Problem Set 11

Economemtrics

1a.

```
simulate <- function(n, beta = 0.5) {
    x <- runif(n, 0, 1)
    u <- rnorm(n, mean = 0, sd = 2)
    y_star <- 1 + beta * x + u
    y <- ifelse(y_star > 0, y_star, 0)
    return(data.frame(x, y))
}

set.seed(42)

simulated_data <- simulate(1000)
fraction_censored <- sum(simulated_data$y == 0) / nrow(simulated_data)

print(fraction_censored)
```

```
## [1] 0.275
```

1b.

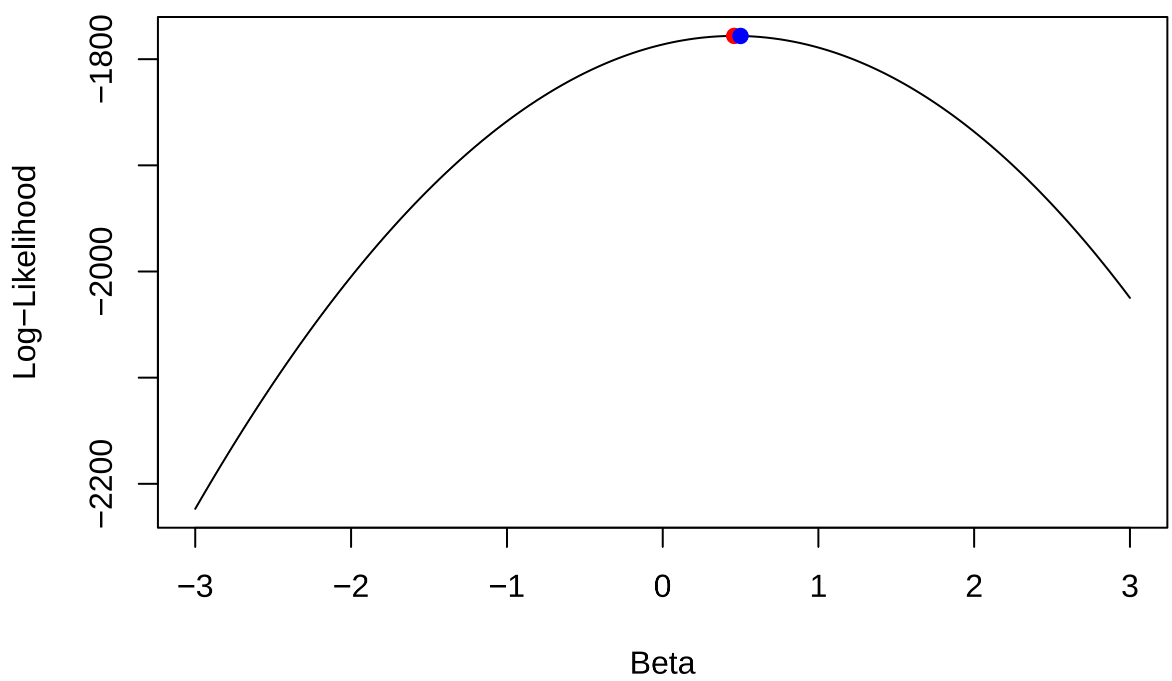
```
log_likelihood <- function(beta, sigma = 2, x, y) {
    y_star <- 1 + beta * x

    uncensored <- y > 0
    log_likelihood_uncensored <- sum(
```

```
        dnorm(y[uncensored], mean = y_star[uncensored], sd = sigma, log = TRUE)
    )
    log_likelihood_censored <- sum(
        pnorm(
            y_star[!uncensored],
            mean = 0, sd = sigma,
            lower.tail = FALSE, log = TRUE
        )
    )

    total_log_likelihood <- log_likelihood_uncensored + log_likelihood_censored

    return(total_log_likelihood)
}
```

1c.

```
beta_grid <- seq(-3, 3, by = 0.01)

x <- simulated_data$x
y <- simulated_data$y

log_likelihood_values <- sapply(
    beta_grid, function(beta) log_likelihood(beta, x = x, y = y)
)

max_index <- which.max(log_likelihood_values)
max_beta <- beta_grid[max_index]

plot(
    beta_grid, log_likelihood_values,
```

```
    type = "l", xlab = "Beta", ylab = "Log-Likelihood"
)
points(
    max_beta, log_likelihood_values[max_index],
    col = "red", pch = 19
)

true_beta <- 0.5
true_log_likelihood <- log_likelihood(true_beta, x = x, y = y)
points(
    true_beta, true_log_likelihood,
    col = "blue", pch = 19
)
```



1d.

```r
optimize(log_likelihood, interval = c(-3, 3), maximum = TRUE, x = x, y = y)
```

```
## $maximum
## [1] 0.4624022
##
## $objective
## [1] -1778.088
```

1e.

```r
set.seed(42)

mle_estimates <- numeric(1000)
ols_all_estimates <- numeric(1000)
ols_non_censored_estimates <- numeric(1000)

for (i in 1:1000) {
    simulated_data <- simulate(1000)
    x <- simulated_data$x
    y <- simulated_data$y

    mle_estimates[i] <- optimize(
        log_likelihood,
        interval = c(-3, 3), maximum = TRUE, x = x, y = y
    )$maximum

    ols_all_estimates[i] <- lm(y ~ x)$coefficients[2]

    ols_non_censored_estimates[i] <- lm(y[y > 0] ~ x[y > 0])$coefficients[2]
}
```

```r
mle_bias <- mean(mle_estimates) - true_beta
ols_all_bias <- mean(ols_all_estimates) - true_beta
ols_non_censored_bias <- mean(ols_non_censored_estimates) - true_beta

print(c(mle_bias, ols_all_bias, ols_non_censored_bias))
```

```
## [1] -5.210238e-05 -1.230968e-01 -2.368117e-01
```

Clearly the MLE produces the least biased estimates. When comparing the two OLS estimates, the one that only uses all observations is less biased than the one that uses only the non-censored observations. This is because the non-censored observations are a subset of the full sample, and thus the OLS estimate using only the non-censored observations has a smaller sample size and is thus less precise.