

Analysis of Employee attrition

Zeralm

10/14/2021

The following code generates descriptive statistics and basic plots, cleans data and performs some hypothesis tests. Finally three Survival Analysis models are applied in order to ascertain the predictors of employee turnover.

The data is real, provided by Edward Babushkin - <https://edwvb.blogspot.com/2017/10/employee-turnover-how-to-predict-individual-risks-of-quitting.html?m=1>

Cleaning

```
##          stag          event  gender      age          industry
##  Min.    : 0.3942   Quit   :553  m:275   Min.    :18.00   Retail      :280
##  1st Qu.: 11.7125   Stayed:554 f:832   1st Qu.:25.00   manufacture:143
##  Median : 24.4107                                Median :30.00   IT          :122
##  Mean    : 36.6903                                Mean    :31.03   Banks       :111
##  3rd Qu.: 51.4497                                3rd Qu.:36.00   etc         : 92
##  Max.    :179.4497                                Max.    :58.00   Consult     : 73
##                                           (Other)    :286
##
##          profession      traffic
##  HR              :739   youjs      :311
##  IT              : 74   empjs      :247
##  Sales           : 65   rabrecNErab:206
##  etc             : 37   friends    :115
##  Marketing       : 30   referral   : 94
##  BusinessDevelopment: 27   KA         : 65
##  (Other)         :135   (Other)    : 69
##
##          coach  head_gender  greywage  way  extraversion
##  no          :667  f:536    white:984  bus :668  Min.    : 1.000
##  yes         :130  m:571    grey :123  car :325  1st Qu.: 4.600
##  my head:310                                foot:114  Median : 5.400
##                                           Mean    : 5.578
##                                           3rd Qu.: 7.000
##                                           Max.    :10.000
##
##          independ  selfcontrol  anxiety  novator
##  Min.    : 1.00  Min.    : 1.000  Min.    : 1.700  Min.    : 1.000
##  1st Qu.: 4.10  1st Qu.: 4.100  1st Qu.: 4.800  1st Qu.: 4.400
##  Median : 5.50  Median : 5.700  Median : 5.600  Median : 6.000
##  Mean    : 5.47  Mean    : 5.616  Mean    : 5.674  Mean    : 5.878
##  3rd Qu.: 6.90  3rd Qu.: 7.200  3rd Qu.: 7.100  3rd Qu.: 7.500
##  Max.    :10.00  Max.    :10.000  Max.    :10.000  Max.    :10.000
```

Visualization

```
# -

# Age distr
graph1 <- ggplot(data, mapping = aes(x = age)) + geom_density() + facet_grid(event~.) +
  labs(title = "Age distribution") +
  theme(plot.title = element_text(hjust = 0.5), strip.text.y = element_text(angle = 0), strip.background = element_rect(fill = "white", stroke = "black"))

# experience vs age
graph2.1 <- data[data$event == "Stayed",] %>% group_by(age) %>% summarize(mean(stag)) %>%
  rename(avg_stag = `mean(stag)`) %>%
  ggplot(mapping = aes(x = age, y = avg_stag)) + geom_col(fill = "Grey") +
  labs(title = "Average experience, people who stayed", x = "Age", y = "Experience") +
  theme(plot.title = element_text(hjust = 0.5))

graph2.2 <- data[data$event == "Quit",] %>% group_by(age) %>% summarize(mean(stag)) %>%
  rename(avg_stag = `mean(stag)`) %>%
  ggplot(mapping = aes(x = age, y = avg_stag)) + geom_col(fill = "Grey") +
  labs(title = "Average experience, people who left", x = "Age", y = "Experience") +
  theme(plot.title = element_text(hjust = 0.5))

graph3 <- melt(data %>% select(event, (extraversion:novator))) %>%
  ggplot(mapping = aes(x = variable, y = value, fill = event), size = 10) +
  stat_summary(fun = mean, position = "dodge", geom = "bar", color = "black") + labs(title = "Big 5 scores") +
  theme(plot.title = element_text(hjust = 0.5), axis.title.x = element_text(color = "white"), legend.title = element_text(color = "white")) +
  scale_fill_brewer(palette = "Greys")

graph4 <- data %>% ggplot(mapping = aes(x = age, event)) + geom_boxplot(aes(fill = event), color = "black") +
  coord_flip() + labs(title = "Age distribution", x = "Age") +
  theme(plot.title = element_text(hjust = 0.5), axis.title.x = element_blank(), legend.title = element_text(color = "white")) +
  scale_fill_brewer(palette = "Greys")

graph5 <- data %>% ggplot(mapping = aes(x = industry)) +
  geom_bar(aes(fill = event), color = "black", position = "dodge", show.legend = FALSE) +
  theme(plot.title = element_text(hjust = 0.5), axis.title.y = element_blank(), legend.title = element_text(color = "white")) +
  labs(title = "Number of people per industry", x = "") + scale_x_discrete(guide = guide_axis(n.dodge = 2)) +
  scale_fill_brewer(palette = "Greys")

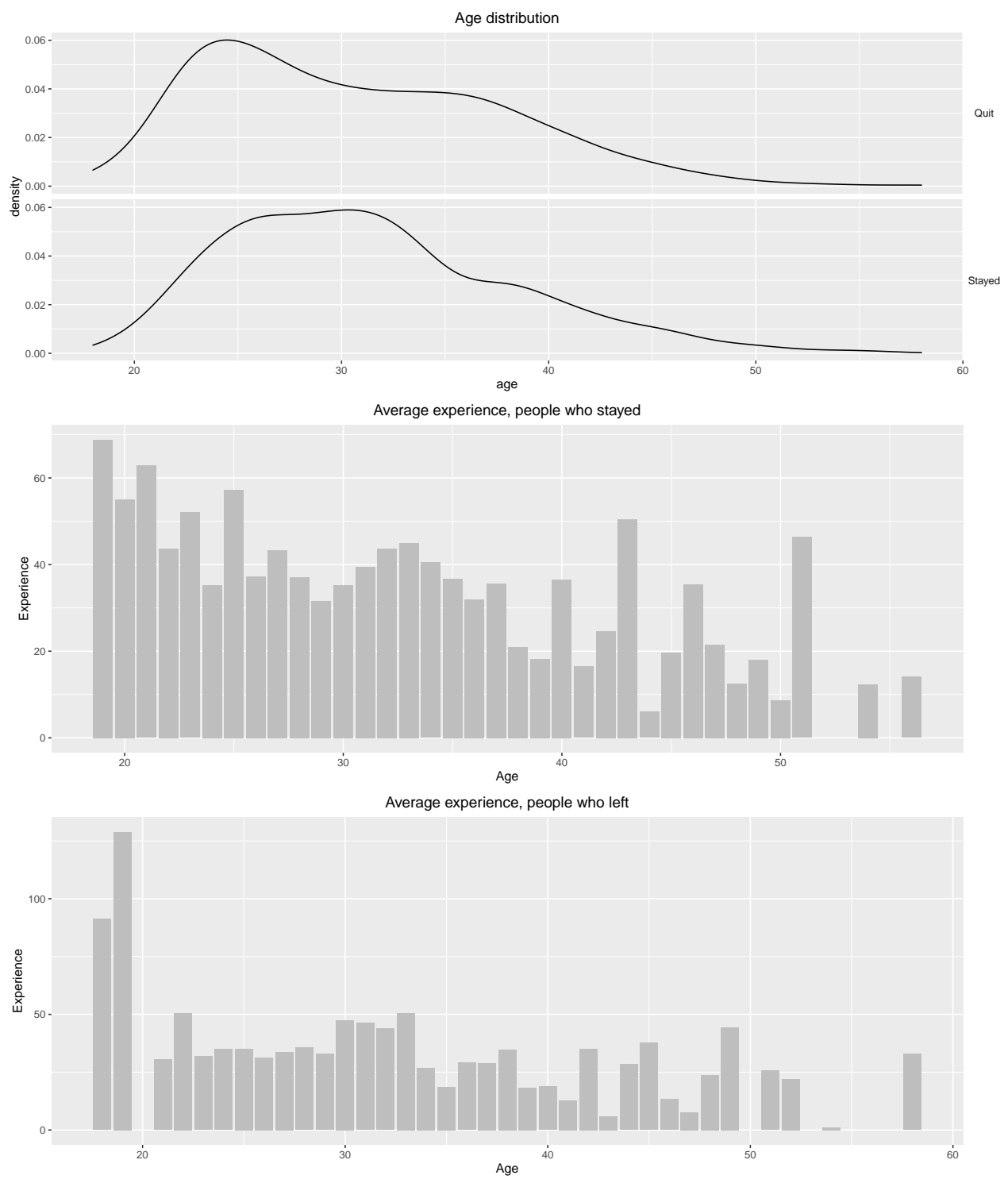
graph6 <- data %>% ggplot(mapping = aes(x = profession)) +
  geom_bar(aes(fill = event), color = "black", position = "dodge", show.legend = FALSE) +
  theme(plot.title = element_text(hjust = 0.5), axis.title.y = element_blank(), legend.title = element_text(color = "white")) +
  labs(title = "Number of people per profession", x = "") + scale_x_discrete(guide = guide_axis(n.dodge = 2)) +
  scale_fill_brewer(palette = "Greys")

graph7 <- data %>% ggplot(mapping = aes(x = way)) +
  geom_bar(aes(fill = event), color = "black", position = "dodge", show.legend = FALSE) +
  theme(plot.title = element_text(hjust = 0.5), axis.title.y = element_blank(), legend.title = element_text(color = "white")) +
  labs(title = "Commute choice", x = "") + scale_fill_brewer(palette = "Greys")

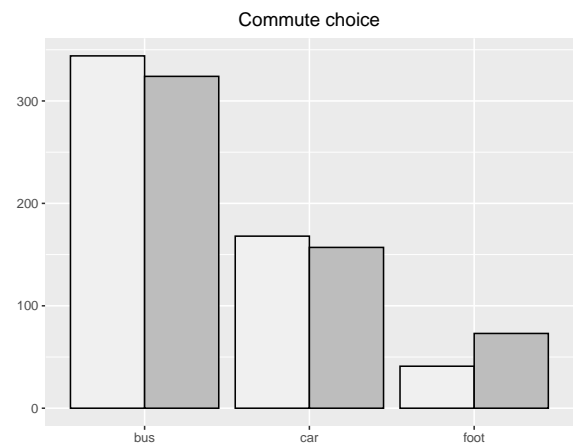
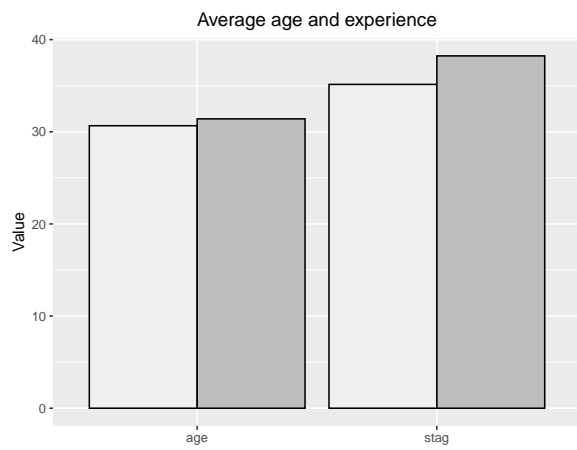
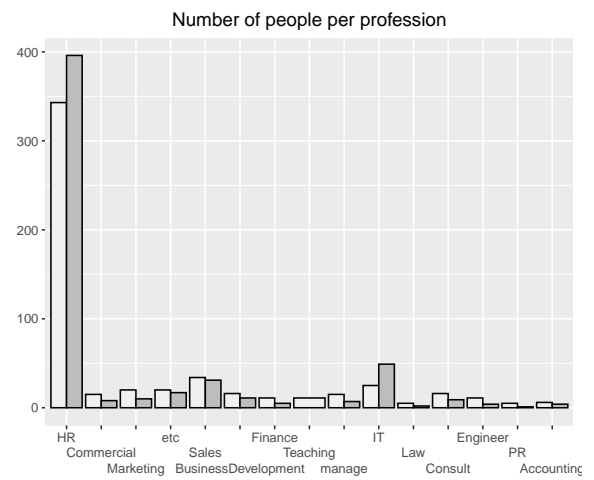
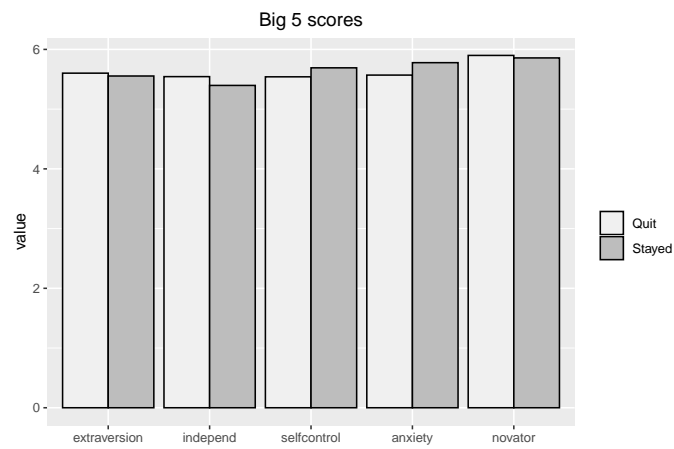
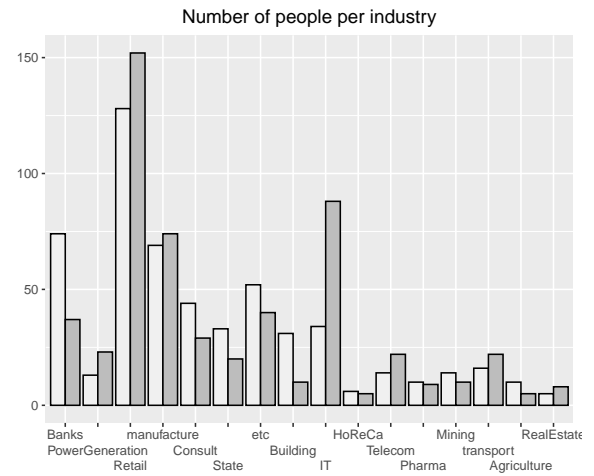
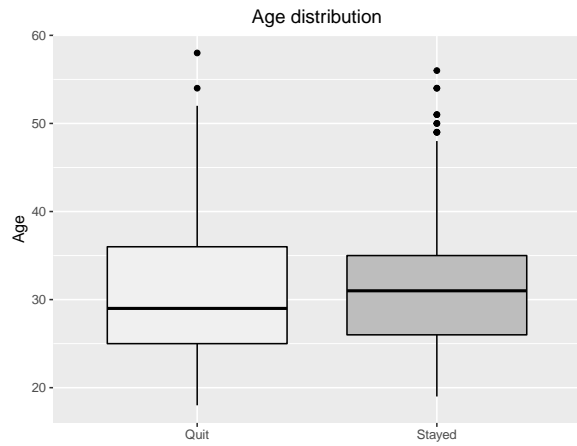
graph8 <- melt(data %>% select(event, age, stag)) %>% ggplot(mapping = aes(x = variable, y = value, fill = event)) +
  stat_summary(fun = mean, color = "black", position = "dodge", geom = "bar", show.legend = FALSE) +
  theme(plot.title = element_text(hjust = 0.5), legend.title = element_blank()) +
```

```
scale_fill_brewer(palette = "Greys")

ggarrange(graph1, graph2.1, graph2.2, nrow = 3, widths = c(2, c(1, 1)))
```



```
ggarrange(graph4 ,graph5, graph3, graph6, graph8, graph7)
```



Statistical tests

- STATISTICAL TESTS

$H_0: B_1 = B_2 = B_3 = B_4 = B_5 = 0$

```

# H1: B1 != 0 or B2 != 0, or both, or..., or all

big5_event <- data[data$event == "Quit", names(data) %in% c("independ", "anxiety", "extraversion", "no
big5_no_event <- data[data$event == "Stayed", names(data) %in% c("independ", "anxiety", "extraversion
big5.pvalue <- HotellingsT2Test(big5_event, big5_no_event)$p.value

# H0: p = 0.5 (probability of leaving)
# H1: p < 0.5

male_event = dim(data[(data$event == "Quit") & (data$gender == "m"), "gender"])[1]
len_male_evth = dim(data[data$gender == "m", "gender"])[1]
male.pvalue <- pbinom(male_event, len_male_evth, 0.5) #Cannot reject

female_event = dim(data[(data$event == "Quit") & (data$gender == "f"), "gender"])[1]
len_female_evth = dim(data[data$gender == "f", "gender"])[1]
female.pvalue <- pbinom(female_event, len_female_evth, 0.5) #Cannot reject

head_male_event = dim(data[(data$event == "Quit") & (data$head_gender == "m"), "head_gender"])[1]
len_head_male_evth = dim(data[data$head_gender == "m", "head_gender"])[1]
headmale.pvalue <- pbinom(head_male_event, len_head_male_evth, 0.5) #Cannot reject

head_female_event = dim(data[(data$event == "Quit") & (data$head_gender == "f"), "head_gender"])[1]
len_head_female_evth = dim(data[data$head_gender == "f", "head_gender"])[1]
headfemale.pvalue <- pbinom(head_female_event, len_head_female_evth, 0.5) #Cannot reject

# H0: B1 = 0
# H1: B1 != 0

age_event = data[data$event == "Quit", "age"]
age_no_event = data[data$event == "Stayed", "age"]
age.pvalue <- t.test(age_event, age_no_event)$p.value #Cannot reject

stag_event = data[data$event == "Quit", "stag"]
stag_no_event = data[data$event == "Stayed", "stag"]
stag.pvalue <- t.test(stag_event, stag_no_event)$p.value #Cannot reject

pvalues = c(big5.pvalue, male.pvalue, female.pvalue, headmale.pvalue, headfemale.pvalue, age.pvalue, stag.pvalue)
testnames = c("Hotelling's t-test: Big 5", "binomial test: males", "binomial test: females", "binomial test: head males", "binomial test: head females", "t-test: age", "t-test: stag")
significant_at_0.05 = pvalues < 0.05

data.frame(testnames, pvalues, significant_at_0.05)

```

```

##          testnames    pvalues significant_at_0.05
## 1 Hotelling's t-test: Big 5 0.21493521          FALSE
## 2   binomial test: males 0.35878148          FALSE
## 3   binomial test: females 0.59586499          FALSE
## 4 binomial test: head males 0.82138355          FALSE
## 5 binomial test: head females 0.18219375          FALSE
## 6           t-test: age 0.07625558          FALSE
## 7           t-test: stag 0.13239632          FALSE

```

Survival analysis

Why survival analysis, instead of logistic regression?

- The coefficients on tenure would not be useful. This variable is mismanaged in logistic regressions: we don't want to predict the turnover of current employees. We want to predict the turnover of future employees, so current tenure is irrelevant and unactionable.
- Survival analysis allows you to be precise and detect breakpoints and trends.
- If you actually want to find breakpoints or changes in slope (splines), you should probably plot the dataset like a distribution of tenure-survival. That would already resemble survival analysis, but in a much less precise way.
- It will be more accurate answering our practical questions on turnover.

```
newdat <- read_csv("data/turnover-data-set_utf.csv", show_col_types = FALSE,
                  col_types = "diifffffffddddd")
newdat <- newdat[!duplicated(newdat) & !apply(is.na(newdat), 1, any),]

train <- newdat[sample(nrow(newdat), nrow(newdat) * 0.7), ]
test <- newdat[setdiff(seq_len(nrow(newdat)), sample(nrow(newdat), nrow(newdat) * 0.7)), ]
```

Kaplan-Meier model

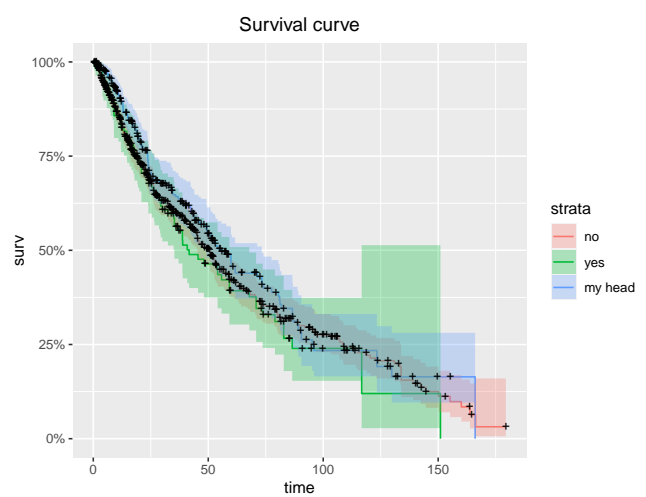
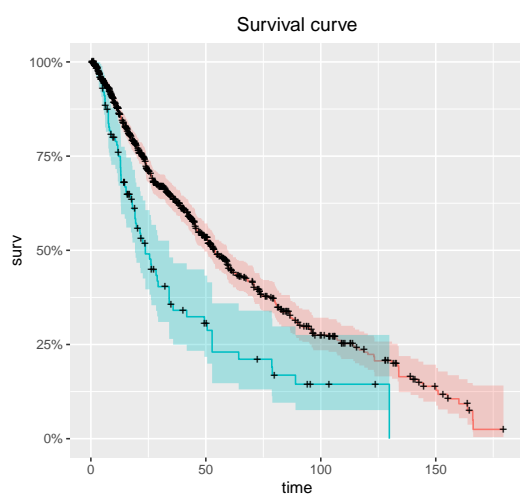
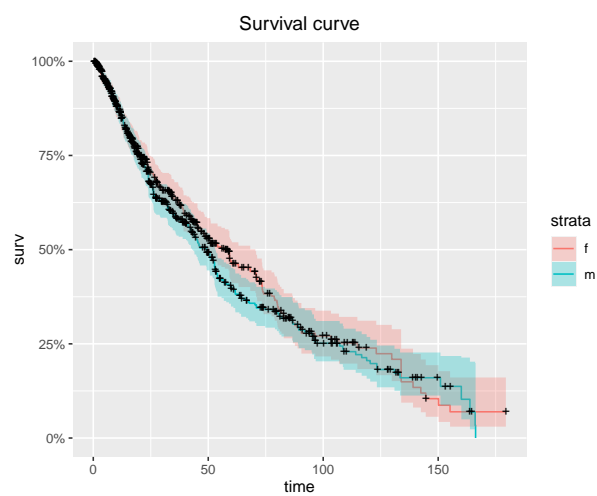
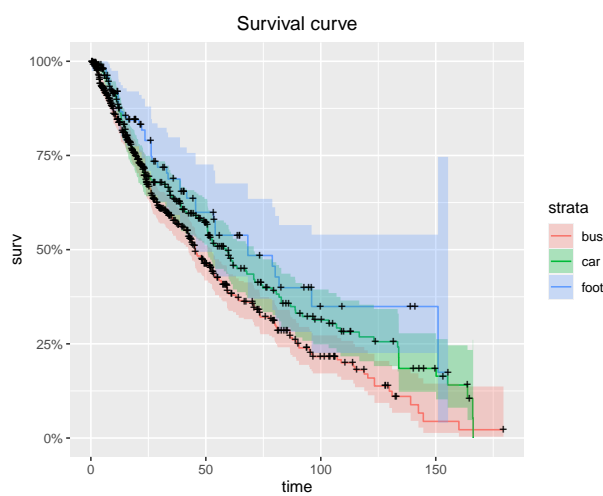
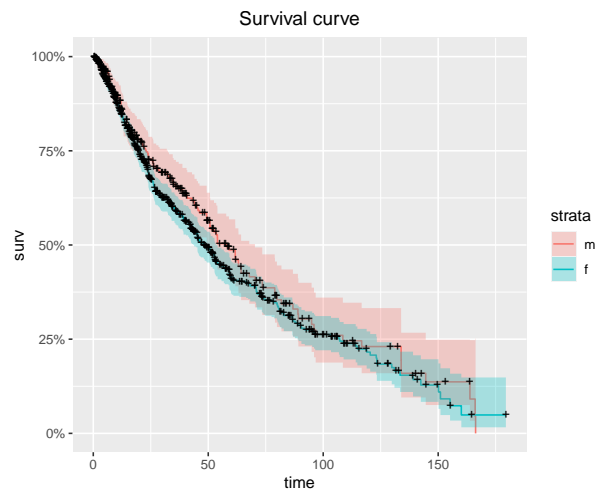
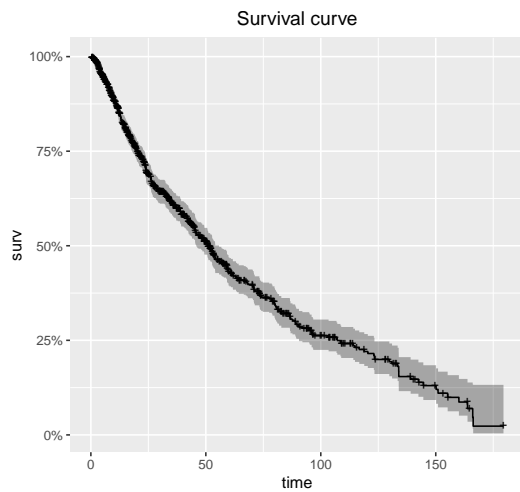
The simplest possible model and our baseline. It takes into account a single independent variable, namely, tenure. However, we can generate many survival curves for every possible realization of another variable. This already provides useful insight.

```
objSurv <- with(train, Surv(stag, event))
survfit(objSurv ~ 1)
```

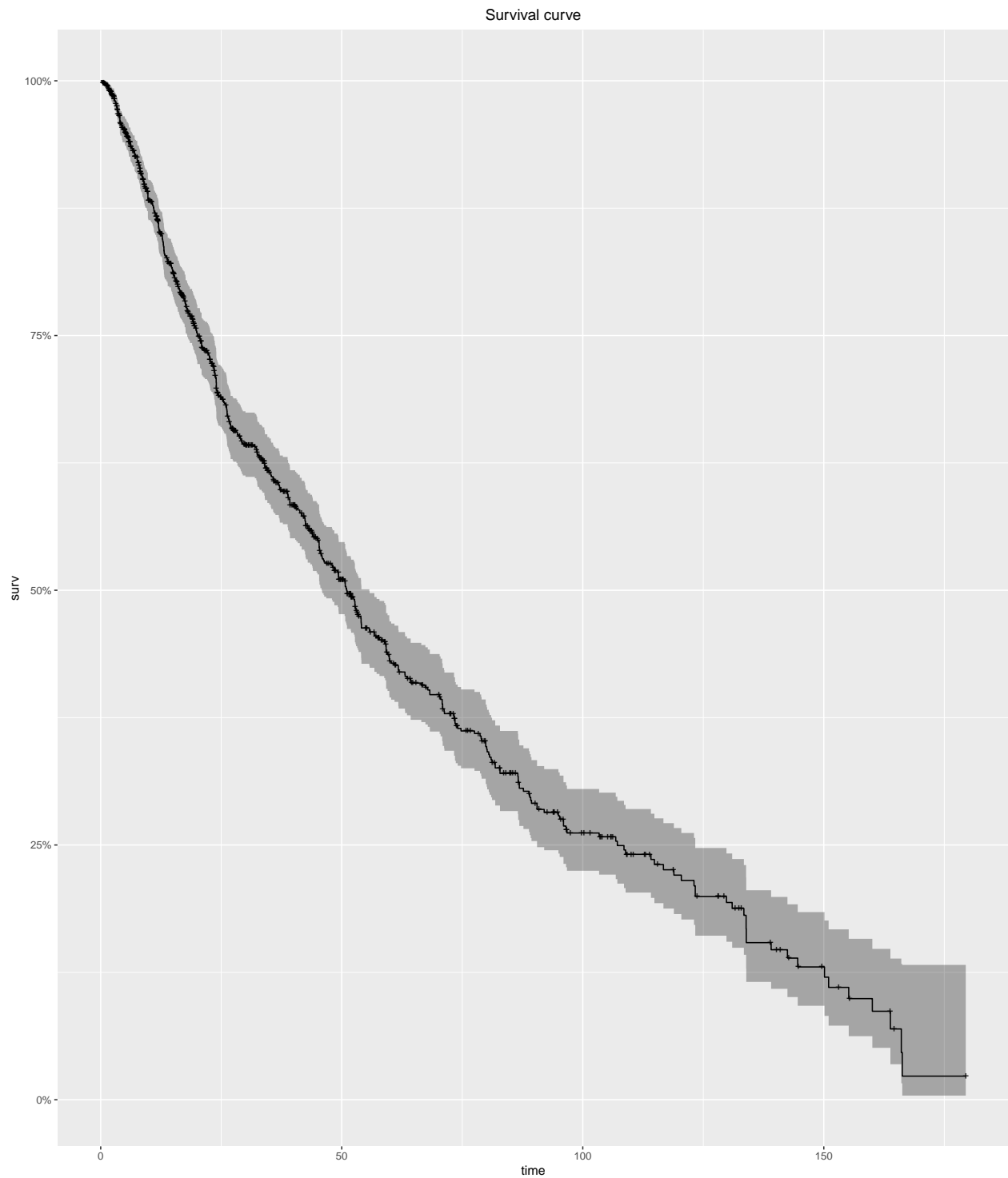
```
## Call: survfit(formula = objSurv ~ 1)
##
##          n events median 0.95LCL 0.95UCL
## [1,] 774      381   52.9    48.1    59.9
```

```
survplot1 <- autoplot(survfit(Surv(stag, event) ~ 1, data = newdat)) + labs(title = "Survival curve") +
survplot2 <- autoplot(survfit(Surv(stag, event) ~ gender, data = newdat)) + labs(title = "Survival curve") +
survplot3 <- autoplot(survfit(Surv(stag, event) ~ way, data = newdat)) + labs(title = "Survival curve") +
survplot4 <- autoplot(survfit(Surv(stag, event) ~ head_gender, data = newdat)) + labs(title = "Survival curve") +
survplot5 <- autoplot(survfit(Surv(stag, event) ~ greywage, data = newdat)) + labs(title = "Survival curve") +
survplot6 <- autoplot(survfit(Surv(stag, event) ~ coach, data = newdat)) + labs(title = "Survival curve") +

ggarrange(survplot1, survplot2, survplot3, survplot4, survplot5, survplot6)
```



```
survplot7 <- autoplot(survfit(Surv(stag, event) ~ 1 ,data = newdat)) + labs(title = "Survival curve") +
survplot7
```



Cox model

Fitting cox model that accounts for all covariates but assumes them to be stable over time:

```
cox <- coxph(Surv(stag, event) ~ 1 + gender + age + industry +  
            + greywage + way + extraversion + independ + selfcontrol +
```



```

anxiety + novator+ coach + head_gender, data = newdat)

summary(cox)

```

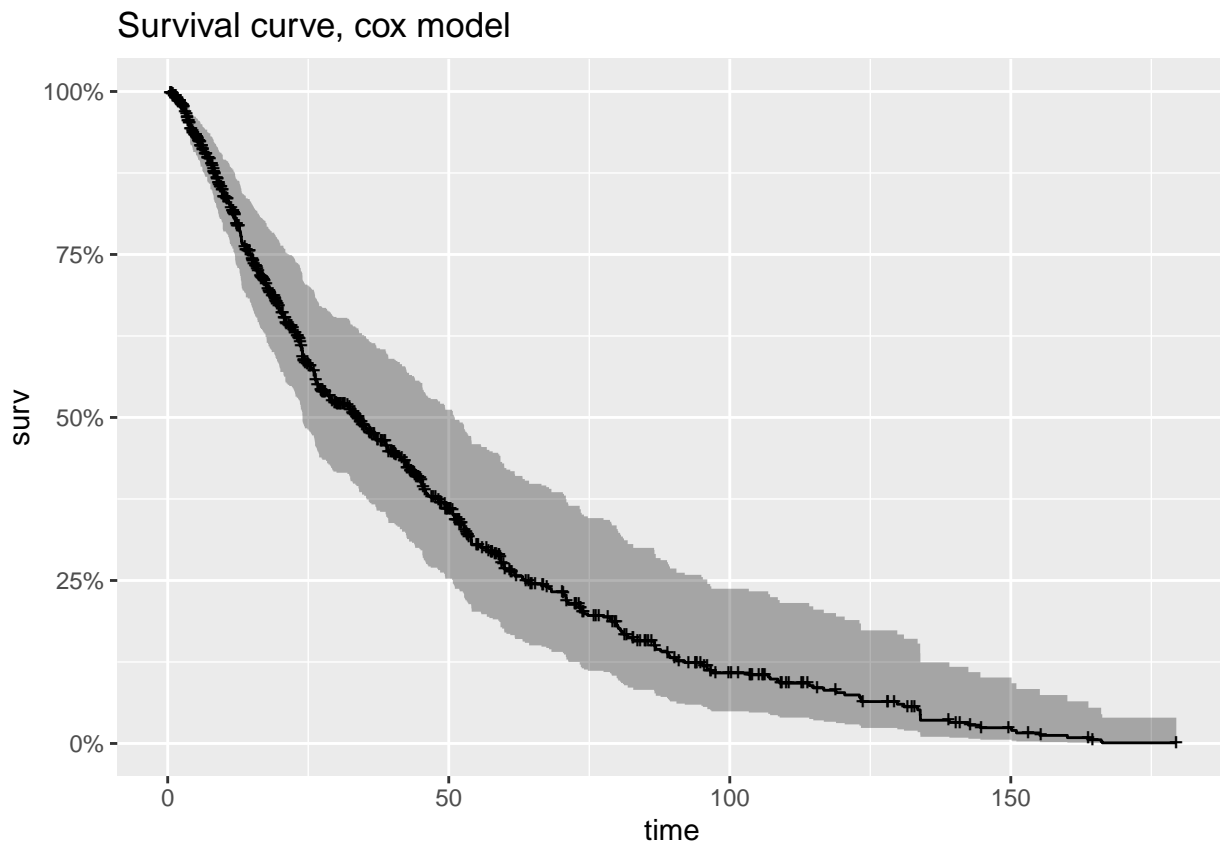
```

## Call:
## coxph(formula = Surv(stag, event) ~ 1 + gender + age + industry +
##       +greywage + way + extraversion + independ + selfcontrol +
##       anxiety + novator + coach + head_gender, data = newdat)
##
## n= 1107, number of events= 553
##
##               coef exp(coef) se(coef)      z Pr(>|z|)
## genderf          0.010213  1.010265  0.111203  0.092 0.926824
## age              0.023624  1.023905  0.006741  3.505 0.000457 ***
## industryPowerGeneration -0.591848  0.553304  0.306951 -1.928 0.053836 .
## industryRetail      -0.642597  0.525925  0.149685 -4.293 1.76e-05 ***
## industrymanufacture -0.526270  0.590804  0.173383 -3.035 0.002403 **
## industryConsult     -0.012943  0.987140  0.197223 -0.066 0.947675
## industryState       -0.220205  0.802355  0.215093 -1.024 0.305947
## industryetc         -0.177630  0.837252  0.185944 -0.955 0.339433
## industryBuilding     0.046920  1.048039  0.220251  0.213 0.831302
## industryIT          -0.884081  0.413094  0.211747 -4.175 2.98e-05 ***
## industryHoReCa      -0.341230  0.710895  0.431231 -0.791 0.428773
## industryTelecom     -0.925861  0.396190  0.295136 -3.137 0.001707 **
## industryPharma      -0.555370  0.573860  0.348911 -1.592 0.111446
## industryMining      -0.244900  0.782783  0.298681 -0.820 0.412250
## industrytransport   -0.405938  0.666351  0.279887 -1.450 0.146957
## industryAgriculture  0.527673  1.694984  0.343511  1.536 0.124510
## industryRealEstate  -1.154774  0.315129  0.475330 -2.429 0.015123 *
## greywagegrey        0.624656  1.867602  0.133713  4.672 2.99e-06 ***
## waycar             -0.211375  0.809470  0.100122 -2.111 0.034757 *
## wayfoot            -0.464456  0.628477  0.171858 -2.703 0.006881 **
## extraversion        0.030479  1.030948  0.034828  0.875 0.381510
## independ           -0.012817  0.987265  0.035085 -0.365 0.714882
## selfcontrol         -0.036709  0.963957  0.035258 -1.041 0.297812
## anxiety            -0.032466  0.968055  0.033551 -0.968 0.333219
## novator            0.009905  1.009954  0.030170  0.328 0.742675
## coachyes           0.167281  1.182087  0.138525  1.208 0.227206
## coachmy head       -0.001242  0.998759  0.107407 -0.012 0.990773
## head_genderm       0.017668  1.017825  0.094271  0.187 0.851334
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##               exp(coef) exp(-coef) lower .95 upper .95
## genderf          1.0103      0.9898   0.8124   1.2563
## age              1.0239      0.9767   1.0105   1.0375
## industryPowerGeneration 0.5533      1.8073   0.3032   1.0098
## industryRetail     0.5259      1.9014   0.3922   0.7052
## industrymanufacture 0.5908      1.6926   0.4206   0.8299
## industryConsult     0.9871      1.0130   0.6707   1.4530
## industryState       0.8024      1.2463   0.5264   1.2231
## industryetc         0.8373      1.1944   0.5815   1.2054
## industryBuilding    1.0480      0.9542   0.6806   1.6138

```

```
## industryIT                0.4131      2.4208      0.2728      0.6256
## industryHoReCa            0.7109      1.4067      0.3053      1.6553
## industryTelecom           0.3962      2.5240      0.2222      0.7065
## industryPharma            0.5739      1.7426      0.2896      1.1371
## industryMining            0.7828      1.2775      0.4359      1.4057
## industrytransport         0.6664      1.5007      0.3850      1.1533
## industryAgriculture       1.6950      0.5900      0.8645      3.3232
## industryRealEstate        0.3151      3.1733      0.1241      0.8000
## greywagegrey              1.8676      0.5354      1.4370      2.4272
## waycar                    0.8095      1.2354      0.6652      0.9850
## wayfoot                   0.6285      1.5911      0.4487      0.8802
## extraversion               1.0309      0.9700      0.9629      1.1038
## independ                   0.9873      1.0129      0.9217      1.0575
## selfcontrol                0.9640      1.0374      0.8996      1.0329
## anxiety                    0.9681      1.0330      0.9064      1.0339
## novator                    1.0100      0.9901      0.9520      1.0715
## coachyes                   1.1821      0.8460      0.9010      1.5508
## coachmy head               0.9988      1.0012      0.8092      1.2328
## head_genderm               1.0178      0.9825      0.8461      1.2244
##
## Concordance= 0.629  (se = 0.014 )
## Likelihood ratio test= 115.3 on 28 df,  p=1e-12
## Wald test               = 120.3 on 28 df,  p=2e-13
## Score (logrank) test = 124.9 on 28 df,  p=3e-14
```

```
autoplot(survfit(cox)) + labs(title = "Survival curve, cox model")
```

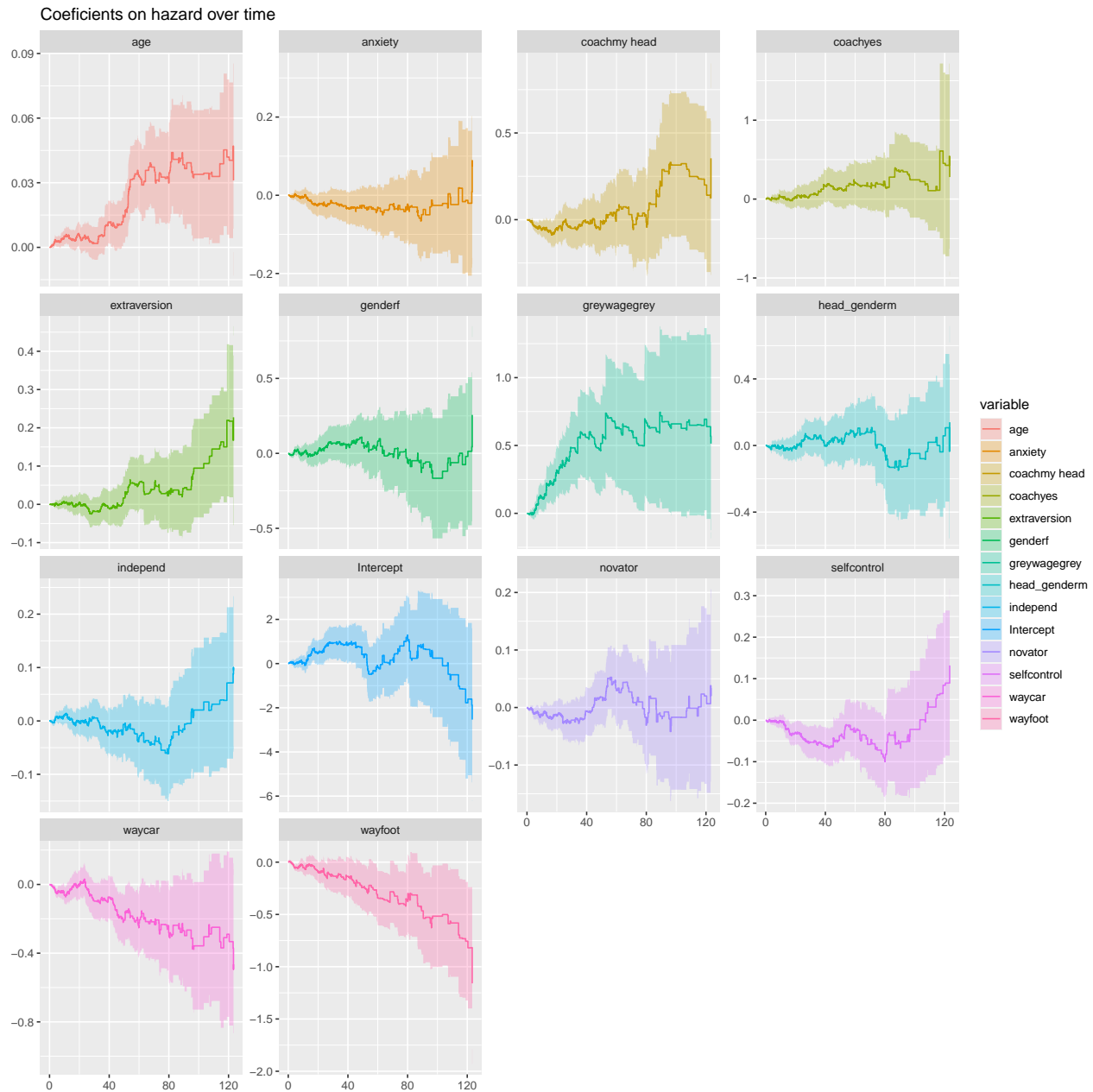


Aalen model

This model doesn't assume coefficients are stable

```
aa_fit = aareg(Surv(stag, event) ~ 1 + gender + age +
               greywage + way + extraversion + independ + selfcontrol +
               anxiety + novator + coach + head_gender, data = newdat)

autoplot(aa_fit) + labs(title = "Coefficients on hazard over time")
```



Some hazard effects clearly vary over time ($a(t) - XB(t)$), so the cox implementation should be taken with a grain of salt.

Grey wage is a predictor of quitting, and going to work by foot is a negative predictor of quitting.

Age is a predictor of quitting on longer tenures. Higher conscientiousness is associated to lower turnover in the first years.

Practical recommendation: hire in surroundings, offer help in accommodation.

People with longer tenure and age are at risk. Maybe senior promotion is not encouraged, or salary for more senior roles is not competitive, so people with more experience are more prone to leave when they acquire seniority.

Look for weak signals of conscientiousness when hiring (psychometric tests are likely to be tricked to favor conscientiousness).

References:

Cox DR. A note on the graphical analysis of survival data. *Biometrika*. 1979;66:188–190

Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc*. 1958;53:457–481