STATISTICS 452/652: Statistical Learning and Prediction

August 20, 2020

Regression Review Part 2: Simple Linear Regression

(Reading: ISLR Sections 3.1–3.4)

1 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 (L2-ProstateData.R) The book I use for the more advanced version of this class[1] has an example in it on measurements made on men with prostate cancer. We will use this example in our work. 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

[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

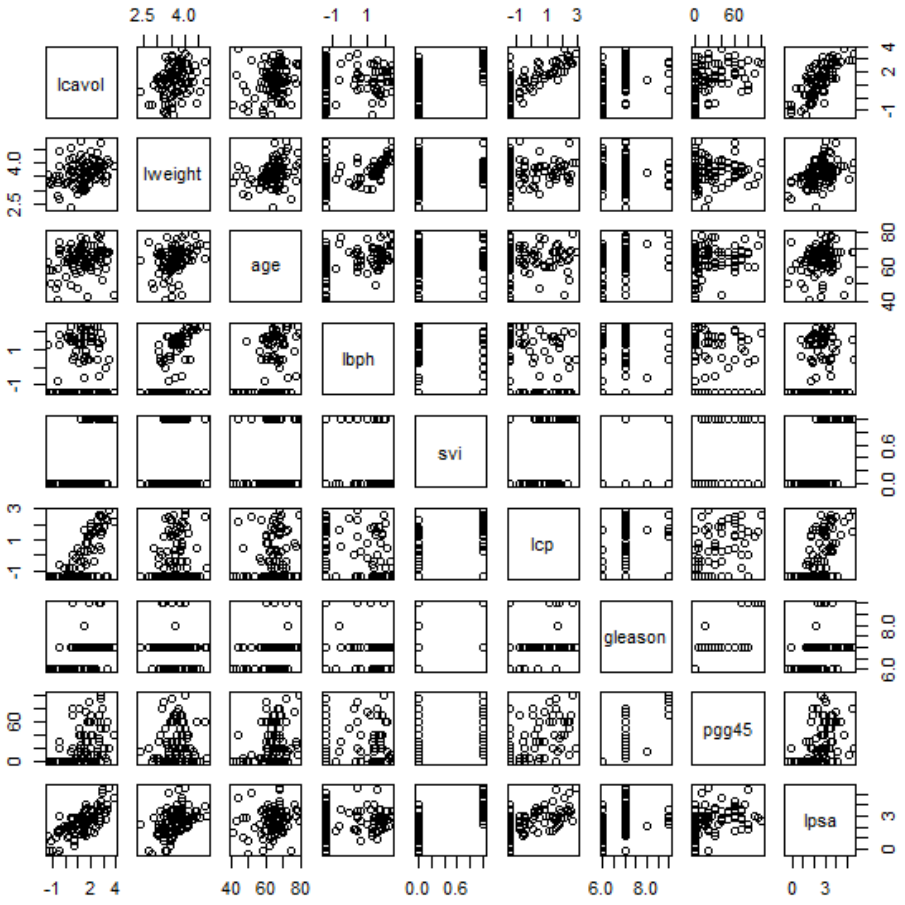
(`lpsa`) from a number of measurements including log cancer volume (`lcavol`), log prostate weight `lweight`, age `age`, log of benign prostatic hyperplasia amount `lbph`, seminal vesicle invasion `svi`, log of capsular penetration `lcp`, Gleason score `gleason`, and percent of Gleason scores 4 or 5 `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("C:\\...\\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
>
> x11()
> 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?*
    - 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?*

Figure 1: Scatterplot matrix for Prostate data.



    * 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.
  - I *always* recommend making plots like this if the size of the data set allows.

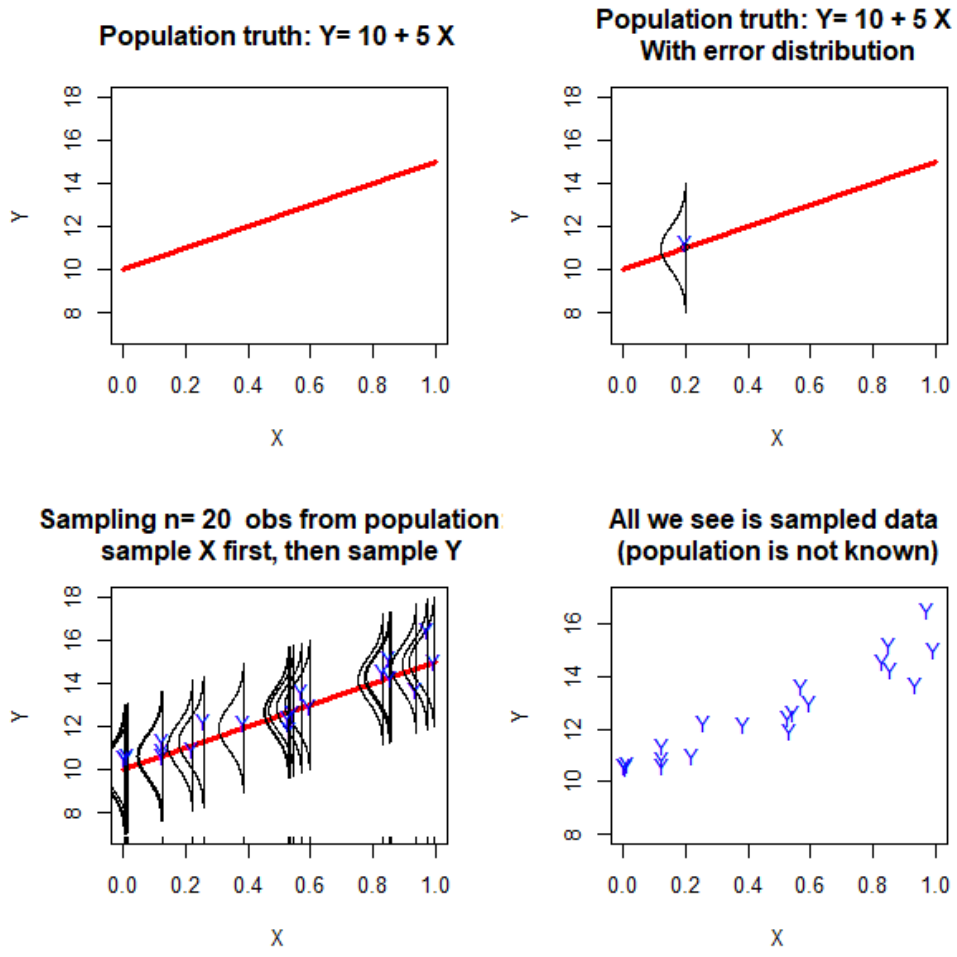
1.1 Linear regression: Straight line relationship

- 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

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

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.



- 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 distributed as Normal 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 like 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 (L2-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.

Figure 3: Example of the sampling distribution for regression estimates. **Top left**: Fitting the simple linear regression to the drawn 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 is 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!).

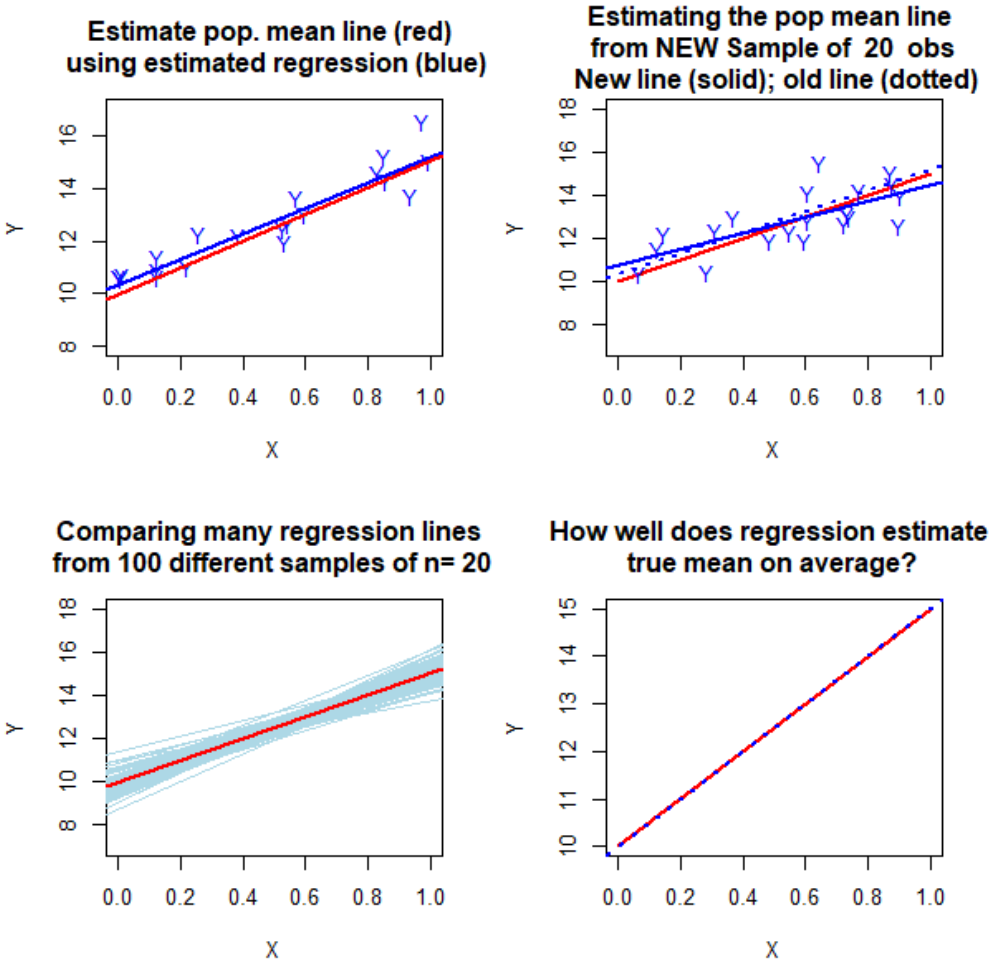
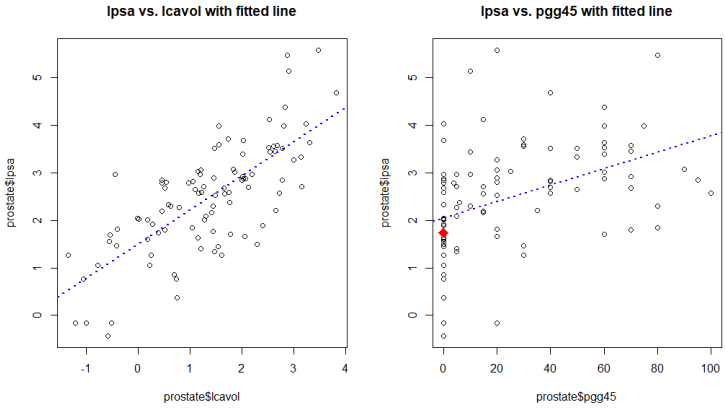


Figure 4: Closeup of two variables from the Prostate data, with estimated regression lines in blue. Right plot also includes mean value of `lpsa` for all data with `pgg45`=0.



1.2 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 for us