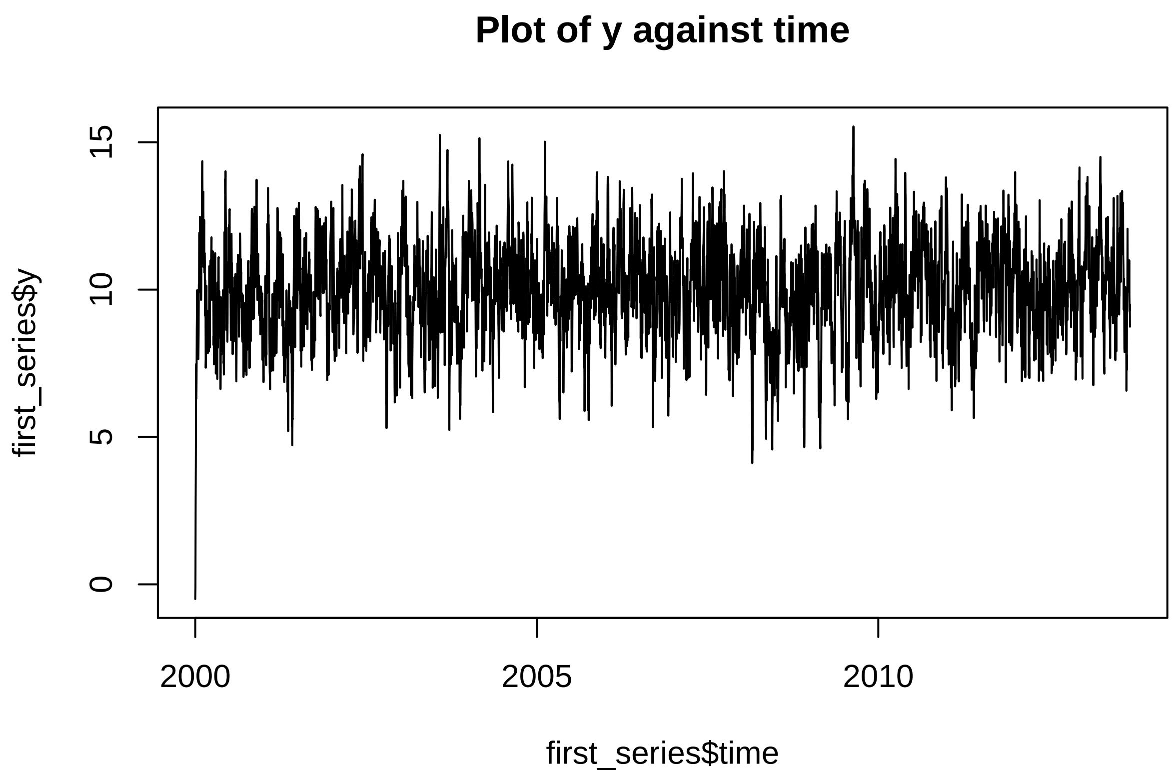
HE3021 Tutorial 6 Submission

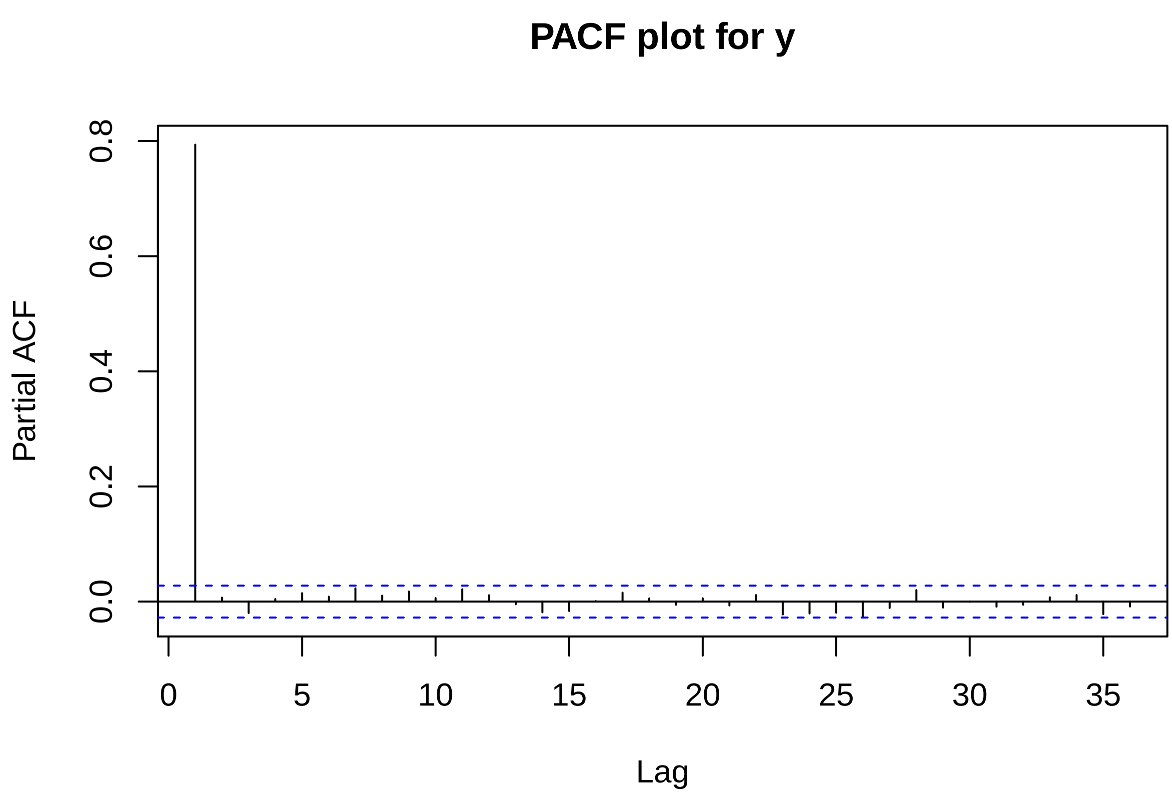
Lui Yu Sen

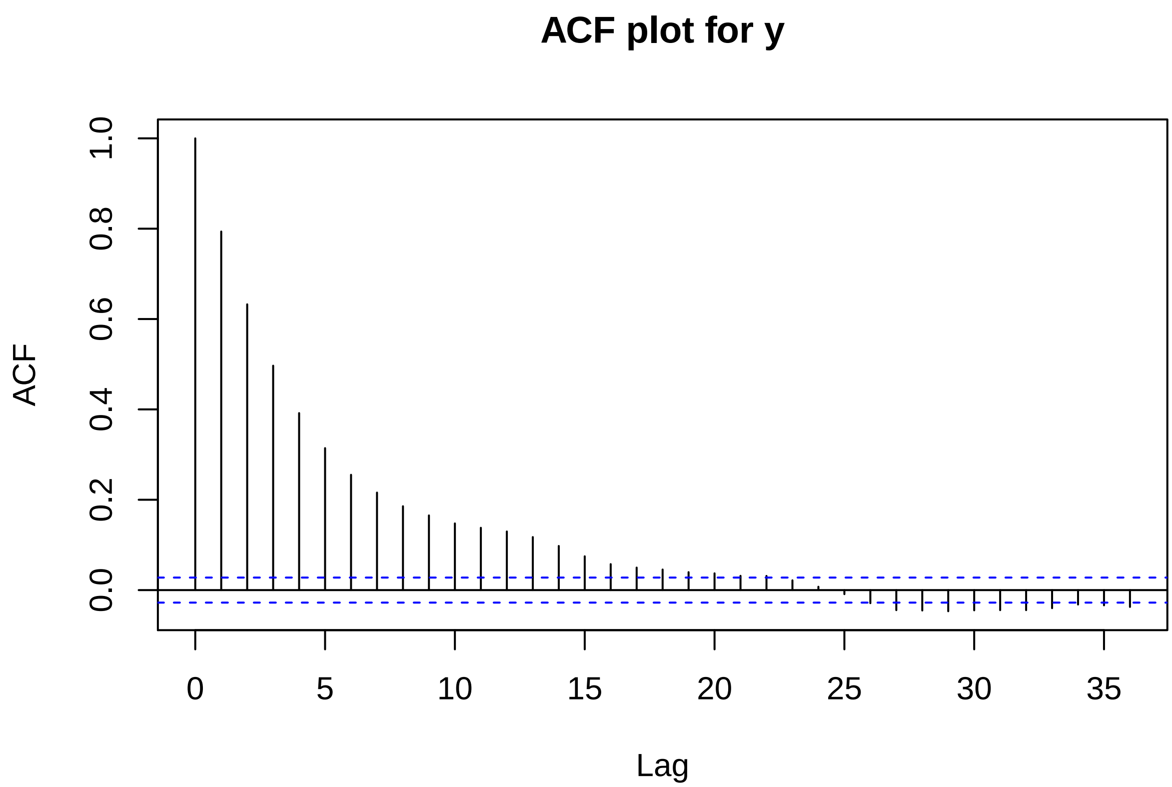
3/16/2021

Q1

a and b


Plot of y against time
first_series$y
0
5
10
15
2000
2005
2010
first_series$time


PACF plot for y
Partial ACF
0.0
0.2
0.4
0.6
0.8
0
5
10
15
20
25
30
35
Lag




```r
adftest <- adf.test(ts(first_series$y))
```

```
## Warning in adf.test(ts(first_series$y)): p-value smaller than printed p-value
```

```r
pp <- pp.test(first_series$y)
```

```
## Warning in pp.test(first_series$y): p-value smaller than printed p-value
```

The ACF plot has an exponentially decaying ACF. ADF test returned 0.01 and PP test returned 0.01, thus we have sufficient evidence to reject the null hypothesis that there exists a unit root at 0.05 significance level. The time series is likely stationary. The ACF is decaying while the PACF has a significant spike at t-1. Thus, it is likely that the appropriate model is AR(1).

c

```r
Arima(first_series$y, order = c(1,0,0), include.mean = TRUE)
```

```
## Series: first_series$y
## ARIMA(1,0,0) with non-zero mean
##
## Coefficients:
```

```
##          ar1    mean
##       0.8003  9.9665
## s.e.  0.0086  0.0696
##
## sigma^2 estimated as 0.9694:  log likelihood=-7016.58
## AIC=14039.17   AICc=14039.17   BIC=14058.72
```

The PACF plot had t-1 as the significant spike, thus I started with and ARIMA(1,0,0) model. The AIC value generated was 1.4039165×10^4.

```
Arima(first_series$y, order = c(2,0,0), include.mean = TRUE)
```

```
## Series: first_series$y
## ARIMA(2,0,0) with non-zero mean
##
## Coefficients:
##          ar1     ar2    mean
##       0.7922  0.0102  9.9666
## s.e.  0.0141  0.0142  0.0704
##
## sigma^2 estimated as 0.9695:  log likelihood=-7016.33
## AIC=14040.65   AICc=14040.66   BIC=14066.72
```

AR(2) generated a larger AIC of 1.4040654×10^4, so AR(1) was the closer fitting model. AR(3) generate even larger AIC of 1.4040342×10^4, thus I chose AR(1).

d

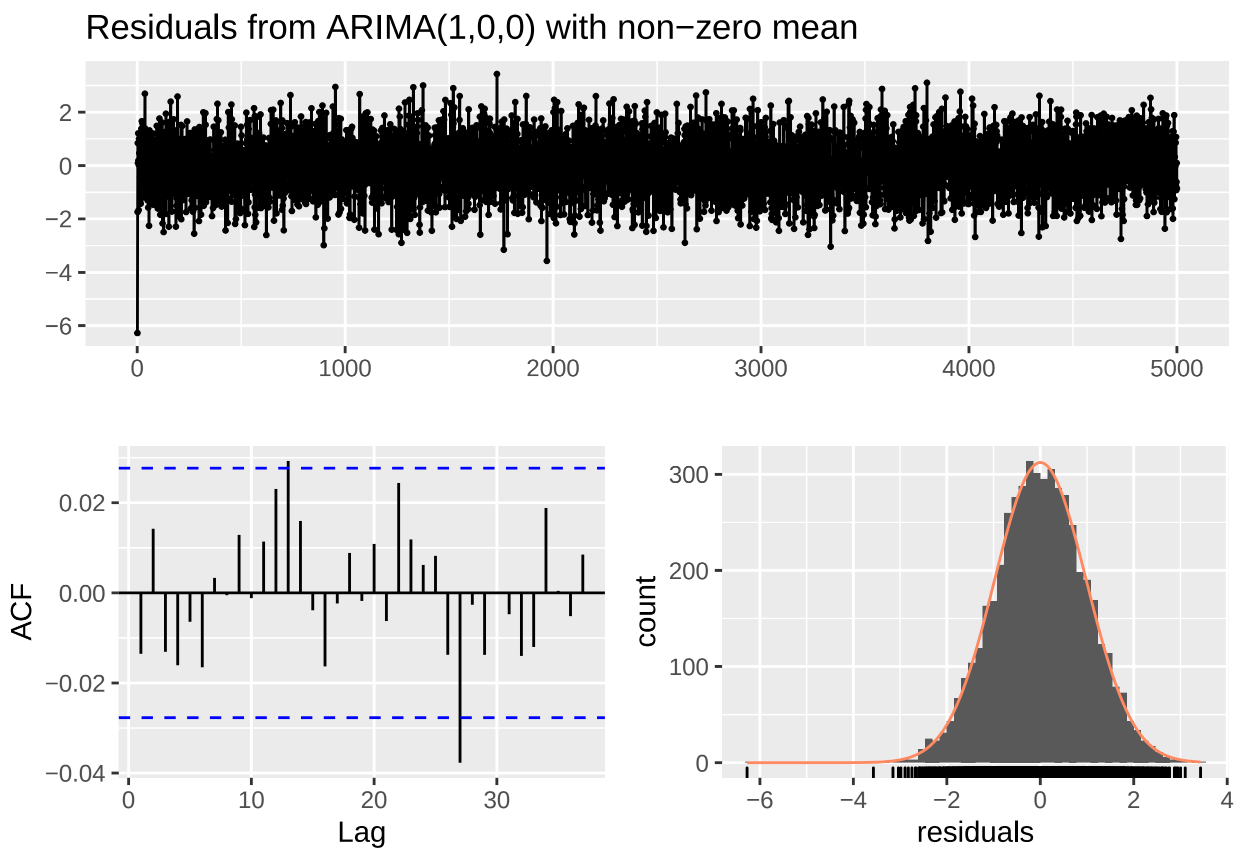
To check for autocorrelation in the residuals, I fitted an AR(2) model for the residuals and did an F-test.

```
residuals <- Arima(first_series$y, order = c(1,0,0), include.mean = TRUE)$residuals
residuals_frame <- data.frame(cbind(residuals[3:5000], residuals[2:4999], residuals[1:4998]))
colnames(residuals_frame) <- c("t","t1","t2")
residuals_lm <- lm(t ~ t1 + t2 - 1, data = residuals_frame) # p-value 0.3232, cannot reject null hyp th
summary(residuals_lm)
```

```
##
## Call:
## lm(formula = t ~ t1 + t2 - 1, data = residuals_frame)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5354 -0.6501  0.0005  0.6680  3.4194
##
## Coefficients:
##    Estimate Std. Error t value Pr(>|t|)
## t1 -0.01569    0.01414  -1.109    0.267
## t2  0.01407    0.01408   0.999    0.318
##
## Residual standard error: 0.9803 on 4996 degrees of freedom
## Multiple R-squared:  0.0004521,   Adjusted R-squared:  5.194e-05
## F-statistic:  1.13 on 2 and 4996 DF,  p-value: 0.3232
```

The p-value is 0.3232, thus we cannot reject the null hypothesis that the partial effects of lags u_{t-1} and u_{t-2} are equal to 0. Thus there is no serial correlation of lag 2. Thus, the residuals resemble white noise. Using the forecast package, a Ljung-Box test was also conducted directly on the ARIMA model object.

```
checkresiduals(Arima(first_series$y, order = c(1,0,0), include.mean = TRUE))
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(1,0,0) with non-zero mean
## Q* = 6.5518, df = 8, p-value = 0.5857
##
## Model df: 2.   Total lags used: 10
```

The high p-value mean that we cannot reject the null hypothesis of no serial correlation under significance level of 0.05. I also conducted a Breusch-Godfrey test using the lmtest package by first manually regressing an AR(1) model.

```
y_lagged <- as.data.frame(cbind(first_series$y[2:5000], first_series$y[1:4999]))
colnames(y_lagged) <- c("y_t", "y_tlag1")
modelar1 <- lm(y_t ~ y_tlag1, data = y_lagged)
bgtest_results <- c()
for (n in 1:24){
    bgtest_results <- c(bgtest_results, bgtest(modelar1, data = y_lagged, order = n)$p.value)
}
range(bgtest_results)
```

```
## [1] 0.2532732 0.5977828
```

Testing lags from 1 to 24 for any seasonal lags returned a range of 24 values all above 0.05 significance level. Thus we cannot reject null hypothesis of no serial correlation. The residuals appear to be white noise.

2

a

$$
\begin{aligned}
&E(y_i|x_{i1}, x_{i2}) = E(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i|x_{i1}, x_{i2}) && (1)\\
&= E(\beta_0|x_{i1}, x_{i2}) + E(\beta_1 x_{i1}|x_{i1}, x_{i2}) + E(\beta_2 x_{i2}|x_{i1}, x_{i2}) + E(u_i|x_{i1}, x_{i2}) && (2)\\
&= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + 0 = P(y_i = 1|x_{i1}, x_{i2}) && (3)
\end{aligned}
$$

b

$$
\begin{aligned}
&Var(y_i|x_{i1}, x_{i2}) = Var(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i|x_{i1}, x_{i2}) && (4)\\
&= 0 + 0 + Var(u_i|x_{i1}, x_{i2}) && (5)\\
&= P(y_i = 1|x_{i1}, x_{i2})(1 - P(y_i = 1|x_{i1}, x_{i2})) && (6)\\
&\implies Var(u_i|x_{i1}, x_{i2}) = f(x_{i1}, x_{i2}) = (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2})[1 - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2})] && (7)
\end{aligned}
$$

Thus a linear probability model cannot be homoskedastic, because it is a Bernoulli process and the error variance is a function of the independent variables.

c

An increase of \$1000 in income leads to increase of 0.08 chance of buying a car, holding education constant. An increase of 1 year of education leads to increase of 0.01 chance of buying a car holding income constant.

d

$$\hat{y}_i = -0.1 + 0.08 \cdot 8 + 0.01 \cdot 16 = 0.7$$

There is a 0.7 chance that this person has a car.

e

$$\hat{y}_i = -0.1 + 0.08 \cdot 15 + 0.01 \cdot 16 = 1.26 > 1$$

There is a 1.26 chance that this person has a car. This result does not make sense, since $P(y = 1|x_{i1}, x_{i2}) \in [0, 1]$. To solve it, let

$$\hat{y}_i = \begin{cases} 0,\ if\ \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} < 0 \\ 1,\ if\ \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} > 1 \\ \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2}, \hat{y}_i \in [0, 1] \end{cases}$$

Then, $\hat{y}_i = -0.1 + 0.08 \cdot 15 + 0.01 \cdot 16 = 1$, there is probability of 1 that this person has a car.

3

a

Let

$$CDF = \Phi(z), PDF = \phi(z), z = \beta_0 + \beta_1 x_1 + beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 \tag{8}$$
$$\frac{\partial P(y=1|X)}{\partial x_1} = \frac{\partial}{\partial x_1} \int_{-\infty}^{z} \phi(z) dx_1 = (2\pi)^{-\frac{1}{2}} \cdot e^{-\frac{z^2}{2}} \cdot \beta_1 \tag{9}$$

b

```
model <- glm(inlf ~ nwifeinc + educ + kidslt6 + age + exper,
             family = binomial(link = "probit"),
             data = mroz)
summary(model)
```

```
##
## Call:
## glm(formula = inlf ~ nwifeinc + educ + kidslt6 + age + exper,
##     family = binomial(link = "probit"), data = mroz)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5942  -0.9371   0.4342   0.8934   2.3229
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.764541   0.439490   1.740   0.0819 .
## nwifeinc    -0.011371   0.004857  -2.341   0.0192 *
## educ         0.131532   0.025082   5.244 1.57e-07 ***
## kidslt6     -0.886208   0.116696  -7.594 3.10e-14 ***
## age         -0.057919   0.007790  -7.435 1.04e-13 ***
## exper        0.069148   0.007556   9.151  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1029.75  on 752  degrees of freedom
## Residual deviance:  813.08  on 747  degrees of freedom
## AIC: 825.08
##
## Number of Fisher Scoring iterations: 4
```

Using the TI-84's normal CDF function:

$$\Phi(0.76454 + (-0.11371)20 + 0.131532 \cdot 10 + (-0.057919)30 + 0.069148 \cdot 10) = \Phi(0.806351) = 0.78998$$

c

$$\frac{\partial P(y=1|X)}{\partial x_{exper}} = \frac{\partial}{\partial x_{exper}} \int_{-\infty}^{z} \phi(z) dx_{exper} = (2\pi)^{-\frac{1}{2}} \cdot e^{-\frac{0.806351^2}{2}} \cdot 0.069148 = 0.019930$$