

*The Effect of Complex Sampling
on Statistical Procedures
in Social Science Research*

Stas Kolenikov

skolenik@unc.edu

February 24, 2005

Personal background

Engineering (MEE, 1996) to Economics (MSc, 1998) to Statistics (Ph.D. expected 2005)
to...

Retrospective of research themes:

- Quality of life: cross-country and cross-regional analysis (NES, 1997–1998, with S. Aivazian)
- Income inequality and poverty with applications to Russia (CEMI, 1998–2000, with S. Aivazian, and WIDER, 2001–2003, with A. F. Shorrocks)
- Active labor market programs in Russia (CEFIR, 2002, with I. Denisova)
- Estimation on the boundary and misspecified models in SEMs (UNC, 2004–2005, with K. A. Bollen)
- Socio-economic status in health economics (UNC-CPC, 2003–2005, with G. Angeles)
- Properties of repeated cluster designs (UNC-CPC, 2003–2005, with G. Angeles)
- Effect of military draft on human capital accumulation in young Russian males (CEFIR, 2005, with I. Denisova)

Outline

1. Linear regression and its assumptions	4
2. Complex surveys:	
• Main concepts	10
• Estimation problems	12
• Software	23
3. Empirical examples with GSS	
• occupational prestige in GSS	25
• attitudes towards race	35
4. References	41

Linear regression, as is

$$y_i = x_i^T \beta + \epsilon_i, \quad i = 1, \dots, N \quad (1)$$

or

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

or

$$\mathbb{E}[Y] = X\beta, \quad \text{or} \quad \mathbb{E}[y_i|x_i] = x_i^T \beta \quad (3)$$

or

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

What are the assumptions 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 (strong) or uncorrelated (weaker)
5. ϵ are identically distributed
6. $\mathbb{E}[\epsilon_i] = 0, \mathbb{E}[\epsilon_i^2] < \infty$
7. functional form: additive errors, linearity w.r.t. X , w.r.t. β
8. feature of the distribution of the dependent variable: expected value
9. (restrictive, needed for MLE and “exact” tests) ϵ are normal

Linear regression: nice stuff

One of the oldest statistical methods!

Easy to estimate by ordinary least squares (OLS):

$$\sum_{i=1}^n [(y_i - x_i^T \beta)^2] \rightarrow \min_{\beta} \Rightarrow \quad (5)$$

$$\text{Normal equations : } \sum_{i=1}^n [x_i (y_i - x_i^T \beta)] = 0, \quad (6)$$

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

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

Easy to test hypotheses on β

Easily available diagnostics

Regression diagnostics

- **Residuals and outliers:** how do the regression results change if a particular observation or a group of observations is excluded from the model? (Belsley, Kuh & Welsch 1980, Atkinson & Riani 2000)
- **Specification:** is the relation between the dependent and explanatory variables really linear? Are the appropriate variables included? Need any of the variables (including the dependent one) be transformed? Are there any endogeneity problems?
- **Model selection:** how does one choose the most appropriate model from the range of the plausible ones? Stepwise methods, information criteria, cross-validation, ... (Hastie, Tibshirani & Friedman 2001)

Books on regression

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

Outline

1. Linear regression and its assumptions	4
2. Complex surveys:	
• Main concepts	10
• Estimation problems	12
• Software	23
3. Empirical examples with GSS	
• occupational prestige in GSS	25
• attitudes towards race	35
4. References	41

Sampling

Most statistical methods are developed for, and most statistics classes are taught with the idea of, simple random samples (SRS):

An SRS of size n from population of size N is the design such that all population subsets of size n have equal probabilities of selection (equal to $\binom{N}{n}^{-1}$)

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

Sampling is not a very popular topic in most programs in mathematical statistics; more emphasis in biostatistics. Mathematical foundations: finite population sampling.

Survey data

Every major survey has some aspects of:

- *stratification*: sampling from well-defined parts of population using the information available prior to the sampling stage
- *clustering*: designs that sample groups of units rather than individual units
- *weighting* for unequal probabilities of selection, if some subpopulations are to be sampled with higher frequencies than others (e.g., minorities), or sampling units differ in size, or ...

Violations of the linear regression assumptions

- Independence (No. 4): clusters are taken as a whole
- Probabilities (No. 5): observations in the same cluster are more likely to go together
- Missing data (No. 2)
- Distributional assumptions (No. 9): there are no normal populations in real life...
- Need to relate the results to the specific population
- What are the population parameters? Regression in population? in metapopulation?

Design effect

Measure of importance of the design: *design effect*

$$DEFF[\bar{x}] = \frac{\mathbb{V}[\bar{x}; \text{design}, n]}{\mathbb{V}[\bar{x}; \text{weighted SRS}, n]} \quad (9)$$

$n/DEFF$ can be viewed as a measure of the effective sample size.

Also encountered: *misspecification effect*

$$MEFF[\bar{x}] = \frac{\mathbb{V}[\bar{x}; \text{design}, n]}{\mathbb{V}[\bar{x}; \text{unweighted SRS}, n]} \quad (10)$$

Good surveys: all variables have $1 < DEFF < 2$. Very good surveys: $DEFF < 1$. Some variables may be problematic with $DEFF > 10$.

Stratification in surveys

- Uses preliminary knowledge that some sampling units are similar to each other
- Helps reduce the variance of the estimates ($DEFF \approx 0.9$ uniformly)
- Allows for valid statistical inference within strata
- If ignored, the variance estimates are conservative
- Practical designs: highly stratified, with several dozen strata
- Relation to weighting

Stratified designs

- *Proportionate*: sampling rates are constant across strata, the proportions of observations in the sample are the same as proportions in populations
- *Optimal*: for populations with variances S_h^2 and costs per unit c_h varying across strata h , the number of units sampled should be $\propto S_h c_h^{-1/2}$
- *Disproportionate*: small subpopulations are oversampled if there is a need to analyze them separately
- Multistage versions

Clustering in surveys

- Sampling groups of units rather than each unit independently
- Used when lists are impossible or prohibitively expensive to obtain
- Units within the group are similar \Rightarrow effective sample size is smaller than the number of sampled units
- Increases the variance, reduces the costs
- $DEFF = 1 + \rho_{ICC}(m - 1)$, ρ_{ICC} is the intraclass correlation (measure of similarity of units in a group), m is the (average) number of units per group. May be as high as 20.
- Also applicable: interviewer effects in personal interview surveys
- **If ignored, leads to biased standard errors** (usually underestimated)
- Also known as: hierarchical, multi-stage sampling

Weighting in surveys

- Used to compensate for unequal probabilities of selection in surveys:

$$w_i = \text{Prob}[\text{selection of unit } i \text{ into the sample} | \text{design}]^{-1} \quad (11)$$

- **If ignored, leads to bias estimates, both for coefficients and the standard errors!** However, if all of the stratification information is used in MLE, there is no need for weighting.
- Another type of weighting is *post-stratification weighting* to make the (marginal) distributions of the data compatible with known (census) distributions, to account for minor incompatibilities between the sample (frame) and population, or for unit non-response
- $DEFF \approx 1 + CV_w$

Point estimation

Most of the time, obtaining the point estimates of the population parameters of interest is not a huge deal:

- totals: $T[x] = \sum_i w_i x_i$
- ratios (including means): $\bar{x} = T[x]/T[1]$
- regression coefficients: $\beta_{\text{WLS}} = (X^T W X)^{-1} X^T W Y$
- maximum (quasi-)likelihood estimation: logistic regression, latent variable models (SEM, LCA, ...)

Estimation of variance, however, is more complicated, as units are not independent of each other, and their selection probabilities are different.

Variance estimation

Design-based vs. model-based methods of variance estimation:

- *Design-based* methods are derived from the assumptions on the design using the methods of finite population sampling
- *Model-based* methods assume the specific form of the dependence imposed by the design, e.g. variance components and multilevel methods.
- *Generalized variance function approach*

For samples with replacement, it suffices to correct only for the first (top) stage level of clustering. For samples without replacement (all major designs), this is a conservative approximation.

Design-based methods of variance estimation

- Analytical methods: sandwich estimator (Huber 1974) (aka robust estimator, linearization estimator, first order Taylor series expansion estimator, White estimator, ...). Regression:

$$\hat{V}[\hat{\beta}] = \left(\sum_k X_k^T W_k X_k \right)^{-1} \left(\sum_k X_k^T W_k e_k e_k^T W_k X_k \right) \left(\sum_k X_k^T W_k X_k \right)^{-1} \quad (12)$$

(k runs over clusters!)

- Resampling methods: BRR, jackknife, bootstrap. **The replicates are taken in whole clusters!** In practice, some data sets come with the sets of the resampling weights (Stat Canada: NLSCY, CCHS, NPHS, GSS, ...). **Scaling factors may be needed** (or fewer PSUs per stratum resampled)

What's left of nice stuff?

A few nice things about linear regression were mentioned above. What is still applicable in surveys?

- Residuals and outlier diagnostics: not routinely performed; concerns with unbiased estimation or interpretation of residuals?
- Model specification: based on auxiliary regressions, hence should be feasible
- Model selection: information criteria are not formally applicable, as no appropriate likelihood results from the estimation procedure.
Cross validation: what part of the sample to be sacrificed?

Difficult situations

- Multilevel and longitudinal data
 - Need for weights at different levels
 - Scaling of weights: minimize bias to the first-level variances (small # obs./cluster)
 - Sample attrition \Rightarrow different weights for the same unit? Which set of weights to use?
- Systematic sampling designs
- Single PSU per stratum
- Case-control studies

Software for survey analysis

- SUDAAN: developed by RTI specifically for the purposes of complex Survey Data ANalysis; handles most situations and a good range of statistical models; SAS-callable
- Stata: general purpose statistical software; supports stratification, clustering and weights for a wide range of statistical models and for user-written maximum likelihood estimation routines through the linearization estimator; resampling methods are available through user-contributed modules
- gllamm: user-contributed Stata module for multilevel and latent variable modelling; handles clustering and weights; very slow
- M-plus: latent variable and multilevel software package developed by B. Muthén; handles clustering and weights
- LISREL: stratification, weighting and clustering in GLMs and SEMs
- SPSS: add-on for complex analysis in SPSS 13

Outline

1. Linear regression and its assumptions	4
2. Complex surveys:	
• Main concepts	10
• Estimation problems	12
• Software	23
3. Empirical examples with GSS	
• occupational prestige in GSS	25
• attitudes towards race	35
4. References	41

Empirical demonstration: GSS

The General Social Survey (GSS) is an omnibus survey of the general US population.

- years from 1972 to 2004 (biennial since 1994)
- about 3000 respondents per year (since 1994), about 1000 items
- face to face interview, RR \sim 70%
- largest sociology project supported by NSF
- part of International Social Survey Program (ISSP)
- split samples for different modules and methodological experiments
- three stages of sampling, top level: 100 PSUs
- 2004 frame is based on 2000 census

Example 1

Occupational prestige as a function of

- demographics: age, gender, education, race
- parents: occupational prestige of father and mother

Complications: missing data on parental prestige (2643 data points, 2165 have father's prestige, 1619 have mother's prestige; out of 2765 observations in 2002 wave)

Regression setup

1. no account for design whatsoever (baseline for MEFF)
2. regression with weights for subsampling within HH (baseline for DEFF)
3. regression with clustering (PSUs)
4. regression with PSUs and weights — the correct use of the design

Also, different regression specifications are tried, with or without parental occupational prestige scores.

Overall results

	Naïve				Weighted			
Age of respondent	0.077 (0.014)**	0.074 (0.016)**	0.091 (0.023)**	0.079 (0.023)**	0.099 (0.015)**	0.094 (0.017)**	0.108 (0.024)**	0.096 (0.025)**
Female	0.364 (0.471)	0.287 (0.534)	0.270 (0.717)	0.249 (0.723)	0.555 (0.505)	0.480 (0.570)	0.517 (0.761)	0.558 (0.769)
Race: black	-1.816 (0.698)**	-0.858 (0.909)	-0.575 (1.154)	-0.920 (1.148)	-1.968 (0.754)**	-1.483 (0.964)	-1.483 (1.209)	-1.836 (1.219)
Race: other	-2.185 (0.994)*	-1.978 (1.117)	-0.787 (1.495)	-1.150 (1.510)	-1.944 (1.005)	-1.833 (1.143)	-1.369 (1.661)	-1.617 (1.676)
Highest year of school	2.359 (0.099)**	2.366 (0.118)**	2.333 (0.169)**	2.459 (0.162)**	2.377 (0.105)**	2.350 (0.125)**	2.262 (0.190)**	2.402 (0.182)**
Father's prestige		0.052 (0.023)*	0.000 (0.031)			0.054 (0.025)*	0.010 (0.034)	
Mother's prestige			0.100 (0.028)**				0.107 (0.031)**	
Constant	8.869 (1.563)**	6.935 (1.895)**	4.880 (2.708)	7.856 (2.591)**	7.322 (1.681)**	5.854 (1.966)**	4.045 (2.876)	7.548 (2.854)**
Observations	2627	2070	1198	1198	2627	2070	1198	1198
R-squared	0.26	0.26	0.25	0.24	0.27	0.27	0.25	0.24

Standard errors in parentheses

* significant at 5%; ** significant at 1%

Overall results

	Cluster-corrected				Design: clustering + weights			
Age of respondent	0.077 (0.013)**	0.074 (0.015)**	0.091 (0.023)**	0.079 (0.023)**	0.099 (0.013)**	0.094 (0.015)**	0.108 (0.025)**	0.096 (0.024)**
Female	0.364 (0.534)	0.287 (0.585)	0.270 (0.734)	0.249 (0.733)	0.555 (0.603)	0.480 (0.655)	0.517 (0.800)	0.558 (0.812)
Race: black	-1.816 (0.633)**	-0.858 (0.890)	-0.575 (1.132)	-0.920 (1.136)	-1.968 (0.697)**	-1.483 (0.971)	-1.483 (1.253)	-1.836 (1.292)
Race: other	-2.185 (0.833)*	-1.978 (0.809)*	-0.787 (1.278)	-1.150 (1.310)	-1.944 (0.852)*	-1.833 (0.841)*	-1.369 (1.394)	-1.617 (1.435)
Highest year of school	2.359 (0.104)**	2.366 (0.123)**	2.333 (0.156)**	2.459 (0.153)**	2.377 (0.108)**	2.350 (0.126)**	2.262 (0.175)**	2.402 (0.171)**
Father's prestige		0.052 (0.023)*	0.000 (0.032)			0.054 (0.026)*	0.010 (0.035)	
Mother's prestige			0.100 (0.027)**				0.107 (0.030)**	
Constant	8.869 (1.593)**	6.935 (1.975)**	4.880 (2.754)	7.856 (2.576)**	7.322 (1.584)**	5.854 (1.937)**	4.045 (2.928)	7.548 (2.750)**
Observations	2627	2070	1198	1198	2627	2070	1198	1198
R-squared	0.26	0.26	0.25	0.24	0.27	0.27	0.25	0.24

Standard errors in parentheses

* significant at 5%; ** significant at 1%

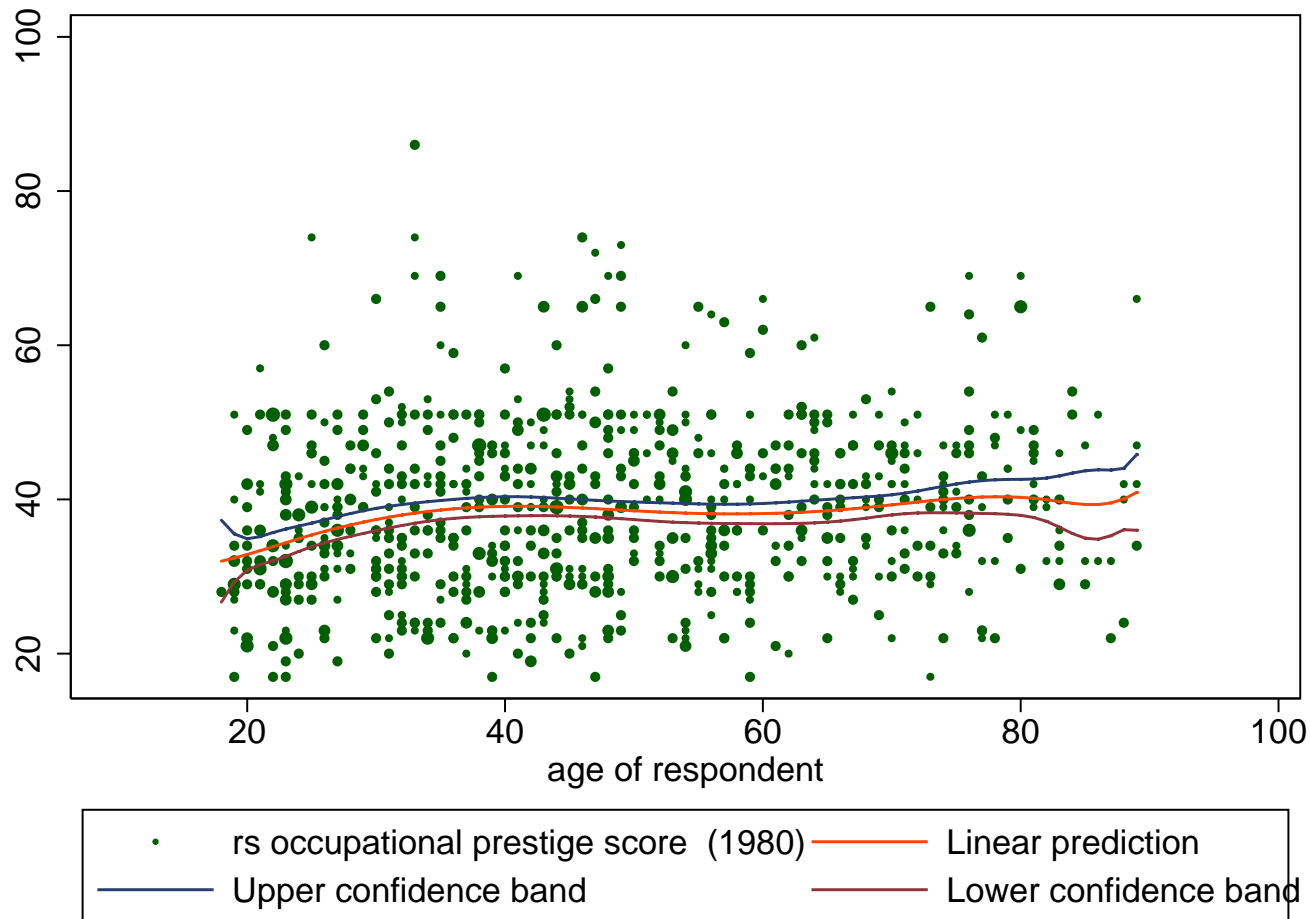
Overall results

Years of schooling	No weights	Weighted
No clustering	$\hat{\beta} = 2.359$ S.e. = 0.099 DEFF = 1	$\hat{\beta} = 2.377$ S.e. = 0.105 DEFF = 1.141
Clustering on PSUs	$\hat{\beta} = 2.359$ S.e. = 0.104 DEFF = 1.099	$\hat{\beta} = 2.377$ S.e. = 0.108 DEFF = 1.216

Specification issues

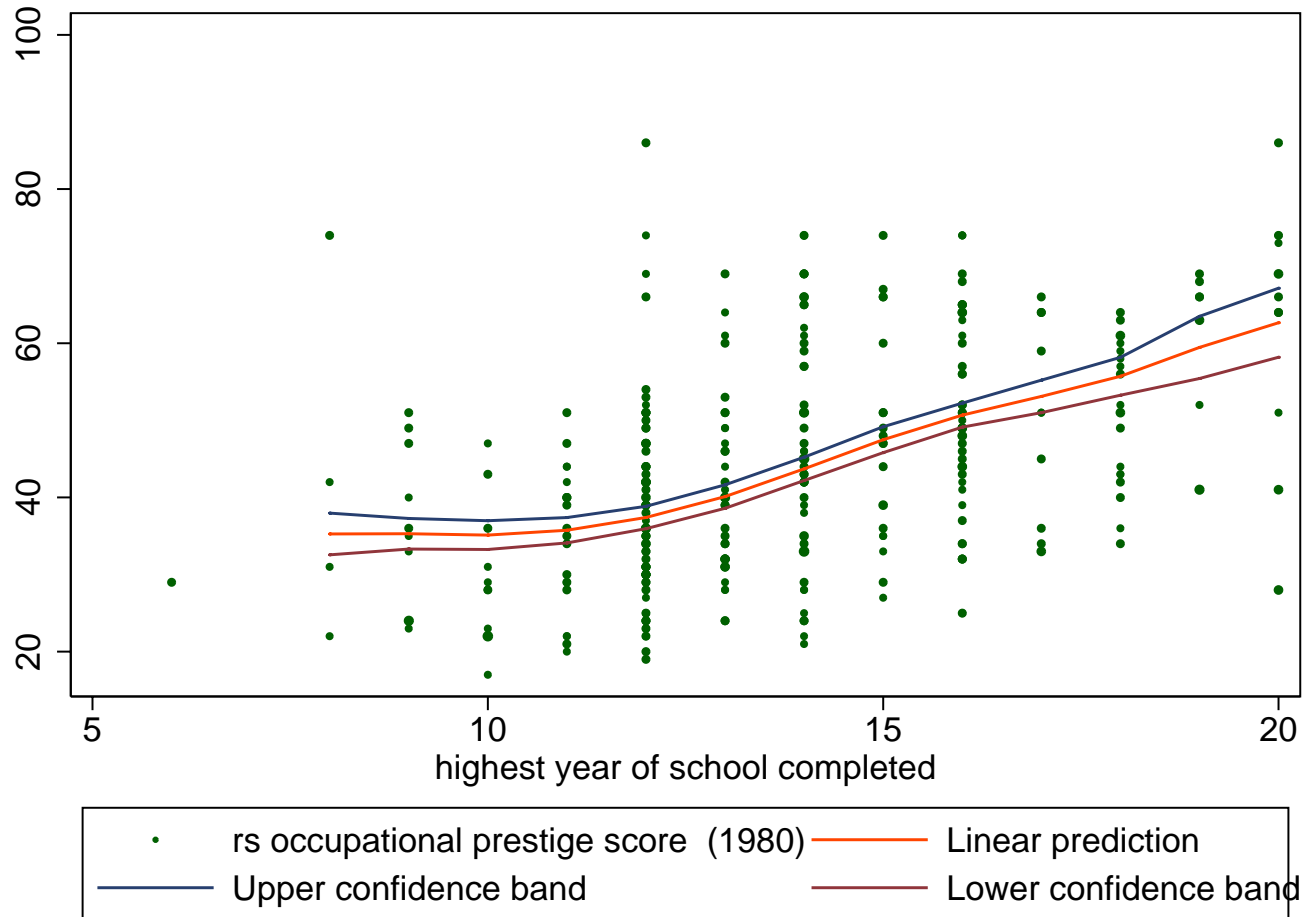
- Are the design corrections important?
- Heteroskedasticity?
- Nonlinearity in age?
- Nonlinearity in education level?
- Endogeneity of education?

Nonlinearity in age



White male, high school. Predicted curve based on the whole sample.

Nonlinearity in education



White male 30 years of age. Predicted curve based the whole sample.

Outline

1. Linear regression and its assumptions	4
2. Complex surveys:	
• Main concepts	10
• Estimation problems	12
• Software	23
3. Empirical examples with GSS	
• occupational prestige in GSS	25
• attitudes towards race	35
4. References	41

Example 2

GSS, 1982 with oversampling of the black subpopulation (relative sampling rate = 3.27)

Q.: Are we spending too much money (3), too little money (1), or about the right amount (2) on improving the conditions of Blacks?

Coefficient	Design			Naïve	
	Estimate	S.e.	DEFF	Estimate	S.e.
Age × 1000	4.7172	3.507	1.568	1.799	3.005
Gender	-0.235*	0.114	1.380	-0.297**	0.103
Years of school	-0.028	0.022	1.641	-0.045**	0.018
Race: black	-3.292**	0.217	0.858	-3.380**	0.163
Race: other	-0.565	0.380	1.552	-0.677	0.399
Thresholds: 1–2	-1.730	0.463	2.132	-2.158	0.336
2–3	0.639	0.447	2.032	0.188	0.331

Conclusion

- Brief review of the linear regression model and its assumptions
- Brief review of common design features
- Complications arising in the survey data
- Empirical example: does the survey structure matter? You don't know until you apply the correction.

P.S.: other problems with survey data

Missing data:

- Pertinent feature of most real life data sets, especially in social sciences
- Unit non-response, item non-response
- MCAR, MAR, NMAR
- Most of the modern methods (EM algorithm, multiple imputation) have some Bayesian flavor
- A common feature in public data sets: multiple imputation
- Closely related concept: plausible values (NAEP)

P.S.: own research in survey sampling

- Design of DHS: several waves of data collection, clusters from the master sample of PSUs are used repeatedly, individuals are sampled anew each time
- **Q:** does one need to account for the re-use of clusters in computing say s.e. of the trend?
- **A:** the design effect is of order $1 - A\rho/n$
 - ρ is the intertemporal correlation of cluster totals
 - n is the number of clusters/PSUs
 - $A > 0$
- The usefulness of the repeated cluster designs comes from the logistics (smaller costs)

P.S.: other extensions of linear regression

- Random regressors: instrumental variables estimator (Wooldridge 2002)
- Non-continuous data: generalized linear models (including logistic and Poisson regression) (McCullagh & Nelder 1989, McCulloch & Searle 2000)
- Nonlinearity in explanatory variables: spline models (Green & Silverman 1994, Hastie et al. 2001)
- Nonlinearity in parameters: nonlinear least squares (Gallant 1987)
- Panel/longitudinal data: fixed and random effects, mixed models (Maddala 1993, Diggle, Heagerty, Liang & Zeger 2002, Hsiao, Hammond & Holly 2002)
- Multilevel models: separate regression models for each level of the model; interaction between levels (Goldstein 2002, Raudenbush & Bryk 2002, Hox 2003). Closely related: latent growth models.
- Distribution features: quantile regression (Koenker 2005)

Outline

1. Linear regression and its assumptions	4
2. Complex surveys:	
• Main concepts	10
• Estimation problems	12
• Software	23
3. Empirical examples with GSS	
• occupational prestige in GSS	25
• attitudes towards race	35
4. References	41

References

- Atkinson, A. C. & Riani, M. (2000), *Robust Diagnostic Regression Analysis*, Springer-Verlag Inc.
- Belsley, D. A., Kuh, E. & Welsch, R. E. (1980), *Regression Diagnostics : Identifying Influential Data and Sources of Collinearity*, Wiley-Interscience, New York.
- 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.

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. & 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. (1974), *Robust Statistics*, Wiley, New York.
- Kish, L. (1965), *Survey Sampling*, John Wiley and Sons, New York.
- Koenker, R. (2005), *Quantile Regression*, Cambridge University Press, Cambridge, UK.
- Korn, E. L. & Graubard, B. I. (1999), *Analysis of Health Surveys*, John Wiley and Sons.

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.

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