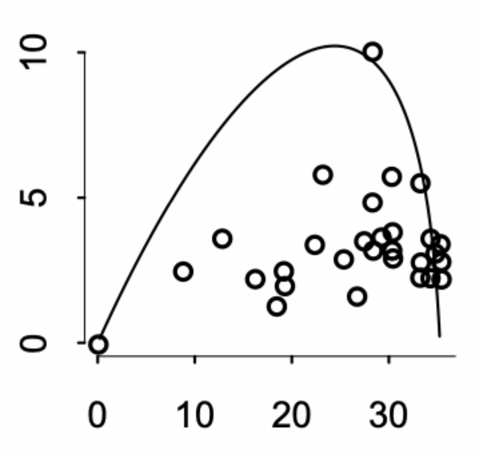
SCI 2025: Homework 5

Setup

The first two homework problems this week are from Chapter 7 of the textbook.

Question 1: 7H1

In 2007, *The Wall Street Journal* published an editorial ("We're Number One, Alas") with a graph of corporate tax rates in 29 countries plotted against tax revenue. A badly fit curve was drawn in (reconstructed below), seemingly by hand, to make the argument that the relationship between tax rate and tax revenue increases and then declines, such that higher tax rates can actually produce less tax revenue.



I want you to actually fit a curve to these data, found in `data(Laffer)`. Consider models that use tax rate to predict tax revenue. Compare, using WAIC or PSIS, a straight-line model to any curved models you like. What do you conclude about the relationship between tax rate and tax revenue?

Question 2: 7H2

In the Laffer data, there is one country with a high tax revenue that is an outlier. Use PSIS and WAIC to measure the importance of this outlier in the models you fit in the previous problem. Then use robust regression with a Student's t distribution to revisit the curve fitting problem. How much does a curved relationship depend upon the outlier point?

Question 3

In machine learning, it is common to use cross-validation for model comparison and tuning of certain parameters. The simplest approach is the "train-test" split, where the data is split into a training set and a test set. The model is fit on the training set, and then the predictions are compared to the true values on the test set. It is typical to use around 70-80% of the data for the training set and the rest for the test set.

Here's how you can do a train-test split in R on the Laffer data:

```r
library(rethinking)
data(Laffer)

set.seed(123)
n <- nrow(Laffer)
train_idx <- sample(1:n, size = round(n * 0.7)) # random sample of 70% of the data
train_data <- Laffer[train_idx, ]
test_data <- Laffer[-train_idx, ]
```

Now, what I would like you to do is fit a model of your choice (informed by the results of the previous questions) to the training data only (train_data). First, make predictions for the *training data* and plot those predictions, as well as the true values as points. Then, make predictions for the *test data* and plot those predictions, as well as the true values as points. Your model predictions should be on the y-axis and the tax rate should be on the x-axis. Be sure to visualize uncertainty in your predictions.

Question 4

Now, repeat the procedure in Question 3, exactly, but *change the random seed* to some new number. I encourage you to do this multiple times. How much variability across seeds (different random splits) is there in: (a) the slope/curve relating tax rate to tax revenue? (b) the discrepancy between the training and test set predictions?

Question 5

Based on what you have learned so far in this course, how do you imagine that model comparison via information criteria and/or cross-validation can support causal inference, and answering scientific questions? Where do you think it could go wrong?