Done by: Freda Xiao (N2202500H), Grace Ong (U1921756L), Wang Anyi (U1922992C)

The aim of our study is to analyze how the different variables collected from the released inmates affect the time until their recidivism. The original aim of the study was to determine the effect of only financial aid on recidivism rates. We hope to go beyond the original study to find if there are certain factors that significantly contributed to the recidivism outcome, or if there are other interactions, covariates that significantly affect the recidivism rate. To do this, we will perform non-parametric and parametric hypothesis tests, as well as regression analyses to see if there are certain models or parameters that appropriately represent the dataset. Recidivism is represented in the dataset by “week”, representing the number of weeks until their arrest and “arrest” which is recorded as a censoring variable with 0 (not arrested) and 1 (arrested). In this report, we analyze the data by using “week” as “survival time” and “arrest” as “censor”.

### 

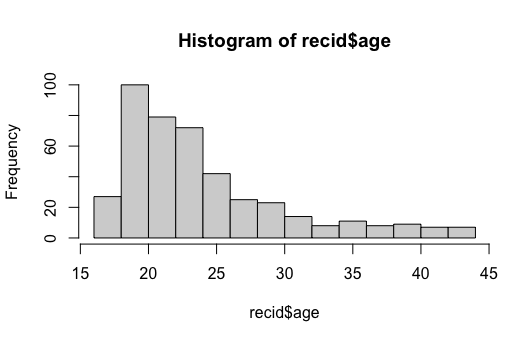
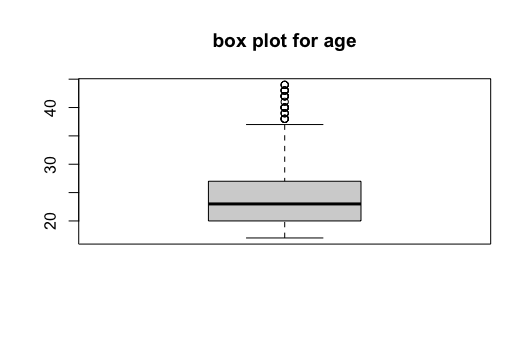
### **Description of dataset**

Initial descriptive summaries of the dataset were performed using R package "psych" as seen in Figure 1.1. Out of 11 variables:

* week, age, prio are **continuous** variables
* arrest, fin, race, wexp, mar, paro are **categorical** variables which are binary
* educ is a **categorical** variable with 5 levels (2-6)
* emp1- emp25 are categorical. Some data from emp1 - emp25 is **right-censored** since there is no information about employment status after the person’s first arrest.

Note: since the information missing is after arrest, we can assume that most of the employment status = 0. However, the duration of the arrest is unknown, so we cannot be sure of how many weeks the status = 0 continues until.

First of all, we can observe the frequencies of the different **age** groups being studied.



From the summary and box plot, the mean age is 24.6 and median age is 23.0. In the histogram, the frequency just below age 20 is significantly higher than the rest. However, most of the frequency is around 20 - 25, which we concluded from the summary above. Since we want to study recidivism, we now observe the age of the inmates who are arrested during the study (arrest=1) and compare them with those who are not arrested (arrest=0).

Compare the box plot of arrest = 1 and arrest = 0 from the dataset.

|  |  |
| --- | --- |

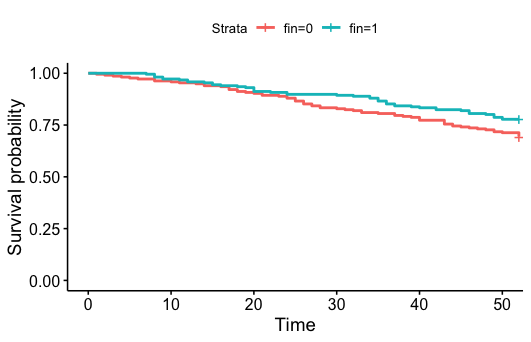
[Arrest=1]: mean = 22.76; median=21.00 [Arrest=0]: mean = 25.25; median=23.00

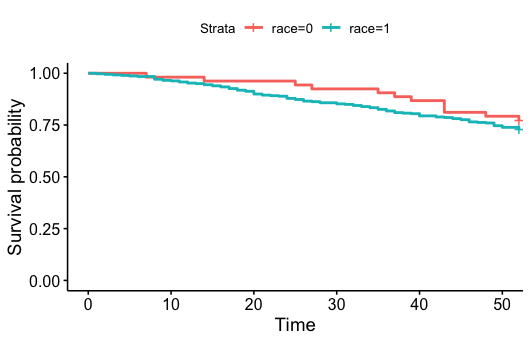
Compared with arrest=0, the mean age of arrested inmates during the study is lower. We can conclude that more people aged around 22 are rearrested after their release. Since we are only interested in rearrested inmates, we extract [arrest=1] from the week and age of the data set. This eliminates the observed persons who are not arrested which makes the data closer to our subject of interest. (1.6)

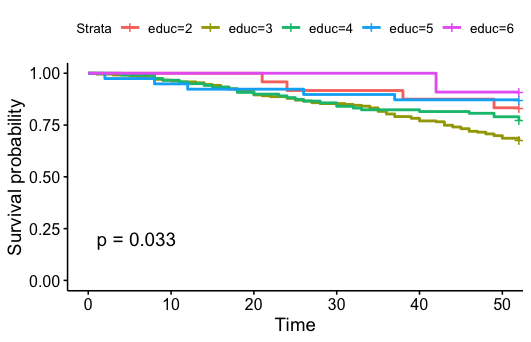
### Hypothesis testing

#### 2.1 KM estimate

We want to study the effect of financial aid, race and education level on recidivism, so we can focus the data with [arrest=1].







From the KM estimates of groups with and without financial aid, groups without financial aid (fin=0) have a slightly higher rate of recidivism.

From the KM estimates of groups depending on race, inmates who are black (race=1) have a slightly higher recidivism rate.

From the KM curves of the 5 groups, generally the groups with higher education levels have higher survival rate compared with the groups with lower education level, though interestingly the survival rate for educ=2 is quite high. However, it is important to note that there are only 11 inmates in educ=6 and 24 in educ=2 (small proportion compared with the population), therefore the survival rate may not be as significant as the other three groups.

To test if there are actual statistical differences amongst the groups, let us perform non-parametric hypothesis tests.

#### 2.2 Non parametric hypothesis testing

**One sided log rank test (fin)**

The goal is to use a one-sided log rank test to check if there is significant difference in the survival rate of inmates who received financial aid vs those who did not.

H0: survival time of fin=0 group == survival time of fin=1 group

H1: survival time of fin=0 group < survival time of fin=1 group

Log rank will be based on group fin=1. According to Figure 2.5, Observed - Expected of group 1 = 48-58.4 = -10.4 (< 0). Since c = -1.959 < -1.64, we reject the null hypothesis that survival / arrest time of two groups are equal and conclude that survival time of group 0 < survival time of group 1. This result is consistent with the KM estimate chart.

**One sided log rank test (race)**

The goal is to use a one-sided log rank test to check if there is significant difference in the survival rate of inmates who were black vs those who were not.

H0: survival time of race=0 group == survival time of race=1 group (black)

H1: survival time of race=0 group > survival time of race=1 group

According to Figure 2.6, since c = -0.75 > -1.64, we do not reject the null hypothesis and conclude that the survival time of the races were equal. In other words, race was not statistically significant in impacting survival time. This result is consistent with the KM curve for race in the earlier section, where race=0 showed generally higher survival probabilities.

**One sided log rank test (paro)**

The goal is to use a one-sided log rank test to check if there is significant difference in the survival rate of inmates who were released on parole vs those who were not.

H0: survival time of paro=0 group == survival time of paro=1 group (parole)

H1: survival time of paro=0 group < survival time of paro=1 group

According to Figure 2.7, since c = -0.57 > -1.64, we do not reject the null hypothesis and conclude that the survival time of the 2 groups were equal.

**One sided log rank test (wexp)**

The goal is to use a one-sided log rank test to check if there is significant difference in the survival rate of inmates who had full time work experience vs those who did not.

H0: survival time of wexp=0 group == survival time of wexp=1 group

H1: survival time of wexp=0 group < survival time of wexp=1 group

Log rank test will be based on group wexp=1. According to Figure 2.8, Observed - Expected = 52-68.4 = -16.4 (<0) Since c = -3.14 < -1.64, we reject the null hypothesis and conclude that the survival time of inmates who had full time work experience had longer survival time.

**Two sided log rank test (educ)**

Performing a log rank on education, with 4 df and 95% significance level.

H0: S2(t)=S3(t)=S4(t)=S5(t)=S6(t)

H1: at least one of the survival rates among the 5 levels of education is different

According to Figure 2.9, since p=0.0327969 < 0.05, we can reject the null hypothesis and conclude that the survival rates between different levels of education are significantly different.

**Two sided log rank test (prio)**

H0: survival time of inmates with different prio (number of prior convictions) were the same

H1: survival time of inmates with different prio (number of prior convictions) were different for at least one pair of category

According to Figure 2.10, since p=0.0004 < 0.05, we reject the null hypothesis and conclude that the difference in survival times of inmates with different numbers of prior convictions were statistically significant.

#### 2.3 Parametric hypothesis testing

The table is a summary of parametric model fitting under 4 different models. Detailed output is shown in Figures 2.11 to 2.14.

|  | log likelihood | AIC | BIC | df |
| --- | --- | --- | --- | --- |
| Exponential | -701.98 | 1405.95 | 1410.02 | 1 |
| Weibull | -696.62 | 1397.25 | 1405.38 | 2 |
| Log normal | -697.91 | 1399.82 | 1407.96 | 2 |
| Log logistic | -696.67 | 1397.35 | 1405.49 | 2 |

Based on AIC/BIC values, Weibull model provides the best fit. From the output of the flexsurvreg on Weibull model (Figure 2.12) , the estimated value of gamma = shape = 1.365 , lambda = 1/scale = 1/123.68 = 0.008. A plot of the fitted Weibull model with the survival step function (Figure 2.15) also shows that the Weibull model (red line) indeed provides a decent fitting.

## Regression Model

#### 3.4 AFT

The table is a summary of AFT under 4 different models. Detailed output is shown in Figures 3.1 to 3.4.

|  | Max log likelihood | XL | P-value | AIC | BIC | Num parameters |
| --- | --- | --- | --- | --- | --- | --- |
| Exponential | -685.3 | 12.86 | 0.00034 | -703.31 | -712.62 | 9 |
| Weibull | -678.88 | - | - | -698.88 | -709.23 | 10 |
| Log normal | -682.91 | - | - | -702.90 | -713.24 | 10 |
| Log logistic | -679.02 | - | - | -699.02 | -709.37 | 10 |

Based on AIC and BIC, Weibull model again gave the best fit.

Based on Weibull model (Figure 3.2):

* Fin: (1=received financial aid, 0=did not receive) mean survival /arrest time (week of first arrest) increases by exp(0.2581)-1 =29.4 % with financial aid provided.
* Age: as age increases by 1 year, the mean survival /arrest time increases by approximately 4.09%.
* Race: (1=black, 0=otherwise) mean survival /arrest time of those who were black is only exp(-0.2473) = 78.0% times of inmates who were not black, although p value of 0.27 indicates that the factor of race is not statistically significant.
* Wexp: (1= had full time work exp) mean survival /arrest time is exp(0.0817)-1 =1.085% higher for inmates who had full time work experience. The p value of 0.59 indicates that work experience is not statistically significant.
* Mar: (1= married) mean survival /arrest time is exp(0.3051)-1= 35.7% higher for inmates who were married at time of release compared to those who were not. The p value of 0.26 indicates that this factor is not statistically significant.
* Paro: (1=parole) mean survival /arrest time is exp(0.0625)-1 = 6.4% higher for inmates who were released on parole compared to those who were not, although the p value of 0.65 indicates that this factor is not statistically significant.
* Educ: mean survival /arrest time increases by exp(0.1318)-1 = 14.1% for every one level increase in schooling level, although the p value of 0.16 indicates that this factor is not statistically significant.
* Prio: mean survival time decreases by exp(-0.0610)-1 = 5.9% for every increase in one time of prior conviction. In other words, the more times an inmate was convicted before, the shorter his time of arrest (higher recidivism rate).

Note: all analysis of covariates of interest above involves holding other covariates constant.

#### 3.5 Semiparametric

##### 3.5.1 Univariate Cox Regression

The output shows the regression beta coefficients, hazard ratios and statistical significance for each variable. After applying univariate Cox analysis on each of the variables and extracting the wald-test p-values testing whether their coefficients are equal to 0, we can see that the p-values for fin, mar, race and paro are more than 0.05, hence conclude that their coefficients are not highly statistically significant. On the other hand, under univariate cox regression analysis, it seems age, wexp, prio and educ have statistically significant coefficients as their p-values are less than 0.05.

Additionally, prio has a positive beta coefficient, while age, wexp and educ have negative beta coefficients. Thus, higher number of prior convictions is associated with higher chance of recidivism, while older age, prior work experience and higher level of education is associated with lower chance of recidivism.

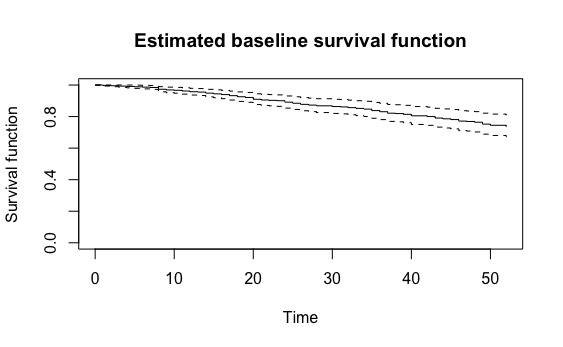
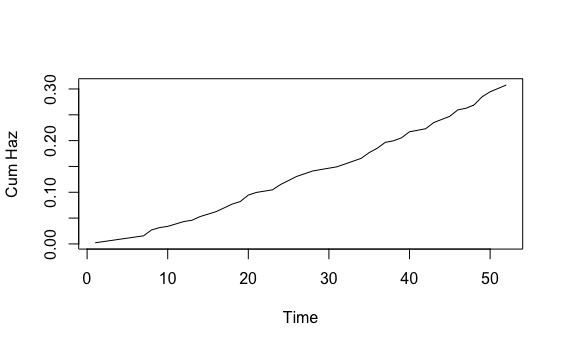
##### 3.5.2 Multivariate Cox Regression

Let us perform a multivariate Cox regression analysis as shown in Figure 3.5.2, with the statistically significant factors age, wexp, prio and educ, to describe how the factors jointly impact recidivism. The Likelihood ratio, Wald, and Logrank tests evaluate the null hypothesis that all of the coefficients are 0. From the summary, the p-values for all of them are less than 0.05, hence we reject the null hypotheses at 5% significance level and conclude that the model is significant. Looking at the individual p-values for each coefficient, the covariates age and prio remain significant, while wexp and educ are not as significant to the multivariate model.

The p-value for age is 0.00461, with a hazard ratio HR = exp(coef) = 0.94, indicating a strong relationship between age and decreased risk of recidivism. The hazard ratios of covariates are interpretable as multiplicative effects on the hazard. For example, holding the other covariates constant, having an increase in age reduces the hazard by a factor of 0.94, or 6%. We conclude that higher age is associated with lower chance of recidivism. Similarly, the p-value for prio is 0.00530, with a hazard ratio HR = 1.08, indicating a strong relationship between number of prior convictions and recidivism. Holding the other covariates constant, a higher number of prior convictions is associated with higher chance of recidivism.

However, holding other covariates constant, work experience and education do not have as strong a relationship with recidivism.

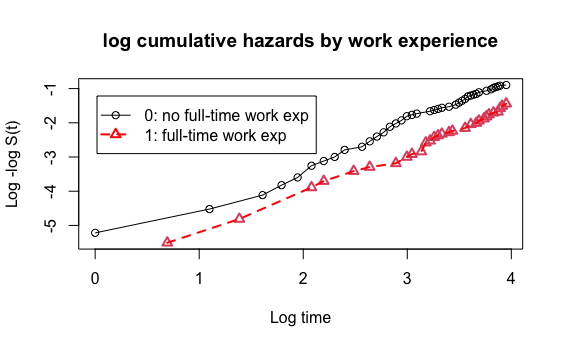
Plotting the estimated cumulative hazards on the left, the function seems to be rather linear, suggesting hazard is quite constant over time. On the right, we plot the survival curve implied by the multivariate Cox regression model.



##### 3.5.3 Interactions

Among the 4 statistically significant variables age, wexp, prio and educ, let us test if there are any interactions between them. After adding interaction terms to paired models, it seems only prio:educ has a low p-value of 0.0524, providing some evidence that interaction exists between prior number of convictions and education. For someone with no prior convictions (prio=0), increasing the level of education by 1 leads to a 35.5% decrease in hazard.

##### 3.5.4 Assessing validity of the PH assumption



Since wexp is a binary data field, we can stratify the dataset and assess the validity of the proportional hazards assumption. Plotting the log-log survival curves, it seems that the proportional hazards assumption is not violated as the log-log survival curves of people with (wexp=0) and (wexp=1) seem to be quite parallel.

### 

### Conclusion

Summarizing the results from the log rank, AFT and Semi parametric models, both the log rank test and the univariate cox regression model agree that 4 covariates, age, wexp, educ, and prio are statistically significant to the rate of recidivism. However as we fit multivariate regression models (AFT and multivariate cox), it is agreeable that only age and prio are statistically significant, with increased age having a positive effect on lowering recidivism rate and higher number of prior convictions having a negative effect on lowering recidivism rate. Both models also agreed, based on the p values, that age was the most significant variable affecting time of arrest, followed by prio.

As such, we would recommend from the results of our findings that more resources should perhaps be allocated to support younger inmates, as well as inmates who have had multiple prior convictions so as to reduce overall recidivism rates.

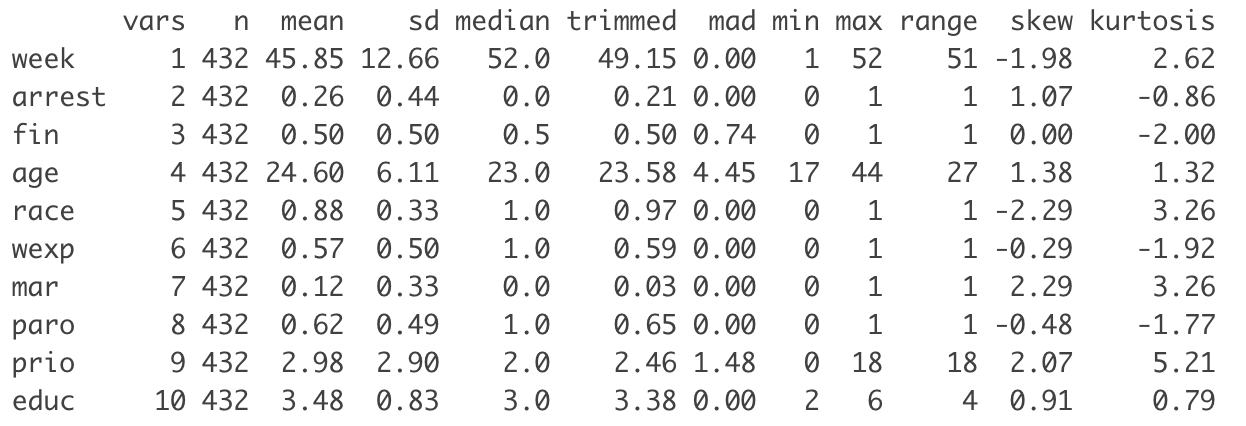
### **Appendix**

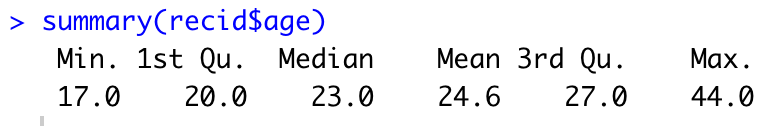
#### Description of Dataset

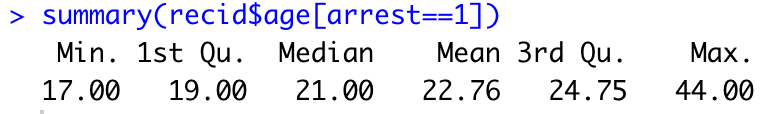
install.packages("psych")

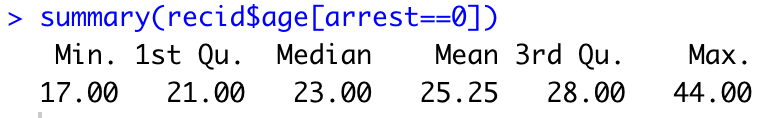
library("psych")

describe(recid)

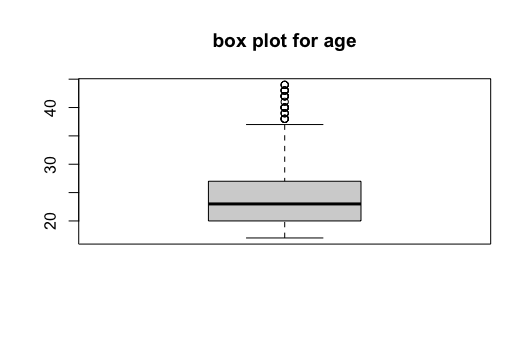






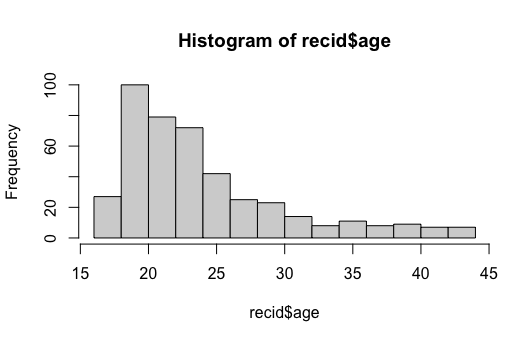


**Figure 1.1 Summary of dataset and variables**

boxplot(recid$age, main = "box plot for age")

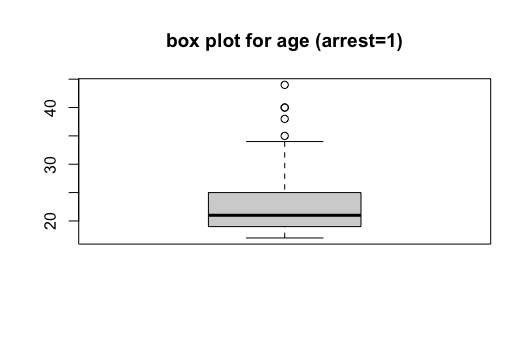
**Figure 1.2 Box plot for age**

hist(recid$age)



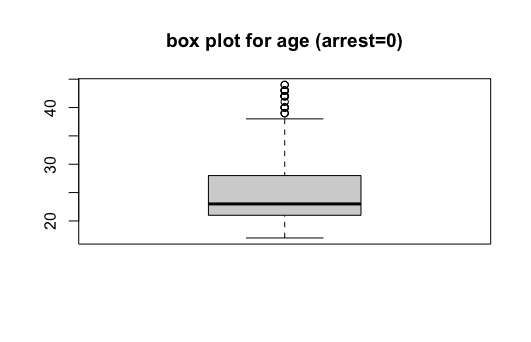
**Figure 1.3 Histogram of age**

boxplot(recid$age[arrest==1], main = "box plot for age (arrest=1)")



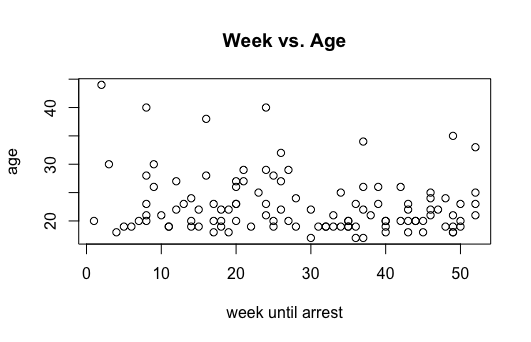
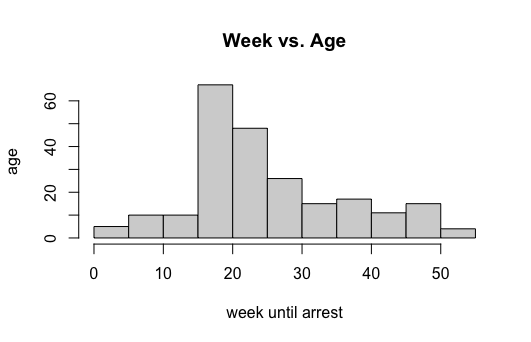
**Figure 1.4 Box plot for age (arrest=1)**

boxplot(recid$age[arrest==0], main = "box plot for age (arrest=0)")

**Figure 1.5 Box plot for age (arrest=0)**

week.age <- cbind(recid$week[arrest==1], recid$age[arrest==1])

hist(week.age, xlab = "week until arrest", ylab = "age", main = "Week vs. Age")



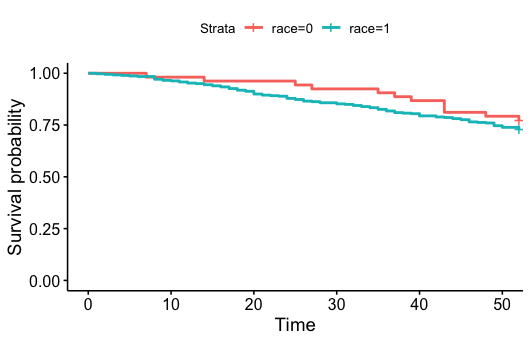
**Figure 1.6 Histogram and scatter plot for week vs. age**

#### Hypothesis Testing

km.1 <- survfit(Surv(week,arrest)~1)

summary(km.1)

ggsurvplot(km.l, data = recid)

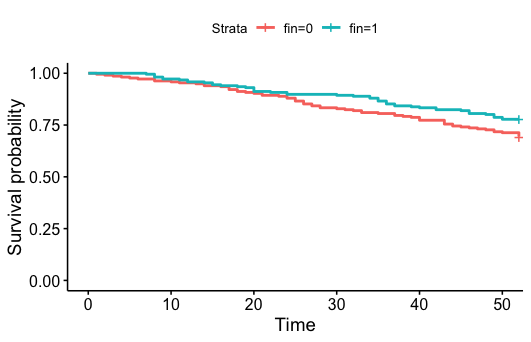


**Figure 2.1 KM Race**

km.fin <- survfit(Surv(week,arrest)~fin)

summary(km.fin)

ggsurvplot(km.fin, data = recid)



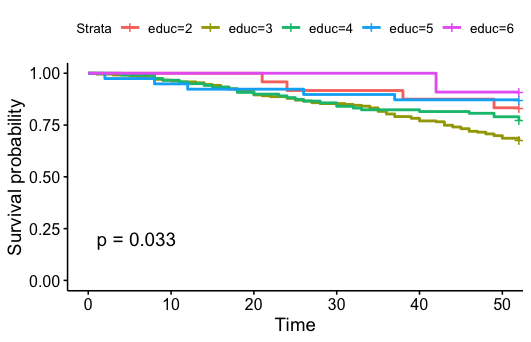
**Figure 2.2 KM Fin**

km.educ <- survfit(Surv(week,arrest)~educ)

km.educ

summary(km.educ)

ggsurvplot(km.educ, data = recid)



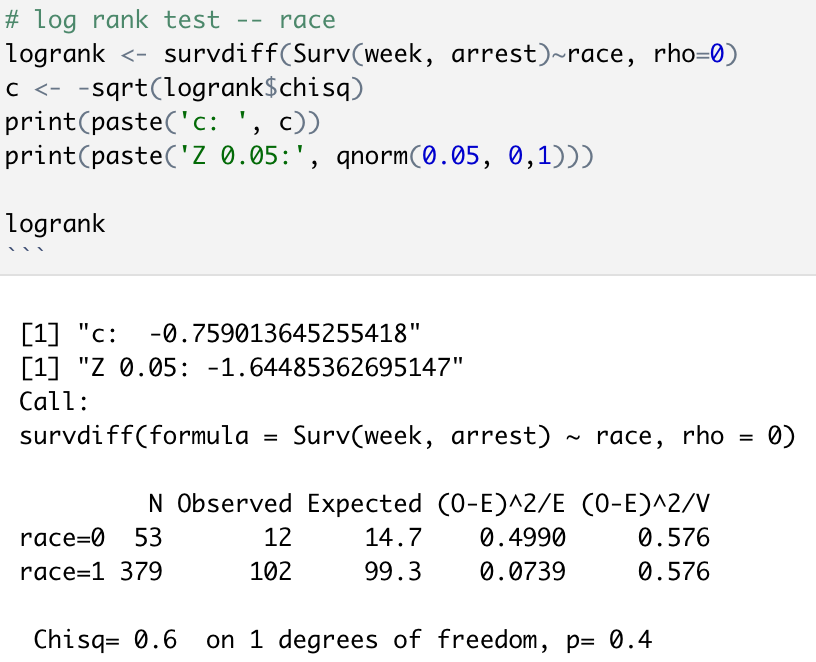
**Figure 2.3 KM Education**

#### 

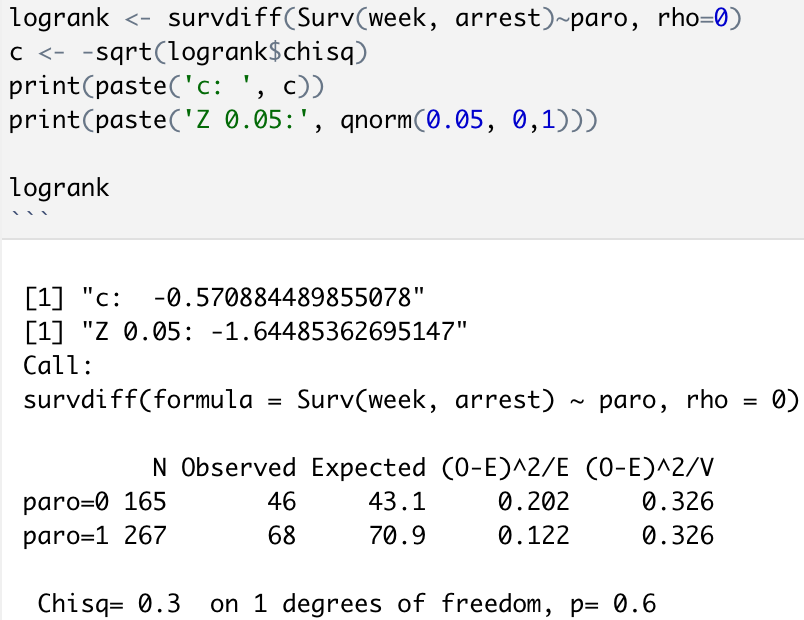
**Figure 2.4 K-M estimate on educ**

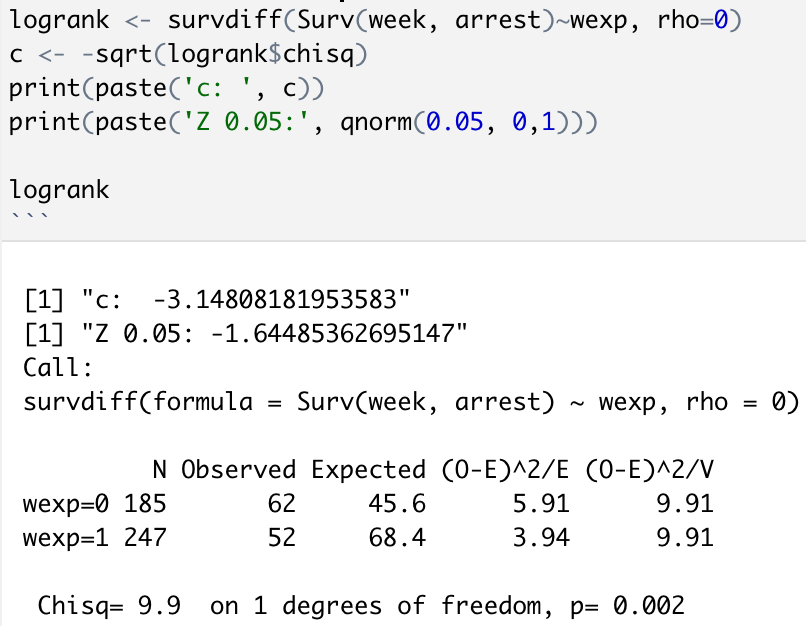
#### Non parametric (log rank)

**Figure 2.5 Log rank test for fin**

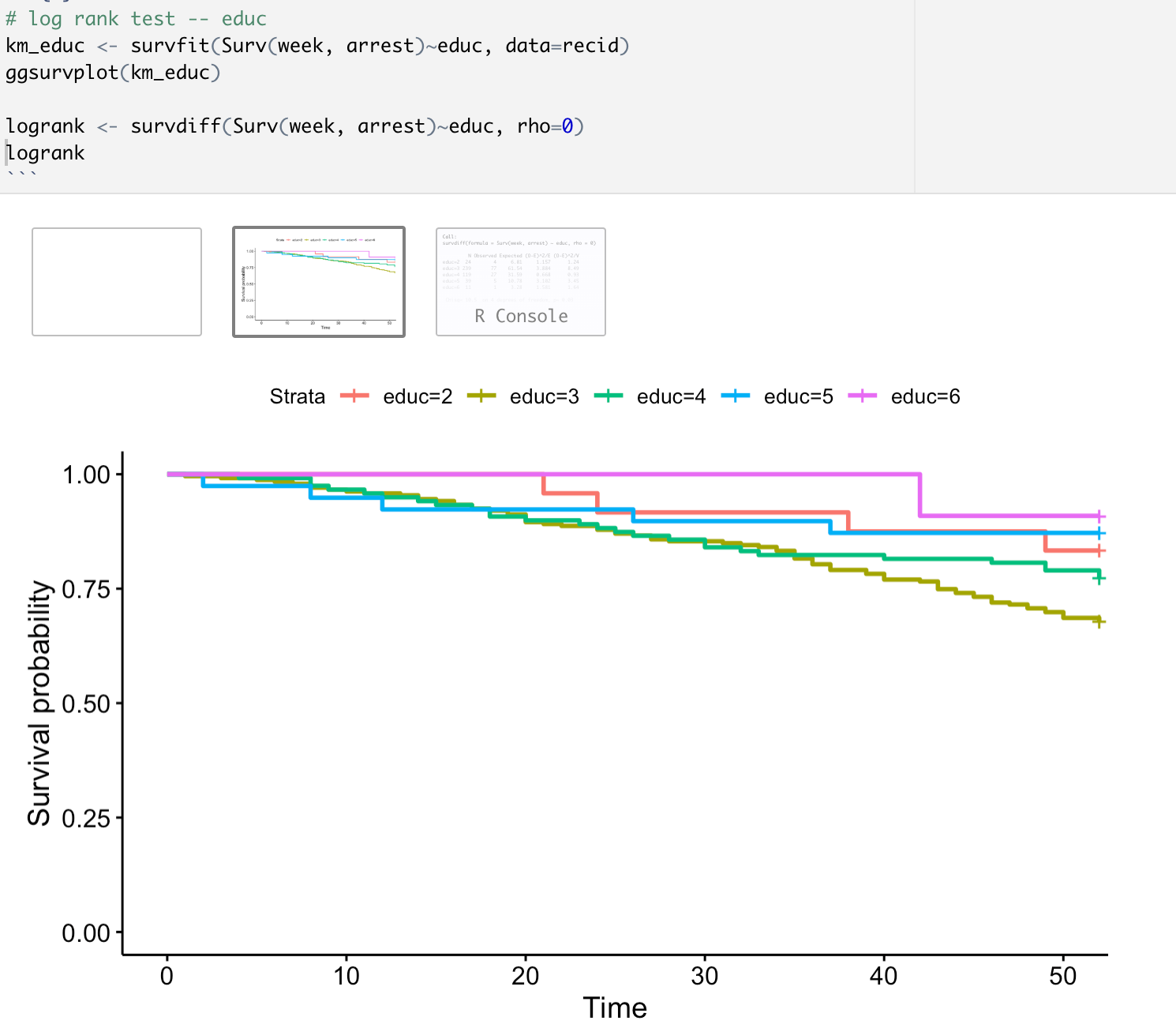
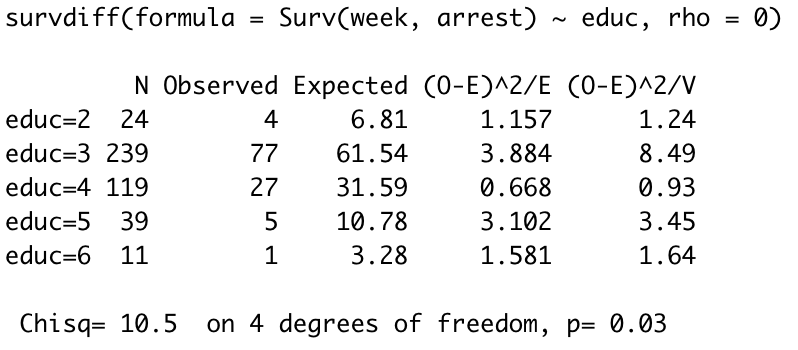


**Figure 2.6 Log rank test for race**

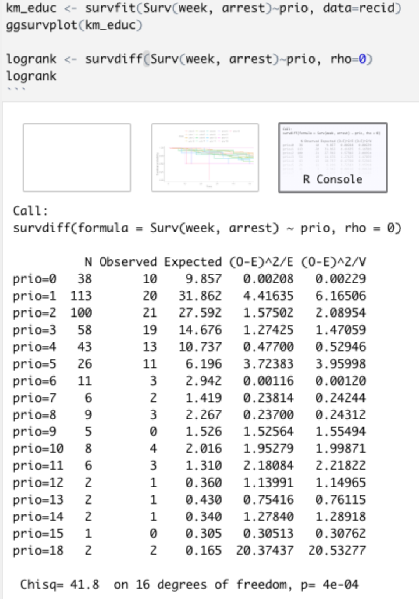
**Figure 2.7 Log rank test for paro**



**Figure 2.8 Log rank test for wexp**

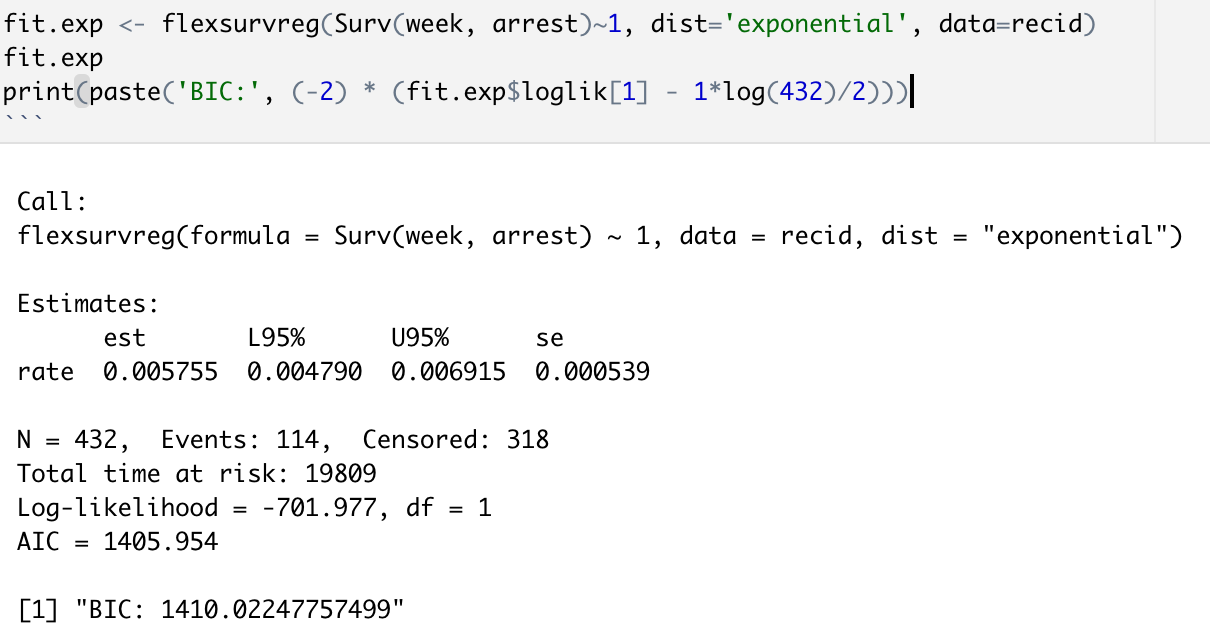


**Figure 2.9 Log rank test for educ**

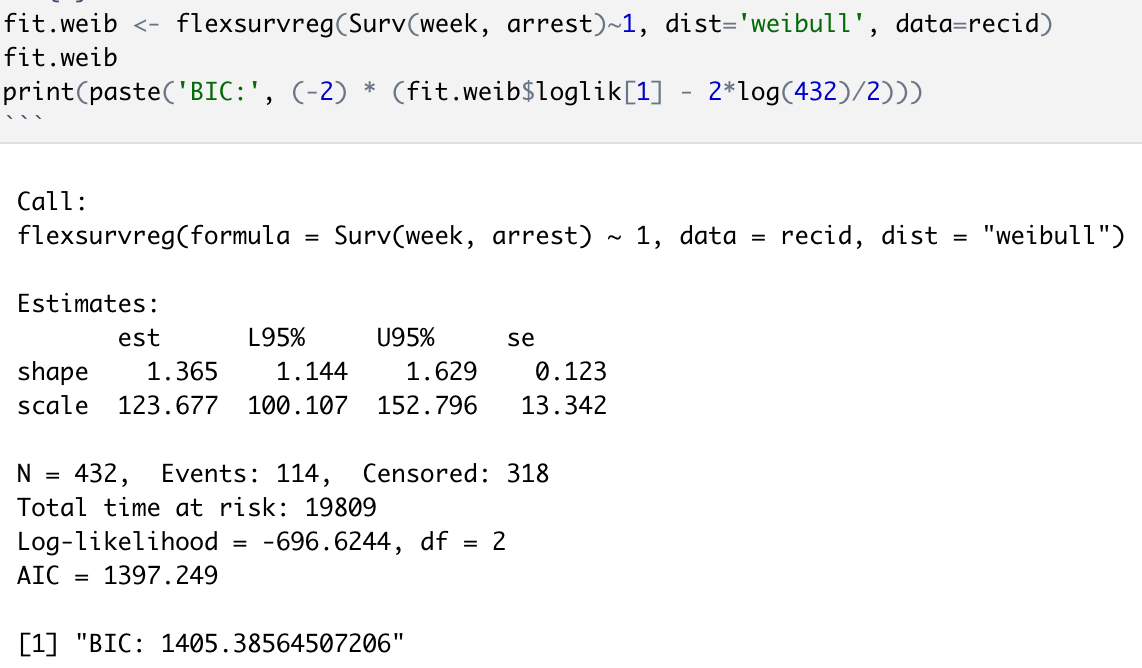


**Figure 2.10 Log rank test for prio**

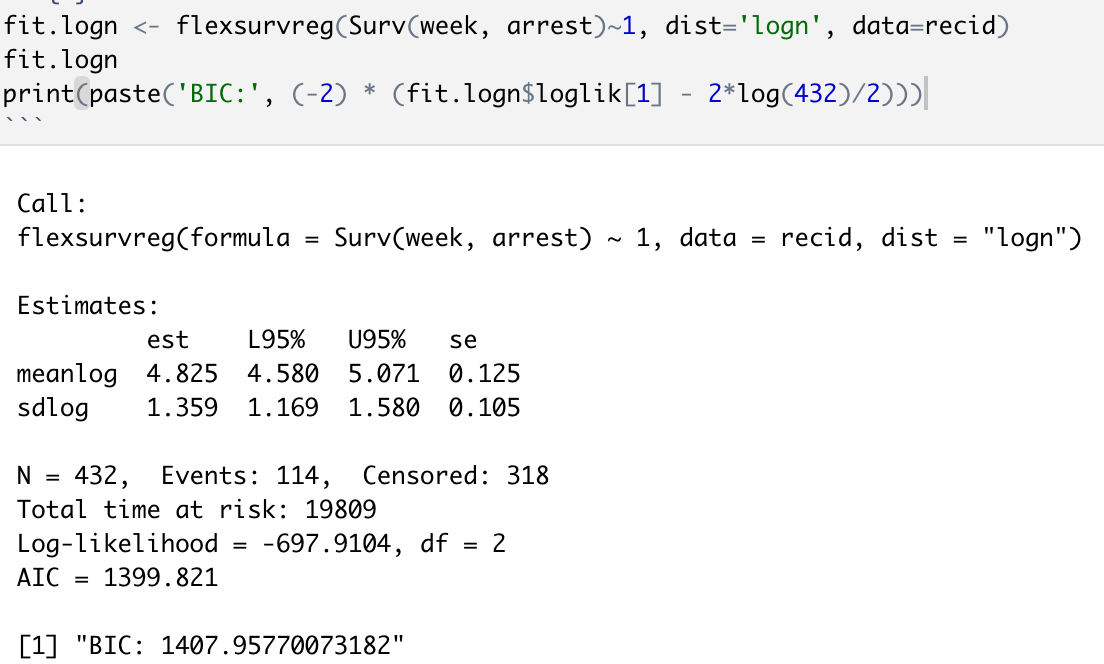
#### Parametric



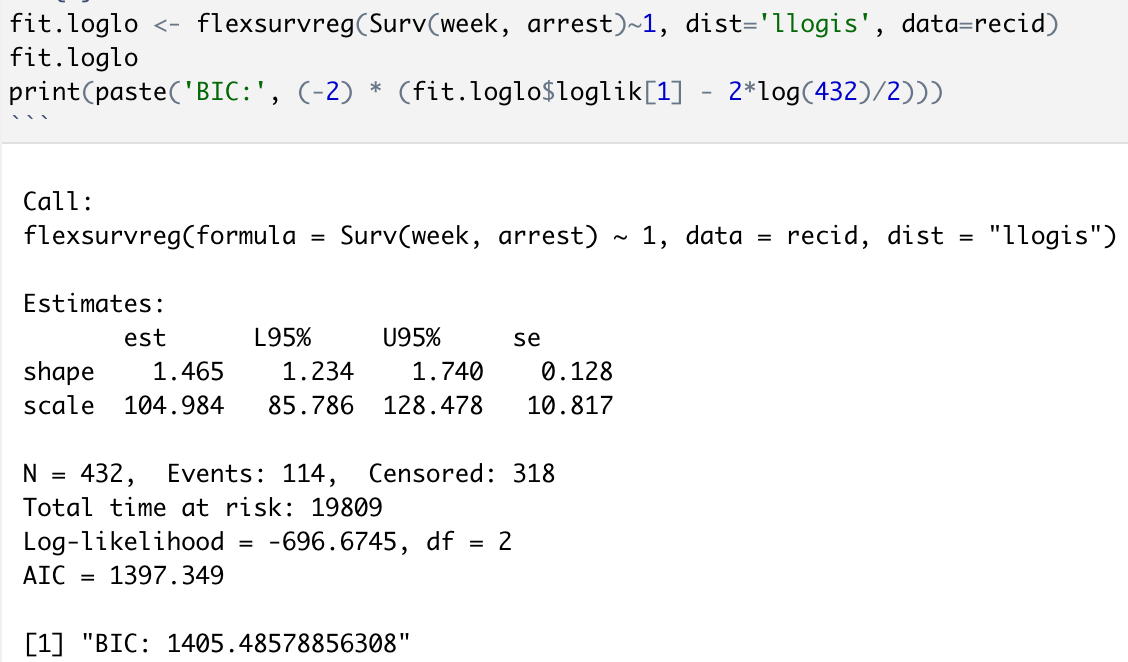
**Figure 2.11 Exponential fitting**



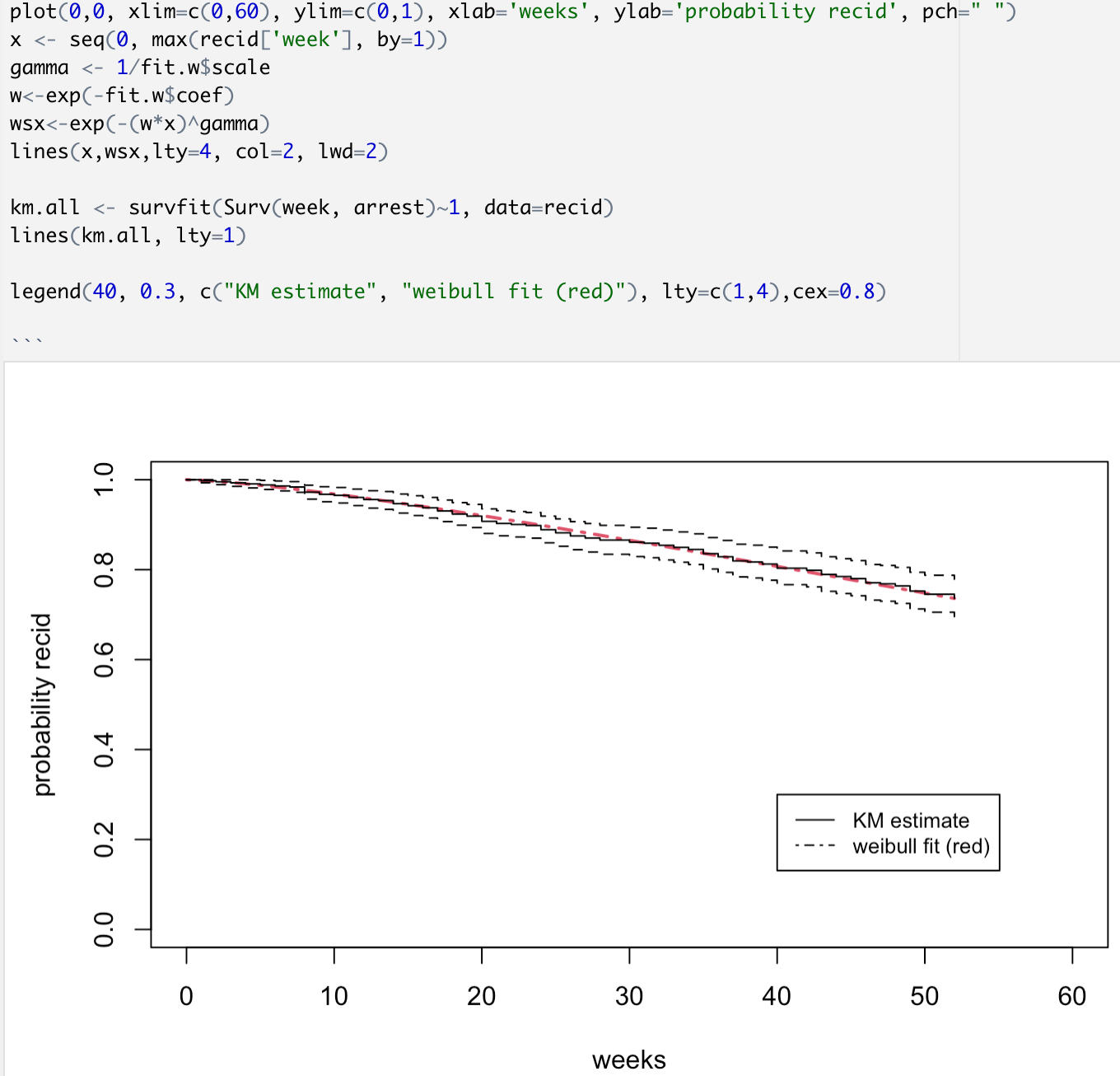
**Figure 2.12 Weibull fitting**



**Figure 2.13 Log normal fitting**

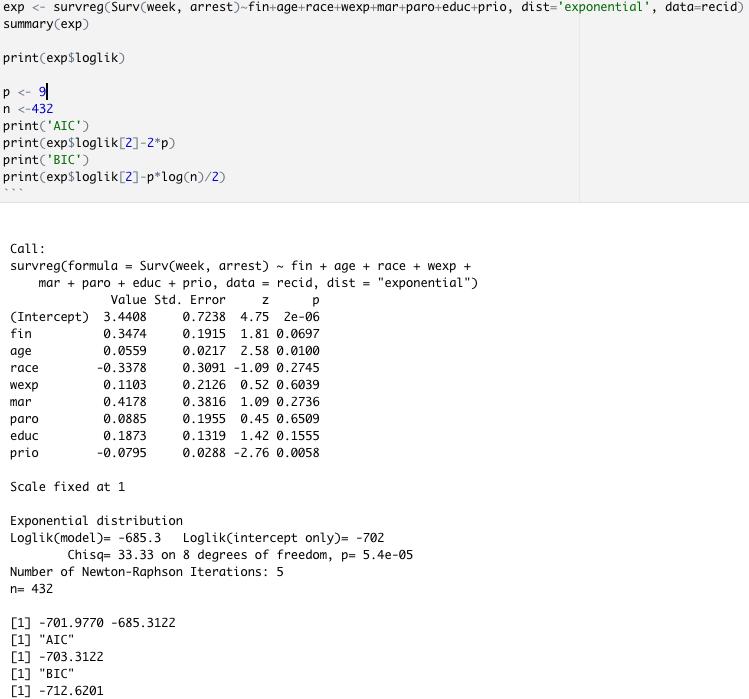


**Figure 2.14 Log logistic fitting**

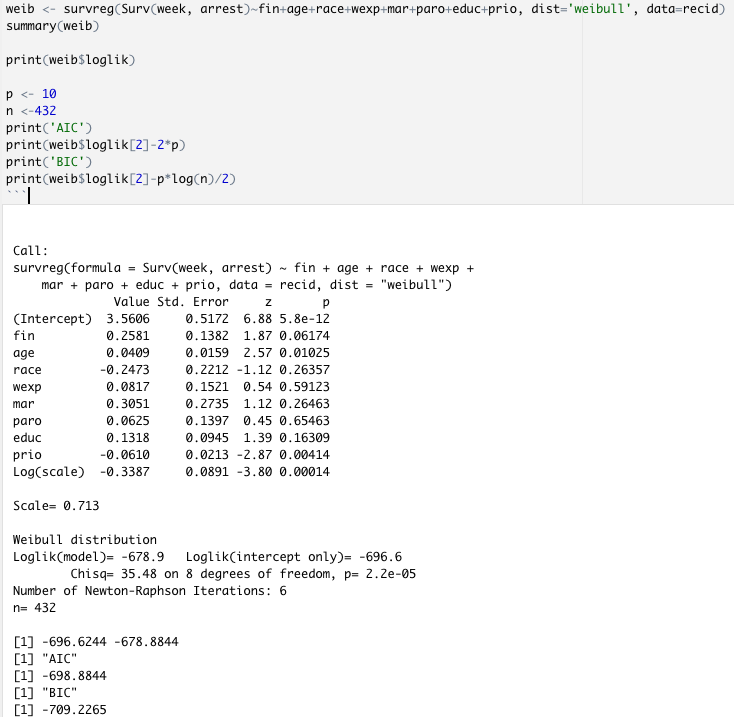


**Figure 2.15 Weibull (optimal model) plotting**

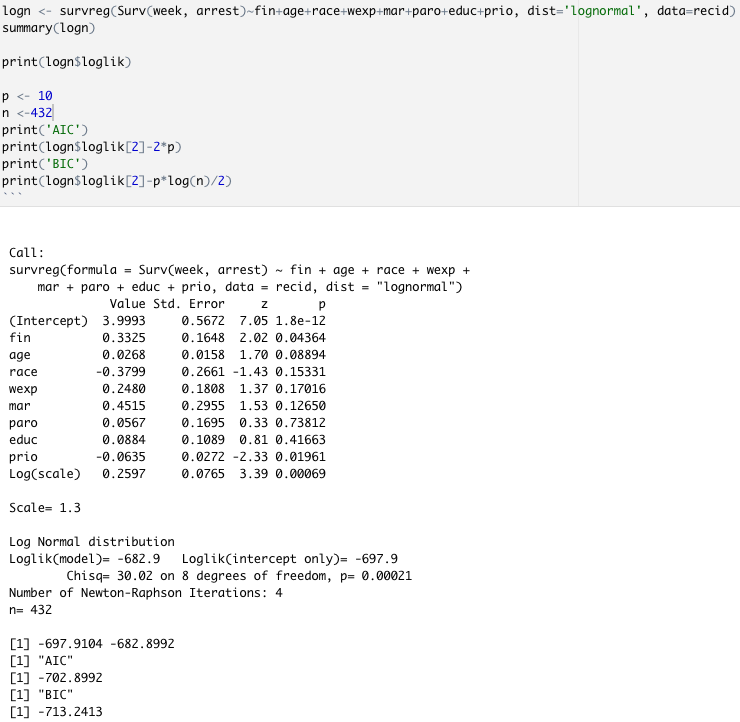
#### AFT



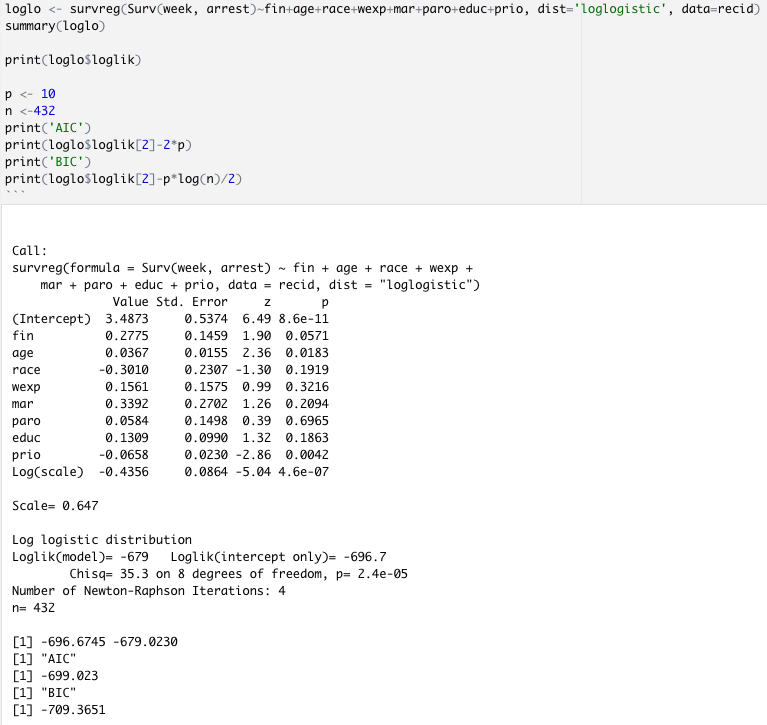
**Figure 3.1 exponential AFT**



**Figure 3.2 Weibull AFT**



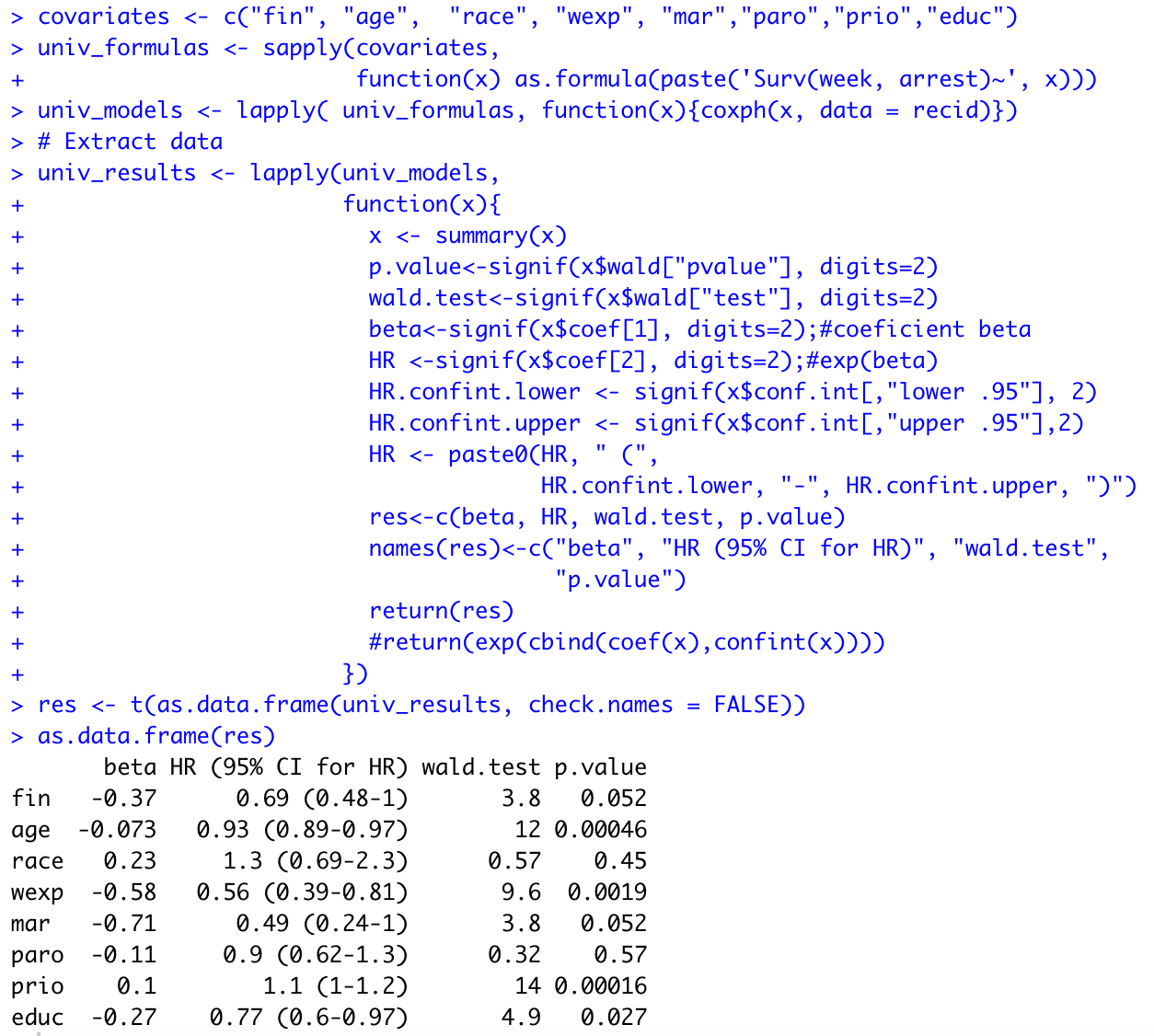
**Figure 3.3 Lognormal AFT**



**Figure 3.4 Log logistic AFT**

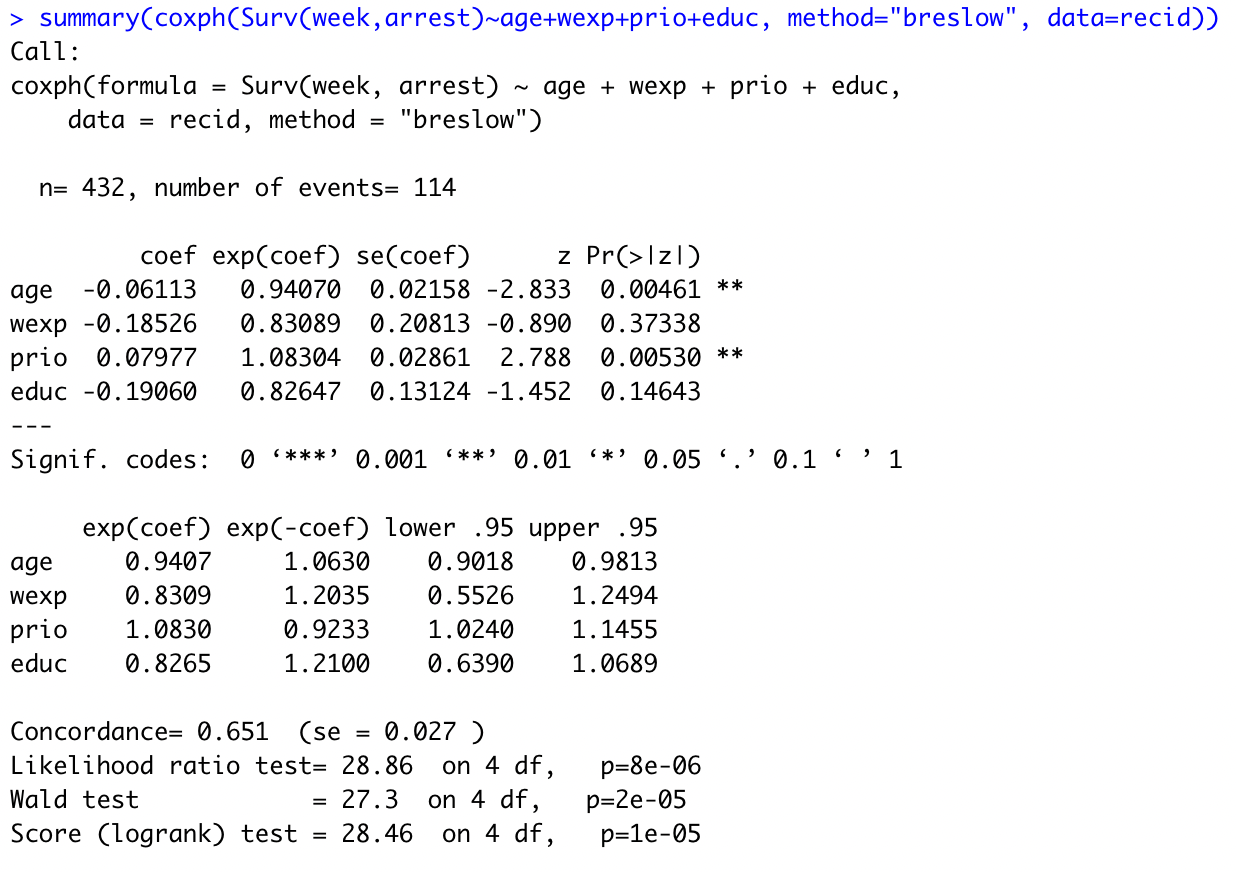
#### Semiparametric

##### Univariate



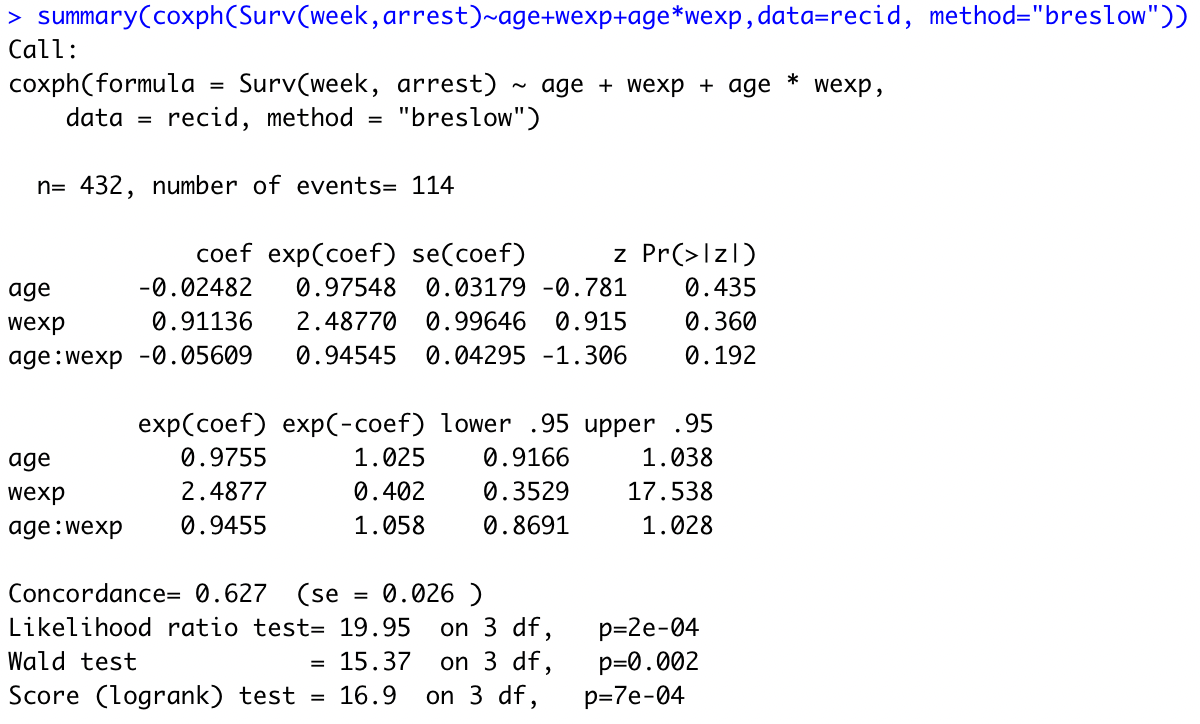
**Figure 3.5.1 Univariate Cox Analysis**

##### Multivariate

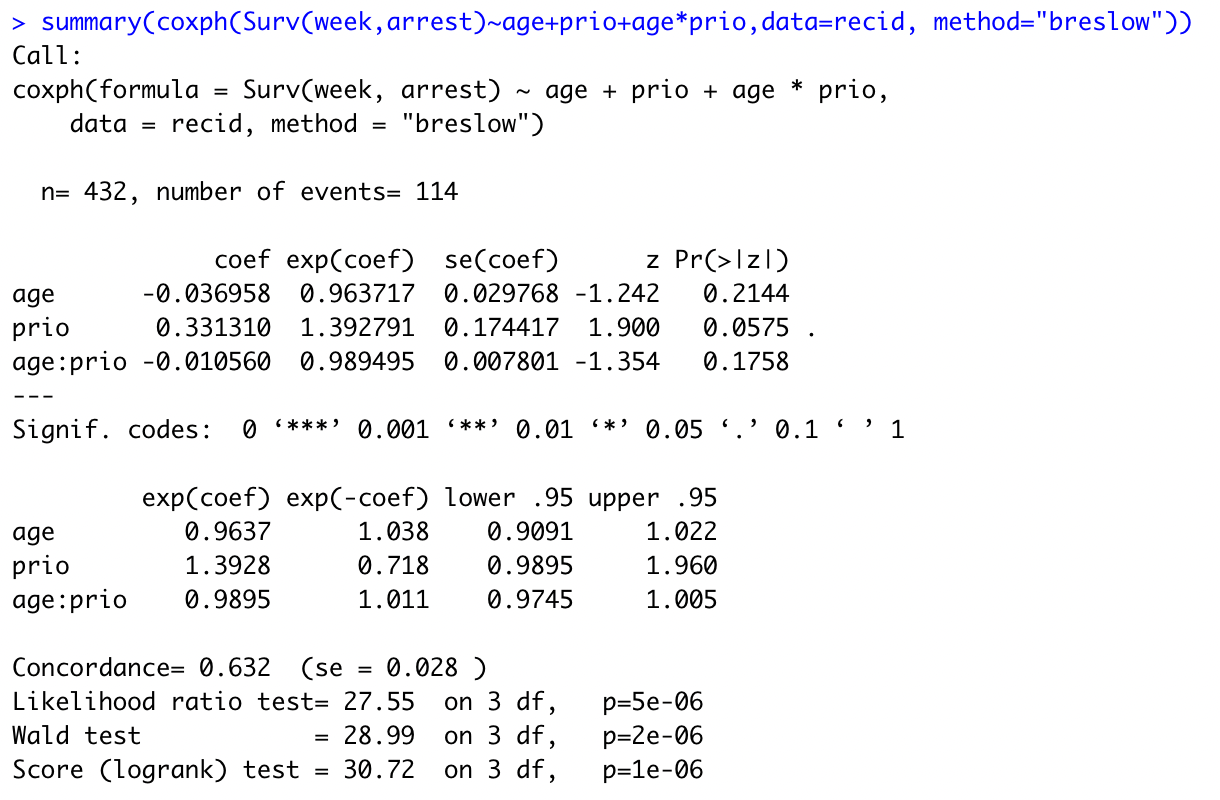


**Figure 3.5.2 Multivariate Cox Regression Analysis**

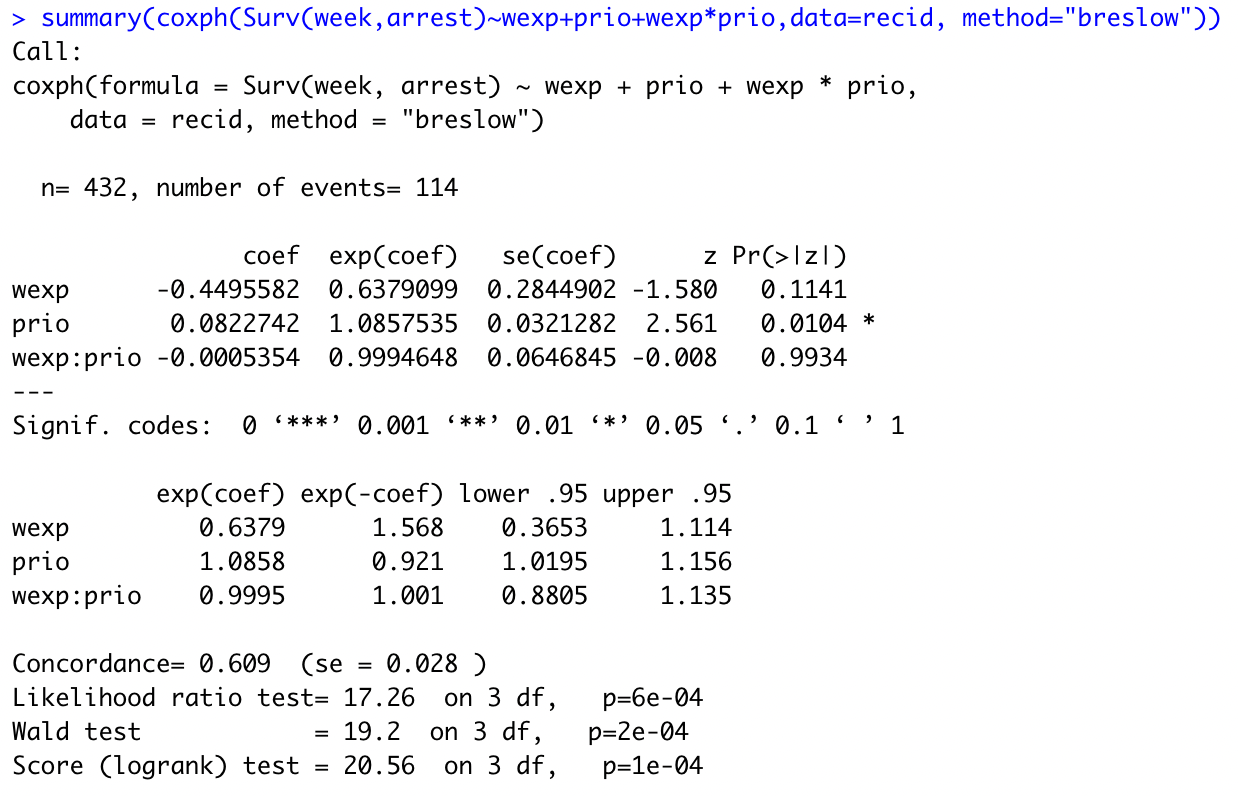
##### Interactions

****

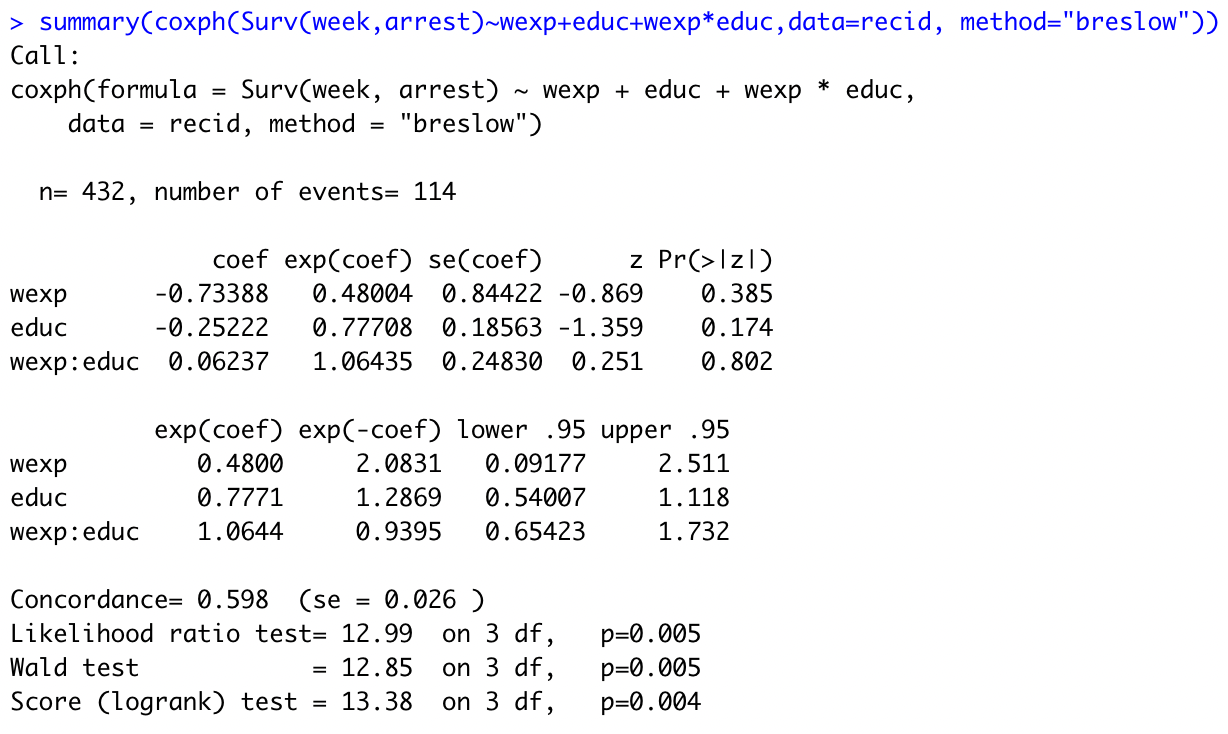
**Figure 3.5.3.1 Age:Wexp Interactions**

****

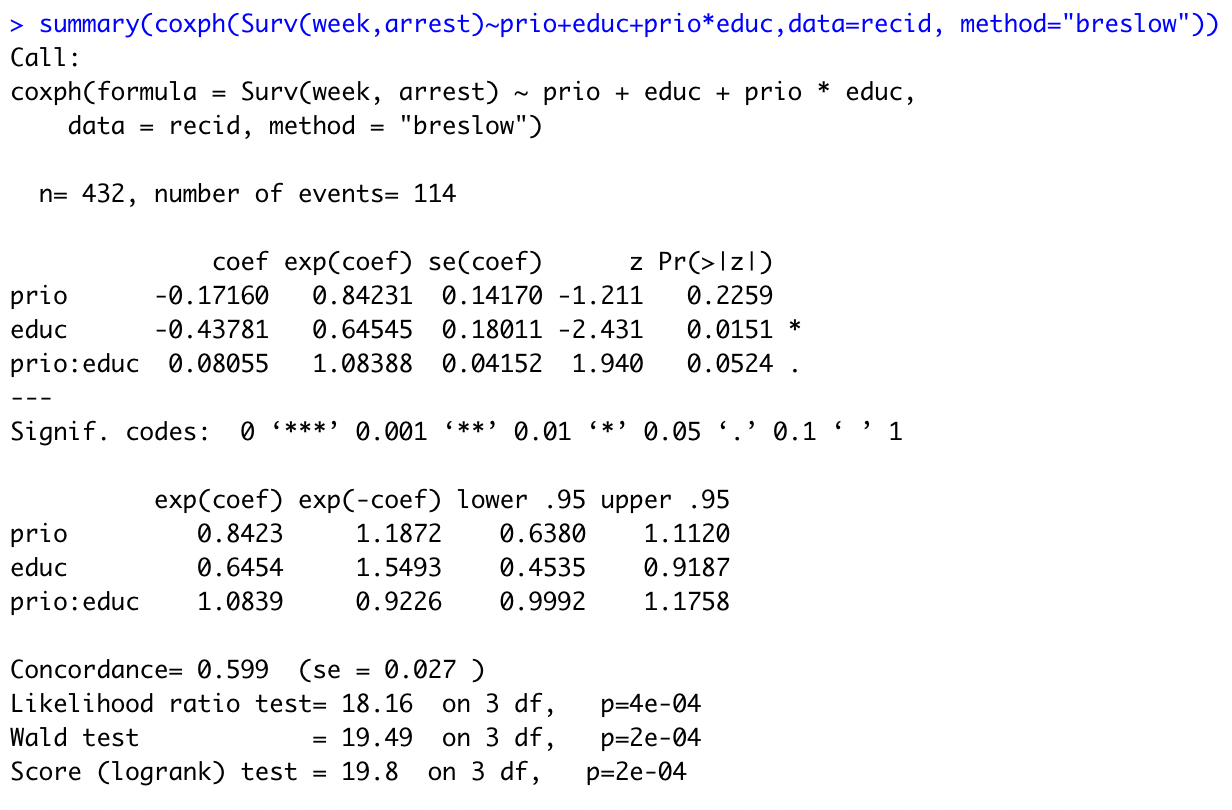
**Figure 3.5.3.2 Age:Prio Interactions**

****

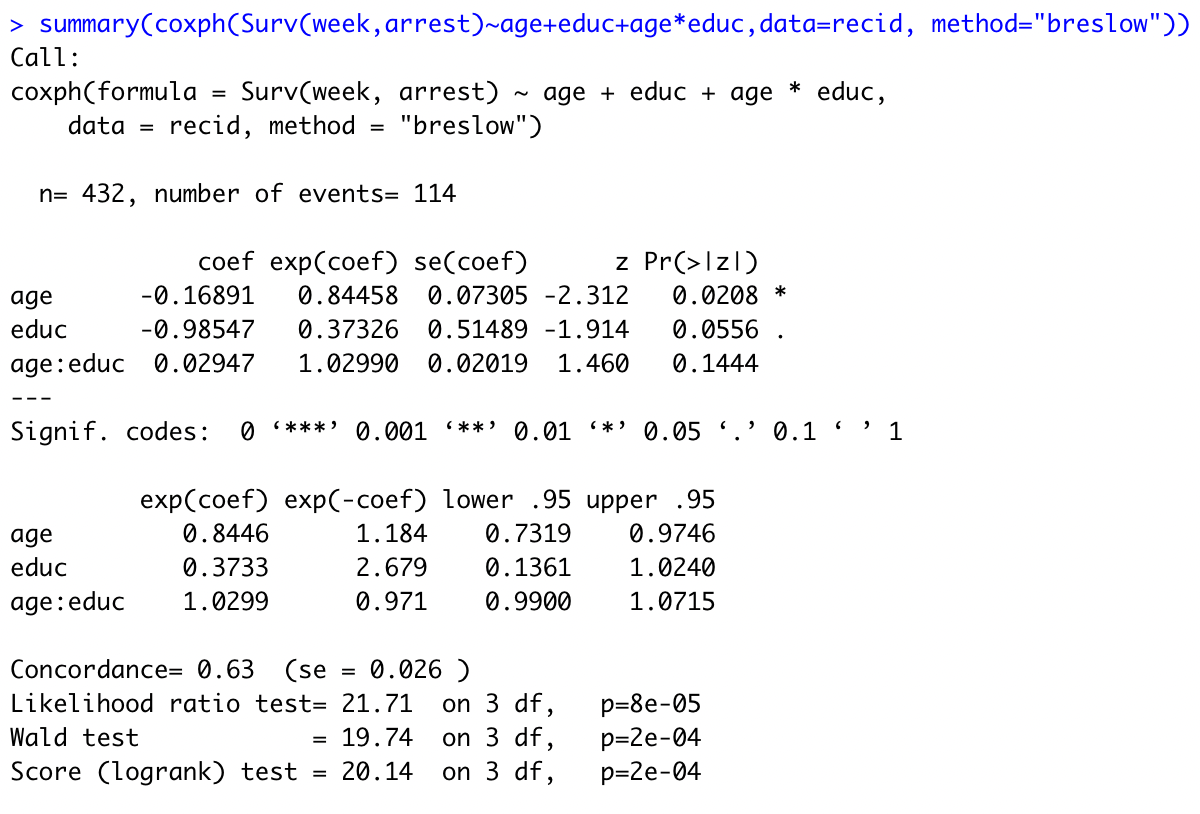
**Figure 3.5.3.3 Wexp:Prio Interactions**

****

**Figure 3.5.3.4 Wexp:Educ Interactions**

****

**Figure 3.5.3.5 Prio:Educ Interactions**

****

**Figure 3.5.3.6 Age:Educ Interactions**