Išgyvenimumo analizė

Vainius Gataveckas, Matas Gaulia, Dovydas Martinkus

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

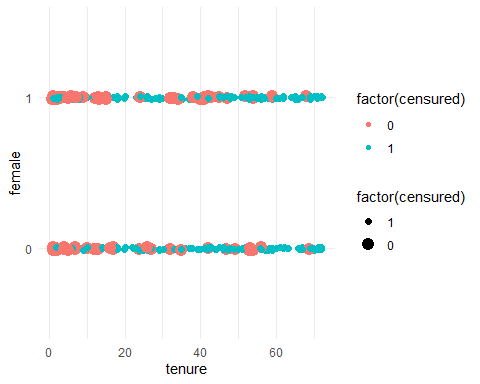
# vidutinis laikas neįskaitant cenzūravimo  
mean(x$tenure[x$censured == 0])

## [1] 19.51163

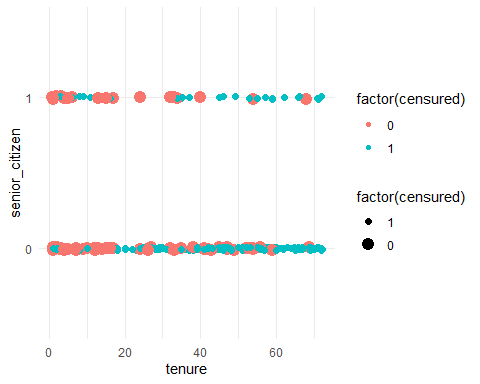
# bendras vidutinis laikas (mažesnis negu yra iš tikrųjų dėl cenzūravimo)  
mean(x$tenure)

## [1] 33.63533

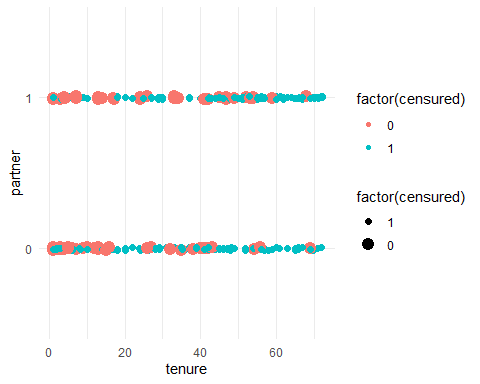
plot <- function(column) {  
 ggplot(x, aes({{ column }}, tenure, color = factor(censured))) +  
 geom\_point(aes(size = factor(censured)), position = position\_jitter(width = 0.01, height = 0.1)) +  
 scale\_x\_discrete() +  
 coord\_flip() +  
 theme\_minimal() +  
 scale\_size\_manual(values = c("1" = 2, "0" = 4))  
}  
  
  
plot(female)



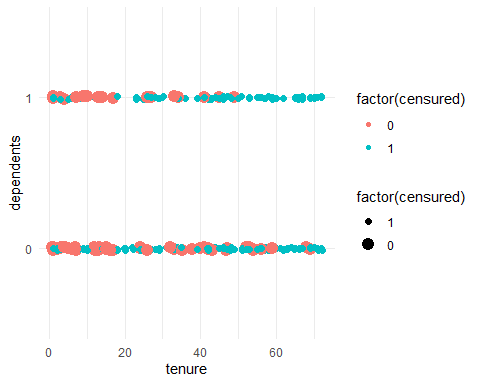
plot(senior\_citizen)



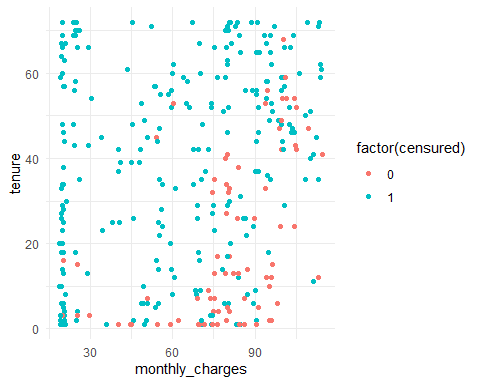
plot(partner)



plot(dependents)



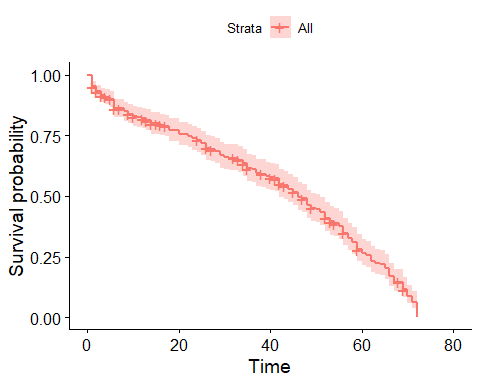
ggplot(x, aes(monthly\_charges, tenure, color = factor(censured))) +  
 geom\_point() +  
 theme\_minimal()



km <- survfit(Surv(tenure, censured) ~ 1, data = x)  
  
print(km, print.rmean = TRUE) # 46 mediana, 41 vidurkis (palyginti su prieš tai gautu vidurkiu)

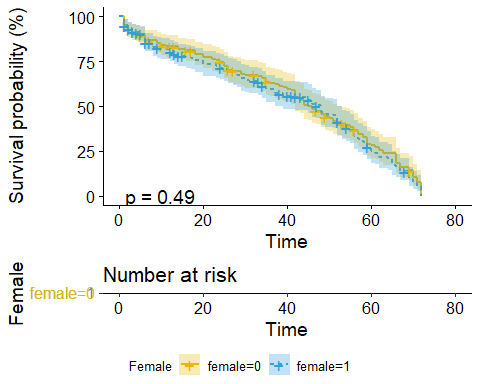
## Call: survfit(formula = Surv(tenure, censured) ~ 1, data = x)  
##   
## n events \*rmean \*se(rmean) median 0.95LCL 0.95UCL   
## 351.00 265.00 41.48 1.34 46.00 42.00 51.00   
## \* restricted mean with upper limit = 72

ggsurvplot(km)



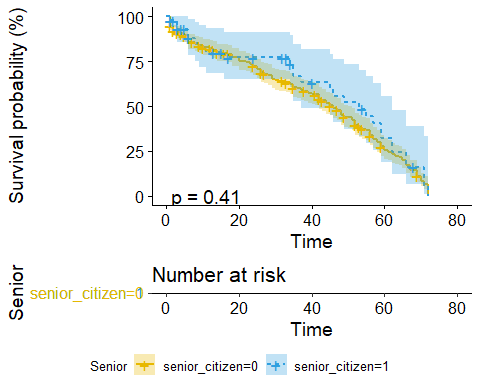
individual <- function(variable, title) {  
 model <-  
 eval(substitute(survfit(Surv(tenure, censured) ~ variable, data = x)))  
  
 print(model, print.rmean = TRUE)  
  
 print(eval(substitute(survdiff(Surv(tenure, censured) ~ variable, data = x, rho = 0))))  
  
 ggsurvplot(model,  
 conf.int = TRUE,  
 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



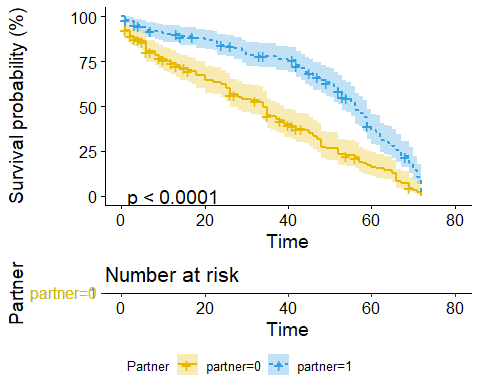
individual(senior\_citizen, "Senior")

## 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



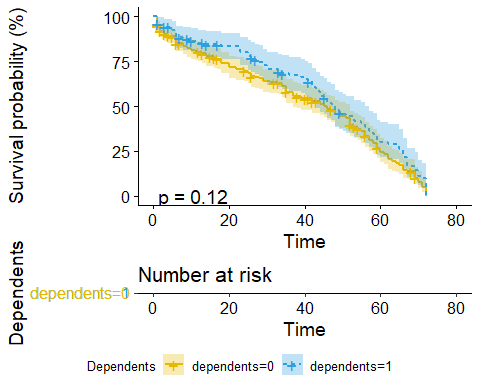
individual(partner, "Partner")

## 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

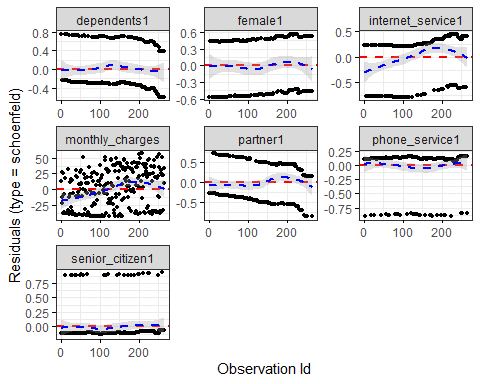
# Cox semiparametrinis modelis  
cox <- coxph(Surv(tenure, censured) ~ phone\_service + dependents + internet\_service +  
 senior\_citizen + monthly\_charges + female + partner, data = x)  
  
summary(cox)

## 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)

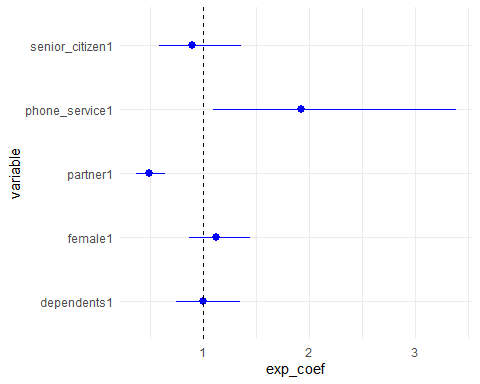
cox.zph(cox2)

## 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 0.042  
## senior\_citizen 0.1165 1 0.733  
## GLOBAL 5.2461 5 0.387

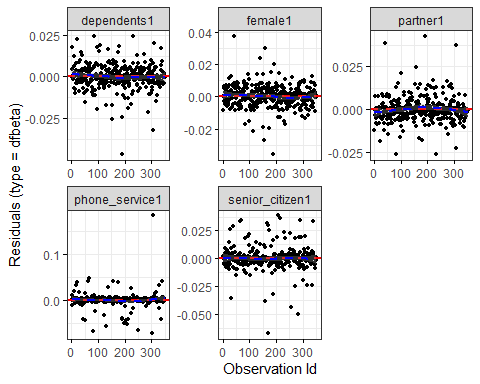
# Kovariančių poveikis multiplikatyvus  
# exp(beta\_i) lygus rizikos funkcijų santykiui  
  
summary(cox2)

## 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  
## <dbl> <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  
## <dbl> <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  
## <dbl> <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  
## <dbl> <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.

library(eha)  
  
  
aft <- aftreg(Surv(tenure, censured) ~ phone\_service + dependents + internet\_service + female +  
 senior\_citizen + monthly\_charges,  
data = x, dist = "weibull", shape = 1  
) # eksponentinis skirstinys  
  
aft2 <- aftreg(Surv(tenure, censured) ~ phone\_service + dependents + internet\_service + female +  
 senior\_citizen + monthly\_charges,  
data = x, dist = "weibull"  
)  
  
  
aft3 <- aftreg(Surv(tenure, censured) ~ phone\_service + dependents + internet\_service + female +  
 senior\_citizen + monthly\_charges,  
data = x, dist = "lognormal"  
)  
  
  
aft4 <- aftreg(Surv(tenure, censured) ~ phone\_service + dependents + internet\_service + female +  
 senior\_citizen + monthly\_charges,  
data = x, dist = "loglogistic"  
)

AIC(aft)

## [1] 2523.455

AIC(aft2)

## [1] 2484.269

AIC(aft3)

## [1] 2583.786

AIC(aft4)

## [1] 2563.306

# geriausi rezultatai gaunami su Weibull skirstiniu

summary(aft2)

## Covariate W.mean Coef Time-Accn se(Coef) LR p  
## phone\_service 0.0042   
## 0 0.114 0 1 (reference)  
## 1 0.886 0.499 1.648 0.177  
## dependents 0.1798   
## 0 0.648 0 1 (reference)  
## 1 0.352 -0.125 0.883 0.093  
## internet\_service 0.0001   
## 0 0.217 0 1 (reference)  
## 1 0.783 0.842 2.321 0.212  
## female 0.2549   
## 0 0.489 0 1 (reference)  
## 1 0.511 0.101 1.106 0.089  
## senior\_citizen 0.3273   
## 0 0.877 0 1 (reference)  
## 1 0.123 -0.140 0.869 0.146  
## monthly\_charges 68.731 -0.018 0.982 0.003 0.0000   
##   
## Events 265   
## Total time at risk 11806   
## Max. log. likelihood -1234.1   
## LR test statistic 40.33   
## Degrees of freedom 6   
## Overall p-value 3.92112e-07

aft2\_step <- step(aft2)

## Start: AIC=2480.27  
## Surv(tenure, censured) ~ phone\_service + dependents + internet\_service +   
## female + senior\_citizen + monthly\_charges  
##   
## Df AIC  
## - senior\_citizen 1 2479.2  
## - female 1 2479.6  
## - dependents 1 2480.1  
## <none> 2480.3  
## - phone\_service 1 2486.4  
## - internet\_service 1 2493.0  
## - monthly\_charges 1 2509.0  
##   
## Step: AIC=2479.23  
## Surv(tenure, censured) ~ phone\_service + dependents + internet\_service +   
## female + monthly\_charges  
##   
## Df AIC  
## - female 1 2478.2  
## - dependents 1 2478.6  
## <none> 2479.2  
## - phone\_service 1 2485.5  
## - internet\_service 1 2491.4  
## - monthly\_charges 1 2507.7  
##   
## Step: AIC=2478.25  
## Surv(tenure, censured) ~ phone\_service + dependents + internet\_service +   
## monthly\_charges  
##   
## Df AIC  
## - dependents 1 2477.8  
## <none> 2478.2  
## - phone\_service 1 2484.6  
## - internet\_service 1 2490.4  
## - monthly\_charges 1 2506.1  
##   
## Step: AIC=2477.75  
## Surv(tenure, censured) ~ phone\_service + internet\_service + monthly\_charges  
##   
## Df AIC  
## <none> 2477.8  
## - phone\_service 1 2484.2  
## - internet\_service 1 2489.6  
## - monthly\_charges 1 2504.9

summary(aft2\_step)

## Covariate W.mean Coef Time-Accn se(Coef) LR p  
## phone\_service 0.0036   
## 0 0.114 0 1 (reference)  
## 1 0.886 0.510 1.665 0.178  
## internet\_service 0.0002   
## 0 0.217 0 1 (reference)  
## 1 0.783 0.820 2.272 0.213  
## monthly\_charges 68.731 -0.017 0.983 0.003 0.0000   
##   
## Events 265   
## Total time at risk 11806   
## Max. log. likelihood -1235.9   
## LR test statistic 36.86   
## Degrees of freedom 3   
## Overall p-value 4.93799e-08

# 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")

