Advanced Regression Analysis

5. Inference and Interpretation of Linear Regression

GOVT 6029 - Spring 2020

Cornell University

Model specification refers to the choice of X_1, X_2, etc.

Model specification refers to the choice of X_1, X_2, etc.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \varepsilon$$

Which X's we include in our model, which we exclude, & how we transform them

Model specification refers to the choice of X_1, X_2, etc.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \varepsilon$$

Which X's we include in our model, which we exclude, & how we transform them

We need to get this right for substantive & statistical reasons

A large % of seminar & referee comments are about specification

Model specification refers to the choice of X_1, X_2, etc.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \varepsilon$$

Which X's we include in our model, which we exclude, & how we transform them

We need to get this right for substantive & statistical reasons

A large % of seminar & referee comments are about specification

First, we'll talk about what different specifications mean

Later, we'll talk about how to choose a specification (fitting the model)

Let's imagine we know that the *true* model for some data Y is

$$Y_i = \beta_0^{\text{true}} + \beta_1^{\text{true}} X_i + \beta_2^{\text{true}} Z_i + \varepsilon_i^{\text{true}}$$

Let's imagine we know that the *true* model for some data Y is

$$Y_i = \beta_0^{\text{true}} + \beta_1^{\text{true}} X_i + \beta_2^{\text{true}} Z_i + \varepsilon_i^{\text{true}}$$

We can think of this as the model which generated the population we sample

Let's imagine we know that the *true* model for some data Y is

$$Y_i = \beta_0^{\text{true}} + \beta_1^{\text{true}} X_i + \beta_2^{\text{true}} Z_i + \varepsilon_i^{\text{true}}$$

We can think of this as the model which generated the population we sample

Which quantitative methods will recover the truth about the model's parameters?

Let's unpack that question:

Let's imagine we know that the *true* model for some data Y is

$$Y_i = \beta_0^{\text{true}} + \beta_1^{\text{true}} X_i + \beta_2^{\text{true}} Z_i + \varepsilon_i^{\text{true}}$$

We can think of this as the model which generated the population we sample

Which quantitative methods will recover the truth about the model's parameters?

Let's unpack that question:

Which quantitative methods will recover the truth about the model's parameters?

Let's unpack that question:

Which quantitative methods will recover the truth about the model's parameters?

Let's unpack that question:

- "recover the truth": yield an unbiased estimate;

Which quantitative methods will recover the truth about the model's parameters?

Let's unpack that question:

- "recover the truth": yield an unbiased estimate;
  one that is right on average over many samples of X, Y, Z.

Which quantitative methods will recover the truth about the model's parameters?

Let's unpack that question:

- "recover the truth": yield an unbiased estimate;
  one that is right on average over many samples of X, Y, Z.
- "methods": includes the method of estimation;

Which quantitative methods will recover the truth about the model's parameters?

Let's unpack that question:

- "recover the truth": yield an unbiased estimate;
  one that is right on average over many samples of X, Y, Z.
- "methods": includes the method of estimation; e.g. least squares

Which quantitative methods will recover the truth about the model's parameters?

Let's unpack that question:

- "recover the truth": yield an unbiased estimate;
  one that is right on average over many samples of X, Y, Z.
- "methods": includes the method of estimation; e.g. least squares
- "model": includes the choice of specification;

Which quantitative methods will recover the truth about the model's parameters?

Let's unpack that question:

- "recover the truth": yield an unbiased estimate;
  one that is right on average over many samples of X, Y, Z.
- "methods": includes the method of estimation; e.g. least squares
- "model": includes the choice of specification;
  e.g., which controls, transformations, and interactions

Because Y_i was constructed by adding together

$\beta_0^{\text{true}} + \beta_1^{\text{true}} X_i + \beta_2^{\text{true}} Z_i \quad + \quad$ a Normal disturbance,

estimation by least squares will not only be unbiased,
but also the best unbiased method.

Because Y_i was constructed by adding together

$\beta_0^{\text{true}} + \beta_1^{\text{true}} X_i + \beta_2^{\text{true}} Z_i \quad + \quad$ a Normal disturbance,

estimation by least squares will not only be unbiased,
but also the best unbiased method.

But it will only be unbiased if we choose the right specification

If we estimate

$$Y_i = \beta_0^* + \beta_1^* X_i + \varepsilon_i^*$$

Because Y_i was constructed by adding together

$\beta_0^{\text{true}} + \beta_1^{\text{true}} X_i + \beta_2^{\text{true}} Z_i \quad + \quad$ a Normal disturbance,

estimation by least squares will not only be unbiased,
but also the best unbiased method.

But it will only be unbiased if we choose the right specification

If we estimate

$$Y_i = \beta_0^* + \beta_1^* X_i + \varepsilon_i^*$$

We won't get any estimate for β_2 (because we assumed it was zero by omitting it)

Moreover, it will usually be the case that $\hat{\beta}_1^*$ is a *biased* estimate of β_1

The source of this bias can be shown formally.

We estimated this

$$Y_i = \beta_0^* + \beta_1^* X_i \varepsilon_i^*$$

leaving out Z_i. Suppose we ran an auxiliary regression of Z_i on X_i:

$$Z_i = \gamma_0 + \gamma_1 X_i + \nu_i$$

The source of this bias can be shown formally.

We estimated this

$$Y_i = \beta_0^* + \beta_1^* X_i \varepsilon_i^*$$

leaving out Z_i. Suppose we ran an auxiliary regression of Z_i on X_i:

$$Z_i = \gamma_0 + \gamma_1 X_i + \nu_i$$

We could substitute this back into the true model,

$$Y_i \;\; = \;\; \beta_0^{\text{true}} + \beta_1^{\text{true}} X_i + \beta_2^{\text{true}} Z_i + \varepsilon_i^{\text{true}}$$

The source of this bias can be shown formally.

We estimated this

$$Y_i = \beta_0^* + \beta_1^* X_i \varepsilon_i^*$$

leaving out Z_i. Suppose we ran an auxiliary regression of Z_i on X_i:

$$Z_i = \gamma_0 + \gamma_1 X_i + \nu_i$$

We could substitute this back into the true model,

$$Y_i = \beta_0^{\text{true}} + \beta_1^{\text{true}} X_i + \beta_2^{\text{true}} Z_i + \varepsilon_i^{\text{true}}$$

$$Y_i = \beta_0^{\text{true}} + \beta_1^{\text{true}} X_i + \beta_2^{\text{true}} (\gamma_0 + \gamma_1 X_i + \nu_i) + \varepsilon_i^{\text{true}}$$

The source of this bias can be shown formally.

We estimated this

$$Y_i = \beta_0^* + \beta_1^* X_i \varepsilon_i^*$$

leaving out Z_i. Suppose we ran an auxiliary regression of Z_i on X_i:

$$Z_i = \gamma_0 + \gamma_1 X_i + \nu_i$$

We could substitute this back into the true model,

$$\begin{aligned} Y_i &= \beta_0^{\text{true}} + \beta_1^{\text{true}} X_i + \beta_2^{\text{true}} Z_i + \varepsilon_i^{\text{true}} \\ Y_i &= \beta_0^{\text{true}} + \beta_1^{\text{true}} X_i + \beta_2^{\text{true}} (\gamma_0 + \gamma_1 X_i + \nu_i) + \varepsilon_i^{\text{true}} \\ Y_i &= \left(\beta_0^{\text{true}} + \beta_2^{\text{true}} \gamma_0\right) + \left(\beta_1^{\text{true}} + \beta_2^{\text{true}} \gamma_1\right) X_i + \left(\varepsilon_i^{\text{true}} + \beta_2^{\text{true}} \nu_i\right) \end{aligned}$$

The parenthetical terms below are what we recover when we run LS omitting Z_i

$$Y_i = \left(\beta_0^{\text{true}} + \beta_2^{\text{true}}\gamma_0\right) + \left(\beta_1^{\text{true}} + \beta_2^{\text{true}}\gamma_1\right) X_i + \left(\varepsilon_i^{\text{true}} + \beta_2^{\text{true}}\nu_i\right)$$

Omitted variable bias

The parenthetical terms below are what we recover when we run LS omitting Z_i

$$
\begin{aligned}
Y_i &= \left(\beta_0^{\text{true}} + \beta_2^{\text{true}}\gamma_0\right) + \left(\beta_1^{\text{true}} + \beta_2^{\text{true}}\gamma_1\right) X_i + \left(\varepsilon_i^{\text{true}} + \beta_2^{\text{true}}\nu_i\right) \\
Y_i &= \beta_0^* + \beta_1^* X_i + \varepsilon_i^*
\end{aligned}
$$

The parenthetical terms below are what we recover when we run LS omitting Z_i

$$
\begin{aligned}
Y_i &= \left(\beta_0^{\text{true}} + \beta_2^{\text{true}}\gamma_0\right) + \left(\beta_1^{\text{true}} + \beta_2^{\text{true}}\gamma_1\right) X_i + \left(\varepsilon_i^{\text{true}} + \beta_2^{\text{true}}\nu_i\right) \\
Y_i &= \beta_0^* + \beta_1^* X_i + \varepsilon_i^*
\end{aligned}
$$

The estimate we get of β_1 is:

$$
\mathrm{E}(\hat{\beta}_1) = \beta_1^{\text{true}} + \beta_2^{\text{true}}\gamma_1
$$

The parenthetical terms below are what we recover when we run LS omitting Z_i

$$
\begin{aligned}
Y_i &= \left(\beta_0^{\text{true}} + \beta_2^{\text{true}}\gamma_0\right) + \left(\beta_1^{\text{true}} + \beta_2^{\text{true}}\gamma_1\right) X_i + \left(\varepsilon_i^{\text{true}} + \beta_2^{\text{true}}\nu_i\right) \\
Y_i &= \beta_0^* + \beta_1^* X_i + \varepsilon_i^*
\end{aligned}
$$

The estimate we get of β_1 is:

$$
\begin{aligned}
\mathrm{E}(\hat{\beta}_1) &= \beta_1^{\text{true}} + \beta_2^{\text{true}}\gamma_1 \\
&= \beta_1^{\text{true}} + \beta_2^{\text{true}} \left(\frac{\sum_i (X_i - \bar{X})(Z_i - \bar{Z})}{\sum_i (X_i - \bar{X})^2} \right)
\end{aligned}
$$

Which is unbiased only if $\beta_2^{\text{true}} = 0$ or $\mathrm{corr}(X_i, Z_i) = 0$.

Thus there are only two conditions under which we can safely omit variable Z from our model:

Thus there are only two conditions under which we can safely omit variable Z from our model:

β is really zero; Z has no effect on Y in the "true" model

or

The correlation of Z with the included X's is zero

$\Rightarrow$ we should include any Z that is correlated with both Y and some included X

Thus there are only two conditions under which we can safely omit variable Z from our model:

β is really zero; Z has no effect on Y in the "true" model

or

The correlation of Z with the included X's is zero

$\Rightarrow$ we should include any Z that is correlated with both Y and some included X

(Some practical caveats to this below)

This is why the complaint "You should have controlled for ..." carries so much weight in criticizing empirical research.

Thus there are only two conditions under which we can safely omit variable Z from our model:

β is really zero; Z has no effect on Y in the "true" model

or

The correlation of Z with the included X's is zero

$\Rightarrow$ we should include any Z that is correlated with both Y and some included X

(Some practical caveats to this below)

This is why the complaint "You should have controlled for …" carries so much weight in criticizing empirical research.

Specification is arguably *the* major concern in most observational research (along with selection & endogeneity)

Does this mean we should include the kitchen sink in the regression?

Does this mean we should include the kitchen sink in the regression?

Is there a penalty to including irrelevant variables?

Does this mean we should include the kitchen sink in the regression?

Is there a penalty to including irrelevant variables?

Yes, but it is smaller. We lose efficiency in two ways:

- Lost degrees of freedom
- Lost variance in relevant covariates after conditioning on irrelevant ones

So is the kitchen sink safer?

So is the kitchen sink safer?

Kevin Clarke emphasizes that you only *solve* OVB when you have the “right” specification”. It is not usually clear whether getting “closer” to the right specification reduces bias.

So is the kitchen sink safer?

Kevin Clarke emphasizes that you only *solve* OVB when you have the “right” specification”. It is not usually clear whether getting “closer” to the right specification reduces bias.

My view: robustness to specification choice is a nice standard.

So is the kitchen sink safer?

Kevin Clarke emphasizes that you only *solve* OVB when you have the “right” specification”. It is not usually clear whether getting “closer” to the right specification reduces bias.

My view: robustness to specification choice is a nice standard. Try to break your findings, and report how easily they are broken.

Is there any other reason why we should omit variables?

Is there any other reason why we should omit variables?

Often, we intentionally omit intermediate steps in a (presumed) causal process

Is there any other reason why we should omit variables?

Often, we intentionally omit intermediate steps in a (presumed) causal process

The usual example is voting. Consider this hypothesis:

Post-college education makes people more likely to vote Democratic

Is there any other reason why we should omit variables?

Often, we intentionally omit intermediate steps in a (presumed) causal process

The usual example is voting. Consider this hypothesis:

Post-college education makes people more likely to vote Democratic

Suppose we test this by regressing self-reported vote choice in the 2004 election on years of education.

Is there any other reason why we should omit variables?

Often, we intentionally omit intermediate steps in a (presumed) causal process

The usual example is voting. Consider this hypothesis:

Post-college education makes people more likely to vote Democratic

Suppose we test this by regressing self-reported vote choice in the 2004 election on years of education.

A critic objects that we should have controlled for party ID, which is correlated with both our response and covariate. Is he right?

Post-college education makes people more likely to vote Democratic

Suppose we test this by regressing self-reported vote choice in the 2004 election on years of education.

A critic objects that we should have controlled for party ID, which is correlated with both our response and covariate. Is he right?

Post-college education makes people more likely to vote Democratic

Suppose we test this by regressing self-reported vote choice in the 2004 election on years of education.

A critic objects that we should have controlled for party ID, which is correlated with both our response and covariate. Is he right?

Probably not.

Same logic says to include vote intention 10 minutes prior to visiting the polls

Post-college education makes people more likely to vote Democratic

Suppose we test this by regressing self-reported vote choice in the 2004 election on years of education.

A critic objects that we should have controlled for party ID, which is correlated with both our response and covariate. Is he right?

Probably not.

Same logic says to include vote intention 10 minutes prior to visiting the polls

We can validly omit variables W which only affect Y through included X's if we want β to absorb the impact of W through X.

As long as this remains a valid statement:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \varepsilon$$

the regression is linear

What makes linear regression "linear"?

As long as this remains a valid statement:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \varepsilon$$

the regression is linear

We can make any algebraic subtitutions we want.

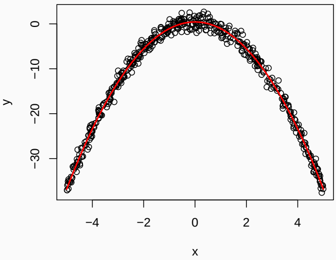
To see this, replace the big X's of the above equation with functions of small x's.

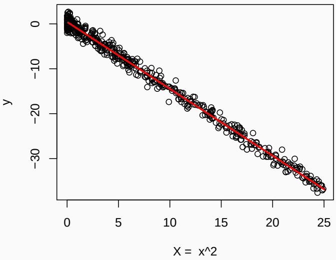
Suppose $X_1 = x_1^2$ and $X_2 = x_1$. Then, by algebraic substitution:

$$Y = \beta_0 + \beta_1 x_1^2 + \beta_2 x_1 + \ldots + \varepsilon$$

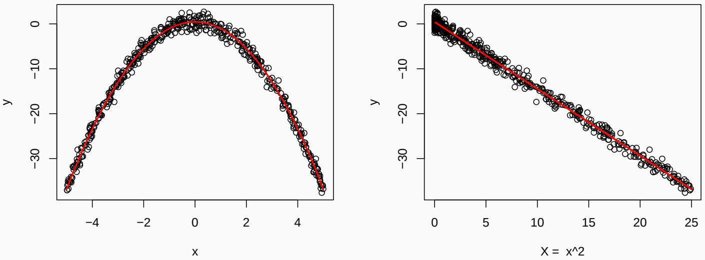
Sample output (note we're leaving out x_1 to make a point; normally we need it, too):

```
Call:
lm(formula = y ~ I(x^2))

Residuals:
    Min      1Q  Median      3Q     Max
-4.3876 -0.1713  0.2031  0.3971  0.7495

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.9008332  0.0310244  29.036   <2e-16 ***
I(x^2)      -0.0006920  0.0007175  -0.965    0.335
```





The same regression; two views



The same regression; two views

Left is the regression in the original, untransformed scale of x

Right is the regression as R sees it, in the transformed scale $x^2 = X$

Linear regression is always linear in the transformed covariate

It may be very curvilinear in the scale we care about

Suppose $X_1 = x_1^3$, $X_2 = x_1^2$, $X_3 = x_1$.

This is a cubic polynomial specification:

$$Y = \beta_0 + \beta_1 x_1^3 + \beta_2 x_1^2 + \beta_3 x_1 + \ldots + \varepsilon$$

Or even add a quartic term, $X_4 = x_1^4$. Then,

$$Y = \beta_0 + \beta_1 x_1^3 + \beta_2 x_1^2 + \beta_3 x_1 + \beta_4 x_1^4 + \ldots + \varepsilon$$

Sample output for a cubic (3rd order) polynomial:

```
Call:
lm(formula = y ~ I(x^3) + I(x^2) + I(x))

Residuals:
    Min      1Q  Median      3Q     Max
-4.3835 -0.1706  0.2006  0.3983  0.7536

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.8840553  0.0830528  10.644   <2e-16 ***
I(x^3)       0.0002337  0.0011134   0.210    0.834
I(x^2)      -0.0044455  0.0168552  -0.264    0.792
I(x)         0.0166821  0.0721171   0.231    0.817
```

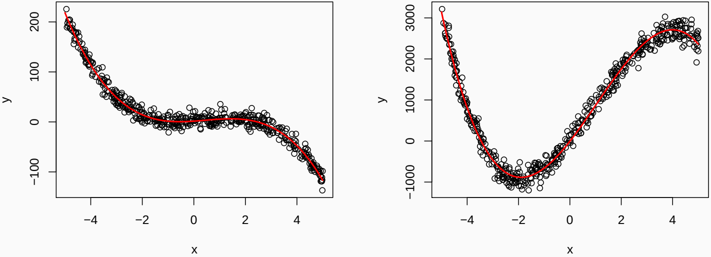
None of the coefficients are significant, but they collectively explain a lot of variance. (How is this possible?)

Sample output for a cubic (3rd order) polynomial:

```
Call:
lm(formula = y ~ I(x^3) + I(x^2) + I(x))

Residuals:
    Min      1Q  Median      3Q     Max 
-4.3835 -0.1706  0.2006  0.3983  0.7536 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.8840553  0.0830528  10.644   <2e-16 ***
I(x^3)       0.0002337  0.0011134   0.210    0.834    
I(x^2)      -0.0044455  0.0168552  -0.264    0.792    
I(x)         0.0166821  0.0721171   0.231    0.817    
```

None of the coefficients are significant, but they collectively explain a lot of variance. (How is this possible?)

With polynomials (and with other interactions)
t-tests of individual parameters are less interesting than CIs around $\hat{Y}$

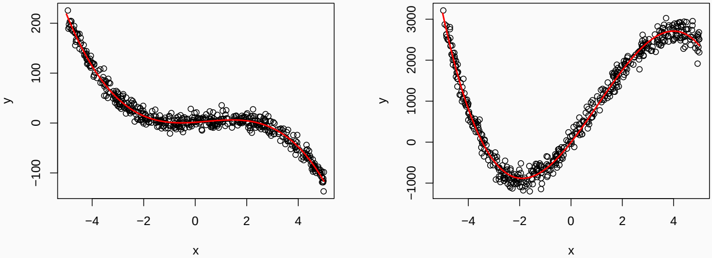


These are two different regressions.

Left is a cubic (3rd order) polynomial specificiation. It has 2 bends

Right is a quartic (4th order) polynomial. It has 3 bends

Each polynomial order we add puts another bend in the line



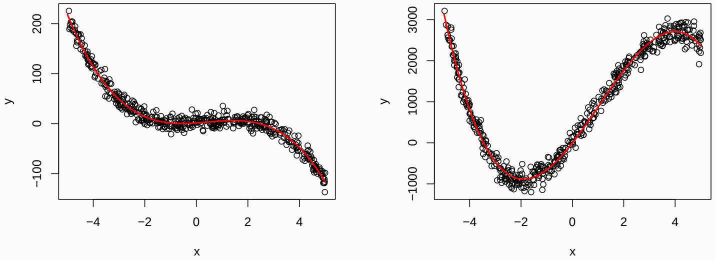
If we include n terms, there will be a bend for every observation

Called “curve-fitting”: a perfect (& perfectly useless) model

Always have a theoretical reason to include polynomial terms

Seldom is more than a quadratic justified by theory

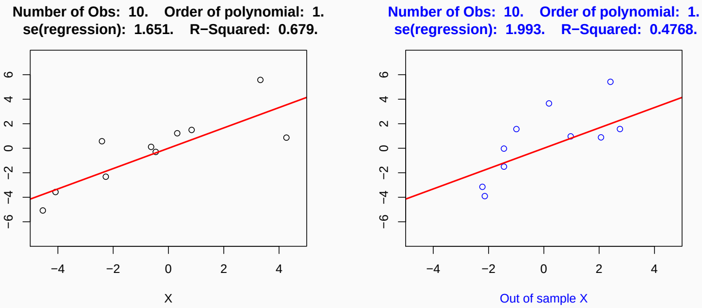
If you include polynomial terms, you need to interpret the result graphically



Warnings about polynomial fits:

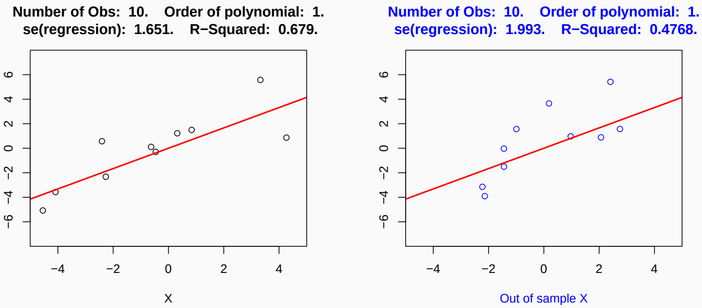
- High order polynomials will *always* fit the sample well, but seldom fit the population well (curve-fitting)
- Extrapolation from polynomial or interactive specifications is dangerous

  These functional forms behave wildly outside the known data



Polynomial overfitting experiment: generate 10 obs from the “true” model:

$$Y = x + \varepsilon, \qquad \varepsilon \sim \mathrm{N}(0, 3)$$

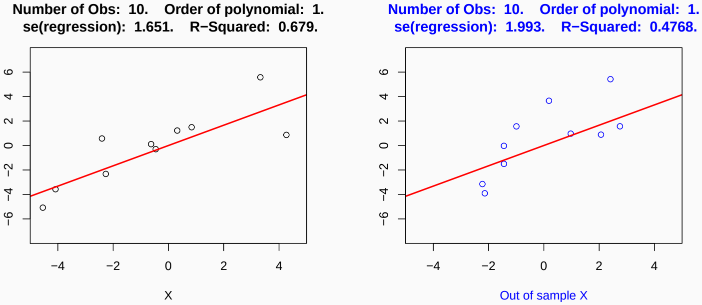


Polynomial overfitting experiment: generate 10 obs from the "true" model:

$$Y = x + \varepsilon, \qquad \varepsilon \sim \mathrm{N}(0, 3)$$

and fit these data using different polynomials of x.

We will show the fit of the model to the original data on the left

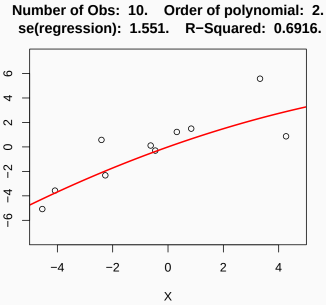


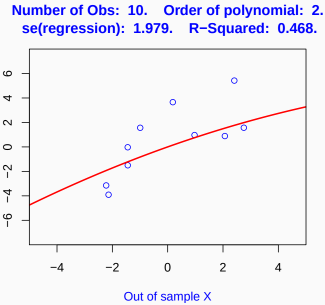
We'll draw new "out of sample data" from the same true model:

$$Y_{\mathrm{OOS}} = x_{\mathrm{OOS}} + \varepsilon_{\mathrm{OOS}}, \qquad \varepsilon \sim \mathrm{N}(0, 3)$$

use the model as fitted on the orig data to predict out-of-sample cases

Then show the fit of the old model to the out-of-sample data on the right

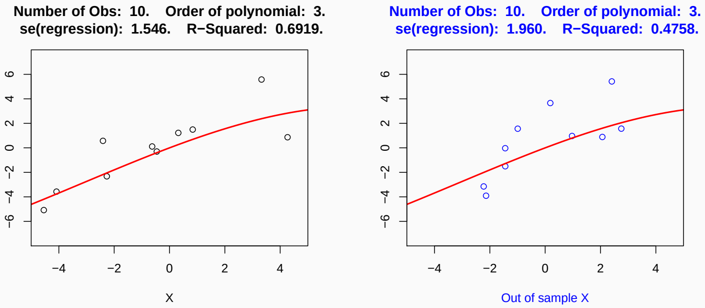




Above is the fit from a quadratic specification of x, ie, we estimated:

$$Y = \hat{\beta}_1 x + \hat{\beta}_2 x^2 + \hat{\varepsilon}$$

Note that we omitted the constant for didactic reasons

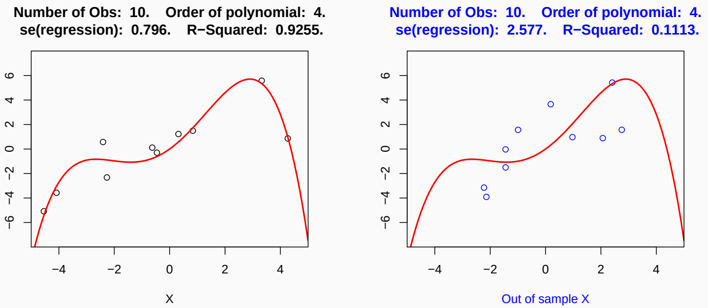


Above is the fit from a cubic specification of x, ie, we estimated:

$$Y = \hat{\beta}_1 x + \hat{\beta}_2 x^2 + \hat{\beta}_3 x^3 + \hat{\varepsilon}$$

How many polynomials can we add and still find $\hat{\beta}$?

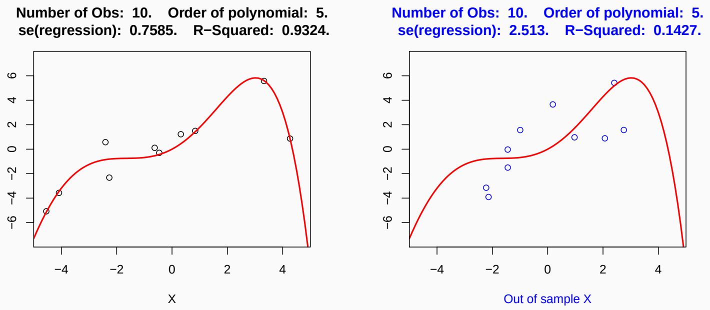
What will happen to the fit in and out of sample as we add polynomials?



Above is the fit from a quartic specification of x, ie, we estimated:

$$Y = \hat{\beta}_1 x + \hat{\beta}_2 x^2 + \hat{\beta}_3 x^3 + \hat{\beta}_4 x^4 + \hat{\varepsilon}$$

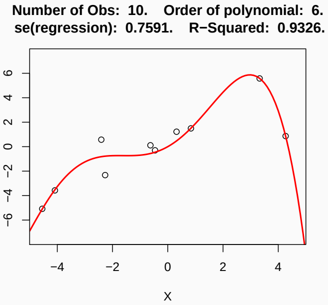
On the left, we see the model finds a curious non-linearity, by which low and high x suppress y, but middle values of x increase y. Do you trust this finding?

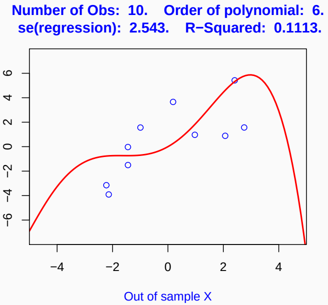


In small samples. outliers can easily create the illusion of complex curves relating x and y

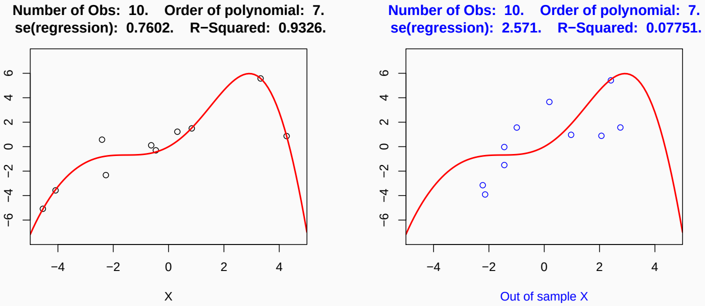
We need *a lot* of data to discern if such curves are more than spurious

(And so we probably need a strong theory, too, to justify the data collection)

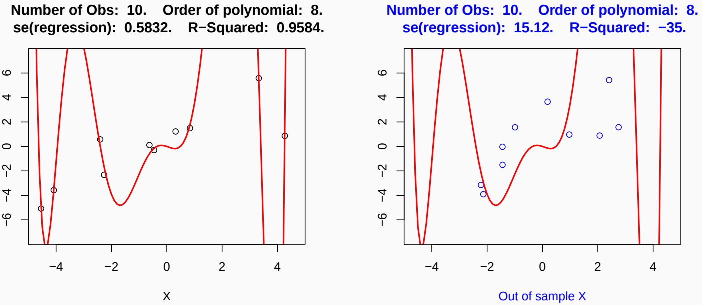




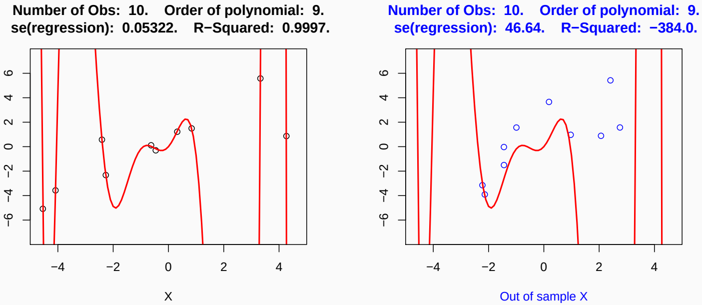
What happens as we approach a tenth-order polynomial?



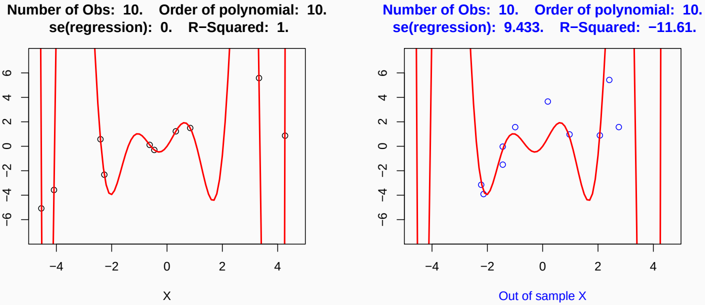
What happens as we approach a tenth-order polynomial?



What happens as we approach a tenth-order polynomial?



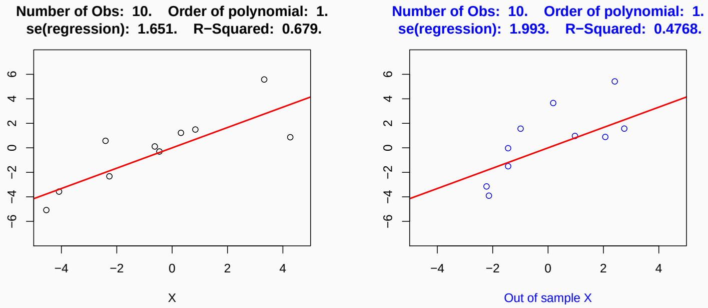
What happens as we approach a tenth-order polynomial?



When the number of parameters in the model equals the number of observations, least squares is able to fit a line through every datapoint

This means the fit will be “perfect”: no error.

And out of sample, it will be completely useless, and worse than guessing that y simply equals its sample mean in every case.



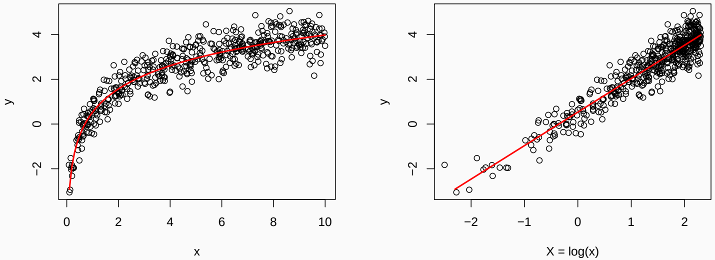
Two lessons:

Beware curve-fitting

Beware good fits in-sample unless they fit well out of sample, too

Suppose $X_1 = \log(x_1)$. Then,

$$Y = \beta_0 + \beta_1 \log(x_1) + \beta_2 X_2 + \ldots + \varepsilon$$



These are the same regression

Left is on the original, untransformed scale

Right is on the transformed, log scale

Log transformations for effects that diminish in per unit potency as x increases

We could keep going, combining transformations and interactions to get very nonlinear models

Suppose $X_1 = x_1 \times x_2$ and $X_2 = x_2$. Then,

$$Y = \beta_0 + \beta_1 x_1 \times x_2 + \beta_2 x_2 + \ldots + \varepsilon$$

Transformations of covariates

We could keep going, combining transformations and interactions to get very nonlinear models

Suppose $X_1 = x_1 \times x_2$ and $X_2 = x_2$. Then,

$$Y = \beta_0 + \beta_1 x_1 \times x_2 + \beta_2 x_2 + \ldots + \varepsilon$$

Suppose $X_1 = x_1 \times \log(x_2) \times \sqrt{x_3}$ and $X_2 = x_2/(x_1 + x_2^2)$. Then,

$$Y = \beta_0 + \beta_1 (x_1 \times \log(x_2) \times \sqrt{x_3}) + \beta_2 x_2/(x_1 + x_2^2) + \ldots + \varepsilon$$

Without strong theoretical support, these models will be silly and produce unreliable results with little predictive power

We could even replace Y.

This is a good idea if you think Y is not a linear function of the regressors, but $g(y)$ *is* a linear function of them

Usually, this is the case for counts, e.g., or money.

We could even replace Y.

This is a good idea if you think Y is not a linear function of the regressors, but $g(y)$ *is* a linear function of them

Usually, this is the case for counts, e.g., or money.

Raising a government budget from \$1 million to \$10 million may be "just as hard" as raising it from \$10 to \$100 million

If X's affect the order of magnitude of Y, you should log Y.

Transformations of the response variable

If $Y = \log(y)$, then,

$$\log(y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \varepsilon$$

Now all X's have diminishing effect

If $Y = \log(y)$, then,

$$\log(y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \varepsilon$$

Now all X's have diminishing effect

In fact, level changes in X yield *percentage* changes in Y

If you log both X and Y, then % changes in X cause % changes in Y

What if Y is bounded but continuous?

Suppose it ranges between 0 and 1 (but doesn't include these values)?

Then we need to "stretch it out" to range from $-\infty$ to ∞

What if Y is bounded but continuous?

Suppose it ranges between 0 and 1 (but doesn't include these values)?

Then we need to "stretch it out" to range from $-\infty$ to ∞

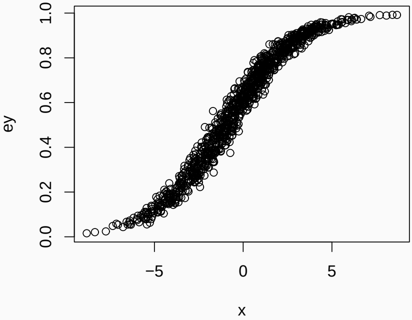
The logit transformation does this: $Y = \log(y/(1-y))$.

$$\log\left(\frac{y}{1-y}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \varepsilon$$

This model doesn't work if the original data include 0s or 1s

It is not "the logit model", which is a non-linear model of 0s and 1s only

It is a linear regression with a logit transformed response variable



Note the curve is S-shaped.

We could rescale it to work for any bounds, not just $(0, 1)$

That is, just transform y to $y^* = \frac{y-a}{b-a}$, then run the regression of $\text{logit}(y^*)$ on your covariates

This works generally, with the caveat that for any bounds (a, b), none of the data can be exactly a or b

All of the above models are examples of linear regression.

They are all linear in the parameters.

They can all be estimated by least squares.

All of the above models are examples of linear regression.

They are all linear in the parameters.

They can all be estimated by least squares.

What specifications *can't* we estimate?

All of the above models are examples of linear regression.

They are all linear in the parameters.

They can all be estimated by least squares.

What specifications *can't* we estimate?

We can't use LS to estimate models that are *non-linear* in the parameters

Example of a model that is non-linear in the parameters:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_1 \beta_2 X_3 + \ldots + \varepsilon$$

No amount of algebra can turn the above into a linear model.

There are advanced methods to deal with this, e.g., non-linear least squares

Doesn't come up that often, because so many specifications *are* linear in the parameters

Lots of flexibility hidden in linear regression.

So just what is linear regression again?

We have expanded linear regression to encompass any model for unbounded continuous functions $f(\cdot)$ and arbitrary functions $g_k(\cdot)$:

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \varepsilon_i$$

We have expanded linear regression to encompass any model for unbounded continuous functions $f(\cdot)$ and arbitrary functions $g_k(\cdot)$:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{1i} + \ldots + \varepsilon_i \\ f(y_i) &= \beta_0 + \beta_1 g_1(x_{1i}) + \ldots + \varepsilon_i \end{aligned}$$

So just what is linear regression again?

We have expanded linear regression to encompass any model for unbounded continuous functions $f(\cdot)$ and arbitrary functions $g_k(\cdot)$:

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1 x_{1i} + \ldots + \varepsilon_i \\
f(y_i) &= \beta_0 + \beta_1 g_1(x_{1i}) + \ldots + \varepsilon_i \\
f(y_i) &= \beta_0 + \sum_{k=1}^{K} \beta_k g_k(x_{ki}) + \ldots + \varepsilon_i
\end{aligned}
$$

"Linearity" in linear regression just refers to the fact that the effect of a one unit change in $g(x_k)$ on $f(y)$ is β_k.

But the relationship between x_k and y themselves could be very non-linear, as a result of $f(\cdot)$ and $g_k(\cdot)$.

Interpreting models with transformed covariates or responses

Various methods:

- For logged response, $\hat{\beta}$ is the % change in Y for a level change in X
- For logged covariate, $\hat{\beta}$ is the level change in Y for a % changes in X
- For both sides logged, $\hat{\beta}$ is the % change in Y for % changes in X,
  known as the elasticity of Y with respect to X
- For polynomial coefficients, make a plot of $\hat{Y}$
- For interactions, make a plot of $\hat{Y}$

So how do we get $\hat{Y}$ for these models?

Could do it by hand fairly easily.

But what if we want confidence intervals around $\hat{Y}$ too?

Use `predict()`.

Key is setting `newdata` input correctly

If `lm()` did the transformations and interactions

So how do we get $\hat{Y}$ for these models?

Could do it by hand fairly easily.

But what if we want confidence intervals around $\hat{Y}$ too?

Use `predict()`.

Key is setting `newdata` input correctly

If `lm()` did the transformations and interactions

Then `predict()` will construct interactions and polynomials from their base terms as needed

Interpreting models with transformed covariates or responses

So how do we get $\hat{Y}$ for these models?

Could do it by hand fairly easily.

But what if we want confidence intervals around $\hat{Y}$ too?

Use `predict()`.

Key is setting `newdata` input correctly

If `lm()` did the transformations and interactions

Then `predict()` will construct interactions and polynomials from their base terms as needed

If not, you need to give predict properly constructed interactions and polynomials

What we've learned about linear regression so far

Now we know how to:

1. Specify a regression model
2. Estimate that model
3. Interpret our findings

Now we know how to:

1. Specify a regression model
2. Estimate that model
3. Interpret our findings

But how do we know if our findings are any good?
That we used the right specification?
That our model explained the data well or poorly?

Need to learn one more skill:

4. Select models with good fit