8 – Nonlinear effects

- Lots of effects in economics are nonlinear
- Examples
- Deal with these in two (sort of three) ways:
  - Polynomials
  - Logarithms
  - Interaction terms (sort of)

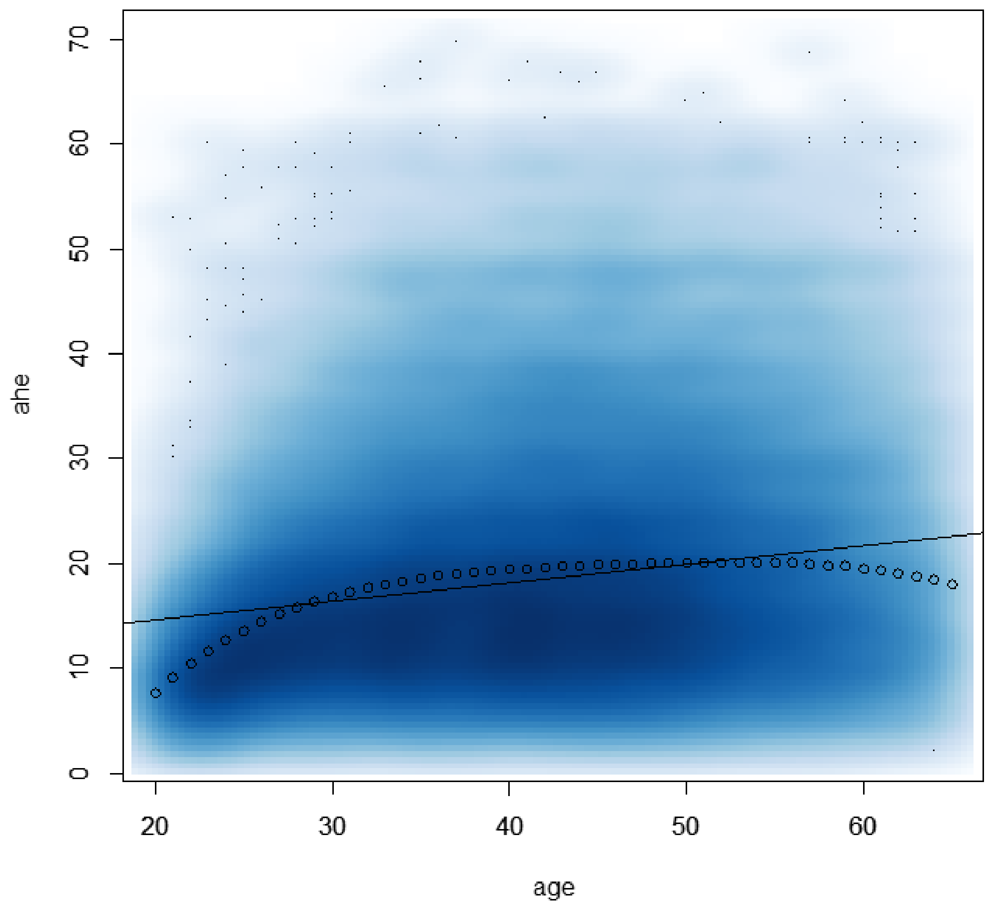
The linear model

Our models so far are linear.

- Change in Y due to change in X?
- See plots for:
  - *age* vs. *ahe*
  - *carats* vs. *diamond price*

If the true relationship is nonlinear, then the linear model is *misspecified*. (A sort of OVB). OLS is biased and inconsistent.

Average hourly earnings (*ahe*) and *age*. CPS data – over 60,000 observations. Linear model vs. polynomial model.



Nonlinear effects

If the relationship between Y and X is nonlinear:

- The effect of X on Y depends on the value of X
- The marginal effect of X is not constant
- Need to *specify* a population model that allows the marginal effect to *change* depending on the value of X

Polynomial regression model

The idea is that non-linear functions can be **approximated** using **polynomials**. For example, a polynomial function is:

$$y = a + bx + cx^2 + dx^3 + ex^4$$

This is a fourth-order polynomial. A second order polynomial is the familiar quadratic equation:

$$y = a + bx + cx^2$$

The validity of the approximation is due to the Taylor series approximation. See:

http://en.wikipedia.org/wiki/Taylor_series#/media/File:Exp_series.gif

We won't discuss the Taylor series here.

The (polynomial) population model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \cdots + \beta_r X_1^r + \epsilon$$

- This is just the linear model, but regressors are powers of X_1
- Other variables can be added as usual
- Estimation, hypothesis testing – same as usual
- NOT a violation of perfect multicollinearity
- Usually just a squared term is enough (quadratic model)
- βs are difficult to interpret

Exercise: For the model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \epsilon$, determine the effect of X_1 on Y.

Determining *r*

The degree of the polynomial can be determined by starting high and use *t* and *F* tests to get it smaller.

For example, in the model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \epsilon$$

The null hypothesis H_0: $\beta_2 = 0$, the null hypothesis says that X_1 has a linear effect, while the alternative hypothesis says it has a nonlinear effect.

Interpreting the estimated βs

In the model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \epsilon$$

β_1 and β_2 don't make much sense by themselves – they kind of go together.

To interpret the estimated regression:

- Plot predicted values
- Consider specific scenarios – take differences

<u>Exercise.</u> Use the diamond data.

a) Regress *price* on *carat*. Interpret your results.
b) Estimate a **quadratic** model.
c) Test the hypothesis that *carat* has a linear effect on *price*.
d) Interpret your results from the quadratic model.
e) Should we have used a **cubic** model?

Answers

a) Load the data:

```
diamond <-
read.csv("https://rtgodwin.com/data/diamond.csv")
```

Estimate:

```
summary(lm(price ~ carat, data=diamond))
```

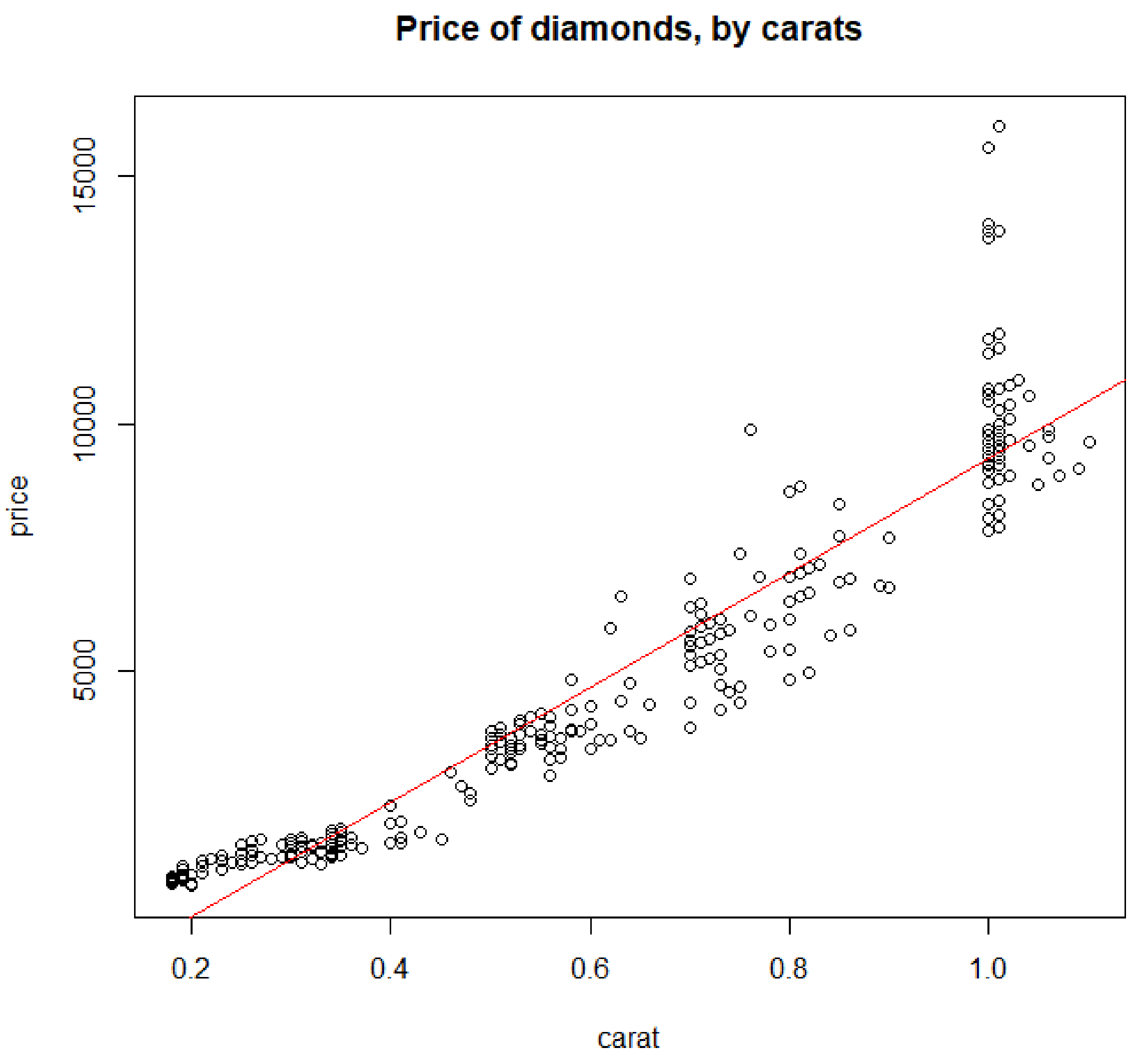
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2298.4      158.5  -14.50   <2e-16 ***
carat        11598.9      230.1   50.41   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1118 on 306 degrees of freedom
Multiple R-squared:  0.8925,  Adjusted R-squared:  0.8922
F-statistic:  2541 on 1 and 306 DF,  p-value: < 2.2e-16
```

Interpretation: when *carats* increases by 1, *price* increases by $**11599**. Or, for each 0.1 increase in *carat*, *price* increases by $**1160**.

Plot it:

```
plot(diamond$carat, diamond$price, main="Price of
    diamonds, by carats")
abline(lm(price ~ carat, data=diamond), col = "red")
```

Doesn’t look very good! The size of the diamond doesn’t matter – same marginal effect everywhere.


Price of diamonds, by carats
price
15000
10000
5000
0.2
0.4
0.6
0.8
1.0
carat

b) The quadratic model is:

$$price = \beta_0 + \beta_1 carat + \beta_2 carat^2 + \epsilon$$

We include the $carat^2$ variable in `lm()` using the `I()` function. We include the term:

```
carat^2
```

where the `^` is the power operator (shift-6).

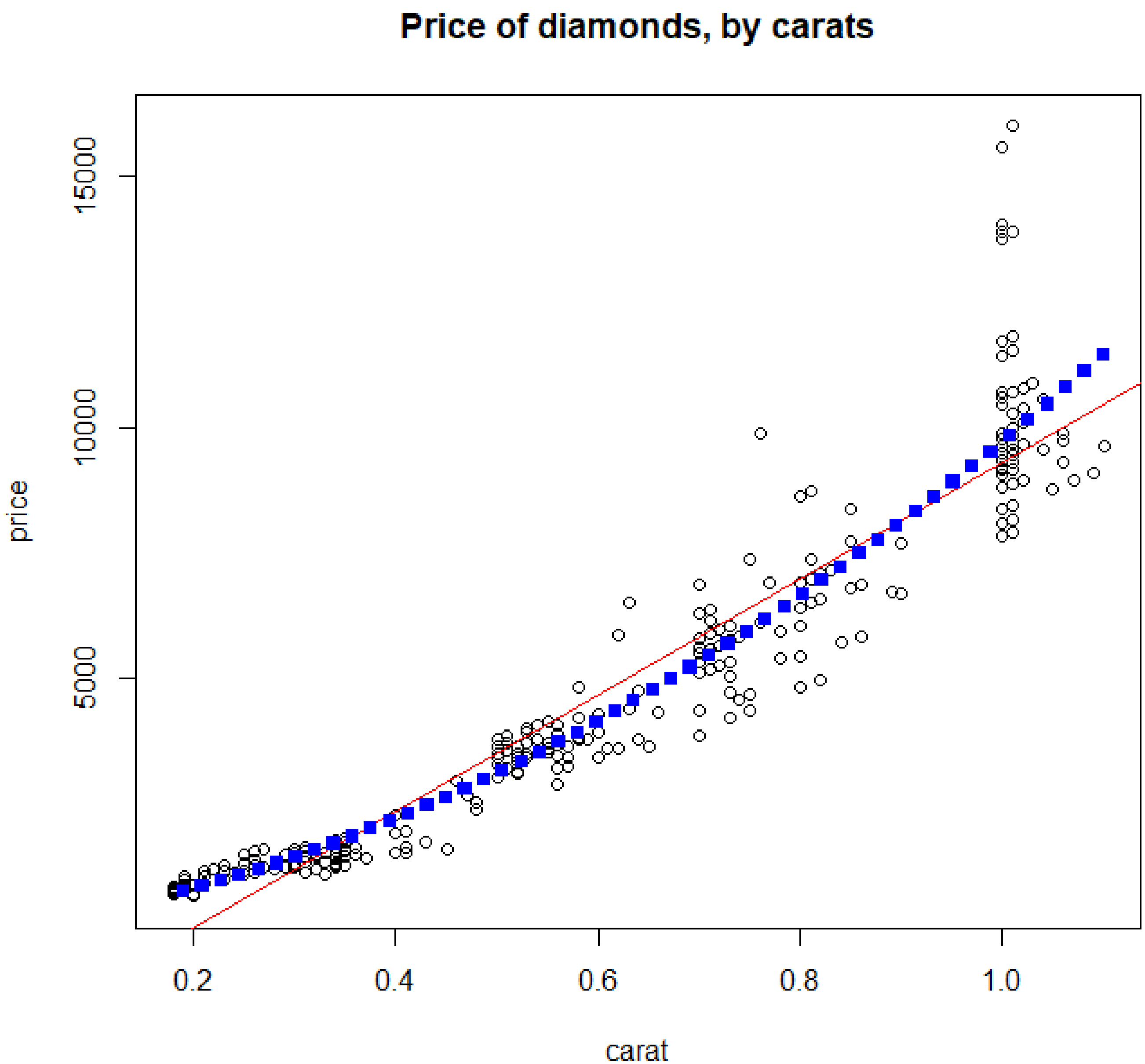
Estimate the quadratic model:

```
summary(lm(price ~ carat + I(carat^2), data=diamond))
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -42.51     316.37  -0.134   0.8932
carat        2786.10    1119.61   2.488   0.0134 *
I(carat^2)   6961.71     868.83   8.013  2.4e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1017 on 305 degrees of freedom
Multiple R-squared:  0.9112,  Adjusted R-squared:  0.9106
F-statistic:  1565 on 2 and 305 DF,  p-value: < 2.2e-16
```

c) Reject! Look at the `***` on `carat2`.

d) Interpretation is tricky. Sign of the squared term? We can draw it! Blue squares are some OLS predicted values.


Price of diamonds, by carats
price
15000
10000
5000
0.2
0.4
0.6
0.8
1.0
carat

The key is to consider specific scenarios (predicted values). For example, we could consider the effect of a 0.1 increase in *carats*, for different *carat* sizes.

$$\widehat{price}|_{carat=0.2} = -42.51 + 2786.10(0.2) + 6961.71(0.2^2)$$
$$= 793.18$$

$$\widehat{price}|_{carat=0.3} = -42.51 + 2786.10(0.3) + 6961.71(0.3^2)$$
$$= 1419.88$$

$$\widehat{price}|_{carat=0.3} - \widehat{price}|_{carat=0.2} = 626.70$$

A 0.1 increase in *carat* increases price by $627, when the diamond is small (0.2 carats). This effect was $1160 in the linear model.

```
predict(quadmod, data.frame(carat = 0.3)) -
     predict(quadmod, data.frame(carat = 0.2))
```

```
626.6952
```

We should consider a change under a different scenario.

$$\widehat{price}|_{carat=1} = -42.51 + 2786.10(1) + 6961.71(1^2) = 9705$$

$$\widehat{price}|_{carat=1.1} = -42.51 + 2786.10(1.1) + 6961.71(1.1^2) = 11446$$

$$\widehat{price}|_{carat=1} - \widehat{price}|_{carat=1.1} = 1741$$

A 0.1 increase in *carat* increases price by $1741, when the diamond is large (1 carat). This effect was $1160 in the linear model.

(In the nonlinear model, the marginal effect depends on the size of the diamond).

e) Estimate a **cubic** model:

$$price = \beta_0 + \beta_1 carat + \beta_2 carat^2 + \beta_3 carat^3 + \epsilon$$

To estimate the model, use:

```
summary(lm(price ~ carat + I(carat^2) + I(carat^3),
  data=diamond))
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    786.3      765.4   1.027   0.3051
carat        -2564.2     4636.9  -0.553   0.5807
I(carat^2)   16638.9     8185.3   2.033   0.0429 *
I(carat^3)   -5162.5     4341.9  -1.189   0.2354
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1017 on 304 degrees of freedom
Multiple R-squared:  0.9116,	Adjusted R-squared:  0.9107
F-statistic:  1045 on 3 and 304 DF,  p-value: < 2.2e-16
```

carat^3 is insignificant. The quadratic specification is good enough.