STATISTICS 452/652: Statistical Learning and Prediction

August 19, 2020

Regression Review Part 4: How the Universe Works

(Reading: ISLR Sections 3.1–3.4)

1 What's really going on in regression

- At the heart of any regression problem is a relationship that we are trying to understand
  - While we use a model to represent this relationship, reality is (almost?) always more complex
  - So what is *really* going on?

1.1 The Universe

- We have a response variable, Y
- There are infinitely many other measurements (variables) that we could measure in addition to Y
- Let $\mathbb{X}$ represent the set of *all other variables in the universe*, except for Y.
  - We don't know and possibly cant even imagine most of these, but they exist.
- We have chosen to measure some of these variables in a sample, $X_1, \ldots, X_p$.
  - There's a lot more out there that we haven't measured.
- **In reality**, there exists some function $g(\mathbb{X})$ that "best" predicts Y among all possible functions in the universe
  - That is, we can say that

$$Y = g(\mathbb{X}) + \delta$$

    * δ represents the part of Y that cannot possibly be explained by all the other knowledge in the universe
      · Often called IRREDUCIBLE ERROR
  - The function $g()$ can take on *any* shape, not necessarily one we have formulas for
  - I may refer to $g(\mathbb{X})$ as the "universal function" or "universal predictor" or things like that, and to $\mathbb{X}$ as "the universe".

- The regression goal is to try to guess the unknown function g of possibly many unknown variables, $\mathbb{X}$, using only the sample we have measured.
  - Sounds hopeless.

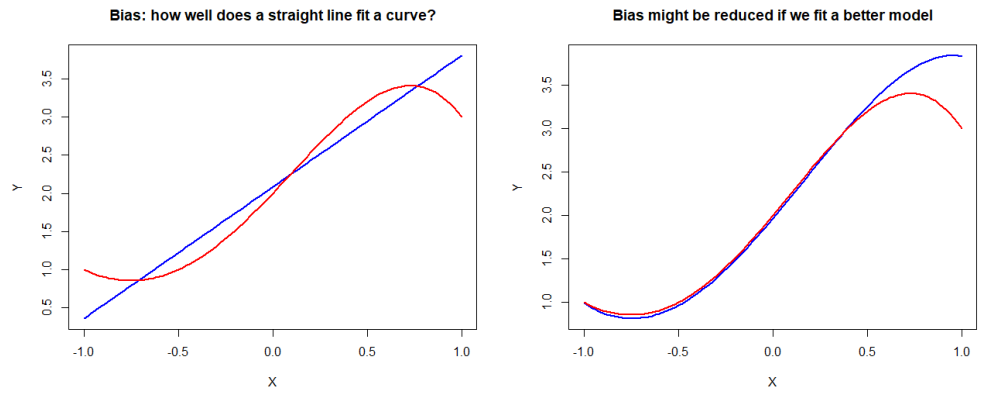
1.2 Modeling

- We cannot possibly guess $g(\mathbb{X})$, so we accept from the start that *everything we do is an approximation.*
- We will try to find a good one with the limited tools we have
  - Historically, tools have focused on flat surfaces (hyperplanes)
  - We will see how to expand the toolkit!
- We *propose a model* for the mean of Y using the available variables, $f(X_1, \ldots, X_p)$, abbreviated for now as $f(X)$
  - Hard to do when you don't know $g(\mathbb{X})$
  - Ideally use something flexible
  - Historically, $f(X)$ was a linear function
    * *NOT* flexible!
    * We will explore more general approximations
- We then write $Y = f(X) + \epsilon$ where we often assume that $\epsilon \sim N(0, \sigma^2)$,
- Let's look at that "error term", ϵ, a little closer
  - If (by some miracle) $f(X) = g(\mathbb{X})$, then $\epsilon = \delta$
    * That practically never happens
  - Otherwise we have two different expressions for Y, reality and model:

$$Y = g(\mathbb{X}) + \delta = f(X) + \epsilon$$

    * δ is the deviation of Y from its *true* mean
    * ϵ is the deviation of Y from its *modeled (measurable)* mean
    * $\epsilon = [g(\mathbb{X}) - f(X)] + \delta$
    * "Random" error that we propose is really a combination of two things:

Figure 1: Bias from fitting a different models $f(X)$, blue, to a curved true relationship $g(\mathbb{X})$, red.



- · The error in our model specification (NOT RANDOM!)
- · The true unexplainable variability inherent in Y (random) .

* So ϵ is only random to the extent that we don't really know what $g(\mathbb{X})$ is, so we can't predict the error in our model specification

1.3 Bias-Variance Tradeoff

- Let's look at $\epsilon = [g(\mathbb{X}) - f(X)] + \delta$ a little closer:
  - The difference $[g(\mathbb{X}) - f(X)]$ is called the BIAS of the model. (Get used to this term!!!)
    * Imagine modeling a curvy relationship $g(\mathbb{X})$, where by luck only one variable in the universe is important, with a straight line using the same variable, $f(X) = \beta_0 + \beta_1 X$ (e.g., see left panel of Figure 1)
    * The bias will be different depending on here we make the comparison
    * The bias will also depend on what model we fit.
  - The δ is "irreducible", meaning that *nothing in the world* can explain it
    * it is the "true" source of randomness in our problem.
- So the thing we can partly control is the bias in our chosen model
  - If we choose a better, more flexible model, we may be able to reduce bias (e.g., see right panel of Figure 1)
  - The more flexible a model is, the more different shapes it may be able to adapt to
- Is there any reason not to always choose flexible models?

Example: Exploring bias (RegressionBiasVarianceHighBias.R) Let's explore that happens when we allow models to be more flexible. In this example:

1. I am creating my universal truth as

$$g(\mathbb{X}) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \delta,\ \delta \sim N(0,1)$$

   with parameter values chosen to create the curve seen red in all four panels of Figure 2.

2. I then use this structure to generate $n = 10$ observations for random values of X (top left).

3. I fit a simple linear model $f(X) = \beta_0 + \beta_1 X$ to these data (blue) and show that it doesn't really fit very well (top right) .

   (a) I should probably not use the same parameter symbols in both equations. In the true structure, they have different meaning from in the model.

4. Then I repeat the process a total of 100 times, generating a new data set from the true structure and fitting the data with a linear model. These 100 estimated lines are in light blue (bottom left).

5. Finally, I take sample average of these 100 lines and compare it to the true structure (bottom right). This shows the average difference between the fitted model and the true structure, which is exactly the model bias.

Comments:

- We can see that the model does not generally do a very good job of estimating the curve. Although it does cross the curve and thus has 0 bias at those points, in other places it is far from the curve and has high bias. The "average bias" of the model is not great.
- We can see the effects of sampling variability on the straight lines. Gathering a different sample does not fix bias. It just allows the wrong model to be estimated differently
  - This makes it clear that bias is a *structural* problem with a model, and has nothing to do with the data that are drawn.
  - In other words, *bias is our fault,* but to be fair, we often don't know any better or can't do any better

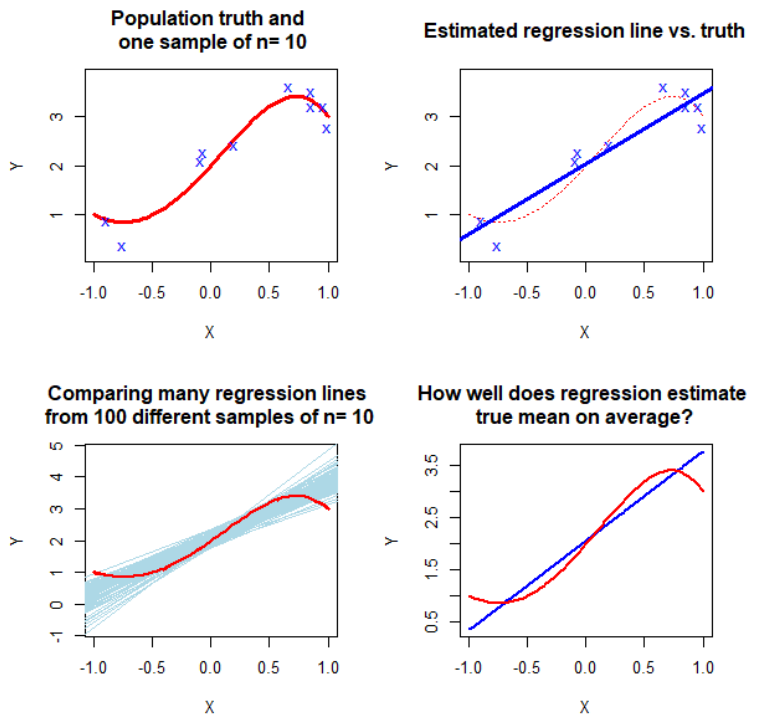
Now let's take a look at what happens if we fit a more flexible model.

- In Figure 3 we repeat the same steps as above, except we fit the model

$$f(X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4$$

- In theory, this model should be able to fit the true structure perfectly

Figure 2: Population true structure is red. Samples and estimated regressions are blue. See full description in text



– Unfortunately, sampling variability has something to say about this.

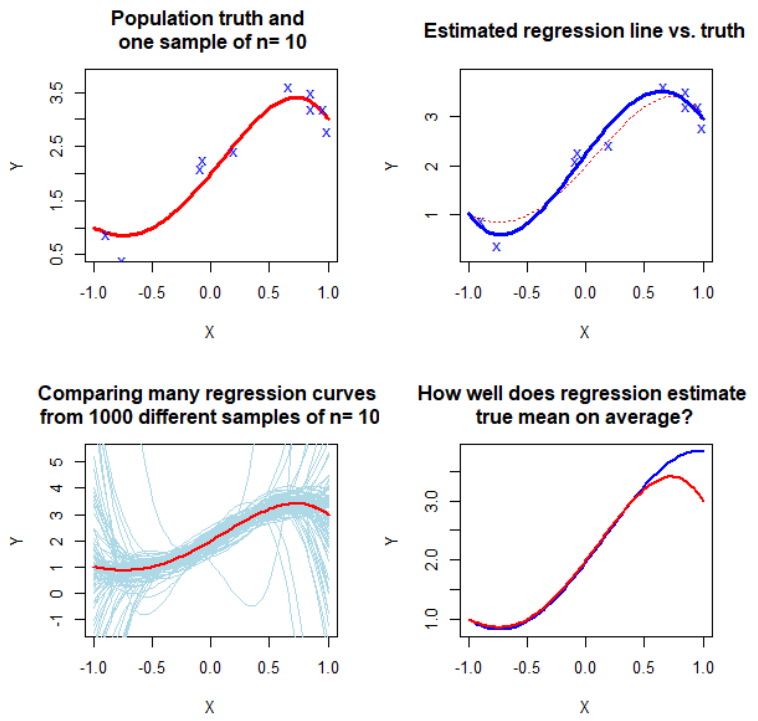
1. The samples are exactly the same ones that we used for the liner model (top left)
2. Any particular model fit to data has potential to be a very good fit (top right)
3. *However, with an odd sample here and there, we can get some bizarre predictions!* (bottom left)
4. Despite this, the *average* line is not a bad estimate for the true structure (bottom right)
   (a) The apparent bias in the top portion of that panel would go away if I averaged together a larger number of curves, each fit to new samples of $n = 10$.

In conclusion, in a problem like this, *we might actually be better off using a simpler, biased model than using a more accurate, but less stable one.*

So what happened here? Why did fitting the "right model" result in such unreliable predictions?

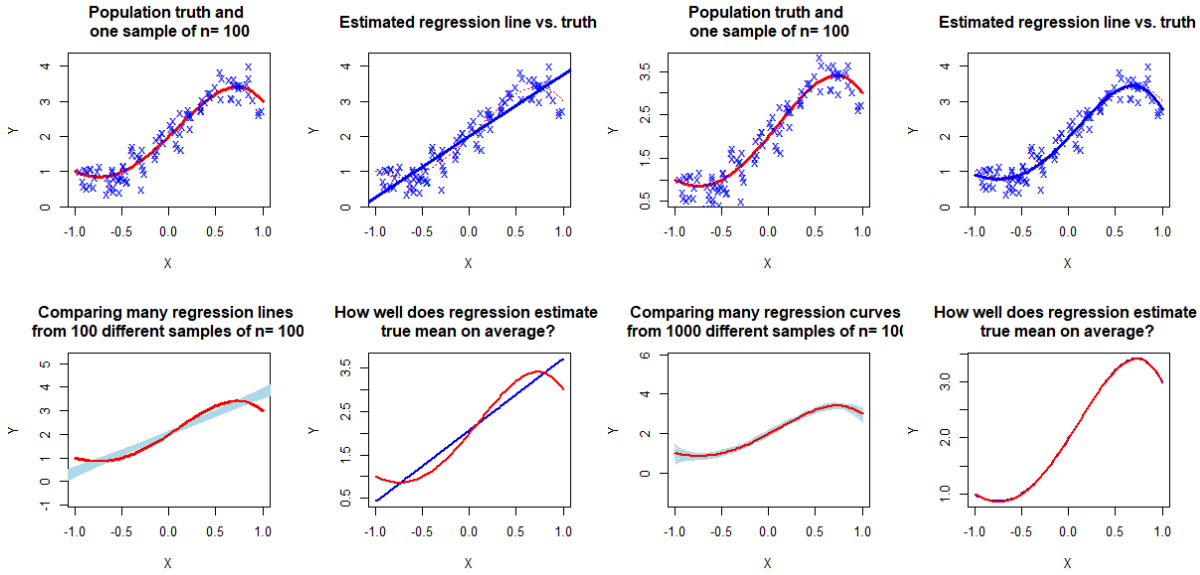
- "Data" consists of a combination of true mean and random error

Figure 3: Repeat of Figure 2 using model that has same structure as truth. Population true structure is red. Samples and estimated regressions are blue. See full description in text.



- Fitting a function to data means fitting a function to *both* the true mean that we are trying to predict *and the particular random errors in the data set*
  - Classic idea of "signal + noise"
  - The model tries to get as close to all the data (including errors) as possible
    - We know that moving an observation up or down will change a linear regression slope and intercept
  - The more flexible a model is, the more it adapts to *the data* (signal AND noise)
  - Models that are "too flexible" can "chase errors"
    - Proper term for this is that the model OVERFITS the data.
    - It overreacts to the random noise, thinking it is signal
  - The result is a model fit that is highly variable from one sample to the next
    - This is referred to as MODEL VARIANCE and is directly related to the idea of a standard error for a fitted model
- The good news is, model variance is something we *can control*...sort of
  - We know that standard errors can be reduced by increasing the sample size
  - The same is true of model variance: larger sample size leads to less variable model fits

Figure 4: Repeating previous example for $n = 100$. True structures in red, sample estimates in blue.



- The larger the sample, the more likely it is that errors average out above and below the true mean
  - The effect of one weird point is diminished by the bulk of the sample
- In small samples, the effect of one weird point can dominate the fit, especially when models are flexible
  - The tendency to overfit data is higher when n is small

Example: Exploring Variance (RegressionBiasVarianceHighBias.R) In this example, we see what happens to bias and variance when we increase the sample size from 10 to 100 in the previous example. See Figure 4.

1. The main things to look at are the bottom two panels of each foursome.
2. The linear model:
   (a) Shows somewhat less variability than before.
   (b) Has the same bias as before.
3. The "correct" model
   (a) Has a LOT less variability than before. There are clearly no bizarre fits anymore.
   (b) The bias is 0 everywhere as expected.

In summary:

- The "correct" model is is now an excellent model that will give good predictions everywhere.

  - It has no bias and relatively low variance.

- The linear approximation is now comparatively poor, even though it hasn't changed much
  - Its high bias is not matched by high variance in the other model

---

- We never get the model perfectly right
  - Bias comes in because reality is more complicated than nearly any model we imagine
  - Variance comes in because we must fit models using data, and the errors that lie within it
- For a given sample (fixed n), different models have different potential for reducing bias and variance
  - Bias is reduced by using models that are flexible
    - Usually means more complex
    - More likely to overfit, leading to high variance
  - Variance is reduces by using models that are less flexible
    - Less prone to chase errors
    - More likely to be biased.
- This is the famous **BIAS-VARIANCE TRADE-OFF**

2 Conclusion

This bring us to, literally, the most important thing you will learn in STAT 452/652:

> **Choosing a good model for any problem is a matter of managing the bias-variance trade-off.**

- There is no one model shape or type that is optimal for every problem
- Model imperfection comes from a combination of bias and variance
  - They combine in opposing ways
  - More flexible models are less biased, more variable
  - Less flexible models are less variable, more biased.
- Modeling is about finding the "sweet spot" where the combination of the two contributions is minimized

  - In small samples, we may need to use models that are less flexible to control variability
  - In large samples, we can afford to give up a little variability in order to reduce bias

- The goal is to find the model that comes closest, most often, to the true mean
  - "closest, most often" still needs to be defined...be patient!
  - BTW, we don't know the true mean, so how can we possibly measure this???
    * There are ways...be patient!

- Choosing the best level of flexibility for a given data set is an art based on science
  - This is what we will work on!

3 Exercises

1. Here look a little more at the effect of sample size on fitting regression models. Use the R program **RegressionBiasVarianceHighBias.R** to do the following. This code produces the two plots as seen in Figure 4 when all the code is run. You do not need to understand how the program works or rewrite any code. A few lines from the top is a line that says "`n = 10`", which controls the sample size. We saw plots for $n = 10$ and $n = 100$ above.

   (a) Change n to be 25 and rerun the entire code. Report these figures. Focusing on the bottom left plots, which model seems to get closer to the true structure across the whole range of X: the linear or the 4th-order polynomial?

   (b) Repeat for $n = 50$, reporting the figures and commenting on how the bottom left plots change.

2. Now we try to better understand bias and variance in modeling. Suppose there were a different problem where the true structure had a little bit of curvature in it but not very much—much less than the one we've been studying so far. Suppose that everything else is as it was in the example.

   (a) Suppose we fit a straight line using the $n = 10$.

      i. Compared to Figure 2 would you expect to see more bias or less bias for this model fit?

      ii. Compared to Figure 2 would you expect to see more variance or less variance for this model fit?

   (b) As we increase the sample size for this situation, which gets smaller: bias, variance, or both?