2021 Advanced Regression Methods for Independent Data

BIOSTAT/STAT 570

Jon Wakefield

Departments of Statistics and Biostatistics
University of Washington
jonno@uw.edu

Chapter 1: Introduction and Motivating Examples

Steps in a Data Analysis

1. Establish the context of the analysis. This includes understanding the data collection procedure, the population sampled and to whom the subsequent inference applies, the background science, and the aims of the analysis.

   On the basis of these considerations, and in particular the scientific background, a model can be formulated.

2. The statistical properties of the combined design/model/estimation procedure should be examined to see if inference is reliable.

3. Once a scientifically reasonable, and mathematically satisfactory model and inference strategy is decided upon, the computational aspects can be considered.

Range of Models

We must consider the deterministic and stochastic parts of the model – I like to think about generative models, that is, models from which one could simulate the micro data (i.e., not just summaries).

Important choices:

- Form of regression model (e.g., the mean model).
- The variance-covariance structure.

Historically, statistical modeling was restricted by the ability to calculate inferential summaries – now we can fit all sorts of weird and wonderful models – what are the implications for consistency/coverage/etc?

For linear models most useful quantities are in closed form.

When Nelder and Wedderburn (1972) introduced Generalized Linear Models (GLMs) – convenience of computation via Iteratively Reweighted Least Squares (IRLS) was stressed.

Range of Models

Now computation much less of a problem – we can now compute estimators from very general model classes, and can simulate standard errors/sampling distributions using techniques such as the Jackknife or the bootstrap (if we are in the frequentist realm).

Not all models are robust to misspecification though, and GLMs have desirable properties in this respect.

There are many model choices!!!

We can be parametric, semi-parametric, or non-parametric[1].

[1] And the latter two terms in particular do not have consistent definitions

Issues in the Analysis

Determine whether the data are observational or experimental in nature:

- In an experimental study units are randomly assigned to exposure (e.g., treatment); in this case, if successfully implemented, any differences in response will (in expectation) be due only to treatment, allowing some hope of causal explanations.

- In an observational study we never know whether observed differences are due to another variable (observed or unobserved) that is related to the observed exposure.

Issues in the Analysis

Determine exactly how sampling was carried out, and from what population the data were collected.

- The latter is vital if we want to understand to whom the conclusions apply.
- The data collection procedure has implications for the analysis, including the models fitted.
- For example, case-control studies in which a binary outcome variable of interest is fixed by design, and the exposures are the random variables, are most easily analyzed using logistic regression models (Chapter 7).

Issues in the Analysis

Carry out explanatory data analysis (EDA):

- Examine univariate and bivariate summaries (and present results in a clear manner!). In particular the data should be checked for errors (are values within correct ranges?).
- Are there outlying/influential observations?
- By influential we mean observations that when perturbed lead to large changes in the inference.

Missing Data

Determine whether any variables were not available, that is consider missing data:

- It is often not safe to ignore such information since the missingness may depend, for example, on the size of the response that would have been observed.
- An extreme example is when the result of an assay is reported as “below the lower limit of detection”.
- Such variables may be stated as this lower limit, and analyzing these data using these values can again lead to large bias (unfortunately probably insufficient time to consider missing data in this course).

Aims of an analysis

Always bear in mind the aim of the analysis.

In general, we may be interested in:

- **Description**.
- **Exploration** (e.g. model formulation, hypothesis generation).
- **Confirmation of a hypothesis/“inferential” analyses.**
- **Prediction**.

There are different ways of slicing our aims...

And we will focus on inference for parameters.

Inferential Procedures

The three main inferential approaches are:

- **Estimating Functions:** motivated through frequentist asymptotic properties, implementation requires maximization/root finding.
- **Likelihood**[2]: motivated through frequentist asymptotic properties, implementation requires maximization.
- **Bayesian:** motivated through decision theory, implementation requires integration (which is often sidestepped through simulation).

Number of assumptions required are different, with the Bayesian approach requiring both a likelihood and a prior.

This additional level of assumptions leads to a greater flexibility in the complexity of questions that may be asked, however.

[2] Not pure likelihood, see Royall (1997)

Where does randomness arise from?

Let's play a hypothetical game!

We begin with a very simple deterministic model for variables observed over time t:

$$y_t = \beta_0^\star + x_t\beta_1^\star + z_t\gamma.$$

Now suppose we only measure y_t, x_t and assume the model

$$Y_t = \mathrm{E}[\,Y_t \mid x_t] + \epsilon_t = \beta_0 + x_t\beta_1 + \epsilon_t.$$

What do the errors ϵ_t represent?

We can always write the unobserved z_t as a linear function of x_t plus 'error' δ_t:

$$z_t = a + bx_t + \delta_t. \tag{1}$$

Where does randomness arise from?

Substitution of z_t gives:

$$\begin{aligned} y_t &= \beta_0^\star + x_t\beta_1^\star + \gamma(a + bx_t + \delta_t) \\ &= \beta_0 + x_t\beta_1 + \epsilon_t \end{aligned}$$

where

$$\begin{aligned} \beta_0 &= \beta_0^\star + a\gamma \\ \beta_1 &= \beta_1^\star + b\gamma \\ \epsilon_t &= \gamma\delta_t. \end{aligned}$$

Hence, β_1 is a combination of:

- the direct effect of x_t on y_t, and
- the effect of z_t, through the linear association between z_t and x_t.

This development illustrates the problems in non-randomized situations of estimating the causal effect of x_t on y_t, that is $\beta_1^\star$.

Where does randomness arise from?

Remember:

$$\begin{aligned} y_t &= \beta_0 + x_t\beta_1 + \epsilon_t \\ \epsilon_t &= \gamma\delta_t. \end{aligned}$$

Turning to the stochastic component we see that properties of ϵ_t are inherited from δ_t.

Hence, assumptions such as constancy of variance of ϵ_t depend on the nature of z_t and, in particular, on the deviation of z_t from linearity.

An idealized deterministic model for a continuous outcome

A more complex model:

$$y = \beta_0^\star + \sum_{j=1}^{p} x_j \beta_j^* + \sum_{k=1}^{q} z_k \gamma_k.$$

Suppose we observe $\boldsymbol{x}$ and Y but not $\boldsymbol{z}$, so Y is now random since $\boldsymbol{z}$ is unknown.

Then

$$\mathsf{E}[Y|\boldsymbol{x}] = \beta_0^\star + \sum_{j=1}^{p} x_j \beta_j^* + \sum_{k=1}^{q} \mathsf{E}[Z_k|\boldsymbol{x}]\gamma_k.$$

An idealized deterministic model for a continuous outcome

If we assume the model

$$Y = \mathsf{E}[Y|\boldsymbol{x}] + \epsilon,$$

then

$$\mathsf{E}[Y|\boldsymbol{x}] = \beta_0 + \sum_{j=1}^{p} x_j\beta_j,$$

so the β_j's again reflect associations between z's and x's.

And the error terms depend on the part of the deterministic model that is not linearly related to $\boldsymbol{x}$.

What are we trying to do?

A principal aim of regression modeling is to “explain” the error using observed covariates.

In general, error terms are representing:

- Unmeasured variables (so this leads to dependent error terms when we have unmeasured variables that are common to different observations, e.g. families, spatial areas).
- Data anomalies (such as inaccurate recording of responses and covariates).
- Measurement error.
- Model misspecification.

What's the distribution of the errors (and should we care)?

Clearly the nature of the randomness, and the probabilities we attach to different events, are conditional upon the information that we have available, and specifically the variables we measure.

- The obvious candidate for the distribution of the error terms is the normal distribution (central limit theorem).
- Lots of other possibilities though: Student's t, Laplacian, Pearson distributions (introduce skewness and kurtosis). Aside: Do we need/want to assume a distribution for the error term?
- Some distributions arise "naturally", for example, the normal, Bernoulli and Poisson, while others are "contrived", for example, Student's t, chi-squared.

Some estimating function approaches do not require the distribution of the data to be specified.

Idealized deterministic model for a binary discrete outcome

Underlying latent trait:

$$w = \alpha^* + x\beta^* + z\gamma.$$

If $w \geqslant w_0$ then appears as $y = 1$, otherwise $y = 0$.

Example 1: $y =$ low birth weight/not low birth weight, $w =$ birth weight, $x, z =$ variables determining weight.

Suppose we don't observe w just the outcome Y:

$$p = \Pr(Y = 1) = \Pr(W \geqslant w_0) = \mathrm{E}[Y = 1].$$

Meaning of p? Limiting frequency of event of interest in population under study.

Finite versus Infinite Samples

Are we interested in:

- the population we actually sample specifically (in which case we may see all of the individuals and no statistics is required!), or
- are we interested in extrapolation to some **superpopulation**, an infinite population of which the population of size *N* from which we sample *n* is assumed to be drawn.

$$\begin{array}{ccccc} \text{Super-population} & \rightarrow & \text{Study Population} & \rightarrow & \text{Sample} \\ \infty & \rightarrow & N & \rightarrow & n \end{array}$$

In this course we will often be interested in the super-population.

If we are interested in finite population characteristics, then survey sampling techniques are relevant.

Again, never forget the design, as doing so may lead to serious bias.

POLYTOMOUS (CATEGORICAL) $1/2/.../k$

For a single outcome $Y = [Y_1, ..., Y_k]^\intercal$ the response must be Generalized Bernoulli, i.e.

$$Y|p \sim \text{GenBern}(p),$$

where $p = [p_1, ..., p_k]^\intercal$ and $\sum_{j=1}^{k} p_j = 1$, with

$$\text{E}[Y|p] = p,$$

$$\Pr(Y_j = 1|p) = p_j$$

and

$$\text{var}(Y_j|p) = p_j(1 - p_j), \quad \text{cov}(Y_j, Y_{j'}|p) = -p_j p_{j'},$$

$j = 1, ..., k$, $j \neq j'$.

Suppose $Y_i|p \sim_{i.i.d.} \text{GenBern}(p)$, $i = 1, ..., n$, and $Y = \sum_{i=1}^{n} Y_i$, then

$$Y|p \sim \text{Multinomial}_k(n, p).$$

Overdispersed Multinomial models are also available.

Count Data 0,1,2,...

Obvious candidate for a model is the Poisson distribution (arises as the natural model for random independent events, and also as an approximation to the binomial for rare events).

If

$$Y|\lambda \sim \text{Poisson}(\lambda)$$

then

$$\text{E}[Y|\lambda] = \text{var}(Y|\lambda) = \lambda,$$

which is restrictive.

But there are many other choices, e.g. the negative binomial arises from assuming that the rate λ arises from a gamma distribution.

Allowing for overdispersion is usually important.

Generalized Linear Models (GLMs)

A convenient pedagogic tool is the GLM which is defined by

1. **Random Component**. $Y_i|\theta_i, \alpha \sim p_Y(\cdot)$ where $p_Y(\cdot)$ is a member of the exponential family, that is

$$p_Y(y_i|\theta_i, \alpha) = \exp[\{y_i\theta_i - b(\theta_i)\}/a(\alpha) + c(y_i, \alpha)].$$

   If α is known this is a one-parameter exponential family model. If α is unknown then the distribution may or may not be a two-parameter exponential family model.

2. **Systematic Component**. We have a linear predictor $\boldsymbol{x}_i\boldsymbol{\beta}$ where $\boldsymbol{x}_i$ is the vector of covariates for observation i.

3. **Link function**. If $\mu_i = \mathrm{E}[Y_i|\theta_i, \alpha]$ then we have a link function $g(\cdot)$ with

$$g(\mu_i) = \boldsymbol{x}_i\boldsymbol{\beta}.$$

Many distributions are members of the exponential family.

Motivating Examples: Prostate Cancer

Data on $n = 97$ men before radical prostatectomy.

Response, Y, the log of prostate specific antigen (PSA); PSA is a concentration and is measured in ng/ml.

Aim: building a model for PSA, using eight covariates.

- `lcavol`: The log of cancer volume (in milliliters (cc)). Areas of cancer were measured from digitized images and multiplied by a thickness to produce a volume.
- `lweight`: The log of the prostate weight (in gms).
- `age`: The age of the patient (in years).
- `lbph`: The log of the amount of benign prostatic hyperplasia (BPH), a non-cancerous enlargement of the prostate gland (in cm^2). Measured as an area in a digitized image.
- `svi`: The seminal vesicle invasion, a 0/1 indicator of whether prostate cancer cells have invaded the seminal vesicle).
- `lcp`: Log of the capsular penetration; r the level of extension of cancer into the capsule (fibrous tissue which acts as an outer lining of the prostate gland). Measure: the linear extent of penetration (in cms).
- `gleason`: Gleason score, a measure of aggressiveness of the tumor. This grading system assigns a grade (1–5) to each of the two largest areas of cancer in the tissue with 1 being the least and 5 the most aggressive; the two grades are added.
- `pgg45`: Percentage Gleason scores 4 or 5.

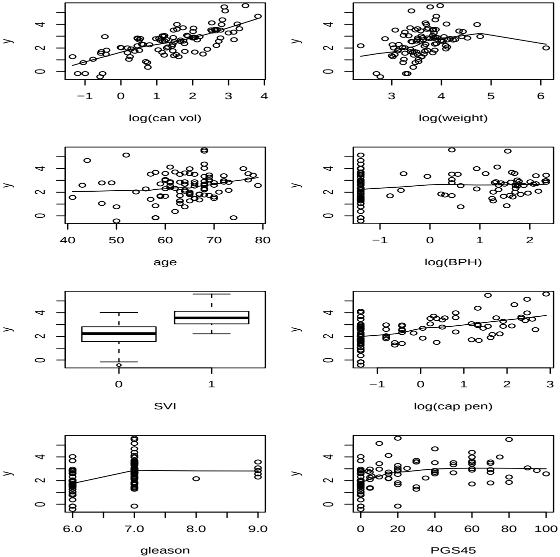

FIGURE 1: log(PSA) (*y*) plotted versus each of explanatory variables (*x*).

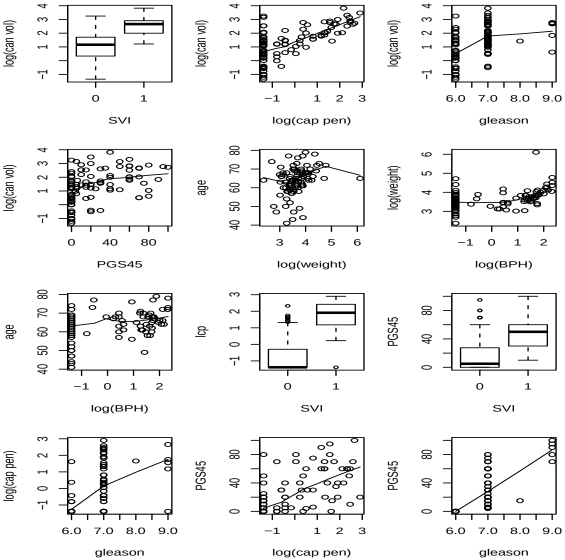

FIGURE 2: Associations between selected explanatory variables.

Multiple Linear Regression

We define Y_i as the log of PSA, and $\boldsymbol{x}_i = [1\ x_{i1}\ ...\ x_{ik}]$ with $k = 8$ as the 1×9 row vector associated with patient i, $i = 1, ..., n = 97$.

A generic mean model is given by

$$\mathrm{E}[Y_i | \boldsymbol{x}_i] = f(\boldsymbol{x}_i, \boldsymbol{\beta})$$

where $f(\cdot, \cdot)$ represents a functional form, and $\boldsymbol{\beta}$ unknown regression parameters.

The most straightforward form is the **multiple linear regression**

$$f(\boldsymbol{x}_i, \boldsymbol{\beta}) = \beta_0 + \sum_{j \in C} x_{ij} \beta_j,$$

where C corresponds to the set of elements of $\{1, 2, \ldots, 8\}$ whose associated covariates we wish to include in the model, and $\boldsymbol{\beta} = \{\beta_j, j \in C\}$.

Lung Cancer and Radon

In this example we examine the association between lung cancer incidence (over the years 1998–2002) and residential radon at the level of the county, in Minnesota.

Radon is a naturally occurring radioactive gas produced by the breakdown of uranium in soil, rock, and water, and is a known carcinogen for lung cancer.

Let:

- Y_i denote the lung cancer incidence count, and
- x_i the average radon in county,

for $i = 1, \dots, n = 87$.

Age and gender are strongly associated with lung cancer incidence.

Lung Cancer and Radon

A standard approach to controlling these factors is to form *expected counts*

$$E_i = \sum_{j=1}^{J} N_{ij} q_j$$

in which we multiply the population in stratum j and county i, N_{ij}, by a "reference" probability of lung cancer in stratum j, q_j, to obtain the expected count in stratum j.

Summing over all J stratum gives the total expected count. Intuitively, these counts are what we would expect if the disease rates in county i conform with the reference. A summary response measure in county i is the standardized morbidity ratio (SMR), given by Y_i/E_i.

Counties with SMRs greater than 1 have an excess of cases, when compared to that expected.

A negative association is seen in Figure 3 where we plot SMRs versus average radon, with a smoother indicating the local trend.

Lung Cancer and Radon

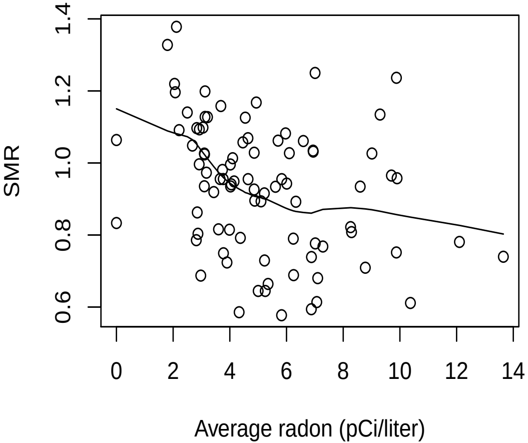

Figure 3: Standardized morbidity ratios versus average radon (pCi/liter) by county in Minnesota.

Pharmacokinetic data

Pharmacokinetics is the study of the time course of a drug and its metabolites after its introduction into the body.

A typical experiment consists of a known dose of drug being administered via a particular route (for example orally or via an injection) at a known time.

Subsequently blood samples are taken and the concentration of the drug is measured.

Hence the data is in the form of *n* pairs of points (x_i, y_i) where x_i, denotes the sampling time at which the *i*-th blood sample was taken and y_i denotes the *i*-th measured concentration.

Pharmacokinetic data

i	Time (Hours) x_i	Concentration (mg/liter) y_i
1	0.1	9.06
2	0.2	10.51
3	0.5	12.97
4	1.0	7.44
5	2.0	10.02
6	3.0	7.68
7	4.0	8.17
8	5.0	5.21
9	6.0	6.27
10	8.0	4.08
11	10.0	4.14
12	12.0	3.07
13	18.0	1.55
14	24.0	0.89

Table 1: Data from a typical pharmacokinetic experiment.

Pharmacokinetic data

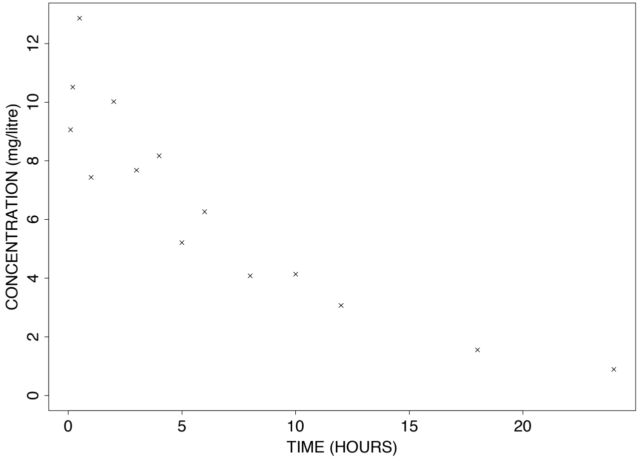

Figure 4: Plot of data from a typical pharmacokinetic experiment.

A simple compartmental system

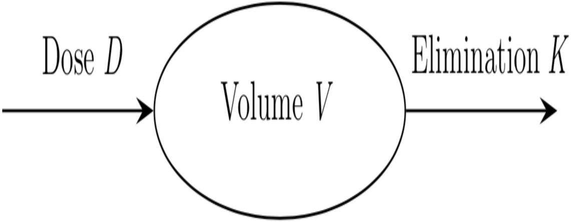

Figure 5: Representation of a one-compartment system with IV dosing.

Pharmacokinetic Data: A Generative Model

Let $w(x)$ be the amount of drug and $y(x)$ the concentration of drug in the body at time x, D the size of the dose, and K an elimination constant in the ODE:

$$\frac{dw}{dt} = -Kw.$$

We can derive the model

$$y(x) = \frac{D}{V} \exp(-Kx),$$

where V is the volume of distribution .

Model is **nonlinear** in the parameters, but notice that if we take the log of the model function we obtain a linear model.

Outcome after head injury

Table 2 reports data collected prospectively by neurosurgeons between 1968 and 1976, and the study was initiated in the Institute of Neurological Sciences in Glasgow.

The original aim was to predict recovery for individual patients on the basis of data collected shortly after the injury.

The data that we consider contain information on the outcome after head injury, which is a binary random variable, and four covariates “Pupils” (with good corresponding to reacting to light, and poor to non-reacting), “Coma score”, “Haematoma present” and “age”.

Outcome after head injury

		Pupils	Good				Poor			
		Haematoma present	No		Yes		No		Yes	
			Coma score							
		Outcome	Low	High	Low	High	Low	High	Low	High
	1–25	Dead	9	5	5	7	58	11	32	12
		Alive	47	77	11	24	29	24	13	16
Age (years)	26–55	Dead	19	6	21	14	45	7	61	15
		Alive	15	44	18	38	11	16	11	21
	$\geqslant$55	Dead	7	12	19	25	20	7	42	17
		Alive	1	6	2	15	0	2	7	7

Table 2: Outcome after head injury as a function of four covariates: pupil, haematoma present, coma score, and age.

Frequentist and Bayesian Inference

There are two principal approaches to inference which we label as **Bayesian** and **frequentist**, and each produce inferential procedures that are optimal with respect to different criteria.

Central to the philosophy of each approach is the interpretation of probability which is taken.

In the frequentist approach probabilities are viewed as limiting frequencies under infinite hypothetical replications of the situation under consideration.

Frequentist and Bayesian Inference

Frequentist:

Inference recipes, such as specific estimators are assessed with respect to their performance under repeated sampling of the data, with model parameters viewed as fixed, albeit unknown, constants.

Bayesian:

In the Bayesian approach, probabilities are viewed as subjective and are conditional on the available information so that, in general, probabilities concerning the same parameter may differ across individuals.

All unknown parameters in a model are treated as random variables, and inference is based upon the posterior distribution for these parameters, given the data.

The posterior distribution is obtained through Bayes theorem, which requires the specification of a prior distribution for the parameters of the model.

Conclusions

What's this course all about?

It's about advanced methods but we never want to lose sight of applications.

In particular:

- Context is important: how were the data collected? Observational versus experimental? Potential sources of selection bias?
- Many important issues attached to regression analysis are independent of philosophical approach, e.g., interpretation, control of confounding, acknowledgment of design/sampling scheme.

Conclusions

- What is the question of interest? Often better to answer this question with a simple, interpretable model, than a fancy one.

- Do we consistently estimate the parameter of interest when model assumptions are not valid?

- If so, is the measure of uncertainty (the standard error, interval estimates) appropriate?

- Where is the information to estimate parameters of interest coming from?

- In a Bayesian model, how dependent on the prior are conclusions?

- We need flexible model classes/estimation procedures, to deal with different types of data.

References

Nelder, J. and Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A*, **135**, 370–384.

Royall, R. (1997). *Statistical Evidence – A Likelihood Paradigm*. Chapman and Hall/CRC, Boca Raton.