Recidivism Data Example

- Experimental study of recidivism of 432 male prisioners
- Observed for a year after being released from prison.
- *WEEK:* week of arrest after being released or censored time.
- *ARREST:* '1' arrested during period. '0' not arrested.
- *FIN:* '1' individual received financial aid. '0' if not.
- *AGE:* in years at the time of release.
- *RACE:* '1' for African American. '0' for others.
- *WEXP:* '1' full time working experience prior to incarceration. '0' if not.

Recidivism Data Example

- *MAR:* '1' individual was married at time of release. '0' if not.
- *PARO:* '1' individual relased on parole. '0' if not.
- *PRIO:* number of prior convictions.
- *EDUC:* Education categorical variable. Code 2 (grade 6 or less), 3 (grade 6-9), 4 (grade 10-11), 5 (grade 12), 6 (post-secondary).
- *EMP1-EMP52:* '1' if individual employed in the coresponding week of study. '0' otherwise.
- Fitted *Cox-Proportional Hazard Model* in R. *coxph*

Concepts behind Cox-Proportional Hazard Model

- Survival times depend on $X_1, X_2, \ldots, X_p$ explanatory variables.
- $h_0(t)$ hazard function when $x_1 = x_2 = \ldots = x_p = 0$ (baseline).
- Suppose for the $i - th$ subject $X_{i,1}, X_{i,2}, \ldots, X_{i,p}$
- Hazard for subject i

$$h_i(t) = exp(\beta_1 X_{i,1} + \beta_2 X_{i,2} + \ldots + \beta_p X_{i,p}) h_0(t)$$

- Log of hazard ratio,

$$log(h_i(t)) = log(h_0(t)) + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \ldots + \beta_p X_{i,p}$$

- Allow $h_0(t)$ to be arbitrary: nonparametric.
- Focus on studying effect of predictors without specifying $h_0(t)$.

- Focus on β_1,

$$log(h(t)) = log(h_0(t)) + \beta_1 X_1$$

- If X_1 increases by 1,

$$log(h_n(t)) = log(h_0(t)) + \beta_1(X_1 + 1)$$

- Or

$$\text{log of new hazard} = \text{log of original hazard} + \beta_1$$

- If exponentiate,

$$\frac{h_n(t)}{h(t)} = exp(\beta_1)$$

- Increasing X_1 by 1, increases hazard by a factor of $exp(\beta_1)$.
- Similar idea applies to other predictors assuming the rest of the $X's$ are fixed.

Fitting the Cox PH model

- Provides estimates of $\beta_1, \beta_2, \ldots, \beta_p$.
- Maximizing the (partial) likelihood function for $L(\beta)$.
- Maximization done numerically with the Newton-Raphson method.

$$\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$$

- R provides $\hat{\beta}_i$, $SE(\hat{\beta}_i)$.
- Z score based on *large samples*, $\hat{\beta}_i \approx N(\beta_i, SE(\hat{\beta}_i))$
- For $H_0 : \beta_i = 0$,

$$Z = \frac{\hat{\beta}_i - 0}{SE(\hat{\beta}_i)} \approx N(0, 1)$$

- An approximate 95% confidence interval: $\hat{\beta}_i \pm 2SE(\hat{\beta}_i)$

- R also provides *Risk ratio estimator*: $\hat{\psi}_i = exp(\hat{\beta}_i)$ and C.I. with *Delta method*.

$$\hat{\psi}_i \pm 1.96\hat{\psi}_i SE(\hat{\beta}_i)$$

- Likelihood Ratio test: $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$ [no regression effects]. vs
  $H_a :$ at least 1 regression coefficient affects hazard
- Alternative one can use: *Wald test* or *Score test*.
- May try to compare various models with AIC

$$AIC = -2L_{max} + 2p.$$

Recidivism example

- 7 variables in model: FIN, AGE, RACE, WEXP, MAR, PARO, PRIO
- $AIC = 1331.495$
- Likelihood ratio test for $\beta_1 = \beta_2 = \ldots = \beta_p = 0$, $LR = 33.27$, 7 dofs and $p - value < 0.0001$
- For 'FIN' variable: $\hat{\beta}_{FIN} = -0.37942$ and $SE(\hat{\beta}_{FIN}) = 0.19138$
- $Z = -1.983$ with $p - value = 0.0472$
- Hazard ratio: $exp(\hat{\beta}_{FIN}) = 0.68426$
- Hazard of arrest for those receiving aid is 0.684 (decrease).
- Confidence interval: $(0.4702, 0.9957)$ (upper limit near 1).