Exploring Variation using Predicted Values and Regression Residuals

ECON 490

Taylor Mackay || Email: tmackay@fullerton.edu

Slides Overview

In these slides, we'll discuss:

- Using predicted values as a way of communicating regression results
- Using AI to help you interpret regression output
- Using residuals to explore relationships

Quick Review of Predicted Values

Suppose we run equation below using `lm()` in R and get estimates for β_0, β_1, and β_2:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

We can generate predicted values or conditional means of Y given our estimates

- Plug in values for Xs and R will give you a ***predicted value*** of Y denoted $\hat{Y}$
- This predicted value $\hat{Y}$ is our "best guess" for Y given what we know about X

Using Predicted Values to Make Regression Intuitive

Using predicted values can make regression feel more "concrete"

- They can help confirm you're interpreting individual coefficients
- Depending on project, predicted values can be an easy way to summarize output

Projects often ask questions like, "What is the gap in Y between groups A and B?"

- Predicted values are an easy way to summarizing this output
- "Someone in group A with X attributes will earn Y^A, while someone in group B..."

Two ways of calculating predicted values: 1) using R and 2) using ChatGPT

Using R to Generate Predicted Values

In coding activities, we saw two approaches:

1. Use coefficients to calculate predicted values (easy for simple equations)
2. Use `predict()` with data set of X values you specify (better for bigger equations)

```
model <- lm(Y ~ X.1 + X.2, data = working.data)

# Create data set of values of X at which to calculate predicted values of Y

prediction.data <- data.frame(X.1 = c(1, 2), X.2 = c(0, 1))

# Now calculate two predicted values of Y for 1) X.1 = 1, X.2 = 0 and 2) X.1 = 2, X.2 = 1

prediction.data <- mutate(prediction.data,
                          predicted.Y = predict(model, newdata = prediction.data))
```

Using AI to Generate Predicted Values

You can also give AI a screenshot of your regression output

- This is a great way of checking your interpretation in general
- ***KEY POINT:*** Interpret things yourself *BEFOREHAND* then check against AI

```
Call:
lm(formula = rate_property_crime ~ unemp_rate * west.coast +
    unemp_rate + west.coast, data = working.data)

Residuals:
     Min       1Q   Median       3Q      Max
-1845.47  -472.00   -23.49   501.91  1695.76

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            1752.36      65.52  26.745  < 2e-16 ***
unemp_rate              121.69      10.65  11.424  < 2e-16 ***
west.coast             1259.66     289.80   4.347 1.57e-05 ***
unemp_rate:west.coast  -128.08      40.07  -3.196  0.00145 **
---
Signif. codes:  0 '***' 0.001 '**' 0.0

Residual standard error: 635.8 on 746
Multiple R-squared:  0.1754,    Adjust
F-statistic: 52.91 on 3 and 746 DF,  p
```

	year	state_name	rate_property_crime	unemp_rate	west.coast
1	2008	Alabama	4084.5150	5.88	0
2	2009	Alabama	3780.3790	10.08	0
3	2010	Alabama	3527.9800	10.37	0
4	2011	Alabama	3605.3960	9.58	0
5	2012	Alabama	3505.4910	8.16	0
6	2013	Alabama	3347.8510	7.33	0
7	2014	Alabama	3179.4040	6.73	0
8	2015	Alabama	2982.8750	6.13	0
			2947.3850	5.91	0
			2949.1380	4.51	0
			2837.7880	3.94	0
			2622.7980	3.18	0

In your prompt, include ***both*** your regression output ***and*** relevant variables from your working data set (so AI knows how they're measured)

Prompting AI for Predicted Values

PROMPT: *I've run the attached regression in R using the attached data set. Can you calculate the predicted value of property crime rates (measured per 100,000 people) for west coast and non-west coast states at an unemployment rate of 4 percent? What happens if unemployment goes to 5 percent?*

Here's the fitted equation (per 100 000):

$$\hat{y} = 1752.36 \; + \; 121.69 \cdot U \; + \; 1259.66 \cdot W \; - \; 128.08 \cdot (U \times W)$$

where U =unemployment rate (%) and $W = 1$ if west-coast (0 otherwise).

1. At $U = 4\%$

- Non-west $(W = 0)$:
  $\hat{y} = 1752.36 + 121.69 \times 4 = 2239.1$
- West-coast $(W = 1)$:
  $\hat{y} = (1752.36 + 1259.66) + (121.69 - 128.08) \times 4 = 2986.5$

2. At $U = 5\%$

- Non-west:
  $\hat{y} = 1752.36 + 121.69 \times 5 = 2360.8$
- West-coast:
  $\hat{y} = 3012.02 + (121.69 - 128.08) \times 5 = 2980.1$

What changes?

- Non-west states: a 1 pp rise in unemployment (4→5) → +121.7 property crimes per 100 000.
- West-coast states: a 1 pp rise → $(121.69 - 128.08) = -6.39$, so a slight drop (~6.4 per 100 000).

Prompting AI for Predicted Values

PROMPT: *I'm not sure how to interpret the interaction term here? Why do we have this? What is it telling us?*

An interaction term like `unemp_rate:west.coast` basically lets the slope on unemployment differ between your two groups (west-coast vs. non-west). In your model

$$\hat{y} = \beta_0 + \beta_1 U + \beta_2 W + \beta_3 (U \times W),$$

- $\beta_1 = 121.69$ is the effect of unemployment on property crime in non-west states ($W = 0$).
- $\beta_3 = -128.08$ is the *extra* effect in west-coast states.

So the marginal effect of U is

$$\frac{\partial \hat{y}}{\partial U} = \begin{cases} 121.69, & W = 0 \text{ (non-west)} \\ 121.69 - 128.08 \approx -6.39, & W = 1 \text{ (west)} \end{cases}$$

In plain English: in non-west states each 1 pp rise in unemployment predicts ~122 more property crimes per 100 000, but on the west coast it predicts about a 6-point drop. We include the interaction to let those slopes be different and test if that difference is statistically real (here it is, $p \approx 0.001$).

Thinking About Practical or Economic Significance

How do we know if a coefficient is practically meaningful? Is this a “big” effect?

- One option – compare to average value of outcome
- Another option – compare to the size of other, more relatable coefficients

Suppose you ran the following regression for people working in a given industry:

$$Wage_i = \alpha_0 + \alpha_1 College\ Grad_i + \alpha_2 Years\ of\ Experience_i + u_i$$

One way to interpret $\alpha_1 \rightarrow$ how many additional years of experience would you need to match the returns to being a college graduate?

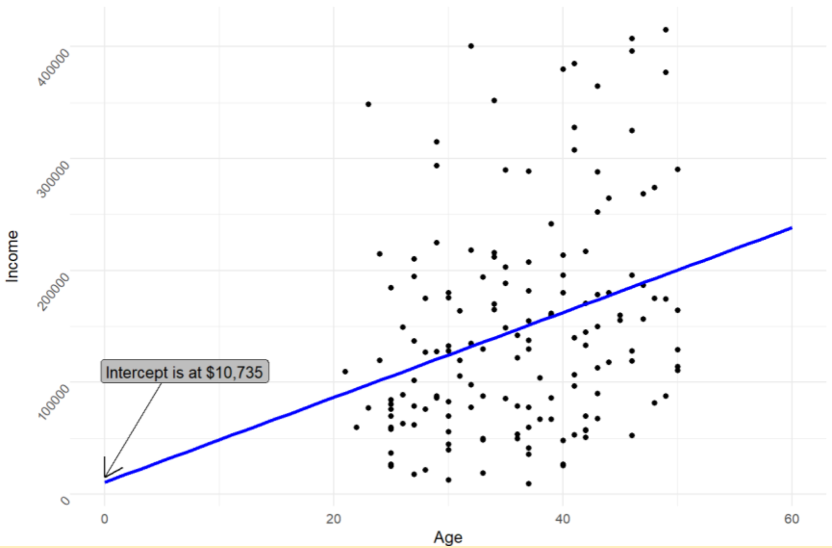
Predicted Values vs. the Intercept Term

My sense is sometimes students try to use the intercept as “home base”

- Use it as a starting point to get a handle on interpreting other βs
- General suggestion – use predicted values instead!

The intercept sets all variables equal to 0 and factor variables to omitted levels…

… even if it doesn’t make intuitive sense to have some of your variables equal 0!



Quick Review of Residuals

We've defined residuals previously using the following:

$$Residual = Actual\ Y - Predicted\ Y = Y - \hat{Y}$$

$\hat{Y}$ is our "best guess" about the value of Y given our X variables

- In other words, everything in Y that our regression ***can*** explain is reflected in $\hat{Y}$
- The residual is all the "left over" variation in Y → it's what we ***can't*** explain

Important Properties of Residuals

Residuals will always have an average value of 0

- This is a general property of OLS whenever you have an intercept term
- That’s why our discussion here focuses on outliers

Residuals are always conditional on a specific set of explanatory variables

- If you change your Xs, you’ll get different residuals
- Residuals may be more or less “directly” interpretable given context

Residuals Example: *Using NBA Data*

Positive residuals mean actual Y is ***higher*** than we'd expect based on Xs

- Conversely, ***negative*** residuals imply actual Y is ***lower*** than we'd expect
- These deviations can provide a way of talking about your regression output

Let's return to our NBA data set for a concrete example

- What's the relationship between points scored (PTS) and number of shots (FGA)?
- Run the simple OLS regression `lm(PTS ~ FGA, nba.data)`

Using residuals, we can see who scores the most and least based on their shot volume

Residuals Example: *Using NBA Data*

From our model output below, each additional shot (FGA) is associated with roughly 1.3 more points (PTS)

```
Call:
lm(formula = PTS ~ FGA, data = nba.data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8915 -0.9854 -0.1359  0.7702  5.6119

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.47216    0.32281  -1.463    0.145
FGA          1.34363    0.02495  53.845   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Notice part of our output is the distribution of residuals – let’s explore this!

Residuals Example: *Using NBA Data*

```
> # Save residuals as a new variable in our data set
> nba.data$residual.points <- resid(model.1)
> # What do our residuals look like? Let's use the summary function:
>
> summary(nba.data$residual.points)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-2.8916 -0.9854 -0.1359  0.0000  0.7702  5.6119
```

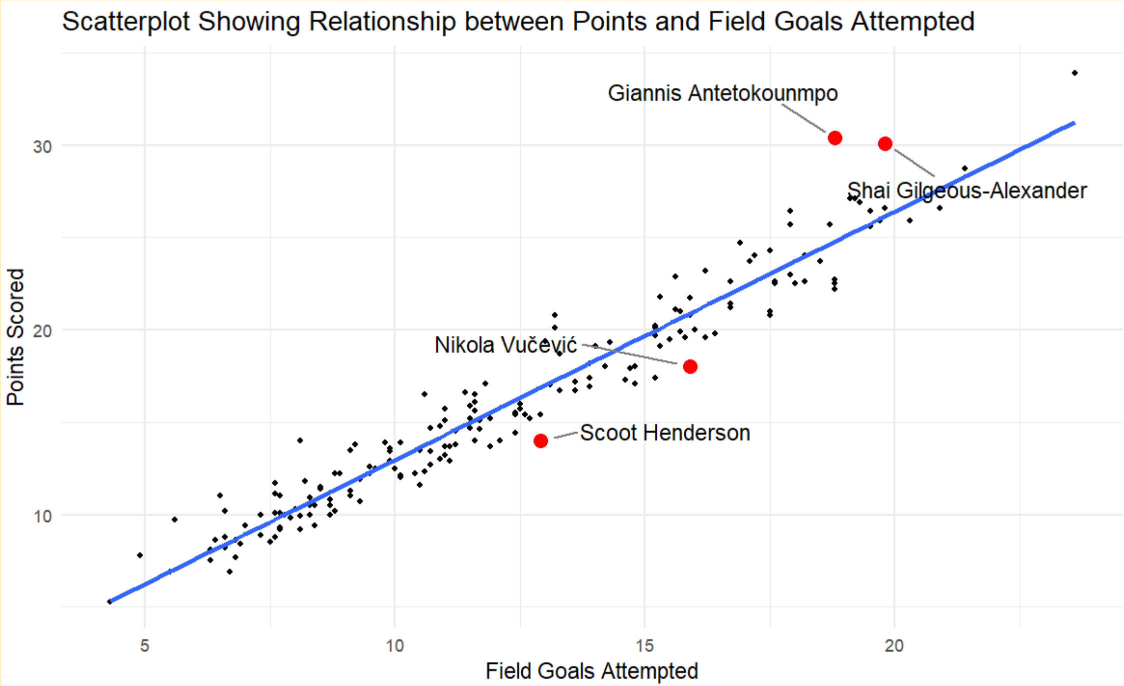
```
> # Let's see who had the largest positive values of residual points:
>
> nba.data %>%
+   arrange(desc(residual.points)) %>%
+   select(Player, Team, Pos, PTS, FGA, residual.points) %>%
+   head(10)
# A tibble: 10 × 6
   Player                  Team  Pos     PTS   FGA residual.points
   <chr>                   <chr> <chr> <dbl> <dbl>           <dbl>
 1 Giannis Antetokounmpo   MIL   PF     30.4  18.8            5.61
 2 Shai Gilgeous-Alexander OKC   PG     30.1  19.8            3.97
 3 Rudy Gobert             MIN   C      14     8.1            3.59
 4 Jimmy Butler            MIA   PF     20.8  13.2            3.54
 5 Kristaps Porziņģis      BOS   C      20.1  13.2            2.84
 6 Nikola Jokić            DEN   C      26.4  17.9            2.82
 7 Daniel Gafford          2TM   PF     11     6.5            2.74
 8 Jarrett Allen           CLE   C      16.5  10.6            2.73
 9 Luka Dončić             DAL   PG     33.9  23.6            2.66
10 Nick Richards           CHO   C       9.7   5.6            2.65
```

```
> # Let's see who had the most negative residual values:
>
> nba.data %>%
+   arrange(residual.points) %>%
+   select(Player, Team, Pos, PTS, FGA, residual.points) %>%
+   head(10)
# A tibble: 10 × 6
   Player          Team  Pos     PTS   FGA residual.points
   <chr>           <chr> <chr> <dbl> <dbl>           <dbl>
 1 Nikola Vučević  CHI   C      18    15.9           -2.89
 2 Scoot Henderson POR   PG     14    12.9           -2.86
 3 Kyle Kuzma      WAS   PF     22.2  18.8           -2.59
 4 Jordan Poole    WAS   SG     17.4  15.2           -2.55
 5 Jordan Clarkson UTA   SG     17.1  14.8           -2.31
 6 Dejounte Murray ATL   SG     22.5  18.8           -2.29
 7 Tyler Herro     MIA   SG     20.8  17.5           -2.24
 8 Cade Cunningham DET   PG     22.7  18.8           -2.09
 9 Miles Bridges   CHO   SF     21    17.5           -2.04
10 Jeremy Sochan   SAS   PF     11.6  10.5           -2.04
```

Residuals Example: *Using NBA Data*

This graph shows PTS ~ FGA relationship with outliers highlighted

Residuals here are given by vertical difference between fitted line and ind. points



General Suggestions for Using Residuals

Opportunity to apply “qualitative” knowledge

- What characteristics do residual outliers share?
- We might not be able to measure this and include it in a regression…
- But we can talk about patterns in intuitive terms

General approach for exploring residuals:

- Run regression then store residuals as a new variable
- Sort data set by residuals and explore biggest and smallest values
- Apply background knowledge to identify patterns – who stands out?

You can describe the results from this process using a table, scatterplot, or verbally