Polynomial Regression: Interpretation and Lower Order Terms

Max H. Farrell
BUS 41100
August 28, 2015

In class we talked about polynomial regression and the point was made that we always keep "lower order" terms whenever we put additional polynomials into the model. This handout explains the intuition and interpretation reasons behind this, with examples. The bottom line is that by assuming a certain coefficient is exactly equal to zero, you are making a strong assumption on how Y responds to X, one that you have no business making.

Contents

1 Building Intuition with the Intercept **1**
1.1 Example: House Prices 2
1.2 Example: Wage Data 3

2 Now Adding Squared Terms **4**
2.1 Example: Call Center Data 5

3 Higher Order Polynomials **6**

4 Connection to Multiple Linear Regression **6**

1 Building Intuition with the Intercept

Let's return to simple linear regression and consider leaving out the intercept. This will give us good intuition for what will happen we run polynomial regression but exclude lower order terms.

The general model is: $\mathbb{E}[Y|X] = \beta_0 + \beta_1 X + \varepsilon$. Remember that β_0, β_1, and σ^2 are *unknown*. In particular, we don't know if $\beta_0 = 0$ or not. But suppose we *force* least squares to set $b_0 = 0$. What have we done?

If $b_0 = 0$, the intercept of the line is at zero: that is, when $X = 0$, we predict $\hat{Y} = 0$. So that means that we force our prediction to be exactly zero at the value $X = 0$, no matter what. *This is a geometric assumption: the graph of the line must pass through (0,0).*

Remember how we usually don't put much stock in the intercept? Now we're giving it a strong interpretation, and requiring a lot of knowledge about it! Do you have a very good reason to believe that is true? Usually not. Moreover, in general there is nothing special about the point $Y = 0$ or $X = 0$. These will change if we measure the variables differently.

Remember that our goal is to extract the *general trend* in how Y changes with X. We used to interpret b_1 as the change in Y as X increases. If the intercept is zero, we don't have this anymore! We can only say that b_1 measures the change in Y as X increase *assuming the intercept fixed at zero*. That is because setting $b_0 = 0$ only gives the "right" answer for b_1 if the *true* $\beta_0 = 0$ too.

That is, you must assume that $\mathbb{E}[Y|X] = \beta_1 X + \varepsilon$. This is a very strong assumption! Let's consider some examples.

1.1 Example: House Prices

Return to the house price data from Lecture 2. We have data on house prices (in thousands of dollars) and size (in thousands of square feet). What is our goal for this analysis? We want to find out how price increases with size. So, I want to be able to answer questions like: On average, how much more expensive is a 3,000 sq. ft. house than a 2,000 sq. ft. house? This is *exactly* how we interpret b_1 from the linear regression

$$\texttt{price}_i = b_0 + b_1 \texttt{size}_i + e_i.$$

What if we force $b_0 = 0$? Now we are assuming $\beta_0 = 0$, i.e. that a zero square foot house costs nothing. Is that reasonable? Let's look at the data:

Including intercept:

```
Call:
lm(formula = price ~ size)

Residuals:
    Min      1Q  Median      3Q     Max
-30.425  -8.618   0.575  10.766  18.498

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   38.885      9.094   4.276 0.000903 ***
size          35.386      4.494   7.874 2.66e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 14.14 on 13 degrees of freedom
Multiple R-squared:  0.8267,    Adjusted R-squared:  0.8133
F-statistic:    62 on 1 and 13 DF,  p-value: 2.66e-06
```

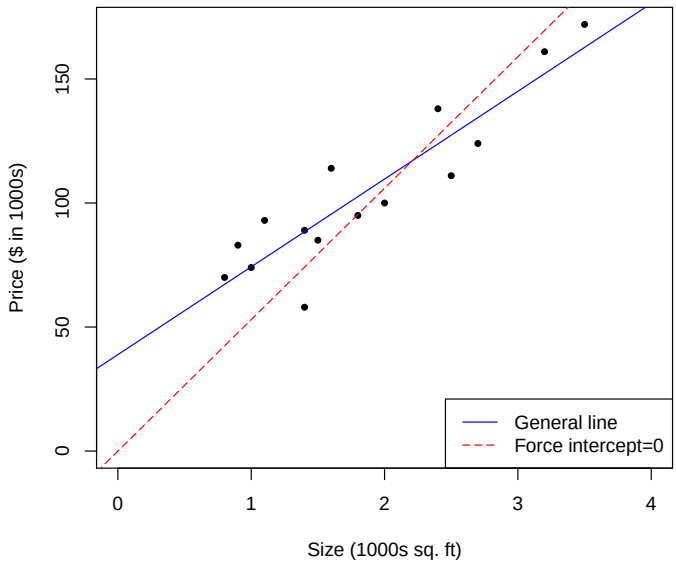
Forcing $b_0 = 0$:

```
Call:
lm(formula = price ~ size - 1)

Residuals:
    Min      1Q  Median      3Q     Max
-21.465 -11.003   5.521  24.313  35.313

Coefficients:
     Estimate Std. Error t value Pr(>|t|)
size   52.986      2.697   19.65 1.37e-11 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 21.13 on 14 degrees of freedom
Multiple R-squared:  0.965,     Adjusted R-squared:  0.9625
F-statistic: 386.1 on 1 and 14 DF,  p-value: 1.368e-11
```



The most important thing is that the slope estimate went up! This makes it look like square footage is much more expensize: the "no intercept" model says that every extra 1,000 square feet costs \$53k, compared to only \$35k from the other regression. Which model is better? (Notice the R^2 is higher on the right, but who cares?!? Remember R^2 doesn't mean anything.)

What happened? By forcing the intercept to be zero, we had to crank the line way up, artificially. Note that I had to manually expand the range of the graph, so we could see both intercepts.

Here's the main question: which one of these would you say better captures the general trend in the response of `price` to `size`?

Now suppose I told you that all house sales are subject to a flat tax of \$5,000. Then, only (`price` - 5000) is under control of the buyer and seller, so we shift all the prices down by 5000. This shouldn't affect the slope of the line at all. But if you are still forcing the intercept to be zero, the slope will have to change![1]

1.2 Example: Wage Data

Let's return to the wage data first used at the end of Lecture 1. Our goal is to find how the number of hours worked (`hours`) responds to hourly pay rate (`pay.rate`). What's different about this example? Here, it makes sense to assume that $\beta_0 = 0$, because that means that if you get paid nothing, you work zero hours. Let's see what the data turns up:

Including intercept:

```
Call:
lm(formula = hours ~ pay.rate)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1913.01      52.89  36.167  < 2e-16 ***
pay.rate       80.94      18.83   4.298  0.00012 ***

Residual standard error: 52.99 on 37 degrees of freedom
Multiple R-squared:  0.333,     Adjusted R-squared:  0.3149
F-statistic: 18.47 on 1 and 37 DF,  p-value: 0.0001203
```

Forcing $b_0 = 0$:

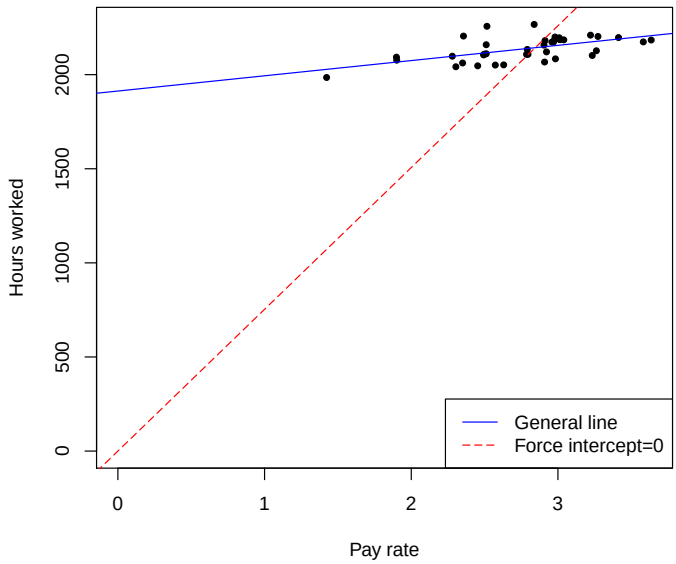
```
Call:
lm(formula = hours ~ pay.rate - 1)

Coefficients:
         Estimate Std. Error t value Pr(>|t|)
pay.rate   753.24      17.98    41.9   <2e-16 ***

Residual standard error: 315.3 on 38 degrees of freedom
Multiple R-squared:  0.9788,    Adjusted R-squared:  0.9783
F-statistic:  1756 on 1 and 38 DF,  p-value: < 2.2e-16
```

The interpretation of the intercept on the left makes no sense: you work almost 2000 hours if you are paid nothing! But the interpretation of the slope is fine: for each extra dollar per hour, you work an extra 80 hours per year. How about on the right? *Assuming you work 0 hours if you aren't paid* you will work an extra 750 hours for each additional dollar per hour. So someone working for \$1/hour (in 1966 remember) will work 750 hours, someone making \$2/hour will work 1500 hours, and so on. Which is more reasonable to you? Look at the picture:

[1] Since $b_0 = \bar{Y} - b_1\bar{X}$, we can immediately solve for $b_1 = \bar{Y}/\bar{X}$. So our forecast for $\hat{Y}$ at the average X, i.e. at $\bar{X}$, is $\hat{Y}(X = \bar{X}) = b_0 + b_1\bar{X} = 0 + \left(\bar{Y}/\bar{X}\right)\bar{X} = \bar{Y}$. So we fit a line that goes through the point $(0, 0)$ and the point $(\bar{X}, \bar{Y})$! If we shift all the prices down by 5,000, the line will rotate toward being flatter, because the new b_1 is the old b_1 minus $5,000/\bar{X}$.



What's going on here? Again, the restriction of $b_0 = 0$ is forcing the line to slope up too fast. Why? Look how far out of the sample you are "predicting" by assuming that $\beta_0 = 0$ (no pay = no work). The data we have are for working people (everyone has positive hours and a positive wage), so this data doesn't tell us anything about people that don't work or aren't paid. And yet we are making a very strong assumption. What about social security or disability (no work, but positive pay)? What about working odd jobs (no formal hours or pay)?

And again, there's nothing special about $Y = 0$ or $X = 0$. It might make sense to measure pay rate as hourly wage above mimimum wage, or measure hours per year relative to a standard work week. Both of these would change the "zero" point.

2 Now Adding Squared Terms

Intuitively, the same problem will crop up for polynomial regression, that is, a *geometric* problem. For now, let's stick to squared terms. We are considering fitting

$$y_i = b_0 + b_1 x_i + b_2 x_i^2 + e_i$$

and setting $b_1 = 0$, that is, leaving out the linear term. *Just like forcing the intercept to be zero was a restriction on the graph of a line, this will also be a geometric/graphical restriction.* Recall the equation of a parabola:

$$y = a(x + v_x)^2 + v_y.$$

The point $(x = -v_x, y = v_y)$ is the vertex (the bottom or the top of the parabola). If $a > 0$ the parabola opens upward (like the letterr "U"); if $a < 0$ the parabola opens downward. Multiplying this out we get

$$y = \underbrace{v_y + v_x^2}_{\beta_0} + \underbrace{2av_x}_{\beta_1} x + \underbrace{a}_{\beta_2} x^2.$$

So if we force least squares to fit $b_1 = 0$ then we are assuming the vertex of the parabola is at $x = 0$ (the bottom if $a > 0$, the top if $a < 0$). Suppose $a > 0$ (that is, $\beta_2 > 0$) so the parabola opens upward. Then the minimum response of Y to X occurs at $X = 0$, by construction. This is again a very strong assumption! This is something you are forcing about the *shape* of how Y responds to X: when $X = 0$, Y responds very little.

Again, there's nothing special about the point $X = 0$. Suppose you measure the variable X, but for some number C, suppose I measure X by $X + C$ instead. For example, if X is wage, I measure it as dollars/hour above minimum wage. This should not affect how Y responds to X at all, that is, the predicted values shouldn't change. Assuming there is no linear term, the response should be

$$Y = \beta_0 + \beta_2(X + C)^2 = \beta_0 + \underbrace{2\beta_2 C}_{\beta_1???} X + \beta_2 X^2,$$

and we have a linear term anyway!

So it boils down to the same intuition: by assuming a certain coefficient is exactly equal to zero, you are making a strong assumption on how Y responds to X, one that you have no business making. The interpretation of b_2 is that as X increases, Y changes by $b_1 + b_2 X$, so that the *rate* of change in Y depends on the particular X value. This makes sense in a lot of applications (like we mentioned in class): diminishing returns to scale, increasing returns to education, etc. When we force $b_1 = 0$, we restrict the response in Y to only have an X-dependent part, and we lose the interpretation of a rate of change plus a factor that changes with X.

2.1 Example: Call Center Data

Let's return the call center data from week 3. The goal is to predict productivity (measured by `calls` per day) using work experience (`months` of employment). In class we had a quadratic fit:

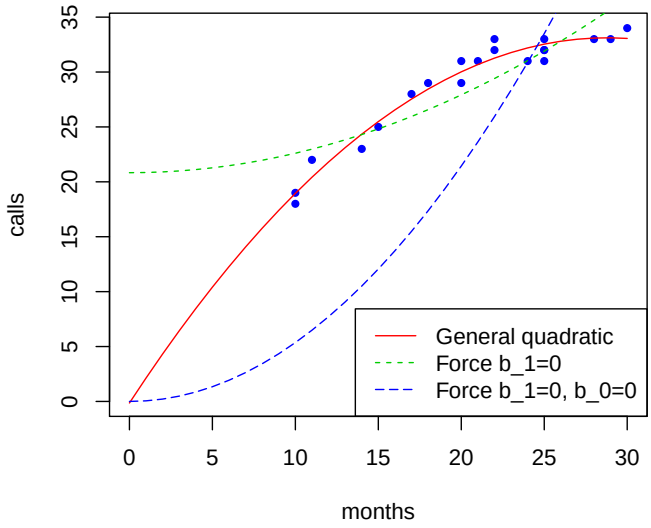
$$y_i = b_0 + b_1\texttt{months} + b_2\texttt{months}^2 + e_i.$$

There's really no great reason to leave out the linear term here. Conceptually, what would that mean in this example? It means that people's productivity gain is the *slowest* when they are brand new. Is that reasonable from an intuitive level? Probably not. And we have no data at `months`=0, so it does not make sense to impose that the minimum productivity is there.

We can also leave out the intercept and the linear term, setting $b_0 = b_1 = 0$. This implies that the parabola goes through (0,0), so that employees with zero experience make zero calls. Is that reasonable? Probably not, even on your first day you could presumably make one call.

I'm omitting the `R` output here, but here is the graph.[2] As you can see, it has all the same issues as before. The blue and green curves do not fit the data at all: they are doing a terrible job of extracting the general trend of how Y changes with X.

[2] It looks like the general curve goes through (0,0) too, but it does not: the intercept is $b_0 = -0.1404712$.



3 Higher Order Polynomials

The story is the same for higher order polynomials, but more intricate. The graphical/geometric interpretations of the above two cases are pretty clear. But what does it mean to leave the linear term out of a cubic fit? To really understand it, you have to go back to the equation for a cubic curve and figure out exactly what restriction you are imposing. I will not delve into the details. The message should be clear: you are making a strong restriction on how Y responds to X.

4 Connection to Multiple Linear Regression

In multiple linear regression we interpret each coefficient *conditional* on what else is in the model. In the first section, when we interpret b_1 from a linear model, the interpretation depends on what we assume about the intercept. If we force b_0, then the slope is interpreted *conditional* on this choice. In the quadratic model, forcing $b_1 = 0$ implies a very specific mechanism for changes in Y.

In multiple linear regression, say with two variables X_1 and X_2, we estimate $y_i = b_0 + b_1 x_{1,i} + b_2 x_{2,i} + e_i$. The interpretation of b_2 is *conditional* on X_1 being in the model. So b_2 measures the change in Y as X_2 increase controlling for X_1, holding it fixed at any given value (this is where the term "controlling" for comes from in the popular press).

For example, suppose $X_1 =$ `education` and $X_2 =$ `experience` and our goal is predict $Y =$ `wages`. If we run the full model, then b_1 measures the return to `education` holding `experience` constant. That is, it gives the wage difference between two people who have exactly the same number of years on the job, but one graduated college and the other only finished high school. If we set $b_2 = 0$ (that is, regress `wages` on only `education`), then b_1 now measures just the returns to `education`. So now it gives the wage difference between two people where one graduated college and the other only finished high school, no matter how many years on the job they have.