POL SCI: Problem Set 6

Spring 2017

University of California, Berkeley

Prof. Thad Dunning/GSI Natalia Garbiras-Díaz

Due Friday, March 20th, 9:00 AM (before lecture)

Each group should turn in its problem set solution by email to Natalia Garbiras-Díaz (nataliagarbirasdiaz+231b@berkeley.edu) by Friday at 9 AM. Please work out the problems on your own, before you meet with your group to agree on solutions.

1. **Multicollinearity**. Let $Y_i = au_i + bv_i + \epsilon_i$ for $i = 1, ..., 100$. The ϵ_i are IID with mean 0 and variance 1. Here, u_i and v_i are fixed, not random. These two data variables have mean 0 and variance 1. The correlation between them is r. Let $M = [u\, v]$ denote the (partitioned) design matrix.

   (a) **Show that the design matrix has rank 1 if $r = 1$ or $r = -1$.**

   (b) **Otherwise, let $(M'M)^{-1}M'Y = (\hat{a}\ \hat{b})'$ be the OLS estimator for a and b. Is the OLS estimator biased or unbiased?**

   (c) **Find the variance of $\hat{a}$; of $\hat{b}$; of $\hat{a} + \hat{b}$; and of $\hat{a} - \hat{b}$. What happens if $r = 0.99$?**

   (d) **What are the implications of multicollinearity for drawing inferences about a and b? What about their sum and their difference? What are the implications of exact collinearity? (Note: *exact collinearity* here means, $r = 1$ or $r = -1$; *multicollinearity* means $r \doteq 1$ or $r \doteq -1$).**

   For questions a-d, see Freedman (2009) p. 249-250.

   (e) **True or false, and explain:**

      i. **Multicollinearity leads to bias in the OLS estimator.** FALSE.

ii. **Multicollinearity leads to bias in the estimated standard errors for the OLS estimates.** FALSE.

iii. **Multicollinearity leads to big standard errors for some estimates.** TRUE.

2. **fGLS vs. panel-corrected standard errors. On p. 175, Freedman (2009) explains that White's method for estimating the SEs in OLS (what Beck and Katz (1995) call "panel-corrected standard errors" in the time-series cross-section context) "may have the same sort of problems as plug-in SEs, because estimated covariance matrices can be quite unstable." Explain why the estimated covariance matrix would be unstable in the settings discussed by Beck and Katz. How this would affect panel-correct standard errors? (Hint: look at p. 638 of Beck and Katz. Where does the covariance matrix of the errors appear in the formula for the covariance matrix of $\hat{\beta}$, and how is it estimated?).**

Note that the variance-covariance matrix of the OLS estimator, when $\text{cov}(\epsilon|X) = G$, is

$$\text{cov}(\hat{\beta}_{OLS}|X) = (X'X)^{-1}X'GX(X'X)^{-1}. \tag{1}$$

(See Freedman 2009: p. 63, equation 8). Then, the White-like standard errors (known as "panel-corrected standard errors" in the time-series cross-section context) are given by (1), with G replaced by $\hat{G}$. That is,

$$\widehat{\text{cov}}(\hat{\beta}_{OLS}|X) = (X'X)^{-1}X'\hat{G}X(X'X)^{-1}. \tag{2}$$

Here, the typical element of $\hat{G}$ is obtained from manipulation of the OLS residuals. For example, if we assume that G has contemporaneously correlated elements (and that this contemporaneous correlation is constant over time), we can estimate the typical element $\text{cov}(\epsilon_{i,t}, \epsilon_{j,t})$ consistently as

$$\widehat{\text{cov}}(\epsilon_{i,t}, \epsilon_{j,t}) = \frac{1}{T}\sum_{i=1}^{T} e_{i,t}e_{j,t}, \tag{3}$$

where $e_{i,t}$ is the residual from the OLS fit for unit i at time t, and T is the number of time periods.

Now, notice that if there are many units, then there are many contemporaneous covariances (correlations) of the errors to estimate. Namely, there are $N(N-1)/2$ contemporaneous covariances, where N is the number of units. This can be a large number, and if T is small, the estimates will be very imprecise, and possibly biased downwards in some cases. (Remember, $\hat{G}$ is a non-linear estimator for G, due to the multiplication in equation 3, so equation 2 is not a linear combination of the data).

This is why the estimated variance-covariance matrix used to find the panel-corrected standard errors can be unstable, just as it is for fGLS.

3. **Measurement error. In this exercise, you will build a plot like the one shown in lecture to depict the consequences of measurement error in dependent and independent variables, in the case of bivariate regression; use the code you wrote on the last problem set to build added variables plots, in the case of multivariate regressions with measurement error; and then run simulations to assess bias and mean squared error of estimators in the presence of measurement error, for the bivariate and multivariate case. (Note: you can write a function for the whole problem, but that is not required). Consider the model**

$$Y^* = \beta X^* + \gamma W^* + \epsilon, \tag{4}$$

which conforms to the usual OLS assumptions. Here, Y^*, X^*, and W^* are true values of the respective variables. But each may be measured with error:

$$Y = Y^* + \delta, \tag{5}$$

$$X = X^* + \eta, \tag{6}$$

and

$$W = W^* + \nu, \tag{7}$$

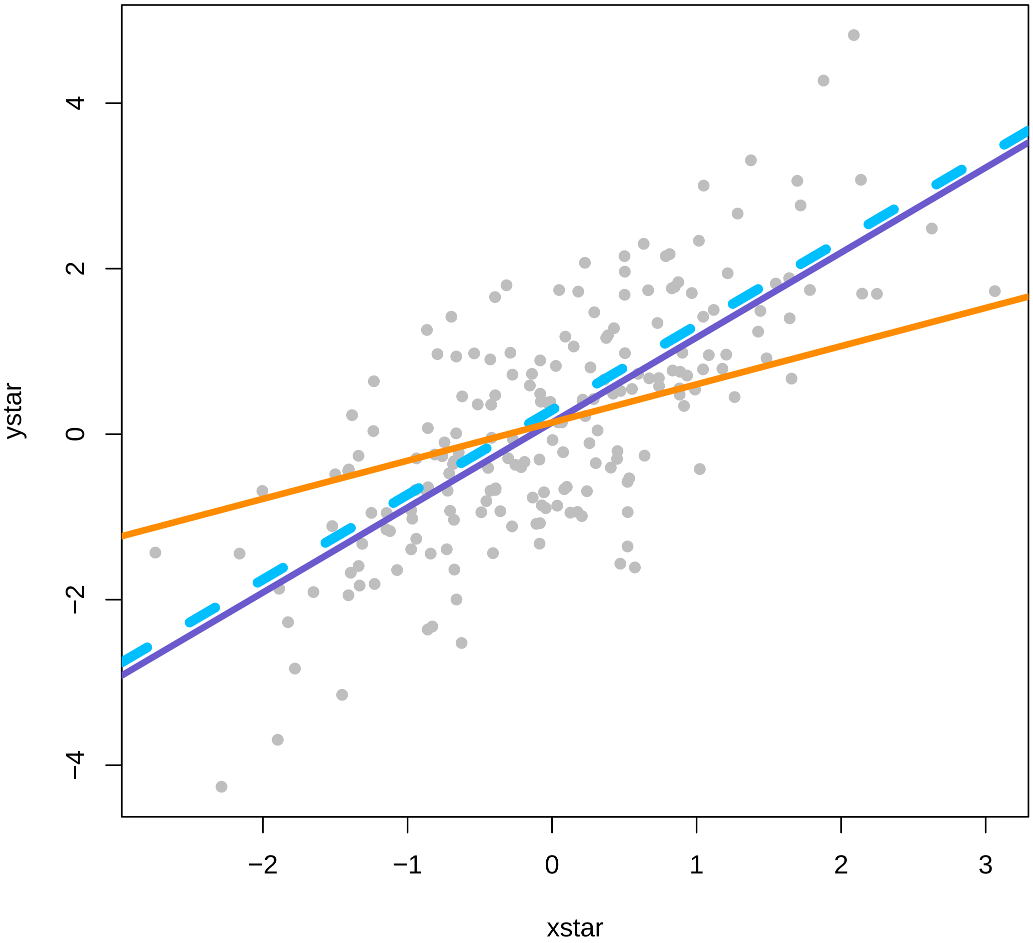
where $E(\delta) = E(\eta) = 0 = E(\nu) = 0$ and the errors are independent of Y^*, X^*, W^*, ϵ, and each other, with variances σ^2_δ, σ^2_η, and σ^2_ν, respectively.

Recicling our old average variable plot function:

```r
av_plot <- function(y, x, z, xlab = "", ylab = "") {

    Z <- cbind(1, z)

    b_y <- solve(t(Z) %*% Z) %*% (t(Z) %*%
        y)
    e_y <- y - Z %*% b_y

    b_x <- solve(t(Z) %*% Z) %*% (t(Z) %*%
        x)
    e_x <- x - Z %*% b_x

    EX <- cbind(1, e_x)
    b_ex <- solve(t(EX) %*% EX) %*% (t(EX) %*%
        e_y)

    predicted_y <- b_ex[1] + e_x * b_ex[2]

    plot(e_x, e_y, col = "grey", pch = 16,
        ylab = ylab, xlab = xlab, xlim = c(-4,
            4), ylim = c(-4, 4))
    lines(e_x, predicted_y, col = "red",
        lwd = 3)

}
```

(a) **First, consider the case where $\gamma = 0$ and $\beta = 1$, so we are back to the bivariate model. Simulate 200 draws of X^*, ϵ, δ, and η, all distributed as $N(0, 1)$ random variables, all independent of each other, and use the draws to construct Y^*, Y, and X. Regress Y^* on X^*, Y on X^*, and Y^* on X. Add these three regression lines to a scatterplot of Y^* against X^*, and also add the line $Y^* = X^*$. What does the plot suggest?**

```r
# setup
gamma <- 0
beta <- 1
xstar <- rnorm(200, 0, 1)
epsilon <- rnorm(200, 0, 1)
delta <- rnorm(200, 0, 1)
eta <- rnorm(200, 0, 1)
wstar <- 0

ystar <- beta * xstar + gamma * wstar + epsilon
x <- xstar + eta
y <- ystar + delta

plot(xstar, ystar, pch = 16, col = "grey")
abline(lm(ystar ~ xstar), col = "slateblue",
    lwd = 4)
abline(lm(y ~ xstar), col = "deepskyblue",
    lty = 2, lwd = 6)
abline(lm(ystar ~ x), col = "darkorange",
    lwd = 4)
```



The plot suggests that measurement error in the dependent variable does not lead to bias but measurement error in the independent variable does.

(b) **Write a simulation to repeat (a) 20,000 times, and create a 3x2 table in R that has "no error", "error in Y", and "error in X" as the row labels and "bias" and "mse" as the column labels. (Here, MSE refers to mean-squared error; refer to the lecture for definitions). Fill in the table with the results of the simulation. What do you conclude about the effects of measurement error in the bivariate case?**

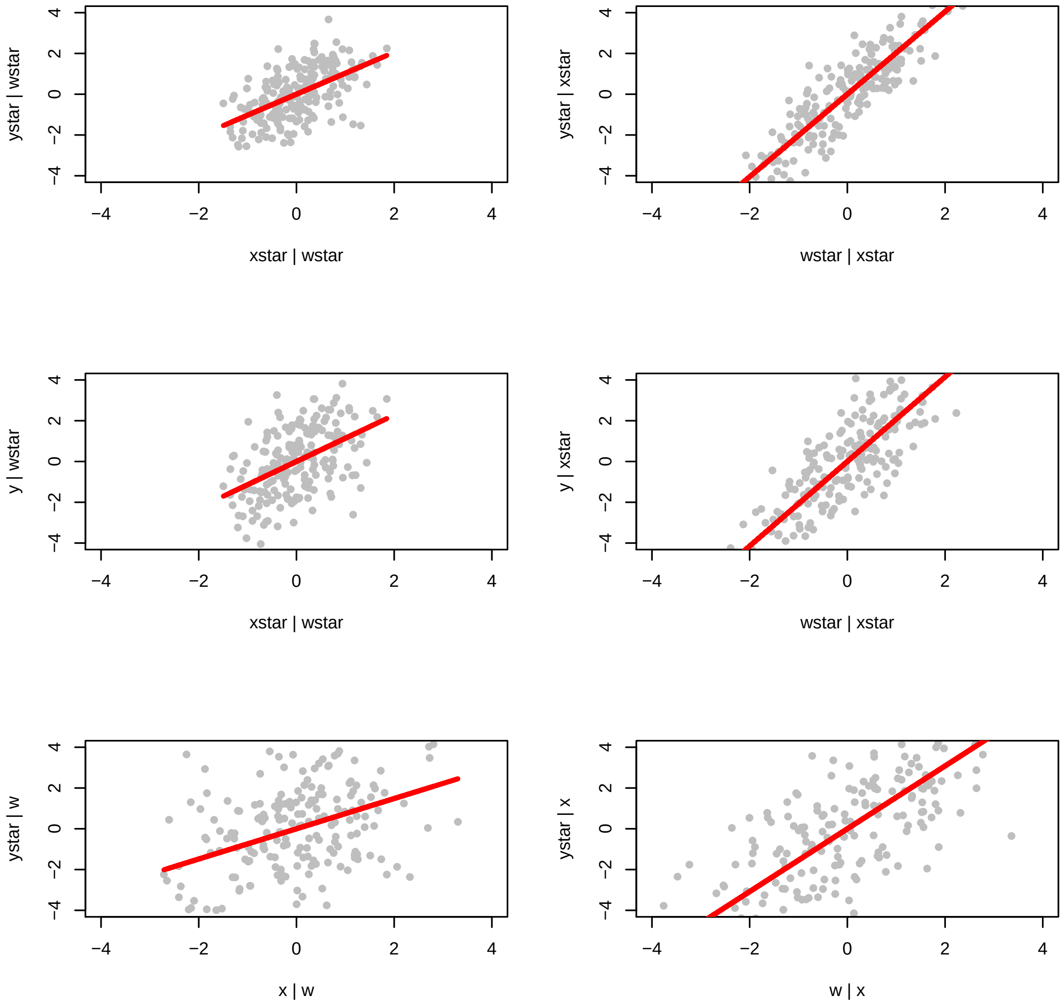
```
sim <- function() {

    gamma <- 0
    beta <- 1
    xstar <- rnorm(200, 0, 1)
    epsilon <- rnorm(200, 0, 1)
    delta <- rnorm(200, 0, 1)
    eta <- rnorm(200, 0, 1)
    wstar <- 0

    ystar <- beta * xstar + gamma * wstar +
        epsilon
    x <- xstar + eta
    y <- ystar + delta

    noerror <- lm(ystar ~ xstar)$coefficients[2]
    error_y <- lm(y ~ xstar)$coefficients[2]
    error_x <- lm(ystar ~ x)$coefficients[2]

    return(c(noerror, error_y, error_x))

}

results <- replicate(20000, sim())

bias <- function(x) {
    mean(x) - 1
}
mse <- function(x) {
    mean((x - 1)^2)
}

table <- as.matrix(rbind(apply(results, 1,
    bias), apply(results, 1, mse)))

colnames(table) <- c("no error", "error in y",
    "error in x")
rownames(table) <- c("bias", "mse")
table

##          no error    error in y
## bias 5.952242e-05 -0.0009112866
## mse  5.106810e-03  0.0102831080
##      error in x
## bias -0.4998880
## mse   0.2537072
```

The simulation confirms the intuition in the plot: in the bivariate case, measurement error in the dependent variable does not lead to bias but measurement error in the independent variable does.

(c) **Now suppose $\gamma = 2$ and $\beta = 1$. Simulate 200 draws of X^*, ϵ, δ, η, and ν, all distributed as $N(0, 1)$ random variables, all independent of each other.**

To create a correlation between the right-hand-side variables, generate $W^* = X^* + \psi$, where ψ is $N(0,1)$. Now, use the draws to construct Y^*, Y, X, and W. Regress Y^* on X^* and W^*, Y on X^* and W^*, and Y^* on X and W. For each regression, construct added variable plots, using the code you wrote on your last problem set, where the (residuals of the) response variable are plotted against the residuals of one independent variable and then the other. (So there should be six plots – two for each of the three regressions). What do the plots suggest?

```r
gamma <- 2
beta <- 1
xstar <- rnorm(200, 0, 1)
epsilon <- rnorm(200, 0, 1)
delta <- rnorm(200, 0, 1)
eta <- rnorm(200, 0, 1)
wstar <- xstar + rnorm(200, 0, 1)

ystar <- beta * xstar + gamma * wstar + epsilon
x <- xstar + eta
y <- ystar + delta
w <- wstar + rnorm(200, 0, 1)


par(mfrow = c(3, 2))

av_plot(y = ystar, x = xstar, z = wstar,
    xlab = "xstar | wstar", ylab = "ystar | wstar")
av_plot(y = ystar, x = wstar, z = xstar,
    xlab = "wstar | xstar", ylab = "ystar | xstar")

av_plot(y = y, x = xstar, z = wstar, xlab = "xstar | wstar",
    ylab = "y | wstar")
av_plot(y = y, x = wstar, z = xstar, xlab = "wstar | xstar",
    ylab = "y | xstar")

av_plot(y = ystar, x = x, z = w, xlab = "x | w",
    ylab = "ystar | w")
av_plot(y = ystar, x = w, z = x, xlab = "w | x",
    ylab = "ystar | x")
```



Just like in (a), the plots suggests that measurement error in the dependent variable does not lead to bias but measurement error in the independent variables does.

(d) **Write a simulation to repeat (c) 20,000 times, and create a 6x2 table in R that has as the row labels "no error–est of beta", " "no error–est of gamma", "error in Y–est of beta", "error in Y–est of gamma," and "error in X and W–est of beta," and "error in X and W–est of gamma" and as the column labels, "bias", "mse" and "average r." (Here, average r is the average correlation between X^* and W^* across the 20,000 simulations; it will be the same for each row of the table). What do you conclude about the effects of measurement error in the multivariate case, when the true values of the right-hand-side variables are correlated?**

```r
sim <- function() {

    gamma <- 2
    beta <- 1
    xstar <- rnorm(200, 0, 1)
    epsilon <- rnorm(200, 0, 1)
    delta <- rnorm(200, 0, 1)
    eta <- rnorm(200, 0, 1)
    wstar <- xstar + rnorm(200, 0, 1)

    ystar <- beta * xstar + gamma * wstar +
        epsilon
    x <- xstar + eta
    y <- ystar + delta
    w <- wstar + rnorm(200, 0, 1)

    out <- c(lm(ystar ~ xstar + wstar)$coefficients[2],
        lm(ystar ~ xstar + wstar)$coefficients[3],
        lm(y ~ xstar + wstar)$coefficients[2],
        lm(y ~ xstar + wstar)$coefficients[3],
        lm(ystar ~ x + w)$coefficients[2],
        lm(ystar ~ x + w)$coefficients[3],
        cor(xstar, ystar))

    return(out)

}

results <- replicate(20000, sim())

# Because now we need to calculate these
# for gamma as well, we create a new
# function
bias_gamma <- function(x) {
    mean(x) - 2
}
mse_gamma <- function(x) {
    mean((x - 2)^2)
}


table <- as.matrix(rbind(apply(results[c(1,
    3, 5), ], 1, bias), apply(results[c(1,
    3, 5), ], 1, mse), rep(mean(results[7,
    ]), 3)))
table_gamma <- as.matrix(rbind(apply(results[c(2,
    4, 6), ], 1, bias_gamma), apply(results[c(2,
    4, 6), ], 1, mse_gamma), rep(mean(results[7,
    ]), 3)))
table <- cbind(table, table_gamma)

colnames(table) <- c("no error--est beta",
    "error in y--est beta", "error in x and w--est beta",
    "no error--est gamma", "error in y--est gamma",
    "error in x and w--est gamma")
rownames(table) <- c("bias", "mse", "average r")
table
```

```
##             no error--est beta
## bias              0.0002356594
## mse               0.0101011006
## average r         0.8009749190
##             error in y--est beta
## bias                0.0006187318
## mse                 0.0202210342
## average r           0.8009749190
##             error in x and w--est beta
## bias                        -0.2014657
## mse                          0.0549791
## average r                    0.8009749
##             no error--est gamma
## bias               0.0006930045
## mse                0.0051205433
## average r          0.8009749190
##             error in y--est gamma
## bias                -2.372971e-05
## mse                  1.018480e-02
## average r            8.009749e-01
##             error in x and w--est gamma
## bias                         -0.5995710
## mse                           0.3690107
## average r                     0.8009749
```

For the multivariate case with correlated independent variables, it is still true that measurement error in the dependent variable does not bias the estimates. Measurement error in the independent variables, however, does generate bias. Moreover, because the variables are correlated, the bias in the estimator of a particular coefficient is due both to the fact that it is measured with error and also because the estimator for the coefficient of the other variable is also biased.

(e) **Now modify the code you wrote for part (d) so that X^{star} and W^* are independent $N(0,1)$ random variables. What do you conclude about the effects of measurement error in the multivariate case, when the true**

values of the right-hand-side variables are generated as independent random variables?

```
sim <- function() {

    gamma <- 2
    beta <- 1
    xstar <- rnorm(200, 0, 1)
    epsilon <- rnorm(200, 0, 1)
    delta <- rnorm(200, 0, 1)
    eta <- rnorm(200, 0, 1)
    wstar <- rnorm(200, 0, 1)

    ystar <- beta * xstar + gamma * wstar +
        epsilon
    x <- xstar + eta
    y <- ystar + delta
    w <- wstar + rnorm(200, 0, 1)

    out <- c(lm(ystar ~ xstar + wstar)$coefficients[2],
        lm(ystar ~ xstar + wstar)$coefficients[3],
        lm(y ~ xstar + wstar)$coefficients[2],
        lm(y ~ xstar + wstar)$coefficients[3],
        lm(ystar ~ x + w)$coefficients[2],
        lm(ystar ~ x + w)$coefficients[3],
        cor(xstar, ystar))

    return(out)

}

results <- replicate(20000, sim())

table <- as.matrix(rbind(apply(results[c(1,
    3, 5), ], 1, bias), apply(results[c(1,
    3, 5), ], 1, mse), rep(mean(results[7,
    ]), 3)))
table_gamma <- as.matrix(rbind(apply(results[c(2,
    4, 6), ], 1, bias_gamma), apply(results[c(2,
    4, 6), ], 1, mse_gamma), rep(mean(results[7,
    ]), 3)))
table <- cbind(table, table_gamma)

colnames(table) <- c("no error--est beta",
    "error in y--est beta", "error in x and w--est beta",
    "no error--est gamma", "error in y--est gamma",
    "error in x and w--est gamma")
rownames(table) <- c("bias", "mse", "average r")
table

##             no error--est beta
## bias              0.0002027819
## mse               0.0050952751
## average r         0.4072152880
##             error in y--est beta
## bias                0.0002971048
## mse                 0.0102623242
## average r           0.4072152880
##             error in x and w--est beta
## bias                        -0.4996737
## mse                          0.2586446
## average r                    0.4072153
##             no error--est gamma
## bias               0.0005352335
## mse                0.0049911003
## average r          0.4072152880
##             error in y--est gamma
## bias                 0.0005471015
## mse                  0.0100406318
## average r            0.4072152880
##             error in x and w--est gamma
## bias                         -0.9990469
## mse                           1.0069870
## average r                     0.4072153
```

For the multivariate case with uncorrelated independent variables, we still have that measurement error in the dependent variable doesn't bias the estimators. Measurement error in the independent variables does generate bias. However, because the variables are uncorrelated, here the bias in the estimator of a particular coefficient is due solely to the fact that the corresponding independent variable is measured with error.

4. **The Neyman model and regression. Let $Y_i(1)$ denote the potential outcome if unit i is treated, and let $Y_i(0)$ denote the potential outcome for the same observation if it is not treated. The unit causal effect is the difference between $Y_i(1)$ and $Y_i(0)$. This causal effect may vary from one unit to the next. The random assignment of units to treatment $(X_i = 1)$ and control $(X_i = 0)$ (equivalently, the random sampling of units from the experimental study group into treatment and control groups) is the only random component in the modeling framework.**

(a) **Show that the potential outcomes model**

$$Y_i = Y_i(0)(1 - X_i) + Y_i(1)X_i \tag{8}$$

may be expressed in the form of a regression model such that b represents the average causal effect of the treatment, and

$$Y_i = a + bX_i + u_i, \tag{9}$$

where the disturbance term $u_i \equiv Y_i(0) - \overline{Y(0)} + ((Y_i(1) - \overline{Y(1)}) - (Y_i(0) - \overline{Y(0)}))X_i$. Here, $\overline{Y(0)}$ is the average potential outcome under control for the study group, and $\overline{Y(1)}$ is the average potential outcome under treatment.

Let $b = \bar{Y}^T - \bar{Y}^C$ be the average causal effect of the treatment and $a = \bar{Y}^C$ be the average potential outcome under control, for all units in the study group. Then, adding and subtracting $\bar{Y}^C + (\bar{Y}^T - \bar{Y}^C)X_i$ to the right-hand side of the potential outcomes model in equation (8), we have

$$\begin{aligned} Y_i &= \bar{Y}^C + (\bar{Y}^T - \bar{Y}^C)X_i + Y_i^C(1 - X_i) + Y_i^T X_i - \bar{Y}^C - (\bar{Y}^T - \bar{Y}^C)X_i \\ &= \bar{Y}^C + (\bar{Y}^T - \bar{Y}^C)X_i + [Y_i^C - \bar{Y}^C + ((Y_i^T - \bar{Y}^T) - (Y_i^C - \bar{Y}^C)X_i] \\ &= a + bX_i + u_i. \end{aligned} \tag{10}$$

(b) **Is it possible for the disturbance term u_i to be statistically independent of the independent variable? Explain your answer.**

No. The value of u_i depends on the realization of X_i. If $X_i = 1$, $u_i = Y_i^T - \bar{Y}^T$. If $X_i = 0$, $u_i = Y_i^C - \bar{Y}^C$. Thus, the conditional distribution of u_i depends on the realization of X_i. Even though $E(u_i|X_i = 1) = E(u_i|X_i = 0) = 0$ (as per part d below), u_i and X_i are not statistically independent.

(c) **Is it possible for the disturbance term u_i to be homoskedastic (i.e., have a constant variance denoted σ^2)? Explain your answer.**

No. The variance of u_i depends on the realization of X_i:

$$\text{Var}(u_i|X = 1) = \text{Var}(Y_i^T - \bar{Y}^T) \neq \text{Var}(Y_i^C - \bar{Y}^C) = \text{Var}(u_i|X = 0). \tag{11}$$

(This precludes special cases where the variance of the potential outcomes under treatment exactly equals the variance of potential outcomes under control).

(d) **In light of your answers to (b) and (c), is the OLS estimator of the model in equation (9) unbiased? Are the OLS standard errors (e.g. those calculated using $\hat{\sigma}^2[X'X]^{-1}$ and stored by R routines such as lm) correct? Explain your answers.**

The OLS estimator is conditionally unbiased. Letting Y denote the $n \times 1$ vector of Y_is, X denote the $n \times 2$ matrix with typical row $[1 \quad X_i]$, u denote the $n \times 1$

vector of u_is, and $\beta = (a \quad b)'$, we have

$$\begin{aligned} E(\hat{\beta}_{OLS}|X) &= E((X'X)^{-1}X'Y|X) && (12)\\ &= (X'X)^{-1}X'E(Y|X) && (13)\\ &= (X'X)^{-1}X'E(X\beta + u|X) && (14)\\ &= \beta + (X'X)^{-1}X'E(u|X). && (15) \end{aligned}$$

Conditional on X, the expectation of u is zero: for each u_i,

$$\begin{aligned} E(u_i|X_i = 1) &= E(Y_i^C - \bar{Y}^C + Y_i^T - \bar{Y}^T - (Y_i^C - \bar{Y}^C)) && (16)\\ &= \bar{Y}^C - \bar{Y}^C + \bar{Y}^T - \bar{Y}^T - \bar{Y}^C + \bar{Y}^C && (17)\\ &= 0 && (18) \end{aligned}$$

and $E(u_i|X_i = 0) = E(Y_i^C - \bar{Y}^C) = \bar{Y}^C - \bar{Y}^C = 0$. Thus, $E(\hat{\beta}_{OLS}|X) = \beta$. However, the nominal OLS standard errors do not apply. For instance, they assume $\text{Var}(u_i|X = 1) = \text{Var}(u_i|X = 0)$, which does not hold (by the answer to c).

5. **Marginal effects plot. Using the Miguel et al. data, plot the marginal effects of lagged rainfall as a function of land crop from the regression of growth (*gdp_g*) on a constant, lagged rainfall (*GPCP_g_l*), land crop (*land_crop*) and a third term with the interaction of lagged rainfall and land crop. This is the plot shown in the lecture slides for week 6. You can write a function though this is not required. To build the plot, follow these steps:**

   (a) **Run the regression and store the coefficients for lagged rainfall and the interaction.**

   (b) **Find the variance for each of these coefficients and their covariance.**

   (c) **Produce a vector of simulated *land_crop* levels. For this, use seq() to get a sequence that goes from the minimum value of *land_crop* to its maximum by small increments.**

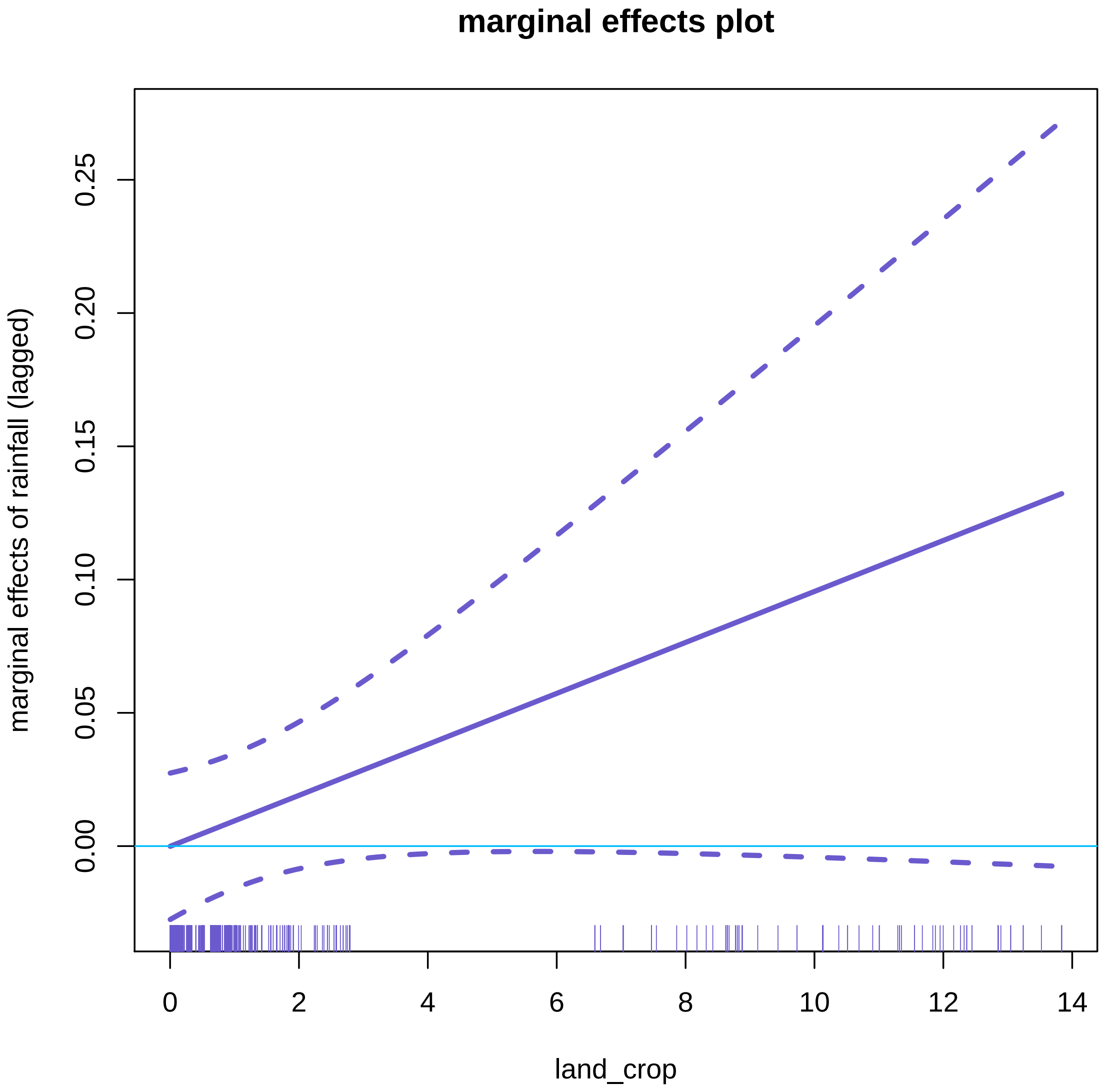
   (d) **Using (a), produce a vector with the marginal effects of lagged rainfall for each value of the vector created in (c).**

   (e) **Using (b), produce a vector with the variance of the marginal effects of lagged rainfall for each value of the vector created in (c). Use the variance to create two new vectors: one for the upper bound of the confidence interval and one for the lower bound.**

   (f) **Plot the marginal effects and the confidence interval lines. Use rug() to add a rug showing the distribution of the data.**

   (g) **Are the effects of rainfall on growth conditional on *land_crop*? Comment. What needs to hold for you to answer in the affirmative?**

```
marginal_eff <- function(y, x, z, xlab = "",
    ylab = "", main = "") {
    # Function for marginal effects plot,
    # where z is the mediator.

    # finding observations with NAs (this is
    # how lm does it!)
    na.x <- 0
    na.z <- 0
    na.y <- 0
    if (sum(is.na(x)) != 0) {
        na.x <- which(is.na(x))
    }
    if (sum(is.na(z)) != 0) {
        na.z <- which(is.na(z))
    }
    if (sum(is.na(y)) != 0) {
        na.y <- which(is.na(y))
    }

    na <- unique(c(na.x, na.z, na.y))
    keep <- c(1:length(x))[-na]

    # building design matrix and keeping only
    # the rows without NAs
    x <- x[keep]
    z <- z[keep]
    X <- cbind(1, x, z, x * z)
    y <- y[keep]

    # getting the coefficients
    coefs <- solve(t(X) %*% X) %*% (t(X) %*%
        y)
    # get coefficients of interest
    beta_1 <- coefs[2]
    beta_3 <- coefs[4]

    # creating fake data for z set range of z
    min_val <- min(z)
    max_val <- max(z)
    # determine intervals between values of z
    increment <- (max_val - min_val)/100
    # create list of moderator values at
    # which marginal effect is evaluated
    z_sim <- seq(from = min_val, to = max_val,
        by = increment)

    # compute marginal effects
    delta <- beta_1 + beta_3 * z_sim

    # and now for the SEs...  residuals
    e <- y - X %*% coefs
    # estimated sigma squared
    hat_sigma2 <- sum(t(e) %*% e)/(nrow(X) -
        length(coefs))
    # and we can find estimated standard
    # errors from
    # $\hat{\sigma^2}(X'X)^{-1}$.
    varcovmat <- hat_sigma2 * solve((t(X) %*%
        X))

    # compute variances of marginal effects
    var <- varcovmat[2, 2] + (z_sim^2) *
        varcovmat[4, 4] + 2 * z_sim * varcovmat[2,
        4]
    # standard errors
    se <- sqrt(var)

    # Upper and lower confidence bounds
    upper_bound <- delta + qnorm(0.975) *
        se
    lower_bound <- delta - qnorm(0.975) *
        se

    plot(z_sim, delta, type = "l", lwd = 3,
        col = "slateblue", xlab = xlab, ylab = ylab,
        ylim = c(min(lower_bound), max(upper_bound)),
        main = main)
    rug(data$land_crop, col = "slateblue")
    lines(y = upper_bound, x = z_sim, lty = 2,
        lwd = 3, col = "slateblue")
    lines(y = lower_bound, x = z_sim, lty = 2,
        lwd = 3, col = "slateblue")
    abline(h = 0, lwd = 1, col = "deepskyblue")
}


marginal_eff(y = data$gdp_g, x = data$GPCP_g_l,
    z = data$land_crop, ylab = "marginal effects of rainfall (lagged)",
    xlab = "land_crop", main = "marginal effects plot")
```



The effects of rainfall on growth are not conditional on $land_crop$, since the marginal effects are never statistically different from zero: the confidence intervals cover zero for all the levels of $land_crop$.

6. **A researcher is analyzing a data set with observations for two countries and twenty time periods (years). She fits the following model to this data set:**

$$Y_{it} = \alpha_1 + \alpha_2 + \gamma Z_{it} + \epsilon_{it}. \tag{19}$$

Here, Y_{it} **is the value of the dependent variable for country** $i = 1, 2$ **in year** $t = 1, \ldots, 20$**, while** Z_{it} **is the key independent variable. The data set is stacked**

by country, so the first twenty observations are for country 1 and the second twenty are for country 2. The OLS fit for this model is $\hat{\beta} = (X'X)^{-1}X'Y$, where X is the design matrix and Y is the vector of observations on the dependent variable. Here, ϵ_{it} is a random error term, with $\epsilon \perp\!\!\!\perp X$, and α_1, α_2, and γ are parameters. The intercepts α_1 and α_2 for countries 1 and 2 are sometimes called "(country) fixed effects."

(a) **What is the size of Y and of X? What are the elements of X? (That is, describe each column of X). Under the model, what parameters does $\hat{\beta}$ estimate?**

The matrix representation of equation (19) is $Y = X\beta + \epsilon$, where $\beta = (\alpha_1\ \alpha_2\ \gamma)'$. Thus, Y is 40×1; X is 40×3. The first column of X has 1s in the first 20 rows and 0s in the last 20 rows; the second column has 0s in the first 20 rows and 1s in the last 20 rows. Here, $\hat{\beta} = (\hat{\alpha}_1\ \hat{\alpha}_2\ \hat{\gamma})'$ estimates β.

(b) **Explain in a few sentences a rationale for including α_1 and α_2 in a model like equation (19), when analyzing time-series cross-section data.**

The concern is typically that there are some unobservable time-invariant features of countries that are related to Z_{it} and Y_{it}. Including the country fixed effects α_1 and α_2 is said to isolate the "effect of within-country variation." To see why, let $\overline{Y}_i$ be the within-country mean of Y_{it} for country i (i.e., the mean across the 20 time periods), $\overline{Z}_i$ be the within-country mean of Z_{it} for country i, and $\overline{\epsilon}_i$ be the within-country mean of the ϵ_{it} for country i. Equation (19) implies

$$\overline{Y}_i = \alpha_1 + \alpha_2 + \gamma\overline{Z}_i + \overline{\epsilon}_i. \tag{20}$$

(This follows simply from summing all 20 equations for $t = 1, \ldots, 20$ in country i and dividing through by 20). Then, subtracting equation (20) from equation (19), we have

$$Y_{it} - \overline{Y}_i = \gamma(Z_{it} - \overline{Z}_i) + (\epsilon_{it} - \overline{\epsilon}_i). \tag{21}$$

The OLS estimator of the coefficient γ in this transformed model (21) is equivalent to the coefficient $\hat{\gamma}$ from the OLS fit to equation (19). Here, we regress $(Y_{it} - \bar{Y}_i)$ on $(Z_{it} - \bar{Z}_i)$. Thus, the slope coefficient relates within-country deviations from the country mean of the independent variable to within-country deviations from the country mean of the dependent variable. This is why people talk about the isolating the "effect of within-country variation" through the inclusion of country fixed effects.

Despite this talk of effects, causality may or may not be in the picture. Notice that according to (19) and (21), γ is the same for both countries: we are "pooling" the slope coefficient. In effect, we're fitting parallel lines: the regression line for each country has a different intercept but the same slope. This may be a good idea if countries have the same slope but different intercepts. Yet, to infer causality from the regression, we may need a response schedule consistent with (19). In particular, we need structural parameters like α_1, α_2, and γ that are invariant to interventions. And we need to assume $(\epsilon_{it} - \bar{\epsilon}_i) \perp\!\!\!\perp (Z_{it} - \bar{Z}_i)$: the value of the error term is assumed independent of deviations from the mean value of Z_{it} *within* countries.

If this doesn't hold, we're just fitting regression lines—with the same slope for each country but different intercepts.

(c) **Denote the vector of residuals after fitting equation (19) by $e = Y - X\hat{\beta}$, and let $\bar{e}_1 = \frac{1}{20}\sum_{t=1}^{20} e_{1t}$ be the average of the residuals for country 1. Is $\bar{e}_1 = 0$? Be specific in your answer (using algebra if appropriate).**

Yes. Recall that $e'X = 0_{1\times p}$: this is an algebraic property of the least-squares regression fit. In particular, the inner product of e and the first column of X is zero. Note that the first column of X has 1s for the first twenty rows (i.e. the observations for country 1) and 0s for the last twenty rows (i.e. the observations for country 2).

Since e is also stacked by country (the first twenty elements are e_{1t} for $t = 1, \ldots, 20$ and the second twenty elements are e_{2t} for $t = 1, \ldots, 20$), the inner product of e and the first column of X is $\sum_{t=1}^{20}(e_{1t} \cdot 1) + \sum_{t=1}^{20}(e_{2t} \cdot 0) = \sum_{t=1}^{20} e_{1t}$. This equals zero by $e'X = 0$. Thus, $\bar{e}_1 = 0$ as well.

(d) **Suppose that the value of the error term for country i in year i is $\epsilon_{it} = \rho\epsilon_{i,t-1}$, where $\rho \in (0, 1)$. (This is called temporal "autocorrelation" of the errors; you can consider the initial value $\epsilon_{i,1}$ to be fixed for $i \in \{1, 2\}$, since we don't observe a realization of $\epsilon_{i,0}$). Is the OLS estimator $\hat{\beta}$ unbiased?**

Yes. Unbiasedness of the OLS estimator follows from $\epsilon \perp\!\!\!\perp X$; dependence among the error terms does not matter for unbiasedness.

(e) **(This continues the previous sub-question). What is the variance-covariance matrix of ϵ, given X? (Be specific: what are the elements of this matrix?) And what is the variance-covariance matrix of $\hat{\beta}$, given X?**

Note that ϵ is a 40×1 vector $(\epsilon_{11}, \ldots, \epsilon_{1,20}, \epsilon_{2,1}, \ldots, \epsilon_{2,20})$, with the first 20 elements being ϵ_{1t} for $t = 1, \ldots, 20$ and the second 20 being ϵ_{2t} for $t = 1, \ldots, 20$. Thus, the variance-covariance matrix of ϵ is a 40×40 matrix. The ϵ_{it} are not i.i.d., since each epsilon in country i depends on the previous period's epsilon: $\epsilon_{it} = \rho\epsilon_{i,t-1}$.

Let's fix the first-period value of the variance in country i at

$$\mathrm{Var}(\epsilon_{i,1}) = \sigma_{i,1}^2. \tag{22}$$

(This in turn depends on the random variable $\epsilon_{i,0}$; here, $\sigma_{i,1}^2$ is simply notational). Then, what is the variance of the second period's error term in country i? We have

$$\mathrm{Var}(\epsilon_{i2}) = \mathrm{Var}(\rho\epsilon_{i1}) = \rho^2\mathrm{Var}(\epsilon_{i1}) = \rho^2\sigma_{i,1}^2.$$

This process continues recursively. For instance,

$$\mathrm{Var}(\epsilon_{i3}) = \mathrm{Var}(\rho\epsilon_{i2}) = \rho^2\mathrm{Var}(\epsilon_{i2}) = \rho^4\sigma_{i,1}^2,$$

and

$$\text{Var}(\epsilon_{i4}) = \text{Var}(\rho\epsilon_{i3}) = \rho^2\text{Var}(\epsilon_{i3}) = \rho^6\sigma^2_{i,1}.$$

In general, the formula for the variance in country i and time t is

$$\text{Var}(\epsilon_{it}) = \rho^{2t-2}\sigma^2_{i,1}.$$

This formula gives us the diagonal elements of the variance-covariance matrix of ϵ. Next, for the off-diagonal elements, note that

$$\text{Cov}(\epsilon_{i2}, \epsilon_{i,1}) = \text{Cov}(\rho\epsilon_{i,1}, \epsilon_{i,1}) = \rho\text{Var}(\epsilon_{i,1}) = \rho\sigma^2_{i,1}.$$

(Here, we substitute $\epsilon_{it} = \rho\epsilon_{i,t-1}$ and use the fact that the covariance of a random variable with itself is its variance). This process also continues recursively for the other covariances between the $\epsilon_{i,t}$s for country i. For example,

$$\text{Cov}(\epsilon_{i3}, \epsilon_{i,1}) = \text{Cov}(\rho\epsilon_{i,2}, \epsilon_{i,1}) = \rho\text{Cov}(\rho\epsilon_{i,1}, \epsilon_{i,1}) = \rho^2\text{Var}(\epsilon_{i,1}) = \rho^2\sigma^2_{i,1}$$

and

$$\text{Cov}(\epsilon_{i3}, \epsilon_{i,2}) = \text{Cov}(\rho\epsilon_{i,2}, \rho\epsilon_{i,1}) = \rho^2\text{Cov}(\rho\epsilon_{i,1}, \epsilon_{i,1}) = \rho^3\text{Var}(\epsilon_{i,1}) = \rho^3\sigma^2_{i,1}$$

We need not assume that the variance $\sigma^2_{i,1}$ is the same for each $i \in \{1, 2\}$.
In total, we have a block-diagonal matrix with two large blocks on the diagonal—the covariances among the error terms for country 1 and the covariances for country 2. (Given the structure of the errors, covariances are zero across countries but are non-zero within countries across years). Denoting the variance-covariance matrix by G, we have

$$G = \begin{pmatrix} \sigma^2_{1,1} & \rho\sigma^2_{1,1} & \rho^2\sigma^2_{i,1} & . & . & 0 & 0 & 0 & 0 & 0 \\ \rho\sigma^2_{1,1} & \rho^2\sigma^2_{1,1} & \rho^3\sigma^2_{1,1} & . & . & 0 & 0 & 0 & 0 & 0 \\ \rho^2\sigma^2_{i,1} & \rho^3\sigma^2_{1,1} & \rho^4\sigma^2_{1,1} & . & . & 0 & 0 & 0 & 0 & 0 \\ . & . & . & \dots & . & 0 & 0 & 0 & 0 & 0 \\ . & . & . & . & \rho^{38}\sigma^2_{1,1} & . & . & . & . & . \\ 0 & 0 & 0 & 0 & 0 & \sigma^2_{2,1} & \rho\sigma^2_{2,1} & \rho^2\sigma^2_{2,1} & . & . \\ 0 & 0 & 0 & 0 & 0 & \rho\sigma^2_{2,1} & \rho^2\sigma^2_{2,1} & \rho^3\sigma^2_{2,1}. & . & . \\ 0 & 0 & 0 & 0 & 0 & \rho^2\sigma^2_{2,1} & \rho^3\sigma^2_{2,1} & \rho^4\sigma^2_{2,1} & . & . \\ 0 & 0 & 0 & 0 & 0 & . & . & . & \dots & . \\ 0 & 0 & 0 & 0 & 0 & . & . & . & . & \rho^{38}\sigma^2_{2,1} \end{pmatrix} \tag{23}$$

A few technical notes about G. First, the question asks about the variance-covariance matrix given X; here, the unconditional matrix is the same as the conditional matrix, by $X \perp\!\!\!\perp \epsilon$. Second, we are implicitly conditioning on the initial value $\epsilon_{i,0}$ on which $\epsilon_{i,1}$ depends; all the ϵ_{it} are random, though they are of course fully determined by the realization of the initial random error term $\epsilon_{i,0}$). Finally, note that the error process $\epsilon_{it} = \rho\epsilon_{i,t-1}$ is not necessarily a typical assumption in models with autocorrelation of the errors. It might be assumed, for instance, that there is some fixed initial value ϵ_{i0} and that $\epsilon_{it} = \rho\epsilon_{i,t-1} + \nu_{it}$, where ν_{it} is i.i.d. for all i and all t.

As for the variance-covariance matrix of $\hat{\beta}$, given X, this is just

$$\mathrm{Cov}(\hat{\beta}_{\mathrm{OLS}}|X) = (X'X)^{-1}X'GX(X'X)^{-1} \tag{24}$$

(See Freedman 2009: 63, equation 8; also equation 4.10 on p. 45 and exercise B2 in Chapter 5).

(f) **Propose two different ways of estimating the model (19), given temporal autocorrelation of the type discussed in parts (d) and (e). Discuss their advantages and disadvantages. Be as concrete as possible: e.g., in each case, give a formula for an estimator of the model parameters, and say how would you estimate the standard errors of those estimators.**

There are two strategies we covered in this course: OLS and fGLS. For one-step GLS, we estimate the covariance matrix G using the residuals from the OLS regression, then fit the fGLS estimator (as discussed in class and in the readings). For OLS, we use the ordinary least-squares estimator but for the standard errors take the square roots of the diagonal elements of (24), using $\hat{G}$ in place of G. Here, OLS is no longer BLUE, so the GLS estimator may have a smaller variance. However, for both approaches, we have to estimate G. There are only three parameters to estimate (ρ, $\sigma^2_{1,1}$, and $\sigma^2_{2,1}$), but the non-linearity in ρ may make things tricky. For example, the fGLS estimator may be badly biased.

(g) **Suppose we included an overall intercept in the model, so that now $Y_{it} = \alpha + \alpha_1 + \alpha_2 + \gamma Z_{it} + \epsilon_{it}$. What is the new design matrix $\tilde{X}$? What are the consequences for the OLS fit $(\tilde{X}'\tilde{X})^{-1}\tilde{X}'Y$? Explain your answer.**

Now, the design matrix has four columns: the first column is all 1s, the second column has 1s in the first 20 positions and 0 in the second 20 positions, and the third column has 0s in the first 20 positions and 1s in the second 20 positions. (The fourth column has the Z_{it}s). Thus, the first column is a linear combination of the second and third column. (Simply add the second and third columns to get the first column). Thus, the design matrix does not have full rank of 4, and so we can't invert $(\tilde{X}'\tilde{X})$. Thus, the OLS estimator is undefined. Lesson: If you want to include an overall intercept α in the model, you'll need to drop α_1 or α_2.

7. **Suppose that one seeks to estimate the parameter β in the time-series equation**

$$Y_t = \alpha + \beta Y_{t-1} + u_t, \tag{25}$$

where the u_t are independent and identically distributed random error terms. Does OLS regression generate unbiased estimates of β? Explain your answer.

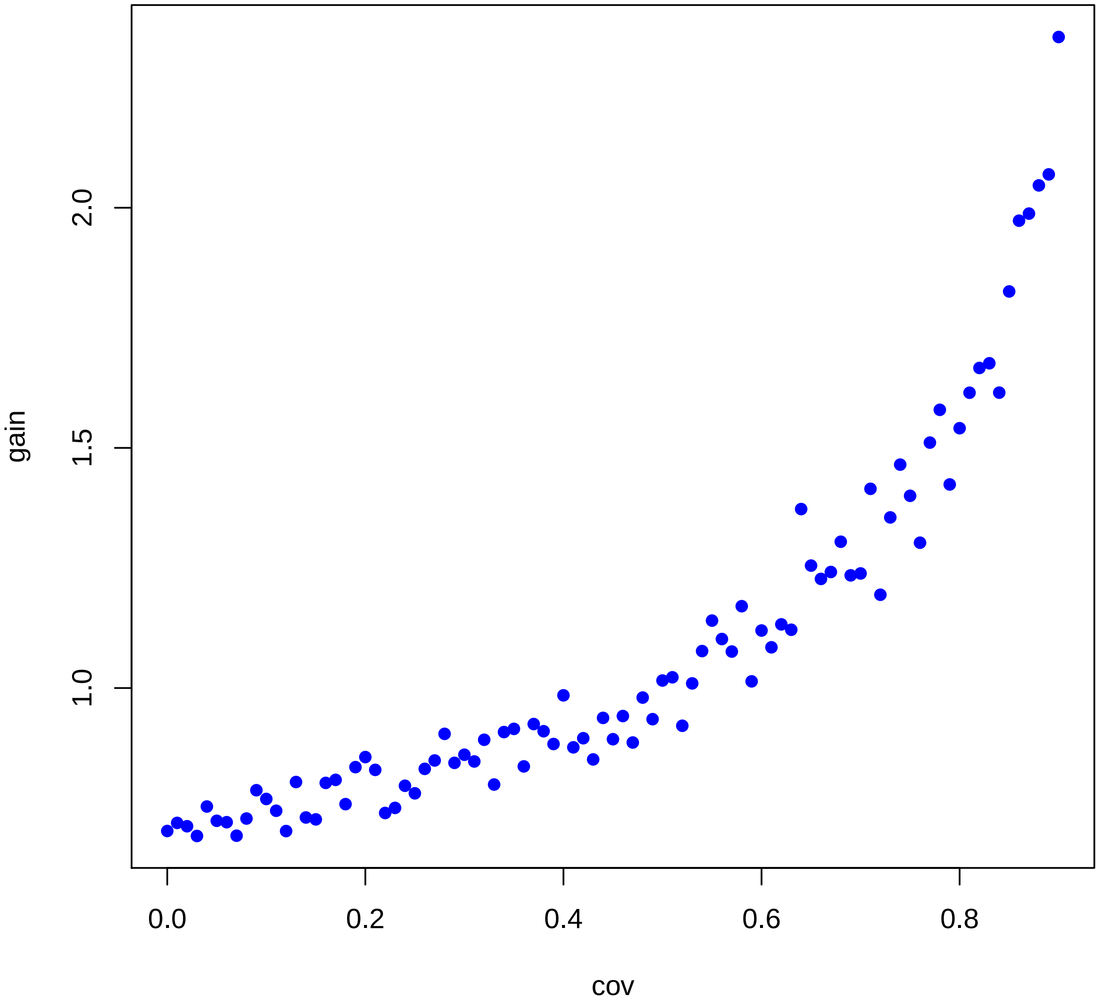
No, it does not. There are T equations here, one for each time period $t = 1, \ldots, T$. We have

$$\begin{array}{rcl} Y_1 & = & \alpha + \beta Y_0 + u_1 \\ Y_2 & = & \alpha + \beta Y_1 + u_2 \\ Y_3 & = & \alpha + \beta Y_2 + u_3 \\ & \ldots & \\ Y_T & = & \alpha + \beta Y_{T-1} + u_T. \end{array}$$

Clearly, Y_1 depends on u_1, Y_2 depends on u_2, and so forth. So the vector of error terms $u = (u_1\, u_2\, \ldots\, u_T)'$ can't be independent of the vector of independent variables Y_{t-1}.

8. **Write a simulation to show that the precision gain from using a difference in differences estimator is a function of the covariance of the potential outcomes and the pre-treatment covariate.**

```
# function to calculate nominal SEs of
# the difference in means
SE_DM <- function(out, treat) {

    var1 <- var(out[treat == 1])/length(out[treat ==
        1])
    var0 <- var(out[treat == 0])/length(out[treat ==
        0])

    return(sqrt(var1 + var0))

}



library(mvtnorm)
set.seed(94705)
cov <- seq(0, 0.9, by = 0.01)
n <- 300

SES <- matrix(NA, length(cov), 2)

for (i in 1:length(cov)) {

    # creating data using cov for the
    # covariance between the pre- treatment
    # variable and the potential outcomes
    cv <- cov[i]
    sigma <- matrix(c(1, cv, cv, cv, 1, cv,
        cv, cv, 1), nrow = 3, ncol = 3)
    data <- as.data.frame(rmvnorm(n, mean = c(10,
        2, 8), sigma = sigma))
    names(data) <- c("pre_treat", "y0", "y1")
    data$treat <- rbinom(n, 1, 0.5)
    data$y_obs <- ifelse(data$treat == 1,
        data$y1, data$y0)
    data$y_diff <- data$y_obs - data$pre_treat

    # cqlculating SEs
    SES[i, ] <- c(SE_DM(out = data$y_obs,
        treat = data$treat), SE_DM(out = data$y_diff,
        treat = data$treat))

}

gain <- SES[, 1]/SES[, 2]  # metric to compare SEs

par(mfrow = c(1, 1))
plot(cov, gain, col = "blue", pch = 16)
```



The plot shows how the precision gain (measured as the ratio between the SE of the difference of means and the SE of the difference of means using a difference in differences estimator) increases substantially as a function of the covariance of the potential outcomes and the pre-treatment covariate. The larger the covariace, the more precision we gain from using the difference in differences estimator.