Group 13 PSTAT 175 Final Project

Grace Nunnelley, Ezra Aguimatang, Tetsuto Katsura

2022-11-30

1.0 Introduction

We chose to analyze the Employee Turnover dataset. This Employer Turnover dataset contains 1129 observations with 17 covariates about how long it took until the employer quit the job. We are using the column `stag` as our time variable, which represents failure and censor time in units months. In the `event` column, 1 equates to an employee quitting while 0 equates to a censor.

The covariates in our model include the categorical variables: `gender`, `industry`, `profession`, `traffic`, `coach`, `head_gender`, `greywage`, and `way`. `traffic` relates to the pipeline an employee came from, `coach` relates to, `head_gender` to the gender of the supervisor, `greywage` to money paid by the company, and `way` to the method of transportation (D. Wijaya, 2020). Our model also includes quantitative variables `age` (at the time of data entry, integers ranging from 18 to 58) as well as personality variables: `extraversion`, `independ`, `selfcontrol`, `anxiety`, and `novator`, all of which are float values ranging from 0 to 10. The personality variables are derived from the big personality model, so `selfcontrol` equates to conscientiousness, `anxiety` to neuroticism, `independ` to agreeableness, and `novator` to openness to experience (L.J.C.H. Hamers, 2021).

With our project, we aim to predict the probability of an employee turning over or leaving versus not leaving based on the events and variables in the dataset. The main motivating factor behind the creation of the dataset is to, "to predict individual risks of quitting of specific applicants." (E. Babushkin, 2017). A HR analyst named Babuskin created this particular dataset and used it to predict the chances of a particular employee's quitting (L.J.C.H. Hamers, 2021). In our case, instead of predicting the individual risk of quitting, we aim to analyze the influence of particular covariate variables on the risk of quitting.

After determining the most relevant variables for our general model, we use procedures to determine the influence of individual covariates on the survival rate. Our chosen covariates are `age`, `industry`, `profession`, `traffic`, `greywage`, `way`, `selfcontrol`, and `anxiety`. We examine each of these covariates to answer this question of how they individually affect the chances of quitting.

For our categorical covariate, we focused on stratification, creating loglog plots, and running a `coxzph`/Schoenfeld test to validate the PH Assumption. We grouped together factors of some of the categorical covariates because they contained too many level factors.

We did further analysis on personality traits to determine if traits above a certain score have a significant effect on our model.

1.1 Citations

Babushkin, E. 2017. Employee turnover: how to predict individual risks of quitting. Retrieved from: https://edwvb.blogspot.com/2017/10/employee-turnover-how-to-predict-individual-risks-of-quitting.html

Broström, Göran. *Event History Analysis with R.* Chapman & Hall/CRC, 2021.

Hamers, L.J.C.H. 2021. Can we predict employee turnover given individual differences and retain talent by coaching? A data-driven approach. Retrieved from: http://arno.uvt.nl/show.cgi?fid=157305

Kleinbaum, David G., and Mitchel Klein. *Survival Analysis, a Self-Learning Text.* New York: Springer, 2012.

Pham H. 2018. How to interpret hazard ratios of Cox output? Answer. Retrieved from: https://stats.stackexchange.com/questions/379469/how-to-interpret-hazard-ratios-of-cox-output

Wijaya, D. 2020. Employee Turnover, Employee Turnover dataset originally used for a Survival Analysis Model. Retrieved from: https://www.kaggle.com/datasets/davinwijaya/employee-turnover

1.2 Turnover Data

Reading in `turnover.csv`

```
turnover <- read_csv("turnover.csv")
## Rows: 1129 Columns: 16
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (8): gender, industry, profession, traffic, coach, head_gender, greywage...
## dbl (8): stag, event, age, extraversion, independ, selfcontrol, anxiety, nov...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
turnover <- turnover %>% mutate(stagSurv = Surv(stag, event),
                                gender = as.factor(ifelse(gender == "m","Male","Female")),
                                coach = as.factor(coach),
                                greywage = as.factor(greywage),
                                head_gender = as.factor(ifelse(head_gender == "m","Male","Female")),
                                industry = as.factor(industry),
                                profession = as.factor(profession),
                                traffic = as.factor(traffic))

roygbiv <- c("red","orange","green","cyan","blue","slateblue2","violet","black")
fourcolor <- function(total, front){
  return(c(rep("white", front), "black", "red", "blue", "green", rep("white", total-front-4)))}
```

2.0 Model Fitting (Forward AIC and BIC Models)

Since there are no more than one event occurring per observation, we are going to look at the normal Cox Proportional Hazards model.

```
CPH.all <- coxph(stagSurv~gender+age+industry+traffic+coach+head_gender+greywage+way+
                  extraversion+independ+selfcontrol+anxiety+novator,turnover)
CPH.all
## Call:
## coxph(formula = stagSurv ~ gender + age + industry + traffic +
##     coach + head_gender + greywage + way + extraversion + independ +
##     selfcontrol + anxiety + novator, data = turnover)
##
##                               coef exp(coef)  se(coef)      z        p
## genderMale                0.014013  1.014112  0.112151  0.125 0.900564
## age                       0.022467  1.022721  0.006716  3.345 0.000822
## industryBanks            -0.467103  0.626815  0.347400 -1.345 0.178765
## industryBuilding         -0.372100  0.689285  0.376330 -0.989 0.322780
## industryConsult          -0.508216  0.601568  0.365056 -1.392 0.163875
## industryetc              -0.614917  0.540686  0.356359 -1.726 0.084427
## industryHoReCa           -0.889154  0.411003  0.527987 -1.684 0.092173
## industryIT               -1.407844  0.244670  0.370011 -3.805 0.000142
## industrymanufacture      -0.942578  0.389622  0.349932 -2.694 0.007068
## industryMining           -0.647478  0.523364  0.426547 -1.518 0.129027
## industryPharma           -1.002758  0.366866  0.455981 -2.199 0.027869
## industryPowerGeneration -1.015771  0.362123  0.423790 -2.397 0.016536
## industryRealEstate       -1.622203  0.197463  0.567472 -2.859 0.004255
## industryRetail           -1.103755  0.331623  0.341024 -3.237 0.001210
## industryState            -0.610662  0.542991  0.372937 -1.637 0.101539
## industryTelecom          -1.331539  0.264071  0.428591 -3.107 0.001891
## industrytransport        -0.937192  0.391726  0.411740 -2.276 0.022836
## trafficempjs              0.736426  2.088457  0.303542  2.426 0.015262
## trafficfriends            0.045459  1.046508  0.330893  0.137 0.890729
## trafficKA                 0.115414  1.122338  0.339457  0.340 0.733859
## trafficrabrecNErab        0.410995  1.508317  0.301576  1.363 0.172939
## trafficrecNErab          -0.035809  0.964824  0.369216 -0.097 0.922736
## trafficreferal            0.280602  1.323926  0.315494  0.889 0.373787
## trafficyoujs              0.516134  1.675538  0.300807  1.716 0.086192
## coachno                   0.079269  1.082495  0.106713  0.743 0.457587
## coachyes                  0.241669  1.273373  0.148363  1.629 0.103332
## head_genderMale           0.091763  1.096105  0.094685  0.969 0.332476
## greywagewhite            -0.538006  0.583912  0.131442 -4.093 4.26e-05
## waycar                   -0.179441  0.835737  0.099589 -1.802 0.071573
## wayfoot                  -0.355279  0.700978  0.169774 -2.093 0.036379
## extraversion              0.023282  1.023555  0.034181  0.681 0.495783
## independ                 -0.003838  0.996169  0.034654 -0.111 0.911808
## selfcontrol              -0.050298  0.950946  0.034399 -1.462 0.143684
## anxiety                  -0.048693  0.952474  0.033483 -1.454 0.145871
## novator                   0.003906  1.003914  0.029808  0.131 0.895743
##
## Likelihood ratio test=140.7  on 35 df, p=1.272e-14
## n= 1129, number of events= 571
```

Looking at the p-values from the Cox summary output, we remove the least relevant covariates according

to their large p-values. These irrelevant covariates include `gender`, `coach`, `head_gender`, `extraversion`, `independ`, and `novator`.

In our output of running our cox model through the forward stepwise AIC function, we can determine that `age`, `industry`, `profession`, `traffic`, `greywage`, `way`, `selfcontrol`, and `anxiety` create significant differences in the survival rates.

Therefore, we determine that the model of best fit is:

```
CPH.AICfit <- coxph(formula = stagSurv ~ age+industry+profession+traffic+
    greywage+way+selfcontrol+anxiety, data = turnover)
```

In addition to finding our model of best fit, we used another coxmodel and forward stepwise AIC function to ensure that interaction within our model is not relevant. Specifically, we created a cox model, shown below, to test interaction between our most relevant continuous variables, `selfcontrol` and `anxiety`.

```
stepAIC(coxph(formula = stagSurv ~ selfcontrol*anxiety, data = turnover))
## Start:  AIC=6935.4
## stagSurv ~ selfcontrol * anxiety
##
##                       Df    AIC
## - selfcontrol:anxiety  1 6933.5
## <none>                   6935.4
##
## Step:  AIC=6933.49
## stagSurv ~ selfcontrol + anxiety
##
##               Df    AIC
## <none>           6933.5
## - anxiety      1 6935.8
## - selfcontrol  1 6938.9
## Call:
## coxph(formula = stagSurv ~ selfcontrol + anxiety, data = turnover)
##
##                 coef exp(coef) se(coef)      z       p
## selfcontrol -0.05734   0.94427  0.02105 -2.723 0.00646
## anxiety     -0.05164   0.94967  0.02477 -2.084 0.03714
##
## Likelihood ratio test=11.11  on 2 df, p=0.003866
## n= 1129, number of events= 571
```

The resulting p-value, which is greater than 0.05, shown in the output of the stepAIC function indicates that interaction is not statistically significant. Thus the interaction between selfcontrol and anxiety is not significant enough to include in our model.

3.0 Check Proportional Hazards Assumptions

We looked through log-log plots for qualitative covariates that have few levels, and run through `coxzph` test for our non-continuous covariates.

For `industry` and `profession`, we will reference the `coxph` function to group together levels into 3 groups. The groups are chosen according to their absolute z value as follows: less than 1.282, 1.282 to 2.576, and beyond 2.576. The z scores were chosen based on 80 and 99 intervals.

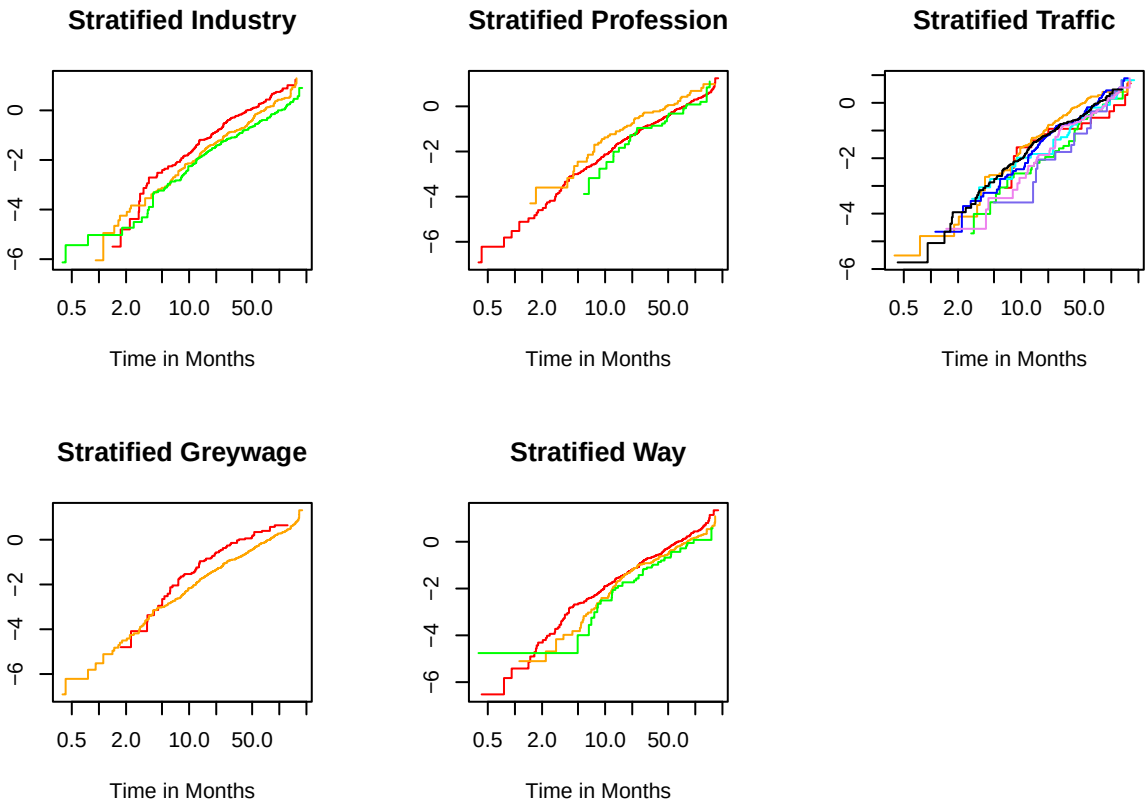
The groups are organized as such. Industry Group 1: `Agriculture`, `Banks`, `Building`, and `Consult` together. Industry Group 2: `Miscellaneous`, `HoReCa`, `Manufacturing`, `Mining`, `Pharmaceuticals`, `State`, `Transportation`. Industry Group 3: `leftovers`. Profession Group 1: `Business Development`, `Consulting`, `Miscellaneous`, `Finance`, `HR`, `IT`, `Law`, `Sales`, `Teaching/Education`. Profession Group 2: `Commercial`, `Engineering`, `Marketing`, `PR`. Profession Group 3: `leftovers`.

```
indG1 <- c("Agriculture","Banks","Building","Consult")
indG2 <- c("etc","HoReCa","manufacture","Mining","Pharma","PowerGeneration","State","transport")
profG1 <- c("BusinessDevelopment","Consult","etc","Finan\xf1e","HR","IT","Law","Sales","Teaching")
profG2 <- c("Commercial","Engineer","Marketing","PR")
turnover <- turnover %>% mutate(ind=ifelse(industry %in% indG1,"G1",
                                           ifelse(industry %in% indG2,"G2","G3")),
                                prof=ifelse(profession %in% profG1,"G1",
                                           ifelse(profession %in% profG2,"G2","G3")))
CPH.Groupfit <- coxph(formula = stagSurv ~ age+ind+prof+traffic+
    greywage+way+selfcontrol+anxiety, data = turnover)
```

We looked through log-log plots for quantitative covariates, and run a `coxzph` test for non-continuous covariates.

3.1 Log(-Log) Plots

```
par(mfrow=c(2,3))

plot(survfit(stagSurv ~ ind, turnover), fun="cloglog", xlab="Time in Months",
     main="Stratified Industry", col = roygbiv)
plot(survfit(stagSurv ~ prof, turnover), fun="cloglog", xlab="Time in Months",
     main="Stratified Profession", col = roygbiv)
plot(survfit(stagSurv ~ traffic, turnover), fun="cloglog", xlab="Time in Months",
     main="Stratified Traffic", col = roygbiv)
plot(survfit(stagSurv ~ greywage, turnover), fun = "cloglog", xlab="Time in Months",
     main="Stratified Greywage", col = roygbiv)
plot(survfit(stagSurv ~ way, turnover), fun = "cloglog", xlab="Time in Months",
     main="Stratified Way", col = roygbiv)

CZPH.fit <- cox.zph(CPH.Groupfit)
CZPH.fit
##                chisq df    p
## age         1.00e+00  1 0.32
## ind         4.33e+00  2 0.11
## prof        6.74e-01  2 0.71
## traffic     9.83e+00  7 0.20
## greywage    3.89e-01  1 0.53
## way         2.69e+00  2 0.26
## selfcontrol 9.88e-06  1 1.00
## anxiety     1.16e+00  1 0.28
## GLOBAL      2.14e+01 17 0.21
```



All the graphs gave mostly parallel lines and cox zph test gives a global p-value of greater than 0.05.

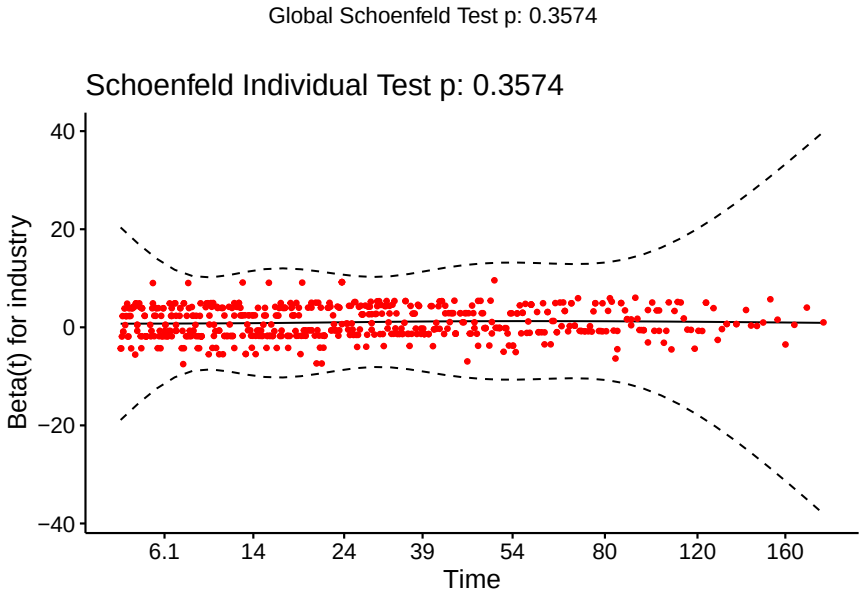
3.2 Goodness-of-Fit Approach

With this approach we utilize a test statistic and p-value in order to more objectively assess the validity of the PH assumption for any given predictor. This is a more concrete and accurate method to test this assumption in comparison to the graphical approach which is based on the tester's subjective notion determining "how parallel is parallel?"(Kleinbaum and Klein, 172). For this study, we will supplement our graphical testing approach of the PH assumption using the `cox.zph()` function to give us the Scaled Residual p-values of the Global and Individual Shoenfeld Tests, which we will also represent with a plot using the `ggcoxzph()` function.

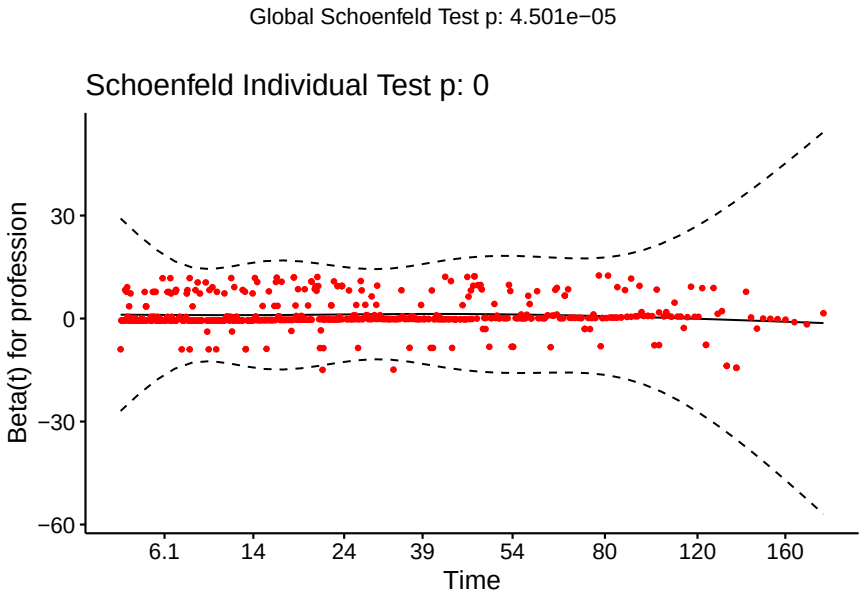
```
# Cox PH Model: Industry
coxph.industry <- coxph(stagSurv~industry, data=turnover)
anova(coxph.industry)
## Analysis of Deviance Table
##  Cox model: response is stagSurv
## Terms added sequentially (first to last)
##
##           loglik  Chisq Df Pr(>|Chi|)
## NULL     -3470.3
## industry -3441.5 57.667 15  6.322e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
# Cox PH Model: Profession
coxph.profession <- coxph(stagSurv~profession, data=turnover)
anova(coxph.profession)
## Analysis of Deviance Table
##  Cox model: response is stagSurv
## Terms added sequentially (first to last)
##
##             loglik  Chisq Df Pr(>|Chi|)
## NULL       -3470.3
## profession -3456.2 28.279 14    0.01306 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
# Cox PH Model: Traffic
coxph.traffic <- coxph(stagSurv~traffic, data=turnover)
anova(coxph.traffic)
## Analysis of Deviance Table
##  Cox model: response is stagSurv
## Terms added sequentially (first to last)
##
##          loglik  Chisq Df Pr(>|Chi|)
## NULL    -3470.3
## traffic -3458.7 23.178  7   0.001587 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
# Cox PH Model: Grey Wage
coxph.greywage <- coxph(stagSurv~greywage, data=turnover)
anova(coxph.greywage)
## Analysis of Deviance Table
##  Cox model: response is stagSurv
## Terms added sequentially (first to last)
##
##           loglik  Chisq Df Pr(>|Chi|)
```

```
## NULL     -3470.3
## greywage -3460.9 18.821  1  1.436e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
# Cox PH Model: Way
coxph.way <- coxph(stagSurv~way, data=turnover)
anova(coxph.way)
## Analysis of Deviance Table
##  Cox model: response is stagSurv
## Terms added sequentially (first to last)
##
##       loglik  Chisq Df Pr(>|Chi|)
## NULL -3470.3
## way  -3464.3 12.063  2   0.002402 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

testzph.industry <- cox.zph(coxph.industry)
testzph.profession <- cox.zph(coxph.profession)
testzph.traffic <- cox.zph(coxph.traffic)
testzph.greywage <- cox.zph(coxph.greywage)
testzph.way <- cox.zph(coxph.way)

ggcoxzph.industry <- ggcoxzph(testzph.industry)
ggcoxzph.profession <- ggcoxzph(testzph.profession)
ggcoxzph.traffic <- ggcoxzph(testzph.traffic)
ggcoxzph.greywage <- ggcoxzph(testzph.greywage)
ggcoxzph.way <- ggcoxzph(testzph.way)
```

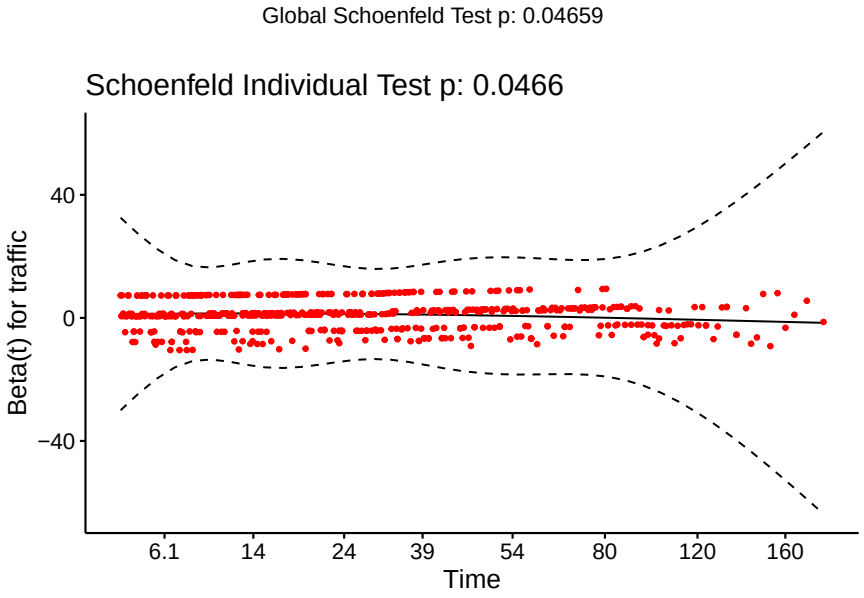
```
ggcoxzph.industry
```



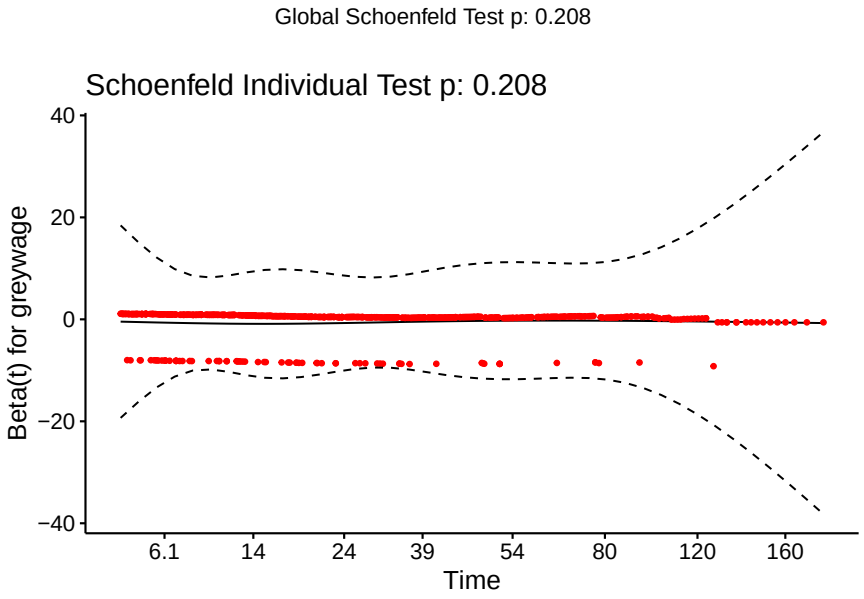

```
ggcoxzph.profession
```



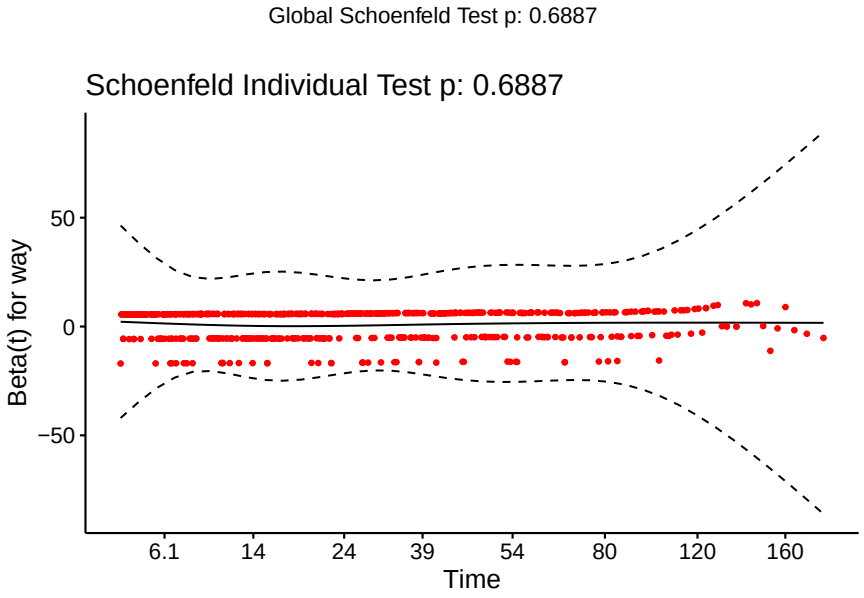
```
ggcoxzph.traffic
```




```
ggcoxzph.greywage
```



```
ggcoxzph.way
```



As can be seen above from our plots of the Scaled Schoenfeld Residuals and resulting p-values of `industry`, `profession`, `traffic`, `greywage`, and `way`, we were able to both corroborate and refute the results of the previous graphical log-log plot approach to testing the Proportional Hazards assumption. With significantly large p-values, we can conclude that `industry` ($p = 0.3574$), `greywage` ($p = 0.208$), and `way` ($p = 0.6887$) all pass the PH assumption individually. However, `profession` ($p = 0$) and `traffic` ($p = 0.0466$) do not present statistically significant p-values and therefore violate the Proportional Hazard assumption.

4.0 Conclusions

```
coxph(formula = stagSurv ~ age + ind + prof + traffic + greywage + way +
        selfcontrol + anxiety, data = turnover)
## Call:
## coxph(formula = stagSurv ~ age + ind + prof + traffic + greywage +
##     way + selfcontrol + anxiety, data = turnover)
##
##                          coef exp(coef)  se(coef)      z        p
## age                  0.021550  1.021784  0.006086  3.541 0.000399
## indG2               -0.427474  0.652154  0.105716 -4.044 5.26e-05
## indG3               -0.779518  0.458627  0.110515 -7.054 1.74e-12
## profG2               0.587924  1.800247  0.150102  3.917 8.97e-05
## profG3               0.188906  1.207928  0.190572  0.991 0.321558
## trafficempjs         0.735896  2.087352  0.298268  2.467 0.013616
## trafficfriends       0.004540  1.004550  0.323877  0.014 0.988816
## trafficKA            0.128817  1.137482  0.331902  0.388 0.697930
## trafficrabrecNErab   0.367932  1.444744  0.297250  1.238 0.215794
## trafficrecNErab     -0.073933  0.928734  0.363228 -0.204 0.838709
## trafficreferal       0.252175  1.286821  0.307796  0.819 0.412620
## trafficyoujs         0.514234  1.672356  0.295901  1.738 0.082236
## greywagewhite       -0.518531  0.595395  0.128531 -4.034 5.48e-05
## waycar              -0.204529  0.815031  0.096335 -2.123 0.033744
## wayfoot             -0.353510  0.702219  0.164702 -2.146 0.031844
## selfcontrol         -0.066259  0.935888  0.021841 -3.034 0.002416
## anxiety             -0.057659  0.943971  0.025033 -2.303 0.021262
##
## Likelihood ratio test=137.1  on 17 df, p=< 2.2e-16
## n= 1129, number of events= 571
# remove exp(coef)
# later will reverse this change and include exp(coef) in stead of exp(-coef)
summary(CPH.Groupfit)$conf.int[,-1]
##                     exp(-coef) lower .95 upper .95
## age                  0.9786805 1.0096681 1.0340451
## indG2                1.5333794 0.5301088 0.8022981
## indG3                2.1804211 0.3693084 0.5695476
## profG2               0.5554792 1.3414215 2.4160123
## profG3               0.8278641 0.8314309 1.7549135
## trafficempjs         0.4790760 1.1633485 3.7452552
## trafficfriends       0.9954703 0.5324597 1.8952068
## trafficKA            0.8791351 0.5935113 2.1800163
## trafficrabrecNErab   0.6921641 0.8068106 2.5870826
## trafficrecNErab      1.0767349 0.4557333 1.8926559
## trafficreferal       0.7771088 0.7039172 2.3524194
## trafficyoujs         0.5979587 0.9363921 2.9867571
## greywagewhite        1.6795587 0.4628062 0.7659678
## waycar               1.2269473 0.6747983 0.9844058
## wayfoot              1.4240566 0.5084853 0.9697663
## selfcontrol          1.0685036 0.8966706 0.9768212
## anxiety              1.0593541 0.8987737 0.9914421
```

```
# adjusted GroupFit (looking at the p-values)
CPH.Groupfit <- coxph(formula = stagSurv ~ gender+age+ind+prof+traffic+
    greywage+way+selfcontrol+anxiety, data = turnover)
```

Looking at the summary output of the cox output on our ideal model, age, group 2 industries, group 3 industries, group 2 professions, and white grey wage have the greatest statistical significance compared to other covariates according to their p-values.

Looking at the table, we can see that the hazard ratios (taken from `exp(coef)`) are `trafficempjs`: 2.0873516, `age`: 1.0217839, `profG2`: 1.8002475, `indG2`: 0.6521543, `indG3`: 0.4586270, `greywagewhite`: 0.5953945

Hazard represents the risk of quitting. `trafficempjs` increases the hazard rate by 108%. `profG2` increases 80%. `age` increases 2%. `indG3` (1-0.46=0.54) decreases hazard rate by 54%. `indG2` decreases `grewwagewhite` decreases 40% (Pham H, 2018).

Using these findings, we can conclude that the traffic factor, `empjs` (whatever it stands for) significantly increases the chances of quitting as well as belonging to profession group 2 (Commercial, Engineering, Marketing, PR).

5.0 Advanced Methods

Now we are going to make adjustments to personality scores to determine if any of the personality scores affect the time. We chose to examine personality scores by altering their range, adding it to our model of best fit, and using Step AIC, coxph, and coxzph processes on our new model to see personality scores within a specific range make a difference.

We changed the range of personality variables by setting values that were less than or equal to a 'threshold' number to zero, and subtracting that threshold number from the values that are greater than that threshold number. We used numbers 3, 5, and 7 as threshold values to alter the range of our personality scores.

Using this process, we determined that scoring greater than 5 on the extraversion scale and anxiety scale, as well as scoring greater than 7 on the self control scale, has a significant effect. Thus, we can argue that being more extraverted, anxious, and self-controlled is relevant to our model.

We can determine the extent of their impact on the chances of quitting by creating hazard ratios out of the old and new personality scores (see below).