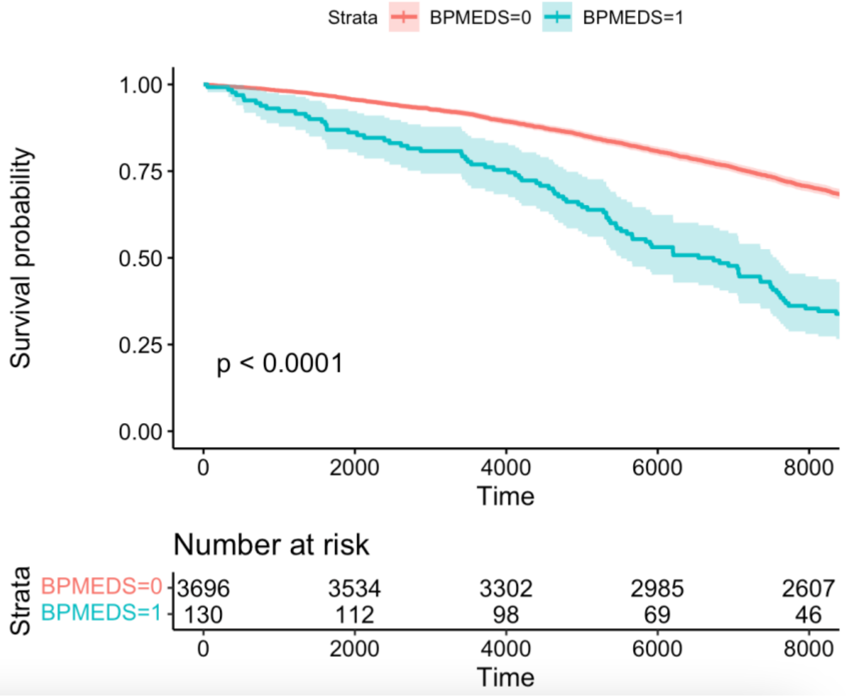
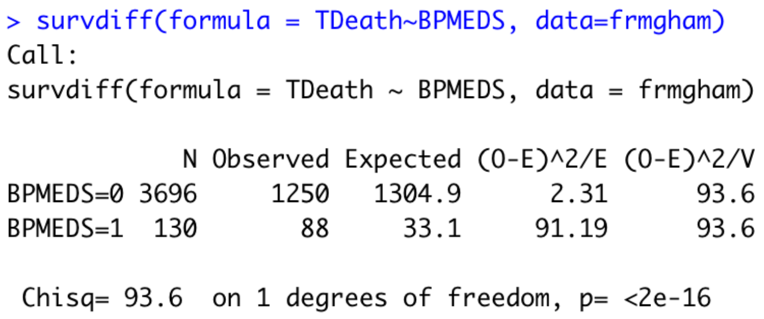
1. **Using frmgham_data_cleaned.RData, we wish to explore if there is an association between anti-hypertensive medication (BPMEDS) with death**:

   a. Produce a Kaplan-Meier plot of time-to-death, to visualize if there appears to be an association between use of anti-hypertensive medication on mortality. Use a log-rank test to formally test for differences



Inferring from the graph above, since the confidence intervals of both curves do not overlap, we can say that patients who take anti-hypertensive medication appear to have worse survival than those who don't take medication. However, to formally test the difference, we set up a hypothesis and perform a log-rank test as below:

H_0: there is no difference in survival between patients who take medication and patients who don't take medication
H_1: there is a difference in survival between patients who take medication and patients who don't take medication



```
> survdiff(formula = TDeath~BPMEDS, data=frmgham)
Call:
survdiff(formula = TDeath ~ BPMEDS, data = frmgham)

            N Observed Expected (O-E)^2/E (O-E)^2/V
BPMEDS=0 3696     1250   1304.9      2.31      93.6
BPMEDS=1  130       88     33.1     91.19      93.6

 Chisq= 93.6  on 1 degrees of freedom, p= <2e-16
```

From the R output above, the p-value is 2x10-16 which means we can confidently reject the null hypothesis and conclude that there is a difference in survival rates between patients who do and don’t take anti-hypertensive medication. This substantiates our initial statement that patients who take medication have worse survival rates.

b. Fit relevant Cox proportional hazards models to test the association between use of anti-hypertensive medication with time-to-death. Based on the models you fit, what can we conclude regarding our primary question of whether use of anti-hypertensive medication at exam time is associated with time-to-death?

The Cox proportional hazard model is as follows:

$$h(t) = h_0(t)\exp(\beta_1 x_1 + \ \beta_2 x_2 + \cdots + \beta_p x_p).$$

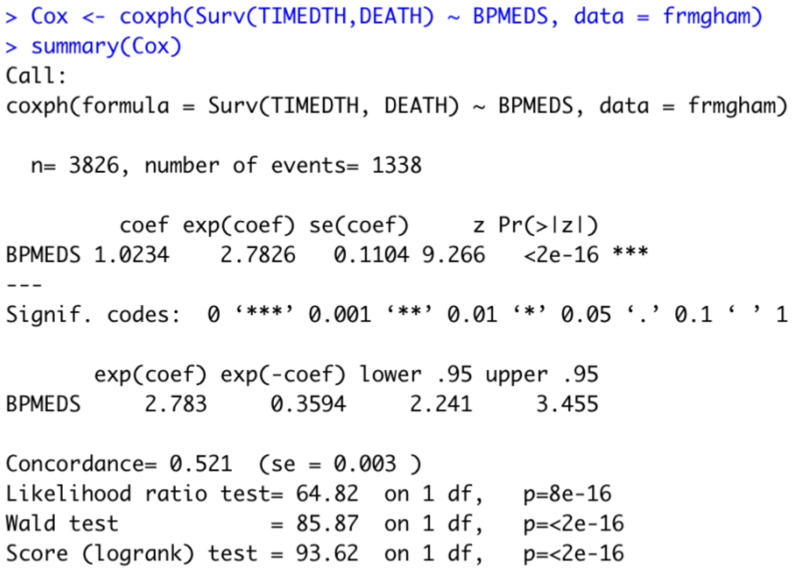
Where $h_0(t)$ = baseline hazard, x = covariates, and β = coefficients

The Cox proportional hazard model assumes that 1) The hazard function for all patients has the same shape and 2) Covariates have the effect of multiplying this hazard function by a constant factor.

For this question, three different Cox models were evaluated: model with one variable (BPMEDS), full model, and backwards selection model. For all three models, we can set up a hypothesis test:

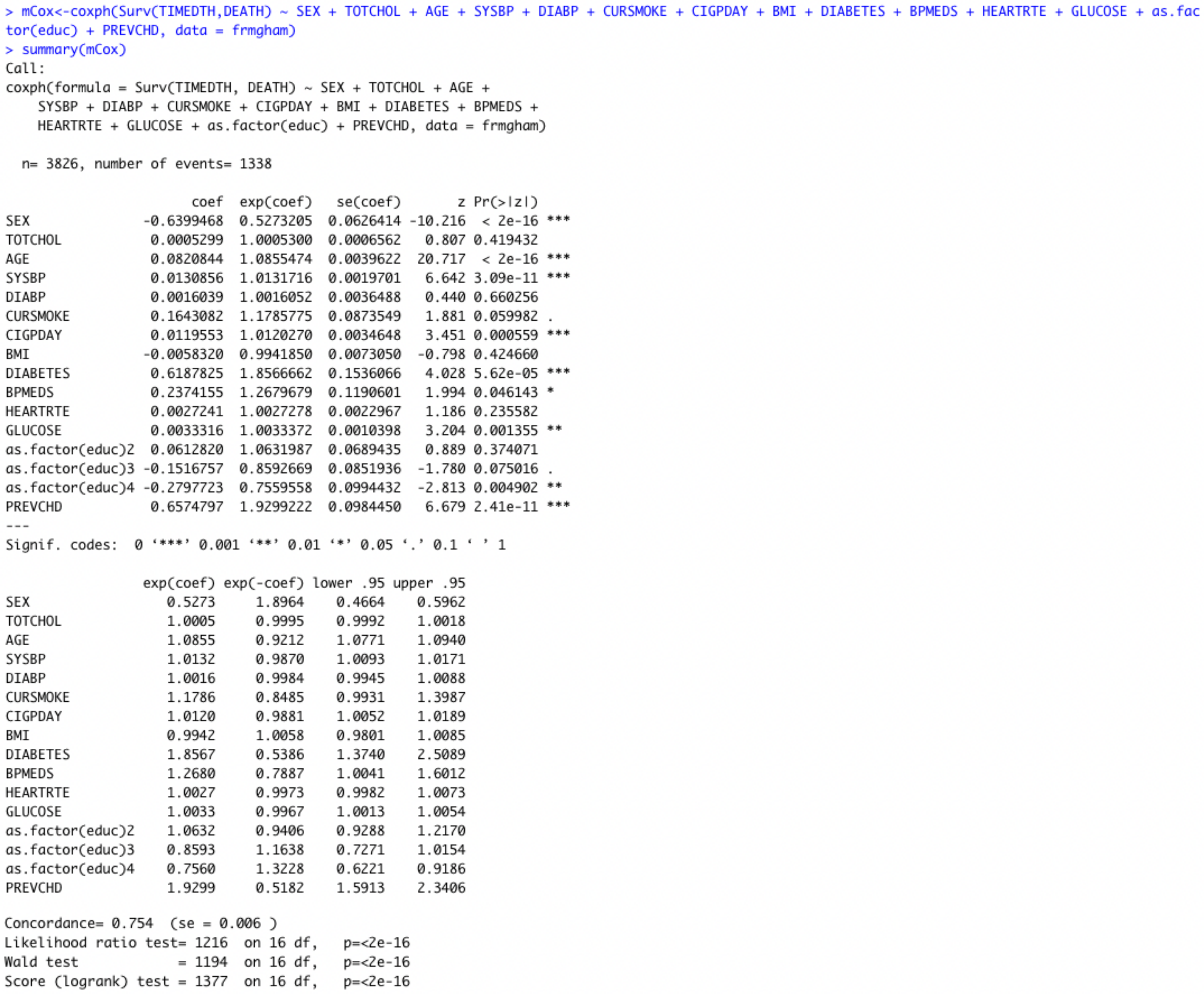
H_0: Patients who take antihypertension medication have the same survival rate as patients who do not take antihypertension medication

H_1: Patients who take antihypertension medication have a different survival rate as patients who do not take antihypertension medication



```
> Cox <- coxph(Surv(TIMEDTH,DEATH) ~ BPMEDS, data = frmgham)
> summary(Cox)
Call:
coxph(formula = Surv(TIMEDTH, DEATH) ~ BPMEDS, data = frmgham)

  n= 3826, number of events= 1338

         coef exp(coef) se(coef)     z Pr(>|z|)
BPMEDS 1.0234    2.7826   0.1104 9.266   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

       exp(coef) exp(-coef) lower .95 upper .95
BPMEDS     2.783     0.3594     2.241     3.455

Concordance= 0.521  (se = 0.003 )
Likelihood ratio test= 64.82  on 1 df,   p=8e-16
Wald test            = 85.87  on 1 df,   p=<2e-16
Score (logrank) test = 93.62  on 1 df,   p=<2e-16
```

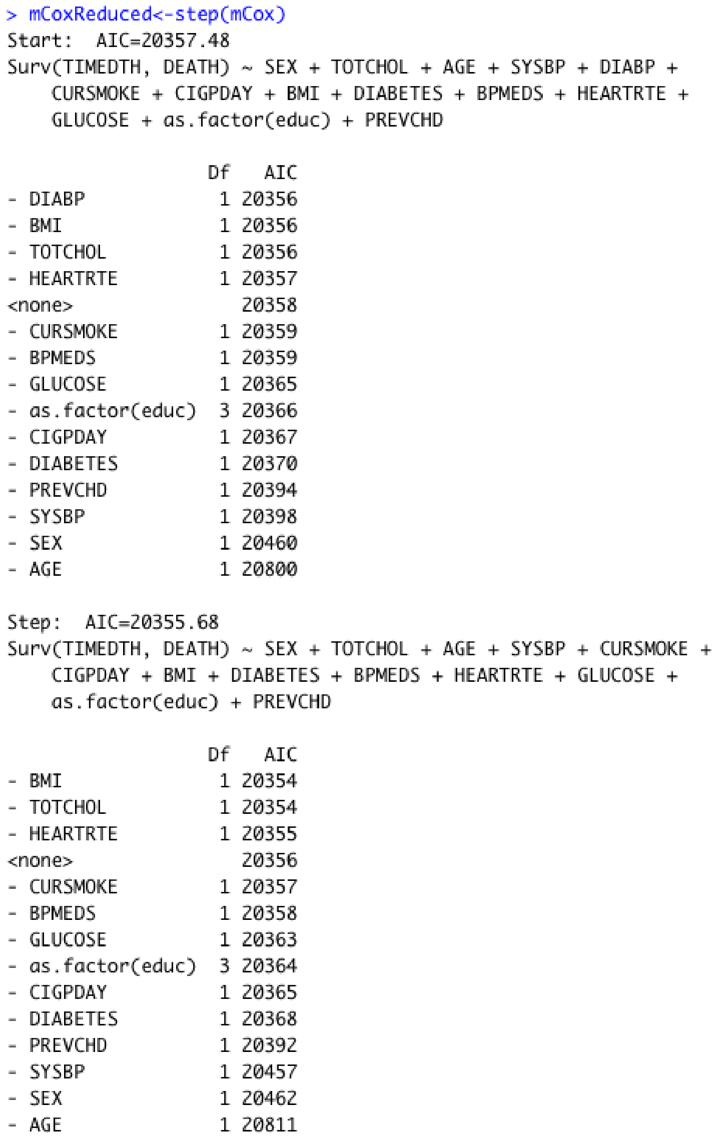
BPMEDS is shown to be significant as it has a p-value of less than 0.05 and the 95% confidence interval does not include 1. This indicates that patients who receive medication has 2.783 times the odds of death than patients who do not receive medication. This preliminary shows that there is a difference in survival rates between the two patient groups. However, we need to take into consideration the impact of other covariates as shown below:



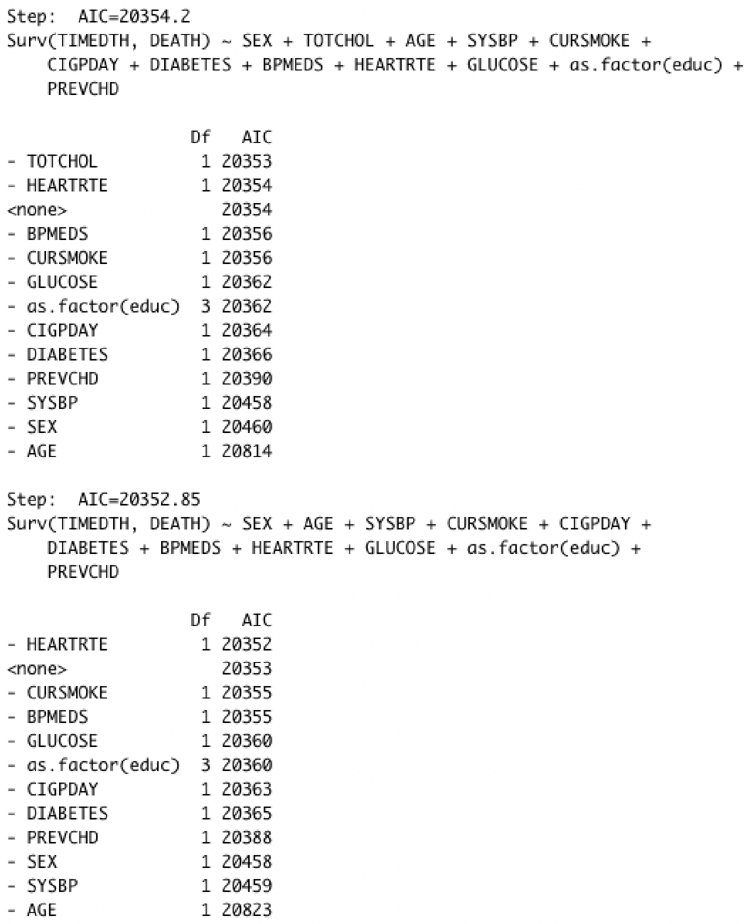
```
> mCox<-coxph(Surv(TIMEDTH,DEATH) ~ SEX + TOTCHOL + AGE + SYSBP + DIABP + CURSMOKE + CIGPDAY + BMI + DIABETES + BPMEDS + HEARTRTE + GLUCOSE + as.fac
tor(educ) + PREVCHD, data = frmgham)
> summary(mCox)
Call:
coxph(formula = Surv(TIMEDTH, DEATH) ~ SEX + TOTCHOL + AGE + 
    SYSBP + DIABP + CURSMOKE + CIGPDAY + BMI + DIABETES + BPMEDS + 
    HEARTRTE + GLUCOSE + as.factor(educ) + PREVCHD, data = frmgham)

  n= 3826, number of events= 1338 

                      coef  exp(coef)   se(coef)       z Pr(>|z|)    
SEX             -0.6399468  0.5273205  0.0626414 -10.216  < 2e-16 ***
TOTCHOL          0.0005299  1.0005300  0.0006562   0.807 0.419432    
AGE              0.0820844  1.0855474  0.0039622  20.717  < 2e-16 ***
SYSBP            0.0130856  1.0131716  0.0019701   6.642 3.09e-11 ***
DIABP            0.0016039  1.0016052  0.0036488   0.440 0.660256    
CURSMOKE         0.1643082  1.1785775  0.0873549   1.881 0.059982 .  
CIGPDAY          0.0119553  1.0120270  0.0034648   3.451 0.000559 ***
BMI             -0.0058320  0.9941850  0.0073050  -0.798 0.424660    
DIABETES         0.6187825  1.8566662  0.1536066   4.028 5.62e-05 ***
BPMEDS           0.2374155  1.2679679  0.1190601   1.994 0.046143 *  
HEARTRTE         0.0027241  1.0027278  0.0022967   1.186 0.235582    
GLUCOSE          0.0033316  1.0033372  0.0010398   3.204 0.001355 ** 
as.factor(educ)2  0.0612820  1.0631987  0.0689435   0.889 0.374071    
as.factor(educ)3 -0.1516757  0.8592669  0.0851936  -1.780 0.075016 .  
as.factor(educ)4 -0.2797723  0.7559558  0.0994432  -2.813 0.004902 ** 
PREVCHD          0.6574797  1.9299222  0.0984450   6.679 2.41e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                 exp(coef) exp(-coef) lower .95 upper .95
SEX                 0.5273     1.8964    0.4664    0.5962
TOTCHOL             1.0005     0.9995    0.9992    1.0018
AGE                 1.0855     0.9212    1.0771    1.0940
SYSBP               1.0132     0.9870    1.0093    1.0171
DIABP               1.0016     0.9984    0.9945    1.0088
CURSMOKE            1.1786     0.8485    0.9931    1.3987
CIGPDAY             1.0120     0.9881    1.0052    1.0189
BMI                 0.9942     1.0058    0.9801    1.0085
DIABETES            1.8567     0.5386    1.3740    2.5089
BPMEDS              1.2680     0.7887    1.0041    1.6012
HEARTRTE            1.0027     0.9973    0.9982    1.0073
GLUCOSE             1.0033     0.9967    1.0013    1.0054
as.factor(educ)2    1.0632     0.9406    0.9288    1.2170
as.factor(educ)3    0.8593     1.1638    0.7271    1.0154
as.factor(educ)4    0.7560     1.3228    0.6221    0.9186
PREVCHD             1.9299     0.5182    1.5913    2.3406

Concordance= 0.754  (se = 0.006 )
Likelihood ratio test= 1216  on 16 df,   p=<2e-16
Wald test            = 1194  on 16 df,   p=<2e-16
Score (logrank) test = 1377  on 16 df,   p=<2e-16
```

Based on this full model, we can observe that the antihypertensive medication covariate has a hazard ratio of 1.2680 with a 95% confidence interval of 1.0041 to 1.6012. As the p-value is less than 0.05 and the confidence interval does not include 1, the hazard ratio is significantly different from 1 – indicating that there is a difference in survival outcome between patients who receive medication and patients who don’t. Hence, patients who receive medication have a 27% higher hazard of death than patients who did not receive medication.

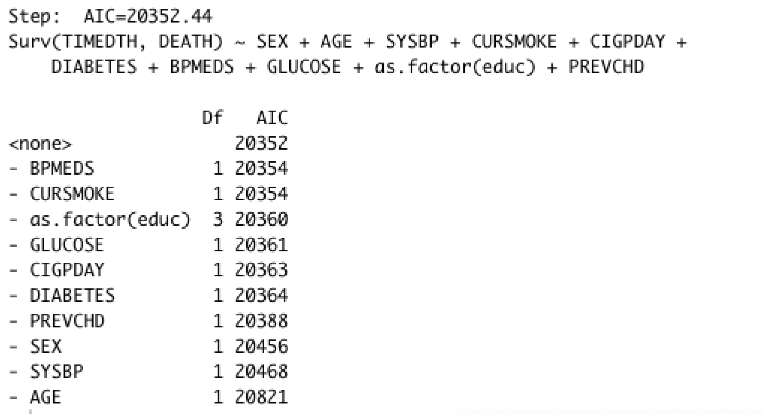
However, to optimize and simplify the Cox regression model, we can perform further model selection through backwards selection as shown below:



```
> mCoxReduced<-step(mCox)
Start:  AIC=20357.48
Surv(TIMEDTH, DEATH) ~ SEX + TOTCHOL + AGE + SYSBP + DIABP +
    CURSMOKE + CIGPDAY + BMI + DIABETES + BPMEDS + HEARTRTE +
    GLUCOSE + as.factor(educ) + PREVCHD

                  Df   AIC
- DIABP            1 20356
- BMI              1 20356
- TOTCHOL          1 20356
- HEARTRTE         1 20357
<none>               20358
- CURSMOKE         1 20359
- BPMEDS           1 20359
- GLUCOSE          1 20365
- as.factor(educ)  3 20366
- CIGPDAY          1 20367
- DIABETES         1 20370
- PREVCHD          1 20394
- SYSBP            1 20398
- SEX              1 20460
- AGE              1 20800

Step:  AIC=20355.68
Surv(TIMEDTH, DEATH) ~ SEX + TOTCHOL + AGE + SYSBP + CURSMOKE +
    CIGPDAY + BMI + DIABETES + BPMEDS + HEARTRTE + GLUCOSE +
    as.factor(educ) + PREVCHD

                  Df   AIC
- BMI              1 20354
- TOTCHOL          1 20354
- HEARTRTE         1 20355
<none>               20356
- CURSMOKE         1 20357
- BPMEDS           1 20358
- GLUCOSE          1 20363
- as.factor(educ)  3 20364
- CIGPDAY          1 20365
- DIABETES         1 20368
- PREVCHD          1 20392
- SYSBP            1 20457
- SEX              1 20462
- AGE              1 20811
```



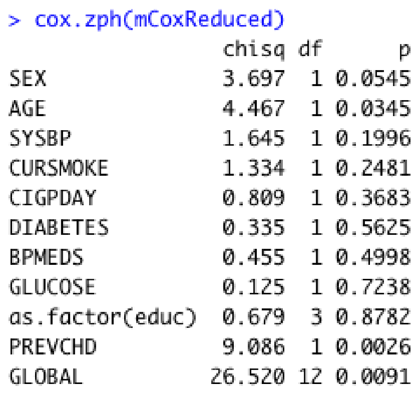
```
Step:  AIC=20354.2
Surv(TIMEDTH, DEATH) ~ SEX + TOTCHOL + AGE + SYSBP + CURSMOKE +
    CIGPDAY + DIABETES + BPMEDS + HEARTRTE + GLUCOSE + as.factor(educ) +
    PREVCHD

                  Df   AIC
- TOTCHOL          1 20353
- HEARTRTE         1 20354
<none>               20354
- BPMEDS           1 20356
- CURSMOKE         1 20356
- GLUCOSE          1 20362
- as.factor(educ)  3 20362
- CIGPDAY          1 20364
- DIABETES         1 20366
- PREVCHD          1 20390
- SYSBP            1 20458
- SEX              1 20460
- AGE              1 20814

Step:  AIC=20352.85
Surv(TIMEDTH, DEATH) ~ SEX + AGE + SYSBP + CURSMOKE + CIGPDAY +
    DIABETES + BPMEDS + HEARTRTE + GLUCOSE + as.factor(educ) +
    PREVCHD

                  Df   AIC
- HEARTRTE         1 20352
<none>               20353
- CURSMOKE         1 20355
- BPMEDS           1 20355
- GLUCOSE          1 20360
- as.factor(educ)  3 20360
- CIGPDAY          1 20363
- DIABETES         1 20365
- PREVCHD          1 20388
- SEX              1 20458
- SYSBP            1 20459
- AGE              1 20823
```



```
Step:  AIC=20352.44
Surv(TIMEDTH, DEATH) ~ SEX + AGE + SYSBP + CURSMOKE + CIGPDAY +
    DIABETES + BPMEDS + GLUCOSE + as.factor(educ) + PREVCHD

                  Df   AIC
<none>               20352
- BPMEDS           1 20354
- CURSMOKE         1 20354
- as.factor(educ)  3 20360
- GLUCOSE          1 20361
- CIGPDAY          1 20363
- DIABETES         1 20364
- PREVCHD          1 20388
- SEX              1 20456
- SYSBP            1 20468
- AGE              1 20821
```

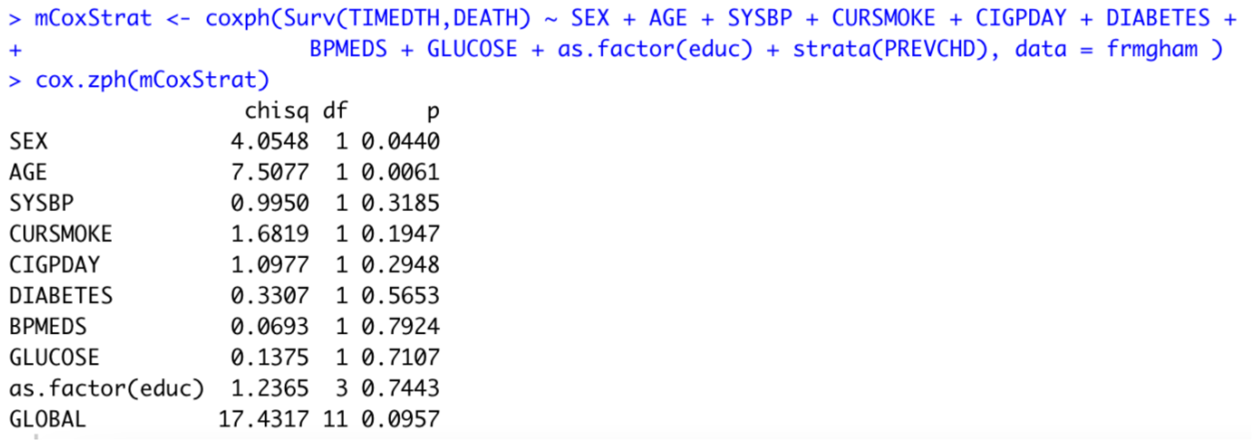
After backwards selection, the model has diastolic blood pressure, BMI, Serum Total Cholesterol (mg/dL), and heart rate removed. To check if this reduced model fits better than the Cox model, we can check if the Cox model’s assumptions hold true by testing if the betas change over time. Setting up the hypothesis test:

H_0: Beta does not change over time / Cox proportional hazards model’s assumption holds true
H_1: Beta changes over time / Cox proportional hazards model’s assumption is violated



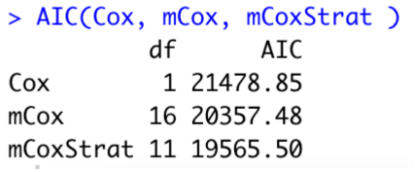
```
> cox.zph(mCoxReduced)
                  chisq df      p
SEX               3.697  1 0.0545
AGE               4.467  1 0.0345
SYSBP             1.645  1 0.1996
CURSMOKE          1.334  1 0.2481
CIGPDAY           0.809  1 0.3683
DIABETES          0.335  1 0.5625
BPMEDS            0.455  1 0.4998
GLUCOSE           0.125  1 0.7238
as.factor(educ)   0.679  3 0.8782
PREVCHD           9.086  1 0.0026
GLOBAL           26.520 12 0.0091
```

From the output above, the ‘global’ p-value of 0.0091 is significant at the 5% level. Thus, we can reject the null hypothesis and say that the Cox proportional hazards model’s assumption is not. However, when we strata PREVCHD (due to its low p-value) we can observe this:



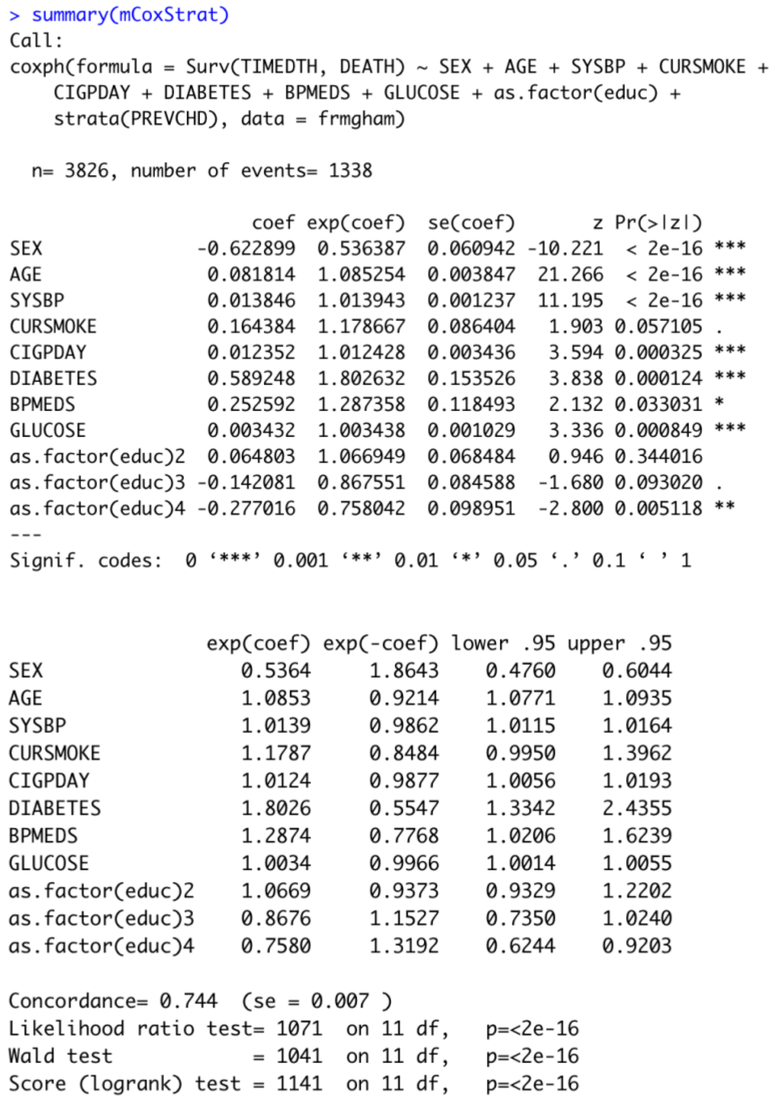
```
> mCoxStrat <- coxph(Surv(TIMEDTH,DEATH) ~ SEX + AGE + SYSBP + CURSMOKE + CIGPDAY + DIABETES +
+                      BPMEDS + GLUCOSE + as.factor(educ) + strata(PREVCHD), data = frmgham )
> cox.zph(mCoxStrat)
                   chisq df      p
SEX               4.0548  1 0.0440
AGE               7.5077  1 0.0061
SYSBP             0.9950  1 0.3185
CURSMOKE          1.6819  1 0.1947
CIGPDAY           1.0977  1 0.2948
DIABETES          0.3307  1 0.5653
BPMEDS            0.0693  1 0.7924
GLUCOSE           0.1375  1 0.7107
as.factor(educ)   1.2365  3 0.7443
GLOBAL           17.4317 11 0.0957
```

To our surprise, the ‘global’ p-value is now greater than 0.05 and we now do not have enough ground to reject the null hypothesis. Now, we can compare which model is statistically the best model using AIC function in R:



```
> AIC(Cox, mCox, mCoxStrat )
          df      AIC
Cox        1 21478.85
mCox      16 20357.48
mCoxStrat 11 19565.50
```

Since the stratified Cox model has the lowest AIC, we can summarise it using R:



```
> summary(mCoxStrat)
Call:
coxph(formula = Surv(TIMEDTH, DEATH) ~ SEX + AGE + SYSBP + CURSMOKE +
    CIGPDAY + DIABETES + BPMEDS + GLUCOSE + as.factor(educ) +
    strata(PREVCHD), data = frmgham)

  n= 3826, number of events= 1338

                      coef exp(coef)  se(coef)       z Pr(>|z|)
SEX              -0.622899  0.536387  0.060942 -10.221  < 2e-16 ***
AGE               0.081814  1.085254  0.003847  21.266  < 2e-16 ***
SYSBP             0.013846  1.013943  0.001237  11.195  < 2e-16 ***
CURSMOKE          0.164384  1.178667  0.086404   1.903 0.057105 .
CIGPDAY           0.012352  1.012428  0.003436   3.594 0.000325 ***
DIABETES          0.589248  1.802632  0.153526   3.838 0.000124 ***
BPMEDS            0.252592  1.287358  0.118493   2.132 0.033031 *
GLUCOSE           0.003432  1.003438  0.001029   3.336 0.000849 ***
as.factor(educ)2  0.064803  1.066949  0.068484   0.946 0.344016
as.factor(educ)3 -0.142081  0.867551  0.084588  -1.680 0.093020 .
as.factor(educ)4 -0.277016  0.758042  0.098951  -2.800 0.005118 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


                 exp(coef) exp(-coef) lower .95 upper .95
SEX                 0.5364     1.8643    0.4760    0.6044
AGE                 1.0853     0.9214    1.0771    1.0935
SYSBP               1.0139     0.9862    1.0115    1.0164
CURSMOKE            1.1787     0.8484    0.9950    1.3962
CIGPDAY             1.0124     0.9877    1.0056    1.0193
DIABETES            1.8026     0.5547    1.3342    2.4355
BPMEDS              1.2874     0.7768    1.0206    1.6239
GLUCOSE             1.0034     0.9966    1.0014    1.0055
as.factor(educ)2    1.0669     0.9373    0.9329    1.2202
as.factor(educ)3    0.8676     1.1527    0.7350    1.0240
as.factor(educ)4    0.7580     1.3192    0.6244    0.9203

Concordance= 0.744  (se = 0.007 )
Likelihood ratio test= 1071  on 11 df,   p=<2e-16
Wald test            = 1041  on 11 df,   p=<2e-16
Score (logrank) test = 1141  on 11 df,   p=<2e-16
```

Based on this summary output, the use of antihypertensive medication covariate has a hazard ratio of 1.2874 and 95% confidence interval of 1.0206 to 1.6239. The p-value of 0.033 and the confidence interval not including 1 provide us enough evidence to reject the null hypothesis. Therefore, we conclude that there is a difference in survival outcome between patients who receive medication and patients who don’t.