1. Suppose that we are interested in estimating the median survival time from a sample of right censored observations. We have a random sample $\{(Y_i, \delta_i), i = 1, \ldots, n\}$ where, as usual,

$$\begin{aligned} Y_i &= \min\{T_i, C_i\} \\ \delta_i &= I(T_i < C_i\} \end{aligned}$$

with the T_i iid F, the C_i iid G, and T_i and C_i independent. We want to estimate $\theta_0 = F^{-1}(1/2)$. An obvious candidate estimator is $\hat{\theta}_{KM}$, the median of the Kaplan-Meier estimate of F. Another candidate estimator is Powell's (1984) censored regression estimator specialized to this simple setting,

$$\hat{\theta}_p = \arg\min_\theta \sum_{i=1}^{n} |y_i - \min\{\theta, c_i\}|.$$

Note that the latter estimator presumes that we know C_i for *all* the observations, not just the censored ones with $\delta_i = 0$, so one might expect the Powell approach to do better, since it uses more information. On the other hand, Kaplan-Meier has known optimality properties in this context.

(a) Asymptotic theory should be helpful in resolving this dispute. Based on Powell's asymptotics for the regression case, show that

$$\sqrt{n}(\hat{\theta}_p - \theta_0) \rightsquigarrow \mathcal{N}(0, (4f^2(\theta_0)(1 - G(\theta_0)))^{-1}).$$

In contrast the theory of $\hat{\theta}_{KM}$ is more involved. Using delta-method arguments show that

$$\sqrt{n}(\hat{\theta}_{KM} - \theta_0) \rightsquigarrow \mathcal{N}(0, \mathrm{Avar}(\hat{S}(\theta_0)/f^2(\theta_0))$$

Unfortunately,

$$\mathrm{Avar}(\hat{S}(t)) = S^2(t) \int_0^t (1 - H(u))^{-2} d\tilde{F}(u)$$

where $1 - H(u) = (1 - F(u))(1 - G(u))$ and $\tilde{F}(u) = \int_0^t (1 - G(u)) dF(u)$ is difficult to evaluate and compare.

(b) Instead, consider a toy Monte-Carlo experiment as a way to make the comparison. The KM estimator of the median can be quite easily computed as,

```
KM.median <- function(y,d){
g <- survfit(Surv(y,d))
        g$time[min(which(g$surv<=.5))]
}
```

Whereas the Powell estimator has a piecewise linear objective function with kinks at the observed Y_i's so it can be evaluated using,

```
Powell.median <- function(y,c){
        R <- function(a,y,c){
                sum(abs(y-pmin(a,c)))
                }
        r <- a <- sort(y)
        for(i in 1:length(a)) {
                r[i] <- R(a[i],y,c)
                }
        mean(a[which(r == min(r))])
        }
```

Verify with some plotting that these functions do what they are claimed to do. Note that you will need to use the `survival` package for the KM estimator. In order to enforce some degree of consistency across experiments, I suggest the following setup with F standard lognormal, G exponential with rate .25. Thus, conveniently, $\theta_0 = 1$.

```
t <- exp(rnorm(n))
cen <- rexp(n,.25)
y <- pmin(t,cen)
```

Design a small experiment to evaluate which of these estimators is better. For laughs try including the naive sample median of the Y's as a third alternative.

2. An estimator may be called an efficient likelihood estimator (ELE) if it asymptotically achieves the CRLB. The following result is central to the theory of "Hausman-type" tests in econometrics and describes the relationship between ELE's and competing estimators.

   Theorem: Suppose $\hat{\theta}_n$ is an ELE, $\tilde{\theta}_n$ is a $\sqrt{n}$-consistent estimator of θ_0, and

   $$\hat{Z}_n = \sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow \hat{Z}_0$$

   $$\tilde{Z}_n = \sqrt{n}(\tilde{\theta}_n - \theta_0) \rightsquigarrow \tilde{Z}_0$$

   where $(\hat{Z}_0, \tilde{Z}_0)$ have a joint normal distribution with mean zero and covariance matrix $\Sigma = (\sigma_{ij})$. Then the asymptotic relative efficiency of $\tilde{\theta}_n$ with respect to $\hat{\theta}_n$, i.e. the ratio of their limiting variances σ_{11}/σ_{22}, is given by $e = \rho^2$ where $\rho = \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$ is the correlation coefficient of $\hat{Z}_n$ and $\tilde{Z}_n$.

   Prove the result by considering $\text{Avar}((1-\alpha)\hat{Z}_n + \alpha\tilde{Z}_n)$ and showing that its limit is minimized at $\alpha = 0$ which implies $\sigma_{11} = \sigma_{12}$.

   Explain the connection of this result to the Hausman test.

3. Generalize the following question from the 2002 Final Exam (available on the web along with a sketchy answer) to the case of regression with $\mu = x_i^\top \beta$.

   Suppose $\{y_1, ..., y_n\}$ are iid random variables, each normally distributed with mean μ and variance μ^2. Find the mle of μ and argue its consistency. Compare the asymptotic efficiency of the mle in this problem with that of the sample mean. This problem is related to estimating models of heteroscedasticity in linear regression which have parameters in common with the model for the conditional mean.