

Linear Regression and Beyond

Stas Kolenikov

skolenik@unc.edu

November 10, 2004

1

Linear regression and beyond

Linear regression, as we know it

$$y_i = x_i^T \beta + \epsilon_i \quad (1)$$

or

$$Y = X\beta + \epsilon \quad (2)$$

or

$$\mathbb{E}[Y] = X\beta \quad (3)$$

or

$$Y \sim N(X\beta, \sigma^2 I) \quad (4)$$

How many assumptions do you see here?

Linear regression: assumptions

1. fixed known non-random X ;
2. all X 's are perfectly observed;
3. β fixed, unknown;
4. ϵ are independent;
5. ϵ are identically distributed;
6. (restrictive) ϵ is normal;
7. (need for variance) $\mathbb{E}[\epsilon_i] < \infty$
8. functional form: additive errors, bilinearity in X, β
9. feature of $\mathcal{D}[Y|X]$: expected value

Linear regression: nice stuff

Easy to estimate by ordinary least squares (OLS):

$$\mathbb{E}_n[(Y - X\beta)^2] \rightarrow \min_{\beta} \Rightarrow \quad (5)$$

$$\text{Normal equations : } \mathbb{E}_n[e(Y - X\beta)] = 0, \quad (6)$$

$$\hat{\beta}_{\text{OLS}} = (X^T X)^{-1} X^T Y, \quad (7)$$

$$\mathbb{V}_{\text{OLS}}[\hat{\beta}] = \sigma^2 (X^T X)^{-1} \quad (8)$$

Easy to test hypotheses on β

Easily available diagnostics: residuals, specification, model selection, ...

Getting started...

Best textbooks:

- Smith & Young (forthcoming): rich with modern methods presented at graduate level; with applications to environmental statistics
- Draper & Smith (1998): also a graduate level mathematical statistics book with somewhat greater coverage of regression topics
- Harrell (2002): an excellent applied book with focus on biostatistical applications, but with important general advice
- Fox (1997) is a reasonable book with social science applications, and even more geared towards social scientists are Cohen, Cohen, West & Aiken (2002) and Pedhazur (1997)

The rest of the talk is...

... about what can happen if the assumptions are violated, or relaxed, or changed, or ...

Random regressors

Suppose regressors X are no longer fixed, as would be the case with most observation-based (non-experimental) data. Then making an additional assumption that

$$\mathbb{E}[\epsilon|X] = 0 \quad (9)$$

or a stronger one,

$$\epsilon \perp\!\!\!\perp X \quad (10)$$

one can still obtain that

$$\mathbb{E}[Y|X] = X^T \beta \quad (11)$$

and thus the OLS estimate is still unbiased.

Random regressors: sandwich estimator

The estimating equations are still identical to (6), and thus by the general M-estimation theory (Huber 1967, Huber 1974, van der Vaart 1998) the variance of the estimator is

$$\begin{aligned} \mathbb{V}_S[\beta_{\text{OLS}}] &= (\mathbb{E}[XX^T])^{-1} \mathbb{V}[X^T \epsilon] (\mathbb{E}[XX^T])^{-1} = \\ &= (\mathbb{E}[XX^T])^{-1} \mathbb{E}[X^T \epsilon \epsilon^T X] (\mathbb{E}[XX^T])^{-1} \end{aligned} \quad (12)$$

The empirical analogues are obtained by replacing the expectation operator \mathbb{E} with empirical expectation \mathbb{E}_n , if the observations are independent. The subscript S stands for *sandwich estimator*, and it is also known as the robust variance estimator or White estimator in econometrics, and as the linearization or first order Taylor series expansion estimator in survey statistics.

Distribution of Y : non-continuous data

The linear regression model can be extended into a larger class of *exponential family* models by assuming a specific distribution of the dependent variable:

$$f(y, \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} - c(y, \phi)\right) \quad (13)$$

and a linear dependence of a certain *link function* of the mean $\mu = \mathbb{E}[y]$ on the regressors:

$$\mathbf{X}\beta = \eta = g(\mu) = g(\mathbb{E}[y]) \quad (14)$$

Those are *generalized linear models* (GLM, GLIM) (McCullagh & Nelder 1989), and they include logit and probit regressions for 0/1 outcomes, Poisson regression for counts, etc.

Distribution of Y

Somewhat related are also multinomial logit model for multiple discrete outcomes:

$$\text{Prob}(Y = j|\mathbf{X}) = \frac{\exp(\mathbf{X}^T \beta_j)}{1 + \sum_{k=1}^{m-1} \exp(\mathbf{X}^T \beta_k)}, j = 1, \dots, m - 1, \quad (15)$$

$$\text{Prob}(Y = m|\mathbf{X}) = \frac{1}{1 + \sum_{k=1}^{m-1} \exp(\mathbf{X}^T \beta_k)} \quad (16)$$

and ordinal logit and probit models for ordered discrete outcomes:

$$\text{Prob}(Y \leq j|\mathbf{X}) = F(\kappa_j - \mathbf{X}^T \beta) \quad (17)$$

popular in econometrics (Maddala 1983, Wooldridge 2002) and social sciences (Long 1997).

Distribution of Y : other features of distribution

If we keep an assumption that Y is continuous, we might be interested in other distribution features, such as (conditional) quantiles of level τ :

$$\text{Prob}[Y \leq |X] = \tau, \quad \text{or} \quad \text{Prob}[Y \leq q(X)] = \tau \quad (18)$$

Those are *quantile regressions*. One special case is $\tau = 50\%$ that corresponds to the *median regression*, or *least absolute deviations* (LAD) regression, one of the versions of robust regression.

Quantile regressions can still be considered in the framework of the M-estimates with the estimating equations

$$\psi_\tau(y, \theta) = -(1 - \tau)I\{y \leq \theta\} + \tau I\{y > \theta\} \quad (19)$$

where $I(\cdot)$ is the indicator function. Thus if $q(x) = x^T \beta$, the estimates $\hat{\beta}$ will be asymptotically normal.

Nonlinearity in X

In general, there is no reason to assume that the dependence on X should be linear, other than for a reasonable first order approximation. Higher order approximations can be obtained by *basis expansions* over regressors:

$$\mathbb{E}[Y|X] = \sum_{k=1}^r \beta_k \psi(X, k) \quad (20)$$

Such models are referred to as *generalized additive models* (Hastie & Tibshirani 1990). They still fall into the linear regression rubric, as OLS is perfectly valid method of estimating the coefficients β !

Nonlinearity in X

A great many classes of basis functions have been proposed, including:

- polynomials: $\psi^{(1)}(x, k) = x^k$
- orthogonal polynomials: $\mathcal{L}(\psi^{(2)}) = \mathcal{L}(\psi^{(1)})$, $\psi_k^{(2)}(x) \perp \psi_l^{(2)}(x)$ in a suitable sense
- B-splines: $\psi^{(3)}(x, k) =$ piecewise cubic in scaled X (Green & Silverman 1994)
- MARS (multivariate adaptive regression splines):
 $\psi^{(4)}(x, k) = I\{x > \alpha_k\}(x - \alpha_k)$

The topic is also related to model selection and statistical learning, and the best source on that is Hastie, Tibshirani & Friedman (2001).

Nonlinearity in β

Also, there is no reason why the dependence on parameters β should be linear. If we instead believe that the relation is essentially nonlinear and cannot be reduced to the linear one by any transformation of the dependent and/or explanatory variables (c.f. $y = \beta_1 x^{\beta_2}$ vs. $y = (x + \beta_1)^{\beta_2}$), then the problem can be solved by *nonlinear least squares* (NLS):

$$\mathbb{E}_n [(Y - h(\mathbf{X}, \beta))^2] \rightarrow \min_{\beta} \quad (21)$$

As soon as this is again an M-estimation problem, the estimates $\hat{\beta}$ should be asymptotically normal. Models of this kind are popular in pharmacokinetics and financial econometrics (Gallant 1987).

Endogenous regressors

Let us return to the problem of regressors that are random. Earlier, we assumed that $\mathbb{E}[\epsilon|X] = 0$. In many applications in economics and social sciences, this is not justified: the error term in the regression combines all unobserved characteristics of an individual, and there is no reason to believe that those unobserved characteristics are uncorrelated with the observed ones.

Solution: find *instrumental variables* Z such that

$$\mathbb{E}[Z^T \epsilon] = 0 \quad (22)$$

Instrumental variable regression

(22) are going to be the new estimating equations, and the solution to them are given by

$$\hat{\beta}_{IV} = (X^T P_Z X)^{-1} (X^T P_Z Y) \quad (23)$$

where

$$P_Z = Z(Z^T Z)^{-1} Z^T \quad (24)$$

is the projector into the space of Z 's. Further,

$$\mathbb{V}[\hat{\beta}_{IV}] = (\mathbb{E}[X^T P_Z X])^{-1} (\mathbb{E}[X^T P_Z \epsilon \epsilon^T P_Z X]) (\mathbb{E}[X^T P_Z X])^{-1} \quad (25)$$

Instrumental variable regression: properties

Instrumental variable estimator hinges on two assumptions: *exogeneity* (22) and *relevance*: $\text{Cov}[Z, X]$ is of full rank. In other words, for every endogenous variable X there is at least one instrument Z (otherwise, the $(X^T P_Z X)$ matrix won't be invertible).

Instrumental variable estimators are consistent and inefficient relative to OLS when the null hypothesis of exogeneity of X (9) holds:

$$H_0 : \mathbb{E}[X\epsilon] = 0 \Rightarrow \text{plim}_{n \rightarrow \infty} (\hat{\beta}_{\text{IV}} - \hat{\beta}_{\text{OLS}}) = 0 \quad (26)$$

For comparison of two estimators, of which one is inefficient under the null, but consistent under alternative, vs. the one that is efficient under the null, but inconsistent under the alternative, there exists *Hausman test*:

$$(\hat{\beta}_{\text{IV}} - \hat{\beta}_{\text{OLS}})^T (\mathbb{V}[\hat{\beta}_{\text{IV}}] - \mathbb{V}[\hat{\beta}_{\text{OLS}}])^{-1} (\hat{\beta}_{\text{IV}} - \hat{\beta}_{\text{OLS}}) \xrightarrow{d} \chi_p^2 \quad (27)$$

where p is the number of linear restrictions being tested.

Measurement error

Another occasion where the regressors are correlated with the error term are the models where regressors are measured with error:

$$Y = X^* \beta + \epsilon, \quad X^* \text{ are unobserved,} \quad (28)$$

$$X = X^* + \delta, \quad X \text{ are observed} \quad (29)$$

Then important OLS matrices like $X^T X$ become $X^{*T} X^* + \mathbb{V}[\delta]$, so the variance-covariance matrix of the measurement errors needs to be estimated and eliminated. The ways to control for measurement error are

- repeated measurements of X^* ;
- instrumental variables Z uncorrelated with δ ;
- known variances of δ ;
- structural equation models for δ

See Fuller (1987) and Carroll, Ruppert & Stefanski (1995).

Panel/longitudinal/repeated measurement data

This is a very popular format of most modern health, economic and educational surveys. The same individuals are observed over time, which helps establishing causal relations and control for unobserved heterogeneity.

In the simplest case, the linear panel model is

$$y_{ij} = x_{ij}^T \beta + u_i + \epsilon_{ij} = x_{ij}^T \beta + \nu_{ij} \quad (30)$$

where $i = 1, \dots, N$ runs over individuals, and $j = 1, \dots, T$ runs over time. (Some other fully cross-sectional hierarchical models fit into the same framework.)

The aim of the estimation procedure is to incorporate u into estimation, or condition on it.

Panel data: random effect

If $u_i \perp\!\!\!\perp \epsilon_{ij} \perp\!\!\!\perp X$, then

$$\Omega \equiv \mathbb{V}[\nu] = I_N \otimes (\sigma_u^2 J_T + \sigma_\epsilon^2 I_T) \quad (31)$$

is a block-diagonal matrix. One can devise the GLS estimator:

$$\hat{\beta}_{\text{RE}} = (X^T \Omega X)^{-1} X^T \Omega Y \quad (32)$$

Panel data: fixed effect

The random effect estimator can suffer from the endogeneity problem: if $\mathbb{E}[Xu] \neq 0$, $\hat{\beta}_{RE}$ is inconsistent. The problem can be somewhat rectified by conditioning on u_i , which may be achieved by forming a contrast of y_{ij} that eliminates u_i :

- take first differences $y_{ij} - y_{i,j-1}$;
- subtract the panel mean $y_{ij} - \bar{y}_i$.

When the same transformation is applied to the regressors, the model can be estimated by OLS (or GLS, as the regression errors are now correlated), and results in the *fixed effect* estimator $\hat{\beta}_{FE}$.

See Maddala (1993), Hsiao, Hammond & Holly (2002), Wooldridge (2002) for econometric applications, and Diggle, Heagerty, Liang & Zeger (2002) for biometric approach.

Random coefficient/mixed models

What if there is more than one random effect associated with each individual, and they affect not only the regression intercept, but also (some of the) slopes?

$$y_{ij} = x_{ij}^T(\beta + u) + \epsilon_{ij} \quad (33)$$

in the *random coefficients* notation, or

$$Y = X\beta + Zu + \epsilon \quad (34)$$

in the *mixed model* (McCulloch & Searle 2000) notation, where β are fixed effects, and u are random effects, $u \perp \epsilon$, $\mathbb{V}[u] = \Sigma$, and Z is the design matrix that links the panel-level random effects to observations.

In some applications (biometrics), Σ is considered a nuisance parameter, and no inference is made on its elements. In other applications (social sciences), the variance parameters may be of substantive interest.

Multilevel models

The mixed models are simple examples of *multilevel models* (Goldstein 2002, Raudenbush & Bryk 2002, Hox 2003) that aim at building regression models at different levels of hierarchy. A typical two-level model will look like

$$\text{level-1 model: } y_{ij} = \sum_{k=1}^{p_1} x_{ijk}^T \beta_{ik} + \epsilon_{ij} \quad (35)$$

$$\text{level-2 model: } \beta_{ik} = \sum_{l=1}^{p_2} \gamma_{lk} z_{il} + \nu_{ik} \quad (36)$$

Just as the mixed effects model, the multilevel models are most commonly estimated by the full maximum likelihood, restricted maximum likelihood, or by Bayesian methods.

Almost done!

But first let us explore a few special cases: survey data, diagnostics, and combinations of some of the aforementioned extension of the linear regression model.

Complex survey estimation

The assumption of independent identically distributed data is rarely satisfied in practice. A practical survey design usually have some of

- *stratification*: using the information available before the sampling stage on what individuals can be expected to be similar, so that we don't have to sample too many of them;
- *clustering*: a practical consideration of sampling individuals in (geographical) clusters, when there are economies of scale to do so
- *weighting* for unequal probabilities of selection, if some subpopulations are to be sampled with higher frequencies than others (often, minorities)

See Kish (1965), Cochran (1977), Thompson (1992), Särndal, Swensson & Wretman (1992), Korn & Graubard (1999), Chambers & Skinner (2003).

Surveys and extensions of the linear model

The complex survey structure can be accounted for reasonably easily in:

- linear regression models;
- *generalized linear models*: probit, logit, Poisson;

Difficulties arise in:

- *multilevel models*, small sample size;
- *panel/longitudinal models*

I have not seen much of other models discussed in this talk being applied in the survey context.

Residual diagnostics

One of the very strong advantages of the linear regression model is the ease of residual testing (Belsley, Kuh & Welsch 1980, Smith & Young forthcoming). Can this be carried over to some of the extensions described here?

- **generalized linear models**: what is a residual?
- **multilevel models, panel/longitudinal models**: what exactly is the residual attributed to — i, ij, \dots ?
- Also, as long as the data are no longer i.i.d., a single residual is far more difficult to isolate
- **survey context**: cannot take out observations, as this would lead to biased/unrepresentative samples

Interactions and combinations

Can some of the extension outlined above be combined in a sensible way?

- **GLM + random effects** → GLMM, generalized linear mixed models
- **GLM + splines**
- **GLM + multilevel models** (IRT?)
- **GLM + measurement error**
- **GLM + panel structure**: panel logit, probit, Poisson
- **quantile regression + panel structure**: cutting edge in econometrics

References

- Belsley, D. A., Kuh, E. & Welsch, R. E. (1980), *Regression Diagnostics : Identifying Influential Data and Sources of Collinearity*, Wiley-Interscience, New York.
- Carroll, R. J., Ruppert, D. & Stefanski, L. A. (1995), *Measurement Error in Nonlinear Models*, Chapman and Hall/CRC.
- Chambers, R. L. & Skinner, C. J., eds (2003), *Analysis of Survey Data*, John Wiley and Sons.
- Cochran, W. G. (1977), *Sampling Techniques*, John Wiley and Sons, New York.
- Cohen, P., Cohen, J., West, S. G. & Aiken, L. S. (2002), *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 3rd edn, Lea.
- Diggle, P., Heagerty, P., Liang, K.-Y. & Zeger, S. (2002), *Analysis of Longitudinal Data*, 2nd edn, Oxford University Press.

- Draper, N. P. & Smith, H. (1998), *Applied Regression Analysis*, 3rd edn, John Wiley and Sons, New York.
- Fox, J. (1997), *Applied Regression Analysis, Linear Models, And Related Methods*, SAGE, Thousand Oaks, CA.
- Fuller, W. A. (1987), *Measurement Error Models*, John Wiley and Sons, New York.
- Gallant, A. R. (1987), *Nonlinear Statistical Models*, John Wiley and Sons, New York.
- Goldstein, H. (2002), *Multilevel Statistical Models*, 3rd edn, Arnold Publishers.
- Green, P. J. & Silverman, B. W. (1994), *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, Vol. 58 of *Monographs on Statistics and Applied Probability*, Chapman & Hall.
- Harrell, F. (2002), *Regression Modeling Strategies*, Springer-Verlag, New York.
- Hastie, T. & Tibshirani, R. (1990), *Generalized additive models*, Chapman & Hall/CRC.

- Hastie, T., Tibshirani, R. & Friedman, J. (2001), *The Elements of Statistical Learning*, Springer-Verlag, New York.
- Hox, J. (2003), *Multilevel Analysis: Techniques and Applications*, Lawrence Erlbaum.
- Hsiao, C., Hammond, P. & Holly, A., eds (2002), *Analysis of Panel Data*, 2nd edn, Cambridge University Press.
- Huber, P. (1967), The behavior of the maximum likelihood estimates under nonstandard conditions, in 'Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability', Vol. 1, University of California Press, Berkeley, pp. 221–233.
- Huber, P. (1974), *Robust Statistics*, Wiley, New York.
- Kish, L. (1965), *Survey Sampling*, John Wiley and Sons, New York.
- Korn, E. L. & Graubard, B. I. (1999), *Analysis of Health Surveys*, John Wiley and Sons.

- Long, J. S. (1997), *Regression Models for Categorical and Limited Dependent Variables*, SAGE, Thousand Oaks.
- Maddala, G. (1983), *Limited Dependent and Qualitative Variables in Econometrics*, Cambridge University Press.
- Maddala, G. (1993), *The Econometrics of Panel Data*, Brookfield.
- McCullagh, P. & Nelder, J. A. (1989), *Generalized Linear Models*, 2nd edn, Chapman and Hall, London.
- McCulloch, C. E. & Searle, S. R. (2000), *Generalized, Linear, and Mixed Models*, John Wiley and Sons, New York.
- Pedhazur, E. J. (1997), *Multiple Regression in Behavioral Research: Explanation and Prediction*, 3rd edn, Holt, Rinehart and Winston, New York.
- Raudenbush, S. & Bryk, A. (2002), *Hierarchical Linear Models*, 2nd edn, SAGE, Thousand Oaks, CA.
- Särndal, C.-E., Swensson, B. & Wretman, J. (1992), *Model Assisted Survey Sampling*, Springer, New York.

Smith, R. L. & Young, K. D. S. (forthcoming), *Linear regression*, Cambridge University Press.

Thompson, S. K. (1992), *Sampling*, John Wiley and Sons, New York.

van der Vaart, A. W. (1998), *Asymptotic Statistics*, Cambridge University Press, Cambridge.

Wooldridge, J. M. (2002), *Econometric Analysis of Cross Section and Panel Data*, MIT Press, Cambridge, MA.