo atsisakė, duomenis cenzūruotais. Didžioji dalis duomenų aibėje esančių duomenų yra cenzūruoti. Kitokio tipo negu dešiniojo cenzūravimo duomenyse nėra. Nubrėžti paslaugų naudojimosi trukmės sklaidos grafikai pagal kiekvienos naudotos kovariantės reikšmes. Pastebėta, kad didžioji dalis turimų necenzūruotų duomenų yra iš trumpai įmonės paslaugomis naudojusių klientų, taip pat iš klientų, kurie už paslaugaus mokėjo daugiau. Apskaičiuoti naivūs (daug mažesni negu yra iš tikrųjų) paslaugų naudojimo trukmės vidurkiai, kurie neatsižvelgia į duomenų cenzūravimą.

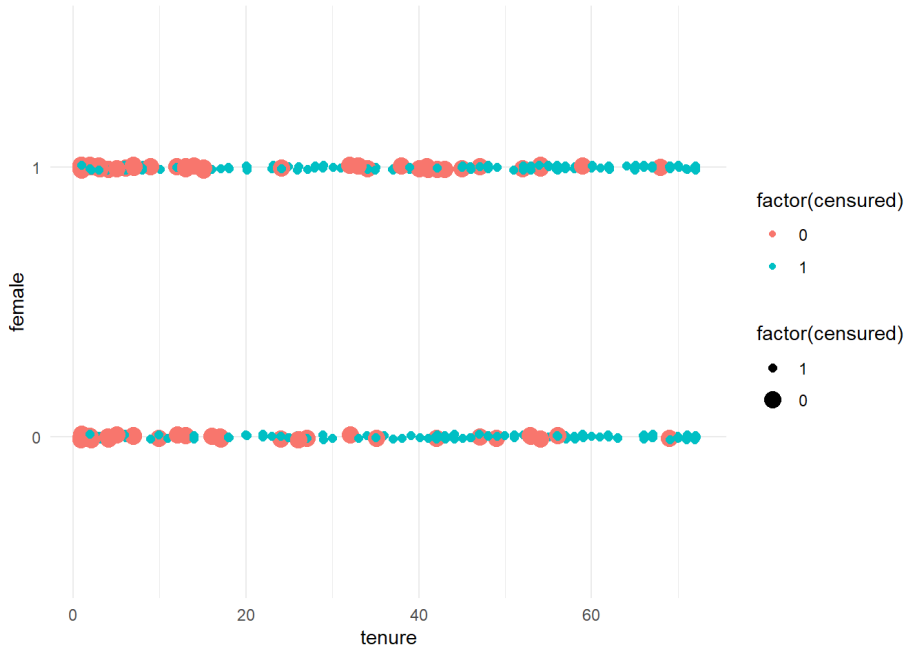
Apskaičiuotas ir grafiškai pavaiduotas Kaplan-Meier išgyvenamumo funkcijos įvertinys visai duomenų aibei. Vidutinis gauta paslaugų naudojimosi trukmė lygi 41.48. Kaip ir tikėtasi, ši reikšmė stipriai didesnė už prieš tai gautus vidurkius.

```
library(tidyverse)
library(modeldata)


x <- read_csv("wa_churn.csv")

library(survival)
library(survminer)


x <- x %>%
 select(
  churn, tenure, female, monthly_charges, phone_service, internet_service, senior_citizen,
  dependents, partner
 ) %>%
 mutate(internet_service = factor(if_else(internet_service == "No", 0, 1))) %>%
 mutate(across(-c(monthly_charges, tenure), ~ as.factor(.)),
  censured = if_else(churn == "No", 1, 0)
 ) %>%
 select(-churn)

table(x$censured)

##
##  0  1
##  86 265

prop.table(table(x$censured))

##
##       0       1
## 0.2450142 0.7549858

# vidutinis laikas neįskaitant cenzūravimo
mean(x$tenure[x$censured == 0])

## [1] 19.51163
```

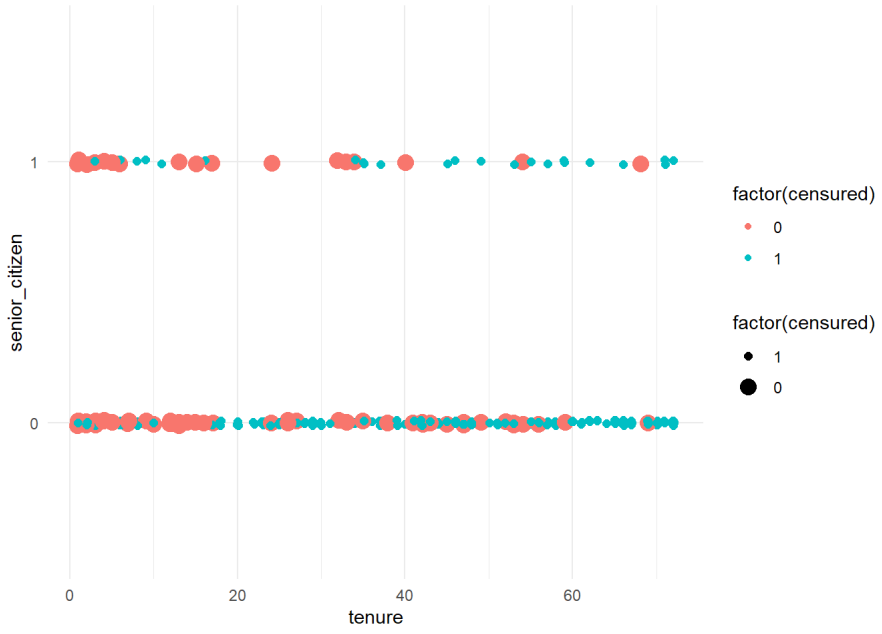
```r
# bendras vidutinis laikas (mažesnis negu yra iš tikrųjų dėl cenzūravimo)
mean(x$tenure)
```

```
## [1] 33.63533
```

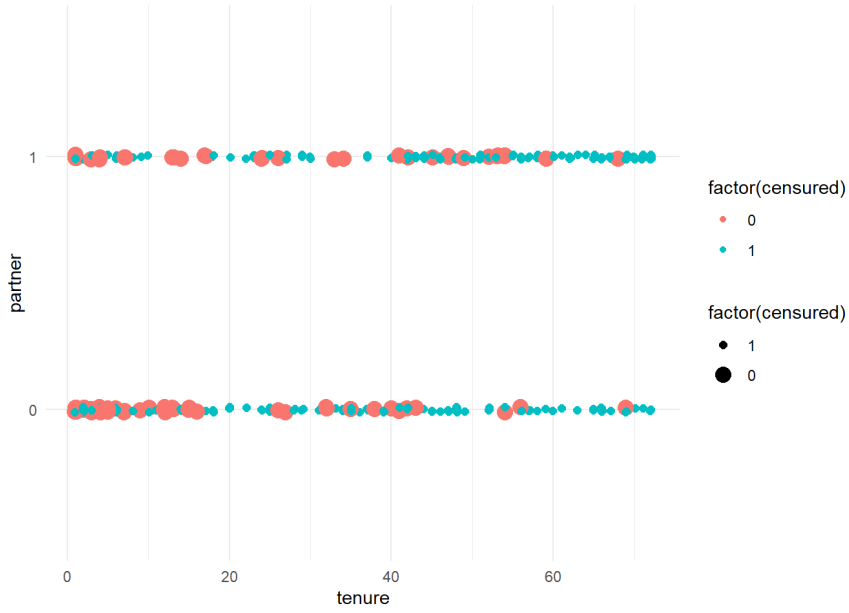
```r
plot <- function(column) {
 ggplot(x, aes({{ column }}, tenure, color = factor(censured))) +
   geom_point(aes(size = factor(censured)), position = position_jitter(width = 0.01, height = 0.1)) +
   scale_x_discrete() +
   coord_flip() +
   theme_minimal() +
   scale_size_manual(values = c("1" = 2, "0" = 4))
}


plot(female)
```



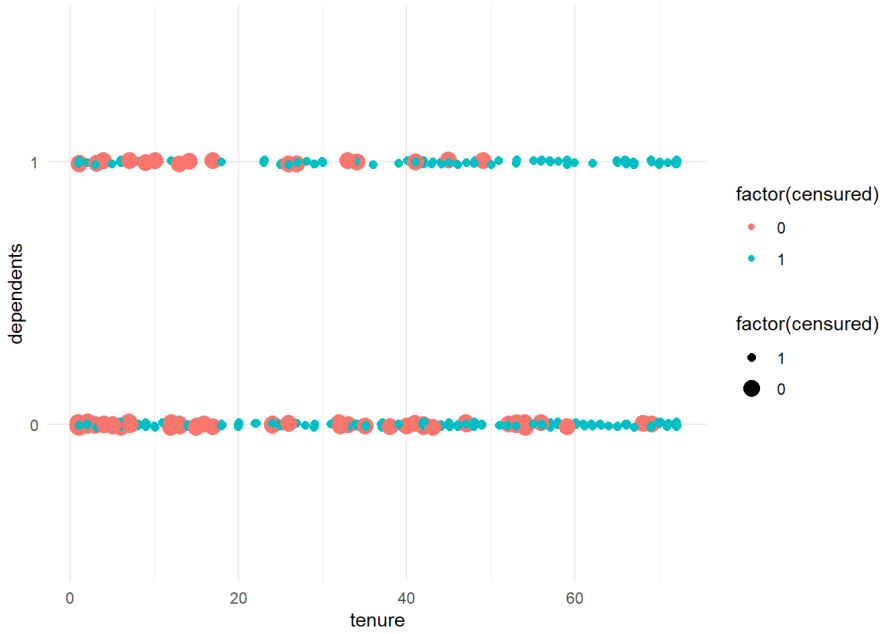

```r
plot(senior_citizen)
```



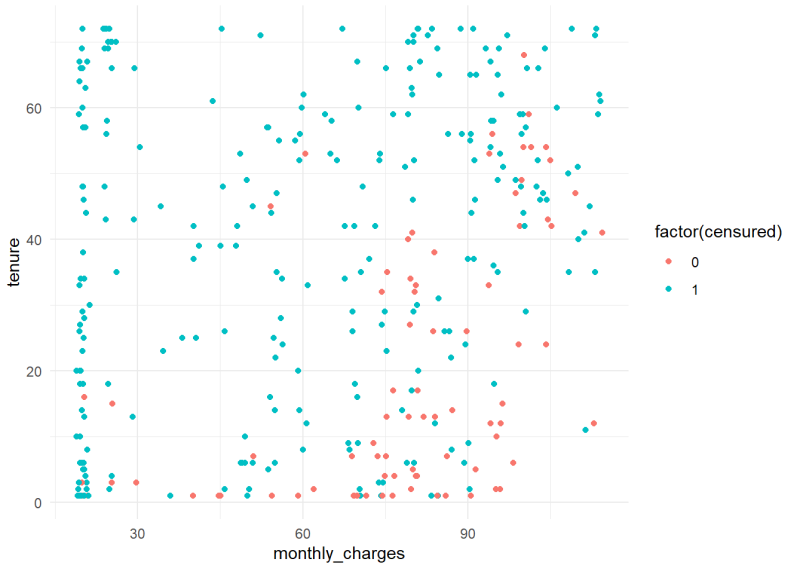

```
plot(partner)
```



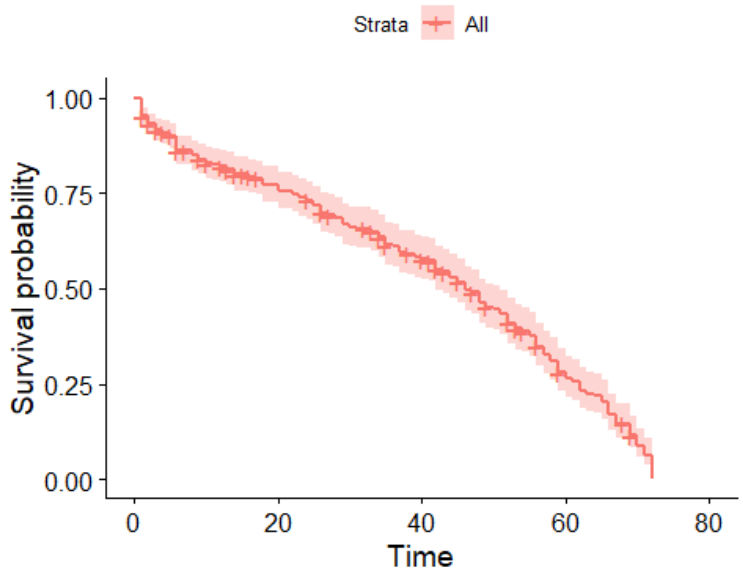

```
plot(dependents)
```




```
ggplot(x, aes(monthly_charges, tenure, color = factor(censured))) +
 geom_point() +
 theme_minimal()
```



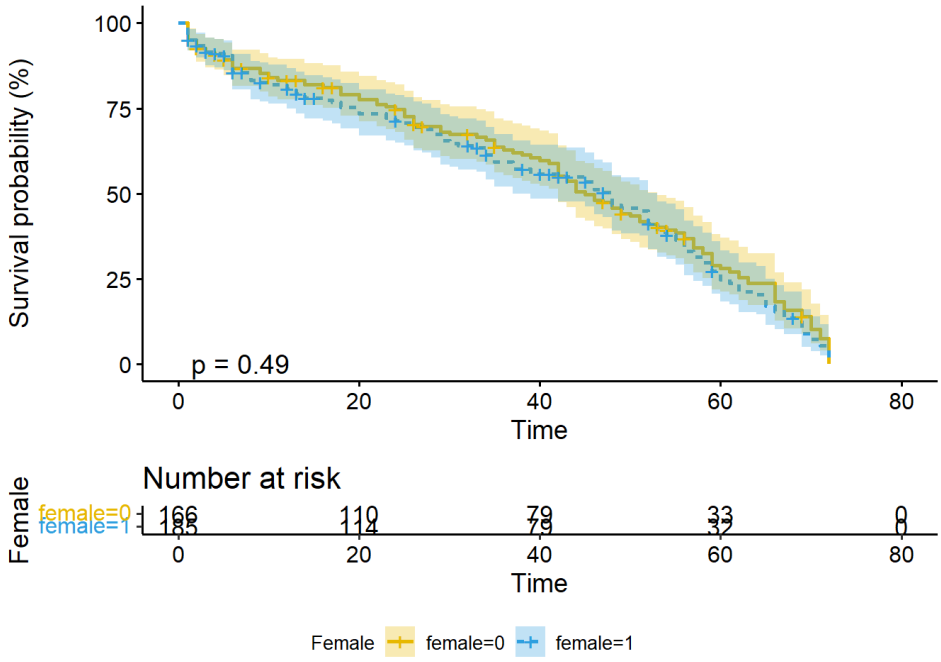

```
km <- survfit(Surv(tenure, censured) ~ 1, data = x)

print(km, print.rmean = TRUE) # 46 mediana, 41 vidurkis (palyginti su prieš tai gautu vidurkiu)

## Call: survfit(formula = Surv(tenure, censured) ~ 1, data = x)
##
##       n    events    *rmean *se(rmean)    median   0.95LCL   0.95UCL
##  351.00    265.00     41.48      1.34     46.00     42.00     51.00
##    * restricted mean with upper limit =  72

ggsurvplot(km)
```



Apskaičiuotas Kaplan-Meier išgyvenamumo funkcijos įvertinys dalijant duomenų aibę atskirai pagal demografinių kovariančių lygmenis. Gauta, kad vyrai paslaugomis vidutiniškai naudojosi ilgiau už moteris, pensijinio amžiaus žmonės - ilgiau už jaunesnius, partnerius ar išlaikytinius turintys žmonės - ilgiau už jų neturinčius. Statistiškai reikšmingas skirtumas naudojant log-rank testą gautas tik tarp turinčių partnerį ir neturinčių ($p<0.001$). Nors šis naudotas išgyvenamumo funkcijos vertinimas leidžia įvertinti demografinių faktorių įtaką, jis neatsižvelgia į kitų kovariančių reikšmių skirtumus, todėl reikia sudaryti pilną regresijos modelį norint tinkamai įvertinti demografinių faktorių įtaką paslaugų naudojimosi trukmei.

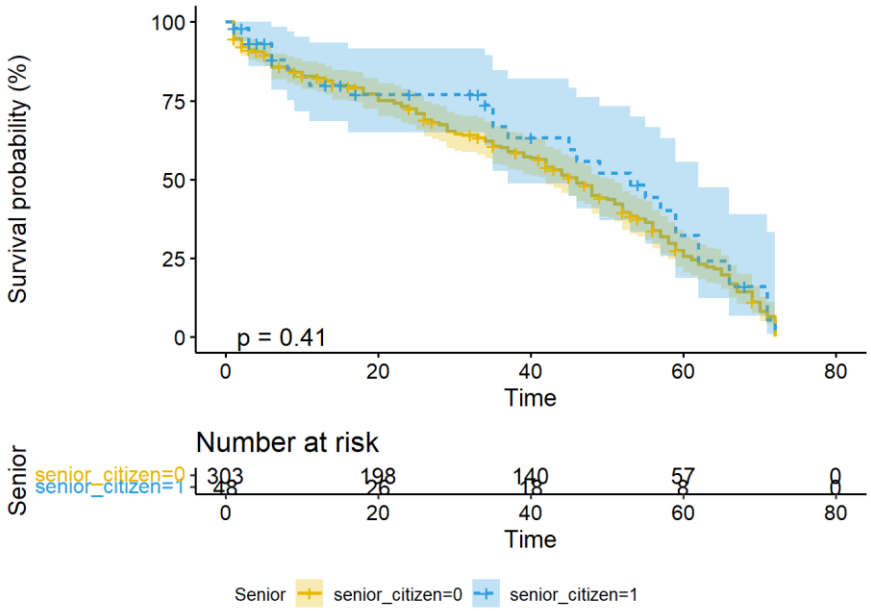
```
individual <- function(variable, title) {
 model <-
  eval(substitute(survfit(Surv(tenure, censured) ~ variable, data = x)))

 print(model, print.rmean = TRUE)

 print(eval(substitute(survdiff(Surv(tenure, censured) ~ variable, data = x, rho = 0))))

 ggsurvplot(model,
  conf.int = TRUE,
  pval = TRUE,
  fun = "pct",
  risk.table = TRUE,
  size = 1,
  linetype = "strata",
  palette = c(
   "#E7B800",
   "#2E9FDF"
  ),
  legend = "bottom",
```

```
  legend.title = title,
 )
}


individual(female, "Female")

## Call: survfit(formula = Surv(tenure, censured) ~ female, data = x)
##
##           n events *rmean *se(rmean) median 0.95LCL 0.95UCL
## female=0 166    128   42.2      1.93     45     42     53
## female=1 185    137   40.8      1.86     48     37     53
##     * restricted mean with upper limit =  72
## Call:
## survdiff(formula = Surv(tenure, censured) ~ female, data = x,
##     rho = 0)
##
##           N Observed Expected (O-E)^2/E (O-E)^2/V
## female=0 166      128      133    0.210    0.469
## female=1 185      137      132    0.213    0.469
##
##  Chisq= 0.5  on 1 degrees of freedom, p= 0.5
```

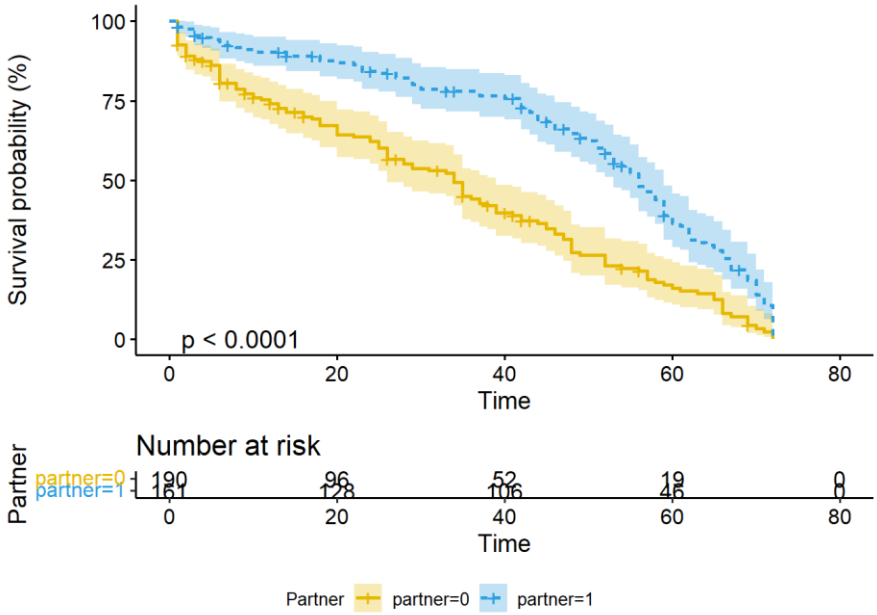



```
individual(senior_citizen, "Senior")

## Call: survfit(formula = Surv(tenure, censured) ~ senior_citizen, data = x)
##
##                   n events *rmean *se(rmean) median 0.95LCL 0.95UCL
## senior_citizen=0 303    237   41.0      1.43     46     42     51
## senior_citizen=1  48     28   44.8      3.89     53     37     62
##     * restricted mean with upper limit =  72
## Call:
## survdiff(formula = Surv(tenure, censured) ~ senior_citizen, data = x,
##     rho = 0)
##
```

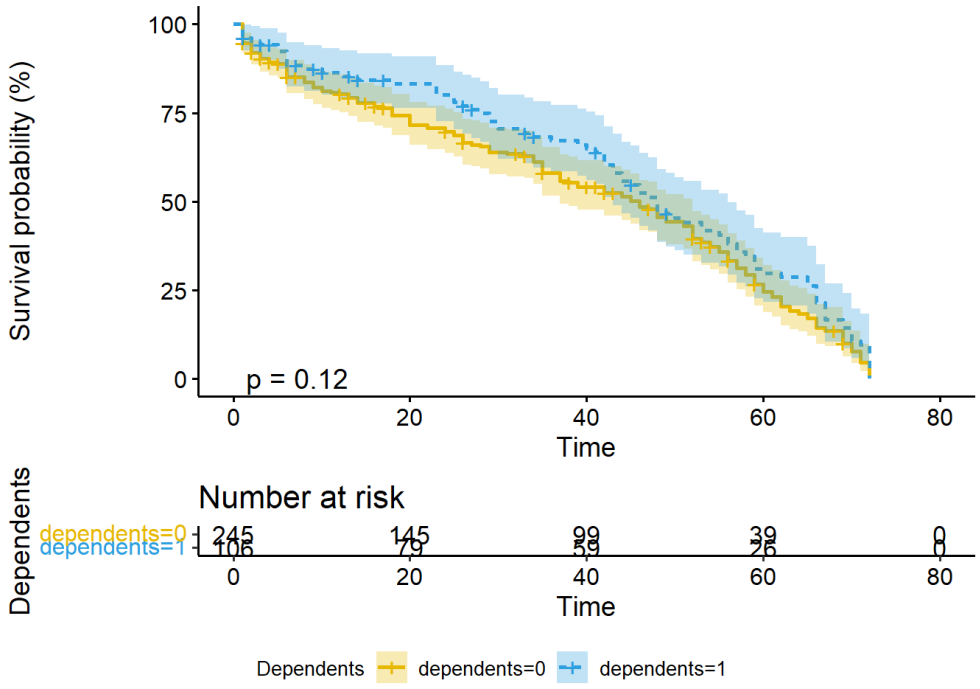
```
##                  N Observed Expected (O-E)^2/E (O-E)^2/V
## senior_citizen=0 303      237    232.8    0.0751     0.672
## senior_citizen=1  48       28     32.2    0.5431     0.672
##
##  Chisq= 0.7  on 1 degrees of freedom, p= 0.4
```




```
individual(partner, "Partner")

## Call: survfit(formula = Surv(tenure, censured) ~ partner, data = x)
##
##           n events *rmean *se(rmean) median 0.95LCL 0.95UCL
## partner=0 190    137   33.3      1.83     34     26     39
## partner=1 161    128   49.8      1.71     56     53     59
##     * restricted mean with upper limit =  72
## Call:
## survdiff(formula = Surv(tenure, censured) ~ partner, data = x,
##     rho = 0)
##
##           N Observed Expected (O-E)^2/E (O-E)^2/V
## partner=0 190      137    93.7     20.0       35
## partner=1 161      128   171.3     10.9       35
##
##  Chisq= 35  on 1 degrees of freedom, p= 3e-09
```




```
individual(dependents, "Dependents")

## Call: survfit(formula = Surv(tenure, censured) ~ dependents, data = x)
##
##                n events *rmean *se(rmean) median 0.95LCL 0.95UCL
## dependents=0 245    176   40.0      1.65     46     37     52
## dependents=1 106     89   44.6      2.26     48     43     56
##     * restricted mean with upper limit =  72
## Call:
## survdiff(formula = Surv(tenure, censured) ~ dependents, data = x,
##     rho = 0)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## dependents=0 245      176      164     0.809      2.41
## dependents=1 106       89      101     1.323      2.41
##
##  Chisq= 2.4  on 1 degrees of freedom, p= 0.1
```

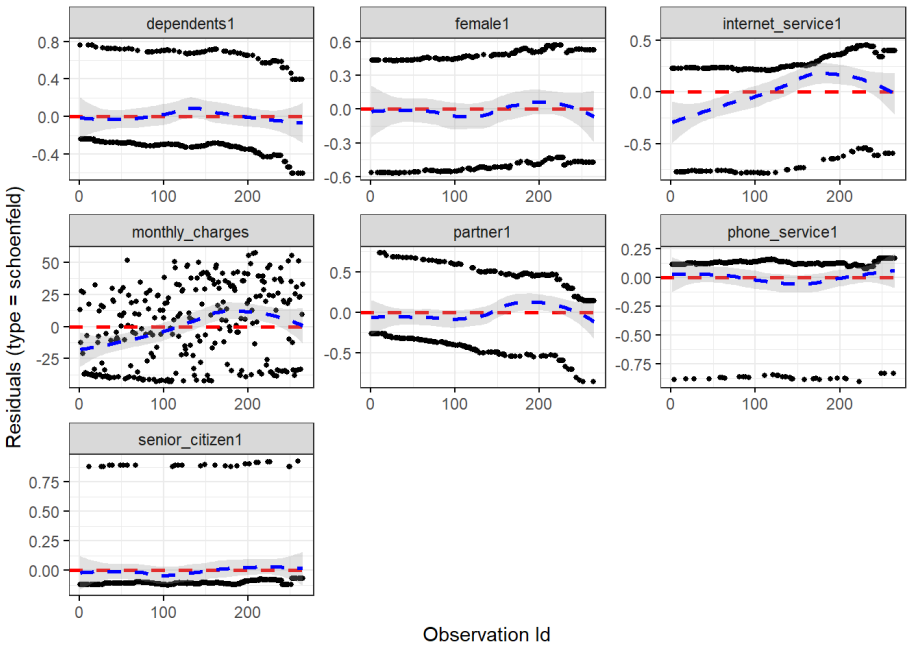



```
# Individualiai imant demografinius faktorius statistiškai reikšmingas skirtumas rastas tik
# su kintamuoju dependents
```

Sudarytas Cox semiparametrinis modelis, naudojantis visas duomenyse esančias kovariantes. Naudojantis schoenfeld liekanų grafikais ir proporcingų rizikos funkcijų statistiniu testu gauta, kad kovariantęs „internet_service“ ir „monthly_charges“ netenkina proporcingos rizikos prielaidos. Kadangi tai nėra tyrimo tikslams pagrindinės kovariantės, naudotas sluoksniavimas (angl. stratification), „monthly_charges“ kovariantę šiam tikslui diskretizuojant dalijant į tris grupes.

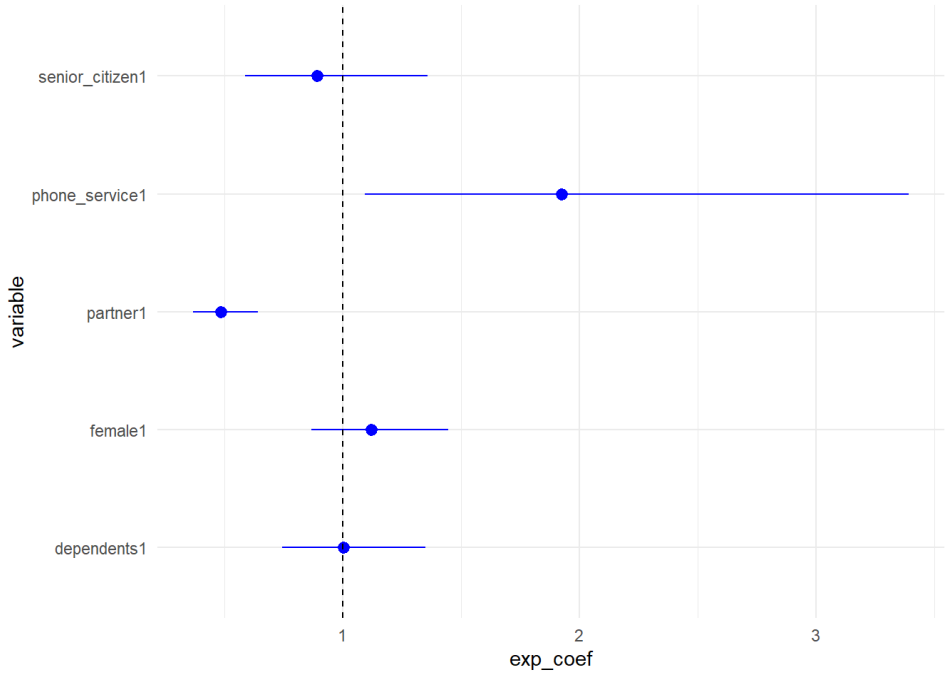
```
# Cox semiparametrinis modelis
cox <- coxph(Surv(tenure, censured) ~ phone_service + dependents + internet_service +
  senior_citizen + monthly_charges + female + partner, data = x)

summary(cox)

## Call:
## coxph(formula = Surv(tenure, censured) ~ phone_service + dependents +
##     internet_service + senior_citizen + monthly_charges + female +
##     partner, data = x)
##
##   n= 351, number of events= 265
##
##                     coef exp(coef)  se(coef)      z Pr(>|z|)
## phone_service1     0.944431  2.571349  0.254881  3.705 0.000211 ***
## dependents1       -0.017813  0.982344  0.149234 -0.119 0.904987
## internet_service1  1.652410  5.219546  0.312656  5.285 1.26e-07 ***
## senior_citizen1   -0.168112  0.845259  0.210612 -0.798 0.424752
## monthly_charges   -0.028832  0.971580  0.004602 -6.265 3.73e-10 ***
## female1            0.162386  1.176315  0.126779  1.281 0.200241
## partner1          -0.729154  0.482317  0.142046 -5.133 2.85e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
##                  exp(coef) exp(-coef) lower .95 upper .95
## phone_service1      2.5713     0.3889    1.5603    4.2376
## dependents1         0.9823     1.0180    0.7332    1.3161
## internet_service1   5.2195     0.1916    2.8281    9.6331
## senior_citizen1     0.8453     1.1831    0.5594    1.2772
## monthly_charges     0.9716     1.0293    0.9629    0.9804
## female1             1.1763     0.8501    0.9175    1.5081
## partner1            0.4823     2.0733    0.3651    0.6372
##
## Concordance= 0.689  (se = 0.016 )
## Likelihood ratio test= 74.27  on 7 df,   p=2e-13
## Wald test            = 74.42  on 7 df,   p=2e-13
## Score (logrank) test = 77.15  on 7 df,   p=5e-14

cox.zph(cox)

##                   chisq df      p
## phone_service     0.00288  1 0.95721
## dependents        0.03993  1 0.84162
## internet_service 19.07878  1 1.3e-05
## senior_citizen    0.72715  1 0.39381
## monthly_charges  22.09409  1 2.6e-06
## female            0.31643  1 0.57376
## partner           2.65362  1 0.10331
## GLOBAL           26.41982  7 0.00042

ggcoxdiagnostics(cox, type = "schoenfeld")
```

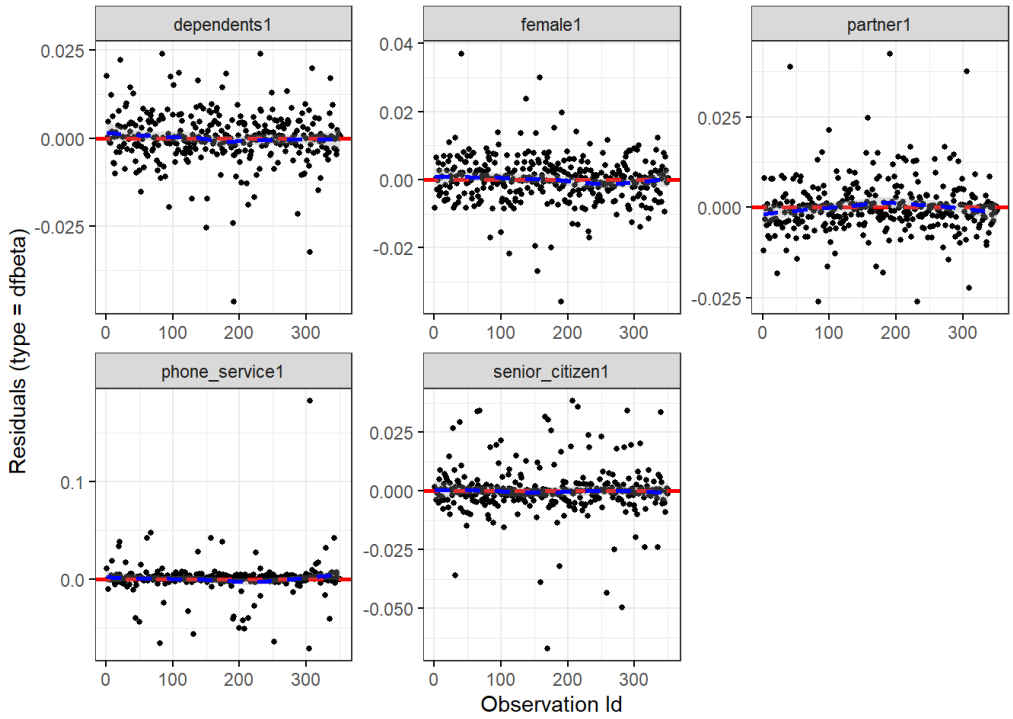



```
# Kadangi kovariantės, kurios pagal modelio diagnostikas pažeidžia proporcingų
# rizikos funkcijų prielaidą pagal tyrimo tikslus yra nepagrindinės (nuisance)
# naudojamas sluoksniavimas
x$monthly_charges_binned <- cut_number(x$monthly_charges, 3)

cox2 <- coxph(Surv(tenure, censured) ~ phone_service + dependents + female + partner +
  senior_citizen + strata(internet_service) + strata(monthly_charges_binned), data = x)
```

```
cox.zph(cox2)

##                chisq df     p
## phone_service  0.3296  1 0.566
## dependents     0.2475  1 0.619
## female         0.0023  1 0.962
## partner        4.1402  1 0.052
## senior_citizen 0.1165  1 0.733
## GLOBAL         5.2461  5 0.387

# Kovariančių poveikis multiplikatyvus
# exp(beta_i) lygus rizikos funkcijų santykiui

summary(cox2)

## Call:
## coxph(formula = Surv(tenure, censured) ~ phone_service + dependents +
##     female + partner + senior_citizen + strata(internet_service) +
##     strata(monthly_charges_binned), data = x)
##
##   n= 351, number of events= 265
##
##                    coef exp(coef)  se(coef)      z Pr(>|z|)
## phone_service1   0.654936  1.925018  0.288634  2.269   0.0233 *
## dependents1      0.003411  1.003417  0.151641  0.022   0.9821
## female1          0.112945  1.119570  0.130259  0.867   0.3859
## partner1        -0.722653  0.485462  0.142596 -5.068 4.02e-07 ***
## senior_citizen1 -0.113261  0.892917  0.213702 -0.530   0.5961
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                exp(coef) exp(-coef) lower .95 upper .95
## phone_service1    1.9250     0.5195    1.0933     3.389
## dependents1       1.0034     0.9966    0.7454     1.351
## female1           1.1196     0.8932    0.8673     1.445
## partner1          0.4855     2.0599    0.3671     0.642
## senior_citizen1   0.8929     1.1199    0.5874     1.357
##
## Concordance= 0.625  (se = 0.02 )
## Likelihood ratio test= 36.14  on 5 df,   p=9e-07
## Wald test            = 35.48  on 5 df,   p=1e-06
## Score (logrank) test = 36.53  on 5 df,   p=7e-07

exp(confint(cox2)) %>%
 cbind(exp(coef(cox2))) %>%
 as.data.frame() %>%
 tibble::rownames_to_column() %>%
 as_tibble() %>%
 set_names(c("variable", "low", "high", "exp_coef")) %>%
 ggplot(aes(variable, exp_coef)) +
 geom_pointrange(aes(ymin = low, ymax = high), color = "blue") +
 coord_flip() +
 theme_minimal() +
 geom_hline(yintercept = 1, color = "black", linetype = "dashed")
```




```
ggcoxdiagnostics(cox2, type = "dfbeta")
```




```
dfbetas <- resid(cox2, "dfbeta")
colnames(dfbetas) <- c("senior", "phone", "partner", "female", "dependents")

dfbetas <- dfbetas[, -c(2)]

ind_max <- function(column) {
  max <- sort(column)[1:4]
```

```
 which(column %in% max)
}

ind <- dfbetas %>%
 abs() %>%
 apply(2, ind_max)

x[ind[, "senior"], ]
```

```
## # A tibble: 4 x 10
##   tenure female monthly_charges phone_service internet_service senior_citizen
##    <int> <fct>           <dbl> <fct>       <fct>           <fct>
## 1     60 0              20.0 1           0               0
## 2     41 1             114.  1           1               0
## 3     42 1             105.  1           1               0
## 4     43 1             105.  1           1               0
## # ... with 4 more variables: dependents <fct>, partner <fct>, censured <dbl>,
## #   monthly_charges_binned <fct>
```

```
x[ind[, "partner"], ]
```

```
## # A tibble: 4 x 10
##   tenure female monthly_charges phone_service internet_service senior_citizen
##    <int> <fct>           <dbl> <fct>       <fct>           <fct>
## 1      1 1              69.2 1           1               1
## 2      1 1              86   1           1               1
## 3     42 1              67.7 1           1               0
## 4      1 1              74.4 1           1               1
## # ... with 4 more variables: dependents <fct>, partner <fct>, censured <dbl>,
## #   monthly_charges_binned <fct>
```

```
x[ind[, "female"], ]
```

```
## # A tibble: 4 x 10
##   tenure female monthly_charges phone_service internet_service senior_citizen
##    <int> <fct>           <dbl> <fct>       <fct>           <fct>
## 1      1 1              35.9 0           1               0
## 2     60 0              20.0 1           0               0
## 3     66 1             101.  1           1               1
## 4     60 1              59.8 0           1               0
## # ... with 4 more variables: dependents <fct>, partner <fct>, censured <dbl>,
## #   monthly_charges_binned <fct>
```

```
x[ind[, "dependents"], ]
```

```
## # A tibble: 4 x 10
##   tenure female monthly_charges phone_service internet_service senior_citizen
##    <int> <fct>           <dbl> <fct>       <fct>           <fct>
## 1     20 0              19.7 1           0               0
## 2      6 1              19.8 1           0               0
## 3     26 0              83.8 1           1               0
## 4     28 1              20.2 1           0               0
## # ... with 4 more variables: dependents <fct>, partner <fct>, censured <dbl>,
## #   monthly_charges_binned <fct>
```

```
# išskirtys išsiskiria labai ilgu (ar labai trumpu) paslaugų naudojimosi laiku,
# tačiau nėra priežasčių jas šalinti iš modelio.
```

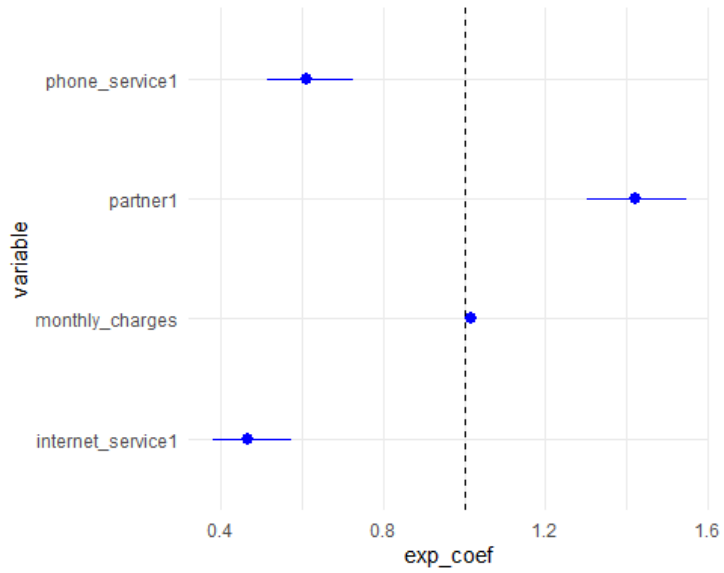
Gautame modelyje gauti pagerėję diagnostinių grafikų ir statistinio testo rezultatai lyginant su prieš tai sudarytu modeliu. Modelyje iš naudotų demografinių kovariančių statistiškai reikšminga tik „partner“. Partnerio turėjimas 51% sumažina riziką atsisakyti telekomunikacijų kompanijos paslaugų. Statistiškai reikšmingai paslaugų naudojimosi trukmę įtakoja ir telefono paslaugų naudojimasis, bet tai nėra tyrimui svarbi kovariantė.

Papildomai ištirtos galimos išskirtys pagal dfbeta kriterijų. Gautos išskirtys visoms dominančioms kovariantėms išsiskiria labai ilga arba labai trumpa paslaugų naudojimosi trukme, tačiau šios reikšmės yra galimos (nėra kilusios dėl duomenų įvedimo klaidų), todėl šių reikšmių pasirinkta nešalinti iš duomenų aibės, naudojamos sudaryti modelius.

```
library(eha)


aft <- aftreg(Surv(tenure, censured) ~ phone_service + dependents + internet_service + female
+ partner +
  senior_citizen + monthly_charges,
data = x, dist = "weibull", shape = 1, param="lifeExp"
) # eksponentinis skirstinys

aft2 <- aftreg(Surv(tenure, censured) ~ phone_service + dependents + internet_service + female
+ partner +
  senior_citizen + monthly_charges,
data = x, dist = "weibull", param="lifeExp"
)


aft3 <- aftreg(Surv(tenure, censured) ~ phone_service + dependents + internet_service + female
+ partner +
  senior_citizen + monthly_charges,
data = x, dist = "lognormal", param="lifeExp"
)


aft4 <- aftreg(Surv(tenure, censured) ~ phone_service + dependents + internet_service + female
+ partner +
  senior_citizen + monthly_charges,
data = x, dist = "loglogistic", param="lifeExp"
)

AIC(aft)

## [1] 2516.884

AIC(aft2)

## [1] 2471.753

AIC(aft3)

## [1] 2569.759

AIC(aft4)

## [1] 2548.15

# geriausi rezultatai gaunami su Weibull skirstiniu
```

```
summary(aft2)

## Covariate            W.mean      Coef Life-Expn  se(Coef)    LR p
## phone_service                                                0.0039
##                0      0.114     0         1   (reference)
##                1      0.886    -0.490     0.613     0.172
## dependents                                                   0.7863
##                0      0.648     0         1   (reference)
##                1      0.352    -0.027     0.973     0.100
## internet_service                                             0.0003
##                0      0.217     0         1   (reference)
##                1      0.783    -0.768     0.464     0.206
## female                                                       0.3278
##                0      0.489     0         1   (reference)
##                1      0.511    -0.085     0.919     0.086
## partner                                                      0.0001
##                0      0.404     0         1   (reference)
##                1      0.596     0.356     1.427     0.093
## senior_citizen                                               0.5631
##                0      0.877     0         1   (reference)
##                1      0.123     0.082     1.085     0.144
## monthly_charges      68.731     0.016     1.016     0.003   0.0000
##
## Events                    265
## Total time at risk         11806
## Max. log. likelihood      -1226.9
## LR test statistic         54.85
## Degrees of freedom        7
## Overall p-value           1.5987e-09

aft2_step <- step(aft2)

## Start:  AIC=2467.75
## Surv(tenure, censured) ~ phone_service + dependents + internet_service +
##     female + partner + senior_citizen + monthly_charges
##
##                    Df    AIC
## - dependents        1 2465.8
## - senior_citizen    1 2466.1
## - female            1 2466.7
## <none>                2467.8
## - phone_service     1 2474.1
## - internet_service  1 2478.8
## - partner           1 2480.3
## - monthly_charges   1 2492.7
##
## Step:  AIC=2465.83
## Surv(tenure, censured) ~ phone_service + internet_service + female +
##     partner + senior_citizen + monthly_charges
##
##                    Df    AIC
## - senior_citizen    1 2464.3
## - female            1 2464.8
## <none>                2465.8
## - phone_service     1 2472.1
## - internet_service  1 2477.1
## - partner           1 2480.1
## - monthly_charges   1 2491.2
##
## Step:  AIC=2464.26
## Surv(tenure, censured) ~ phone_service + internet_service + female +
```

```
##     partner + monthly_charges
##
##                     Df    AIC
## - female             1 2463.0
## <none>                 2464.3
## - phone_service      1 2470.5
## - internet_service   1 2475.2
## - partner            1 2478.6
## - monthly_charges    1 2489.5
##
## Step:  AIC=2463.03
## Surv(tenure, censured) ~ phone_service + internet_service + partner +
##     monthly_charges
##
##                     Df    AIC
## <none>                 2463.0
## - phone_service      1 2469.4
## - internet_service   1 2474.0
## - partner            1 2477.8
## - monthly_charges    1 2487.8

summary(aft2_step)

## Covariate            W.mean      Coef Life-Expn  se(Coef)    LR p
## phone_service                                                0.0038
##                0      0.114     0         1   (reference)
##                1      0.886    -0.490     0.612     0.172
## internet_service                                             0.0003
##                0      0.217     0         1   (reference)
##                1      0.783    -0.760     0.468     0.205
## partner                                                      0.0000
##                0      0.404     0         1   (reference)
##                1      0.596     0.350     1.420     0.085
## monthly_charges      68.731     0.016     1.016     0.003   0.0000
##
## Events                    265
## Total time at risk         11806
## Max. log. likelihood      -1227.5
## LR test statistic         53.57
## Degrees of freedom        4
## Overall p-value           6.47457e-11

# exp(beta_i) parodo kiek kartų padidėjo laikas iki įvykio (išgyvenamumo funkcija)
# exp(beta_i) > 0 -> įvykis įvyksta vėliau
# exp(beta_i) < 0 -> įvykis įvyksta anksčiau

summary(aft2_step)$coefficients %>%
  as.data.frame() %>%
  rownames_to_column() %>%
  select(rowname, coef, `se(coef)`) %>%
  mutate(low = exp(coef - `se(coef)`), high = exp(coef + `se(coef)`), coef = exp(coef)) %>%
  select(-`se(coef)`) %>%
  set_names(c("variable", "exp_coef", "low", "high")) %>%
  ggplot(aes(variable, exp_coef)) +
  geom_pointrange(aes(ymin = low, ymax = high), color = "blue") +
  coord_flip() +
  theme_minimal() +
  geom_hline(yintercept = 1, color = "black", linetype = "dashed")
```



Sudarytas parametrinis AFT (Accelerated Failure Time) modelis naudojantis visas kovariantes. Pagal AIC kriterijų lyginti modeliai, naudojant eksponentinį, Veibulo, loglogistinį ir lognormalujį skirstinius. Geriausi rezultatai gauti naudojant Veibulo skirstinį. Pilname modelyje parnerio turėjimas gauta kaip vienintelė statistiškai reikšminga demografinė kovariantė.

Naudojant pažingsninę regresiją gautas sumažintas modelis apskritai priklauso tik nuo vieno demografinio faktoriaus: partnerio turėjimas 42% pailgina laiką iki paslaugų atsisakymo. Pagal šį modelį išgyvenamumo funkcijai statistiškai reikšmingą poveikį taip pat turi interneto, telefono paslaugų naudojimasis, per mėnesį kliento sumokama suma, bet tai nėra pagrindinės kovariantės šio tyrimo tikslui.

Išvados

Tyrime siekta ištirti, kokia yra demografinių faktorių įtaka telekomunikacijų bendrovės paslaugų atsisakymo trukmei.

Pagal Kaplan-Meier išgyvenamumo funkcijų įverčius gautas statistiškai reikšmingas skirtumas tarp partnerius turinčių ir jų neturinčių klientų. Tarp kitų demografinių faktorių lygmenų statistiškai reikšmingų skirtumų nerasta.

Kadangi toks vertinimo metodas neatsižvelgia į kitų kovariančių reikšmes, sudaryti regresijos modeliai. Sudarant Cox semiparametrinį modelį nepagrindinėms (su klientų demografinėmis grupėmis nesusijusioms) kovariantėms, kurioms negaliojo proporcingų rizikos funkcijų prielaida, naudotas sluoksniavimas (stratifikacija). Šiuo modeliu gauta statistiškai reikšminga partnerio turėjimo įtaka ($p<0.001$). Partnerio turėjimas 51% sumažina riziką atsisakyti telekomunikacijos įmonės paslaugų. Modeliu taip pat gauta, kad moterys klientės turi 11% didesnę riziką atsisakyti įmonės paslaugų, pensijinio amžiaus klientai – 11% mažesnę riziką, tačiau abi šios kovariantės, kaip ir išlaikytinių turėjimas, modelyje nebuvo statistiškai reikšmingos.

Rezultatų palyginimui sudarytas parametrinis AFT modelis. Geriausi rezultatai gauti naudojant Veibulo skirstinį. Panaudojus pažingsninę regresiją galutinis modelis priklausė tik nuo vieno demografinio požymio – parterio turėjimo. Pagal šį modelį partnerį turintys klientai paslaugomis naudojasi 42% ilgiau, negu neturintys klientai.

Abu sudaryti regresijos modeliai patvirtina Kaplan-Maier kreivėmis gautus rezultatus, kad svarbiausias demografinis požymis, veikiantis paslaugų naudojimosi trukmę, yra ar klientas turi partnerį ar ne.