INSTRUCTOR NOTES FOR M358K FOR PART VII (CHAPTER 23: INFERENCES FOR REGRESSION, ONLY) OF DEVEAUX, VELLEMAN AND BOCK, *STATS: DATA AND MODELS*, 3RD ED.

Chapter 27: Inferences for Regression

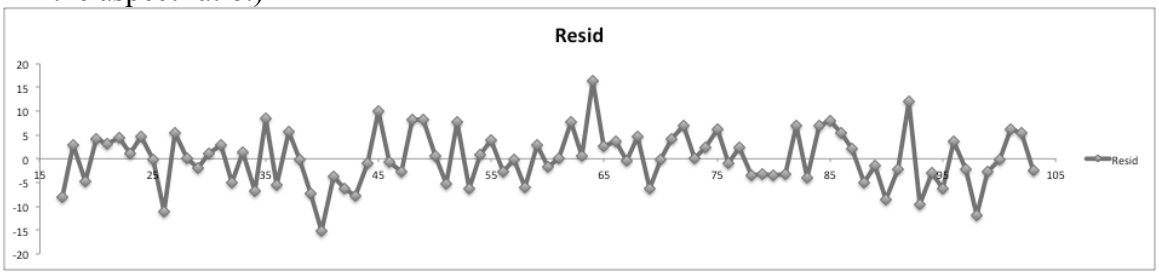
- Suggest that students review Chapters 8 and 9 before starting this chapter.
- p. 674
  - The paragraph right before The Population and the Sample (starting "How useful is …" makes the important distinction between "description of data" and "inference to the population."
  - Although the ActivStats activity "Simulate the Sampling Distribution of a Regression Slope" is mentioned on p. 680, it might be good to use at this point.
- pp. 674 – 675 The Population and the Sample
  - The idea behind Figure 27.2 is good, but the rendering is difficult, with the resulting being less than optimal. You may need to point out that the histograms are intended to be perpendicular to the plane of the axes. You may want to show (or suggest that students make) a three-dimensional model. One method: Use legos or something similar. Another: Make a pop-up model using the template and instructions at http://www.ma.utexas.edu/users/mks/M358KInstr/M358KInstructorMaterials.html. (The latter also has some more information that may be useful in helping students think about what is going on.)
  - Be sure to emphasize the highlighted sentence "The model assumes that the *means* …" (paragraph below Figure 27.3). Point out that these are conditional means: μ_y = E(Y|x). (Simple notation is inherently problematical here: From one perspective, using μ_x rather than μ_y would make sense, to show that this number depends on x. On the other hand, it is the mean of a Y variable, so using subscript y makes sense.)
  - It might be helpful at this point to bring in the picture in Figure 27.6.
    - Be sure to compare and contrast with Figure 27.3.
    - I've found it useful to use a 3-D model of Figure 27.6. It was easy to make: I just drew the mean line on the back of a pad of paper, then marked four x-values and taped on at each one a pipe cleaner bent in the shape of a normal distribution. (Any other easily bendable but easily visible type of wire should work.) Using the back of a pad allows the model to fit in one of the boxes that publishers often use to send textbooks (especially when they're sending more than one), which makes it easy to carry to class and to store for use in a later class.
  - Emphasize using the notation μ_y instead of $\hat{y}$ (top of p. 675) when talking about the model. Among other things, this may help prevent confusion when you get to prediction intervals.
  - Point out that the errors ε are random variables. (There is some abuse of notation here as well. To be more precise, for each x, $Y - \beta_0 - \beta_1 x$ is a random variable, which we might call ε_x, but one of the model

assumptions is that all of the ε_x's have the same distribution, so we suppress the x. Figure 27.7 or a 3-D model of it illustrates this: the normal distributions shown all have exactly the same shape and size, but are centered at different points on the line of conditional means.)
    - The sentence, "But the advantage of a model is …" makes a good point.
    - Emphasize the difference between the residuals e and the errors ε.
    - The sentence, "Our challenge is to account for our uncertainty …" makes in important point.
- pp. 675 – 678 Assumptions and Conditions
    - Introduction (p. 676)
        - The ActivStats activity Learn the Assumptions for Regression Inference might be helpful for some students as either an introduction to or a review of this section.
        - Emphasize the second paragraph, "Also, we need to be careful about the order in which we check conditions. …"
    - 1. Linearity (pp. 675 – 676) Emphasize:
        - The first sentence.
        - The caution about not drawing a straight line through the scatterplot.
        - Plotting residuals against x or predicted values.
        - Stopping or re-expressing if the scatterplot is not straight enough.
        - Only use quantitative data for regression
    - 2. Independence (p. 676) Add that spatial correlation is also a common source of lack of independence, so needs to be checked if appropriate. (Examples: If measurements are made every so many feet, or at increments of altitude; plants grown in a greenhouse or the same field)
    - 3. Equal Variance Assumption (p. 676)
        - Emphasize:
            - The importance of constant variance.
            - Plotting residuals against predicted values can be helpful.
        - You might add that sometimes re-expressing one or both variables can remove non-constant variance (but you need to be careful not to mess up linearity in the process.)
    - 4. Normal Population Assumption (pp. 676 – 677) Lot of good points (including footnote 1).
- p. 678 "Which Come First? …" is a nice summary of good statistical practice for simple regression – and also a good introduction to a common conundrum in doing statistics: You may need to look at the data to form a reasonable model, but looking at the data too much results in over-fitting to the sample, which defeats the purpose of being able to do inference to the larger population.
- pp. 679 – 680 The Step-By-Step Example (as usual) models good statistical practice.
- p. 680 If you haven't used the ActivStats activity "Simulate the Sampling Distribution of a Regression Slope" yet, you might want to at this point (or even repeat it if you've used it previously).
- pp. 681 - 683

- Figures 27.8, 27.9, and 27.10 are a good idea for building intuition about why the residual standard deviation, standard deviation of x, and sample size should all appear in the formula for the standard error of the slope.
- Be sure to show (or ask) how the formula for $SE(b_1)$ reflects each of the three observations suggested by Figures 27.8, 27.9, and 27.10.
- Be sure to draw attention to the Notation Alert at the bottom of p. 682.
- The ActivStats activities "Learn About Regression Slope Standard Error" and "Simulate Regression Slope by Changing the Standard Deviation of X" might be helpful or interesting for students to go through. (They might also help with the above or the following questions.)
- Additional questions you may want to pose:
  - If the x_i's are close together, what does this say about $SE(b_1)$?
  - How might this guide you in designing an experiment where you can choose which x's to use for taking observations?
- The optional supplement *Outlines of Proofs of Formulas on pp. 681-682* might be of interest to you or some of your students.

- p. 683 The optional supplement *Properties of the Intercept* has some facts about the least squares intercept and the covariance between the slope and intercept, along with some questions about what they imply.
- p. 683 – 690 Inference for regression brings with it a problem that is not addressed at this point in the text: The problem of multiple testing (AKA multiple testing, multiple comparisons, multiplicities, or The Curse of Multiplicity).
  - This problem is mentioned later in the text, as a starred section, pp. 733 – 735, in Chapter 28 (Analysis of Variance). That is a decent introduction, although it mostly focuses on the problem in the context of ANOVA.
  - However, the Bonferroni method (p. 734) does apply more generally.
  - If you have the time, I recommend that you mention the problem of multiple inference – it is something which is often neglected, but is receiving more and more attention recently in research in a number of (not all, unfortunately) fields.
  - My custom was not to discuss it in class, but to bring it up if students wanted to do a project that involved more than one hypothesis test on the same data. However, now that I am aware that it is a big problem leading to many "false discoveries", I would try to discuss it in class.
  - The Wikipedia page Multiple Comparisons, http://en.wikipedia.org/wiki/Multiple_comparisons, gives more information. I've got some discussion and references at http://www.ma.utexas.edu/users/mks/statmistakes/multipleinference.html (See also the link there to Data Snooping.)
  - For inference just involving regression coefficients, confidence regions can be used. However, since (as noted in the text) inference for the constant coefficient is rarely done, this is mainly relevant to multiple regression. The Wikipedia page http://en.wikipedia.org/wiki/Confidence_region has some (not the best) discussion.

- p. 684 Note the points in the blue box, and in the last paragraph of the "For Example".
- pp. 684 – 685
  - The "Just Checking" is good.
  - The ActivStats activity Compute a Hypothesis Test for the Regression Slope gives an overview of what goes into a hypothesis test for whether the regression slope is zero; some students might find it helpful.
- pp. 685 – 688 An interesting example.
  - p. 686 Here's a different plot of the residuals that might help in judging independence. (I found the data easily on the web, made a connecting line plot of residuals vs time, then dragged horizontally and vertically to adjust the aspect ratio.)



  - Note the comment "These are not a random sample, so …" (top of p. 687)
  - p. 688 Points to note:
    - The final sentence under "Tell More"
    - The box after the example
    - The footnote
- pp. 688 – 689: Standard Errors for Predicted Values
  - p. 688 A good first sentence.
  - Bottom of p. 688 – top of 689 The distinction between the two questions is important, but they are often confused.
  - The blue box on p. 689 helps explain the distinction.
  - A picture such as that at http://www.ma.utexas.edu/users/mks/statmistakes/CIvsPI.html (plus the paragraph preceding the picture) can also help.
  - Middle of p. 689: The paragraph starting "The standard errors …" is useful, especially the last two sentences.
  - Last sentence of p. 689: Good advice
  - You may want to assign filling in the details of Footnote 7 as an exercise.
- p. 690: Confidence Intervals for Predicted Values
  - I'm not in love with the use of the terminology "confidence interval for predicted value"; I prefer the terminology "prediction interval," since it helps make a distinction between (an ordinary) confidence interval and a prediction interval: A confidence gives an interval estimate of a *parameter* (e.g., the conditional mean), whereas a prediction interval is used to estimate the *value of a distribution* (in this case, conditional distribution).
  - Figure 27.11 is worth some study or explanation – it can help make the distinction between the two types of interval. (Note: The green lines are called "confidence bands" and the red lines "prediction bands". Note how the

confidence bands are curved. The prediction bands are, too, but the curvature is so slight that it doesn't show up.)

- The technical explanation of the prediction interval would go something like: "The interval (12.6, 31.2) was obtained by a process which, for 95% of all independent random samples $y_1, \ldots, y_{250}$, y taken from $Y|x_1, \ldots, Y|x_n$, $Y|_{38}$, respectively (where Y is percent body fat, x is waist size, and x_i is the waist size of the i^{th} man in the data set used to make calculations), would result in an interval containing the true percent body fat of a particular man with waist size 38 inches (assuming that all the model assumptions fit)."
- Prediction intervals occur in other situations as well as in regression –for example in some industrial settings. See http://www.astm.org/standardization-news/data-points/statistical-intervals-part-2-so11.html for some discussion if you're interested.

- p. 691 There is a *mistake* in the "For Example": The first step (calculation of a 95% confidence interval (0.452, 1.00) for the mean of logDiam when Age is 5 million years) is OK. The problem enters when transforming back to diameter:
  - If logDiam|(Age = 5) is symmetric, then $10^{\text{mean(logDiam|(Age = 5))}}$ is the *median* of Diam|(Age = 5) (because raising to a power is a monotonic function).
  - However, this will *not* in general be the mean of Diam|(Age = 5).
  - In particula, if the model fit exactly, logDiam|(Age = 5) would be normal (so Diam|(Age = 5) would be called *lognormal*). If the mean and standard deviation of logDiam|(Age = 5) are μ and σ, respectively, then the mean of Diam|(Age = 5) would be $10^{\mu+\frac{\sigma^2}{2}}$, which is *larger* than 10^{μ}.
  - The confidence interval (2.8, 10) obtained *is* a confidence interval for the *median* of Diam|(Age = 5). (Note that it is *not* symmetric about the estimated median $10^{0.726} = 5.32$.)
- Math Box (pp. 691 – 692): This seems like a pretty good semi-rigorous/semi-intuitive explanation of the standard errors for predicting means and individual values, but there a couple of points that could be given a little more explanation; namely:
  - "… the slope, b_1 and mean, $\bar{y}$, should be independent …": To see this intuitively, suppose b_1 is the least squares slope calculated from the data set (x_i, y_i). Now consider the data set $(x_i, y_i + c)$, where c is a constant. If we graph the two data sets, the second will just be the first shifted upward c units. So the least squares slope will be the same, and the mean of the y's for the second data set will just be c + (the mean for the first data set). So we can alter the mean while keeping the slope constant, which intuitively suggests the mean and slope are independent.
  - Last line: The reasoning is that the error term, e, is independent of both b_1 and $\bar{y}$; this follows from the model assumption of independence of errors and the fact (mentioned at the top of p. 2 of the handout *Outlines of Proofs of Formulas on pp. 681- 682*) that b_1 can be written as a linear combination of the sampled values $y_1, y_2, \ldots, y_n$
- After p. 692: See the supplement More Regression Tests -- F-test and Test for Zero Population

  - You should probably talk about the F-test, since it is often used and (at least in the past) was requested for Actuarial students. It also gives a little introduction to ideas used in Analysis of Variance and Multiple Regression.
  - Consider the Test for Zero Population to be optional. The same end can be achieved by the standard t-test for β_1, so the test is only useful if for some bizarre reason you know r^2 but don't have the data.
- pp. 693 – 696: Omit logistic regression (unless you have time to spare).
- pp. 696 – 697 (What Can Go Wrong)
  - Some of this just reiterates what was said in Chapter 9.
  - The second paragraph under "Watch out for the plot thickening" is either wrong or poorly stated. It may be that this is assuming that the plot is increasing in height as x increases.
  - Also, the statement, "we can predict y precisely" might be misinterpreted: "precisely" here does *not* mean "exactly", but "with small margin of error."
  - Emphasize "Watch out for extrapolation" and the discussion following it.
  - "Watch out for one-tailed tests" makes a good point.
- Recommend that they read the Connections section, to help get the big picture of themes-with-variations that runs through statistics.
- pp. 699 – 700: Have them read the general part of Regression Analysis on the Computer.
- Exercises (pp. 702 – 712)
  - Suggestions to select from for self-check:
    - #1 A good exercise for dealing with possible confusion when "error" is part of the context of a problem: the student needs to keep the context use and the regression use straight.
    - #3 plus #5. 3(b) is especially good. Note the wording in 3(c).
    - #7, 9, and 11. (Although #7 says "assuming that the data satisfy the conditions …", these exercises ask some good questions.)
    - #13 plus 15
    - #17, 19, and 21
    - #23 plus 25
    - #29 plus 31
    - #33 Generally good, but part (a) is particularly nice.
    - #37 Part (c) is particularly nice. But be aware that if you also assign #38 as well, you should deal with multiple testing
    - #39
    - #41 Part (c) is particularly good.
  - Possibilities to choose from for handing in:
    - #2 Even though part (c) says, "Assuming the conditions for inference are satisfied," part (e) is good.
    - #4 plus 6 But note the misprint: the units for size are 1000 ft^2, not \$1000's ft^2. Part (4b) is especially good.
    - I wouldn't give #8 alone (because of the "Assuming …), but #8, 10, 12 together might be reasonable. (But see comment on #18, 20, 22 below)
    - #14, 16, 18 together? Or perhaps just #14 (since it shows that there is some question about the model assumptions.)

        - #18, 20, 22 Of these last three combinations, this would probably be my first choice – but I might also assign #8. Also, another good question to ask in #18: What might account for the outlier? (A nice question to tie context with analysis.)
        - #24
        - #30 plus 32
        - #34 part (a) is especially good. But note that the answer to (c) is not complete!
        - #38 Part (c) is particularly nice – but be aware that if you assign both #37 and #38, you need to deal with multiple inference!
        - #40
    - Additional possibilities for exercises:
        - The exercise suggested above re p. 689: filling in the details of Footnote 7
        - Exercises posing a question, where students have to decide whether a confidence interval for conditional mean or a prediction interval is needed.