**Recidivism in the U.S.**

The dataset considered here is analyzed in Wooldridge (2002) and credited to Chung, Schmidt and Witte (1991). The data pertain to a random sample of convicts released from prison between July 1, 1977 and June 30, 1978. Of interest is the time until they return to prison. The information was collected retrospectively by looking at records in April 1984, so the maximum possible length of observation is 81 months. The data are available from the Stata website and can be accessed using the command

. use http://www.stata.com/data/jwooldridge/eacsap/recid

. desc, short

Contains data from http://www.stata.com/data/jwooldridge/eacsap/recid.dta

 obs: 1,445

 vars: 18 28 Sep 1998 13:28

 size: 39,015 (99.6% of memory free)

Sorted by:

You should have 1445 observations on 18 variables. The duration variable is called durat and represents time in months until return to prison or end of follow up. The censoring indicator is called cens and is coded 1 if the observation was censored (i.e. the individual had *not* returned to prison).

**Setting the Data**

Before using any of Stata's survival command's one has to **stset** the data. This tells Stata that we have duration data and specifies the *time variable* and the *failure indicator*. For this example we need to calculate the latter:

. gen fail = 1 - cens

. stset durat, failure(fail)

 failure event: fail != 0 & fail < .

obs. time interval: (0, durat]

 exit on or before: failure

------------------------------------------------------------------------------

 1445 total obs.

 0 exclusions

------------------------------------------------------------------------------

 1445 obs. remaining, representing

 552 failures in single record/single failure data

 80013 total analysis time at risk, at risk from t = 0

 earliest observed entry t = 0

 last observed exit t = 81

Stata runs a few sanity checks. Note that we have 552 events on 80,013 weeks of exposure, which gives a crude annual recidivism rate of 35.9% (or 36 events per 100 person-years of exposure). A crude rate should always be interpreted cautiously, particularly if there is duration dependence.

**A Proportional Hazards Weibull**

Wooldridge fits a Weibul model using as predictors

|  |  |
| --- | --- |
| workprg | an indicator of participation in a work program |
| priors | the number of previous convictions |
| tserved | the time served rounded to months |
| felon | and indicator of felony sentences |
| alcohol | an indicator of alcohol problems |
| drugs | an indicator of drug use history |
| black | an indicator for African Americans |
| married | an indicator if married when incarcerated |
| educ | the number of years of schooling, and |
| age | in months.  |

Let us first fit a proportional hazards model, which we can do using the streg command with the option distrib(weibull) to specify a Weibull distribution.

. local predictors workprg priors tserved felon alcohol drugs ///

> black married educ age

. streg `predictors', distrib(weibull)

 failure \_d: fail

 analysis time \_t: durat

Fitting constant-only model:

Iteration 0: log likelihood = -1739.8944

Iteration 1: log likelihood = -1716.1367

Iteration 2: log likelihood = -1715.7712

Iteration 3: log likelihood = -1715.7711

Fitting full model:

Iteration 0: log likelihood = -1715.7711

Iteration 1: log likelihood = -1669.1785

Iteration 2: log likelihood = -1634.3693

Iteration 3: log likelihood = -1633.0405

Iteration 4: log likelihood = -1633.0325

Iteration 5: log likelihood = -1633.0325

Weibull regression -- log relative-hazard form

No. of subjects = 1445 Number of obs = 1445

No. of failures = 552

Time at risk = 80013

 LR chi2(10) = 165.48

Log likelihood = -1633.0325 Prob > chi2 = 0.0000

------------------------------------------------------------------------------

 \_t | Haz. Ratio Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

 workprg | 1.095148 .0992728 1.00 0.316 .9168814 1.308074

 priors | 1.092848 .014683 6.61 0.000 1.064445 1.122008

 tserved | 1.013655 .0017037 8.07 0.000 1.010321 1.017

 felon | .7412054 .0785485 -2.83 0.005 .6021898 .9123128

 alcohol | 1.564179 .165389 4.23 0.000 1.271406 1.92437

 drugs | 1.325064 .1296765 2.88 0.004 1.093791 1.605237

 black | 1.574149 .1390031 5.14 0.000 1.32398 1.871587

 married | .8593436 .0938794 -1.39 0.165 .6937084 1.064527

 educ | .9769709 .0189724 -1.20 0.230 .9404845 1.014873

 age | .9962823 .000523 -7.09 0.000 .9952577 .997308

-------------+----------------------------------------------------------------

 /ln\_p | -.2158398 .0389149 -5.55 0.000 -.2921115 -.1395681

-------------+----------------------------------------------------------------

 p | .8058644 .0313601 .7466852 .8697338

 1/p | 1.240904 .0482896 1.149777 1.339252

------------------------------------------------------------------------------

Note that we do not specify the outcome, as this has been done with stset; we just list the predictors.

The Weibull parameter *p* is 0.8, indicating that the risk of recidivism declines over time (about 21% per week!). The hypothesis that the risk is constant over time would be soundly rejected.

BY default Stata exponentiates the coefficients to show relative risks. Use the option nohr, for no hazard ratios, to obtain the coefficients. This can be done issuing the streg command with no predictors, and reproduces Table 20.1 in Wooldridge:

. streg, nohr

Weibull regression -- log relative-hazard form

No. of subjects = 1445 Number of obs = 1445

No. of failures = 552

Time at risk = 80013

 LR chi2(10) = 165.48

Log likelihood = -1633.0325 Prob > chi2 = 0.0000

------------------------------------------------------------------------------

 \_t | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

 workprg | .0908893 .0906478 1.00 0.316 -.0867772 .2685558

 priors | .0887867 .0134355 6.61 0.000 .0624535 .1151198

 tserved | .0135625 .0016808 8.07 0.000 .0102682 .0168567

 felon | -.2994775 .105974 -2.83 0.005 -.5071826 -.0917723

 alcohol | .4473611 .1057353 4.23 0.000 .2401236 .6545985

 drugs | .2814605 .0978644 2.88 0.004 .0896499 .4732711

 black | .4537147 .0883037 5.14 0.000 .2806426 .6267867

 married | -.1515864 .1092454 -1.39 0.165 -.3657035 .0625307

 educ | -.0232984 .0194196 -1.20 0.230 -.0613601 .0147633

 age | -.0037246 .000525 -7.09 0.000 -.0047536 -.0026956

 \_cons | -3.402094 .3010177 -11.30 0.000 -3.992077 -2.81211

-------------+----------------------------------------------------------------

 /ln\_p | -.2158398 .0389149 -5.55 0.000 -.2921115 -.1395681

-------------+----------------------------------------------------------------

 p | .8058644 .0313601 .7466852 .8697338

 1/p | 1.240904 .0482896 1.149777 1.339252

------------------------------------------------------------------------------

All but three of the predictors affect recidivism, the exceptions being participation in a work program, marital status and education.

The coefficient of drugs indicates that former inmates with a history of drug use have 31% higher risk or returning to jail at any given time that peers with identical characteristics but no history of drug use.

**Accelerated Failure Time Weibull**

Let us fit the Weibull model in the accelerated failure time framework. We can do this simply adding the time option:

. streg `predictors', distrib(weibull) time

 failure \_d: fail

 analysis time \_t: durat

Fitting constant-only model:

Iteration 0: log likelihood = -1739.8944

Iteration 1: log likelihood = -1716.1367

Iteration 2: log likelihood = -1715.7712

Iteration 3: log likelihood = -1715.7711

Fitting full model:

Iteration 0: log likelihood = -1715.7711

Iteration 1: log likelihood = -1669.1785

Iteration 2: log likelihood = -1634.3693

Iteration 3: log likelihood = -1633.0405

Iteration 4: log likelihood = -1633.0325

Iteration 5: log likelihood = -1633.0325

Weibull regression -- accelerated failure-time form

No. of subjects = 1445 Number of obs = 1445

No. of failures = 552

Time at risk = 80013

 LR chi2(10) = 165.48

Log likelihood = -1633.0325 Prob > chi2 = 0.0000

------------------------------------------------------------------------------

 \_t | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

 workprg | -.1127848 .1125346 -1.00 0.316 -.3333486 .107779

 priors | -.1101757 .0170675 -6.46 0.000 -.1436273 -.0767241

 tserved | -.0168297 .0021303 -7.90 0.000 -.021005 -.0126544

 felon | .3716227 .1319951 2.82 0.005 .112917 .6303284

 alcohol | -.555132 .1322427 -4.20 0.000 -.8143229 -.295941

 drugs | -.3492654 .1218801 -2.87 0.004 -.5881461 -.1103847

 black | -.5630162 .110817 -5.08 0.000 -.7802135 -.3458189

 married | .1881041 .1357519 1.39 0.166 -.0779647 .4541729

 educ | .0289111 .0241153 1.20 0.231 -.0183541 .0761763

 age | .0046219 .0006648 6.95 0.000 .0033189 .0059249

 \_cons | 4.22167 .3413114 12.37 0.000 3.552712 4.890628

-------------+----------------------------------------------------------------

 /ln\_p | -.2158398 .0389149 -5.55 0.000 -.2921115 -.1395681

-------------+----------------------------------------------------------------

 p | .8058644 .0313601 .7466852 .8697338

 1/p | 1.240904 .0482896 1.149777 1.339252

------------------------------------------------------------------------------

By default Stata does *not* exponentiate the coefficients in AFT models. You *can* exponentiate them using the option tr, which stands for *time ratios*.

The substantive results are the same as before, which is not surprising because we have fitted exactly the same model. You may want to verify that the AFT parameters are exactly the same as the PH parameters with opposite sign and divided by p. For example the coefficient for drugs is -0.28/0.8 = -0.35.

However, we have two new interpretations of these effects. Exponentiating the drug coefficient we see that former inmates with a history of drug use spend 29% less time out of prison than peers with the same characteristics but no history of drug use. This is because

. di 1-exp(\_b[drugs])

.29479404

Also, we can say that time outside of prison passes 42% faster for former inmates with a history of drug use than for those without, everything else being equal. (So they get into trouble more quickly.) This is because

. di exp(-\_b[drugs])

1.4180255

**A Log-Normal AFT Model**

The Weibull allows the hazard to increase or decrease with time but at a constant rate. Wooldridge notes that the log-normal distribution provides a better fit to the data. We can fit a log-normal in Stata just changing the distrib option to lognormal:

. streg `predictors', distrib(lognormal)

 failure \_d: fail

 analysis time \_t: durat

Fitting constant-only model:

Iteration 0: log likelihood = -1999.58

Iteration 1: log likelihood = -1695.747

Iteration 2: log likelihood = -1681.0153

Iteration 3: log likelihood = -1680.4273

Iteration 4: log likelihood = -1680.427

Iteration 5: log likelihood = -1680.427

Fitting full model:

Iteration 0: log likelihood = -1680.427

Iteration 1: log likelihood = -1608.1657

Iteration 2: log likelihood = -1597.1838

Iteration 3: log likelihood = -1597.0591

Iteration 4: log likelihood = -1597.059

Lognormal regression -- accelerated failure-time form

No. of subjects = 1445 Number of obs = 1445

No. of failures = 552

Time at risk = 80013

 LR chi2(10) = 166.74

Log likelihood = -1597.059 Prob > chi2 = 0.0000

------------------------------------------------------------------------------

 \_t | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

 workprg | -.0625714 .1200369 -0.52 0.602 -.2978394 .1726965

 priors | -.1372528 .0214587 -6.40 0.000 -.179311 -.0951946

 tserved | -.0193305 .0029779 -6.49 0.000 -.0251671 -.0134939

 felon | .4439944 .1450865 3.06 0.002 .1596302 .7283586

 alcohol | -.6349088 .1442165 -4.40 0.000 -.9175681 -.3522496

 drugs | -.2981599 .1327355 -2.25 0.025 -.5583168 -.0380031

 black | -.5427175 .1174427 -4.62 0.000 -.772901 -.312534

 married | .3406835 .139843 2.44 0.015 .0665962 .6147707

 educ | .0229195 .0253974 0.90 0.367 -.0268584 .0726975

 age | .0039103 .0006062 6.45 0.000 .0027221 .0050984

 \_cons | 4.099386 .3475349 11.80 0.000 3.41823 4.780542

-------------+----------------------------------------------------------------

 /ln\_sig | .5935861 .0344122 17.25 0.000 .5261395 .6610327

-------------+----------------------------------------------------------------

 sigma | 1.810469 .0623022 1.692386 1.936791

------------------------------------------------------------------------------

We do not need to specify time, as this distribution is available in Stata only in the AFT framework.

We see that the log-likelihood is indeed higher, -1597.1 compared to -1633.0 for the Weibull, so the model provides a better fit to the data.

Most of the effects are robust to the choice of distribution, but note that the protective effect of marriage is now significant. The coefficient for drugs, at -0.30 is smaller in magnitude and less significant than before.

The command stcurve can plot some aspects of the fit. Try the hazard option to have a look at the log-normal hazard evaluated at the mean of all predictors. You'll see that it raises very rapidly in the first seven weeks or so and then declines.

Fitting a generalized gamma model leads to similar conclusions except that the effect of drugs looses significance. These results suggests that there may be an interaction between drug history and duration, as the effect depends on how the hazard is specified. We will return to this issue.