ECON 710B - Problem Set 12

Alex von Hafften*

4/27/2021

Exercise 27.1

Derive (27.2) and (27.3). Hint: Use Theorems 5.7 and 5.8 of Introduction to Econometrics.

From pg. 849 of Hansen (on Censored Regression Functions):

The conditional mean of the uncensored, censored, and truncated distributions are

$$
\begin{aligned}
m^*(X) &= E[Y^*|X] = X'\beta \\
m(X) &= E[Y|X] = X'\beta\Phi(\frac{X'\beta}{\sigma}) + \sigma\phi(\frac{X'\beta}{\sigma}) && (27.2) \\
m^{\#}(X) &= E[Y^{\#}|X] = X'\beta + \sigma\lambda(\frac{X'\beta}{\sigma}) && (27.3)
\end{aligned}
$$

The function $\lambda(x) = \phi(x)/\Phi(x)$ in (27.3) is called the inverse Mills ratio.

Theorem 5.8.4: If $X \sim N(\mu, \sigma^2)$ then for $c^* = (c-\mu)/\sigma$, $E[X\mathbb{1}\{X > c\}] = \mu(1-\Phi(c^*)) + \sigma\phi(c^*)$.

Theorem 5.8.6: If $X \sim N(\mu, \sigma^2)$ then for $c^* = (c-\mu)/\sigma$, $E[X|X > c] = \mu + \sigma(-c^*)$.

$$
\begin{aligned}
m(X) &= E[Y|X] \\
&= E[Y^*\mathbb{1}\{Y^* > 0\}|X] \\
&= E[(X'\beta + e)\mathbb{1}\{X'\beta + e > 0\}|X] \\
&= X'\beta E[\mathbb{1}\{\frac{e}{\sigma} > \frac{-X'\beta}{\sigma}\}|X] + E[e\mathbb{1}\{e > -X'\beta\}|X] \\
&= X'\beta\Phi(\frac{X'\beta}{\sigma}) + (0)(1-\Phi(e/\sigma)) + \sigma\phi(-X'\beta/\sigma) \\
&= X'\beta\Phi(\frac{X'\beta}{\sigma}) + \sigma\phi(\frac{X'\beta}{\sigma})
\end{aligned}
$$

*This problem set is for ECON 710B Econometrics taught by Bruce Hansen with assistance from Jian Zhang at UW-Madison. Exercises come from Bruce's textbook. I worked on this problem set with a study group of Michael Nattinger, Andrew Smith, and Ryan Mather. I also discussed problems with Sarah Bass, Emily Case, Danny Edgel, and Katherine Kwok.

$$
\begin{aligned}
m^{\#}(X) &= E[Y^{\#}|X] \\
&= E[Y^*|X, Y^* > 0] \\
&= E[X'\beta + e|X, X'\beta + e > 0] \\
&= X'\beta + E[e|e > -X'\beta] \\
&= X'\beta + (0) + \sigma\lambda(X'\beta) \\
&= X'\beta + \sigma\lambda(\frac{X'\beta}{\sigma})
\end{aligned}
$$

Exercise 27.2

Take the model:

$$
\begin{aligned}
Y^* &= X'\beta + e \\
e &\sim N(0, \sigma^2) \\
Y &= \begin{cases} Y* & \text{if } Y^* \leq \tau \\ \text{missing} & \text{if } Y^* > \tau \end{cases}
\end{aligned}
$$

In this model, we say that Y is capped from above. Suppose you regress Y on X. Is OLS consistent for β? Describe the nature of the effect of the mis-measured observation on the OLS estimator.

No, OLS is not consistent for β. The OLS estimator is bias downward. This is a truncated regression with the truncation at τ and above opposed to zero and below. In truncated regressions (with truncation below), OLS is biased upward.

Exercise 27.4

For the censored conditional mean (27.2) propose a NLLS estimator of (β, σ).

$$
(\hat{\beta}, \hat{\sigma})' = \arg\min_{(\beta,\sigma)} \sum_{i=1}^{n} \left(Y_i - X_i'\beta\Phi(\frac{X_i'\beta}{\sigma}) - \sigma\phi(\frac{X_i'\beta}{\sigma}) \right)^2
$$

Exercise 27.8

Show (27.7).

From pg. 856 of Hansen (about the Heckman Model): Based on the same calculations as discussed in the previous section, the conditional mean of Y in the selected sample is

$$E[Y|X, Z, S = 1] = X'\beta + \sigma_{21}\lambda(Z'\gamma)$$

where $\lambda(x)$ is the inverse Mills ratio.

Notice that the projection of e on u: $e = \frac{\sigma_{21}}{1}u + \varepsilon = \sigma_{21}u + \varepsilon$:

$$\begin{aligned} E[Y|X, Z, S = 1] &= X'\beta + E[e|X, Z, S = 1] \\ &= X'\beta + E[e|S^* > 0] \\ &= X'\beta + E[e|u > -Z'\gamma] \\ &= X'\beta + \sigma_{21}\lambda(Z'\gamma) \end{aligned}$$

Exercise 27.9

Take the `CHJ2004` dataset. The variables *tinkind* and *income* are household transfers received in-kind and household income, respectively. Divide both variables by 1000 to standardize. Create the regressor $Dincome = (income - 1) \times \mathbb{1}\{income > 1\}$.

```
CHJ2004 <- read_dta("CHJ2004.dta")

data_27_9 <- CHJ2004 %>%
  transmute(income = income / 1000,
            tinkind = transfers / 1000,
            Dincome = (income - 1) * as.numeric(income > 1))
```

(a) Estimate a linear regression of `tinkind` on `income` and `Dincome`. Interpret the results.

```
ols <- lm(tinkind ~ income + Dincome, data = data_27_9)

coeftest(ols, vcov. = vcovHC, type = "HC1")
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  49.7156     3.8095  13.050 < 2.2e-16 ***
## income      -42.6706     3.8266 -11.151 < 2.2e-16 ***
## Dincome      42.6725     3.8270  11.150 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(b) Calculate the percentage of censored observations (the percentage for which $tinkind = 0$. Do you expect censoring bias to be a problem in this example?

```
sum(data_27_9$tinkind == 0) / nrow(data_27_9)
```

```
## [1] 0.1197605
```

(c) Suppose you try and fix the problem by omitting the censored observations. Estimate the regression on the subsample of observations for which $tinkind > 0$.

```
ols <- data_27_9 %>%
  filter(tinkind > 0) %>%
  lm(tinkind ~ income + Dincome, data = .)

coeftest(ols, vcov. = vcovHC, type = "HC1")
```

```
##
## t test of coefficients:
##
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  50.5303     3.7741  13.389 < 2.2e-16 ***
## income      -42.8158     3.8026 -11.260 < 2.2e-16 ***
## Dincome      42.8563     3.8035  11.268 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(d) Estimate a Tobit regression of *tinkind* on *income* and *Dincome*.

```
tobit(tinkind ~ income + Dincome, data = data_27_9) %>% summary()
```

```
##
## Call:
## tobit(formula = tinkind ~ income + Dincome, data = data_27_9)
##
## Observations:
##          Total  Left-censored     Uncensored Right-censored
##           8684           1950           6734              0
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  49.833629   3.064117   16.26   <2e-16 ***
## income      -44.999705   3.094310  -14.54   <2e-16 ***
## Dincome      44.965210   3.095011   14.53   <2e-16 ***
## Log(scale)    3.135919   0.008797  356.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Scale: 23.01
##
## Gaussian distribution
## Number of Newton-Raphson Iterations: 3
## Log-likelihood: -3.212e+04 on 4 Df
## Wald-statistic: 256.3 on 2 Df, p-value: < 2.22e-16
```

(e) Estimate the same regression using CLAD.

```
crq(Surv(data_27_9$tinkind, 0) ~ data_27_9$income,
    tau = 0.5)
```

(f) Interpret and explain the differences between your results in (a)-(e).

...

Exercise 28.12

Using the `cps09mar` dataset perform an analysis similar to that presented in Section 28.18 but instead use the sub-sample of Hispanic women. This sample has 3003 observations. Which models are selected by BIC, AIC, CV and FIC? (The precise information criteria you examine may be limited depending on your software.) How do you interpret the results? Which model/estimate would you select as your preferred choice? [1]

```
cps09mar <- read_delim(file = "cps09mar.txt",
                 delim = "\t",
                 col_names = c("age", "female", "hisp", "education", "earnings", "hours",
                               "week", "union", "uncov", "region", "race", "marital"),
                 col_types = cols())

sample <- cps09mar %>%
  filter(female == 1, hisp == 1) %>%
  mutate(l_wage = log(earnings/hours/week),
         experience = age - education - 6,
         college = education >= 16,
         education_9 = case_when(education > 9 ~ education, TRUE ~ 0),
         education_dummies = case_when(education %in%  c(12, 13, 14, 16, 18, 20)~education,
                                       TRUE ~ 0),
         married = marital == 7)

dim(sample)
```

[1] 3003 18

```
model_1 <- lm(l_wage ~ experience + I(experience^2) +
                college +
                as.factor(region) + married,
              data = sample)
model_2 <- lm(l_wage ~ experience + I(experience^2) +
                education + education_9 +
                as.factor(region) + married,
              data = sample)
model_3 <- lm(l_wage ~ experience + I(experience^2) +
                as.factor(education_dummies) +
                as.factor(region) + married, data = sample)
model_4 <- lm(l_wage ~ experience + I(experience^2) + I(experience^3) + I(experience^4) +
                college +
                as.factor(region) + married,
              data = sample)
model_5 <- lm(l_wage ~ experience + I(experience^2) + I(experience^3) + I(experience^4) +
                education + education_9 +
                as.factor(region) + married,
              data = sample)
model_6 <- lm(l_wage ~ experience + I(experience^2) + I(experience^3) + I(experience^4) +
                as.factor(education_dummies) +
                as.factor(region) + married, data = sample)
model_7 <- lm(l_wage ~ experience + I(experience^2) + I(experience^3) + I(experience^4) +
                I(experience^5) + I(experience^6) + college +
                as.factor(region) + married,
              data = sample)
model_8 <- lm(l_wage ~ experience + I(experience^2) + I(experience^3) + I(experience^4) +
```

[1] You only need to compare AIC and BIC selection.

```
                I(experience^5) + I(experience^6) + education + education_9 +
                as.factor(region) + married,
              data = sample)
model_9 <- lm(l_wage ~ experience + I(experience^2) + I(experience^3) + I(experience^4) +
                I(experience^5) + I(experience^6) + as.factor(education_dummies) +
                as.factor(region) + married, data = sample)

predicted_df <- tibble(experience = c(0, 30),
                       education = 0,
                       education_9 = 0,
                       education_dummies = 0,
                       college = TRUE,
                       region = 1L,
                       married = TRUE)

return_df <- rbind(predict(model_1, newdata = predicted_df),
      predict(model_2, newdata = predicted_df),
      predict(model_3, newdata = predicted_df),
      predict(model_4, newdata = predicted_df),
      predict(model_5, newdata = predicted_df),
      predict(model_6, newdata = predicted_df),
      predict(model_7, newdata = predicted_df),
      predict(model_8, newdata = predicted_df),
      predict(model_9, newdata = predicted_df)) %>%
  as_tibble() %>%
  rename(l_wage_0 = `1`,
         l_wage_30 = `2`) %>%
  mutate(wage_0 = exp(l_wage_0),
         wage_30 = exp(l_wage_30),
         return = (wage_30-wage_0)/wage_0*100)

tibble(education = rep(c("College", "Spline", "Dummy"), 3),
       experience = rep(c(2, 4, 6), each = 3),
       return = return_df$return,
       aic = c(AIC(model_1), AIC(model_2), AIC(model_3),
               AIC(model_4), AIC(model_5), AIC(model_6),
               AIC(model_7), AIC(model_8), AIC(model_9)),
       bic = c(BIC(model_1), BIC(model_2), BIC(model_3),
               BIC(model_4), BIC(model_5), BIC(model_6),
               BIC(model_7), BIC(model_8), BIC(model_9))) %>%
  kable(digits = 0)
```

education	experience	return	aic	bic
College	2	24	4763	4817
Spline	2	39	4482	4542
Dummy	2	41	4392	4476
College	4	40	4755	4821
Spline	4	56	4471	4543
Dummy	4	58	4382	4478
College	6	41	4759	4837
Spline	6	60	4475	4559
Dummy	6	59	4385	4494

The model with the lowest AIC is the model with education dummies and 4 powers of experience. The model with the lowest BIC is the model with education dummies and 2 powers of experience. I think my preferred model is the latter. It seems like educations dummies provide a lot of value. It seems like the return to experience is on the order of 50%.