Specification

Introduction

Now we are going to turn our attention to Assumption 1 (Linearity), which says that

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

We're not going to *relax* this assumption so much as *build on it* in order to get more from our regressions

Specification refers to the variables included in a regression, and the way that they are included

As we'll see, we can use econometric specification techniques to learn even more from our regrssions

Dummy variables

A **dummy** variable (aka **indicator** or **binary** variable) takes the value one if a condition is true and zero otherwise

E.g.s

- $female_i$ is a dummy variable for whether person i is a woman

- $south_i$ is a dummy variable for whether person i lives in the South

- $college_i$ is a dummy variable for whether person i graduated from college

Here is why dummy variables are useful

Suppose d_i is a dummy variable, and consider the regression

$$y = \alpha + \beta_1 x + \beta_2 d + \varepsilon$$

The prediction for units with $d = 1$ is

$$\widehat{y}_i\big|_{d=1} = a + b_1 x_i + b_2$$

The prediction for units with $d = 0$ is

$$\widehat{y}_i\big|_{d=0} = a + b_1 x_i$$

So the *difference* in predictions between those with $d = 1$ and those with $d = 0$ is

$$\hat{y}_i\big|_{d=1} - \hat{y}_i\big|_{d=0} = (a + b_1 x_i + b_2) - (a + b_1 x_i) = b_2$$

To summarize, the coefficient on the dummy tells us the difference in preditions for those where the dummy is true relative to those where it isn't

E.g.: Suppose we estimate the model

$$\widehat{wage} = a + b \cdot south$$

b represents the difference in predicted wages for those who live in the South and those who live in the North

The dummy variable trap

You might think, why only include a variable for living in the South? Don't we want to know wages for people living in the North, too?

If you tried to estimate the model

$$\widehat{wage} = a + b_1 \cdot south + b_2 \cdot north$$

You would run into a **perfect multicollinearity** problem because $south = 1 - north$ (any time $south = 1, north = 0$, and vice versa)

Recall that when we have perfect multicollinearity, we can't even estimate the regression coefficients

The **dummy variable trap** is when dummies for a set of categories are perfectly correlated

The solution to this problem is to *omit* the dummy for one category

In this case, the included dummies tell us predictions or effects relative to the **omitted** or **reference** category

The regression that only includes *south* tells us everything we need to know:

$$\widehat{wage} = a + b \cdot south$$

The prediction for people in the North is $\widehat{wage}\big|_{south=0} = a$

The prediction for people in the South is
$\widehat{wage}\big|_{south=1} = a + b$

Here, *north* is the *omitted category*, so b tells us the effect of living in the South *relative* to living in the North

Interaction terms

We can get even more out of a regression by **interacting** dummy variables with other regressors

For example, suppose we use the model

$$\widehat{wage} = a + b_1 \cdot educ + b_2 \cdot south + b_3 \cdot educ \cdot south$$

Here, $educ \cdot south$ is an **interaction term** between $south$ and $educ$

This really two regression lines in one

For those in the North, the estimated line is

$$\widehat{wage}\,\big|_{south=0} = a + b_1 \cdot educ$$

For those in the South, it's

$$\widehat{wage}\,\big|_{south=1} = a + b_1 \cdot educ + b2 + b_2 \cdot ecuc$$
$$= (a + b_2) + (b_1 + b_2)educ$$

That is, the intercept for those in the South is $(a + b_1)$ and the slope is $(b_1 + b_2)$

So b_1 is the *difference* in intercepts for those in the South, and b_2 is the difference in slopes

We could also do this by running separate regressions for those in the South and those in the North

One advantage of doing this with interaction terms is that it allows us to test hypotheses about the differences in regression lines

For example, suppose we wanted to test the null that the regression lines were the same in the South and the North

We could do this using an F-test of the joint null that $\beta_1 = 0$ and $\beta_2 = 0$ (note that these are the *population* versions of our regression estimates b_1 and b_2)

Let's look an example. In the Griliches dataset, the variable rns is a dummy for living in the South:

```
1  library(tidyverse)
2  gril <- read_csv("griliches.csv")
3  summary(gril$rns)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.     Max.
 0.0000  0.0000  0.0000  0.2691  1.0000  1.0000
```

The mean of rns is .27, which means that about 27% of the sample lives in the South

```
1  gril$s_rns <- gril$rns * gril$s
2  gril$w <- exp(gril$lw)
3  ixn.model <- lm(w ~ rns + s + s_rns, data=gril)
4  summary(ixn.model)
```

Call:
lm(formula = w ~ rns + s + s_rns, data = gril)

Residuals:
    Min   1Q Median   3Q   Max
-339.09 -86.93 -21.81   55.19 676.42

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -94.575   33.906 -2.789 0.00541 **
rns   -90.276   61.992 -1.456 0.14574
s   31.787   2.480 12.816 < 2e-16 ***
s_rns   4.707   4.607   1.022 0.30724

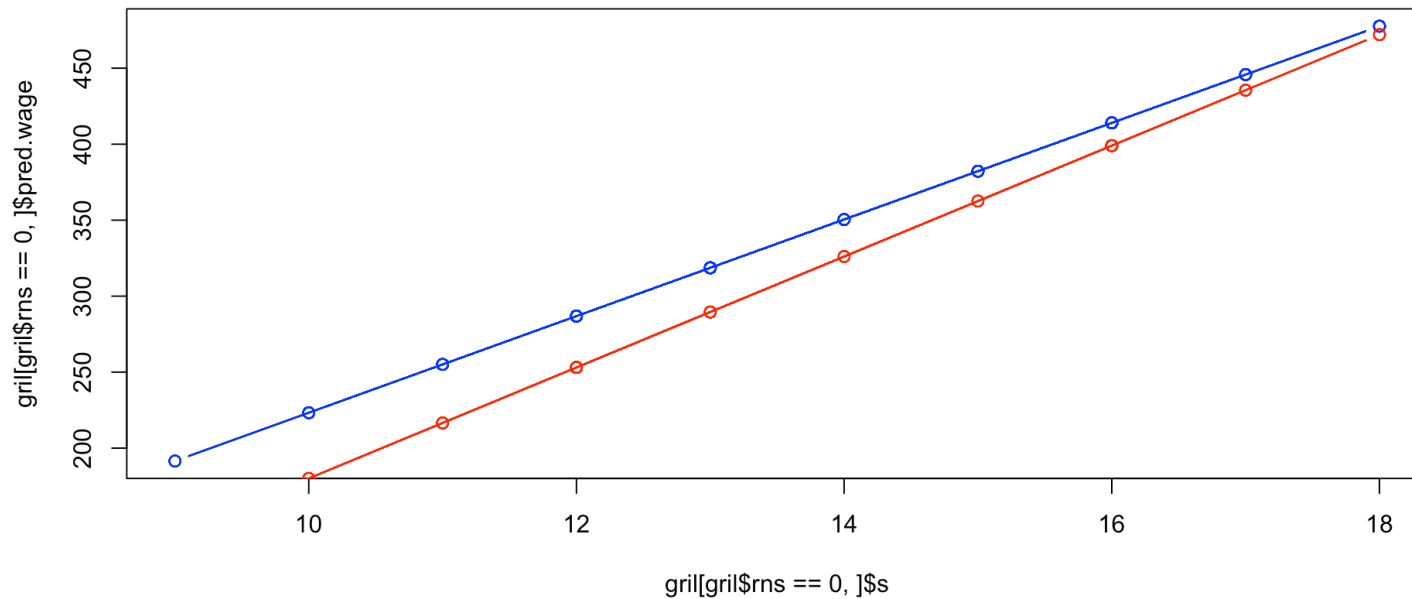Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 ' ' 0.1 ' ' 1
```

# Let's compare this to what we would get from two separate regressions:

```
1  north.model <- lm(w ~ s, data=gril[gril$rns==0,])
2  south.model <- lm(w ~ s, data=gril[gril$rns==1,])
3  south.model$coef - north.model$coef
```

```
(Intercept)            s
 -90.275615    4.707087
```

We can also plot separate regression lines for those in the North and South

```
1  gril$pred.wage <- ixn.model$fitted.values
2  plot(gril[gril$rns==0,]$s, gril[gril$rns==0,]$pred.wage, type="b", col="blu
3  lines(gril[gril$rns==1,]$s, gril[gril$rns==1,]$pred.wage, type="b", col="re
```

The lines appear to have different slopes and intercepts

However, in our interacted regression, neither $south$ nor $south \cdot educ$ were significant

So we can't reject the null that the intercepts are the same, or the null that the slopes are the same

Let's do a joint test of the null that the intercept and slope are *both* the same

```
1  library(car)
2  linearHypothesis(ixn.model, c("rns=0", "s_rns=0"))
```

```
Linear hypothesis test:
rns = 0
s_rns = 0

Model 1: restricted model
Model 2: w ~ rns + s + s_rns

  Res.Df       RSS Df Sum of Sq     F  Pr(>F)
1    756 12501021
2    754 12368706  2    132315 4.033 0.01811 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We reject the null that the lines are the same

Fixed effects

Suppose that we have **panel data**, which consists of observations on the same set of people at different points in time

Now, our observations are indexed by units and time:

$$y_{it} = \alpha + \beta x_{it} + \varepsilon_{it}$$

As usual, we might be concerned that ε contains some omitted variables that are correlated with x, in which case estimates of β will be biased and inconsistent

However, suppose that the error term can be can be decomposed into a *time-invariant* component c_i and a *time-varying* component u_{it} :

$$\varepsilon_{it} = c_i + u_{it}$$

Also suppose that only the time-invariant part c_i is correlated with the error term (in other words, the omitted variables that we're worried about don't change over time, and are contained in c_i)

In this case, we can control for these time-invariance components by including dummy variables for each individual in our sample:

$$y_{it} = \alpha + \beta x_{it} + \sum_{j=1}^{N} c_j d_{it}^j + u_{it}$$

where d_{it}^j is a dummy for whether observation it "belongs" to unit j

The unit dummies d_{it}^j are known as **fixed effects**

Intuitively, the fixed effects capture anything about a particular unit that doesn't change over time (including any omitted variables)

Fixed effects are a nice way to try to address endogeneity when you have panel data[1]

However, they only completely address endogeneity if the omitted-variables of concern are truly time invariant, which is a strong assumption

Since FEs capture *everything* that doesn't change over time, they also capture the effect of time-invariant *observable* variables, so we can't include those in a regression that has FEs

---

1. They can also be used in other situataions where there are multiple observations belonging to the same "group." For example, you could include fixed effects for the states where individuals live

I have already slyly shown you an example of fixed effects

In our example of the effect of capital punishment on crime rates, we were concerned that states might *always* have higher crime rates and more executions

To account for this, we included dummies for each state

```
1  library(haven)
2  mur <- read_dta("MURDER.dta")
3  mur.mod4 <- lm(mrdrte ~ exec + unem + as.factor(year) + as.factor(state), d
4  summary(mur.mod4)
```

Call:
lm(formula = mrdrte ~ exec + unem + as.factor(year) + as.factor(state),
    data = mur)

Residuals:
     Min   1Q   Median   3Q   Max
-26.6858 -0.6584 -0.0657   0.6747 13.3941

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.9038   3.3143   1.781 0.07796 .
exec   -0.1383   0.1770 -0.781 0.43642
unem   0.2213   0.2964   0.747 0.45701
as.factor(year)90   1.5562   0.7453   2.088 0.03939 *
as.factor(year)93   1.7332   0.7004   2.475 0.01506 *
```

The linear probability model

We've seen what happens when we have a dummy variable on the right-hand side of a regression, but what about when it's on the left-hand side?

To see how that works, note that when y is a dummy variable,

$$E(y) = P(y = 0) \cdot 0 + P(y = 1) \cdot 1 = P(y = 1)$$

Also, recall that under the assumptions of the linear regression model,

$$E(y) = \alpha + \beta x$$

Putting things together, for a binary/dummy dependent variable:

$$P(y = 1) = \alpha + \beta x$$

In other words, we are assuming that the *probability* that $y = 1$ is a function of x

Hence, we can interpret the population slope coefficient as the effect of x on $P(y = 1)$

$$\beta = \frac{\Delta P(y = 1)}{\Delta x}$$

And of course, we always interpret b as our estimate of β

To give an example of this, let's look at the file `MROZ.dta`, which has data on labor-market characteristics for a sample of 753 married women

```r
1  library(haven)
2  mroz <- read_dta("MROZ.dta")
3  summary(mroz)
```

```
      inlf                hours             kidslt6           kidsge6
 Min.   :0.0000    Min.   :   0.0    Min.   :0.0000    Min.   :0.000
 1st Qu.:0.0000    1st Qu.:   0.0    1st Qu.:0.0000    1st Qu.:0.000
 Median :1.0000    Median : 288.0    Median :0.0000    Median :1.000
 Mean   :0.5684    Mean   : 740.6    Mean   :0.2377    Mean   :1.353
 3rd Qu.:1.0000    3rd Qu.:1516.0    3rd Qu.:0.0000    3rd Qu.:2.000
 Max.   :1.0000    Max.   :4950.0    Max.   :3.0000    Max.   :8.000


      age               educ              wage             repwage
 Min.   :30.00     Min.   : 5.00     Min.   : 0.1282   Min.   :0.00
 1st Qu.:36.00     1st Qu.:12.00     1st Qu.: 2.2626   1st Qu.:0.00
 Median :43.00     Median :12.00     Median : 3.4819   Median :0.00
 Mean   :42.54     Mean   :12.29     Mean   : 4.1777   Mean   :1.85
 3rd Qu.:49.00     3rd Qu.:13.00     3rd Qu.: 4.9708   3rd Qu.:3.58
 Max.   :60.00     Max.   :17.00     Max.   :25.0000   Max.   :9.98
                                     NA's   :325
```

Let's estimate a linear model of the probability that a woman participates in the labor force as a function of whether she has kids, her education and experience, and her husband's education and wage:

```
1  lf.model <- lm(inlf ~ kidslt6 + kidsge6 + educ + exper + huseduc + huswage,
2  summary(lf.model)
```

Call:
lm(formula = inlf ~ kidslt6 + kidsge6 + educ + exper + huseduc +
    huswage, data = mroz)

Residuals:
    Min   1Q Median   3Q   Max
-1.2291 -0.4197 0.1138 0.3867 0.8378

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.152466   0.096458 -1.581 0.114383
kidslt6   -0.171099   0.032028 -5.342 1.22e-07 ***
kidsge6   0.044983   0.012973   3.467 0.000556 ***
educ   0.051182   0.009131   5.605 2.93e-08 ***
exper   0.019536   0.002176   8.977 < 2e-16 ***
```

According to this model, inter alia,

- A woman with kids less than 6 is 17% less likely to be in the labor force

- Each additional year of education increases the probability of being n the labor force by 5%

- Every additional dollar that her husband makes per hour decreases her probability of beig in the labor force by about 1%

# Differences in differences

Differences in differences (DD for short) is a way of estimating the effects of a policy

It's widely used, and all it takes is a regression with some dummy variables

We'll illustrate the idea using an example from Card and Krueger (1994)

Suppose that we want to figure out the effect of the minimum wage on employment

If you had panel data on multiple states over time, you might try running a regression like

$$emp_{it} = \alpha + \beta \cdot minwage_{it} + \varepsilon_{it}$$

But you might run into some omitted-variable-style endogeneity problems with this approach:

- Maybe states with higher minimum wages *always* tend to have higher/lower employment
- Maybe states tend to increase their minimum wages in periods when employment is higher/lower

DD is one way to address this

In 1992, New Jersey increased its minimum wage from $4.25/hour to $5.05/hour

In nearby Pennsylvania, there was no change in the minimum wage

A comparison of wages in NJ before and after the min. wage change won't necessarily tell us the effect of the min. wage, because employment might have changed for other reasons as well:

$$emp|_{NJ,Post} - emp|_{NJ,Pre} = \text{Effect of minimum wage} + \text{Other}$$

Now suppose we do the same thing for PA

Since there was no change in the min. wage there, employment could only change for other reasons, so:

$$emp|_{PA,Post} - emp|_{PA,Pre} = \text{Other factors}$$

Now, NJ and PA are neighboring states, so it seems reasonable to believe that they experience similar economic conditions

The **parallel trends** assumption holds that, if it weren't for the min. wage, employment would have changed by the same amount in NJ in PA during this period

In other words, the "other factors" affecting employment are the same in both places

Under parallel trends, we can estimate the effect of the minimum wage by looking at the *difference* between NJ in PA in *differences* in employment before and after NJ changed its min. wage:

$$(emp|_{NJ,Post} - emp|_{NJ,Pre}) - (emp|_{PA,Post} - emp|_{PA,Pre})$$
$$= (\text{Effect of min. wage} + \text{Other factors}) - (\text{Other factors})$$
$$= \text{Effect of min. wage}$$

This is the **difference in difference estimator**

We could do this "by hand," calculating the employment rate in NJ and PA in the pre and post periods

We can also use dummy variables to do this with a regression

Let

- $Post$ be a dummy for whether an observation is in the "post period" (after NJ changed its min. wage)
- $NJ$ be a dummy for whether an observation is for NJ, and
- $Post \cdot NJ$ be an interaction between the two

Now consider the regression

$$\widehat{emp} = a + b_1 \cdot Post + b_2 \cdot NJ + b_3 \cdot Post \cdot NJ$$

From what we know about dummy variables,

- $\widehat{emp}\big|_{PA,Pre} = a$
- $\widehat{emp}\big|_{PA,Post} = a + b_1$
- $\widehat{emp}\big|_{NJ,Pre} = a + b_2$
- $\widehat{emp}\big|_{NJ,Post} = a + b_1 + b_2 + b_3$

Therefore, the DD estimator is

$$(\widehat{emp}|_{NJ,Post} - \widehat{emp}|_{NJ,Pre}) - (\widehat{emp}|_{PA,Post} - \widehat{emp}|_{PA,Pre})$$
$$= [(a + b_1 + b_2 + b_3) - (a + b_2)] - [(a + b_1) - a]$$
$$= [(b_1 + b_3) - b_1]$$
$$= b_3$$

So the DD estimator is also the coefficient on $Post \cdot Treat$

Card and Krueger collected employment data from a number of fast food restaurants in PA and NJ, before and after NJ's min. wage hike

First, we need to do some annoying data cleanup (to get it into the stadndard panel data format)

This is a little advanced, but I'll leave it here so you can see how things like this work

```
1  library(haven)
2  ff <- read_dta("fastfood.dta")
3  ff <- ff |>
4       select(state, empft, empft2, emppt, emppt2) |>
5       rename(empft_1=empft, emppt_1=emppt, empft_2=empft2, emppt_2 = emppt2
6       pivot_longer(starts_with("emp"), names_to=c(".value", "period"), valu
7                    names_sep="_")
```

Now let's run the DD regression. Note that `state` is a dummy variable for NJ and `period` is a dummy for the post period (here, I am using the `factor` command to tell R that `state` and `period` are dummy variables)

```
1  ck <- lm(empft ~ factor(state) + factor(period) + factor(state)*factor(peri
2  summary(ck)
```

```
Call:
lm(formula = empft ~ factor(state) + factor(period) + factor(state) *
    factor(period), data = ff)

Residuals:
    Min      1Q  Median      3Q     Max
-10.205  -6.445  -2.445   3.555  52.276

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                     10.2051     0.9377  10.883   <2e-16 ***
factor(state)1                  -2.4812     1.0439  -2.377   0.0177 *
factor(period)2                 -2.6402     1.3304  -1.985   0.0475 *
factor(state)1:factor(period)2   3.3617     1.4812   2.270   0.0235 *
---
```

The coefficient on $Post \cdot State$ (i.e., the interaction between `state` and `period`) is 3.36, and is statistically significant

This suggests that the min. wage increase actually *increased* employment in NJ

This contrasts with textbook economic theory, which says that a higher minimum wage should *decrease* employment

Economists found this result surprising, especially since the study is so well done, and continue to work on (and argue about) the effects of the minimum wage

The key assumption behind DD is **parallel trends,** which says that the **treatment group** (NJ in our example) would have experienced the same change in outcomes as the **control group** (PA) if it were not for the policy (the min. wage increase)

This is not directly testable, because we don't have data what *would have* happened in NJ if the min. wage was never increased there

However, we *can* do historical comparisons between the treatment and control groups to see if they tend to experience similar changes in outcomes *before* the **treatment/policy change**

Quadratics

So far, all of our analysis has assumed that the relationship between y and x is linear

However, economics is rife with non-linear relationships

For example, we often assume that a short-run production function exhibits *dimishing marginal product*, meaning that each additional unit of an input produces a little less output

Formally, this means that the production function increases at a decreasing rate (i.e., the second derivative is negative, so the function is *concave*)

Here's an example of diminishing marginal product:

```r
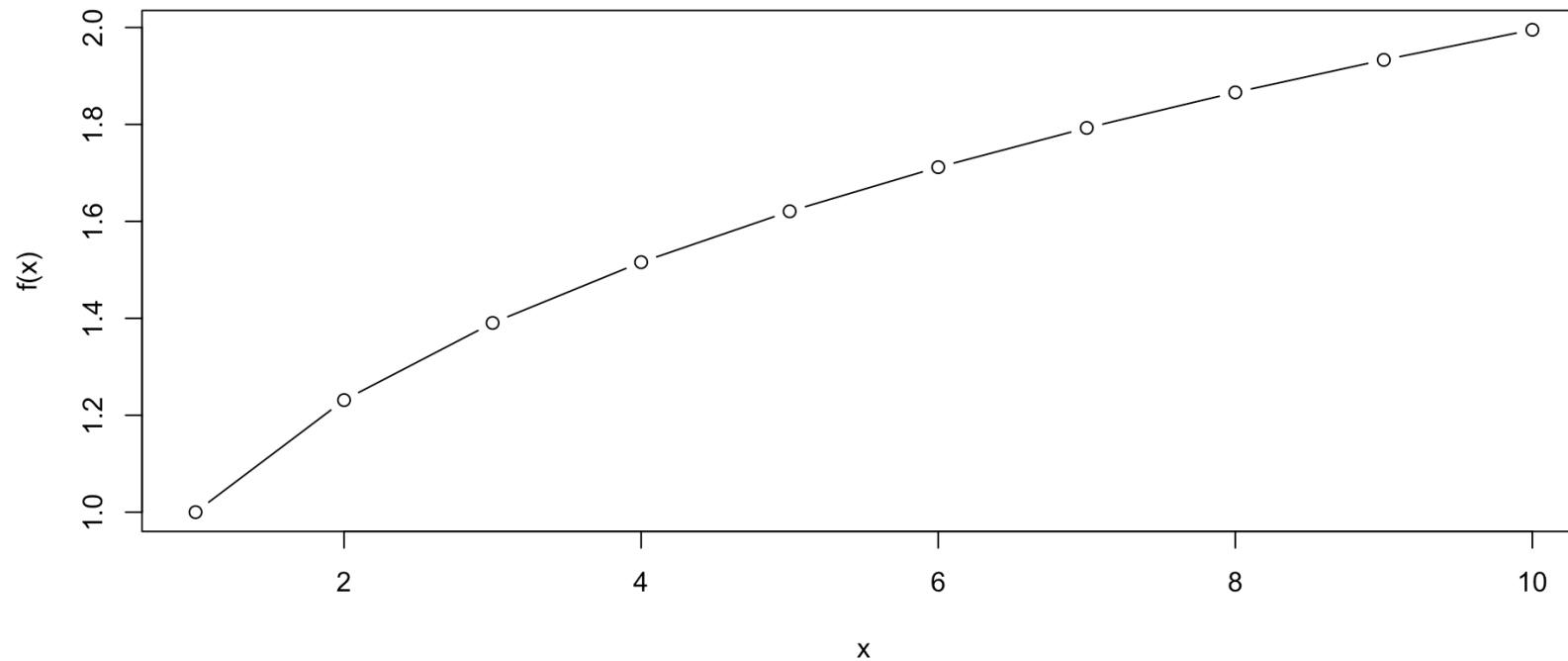1  x <- 1:10
2  y <- x^{.3}
3  plot(x, y, type="b", xlab="x", ylab="f(x)")
```

We can use **quadratics** to capture these kinds of non-linear relationshps using a regression

Suppose we estimate the model

$$\widehat{y} = a + b_1 x + b_2 x^2$$

This is a linear function of x and x^2 (if we think of those as two separate variables), but a non-linear function of x alone

Using calculus,

$$\frac{\Delta \widehat{y}}{\Delta x} = b_1 + 2b_2 x$$

$$\frac{\Delta \widehat{y}}{\Delta x} = b_1 + 2b_2 x$$

Thus, the *slope* changes with the value of x

This allows us to capture a lot of different types of relationships:

- Increasing at an increasng rate: $b_1 > 0, b_2 > 0$ (convex increasing)

- Increasing at a decreasing rate: $b_1 > 0, b_2 < 0$ (concave increasing, like the diminishing marginal product example)

- Decreasing at an increasing rate: $b_1 < 0, b_2 < 0$ (concave decreasing)

- Decreasing at a decreasing rate: $b_1 < 0, b_2 > 0$ (convex decreasing)

Let's see an example of this

```
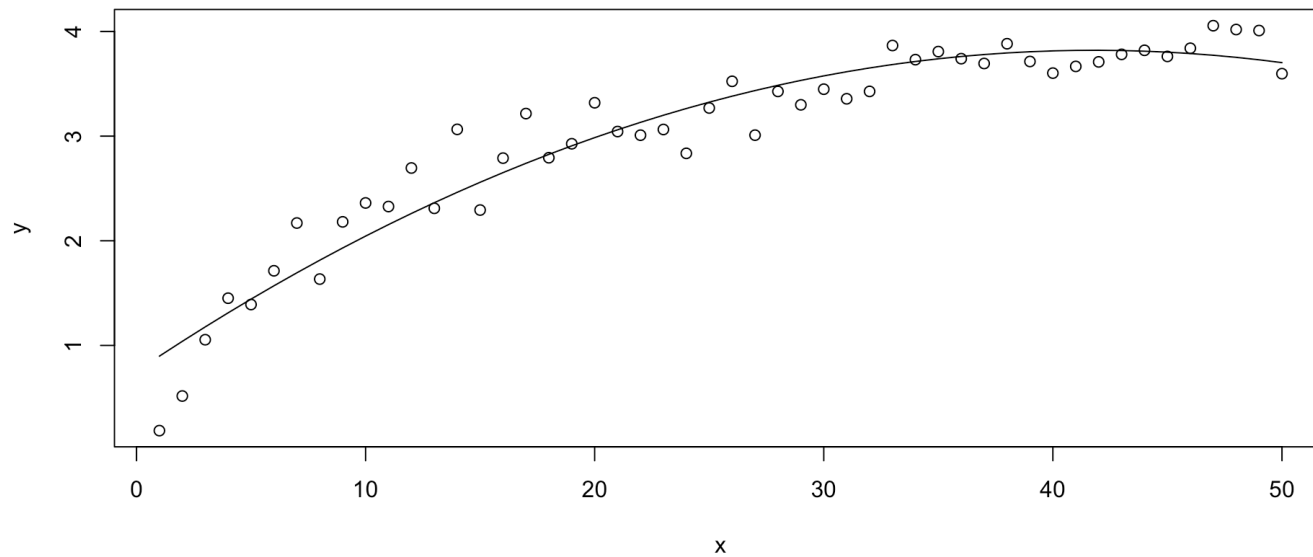1  x <- 1:50
2  y <- log(x) + rnorm(50)/5
3  quad <- lm(y ~ x + I(x^2))
4  y.hat <- quad$fitted.values
5  plot(x, y)
6  lines(x, y.hat, type="l")
```

Even though the *true* relationship between x and y is not quadratic, the quadratic specification provides a pretty good approximation

It's also worth noting that a linear relationship might also provide a good approximation if we're mainly interested in knowing how x affects y "on average"

```r
1  quad <- lm(y ~ x)
2  y.hat.lin <- quad$fitted.values
3  plot(x, y)
4  lines(x, y.hat.lin, type="l")
```
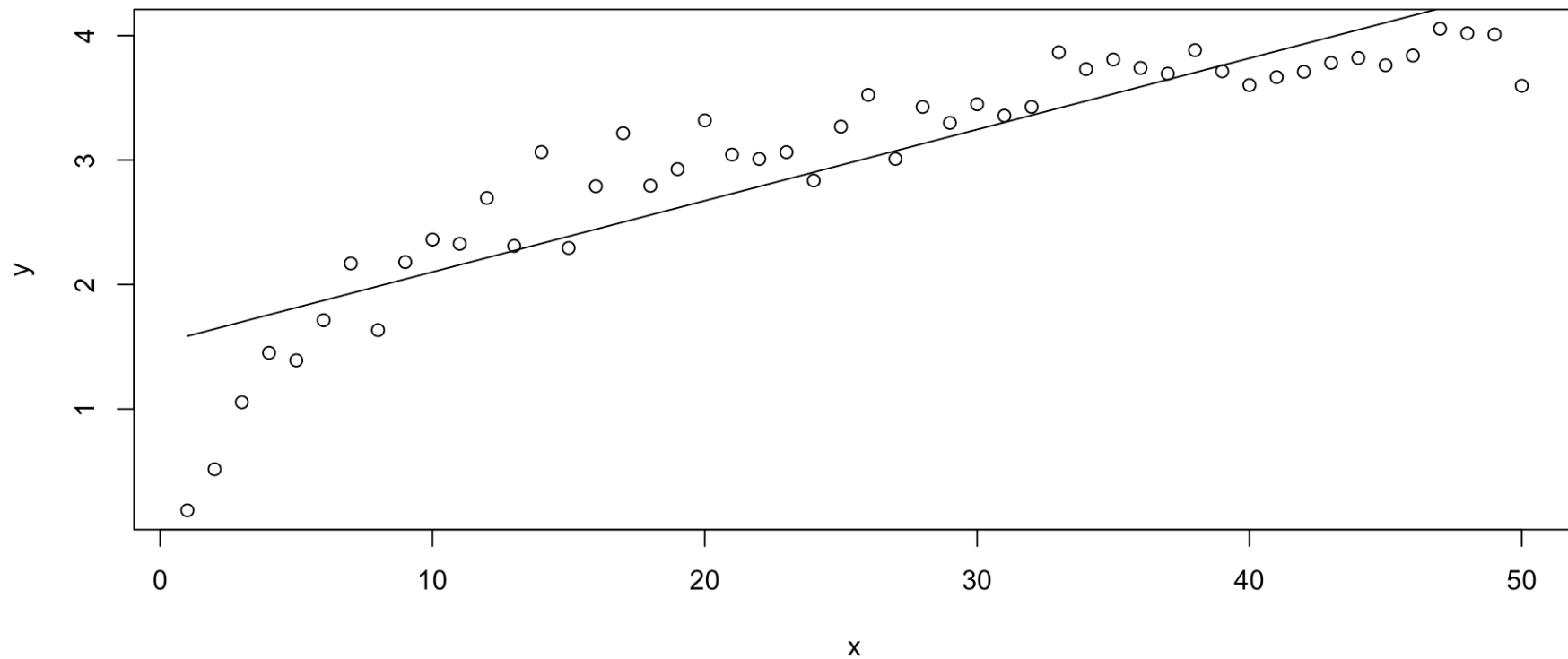
Logs

So far, all of our analysis has looked at the effect of the *level* of x on the *level* of y

I.e., if x increases by 1, $\hat{y}$ increases by b

Sometimes we want to know the proportional effect of variables

For example, we might want to know the percent by which $\hat{y}$ will increase if x increases by 1%

We can do this using logs (logarithms)[1]

There are three important log specifications:

1. Log linear. Suppose that

$$\log \widehat{y} = a + b \log x$$

How do we interpret b?

From our existing interpretation of regression coefficients, we know that

$$b = \frac{\Delta \log \widehat{y}}{\Delta \log x}$$

---

1. Recall that $\log(v)$ is the solution to the equation $e^x = v$

Now, recall from calculus that

$$\frac{\Delta \log y}{\Delta y} = \frac{1}{y}$$

Or, rearranging,

$$\Delta \log y = \frac{\Delta y}{y}$$

Similarly, $\Delta \log x = \Delta x / x$

Plugging this into our expression for b:

$$b = \frac{\Delta \log \hat{y}}{\Delta \log \hat{x}} = \frac{\Delta \hat{y}/\hat{y}}{\Delta x/x} = \frac{100 * (\Delta \hat{y}/\hat{y})}{100 * (\Delta x/x)} = \frac{\%\Delta \hat{y}}{\%\Delta x}$$

Thus, we can interpret b as the **elasticity** of y with respect to x

That is, if x increases by 1%, we expect y to increase by b%

2. Semilog. Now consider the specification

$$\log \hat{y} = a + bx$$

As always, we know that

$$b = \frac{\Delta \log \hat{y}}{\Delta x}$$

Since $\Delta \log \hat{y} = \Delta \hat{y} / \hat{y}$, this becomes

$$b = \frac{\Delta \hat{y} / \hat{y}}{\Delta x}$$

If we multiply both sides by 100, we find that

$$100 \cdot b = \frac{100(\Delta \hat{y}/\hat{y})}{\Delta x} = \frac{\%\Delta \hat{y}}{\Delta x}$$

Therefore, we can interpret b as the % change in y that will occur if x increases by 1 unit

3. "I don't have a name for this one." Now suppose that

$$\hat{y} = a + b \log x$$

Now,

$$b = \frac{\Delta \hat{y}}{\Delta \log x} = \frac{\Delta \hat{y}}{\Delta x / x}$$

Now if we divide by 100,

$$\frac{b}{100} = \frac{\Delta \widehat{y}}{100(\Delta x / x)}$$

so we can interpret $b/100$ as the change in $\widehat{y}$ when x increases by 1%

Let's see examples of these specifications:

```
1  gril$ls <- log(gril$s)
2  b1 <- lm(lw ~ ls, data=gril)$coef[2]
3  b2 <- lm(lw ~ s, data=gril)$coef[2]
4  b3 <- lm(w ~ ls, data=gril)$coef[2]
5  c(b1, b2, b3)
```

```
        ls             s            ls
1.30005619    0.09662449  446.74413585
```

So

- If education increases by 1%, the wage increases by 1.3%

- If education increases by 1 year, the wage increases by 9.6%

- If education increases by 1%, the wage increases by $4.46