Problem Set 1

Eric Karsten

January 17, 2020

Statistical and Machine Learning (25 points)

Describe in 500-800 words the difference between supervised and unsupervised learning. As you respond, consider the following few questions to guide your thinking, e.g.:

- What is the relationship between the X's and Y?
- What is the target we are interested in?
- How do we think about data generating processes?
- What are our goals in approaching data?
- How is learning conceptualized?

Linear Regression Regression (35 points)

Problem Statement

Using the mtcars dataset in R (e.g., run names(mtcars)), answer the following questions:

a. (10) Predict miles per gallon (mpg) as a function of cylinders (cyl). What is the output and parameter values for your model?

b. (5) Write the statistical form of the simple model in the previous question (i.e., what is the population regression function?).

c. (10) Add vehicle weight (wt) to the specification. Report the results and talk about differences in coefficient size, effects, etc.

d. (10) Interact weight and cylinders and report the results. What is the same or different? What are we theoretically asserting by including a multiplicative interaction term in the function?

Solution

```r
# Loading Packages
library(tidyverse)
library(stargazer)

# Pretty table courtesy of
# Hlavac, Marek (2018).
# stargazer: Well-Formatted Regression and Summary Statistics Tables.
# R package version 5.2.2. https://CRAN.R-project.org/package=stargazer

# Loading Data
data("mtcars")

# Running Regressions
m1 <- lm(mpg ~ cyl, data = mtcars)
m2 <- lm(mpg ~ cyl + wt, data = mtcars)
m3 <- lm(mpg ~ cyl * wt, data = mtcars)
```

```r
# Plotting Regressions
stargazer(m1,m2,m3, omit.stat = c("f", "ser"),
          header = F,
          covariate.labels = c("Cylinder Count", "Weight (Tons)",
                               "Cylinder:Weight Interaction", "Constant"),
          dep.var.caption = "MPG",
          dep.var.labels.include = F,
          table.placement = "H",
          column.sep.width = "0pt",
          title = "Predicting MPG from Cylinder count and Vehicle Weight")
```

Table 1: Predicting MPG from Cylinder count and Vehicle Weight

	MPG		
	(1)	(2)	(3)
Cylinder Count	-2.876^{***}	-1.508^{***}	-3.803^{***}
	(0.322)	(0.415)	(1.005)
Weight (Tons)		-3.191^{***}	-8.656^{***}
		(0.757)	(2.320)
Cylinder:Weight Interaction			0.808^{**}
			(0.327)
Constant	37.885^{***}	39.686^{***}	54.307^{***}
	(2.074)	(1.715)	(6.128)
Observations	32	32	32
R^2	0.726	0.830	0.861
Adjusted R^2	0.717	0.819	0.846
Note:			*p<0.1; **p<0.05; ***p<0.01

The regression output for parts (a), (c), and (d) is included in the table above.

a. To provide an interpretation of the regression in (a), a car with no cylinders (an ill-defined notion) would have about 38 miles per gallon, and then each aded cylinder reduces that efficiency of the car by about 2.8 miles per gallon.

b. The model that is being fit in the first regression is

$$mpg_i = \beta_0 + \beta_1 cyl_i + \varepsilon_i$$

where ε_i is a mean zero error term that is independent of the number of cylinders.

c. Regression (2) in the table above reflects a specification where we add in a coefficient for weight. We notice that the cylinder count coefficient decreases when we add the weight coefficent, an indication that weight and cylinder count move together to some extent. We also note that both coefficents are negative. This means that cars with more cylinders tend to be less efficient and that the marginal cylincer reduces MPG by about 1.5. We also see that cars that weight more tend to be less efficient, that is a car that is one ton heavier than a car with an equal number of cylinders will tend to be 3 MPG less efficient. The intercept coefficent increases slightly under this specifiaction.

d. Regression (3) in the table above reflects the specification with an interaction term. By including the interaction term, we are asserting that there is something non-linear in the effects of cylinder count and weight on vehichle milage. That is, we are making the claim that the effect of incresed weight and increased cylinder count may not move independently of one another, but rather that the magnitude of the effect of an increase in cylinder count on milage depends on the weight of the vehicle and vice versa. We notice when we include this term in the regression that the clinder count effect and weight effect increase dramatically in magnitude (but remain negative). The interaction term is positive indicating that heavier cars will have a smaller reduction in MPG due to an increase in cyclinder count and than lighter cars would (or that cars with more cylinders will have less of a reduction in MPG due to an increase in weight than those with fewer cylinders would have). We also notice that in reaction to the increase in the magnitude of the slope coefficients relative to prior regressions, this regression has a much large intercept coefficient. This is reasonable because the intercept is the centering term, but it is a centering term for a car weighing nothing and with no cylincers, so it does not really have a well-defined interpetation since such a car doesn't exist.

Non-linear Regression (40 points)

1. Using the wage_data file, answer the following questions:

a. (10) Fit a polynomial regression, predicting wage as a function of a second order polynomial for age. Report the results and discuss the output.

```
wage <- read_csv("wage_data.csv")

w1 <- lm(wage ~ age, data = wage)
w2 <- lm(wage ~ age + I(age^2), data = wage)

# Plotting Regressions
stargazer(w1, w2, omit.stat = c("f", "ser"),
          header = F,
          covariate.labels = c("Age", "Age$^2$", "Constant"),
          dep.var.caption = "Wage",
          dep.var.labels.include = F,
          table.placement = "H",
          column.sep.width = "0pt",
          title = "Predicting Wages as a function of Age")
```

Table 2: Predicting Wages as a function of Age

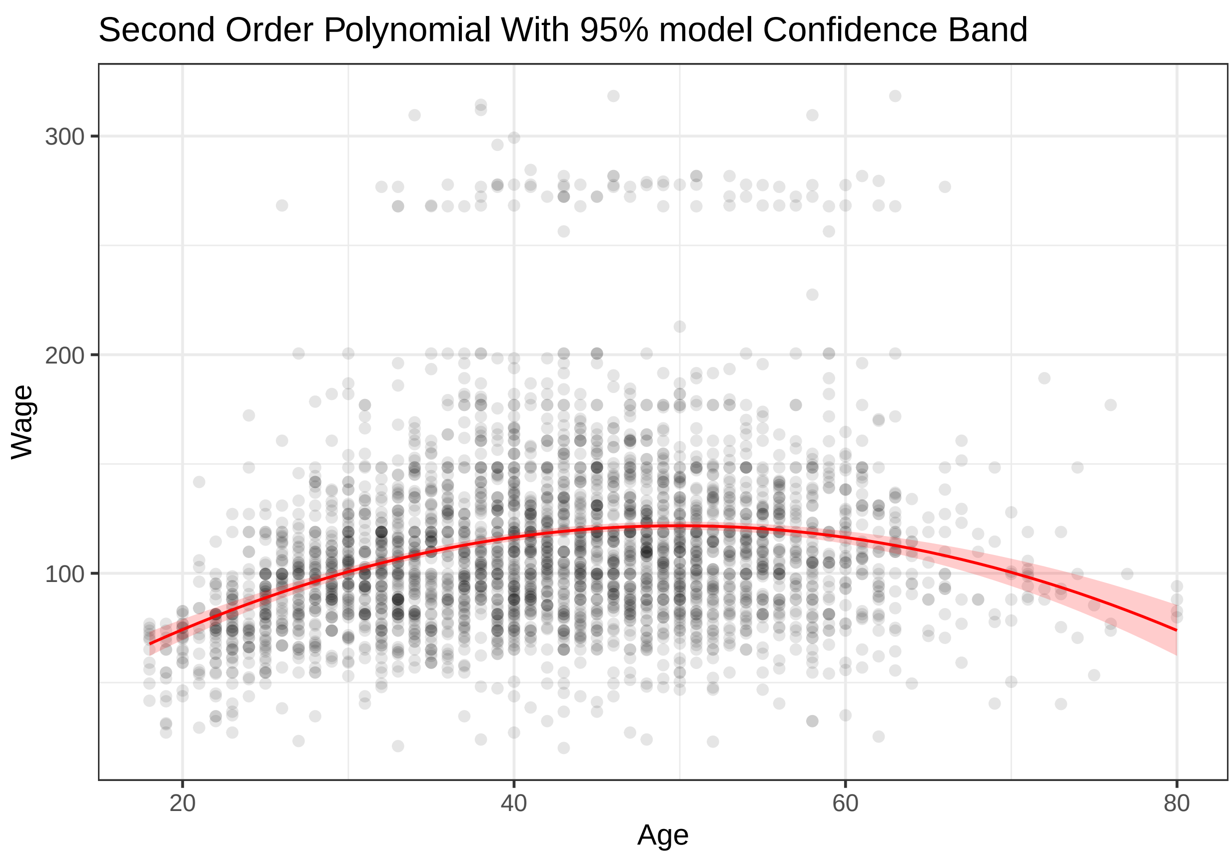
	Wage	
	(1)	(2)
Age	0.707***	5.294***
	(0.065)	(0.389)
Age^2		−0.053***
		(0.004)
Constant	81.705***	−10.425
	(2.846)	(8.190)
Observations	3,000	3,000
R^2	0.038	0.082
Adjusted R^2	0.038	0.081
Note:	*p<0.1; **p<0.05; ***p<0.01	

In the table above, equation (2) fits a second order polynomial for wages as a function of age. We see that the coefficients for both Age and Age^2 are significantly different from zero. Addionally, we see that there is a serious bump in predictive power from a naive linear model (equation (1) reported as a reference). That said, the function overall doesn't have a lot of prectivive power, explaining only 8% of the variance in wages. This isn't surprising because we know that structutally, there is a lot more that goes into wage than just age, so we wouldn't expect to get fantastic predictive power from such a simple and plainly incorrect model.

b. (10) Plot the function with 95% confidence interval bounds.

```
# Using Predict Function to give us confidence intervals:
predictions <-
  predict(w2, newdata = tibble(age = 18:80), interval = 'confidence') %>%
  as_tibble() %>%
  mutate(age = 18:80)

# Plotting the function with confidence intervals as well as original data
wage %>%
  ggplot(aes(x = age, y = wage)) +
  geom_point(alpha = .1) +
  geom_line(data = predictions, aes(x = age, y = fit), color = "red") +
  geom_ribbon(data = predictions, aes(x = age, y = fit, ymin = lwr, ymax = upr),
              alpha = .2, fill = "red") +
  theme_bw() +
  labs(x = "Age",
       y = "Wage",
       title = "Second Order Polynomial With 95% model Confidence Band")
```



c. (10) Describe the output. What do you see substantively? What are we asserting by fitting a polynomial regression?

In the figure above, we see that our model isn't doing an particularly amazing job of fitting the data. As I have already discussed, there are a lot of things that affect wages aother than age, and so it's natural that we see some unexplained within-age variation. We see above that the model slopes upwards early in peoples careeers, peaking around the age of 50 and then declining through retirement. This model is somewhat imperfect in that it assumes (by being quadratic and thus symmetrical) that wages decline in retirement in the same way they rise during the early career stage. This is not necessarily a plausible assumption of the functional form of the polynomial regression. We might obtain a better prediction by using a regression discontinuity design at the retirement age of 65, this bring snew information into our model that is informed by the structure of the world. Additionally, we see that the error band on our estimate widens out in the early career and in the late career. This makes sense because there are fewer data points in this area to tightly calibrate the model. A finial note is that we are plotting a 95% confidence band for the model, not a 95% confidence band for the data (which would be much much wider given the lack of predictive power of age).

d. (10) How does a polynomial regression differ both statistically and substantively from a linear regression (feel free to also generalize to discuss broad differences between non-linear and linear regression)?

The most obvious difference between to two models is one of flexibility to fit the data. By virtue of having a non-linear parameter, the polynomial regression is able to fit a curve closer to data that is clearly not purely linear. A linear model is of course constrained to being a line. Statistically, the linear model is easier to interpret because we can say the coefficeint represents the average ernings bump every year. On the other hand, the coefficients in the non-linear regression don't have as clean estimates. On the other hand non-linear regression can have higher predictive power (assuming it has not been overfit to the data), so the kind of model you fit depends on what you are trying to do. If you wan't to interpret average structural parameters, the linear regression is a good tool, on the other hand, if you want good predctions of the effect of age and the parameter of interest is something else, then a model that is non-linear in age is superior because it won't spit out absurd estimates for older workers.