POLS 503: Advanced Quantitative Political Methodology: The Notes

Jeffrey B. Arnold

2016-05-25

Contents

Chapter 1

Introduction

Notes for POLS 503.

Chapter 2

Linear Regression and the Ordinary Least Squares (OLS) Estimator

Since we will largely be concerned with using linear regression for inference, we will start by discussion the population parameter of interest (population linear regression function), then the sample statistic (sample linear regression function) and estimator (ordinary least squares).

We will then consider the properties of the OLS estimator.

2.1 Linear Regression Function

The **population linear regression function** is

$$r(x) = \mathrm{E}[Y|X = x] = \beta_0 + \sum_{k=1}^{K} \beta_k x_k.$$

The population linear regression function is defined for random variables, and will be the object to be estimated.

Names for $\boldsymbol{y}$

- dependent variable
- explained variable
- response variable
- predicted variable
- regressand
- outcome variable

Names for $\boldsymbol{X}$,

- indpendent variables
- explanatory varaibles
- treatment and control variables
- predictor variables
- covariates
- regressors

To estimate the unkonwn population linear regression, we will use the **sample linear regression function**,

$$\hat{r}(x_i) = \hat{y}_i = \hat{\beta}_0 + \sum_{k=1}^{K} \hat{\beta}_k x_k.$$

However, we

$\hat{Y}_i$ are the fitted or predicted value The **residuals** or **errors** are the prediction errors of the estimates

$$\hat{\epsilon}_i = y_i - \hat{y}_i$$

$\boldsymbol{\beta}$ are the parameters; β_0 is called the *intercept*, and $\beta_1, \ldots, \beta_K$ are called the *slope parameters*, or *coefficients*.

We will then consider the properties of the OLS estimator.

The linear regression can be more compactly written in matrix form,

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} & \cdots & x_{K,1} \\ 1 & x_{1,2} & x_{2,2} & \cdots & x_{K,2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1,N} & x_{2,n} & \cdots & x_{K,N} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix} .$$

More compactly, the linear regression model can be written as,

$$\underbrace{\boldsymbol{y}}_{N\times 1} = \underbrace{\boldsymbol{X}}_{N\times K} \underbrace{\boldsymbol{\beta}}_{K\times 1} + \underbrace{\boldsymbol{\varepsilon}}_{N\times 1}.$$

The matrix $\boldsymbol{X}$ is called the *design* matrix. Its rows are each observation in the data. Its columns are the intercept, a column vector of 1's, and the values of each predictor.

2.2 Ordinary Least Squares

Ordinary least squares (OLS) is an estimator of the slope and statistic of the regression line[1]. OLS finds values of the intercept and slope coefficients by minimizing the squared errors,

$$\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_K = \underset{b_0, b_1, \ldots, b_k}{\arg\min} \sum_{i=1}^{N} \underbrace{\left(y_i - b_0 - \sum_{k=1}^{K} b_k x_{i,k} \right)^2}_{\text{squared error}},$$

or, in matrix notation,

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \arg\min_{\boldsymbol{b}} \sum_{i=1}^{N} (y_i - \boldsymbol{b}'\boldsymbol{x}_i)^2 \\ &= \arg\min_{\boldsymbol{b}} \sum_{i=1}^{N} u_i^2 \\ &= \arg\min_{\boldsymbol{b}} \boldsymbol{u}'\boldsymbol{u} \end{aligned}$$

[1]Ordinary least squares is distinguished from, and a special case of *generalized least squares* (GLS), which adds an additional $N \times N$ matrix to the objective function,

$$\hat{\beta}_{WLS} = \arg\min_{\boldsymbol{b}} \sum_{i=1}^{N} (y_i - \boldsymbol{x}'\boldsymbol{b})' \Omega (y_i - \boldsymbol{x}'\boldsymbol{b}).$$

Weighted least squares (WLS) is another special case of GLS.

where $\boldsymbol{u} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$.

In most statistical models, including even generalized linear models such as logit, the solution to this minimization problem would be solved with optimization methods that require iteration. One nice feature of OLS is that there is a closed form solution for $\hat{\beta}$ even in the multiple regression case, so no iterative optimization methods need to be used.

In the bivariate regression case, the OLS estimators for β_0 and β_1 are

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 7 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

$$= \frac{\text{Cov}(\boldsymbol{x}\boldsymbol{y})}{\text{Var}\,\boldsymbol{x}} \quad = \frac{\text{Sample covariance betweeen } \boldsymbol{x} \text{ and } \boldsymbol{y}}{\text{Sample variance of } \boldsymbol{x}}.$$

In the multiple regression case, the OLS estimator for $\hat{\boldsymbol{\beta}}$ is

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

The term $\boldsymbol{X}'\boldsymbol{X}$ is similar to the variance of $\boldsymbol{x}$ in the bivariate case. The term $\boldsymbol{X}'\boldsymbol{y}$ is similar to the covariance between $\boldsymbol{X}$ and $\boldsymbol{y}$ in the bivariate case.

The sample linear regression function estimated by OLS has the following properties:

1. Residuals sum to zero,
$$\sum_{i=1}^{N} \hat{\epsilon}_i = 0.$$
This implies that the mean of residuals is also 0.
2. The regression function passes through the point $(\bar{\boldsymbol{y}}, \bar{\boldsymbol{x}}_1, \ldots, \bar{\boldsymbol{x}_K})$. In other words, the following is always true,
$$\bar{\boldsymbol{y}} = \hat{\beta}_0 + \sum_{k=1}^{K} \hat{\beta}_k \bar{\boldsymbol{x}}_k.$$
3. The residuals are uncorrelated with the predictor
$$\sum_{i=1}^{N} x_i \hat{\epsilon}_i = 0$$
4. The residuals are uncorrelated with the fitted values
$$\sum_{i=1}^{N} \hat{y}_i \hat{\varepsilon}_i = 0$$

2.3 Properties of the OLS Estimator

2.3.1 What makes an estimator good?

Estimators are evaluated not on how close an estimate in a given sample is to the population, but how their sampling distributions compare to the population. In other words, judge the *methodology* (estimator), not the *result* (estimate).[^ols-properties-references]

Let θ be the population parameter, and $\hat{\theta}$ be an estimator of that population parameter.

Bias The bias of an estimator is the difference between the mean of its sampling distribution and the population parameter,

$$\text{Bias}(\hat{\theta}) = \text{E}(\hat{\theta}) - \theta.$$

Variance The variance of the estimator is the variance of its sampling distribution, $\text{Var}(\theta)$.

Efficiency (Mean squared error) An efficient estimator is one that minimizes a given "loss function", which is a penalty for missing the population average. The most common loss function is squared loss, which gives the *Mean Squared Error (MSE)* of an estimator.

$$\text{MSE}(\hat{\theta}) = \text{E}\left[(\hat{\theta} - \theta)^2\right] = (\text{E}(\hat{\theta}) - \theta)^2 + \text{E}(\hat{\theta} - \text{E}(\hat{\theta}))^2 = \text{Bias}(\hat{\theta})^2 + \text{Var}(\hat{\theta})$$

The mean squared error is a function of both the bias and variance of an estimator.

This means that some biased estimators can be more efficient : than unbiased estimators if their variance offsets their bias.[2]

Consistency is an asymptotic property[3], that roughly states that an estimator converges to the truth as the number of observations grows, $\text{E}(\hat{\theta} - \theta) \to 0$ as $N \to \infty$. Roughly, this means that if you had enough (infinite) data, the estimator will give you the true value of the parameter.

2.3.2 Properties of OLS

Assumption	Formal statement	Consequence of violation
No (perfect) collinearity	$\text{rank}(\boldsymbol{X}) = K, K < N$	Coefficients unidentified
$\boldsymbol{X}$ is exogenous	$\text{E}(\boldsymbol{X}\boldsymbol{\varepsilon}) = 0$	Biased, even as $N \to \infty$
Disturbances have mean 0	$\text{E}(\varepsilon) = 0$	Biased, even as $N \to \infty$
No serial correlation	$\text{E}(\varepsilon_i \varepsilon_j) = 0$, $i \neq j$	Unbiased, wrong se
Homoskedastic errors	$\text{E}(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon})$	Unbiased, wrong se
Gaussian errors	$\varepsilon \sim \mathcal{N}(0, \sigma^2)$	Unbiased, se wrong unless $N \to \infty$

Note that these assumptions can be sometimes be written in largely equivalent, but slightly different forms.

When is a variable *endogenous*

1. Omitted variables
2. Measurement error
3. Simultaneity

Assumptions of CLR models

1. No perfect collinearity: No exact linear relationships in the predictors. X is full rank.
2. Linearity: Outcome variable is a linear function of a specific set of independent variables and a disturbance:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

.

[2]It follows from the definition of MSE, that biased estimator, $\hat{\theta}_B$, has a lower MSE than an unbiased estimator, $\hat{\theta}_U$, if $\text{Bias}(\theta_B)^2 < \text{Var}(\theta_U) - \text{Var}(\theta_B)$.

[3]As the number of observations goes to infinity.

3. Observations on independent samples can be considered fixed in repeated samples or X is uncorrelated with the errors.
4. Expected value of the disturbance term is zero.
5. Homoskedasticity: Disturbances have the same variance and are uncorrelated: $\text{Var}(\varepsilon_i) = \sigma^2$, $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j$.
6. Error terms are distributed normal.

- OLS solution exists with unique β: 1
- OLS is unbiased and consistent: 1-4
- OLS is best-linear unbiased estimator (BLUE) Gauss-Markov. Large scale inference. 1-5.
- OLS small scale inference: 1-6. Best unbiased estimator (not just among linear)

Why OLS?

- Computational cost: There exists a closed form solution to the OLS estimate and standard errors.
- Least squares loss: OLS minimizes least squared residuals, and thus is optimal for this criteria. Note, that this is only for within sample.
- Hightest R^2: Follows from the previous.
- Unbiased:
- Best unbiased:
- Mean squared error: OLS is **not** the minimum MSE model.
- Asymptotic criteria: Asymptotically unbiased and consitent.
- Maximum likelihood: OLS is equivalent to the MLE estimator for β.

2.4 References

- Wooldridge (2013), Ch 3.
- Fox (2016), Ch 6, 9.

Chapter 3

OLS Inference

3.1 Sampling Distribution

The sampling distribution of the OLS parameters is

$$\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}).$$

Thus, the variance of the coefficients is

$$\text{Var}(\hat{\beta}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

which is a symmetric matrix,

$$\text{Var}(\hat{\beta}) = \begin{bmatrix} \text{Var}(\hat{\beta}_0) & \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) & \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) & \cdots & \text{Cov}(\hat{\beta}_0, \hat{\beta}_K) \\ \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) & \text{Var}(\hat{\beta}_1) & \text{Cov}(\hat{\beta}_1, \hat{\beta}_2) & \cdots & \text{Cov}(\hat{\beta}_1, \hat{\beta}_K) \\ \text{Cov}(\hat{\beta}_0, \hat{\beta}_2) & \text{Cov}(\hat{\beta}_1, \hat{\beta}_2) & \text{Cov}(\hat{\beta}_2) & \cdots & \text{Cov}(\hat{\beta}_2, \hat{\beta}_K) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(\hat{\beta}_0, \hat{\beta}_K) & \text{Cov}(\hat{\beta}_1, \hat{\beta}_K) & \text{Cov}(\hat{\beta}_K) & \cdots & \text{Var}(\hat{\beta}_k) \end{bmatrix}$$

On the diagonal are the variances of the parameters, and the off-diagonal elements are the covariances of the parameters.

3.2 t-tests for single parameters

The null hypothesis and alternative hypotheses for two-sided tests are,

$$H_0 : \beta_k = \beta_0$$
$$H_a : \beta_k \neq \beta_0$$

Then in large samples,

$$\frac{\hat{\beta}_k - \beta_k}{\text{se}(\widehat{\beta}_k)} \sim \mathcal{N}(0, 1)$$

In small samples,

$$\frac{\hat{\beta}_k - \beta_k}{\text{se}(\widehat{\beta}_k)} \sim \mathcal{T}_{N-(K+1)}$$

The estimated standard errors of $\hat{\beta}$ come from

$$\begin{aligned} \text{Var}(\hat{\boldsymbol{\beta}}) &= \hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ \hat{\sigma}^2 &= \frac{\boldsymbol{\epsilon}'\boldsymbol{epsilon}}{(N-(K+1))} \end{aligned}$$

So, under the common null hypothesis test for $\beta_k = 0$,

$$\frac{\hat{\beta}_k}{\text{se}(\widehat{\beta}_k)} \sim \mathcal{T}_{N-(K+1)}$$

And the confidence intervals for a $(1-\alpha) \times 100$ confidence interval for $\hat{\beta}_k$ are,

$$\hat{\beta}_k \pm t^*_{\alpha/2} \times \text{se}(\hat{\beta}_K)$$

where $t^*_{\alpha/2}$ is the quantile of the $\mathcal{T}_{n-(K+1)}$ distribution such that $P(T \leq t^*) > 1 - \alpha/2$.

3.3 F-tests of Multiple Hypotheses

TODO

Chapter 4

Omitted Variable Bias and Measurement Error

4.1 Omitted Variable Bias

$\text{Cov}(X_1, X_2)$	$\text{Cov}(X_2, Y) > 0$	$\text{Cov}(X_2, Y) = 0$	$\text{Cov}(X_2, Y) < 0$
> 0 $0 < 0$	+ 0 -	0 0 0	- 0 +

4.1.1 What's the problem?

4.1.2 What to do about it?

Summary:

1. OVB is intrinsic to observational methods relying on selection on observables—not just regression.
2. Control for all plausible "pre-treatment" variables
3. Reason about possible biases due to OVB
4. Sensitivity of coefficients to inclusion of control variables is an indication of the plausibility of OVB. Altonji, Elder, and Taber (2005). formalize this.

In practice, this is a primary problem of many papers and papers; and for good reason, it biases the coefficient of interest. Reviewers and discussants will often ask about whether you have considered controlling for *foo*. Although these may be legitimate concerns, not all commenters understand the purpose of control variables. There two arguments to consider when addressing these arguments.

1. The omitted variable has to plausibly be correlated with *both* the variable of interest *and* the outcome variable, and the burden is on the commenter to provide at a confounding variable and plausible relationships. Simpy stating that there could be an unobservable variable is trivially true, uninteresting, and not a fatal critique. That said, the evidentiary content of your methods would be higher if you used methods less susceptible to potential unobserved confounders.
2. The omitted variable should be a *good* control and not a "post treatment" variable. If the omitted variable should not be one of the causal pathways by which X affects Y, it should not be controlled for. If Z affects the values of X and also affects Y, then it needs to be controlled for.

There are two common ways of assessing plausibility.

1. **Informal method**. This is what you see in many empirical papers. Estimate the model including different control variables. The less sensitive the coefficient(s) of the variables of interest are to the inclusion of control variables, the more plausible it is that the variable of interest is not sensitive to unobserved variables (Angrist and Pischke 2014). Oster (2013) states

   > A common heuristic for evaluating the robustness of a result to omitted variable bias concerns is to look at the sensitivity of the treatment effect to inclusion of observed controls. In three top general interest economics journals in 2012, 75% of non-experimental empirical papers included such sensitivity analysis. The intuitive appeal of this approach lies in the idea that the bias arising from the observed controls is informative about the bias that arises from the unobserved ones.

   Note that what is important is that the *coefficient* is stable to the inclusion of controls, not that the coefficient remains statistically significant (which seems to be what many authors focus on).

2. **Formal method** Several papers, including Altonji, Elder, and Taber (2005), Bellows and Miguel (2009), and Oster (2013), formalize the intuition behind the heuristic of coefficient stability to assess the sensitivity of the treatment to OVB.

Altonji, Elder, and Taber (2005) propose a method for assessing the potential impact of the omitted variable bias as the importance of the omitted variable needed to explain away the entire effect. That work addresses the case for a dichotomous treatment variable, and assumed joint normality. Bellows and Miguel (2009) extend the method to continuous treatment variables.

The statistic proposed by Bellows and Miguel (2009) is simple,

$$\delta = \frac{\hat{\beta}_F}{\hat{\beta}_R - \hat{\beta}_C},$$

where δ is

$$\operatorname{Cov} x, \tilde{w}x, w'\gamma$$

δ is interpreted as the how strong the covariance between the unobserved part of the controls and the treatment variable must be relative to the covariance between the observed part of the controls and the treatment variable to explain away the entire effect of x on y. A larger ratio suggests it is implausible that omitted variable bias could explain away the entire observed effect.

Suppose we would like to estimate

$$y = \beta x + \gamma z$$

If z is left out of the OLS estimation, then the estimates of β will have omitted variable bias,

$$\operatorname{plim} \hat{\beta}_{OLS,NC} = \beta + \gamma \frac{\operatorname{Cov}(x, z)}{\operatorname{Var} x}$$

Suppose that instead of z we observe a set of controls w^* that are related to the full set of controls,

$$z = w'\beta + \omega$$

See Appendix A of Bellows and Miguel (2009) for the derivation.

TODO Insert path diagram.

OVB is a intrinsic problem in observational research, and there is nothing you can do to ever ensure that you have controlled for all relevant variables (however, all inference is uncertain, even the designs discussed next, so people should learn to deal with uncertainty). Also, methods such as matching, propensity scores, or inverse weighting still depend on assumptions about selection on observables, even if they may be less sensitive to certain kinds of modeling assumptions. The alternative is to use designs which do not require directly controlling for observable differences. Examples of these designs include: experiments (obviously), natural experiments, instrumental variables, and regression discontinuity.

4.2 Measurement Error

4.2.1 What's the problem?

It biases coefficients:

1. Variable with measurement error: biases β towards zero (**attenuation bias**)
2. Other variables: Biases β similarly to omitted variable bias. In other words, when a variable has measurement error it is an imperfect control. You can think of omitted variables as the limit of the effect of measurement error as it increases.

4.2.2 What to do about it?

There's no easy fix within the OLS framework.

1. If the measurement error is in the variable of interest, then the variable will be biased towards zero, and your estimate is too large.
2. Find better measures with lower measurement errors. If the variable is the variable of interest, then perhaps combine multiple variables into a single index. If the measurement error is in the control variables, then include several measures. That these measure correlate closely increases their standard errors, but the control variables are not the object of the inferential analysis.
3. More complicated methods: errors in variable models, structural equation models, instrumental variable (IV) models, and Bayesian methods.

Chapter 5

Collinearity and Multicollinearity

5.1 (Perfect) Collinearity

In order to estimate unique $\hat{\beta}$ OLS requires the that the columns of the design matrix $\boldsymbol{X}$ are linearly independent.

Common examples of groups of variables that are not linearly independent:

- Categorical variables in which there is no excluded category. You can also include all categories of a categorical variable if you exclude the intercept. Note that although they are not (often) used in political science, there are other methods of transforming categorical variables to ensure the columns in the design matrix are independent.
- A constant variable. This can happen in practice with dichotomous variables of rare events; if you drop some observations for whatever reason, you may end up dropping all the 1's in the data. So although the variable is not constant in the population, in your sample it is constant and cannot be included in the regression.
- A variable that is a multiple of another variable. E.g. you cannot include log(GDP in millions USD) and log($GDPinUSD$) since log(GDP in millions USD) = log($GDPinUSD$)/1,000,000. in
- A variable that is the sum of two other variables. E.g. you cannot include log($population$), log(GDP), log($GDPpercapita$) in a regression since

$$\log(\text{GDP per capita}) = \log(\text{GDP/population}) = \log(\text{GDP}) - \log(\text{population})$$

.

5.1.0.1 What to do about it?

R and most statistical programs will run regressions with collinear variables, but will drop variables until only linearly independent columns in $\boldsymbol{X}$ remain.

For example, consider the following code. The variable `type` is a categorical variable with categories "bc", "wc", and "prof". It will

```
data(Duncan, package = "car")
# Create dummy variables for each category
Duncan <- mutate(Duncan,
                 bc = type == "bc",
                 wc = type == "wc",
```

```
                prof = type == "prof")
lm(prestige ~ bc + wc + prof, data = Duncan)
```

```
##
## Call:
## lm(formula = prestige ~ bc + wc + prof, data = Duncan)
##
## Coefficients:
## (Intercept)       bcTRUE       wcTRUE     profTRUE
##       80.44       -57.68       -43.78           NA
```

R runs the regression, but coefficient and standard errors for `prof` are set to `NA`.

You should not rely on the software to fix this for you; once you (or the software) notices the problem check the reasons it occurred. The rewrite your regression to remove whatever was creating linearly dependent variables in $\boldsymbol{X}$.

5.2 Multicollinearity

Multicollinearity is the (poor) name for less-than-perfect collinearity. Even though there is enough variation in $\boldsymbol{X}$ to estimate OLS coefficients, if some set of variables in $\boldsymbol{X}$ is highly correlated it will result in large, but unbiased, standard errors on the esimates.

What happens if variables are not linearly dependent, but nevertheless highly correlated? If $\text{Cor}(\boldsymbol{x}_1, vecx_2) = 1$, then they are linearly dependent and the regression cannot be estimated (see above). But if $\text{Cor}(\boldsymbol{x}_1, vecx_2) = 0.99$, the OLS can estimate unique values of of $\hat{\beta}$. However, it everything was fine with OLS estimates until, suddenly, when there is linearly independence everything breaks. The answer is yes, and no. As $|\text{Cor}(\boldsymbol{x}_1, \boldsymbol{x}_2)| \to 1$ the standard errors on the coefficients of these variables increase, but OLS as an estimator works correctly; $\hat{\beta}$ and $\text{se}\,\hat{\beta}$ are unbiased. With multicollinearly, OLS gives you the "right" answer, but it cannot say much with certainty.

Insert plot of highly correlated variables and their coefficients.

Insert plot of uncorrelated variables and their coefficients.

5.2.1 What to do about it?

Remember multicollinearity does not violate the assumptions of OLS. If all the other assumptions hold, then OLS is giving you unbiased coefficients and standard errors. What multicollinearity is indicating is that you may not be able to answer the question with the precision you would like.

1. If the variable(s) of interest are highly correlated with other variables, then it means that there is not enough variation, controlling for other factors. You may check that you are not controlling for "post-treatment" variables. Dropping control variables if they are correctly included will bias your estimates. But otherwise, there is little you can do other than get more data. You could re-consider your research design and question. What does it mean if there is that little variation in the treatment variable after controlling for other factors?
2. If control variables are highly correlated with each other, it does not matter. You should not be interpreting their coefficients, so their standard errors do not matter. In fact, controlling for several similar, but correlated variables, may be useful in order to offset measurement error in any one of them.

Chapter 6

Functional Form and Non-linearity

6.1 Non-linearity

6.1.1 What's the problem?

If the relationship between the regression surfacne and $E(Y|X)$ is not captured well, then the results of the regression may be misleading, although this depends on the modeling approach regression is being used for.

The extent of the problem varies with which variables are affected, and the purpose of the analysis.

1. If the analysis is interested in the average marginal effect of the treatment variable, then using the OLS coefficient to estimate the AME is not a bad approximation. The values of the individual marginal effects will be incorrect, but the average should be a reasonable approximation. If you are interested in the AME of sub-populations or other estimands, then you will need to account for the non-linearity.
2. If the non-linearity is in the control variables, then it is another form of omitted variable bias.

6.1.2 What to do about it? And How to Solve it?

The general approaches to identifying non-linearity include:

- Residual plots with curvature tests: **car** function `residualPlots`.
- Added-variable (AV) plot: **car** function `avPlots`.
- Component+residual (CERES) plot: **car** functions `crPlots` and `ceresPlots`.
- Ramsay RESET test. **lmtest** function `resettest`
- Compare Robust SE and classical OLS SE. King and Roberts.

In general, I think most of these approaches are time consuming, sub-optimal given new methods and computation, and open up the regression model to too many researcher degrees of freedom that will not be represented in the uncertainty of the model

There now exist many models (notably semi-parametric and non-parametric) which allow for more flexible functional forms with less-model dependence. Some of these include:

1. GAM and spline models
2. K-nearest neighbor models
3. Matching methods
4. LASSO, Ridge and other Shrinkage Regression (especially with basis functions)

6.2 Logarithm

6.2.1 Examples of Relevant Theories

- Converts multiplicative theories to additive theories. Theories with diminishing returns to scale. Theories about percentage changes or growth.
- Most uses of (per capita) GDP, population:
- Cobb-Douglas growth

$$y = \alpha(K^(\delta)L^(1-\delta))^{\nu}$$

Linearized,

$$\log y = \log \alpha + \log k$$

- Gravity trade equation
- Lanchester law for casualties

$$\Delta x = \alpha x^{\beta} y^{\gamma}$$

where Δx are casualties per period, x is the initial size of forces forces, and y are opposing forces. This can be linearized and estimated with OLS,

$$\log \Delta x = \log \alpha + \beta \log x + \gamma \log y$$

as long as $x, y > 0$ (and preferrably large).

6.3 Miscellaneous

6.3.1 Square Root and Variance Stabalizing Transformations

6.3.2 Power-Transformation

6.4 Polynomials

6.4.1 Squared

6.4.1.1 What theories?

- Kuznets curve: economic development and inequality
- Environmental Kuznets curve: environmental quality and economic development
- Democratic Civil Peace: intermediate regimes prone to civil war, democracies and autocracies are less prone to civil war.

6.4.2 Higher-Order Polynomials

- Time cubed
- Seat-Vote curves? Other old examples in Tufte 1975
- These are generally

6.5 Interactions

Standard errors are more difficult to calculate. See Golder's page, and Aiken and West (1991).

6.5.1 Theories

Golder et. al recommend that for simple interaction model such as:

$$\boldsymbol{y} = \beta_0 + \beta_x \boldsymbol{x} + \beta_z \boldsymbol{z} + \beta_{xz} \boldsymbol{xz} + \boldsymbol{varepsilon}$$

the reseearcher make as many of the following predictions as possible

1. The marginal effect of X is (positive, negative, zero) when Z is at its **lowest** level.
2. … when Z is at its **highest** level.
3. The marginal effect of Z is (positive, negative, zero) when X is at its lowest level.
4. … when X is at its **highest** level.
5. The marginal effect of each X and Z is (positively, negatively) related to the other variable

6.5.2 Recommendations

Golder et al recommend

1. Use multiplicative interaction models for conditional hypotheses.
2. Include all constituent terms of the interaction in the model.
3. Do not interpret coefficients on terms seperately, or as if they are unconditional marginal effects.
4. Calculate substantively meaningful marginal effects and their standard errors.

6.5.3 Plots

Golder et al recommend:

1. Construct marginal effect plots for both X and Z.
2. The range of the horizontal axis should extend from the minimum to the maximum value of variable in the sample.
3. The plot should include a frequency distribution of the variable of interest, as either a rug plot, histogram, or density.
4. Report the product term coefficient and its t-statistic or standard error.

6.6 Flexible Functional Forms

- Splines
- GAM
- Gaussian Processes
- Random Forests
- Neural Networks

6.7 References

- Matt Golder Interactions
- Golder's papers

Chapter 7

Non-constant Variances and Correlated Errors

7.1 iid errors

The OLS coefficient standard errors,

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

depends on the assumption of homoskedastic errors. Homoskedasticity has two components,

1. Disturbances have the same variance, $\text{Var}(\varepsilon_i) = \sigma^2$ for all i.
2. No correlation between disturbances, $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j$.

Either or both of these components can be violated, and when they are, the standard errors of the OLS estimator are incorrect.

The general OLS variance-covariance matrix of the coefficients is,

$$\text{Var}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}\Sigma\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

where ***Sigma*** is the correlation of the disturbances, $\boldsymbol{\varepsilon}$,

$$\boldsymbol{\Sigma} = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2 & \cdots & \sigma_1\sigma_N \\ \sigma_2\sigma_1 & \sigma_2^2 & \cdots & \sigma_2\sigma_N \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_N\sigma_1 & \sigma_N\sigma_2 & \cdots & \sigma_N^2 \end{bmatrix}$$

When we assume homoskedasticity, the variance-covariance matrix of ε is

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = \sigma^2 \boldsymbol{I}_N$$

,

Under homoskedasticity, the sampling distribution of

$$\begin{aligned} \text{Var}(\beta|\boldsymbol{X}) &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\sigma^2\boldsymbol{I}_N\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} \end{aligned}$$

To estimate $\text{Var}(\hat{\boldsymbol{\beta}}|\boldsymbol{X})$, replace σ^2 with $\hat{\sigma}^2$, where

$$\hat{\sigma}^2 = \frac{1}{N-K-1}\sum \varepsilon_i^2$$

Q. What if the assumption of homoskedasticity isn't true?

A. The coefficients $\hat{\boldsymbol{\beta}}$ are unbiased, but the standard errors $\hat{\text{se}}(\hat{\boldsymbol{\beta}})$ are biased.

Since we don't know $\boldsymbol{\Sigma}$, why not simply estimate the elements of $\boldsymbol{\Sigma}$ along with $\boldsymbol{\beta}$? The problem is that there are N observations, and $\boldsymbol{\Sigma}$ is an $N \times N$ matrix with $(N * (N+1))/2$ elements since it is symmetric. So some structure needs to be put on $\boldsymbol{\Sigma}$, i.e. we need to make additional assumptions, to estimate $\boldsymbol{\Sigma}$. As we will see, we can make less restrictive assumptions than homoskedasticity, i.e. $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}$, but will always have to assume some sort of structure in $\boldsymbol{\Sigma}$.

The consequences of violating homoskedasticity are,

- Biased (often downward) standard errors, $\text{E}(\text{se}\,\hat{\beta}) \neq \text{sd}(\beta)$.
- Test statistics do hot have t or F distributions.
- α-level tests have the wrong level of Type I error. E.g. a 5% test will not have 5% Type I errors.
- Confidence intervals do not have the correct coverage. E.g. a 95% confidence does not contain the true mean in 95% of samples.
- OLS is not BLUE.
- $\hat{\boldsymbol{\beta}}$ is still and unbiased and consistent estimator for $\boldsymbol{\beta}$.

So why don't we estimate all the elements of $\boldsymbol{\Sigma}$ like we estimate β? Since $\boldsymbol{\Sigma}$ is an $N \times N$ symmetric matrix, it has $(N \times (N+1))/2$ elements. But we have only N data points to estimate them. In order to estimate $\boldsymbol{\Sigma}$ we cannot estimate arbitrary correlations in $\boldsymbol{\Sigma}$, but we need to apply some structure to the variance-covariance matrix in order to reduce the number of elements to estimate.

1. Heteroskedasticity
2. Clustered Standard Errors
3. Serial Correlation

In general there are two types of methods to deal with issues in the error,

1. New estimators that model the error process and estimate elements of $\boldsymbol{\Sigma}$ simultaneously with the coefficients β. This includes weighted least squares (heteroskedasticity), Prais-Wiston (AR(1) errors). These methods produce $\hat{\beta} \neq \hat{\beta}_{OLS}$.
2. Since OLS produces unbiased and consistent estimates of $\hat{\beta}$, we keep the coefficient estimates, but correct the variance-covariance matrix $\hat{\text{Var}}\beta$.

There's only one way for homoskedasticity to be correct ($\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}$), and many ways for it to be wrong. We'll consider a few of the most common, and methods to deal with them.

1. Heteroskedasticity
2. Autocorrelation
3. Clustering

7.2 Heteroskedasticity

The homoskedastic case assumes that each error term has its own variance. In the heteroskedastic case, each disturbance may have its own variance, but they are still uncorrelated ($\mathbf{\Sigma}$ is diagonal)

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_N^2 \end{bmatrix} = \sigma^2 \boldsymbol{I}_N$$

With homoskedasticity the estimator of the variance covariance matrix takes a particularly simple form,

$$\begin{aligned} \text{Var}(\hat{\beta}|\boldsymbol{X}) &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbf{\Sigma}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= \hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1} \end{aligned}$$

where

$$\hat{\sigma}^2 = \frac{\sum \hat{\epsilon}^2}{N - K - 1}$$

7.2.1 Diagnostics

- Plot residuals vs. fitted values
- Spread-level plots (`car::spreadLevel`),
- Compare Robust SE vs. non-robust SE. If they are different
- Tests: Breusch-Pagan (`lmtest::bptest`, `car::ncvTest`),

7.2.2 Dealing with Heteroskedasticity

1. Transform the dependent variable. For example, log the dependent variable.
2. Model heteroskedasticity using Weighted Least Squares (WLS)
3. Use OLS with an estimator of $\text{Var}(\hat{\boldsymbol{\beta}})$ that is **robust** to heteroskedasticity
4. Admit that OLS is insufficient, and use a different model

- If the form of heteroskedasticity follows a particularly simple form, transform the dependent variable. For example, log the dependent variable.
- If the form of the heteroskedasticity is known: weighted least squares. `lm()` with the `weights` argument.
- If the form of the heteroskedasticity is unknown: Huber-White heteroskedasticity consistent standard errors. See **sandwich** package. You can calculate the heteroskedasticity correct covariance matrix using `sandwich::vcovHC` and then use `lmtest::coeftest` to calculate p-values and standard-errors.

7.2.3 Advice

In practice, often diagnostics are not conducted and robust standard errors are used. This is partially due to the ease with which heteroskedasticity consistent standard errors can be calculated in Stata (see `, robust`).

Robust standard errors, especially when used with MLE estimators, is controversial.

- See Freedman
- See King and Roberts

But this depends on how they are being used, see Angrist.

7.3 Clustered Errors

- Clusters: $g = 1, \ldots, G$.
- Units: $i = 1, \ldots, N_g$.
- N_g is the number of observations in cluster g
- $N = \sum_g N_g$ is the total observations
- Units (usually) belong to a single cluster
  - voters in household
  - individuals in states
  - students in classes
- This is particularly important when the outcome varies at the unit-level, $y_i j$ and the main independent variable varies at the cluster level.
- Ignoring clustering overstates the effective number of individuals in the data.

Clustered dependence

$$\begin{aligned} y_{ig} &= \beta_0 + \beta_1 x_{ig} + \varepsilon_{ig} \\ &= \beta_0 + \beta_1 x_{ig} + \nu_g + \eta_{ig} \end{aligned}$$

Then the cluster error is

$$\nu \sim N(0, rho\sigma^2),$$

and the individual error is

$$\eta_{ig} \sim N(0, (1-\rho)\sigma^2).$$

The cluster and unit errors are assumed to be independent of each other. $\rho \in (0, 1)$ is the *within-cluster correlation*. If we ignore the cluster, and use η_{ig} as the error, the variance is σ^2

$$\begin{aligned} \text{Var}(\eta_{ig}) &= \text{Var}(\nu_g + \eta_{ig}) \\ &= \text{Var}(\nu_g) + \text{Var}(\varepsilon_{ig}) \\ &= \rho\sigma^2 + (1-\rho)\sigma^2 = \sigma^2 \end{aligned}$$

The Covariance between units in the same cluster is

$$\text{Cov}(\varepsilon_{ig}, \varepsilon_{ig}) = \rho\sigma^2,$$

meaning that the correlation for units within a group is

$$\text{Cor}(\varepsilon_{ig}, \varepsilon_{ig}) = \rho.$$

But, there is zero covariance and correlation between units in different clusters. For example, the covariance matrix of

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_{1,1} & \varepsilon_{2,1} & \varepsilon_{3,1} & \varepsilon_{4,2} & \varepsilon_{5,2} \end{bmatrix}'$$

is

$$\text{Var}(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 & \sigma\rho & \sigma\rho & 0 & 0 \\ \sigma\rho & \sigma^2 & \sigma\rho & 0 & 0 \end{bmatrix}$$

More generally, the variance-covariance matrix of a the errors is **block diagonal**,

$$\text{Var}(\varepsilon|\boldsymbol{X}) = \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Sigma}_G \end{bmatrix}$$

where $\boldsymbol{\Sigma}_g$ are the covariance matrices of each cluster, and $\mathbf{0}$ are matrices of zeros of the appropriate sizes.

There are several ways to address clustering, including:

1. Include an indicator variable for each cluster
2. Random effects models
3. Cluster-robust ("clustered") standard error
4. Aggregate data to the cluster data and use WLS with $\bar{y}_g = \frac{1}{N_g} \sum_i y_{ig}$ where the clusters are weighted by N_g.

Cluster-robust standard errors uses the observed residuals, $\hat{\varepsilon}_i$, to estimate a the variance-covariance matrix $\hat{\text{Var}}(\hat{\boldsymbol{\beta}})$ which allows units to be independent across clusters and dependent within clusters.

$$\hat{\Sigma} = \begin{bmatrix} \hat{\varepsilon}_1^2 & 0 & \cdots & 0 \\ 0 & \hat{\varepsilon}_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{\varepsilon}_N^2 \end{bmatrix} = \hat{\boldsymbol{\varepsilon}} \boldsymbol{I}_N \hat{\boldsymbol{\varepsilon}}'$$

- cluster robust standard errors do not change $\hat{\boldsymbol{\beta}}$. Thus they do not fix bias in the coefficients.
- cluster robust standard errors is a consistent estimator of $\text{Var}\,\hat{\boldsymbol{\beta}}$ given the clustered dependence.
  - This relies on the assumption of independence between clusters
  - Does not rely on the model form
  - CRSE are usually larger than classic standard errors
- Consistency of the CRSE are in the number of groups, not the number of individuals
  - CRSE work well when the number of **clusters** is large (> 50)
  - Alternative: use a block bootstrap

See the R package **plm** (Panel linear models in R).

See Cameron and Miller, Practioner's Guide to Cluster-Robust Inference.

See clusterSEs package for implementations of several clustered standard error methods appropriate for small numbers of clusters.

7.4 Serial Correlation

More general case allows for heteroskedasticity, and auto-correlation ($\text{Cov}(\varepsilon_i, \varepsilon_j) \neq 0$),

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \cdots & \sigma_{1,N} \\ \sigma_{2,1} & \sigma_2^2 & \cdots & \sigma_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{N,1} & \sigma_{N,2} & \cdots & \sigma_N^2 \end{bmatrix}$$

As with heteroskedasticity, OLS with be unbiased, but the standard errors will be incorrect.

Tests

- Breusch-Godfrey Test (`lmtest::bgtest`)

Solution

- If the form is known: Prais-Wiston, include lagged dependent variable.
- Huber-White Heteroskedasticity and Autocorrelation Robust standard errors. These are an extension of the heteroskedasticity robust standard errors to also include autocorrelation. See **sandwich** function `hcacVCOV`.

7.5 Non-Normal Errors

In addition to assuming that the errors are iid, the classical linear regression assumptions include that the errors are distributed normal,

$$\varepsilon_i \sim N(0, \sigma^2).$$

What happens if this assumption is violated?

1. If the sample size is small, normal errors are required for correct confidence interval coverage and p-values in tests. In large samples, CLT properties of OLS kick in and the normality of errors assumption is not needed to justify the sampling distributions of the test statistics.
2. Heavy-tailed errors threaten the efficiency of OLS estimate.
3. Skewed or multi-modal errors suggest that the conditional mean $E(Y|\boldsymbol{X})$, estimated by OLS may not be a good summary of the data.

How to diagnose non-normal errors? Plot the quantiles of the *studentized* residuals (see the section on outliers) against the expected quantiles of a normal distribution using a QQ-plot. A visual test should be sufficient; there is no need for a formal test.

In general, non-normal errors are a minor issue, and towards the bottom in priority of problems in inference.

How to fix it?

1. Transform the dependent variable so that the erorrs are closer to normally distributed and use OLS.
2. Add different sets of covariates. This is especially likely with multi-modal error distributions, which could suggest an omitted categorical variable.
3. Use a different model other than OLS.

R Calculate Studentized residuals with the function `rstudent` and a QQ-plot using stat_qq in **ggplot2**.

Diagnostics

- QQ-plot of the Studentized residuals

Important things to remember:

- The assumption is not that Y has a normal distribution, it is that the errors *after* including covariates are normal.
- While non-normal errors will not bias β and have little effect on the standard errors unless the sample size is small, they could serve as a warning that your model is mis-specified, or that the conditional expectation of Y is not good summary.

Chapter 8

Weighting in Regression

8.1 Weighted Least Squares (WLS)

Ordinary least squares estimates coefficients by finding the coefficients that minimize the sum of squared errors,

$$\hat{\beta}_{OLS} = \arg\min_{\boldsymbol{b}} \sum_{i=1}^{N} (y_i - \boldsymbol{x}'\boldsymbol{b})^2.$$

In the objective function, it treats the errors of all observations equally. However, there may be situations where we are more concerned about minimizing some errors more than others. For example, suppose we know that some y_i have more measurement error than others, then we may care more about minimizing errors for those y_i which we are more certain about.

In weighted least squares (WLS) we estimate the coefficients by finding the values that minimize the *weighted* sum of squared errors,

$$\hat{\beta}_{WLS} = \arg\min_{\boldsymbol{b}} \sum_{i=1}^{N} w_i (y_i - \boldsymbol{x}'\boldsymbol{b})^2,$$

where w_i are the weights for each observation. Note that OLS is a special case of WLS where $w_i = 1$ for all the observations.

8.2 When should you use WLS?

The previous section showed what WLS is, but when should you use weighted regression?

Suppose we have weights for observations, when should we use them? Well, it depends on the purpose of your analysis:

1. If you are estimating population descriptive statistics, then weighting is needed to ensure that the sample is representative of the population.
2. If you are concerned with causal inference, then weighting is more nuanced. You may or may not need to weight, and it will often be unclear which is better.

There are three reasons for weighting in causal inference (Solon, Haider, and Wooldridge 2015):

1. Correct standard errors for heteroskedasticity
2. Get consistent estimates by correcting for endogenous sampling

3. Identify average partial effects when there is unmodeled heterogeneity in the effects.

Heteroskedasticity: Estimate OLS and WLS. If the model is misspecified or there is endogenous selection, then OLS and WLS have different probability limits. The constrast between OLS and WLS estimatates is a diagnostic for model misspecification or endogenous sampling. Always use robust standard errors.

Endogenous sampling: If the sample weights vary exogenously instead of endogenously, then weighting may be harmful for precision. The OLS still specifies the conditional mean. Sampling is exogenous if the sampling probabilities are independent of the error - e.g. if they are only functions of the explanatory variables. If the probabilities are a function of the dependent variable, then they are endogenous. (1) if sampling rate is endogenous, weight by inverse selection. (2) use robust standard errors. (3) if sampling rate is exogenous, then OLS and WLS are consistent. Use OLS and WLS as test of model mispecification.

Heterogeneous effects: Identifying average partial effects. WLS estimates the linear regression of the population, but this is not the same as the average partial effects. But that is because OLS does not estimate the average partial effect, but weights according to the variance in X.

8.2.1 Correcting for Known Heteroskedasticity

When are there cases with known heteroskedasticity? This is probably rare, but it arises in a few circumstances:

1. The outcome variable consists of measurements with a given measurement error. For example, the y come from instruments or are estimated themselves.
2. The error of output depends on input variables in known ways. For example, the sampling error of polls.

Suppose that the heteroskedasticity is known up to a multiplicative constant,

$$\text{Var}(\varepsilon_i|\boldsymbol{X}) = a_i\sigma^2,$$

where $a_i = a_i\boldsymbol{x}_i'$ is a positive and known function of $\boldsymbol{x}_i$.

Then in weighted least squares multiply y_i by $1/\sqrt{a_i}$,

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K + \varepsilon_i \\ y_i/\sqrt{a_i} &= \beta_0/\sqrt{a_i} + \beta_1 x_1/\sqrt{a_1} + \beta_2 x_2/\sqrt{a_2} + \cdots + \beta_K x_K/\sqrt{a_K} + \varepsilon_i^* \end{aligned}$$

where $\varepsilon_i^* \sim N(0, \sigma^2)$. This rescales errors to $\varepsilon_i/\sqrt{a_i}$ which keeps $\text{E}(\varepsilon_i) = 0$, but makes the variance constant,

$$\text{Var}\left(\frac{1}{\sqrt{a_i}}\varepsilon_i|\boldsymbol{X}\right) = \frac{1}{a_i}\text{Var}(\varepsilon_i|\boldsymbol{X}) = \frac{1}{a_i}a_i\sigma^2 = \sigma^2$$

If a_i is known, then the model is homoskedastic and the estimator is BLUE.

Define the weighting matrix,

$$\boldsymbol{W} = \begin{bmatrix} 1/\sqrt{a_1} & 0 & \cdots & 0 \\ 0 & 1/\sqrt{a_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & 1/\sqrt{a_N} \end{bmatrix}.$$

Then run the regression,

$$\begin{aligned} \boldsymbol{W}y &= \boldsymbol{W}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{W}\boldsymbol{\varepsilon} \\ \boldsymbol{y}^* &= \boldsymbol{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}^*. \end{aligned}$$

Run the regression of $\boldsymbol{y}^*$ on $\boldsymbol{X}^*$, and the Gauss-Markov assumptions are satisfied. Then using the usual OLS formula,

$$\hat{\boldsymbol{\beta}}_{WLS} = ((\boldsymbol{X}^*)'\boldsymbol{X}^*)(\boldsymbol{X}^*)'\boldsymbol{y}^* = (\boldsymbol{X}'\boldsymbol{W}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}'\boldsymbol{W}\boldsymbol{y}.$$

In R Use `lm()` with the `weights` argument.

8.3 Sampling Weights

Using sampling weights is most important for univariate statistics which are estimates of population parameters. However, whether to use them when estimating a regression is less clear.

- if sample weights are a function of X only, estimates are unbiased and more efficient without weighting
- if the sample weights are a function of $Y|X$, then use the weights

With fixed X, regression does not require random sampling, so the sampling weights of the X are irrelevants.

If the original unweighted data are homoskedastic, then sampling weights induces heteroskedasticity. Suppose the true model is,

$$Y_i = \boldsymbol{x}'\boldsymbol{\beta} + \varepsilon_i$$

where $\varepsilon_i \sim N(0, \sigma^2)$. Then the weighted model is,

$$\sqrt{w_i}Y_i = \sqrt{w_i}\boldsymbol{x}'\boldsymbol{\beta} + \sqrt{w_i}\varepsilon_i$$

and now $\sqrt{w_i}\varepsilon_i \sim N(0, w_i\sigma^2)$.

If the sampling weights are only a function of the X, then controlling for X is sufficient. In fact, OLS is preferred to WLS, and will produce unbiased and efficient estimates. The choice between OLS and WLS is a choice between different distributions of $\boldsymbol{X}$. However, if the model is specified correctly the coefficients should be the same, regardless of the distribution of $\boldsymbol{X}$. Thus, if the estimates of OLS and WLS differ, then it is evidence that the model is misspecified.

Winship and Radbill (1994) suggest using the method of Dumouchel and Duncan (1983) to test whether the OLS and WLS are difference.

1. Estimate $E(Y) = \boldsymbol{X}\beta$
2. Estimate $E(Y) = \boldsymbol{X}\beta + \delta\boldsymbol{w} + \boldsymbol{\gamma}\boldsymbol{w}\boldsymbol{X}$, where all X
3. Test regression 1 vs. regression 2 using an F test.
4. If the F-test is significant, then the weights are not simply a function of X. Either try to respecify the model or use WLS with robust standard errors. If the F-test is insignificant, then the weights are simply a function of X. Use OLS.

Modern survey often use complex multi-stage sampling designs. Like clustering generally, this will affect the standard errors of these regressions. Clustering by primary sampling units is a good approximation of the standard errors from multistage sampling.

8.4 References

- WLS derivation Fox (2016) 304–306

Textbook discussions: Angrist and Pischke (2009) 91–94, Angrist and Pischke (2014) 202–203,

Solon, Haider, and Wooldridge (2015) is a good (and recent) overview with practical advice of when to weight and when not-to weight linear regressions. Also see the advice from the World Bank blog.

Gelman (2007b), in the context of post-stratification, proposes controlling for variables related to selection into the sample instead of using survey weights; also see the responses (Bell and Cohen 2007; Breidt and Opsomer 2007; Little 2007; Pfeffermann 2007), and rejoinder (Gelman 2007a) and blog post. Gelman's approach is similar to that earlier suggested by Winship and Radbill (1994).

See also Deaton (1997), Dumouchel and Duncan (1983), and Wissoker (n.d.).

For survey weighting, see the R package survey and its

Chapter 9

Interpreting Regression Coefficients

9.1 Interpreting Coefficients

Consider the regression,

$$Y_i = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon$$

The regression coefficient β_1 it the change in the expected value of Y associated with a one-unit change in X holding Z constant,

$$\begin{aligned} E(Y|X = x, Z = z) - E(Y|X = x + 1, Z = z) &= (\beta_0 + \beta_1 X + \beta_2 z) - (\beta_0 + \beta_1(x + 1) + \beta_2 z) \\ &= \beta_1 x - \beta_1(x + 1) \\ &= \beta_1(x - x + 1) \\ &= \beta_1 \end{aligned}$$

More formally, the coefficient β_k is the partial derivative of $E(Y|X)$ with respect to X_k,

$$\frac{\partial\, E(Y|X)}{\partial\, X_k} = \frac{\partial}{\partial\, X_k}\left(\beta_0 + \sum_{k=1}^{K} \beta_k X_k)\right) = \beta_k$$

Implications:

1. If X is multiplied by a constant scalar a,

$$E(Y|X) = \tilde{\beta}_0 + \tilde{\beta}_1 aX = \beta_0 + (a\beta_1)X.$$

2. If X_k has a scalar a added to it,

$$E(Y|X) = \tilde{\beta}_0 + \tilde{\beta}_1(X + a) = (\beta_0 + \tilde{\beta}_1 a) + \tilde{\beta}_1 X$$

   Thus, $\tilde{\beta}_0 = (\beta_0 + \beta_1 a)$ and $\tilde{\beta}_1 = \beta_1$.

9.2 Finite Differences and Marginal Effects

A marginal effect is the effect on the conditional mean of the outcome variable, y, from a change in a predictor, x.

The marginal effect is the derivative of the regression function with respect to a predictor. The marginal effect of X is

$$\begin{aligned}\frac{\partial E(Y)}{\partial X} &= \frac{1}{\partial X}(\beta_0 + \beta_1 X + \beta_2 Z) \\ &= \beta_1\end{aligned}$$

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \varepsilon_i$$

The partial effects , or first differences, of a regression is the change in the expected value of Y with respect to a discrete change of X:

$$\begin{aligned}\frac{\Delta E(Y)}{\Delta X} &= E(Y|X = x + d) - E(Y|X = x) \\ &= \beta_0 + \beta_1(x + d) + \beta_2 Z_i - (\beta_0 + \beta_1 x + \beta_2 Z_i) \\ &= \beta_1 d\end{aligned}$$

When $d = 1$ (a one unit change in X) the first difference is β_1, the coefficient of X.

That the coefficients are the marginal effects of each predictor makes linear regression particularly easy to interpret. However, this interpretation of predictors becomes more complicated once a variable is included in multiple terms through interactions or nonlinear functions, such as polynomials.

9.3 Standardized Coefficients

A standardized coefficient is the coefficient on X, when X is standardized so that $\text{mean}(X) = 0$ and $\text{Var}(X) = 1$. In that case, β_1 is the change in $\text{E}(Y)$ associated with a one standard deviation change in X.

Additionally, if all predictors are set so that $\text{mean}(X) = 0$, β_0 is the expected value of Y when all X are at their means. However, if any variables appear in multiple terms, then the standardized coefficients are not particularly useful.

Standardized coefficients are generally not used in political science. (King How Not to Lie with Statistics, p. 669) More often, the effects of variables are compared by the first difference between the value of the variable at the mean, and a one standard deviation change. While, this is equivalent to the standardized coefficient

Note, that standardizing variables can help computationally in some cases. In OLS, there is a closed-form solution, so iterative optimization algorithms are not needed in to find the best parameters. However, in more complicated models which require iterative optimization, standardizing variables can often improve the performance of the optimization. Thus standardizing variables before analysis is common in machine learning. However, the purpose is for ease of computation, not for ease of interpretation.

9.4 Marginal Effects and First Difference

The marginal effect of a regressor is the change in the outcome variable associated with a small change in the predictor,

The **first difference** of a regression of a variable with respect to the dependent variable is for two values of a variable x_j and $x_j + h$ as,

$$\frac{\Delta \boldsymbol{y}}{\Delta x} = y(x_j) - y(x)$$

The **marginal effect** is the regression coefficient,

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \beta_1.$$

For a predictor with only a linear term,

$$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x}_1,$$

However, if a predictor appears in multiple terms, the marginal effect will be a more complicated function. For example, if x appears as a squared term and an interaction,

$$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x} + \beta_2 \boldsymbol{x}^2 + \beta_3 \boldsymbol{z} + \beta_4 \boldsymbol{x}\boldsymbol{z} + \beta_5 \boldsymbol{x}\boldsymbol{z}^2$$

then its marginal effect is a function of both its current value, $\boldsymbol{x}$, and the value of $\boldsymbol{z}$,

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \beta_1 + 2\beta_2 \boldsymbol{x} + \beta_4 \boldsymbol{z} + 2\beta_5 \boldsymbol{x}\boldsymbol{z}$$

9.5 References

The R package marfx is still under development, but includes the ability to estimate average marginal and finite difference effects.

Some useful R packages

- **ggplot2** and **broom**: Once point estimates and the confidence interval are calculated, it is easy to plot them.
- coefplot: Plots point estimates and confidence intervals for fitted models.
- dotwhisker: Another point-estimate and confidence interval plot pacakge for fitted models.

Packages for marginal effects or similar

- margins port of the margins command
- Built-in `predict()` function calculates predictions and confidence intervals
- mfx marginal effects for beta, logit, negative binomial, poisson and probit models.
- **erer**: `maBina`: marginal effects for binary choice models; `ocME`: marginal effects for ordered logit or probit.
- **car**: `mmps`: Marginal model plots.

Chapter 10

Regression Diagnostics

Several packages in R provide large collections of regression diagnostics:

- lmtest
- car

Reading the vignettes or documentation of these packages is a good overview of available regression diagnostics. Also see the Econometrics Task View.

Fox (2016) has a particularly extensive overview of regression diagnostics.

Though for Stata, this tutorial has an overview of many regression diagnostics.

Chapter 11

Resampling Methods

11.1 Bootstrap

TODO

11.2 Cross-Validation

11.2.1 Bias-Variance Tradeoff

Error in a model can come from two places

1. Bias: On average, how close is a model to the estimand.
2. Variance: Between samples, how variable are the estimates of a model.

In a regression model we are trying to pedict a Y with covariates X, given a function $Y = f(X) + \varepsilon$, where $f(X)$ is the conditional expectation function, $f(X) = E(Y|X)$, and ε is an error term. Now suppose we estimate $f(X)$ by $\hat{f}(X)$. In OLS, $\hat{y} = \hat{f}(X) = \hat{\beta}(X)$.

TODO MSE and bias-variance trade-off

Diagram and examples of bias variance trade off

11.2.2 Cross-Validation

Diagram of cross validation

Example of cross validation

Difficulties of cross validation

- time series
- categorical variables / fixed effects
- missing data

11.2.3 Other Quantities

11.2.3.1 Information Criteria

There are a class of criteria that we'll call information criteria. They all take the the log-likelihood of a model but apply a penalty increasing with the complexity of the model,

$$IC* = -\mathcal{L} + \text{penalty}$$

Models with lower IC* are preferred.

The *Akaike Information Criterion* (AIC) is defined as

$$AIC = -2 \log \mathcal{L} + 2K$$

where K is the number of parameters. As $N \to \infty$, the AIC is equivalent to minimizing the leave-one-out cross validation value. So the AIC is useful for model selection when the purpose is prediction.

The *Bayesian Information Criterion* (BIC) is defined as,

$$BIC = -2 \log \mathcal{L} + K \log N$$

where K is the number of parameters, and N is the number of observations. Note that the penalty for BIC is larger than the one for AIC, since $2K < K \log N$ for any $N > 7$. The penalty for BIC also increases with the sample size, unlike the AIC. Like AIC, BIC is equivalent to a leave v out cross-validation where $v = n(1 - 1/(\log N - 1))$. The nice feature of the BIC is that, unlike AIC, it is consistent; if there is enough data, the BIC will select the true model. However, the probability that your model is the true model is about zero, so I am skeptical about the importance of this property in practice.

11.2.4 Others

There are a few other statistics to keep in mind:

- Mallows C_p
- Generalized Cross Validation (GCV)

These statistics generally only apply to linear regression. Since they are approximations to cross-validation, their relative advantage has declined as computational power has increased. However, as our data has also increased, these approximations may be useful for big data or computationally intensive jobs.

11.2.5 References

- Hyndsight

Chapter 12

Panel (Longitudinal) Data

In these methods there are repeated measurements of the same unit over time. This requires different methods and also has implications for causal inference. While simply having panel data does not identify an effect, it allows the researcher to claim identification using different assumptions than simply selection on observables (as in the cross-sectional case).

12.1 Terminology

There are several closely related concepts and terminology to cover.

Panel (lognitudinal) data small T, large N. Examples: longitudinal surveys with a few rounds.

Time series cross-section data large T, medium N. Examples: most country-year panels in CPE/IPE with several decades of data.

For the purposes of causal inference, identification relies on the same assumptions. However, different estimators work differently under different data types. Some estimators work well as $N \to \infty$, some as $T \to \infty$, and usually these are not the same. Additionally, longer time series may require and/or have enough data for the researcher to estimate serial correlation in the errors.

There are some additional related concepts that should also be mentioned at this time, hopefully to spare the reader future confusion (and not to add to it):

Hierarchical Models units nested within groups. E.g. children in schools, districts within states

Time-series Models large T, usually $N = 1$, or the different units modeled separately.

Terminology can be confusing and varies across fields and literatures. In particular, fixed effects and random effects are used differently and often estimated differently in statistics and econometrics. This is easily seen by comparing the **lme4** and **plm** packages in R which bost estimate fixed and random effects models. Hierarchical models will often used fixed and random effects even though there is no *time* component, and thus they are not longitudinal models. The reason that I bring up this terminology is that if you search for fixed and random effects you can quickly be confused when it seems that people are talking about seemingly different concepts; they more of less may be.

12.2 Fixed Effects

In a fixed effects model, each unit i has its own effect,

$$Y_{i,t} = X'_{i,t}\beta + u_i + \varepsilon_{i,t}.$$

This means that if instead of estimating the equation above, we estimate the pooled OLS model,

$$y_{i,t} = \boldsymbol{x}'_{i,t}\hat{\boldsymbol{\beta}}_{\text{pool}} + \tilde{\varepsilon}_{i,t},$$

the estimate of $\hat{\boldsymbol{\beta}}_{\text{pool}}$ will be biased if $\text{Cov}(\boldsymbol{u}, \boldsymbol{x}_k)$ for any of the covariates and $\text{Cov}(\boldsymbol{u}, \boldsymbol{y})$. This is a case of omitted variable bias, where the unit effects, u_i, are the omitted variables.

12.2.1 Estimating

There are a couple of approaches

12.2.2 Causal Inference

TODO

12.3 Lagged Dependent Variables

A different model is to assume a lagged dependent variable,

$$Y_{i,t} = \rho Y'_{i,t-1} + X'_{i,t}\beta + \varepsilon_{i,t}$$

This captures some of the unit-specific aspects that the fixed effects capture. However, the LDV model is making a different assumption than fixed effects. The FE model assumes that each unit has a seperate effect that is constant over time, while the LDV model assumes that anything specific about a unit is captured through the value of the dependent variable in the previous period.

Beck and Katz recommendation of LDV with PCSE.

The LDV and Fixed Effects models make different assumptions, and they are not nested. So why not combine them into a single model?

$$Y_{i,t} = \rho Y'_{i,t-1} + X'_{i,t}\beta + \alpha_i + \varepsilon_{i,t}.$$

There is a problem with this approach. OLS is biased. The fixed effect estimator includes demeaned values of the outcome variable and covariates. So the FE model with a LDV will use $Y_{i,t-1} - \bar{Y}_{i,t-1}$. This average includes $Y_{i,t}$ and $Y_{i,t} = ... + \varepsilon_{i,t}$. Thus by construction, $Y_{i,t} - \bar{Y}_{i,t-1}$ is correlated with the errors.

So what can we do about this? There are two options.

1. Ignore it. The bias is proportional to $1/T$. In panels with 20 or more periods, the bias may be small. Moreover, the bias is generally largest in the coefficient of the lagged dependent variable itself, which may not be of primarny interest. Accept the bias.
2. Use both LDV and FE models. The LDV and FE methods can bound the effects of the coefficient of interest. See Angrist and Pischke.
3. Use IV methods to instrument the lagged dependent variable. See Arrellano-Bond methods.

This is a case where the difference between panel and TSCS is important. In many TSCS settings with larger T it is probably fine to estimate fixed effects with LDV. However, if you have panel data model with few T, then you should use either method 2 or 3.

12.4 Random Effects

Consider the panel data model,

$$Y_{i,t} = \alpha + X'_{i,t}\beta + u_i + \varepsilon_{i,t}$$

In fixed effect, the errors are assumed to be uncorrelated with both the unit effects and the covariates,

$$\mathrm{E}(\varepsilon_{i,t}|X_i, u_i) = 0.$$

With random effects we make an additional assumption, the unit effects are uncorrelated with the covariates,

$$\mathrm{E}(u_i|X_i) = \mathrm{E}(u_i) = 0.$$

What this means that under the assumptions of random effects, omitting u_i would not bias β since they are assumed to be uncorrelated with X. Thus, there's no omitted variable bias.

So why use random effects? To fix standard errors.

$$Y_{i,t} = X'_{i,t}\beta + \nu_i$$

where $\nu_i = u_i + \varepsilon_{i,t}$. However, this means that

$$\mathrm{Cov}(Y_{i,1}, Y_{i,2}|X_{i,t}) = \sigma_u^2.$$

This violates the OLS assumption of non-autocorrelation. Using random effects gets consistent standard errors.

12.4.1 How to estimate random effects?

There are a variety of methods, but the econometric method is to use **quasi-demeaning** or **partial pooling**,

$$(Y_{i,t} - \theta\bar{Y_i}) = (X_{i,t} - \bar{X_i})'\beta + (\nu_{i,t} - \theta \operatorname{Var} \nu_i)$$

where $\theta \in [0, 1]$ where $\theta = 0$ is OLS, and $\theta = 1$ is fixed effects. Some math (TM) shows,

$$\theta = 1 - \left(\sigma_u^2/(\sigma_u^2 + T\sigma_e^2 psilon)\right)^{1/2}.$$

The **random effects estimator** runs pooled OLS on this model, but replaces θ with the estimate $\hat{\theta}$.

See the R package **plm**.

The R package **lme4** and Bayesian methods, e.g. Gelman and Hill, take a different approach to estimating random effects.

12.5 Difference in Difference Estimators

TODO

12.6 Non-standard Error Issues

- Panel Corrected Standard Errors
- Clustered Standard Errors

12.7 References

- Matt Blackwell Gov 2002: 8. Panel Data

Chapter 13

Appendix

13.1 Multivariate Normal Distribution

The multivariate normal distribution is the generalization of the univariate normal distribution to more than one dimension.[1] The random variable, $\boldsymbol{x}$, is a length k vector. The k length vector $\boldsymbol{\mu}$ are the means of $\boldsymbol{x}$, and the $k \times k$ matrix, $\boldsymbol{\Sigma}$, is the variance-covariance matrix,

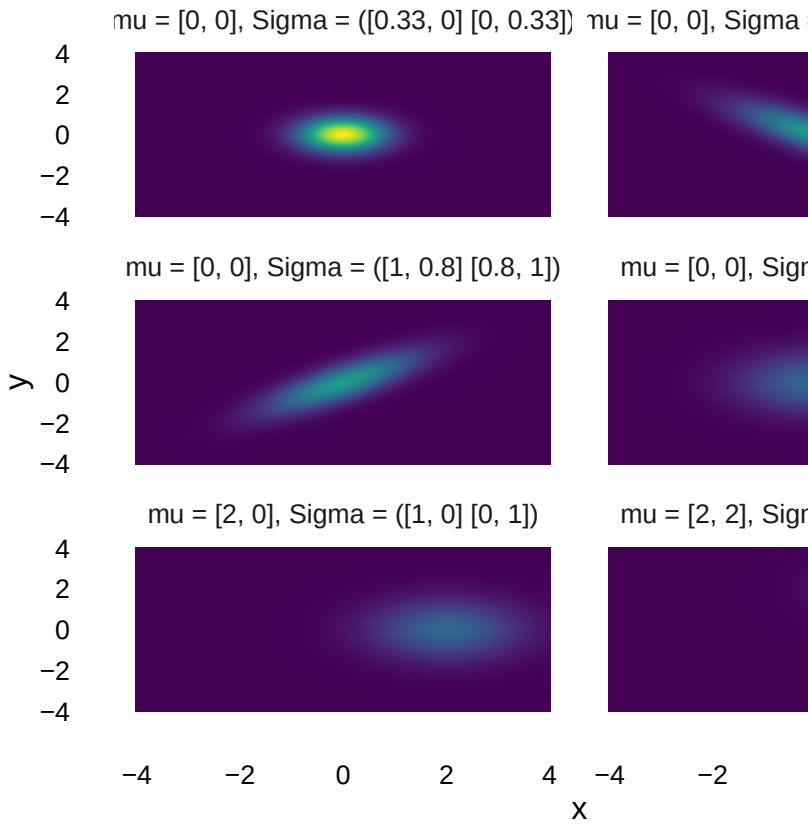
$$\begin{aligned} \boldsymbol{x} &\sim \mathcal{N}_k\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \\ \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix} &\sim \mathcal{N}_k \left(\begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \cdots & \sigma_{1,k} \\ \sigma_{2,1} & \sigma_2^2 & \cdots & \sigma_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{k,1} & \sigma_{k,2} & \cdots & \sigma_{k,k} \end{bmatrix} \right) \end{aligned}$$

The density function of the multivariate normal is,

$$p(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2k)^{-\frac{k}{2}} \left|\boldsymbol{\Sigma}\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right).$$

You can sample from and calculate the density for the multivariate normal distribution with the functions `dmvnorm` and `rmvnorm` from the package **mvtnorm**.

[1] See Multivariate normal distribution and references therein.



Density plots of different bivariate normal distributions,

Chapter 14

References

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber. 2005. *Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools. Journal of Political Economy.* Vol. 113. 1. University of Chicago Press. http://www.jstor.org/stable/10.1086/426036.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion.* Pr.

———. 2014. *Mastering 'Metrics.* Princeton UP.

Bell, Robert M., and Michael L. Cohen. 2007. "Comment: Struggles with Survey Weighting and Regression Modeling." *Statist. Sci.* 22 (2). The Institute of Mathematical Statistics: 165–67. doi:10.1214/088342307000000177.

Bellows, John, and Edward Miguel. 2009. "War and Local Collective Action in Sierra Leone." *Journal of Public Economics* 93 (11–12): 1144–57. doi:http://dx.doi.org/10.1016/j.jpubeco.2009.07.012.

Breidt, F. Jay, and Jean D. Opsomer. 2007. "Comment: Struggles with Survey Weighting and Regression Modeling." *Statist. Sci.* 22 (2). The Institute of Mathematical Statistics: 168–70. doi:10.1214/088342307000000195.

Deaton, Angus. 1997. *The Analysis of Household Surveys : A Microeconometric Approach to Development Policy.* The World Bank. http://documents.worldbank.org/curated/en/1997/07/694690/analysis-household-surveys-microeconometric-approach-development-policy.

Dumouchel, William H., and Greg J. Duncan. 1983. "Using Sample Survey Weights in Multiple Regression Analyses of Stratified Samples." *Journal of the American Statistical Association* 78 (383): 535–43. doi:10.1080/01621459.1983.10478006.

Fox, John. 2016. *Applied Regression Analysis & Generalized Linear Models.* 3rd ed. Sage.

Gelman, Andrew. 2007a. "Rejoinder: Struggles with Survey Weighting and Regression Modeling." *Statist. Sci.* 22 (2). The Institute of Mathematical Statistics: 184–88. doi:10.1214/088342307000000203.

———. 2007b. "Struggles with Survey Weighting and Regression Modeling." *Statist. Sci.* 22 (2). The Institute of Mathematical Statistics: 153–64. doi:10.1214/088342306000000691.

Little, Roderick J. 2007. "Comment: Struggles with Survey Weighting and Regression Modeling." *Statist. Sci.* 22 (2). The Institute of Mathematical Statistics: 171–74. doi:10.1214/088342307000000186.

Oster, Emily. 2013. "Unobservable Selection and Coefficient Stability: Theory and Validation." Working Paper 19054. NBER.

Pfeffermann, Danny. 2007. "Comment: Struggles with Survey Weighting and Regression Modeling." *Statist. Sci.* 22 (2). The Institute of Mathematical Statistics: 179–83. doi:10.1214/088342307000000168.

Solon, Gary, Steven J. Haider, and Jeffrey M. Wooldridge. 2015. "What Are We Weighting for?" *Journal*

of Human Resources 61 (2). [Wiley, International Statistical Institute (ISI)]: 317–37. http://www.jstor.org/stable/1403631.

Winship, Christopher, and Larry Radbill. 1994. "Sampling Weights and Regression Analysis." *Sociological Methods & Research* 23 (2): 230–57. doi:10.1177/0049124194023002004.

Wissoker, Douglass. n.d. "Notes on Weighting in Regression." http://anfdata.urban.org/sdaweb/nsaf_tutorial/reg_weights.pdf.

Wooldridge, Jeffrey M. 2013. *Introductory Econometrics: A Modern Approach*. 5th ed. South-Western.