Villanova MAT 8406

Regression Analysis Roadmap

November 12, 2015

Introduction

This document is a comprehensive guide to executing and reporting on your regression analysis project. Its components are:

- **Part I:** A review of the techniques you have learned and can immediately apply to analyzing your data.
- **Part II:** An outline of the steps to take in a regression study. What to do, in what order, with some specific tips and cautions.
- **Part III:** Communicating the results. This has two parts. First is a list of things to remember to include (to the extent they are relevant). Second is a model outline for a written report, paper, or presentation.

Although the length may seem formidable, please understand that is because this is meant to cover a gamut of individual projects as well as to serve as a guide in the future. Your project will not involve all the techniques nor will it include all of the items that can be reported. Because one (unusual) purpose of the project is to demonstrate your mastery of this subject and its techniques, you should attempt to apply methods covered in each of the textbook chapters we have worked on in class, to the extent they are applicable. If you believe that *no* technique in a particular chapter is needed, then consider including a brief explanation why not. The explanation should exhibit an understanding of what that technique is and what is good for, along with a clear analysis of why it is *not* applicable to your study.

Evaluation

Your project will be evaluated on how well it achieves the following objectives:

1. Exhibit a practical knowledge of the principal techniques covered in the course.
    - Exercises you have submitted during the course will provide substantial support for this.
2. Perform an appropriate and full analysis of a significant dataset.
3. Provide a correct analysis.
4. Present a well-reasoned interpretation of the results.
5. Produce a clear, organized written report and presentation.
6. Generate informative, appropriate tables and graphics.

Evidence of *creative analysis, additional research,* or use of a *variety of appropriate techniques* will enhance the evaluation.

Part I: What You Have Learned

This list is a review of the introductory presentation from our first meeting. Topics we have not (yet) covered have been removed. Additional topics have been added.

1: Introduction

Purposes of Regression

1. To get a summary of multivariate data.
2. To set aside the effect of a variable that might confuse the issue.
3. Contribute to attempts at causal analysis.
4. Measure the size of an effect.
5. Try to discover a mathematical or empirical law.
6. Prediction.
7. Exclusion: getting x "out of the way" when we want to study the relationship between two other variables that might be affected by x.

---

2. Simple Linear Regression

The *simple linear regression model* is

$$y = f(x;\beta) + \varepsilon = \beta_0 + \beta_1 x + \varepsilon.$$

It is *linear* because $f(x;\cdot)$ is a linear function of the *parameters* $\beta = (\beta_0, \beta_1)$. It can be fit with Ordinary Least Squares (OLS). With it you now know how to

- **Estimate the coefficients** with estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.
- **Estimate the variance** of ε as $\hat{\sigma}^2$.
- **Estimate the value** of $f(x;\beta)$ for any given x as $\hat{y}$.
- **Construct confidence limits and intervals** for $E[y|x]$.
- **Construct *simultaneous* confidence limits** for $\beta_0 + \beta_1 x$.
- **Construct prediction limits** for linear functions of $y|x$.

Principles of Exploratory Data Analysis (EDA)

Above all, *look* at the data.

The Three R's

- **Resistance/robustness.**
  - *Resistance:* Use methods to summarize and analyze the data whose results are not appreciably influenced by unusual or erroneous values.
  - *Robustness:* Use methods that do not rely on any restrictive assumption, such as normality of residuals.
- **Residuals.** Compute and look at residuals to probe more deeply into the data. Do not be satisfied with a summary of conclusion until you have studied how the data depart from that summary.
- **Re-expression.** Take control over how the data are written down (expressed). Do not assume that the numbers in which they were originally recorded are the right numbers for data analysis. Systematically look for re-expressions that *simplify* data description and *yield insight*.

Basic techniques

- Letter statistics.
- N-letter summaries.
- Diagnostic plots: using graphics of *derived information* (such as residuals or spreads) to point the way towards another form of analysis.
- Robust regression: "slicing and dicing" bivariate data (scatterplots) by values of the regressors (the x-values).
- Box-and-whisker plots.
- Scatterplot matrices. (Introduced in chapter 3.)

Communication

You have practiced communicating your results to non-statisticians. The techniques included:

- Providing clear, correct interpretations in non-technical language of the meanings of regression coefficients, predicted values, confidence intervals, prediction intervals and interactions.
- Writing in good English prose (instead of just reproducing software output).
- Identifying and focusing on which elements of your work are of interest and value (perhaps relegating many technical details to appendices).
- Organizing your report.

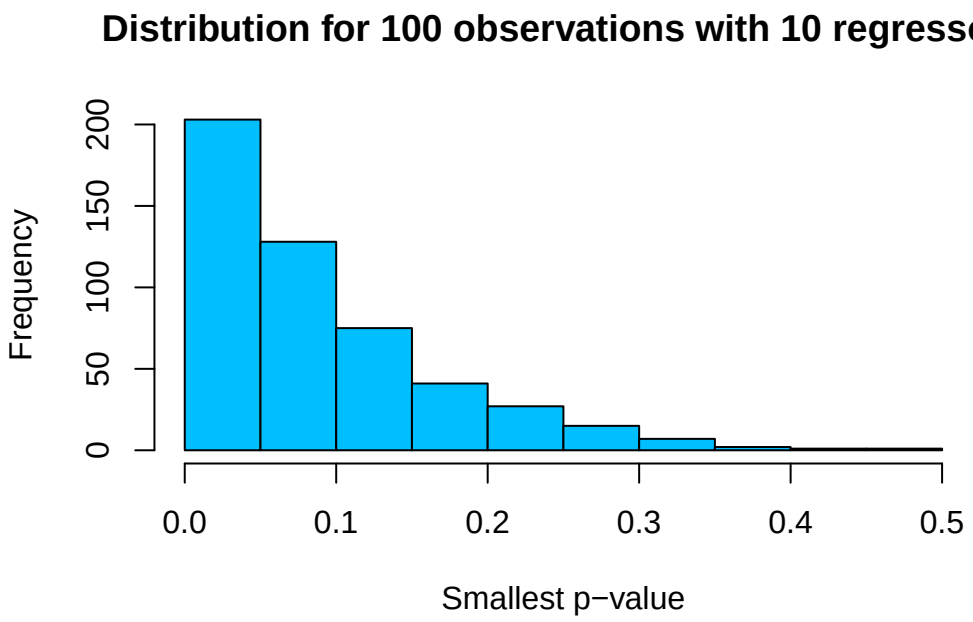
Simulation

Stochastic simulation provides you a lifetime of insight in a short time. You have seen how to use it to

- Enhance your understanding of a procedure you are using.
- Check on your calculations,
- Assess the implications of variation and uncertainty in data.

R in particular makes simulation easy.

```
#
# What is the smallest p-value in a multiple regression where all
# the coefficients are truly zero?
#
n <- 1e2           # Number of observations
k <- 10            # Number of variables
n.sim <- 500       # Number of simulation iterations
sim <- replicate(n.sim, {
  #
  # Generate random data.
  #
  x.df <- as.data.frame(matrix(rnorm(n*k), n))
  y <- rnorm(n)
  #
  # Compute statistics of interest.
  #
  p.values <- coef(summary(lm(y ~ ., x.df)))[, "Pr(>|t|)"]
  min(p.values)
})
#
# Explore and summarize the results.
#
hist(sim, col="DeepSkyBlue",
     xlab="Smallest p-value",
     main=paste("Distribution for", n, "observations with", k, "regressors"))
```




```
cat(format(100 * mean(sim < 0.05), mdigits=2),
    "% of the datasets had at least one 'significant' coefficient.\n")
```

```
# 40.6 % of the datasets had at least one 'significant' coefficient.
```

Note that this is not a routine element of a regression analysis. It is, however, the computational basis of many modern statistical methods, including bootstrapping, permutation tests, and Bayesian analysis.

---

3. Multiple Regression

The *multiple linear regression model* is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon.$$

It can be fit by a sequence of one-variable regressions of the form

$$y_{.*} = \gamma x_{.*}$$

where the variables $y_{.*}$ and $x_{.*}$ are the *residuals* of other regressions.

Rarely is the model completely correct. **You have learned to *detect* how the data appear to be inconsistent with the model, *diagnose* the inconsistency, and *correct* it when possible.**

Techniques

- Generalizations of those for simple linear regression: coefficient estimates, error variance estimates, predictions, confidence intervals, simultaneous confidence bands, and prediction limits.

- Added-variable (partial) F tests.
- VIF to detect and assess collinearity.
- Understanding extrapolation and its problems.
- Reading and interpreting the output of regression software. Knowing the arithmetical relationships among the values (such as how a p-value is obtained from a coefficient estimate and its standard error and how R^2 is related to F.)

---

4. Model Adequacy Checking

Recall that the *residuals* of any model $y = f(x; \beta)$ with estimated coefficients $\hat{\beta}$ are the differences between the observed responses y_i and the fitted responses $f(x_i; \hat{\beta})$,

$$e_i = y_i - f(x_i; \hat{\beta}).$$

An important underlying concept concerns analyzing how model outputs–residuals, coefficients, estimated responses, and more–change as each observation is systematically deleted from the dataset.

You have used this concept to *standardize* the residuals in several different ways: that is, to re-express them as multiples of how much they could be expected to vary.

You have become familiar with making and interpreting various *plots* of these standardized residuals to detect and diagnose problems.

- Residuals vs. predicted values.
- Normal QQ plot of residuals.

---

5. Transformations and Weighting

Two common, correctable ways in which a model may be wrong are

- The variances of the errors ε can change appreciably in association with one or more variables (*heteroscedasticity*).
- The underlying relation betwen y and the x_k may be monotonic but *non-linear.*

These are best diagnosed with suitable plots, usually involving residuals:

- The spread-vs-level plot.
- Diagnostic plots for linearity.
- Guessing and checking!

Although many of you have found software that automatically finds re-expressions, you have also discovered that (a) it doesn't always work and (b) does not afford any insight into *why* it works or what it accomplishes. *Use the EDA methods when you can.*

You have learned how to correct these problems in various ways, including re-expressing y and/or one or more of the x_k in a nonlinear way.

(The text discusses using weighted least squares but we have not covered that in class or the exercises.)

---

6. Diagnostics for Leverage and Influence

A **leverage** point is a datum $(\mathbf{x}, y) = (x_1, x_2, \ldots, x_k, y)$ that has an unusual $\mathbf{x}$ value "and may control certain properties" of the fit.

An **influence** point is a leverage point with a "noticeable impact on the model coefficients."

You have seen how these can be detected and measured systematically by deleting each point from the dataset, refitting the model, and comparing the results. (As always, the comparisons are made relative to the amount of change *expected* to occur for non-leverage points: that is, they use *standardized* statistics.) This is done with efficient algorithms–you need only learn how to use them and interpret the results.

- You have seen how a simple graphical display of these diagnostics (DFBETAS, DFFITS, COVRATIO, Cook's distance, hat matrix diagonal) allows for almost immediate and comprehensive identification of leverage and influential points.
- Nobody really knows what to *do* about such data. You have to use judgment and, often, perform analyses both with and without them to see how much they matter.

---

8. Indicator Variables

Regression models can handle discrete regressors x_j–"factors"–by encoding them numerically using *indicator variables.* Such variables have values of just 0 or 1. They are often called *dummy variables.*

This mathematical unification creates both a computational and conceptual simplification.

- You have learned how to interpret regression coefficients of indicator variables.

---

9. Collinearity

Sometimes–especially in economic and social datasets–collections of variables provide nearly-redundant information. Regression procedures are then trying to estimate many coefficients to describe just one quantity. They break down–or, in the worst cases, they *seem* to work but give results that will be misinterpreted.

- You have learned some methods to detect these problems (such as computing the VIF).
- Soon you will learn methods to avoid and deal with these problems.

---

10. Variable Selection

When you have many variables to choose from, using too many creates problems ranging from collinearity (*qv*) to overfitting.

- There is a rich collection of *ad hoc* methods to help you choose: stepwise regression (forward and backward), all-subsets regression, AIC and BIC comparisons, etc.
  - Much of it is black art.
  - Carried too far, these methods can result in extreme instances of "data snooping," with disastrous (but often hidden) consequences.
- Currently the preference appears to be to apply some form of *cross-validation* technique.
- How you cope with this depends partly on the purpose of the regression (see the first two slides).

Part II: An Outline of a Regression Study

What would a complete regression study include? In what order would the work be conducted? What would a full report look like?

Part I shows what *you* are now able to do.

Let's address the second and third questions.

Order of Work

This outline parallels, and expands on, the specific work you were asked to do to demonstrate progress in your project.

Please understand that the techniques discussed here are *minimal* in that they are limited to what is addressed in the portions of the text you have studied. At various junctures additional (better, more powerful, or more defensible) methods have been mentioned, including bootstrapping and cross-validation. The extent to which you use such methods will depend not only on their applicability to your project but also on your awareness of them, ability to learn how to use them well, availability of software, and the amount of time available. **The methods you have already learned will carry you a long way and are sufficient to complete your project successfully.**

1. **Develop an interest or idea.**
   - Find something you want to study.
   - Read about it, discuss it, explore it, learn what you can about the subject.
   - Formulate an investigation question.
   - Decide what you need to accomplish: Estimate coefficients? Predict something? Control a process?
2. **Collect data.**
   - If possible, find original data sources. Evaluate their quality, completeness, and cost to obtain. ("Cost" can be any combination of price, your effort, and your time.)
   - Determine what level of aggregation of observations you need to study (or can study). Remember the Ecological Fallacy!
   - Assemble the data into a usable, organized format.

- Document your data. Create a data dictionary explaining what the variables mean.
- Unless you have a definite, unmodifiable set of hypotheses to check, *hold out a randomly chosen testing subset of the data.* Don't look at these until the end! Typical amounts of holdout are 10% to 50%. The data you will analyze next are the "training data."

3. **Explore the (training) data.**
   - Start with *univariate* exploration. Look at plots of distributions of individual variables. Summarize them. Identify unusual values and fix them if they are erroneous.
   - Look for patterns of *missingness* in the data to see whether
     a. They might preclude further analysis or
     b. They might bias the results.
   - Move to *bivariate* exploration, such as scatterplots, scatterplot matrices, and side-by-side boxplots, to study relationships among pairs of variables.
   - For *multivariate* exploration, begin slicing data along one or more regressors and performing univariate and bivariate exporation of the remaining variables within each slice.
4. **Create an initial model.**
   - Write down the model abstractly, as in $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$, $\varepsilon \sim \text{NID}(\mu, \sigma)$.
   - Find software that can fit the *full* model and apply it.
   - Review and interpret the output. Compare it to your investigation questions (#1).
   - Review and interpret diagnostic plots to check for violations of the model assumptions.

From this point on, what you do is iterative, creative, and open-ended. It will routinely include some of the following:

5. **Re-express the response variable(s).**
   - Use spread-vs-level plots and other techniques to achieve constant variance, if possible.
6. **Re-express the regressors.**
   - Do this to make their relationships with the response more linear *after controlling for all the other regressors.*
   - Added-variable plots can be helpful.
   - Often, experimentation guided by experience is good enough.
7. **Select an appropriate subset of the regressors.**
   - You have learned many techniques that *might* work: all-subsets regression, stepwise regression, backwards elimination; based on F-values, p-values, AIC, BIC, Mallows' C_p, *etc.*
   - Keep any variables that must be included in the final model for theoretical reasons.
8. **Check whether any two-way interactions may be important.**
   - This amounts to adding them as additional variables and re-performing the variable selection step #7.
9. **Assess the effects of collinearity.**
   - This is most important for estimating regression coefficients in small datasets.
   - When interactions are introduced, the potential for collinearity becomes high.
   - The result of this step is either to discard some variables (as effectively redundant) or to create new "composite" variables using combinations of two or more original variables.

At each of the steps 5 - 8, you may need to turn back to an earlier step and start over from there. You will always be on the lookout for outliers–they can show up any time–and will be considering the effects of retaining them, excluding them, or downweighting them.

10. **Fit a tentative "final" model.**
    - Redo all the checks:
      - Re-evaluate how all variables are expressed;
      - Check the residuals;
      - Test for linearity;
      - Evaluate goodness of fit.
    - Interpret the results.
      - Do they make sense?
      - Are there surprises? If so, can you explain them?
      - Think how you would communicate them.
      - Study summaries of the model quality, such as R^2, PRESS, and the distribution of the residuals.
11. **Check your model against the test dataset.**
    - Apply your model to fit the responses in the test data to the regressors.
    - Plot the predicted against the actual responses.
    - Compute measures of the fit, such as the R^2 of this plot, or summaries of the differences between the actual and predicted values.
    - Compare the summaries to what your tentative results led you to expect.
12. **Fit a final model.**
    - Unless there are large, important differences between the performance on the training and test sets (which suggest you made a huge mistake somewhere, such as overfitting the training data), *combine the training and testing data.* Refit the model on the combined dataset.

Part III: Communicating the Results

You can find many guides to writing a report of a regression analysis. Here, as a point of departure for discussion, is an outline provided by Lehana Thabane, Director of the Biostatistics Unit at McMaster University (Canada). Because it is oriented towards a particular kind of study (biostatistical), some of the recommendations might not apply and others will be questionable in particular cases, but it provides a good idea of the kinds of things people look for in such reports. Thabana usefully follows this outline by lists of common errors and additional remarks.

Reporting the Methods

- State how the sample size was determined (if regression was a primary method of analysis).
- Identify the variables and summarize them descriptively.
- Specify how the (explanatory) variables that appear in final model were selected.
- Provide test of the model goodness-of-fit and methods assessing model assumptions.
  - Specify whether explanatory variables were tested for interaction effects.
  - Specify whether all potential variables were tested for collinearity and how it was handled.
  - Normality/constant variance assessments.
- Specify how model was validated.
- Specify how
  - outliers were handled (Diagnostics: Boxplots of residuals).

  - influential observations are handled (Diagnostics: Cook's statistic).
  - missing data are handled.
- Specify how results are summarized:
  - Coefficient, standard error, 95% CI and associated p-value.
  - Use OR (odds ratio), 95% CI and p-value for logistic regression.
  - P-values reported to 3 decimal places.
- Specify any planned sensitivity analyses.

Reporting the Results

- Report coefficient, standard error, 95% CI and associated p-value.
- Report results of goodness-of-fit assessments.
  - Provide the coefficient of determination, R^2. This provides the amount of variability in the response accounted for by the explanatory variables included in the model.
  - Provide LR (likelihood ratio) statistic, degrees of freedom and associated p-value.
- Report results of model assumptions.
  - Qqplots
  - Residual plots
  - Collinearity statistics (VIF)
- Report results of sensitivity analysis
  - Methods of handling missing data
  - Different methods of analysis (different assumptions)
  - Outliers (analysis with and without outliers)
  - Different definitions of outcomes (*e.g.*, different cut-off points for binary outcomes).
  - Any twists based on variations in assumptions.

Common Errors in Regression Analyses

- Multivariable versus multivariate analysis
  - Multivariable/multiple regression: single dependent/outcome variable with multiple independent variables/predictors.
  - Multivariate: multiple dependent/outcome variables with single or multiple independent variables/predictors.
- No reporting of the assessment of model assumptions.
- Poor reporting of the methods used to select predictor variables for inclusion in multivariable analysis.
- No clearly stated hypotheses and justification.
- Poor reporting of the results.
- Limited Scope
  - The model may be applicable only to the range for which the data were available.
  - Cover a wide spectrum of the data
- Form of the relationship
  - A statistically significant linear relationship does not necessarily mean that the relationship is a straight line.
- It is important to have a clear hypothesis and justification for it.

- Confounding
  - Undefined confounding variables may create the illusion of the existence of a relationship or mask it.
- Check whether uncontrolled variables are accounted for.
- Inadequacy
  - Goodness-of-fit is not the same as prediction.
  - A model with very good fit may not do well in predicting unobserved responses... Why?
- The data used to develop the model may have shown a spurious relationship.
  - Review the literature to make sure that the model is plausible and has causal basis (biological hypothesis).
- The relationship may genuine, but the data/sample was not representative of the target population.
  - Use proper probability sampling techniques.
- The relationship may have changed over time:
  - It is important to check that the relationship remains unchanged during data collection and in the near future

Other Important Remarks

- Always use a scatter plot to display relationship among variables before you model the relationship.
- Think about the goal of modeling prior to "turning on the computer."
- Is the goal to determine the cause-and-effect mechanism? (Standard regression techniques won't be helpful!)
- Is it to derive a formula for prediction?
- Pearson's correlation coefficient is important for bivariate Normal variables.
- Correlation coefficients without scatter plots can be misleading.
  - Always report correlation coefficients along with scatter plots.

A Model Outline

A standard way to organize a scientific or statistical paper proceeds like this.

1. **Introduction.** State the investigation objective. Provide a brief description of the contents of the paper.
2. **Background.** Introduce theory and terminology. Motivate the investigation questions. Describe previous work (by you and others). Explain how this analysis builds on, extends, or modifies that work.
3. **Data description.** Explain how the data were obtained, processed, and selected for analysis. Describe all variables: their meanings, units of measurement, degrees of precision and accuracy of measurement. Provide summaries (in the form of tables, statistics, and/or grpahics) of the data. (Most such summaries tend to be univariate.) Identify outlying or unusual data, explaining why they are considered outliers and how they were handled in the analysis.

4. **Statistical methods.** Describe the statistical methods applied. Explicitly state the underlying models and assumptions. Reference the software procedures used. For unusual or novel methods, provide the mathematical background and justification. Explain any conventions used within the report, such as what a "±" value represents (a standard error or a 95% confidence interval, for instance), whether (and how) corrections for multiple comparisons we applied, *etc.*

5. **Results.** Present the *key* results, using narrative, tables, and figures. Relegate supporting information (such as diagnostic tests and plots, alternative analyses that were performed, *etc*) to appendices.

6. **Discussion.** Interpret the results. Use them to answer the investigation questions. Connect these results to those obtained by other investigators. State new questions that are raised by these results. Identify limitations in the data and analysis, explaining what their potential effects are and how they might be overcome.

7. **Conclusions.** Summarize what was done. State the principal conclusions.

8. **References.**

9. **Appendices** Provide extended tables and additional figures. Make sure they are all clearly labeled and titled. Provide a caption with each explaining what information or data it is presenting. For a graphic, also explain *how* it represents the data. State the principal conclusions that can be drawn from it Describe the specific aspects of the graphic that lead to those conclusions.

The length of such a document may vary from two pages to many thousands of pages. A report in the scientific or statistical literature will usually be between 6 and 50 pages (in a word processing document), not including the appendices, and often is around 25 pages or so (which is about 10 pages when it appears in print).