

# Assignment 2

630516am and 590049bm

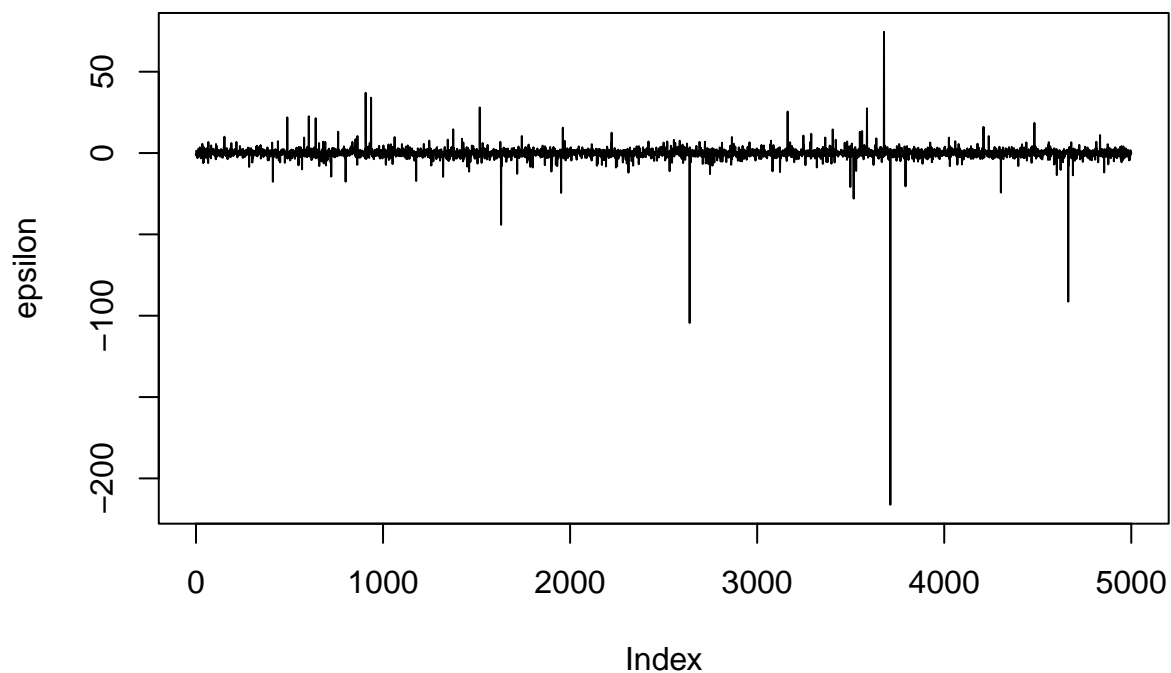
30 Nov 2021

## Question 1:

- a) Assume that  $\gamma = 1$ . Generate 5000 observations and then use OLS to estimate the parameters in the model and calculate the related OLS standard errors, t-values and p-values. Then use OLS to estimate the model with White standard errors<sup>1</sup> and calculate the related White standard errors, t-values and p-values. Compare the results.

```
n = 5000
b0 = 3
b1 = 5
b2 = 8
x1 = rnorm(n, 1, 1)
x2 = rnorm(n, 2, 1)
z = rgamma(n, 1.2, 1.1)
sigma = 1
gamma = 1
```

```
sigmaV = sigma*exp(gamma*z)
epsilon = rnorm(n, 0, sqrt(sigmaV))
plot(epsilon, type = "l")
```



```

y = b0 + b1*x1 + b2*x2 + epsilon
data = data.frame(cbind(y, x1, x2))
model <- lm(y ~ x1 + x2, data=data)
summary(model)

##
## Call:
## lm(formula = y ~ x1 + x2, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -216.090   -1.007    0.104    1.186   74.573
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.83388    0.16279   17.41  <2e-16 ***
## x1           5.00369    0.06495   77.03  <2e-16 ***
## x2           8.03097    0.06627  121.18  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.674 on 4997 degrees of freedom
## Multiple R-squared:  0.805, Adjusted R-squared:  0.8049
## F-statistic: 1.032e+04 on 2 and 4997 DF, p-value: < 2.2e-16

coeftest(model, vcov.= vcovHC(model, type="HCO"))

##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  2.833880    0.137610  20.593 < 2.2e-16 ***
## x1           5.003686    0.044425  112.631 < 2.2e-16 ***
## x2           8.030965    0.046046  174.412 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

We used the `coeftest` function from the `lmtest` library in combination with the function `vcovHC` from the `sandwich` package to calculate a model with the White standard errors. The `coeftest` function calculates the t test using a heteroskedasticity robust variance-covariance matrix produced by the `vcov` function. With “HCO” we indicate that we want to obtain a White standard error (Source: <https://www.r-econometrics.com/methods/hcrobusterrors/>). \

Corresponding to the theory, the estimates are unbiased despite heteroscedasticity. Hence, it is no surprise that the estimates are the same for both regressions. However, heteroscedasticity causes inefficiency of the variance which is why the standard error is higher in the basic OLS regression than in the OLS model with White standard errors.

- b) What are the procedures to perform the Breusch-Pagan test for heteroskedasticity? Perform a Breusch-Pagan test for heteroskedasticity. Provide the value of the test statistic and explain if the null hypothesis is rejected.

In a Breusch-Pagan test the squared OLS residuals are regressed on variables that may relate to the variance. It is assumed that heteroscedasticity is driven by  $z_i$ . Hence, the null hypothesis is that  $\gamma_2, \dots, \gamma_n$  are zero. \

To perform the test, we have to estimate  $y$  with OLS and compute the residuals in the first step. Thereafter, we perform an auxiliary regression of the form  $e_i^2 = \gamma_1 + \gamma_2 z_{2i} + \dots + \gamma_p z_{pi} + \eta_i$ . Taking  $R^2$  of our auxiliary

regression, we can calculate  $LM = nR^2$  for our test statistic.

```
usq <- resid(model)^2
# auxiliary regression: dependent variable squared residuals of first regression, explanatory variables
res <- lm(usq ~ z, data)
summary(res)

##
## Call:
## lm(formula = usq ~ z, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
##    -704    -68      19      76   45517
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -125.732     14.315  -8.783  <2e-16 ***
## z             135.381      9.714  13.937  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 681.2 on 4998 degrees of freedom
## Multiple R-squared:  0.03741,    Adjusted R-squared:  0.03722
## F-statistic: 194.2 on 1 and 4998 DF,  p-value: < 2.2e-16

nres <- nobs(res)
Rsqr<-summary(res)$r.squared

# test statistic
BP <- nres*Rsqr
BP

## [1] 187.0461
1-pchisq(BP,2)
```

```
## [1] 0
```

```
# reject H0
```

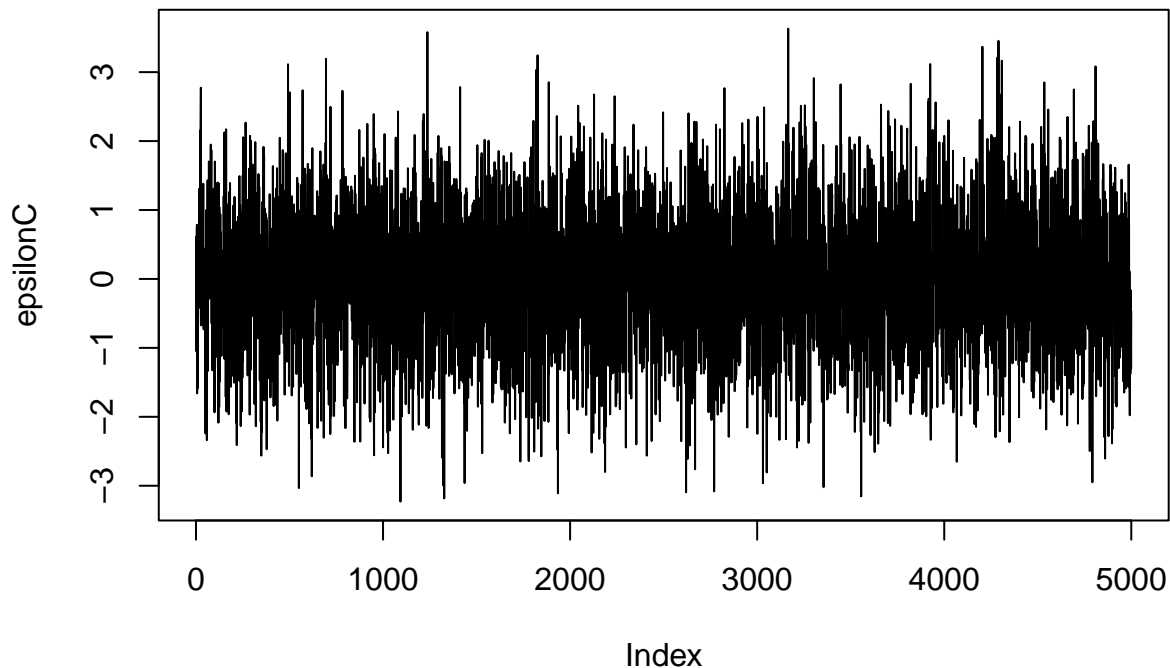
If  $z_i$  would not drive the heteroscedasticity, the model would not have any explanatory power. Hence,  $R^2$  would be close to zero as well as LM (or BR as denoted in the code). However, as the result of the test statistic shows, BR is not close to zero which gives us evidence that  $z_i$  indeed drives the heteroscedasticity. Therefore, we reject the null hypothesis. \

c) Assume that  $\gamma = 0$ . Estimate  $\beta_0, \beta_1$  and  $\beta_2$  separately using \

1. OLS \
2. WLS with known  $\gamma = 0$  \
3. FWLS with estimated  $\gamma$  (i.e.  $\gamma$  is unknown) \

Explain the weights you use for WLS and FWLS. Provide the coefficients and standard errors of the three estimators for three methods and compare the results. Are the estimators close to the their true values? \

```
gammaC = 0
sigmaC = sigma*exp(gammaC*z)
epsilonC = rnorm(n, 0, sqrt(sigmaC))
plot(epsilonC, type = "l")
```



```
yC = b0 + b1*x1 + b2*x2 + epsilonC
dataC = data.frame(cbind(yC, x1, x2))
modelC <- lm(yC ~ x1 + x2, data = dataC)
summary(modelC)
```

```
##
## Call:
## lm(formula = yC ~ x1 + x2, data = dataC)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.2104 -0.6910 -0.0033  0.6863  3.6539
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.93799    0.03510   83.71  <2e-16 ***
## x1             5.00834    0.01400  357.64  <2e-16 ***
## x2             8.01890    0.01429  561.22  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.008 on 4997 degrees of freedom
## Multiple R-squared:  0.9889, Adjusted R-squared:  0.9888
## F-statistic: 2.216e+05 on 2 and 4997 DF, p-value: < 2.2e-16
```

For the weighted least square model, we use  $\exp(\gamma z_i)$  as our weight, as this is the component that drives the heteroskedasticity in our errors:  $w_i = \frac{1}{e^{\gamma z_i}}$ . As  $\gamma = 0$ , the regression should result in the same estimators and same variances as in the basic OLS regression.

```
# ----- WLS
w <- 1/exp(gammaC*z) # specify weight
modelWLS <- lm(yC ~ x1 + x2, data = dataC, weights=w)

summary(modelWLS)$coeff
```

```
##           Estimate Std. Error   t value Pr(>|t|)
## (Intercept) 2.937985 0.03509769  83.70881      0
## x1          5.008338 0.01400366 357.64500      0
## x2          8.018896 0.01428827 561.22231      0
```

For the feasible weighted least square model, we first use the residuals of the basic OLS regression to estimate  $\gamma$ . Then we use this estimator to calculate  $\exp(\gamma z_i)$  which is then used as our weight as this is the component which drives the heteroskedasticity in our errors.

```
# --- FWLS
eps <- modelC$residuals
modelEps <- lm(eps ~ z)
gammaFWLS <- modelEps$coefficient[2]
w <- 1/exp(gammaFWLS*z) # specify weight
modelFWLS <- lm(yC ~ x1 + x2, data = dataC, weights=w)

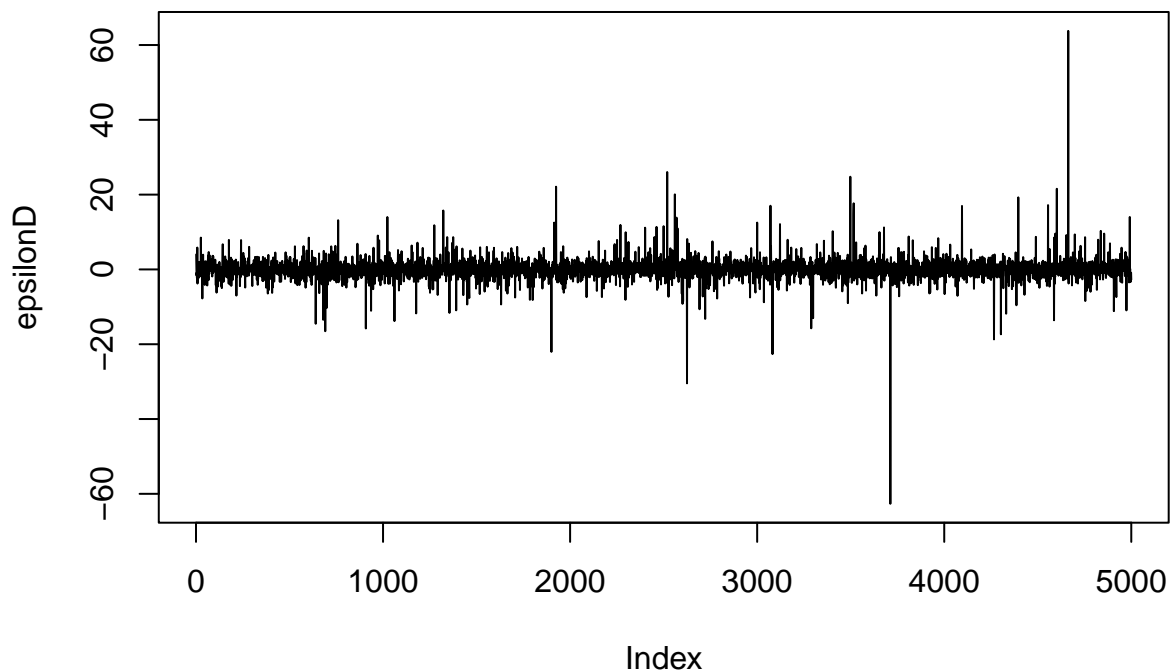
summary(modelFWLS)$coeff
```

```
##           Estimate Std. Error   t value Pr(>|t|)
## (Intercept) 2.937786 0.03511869  83.65304      0
## x1          5.008082 0.01400596 357.56783      0
## x2          8.018912 0.01429610 560.91621      0
```

The estimators and variances are approximately equal in all estimated models (there are minor differences from the fourth digit on). Only the variance in the feasible weighted least square model is slightly higher. However, the difference is minimal. For  $x_1, x_2$  the estimators are very close to their true value. The intercept also close although there is a higher difference then for the other estimators.\

- d) Now assume that  $\gamma = 1$ . Repeat sub-question (c). Provide the coefficients and standard errors of the three estimators for three methods and compare the results. Are the estimators close to the true ones

```
gammaD = 1
sigmaD = sigma*exp(gammaD*z)
epsilonD = rnorm(n, 0, sqrt(sigmaD))
plot(epsilonD, type = "l")
```



```

yD = b0 + b1*x1 + b2*x2 + epsilonD
dataD = data.frame(cbind(yD, x1, x2))
modelD <- lm(yD ~ x1 + x2, data = dataD)
summary(modelD)

##
## Call:
## lm(formula = yD ~ x1 + x2, data = dataD)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -62.696  -1.117  -0.031   1.051  63.751
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.04556    0.10044   30.32  <2e-16 ***
## x1           4.98751    0.04008  124.45  <2e-16 ***
## x2           7.99606    0.04089  195.55  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.884 on 4997 degrees of freedom
## Multiple R-squared:  0.915, Adjusted R-squared:  0.9149
## F-statistic: 2.688e+04 on 2 and 4997 DF, p-value: < 2.2e-16

# ----- WLS
w <- 1/exp(gammaD*z) # specify weight
modelWLSd <- lm(yD ~ x1 + x2, data = dataD, weights=w)

summary(modelWLSd)$coeff

##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  3.007112 0.05114675  58.79381      0
## x1           5.019518 0.02065300 243.04060      0
## x2           7.987665 0.02086295 382.86359      0

# --- FWLS
epsD <- modelD$residuals
modelEpsD <- lm(epsD ~ z)
gammaFWLSD <- modelEpsD$coefficient[2]
w <- 1/exp(gammaFWLSD*z) # specify weight
modelFWLSD <- lm(yD ~ x1 + x2, data = dataD, weights=w)

summary(modelFWLSD)$coeff

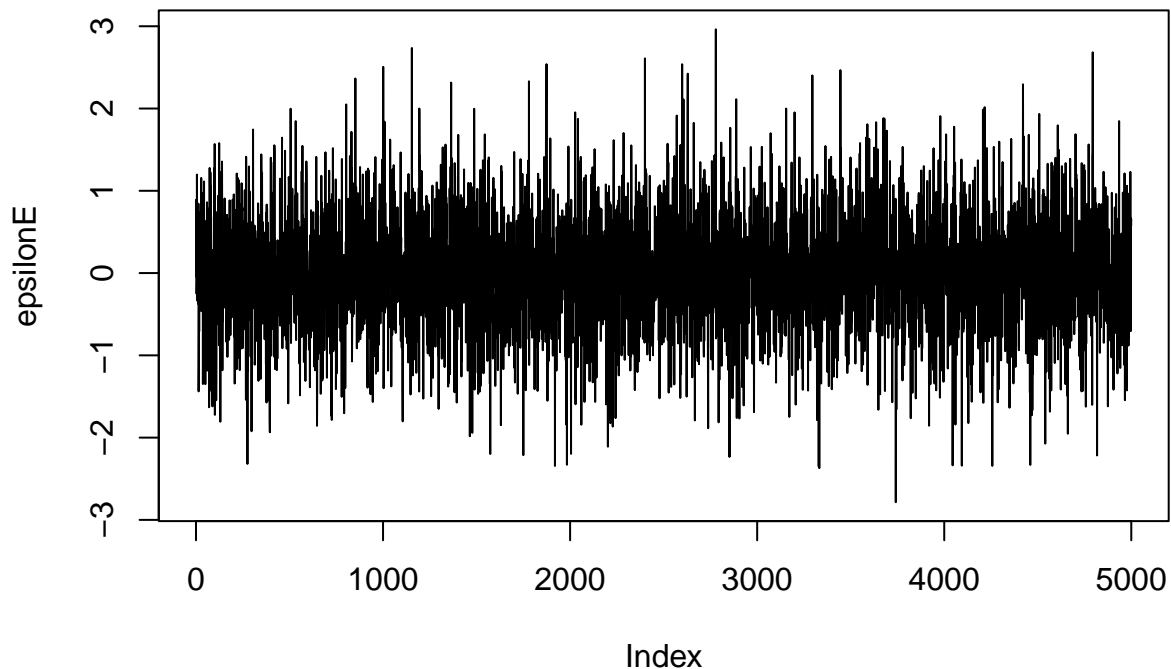
##              Estimate Std. Error  t value      Pr(>|t|)
## (Intercept)  3.041220 0.09000512  33.78941 1.347516e-225
## x1           4.992763 0.03596599 138.81900 0.000000e+00
## x2           7.992923 0.03664916 218.09295 0.000000e+00

```

The estimators of all three models are very close to the true values. However, the variance is higher in the basic OLS model than in the WLS and FWLS models which is not surprising given the heteroskedasticity of the errors.

- e) Now assume that  $\gamma = -1$ . Repeat sub-question (c). Provide the coefficients and standard errors of the three estimators for three methods and compare the results. Are the estimators close to the true ones?

```
gammaE = -1
sigmaE = sigma*exp(gammaE*z)
epsilonE = rnorm(n, 0, sqrt(sigmaE))
plot(epsilonE, type = "l")
```



```
yE = b0 + b1*x1 + b2*x2 + epsilonE
dataE = data.frame(cbind(yE, x1, x2))
modelE <- lm(yE ~ x1 + x2, data = dataE)
summary(modelE)
```

```
##
## Call:
## lm(formula = yE ~ x1 + x2, data = dataE)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.79057 -0.38847 -0.00826  0.37916  2.96861
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.984309   0.023259   128.3  <2e-16 ***
## x1           4.979190   0.009280   536.5  <2e-16 ***
## x2           8.018769   0.009469   846.9  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6678 on 4997 degrees of freedom
## Multiple R-squared:  0.9951, Adjusted R-squared:  0.9951
## F-statistic: 5.028e+05 on 2 and 4997 DF, p-value: < 2.2e-16

# ----- WLS
w <- 1/exp(gammaE*z) # specify weight
modelWLSe <- lm(yE ~ x1 + x2, data = dataE, weights=w)
```

```
summary(modelWLSE)$coeff

##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  2.998324  0.011616706  258.1045      0
## x1           5.000555  0.004841276 1032.9000      0
## x2           8.004173  0.005105967 1567.6116      0

# --- FWLS
epsE <- modelE$residuals
modelEpsE <- lm(epsE ~ z)
gammaFWLSE <- modelEpsE$coefficient[2]
w <- 1/exp(gammaFWLSE*z) # specify weight
modelFWLSE <- lm(yE ~ x1 + x2, data =dataE, weights=w)

summary(modelFWLSE)$coeff

##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  2.984264  0.023276573  128.2089      0
## x1           4.979127  0.009287794  536.0936      0
## x2           8.018818  0.009475978  846.2259      0
```

The estimators are still close to their true values. However, the difference is slightly higher than in the previous estimations. A very interesting aspect is that the standard errors are approximately the same across all models (it this maybe a mistake?).



## Question 2

1. Show that the OLS estimator of the parameter  $\beta$  is not consistent.
2. Derive  $\text{plim}(\hat{\beta})$  where  $\hat{\beta}$  is the OLS estimator of  $\beta$ . Determine the sign of the magnitude of the inconsistency when  $0 < \beta < 1$ , that is, the sign of  $\text{plim}(\hat{\beta}) - \beta$  when  $0 < \beta < 1$ .

In order to evaluate consistency, we must derive the probability limit. Hence, we answer two questions at once.

First, we demean the two variables so that the constant-term  $\alpha$  equals zero. Then we regress  $\tilde{C} = \beta\tilde{D} + \epsilon$ . We can do this because of Frisch-Waugh-Lovell. The estimate that we get is:

$$\hat{\beta} = (\tilde{D}^T \tilde{D})^{-1} \tilde{D}^T C = (\tilde{D}^T \tilde{D})^{-1} (\beta \tilde{D} + \epsilon)$$

and

$$\mathbb{E}[\hat{\beta}] = \beta + (\tilde{D}^T \tilde{D})^{-1} \tilde{D}^T \epsilon$$

Evaluating the probability limit gives:

$$\text{plim}_{n \rightarrow \infty}(\hat{\beta}) = \beta + \text{plim}\left(\frac{1}{n} \tilde{D}^T \tilde{D}\right)^{-1} \cdot \text{plim}\left(\frac{1}{n} \tilde{D}^T \epsilon\right)$$

which simplifies to:

$$\beta + \frac{1}{\text{Var}(D)} \cdot \frac{1}{1 - \beta} \sigma^2$$

by the fact that variances and covariances are the same after demeaning, and by the reduced form equation for  $D$  made explicit below. Under  $0 < \beta < 1$ , since variances are positive, the right term can only be positive and thus the bias is always positive.

Substituting equation (2) into equation (1) and solving for  $C$  gives:

$$C = \frac{\alpha}{1 - \beta} + \frac{\beta}{1 - \beta} Z_i + \frac{1}{1 - \beta} \epsilon_i$$

substituting this back in the definition for  $D$  gives:

$$D = \frac{\alpha}{1 - \beta} + \left( \frac{\beta}{1 - \beta} + 1 \right) Z_i + \frac{1}{1 - \beta} \epsilon_i$$

From this, we can calculate  $\text{Cov}(D, \epsilon_i)$ , which is  $\frac{1}{1 - \beta} \text{Var}(\epsilon) = \frac{1}{1 - \beta} \sigma^2$ .

3. Find an instrumental variable (IV) for the endogenous variable  $D$  and argue why it could be an IV.

The instrumental variable could be  $Z$ , because it is relevant, i.e.  $\text{Cov}(D, Z) \neq 0$ . Also, it is exogenous (valid), as it is exogenously generated and has no correlation with the error term  $\epsilon$  according to the DGP sketched out here.

4. Derive  $b_{IV}$ , the IV estimator of  $\beta$  in terms of the variables  $C$ ,  $D$ , and  $Z$  step by step.

First, suppose  $X$  is a matrix consisting of a column of 1's and  $D$ , so we can write:

$$\mathbb{E}[z_i \epsilon_i] = 0 = \frac{1}{n} \sum z_i (c_i - \alpha - \beta D_i) = \frac{1}{n} \sum z_i (c_i - x_i \beta)$$

Using this moment condition to solve for  $\beta$ , we retrieve the  $b_{IV}$  estimator:

$$\hat{b}_{IV} = \left( \sum z_i^T x_i \right)^{-1} (z_i^T c_i) = (Z^T X)^{-1} Z^T C$$

5. Use the expression of  $b_{IV}$  to show that it is consistent.

$$b_{IV} = (Z^T X)^{-1} Z^T C = (Z^T X)^{-1} Z^T (X\beta + \epsilon) = (Z^T X)^{-1} Z^T X\beta + (Z^T X)^{-1} Z^T \epsilon = \beta + (Z^T X)^{-1} Z^T \epsilon$$

Evaluating the plim of this estimator then gives:

$$\text{plim}(b_{IV}) = \beta + \text{plim}(Z^T X)^{-1} \cdot \text{plim}(Z^T \epsilon)$$

where the last factor goes to zero as  $n \rightarrow \infty$ .

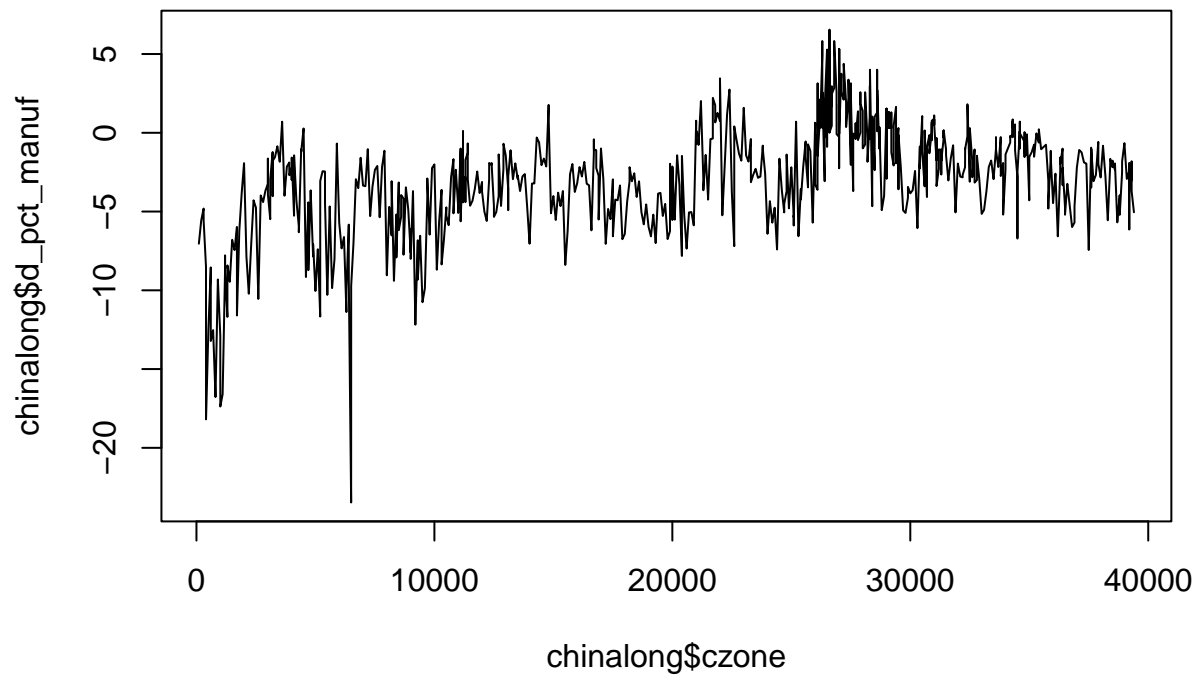
### Question 3

```
china <- read_dta("workfile_china.dta")
chinalong <- read_dta("workfile_china_long.dta")
chinapreperiod <- read_dta("workfile_china_preperiod.dta")
```

- a) Plot the distribution of the growth rate of employment and of import exposure 1990-2007 across US commuting zones.

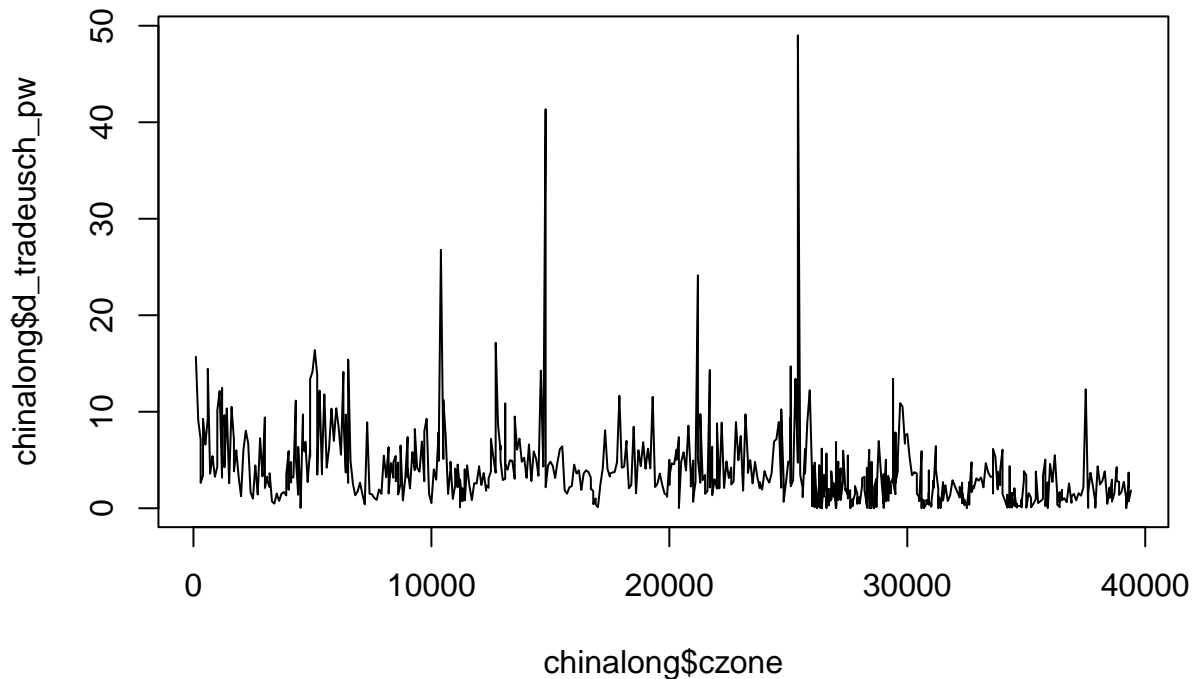
```
plot(chinalong$czone, chinalong$d_pct_manuf, type="l", main = "distribution of the growth rate of employment")
```

#### distribution of the growth rate of employment



```
plot(chinalong$czone, chinalong$d_tradeusch_pw, type="l", main = "distribution of import exposure")
```

## distribution of import exposure



- b) Regress import exposure on the growth rate of employment from 1990- 2007. Plot your results. You should be able to reproduce panel B of Figure 2. Compute normal OLS standard errors and HAC standard errors clustered by the state levels (hint use the `vcovHAC` command from the `sandwich` package) and compare them.

```
model <- lm(d_pct_manuf ~ d_tradeusch_pw, data=chinalong)

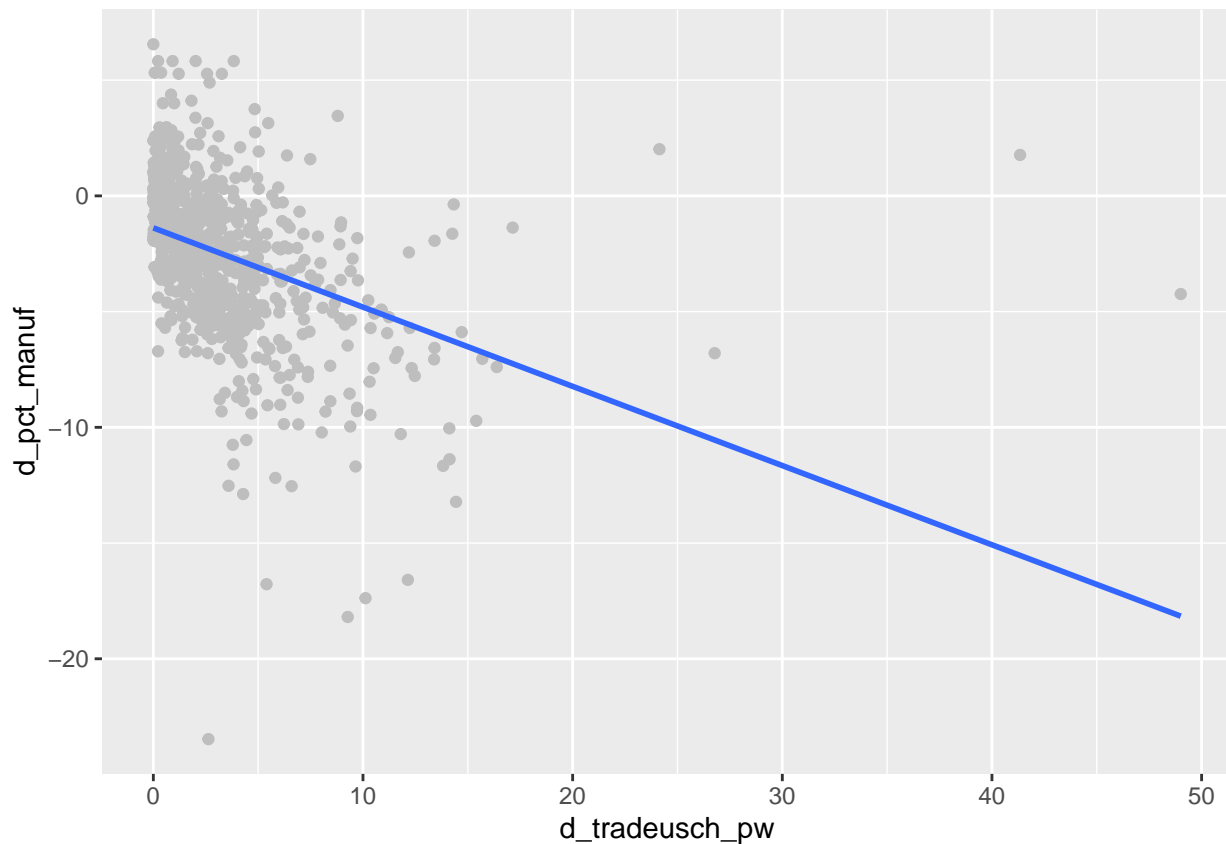
summary(model)

##
## Call:
## lm(formula = d_pct_manuf ~ d_tradeusch_pw, data = chinalong)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.1873  -1.6181   0.1655   1.6721  17.3031
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.38292    0.15619  -8.854  <2e-16 ***
## d_tradeusch_pw -0.34231    0.02981 -11.484  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.176 on 720 degrees of freedom
## Multiple R-squared:  0.1548, Adjusted R-squared:  0.1536
## F-statistic: 131.9 on 1 and 720 DF, p-value: < 2.2e-16

chinalong %>%
  ggplot(aes(x = d_tradeusch_pw, y = d_pct_manuf )) +
  geom_point(colour = "grey") +
```

```
geom_smooth(method = "lm", fill = NA) #+
```

```
## `geom_smooth()` using formula 'y ~ x'
```



```
#hrbrthemes::theme_ipsum()
```

```
# controlling for clustered errors
coeftest(model, vcov = vcovCL, type = "HC1", cluster = ~statefip)
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.38292    0.45014  -3.0722 0.0022051 **
## d_tradeusch_pw -0.34231    0.10356  -3.3054 0.0009953 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# HAC - controlling for heteroskedasticity- and autocorrelation-consistent errors
coeftest(model, vcov = vcovHAC, type = "HC1", cluster = ~statefip)
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.382917   0.305007 -4.5340 6.776e-06 ***
## d_tradeusch_pw -0.342313   0.091021 -3.7608 0.0001831 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The estimators are approximately the same. However, the standard errors are smaller for the model with HAC standard errors clustered by the state levels as we control for heteroskedasticity and autocorrelation.

- c) Is this a good causal estimate of the effect of import exposure on employment? Give a reason why or why not.

It is rather seldom that a regression with only one explanatory variable is a sufficient causal estimate. While a higher import exposure indeed correlates with higher unemployment as a higher part of production is outsourced to other countries, we also could consider the export exposure (as exports correlate positively with employment). The question is also what determines the amount of imports in the USA (add more!).

- d) The authors construct an instrument for import exposure using the growth rate of Chinese imports in eight other similar countries. Construct the instrumental variable estimate of the effect of the growth of import exposure on the growth of employment using the instrument from the data in "workfile china.dta". Do so in two ways. First, use a package. Then use matrix multiplication. Present regression results for both. Do not include any additional controls for now. To show that you have done the matrix multiplication is correct, report the third entry of the projection matrix of the instrument times the endogenous variable i.e of  $P_z X$ .

```
# Package
# IV regression using RPT, RPN and RPU as instruments
modellIV <- ivreg(d_pct_manuf ~ d_tradeusch_pw | d_tradeotch_pw_lag, data=chinalong)
summary(modellIV)
```

```
##
## Call:
## ivreg(formula = d_pct_manuf ~ d_tradeusch_pw | d_tradeotch_pw_lag,
##       data = chinalong)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.35972  -1.70983   0.04751   1.62787  25.50194
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.64208    0.18387  -3.492 0.000509 ***
## d_tradeusch_pw -0.55855    0.04004 -13.949 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.29 on 720 degrees of freedom
## Multiple R-Squared:  0.09303, Adjusted R-squared:  0.09177
## Wald test: 194.6 on 1 and 720 DF,  p-value: < 2.2e-16
```

```
#-----
# ---- 2SLS (without packages)
n      = nobs(modellIV)
Z      <- cbind(rep(1,n),chinalong$d_tradeotch_pw_lag) # instruments
X      <- cbind(rep(1,n),chinalong$d_tradeusch_pw)
y      <- as.matrix(chinalong$d_pct_manuf)
Xhat   <- Z%*%solve(t(Z)%*%Z)%*%t(Z)%*%X
B      <- solve(t(Xhat)%*%Xhat)%*%t(Xhat)%*%y # 2SLS estimate
eIV    <- y-X%*%B
k2     <- 2 # number of regressors of second step
sigmasq <- as.numeric((t(eIV)%*%eIV)/(n-k2))
```

```
sder    <- sqrt(diag(sigmasq*solve(t(Xhat)%*%Xhat))) # standard errors of B
#-----
```

Third entry of  $P_z X$ :

```
Xhat[3,2]
```

```
## [1] 2.213496
```

- e) You might notice that your results are different from the results in the paper. The authors use weighted estimates, where the weights are shares of manufacturing employment. Now reproduce the results in Table 2 and Table 3 in the paper exactly as they do (so two tables containing all the coefficient estimates). For the first, you will need to use `workfile china preperiod.dta`. For the second, use `"workfile china.dta"` instead. Report the first-stage F statistics. Is the instrument a good instrument?

```
model1 <- coefest(ivreg(d_sh_empl_mfg ~ d_tradeusch_pw | d_tradeotch_pw_lag, data=subset(chinapreperiod,
model2 <- coefest(ivreg(d_sh_empl_mfg ~ d_tradeusch_pw | d_tradeotch_pw_lag, data=subset(chinapreperiod,
model3 <- coefest(ivreg(d_sh_empl_mfg ~ d_tradeusch_pw + t2000 | d_tradeotch_pw_lag + t2000, data=subset(chinapreperiod,
model4 <- coefest(ivreg(d_sh_empl_mfg ~ d_tradeusch_pw_future | d_tradeotch_pw_lag_future, data=subset(chinapreperiod,
model5 <- coefest(ivreg(d_sh_empl_mfg ~ d_tradeusch_pw_future | d_tradeotch_pw_lag_future, data=subset(chinapreperiod,
model6 <- coefest(ivreg(d_sh_empl_mfg ~ d_tradeusch_pw_future + t1980 | d_tradeotch_pw_lag_future + t1980, data=subset(chinapreperiod,

stargazer(model1, model2, model3, model4, model5, model6,
            header=FALSE,
            column.sep.width="0pt")
```

Table 1:

	<i>Dependent variable:</i>					
	(1)	(2)	(3)	(4)	(5)	(6)
d_tradeusch_pw	-0.888*** (0.183)	-0.718*** (0.065)	-0.746*** (0.069)			
t2000			0.444 (0.327)			
d_tradeusch_pw_future				0.431*** (0.151)	-0.130 (0.127)	0.148 (0.096)
t1980						-1.945*** (0.250)
Constant	-1.056*** (0.195)	-0.846*** (0.258)	-1.218*** (0.140)	-0.954*** (0.319)	-1.832*** (0.334)	-0.415 (0.302)

Note:

\*p<0.1; \*\*p<0.05; \*\*\*p<0.01

```
model1Stage1 <- lm(d_tradeusch_pw ~ d_tradeotch_pw_lag, data=subset(chinapreperiod, yr==1990), vcov = v)

## Warning: In lm.wfit(x, y, w, offset = offset, singular.ok = singular.ok,
##      ...):
## extra arguments 'vcov', 'cluster' will be disregarded
```

```

model2Stage1 <- lm(d_tradeusch_pw ~ d_tradeotch_pw_lag, data=subset(chinapreperiod, yr==2000), vcov = v

## Warning: In lm.wfit(x, y, w, offset = offset, singular.ok = singular.ok,
##      ...):
## extra arguments 'vcov', 'cluster' will be disregarded
model3Stage1 <- lm(d_tradeusch_pw + t2000 ~ d_tradeotch_pw_lag+ t2000, data=subset(chinapreperiod, yr==1

## Warning: In lm.wfit(x, y, w, offset = offset, singular.ok = singular.ok,
##      ...):
## extra arguments 'vcov', 'cluster' will be disregarded
model4Stage1 <- lm(d_tradeusch_pw_future ~ d_tradeotch_pw_lag_future, data=subset(chinapreperiod, yr==1

## Warning: In lm.wfit(x, y, w, offset = offset, singular.ok = singular.ok,
##      ...):
## extra arguments 'vcov', 'cluster' will be disregarded
model5Stage1 <- lm(d_tradeusch_pw_future ~ d_tradeotch_pw_lag_future, data=subset(chinapreperiod, yr==1

## Warning: In lm.wfit(x, y, w, offset = offset, singular.ok = singular.ok,
##      ...):
## extra arguments 'vcov', 'cluster' will be disregarded
model6Stage1 <- lm(d_tradeusch_pw_future + t1980 ~ d_tradeotch_pw_lag_future + t1980, data=subset(chinap

## Warning: In lm.wfit(x, y, w, offset = offset, singular.ok = singular.ok,
##      ...):
## extra arguments 'vcov', 'cluster' will be disregarded
stargazer(model1Stage1, model2Stage1, model3Stage1, model4Stage1, model5Stage1, model6Stage1,
  header = FALSE,
  column.sep.width = "0pt",
  font.size = "tiny",
  df = FALSE)

```

Table 2:

	Dependent variable:					
	d_tradeusch_pw (1)	d_tradeusch_pw (2)	d_tradeusch_pw + t2000 (3)	d_tradeusch_pw_future (4)	d_tradeusch_pw_future (5)	d_tradeusch_pw_future + t1980 (6)
d_tradeotch_pw_lag	0.950*** (0.040)	0.767*** (0.032)	0.792*** (0.023)			
t2000			1.237*** (0.072)			
d_tradeotch_pw_lag_future				0.929*** (0.028)	0.926*** (0.028)	0.927*** (0.020)
t1980						0.996*** (0.045)
Constant	0.187*** (0.048)	0.647*** (0.099)	0.346*** (0.050)	0.218*** (0.060)	0.219*** (0.060)	0.220*** (0.048)
Observations	722	722	1,444	722	722	1,444
R <sup>2</sup>	0.444	0.446	0.655	0.603	0.608	0.651
Adjusted R <sup>2</sup>	0.443	0.446	0.655	0.602	0.607	0.651
Residual Std. Error	0.028	0.056	0.044	0.032	0.032	0.032
F Statistic	574.852***	580.768***	1,369.648***	1,091.543***	1,114.909***	1,344.611***

Note:

\*p<0.1; \*\*p<0.05; \*\*\*p<0.01

The instruments are good, as they are significant in the first stage and the F statistics are above 10.



## Question 4

1. Imagine we fit a linear probability model of  $y_i = \alpha + \beta x_i + \epsilon_i$ . Derive the distribution of the error terms. Will our least-squares parameter estimate  $\beta$  be unbiased? Will it still be the most efficient estimator?

We know that  $y_i$  is distributed with probability  $p$ . If we use a linear model to estimate a  $y$ , we impose that  $p(y_i = 1) = \mathbb{E}[y_i] = \alpha + \beta x_i$ . Then, we can characterize the distribution of the error term:

$$\epsilon_i = \begin{cases} 1 - \hat{\alpha} - \hat{\beta}x_i & \text{with } p = \alpha + \beta x_i \\ -\hat{\alpha} - \hat{\beta}x_i & \text{with } p = 1 - (\alpha + \beta x_i) \end{cases}$$

Then, since  $\epsilon_i$  is now a shifted Bernoulli variable, we can calculate the expected value as:

$$\mathbb{E}[\epsilon_i] = (1 - \alpha - \beta x_i) \cdot (\alpha + \beta x_i) + (-\alpha - \beta x_i) \cdot (1 - \alpha - \beta x_i) = 0$$

The fact that  $\mathbb{E}[\epsilon] = 0$  also means that the OLS estimator is unbiased. However, the variance  $\sigma_\epsilon^2$  as  $p(1-p) = (\alpha + \beta x_i) \cdot (1 - (\alpha + \beta x_i)) = f(x_i)$ . This means that the variance of the error term is heteroskedastic! Hence, the estimator will not be the most efficient estimator, as one of the Gauss-Markov assumptions is violated.

2. Now imagine that we want to estimate this regression model for a given distribution of the errors  $F$  (e.g the logistic distribution) using maximum likelihood. Write out the distribution of  $y_i$ .

In this case, more generally,  $y_i$  is distributed as:

$$y_i = \begin{cases} 1 & \text{with } p = F(X_i\beta) \\ 0 & \text{with } 1 - p = 1 - F(X_i\beta) \end{cases}$$

The likelihood of one observation (which is the pdf) is then simply:

$$l_1(y_i|x_i) = (F(X_i\beta))^{y_i} (1 - F(X_i\beta))^{1-y_i}$$

3. Use the distribution to write out the log-likelihood function. Then, write out the first-order condition for maximisation with respect to  $\beta$ .

The log-likelihood for  $n$  observations is:

$$\mathcal{L}_n(y_i|x_i) = \sum_{i=1}^n y_i \log(F(X_i\beta)) + (1 - y_i) \log(1 - F(X_i\beta))$$

Taking the first derivative with respect to the parameters  $\beta$  gives:

$$\frac{\partial \log \mathcal{L}_n(y_i|x_i)}{\partial \beta} = \sum y_i \frac{1}{F(X_i\beta)} f(X_i\beta) X_i - (1 - y_i) \frac{1}{1 - F(X_i\beta)} f(X_i\beta) X_i = 0$$

This can be rewritten as:

$$\sum \frac{y_i - F(X_i\beta)}{F(X_i\beta)(1 - F(X_i\beta))} f(X_i\beta) X_i = 0$$

4. Imagine we assume a logistic distribution of the errors. Show that our expression above simplifies to ...

We use the fact from Heij et al., p. 449, that for the logistic distribution,  $F(\cdot)(1 - F(\cdot)) = f(\cdot)$ . Then, our expression simplifies to:

$$\sum y_i - F(X_i\beta) x_i = 0$$

Then, substituting the logit cdf for  $F$  gives:

$$\sum y_i - \left( \frac{1}{1 + \exp^{-x_i\beta}} \right) x_i = 0$$

which is what we were required to show.

5. Finally, use the value of  $F(x_i\beta)$  to write the log of the odds ratio as a function of the parameters of the model. Thus, give an interpretation of the value of  $\beta$ .

The log odds ratio is defined as:

$$OR = \frac{\frac{1}{1 + \exp^{-x_i\beta}}}{1 - \frac{1}{1 + \exp^{-x_i\beta}}} = e^{x_i\beta}$$

The log-odds ratio is then:

$$\log OR = x_i\beta$$

Beta is then equal to the derivative of the log odds ratio with respect to a regressor. This means that the strength of  $\beta$  is indicative of the relative likelihood of  $P(Y_i = 1)$  occurring versus  $P(Y_i = 0)$  occurring. In other words, if  $\beta > 0$ , then an increase in the independent variable makes the event more likely, and a decrease in the independent variable makes the event less likely.