ECON 221 – PSET 1

Group: Aman Puri, James Wang, and Catherine Darin
09/17/2015

A.

Given that we know that the dependent variable is current earnings for the healthcare industry, we can make some inferences about what lagged variable could be a reasonable predictor of earnings. Looking at the lagged-x values, we see that the mean is .1758, the minimum is -1.686, and the maximum is 2.68. The fact that this variable can take on negative values is an important consideration. One possible x-variable could be the quarterly percentage change in the value of a portfolio of the 10 largest healthcare stocks - the reasoning here would be that if the biggest companies in the healthcare industry are doing well, then possibly these profits are trickling down and benefiting the industry as a whole. Another possible x-variable could be the percentage change in GDP growth from the previous quarter, as the overall economy is closely related to the healthcare industry. A reduction or increase in the GDP growth rate in one quarter is likely to have knock-on effects on the broader macro-economy in the subsequent quarter. Another idea is that the x-variable could be the quarterly percentage change in National Healthcare Expenditure, as spending on healthcare could reasonably be a predictor of increased usage of healthcare facilities and investment, leading to subsequent revenues.

The fact that the x value is lagged implies that there is a delayed relationship between x and healthcare earnings. That is, a change in x may not immediately relate to a change in y - rather, it make take several months for the change in healthcare earnings to be realized. Additionally, for forecasting purposes, it can be useful to include a lagged variable, because if we are trying to predict an outcome for the current quarter, we can utilize data from the previous quarter. This way, we can also conveniently test the accuracy of our model's predictions over time (comparing our predictions to what actually happened) to improve our model.

B.

Looking at the Y_t vs. X_{t-1} graph, we can see a positive, generally linear association between healthcare earnings and the unknown lagged x variable. The correlation between Y_t and X_{t-1} is .6676, which also suggests a general positive linear relationship between the two variables. However, as there is not a precise relationship and some data points diverge from this linear trend, we should not assume that the DGP is strictly linear.

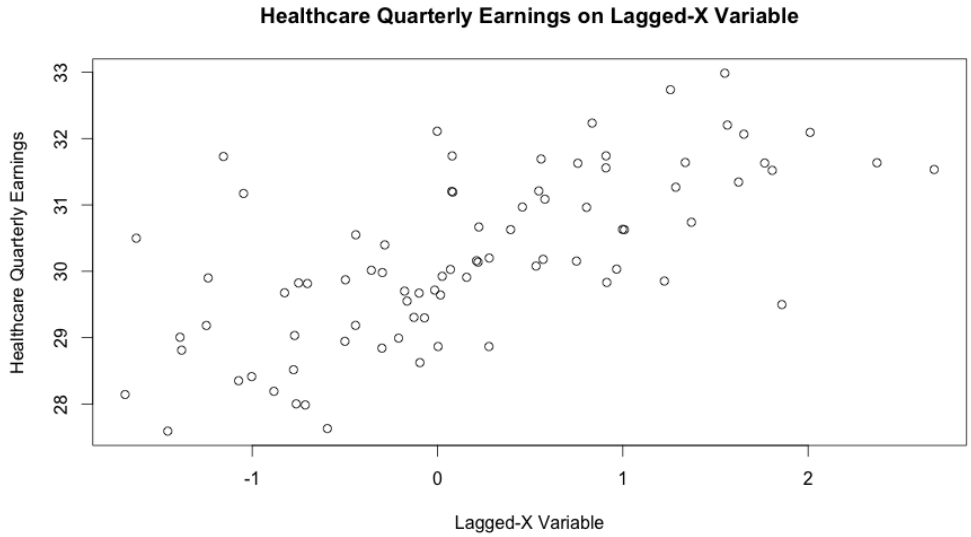
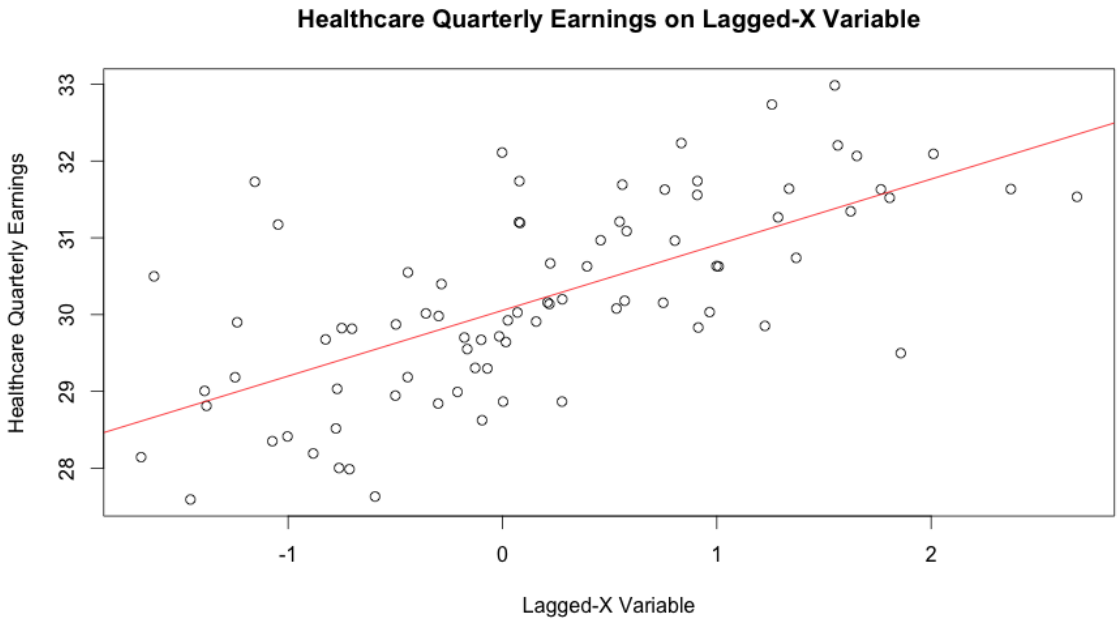


Table 1: lm (y ~ XLAG1) (FIT1)					RSE	.9456
Regressor	**Coefficient estimate**	**Std. error**	**T-value**	**P-value**	**R^2**	0.4388
(Intercept)	30.0542	.1061	283.28	< 2e-16 ***	**Adjusted R^2**	0.4457
XLAG1	.8552	.1066	8.021	7.37e-12 ***	**F-statistic**	64.33
					F-stat p-value	< 7.37e-12
					BIC	234.7273

C.

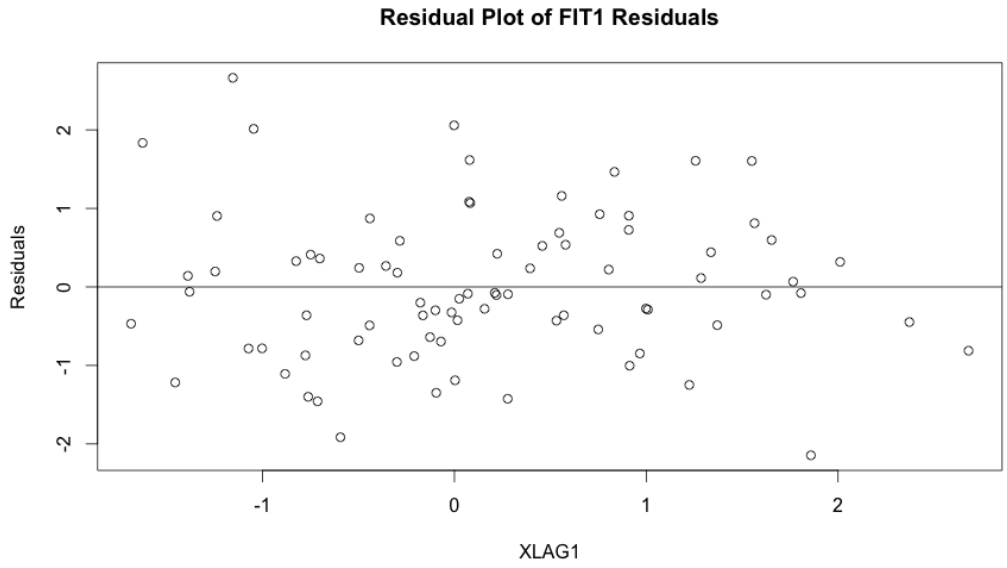
Above, we can see the linear regression of Y_t on X_{t-1}. Based on the outputs, the regression equation can be deduced: $Y_t = 0.8552X_{t-1} + 30.0542$. The slope of 0.8552 implies that when X_{t-1} increases by 1 unit, Y_t (healthcare earnings) increases by 0.8552. The intercept of 30.0542 suggests that when the lagged variable $X_{t-1} = 0$, Y_t (healthcare earnings) assumes a default value of 30.0542. This interpretation of the intercept may or may not be meaningful, depending on the true nature of X_{t-1} and Y_t, as well as the trend in the variables over time.

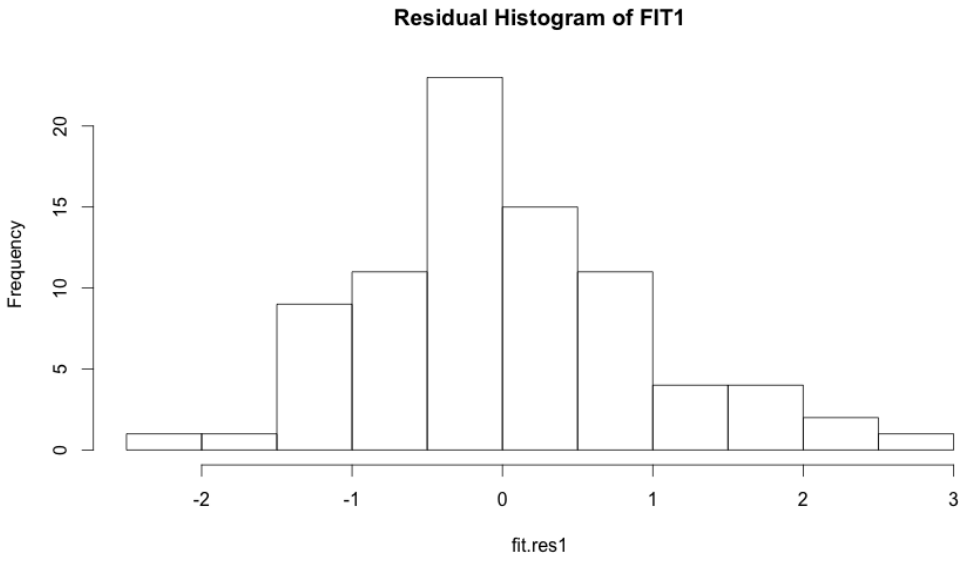
Looking at the standard errors of our regression, they seem to be relatively small. Both of our estimates for XLAG1 and the intercept are statistically significant. For the intercept, the standard error of .1061 is particularly small compared to the estimate of the intercept being 30.054. For the



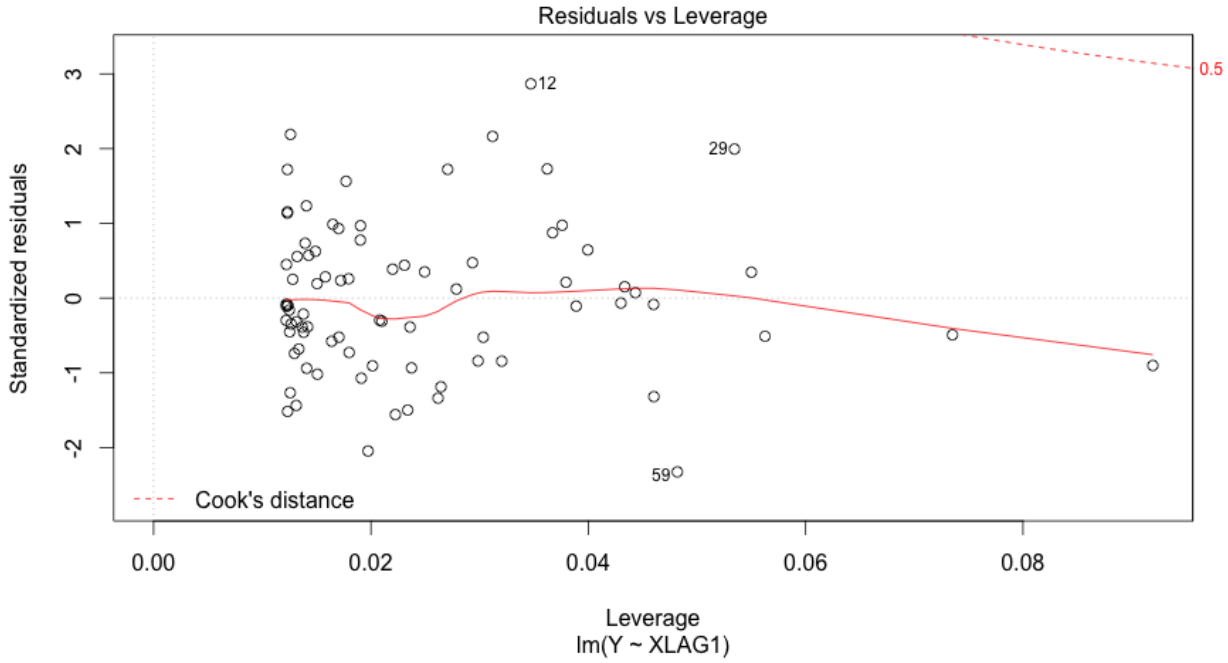
coefficient on XLAG1, the SE is about the same size (but larger when thought about in comparison to the estimate), and a 95% confidence interval for our coefficient would be about .8552 +/- .21, or (.646, 1.06).

We also see that the R^2 is about .4388, which is relatively high, especially when we are using only one regressor. This, as well as the statistically significant F-Stat, leads us to believe that x is a relatively good predictor of healthcare earnings. Additionally, we calculated a BIC statistic as a comparison point for other models.





To verify that the assumptions of homoscedasticity and a normal distribution of the error terms are being upheld, the residuals of the regression can be analyzed. The normal quantile plot and histogram of residuals demonstrate that the residuals do follow an approximate normal yet slightly skewed distribution, though this slight skewness does not give us much cause for concern. However, the plot of the residuals against X_{t-1} reveals that the error terms may not be completely random noise. Based on the residual plot, there doesn't seem to be any apparent trend or serial correlation in the residuals, but it appears as though the errors may be slightly heteroskedastic. The residuals appear more dispersed for lower, negative values of X_{t-1} than for higher, positive values of the lagged variable, thus producing a very slight megaphone shape in the plot. However, this is not a strong pattern and we therefore need to later test for heteroskedasticity to verify if it is indeed making the standard errors of the OLS estimators inefficient. This irregularity in the pattern of the error terms may also suggest that the appropriate model is nonlinear, which we will also later test for.



In order to test for the potential presence of outliers, which may be biasing the OLS estimates of the coefficient and intercept, we can analyze the leverage plot for the linear regression. The plot above shows that there are no clear outliers present. While the data in rows 12, 29 and 59 are fairly distant from the fitted line, it is unlikely that they are having any strong impact on the OLS estimates and it would not be reasonable to exclude these values from the dataset.

D.

	FIT1	FIT2	FIT3	FIT4	FIT5	FIT6	FIT7
BIC	234.7273	239.0946	907.4008	-323.8864	239.1339	-319.4803	-308.3462

For this question, we looked at seven different models to consider non-linear and other relationships. These models were:

FIT1: $Y_t \sim X_{t-1}$
FIT2: $Y_t \sim X_{t-1} + (X_{t-1})^2$
FIT3: $(Y_t)^2 \sim X_{t-1}$
FIT4: $\log(Y_t) \sim X_{t-1}$
FIT5: $Y_t \sim X_{t-1} + t$
FIT6: $\log(Y_t) \sim X_{t-1} + t$
FIT7: $\log(Y_t) \sim X_{t-1} + t + q1 + q2 + q3 + q4$ (quarterly dummies)

We picked the 3 best-performing models to discuss, which were FIT4, FIT6, and FIT7 (see output summaries below), based on the BIC statistic. We will also occasionally show our FIT1 model from above as a comparison.

Table 2: lm (log(Y) ~ XLAG1) (FIT4)					**RSE**	.03136
Regressor	**Coefficient estimate**	**Std. error**	**T-value**	**P-value**	**R^2**	0.4456
(Intercept)	3.4021	.003519	966.838	< 2e-16 ***	**Adjusted R^2**	0.4387
XLAG1	.028358	.003356	8.019	7.47e-12 ***	**F-statistic**	64.3

Table 3: lm ($\log(Y_t) \sim X_{t-1} + t$) (FIT6)					**RSE**	0.0316
Regressor	**Coefficient estimate**	**Std. error**	**T-value**	**P-value**	**R^2**	0.4456
XLAG1	2.835e-02	1.477e-04	7.942	1.13e-11 ***	**F-statistic**	31.75
T	3.521e-06	3.570e-03	.024	.981	**F-stat p-value**	< 7.597e-11
Intercept	3.402e-00	7.036e-03	483.49	< 2e-16 ***	**BIC**	-319.4803

Table 4: lm (log(Y_t) ~ X_{t-1} + t + q1 + q2 + q3 + q4) (FIT7)					**BIC**	-308.346
Regressor	**Coefficient estimate**	**Std. error**	**T-value**	**P-value**		
(Intercept)	3.505e+00	9.492 e-03	358.7	< 2e-16 ***		
XLAG1	2.77e-02	3.791e-03	7.320	2.17e-10***		
T	6.036e-06	1.488e-04	0.041	.968		
Q1	-8.256-e03	9.929-03	-0.832	.408		
Q2	-7.201e-03	1.013e-02	-0.711	.479		
Q3	3.613e-03	1.035e-02	.349	.728	***Can't have intercept + full seasonal dummies**	
Q4	NA	NA	NA	NA		

i. It does appear necessary to include an intercept in the linear regression (FIT1), as we found that the coefficient on the intercept was statistically significant (p = 2e-16). Even just looking at our plot, we can see that the healthcare earnings range from around 28 to 33, and there is no logical reason to assume that we should put an artificial 0,0 point into our regression. If we would remove the intercept, this would clear bias our least squares estimators of the slope and coefficient. This reasoning applies to other models we tested as well.

ii. Our first indication that the functional form was linear was simply observing a plot of the data. We then tested a polynomial regression, FIT2: Y ~ XLAG1 + XLAG^2 (below), to check if there was a non-linear relationship that we could not see. However, we found that the coefficient for the second-order polynomial was insignificant, and thus, this reaffirmed our belief that the functional form was linear (see output to right)

```
Call:
lm(formula = Y ~ XLAG1 + I(XLAG1^2))

Residuals:
    Min      1Q  Median      3Q     Max
-2.1197 -0.6292 -0.0975  0.5170  2.6841

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 30.06982    0.13358 225.115  < 2e-16 ***
XLAG1        0.86625    0.12140   7.135 4.11e-10 ***
I(XLAG1^2)  -0.01779    0.09121  -0.195    0.846
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9513 on 79 degrees of freedom
Multiple R-squared:  0.446,     Adjusted R-squared:  0.432
F-statistic:  31.8 on 2 and 79 DF,  p-value: 7.402e-11
```

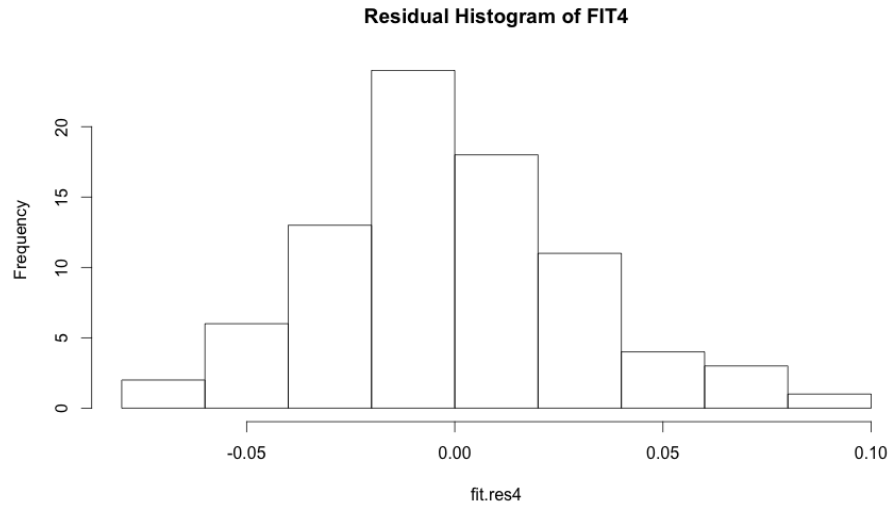
```
Call:
lm(formula = I(Y^2) ~ XLAG1)

Residuals:
     Min       1Q   Median       3Q      Max
-130.618  -38.136   -6.499   31.377  161.615

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  904.806      6.412 141.104  < 2e-16 ***
XLAG1         51.643      6.444   8.014  7.6e-12 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 57.15 on 80 degrees of freedom
Multiple R-squared:  0.4453,    Adjusted R-squared:  0.4384
F-statistic: 64.22 on 1 and 80 DF,  p-value: 7.601e-12
```

We also tried another model, FIT3: $Y^2 \sim$ XLAG1 (to left), which was an overall poor model when compared to the BIC statistics of other models. As another method of verification, we conducted the RESET test on our three best models (as chosen by the BIC test), and we failed to reject the null hypothesis of "all coefficients of the various powers of the fitted values are zero".

Additionally, we also decided to introduce log into the model (see FIT4 above) to further test for nonlinearity and observe whether the relationship between Y_t and X_{t-1} is multiplicative rather than additive. Log models also help correct for a positively skewed distribution of the residuals, which we observe in the residuals histogram for the linear regression of Y_t on X_{t-1} (as discussed in part C).

As X_{t-1} contains negative values, it is not possible to take the log of this independent variable (without transforming the variable in another way to make all values positive, such as taking its square), so we will take the log of Y_t and regress this on X_{t-1}.

For this logged model, (the summary statistics for FIT4 are in Table 2 above) show that the coefficients of this log Y_t on X_{t-1} regression are statistically significant, indicating that the log model may be a good fit for the data. The standard errors additionally are of fairly similar relative magnitude (in comparison to the estimate) as our estimators did in FIT1. Additionally, the BIC of log Y_t on X_{t-1} is actually -323.8864, which is 558.6137 points better than the simple linear regression (FIT1), indicating that logging Y_t substantially improved the fit of the model on the data. Below, we can also see that logging the data helped normalize the residuals.



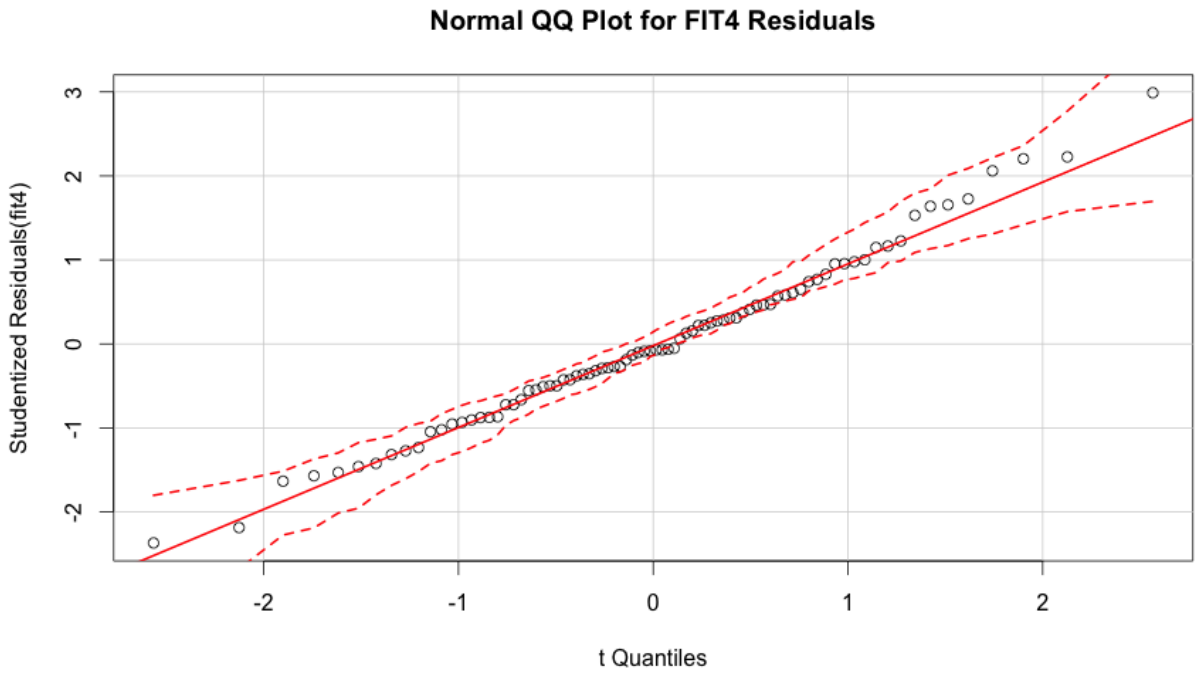
iii. Upon first observing the residual plot for FIT1 (see above), we thought that the residuals seemed slightly heteroskedastic (with the spread of the residuals decreasing as XLAG1 increased). In order to test for heteroskedasticity, we conducted Breusch-Pagan tests on our best models (FIT4, FIT6, FIT7) and FIT1.

	FIT1	FIT4	FIT6	FIT7
BP-test (p-value)	0.2255	0.1436	0.3165	0.4049

The BP-test tests whether the estimated variance of the residuals from a regression are dependent on the values of the independent variables. Each BP-test failed to reject the null hypothesis of homoscedasticity, thus indicating that it is safe to say the residuals are not heteroskedastic.

iv. As mentioned earlier, it does not appear as though there are any clear outliers in the dataset that might be strongly biasing the results of our regression. For this reason we did not remove any points from the dataset.

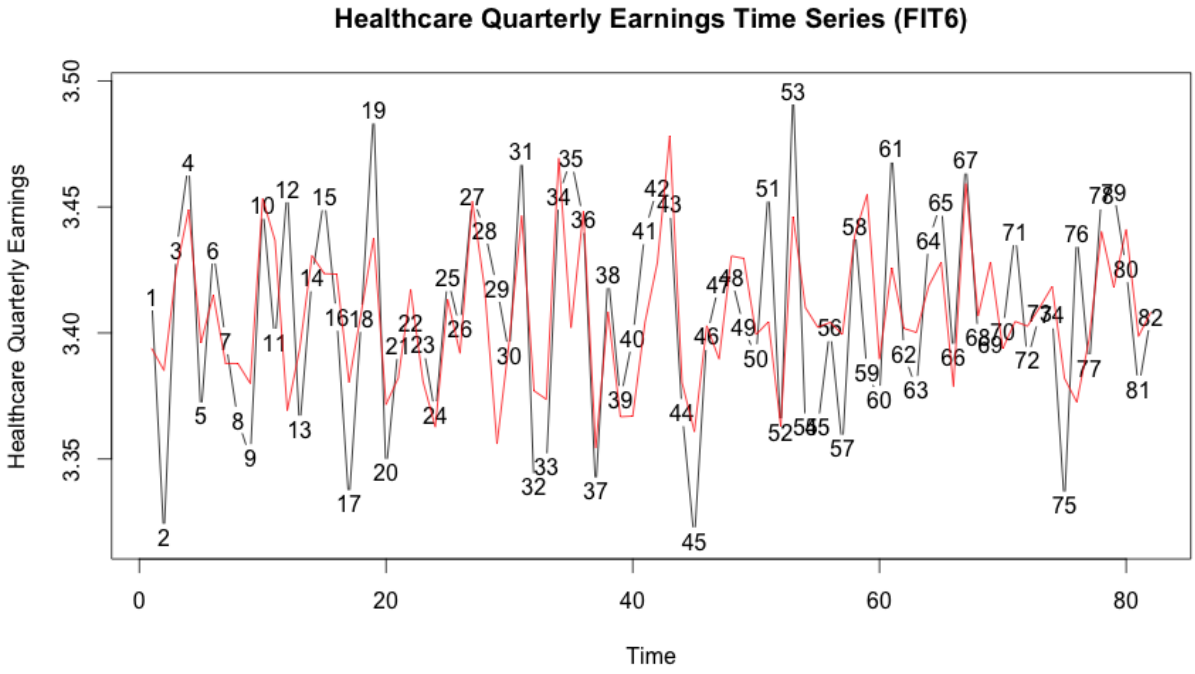
v. Based on the normal qq plot of residuals for FIT4 (see below), as well as the histogram of residuals for the log Y_t vs. X_{t-1} model (see above), we can see that the errors fairly closely follow a normal distribution. Thus, the condition of normality is justified and the estimates for the regressors are consistent, and perhaps efficient (though other violated conditions may be making our estimates inefficient).

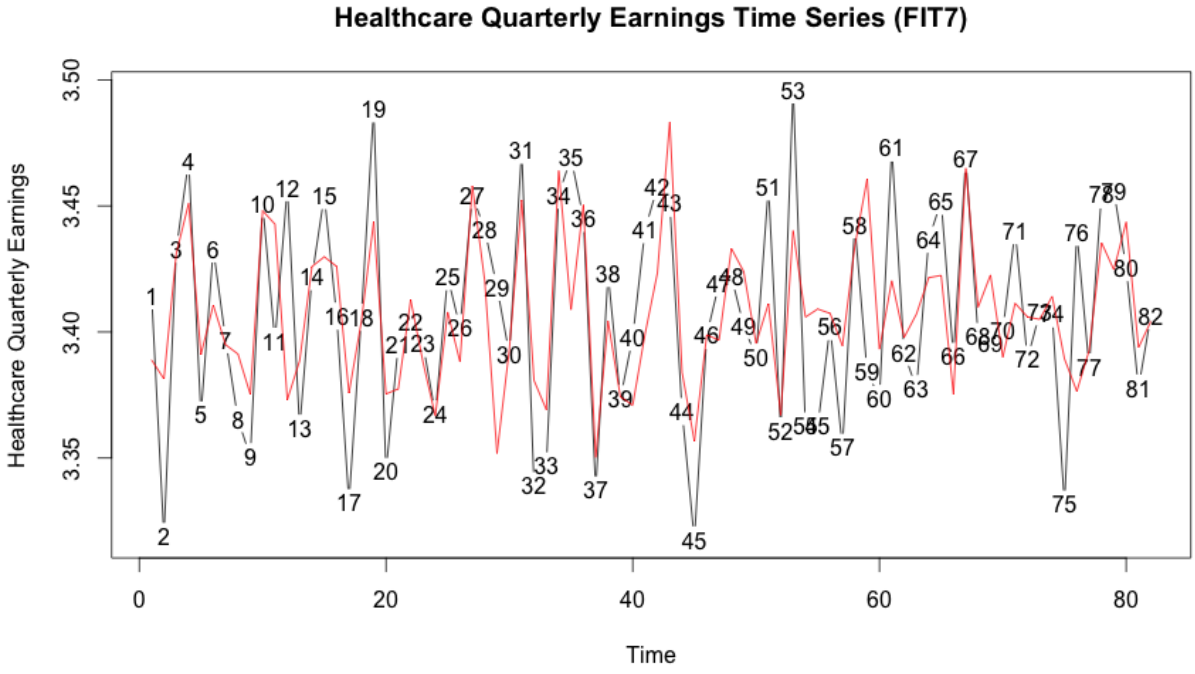


vi. Looking at our regression plots and diagnostics so far, there does not appear to be any structural change. That is, there does not seem to be a point within the data set where a new or unprecedented trend materializes. Nevertheless, we created a function that ran a test for structural change on each point (essentially a Chow test) in both our regular dataset & time series dataset for our best models FIT4, FIT6, and FIT7, as well as FIT1 for comparison. For each of these tests, we did not find a statistically significant p-value (alpha level = 0.05), suggesting that we can retain our null hypothesis of no structural change.

	FIT1	FIT4	FIT6	FIT7
Test for Structural Change	No p-value for any data point exceeded 0.05	No p-value for any data point exceeded 0.05	No p-value for any data point exceeded 0.05	No p-value for any data point exceeded 0.05

Other considerations. In addition to the "bare minimum considerations", we further investigated other models to see if we could find a better fit for the data. In particular, models including the time variable t (FIT5 and FIT6) and seasonal dummies (FIT7) were considered to observe if there was any pattern in the data with respect to time.





When we ran regressions (FIT5 and FIT6) that included a time variable, we did not find a statistically significant coefficient on t, leading us to believe that the data may have been de-trended over time. However, the plot of the time series Y_t seems to demonstrate periodic

oscillations, suggesting the potential presence of a seasonal pattern. To test this, we ran a model with quarterly seasonal dummies (FIT7). As shown above, these regressors were not statistically significant, suggesting that the time series Y_t does not follow a strong seasonal pattern.` that X represents a percentage change, it makes sense that the effect of an increase in X on Y is multiplicative rather than additive. For this reason, taking the log of Y allows us to analyze the effect of X on Y in percentage terms, thus producing an additive, linear model that best fits the data.

R Code

```
#Set work directory
setwd("/Users/jameswang/Documents/Dropbox/Notability/ECON221/P.Set_1")
getwd()

#import data
data <- read.table("PS0_2013_data.txt", header=TRUE)
attach(data)
hist(XLAG1)

###fit 8 different models
fit1 <- lm(Y~XLAG1)                                  #linear
fit2 <- lm(Y~XLAG1+I(XLAG1^2))                       #second-order polynomial
fit3 <- lm(I(Y^2)~XLAG1)                             #Y^2
fit4 <- lm(log(Y)~XLAG1)                             #log(Y)

data.ts = ts(data$Y)
t = seq(1,length(data.ts))
fit5 = lm(data.ts~t+XLAG1)                           #time-series

log.data.ts = log(data.ts)
fit6 = lm(log.data.ts~t+XLAG1)                       #time-series w/log(Y)

q1.mat = matrix(data = rep(c(rep(1,1),rep(0,3)),20),nrow=80,ncol=1)
q2.mat = matrix(data = rep(c(rep(0,1),rep(1,1),rep(0,2)),20),nrow=80,ncol=1)
q3.mat = matrix(data = rep(c(rep(0,2),rep(1,1),rep(0,1)),20),nrow=80,ncol=1)
q4.mat = matrix(data = rep(c(rep(0,3),rep(1,1)),20),nrow=80,ncol=1)

q1.temp = rbind(q1.mat,c(1))
q2.temp = rbind(q2.mat,c(0))
q3.temp = rbind(q3.mat,c(0))
q4.temp = rbind(q4.mat,c(0))

q1 = rbind(q1.temp,c(0))
q2 = rbind(q2.temp,c(1))
q3 = rbind(q3.temp,c(0))
q4 = rbind(q4.temp,c(0))

n = length(log.data.ts)
fit7 = lm(log.data.ts~t+XLAG1+q1+q2+q3+q4)            #time series + seasonal
 dummies

#perform model selection, use Bayesian Information Criterion (BIC/SIC)
BIC(fit1)      #234.7273
BIC(fit2)      #239.0946
BIC(fit3)      #907.4008
BIC(fit4)      #-323.8864 ***best
BIC(fit5)      #239.1339
BIC(fit6)      #-319.4803 ***second-best
BIC(fit7)      #-308.3462 ***third-best

#summary statistics for best models & fit1
```

```
summary(fit1)
summary(fit4)
summary(fit6)
summary(fit7)

#plot the best models
plot(XLAG1,Y,ylab="Healthcare Quarterly Earnings",xlab="Lagged-X Variable",
 main="Healthcare Quarterly Earnings on Lagged-X Variable")
abline(fit1,col="red")

plot(XLAG1,log(Y),ylab="Y",xlab="XLAG1")
abline(fit4,col="red")

plot.ts(t,log.data.ts,xlab="Time",ylab="Healthcare Quarterly
 Earnings",main="Healthcare Quarterly Earnings Time Series (FIT6)")
lines(fitted(fit6),col="red")

plot.ts(t,log.data.ts,,xlab="Time",ylab="Healthcare Quarterly
 Earnings",main="Healthcare Quarterly Earnings Time Series (FIT7)")
lines(fitted(fit7),col="red")

#plot residuals
fit.res1 <- resid(fit1)
plot(XLAG1, fit.res1, ylab="Residuals", xlab="XLAG1",main="Residual Plot of
 FIT1 Residuals")
abline(0,0)
hist(fit.res1,main="Residual Histogram of FIT1")

fit.res4 <- resid(fit4)
plot(XLAG1, fit.res4, ylab="Residuals", xlab="XLAG1")
abline(0,0)
hist(fit.res4,main="Residual Histogram of FIT4")

fit.res6 <- resid(fit6)
plot(XLAG1, fit.res6, ylab="Residuals", xlab="XLAG1")
abline(0,0)
hist(fit.res6)

fit.res7 <- resid(fit7)
plot(XLAG1, fit.res7, ylab="Residuals", xlab="XLAG1")
abline(0,0)
hist(fit.res7)

#plot normal qq plots
qqPlot(fit1,main="Normal QQ Plot for FIT1 Residuals")
qqPlot(fit4,main="Normal QQ Plot for FIT4 Residuals")
qqPlot(fit6,main="Normal QQ Plot for FIT6 Residuals")
qqPlot(fit7,main="Normal QQ Plot for FIT7 Residuals")

#plot leverage plot ... identify influential points
plot(fit1) #see 4th plot
plot(fit4) #see 4th plot
plot(fit6) #see 4th plot
plot(fit7) #see 4th plot
```

```
#conduct Bonferroni-adjusted outlier test ... test for outliers ...
outlierTest(fit1) #p-value = .28974 ... no outliers
outlierTest(fit4) #p-value = .30527 ... no outliers
outlierTest(fit6) #p-value = .29547 ... no outliers
outlierTest(fit7) #p-value = .40510 ... no outliers

#conduct RESET test ... test for linearity
reset(fit1,power=2:3,type="regressor",data=data)  #.1222 not significant,
 functional form is linear
reset(fit4,power=2:3,type="regressor",data=data)  #.1284 not significant,
 functional form is linear
reset(fit6,power=2:3,type="regressor",data=data)  #.2568 not significant,
 functional form is linear
reset(fit7,power=2:3,type="regressor",data=data)  #.9396 not significant,
 functional form is linear

#conduct Durbin-Watson Test ... test for serial correlation
#install.packages("car")
#library(car)
dwt(fit1,simulate=TRUE)     #2.2982
dwt(fit4,simulate=TRUE)     #2.2912 ***best
dwt(fit6,simulate=TRUE)     #2.2913 ***second-best
dwt(fit7,simulate=TRUE)     #2.2978 ***third-best

#conduct Breusch-Pagan Test ... test for heteroskedasticity
#install.packages("lmtest")
#library(lmtest)
bptest(fit1)     #p-value = .2255 ... fail to reject null hypothesis of
 homoskedasticity
bptest(fit4)     #p-value = .1436 ... fail to reject null hypothesis of
 homoskedasticity
bptest(fit6)     #p-value = .3165 ... fail to reject null hypothesis of
 homoskedasticity
bptest(fit7)     #p-value = .4049 ... fail to reject null hypothesis of
 homoskedasticity

#conduct test ... test for structural change
#install.packages("strucchange")
#library(struccchange)
struc.change <- function(y,variables,data.set) {
  vec1 <- vector("numeric")
  vec2 <- logical()
   for (i in 1:82) {
      result <- sctest(y~variables,data=data.set,type="Chow",i)
      vec1 <- c(vec1,result$ p.value)
   }
  for (j in vec1)
  {
    if (j <= 0.05) {
      j = FALSE
      vec2 <- c(vec2,j)
    } else {
      j = TRUE
```

```
     vec2 <- c(vec2,j)
    }
  }
  return(vec2)
}

test1 <- struc.change(Y,c(XLAG1),data) #no rejections of the null...no
 structural change
test2 <- struc.change(log(Y),c(XLAG1),data) #no rejections of the null...no
 structural change
test3 <- struc.change(log(Y),c(XLAG1+t),data.ts) #no rejections of the
 null...no structural change
test4 <- struc.change(log(Y),c(XLAG1+t+q1+q2+q3+q4),data.ts) #no rejections
 of the null...no structural change
```