Assignment 2

630516am and 590049bm

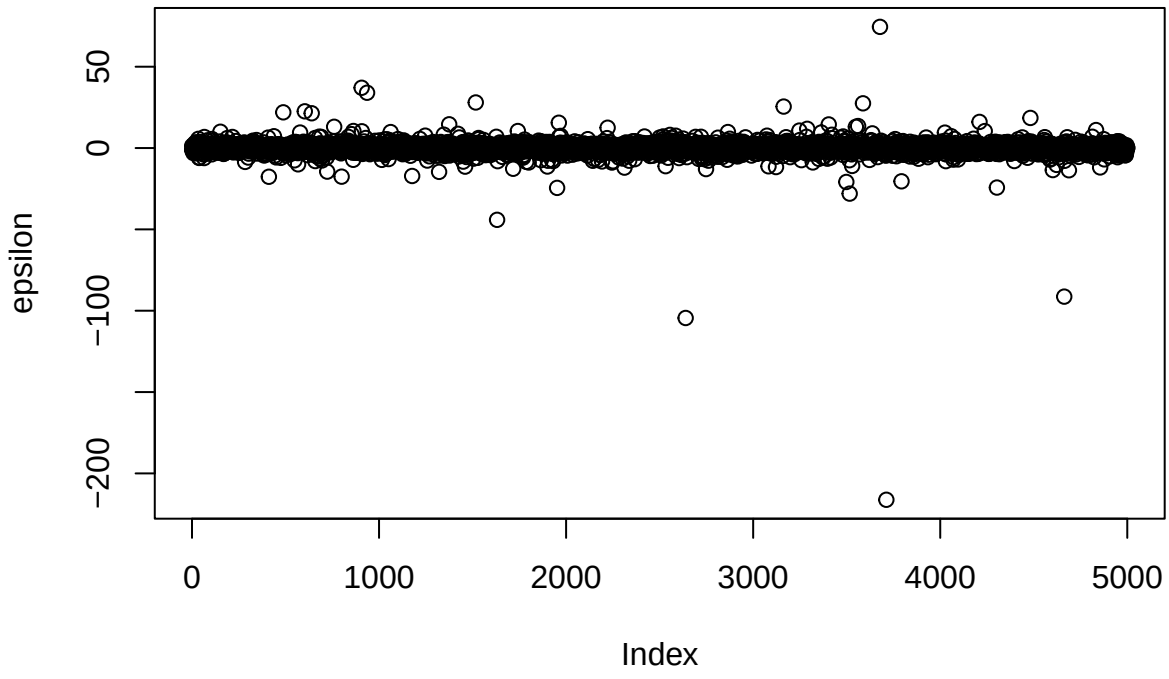
30 Nov 2021

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 errors1 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 = "p")
```



```
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="HC0"))
```

```
##
## 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 $\gamma_2, ..., \gamma_n$ 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 $e_i^2 = \gamma_1 + \gamma_2 z_{2i} + ... + \gamma_p zpi + \eta_i$. 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)
Rsq<-summary(res)$r.squared

# test statistic
BP <- nres*Rsq
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 β_0, β_1 and β_2 separately using\

1. OLS\
2. WLS with known $\gamma = 0$\
3. FWLS with estimated γ (i.e. γ 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?\

Firstly, we generate the new epsilons again, while leaving the variables generated x_1 and x_2 untouched. In the question, it is not specified whether we show generate new x_1 and x_2's, in addition to the ϵs. Therefore, we leave them untouched, while generating all the variables that are dependent on σ (directly and indirectly) again.

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

```
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 γ. 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$residuals
modelEps <- lm(log(eps^2) ~ z)
gammaFWLS <- modelEps$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) 2.938167 0.03507559  83.76672        0
## x1          5.008609 0.01400149 357.71967        0
## x2          8.018884 0.01428016 561.54009        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))
```

```
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$residuals
modelEpsD <- lm(log(epsD^2) ~ z)
gammaFWLSD <- modelEpsD$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) 3.007161 0.05167999  58.18811        0
## x1          5.019272 0.02086402 240.57078        0
## x2          7.987736 0.02108072 378.91200        0
```

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))
```

```
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
```

```r
# --- FWLS
epsE <- modelE$residuals
modelEpsE <- lm(log(epsE^2) ~ z)
gammaFWLSE <- modelEpsE$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) 2.997396 0.012593523  238.0109        0
## x1          4.999755 0.005184879  964.2955        0
## x2          8.004795 0.005464376 1464.9055        0
```

The estimators are still close to their true values. However, the difference is slightly higher than in the previous estimations.

Question 2

1. Show that the OLS estimator of the parameter β is not consistent.

We have $C = D\beta + \epsilon$ and $D = C + Z$. For the OLS estimator we get:

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

$$b = \beta + (D^T D)^{-1} D^T \epsilon$$
$$b = \beta + (\frac{1}{n} D^T D)^{-1} \frac{1}{n} D^T \epsilon$$
$$\text{plim}(b) = \beta + \text{plim}((\frac{1}{n} D^T D)^{-1}) \text{plim}(\frac{1}{n} D^T \epsilon)$$

We assume that $\text{plim}(\frac{1}{n} D^T D) = Q_{DD}$ and that Q_{DD} is of full rank (so we can take the inverse). Hence:

$$\text{plim}(b) = \beta + Q_{DD}^{-1} \text{plim}(\frac{1}{n} D^T \epsilon)$$
$$\text{plim}(b) = \beta + Q_{DD}^{-1} \text{plim}(\frac{1}{n} (C + Z)^T \epsilon)$$
$$\text{plim}(b) = \beta + Q_{DD}^{-1} (\text{plim}(\frac{1}{n} C^T \epsilon) + \text{plim}(\frac{1}{n} Z^T \epsilon))$$

For $n \to \infty$, $\text{plim}(\frac{1}{n} Z^T \epsilon) = 0$ as $\mathbb{E}[Z_i \epsilon_i] = 0$. $\text{plim}(\frac{1}{n} C^T \epsilon$, on the other side, is unequal zero as ϵ is inside the DGP ($C = \beta D + \epsilon$) and therefore the covariance is unequal to zero.

2. Derive plim (b) where b is the OLS estimator of β. Determine the sign of the magnitude of the inconsistency when $0 < \beta < 1$, that is, the sign of plim(b) - β 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 α 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:

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

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 Cov(D, ϵ_i), which is $\frac{1}{1-\beta}\text{Var}(\epsilon) = \frac{1}{1-\beta}\sigma^2$:

$$\begin{aligned}\text{plim}(\frac{1}{n} C\prime\epsilon) = \frac{1}{n}\mathbb{E}\left[C\prime\epsilon\right] = \frac{1}{n}\mathbb{E}\sum[(C_i\prime\epsilon_i)\epsilon_i] = \frac{1}{n}\mathbb{E}\left[\sum(\frac{\alpha}{1-\beta} + \frac{\beta}{1-\beta}Z_i + \frac{1}{1-\beta}\epsilon_i)\right] \\ \Rightarrow= \frac{1}{n}\mathbb{E}\left[\frac{1}{1-\beta}\sum\epsilon_i\right] \\ \Rightarrow= \frac{1}{1-\beta}\sigma^2\end{aligned}$$

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

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

4. Derive b_{IV}, the IV estimator of β 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, and x_i is a row out of this data matrix, 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 β (what we can do as $m = k$), we retrieve the b_{IV} estimator:

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

The second element of this vector is the coefficient for β in the consumption equation.

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:

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

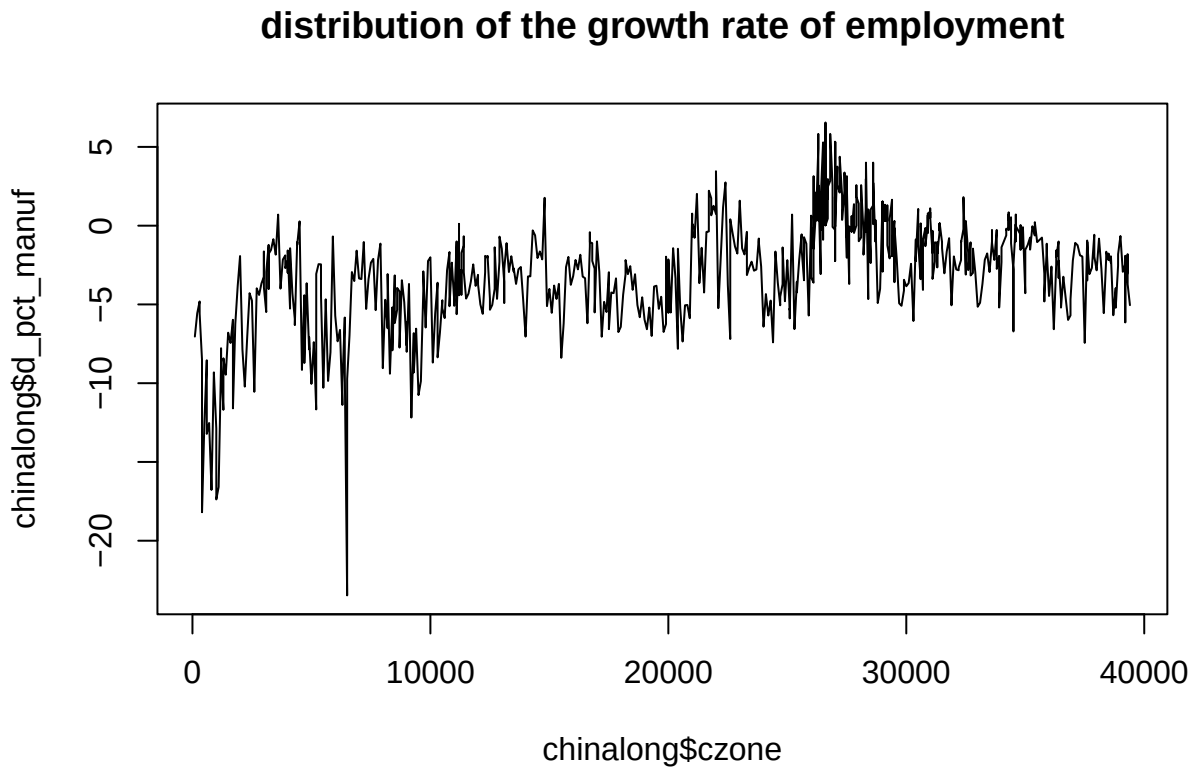
where the last factor goes to zero as $n \to \infty$.

Question 3

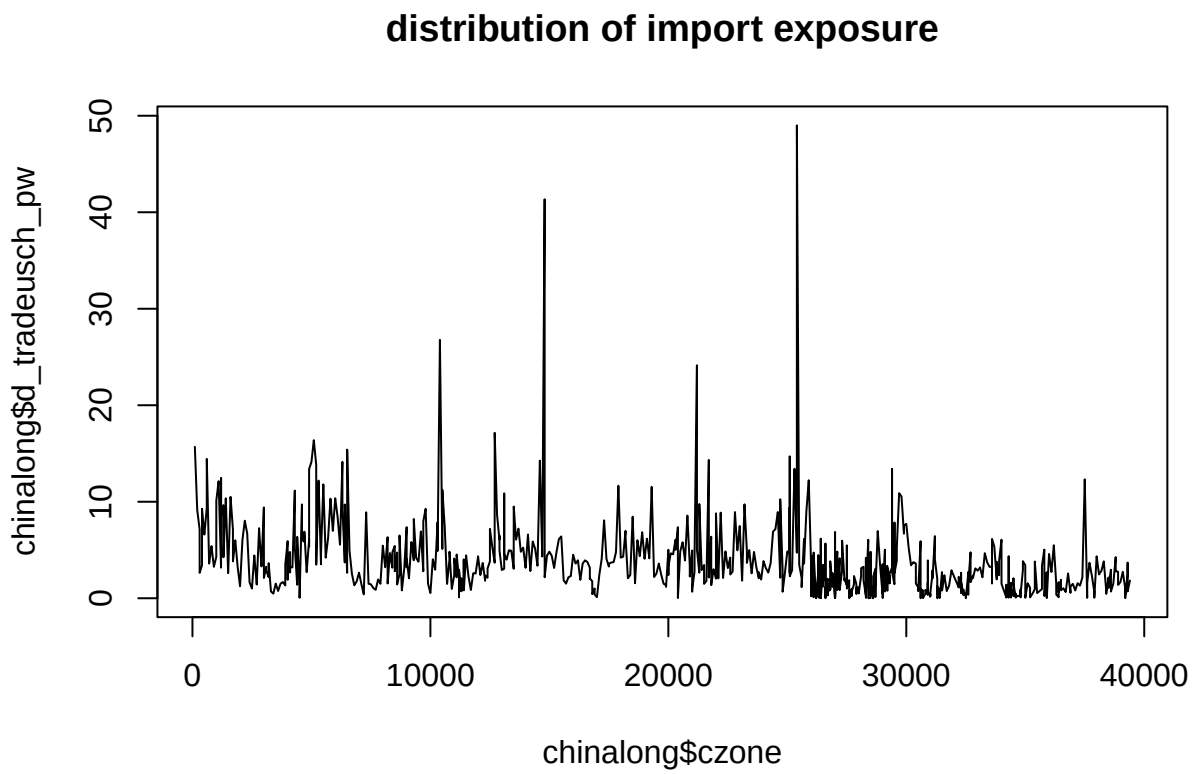
```
china <- read_dta("workfile_china.dta")
chinalong <- read_dta("workfile_china_long.dta")
chinapreperiod <- read_dta("workfile_china_preperiod.dta")
```

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")
```




```
plot(chinalong$czone, chinalong$d_tradeusch_pw, type="l",
     main = "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.

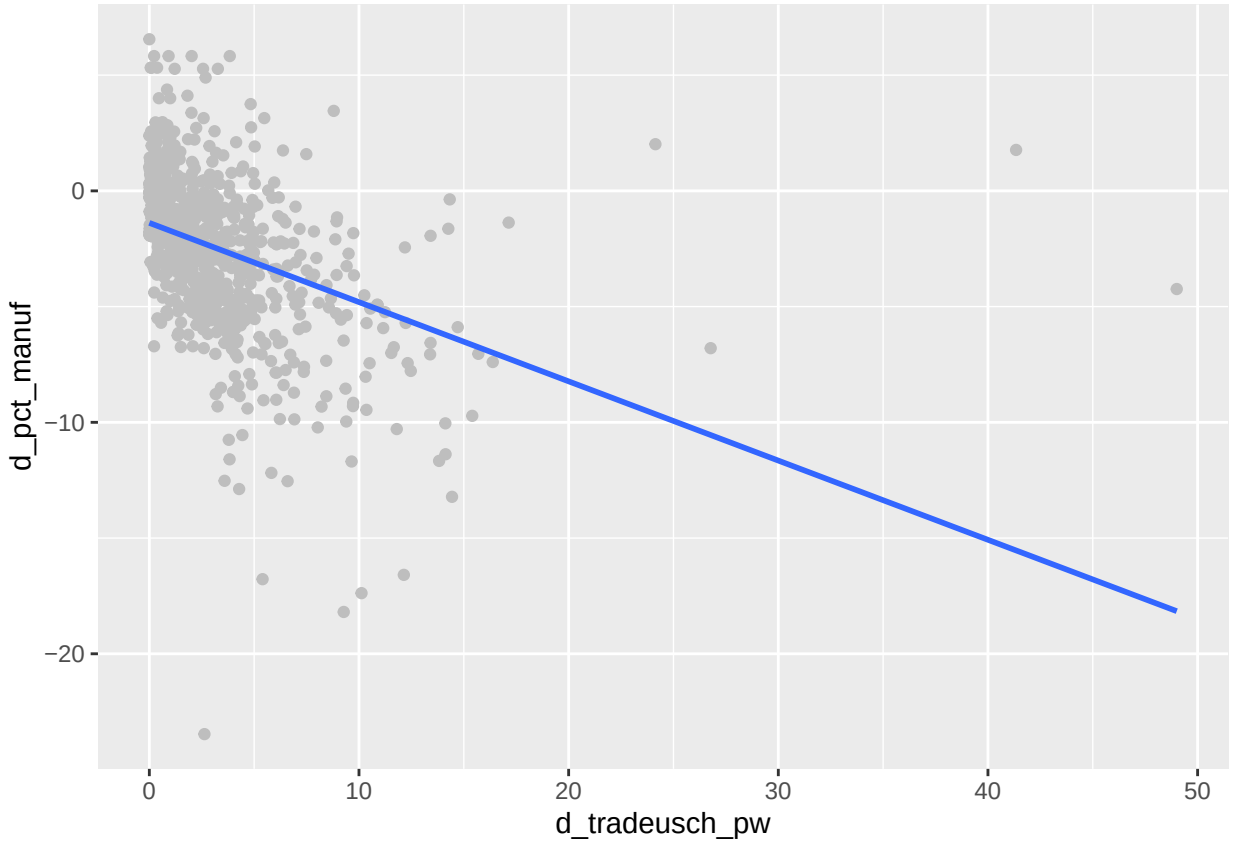
```
model <- lm(d_pct_manuf  ~ d_tradeusch_pw, data=chinalong)

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'
```




```
# 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 heteroskedastitiy- 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
modelIV <- ivreg(d_pct_manuf  ~ d_tradeusch_pw| d_tradeotch_pw_lag, data=chinalong)
summary(modelIV)
```

```
##
## 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(modelIV)
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 <- coeftest(ivreg(d_sh_empl_mfg ~ d_tradeusch_pw| d_tradeotch_pw_lag,
                         data=subset(chinapreperiod, yr==1990),
                         weights=timepwt48), vcov = vcovCL, cluster = ~statefip)
model2 <- coeftest(ivreg(d_sh_empl_mfg ~ d_tradeusch_pw| d_tradeotch_pw_lag,
                         data=subset(chinapreperiod, yr==2000),
                         weights=timepwt48), vcov = vcovCL,
                   cluster = ~statefip)
model3 <- coeftest(ivreg(d_sh_empl_mfg ~ d_tradeusch_pw + t2000| d_tradeotch_pw_lag+ t2000,
                         data=subset(chinapreperiod, yr==1990|yr==2000),
                         weights=timepwt48), vcov = vcovCL, cluster = ~statefip)
model4 <- coeftest(ivreg(d_sh_empl_mfg ~ d_tradeusch_pw_future| d_tradeotch_pw_lag_future,
                         data=subset(chinapreperiod, yr==1970),
                         weights=timepwt48), vcov = vcovCL,
                   cluster = ~statefip)
model5 <- coeftest(ivreg(d_sh_empl_mfg  ~ d_tradeusch_pw_future| d_tradeotch_pw_lag_future,
                         data=subset(chinapreperiod, yr==1980),
                         weights=timepwt48), vcov = vcovCL, cluster = ~statefip)
model6 <- coeftest(ivreg(d_sh_empl_mfg  ~ d_tradeusch_pw_future + t1980|
                           d_tradeotch_pw_lag_future + t1980,
                         data=subset(chinapreperiod, yr==1970|yr==1980),
                         weights=timepwt48),
                   vcov = vcovCL, cluster = ~statefip)
```

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

```
model1Stage1 <- coeftest(lm(d_tradeusch_pw ~ d_tradeotch_pw_lag,
                            data=subset(chinapreperiod, yr==1990),
                            weights=timepwt48), vcov = vcovCL, cluster = ~statefip)
model2Stage1 <- coeftest(lm(d_tradeusch_pw ~ d_tradeotch_pw_lag,
                            data=subset(chinapreperiod, yr==2000),
                            weights=timepwt48), vcov = vcovCL, cluster = ~statefip)
model3Stage1 <- coeftest(lm(d_tradeusch_pw + t2000 ~ d_tradeotch_pw_lag+ t2000,
                            data=subset(chinapreperiod, yr==1990|yr==2000),
                            weights=timepwt48), vcov = vcovCL, cluster = ~statefip)
model4Stage1 <- coeftest(lm(d_tradeusch_pw_future ~ d_tradeotch_pw_lag_future,
                            data=subset(chinapreperiod, yr==1970),
                            weights=timepwt48), vcov = vcovCL, cluster = ~statefip)
model5Stage1 <- coeftest(lm(d_tradeusch_pw_future ~ d_tradeotch_pw_lag_future,
                            data=subset(chinapreperiod, yr==1980),
                            weights=timepwt48), vcov = vcovCL, cluster = ~statefip)
model6Stage1 <- coeftest(lm(d_tradeusch_pw_future + t1980 ~ d_tradeotch_pw_lag_future + t1980,
                            data=subset(chinapreperiod, yr==1970|yr==1980),
```

Table 1:

	Dependent variable:					
	(1)	(2)	(3)	(4)	(5)	(6)
d_tradeusch_pw	−0.888***	−0.718***	−0.746***			
	(0.183)	(0.065)	(0.069)			
t2000			0.444			
			(0.327)			
d_tradeusch_pw_future				0.431***	−0.130	0.148
				(0.151)	(0.127)	(0.096)
t1980						−1.945***
						(0.250)
Constant	−1.056***	−0.846***	−1.218***	−0.954***	−1.832***	−0.415
	(0.195)	(0.258)	(0.140)	(0.319)	(0.334)	(0.302)
Note:						*p<0.1; **p<0.05; ***p<0.01

```
                          weights=timepwt48), vcov = vcovCL, cluster = ~statefip)
```

```
stargazer(model1Stage1, model2Stage1, model3Stage1, model4Stage1,
          model5Stage1, model6Stage1,
          header = FALSE,
          column.sep.width = "0pt",
          font.size = "tiny",
          df = FALSE)
```

Table 2:

	Dependent variable:					
	(1)	(2)	(3)	(4)	(5)	(6)
d_tradeotch_pw_lag	0.950***	0.767***	0.792***			
	(0.112)	(0.088)	(0.080)			
t2000			1.237***			
			(0.139)			
d_tradeotch_pw_lag_future				0.929***	0.926***	0.927***
				(0.079)	(0.079)	(0.079)
t1980						0.996***
						(0.008)
Constant	0.187*	0.647***	0.346***	0.218	0.219	0.220
	(0.110)	(0.220)	(0.079)	(0.135)	(0.134)	(0.134)
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.

Note: As the authors did not include a code book, it is not possible to replicate table 3 unless we use their codes. However, this does not fulfill the purpose of replicating the findings as we would just copy their code. Authors should ensure that it is possible to replicate their findings by including all necessary information!

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 β 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 ϵ_i is now a shifted Bernouilli 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 σ_ϵ^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.

We have $y_i = \alpha + \beta x_i + \epsilon_i = x_i\prime\beta + \epsilon_i$. From here on, we continue with matrix notation. In this case, more generally, y_i is distributed as:

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

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

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

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

The log-likelihood for n observations is:

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

Taking the first derivative with respect to the parameters β gives:

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

This can be rewritten as:

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

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, $F(.)(1-F(.)) = f(.)$. Then, our expression simlifies to:

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

Then, substituting the logit cdf for F gives:

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

which is what we were required to show.

5. Finally, use the value of $F(x_i\prime\beta)$ 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 β.

The log odds ratio is defined as:

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

The log-odds ratio is then:

$$\log OR = x_i\prime\beta$$

Beta is then equal to the derivative of the log odds ratio with respect to a regressor. This means that the the strength of β 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.