STAT 452/652: Statistical Learning and Prediction

Owen G. Ward
owen_ward@sfu.ca

Fall 2023

These notes are largely based on previous iterations of this course, taught by Prof. Tom Loughin and Prof. Haolun Shi.

1 Introduction and Review

(Reading: ISLR 3.1–3.4)

Goals of This Section

You should previously have taken at least two courses in statistics: an intro course that is pretty standard and a course featuring regression. In this section we will:

1. Review some of the most important ideas from your *first* course in statistics
   - (a) What is a "model"
   - (b) sampling distributions
   - (c) standard errors
2. Review important features about linear regression, from your second course
   - (a) What the model means
   - (b) How model parameters are estimated
   - (c) Sampling variability
   - (d) Ways to extend the model to fit other shapes
3. Use these ideas to reconstruct what "regression modeling" is meant to achieve
   - (a) Why are models wrong?
   - (b) How are models wrong?
   - (c) How do we use this information to guide our efforts?

1.1 Review of Introductory Statistics

The typical intro-stat course contains the following. You should recall and remember all of it, or review if there is anything you don't recognize.

- Basic probability
- Parameters: population quantities of interest
  - Means, standard deviations, probabilities of events
- Statistics: estimates of parameters
  - Sample mean, sample variance and standard deviation, sample proportions
- Distributions, including normal distribution
  - Empirical rule
- Measuring variability of statistics
  - Standard errors
  - Sampling distributions, central limit theorem
- Hypothesis tests
  - Null/alternative hypotheses
  - test statistics
  - Type 1 errors, α, p-value
  - Interpretations
- Confidence intervals
  - Coverage
  - Interpretations

Distributions as Models

Classical statistics starts with a MODEL. So what does that mean?

- Classical statistics starts with POPULATION of units and some particular measurement on those units
  - A population is just a word that means "the collection [or SET] of all possible units"
    - Equivalently it is the collection of all possible values that the particular measurement can take on
  - Ex: CGPA in first year for students entering Canadian universities from high school in 2010s
    - Population is students entering Uni from HS in these years
    - *OR* population is the CGPA after first year for these students

- A STATISTICAL MODEL is a probability distribution that we *assume* describes the measurements in the population
    - The probability distribution is just a **mathematical formula** that describes the shape of the histogram of measurements in the population
    - May be discrete (takes on only certain values) or continuous (takes on any value within an interval)
    - Usually chosen for combination of mathematical convenience and good fit
- **A MODEL IS JUST AN APPROXIMATION!**
    - There is no reason that real populations must adhere to man-made mathematical constructs
    - *No* population truly follows a normal distribution *exactly*
        - Implies that every measurement is taken to infinite number of decimals
        - Implies no lower or upper limits to size of measurement
    - **"All models are wrong. Some are useful." -George Box.**
    - *Many* populations are reasonably approximated by normal distributions
        - Central mound
        - Roughly symmetric tails on each side
- The model allows us to do probability calculations that would otherwise be impossible
    - Empirical rule
        - 68% ($\approx 2/3$) of population is within 1 standard deviation (SD) of the mean
        - 95% is within 2 SD of the mean
        - 99.7% is within 3 SD of the mean
    - Makes it possible to do tests with specific α and to compute p-values
    - Allows us to have XX% confidence that a confidence interval will cover a parameter.
    - ALL of classical statistical inference is based on models
    - While these assumptions were important when it was impossible to compute things quickly, many of these issues are circumvented with modern computers
- Calculations based on models give great answers when the approximation is good
    - Reliability of answers deteriorates when approximation becomes worse
    - Different statistics are more (or less) sensitive to violations from the model assumptions than others
- It's easy to check the quality of a model fit in simple problems
    - It can be impossible to check it in complex ones.
    - *Guess which type we will study?*

Effects of Sampling on Statistics

- When we "do statistics" we start by identifying one or more PARAMETERS we are interested in learning about
  - Parameter is just a quantity that could be computed from the population
  - Ex: Average CGPA or fraction of CGPAs below 2.0 among all students in the population.
- We draw a sample of "n" units from this population and measure them
- We use the sample to compute a statistic that estimates the parameter
  - Call this an "estimate"
  - Ex: average CGPA in sample or proportion of sample with $CGPA < 2.0$
- **It is important to understand that the estimate we compute depends on the sample we draw!**
  - If we sampled a different set of n units, the estimate would be a little different
  - In fact, every different sample of n units has the potential to give a different estimate
- We could consider making a histogram of all the different possible values that a statistic can take on
  - The population histogram of all possible values that a statistic could take from a sample of size n is called THE SAMPLING DISTRIBUTION OF THE STATISTIC.
  - See the demonstration here: http://onlinestatbook.com/stat_sim/sampling_dist/
- In some cases, mathematical analysis of a model tells us about the sampling distributions for certain statistics
  - The sample mean of a sample from a normal distribution also has a normal distribution
  - The CENTRAL LIMIT THEOREM (CLT) tells us that that sample means of samples from *(almost) ANY* distribution have an approximately normal sampling distribution.
    - The approximation is closer if the population distribution is closer to normal
    - The approximation is closer when the sample size is larger.
  - This is what we use for all of our inferences—tests and confidence intervals—in classical statistics.
- Even when we don't need to do formal inference, we still should have some measure of uncertainty of a statistic
  - A single "point estimate", like "100" is of limited use
  - Is it 100 ± 0.1 or 100 ± 50?
  - The difference is very meaningful for how we interpret the results and how we act upon them.
- We can get a measure of uncertainty of a statistic from its sampling distribution

  - If the sampling distribution of an estimate is very concentrated, then we have a good idea where the true parameter is.
  - If it is spread out, then we aren't so sure

- The standard deviation of a statistic's sampling distribution is called the STANDARD ERROR (SE) OF THE STATISTIC
  - Some simple statistics (means, proportions, estimated linear regression parameters) have simple formulas for standard errors
  - The SE for MANY more complicated statistics have complicated formulas or no formulas

- **In all cases, though, the larger n is, the smaller the SE.**
  - Estimates get more precise (less variable) as the sample size increases
  - *This is the key thing I want you to remember about SEs, because this principle will carry through everything we do.*

Problem Set 1: Sampling Distributions

1. Visit http://onlinestatbook.com/stat_sim/sampling_dist/ and play with this app as described below. This is meant to help you better understand how sampling distributions work.

   (a) Start with the normal distribution that is represented on the initial screen as your "parent population". Select "Mean" in the third plot, "Variance" in the fourth plot, and "N=5" for both. Run 10,000 samples through this simulation. Take a screenshot of the results showing all four plots, including the statistics on the left and the settings on the right. Present it as your response. No comment needed yet.

   (b) Clear everything and repeat with "N=25" in both places.

   (c) Clear everything and repeat with "Skewed" in the first plot and "N=5" in the third and fourth.

   (d) Clear everything and repeat with "Skewed" in the first plot and "N=25" in the third and fourth.

   (e) Comment on the following, explaining what evidence these plots provide to support your answers. Please note that the scales on the X-axis sometimes change, so factor this into your explanations when necessary. You shouldn't need more than a sentence, or *maybe* two for each. If you say too much, you are missing the big points and will be penalized.

      i. Do different statistics computed on the same samples have to have the same sampling distribution?

      ii. What effects does increasing sample size have on the sampling distributions of statistics?

      iii. What effects does changing the parent distribution (the distribution from which data are sampled) from normal to skewed have on the sampling distributions of statistics?

1.2 Review of Simple Linear Regression

- The notion of regression is simple: try to use one measurement (or set of measurements) to predict another one
- All of these measurements are represented as variables
  - The variable being predicted is denoted by Y
    - Called the response variable (also target, output, dependent variable)
  - The variables doing the prediction are denoted by X with optional subscripts when there are more than one of them $(X_1, X_2, \ldots X_p)$
    - Called explanatory variables (also predictor, input, independent variables)

Example: Prostate data (`Sec1_ProstateData.R`) The book used for the more advanced version of this class [1] has an example on measurements of men with prostate cancer. We will use this example also. The book description is quoted below:

> The data for this example, displayed in Figure 1.11, come from a study by Stamey et al. (1989) that examined the correlation between the level of prostate specific antigen (PSA) and a number of clinical measures, in 97 men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (`lpsa`) from a number of measurements including log cancer volume (`lcavol`), log prostate weight `lweight`, age `age`, log of benign prostatic hyperplasia amount `lbph`, seminal vesicle invasion `svi`, log of capsular penetration `lcp`, Gleason score `gleason`, and percent of Gleason scores 4 or 5 `pgg45`.

So in our context, Y=`lpsa` and the other 8 variables are all potential X's $(X_1, \ldots, X_8)$. A scatterplot matrix is seen in Figure 1 based on the code below. We will focus on the relationship between `lpsa` and `lcavol`.

```
> prostate <-  read.table("Prostate.csv",
+                      header=TRUE, sep=",", na.strings=" ")
> round(head(prostate), digits=3)
  lcavol lweight age   lbph svi    lcp gleason pgg45   lpsa
1 -0.580   2.769  50 -1.386   0 -1.386       6     0 -0.431
2 -0.994   3.320  58 -1.386   0 -1.386       6     0 -0.163
3 -0.511   2.691  74 -1.386   0 -1.386       7    20 -0.163
4 -1.204   3.283  58 -1.386   0 -1.386       6     0 -0.163
5  0.751   3.432  62 -1.386   0 -1.386       6     0  0.372
6 -1.050   3.229  50 -1.386   0 -1.386       6     0  0.765
>
>
> pairs(prostate)
```

- When we gather data like this, we may have a number of questions in mind:
  - *Is there any relationship between X and Y? If so, how strong is it?*

[1] Hastie, T.; Tibshirani, R.; and Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition.* Springer. Available for free https://web.stanford.edu/~hastie/Papers/ESLII.pdf

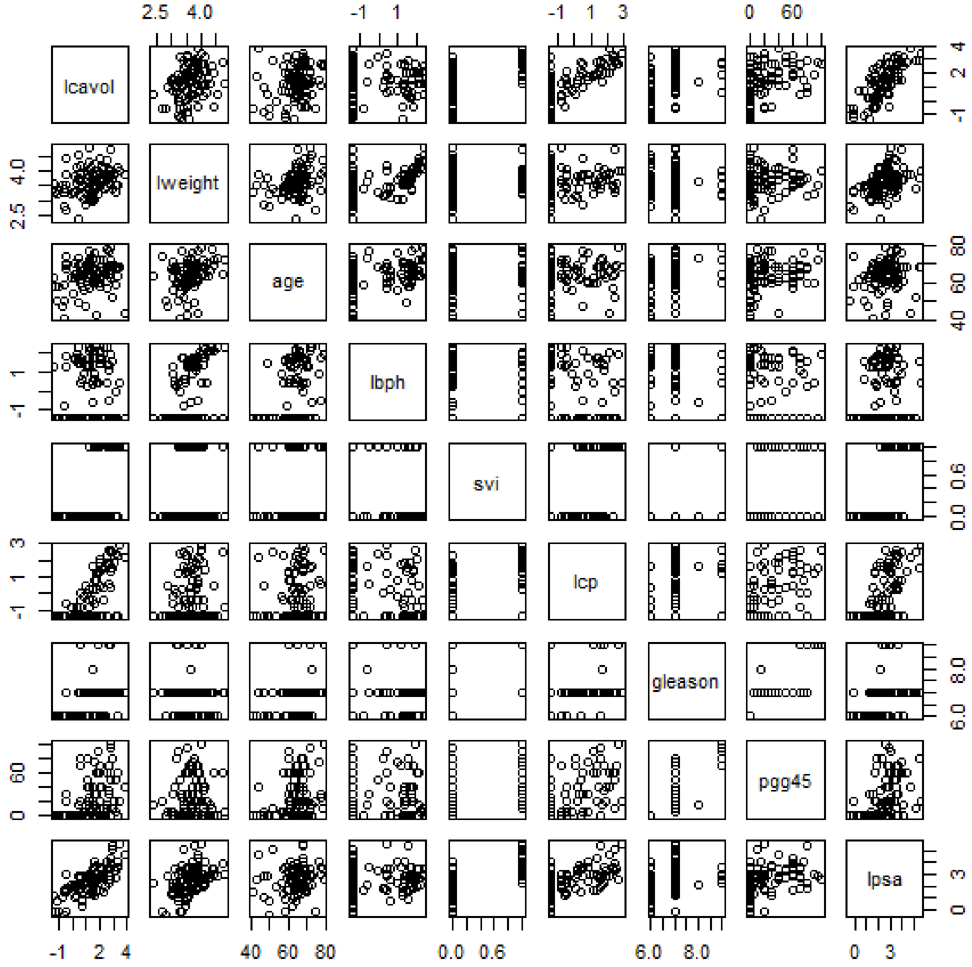


Figure 1: Scatterplot matrix for Prostate data.

    - If there is no relationship, then there is no point in studying the relationship further
    - If there is a relationship, is knowing X a very good substitute for knowing Y, or is it not much better than a random guess?
  - *If we have several explanatory variables, which one(s) relate to Y? How strong are these relationships?*
    - This deals with our ability to separate out the effects of different explanatory variables, which may themselves be correlated
    - We especially want to focus on the variables that have the most effect.
  - *What does the relationship between Y and X look like? Is it linear?*
    - If we want to use X to predict Y, we will have to try to mimic their relationship somehow.
    - The easiest relationship to mimic is a linear one, as we will soon see.
  - *How accurately can we predict Y?*
    - Remembering that a point estimate by itself is not ideal, if we are trying to predict Y, we really should try to attach a measure of uncertainty to our predictions

- Very rough initial clues toward these answers can often be found from a scatterplot matrix like that in Figure 1.
  - It is a very good idea to make plots like this if the size of the data set allows.

Linear regression: Straight line relationship

- Mathematically, a straight line is the simplest possible relationship between two variables.
  - In the early days, simple was good.
  - Many relationships are more or less monotonically (constantly) increasing or decreasing
  - Approximating these relationships as linear is a reasonable place to start.
- Equation for a straight line is $Y = \beta_0 + \beta_1 X$.
  - β_0 is the INTERCEPT; i.e., the value of Y when $X = 0$
  - β_1 is the SLOPE; i.e., the change in Y for each 1-unit increase in X
    - Depends on the units in which Y and X are measured
    - $\beta_1 > 0$ is an increasing relationship, $\beta_1 < 0$ is a decreasing relationship
    - See Figure 2, top left
- In a regression model, β_0 and β_1 are *parameters*
  - Their values are unknown in advance
  - We need to take a sample and estimate the parameters using the data
  - Obviously, those estimates will change depending on the particular sample chosen

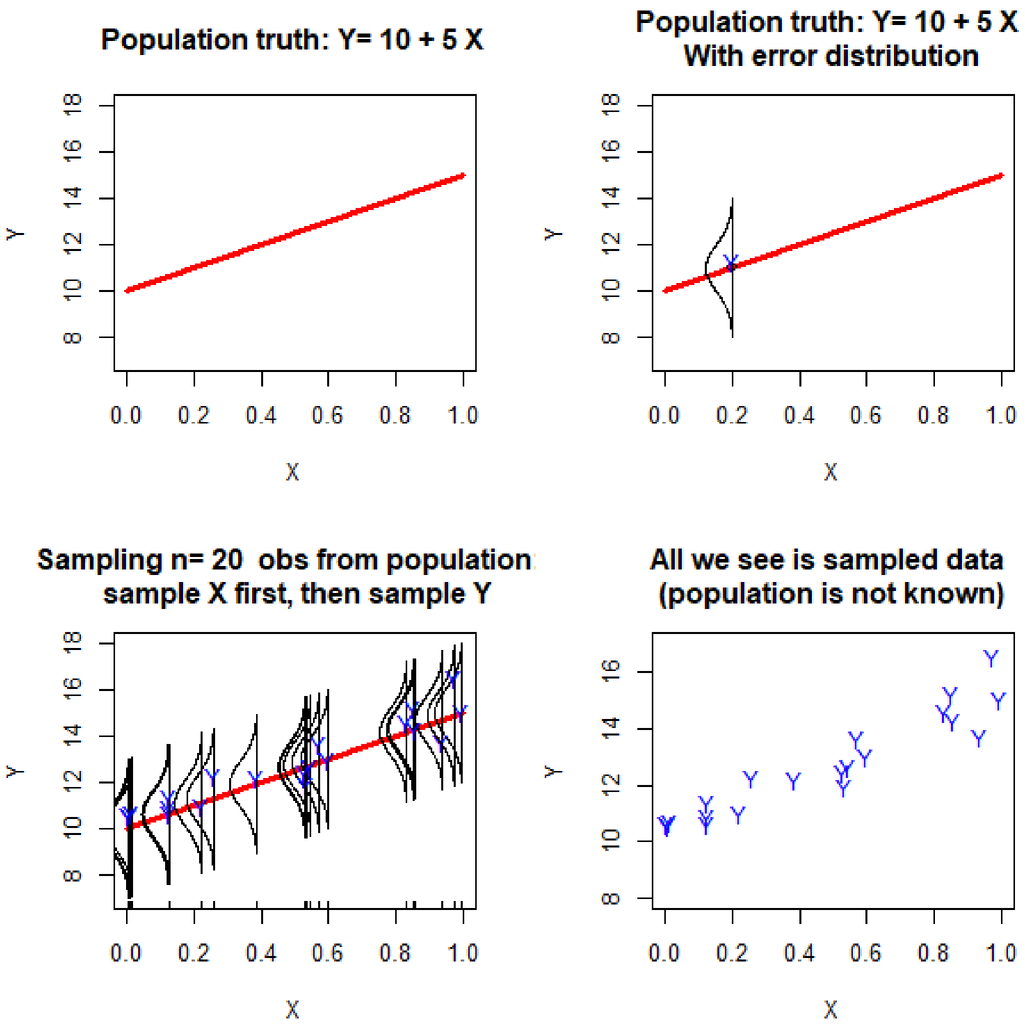


Figure 2: Example of how regression models are built. **Top left**: A straight line, here with intercept $\beta_0 = 10$ and slope $\beta_1 = 5$. **Top right**: The normal distribution for ϵ shown around the line at a particular value of X, along with one random draw from that distribution (blue "Y"). **Bottom left**: Repeating the sampling process for $n = 20$ observations drawn for Y at 20 random values of X. **Bottom right**: The data, when all of the invisible model elements are removed.

The regression model and estimation

- The first thing to note is that the data never lie *perfectly* on the straight line
  - If they do, be suspicious!
  - Instead, there is variability around any line you might imagine
  - We might wonder where that variability comes from (more later)
- In classical statistics, we *assume* that the sample we collect is generated from some model for the population
  - Standard model is to assume that, for a given X value, Y originates from this model:

    $$Y = \beta_0 + \beta_1 X + \epsilon, \quad \epsilon \sim N(0, \sigma^2), \tag{1}$$

    where "$\sim N(0, \sigma^2)$" means "is Normally distributed with mean of 0 and variance of σ^2".
  - Since the errors ϵ have mean 0, then they do not add anything to the regression, on average.
    - They just allow there to be variability around the line
    - Thus, the mean value of Y at any value of X is exactly the line, $\beta_0 + \beta_1 X$
  - **This is true in general: when we say we are modeling Y, we usually mean we are modeling the *mean* of Y.**
    - We often mean just the line part when we talk about "the model for Y"
  - See Figure 2 to see how the model generates data.
- Once we have data, we need to use it to estimate the parameters β_0 and β_1
  - There are infinitely many possible combinations of β_0 and β_1—which one should we use?
- We want to choose a line that comes "as close as possible" to the points
  - Can be hard to get close to all of them—often, moving a line closer to one point moves it farther from another
  - We need an *aggregate* measure of closeness that applies to the whole data set
  - Then find the best line that minimizes the measure of closeness
- For historical reasons, we use the LEAST SQUARES (LS) CRITERION
  - Represent a sample from variables Y and X with small letters and subscripts
    - We observe pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_n y_n)$. These are the plotted points
  - The LS criterion measures the sum of squared deviations between observed points y_i, $i = 1, \ldots, n$ and the respective locations on a line, $\beta_0 + \beta_1 x_i \, i = 1, \ldots, n$:

    $$\sum_{i=1}^{n} (y_i - [\beta_0 + \beta_1 x_i])^2 \tag{2}$$

- This gives us an OBJECTIVE FUNCTION, a formula that we want to optimize

  - Through calculus, we can differentiate with respect to the parameters, set equal to zero, and solve for the parameters
  - This gives formulas for parameter estimates that you may have learned, but I won't care about here
    * They are called the LEAST SQUARES ESTIMATES (LSEs)
  - We will refer to the estimates as $\hat{\beta}_0$ and $\hat{\beta}_1$, where a "hat" (ˆ) always means a quantity estimated from the data.

- Now we have a "prediction function": an estimated regression line, $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$
  - We can plug in any value for X and come up with a prediction $\hat{Y}$.
  - See Figure 3, top left.

Example: Prostate data (`Sec1_ProstateData.R`) Figure 4 shows the relationship between Y =`lpsa` and X_1 =`lcavol` (left) and X_8 =`pgg45` (right). The relationship with `lpsa` looks pretty close to linear for `lcavol` but not as clear for `pgg45`. Estimated regression lines are shown in blue. The mean value for `lpsa` at `pgg45`=0 does not appear to be well estimated by the regression line at `pgg45=0`, although both estimates have variability that is not shown here. The code that does this is in the program for this example. The R function `lm()` fits linear regression using least squares and reports back the LSEs.

- You would have learned in your previous regression class about using residuals to examine the fit of the linear regression model.
  - We won't focus on that here, but it is always a good idea to make sure that your models are reasonable approximations

- There are many ways in which the model can be wrong:
  1. Nonlinear relationship
  2. Non-normal error distribution
  3. Non-constant variance
  4. Non-independent errors

- We will focus on ways to address #1, which is the most serious concern
  - #2,3,4 mainly affect inferences you make on the model
  - Being wrong about the model shape renders ALL inferences irrelevant.

The sampling distribution of regression estimates

- Of course, gathering data and fitting a regression line doesn't mean you have the right answer
  - The estimated parameters $\hat{\beta}_0$ and $\hat{\beta}_1$ are statistics and would change with a new sample
  - This means that the entire regression line and predicted values will change with a new sample
  - See Figure 3

- Understanding the sampling variability of predicted values will be important throughout this course, so keep it in mind

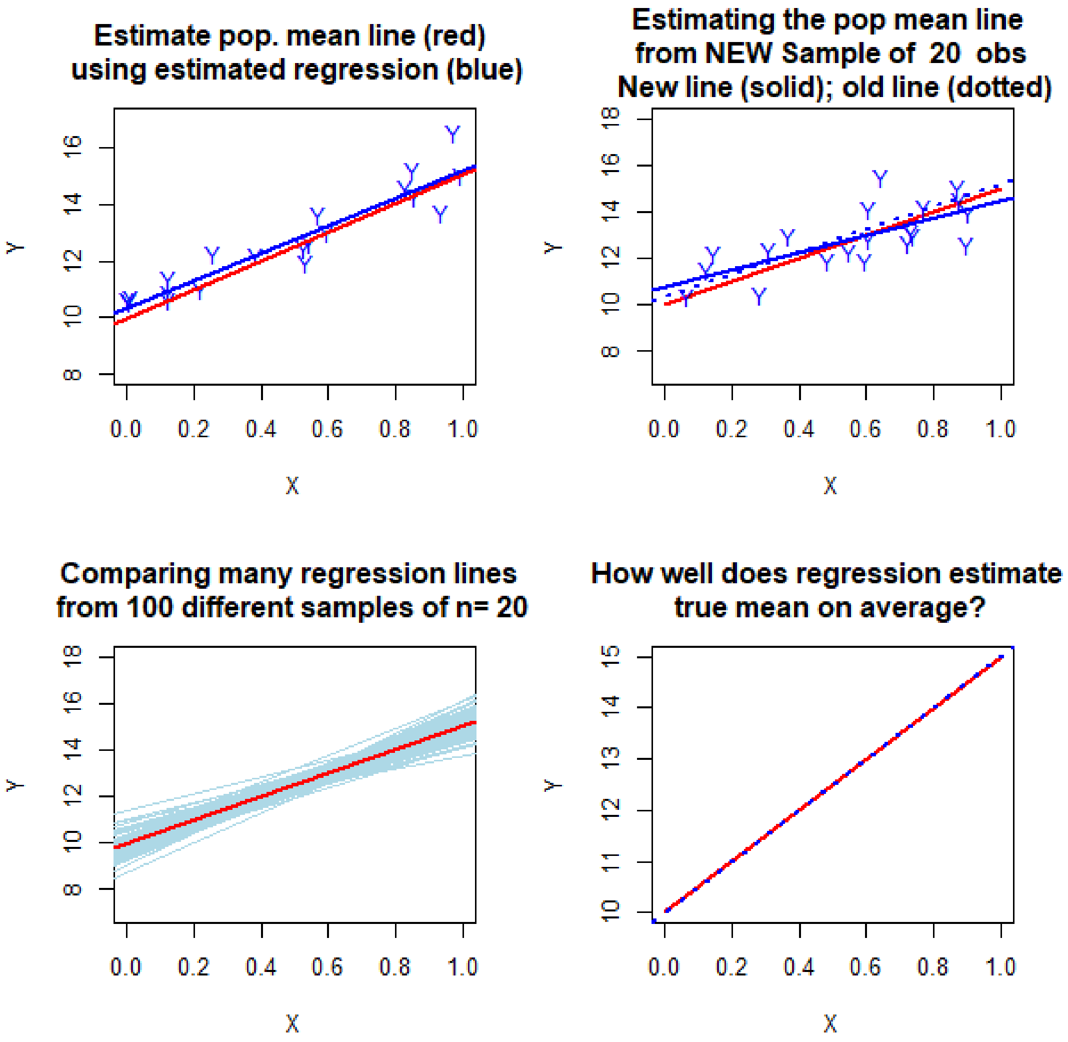


Figure 3: Example of the sampling distribution for regression estimates. **Top left**: Fitting the simple linear regression to the data in Figure 2 (blue line) and comparing it to the population "true line" (red line). **Top right**: A new set of data is drawn and the new line is fitted (solid blue). The old line from the previous plot is shown as dotted blue. Notice that they differ! **Bottom left**: Repeating the sampling process for 100 separate samples of size 20, fitting regression line to each sample. Each line is represented in light blue. Notice the variability among the lines, but also they are somewhat stable and close to the actual line. **Bottom right**: The average of the 100 estimated lines (blue dotted line) compared to the true population line. Notice that, on average, the linear regression is practically perfect (no bias!).

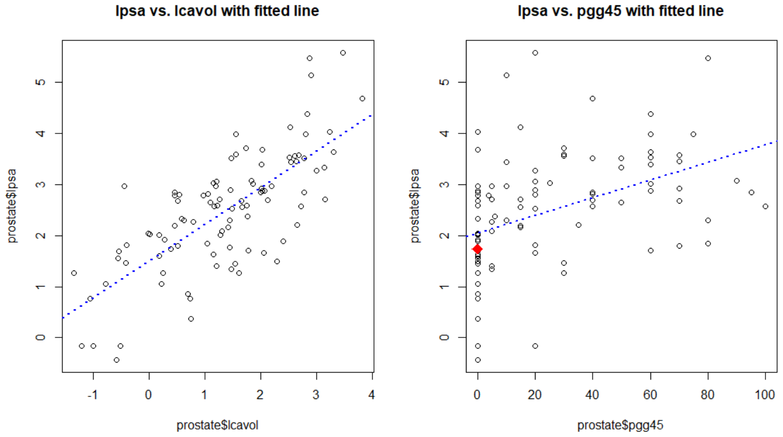


Figure 4: Closeup of two variables from the Prostate data, with estimated regression lines in blue. Right plot also includes mean value of `lpsa` for all data with `pgg45`=0.

1.3 Review of Multiple Linear Regression

- Simple linear regression is rarely used as a final analysis
- Most problems have more than one variable
  - Want to understand all of their effects
  - Want to build regression model
  - Want to know which variables are important
- Model structures is harder to visualize, except when there are only 2 variables
  - Generally referred to as "surfaces"
- Multiple linear regression model is just like simple linear model, but with more variables.
  - Main new element is p explanatory variables, $X_1, \ldots, X_p$
- Standard model is to assume that, for X value, Y originates from an expanded version of the simple linear regression model (1):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon, \quad \epsilon \sim N(0, \sigma^2) \tag{3}$$

  - $\beta_0, \ldots, \beta_p$ are regression parameters or coefficients
    - β_0 is the intercept (mean value of Y when all $X_j = 0$)
    - β_j, $j = 1, \ldots, p$ are "(partial) regression coefficients"
    - Change in mean value of Y for 1-unit change in X_j *holding all other variables constant.*
- Need to estimate the parameters

  - Minimize the LS criterion (2) again, expanded for full model

$$\sum_{i=1}^{n}(y_i - [\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}])^2,$$

    where $x_{i1}, x_{i2}, \ldots, x_{ip}$, $i = 1, \ldots, n$ are the n observed values of the variables $X_1, \ldots, X_p$

- Mean value of Y for given values of $X_1, \ldots, X_p$ is $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$
  - The shape of this function is a p-dimensional hyperplane.
  - This is a fancy term for something that reduces to a straight line in every direction
- When $p = 2$ we can visualize this
  - X_1, X_2, Y are points in 3 dimensions, so picture a cube
  - X_1,X_2 form the bottom surface of the cube
  - Y values are points floating above the surface in a cloud
  - Multiple linear regression model will fit a plane (a board) through the points (see example)
- Multiple regression has several challenges that are not faced when $p = 1$
  - When $p > 2$,we can no longer fully visualize the surface
  - We may not be sure that all variables are necessary
    - There is uncertainty about what is the “right” model
  - Multicollinearity—correlated explanatory variables—is an incredible nuisance
    - Can lead to variables having unexpected (and impossible) coefficients as variables “explain overlapping information” in Y
  - Harder to tell when there are problems with the fit of the model
    - Residual analysis is less reliable
- Despite these challenges, we *must* move forward
  - We can’t control what the data give us; we just have to develop tools to extract what we can from it

Example: Prostate data (`Sec1_ProstateData.R`) Let’s look at a 3D plot of the three variables we looked at in the previous example, `lpsa` vs. `lcavol` and `pgg45`. This is best done live with the program and code.

```
> prostate <-  read.table("Prostate.csv",
+                        header = TRUE, sep = ",", na.strings = " ")
> round(head(prostate), digits=3)
  lcavol lweight age   lbph svi    lcp gleason pgg45   lpsa
1 -0.580   2.769  50 -1.386   0 -1.386       6     0 -0.431
2 -0.994   3.320  58 -1.386   0 -1.386       6     0 -0.163
3 -0.511   2.691  74 -1.386   0 -1.386       7    20 -0.163
4 -1.204   3.283  58 -1.386   0 -1.386       6     0 -0.163
5  0.751   3.432  62 -1.386   0 -1.386       6     0  0.372
```

```
6 -1.050   3.229  50 -1.386   0 -1.386       6     0  0.765
>
>
> pairs(prostate)
```

Problem Set 2: Simple and Multiple Linear Regression

Refer to the Air Quality data available in R as the data frame “airquality”.
Run `help(airquality)` to learn a little more about this data set. We will treat `Ozone` as the response variable and use `Temp`, `Wind`, and `Solar.R` as explanatory. We won’t use `Month` or `Day`.

1. Create a separate data frame for these data containing only the variables we will need. You can use something like
   `AQ = airquality[,1:4]`.
   Then create a scatterplot matrix of these four variables. Comment on
   (a) Relationships of each X with Y
   (b) Relationships among the three explanatories.
2. Run separate simple linear regressions of `Ozone` against each explanatory variable.
   (a) Report the three slopes and t-values in a table.
   (b) Make three separate scatterplots and add the respective regression lines to each plot. Present the plots and comment on how well the lines seem to fit each variable.
3. Make a 3D plot of `Ozone` against temperature and wind speed. Rotate it around and notice to yourself what relationship the ozone might have jointly with temperature and wind. Take a screenshot from any angle you think helps you to see most of this relationship. No comments are needed.
4. Fit the multiple linear regression that corresponds to this 3D plot.
   (a) Report the slopes and t-values. Are they much different from when they were computed in simple linear regressions?
   (b) Add the plane surface to the 3D plot. Rotate it around and comment on the quality of the fit. Show a screenshot from some angle that helps to support your comment

1.4 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?

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 can't 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
  - We may refer to $g(\mathbb{X})$ as the "universal function" or "universal predictor" or things like that.
- 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.
  - Does this sound impossible?

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 soon see how to go beyond this
- 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* really that 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:
            - · 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

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. We will talk a lot about bias in this class!
        - ∗ Imagine modeling a curved 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 5)
        - ∗ 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 try to 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 5)
    - 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?

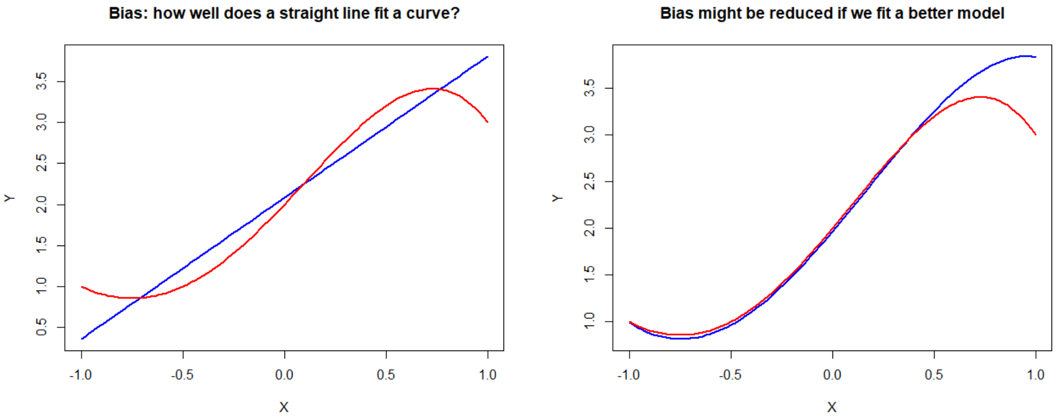


Figure 5: Bias from fitting a different models $f(X)$, blue, to a curved true relationship $g(\mathbb{X})$, red.

Example: Exploring bias (`Sec1_RegressionBiasVarianceHighBias.R`) Let's explore that happens when we allow models to be more flexible. In this example:

1. We are creating the universal truth to be

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

   with parameter values chosen to create the curve seen in red in all four panels of Figure 6.

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

3. We can 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) We 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 we can 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, take the 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.

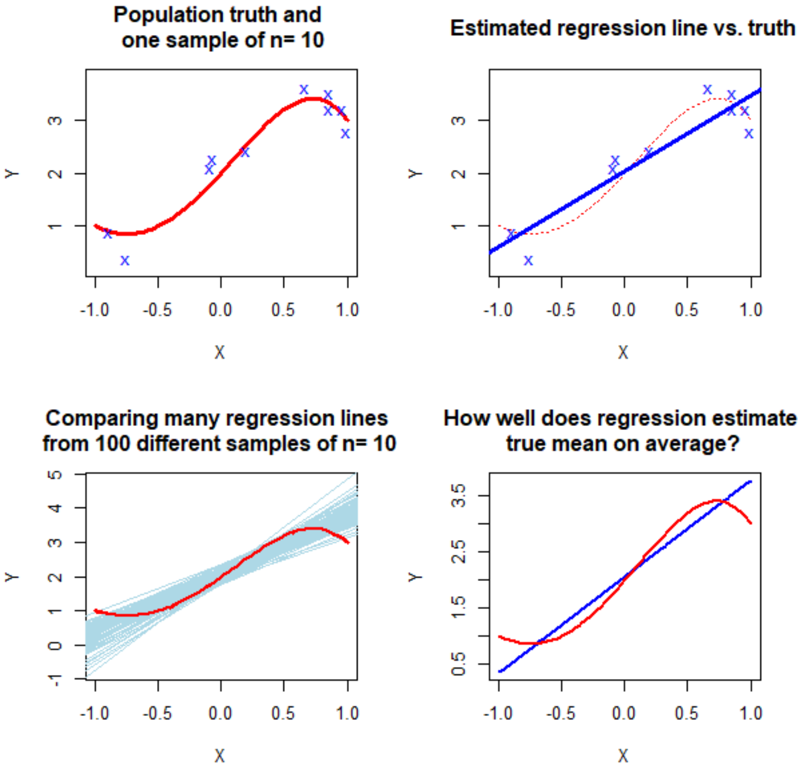


Figure 6: Population true structure is red. Samples and estimated regressions are blue. See full description in text

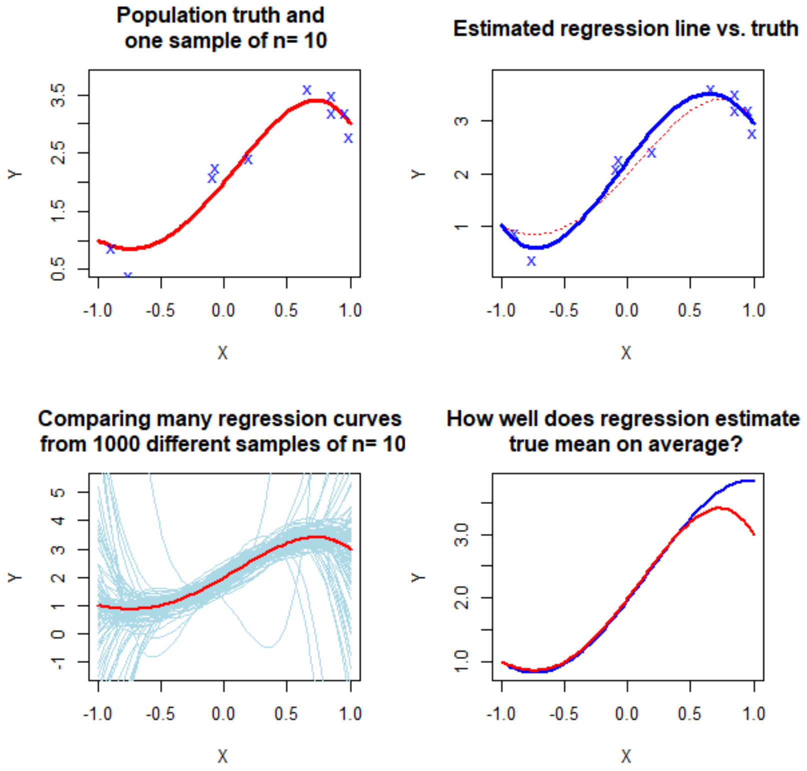


Figure 7: Repeat of Figure 6 using model that has same structure as truth. Population true structure is red. Samples and estimated regressions are blue. See full description in text.

- 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 7 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
  - 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?

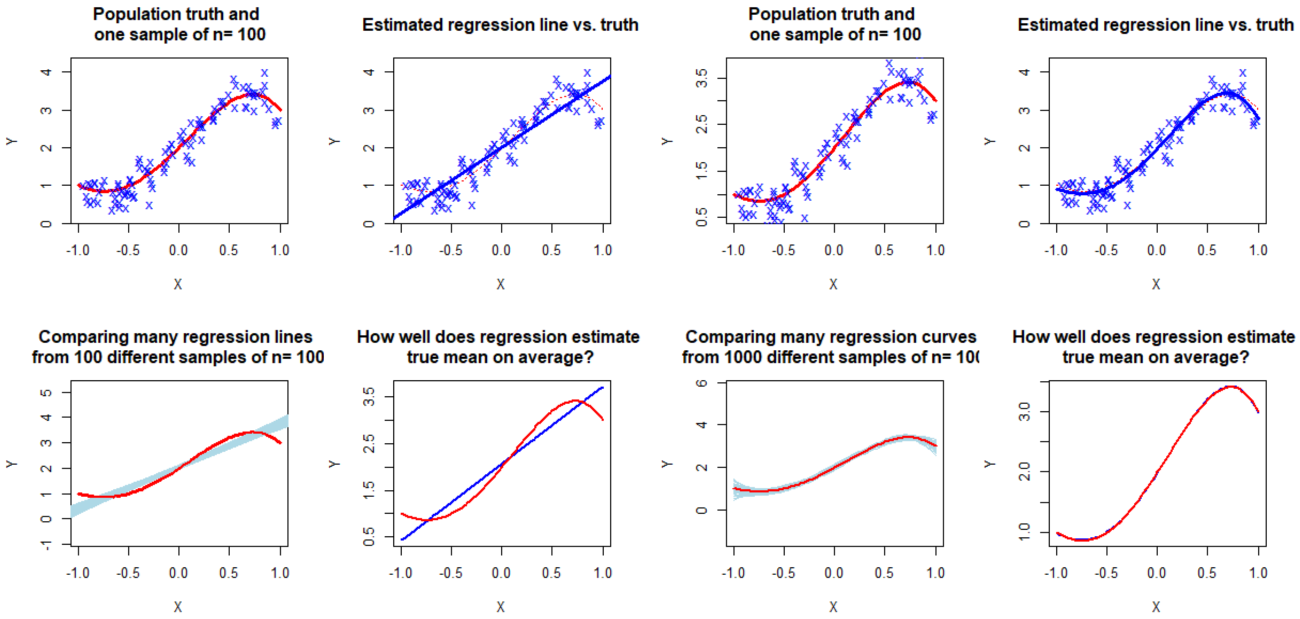
- "Data" consists of a combination of true mean and random error
- Fitting a function to data means fitting a function to *both* the true mean that we are trying to predict *and the particular random errors in the data set*
    - Classic idea of "signal + noise"
    - The model tries to get as close 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
        - 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 (`Sec1_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 8.

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:

Figure 8: Repeating previous example for $n = 100$. True structures in red, sample estimates in blue.



- 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**

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 trade-off between bias and variance.**

- 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!

Next we will spend time examining ways to compare multiple models in a principled way, that will allow us to select a "best" model.

Problem Set 3: More Regression

1. Here we look a little more at the effect of sample size on fitting regression models. Use the R program **`Sec1_RegressionBiasVarianceHighBias.R`** to do the following. This code produces the two plots as seen in Figure 8 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 6 would you expect to see more bias or less bias for this model fit?

      ii. Compared to Figure 6 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?