Discussion Section

Week 2

Bret Stevens

University of California, Davis

Admin

- Thanks for coming to the make up!
- We still have normal discussion on Thursday
- The midterm is this upcoming Monday
  - Focus on the theoretical material from the homeworks
  - The test will be largely based off of that material
  - They will not be the same questions
  - You should feel **very** comfortable with the homework material
- I will have extra office hours this week, as well as the morning of the exam

R^2 Review

- You have seen this material 3 times now, so I want you to tell me what R^2 is

R^2 Review

- You have seen this material 3 times now, so I want you to tell me what R^2 is
- Can anyone give me the basic textbook definition?

R^2 Review

- You have seen this material 3 times now, so I want you to tell me what R^2 is
- Can anyone give me the basic textbook definition?
    - I'm not afraid to call on unsuspecting volunteers, so please speak up for your classmates sake

R^2 Review

- You have seen this material 3 times now, so I want you to tell me what R^2 is
- Can anyone give me the basic textbook definition?
  - I'm not afraid to call on unsuspecting volunteers, so please speak up for your classmates sake
- Now we're going to do the derivation of R^2, like in the HW
  - Since you've done this before, I expect you to all be able to do it
  - We're going to crowd-source the answer
  - Before we do, let's remember some important facts

R^2 Review - Fact 1

$\bar{e} =$?, why?

R^2 Review - Fact 1

$\bar{e} =$?, why?

- Remember when we derived b_0 and b_1 in the basic case? We minimized the sum of the squared errors:

$$\min_{b_0, b_1} \sum_i (y_i - b_0 - b_1 x_i)^2$$

When we take the first order condition for b_0 and find:

$$\sum_i -2(y_i - b_0 - b_1 x_i) = 0$$

$$\sum_i (y_i - b_0 - b_1 x_i) = 0$$

Remember,

$$e_i = y_i - b_0 - b_1 x_i$$

R^2 Review - Fact 1

- So we have:

$$\sum_i e_i = 0$$

If we multiply each side by $\frac{1}{n}$:

$$\frac{1}{n}\sum_i e_i = 0$$

$$\Rightarrow \bar{e} = 0$$

This is true for all regressions

R^2 Review - Fact 2

$\bar{\hat{y}} =$?, why?

R^2 Review - Fact 2

$\bar{\hat{y}} =$?, why?

- What is $\hat{y}$?

R^2 Review - Fact 2

$\bar{\hat{y}} =$?, why?

- What is $\hat{y}$?
  - Its the predicted value from our regression
  - So if we used our b_0 and b_1 to predict someone's value of y_i we would get $\hat{y}_i$
  - We can say:

$$y_i = \hat{y}_i + e_i$$

Lets take the average of both sides:

$$\frac{1}{n}\sum_i y_i = \frac{1}{n}\sum_i \hat{y}_i \frac{1}{n}\sum_i e_i$$

$$\bar{y} = \bar{\hat{y}} + \bar{e}$$

$$\bar{y} = \bar{\hat{y}}$$

R^2 Review - Derivation

Okay, now lets derive R^2

Appendix

Multiple Regression

How is multiple regression different than simple regression?

Multiple Regression

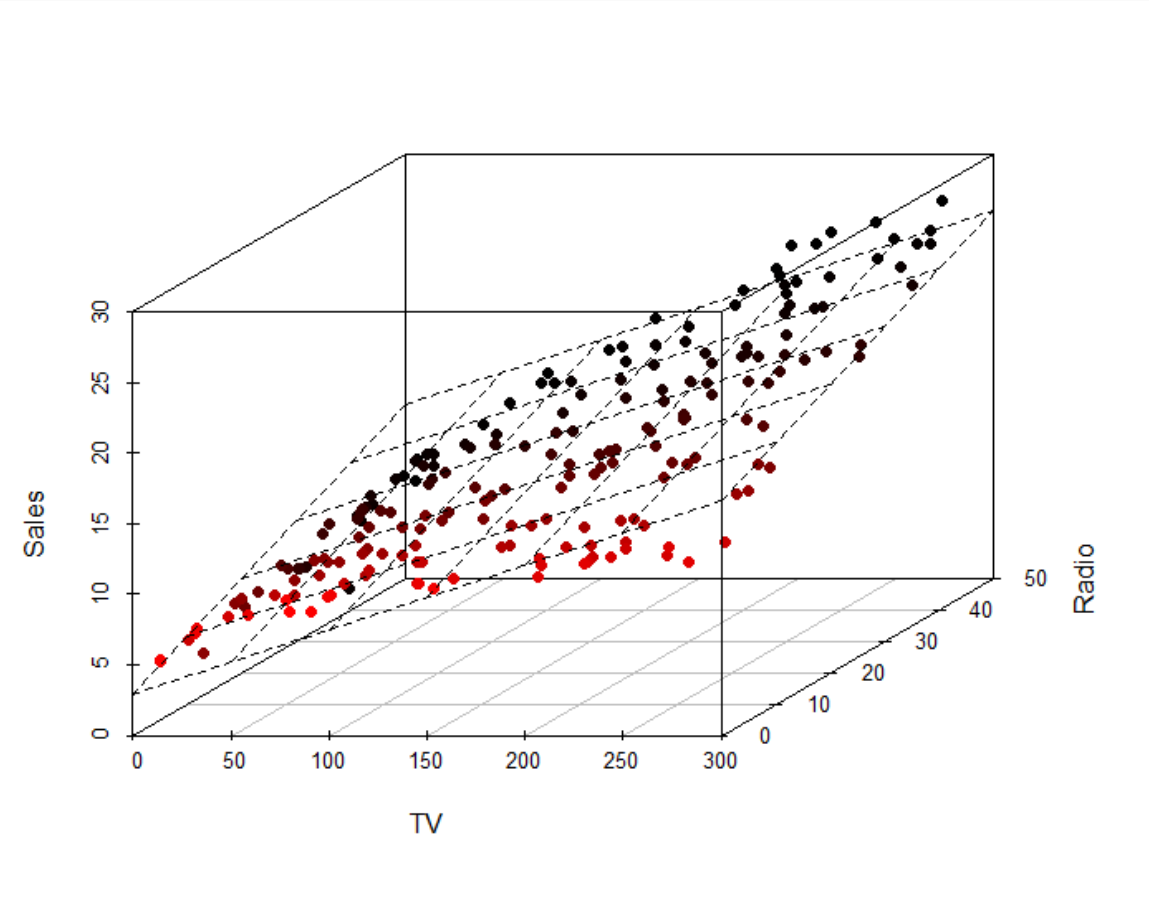
How is multiple regression different than simple regression?

- In simple regression we define the “best” line to fit the data
- In multiple regression we define the “best” plane or hyperplane to fit the data

Multiple Regression

How is multiple regression different than simple regression?

- In simple regression we define the “best” line to fit the data
- In multiple regression we define the “best” plane or hyperplane to fit the data



Multiple Regression

Before we wrote our econometric models like:

$$y_i = b_0 + b_1 x_i + e_i$$

With multiple regression with k regressors:

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_k x_{ki} + e_i$$

Multiple Regression

Before we wrote our econometric models like:

$$y_i = b_0 + b_1 x_i + e_i$$

With multiple regression with k regressors:

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_k x_{ki} + e_i$$

With multiple regression we can ask many more questions than we can with simple regression. Instead of "How does years of schooling affect income" we can ask "How does years of schooling affect income, accounting for age, race, and parents education". This lets us feel more confident that we are really "identifying" the effect of the variable we are interested in. Using multivariate regression can also help us make better predictions and compare the magnitude of effects between different variables.

Multiple Regression

In simple regression we had:

$$
\begin{aligned}
b_0 &= \bar{y} - b_1 \bar{x} \\
&= \bar{y} - \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} \bar{x} \\
&= \bar{y} - \frac{Cov(x, y)}{Var(x)} \bar{x}
\end{aligned}
$$

$$
\begin{aligned}
b_1 &= \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} \\
&= \frac{Cov(x, y)}{Var(x)}
\end{aligned}
$$

Multiple Regression

In a bivariate regression we have:

$$b_0 = \bar{y} - b_1\bar{x_1} - b_2\bar{x_2}$$

$$b_1 = \frac{\sum_i x_{1i}y_i \sum_i x_{2i}^2 - \sum_i x_{1i}x_{2i} \sum_i x_{2i}y_i}{\sum_i x_{1i}^2 \sum_i x_{2i}^2 - (\sum_i x_{1i}x_{2i})^2}$$

$$b_2 = \frac{\sum x_{2i}y_i \sum x_{1i}^2 - \sum_i x_{1i}x_{2i} \sum_i x_{1i}y_i}{\sum_i x_{1i}^2 \sum_i x_{2i}^2 - \left(\sum_i x_{1i}x_{2i}\right)^2}$$

These will be the same as simple regression if the variables are uncorrelated.

Multiple Regression

When we had simple regression, we would intemperate b_1 as:

"A 1 unit change in x_1 is correlated with a b_1 unit change in y"

With multiple regression our interpretation changes to:

"Holding all else constant, A 1 unit change in x_1 is correlated with a b_1 unit change in y"

This is because b_1 and b_2 are the partial derivatives of y:

$$y_i = b_0 + b_1 x_1 + b_2 x_2 + e_i$$
$$\Rightarrow \frac{\partial y}{\partial x_1} = b_1$$
$$\Rightarrow \frac{\partial y}{\partial x_2} = b_2$$

Multiple Regression

There becomes an issue when x_1 and x_2 are perfectly multicolinear.

Multiple Regression

There becomes an issue when x_1 and x_2 are perfectly multicolinear. What is that?

Multiple Regression

There becomes an issue when x_1 and x_2 are perfectly multicolinear. What is that? This means that x_2 is a "linear combination" of x_1 or vice versa. A linear combination is a relationship that can be described with a line:

$$x_1 = mx_2 + b$$

You can think of adding someone's height in inches and someone's height in feet to a regression. In that case:

$$height_{inches} = 12 \cdot height_{feet}$$

Thus, $height_{inches}$ is a linear combination of $height_{feet}$ and the regression would fail as we would be dividing by zero when we find b_1 and b_2. If you want to see why, go to Lecture 3 slide 22.

Multiple Regression

Multicolinearity is a big problem when using binary or categorical variables. First, I should note that these are essentially the same thing.

- A binary variable is a variable that takes either a 0 or 1. This could be data that can be represented by a “yes” or “no”. In some cases “yes” can be signified by the 1 and “no” the 0, but that does not have to be the case.
- A factor or categorical variable can take on any integer value (1,2,3, etc), but this integer stands for some category. Think of race or car brand.

Multiple Regression

Typically in our data sets we leave factor variables the way they are because its easier to understand. However, when we actually use them in a regression we break them into separate binary variables. Instead of

$$\text{Type} : 1 = \text{Ford}, 2 = \text{Chevy...}$$

We transform the data so that it reads:

$$\text{Ford} : 1 = \text{Yes}, 0 = \text{No}$$

$$\text{Chevy} : 1 = \text{Yes}, 0 = \text{No}$$

This makes it so the regression understands that these are categories and not a variable that is an integer.

Multiple Regression

Categorical/binary variables create an issue with multicolinearity. Lets say you have a variable like hair color which can only take 4 values [Black, Brown, Blonde, Red]. If we know that someone does not have Black, Brown, or Blonde hair, we know they must have red hair. Further, if they have Black hair they cannot have Brown, Blonde, Red hair. Thus we can perfectly define each variable by the other variables:

$$Red = 1 - Black - Brown - Blonde$$

Thus, if you include all 4 types in the regression, you will run into multicolinearity. Thus, we have to leave one category out and use it as a "base". We then interpret each b for the categories as "compared to the base category". For example if we were running a regression of income on hair color and we left out Red from our regression we would interpret b_{brown} as "having brown hair is correlated with earning b_{brown} more/less than someone with red hair."

Multiple Regression

We can also do residual regression to find similar results to multiple regression. Lets say we want to know b_1 from the following regression:

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + e_i$$

We can find:

$$x_{1i} = c_0 + c_1 x_{2i} + v_{1i}$$

This is a regression of x_1 on x_2. The parameter c_1 will tell us something about the relationship between x_1 and x_2, but that's not what we're interested in. The error term, v_i, tells us all of the variation in x_{1i} that cannot be explained by x_{2i}. Thus, if we use that error term as the independant variable in a regression like this one:

$$y_i = d_0 + d_1 v_i + w_i$$

We will get that $d_1 = b_1$. This is because we have essentially "accounted for" the relationship between x_1 and x_2 by using, v_i, the variation in x_{1i} that cannot be explained by $x_2 i$.

R^2 Review - Derivation

This is for Canvas:

$$
\begin{aligned}
y_i &= \hat{y}_i + e_i \\
\Rightarrow Var(y_i) &= Var(\hat{y}_i + e_i) \\
\Rightarrow Var(y_i) &= Var(\hat{y}_i) + 2Cov(\hat{y}_i, e_i) + Var(e_i)
\end{aligned}
$$

What about that Cov?

$$
\begin{aligned}
Cov(\hat{y}_i, e_i) &= Cov(y_i - e_i, e_i) \\
&= Cov(b_0 + b_1 x_i + e_i - e_i, e_i) \\
&= Cov(b_0, e_i) + Cov(b_1 x_i, e_i) \qquad = 0 + b_1 Cov(x_i, e_i)
\end{aligned}
$$

If we assume $Cov(x_i, e_i) = 0$, then we are left with:

$$Var(y_i) = Var(\hat{y}_i) + Var(e_i)$$

R^2 Review - Derivation

Translate to sums:

$$\frac{1}{n}\sum_i (y_i - \bar{y})^2 = \frac{1}{n}\sum_i (\hat{y}_i - \bar{\hat{y}})^2 + \frac{1}{n}\sum_i (e_i - \bar{e})^2$$

We know $\bar{e} = 0$ and $\bar{\hat{y}} = \bar{y}$. We can also multiply everything by n. This gives us:

$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i e_i^2$$

$$\Rightarrow TSS = SSE + SSR$$

$$\Rightarrow 1 = \frac{SSE}{TSS} + \frac{SSR}{TSS}$$

$$\Rightarrow R^2 = 1 - \frac{SSR}{TSS}$$

back