POLS 503: Advanced Quantitative Political Methodology: The Notes

Jeffrey B. Arnold

2016-05-04

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 $\beta_1, \dots, \beta_K$ 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,

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

or, in matrix notation,

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

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

[1] Ordinary least squares is distinguished from *generalized least squares* (GLS).

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

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

$$\begin{aligned} \hat{\beta}_1 7 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \\ = \frac{\mathrm{Cov}(\boldsymbol{x}\boldsymbol{y})}{\mathrm{Var}\,\boldsymbol{x}} &= \frac{\text{Sample covariance betweeen } \boldsymbol{x} \text{ and } \boldsymbol{y}}{\text{Sample variance of } \boldsymbol{x}}. \end{aligned}$$

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 $(\bar{\boldsymbol{y}}, \bar{\boldsymbol{x}}_1, \dots, \bar{\boldsymbol{x}_K})$. 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,
$$\mathrm{Bias}(\hat{\theta}) = \mathrm{E}(\hat{\theta}) - \theta.$$

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

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.

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

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, $\mathrm{E}(\hat{\theta} - \theta) \to 0$ 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

- When is OLS unbiased?
- When is OLS consistent?
- When is OLS efficient?

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

2.4 Multi-Collinearity

2.4.1 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.

[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 $\mathrm{Bias}(\theta_B)^2 < \mathrm{Var}(\theta_U) - \mathrm{Var}(\theta_B)$.

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

- 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 regresion since $\log(GDPpercapita) = \log(GDP/pop) = \log(GDP) - \log(pop)$.

2.4.1.1 What to do about it?

R and most statistical programs will drop variables from the regression until only linearly independent columns in $\boldsymbol{X}$ remain. You should not rely on the softward to fix this for you; once you (or the software) notices the problem check the reasons it occured. The rewrite your regression to remove whatever was creating linearly dependent variables in $\boldsymbol{X}$.

2.4.2 Less-than Perfect Collinearity

What happens if variables are not linearly dependent, but nevertheless highly correlated. If $\mathrm{Cor}(\boldsymbol{x}_1, vecx_2) = 1$, then they are linearly dependent and the regression cannot be estimated (see above). But if $\mathrm{Cor}(\boldsymbol{x}_1, vecx_2) = 0.99$, 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 $|\mathrm{Cor}(\boldsymbol{x}_1, \boldsymbol{x}_2)| \to 1$ the standard errors on the coefficients of these variables increase, but OLS as an estimator works correctly; $\hat{\beta}$ and $\mathrm{se}\,\hat{\beta}$ 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.

2.5 Weighted Least Squares

In weighted least squares (WLS), instead of minimizing the sum of squared errors, minimize a weighted sum of squared errors,

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

OLS is the special case of WLS in which all observations are weighted equally.

What reasons are there to use WLS?

- Heteroskedasticity: The observations have different levels of precision. This works well if the form of the heteroskedasticity is known, perhaps measurement error in a meta-analysis. Feasible GLS is when errors from OLS are used as weights in WLS: there are better ways to do this.
- Aggregation: The observations represent the sizes of different groups, or the probability of being selected into the sample. E.g. weighting countries or states by their population.

How to run WLS in R? Use `lm` with the `weights` argument.

2.6 References

- Wooldrige, Ch 3.
- Fox, Ch 6, 9.

Chapter 3

OLS Troubleshooting and Diagnostics

3.1 Multi-Collinearity

3.1.1 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 $\log(\text{GDP in millions USD})$ and $\log(GDPinUSD)$ since $\log(\text{GDP in millions USD}) = \log(GDPinUSD)/1,000,000$. 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

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

.

3.1.1.1 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 $\boldsymbol{X}$.

3.1.2 Less-than Perfect Collinearity

What happens if variables are not linearly dependent, but nevertheless highly correlated? If $\mathrm{Cor}(\boldsymbol{x}_1, vecx_2) = 1$, then they are linearly dependent and the regression cannot be estimated (see above). But if $\mathrm{Cor}(\boldsymbol{x}_1, vecx_2) = 0.99$, 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 $|\mathrm{Cor}(\boldsymbol{x}_1, \boldsymbol{x}_2)| \to 1$ the standard errors on the coefficients of these variables increase, but OLS as an estimator works correctly; $\hat{\beta}$ and $\mathrm{se}\,\hat{\beta}$ 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.

3.1.2.1 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.

3.2 Omitted Variable Bias

3.2.1 What's the problem?

3.2.2 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 Oster2013, formalize the intuition behind the heuristic of coefficient stability to assess the sensitivity of the treatment to OVB.

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.

3.3 Measurement Error

3.3.1 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.

3.3.2 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.

3.4 Non-linearity

3.4.1 What's the problem?

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.

3.4.2 What to do about it?

Visual diagnostics

- Residual plots with curvature tests: **car** function `residualPlots`.
- Added-variable (AV) plot: **car** function `avPlots`.

- Component+residual (CERES) plot: **car** functions `crPlots` and `ceresPlots`.

Tests

- Ramsay RESET test. **lmtest** function `resettest`
- Compare Robust SE and classical OLS SE. King and Roberts.

3.5 Non-constant Variances

3.5.1 Heteroskedasticity

Note, that OLS assumes that the variance of the the disturbances is constant $\hat{Y} - Y = \varepsilon = \sigma^2$. What happens if it isn't?

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)

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

The problem is that now there are N variance parameters to estimate, in addition to the K slope coefficients. Now, there are more parameters than we can estimate. With heteroskedasticity, OLS with be unbiased, but the standard errors will be incorrect.

3.5.1.1 What to do about it?

Diagnostics

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

Solution

- 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 consisten 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.

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

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

3.5.2 Auto-correlation

More general case allows for heteroskedasticity, and autocorrelation ($\mathrm{Cov}(\varepsilon_i, \varepsilon_j) \neq 0$),

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

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`.

3.5.3 Clustered Standard Errors

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

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

3.6 Non-Normal Errors

This is not particularly important problem, and only relevant for inference with small samples since OLS has CLT properties.

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 4

Appendix

4.1 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,

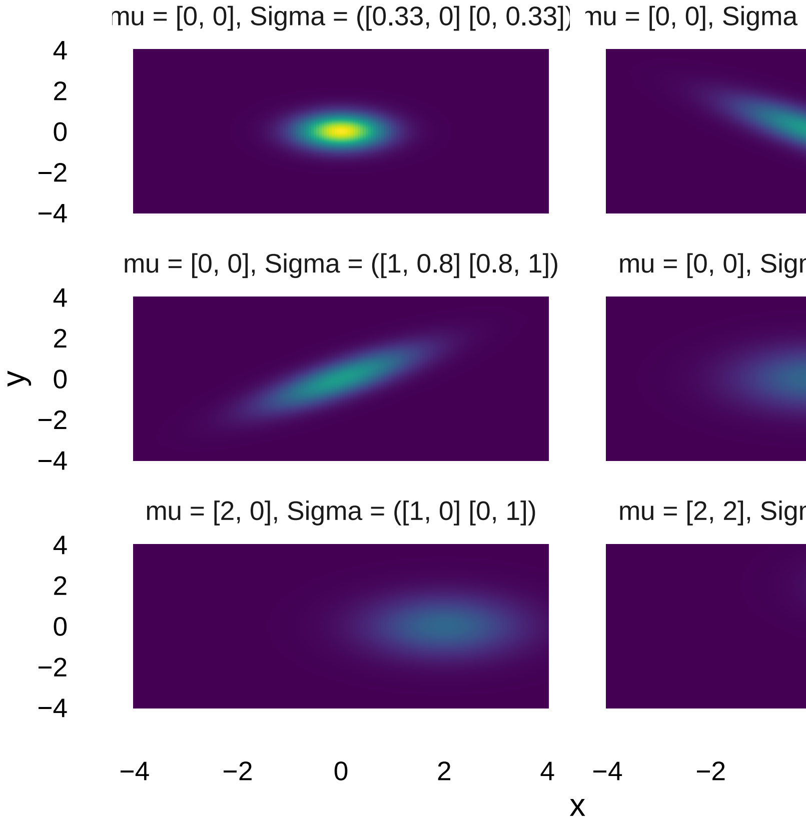
$$\begin{aligned} \boldsymbol{x} &\sim \mathcal{N}_k\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \\ \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix} &\sim \mathcal{N}_k\left(\begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \cdots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

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 5

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* 113 (1): 151–84. doi:10.1086/426036.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2014. *Mastering 'Metrics*. Princeton UP.

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.

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