ECON 710B - Problem Set 9

Alex von Hafften*

4/6/2021

Exercise 20.1

Take the estimated model:

$$Y = -1 + 2X + 5(X-1)\mathbb{1}\{X \geq 1\} - 3(X-2)\mathbb{1}\{X \geq 2\} + e.$$

What is the estimated marginal effect of X on Y for X = 3?

For $X = 3 \implies \mathbb{1}\{X \geq 1\} = \mathbb{1}\{X \geq 2\} = 1$, so

$$Y = -1 + 2X + 5(X-1) - 3(X-2) + e = 4X + e$$

$$\implies \frac{\partial Y}{\partial X} = 4$$

*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.

Exercise 20.3

Take the linear spline from the previous question:

$$m_K(x) = \beta_0 + \beta_1 x + \beta_2(x-\tau_1)\mathbb{1}\{x \geq \tau_1\} + \beta_3(x-\tau_2)\mathbb{1}\{x \geq \tau_2\} + \beta_4(x-\tau_3)\mathbb{1}\{x \geq \tau_3\}$$

Find the (inequality) restrictions on the coefficients β_j so that $m_K(x)$ is concave.

The slopes of the splines:

$x \in$	slope
$(-\infty, \tau_1]$	β_1
$(\tau_1, \tau_2]$	$\beta_1 + \beta_2$
$(\tau_2, \tau_3]$	$\beta_1 + \beta_2 + \beta_3$
$(\tau_3, -\infty)$	$\beta_1 + \beta_2 + \beta_3 + \beta_4$

To be concave, the following inequalities need to hold:

$$\beta_1 \geq \beta_1 + \beta_2 \geq \beta_1 + \beta_2 + \beta_3 \geq \beta_1 + \beta_2 + \beta_3 + \beta_4$$

These inequalities imply:

$$\beta_2 \leq 0, \beta_3 \leq 0, \beta_4 \leq 0$$

Exercise 20.11

Take the `cps09mar` dataset (full sample).

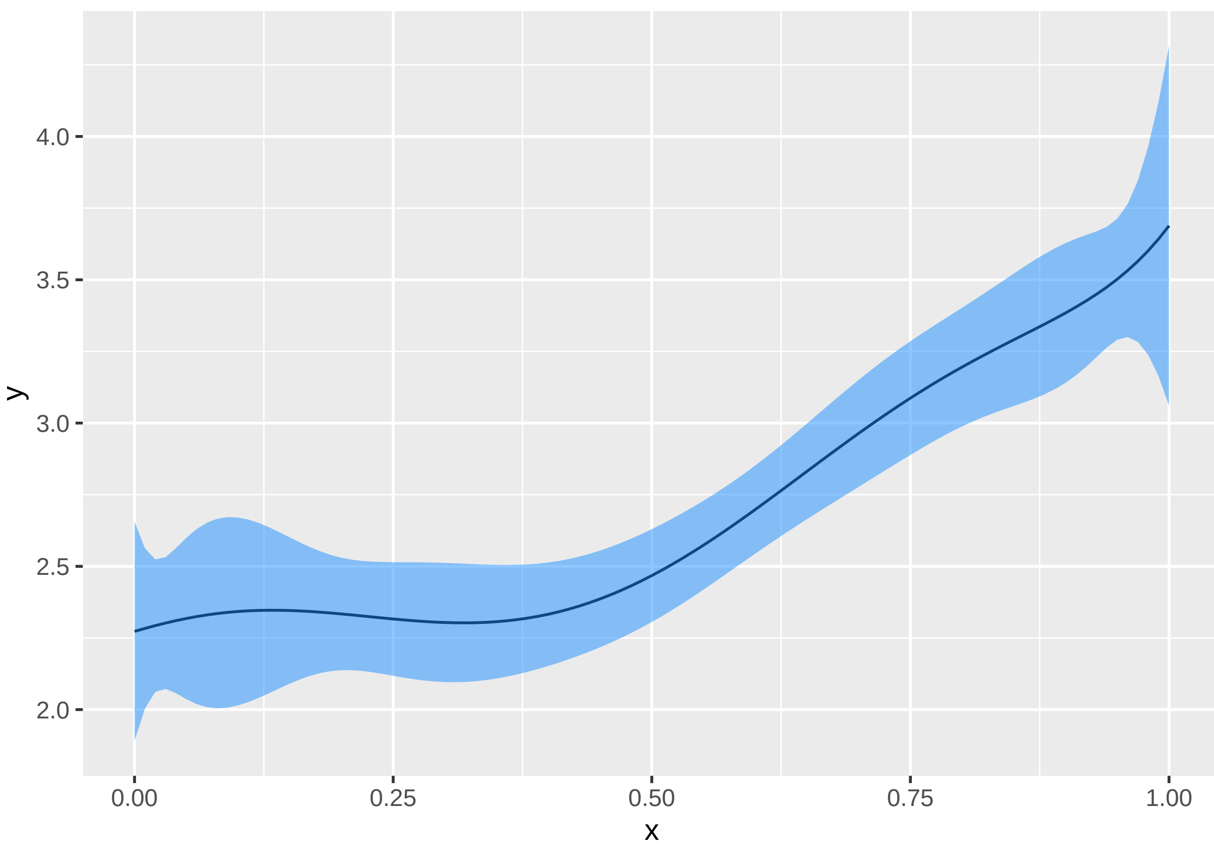
(a) Estimate a 6th order polynomial regression of $\log(wage)$ on *education*. To reduce the ill-conditioned problem first rescale education to lie in the interval $[0, 1]$.

```
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()) %>%
  mutate(education_r = education/max(education),
         l_wage = log(earnings / (hours * week))) %>%
  arrange(education_r)

lm_1 <- lm(l_wage ~ education_r, data = cps09mar)

lm_2 <- lm(l_wage ~ education_r + I(education_r^2), data = cps09mar)

lm_3 <- lm(l_wage ~ education_r + I(education_r^2) + I(education_r^3), data = cps09mar)

lm_4 <- lm(l_wage ~ education_r + I(education_r^2) + I(education_r^3) +
             I(education_r^4), data = cps09mar)

lm_5 <- lm(l_wage ~ education_r + I(education_r^2) + I(education_r^3) +
             I(education_r^4) + I(education_r^5), data = cps09mar)

lm_6 <- lm(l_wage ~ education_r + I(education_r^2) + I(education_r^3) +
             I(education_r^4) + I(education_r^5) + I(education_r^6), data = cps09mar)

stargazer(lm_1, lm_2, lm_3, lm_4, lm_5, lm_6,
          header = FALSE, float = FALSE,
          omit.stat = c("f", "ser", "rsq"))
```

	Dependent variable:					
	l_wage					
	(1)	(2)	(3)	(4)	(5)	(6)
education_r	2.164***	0.230**	−2.268***	−1.631***	4.863***	1.050
	(0.020)	(0.106)	(0.290)	(0.570)	(0.965)	(1.801)
I(education_r^2)		1.420***	6.013***	3.892**	−36.294***	−2.102
		(0.077)	(0.502)	(1.706)	(5.116)	(14.564)
I(education_r^3)			−2.546***	0.135	93.769***	−22.855
			(0.275)	(2.081)	(11.430)	(47.895)
I(education_r^4)				−1.156	−94.906***	94.623
				(0.889)	(11.288)	(76.424)
I(education_r^5)					33.998***	−113.848*
					(4.081)	(59.104)
I(education_r^6)						44.548**
						(17.766)
Constant	1.440***	2.071***	2.456***	2.406***	2.251***	2.273***
	(0.014)	(0.037)	(0.056)	(0.068)	(0.070)	(0.071)
Observations	50,742	50,742	50,742	50,742	50,742	50,742
Adjusted R^2	0.193	0.198	0.200	0.200	0.201	0.201
Note:						*p<0.1; **p<0.05; ***p<0.01

(b) Plot the estimated regression function along with 95% pointwise confidence intervals.

```
# values to plot estimated regression function
x <- seq(0, 1, by = .01)
x_k <- cbind(1, x, x^2, x^3, x^4, x^5, x^6)

# Estimating point-wise confidence intervals
omega <- 0
for (i in 1:nrow(x_k))  omega <- omega + x_k[i, ] %*% t( x_k[i, ] ) * lm_6$residuals[i]^2

meat <- solve(t(x_k) %*% x_k) %*% omega %*% solve(t(x_k) %*% x_k)

v_hat_x <- NULL
for (i in 1:length(x)) {
  x_k_x <- c(1, x[i], x[i]^2, x[i]^3, x[i]^4, x[i]^5, x[i]^6)
  v_hat_x <- c(v_hat_x, t(x_k_x) %*% meat %*% x_k_x)
}

tibble(x, v_hat_x) %>%
  mutate(y = as.numeric(x_k %*% lm_6$coefficients),
         ci_lower = y - 1.96 *sqrt(v_hat_x),
         ci_upper = y + 1.96 *sqrt(v_hat_x)) %>%
  ggplot() +
  geom_line(aes(x = x, y=y)) +
  geom_ribbon(aes(x=x, ymin = ci_lower, ymax = ci_upper), fill = "dodgerblue", alpha=0.5)
```



Exercise 20.15

The `RR2010` dataset is from Reinhart and Rogoff (2010). It contains observations on annual U.S. GDP growth rates, inflation rates, and the debt/gdp ratio for the long time span 1791-2009. The paper made the strong claim that gdp growth slows as debt/gdp increases, and in particular that this relationship is nonlinear with debt negatively affecting growth for debt ratios exceeding 90%. Their full dataset includes 44 countries, our extract only includes the United States. Let Y_t denote GDP growth and let D_t denote debt/gdp. We will estimate the partial linear specification

$$Y_t = \alpha Y_{t-1} + m(D_{t-1}) + e_t$$

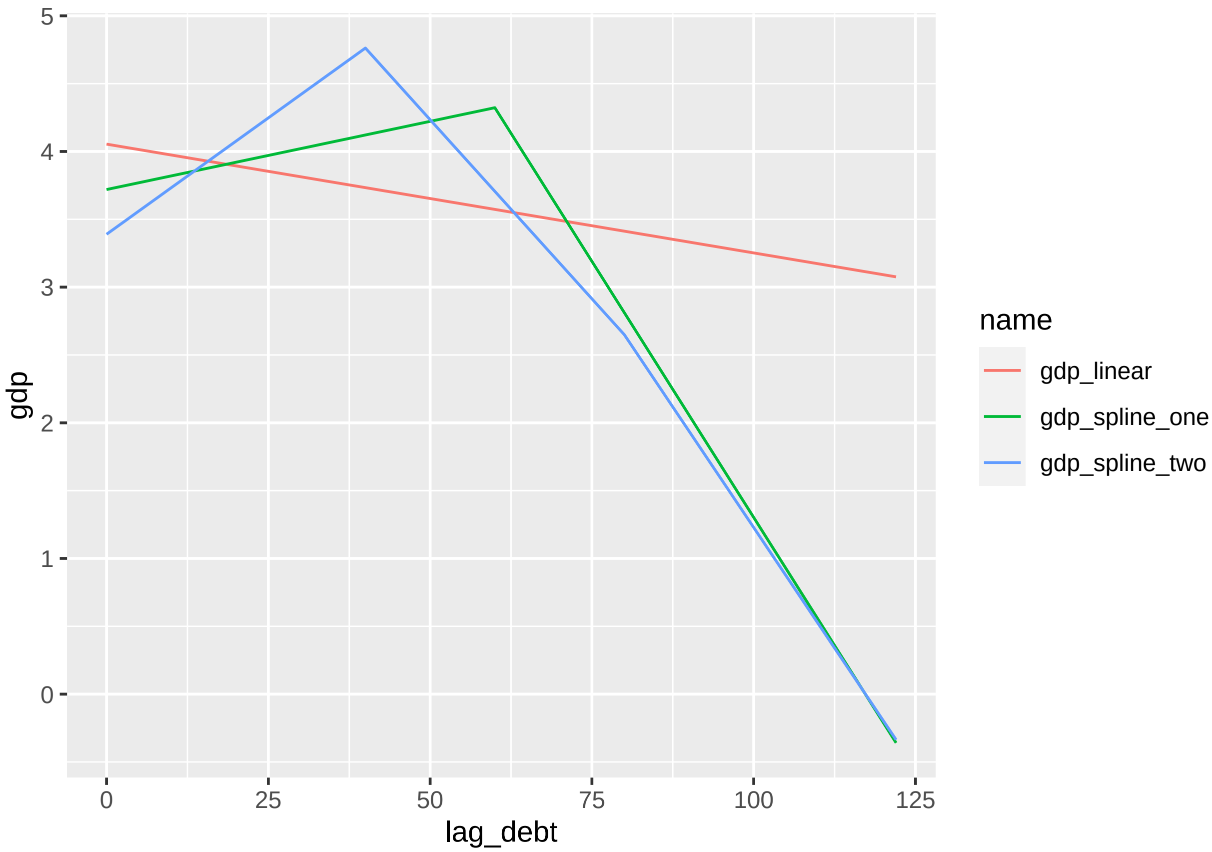
using a linear spline for $m(D)$.

(a) Estimate (1) linear model; (2) linear spline with one knot at $D_{t-1} = 60$; (3) linear spline with two knots at 40 and 80. Plot the three estimates.

```
RR2010 <- read_delim("RR2010.txt", delim = "\t", col_types = cols()) %>%
  mutate(debt40 = (debt - 40) * as.numeric(debt >= 40),
         debt60 = (debt - 60) * as.numeric(debt >= 60),
         debt80 = (debt - 80) * as.numeric(debt >= 80))

linear <- lm(gdp ~ lag(gdp) + lag(debt), data = RR2010)
spline_one <- lm(gdp ~ lag(gdp) + lag(debt) + lag(debt60), data = RR2010)
spline_two <- lm(gdp ~ lag(gdp) + lag(debt) + lag(debt40) + lag(debt80), data=RR2010)

stargazer(linear, spline_one, spline_two,
          header = FALSE, float =  FALSE, omit.stat = c("f", "ser", "rsq"))
```

	Dependent variable:		
	gdp		
	(1)	(2)	(3)
lag(gdp)	0.300***	0.284***	0.280***
	(0.065)	(0.066)	(0.066)
lag(debt)	−0.008	0.010	0.034
	(0.012)	(0.016)	(0.024)
lag(debt60)		−0.086*	
		(0.049)	
lag(debt40)			−0.087
			(0.056)
lag(debt80)			−0.018
			(0.104)
Constant	2.898***	2.628***	2.312***
	(0.515)	(0.536)	(0.592)
Observations	218	218	218
Adjusted R^2	0.085	0.094	0.098
Note:	*p<0.1; **p<0.05; ***p<0.01		

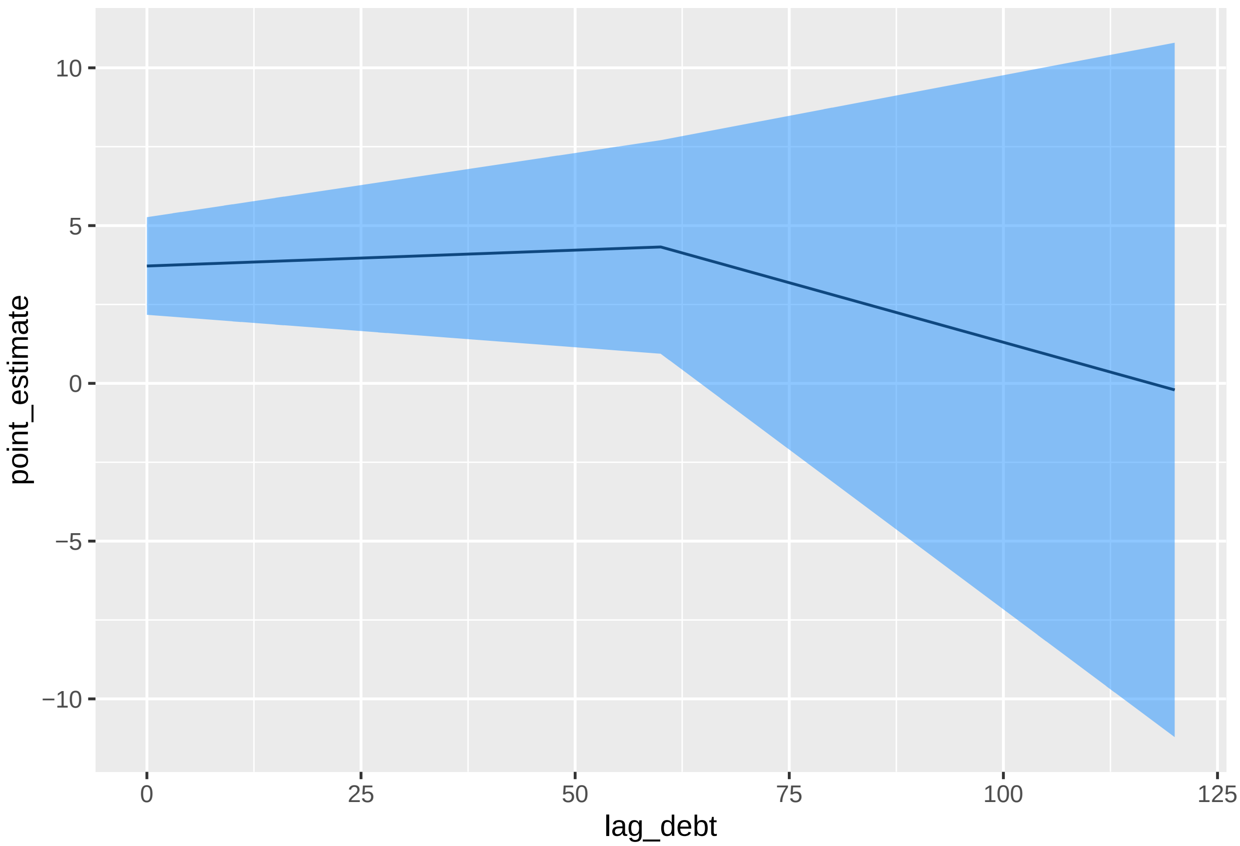
```r
# plot
lag_gdp <- mean(RR2010$gdp)

fitted <- tibble(lag_debt= 0:122) %>%
  mutate(debt40 = (lag_debt - 40) * as.numeric(lag_debt >= 40),
         debt60 = (lag_debt - 60) * as.numeric(lag_debt >= 60),
         debt80 = (lag_debt - 80) * as.numeric(lag_debt >= 80),
         gdp_linear = linear$coefficients[1] + linear$coefficients[2]*lag_gdp +
           linear$coefficients[3]*lag_debt,
         gdp_spline_one = spline_one$coefficients[1]+spline_one$coefficients[2]*lag_gdp +
           spline_one$coefficients[3]*lag_debt+spline_one$coefficients[4]*debt60,
         gdp_spline_two = spline_two$coefficients[1]+spline_two$coefficients[2]*lag_gdp +
           spline_two$coefficients[3]*lag_debt+spline_two$coefficients[4]*debt40 +
           spline_two$coefficients[5]*debt80)

fitted %>%
  pivot_longer(cols = starts_with("gdp")) %>%
  ggplot() +
  geom_line(aes(x=lag_debt, y=value, color = name)) +
  ylab("gdp")
```



The plot is the fitted value of GDP growth if GDP growth was at the mean level in the prior quarter.

(b) For the model with one knot plot with 95% confidence intervals.

```
beta <- spline_one$coefficients
se <- sqrt(diag(vcov(spline_one, method = "HC1")))
se <- summary(spline_one)$coefficients[, 2]

fitted_b <- tibble(lag_debt= 0:120) %>%
  mutate(debt60 = (lag_debt - 60) * as.numeric(lag_debt >= 60),
         error = se[1] + se[2]*lag_gdp + se[3]*lag_debt + se[4]*debt60,
         point_estimate = beta[1] + beta[2]*lag_gdp + beta[3]*lag_debt + beta[4]*debt60,
         ci_lower = point_estimate - 1.96*error,
         ci_upper = point_estimate + 1.96*error)

fitted_b %>%
  ggplot() +
  geom_line(aes(x=lag_debt, y=point_estimate)) +
  geom_ribbon(aes(x=lag_debt, ymin = ci_lower, ymax = ci_upper), fill = "dodgerblue", alpha = .5)
```



(c) Compare the three splines models using either cross-validation or AIC. Which is the preferred specification?

```
AIC(linear)
```

```
## [1] 1251.774
```

```
AIC(spline_one)
```

```
## [1] 1250.711
```

```
AIC(spline_two)
```

```
## [1] 1250.688
```

We find that AIC values decrease with the number of knots. Thus, since the lower the AIC value the better, the preferred specification is the model with two knots. The model with one spline is very close to the model with two splines, which matches the plot in (b).

(d) Interpret the findings.

With adding splines, we found a better model from the AIC perspecitive. This generally supports Reinhart and Rogoff's hypothesis that there is a nonlinear relationship between gdp growth and debt/gdp.

Exercise 21.1

We have described the RDD when treatment occurs for $T = \mathbb{1}\{X \geq c\}$. Suppose instead that treatment occurs for $T = 1\{X \leq c\}$. Describe the differences (if any) involved in estimating the conditional ATE $\bar{\theta}$.

...

Exercise 21.2

Suppose treatment occurs for $T = \mathbb{1}\{c_1 \leq X \leq c_2\}$ where both c_1 and c_2 are in the interior of the support of X. What treatment affects are identified?

...

Exercise 21.3

Show that (21.1) is obtained by taking the conditional expectation as described.

. . .

Exercise 21.4

Explain why equation (21.4) estimated on the subsample for which $|X - c| \leq h$ is identical to a local linear regression with a Rectangular bandwidth.

...

Exercise 21.6

Use the datafile LM2007 on the textbook webpage. Replicate the baseline RDD estimate as reported in Table 21.1. Repeat with a bandwidth of $h = 4$ and $h = 12$. Report your estimates of the conditional ATE and standard error.

...

Exercise 21.8

Do a similar estimation as in the previous exercise, but using the dependent variable `mort_age25plus_related_postHS` (mortality due to HS-related causes in the 25+ age group).

...