Adrienne Scott
GOVT 6029: Advanced Regression
Homework 3
May 2, 2018

Question A

By changing the correlation from 0 to .5, .9, and .99, the standard errors increase substantially. Raising the correlation adds bias to all the estimates, but it mainly affects the precision of the estimates for B1, and B2, it has minimal effect on B3. For example, x1 increases from a standard error of .15 at the lowest value (when set at 0) to 1.02 at the highest (when set to .99). X2 increases from a standard error of .15 at the lowest value (when set at 0) to 1.02 at the highest (when set to .99). X3 only increases from a standard error of .14 at the lowest value (when set at 0) to .15 at the highest (when set to .99). As the standard errors approach 1, they predictions become increasingly unstable.

Intercept	**X1**	**X2**	**X3**
True standard errors across 1000 simulation runs			
0.1429859	0.1466162	0.1459070	0.1449920
True standard errors across 1000 simulation runs (.5)			
0.1406227	0.1611374	0.1652900	0.1403782
True standard errors across 1000 simulation runs (.9)			
0.1412854	0.3274383	0.3267843	0.1484614
True standard errors across 1000 simulation runs (.99)			
0.1507659	1.0174643	1.0114101	0.1504241

Question B

After setting the correlation of X1 and X2 to 1, I find that an error occurs, informing me that x has missing values. Because the regression is perfectly correlated than RStudio drops one of the variables and you receive and error (omitted variable bias). The regression summary displays the following results for the coefficients compared to the original:

Standard Errors (1)			
Intercept	**X1**	**X2**	**X3**
0.1354	0.1366	NA	0.1349
Standard Errors (0)			
0.1429859	0.1466162	0.1459070	0.1449920

The table shows that the value of the standard error for X2 is missing and for the values of X1 and X3 that are available are poor estimates of the coefficients.

Question C

Because we omitted X2, the results are imprecise and show a great increase in the values of the standard errors of the original. It displays errors of .33 (which is more than twice the original).

Standard Errors (X2 omitted)

Intercept	X1	X3
0.3304900	0.3321876	0.3354280

Question D

After setting the correlation of X1 and X2 to .9, the variables standard errors reduce significantly. Although it is not as high as the values shown in Question C, they are still imprecise, but much closer to those of the original example. The results are below:

Standard Errors (X2 omitted & .9)		
Intercept	**X1**	**X3**
0.1867691	0.1998613	0.1978990

Question E

Even though I kept the correlation set at .9, when I omit X3 there is no correlation between X3 and the other variables, therefore, the standard increase substantially from the previous example (even higher than the values in Question C when I omitted X2 alone). The results are very inaccurate, with the standard errors approaching 1, as shown below:

Standard Errors (X3 omitted & .9)		
Intercept	**X1**	**X3**
0.4158694	0.9959322	0.9835622

Question F

Correlated covariates explain the differences across the results for Questions C, D, and E. As we manipulate the variables, the results increase in the degree of collinearity. This breaks assumptions in the Gauss-Markov principle, causing the least squares estimates to be biased and unstable. A simple way to deal with the problem of highly correlated covariates, is to only use one of the variables in the regression model (and/or replace one of the highly correlated covariates with one that is not).

Question G

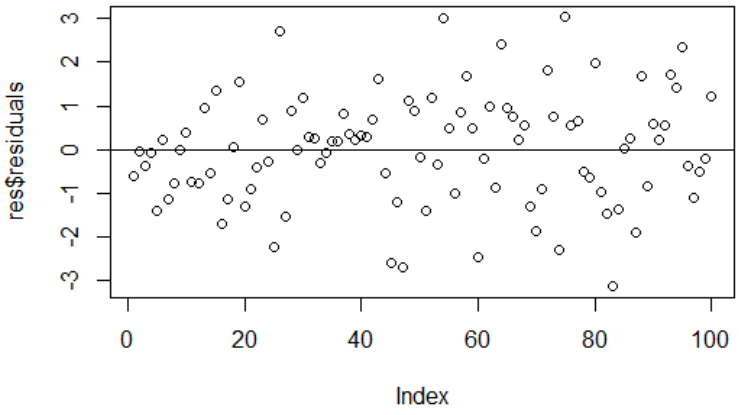
The selection of y increases the standard error of the beta coefficients. This means that for this instance, where all observations in which y is greater than its sample mean are deleted, the standard errors are less accurate than the original (mcls.r). The estimate is biased because the distance between the true and expected value is greater than zero.

Standard Errors (y > mean deleted)			
Intercept	**X1**	**X2**	**X3**
0.3219479	0.2315343	0.2531286	0.2784360
Standard errors (mcls.r)			
0.1429859	0.1466162	0.1459070	0.1449920

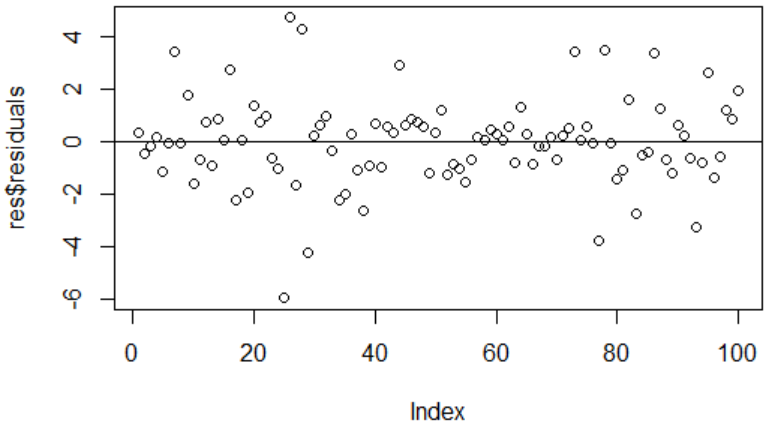
Question H

When y1 is equal to 0, it is homoscedastic—all the error terms are about equal. However, when y1 is equal to 1, it is heteroskedastic—the error terms vary. This added heteroskedasticity skews the results because it weighs the data points with larger discrepancies more heavily. In addition, it biases the standard errors and makes the beta coefficients less precise. You can see that in graph 1 (homoscedastic) for the most part the residuals are evenly dispersed, whereas in graph 2 (heteroskedastic) the residuals are either very close to the mean or very far from it (there is more variance in the dispersion).

Standard Errors (y1=0)			
Intercept	**X1**	**X2**	**X3**
0.1412854	0.1423187	0.1484614	0.1475454



Standard Errors (y1=1)			
Intercept	**X1**	**X2**	**X3**
0.1943577	0.2563994	0.1833084	0.1797310



Question I

Serial correlation occurs when the standard errors at one point are correlated with the errors at another point. The errors in the first instance continue to occur the next set of repetitions.

Intercept	X1	X2	X3
True standard errors across 1000 simulation runs (0)			
0.1451464	0.1540434	0.1474923	0.1475439
True standard errors across 1000 simulation runs (.5)			
0.1360513	0.1662536	0.1706794	0.1459816
True standard errors across 1000 simulation runs (.9)			
0.1439312	0.3487373	0.3404913	0.1461843

Question J

Collecting more data can only help solve some problems. By increasing the number of observations, you can avoid problems of imprecise estimates due to a small sample size, however, once your sample is large enough this will no longer help you get more precise estimates. This will not solve problems of heteroskedasticity, omitted variable bias, or partial/perfect collinearity. You can solve problems of high collinearity by substituting the variable with another variable that is not collinear to the other. One way to avoid problems of heteroskedasticity is to place the variable on a logarithmic scale.