Computational Statistics 9.3: Model specification

Categories, quadratics, interactions

Independent variables

Categories and dummies
- **Factors in R**
- **Factors 2**
- **Factors 3**

Quadratic effects
- **Quadratic example**
- **Quadratic 2**
- **Quadratic 3**
- **Quadratic 4**
- **Quadratic 5**

<
>

Computational Statistics 9.3: Model specification

Categories, quadratics, interactions

Overview

This lesson introduces a variety of independent variable types.

Objectives

After completing this module, students should be able to:

1. Estimate and interpret regressions using categorical independent variables.
2. Estimate and interpret regressions using quadratic independent variables.
3. Estimate and interpret regressions using interactions between independent variables.

Readings

Schumacker, Chapter 17.

Independent variables

So far we have treated our variables as all of the same type: continuous numerical measures. But often we have different types of variables, such as binary or categorical. We can also have different ways that the independent variables affect the dependent variable: perhaps two variables interact, or instead of having a simple linaer effect, the effect of a variable rises and then falls as the variable increases. Many of these variants can be handled using basic multiple regression, but others require special recoding of variables or otherwise tweaking the regression.

Categories and dummies

Binary independent variables are not a problem for the default regression methods. It is easy to interpret the change in x, and to measure the change in y. Binary dependent variables are a whole other problem though – the problem being that you might sometimes have a binary dependent variable (went to college = 1, didn't = 0), and then get predicted values outside this range ($\hat{y} = 3.6$ means what?). So you need to constrain your model to stay within 0 and 1 for the dependent variable. This requires a special modeling approach (the *logit*) that we will return to next week, but for binary independent variables, we are ok.

Categorical independent variables (ie, more than 2 categories) are a little trickier. Say we are interested in the effect of religious affilitation on happiness, and we have five different religions that we have measured. What would a 1-unit change in X mean? We of course could change X to a different question, like "How religious are you on a 1-10 scale?", but that's not the same as measuring the effect of being Christian vs Hindu, say, on happiness.

When you have an IV with relatively few categories, the solution is usually *dummies*. That is, you create a bunch of binary variables, one for each category, out of your categorical variable (ie, Hindu=1, non-Hindu=0, and the same for each category). This approach is so common that R has a built-in method to do it automatically.

There's just one thing to watch out for. If you do it yourself, or you are interpreting R, you should always have exactly $d - 1$ dummies if there are d categories. Say instead you have 10 religions and 10 0/1 dummies. What does a change in Hindu=0 to Hindu=1 mean? Well, we know that Hindu=1 means the person is Hindu. But if there are 10 options, what does Hindu=0 mean? Does it mean that the person is Christian? Muslim? Atheist? A one-unit change means a change from something to something else, but this makes no sense if we don't know what the 0 means.

So for this reason, there are always $d - 1$ dummies. That means that one of the categories is omitted, and becomes the 0 – the basis against which the others are compared. For religions, we might make "aetheistic" or "no religion" the omitted category, and then treat everything relative to that. For other categorical variables, it is less clear that one is clearly the default, and in any case, it doesn't matter, because it's all relative. Say our three-category independent variable is people's favorite color, and there are only three options: red, green, or blue. Say "green" is the omitted dummy. Then the effect of the "red" dummy is the effect of switching from green to red, and the "blue" dummy is the effect of switching from green to blue; the effect of red->green or blue->green is just the opposite. What about the effect of

blue->red? That's just the difference between $\beta_{red} - \beta_{blue}$. There isn't really an "effect of green," since it's all relative, so the omitted dummy is not a problem. We'll see one more way to look at it in the next example.

Factors in R

Categorical variables are known as factors in R. Let's create some simulated data for a y that is a function of a numerical variable x1 and a categorical variable x2, where x2 has three categories (eg, three religions), where each category has a different effect on y.

```
x1 <- rnorm(1000)
x2 <- floor(runif(1000,1,4))
y <- 2* x1 + 2*rnorm(1000)
for(i in 1:length(y)){
  if(x2[i] ==1){
    y[i] <- y[i] + 1
  }
  if(x2[i] ==2){
    y[i] <- y[i] + 5
  }
  if(x2[i] ==3){
    y[i] <- y[i] + 2
  }
}
df <- data.frame(y=y,x1=x1,x2=as.factor(x2))
head(df)
```

```
          y          x1 x2
1 3.3828831  0.45742442  2
2 7.1737044  1.72282753  2
3 0.5073509 -0.05300903  3
4 0.7851293  0.05970947  1
5 5.3772296  0.45222292  1
6 0.2717828  0.28928403  1
```

So being in category 1 gives you a boost of +1 for y; being in category 2 gives you a boost of +5; and category 3 has a boost of +2. Note that each of the category values for x2 as stored in the data frome looks like a number, but internally these are stored as categories – they could as well be "apple," "orange" and "pear" for all R cares.

Factors 2

Now lets run our regression:

```
summary(lm(y~x1+x2,data=df))
```

```
Call:
lm(formula = y ~ x1 + x2, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-7.3513 -1.2735  0.0169  1.3581  6.1431

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.85267    0.10996   7.754 2.19e-14 ***
x1           1.96593    0.06134  32.048  < 2e-16 ***
x22          4.12391    0.15320  26.918  < 2e-16 ***
x23          1.16131    0.15396   7.543 1.03e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.971 on 996 degrees of freedom
Multiple R-squared:  0.6332,    Adjusted R-squared:  0.6321
F-statistic:   573 on 3 and 996 DF,  p-value: < 2.2e-16
```

So what's going on? The coefficient for x1 is straight-forward enough – about 2, as it should be. R, knowing x2 is a factor, has automatically turned it into two dummies, omitting the third, which by (arbitrary) default is the first category. Thus x22 is the effect of shifting from category 1 to category 2, and x23 is the effect of shifting from category 1 to category 3. The effect of the first shift is going from +1 to +5, or +4; the second is going from +1 to +2, or +1. The effect of shifting from category 2 to category 3 is then 2 - 5, or -3.

Incidentally, if we wanted to choose a different category as our reference (omitted) category, we can use `relevel`:

```
df$x2 <- relevel(df$x2,ref=3)
```

This sets "3" as the omitted category. What would you think the coefficients on $\beta_{x2=1}$ and $\beta_{x2=2}$ to be now?

- **4 and 1**
- **-4 and -3**
- **-1 and 3**

Factors 3

Finally, let's think about the intercept. When we created y, we didn't add any constant anywhere. But when category 2 and category 3 are both 0, that means (the omitted) category 1 must be true – ie, the default state is +1, so that gets folded into the intercept. Another way to think about it is that each category is like its own intercept. Here's our prediction equation for y:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_{22} + \beta_3 x_{23}$$

When $x_{22} = 1$ (ie, for someone of category 2), our prediction is:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2$$

When $x_{23} = 1$, our prediction is:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_3$$

And when the predicted individual/observation is of category 1, both x22 and x23 are 0, so we have:

$$\hat{y} = \beta_0 + \beta_1 x_1$$

Thus that β_0 contains whatever boost you get by being category 1. Of course, the intercept without that boost could be anything – but since we've defined our categories such that you can never be *none* of them, it's meaningless to say "What's β_0 if all three categories are 0?" It's all relative with categories. To sum up, the coefficients we see on the regression are always the effect of shifting from the omitted category to the dummy category.

Quadratic effects

Often it's the case that we think that the effect of some variable on our dependent variable is not simply linear, but maybe goes up and then back down, or levels off, or accelerates as it increases. For instance, we might expect that the effect of age on voter turnout is that your tendency to turn out to vote increases as you get older, until you get quite old, at which point increased age (and infirmity) starts decreasing your likelihood of voting again.

One standard way of modeling these non-linear effects is through quadratic variables. If you recall from basic math, our quadratic equation is of the form:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

where β_0, β_1 and β_2 determine the shape of the curve, including whether it is concave up or concave down.

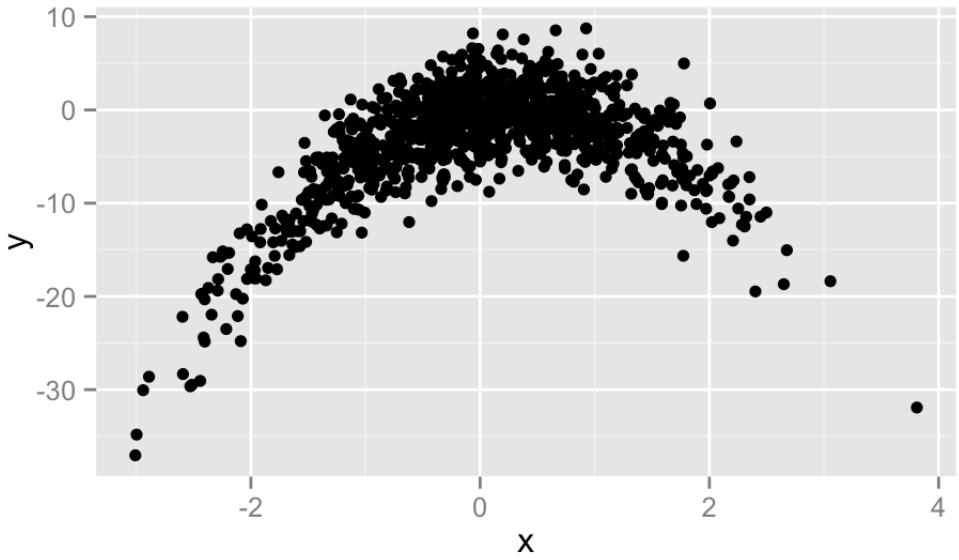
Quadratic example

As usual, it's easiest to see how this works when we play god, and create the data ourselves. Let's create y as a quadratic function of x:

```
set.seed(1)
x <- rnorm(1000)
y <- 2*x - 3*x^2 + 3*rnorm(1000)
df <- data.frame(y=y,x=x)
```

Now let's examine the data:

```
library(ggplot2)
ggplot(data=df,aes(x=x,y=y)) + geom_point()
```



Clearly we have a curvilinear relationship, not just a linear relationship.

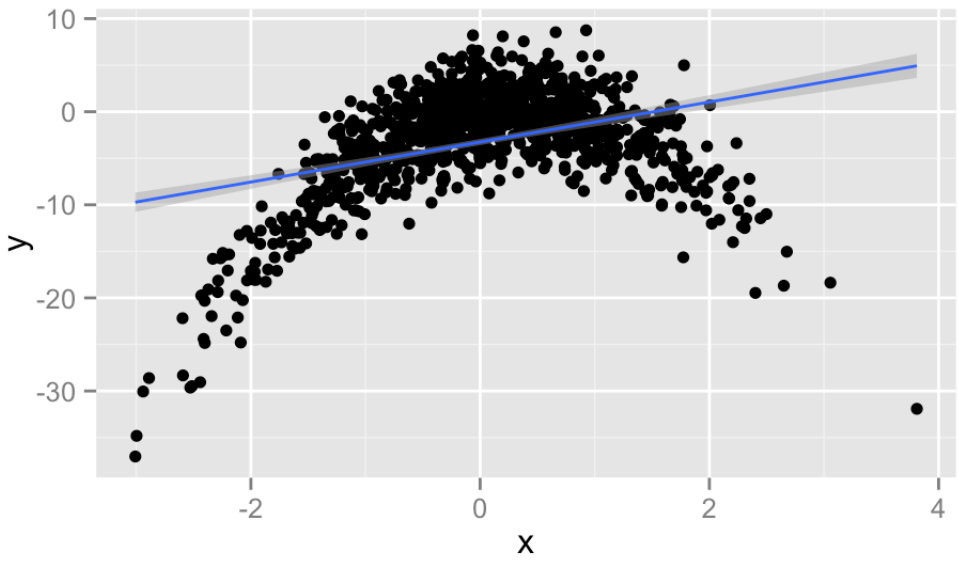
Quadratic 2

We can treat this is a linear relationship and just run a linear regression, and even get a decent approximation:

```
summary(lm(y~x,data=df))
```

```
Call:
lm(formula = y ~ x, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-36.842  -2.240   0.810   3.534  11.599

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3.2574     0.1720  -18.94   <2e-16 ***
x             2.1487     0.1663   12.92   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.439 on 998 degrees of freedom
Multiple R-squared:  0.1433,    Adjusted R-squared:  0.1425
F-statistic:   167 on 1 and 998 DF,  p-value: < 2.2e-16
```

```
ggplot(data=df,aes(x=x,y=y)) + geom_point() + geom_smooth(method=lm)
```



But clearly we are missing something in this relationship.

Quadratic 3

Even though y is a function of just one variable, x, if we want to capture the quadratic effect of x, we essentially have to create a new variable equal to x^2. If we create this new variable and add it into our regression, we capture exactly the relationship behind the scenes:

```
x2 <- x^2
summary(lm(y~x+x2))
```

```
Call:
lm(formula = y ~ x + x2)

Residuals:
    Min      1Q  Median      3Q     Max
-9.8683 -2.0394 -0.0461  2.2692 10.8088

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.10976    0.12092  -0.908    0.364
x            2.02177    0.09548  21.175   <2e-16 ***
x2          -2.94278    0.06528 -45.083   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.122 on 997 degrees of freedom
Multiple R-squared:  0.7181,    Adjusted R-squared:  0.7175
F-statistic:  1270 on 2 and 997 DF,  p-value: < 2.2e-16
```

As usual, R has an easier way to handle quadratic variables, though it's not quite just putting in a `+ x^2` into the `lm()` function. Instead, we have to wrap the `x^2` inside the indicator function `I()` to let R know it should create this new variable on the fly:

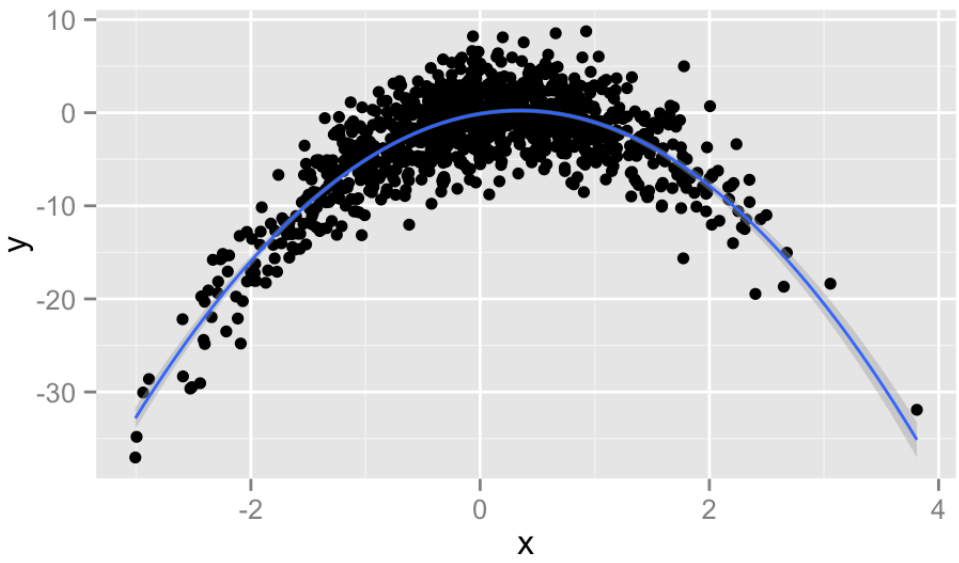
```
summary(lm(y~x+I(x^2)))
```

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

Residuals:
    Min      1Q  Median      3Q     Max
-9.8683 -2.0394 -0.0461  2.2692 10.8088

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.10976    0.12092  -0.908    0.364
x            2.02177    0.09548  21.175   <2e-16 ***
I(x^2)      -2.94278    0.06528 -45.083   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.122 on 997 degrees of freedom
Multiple R-squared:  0.7181,    Adjusted R-squared:  0.7175
F-statistic:  1270 on 2 and 997 DF,  p-value: < 2.2e-16
```

Quadratic 4

We can also plot this a bit more nicely using the “polynomial” formula option in ggplot, where a quadratic equation is a polynomial of degree 2.

```
ggplot(data=df,aes(x=x,y=y)) + geom_point() + geom_smooth(method = "lm", formula = y ~ poly(x, 2))
```



Here we can see that unlike our bad linear approximation, the regression with a quadratic term matches the data very nicely.

Quadratic 5

So how do we interpret our coefficients in the regression output? One shorthand is that if the coefficient on the x^2 term is negative, then generally that means the curve is concave down (ie, it rises and then falls); if it's positive, then the curve is concave up; and if it's not significant, that means that there's no significant curved effect, and the effect of X is (at best) linear.

But more importantly, you have to be careful about what your claims are of the effect of X. A one-unit change in X now has two pathways for influencing y – via the linear effect, and via the quadratic effect. And as you can see from the picture in the previous slide, the effect of increasing X by 1 on Y varies depending on where you are along the curve! Sometimes it can be positive, sometimes negative.

A common way to illustrate the effect of x is to examine the effect of increasing x by one unit near the center of X, since as from $\bar{x}$ to $\bar{x}+1$. We can just estimate this effect by plugging the mean value $\bar{x}$ into our prediction for $\hat{y}$, and then plugging in $\bar{x}+1$, and taking the difference. Since in this case the mean of x is 0 (since x was constructued using `rnorm(1000)`), $\hat{y}$ at $x=0$ is 0 (because we know $y=2x-3x^2$), and $\hat{y}$ at $x=1$ is -1, so the effect of a one-unit change from $\bar{x}$ to $\bar{x}+1$ is -1. Of course, we usually do these calculations not using the exact coefficients (because we never know them except for simulated data), but rather using the regression coefficients, but the math is the same.

For those who know a little calculus, we can also calculate the peak or trough of the curve, so that we can identify where the effect of X switches from negative to positive, or vice versa. Since our equation for y is $y=\beta_0+\beta_1 x+\beta_2 x^2$, if we take the derivative and set it to 0 (which is where the slope of the quadratic curve is 0), we have $0=\beta_1+2\beta_2 x$. If we solve that for x to find the x value where the slope is 0 (ie, the peak or trough of the curve), we have $x=-\beta_1/2\beta_2$. We'll show an example of using this in the next slide.

Quadratic example

Just to see an example with real data, here is ideology as a function of all our usual variables plus education squared, to see whether the increasing conservativism that comes with education perhaps reverses at higher education levels:

```
anes_2008tr <- read.table("anes_2008tr.csv",sep=",",header=TRUE,stringsAsFactors=FALSE)
summary(lm(ideology_con ~ age + gender_male + race_white +
              education + income + I(education^2),data=anes_2008tr))
```

```
Call:
lm(formula = ideology_con ~ age + gender_male + race_white +
    education + income + I(education^2), data = anes_2008tr)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5521 -0.4285 -0.0408  0.6096  3.3077

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.241690   0.201589  16.081  < 2e-16 ***
age             0.006238   0.001478   4.222 2.51e-05 ***
gender_male     0.113361   0.053602   2.115 0.034546 *
race_white      0.266574   0.055299   4.821 1.52e-06 ***
education       0.158078   0.089168   1.773 0.076393 .
income          0.103788   0.026963   3.849 0.000122 ***
I(education^2) -0.028015   0.010575  -2.649 0.008123 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.271 on 2315 degrees of freedom
Multiple R-squared:  0.03494,   Adjusted R-squared:  0.03244
F-statistic: 13.97 on 6 and 2315 DF,  p-value: 1.135e-15
```

As we can see, while the coefficient on education remains positive, now there is a negative coefficient on education^2, suggesting that there is indeed a liberalizing effect for higher education levels. We can use the equation from the previous slide to see exactly where the effect of education shifts from being pro-conservative to pro-liberal: $x = -\beta_1/2\beta_2 = -0.158078/(2 * -0.028015) = 2.82$ As it happens, since 3 is high school, the effect of education seems to boost conservatism up through high school, but then any additional education after high school seems to boost liberalism instead. To put it another way, people with a high school or incomplete college education tend to be the most conservative, with those with either more or less education than that being more liberal.

Interaction terms

The last important independent variable characteristic to cover here is interactions. Often the effects of two variables are not independent, but interact. For instance, the effect of talking on a cell phone on driving might be negative, and the effect of being drunk while driving might be negative, but there might be an especially large negative effect if one is doing both – not just the sum of the two, but more than the sum, as the two variables interact.

Let's construct a system where two variables interact – ie, where y gets a boost not just from them separately, but also from a synergistic interaction between the two. We can think of y as education, and the two x variables as income and hard work: both will independently boost education, but when you

have both at the same time, higher education levels are much more likely. Our equation (ignoring all the other variables that might be affecting y) is now:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$$

We can create some simulated data and then run a regression quite easily:

```
x1 <- rnorm(1000)
x2 <- rnorm(1000)
y <- x1 + 2*x2 + 3*x1*x2 + 5*rnorm(1000)
df <- data.frame(y=y,x1=x1,x2=x2)
intreg <- lm(y ~ x1 + x2 + x1*x2,data=df)
summary(intreg)
```

```

Call:
lm(formula = y ~ x1 + x2 + x1 * x2, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-18.7618  -3.2348   0.0194   3.1611  13.7726

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.09857    0.15593  -0.632    0.527
x1           0.64849    0.15132   4.286    2e-05 ***
x2           2.01100    0.15017  13.392   <2e-16 ***
x1:x2        3.29559    0.14145  23.299   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.93 on 996 degrees of freedom
Multiple R-squared:  0.4259,    Adjusted R-squared:  0.4241
F-statistic: 246.3 on 3 and 996 DF,  p-value: < 2.2e-16
```

This time it really is as easy as just adding `x1*x2` to the regression. Again, having constructed the data, the interpretation of the regression results is straightforward. But once again, it's not a trivial task to say what the effect of x1 is – after all, the effect of x1 now depends on x2!

Illustration

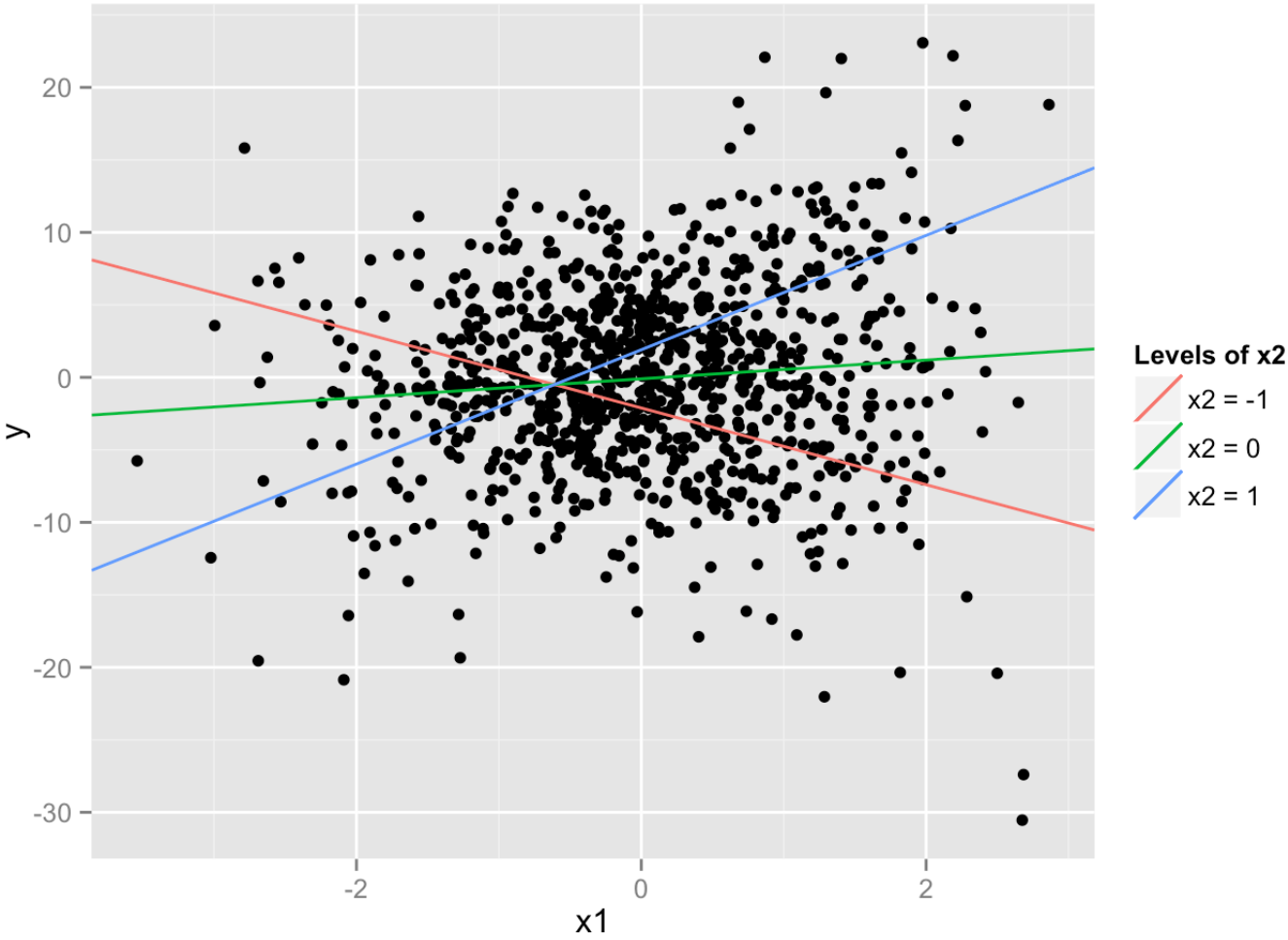
We can again illustrate this with ggplot, although it helps to think about our formula briefly. What's the effect of a change in x1? Well, we can re-arrange our terms in the formula above and get:

$$y = \beta_0 + (\beta_1 + \beta_3 x2)x_1 + \beta_2 x_2 + \epsilon$$

Thus the effect of changing x1 by +1 unit is to increase y by β_1 units, plus $\beta_3 x2$ units, which will depend on the size of x2 – if x2 is 1, the total effect is $\beta_1 + \beta_3$, whereas when x2 is, say, -5, the total effect is $\beta_1 - 5 * \beta_3$.

To visualize this, what we can do is plot the effect of x1 on y for various levels of x2, to see how the effect of x1 varies as it interacts with x2. Keep in mind the equation from the previous slide – we are just plugging in different values if x2 (-1, 0, and 1, in this case), and plotting the remaining equation as a function of x1, via a series of different lines, each one corresponding to a different value of x2. We collapse $\beta_0 + \beta_2 x_2$ into a single intercept, because x2 is set at a specific value (eg, -1), which makes this a single number. The remainder – $\beta_1 x_1 + \beta_3 x_1 x_2$ – is the slope, because both parts are a function of x1, which is the variable we are considering.

```
beta0 <- intreg$coef[1]
beta1 <- intreg$coef[2]
beta2 <- intreg$coef[3]
beta3 <- intreg$coef[4]
ggplot(df, aes(x = x1, y = y)) + geom_point() +
geom_abline(aes(color = "x2 = -1"),  intercept = beta0 + -1 * beta2,
            slope = beta1 - 1 * beta3, show_guide = TRUE) +
geom_abline(aes(color = "x2 = 0"),  intercept = beta0 + 0 * beta2,
            slope = beta1 + 0 * beta3, show_guide = TRUE) +
geom_abline(aes(color = "x2 = 1"), intercept = beta0 + 1 * beta2,
            slope = beta1 +  1 * beta3, show_guide = TRUE) +
guides(color = guide_legend(title = "Levels of x2"))
```



What we see here is that when x2 is positive (eg, x2 = 1), the effect of x1 is positive, and vice versa when x2 is negative. If this were a real-world example, it would mean that with high levels of income (x2), harder work (x1) got you more education, but with low levels of income, harder work actually got you less education. Let's hope that's not the case!

Interaction example

Once again, we can always calculate the effect of a 1 unit (or any change) in x1 on x2 by choosing a set value for the other variable (eg, setting x2 to its mean) and then taking the difference in the predicted $\hat{y}$ values at two different values of x1.

Here's our ideology data again, this time with an interaction between race and income:

```
summary(lm(ideology_con ~ age + gender_male + race_white +
               education + income + income*race_white,data=anes_2008tr))
```

```
Call:
lm(formula = ideology_con ~ age + gender_male + race_white +
    education + income + income * race_white, data = anes_2008tr)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5788 -0.4316 -0.0208  0.5804  3.3603

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        3.843333   0.132143  29.085  < 2e-16 ***
age                0.005534   0.001454   3.806 0.000145 ***
gender_male        0.109589   0.053618   2.044 0.041079 *
race_white        -0.057007   0.146144  -0.390 0.696518
education         -0.073358   0.017991  -4.077 4.71e-05 ***
income             0.036364   0.037571   0.968 0.333202
race_white:income  0.124357   0.050267   2.474 0.013435 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.271 on 2315 degrees of freedom
Multiple R-squared:  0.03457,   Adjusted R-squared:  0.03207
F-statistic: 13.82 on 6 and 2315 DF,  p-value: 1.74e-15
```

Ignore, for the moment, the question of which of these variables are statistically significant. If we want to examine the effect of income, we now see it in two ways: the direct effect, and the interaction effects. What this means is that, for different races (white and non-white, in this case), the effect of income may differ. What is the effect of income on ideology for white people? Well, for white people, `race_white` = 1. So the effect is $\beta_1 + \beta_3 * x_2 = \beta_1 + \beta_3 = 0.036 + 0.124 = 0.16$ For non-white respondents, the estimated effect is $\beta_1 + \beta_3 * x_2 = \beta_1 = 0.036$ Thus while it appears that income makes people more conservative, this effect is much stronger for whites than non-whites: every additional dollar of income has a much more conservativizing effect for white than it does for non-whites – over 4 times as strong!

Testing nested models

Note that while the interaction term in the previous regression was statistically significant, the addition of the interaction has rendered both income and race alone insignificant. But that doesn't mean they don't matter: when you have an interaction, the effect of the interacted variables gets split among the various pathways – the direct effect, and the interaction effect. This can also happen when you add a quadratic term, where sometimes the total effect can be split between the linear and the quadratic effects. What this means is that you can't just use the p value on a single coefficient to test whether a pair of variables (or any group of them) signficantly contribute to the regression.

Instead, we need a test that tests a bunch of variables at once: the complete regression with all the variables, versus a reduced regression without income and income*race (for instance); or a regression with everything, versus a regression without education and education^2.

Luckily, it turns out this test is both easy and familiar: it's just another F test. This time, we are not testing *all* the variables at once, as we did in the previous module; instead, we are testing one set of variables (the complete model) against a subset of variables (the reduced model, where we have droped 1 or more variables). The null hypothesis is that that complete model does no better than the reduced model (ie, that the variables you are debating including are not significant; this is the same null as we have for a sigle variable when we examine the t statistic on its beta coefficient). And as usual, if we get a F statistic that is sufficiently large, then we conclude that the complete model with the extra variables is signficantly better than the reduced model in explaining y, and thus we are justified in including all the variables under debate.

The F statistic equation for testing these nested models (where the reduced model is a subset of the complete one) is actually quite similar to what we saw for the F test of all the variables:

$$F = \frac{R_c^2 - R_r^2 / df_1}{(1 - R_c^2)/df_2}$$

Where $df_1 =$ number of additional variables in the complete model (eg, 2) and $df_2 = n - k - 1$ for the complete model (ie, k is the total number of independent variables in the complete model). R_c^2 is the R^2 for the complete model, and similarly for the reduced. Basically, the bigger $R_c^2 - R_r^2$ is, the bigger the gain in explanatory power of the complete model over the reduced, and thus the more likely the extra variables in the complete model are significant.

Testing nested models in R

We can test for whether the pair of variables, income and income*race, belong in our full model, by doing an F test on the complete model vs the reduced model without either term:

```
complete <- lm(ideology_con ~ age + gender_male + race_white +
               education + income + income*race_white,data=anes_2008tr)
reduced <- lm(ideology_con ~ age + gender_male + race_white +
               education,data=anes_2008tr)
anova(reduced, complete)
```

```
Analysis of Variance Table

Model 1: ideology_con ~ age + gender_male + race_white + education
Model 2: ideology_con ~ age + gender_male + race_white + education + income +
    income * race_white
  Res.Df    RSS Df Sum of Sq    F    Pr(>F)
1   2317 3771.6
2   2315 3738.9  2    32.625 10.1 4.292e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We do the test with the `anova()` function, which as we covered earlier, is just an F test. We can verify this by calculating it by hand:

```
r2c <- summary(complete)$r.squared
r2r <- summary(reduced)$r.squared
fstat <- ((r2c - r2r) / 2) / ((1 - r2c) / (2322 - 6 - 1))
fstat
```

```
[1] 10.10008
```

```
pf(fstat,2,(2322-6-1),lower.tail=F)
```

```
[1] 4.291568e-05
```

This shows that the income and income*race terms definitely both belong in the regression. If we were extra worried, we could also test the complete model against a reduced one with just the elimination of the interaction term, since maybe it's ambiguous about the importance of that term, given that when it is added, the interaction is signficant, but the income variable no longer is. If you do that F test (try it!) you will find that the addition of just the interaction still boosts the model signficantly, with an F test p value of 0.013 – exactly what one sees on the interaction coefficient back when we ran the original complete regression.

When to include quadratic or interaction terms

There is no definite answer to this question. Mainly you need to be guided by your substantive knowledge of the material. Sometimes it seems plausable that there might be a quadratic effect (such as the effect of age on turnout) or an interaction effect (such as the interaction between race and education in affecting ideology), but sometimes you can't think of a story for why that should be the case, and then you should be careful. What you don't want to do is try a quadratic term for every variable, or interact every pair of variables you have. That's fishing, and you are likely to find something significant just by chance if you do that. The best is to be a bit exploratory, and use t statistics on coefficients and F tests on nested models to aggressively reject these additional variables if they don't strongly contribute to

your model. In the end, you are looking for true results that will guide policy or other actions without wasting your time. Be careful of false positives – but be mindful that interactions or non-linear effects can tell both important and interesting stories!