Prediction and Explanation in Linear Regression

Nicole McCarthy

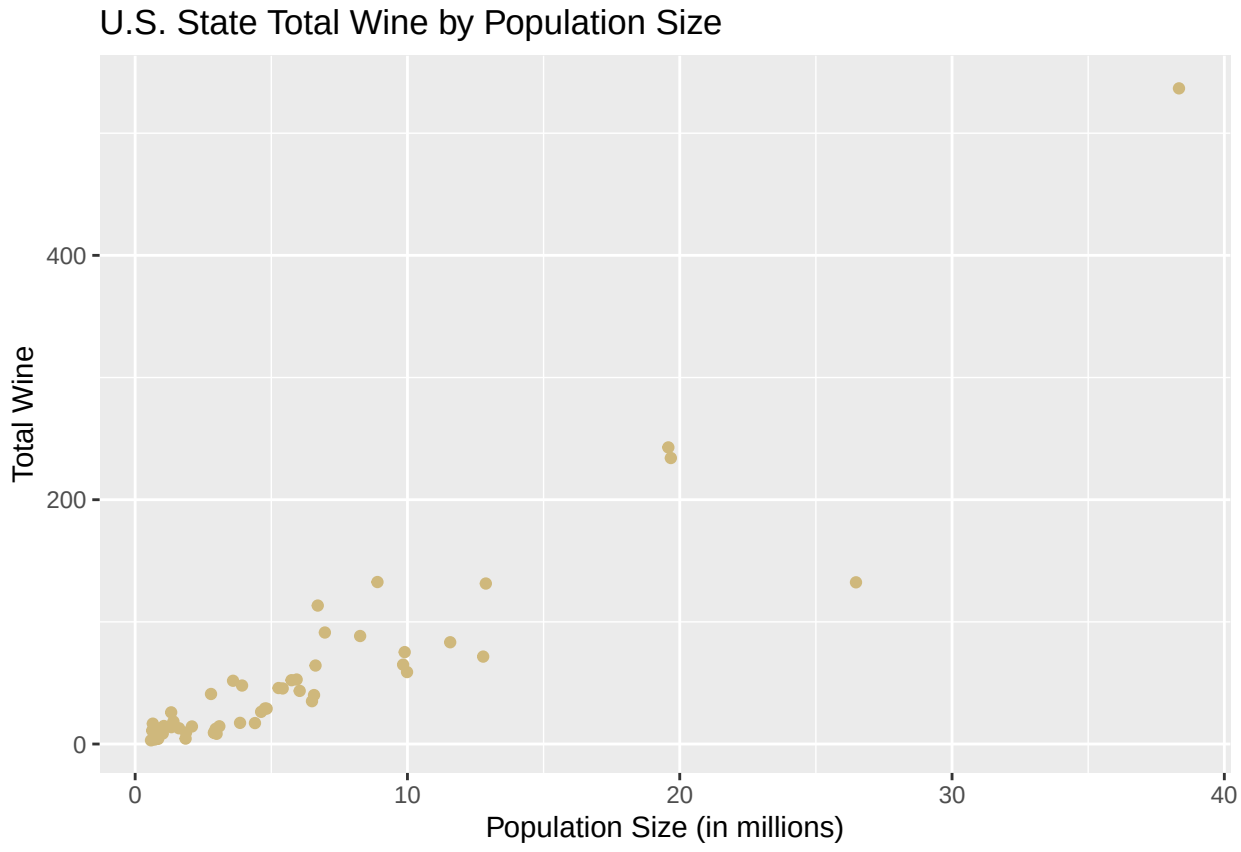
October 2, 2025

Problem 1: Interpreting Intervals

For this problem, we're going to practice creating and interpreting Confidence (Mean) Intervals and Prediction Intervals. To do so, we're going to use data in U.S. State Wine Consumption (millions of liters) and Population (millions).

1. (a) Initial Inspections

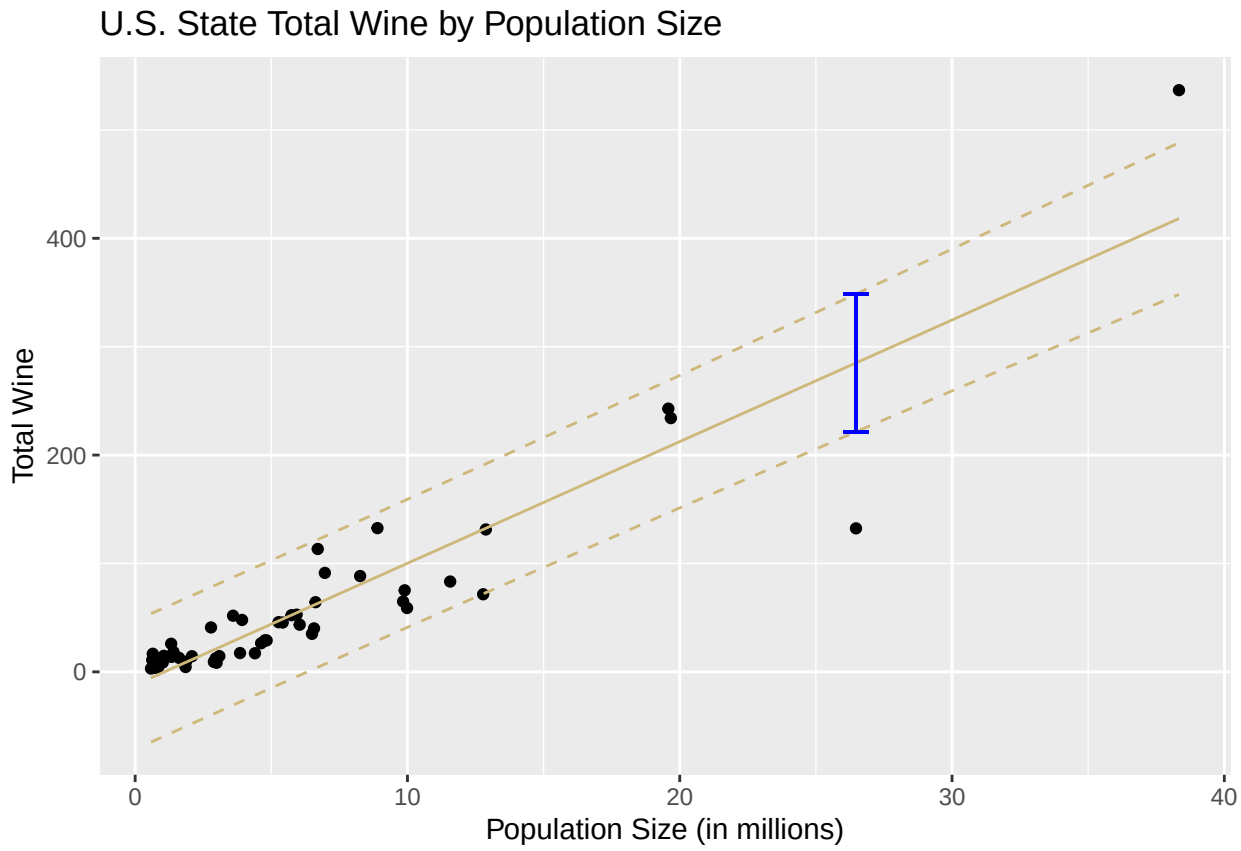
Load in the data and create a scatter plot with `population` on the x-axis and `totWine` on the y-axis. For fun, set the color of the point to be `#CFB87C`.



1. (b) Confidence Intervals

Fit a linear regression with `totWine` as the response and `pop` as the predictor. Add the regression line to your scatter plot. For fun, set its color to gold with `col=#CFB87C`. Add the 90% Confidence Interval for the regression line to the plot.

Then choose a single point-value population and display the upper and lower values for the Confidence Interval at that point. In words, explain what this interval means for that data point.

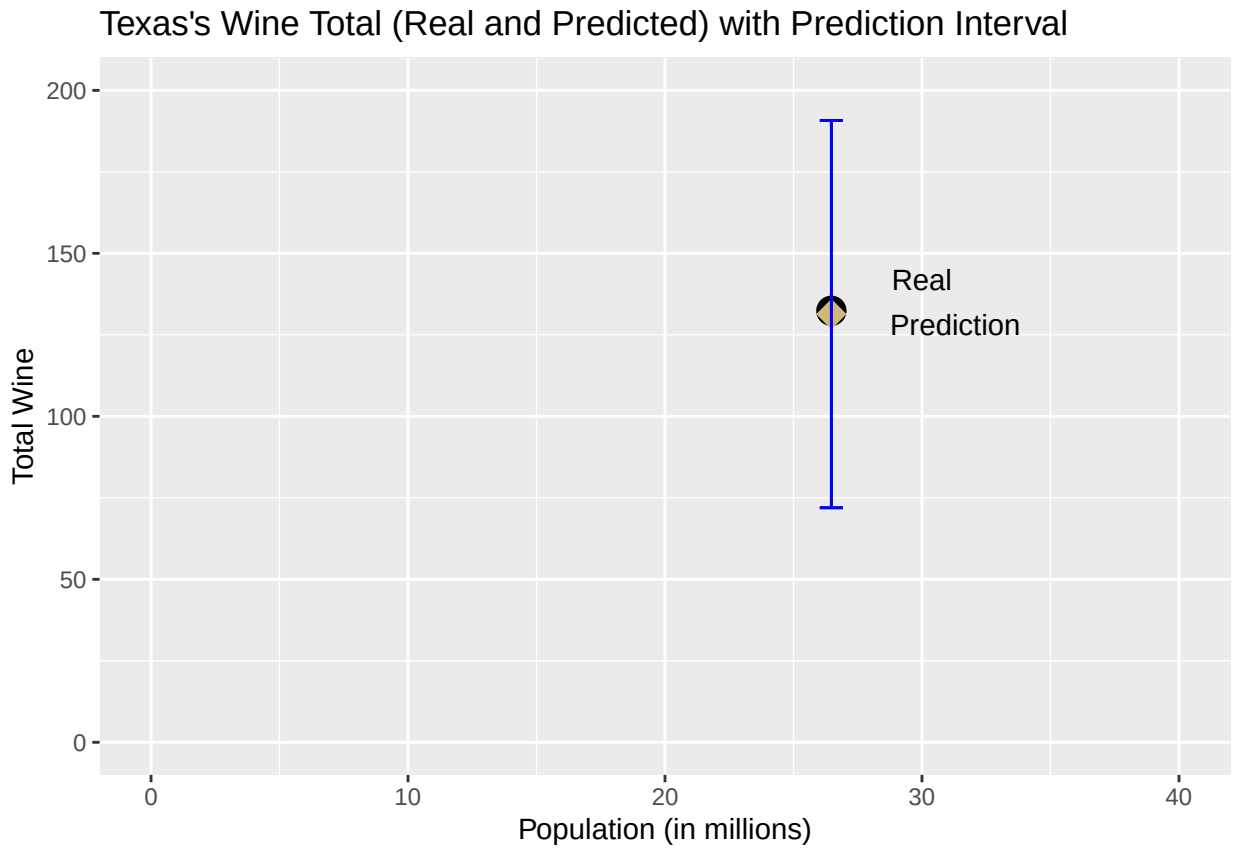


The 90% prediction interval (seen in blue) for the states with the same population size as Texas (about 26 million people) shows the range of total wine that includes 90% of the possible values we predict for a state of that size according to this model. In other words, if 100 states with that population size, we can expect 90 of them to have a total wine value within that range (around 225 - 350).

Note: The true data point lies far below the prediction interval. This shows that the model is either not well fitted for populations of that size, or we happened to sample a datum that is within the 10% chance of falling outside of that interval.

1. (c) Prediction Intervals

Using the same `pop` point-value as in **1.b**, plot the prediction interval end points. In words, explain what this interval means for that data point.



If the wine total was sampled for multiple states with the same population size as Texas, 90% of the samples would fall within the prediction interval.

Note: The predicted value and interval fall within the same space as the real data point here (unlike in the previous graph) because we are using the interval specific to this point, rather than a projected interval for the entire model.

1. (d) Some "Consequences" of Linear Regression

As you've probably gathered by now, there is a lot of math that goes into fitting linear models. It's important that you're exposed to these underlying systems and build an intuition for how certain processes work. However, some of the math can be a bit too... tedious for us to make you go through on your own. Below are a list of "consequences" of linear regression, things that are mathematically true because of the assumptions and formulations of the linear model (let $\widehat{\varepsilon}_i$ be the residuals of the regression model):

1. $\sum \widehat{\varepsilon}_i = 0$: The sum of residuals is 0.
2. $\sum \widehat{\varepsilon}_i^2$ is as small as it can be.
3. $\sum x_i \widehat{\varepsilon}_i = 0$
4. $\sum \hat{y}_i \widehat{\varepsilon}_i = 0$: The Residuals are orthogonal to the fitted values.
5. The Regression Line always goes through $(\bar{x}, \bar{y})$.

Check that your regression model confirms the "consequences" $1, 3, 4$ and 5. For consequence 2, give a logical reason on why this formulation makes sense.

Note: even if your data agrees with these claims, that does not prove them as fact. For best practice, try to prove these facts yourself!

Answers

Claim 1: The sum of residuals is 0. Let's start with our basic matrix formula for a linear equation:

$$\widehat{Y} = X\widehat{\beta} + E$$

Now, let's solve for E, the matrix of residuals (ε).

$$E = \widehat{Y} - X\widehat{\beta}$$

If $\sum E = 0$, then $\widehat{Y} = X\widehat{\beta}$. This is the goal of our linear model! Therefore, the values of $\widehat{\beta}$ are specifically chosen such that $\sum E = 0$.

Claim 2: $\sum \widehat{\varepsilon_i^2}$ is as small as it can be.

Assuming that **Claim 1** is true, then squaring the residuals (ε), will not change the sum very far from 0. This is because the residuals themselves are minimized (which is why their sum is 0). When the residuals are minimized, the sum of their squares is also minimized.

Claim 3: $\sum x_i\widehat{\varepsilon_i} = 0$

Assuming that **Claim 1** is true, then multiplying x_i by the residuals (ε), will not change the sum from 0.

$$\sum x_i\widehat{\varepsilon_i} = 0$$

$$\sum x_i \sum \widehat{\varepsilon_i} = 0$$

$$\sum x_i * 0 = 0$$

$$0 = 0$$

Claim 4: $\sum \hat{y}_i\widehat{\varepsilon_i} = 0$

Assuming that **Claim 1** is true, multiplying $\hat{y}_i$ by the residuals (ε), will not change the sum from 0.

$$\sum \hat{y}_i\widehat{\varepsilon_i} = 0$$

$$\sum \hat{y}_i \sum \widehat{\varepsilon_i} = 0$$

$$\sum \hat{y}_i * 0 = 0$$

$$0 = 0$$

Claim 5: The Regression Line always goes through $(\bar{x}, \bar{y})$.

Let us start again with the basic matrix equation for a linear model.

$$\widehat{Y} = X\widehat{\beta} + E$$

Assuming **Claim 1** is true, we can ignore E, which is essentially equal to 0.

$$\widehat{Y} = X\widehat{\beta}$$

Let's move this from matrix form to a linear equation.

$$\hat{y} = x\widehat{\beta_1} + \widehat{\beta_0}$$

Now, let's transform $\hat{y}$ and x into $\bar{y}$ and $\bar{x}$. First, sum all y_i and x_i.

$$\sum_{i=0}^{n} y_i = \sum_{i=0}^{n} [x\widehat{\beta_1} + \widehat{\beta_0}]$$

$$\sum_{i=0}^{n} y_i = \sum_{i=0}^{n} x\widehat{\beta_1} + \sum_{i=0}^{n} \widehat{\beta_0}$$

$$\sum_{i=0}^{n} y_i = \widehat{\beta_1} \sum_{i=0}^{n} x + n\widehat{\beta_0}$$

Second, divide by n.

$$\frac{1}{n}\sum_{i=0}^{n} y_i = \frac{1}{n}\widehat{\beta_1} \sum_{i=0}^{n} x + \widehat{\beta_0}$$

$$\bar{y} = \widehat{\beta_1}\bar{x} + \widehat{\beta_0}$$

This equation tells us that the predicted value $\hat{y}$ of $\bar{x}$ is $\bar{y}$. That is,

$$\hat{y}(\bar{x}) = \widehat{\beta_1}\bar{x} + \widehat{\beta_0}$$

$$\hat{y}(\bar{x}) = \bar{y}$$

Because the regression line always passes through $\bar{x}$, and it's solution is $\bar{y}$, **Claim 5** is correct.

Problem 2: Explanation

Did our wine drinking data come from an experiment or an observational study? Do you think we can infer causation between population and the amount of wine drank from these data?

Answer

Our wine drinking data came from an **observational study**. Therefore, we **cannot** infer causation between the population size and the amount of wine drank from these data.

Problem 3: Even More Intervals!

We're almost done! There is just a few more details about Confidence Intervals and Prediction Intervals which we want to go over. How does changing the data affect the confidence interval? That's a hard question to answer with a single data set, so let's simulate a bunch of different data sets and see what they intervals they produce.

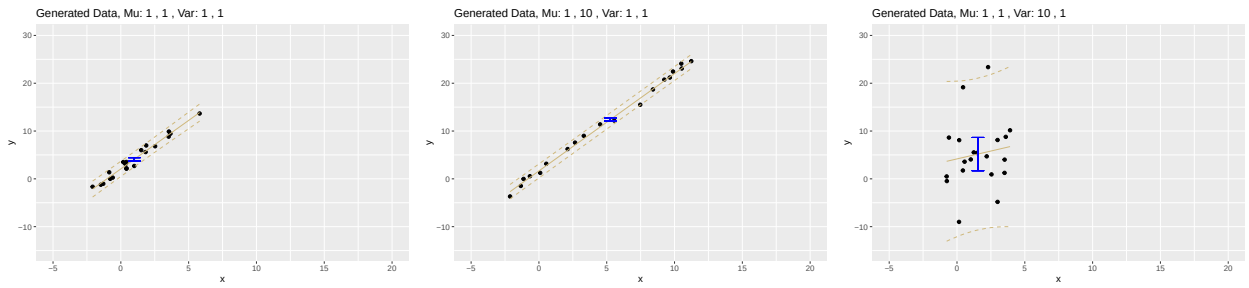
3. (a) Visualize the data

The code cell below generates 20 data points from two different normal distributions. Finish the code by fitting a linear model to the data and plotting the results with ggplot, with Confidence Intervals for the mean and Prediction Intervals included.

Experiment with different means and variances. Does changing these values affect the CI or PI?

```
gen_data <- function(mu1, mu2, var1, var2){
    # Function to generate 20 data points from 2 different normal distributions.
    x.1 = rnorm(10, mu1, 2)
    x.2 = rnorm(10, mu2, 2)
    y.1 = 2 + 2*x.1 + rnorm(10, 0, var1)
    y.2 = 2 + 2*x.2 + rnorm(10, 0, var2)

    df = data.frame(x=c(x.1, x.2), y=c(y.1, y.2))
    return(df)
}
```

```r
set.seed(0)

inputs <- data.frame(try1 = c(1, 1, 1, 1),
                     try2 = c(1, 10, 1, 1),
                     try3 = c(1, 1, 10, 1))

for (row in inputs){
  # get values out of row
  m1 = row[1]
  m2 = row[2]
  v1 = row[3]
  v2 = row[4]

  # generate data
  data <- gen_data(m1, m2, v1, v2)

  # fit the model
  mod <- lm(y~x, data)
  predictions <- predict(mod, data, interval = 'prediction', level = 0.95)
  data <- merge(data, predictions, by.x = 0, by.y = 0) # merge on indices

  # get mean data (for confidence interval)
  xbar <- data.frame(x = mean(data$x), y = mean(data$y))
  ybar.preds <- predict(mod, xbar, interval = 'confidence', level = 0.95)

  # plot generated data with model lines and confidence interval for xbar
  plt <- ggplot(data, aes(x = x, y = y)) +
    geom_point() +
    labs(title = paste("Generated Data, Mu:", m1, ",", m2, ", Var:", v1, ",", v2)) +
    geom_line(aes(y = fit), color = '#CFB87C') +
    geom_line(aes(y = lwr), linetype = "dashed", color = '#CFB87C') +
    geom_line(aes(y = upr), linetype = "dashed", color = '#CFB87C') +
    geom_errorbar(aes(x = xbar[1, 1], ymin = ybar.preds[1, 2], ymax = ybar.preds[1, 3]), color = 'blue'
    ylim(-15, 30) +
    xlim(-5, 20)

  # show plot
  print(plt)
}
```



Increasing μ elongates the regression line and prediction interval lines along the x-axis, and moves the confidence interval of $\bar{x}$ up. Increasing σ^2 widens the prediction intervals and confidence interval along the y-axis.Both of these changes in modeling/statistical outputs follow the changes seen in the data itself.

3. (b) The Smallest Interval

Recall that the Confidence (Mean) Interval, when the predictor value is x_k, is defined as:

$$\hat{y}_h \pm t_{\alpha/2,n-2}\sqrt{MSE \times \Big(\frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum(x_i - \bar{x})}\Big)}$$

where $\hat{y}_h$ is the fitted response for predictor value x_h, $t_{\alpha/2,n-2}$ is the t-value with $n-2$ degrees of freedom and $MSE \times \Big(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum(x_i - \bar{x})}\Big)$ is the standard error of the fit.

From the above equation, what value of x_k would result in the CI with the shortest width? Does this match up with the simulated data? Can you give an intuitive reason for why this occurs?

Answer

To get the shortest width confidence interval possible, we need to minimize the standard error, $MSE \times \Big(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum(x_i - \bar{x})}\Big)$. Since we can only manipulate the portion that is affected by x_k, we'll focus our attention on this portion:

$$\frac{(x_k - \bar{x})^2}{\sum(x_i - \bar{x})}$$

To minimize this equation, x_k will need to be as close to $\bar{x}$ as possible so that the numerator can stay close to 0. Therefore, the confidence interval width is minimized when the predictor value, x_k, is equal to $\bar{x}$.

$$x_k = \bar{x}$$

3. (c) Interviewing the Intervals

Recall that the Prediction Interval, when the predictor value is x_k, is defined as:

$$\hat{y}_h \pm t_{\alpha/2,n-2}\sqrt{MSE\Big(1 + \frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum(x_i - \bar{x})}\Big)}$$

Does the "width" of the Prediction Interval change at different population values? Explain why or why not.

Answer

No, the width of the prediction interval does not change at different population values. This is because the prediction interval is only affected by the variance, not the mean, of the predictors.

Problem 4: Causality

Please answer the following three questions. Each answer should be clearly labeled, and a few sentences to a paragraph long.

1. In your own words, describe the fundamental problem of causal inference. How is this problem related to the counter-factual definition of causality?
2. Describe the use of "close substitutes" as a solution to the fundamental problem of causal inference. How does this solve the problem?
3. What is the difference between a *deterministic* theory of causality and a *probabilistic* theory of causality?

Answers

1. The fundamental problem of causal inference is that both the presence and absence of the cause needs to be observed for the effects (and therefore causal relationship) to be clear. This is related to the counter-factual definition of causality, which states that a causal relationship can only exist if the absence of the cause means there is an absence of the effect. This is a problem because this type of data can only be collected through empirical experimentation with control variables. Otherwise, it is impossible to know what would have happened if something did not happen.
2. The use of close substitutes is a potential fix to the counter-factual definition of causality. In settings where experiments are not possible, we can find data on what happened to subjects that we consider similar to the one in question and see whether the causal variable was in place and if the effect followed. This solves the problem because we don't need the exact subject to undergo multiple, impossible trials. Instead, we "substitute" these trials with other subjects.
3. A *deterministic* theory of causality states that the presence of the cause means the effect will also be present. A *probabilistic* theory of causality states that the presence of the cause means the effect is more likely to be present. In other words, the *deterministic* theory is a certainty, whereas the *probabilistic* theory is simply a higher chance.

Problem 5: Causal inference and ethics

How we think about causality, and the statistical models that we use to learn about causal relationships, have ethical implications. The goal of this problem is to invite you to think through some of those issues and implications.

Statisticians, data scientists, researchers, etc., are not in agreement on the best ways to study and analyze important social problems, such as racial discrimination in the criminal justice system. Lily Hu, a PhD candidate in applied math and philosophy at Harvard, wrote that disagreements about how to best study these problems "well illustrate how the nuts and bolts of causal inference... about the quantitative ventures to compute 'effects of race'... feature a slurry of theoretical, empirical, and normative reasoning that is often displaced into debates about purely technical matters in methodology."

Here are some resources that enter into or comment on this debate:

1. Statistical controversy on estimating racial bias in the criminal justice system
2. Can Racial Bias in Policing Be Credibly Estimated Using Data Contaminated by Post-Treatment Selection?
3. A Causal Framework for Observational Studies of Discrimination

Please read Lily Hu's blog post and Andrew Gelman's blog post "Statistical controversy on estimating racial bias in the criminal justice system" (and feel free to continue on with the other two papers!) to familiarize yourself with some of the issues in this debate. Then, write a short essay (300-500 words) summarizing this debate. Some important items to consider:

1. How does the "fundamental problem of causal inference" play out in these discussions?
2. What are some "possible distortionary effect[s]" of using arrest data from administrative police records to measure causal effects of race"?
3. What role do assumptions (both statistical and otherwise) play in this debate? To what extent are assumptions made by different researchers falsifiable?

Answer

Lily Hu argues in her blog post that the empirical methodologies used in research, and especially statistics, are not in fact free of human biases or assumptions. They are based on them. Because of this, statistical methodologies can be set up and used to help reduce inequality in areas of research relating to human interaction and lives by taking a progressive, liberal approach to the assumptions made in the statistical testing.

Andrew Gelman's post studies two articles that discuss statistical methodologies in researching racial discrimination in policing. The first, Knox et al., claims that a causal effect cannot be found from the data available because there is bias in the collection of the data. In other words, the traffic stops and arrests made by police do not record who they observed but did not interact with. The second article, Gaebler et al., argues that you can determine causal effects at different points in the policing process if you are careful with your assumptions. Gelman argues that both papers actually are just pointing out the same issue that statisticians need to be aware of concerning causal relationships and their identification.

1. The fundamental problem of causal inference is the core of the issue in these debates. The main problem that sparks the argument is that we cannot know what would have happened in any given case if the person arrested was a different race. Gaebler argues (and Gelman agrees) that this issue can be circumvented with the use of proper assumptions and close substitutes. Hu essentially claims the same. All warn about the pitfalls that researchers are bound to come across when taking these routes.

2. Some possible distortionary effects of using arrest data from administrative police records to measure causal effects of race include what Gaebler (and Hu) mentioned on not having data on who police did not arrest. Essentially every decision that led to a police officer's location and actions needs to be considered. The problem is that only some of this data is recorded.

3. Assumptions are at the core of these research problems. Statistical assumptions are made as soon as data begins to be analyzed, whether the researcher is aware of them or not. Political and social assumptions are made in forming the research question itself. Therefore, some assumptions, particularly statistical ones, are falsifiable, but many others are not. This is the nature of conducting socioligical, political, or anthropological research.