Simulation demonstration of regression assumptions

This example answer is compiled from an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

I demonstrate the effects of violations of each of the classical linear model assumptions on OLS regression using Monte Carlo simulations. (MLR.1-MLR.6, see Wooldridge, 2009, pp. 157–158). A monte Carlo Simulation runs a function that generates random data sample from a model and calculates a statistic from the data. This is repeated multiple times to see the behavior of the statistic over replications.

In this example, I will generate data based on a number of population models and always estimate the following regression model:

$y = \beta_0 + \beta_1 x1 + \beta_2 x2 + \beta_3 x3 + \epsilon$

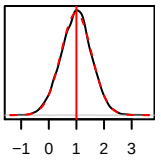
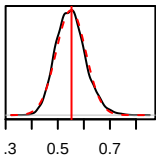
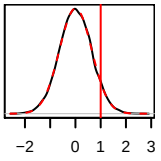
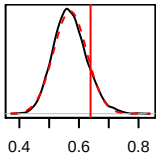
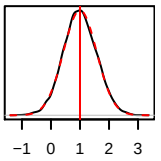
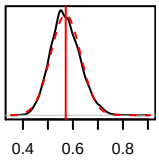
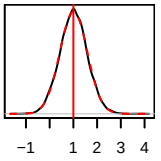
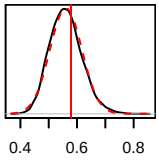
All independent variables are standardized (mean = 0, SD = 1) and correlated at 0.3 in the population, andI construct different populations by modifying the functional form and the error distribution to get different population models. The population variance of the fitted values is 4.8 and the variance of the error term is in most casea about 25 (SD = 5), which means that the population R^2 is about 16%

I use each population model to generate 10000 samples of data and use each sample to estimate the model. The sample size (N) is 100. Then I collect the estimates and standard errors for β_1 and analyze bias of these statistics. The OLS estimates of β_1 are unbiased if the mean estimate equals the population value, which in these scenarios is always 1. The standard errors are unbiased if the mean standard error is equal to the standard deviation of the estimates. (Recall that the standard error is an estimate of the standard deviation of the estimates over repeated samples.) I also plot the distribution of both the estimates and standard errors and compare with the normal distrbution. Normality is important because the t-test that are used for null hypothesis significance testing assume normality and may give incorrect results if this assumption is violated.

MLR.1 Linear in parameters

This assumption is about correct model specification. A good way to diagnose this is the residuals versus fitted plot and added variable plots (also known as partial regression plots.)

In the table below th n()-function provides the standard normal distribution.

	Population.model	Est.mean	Est.SD	SE.mean	Est.density	SE.density
1	y = x1 + x2 + x3 + 5*n(N)	0.994	0.553	0.550		
2	y = x1^2 + x2 + x3 + 5*n(N)	0.006	0.639	0.570		
3	y = x1 + x2^2 + x3 + 5*n(N)	1.000	0.571	0.571		
4	y = x1 + x2 + x1 * x2 + x3 + 5*n(N)	1.009	0.579	0.561		

Model 1 is a model where all assumptions hold. Both estimates and standard errors are unbiased and normal. Model 2 includes a quadratic function of x1 and results in biased relationship. Model 3 includes a quadratic function of x2, but this does not affect the estimates of x1, which remain unbiased. Model 4 contains an unmodeled interaction term. This causes the standard errors to become biased but does not bias the estimates.

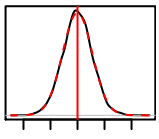
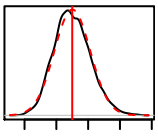
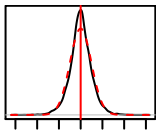
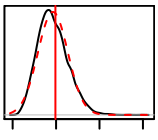
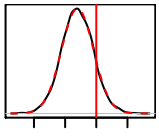
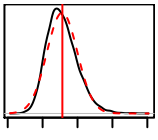
Functional form misspecification can therefore affect the biasedness of both estimates and standard errors, but mispecification of one functional form does not necessarily affect the estimates of other relationships.

MLR.2 Random sampling

The assumption MLR.2 can fail in at least two important ways: non-representative sample or clustered observations.

Wooldridge discusses the representativeness of the sample and mentions two mechanisms of how this can arise (Wooldridge, 2009, sec. 9.5). First, it is possible that data on some variables are missing in a systematic way, which will influence the results. Second, it is possible that the sample is selected in a particular way. This links also to Antonakis' discussion on omitted selection.

Clustering of observations refers to scenarios where the observations are grouped in some way. We discuss this kind of designs more during the 5th lecture. Clustering of the data leads to violations of the assumption of independent and identically distributed (iid) error term, which may lead to biased standard errors and non-normal distribution of the parameter estimates (Wooldridge 2009, p. 457). What independent and identically distributed means is best understood by examining cases where it fails, which we will do during lecture 5.

	Population.model	Est.mean	Est.SD	SE.mean	Est.density	SE.density
5	y = x1 + x2 + x3 + 5*n(N)	0.997	0.550	0.550		
6	y = x1 + x2 + x3 + 5*n(5)	1.000	0.493	0.464		
7	y = x1 + x2 + x3 + 5*n(N),y>0	0.397	0.513	0.512		

The first model (5) is again a model where all assumptions hold. Model 6 represents and extreme case of clustered error term. In this scenario, the errors are normally distributed in the population, but the data are clustered as 5 groups of 20 and all observations in a cluster have the same error term value. In other words, the error term is perfectly correlated within clusters. This is implemented by generating just 5 random normal variables, which are then recycled so that each is used as error term for 20 observations. In this sceanario the estimates remain unbiased because the error term is still uncorrelated with the dependent variables. The distribution of the estimates is slightly more peaked tahn the normal dsitribution and teh standard errors are skewed and biased negatively.

Model 6 shows the effects of selection effect. If our data is not a random sample, but selected based on the dependent variable such that only observations with positive value of y are included in the analysis, estimates will be have a large negative bias. The direction of this bias depends on how the data are selected.

MLR.3 No perfect collinearity

If any of the independent variables are perfectly collinear, the regression model cannot even be estimated. Note that perfect collinearity does not require that two of the independent variables are perfectly correlated. For example, if two independent variables, x1 and x2 are uncorrelated and have same standard errors, then defining x3 = x1 + x2 and using x3 as independent variable in regression with x1 and x2 would lead to perfect collinearity. However, the correlation between x1 and x3 would only be about 0.7.

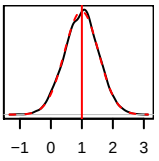
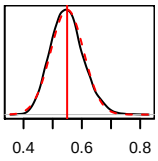
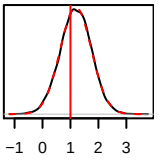
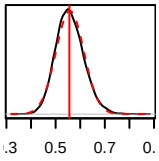
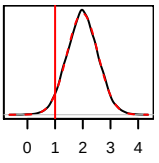
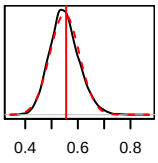
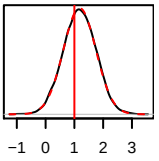
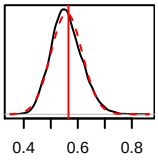
This can be easily demonstrated with just one sample.

```
x1 <- n(N)
x2 <- n(N)
x3 <- x1 + x2

y <- x1 + x2 + x3 + n(N)

summary(lm(y ~ x1 + x2 + x3))
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.98104 -0.71287 -0.04034  0.64446  2.38611
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.11637    0.09957  -1.169    0.245
## x1           1.88685    0.11656  16.188   <2e-16 ***
## x2           2.13567    0.11068  19.296   <2e-16 ***
## x3                NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9931 on 97 degrees of freedom
## Multiple R-squared:  0.8787, Adjusted R-squared:  0.8762
## F-statistic: 351.2 on 2 and 97 DF,  p-value: < 2.2e-16
```

MLR.4 Zero conditional mean

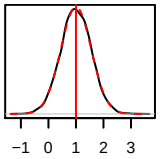
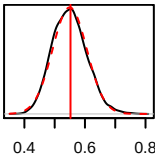
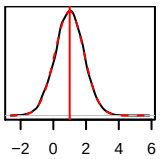
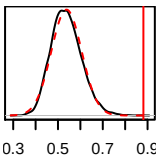
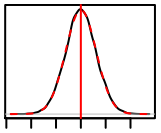
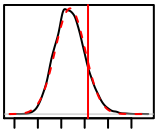
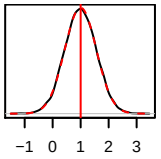
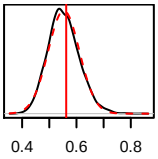
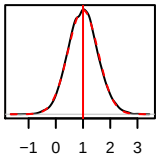
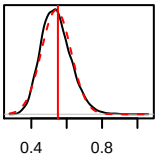
This is the no endogeneity assumption because it implies the error is uncorrelated with all independent variables. This becomes more clear if you look at the MLR.4' (Wooldridge, 2009, p. 169) assumption, which is a weaker form of this assumption. Unfortunately, there is no general test of this assumption. However, there are specific mechanism that can lead to violation of his assumption and these mechanisms can be tested.

	Population.model	Est.mean	Est.SD	SE.mean	Est.density	SE.density
8	y = x1 + x2 + x3 + 5*n(N)	0.988	0.549	0.550		
9	y = x1 + x2 + x3 + x4 + 5*n(N)	1.192	0.556	0.558		
10	y = x1 + x2 + x3 + (5*n(N)+x1)	1.989	0.555	0.550		
11	y = x1 + x2 + x3 + (5*n(N)+x4)	1.182	0.566	0.559		

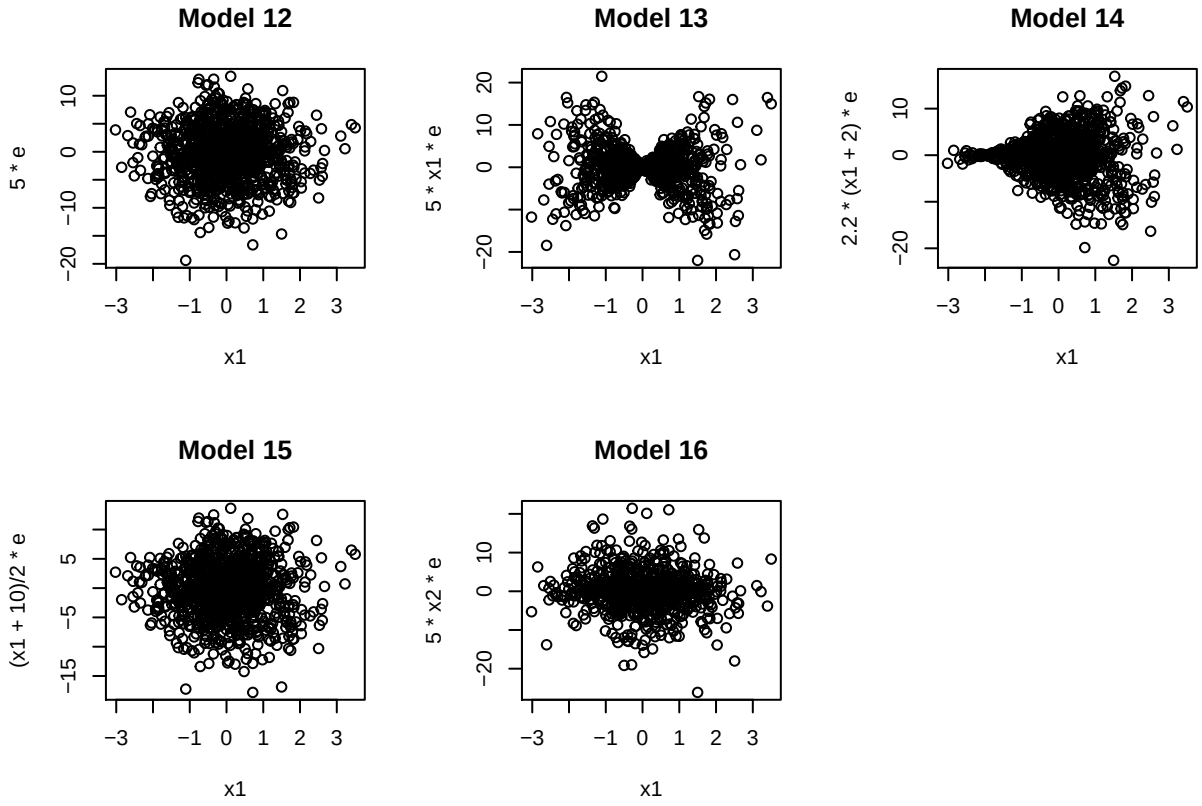
The first model (8) is again a model where all assumptions hold. In Model 9 the population contains a fourth variable, x4, which correlates with x1, x2, and x3 at 0.3. Because x4 has an effect on the dependent variable and it is correlated with the other independent variables, it should be included in the model as a control. Because the variable is omitted, the effects of x4 are attributed to other independent variables causing a positive bias. Model 10 and Model 11 represent scenarios where the error term correlates with x1 and x4 respectively. This endogeneity causes positive bias in the parameter estimates.

MLR.5 Homoskedasticity

Failure of this assumption lead to inefficient estimation and more importantly biased standard errors.

	Population.model	Est.mean	Est.SD	SE.mean	Est.density	SE.density
12	y = x1 + x2 + x3 + 5*n(N)	1.003	0.552	0.550		
13	y = x1 + x2 + x3 + 5*x1*n(N)	0.988	0.881	0.538		
14	y = x1 + x2 + x3 + 2.2*(x1+2)*n(N)	1.004	0.614	0.538		
15	y = x1 + x2 + x3 + ((x1+10)/2)*n(N)	0.996	0.562	0.553		
16	y = x1 + x2 + x3 + 5*x2*n(N)	0.998	0.550	0.541		

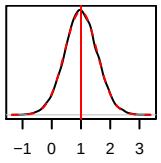
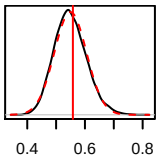
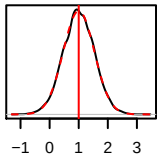
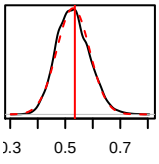
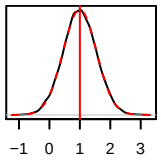
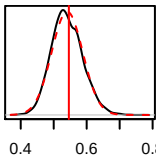
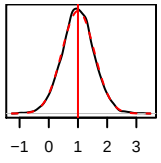
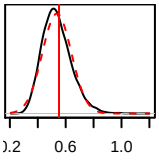
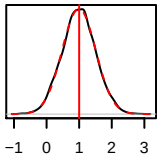
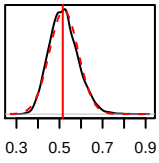
The first model (12) is again a model where all assumptions hold and the rest have various heteroskedasticity problems. To make the nature of these problems more clear, I have plotted the values of the error term over the values of x1 for a sample of 1000 for all five models.



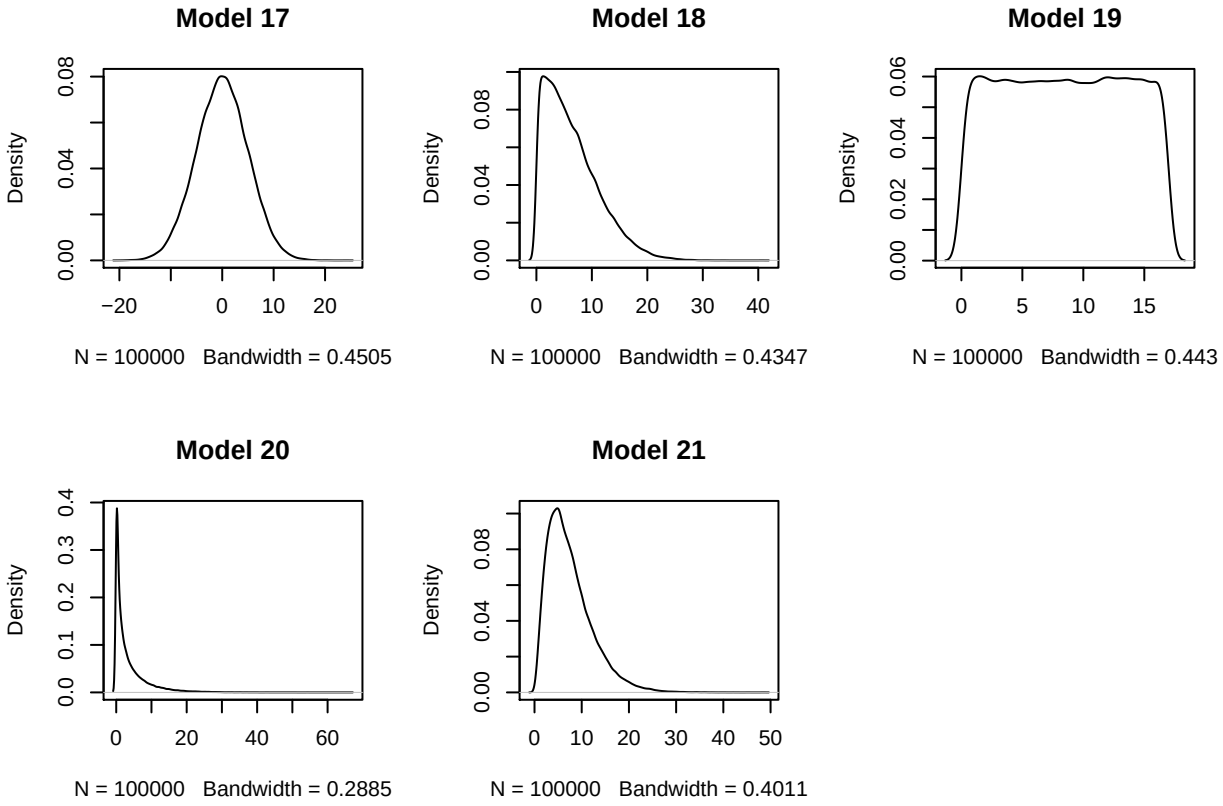
In all cases the estiametes remain unbiased and normal, which is expected because heteroskedasticity affects only efficiency (the standard deviation of the estimates) and bias of standard errors. Model 13 has severe heteroskedasticity problem in the form of butterfly shaped residuals and has the most biased SEs. Model 14 has cone shaped residuals which bias SEs, but not as severely. The heteroskedasticity problems in Model 15 and Model 16 are barely noticeable with plain eye, and also biase the SEs only little.

MLR.6 Normality

The normality of the error term assumption is required for the t-test for the path coefficients and F-test for the full model to produce correct results.

	Population.model	Est.mean	Est.SD	SE.mean	Est.density	SE.density
17	y = x1 + x2 + x3 + 5*n(N)	0.996	0.558	0.550		
18	y = x1 + x2 + x3 + 8*abs(n(N))	1.001	0.535	0.529		
19	y = x1 + x2 + x3 + 17*runif(N)	0.997	0.546	0.541		
20	y = x1 + x2 + x3 + 3.5*n(N)^2	1.004	0.552	0.537		
21	y = x1 + x2 + x3 + 1.5*rchisq(N,df=5)	0.992	0.515	0.520		

The first model (17) is again a model where all assumptions hold. To quantify the degree of non- normality, I have plotted the distributions of the five error terms below using a sample size of 100000 to draw the distribution. (All of these have known propability density functions, but doing actual samples is a bit easier to do than to draw the known distributions.)



Again, in all models the estimates are unbiased. The standard errors are clearly biassed down for Model 20, which has extremely skewed error distribution. SEs of other models are unbiased or at least close to unbiased. Indeed, when sample size increases the normality assumption becomes less important. Wooldridge notes that as sample size and adds that "Some econometricians think that n = 30 is satisfactory, but this cannot be sufficient for all possible distributions of u." (Wooldridge, 2009, p. 175) Lack of normality is not necessarily a big problem with the sample sizes typical for management research.