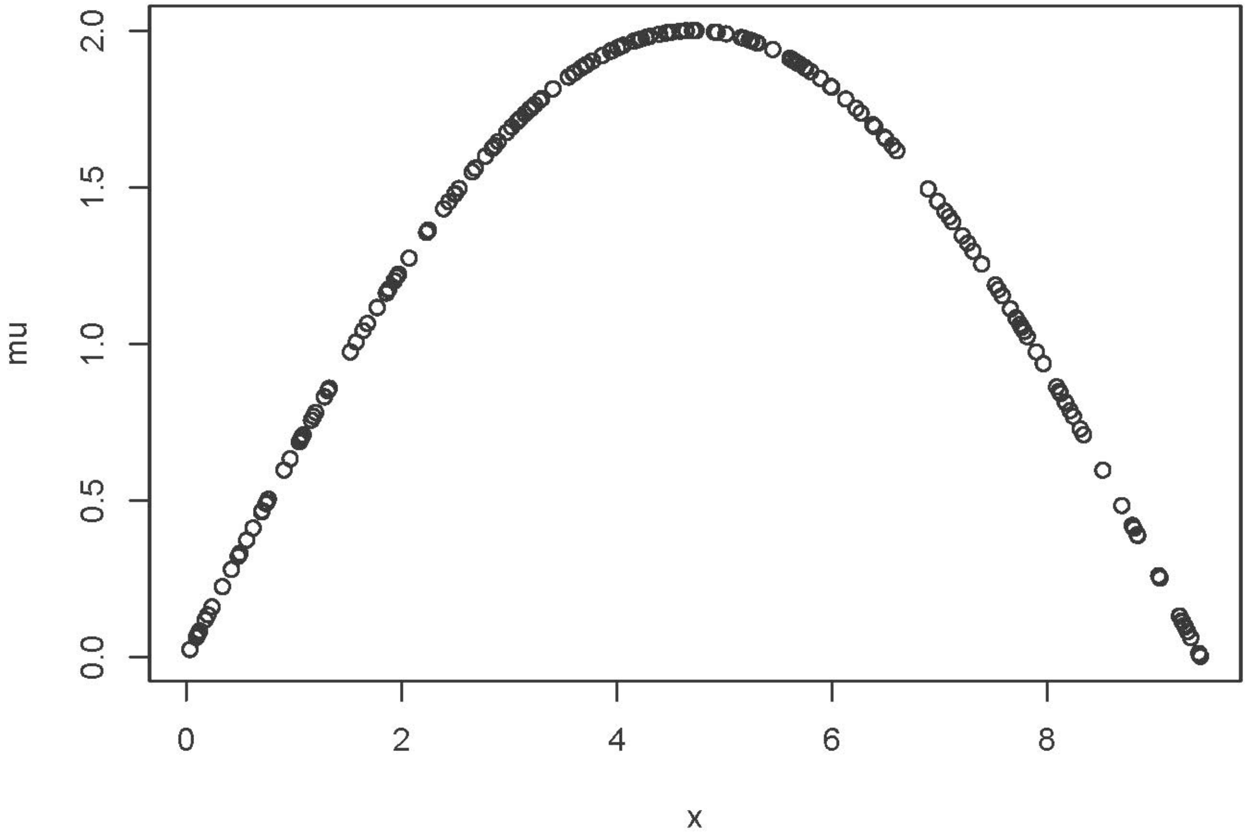
Model Diagnostics with R[*]

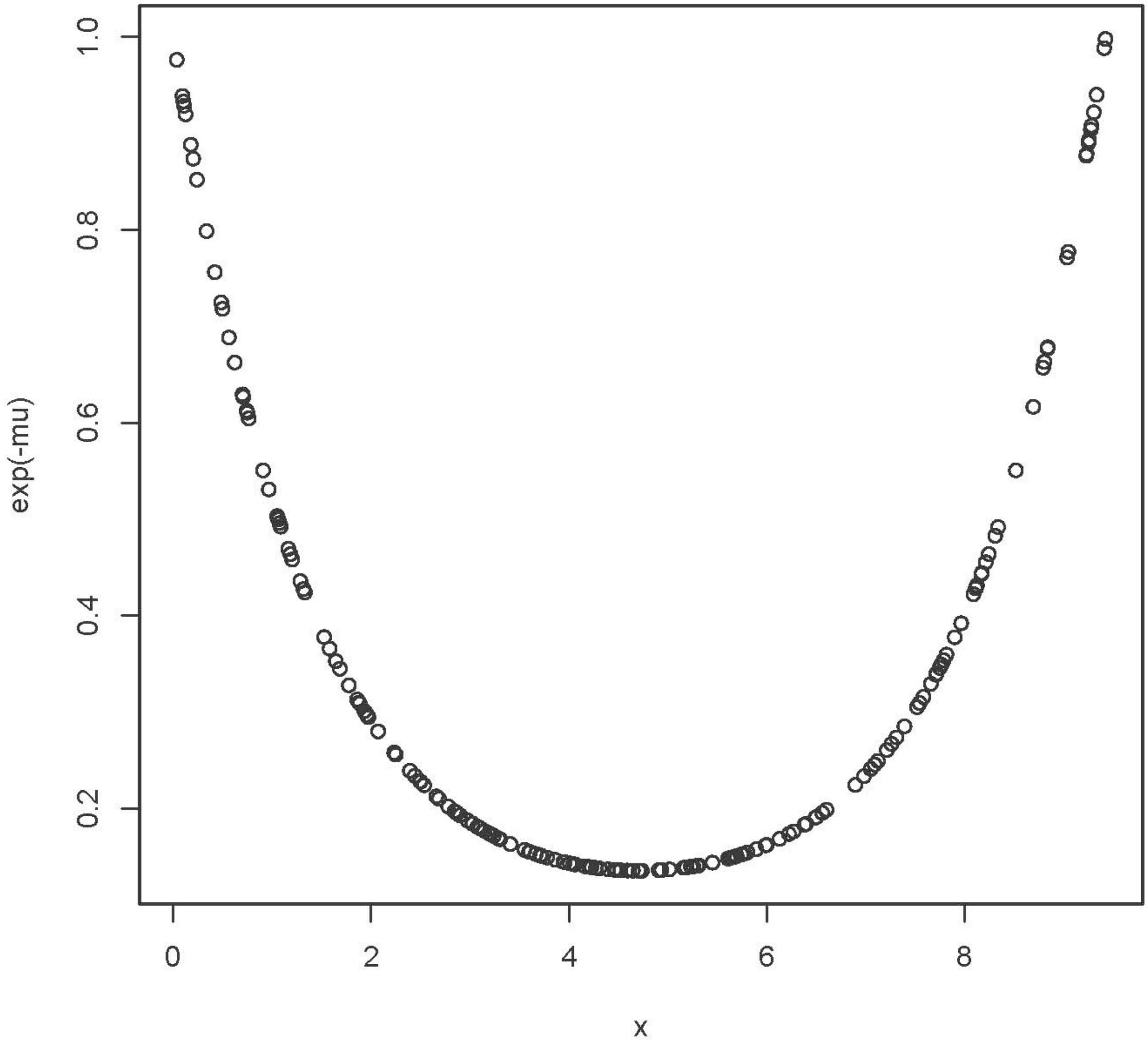
First, experiment with simulated data, where we know the truth.

```
> # Try a proportional hazards (Weibull) model for which the functional
> # form is curvy, not a straight line.
>
> rm(list=ls()); # options(scipen=999)
> top = 3*pi # Upper limit of uniform distribution on x
> sigma = 1
> # Simulate
> set.seed(9999); n = 200
> delta = numeric(n) # Indicator for uncensored, initially zero
> x = runif(n,0,top)
> mu =  2*sin(x/3)
> plot(x,mu)
```

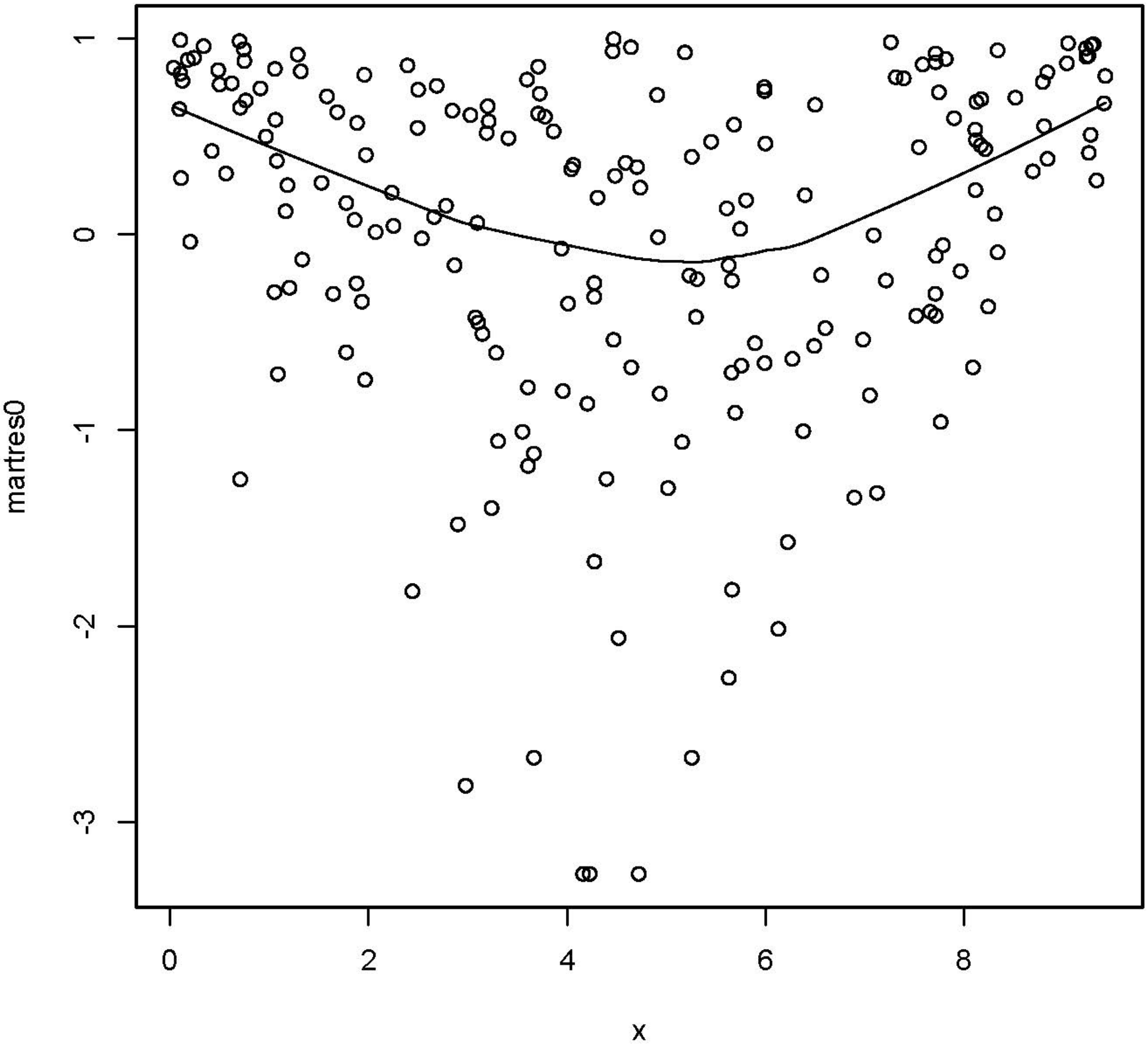


[*] Copyright information is on the last page.

```
> epsilon = rexp(n)
> lifetime = exp(mu)*epsilon^sigma # Weibull regression with a funny
>                                  # functional form
> plot(x,exp(-mu)) # The hazard function is proportional to exp(-mu)
```

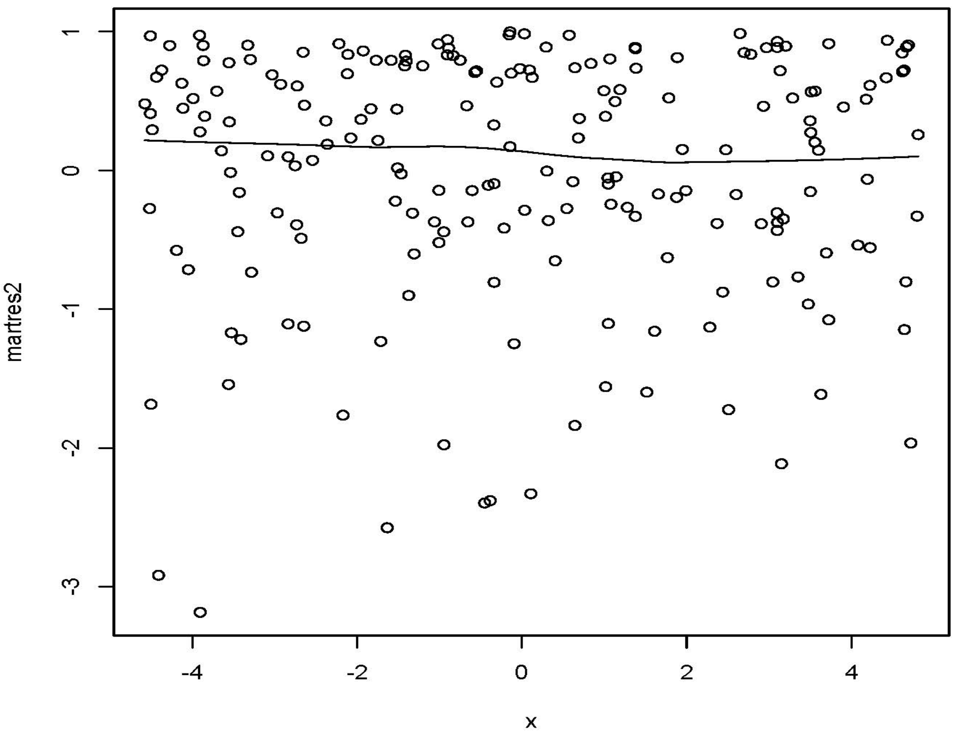


```
>  censortime = abs(rnorm(n,0,20)) # Absolute normal censoring time
> # If censoring time is greater than lifetime, then it's NOT censored.
> delta[censortime>lifetime] = 1;  table(delta)
delta
  0   1 
 32 168 
> # Minumum of censortime and lifetime is what we can observe.
> Time = pmin(censortime,lifetime) # pmin is parallel minimum.
> Time = round(Time,3)
> # round(cbind(x,lifetime,censortime,Time,delta)[1:10,],3) # Take a look
> wdata = cbind(x,Time,delta); # wdata # This is all you can see in practice.
> head(wdata)
            x   Time delta
[1,] 8.113222  3.351     1
[2,] 6.222630 11.620     1
[3,] 7.543745  2.194     1
[4,] 1.957017  0.576     1
[5,] 6.498236  1.183     1
[6,] 7.962717  5.430     1

> # Fit the model
> library(survival)
> stime = Surv(Time,delta)
> ph1 = coxph(stime~x); summary(ph1)
Call:
coxph(formula = stime ~ x)

  n= 200, number of events= 168 

       coef exp(coef)  se(coef)    z Pr(>|z|)
x -0.003535  0.996471  0.035198 -0.1     0.92

  exp(coef) exp(-coef) lower .95 upper .95
x    0.9965      1.004      0.93     1.068

Concordance= 0.511  (se = 0.026 )
Rsquare= 0   (max possible= 0.999 )
Likelihood ratio test= 0.01  on 1 df,   p=0.92
Wald test            = 0.01  on 1 df,   p=0.92
Score (logrank) test = 0.01  on 1 df,   p=0.92

> # help(residuals.coxph)
> 
> # Fit a model with no explanatory variables
> ph0 = coxph(stime ~ 1); summary(ph0)
Call:  coxph(formula = stime ~ 1)

Null model
  log likelihood= -734.5636 
  n= 200 

> martres0 = residuals(ph0,type='martingale')
> plot(x,martres0)
> smooth = lowess(x,martres0); lines(smooth)
> martres0 = residuals(ph0,type='martingale')
> plot(x,martres0)
> smooth = lowess(x,martres0); lines(smooth)
```



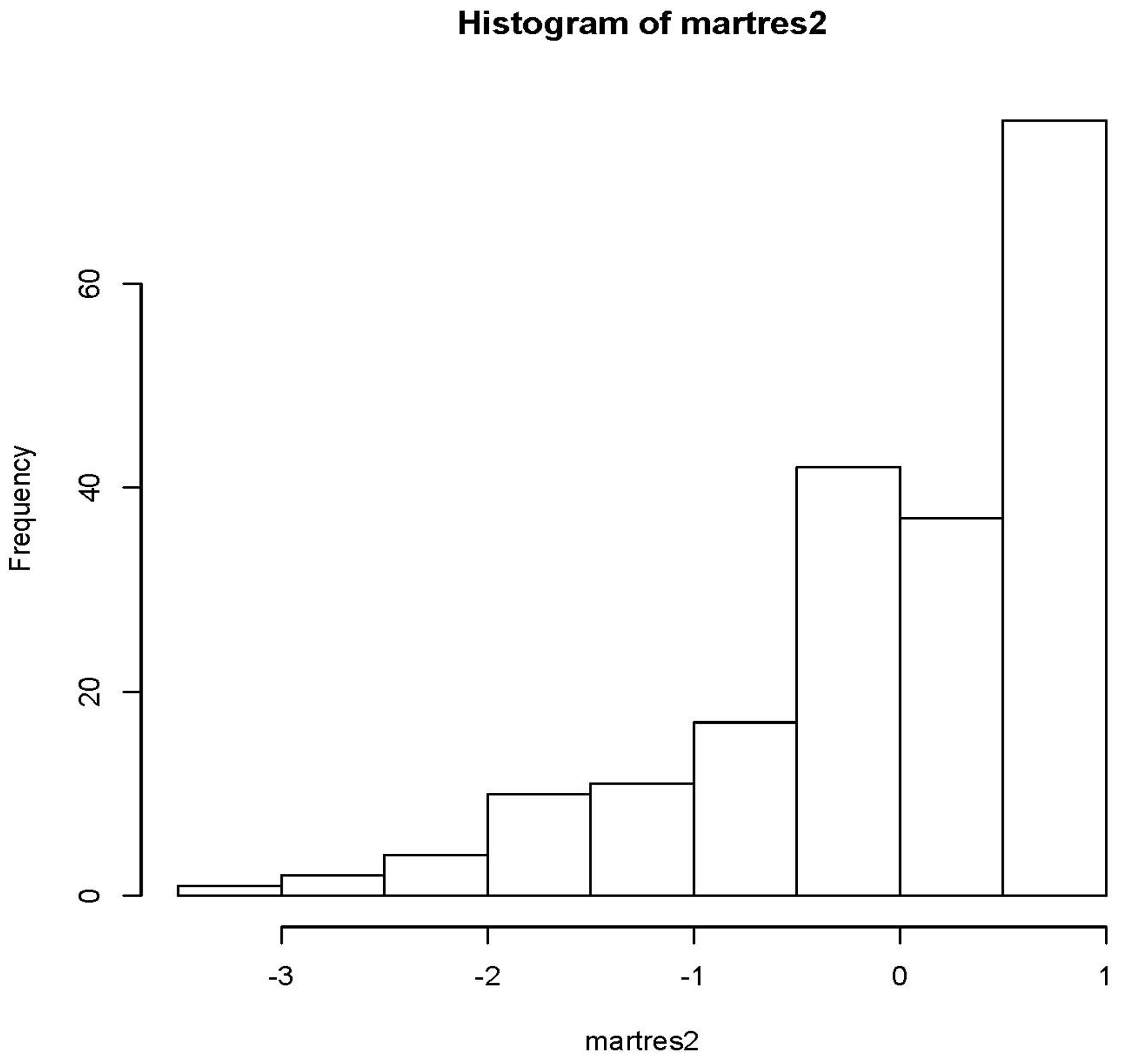
This suggests a U-shaped function. No way to guess the truth. Try polynomial regression.

```
> x = x-mean(x) # centered
> x2 = x^2 # Quadratic term
> ph2 = coxph(stime~x+x2); summary(ph2)
Call:
coxph(formula = stime ~ x + x2)

  n= 200, number of events= 168

       coef exp(coef) se(coef)      z Pr(>|z|)
x  -0.01431   0.98579  0.02695 -0.531    0.595
x2  0.10427   1.10990  0.01269  8.220 2.22e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

   exp(coef) exp(-coef) lower .95 upper .95
x     0.9858      1.014    0.9351     1.039
x2    1.1099      0.901    1.0826     1.138

Concordance= 0.672  (se = 0.026 )
Rsquare= 0.273   (max possible= 0.999 )
Likelihood ratio test= 63.71  on 2 df,   p=1.465e-14
Wald test            = 67.58  on 2 df,   p=2.109e-15
Score (logrank) test = 74.39  on 2 df,   p=1.11e-16

>
> martres2 = residuals(ph2,type='martingale')
> plot(x,martres2)
> smooth = lowess(x,martres2); lines(smooth)
```



Looks clean.

```
> martres2 = residuals(ph2,type='martingale')
> plot(x,martres2)
> smooth = lowess(x,martres2); lines(smooth)
>
> hist(martres2) # Educational. Max is one.
```

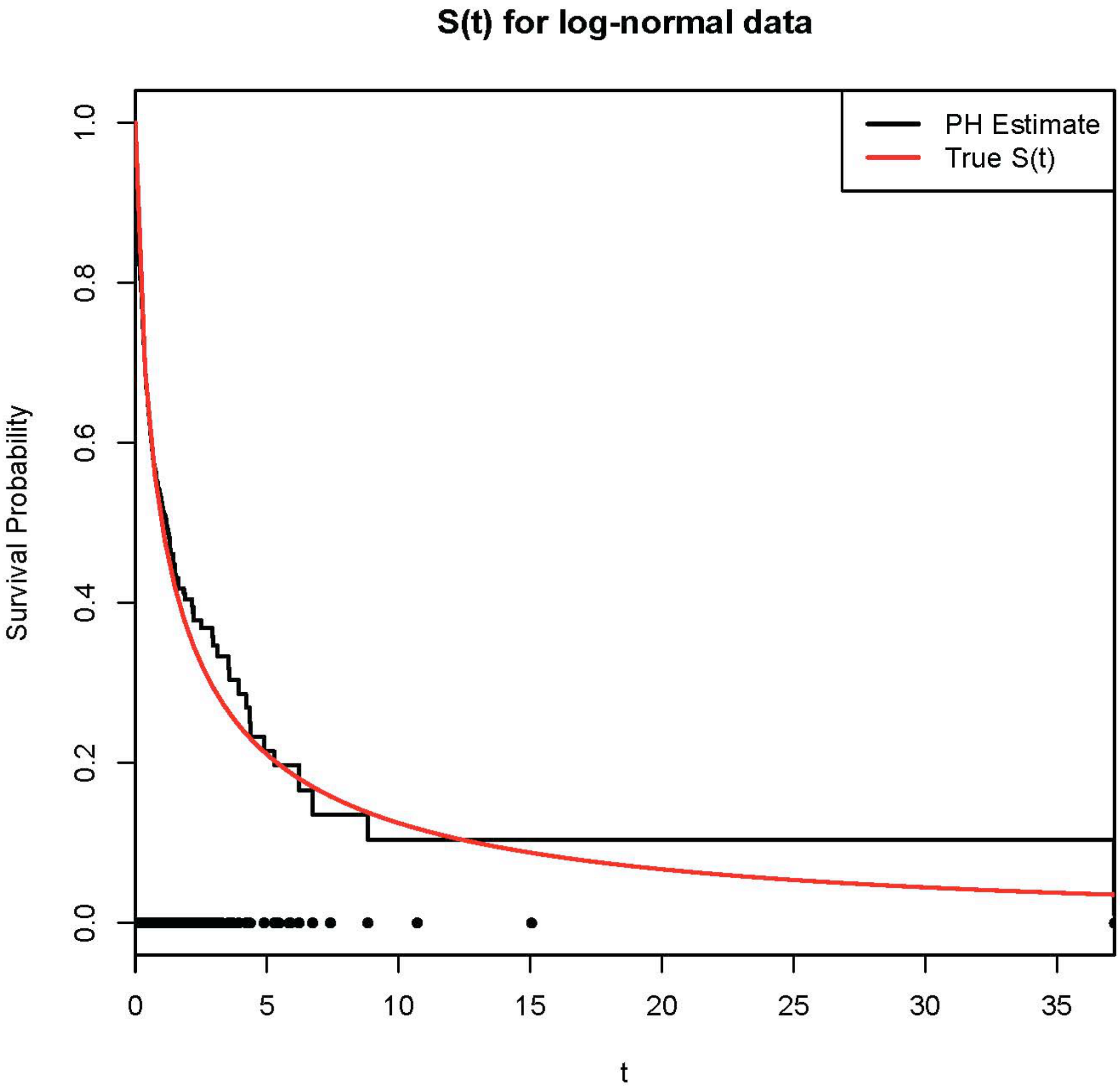


```
> sum(martres2)
[1] -5.827804e-15
>
> cox.zph(ph2) # Test proportional hazards (H0 is true)
           rho  chisq     p
x      0.00814 0.0109 0.917
x2     0.03671 0.2417 0.623
GLOBAL      NA 0.2589 0.879
```

Another experiment. This time, the truth is log-normal, not proportional hazards.

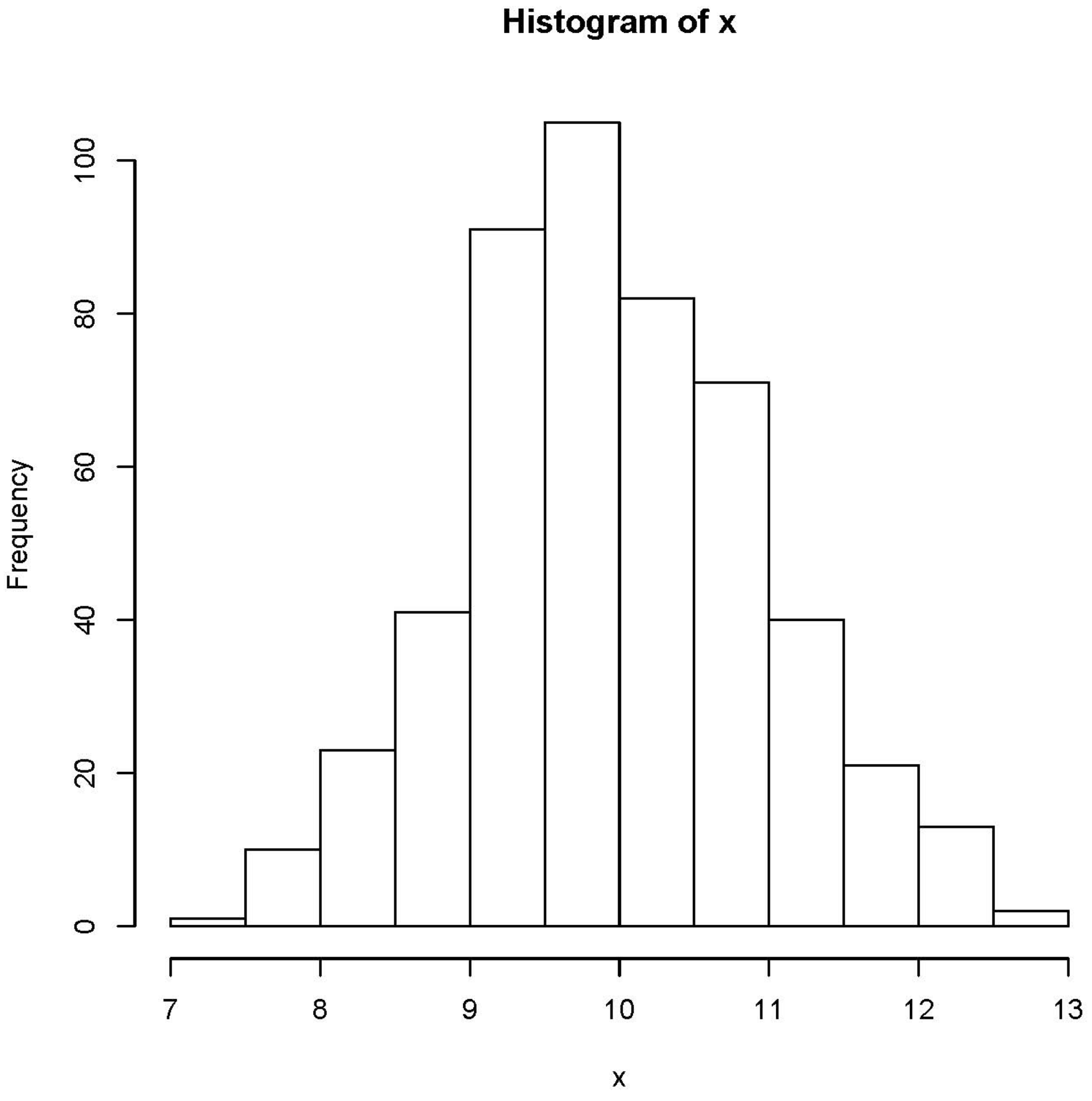
```
> rm(list=ls());  # options(scipen=999)
> Ex = 10; SDx = 1 # Parameters of (normal) explanatory variable X
> beta0 = -10; beta1 = 1; sigma = 2 # Regression parameters
> n = 500; id = 1:n
> delta = numeric(n) # Indicator for uncensored, initially zero
> # install.packages("survival",dependencies=TRUE) # Only need to do this once
> library(survival)
> 
> # Simulate one data set
> 
> set.seed(9999)
> x = rnorm(n,Ex,SDx)
> mu = beta0 + beta1*x
> y = rnorm(n,mu,sigma); lifetime = exp(y)
> # sort(lifetime)
> # hist(sort(lifetime)[1:(n-2)],breaks=20)
> censortime = abs(rcauchy(n)) # Absolute Cauchy censoring time
> # censortime = 1/runif(n) - 1 # Shifted Pareto censoring time
> # If censoring time is greater than lifetime, then it's NOT censored.
> delta[censortime>lifetime] = 1;  table(delta)
delta
  0   1 
253 247 
> # Minumum of censortime and lifetime is what we can observe.
> Time = pmin(censortime,lifetime) # pmin is parallel minimum.
> 
> phmodel = coxph(Surv(Time,delta) ~ x); summary(phmodel)
Call:
coxph(formula = Surv(Time, delta) ~ x)

  n= 500, number of events= 247 

      coef exp(coef) se(coef)      z Pr(>|z|)    
x -0.59968   0.54899  0.06854 -8.749   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

  exp(coef) exp(-coef) lower .95 upper .95
x     0.549      1.822      0.48    0.6279

Concordance= 0.688  (se = 0.02 )
Rsquare= 0.147   (max possible= 0.995 )
Likelihood ratio test= 79.42  on 1 df,   p=0
Wald test            = 76.55  on 1 df,   p=0
Score (logrank) test = 77.91  on 1 df,   p=0
> # Test proportional hazards (H0 is that PH is correct: False for these data)
> cox.zph(phmodel)
       chisq df       p
x       12.4  1 0.00043
GLOBAL  12.4  1 0.00043

> ave = data.frame(x=Ex) # Average (True population mean) x value
> S = survfit(phmodel,newdata=ave,se.fit=FALSE); S
Call: survfit(formula = phmodel, newdata = ave, se.fit = FALSE)

     n events median 
500.00 247.00   1.16 
> truemedian = exp(beta0 + beta1*Ex)
> cat("\nTrue median survival time = exp(beta0+beta1*Ex) =",truemedian,"\n\n")

True median survival time = exp(beta0+beta1*Ex) = 1 
```

```
>
> top=max(Time) # Upper limit of x in plot
> plot(S,xlim=c(0,top),xlab='t',ylab='Survival Probability', lwd=2)
> title("S(t) for log-normal data")
> # Plot points at observed time values
> zero = Time-Time; points(Time,zero,pch=20)
> # Plot true S(t)
> tt = seq(from=0,to=top,length=101)
> trueS = 1-pnorm(log(tt), mean = beta0+beta1*Ex, sd = sigma)
> lines(tt,trueS, col='red', lwd=2)
> truered = expression('True S(t)',col='red')
> legend('topright', col=c(1,2), lwd=2, legend=c('PH Estimate','True S(t)'))
```

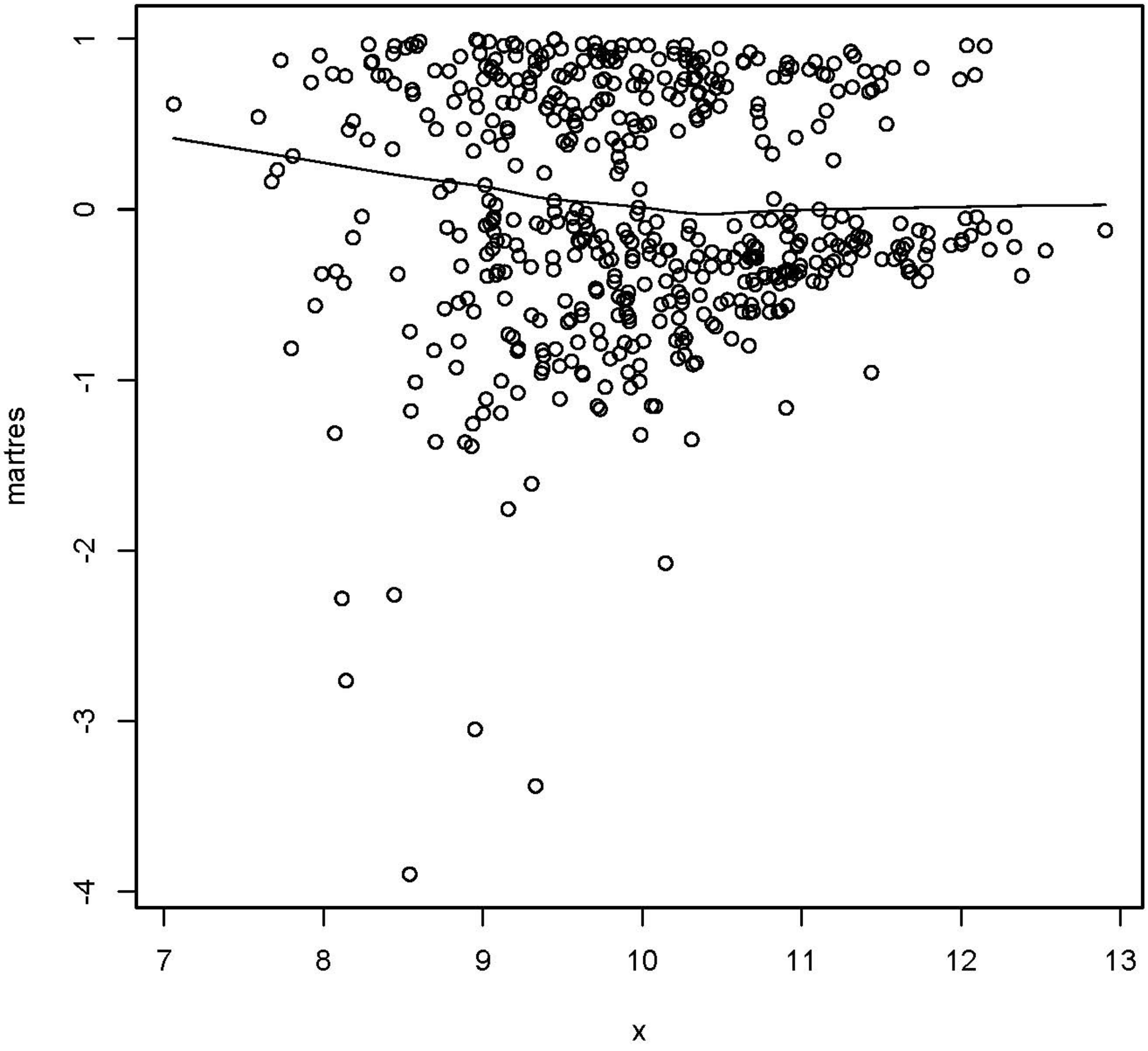


My comment is that the estimate of S(t) is quite good where there are data.
Another comment is that the largest survival time looks like an outlier, but it is absolutely ok for a log-normal model.

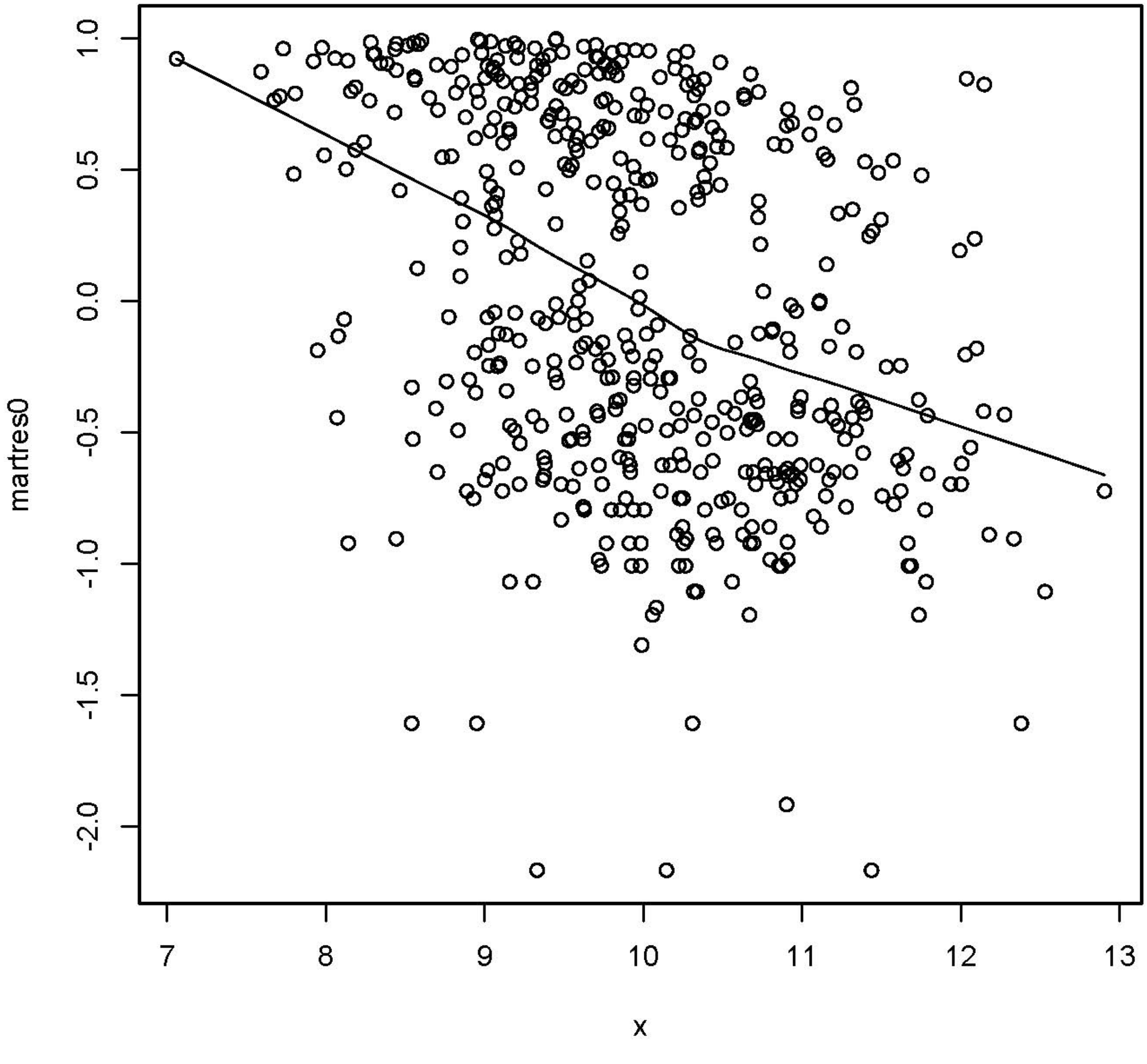
```
> # Look at that observation
> id[Time>30]
[1] 419
> c(x[419],Time[419])
[1] 11.43622 37.18183
> hist(x)
```



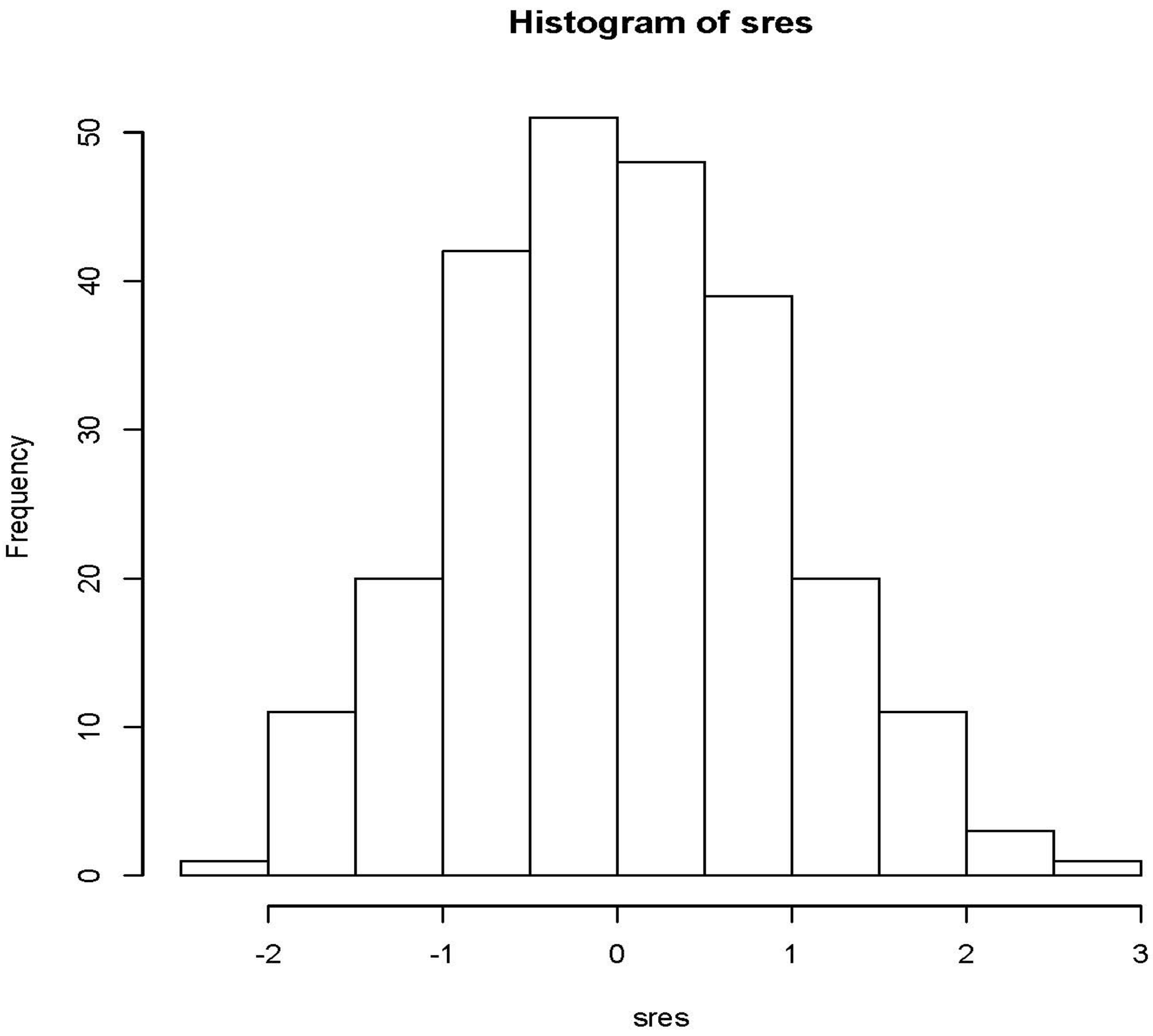
```
> # Look at martingale residuals
>
> martres = residuals(phmodel,type='martingale')
> plot(x,martres); smooth = lowess(x,martres); lines(smooth)
```



```
> # Look at residuals for a model with no expl vars (recommended)
> ph0 = coxph(Surv(Time,delta) ~ 1)
> martres0 = residuals(ph0,type='martingale')
> plot(x,martres0); smooth = lowess(x,martres0); lines(smooth)
```

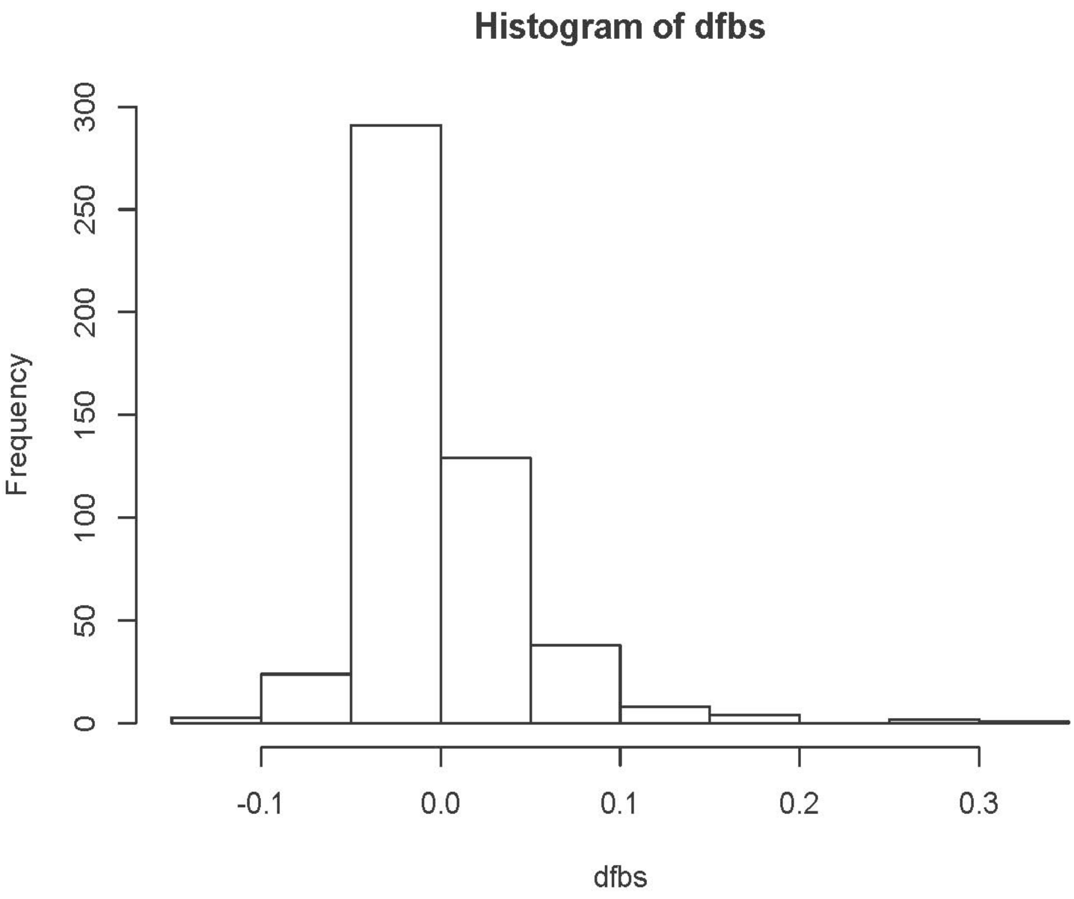


```
> # Look at Schoenfeld residuals
> sres = residuals(phmodel, type = 'schoenfeld')
> hist(sres)
```



They look beautiful; they are supposed to be normal. No outliers.

```
> # Look at bfbetas (beta-hat with one left out, standardized)
> dfbs = residuals(phmodel, type = 'dfbetas')
> summary(dfbs)

     Min.   1st Qu.    Median      Mean   3rd Qu.      Max. 
-0.104323 -0.026239 -0.007703  0.000000  0.013838  0.323013 
>  hist(dfbs)
```



```
> q = id[dfbs>0.25]; q
[1] 215 334 414
```

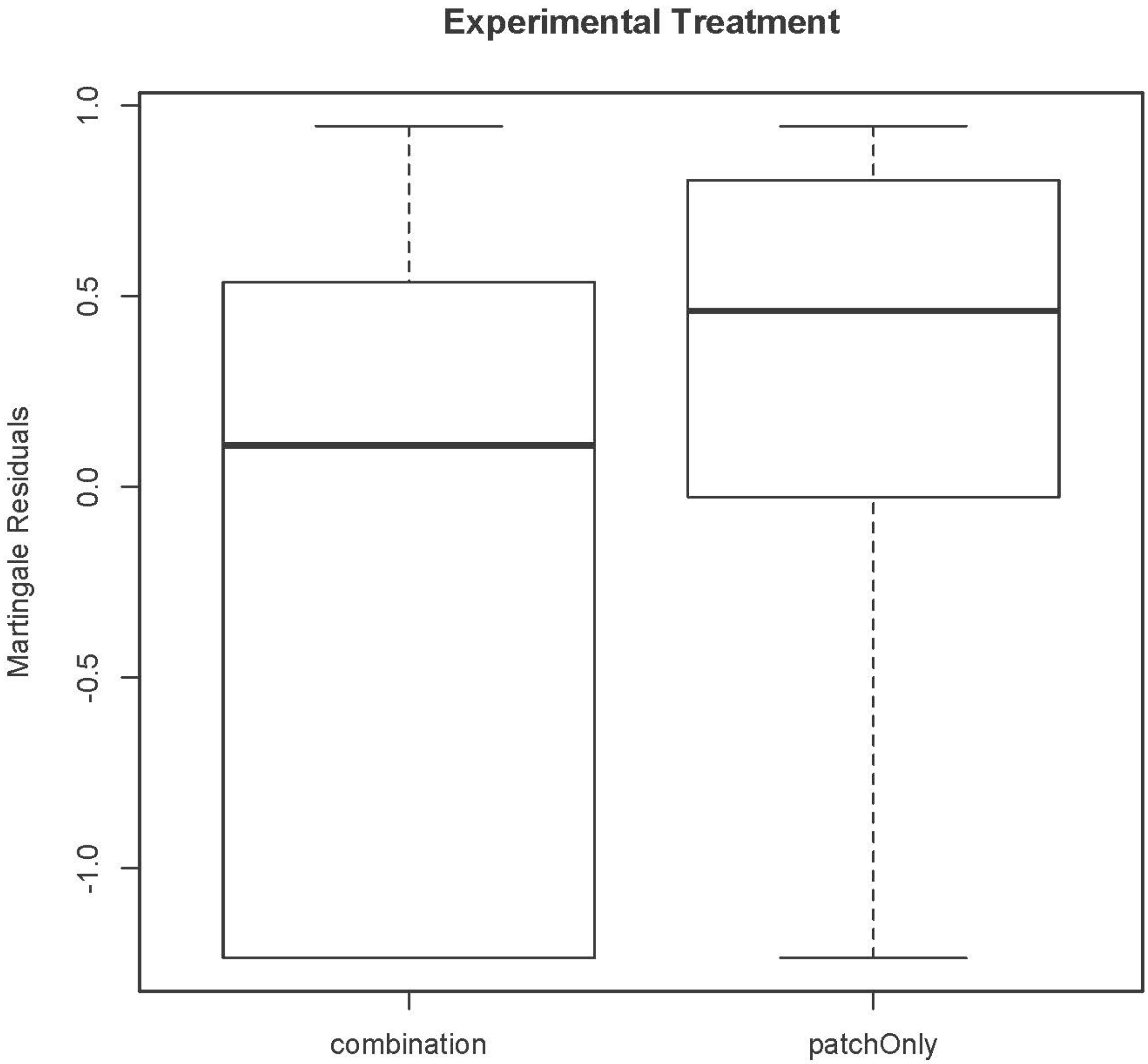
Does not include id = 419

Real Data (pharmacoSmoking)

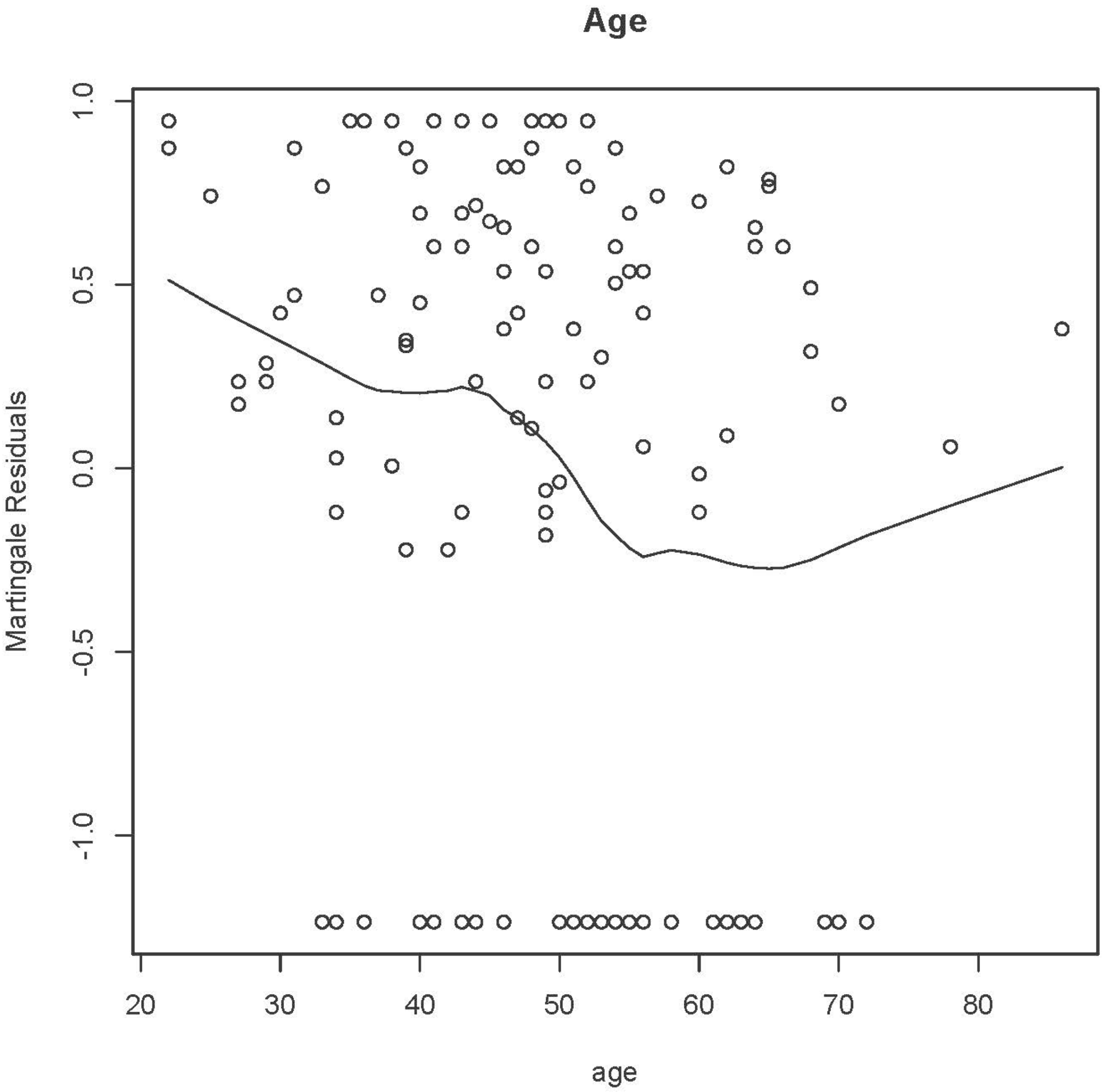
```
> rm(list=ls()); options(scipen=999)
> # install.packages("survival",dependencies=TRUE) # Only need to do this once
> library(survival) # Do this every time
> # install.packages("asaur",dependencies=TRUE) # Only need to do this once
> library(asaur)
>
> # help(pharmacoSmoking)
> summary(pharmacoSmoking)
       id              ttr            relapse              grp    
 Min.   :  1.00   Min.   :  0.00   Min.   :0.000   combination:61  
 1st Qu.: 33.00   1st Qu.:  8.00   1st Qu.:0.000   patchOnly  :64  
 Median : 67.00   Median : 49.00   Median :1.000                   
 Mean   : 66.15   Mean   : 77.44   Mean   :0.712                   
 3rd Qu.: 99.00   3rd Qu.:182.00   3rd Qu.:1.000                   
 Max.   :130.00   Max.   :182.00   Max.   :1.000                   
      age           gender          race     employment  yearsSmoking  
 Min.   :22.00   Female:81   black   :38   ft   :72    Min.   : 9.00  
 1st Qu.:41.00   Male  :44   hispanic: 8   other:39    1st Qu.:22.00  
 Median :49.00               other   : 2   pt   :14    Median :30.00  
 Mean   :48.84               white   :77               Mean   :30.88  
 3rd Qu.:56.00                                         3rd Qu.:39.00  
 Max.   :86.00                                         Max.   :56.00  
 levelSmoking ageGroup2  ageGroup4  priorAttempts     longestNoSmoke
 heavy:89     21-49:66   21-34:16   Min.   :   0.00   Min.   :   0.0  
 light:36     50+  :59   35-49:50   1st Qu.:   1.00   1st Qu.:   7.0  
                         50-64:48   Median :   2.00   Median :  90.0  
                         65+  :11   Mean   :  12.68   Mean   : 539.7  
                                    3rd Qu.:   5.00   3rd Qu.: 365.0  
                                    Max.   :1000.00   Max.   :6205.0  
>
> attach(pharmacoSmoking) # More convenient for exploratory analysis and plotting
>
> # Make an indicator dummy variable for combination therapy: Reference is patch
only
> n = length(grp); combo = numeric(n);
> combo[grp=='combination'] = 1; rm(n)
> DayOfRelapse = Surv(ttr+1,relapse) # Day of relapse starts with one.
> # Collapse race categories
> Race = as.character(race) # Small r race is a factor. This is easier to modify.
> Race[Race!='white'] = 'blackOther'; Race=factor(Race)
> race = Race
>
> # Exploratory strategy: Fit a model with no explanatory variables, and plot the
martingale residuals against potential explanatory variables.
> # A smooth curve through the points really helps. "lowess" stands for locally
weighted scatterplot smoothing.
>
> # Earlier, we settled on a model with treatment group, age and employment status.
>
> model0 = coxph(DayOfRelapse ~ 1)
> martres0 = residuals(model0,type='martingale')
>
```

A value of martingale residuals near 1 represents individuals that “died too soon”, and large negative values correspond to individuals that “lived too long”.

```
> # Fig 1
> plot(grp,martres0,ylab='Martingale Residuals')
> title('Experimental Treatment')
```

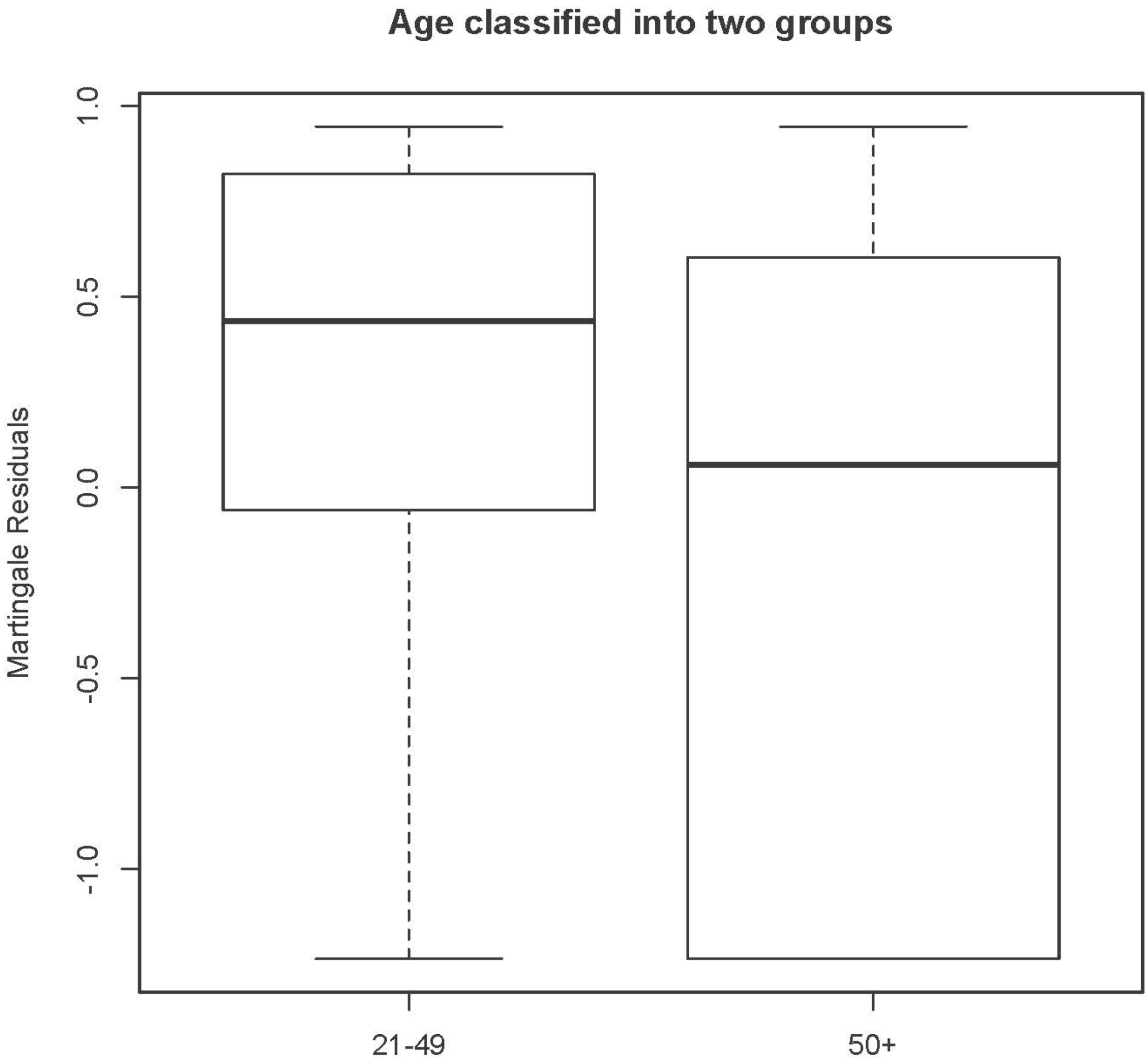


```
>
> # Fig 2
> plot(age,martres0,ylab='Martingale Residuals',main='Age')
> lines(lowess(age,martres0)) # Plots a smooth curve
>
```

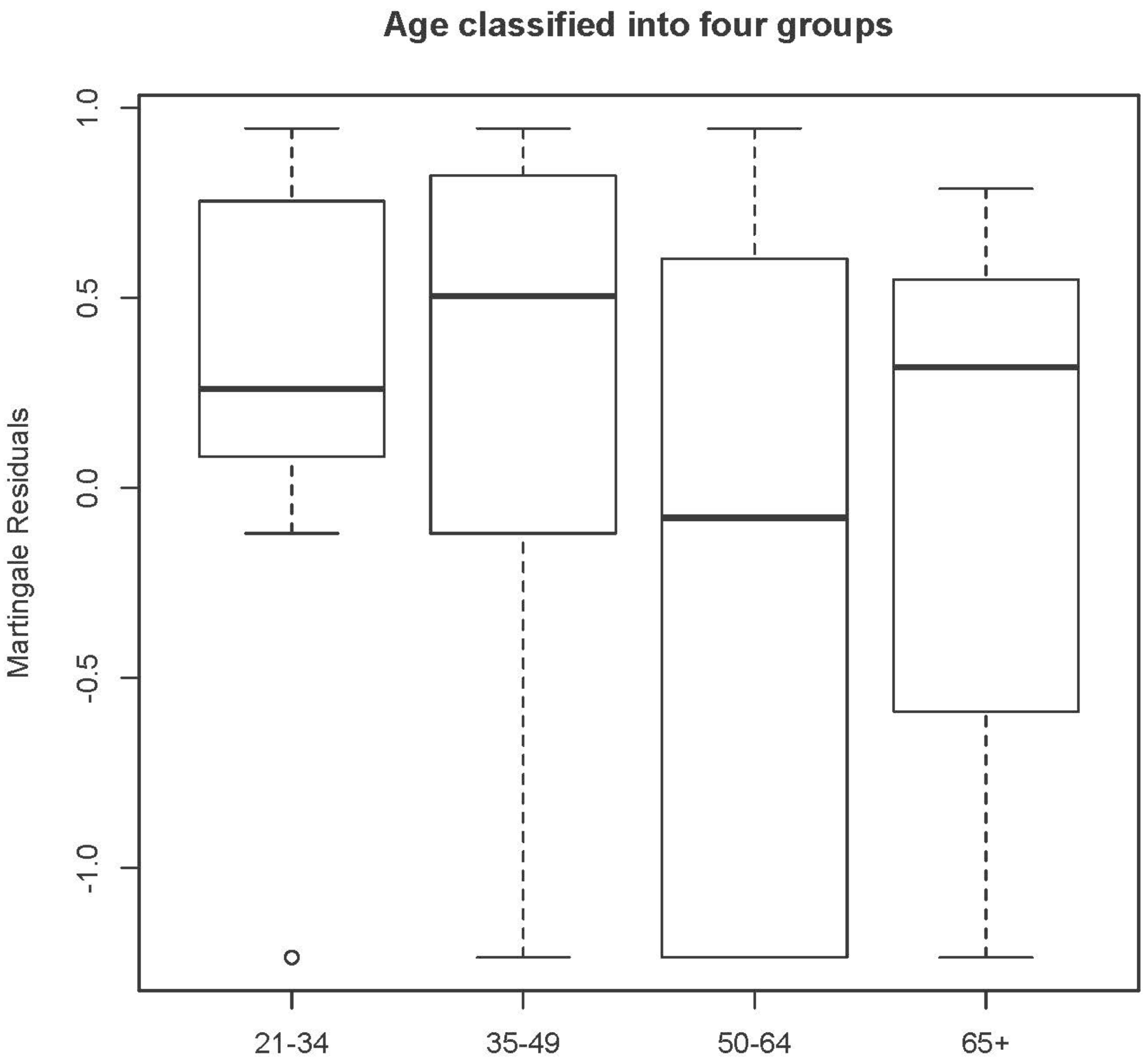


Now maybe we see why the data frame has categorical versions of age.

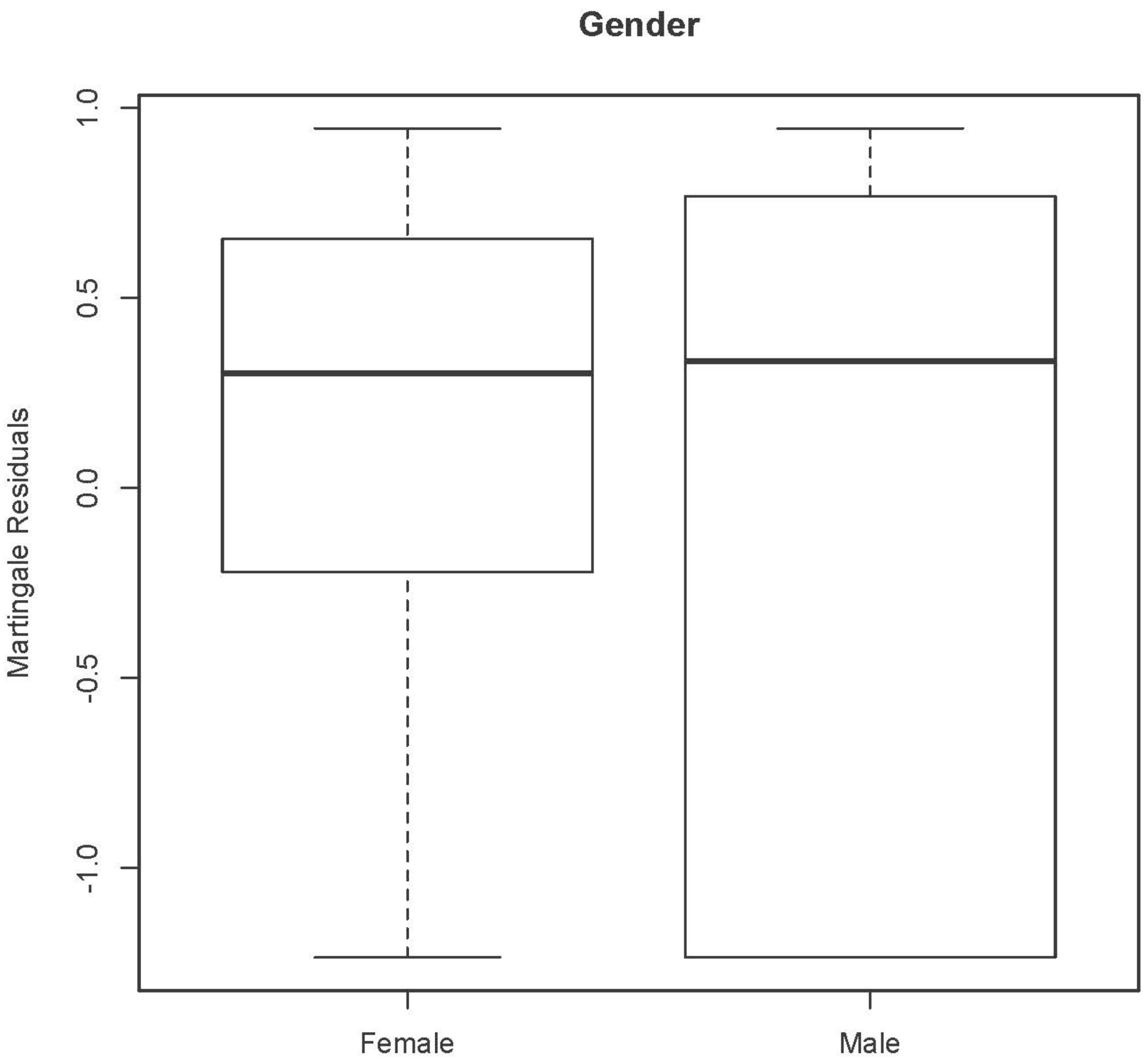
```
> # Fig 3
> plot(ageGroup2,martres0,ylab='Martingale Residuals')
> title('Age classified into two groups')
>
```



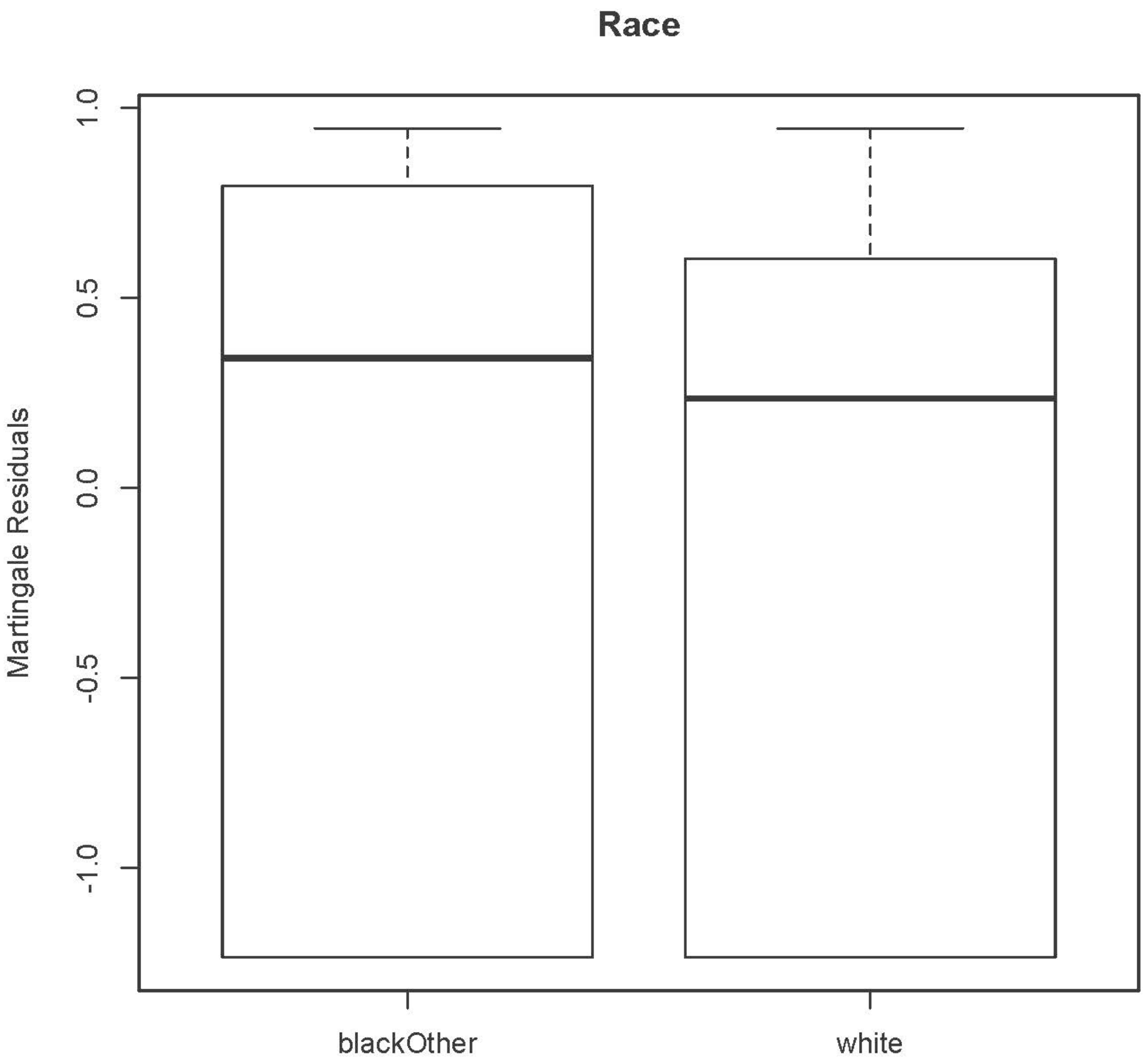
```
>
> # Fig 4
> plot(ageGroup4,martres0,ylab='Martingale Residuals')
> title('Age classified into four groups')
>
```



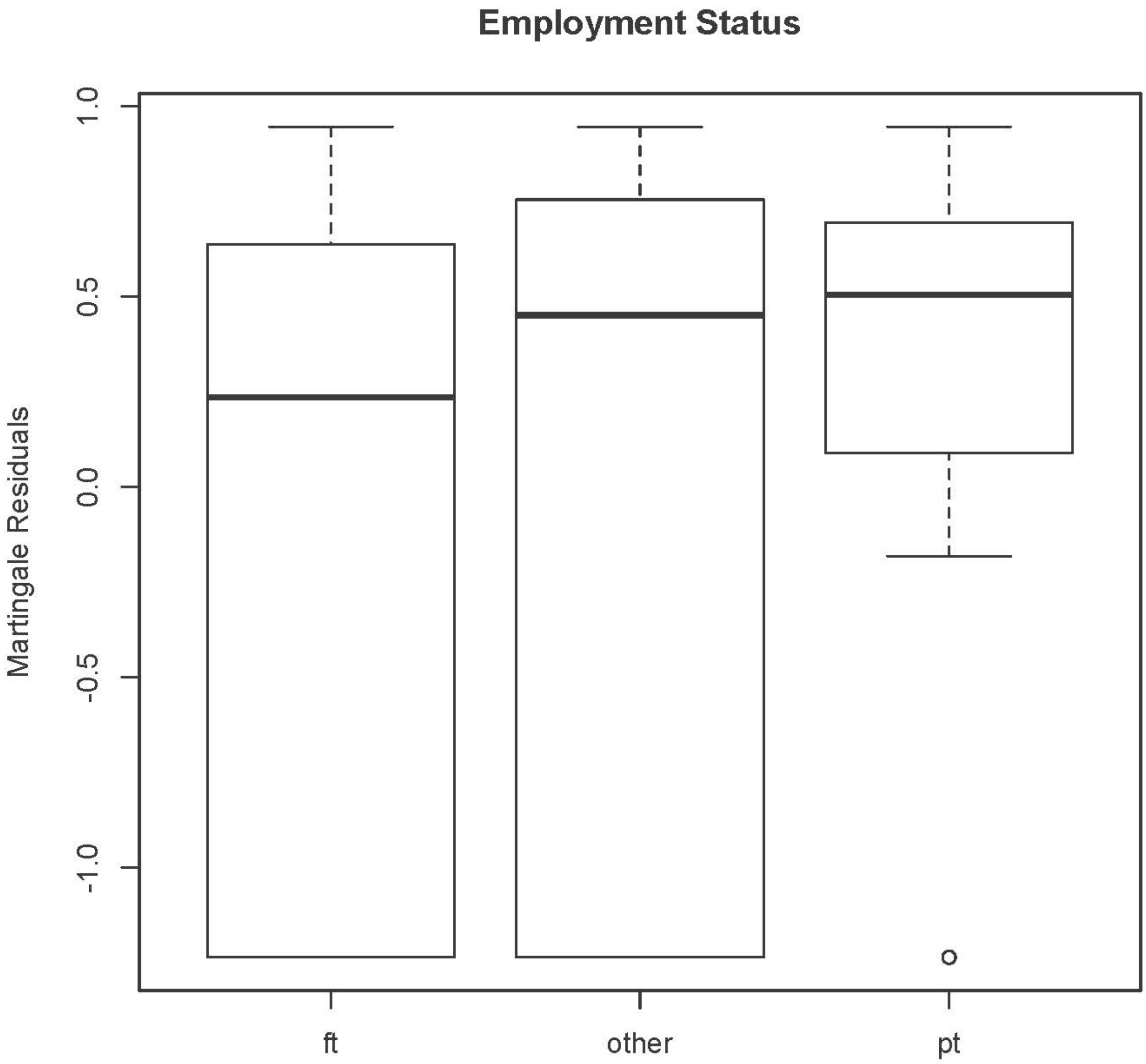
```
>
> # Fig 5
> plot(gender,martres0,ylab='Martingale Residuals')
> title('Gender')
>
```



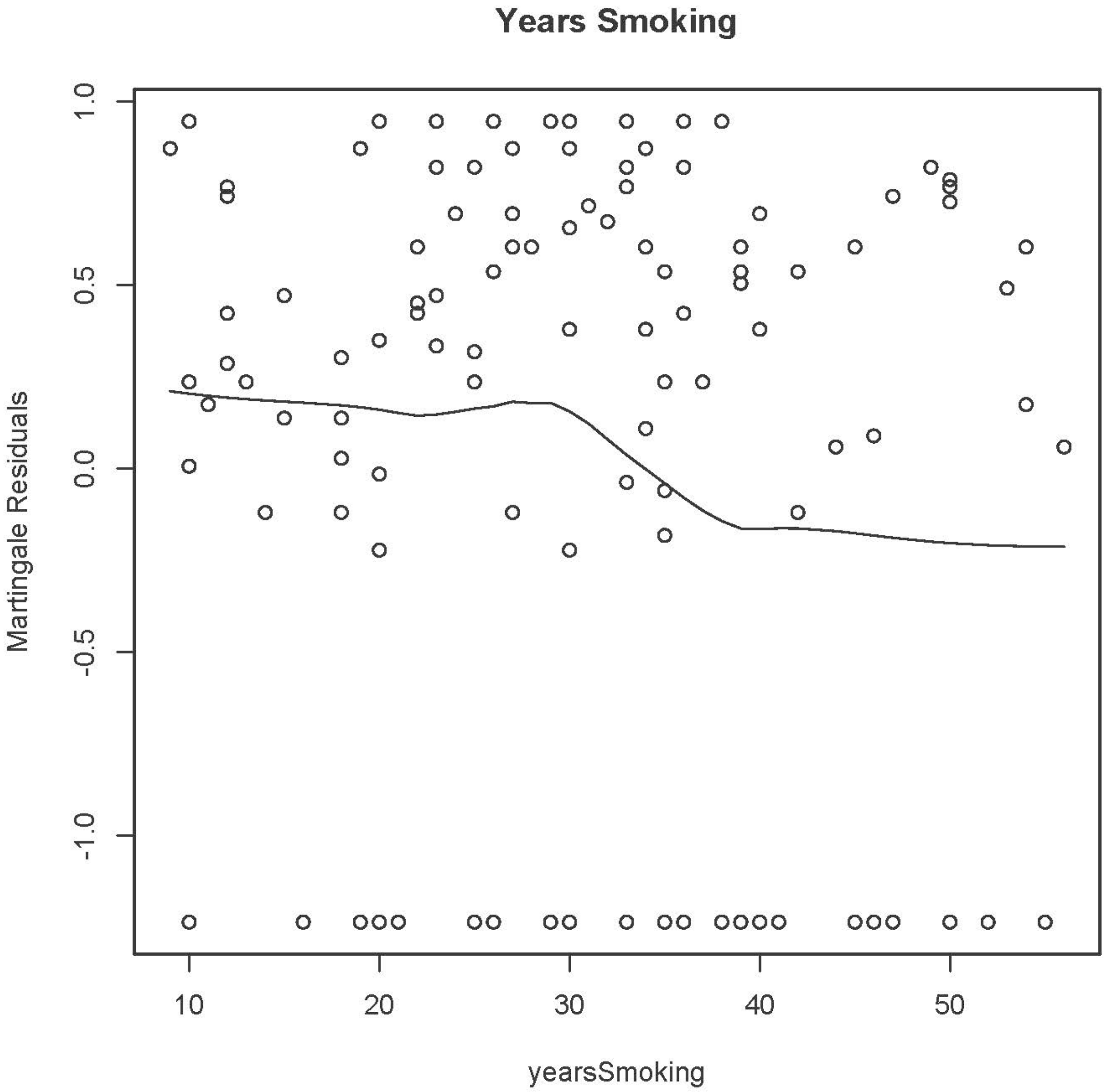
```
>
> # Fig 6
> plot(race,martres0,ylab='Martingale Residuals')
> title('Race')
>
```



```
>
> # Fig 7
> plot(employment,martres0,ylab='Martingale Residuals')
> title('Employment Status')
>
```

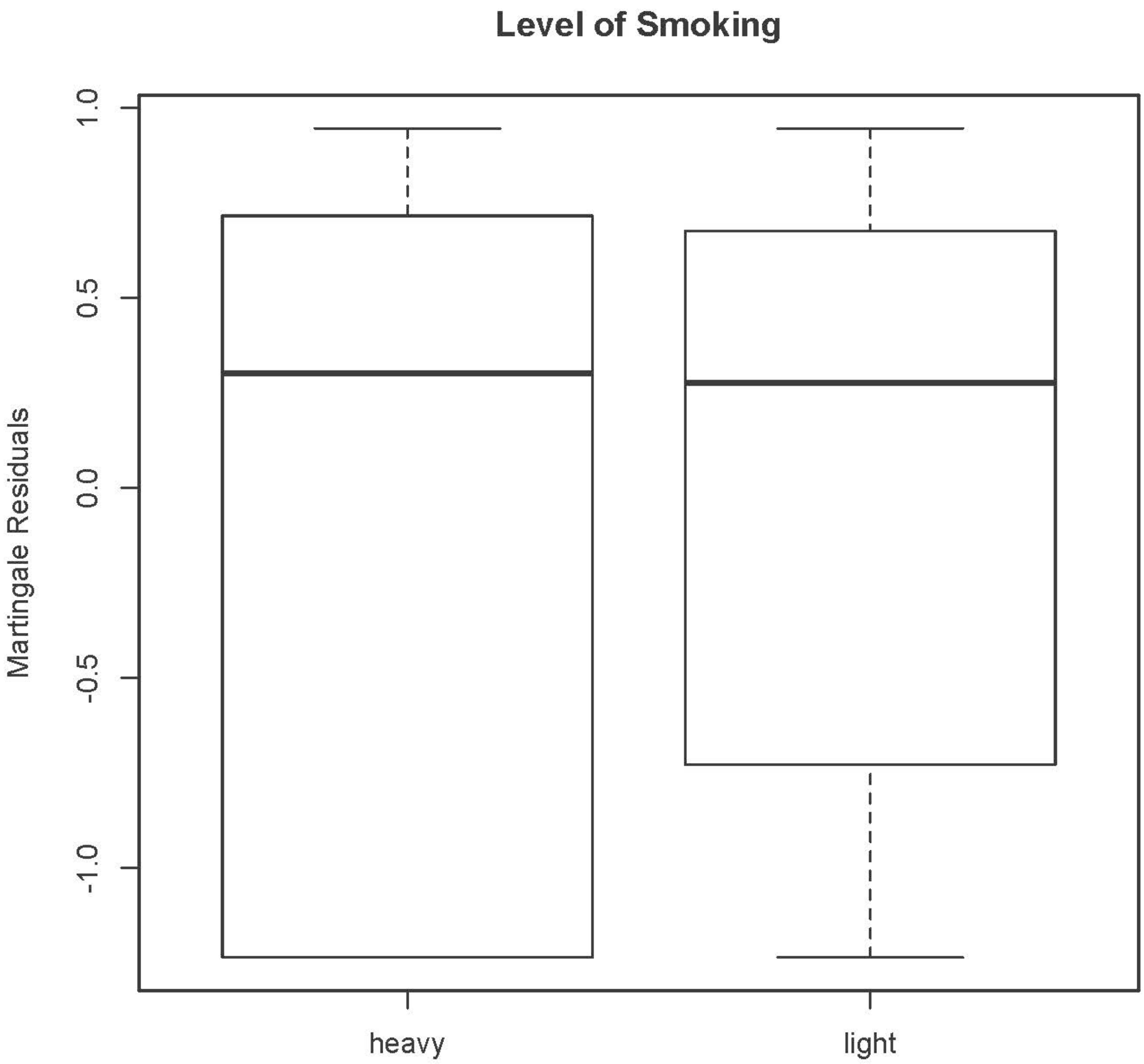


```
>
> # Fig 8
> plot(yearsSmoking,martres0,ylab='Martingale Residuals')
> title('Years Smoking')
> lines(lowess(yearsSmoking,martres0))
>
```

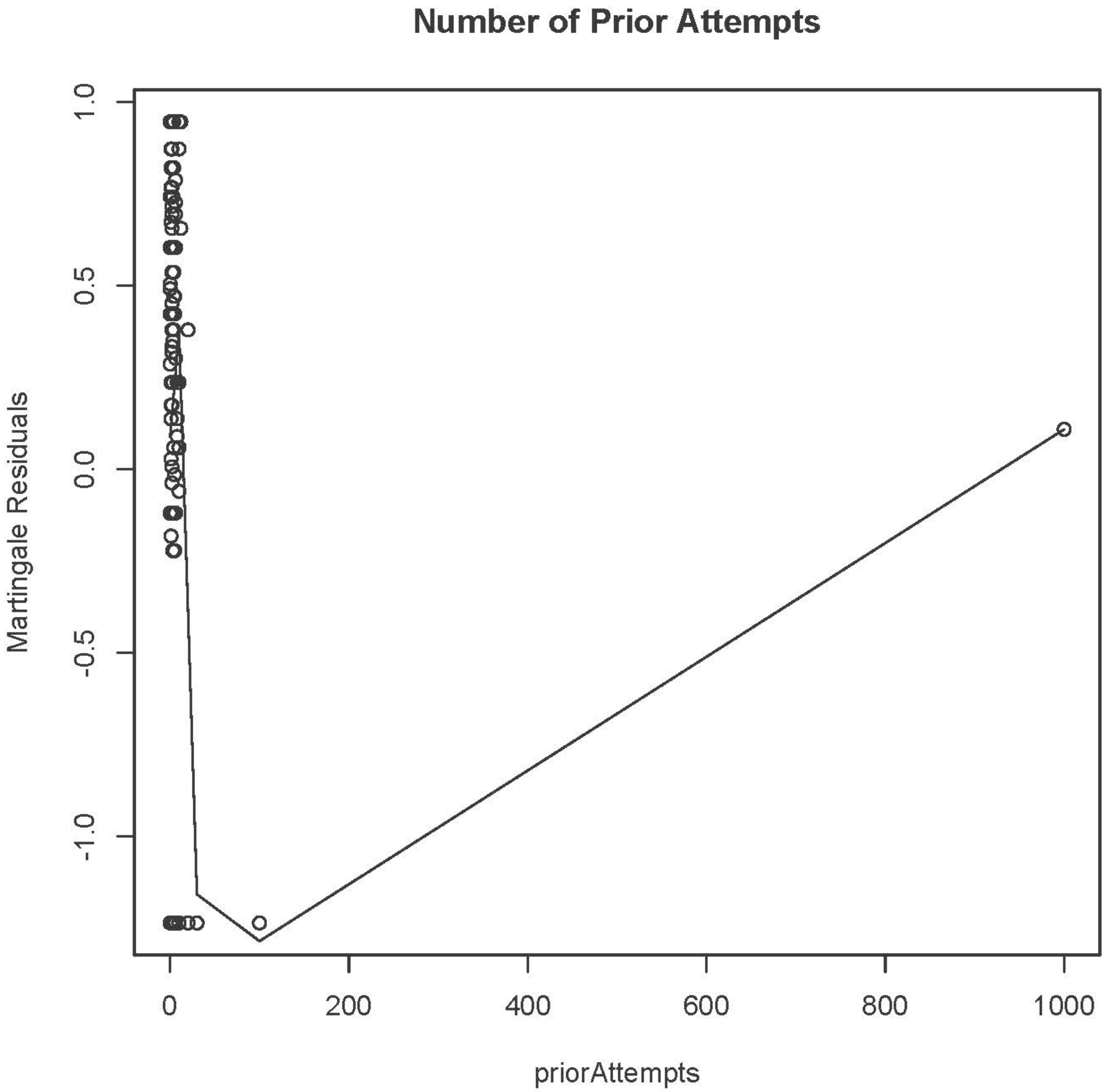


Possible curve -- re-visit. Also it's unclear what this will look like once we control for age.

```
> # Fig 9
> plot(levelSmoking,martres0,ylab='Martingale Residuals')
> title('Level of Smoking')
>
>
```



```
> # Fig 10
> plot(priorAttempts,martres0,ylab='Martingale Residuals')
> title('Number of Prior Attempts')
> lines(lowess(priorAttempts,martres0))
>
```

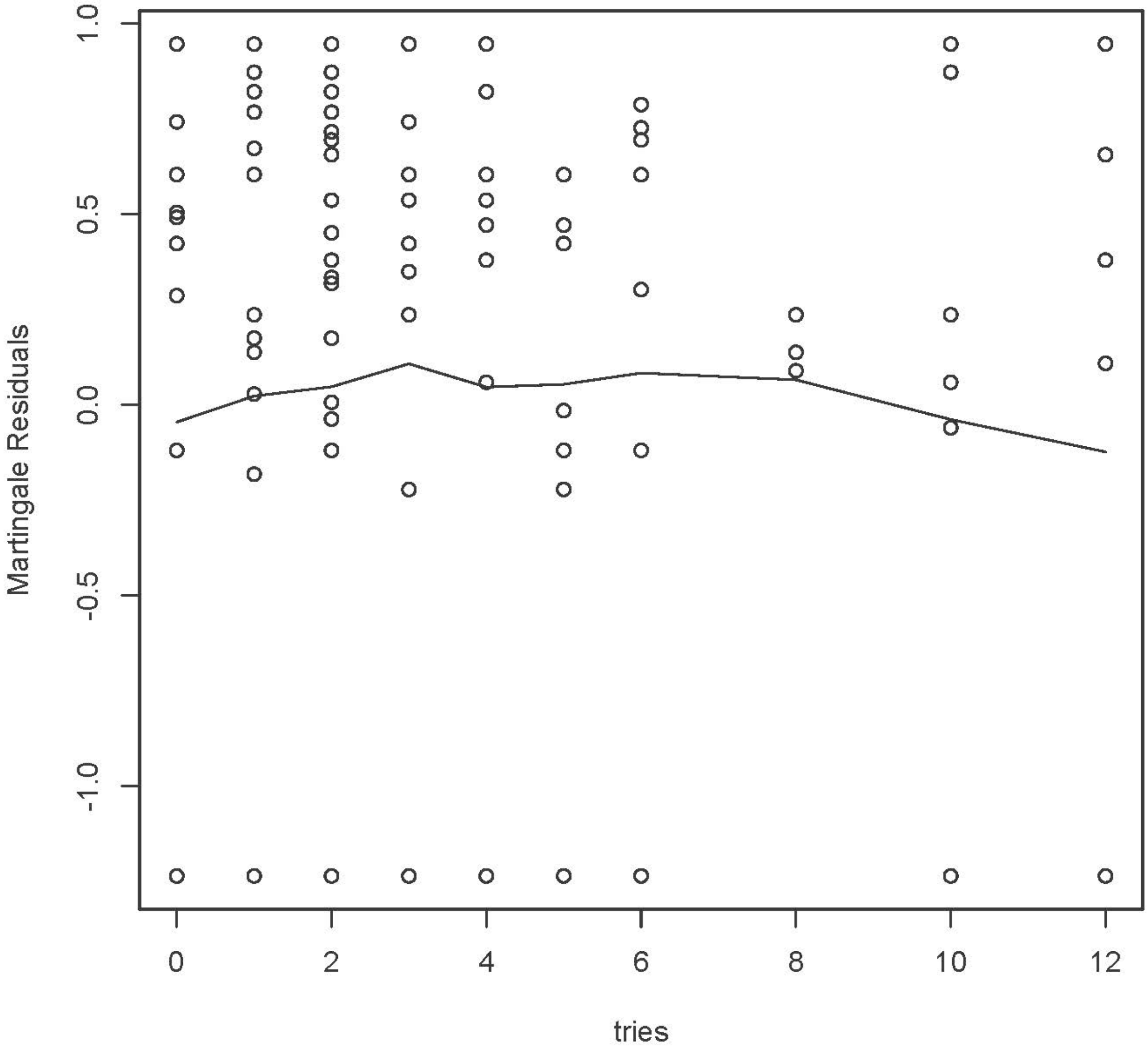


```
>
> id[priorAttempts==1000]
[1] 98
> pharmacoSmoking[98,]

   id ttr relapse         grp age gender  race employment yearsSmoking
98 14 182       0 combination  52 Female white      other           33
   levelSmoking ageGroup2 ageGroup4 priorAttempts longestNoSmoke
98        heavy       50+     50-64             5            270

> id[1:10]
 [1]  21 113  39  80  87  29  16  35  54  70

> loc = 1:length(age)
> loc[priorAttempts==1000]
[1] 105
> pharmacoSmoking[105,]

    id ttr relapse         grp age gender  race employment yearsSmoking
105 98  65       1 combination  48 Female white         ft           34
    levelSmoking ageGroup2 ageGroup4 priorAttempts longestNoSmoke
105        heavy     21-49     35-49          1000            548
```

Okay, it's believable.

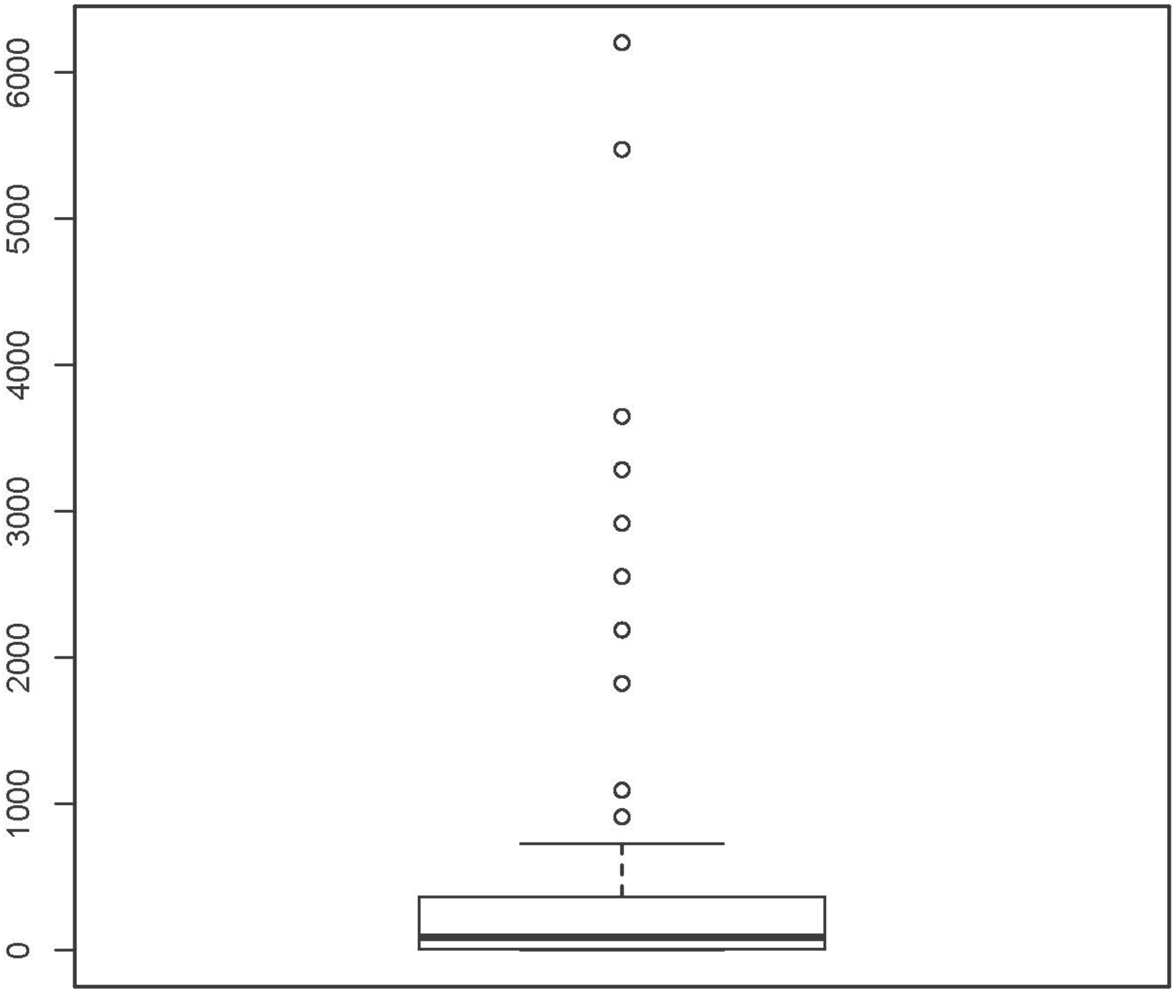
```
>
> sort(priorAttempts)
  [1]    0    0    0    0    0    0    0    0    0    0    0    1    1    1    1
 [16]    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
 [31]    1    1    1    1    1    1    1    1    2    2    2    2    2    2    2
 [46]    2    2    2    2    2    2    2    2    2    2    2    2    2    2    2
 [61]    2    2    2    3    3    3    3    3    3    3    3    3    3    3    3
 [76]    3    3    3    4    4    4    4    4    4    4    4    4    4    4    5
 [91]    5    5    5    5    5    5    5    5    6    6    6    6    6    6    6
[106]    6    8    8    8    8   10   10   10   10   10   10   10   12   12   20
[121]   20   30   30  100 1000

>
> # Recode the outliers and take another look
> tries = priorAttempts
> tries[tries>12] = 12
>
```

```
> # Fig 11
> plot(tries,martres0,ylab='Martingale Residuals')
> lines(lowess(tries,martres0))
>
```



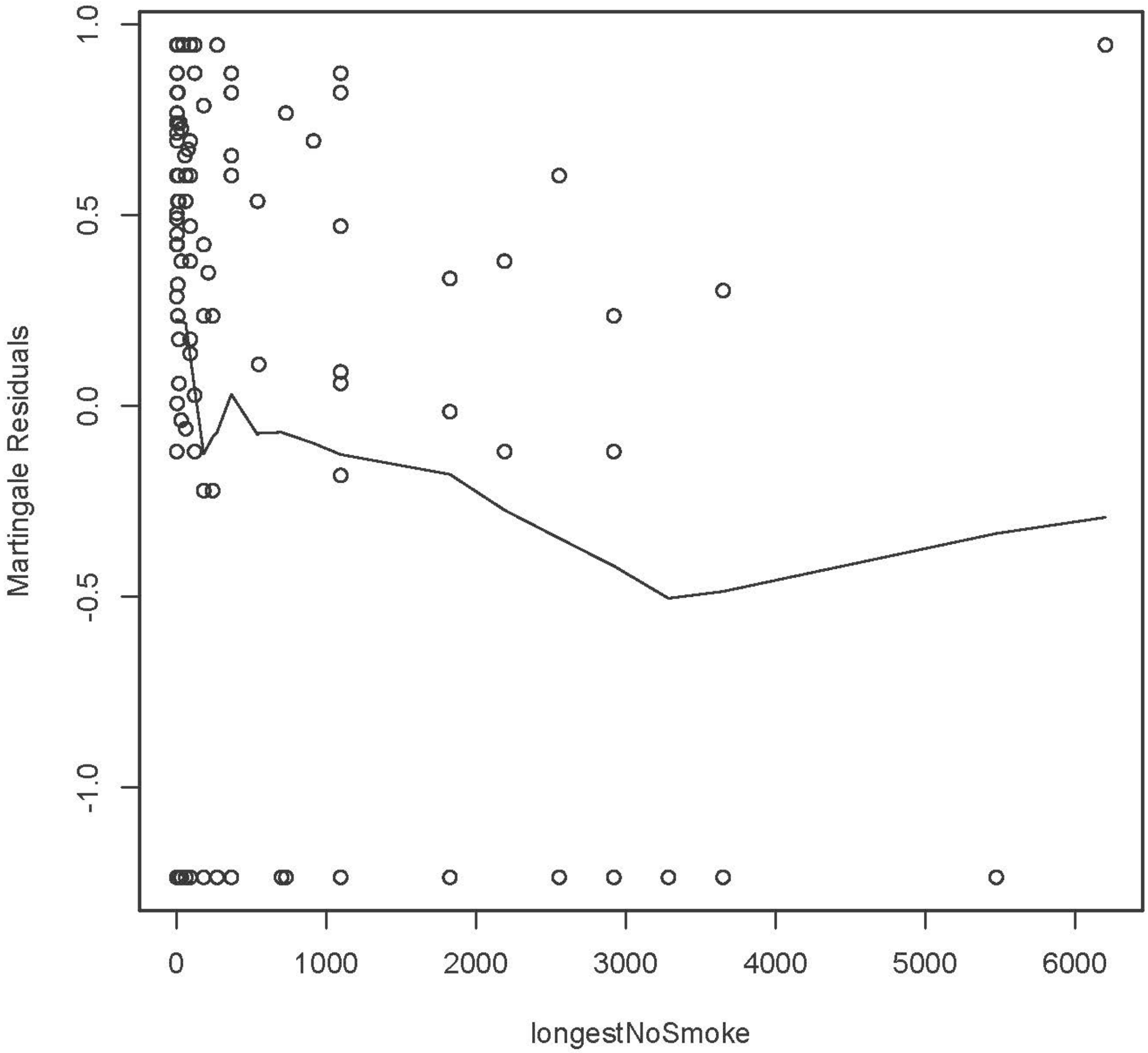
Looks like nothing to me.

```
> # Fig 12
> boxplot(longestNoSmoke)
```

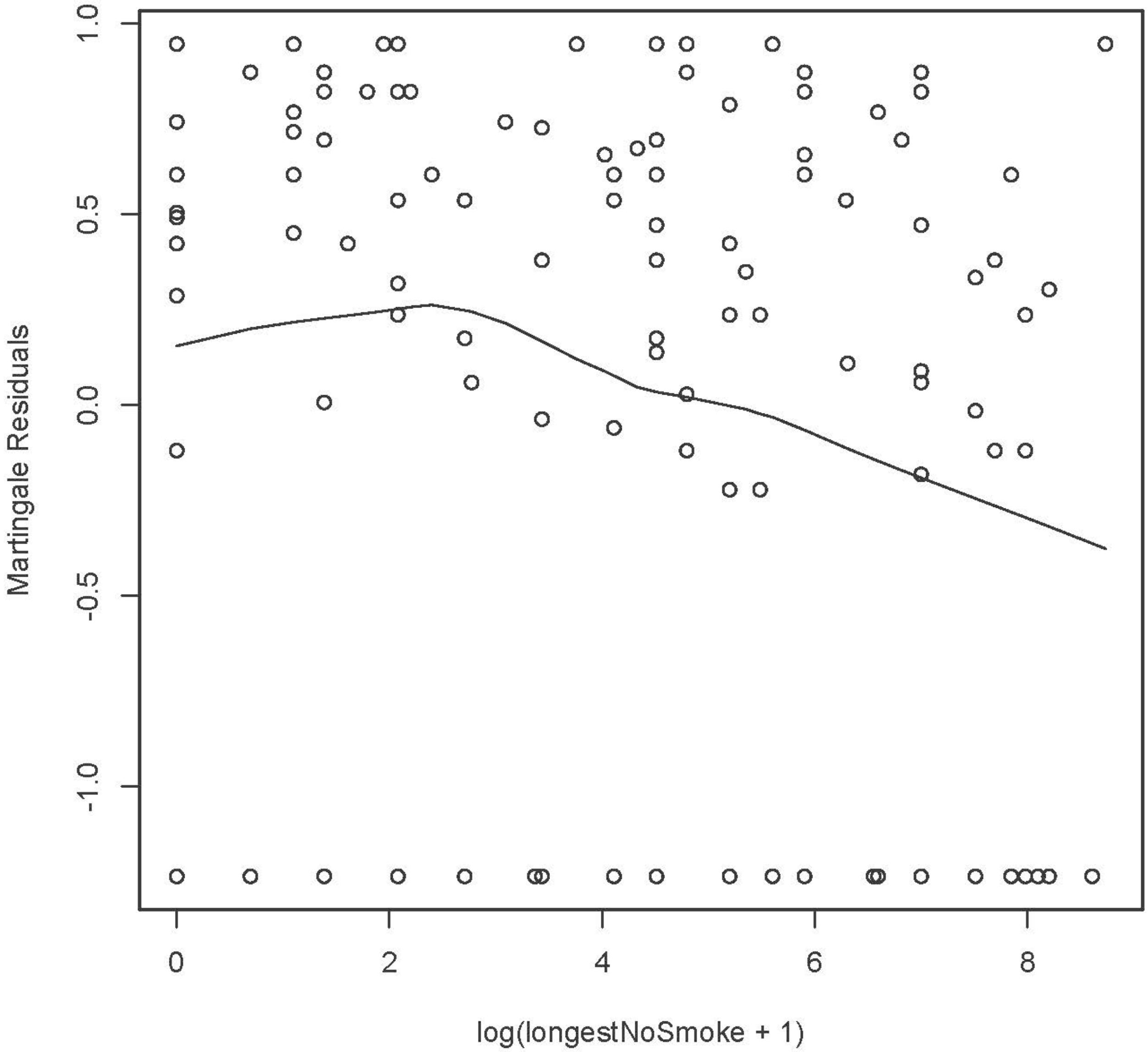


```
> sort(longestNoSmoke)
  [1]    0    0    0    0    0    0    0    0    0    0    0    0    1    1    1
 [16]    2    2    2    2    2    2    3    3    3    3    3    4    5    6    7
 [31]    7    7    7    7    7    7    7    7    7    8   10   14   14   14   15
 [46]   21   28   28   30   30   30   30   42   55   60   60   60   60   60   75
 [61]   90   90   90   90   90   90   90   90   90   90   90  120  120  120  120
 [76]  180  180  180  180  180  180  180  180  210  240  240  270  270  270  365
 [91]  365  365  365  365  365  540  548  700  730  730  913 1095 1095 1095 1095
[106] 1095 1095 1095 1095 1095 1825 1825 1825 2190 2190 2555 2555 2920 2920 2920
[121] 3285 3650 3650 5475 6205
>
```

```
>
> # Fig 13
> plot(longestNoSmoke,martres0,ylab='Martingale Residuals')
> lines(lowess(longestNoSmoke,martres0))
>
```



```
>
> # Fig 14
> plot(log(longestNoSmoke+1),martres0,ylab='Martingale Residuals')
> lines(lowess(log(longestNoSmoke+1),martres0))
>
>
```



It does not look like much, but try the log version later.

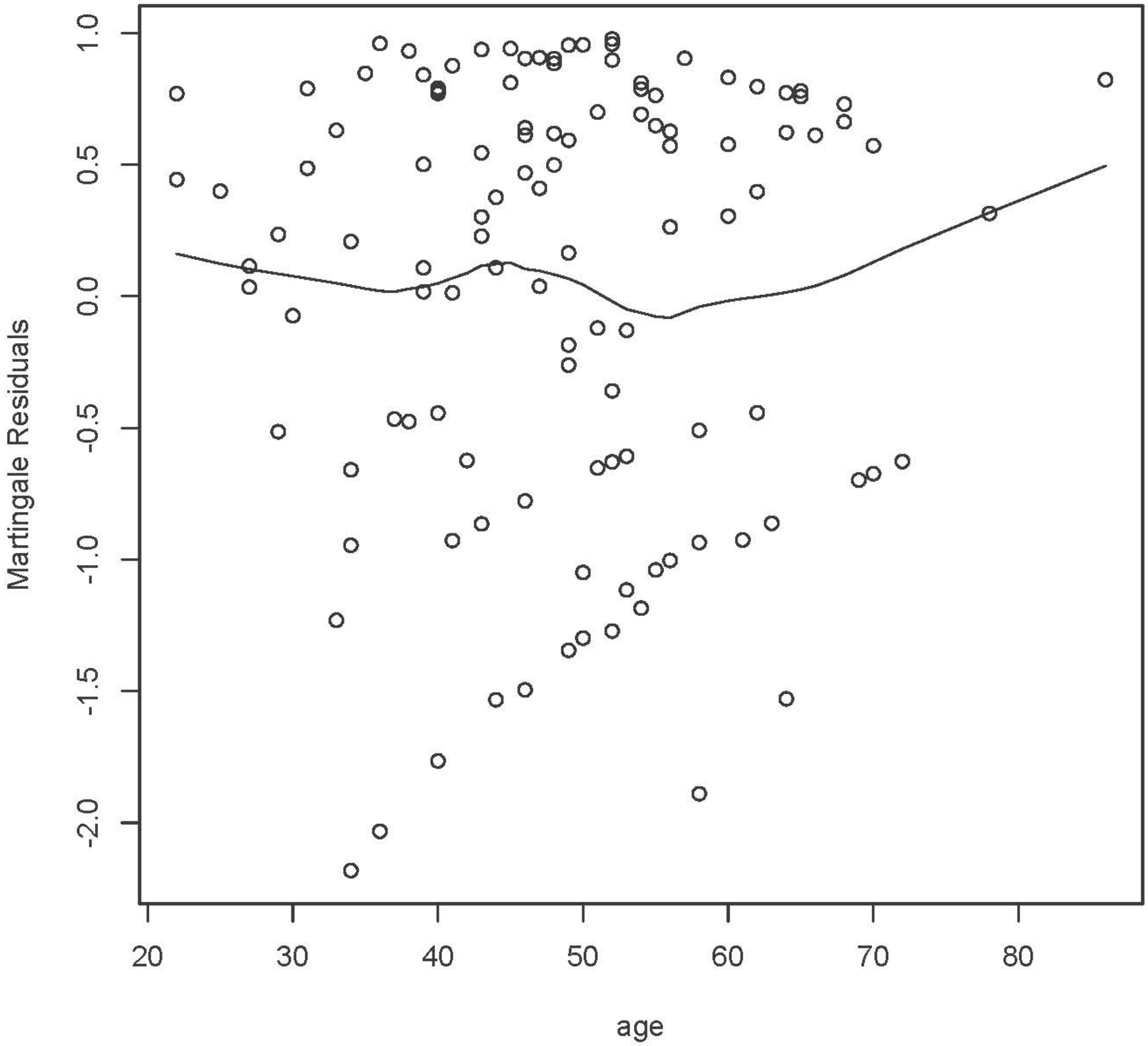
```
> # Plots suggest treatment group, age and employment status, with a possible
> # curve for age.
> # Check this first, and then possible curves for yearsSmoking and longestNoSmoke.
>
> agesq = age^2
> model1 = coxph(DayOfRelapse ~ combo + age + agesq + employment); summary(model1)
Call:
coxph(formula = DayOfRelapse ~ combo + age + agesq + employment)

  n= 125, number of events= 89

                      coef  exp(coef)   se(coef)      z Pr(>|z|)
combo           -0.6206075  0.5376177  0.2188288 -2.836  0.00457 **
age             -0.1001902  0.9046654  0.0549849 -1.822  0.06843 .
agesq            0.0006729  1.0006732  0.0005572  1.208  0.22713
employmentother  0.6800741  1.9740240  0.2754600  2.469  0.01355 *
employmentpt     0.6757762  1.9655581  0.3278821  2.061  0.03930 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                exp(coef) exp(-coef) lower .95 upper .95
combo              0.5376     1.8601    0.3501    0.8255
age                0.9047     1.1054    0.8122    1.0076
agesq              1.0007     0.9993    0.9996    1.0018
employmentother    1.9740     0.5066    1.1505    3.3871
employmentpt       1.9656     0.5088    1.0337    3.7375

Concordance= 0.633  (se = 0.034 )
Rsquare= 0.17   (max possible= 0.998 )
Likelihood ratio test= 23.36  on 5 df,   p=0.0002886
Wald test            = 24.19  on 5 df,   p=0.0001995
Score (logrank) test = 24.68  on 5 df,   p=0.0001605
```

Age and age-squared are highly correlated and may be washing each other out.

```
> cage = age-mean(age); cagesq = cage^2
> model2 = coxph(DayOfRelapse ~ combo + cage + cagesq + employment)
> summary(model2)
Call:
coxph(formula = DayOfRelapse ~ combo + cage + cagesq + employment)

  n= 125, number of events= 89

                      coef  exp(coef)   se(coef)      z Pr(>|z|)
combo           -0.6206075  0.5376177  0.2188288 -2.836 0.004568 **
cage            -0.0344582  0.9661287  0.0101552 -3.393 0.000691 ***
cagesq           0.0006729  1.0006732  0.0005572  1.208 0.227128
employmentother  0.6800741  1.9740240  0.2754600  2.469 0.013554 *
employmentpt     0.6757762  1.9655581  0.3278821  2.061 0.039300 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                exp(coef) exp(-coef) lower .95 upper .95
combo              0.5376     1.8601    0.3501    0.8255
cage               0.9661     1.0351    0.9471    0.9856
cagesq             1.0007     0.9993    0.9996    1.0018
employmentother    1.9740     0.5066    1.1505    3.3871
employmentpt       1.9656     0.5088    1.0337    3.7375

Concordance= 0.633  (se = 0.034 )
Rsquare= 0.17   (max possible= 0.998 )
Likelihood ratio test= 23.36  on 5 df,   p=0.0002886
Wald test            = 24.19  on 5 df,   p=0.0001995
Score (logrank) test = 24.68  on 5 df,   p=0.0001605
```

```
>
> # We do not have good evidence of departure from a straight-line relationship.
> # Drop the quadratic term. This is the model from past lectures.
>
> model3 = coxph(DayOfRelapse ~ combo + age + employment); summary(model3)

Call:
coxph(formula = DayOfRelapse ~ combo + age + employment)

  n= 125, number of events= 89

                    coef exp(coef) se(coef)      z Pr(>|z|)
combo           -0.60788   0.54450  0.21837 -2.784  0.00537 **
age             -0.03529   0.96533  0.01075 -3.282  0.00103 **
employmentother  0.70348   2.02077  0.26929  2.612  0.00899 **
employmentpt     0.65369   1.92262  0.32732  1.997  0.04581 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                exp(coef) exp(-coef) lower .95 upper .95
combo              0.5445     1.8365    0.3549    0.8354
age                0.9653     1.0359    0.9452    0.9859
employmentother    2.0208     0.4949    1.1920    3.4256
employmentpt       1.9226     0.5201    1.0122    3.6518

Concordance= 0.638  (se = 0.034 )
Rsquare= 0.162   (max possible= 0.998 )
Likelihood ratio test= 22.03  on 4 df,   p=0.0001979
Wald test            = 21.91  on 4 df,   p=0.0002084
Score (logrank) test = 22.48  on 4 df,   p=0.0001608
```

```
>
> # Martingale residual plot too
>
> # Fig 15
> martres3 = residuals(model3,type='martingale')
> plot(age,martres3,ylab='Martingale Residuals')
> lines(lowess(age,martres3))
```

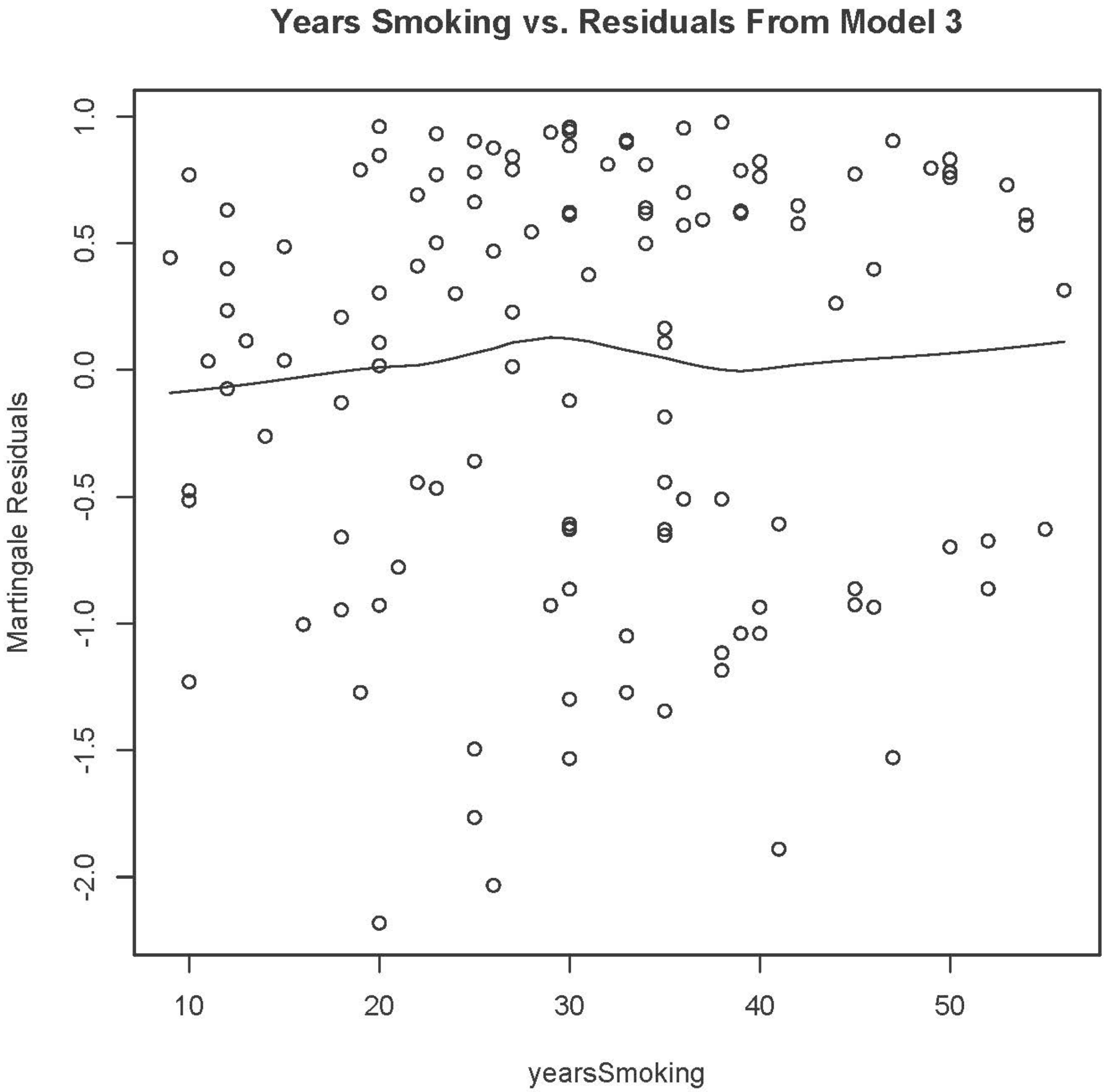


Does not look like much either.

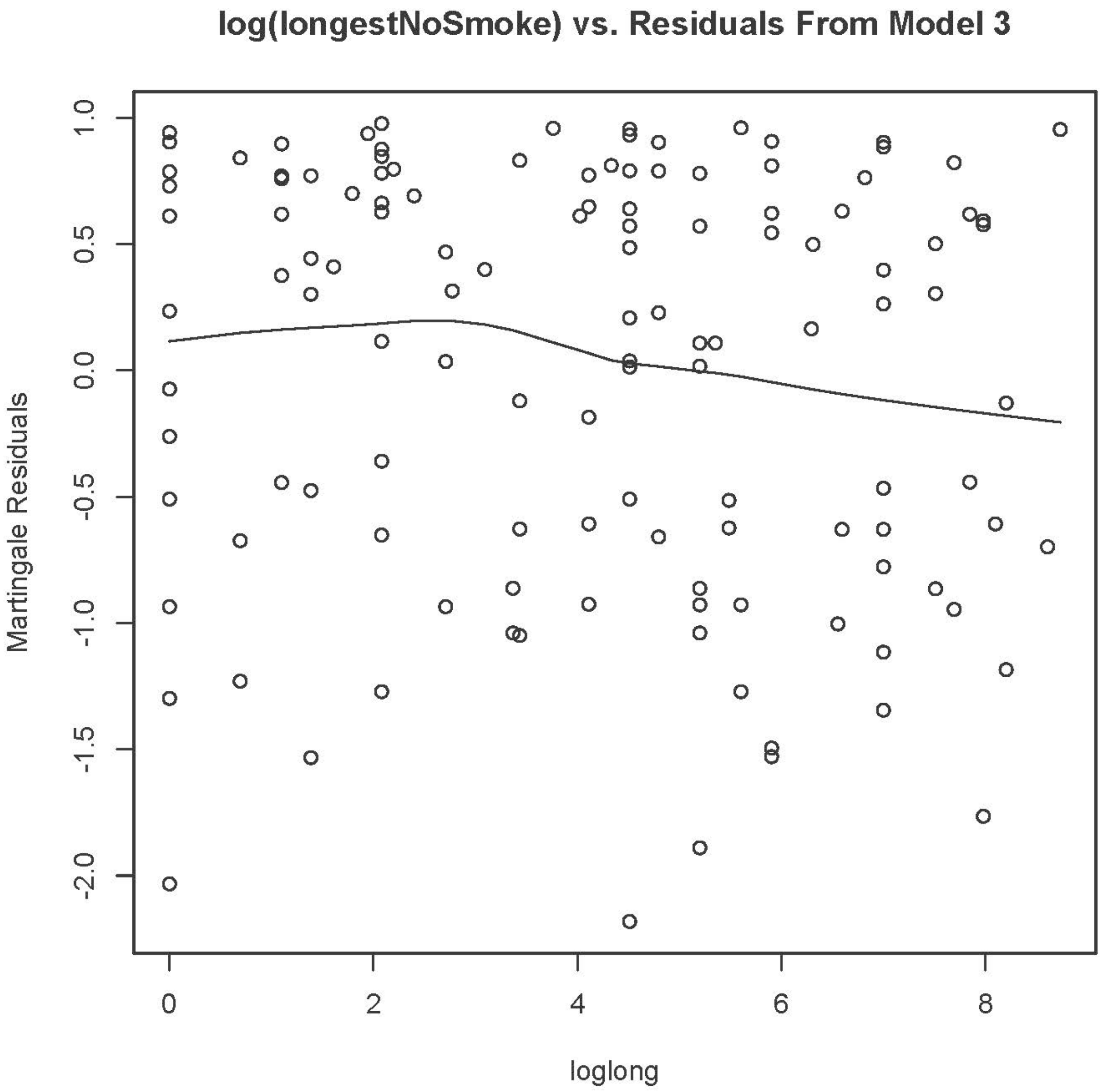
```
>
> # Try 4-category age. Based on the plot, make 50-64 the reference category.
>
> table(ageGroup4)
ageGroup4
21-34 35-49 50-64   65+
   16    50    48    11
>
> agecat = ageGroup4; contrasts(agecat) = contr.treatment(4,base=3)
> colnames(contrasts(agecat)) = c('21-34', '35-49', '65+')
> contrasts(agecat)
      21-34 35-49 65+
21-34     1     0   0
35-49     0     1   0
50-64     0     0   0
65+       0     0   1
>
> model4 = coxph(DayOfRelapse ~ combo + agecat + employment); summary(model4)
Call:
coxph(formula = DayOfRelapse ~ combo + agecat + employment)

  n= 125, number of events= 89

                   coef exp(coef) se(coef)      z Pr(>|z|)
combo           -0.6564    0.5187   0.2198 -2.986 0.002831 **
agecat21-34      1.0233    2.7825   0.3597  2.845 0.004437 **
agecat35-49      0.9115    2.4880   0.2637  3.456 0.000548 ***
agecat65+        0.3162    1.3720   0.4216  0.750 0.453141
employmentother  0.6231    1.8648   0.2764  2.254 0.024177 *
employmentpt     0.5214    1.6844   0.3320  1.570 0.116314
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                exp(coef) exp(-coef) lower .95 upper .95
combo              0.5187     1.9278    0.3371    0.7981
agecat21-34        2.7825     0.3594    1.3749    5.6308
agecat35-49        2.4880     0.4019    1.4837    4.1718
agecat65+          1.3720     0.7289    0.6005    3.1345
employmentother    1.8648     0.5363    1.0848    3.2057
employmentpt       1.6844     0.5937    0.8787    3.2289

Concordance= 0.647  (se = 0.034 )
Rsquare= 0.187   (max possible= 0.998 )
Likelihood ratio test= 25.89  on 6 df,   p=0.0002333
Wald test            = 24.59  on 6 df,   p=0.000406
Score (logrank) test = 25.54  on 6 df,   p=0.0002709
```

This supports either 2 categories for age, or just straight line. I am back to model 3.

```
>
> # Look at plots for yearsSmoking and log(longestNoSmoke+1) against martingale
residuals from model 3
>
> # Fig 16
> plot(yearsSmoking,martres3,ylab='Martingale Residuals')
> title('Years Smoking vs. Residuals From Model 3')
> lines(lowess(yearsSmoking,martres3))
```



```
>
> # Fig 17
> loglong = log(longestNoSmoke+1)
> plot(loglong,martres3,ylab='Martingale Residuals')
> title('log(longestNoSmoke) vs. Residuals From Model 3')
> lines(lowess(loglong,martres3))
```



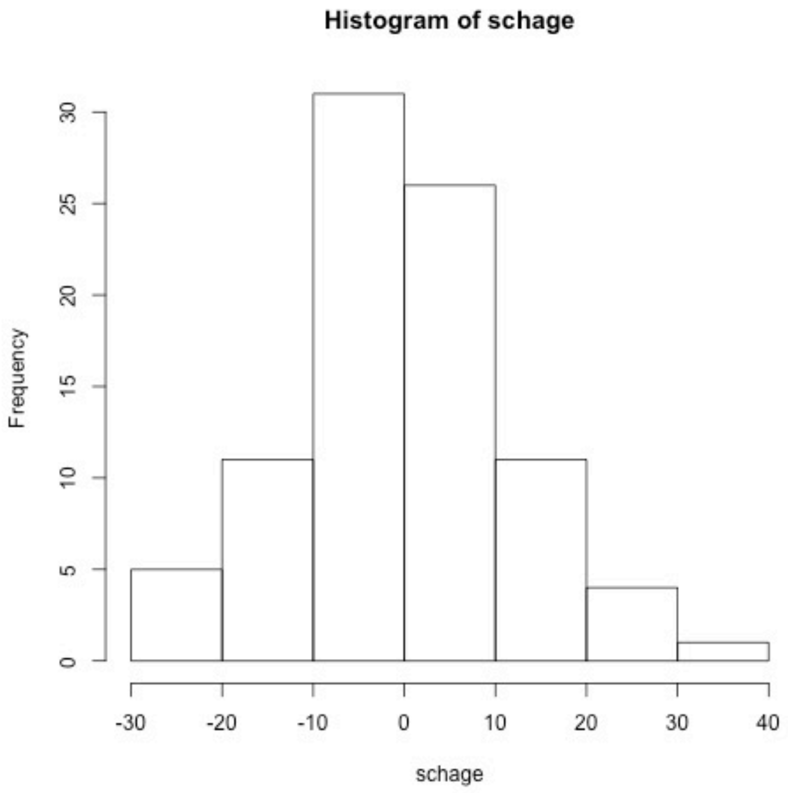
Neither looks particularly promising.

```
> 
> # Now fit a big model, including the variables that are not too promising
> 
> model5 = update(model1, . ~ . +
+          gender + race + yearsSmoking + levelSmoking + priorAttempts + loglong)
> summary(model5)

Call:
coxph(formula = DayOfRelapse ~ combo + age + agesq + employment + 
    gender + race + yearsSmoking + levelSmoking + priorAttempts + 
    loglong)

  n= 125, number of events= 89 

                        coef exp(coef)  se(coef)      z Pr(>|z|)   
combo             -0.6162036 0.5399905 0.2213444 -2.784  0.00537 **
age               -0.1041271 0.9011108 0.0625034 -1.666  0.09572 . 
agesq              0.0006036 1.0006038 0.0006163  0.979  0.32736   
employmentother    0.6317623 1.8809224 0.2808459  2.249  0.02448 * 
employmentpt       0.7378442 2.0914221 0.3412891  2.162  0.03062 * 
genderMale        -0.0136972 0.9863962 0.2461556 -0.056  0.95563   
racewhite         -0.1231556 0.8841261 0.2321666 -0.530  0.59579   
yearsSmoking       0.0147983 1.0149084 0.0189560  0.781  0.43500   
levelSmokinglight  0.0331515 1.0337071 0.2672959  0.124  0.90130   
priorAttempts      0.0006752 1.0006754 0.0011392  0.593  0.55337   
loglong           -0.0539605 0.9474696 0.0464761 -1.161  0.24563   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

                  exp(coef) exp(-coef) lower .95 upper .95
combo                0.5400     1.8519    0.3499    0.8333
age                  0.9011     1.1097    0.7972    1.0185
agesq                1.0006     0.9994    0.9994    1.0018
employmentother      1.8809     0.5317    1.0847    3.2616
employmentpt         2.0914     0.4781    1.0714    4.0827
genderMale           0.9864     1.0138    0.6089    1.5980
racewhite            0.8841     1.1311    0.5609    1.3936
yearsSmoking         1.0149     0.9853    0.9779    1.0533
levelSmokinglight    1.0337     0.9674    0.6122    1.7455
priorAttempts        1.0007     0.9993    0.9984    1.0029
loglong              0.9475     1.0554    0.8650    1.0378

Concordance= 0.643  (se = 0.034 )
Rsquare= 0.187   (max possible= 0.998 )
Likelihood ratio test= 25.92  on 11 df,   p=0.00668
Wald test            = 26.26  on 11 df,   p=0.005926
Score (logrank) test = 26.86  on 11 df,   p=0.004824
```

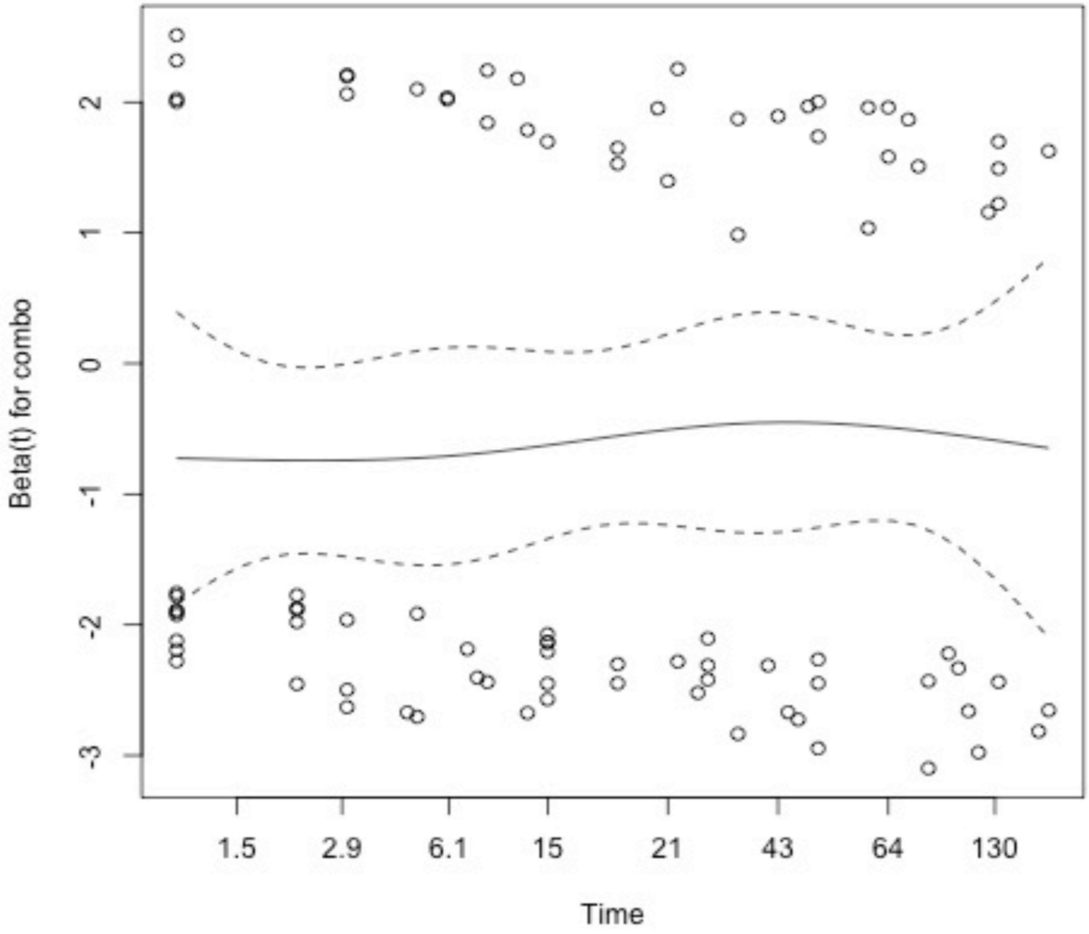
My conclusion is that I like model 3: DayOfRelapse ~ combo + age + employment

```
> # Test proportional hazards
> cox.zph(model3)
                    rho  chisq     p
combo            0.0394 0.1412 0.707
age              0.0176 0.0376 0.846
employmentother -0.0544 0.3111 0.577
employmentpt     0.0619 0.3497 0.554
GLOBAL               NA 1.0746 0.898
```

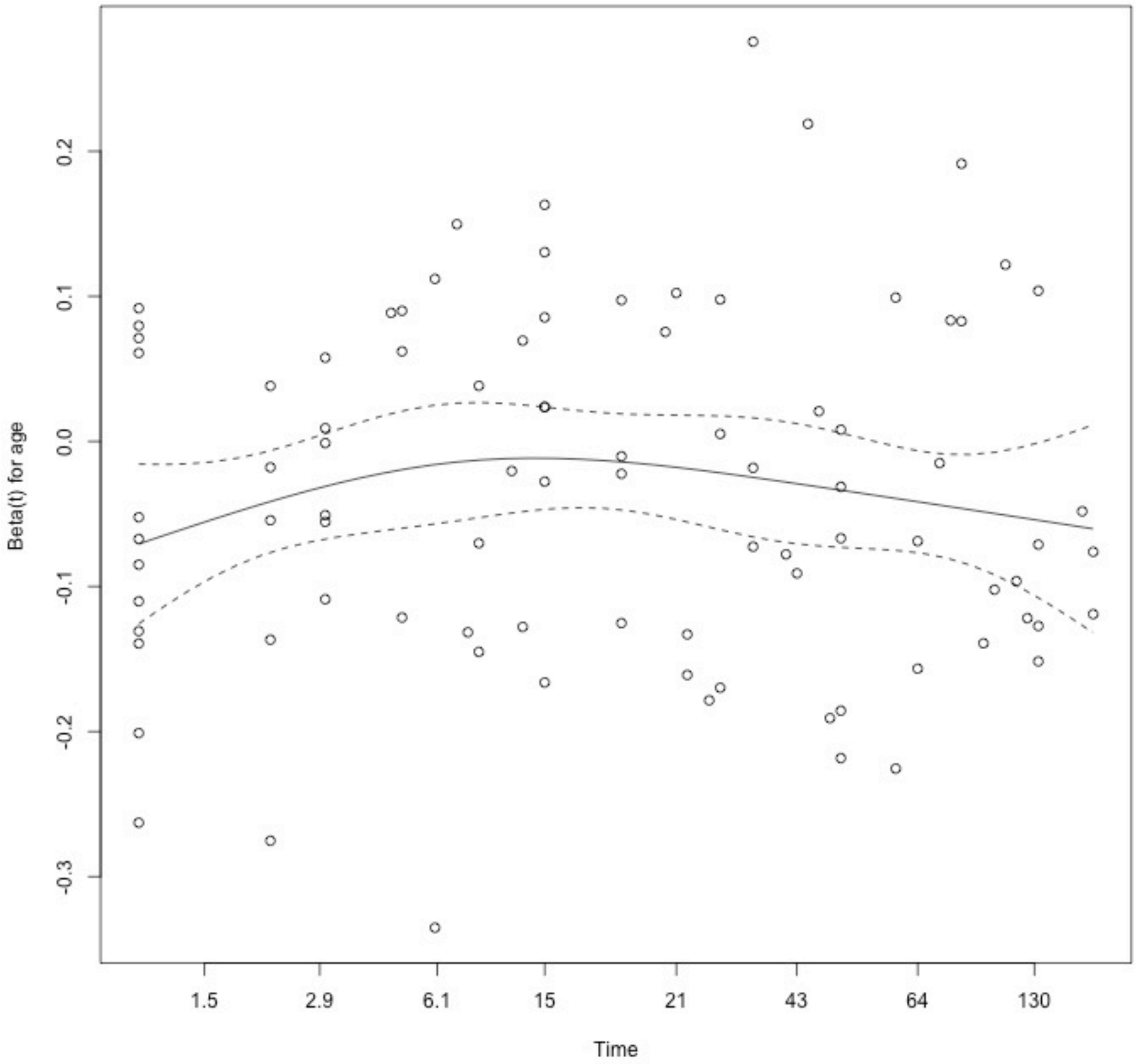
```
>
> # Check for influential observations
>
> # Look at Schoenfeld residuals
> sres = residuals(model3, type = 'schoenfeld')
> dim(sres); head(sres)

[1] 89  4
       combo         age employmentother employmentpt
1  0.6666922  0.06344131       0.6409129   -0.1744147
1 -0.3333078 -6.93655869      -0.3590871   -0.1744147
1 -0.3333078  3.06344131       0.6409129   -0.1744147
1 -0.3333078  7.06344131      -0.3590871   -0.1744147
1  0.6666922  7.06344131      -0.3590871   -0.1744147
1 -0.3333078 -9.93655869       0.6409129   -0.1744147

> schage = sres[,2]
>
>
> # Fig18
> hist(schage)
> # Write jpeg file to desktop
> jpeg('/Users/brunner/Desktop/Fig18.jpg'); hist(schage); dev.off()
quartz 
     2 
```



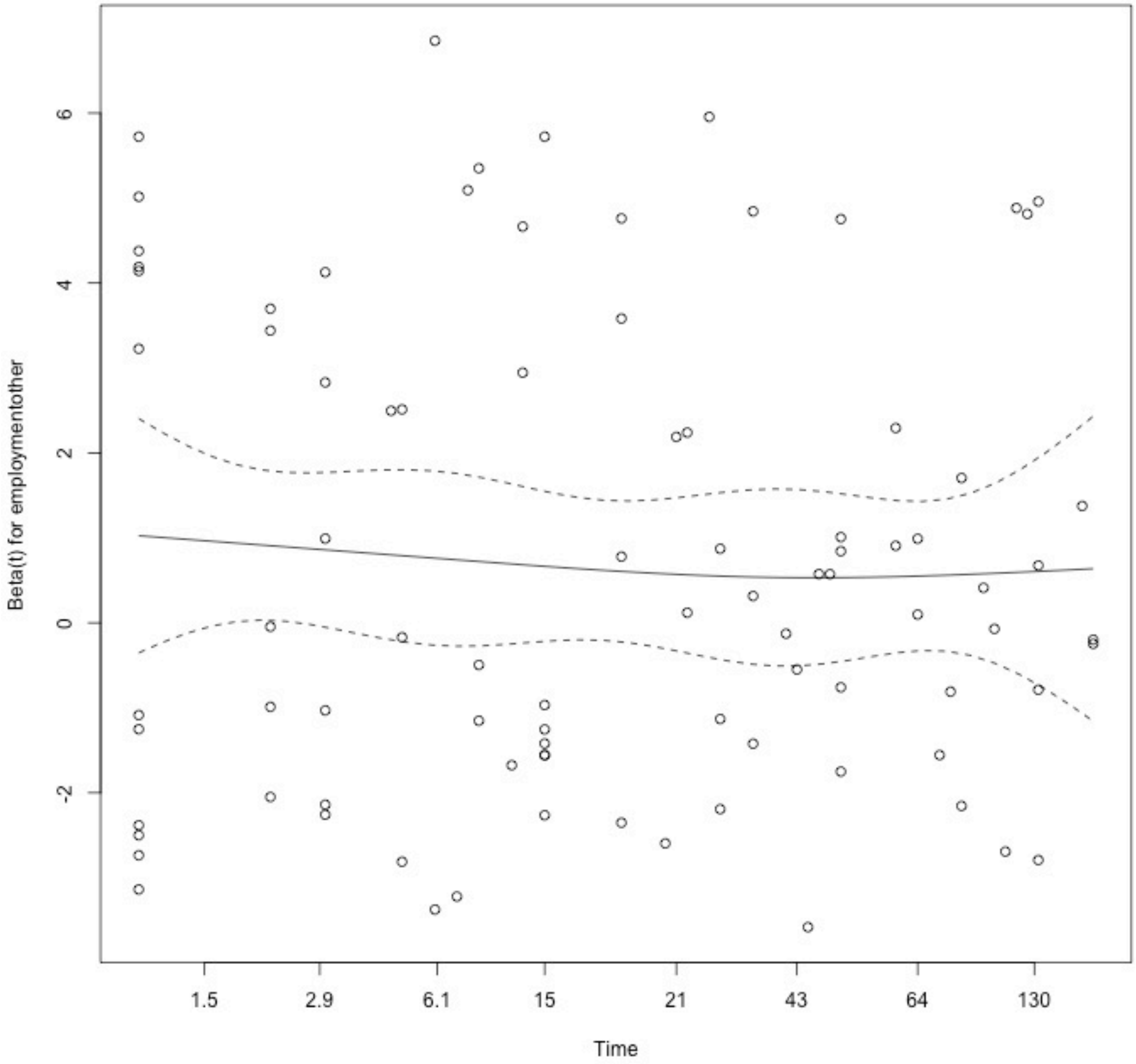
```
>
> # Plots of scaled Schoenfeld residuals (divided by estimated sd) against
> # time are can indicate departure from proportional hazards.
>
> # Therneau and Gramsch showed that, if the hazard ratio is a function g(t),
> # then the expected value of the scaled residual is beta(t) + c
>
> # The test for proportional hazards is a test of horizontal slope for
> # each x variable.
>
> # Get scaled residuals from cox.zph.
> ssr = cox.zph(model3) # Scaled residuals (also test)

> # Fig 19
> plot(ssr[1])
> jpeg('/Users/brunner/Desktop/Fig19.jpg'); plot(ssr[1]) ; dev.off()
quartz
     2
```



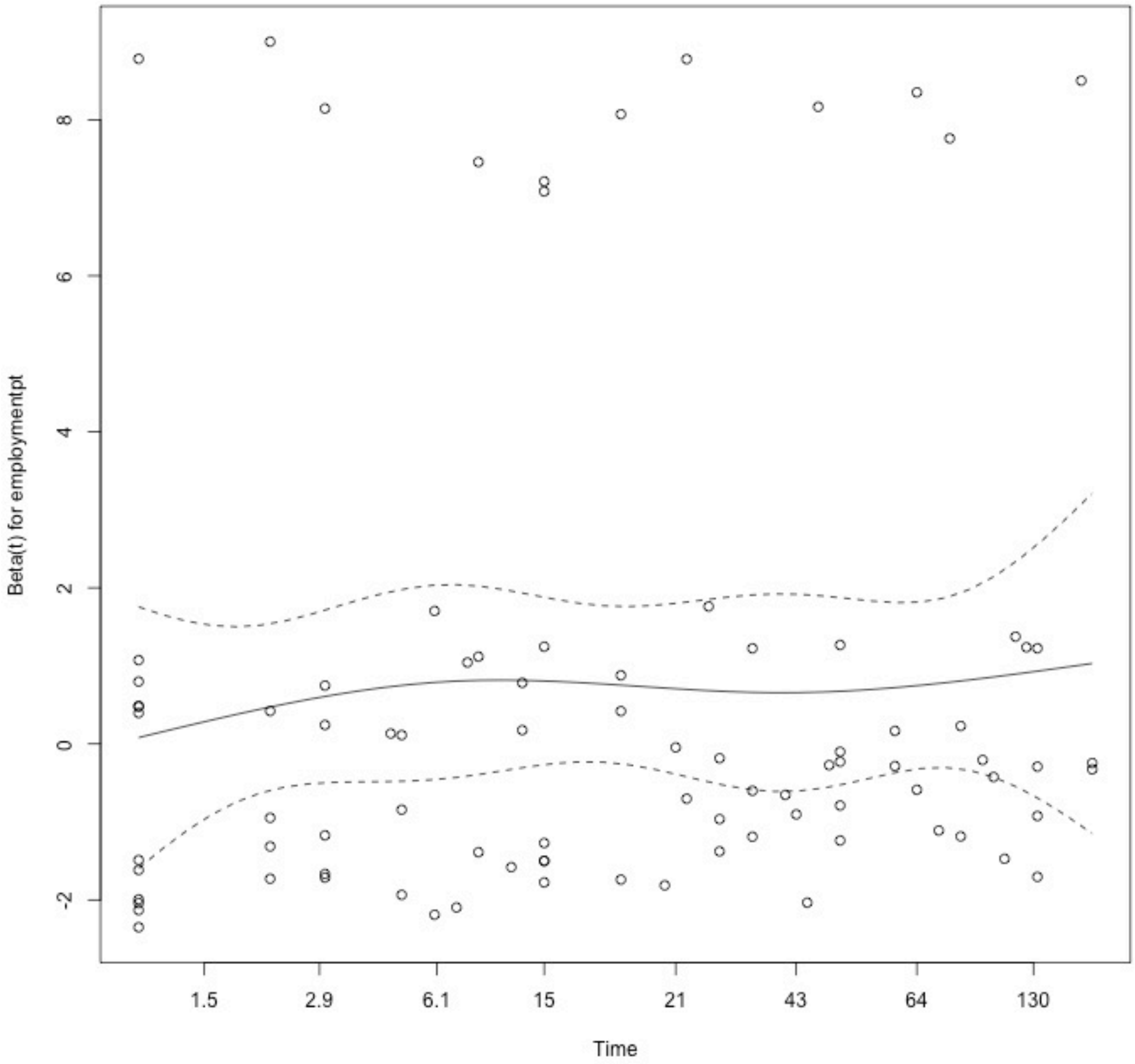
```
>
> # Fig 20
> plot(ssr[2])
> jpeg('/Users/brunner/Desktop/Fig20.jpg'); plot(ssr[2]) ; dev.off()
quartz
     2
```



```
> # Fig 21
> plot(ssr[3])
> jpeg('/Users/brunner/Desktop/Fig21.jpg', width = 700, height = 700)
> plot(ssr[3]) ; dev.off()
quartz
     2
```

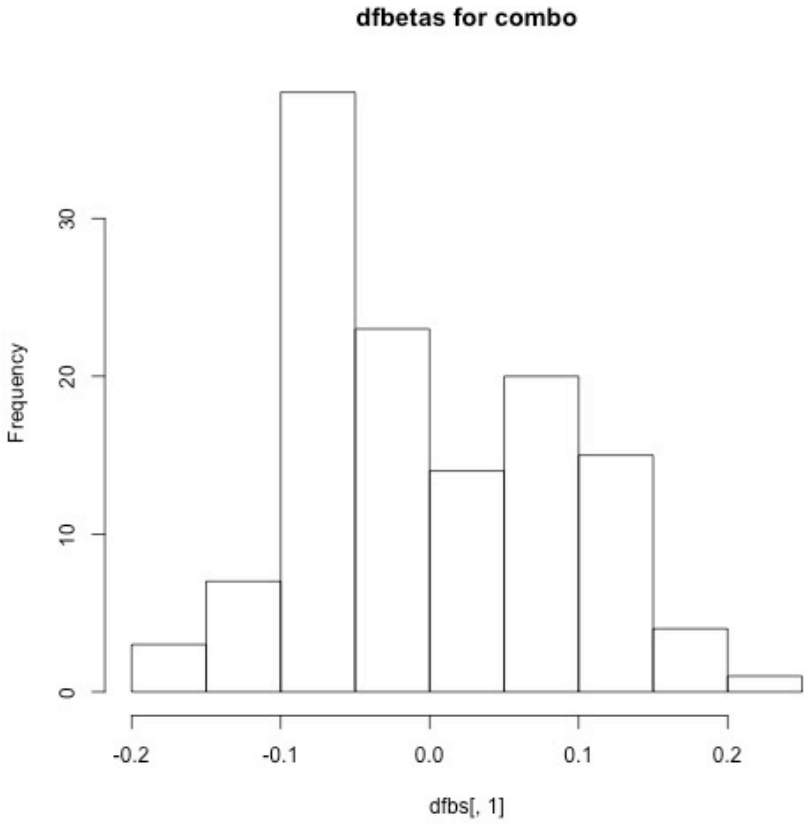


```
> # Fig 22
> plot(ssr[4])
> jpeg('/Users/brunner/Desktop/Fig22.jpg', width = 700, height = 700)
> plot(ssr[4]) ; dev.off()
quartz
     2
```

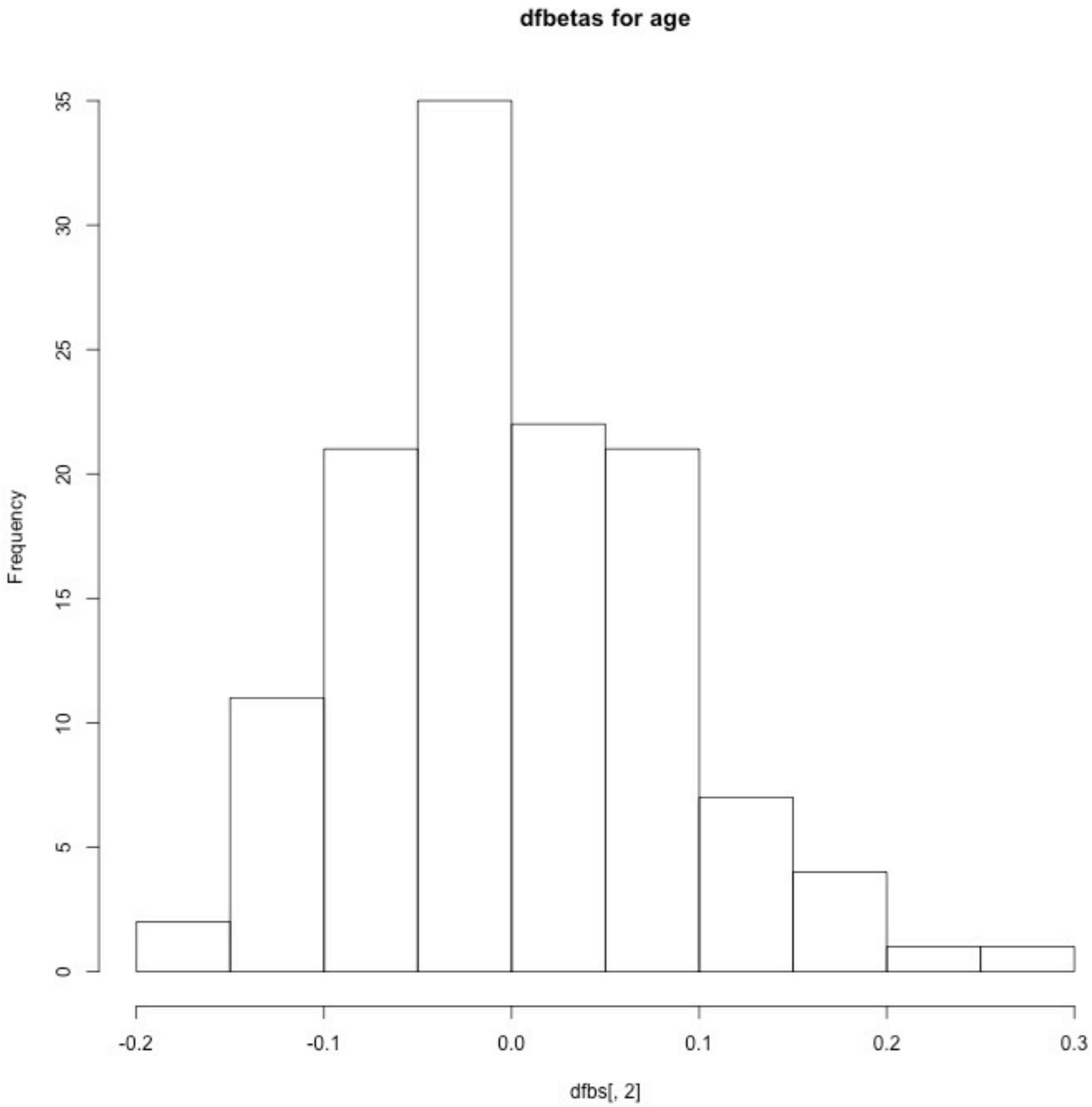


```
>
> ssr
                      rho  chisq     p
combo              0.0394 0.1412 0.707
age                0.0176 0.0376 0.846
employmentother   -0.0544 0.3111 0.577
employmentpt       0.0619 0.3497 0.554
GLOBAL                 NA 1.0746 0.898
```

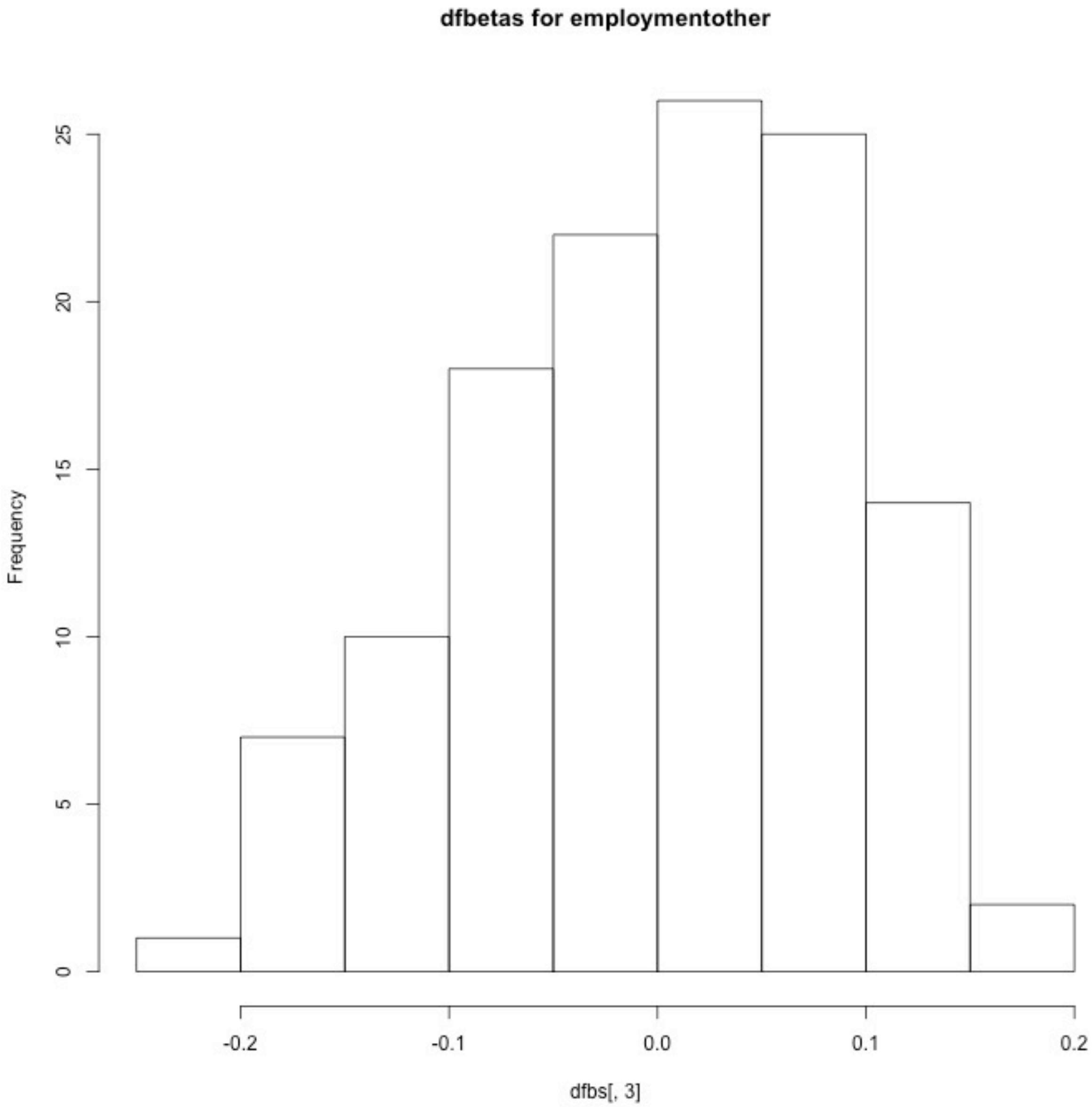
```
>
> # Look at bfbetas (beta-hat with one left out, standardized)
> dfbs = residuals(model3, type = 'dfbetas')
> dim(dfbs); head(dfbs)

[1] 125   4
          [,1]        [,2]        [,3]         [,4]
1  0.164823287  0.15809557  0.05955969  0.089952431
2 -0.006822493 -0.01266160  0.01817193  0.005336660
3  0.050525661 -0.13050951  0.10683305  0.017798812
4  0.101956037  0.08958552 -0.10516817 -0.064573127
5  0.126362316 -0.07386963  0.13516905 -0.008288001
6 -0.114896122  0.01309695  0.06960045  0.054304287
> colnames(dfbs) = names(model3$coefficients)
> summary(dfbs)
     combo                age            employmentother     employmentpt
 Min.   :-0.17904   Min.   :-0.18359   Min.   :-0.24900   Min.   :-0.388361
 1st Qu.:-0.06428   1st Qu.:-0.05136   1st Qu.:-0.05825   1st Qu.:-0.032523
 Median :-0.02301   Median :-0.01126   Median : 0.01096   Median : 0.005337
 Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.000000
 3rd Qu.: 0.06836   3rd Qu.: 0.05283   3rd Qu.: 0.06960   3rd Qu.: 0.044017
 Max.   : 0.21821   Max.   : 0.26434   Max.   : 0.17770   Max.   : 0.215041
>
> # Fig 23
> hist(dfbs[,1],main='dfbetas for combo')
> jpeg('/Users/brunner/Desktop/Fig23.jpg', width = 500, height = 500)
> hist(dfbs[,1],main='dfbetas for combo')
> dev.off()
```



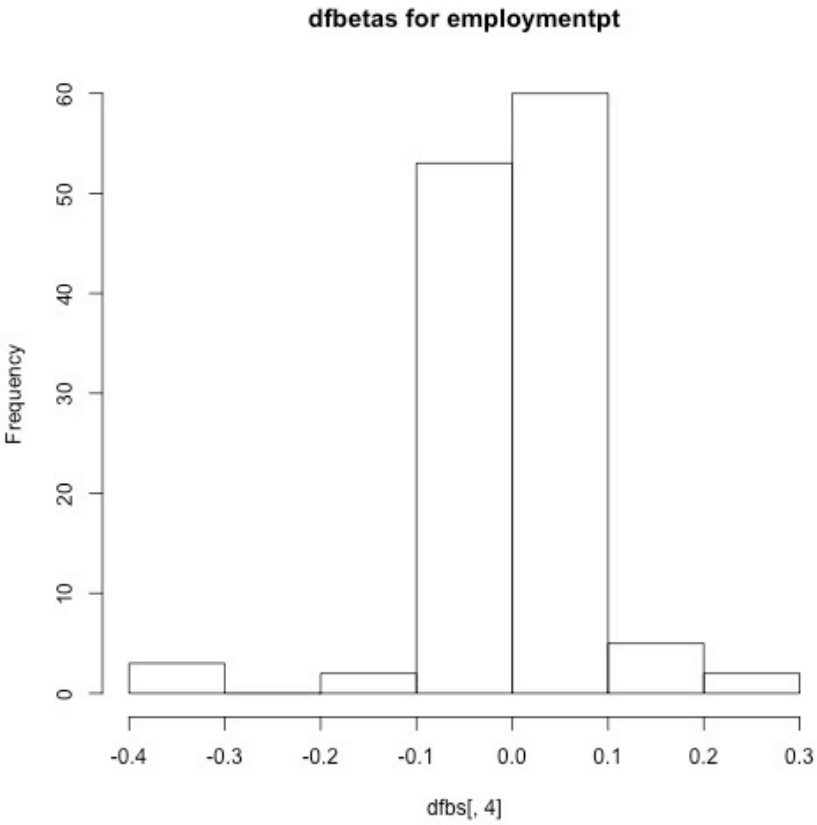
```
>
> # Fig 24
> hist(dfbs[,2],main='dfbetas for age')
> jpeg('/Users/brunner/Desktop/Fig24.jpg', width = 700, height = 700)
> hist(dfbs[,2],main='dfbetas for age') ; dev.off()
quartz
     2
```



```
>
> # Fig 25
> hist(dfbs[,3],main='dfbetas for employmentother')
> jpeg('/Users/brunner/Desktop/Fig25.jpg', width = 700, height = 700)
> hist(dfbs[,3],main='dfbetas for employmentother') ; dev.off()
quartz
     2
```



```
> # Fig 26
> hist(dfbs[,4],main='dfbetas for employmentpt')
> jpeg('/Users/brunner/Desktop/Fig26.jpg', width = 500, height = 500)
> hist(dfbs[,4],main='dfbetas for employmentpt') ; dev.off()
quartz 
     2 
```



```
> # Find those observations with low dfbetas for employmentpt
> loc = 1:nrow(dfbs); low = loc[dfbs[,4]< -0.25]; low
[1]  33  84 125
> 
> pharmacoSmoking[low,]
     id ttr relapse         grp age gender  race employment yearsSmoking
33  130 182       0 combination  46 Female white         pt           25
84   81 155       1   patchOnly  49 Female white         pt           35
125 128 182       0 combination  50 Female black         pt           30
    levelSmoking ageGroup2 ageGroup4 priorAttempts longestNoSmoke
33         heavy     21-49     35-49             3            365
84         heavy     21-49     35-49             1           1095
125        heavy       50+     50-64             0              0
```
