Chapter 1 Linear Regression

Paul E. Johnson

January 25, 2016

This is a nice survey of regression for students who are already familiar with regression.

1 Highlights

1. p. 12: population "infinite population from which the data can be be viewed as sampled or to the statistical model that is viewed as the data-generating mechnanism..."

   Vital to disconnect ordinary word "population" from this infinite "super-population" way of thinking.

2. p. 13 OLS definition. As a conditional probability model, could say "y given x is Normal"

$$y_i|x_i \sim N(\mu_x, \sigma_x^2) \tag{1}$$

   Researcher has to supply a model to predict the mean and variance. In the Normal, this is direct and easy to see:

$$\mu_x \equiv E[y_i|x_i] \tag{2}$$

$$\sigma_{x_i}^2 \equiv Var(y|x_i) \tag{3}$$

   Note: $\equiv$ means "is defined as".

   a) Terminology: y_i = "response variable" (dependent, criterion) or endogenous: "determined inside"

   b) x_i= "explanatory variable or covariate" or "independent" or "predictor" or "regressor".

3. p. 15. t-test. Given data for 2 groups, with hypothesized μ_0 and μ_1, does the data lead us to suspect that $\mu_0 \neq \mu_1$?

   a) Variance assumption issue

      It was common to assume the variance within each group is the same, that $\sigma_0^2 = \sigma_1^2$. If we assume that, we can calculate sample variance using whole sample as if they are equal.

In R, in about 2000, they dropped the common variance default and started using Welch's correction. Stata allows same with "unequal" option (p. 16)

b) Null hypothesis

$$\begin{aligned} H_0 &= \mu_0 = \mu_1 && (4) \\ &\Rightarrow \mu_o - \mu_1 = 0 && (5) \end{aligned}$$

c) Notational wrinkle

In my notes on the t-test, I write this

$$\hat{t} = \frac{\hat{\mu}_0 - \hat{\mu}_1}{std.err.(\hat{\mu}_0 - \hat{\mu}_1)} = \frac{\hat{\mu}_0 - \hat{\mu}_1}{\widehat{std.dev.}(\hat{\mu}_0 - \hat{\mu}_1)} \tag{6}$$

That is to say, the denominator is an "estimate of the standard deviation" of a quantity, and, in my opinion, the standard error is an *estimate of the standard deviation*. Hence, I do not have a hat on *std.err.* because it is, by definition, an estimate!

In contrast, this book puts a hat on the standard error, which I cannot quite understand, and they don't put a hat on t , which seems wrong to me. See p. 16

$$t = \frac{\hat{\mu}_0 - \hat{\mu}_1}{\widehat{\mathrm{SE}}(\hat{\mu}_0 - \hat{\mu}_1)} \tag{7}$$

The Wald-type "95% confidence interval"

d) p. 17. Assumption that y is normal is needed for a t-test, except if the sample is large enough to invoke the Central Limit Theorem. That's what they mean about "large samples."

e) A regression of same will lead to identical conclusions. Seems to me the t-test was introduced here with the idea that students are likely to be familiar with it.

4. p. 17 One-Way Anova

a) My advisors said to me in 1981, ANOVA is not useful anymore, just get good at doing regression with dummy variables. As a result, I do not have patience to exert much energy on this.

b) p. 17. The partitioning of the sum of squares (variance decomposition)

$$TSS = MSS + SSE \tag{8}$$

c) The "between group" sum of squares is the difference between the *center* of the group and the whole data collection. ANOVA is about finding whether the group's center is far from the whole data center. Of course, quantifying "far" is the problem and that's mostly what ANOVA is about. Is the gap from group-to-whole data large compared to the individual level noise observed within each group.

d) p. 19. F test

i. The null hypo is that the mean within each group is the same

$$H_0 = \mu_0 = \mu_1 = \mu_2 \ldots = \mu_g \tag{9}$$

$$F = \frac{MMS}{MSE} \tag{10}$$

i. with degrees of freedom $g - 1$ and $n - g$ (n = number of rows in data, g is number of groups).

MMS summarizes the difference of group means from the whole data center value.

MSE is "within group variance", it is variation left over after taking into account the group center position.

e) Why did my advisors say ANOVA is not worth studying?

i. Cannot include numeric predictors

ii. Rejecting null is uninformative. Is group 2 different from group 3? Maybe, we need a follow-up test to find out. Why not learn to do those followup t or F tests with regression? They can co-exist with continuous predictors.

5. Remember the laws of logarithyms, where I write 'log' to mean log_e or ln, the "natural logarithm".

- $log(xyz) = log(x) + log(y) + log(z)$
- $log(z - y) = log(z) - log(y)$
- $log(e^x) = x$

6. p. 19 Linear Regression

a) Remember conditional y in equation (1). That is restated on p. 20, except with assumption that variance of y_i conditional on x_i is same for all i

$$y_i|x_i \sim N(\mu_{x_i}, \sigma^2) \tag{11}$$

b) p. 20. Now you supply a more interesting prediction formula

$$\mu_{x_i} = E[y_i|x_i] = \beta_1 + \beta_2 x_i \tag{12}$$

c) Notation difference. In my notes I always write $\beta_0 + \beta_1 x_i$. That's used more broadly in the literature, IMHO.

d) It is mathematically equivalent to write

$$y_i = \beta_1 + \beta_2 x_i + \epsilon_i \tag{13}$$

Thats the "fixed" or "systematic" part plus the "stochastic" or "random" part.

RHS uses "epsilon" ϵ for random error that is centered on 0 , as in

$$\epsilon_i|x_i \sim N(0, \sigma^2) \tag{14}$$

e) Recall the law of variance calculations. For some constants k_j and variables u and v

$$Var(k_1 + k_2 v_i + k_3 u_i) = k_2^2 Var(v_i) + k_3^2 Var(u_i) + 2k_2 k_3 Cov(u_i, v_i) \tag{15}$$

That's always correct, whether variables are normal or not. This is written out in my probability notes for the regression course.

i. The variance of a constant k_1 is 0, $Var(k_i) = 0$

ii. The variance of a constant times a variable is the constant squared time variance, $Var(k_2 v_i) = k_2^2 Var(v_i)$.

iii. The variance of a sum is the variance of each individual term plus the 2 times the covariance with the constants multiplied together.

f) p. 20. For a given value of x_i, then x_i is treated as a constant. Thus they claim toward the bottom of the page:

$$Var(y_i|x_i) = Var(\epsilon_i|x_i) = \sigma^2 \tag{16}$$

That is derived keeping x_i as fixed, as though it is a constant. So β_1 and $\beta_2 x_i$ are seen as constants applying rules above, so

$$Var(\beta_1 + \beta_2 x_i + \epsilon_i|x_i) = Var(\epsilon_i) = \sigma_\epsilon^2 \tag{17}$$

g) Predicted value

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i \tag{18}$$

7. Ordinary Least Squares versus Maximum Likelihood

a) OLS derives estimate as minimizer of

$$Sum\,of\,Squared\,Errors(\hat{\beta}) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_1 + \hat{\beta}_2 x_i)^2 \tag{19}$$

b) The estimate of the variance of the error term is usually taken as

$$\widehat{\sigma_\epsilon^2}^{OLS} = Mean\,Squared\,Error = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} \tag{20}$$

c) p. 21. Maximum Likelihood derives the estimators from different premise, MLE: choose parameter estimates for σ^2 and β_j that make the sample *most likely* to have been produced by the estimated model you propose.

i. Likelihood is the PRODUCT of probabilities that your model produced the data.

ii. If you say $y_i|x_i$ is Normal, then you know the probability of observing one particular case is something like

$$\frac{1}{some\,constant} e^{-\frac{1}{2}\left(\frac{y_i - \hat{y}_i}{\sigma}\right)^2} \tag{21}$$

iii. If you multiply together a lot of terms like that, you end up with a very unstable calculation.

iv. So log that and maximize the *Log Likelihood.*

A. With those laws in mind, the big product of individual probabilities boils down to a sum

$$\frac{1}{some\,constant}e^{-\frac{1}{2}\left(\frac{y_1-\hat{y}_1}{\sigma}\right)^2} \times \frac{1}{some\,constant}e^{-\frac{1}{2}\left(\frac{y_2-\hat{y}_2}{\sigma}\right)^2} \ldots (n\,terms) \tag{22}$$

B. and the log of that boils down to

$$\sum n \times \frac{1}{some\,constant} + \sum -\frac{1}{2}\left(\frac{y_i-\hat{y}_i}{\sigma}\right)^2 \tag{23}$$

C. And when you throw out the constants, you see the numerator boils down to the same objective as OLS.

d) Compared to OLS

i. The MLE slope estimates are identical

ii. The estimate of the variance of the error term from MLE is different. It is known to be biased because it does not subtract from the denominator

$$\widehat{\sigma_\epsilon^2}^{MLE} = \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n} \tag{24}$$

iii. It turns out that this difference b/t MLE variance estimates and other "unbiased" ones is an enduring, troublesome issue that shapes a huge investment in research, debate, and vitriolic dispute about ML versus REML that we'll see later.

iv. Note: it is possible to derive OLS estimates without assuming ϵ is Normal. It is NOT POSSIBLE to derive MLE without giving a particular distribution to $y_i$or ϵ_i.

8. p. 22. Coefficient of determination, R^2, is "not a measure of model fit" and is "not a meaure of the magnitude of the effect of x_i (or effect size)".

9. p. 25. Mean Centering a numeric predictor. When does it help? God, what a mess this has become.

a) Scaling is not helpful and does not matter generally. Look down to item 7 in this lecture

`http://pj.freefaculty.org/guides/stat/Regression/ElementaryOLS/Regression-2-lec`
`pdf`

b) Mean-Centering is thought to be helpful by people who don't understand much about regression

`http://pj.freefaculty.org/guides/stat/Regression-Nonlinear/Interaction-Continuo`
`Interaction-Continuous-1-lecture.pdf`

c) In the vignette for rockchalk, this is all beaten to death, over and over:

`http://pj.freefaculty.org/R/rockchalk.pdf`

d) I don't mind if you center or subtract something from a predictor to bring the y-axis into view in a plot. I do object if you think this is somehow making a regression "better".

10. p. 25. Don't use standardized regression coefficients

11. p. 28: SOMETHING I WANT TO POINT OUT

a) model matrix, AKA design matrix. This is the numeric matrix used in calculating regression coefficients.

b) always an intercept, column of 1's

c) If we want a prediction using one "dummy variable" $x_i \in \{male, female\}$. We can choose to estimate coefficients for 2 out of these 3 columns.

$$\begin{bmatrix} "intercept" & x_i = male & x_i = female \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \tag{25}$$

i. Use columns 1 and 2, then the estimate of the intercept is the "baseline" value that applies to all cases and coefficient on second column is "difference from the baseline".

ii. Use columns 1 and 3, then baseline is compared against the group defined by $x_i = 1$

iii. Suppress the intercept? OK, use columns 2 and 3

12. p. 29: "robust" or *sandwich estimator* of standard error. vce(robust)

a) This may deserve more caution, most Stata users seem to place a lot more faith in these things than they ought to.

b) Reasons for caution related to asymptotic assumptions.

13. p. 30: Multiple regression

a) add more predictors

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i \tag{26}$$

In my course notes, I instead write about columns $x1$, $x2$, $x3$ like this

$$y_i = \beta_0 + \beta_1 x1_i + \beta_2 x2_i + \beta_3 x3_i + \varepsilon_i \tag{27}$$

b) "controlling for", "adjusting for", "partialling out", or "keeping constant" other predictors. "Other things equal".

c) p. 31. Nice parallel lines graph! You understand that, you understand this section.

d) p. 32. See elaborate use of twoway. Achieves same result as R rockchalk::plotSlopes.

14. p. 32: Foreshadow "causality" discussion, "matching models" with terminology "overlap". I'll sketch something to explain. I don't understand what we are supposed to conclude from graph on p. 33. Except to note that Stata can make density plots with twoway, which is a good thing.

    a) Randomized Treatment assignment would solve everything, but we don't generally get to assign predictors in quasi-experimental data. (Can't make 435 randomly chosen people into House of Representatives, can we? Or assign membership in Republican party to Representatives).

    b) p. 33. "confounding" p. 34. "spurious" relationship. The effect we think is "Treatment" is really due to an unmeasured confounding variable.

    c) Omitted variable bias versus

    d) Too many variables cause inflated standard errors

        i. Remember: If columns are orthogonal, there is no danger here.

        ii. Only if columns share variance is this a concern ("multicollinearity") and it is a dicey problem.

15. Manual recoding or factor variables?

    a) R uses an automatic "factor variable" framework, some are very enthusiastic about it

        i. We have a variable declared as categorical (factor, ordered) and R manufactures the model matrix that goes with it.

        ii. eliminates the danger that a categorical variable like "gender", $x_i = 0$ or $x_i = 1$, will be mistaken for a numeric thing.

        iii. eliminates the danger/confusion that results when one person wants to use coding 1 and 2 rather than 0 and 1.

        iv. May "automatically" give better labels in output, like $genderMale$ and $genderFemale$.

    b) Other stat packs, incl Stata 11, have inserted ability to use "factors".

16. Interactions

    a) moderator, same as "effect modifier"

    b) p. 37: see make as moderator

2 Stata Commands I circled

1. p. 12: `use http://`
2. p. 13: `graph box` (box plots)
3. p. 14: `histogram y, by(x, rows(), xtitle().`
4. p. 14: `generate` to transform/recode a variable
5. p. 15: `ttest y, by (x)`
6. p. 24: `regress y x`
7. p. 26: `predict yhat, xb` (yhat is the new name for predicted values, will appear in data, whereas is xb is short name)
8. p. 26: `twoway (scatter y x) (line yhat x, sort), ytitle(yname here), xtitle(x name here)`. Don't know why "sort" there.
9. p. 30: `tabstat x1, by (x2) statistics(mean, sd)` produced aggregates for subgroups. Handy!
10. p. 32-33: more elaborate uses of `twoway`
11. p. 39: `lincom`. Can be used with regressions to test hypos about linear predictions involving coefficients. Have not seen this before, looks handy. Additional usage p. 46.
12. p. 39: please DO NOT use twoway with typed-in coefficients as on bottom of page.
13. p. 40: please DO access coefficients with `_b[]` notation. Another good usage of same p. 42.
14. p. 43: `tabulate x, generate(r)` , a way of manually creating the dummy-variable columns in the data frame, will be in design matrix. This is better than crude generate (p. 42), but I'd suggest doing neither. Let the factor facility in Stata handle it.
15. p. 47: `testparm i.x` carries out the F test that all of the dummy columns have 0 slope coefficients.
16. p. 56: Causality and Treatment
    a) This is valuable, but probably out of place
    b) There are other books about checking for causal effects and synthetically balancing treatment and non-treatment subjects.
    c) Claims that econometrics books are more likely to venture into discussion of causality and assumptions needed to achieve it.
    d) "strict exogeneity": treatment and error are uncorrelated is one sufficient assumption.

3 Factor Variables.

1. Software that can handle categorical variables gracefully will notice the scores of gender are $\{male, female, female, male, \ldots\}$ and then manufacture a suitable design matrix for us.

2. If a predictor has g categories, it will manufacture $g - 1$ categories. See:

   "Categorical Predictors" `http://pj.freefaculty.org/guides/stat/Regression/CategoricalPre`

   "Interactions: Continuous" `http://pj.freefaculty.org/guides/stat/Regression-Nonlinear/Interaction-Continuous/Interaction-Continuous-1-lecture.pdf`

   "Interactions: Categorical" `http://pj.freefaculty.org/guides/stat/Regression-Nonlinear/Interaction-Categorical/Interaction-Categorical-1-lecture.pdf`

3. Stata has ways to do same, since version 11

   Consider a variable "gender" coded 1 for male and 2 for female.

   a) `i.gender` If we enter "i.gender" as a predictor, then Stata will try to do the right thing (add the correct columns to the design matrix).

   b) `c.yearsdg`

      i. Generally, it is not needed to do this with numeric predictors, except when they are interacted with categorical predictors.

   c) Rather than manually manufacturing an interaction, as RHS suggest p. 38 with

      `generate male_years = male*yearsdg`

      Instead we add into the regression:

      `i.male##c.yearsdg`

      The `##` usage is quivalent to R's "`*`", where, for example, `x1*x2` would cause the insertion of x1, x2, and x1:x2 (where x1:x2 is the product of x1 and x2).

   d) p. 31 Stata allows # notation, i.x1#c.x2 meaning "product of following terms", but it does not include `i.x1` or `c.x2` in the regression model for us.

      i. I believe using # is bad form and won't do it, note danger in RHS p. 41 that the variable "male" is treated as numeric while "i.male" is treated as a factor. Could cause miscoding if male = 1,2, for example.

      ii. The R equivalent for this would be y ~ x1 + x2 + x1:x2.

   e) p. 46. Determine the baseline category:

      `ib3.rank` or `b3.rank` are equally good? b is the base level.

   f) TODO list: find out how to see the model matrix from a fitted Stata regression.

   g) p. 46: RHS luke warm on factors: "Although factors are very convenient, an advantage of constructing your own dummy variables is that you can give them meaningful names." Seems week to me, especially if categorical variables have value labels.

h) p. 50, 51: have both styles, the old-fashioned “recode your own columns” approach and the Stata factor variables approach.

i) p. 54: Note multiple interactions in ## model formula