Keun Gang Kyen

Stats 153

Getting into the Stomach of the Question: Watching Rice Grow

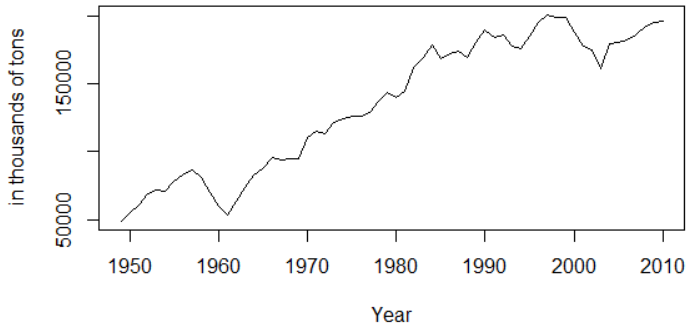
INTRODUCTION: Rice is the most important grain for human consumption where there are more than 40,000 varieties of rice! Before getting to the heart (stomach) of the question, here are some fun facts about rice: Rice originated from India 4,000 years ago, but the first known account was in China around 2,800 BC. Today an American consumes 20 pounds of rice annually with 4 pounds going towards beer brewing. Alas, Canada produces no rice of its own! How shocking! After being cooked, rice swells to three times its original weight. The real question should be: why does processed rice cost more than whole, "natural" rice?

BACKGROUND: My data comes from the US Department of Agriculture's website focusing on China's annual crop production of rice (in units of thousands of tons) from 1949 to 2010. My second time series is annual GDP per capita of China from 1952 to 2009 in Yuan (http://www.ers.usda.gov/data-products/china-agricultural-and-economic-data.aspx#.UqnyY_RDuSo). The USDA website acknowledges the collection of data originates from "official statistical publications of the People's Republic of China." Thus in this report I assume that the data are not inherently "massaged" to create a unrealistic trend of growth or any artifact not arriving from actual data in order to make a strong model.

QUESTION: The question I am studying is whether a univariate model (solely utilizing rice production) or a bivariate model (with GDP per capita as the explanatory variable and rice production as the response variable) will better predict future rice production.

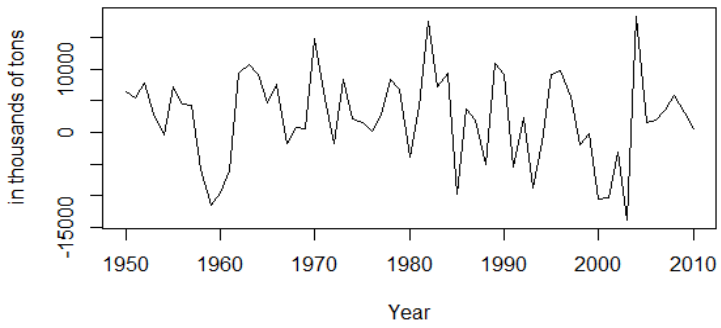
HYPOTHESIS: Adding another variable that should trend closely with rice data, I expect to find that a bivariate model is a stronger model to predict future annual rice production.

Annual Rice Production in China (Fig. 1)



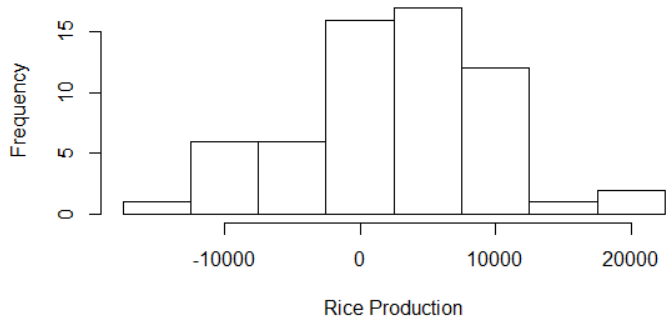
METHOD: Taking a look at rice production (Fig. 1), we see that annual rice production is obviously not stationary, so we can take a first difference to remove the upward trend.

First Difference of Rice Production (Fig. 2)



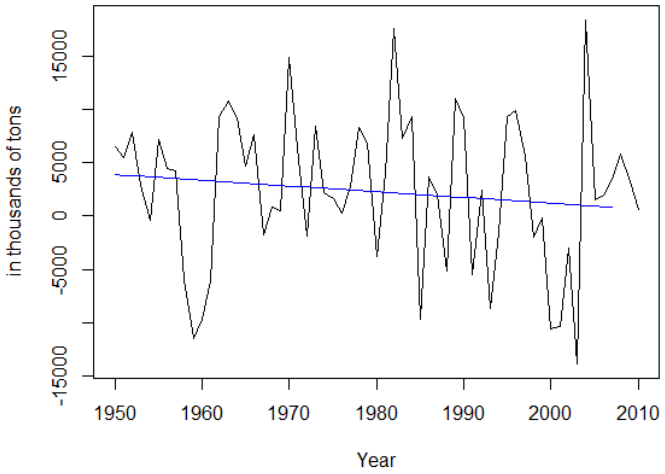
With the first difference (fig. 2), we see that the transformed time series is relatively stationary with a constant mean and adequately constant variance. Hence, we see that there is no need for a GARCH model.

Fig. 3



In figure 3, we created a histogram of the first difference of the rice time series. Hence, annual growth of rice production (essentially the first derivative) is not standard normally distributed nor spread out like a bell curve due to the "fat tails."

First Difference of Rice Production (Fig. 4)

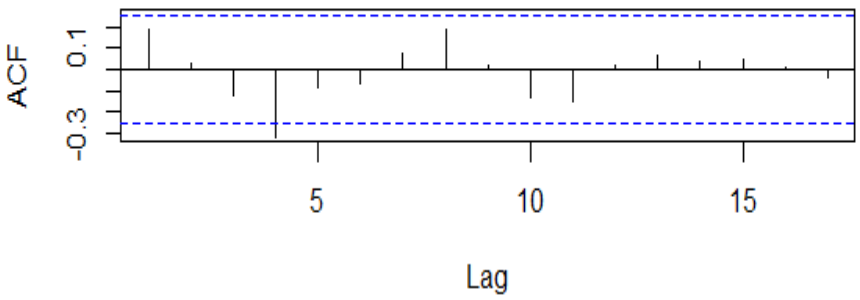


Regressing the first difference on year (fig. 4), there seems to be a slightly download sloping where the intercept is 3948 thousands of tons and slope is -53.56 thousand tons per year. Though the p-value for the intercept is statistically significant (p-value=.0447) with alpha=.05, the p value for the slope is not significant (p value is .3489). Hence, R-squared is essentially 0 as just taking the mean of the first difference gives you a decent prediction! Had the p value for slope been more statistically significant, it still would have not been practically significant as the slope is -53

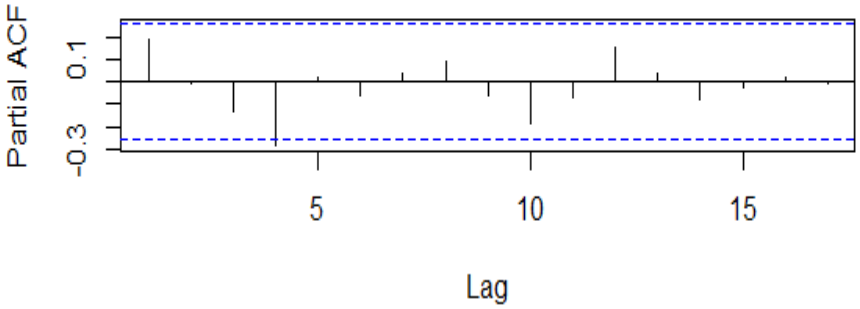
where the mean of the first difference is 2411 (only about 2% of the mean). The negative slope can be explained by a variety of interrelated factors: diminishing marginal returns where growth rate should not persistently remain high. Another explanation is that in a separate time series, the crop land area for rice has been kept relatively stationary, albeit decreasing from the late 1990s—hence there is less land to grow rice.

Fun Fact: In 1958 to 1961, China embarked on the Great Leap Forward where agricultural changes led to the Great Chinese Famine. Among the "scientific" agricultural policies of growing crops closer to increase crop yield or plowing deeper to reach more fertile soil, these techniques failed drastically and backfired, leading to millions dying from starvation—hence the decreasing in rice production from 1958 to 1961.

Difference of Rice Time Series (Fig. 5)



Difference of Rice Time Series (Fig. 6)



Looking at the ACF plot of the residuals (for linear model for the first difference), fig. 5 shows that the tail cuts to non-significance immediately so we do not need to use a moving average process. Looking at the PACF plot of the residuals, fig. 6 shows that the residuals do not seem to be regressing onto previous year's data points, so we do not need to use an autoregressive process. At lag 4 for both plots, we do have a "significant" bar that lies outside our null hypothesis bands, but at a 95% confidence level, we do

expect 1 bar to extend beyond our confidence interval. Fortunately for us, the Ljung Test has a p-value of .11, indicating that we do not have significant autocorrelations. One could argue that there may exist a small seasonal trend in the residuals and that a marginally better model can be constructed.

Fig. 7

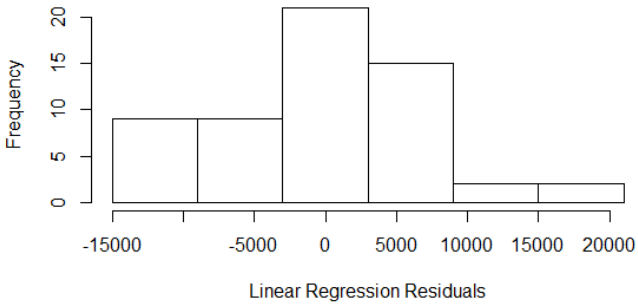
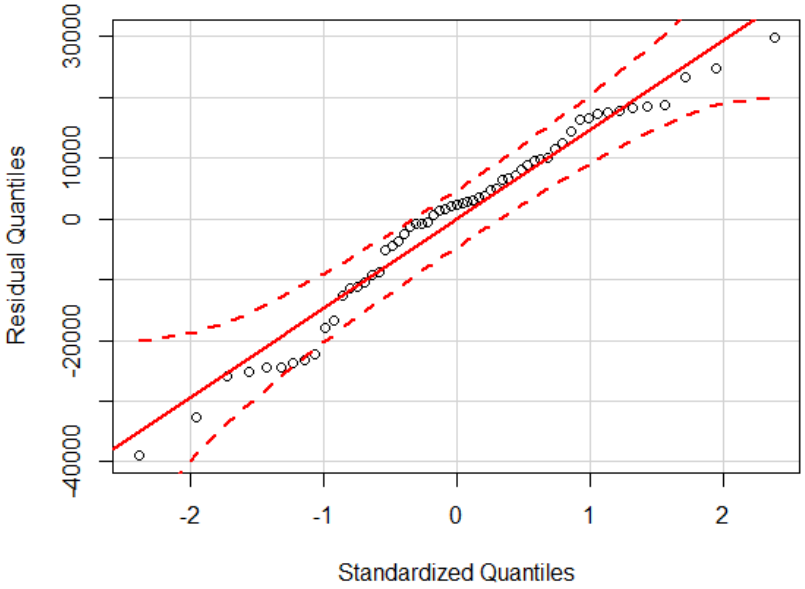


Fig. 7 and Fig. 8 show the residuals of the linear regression of the first difference do exhibit some properties of normality. The histogram shows that the residuals are centered around 0, but are not distributed symmetrically and does has a long right tail. The Q-Q plot has only 2 points outside of the 95% confidence bands when we can expect 3. The points outside of the bands may arrive from Great Famine from 1958 to 1961. First difference regressed on year seems to produce a strong model with normally distributed residuals despite a low R-squared.

Fig. 8



Using the eacf, we see that our previous conclusions are supported as an appropriate model would be an "ARMA" function with order (0,0)—the equation simply being Y(t)= e(t). We have strong evidence that the first difference of rice production exhibits a random walk. Hence, let's first difference linear regression to make a prediction on the rice production in years 2008 and 2009 (fig. 9). Pretty close!

AR/MA

	0	1	2	3	4	5
0	o	o	o	x	o	o
1	o	o	o	o	o	o
2	o	o	o	o	o	o
3	x	o	x	o	o	o
4	o	o	o	x	o	o
5	o	o	o	x	o	o

For more data supporting our conclusion, we can "undo" the first difference and look back at the raw data. Here is the residuals for a linear model of rice production regressed on time (fig. 10). The histogram of residuals in not very normal, only centered at 0, asymmetric, and long left tail. However, histogram conflicts with the Q-Q plot (fig. 11), which shows that the residuals are in fact relatively normal! Only 1 point lies outside the 95% confidence level bands, which may the left tail seen in the histogram. In the middle of the Q-Q plot, we see that points are on top of the theoretical line, hinting at some "irregularities" with the shape of the bell curve.

Using a linear model where rice production is regressed upon year (fig. 12), the R squared is .89 (using the Box-Cox transform, I found lambda to be 1.071 and the resulting R-

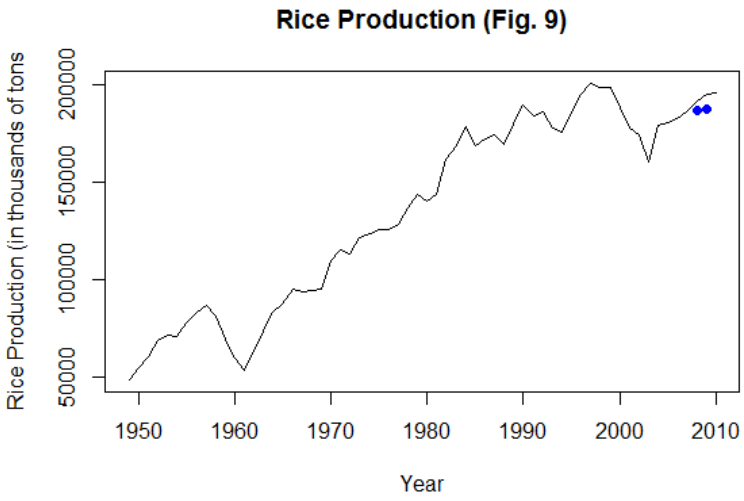


Fig. 10

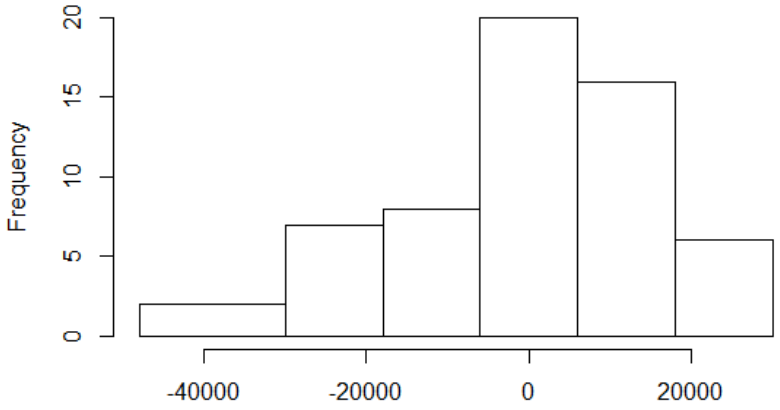
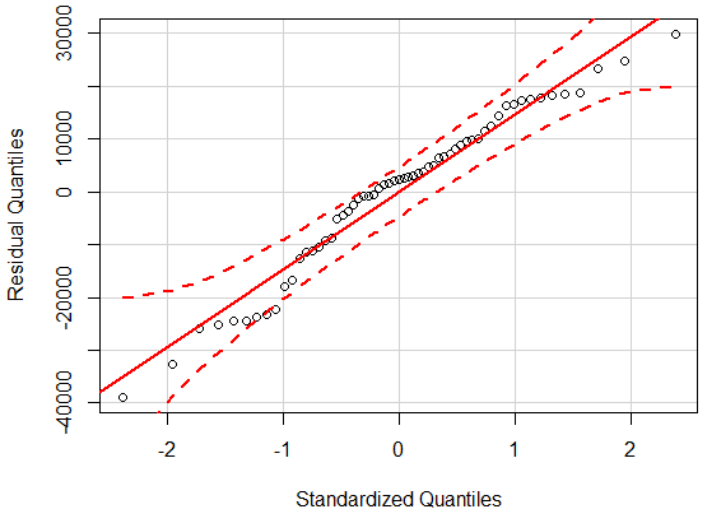
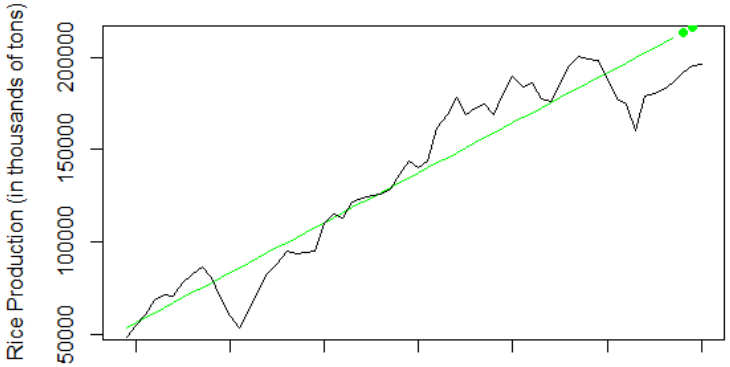


Fig. 11



Rice Production Linear Model (Fig. 12)



squared to essentially remain the same.) Though I chose to ignore the transform as I did not deem it to be significant, the Box-Cox procedure verified that our linear model is sufficient robust with nearly 90% of the variation explained.

We quickly determine that rice production regressed on time is not as good of a model for predicting 2008 and 2009 compared to the first difference regression despite this R-squared is superior to that of the first difference. Looking at the predictions for 2008 and 2009, we see that the predictions are not very close to the actual data. This can be explained by interference from events such as the 2001 Tech Bubble or other natural disasters affecting crop land.

```
AR/MA
  0 1 2 3 4 5
0 x x x x x x
1 o o o x o o
2 o o o x o o
3 x o o x o o
4 x x o o o o
5 x o o o o o
```

Let's try an ARMA process to model rice production. Here is the eacf on raw rice production data. Using an AR(1), the coefficients are revealed—the coefficient for Y(t-1) is nearly identical to 1 and intercept is around 120,000: Y(t)=.993Y(t-1)+e(t)+119,667. We expected this result as we already showed the time series to be a random walk. Fortunately, our AR(1) model (fig. 13) predicts quite accurately for 2008 and 2009 because it already took in account of the previous years where rice production fell below the linear trend. It helps that the AR and the actual data lie nearly on top of each other. Take a look at the histogram (fig. 14) and the Q-Q plot (fig 15) of

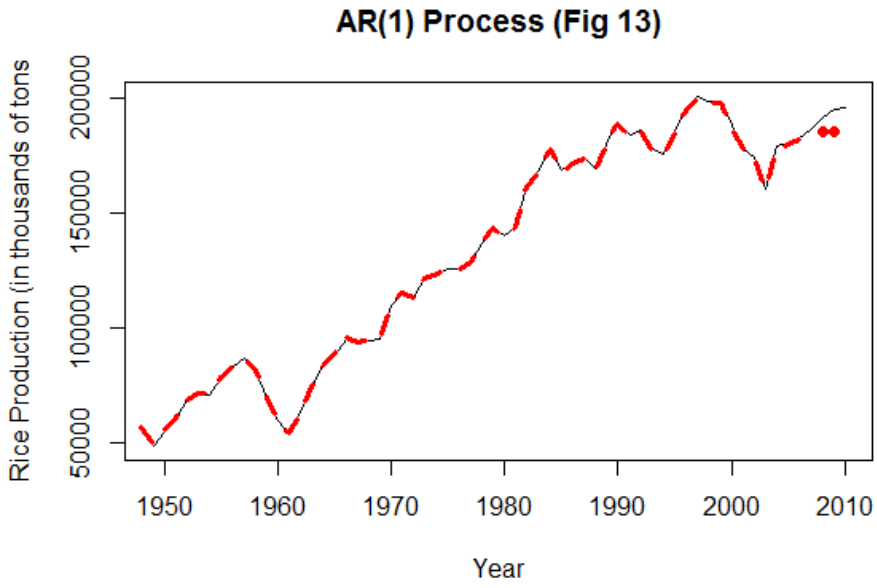


Fig. 14

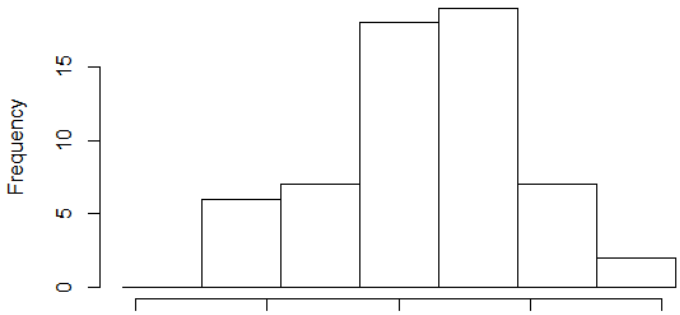
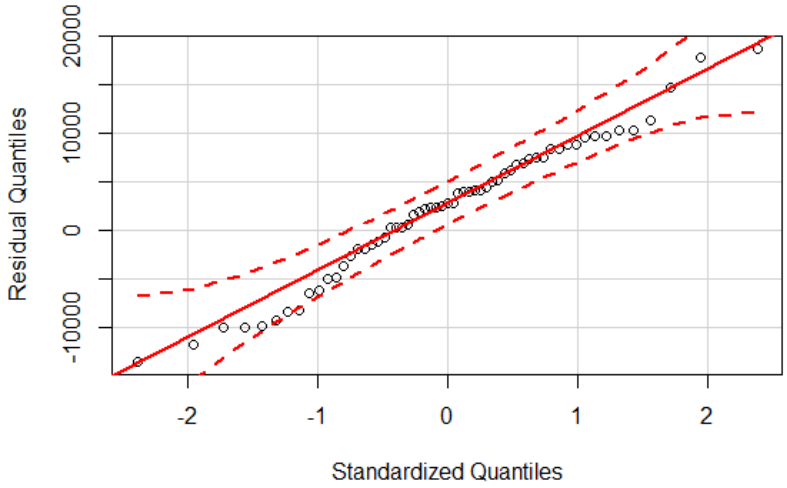
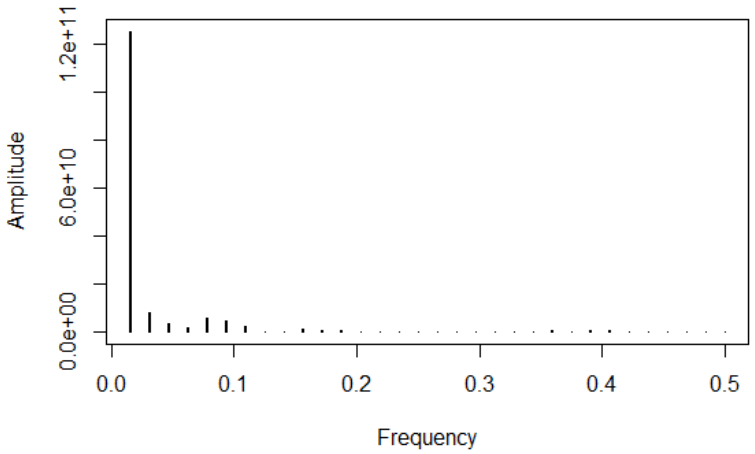


Fig. 15



the residuals of the AR(1)—they are nearly perfect. The histogram shows symmetry and a bell shape while the Q Q plot only has 1 point outside the 95% confidence level bands. With the Ljung test, the p-value of the residuals is .17, indicating that we do not have significant auto-correlations. To conserve space, I give you my assurance that nothing undesirable (patterns or significance) in the ACF/PACF is occurring.

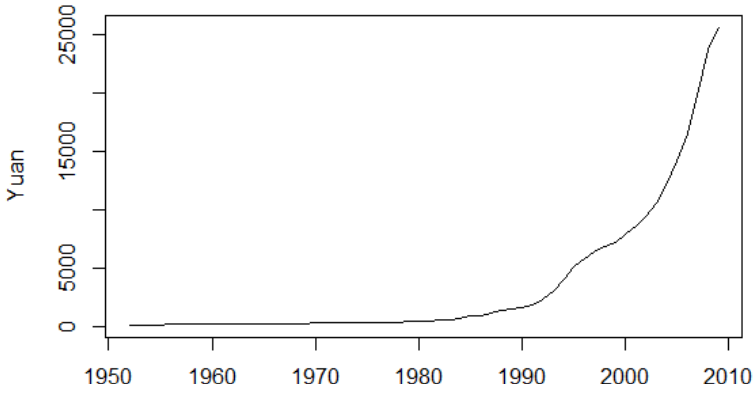
Periodogram (Fig. 16)



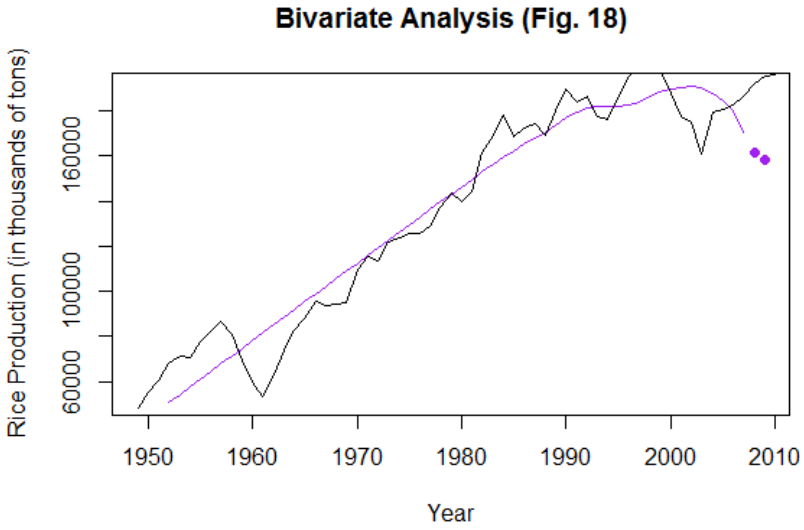
The periodogram of the rice data in fig. 16 further justifies our conclusion that it is a random walk, supporting our autoregressive process. It shows that the period is very short, essentially that every point is a new "period" so that there is effectively no "period."

Let's now transition to a bivariate model, which takes in account of GDP per person in China. GDP per capita seems to grow at an exponential rate (fig. 17). Just by regressing rice on years and GDP per capita, our adjusted R-squared is .93! Running a Box-Cox transform, we find lambda to be .87 with the transformed model having essentially the same R-squared—.93.Though I chose to again ignore the transform as I did not deem it to be significant, the Box-Cox procedure verified that our linear model is sufficient robust with nearly 93% of the variation explained.

GDP per capita in China (Fig. 17)



Although GDP per capita seems to be exponential growth, taking the log of GDP per capita for the regression actually results in a lower R-squared. I have tried different transforms using a combination of first differences and log for the GDP per capita and the R-squared remains around .89. Trying again while removing the effect of time, the R-squared falls between .32 and -01. Hence, I have retained rice regressed onto year and GDP per capita, despite it not make intuitive sense.



Unfortunately, our prediction is rather weak and even worse than our previous linear regression onto only time. One possible explanation is that the coefficient for time is positive but the coefficient for GDP per capita is negative: in 2008 and 2009, GDP per capita soared, thus actually pushing the predictions for rice production even lower!

From the high R-squared, we expected the residuals from the regression onto time and GDP per capita to be quite normal (fig. 19), and indeed they are. In the Q-Q plot (fig. 20), only 5 points are outside of the 95% confidence level bands when we expected 3.

Fig. 19

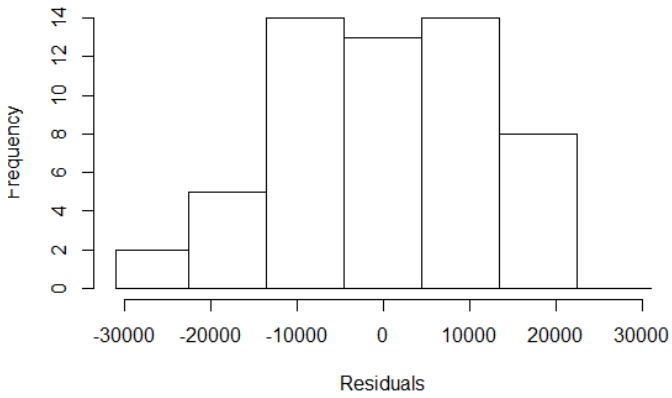
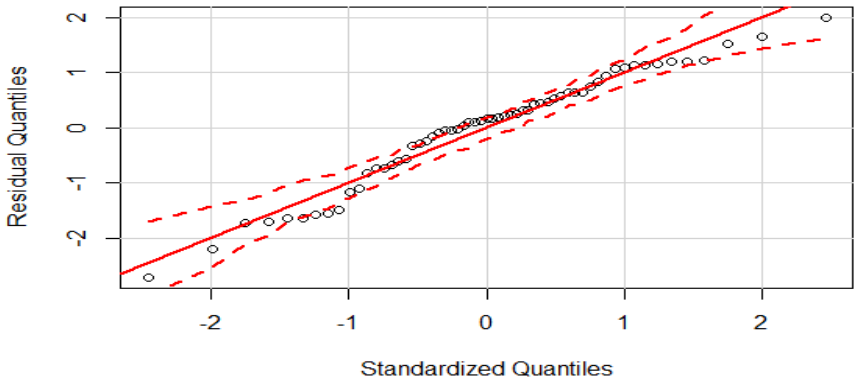
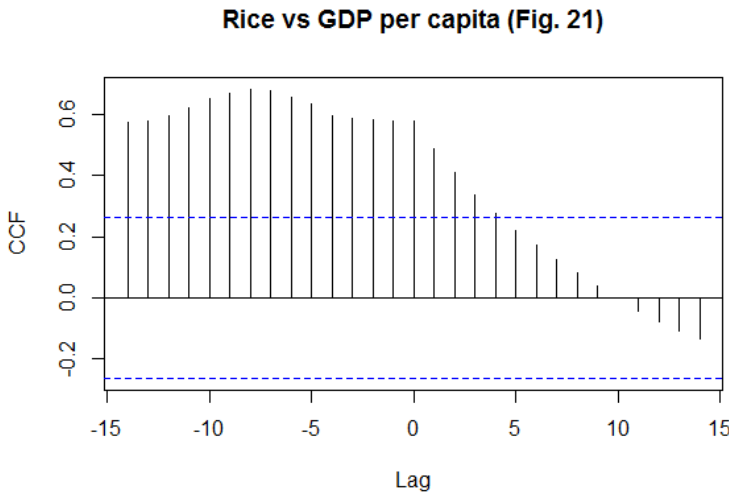


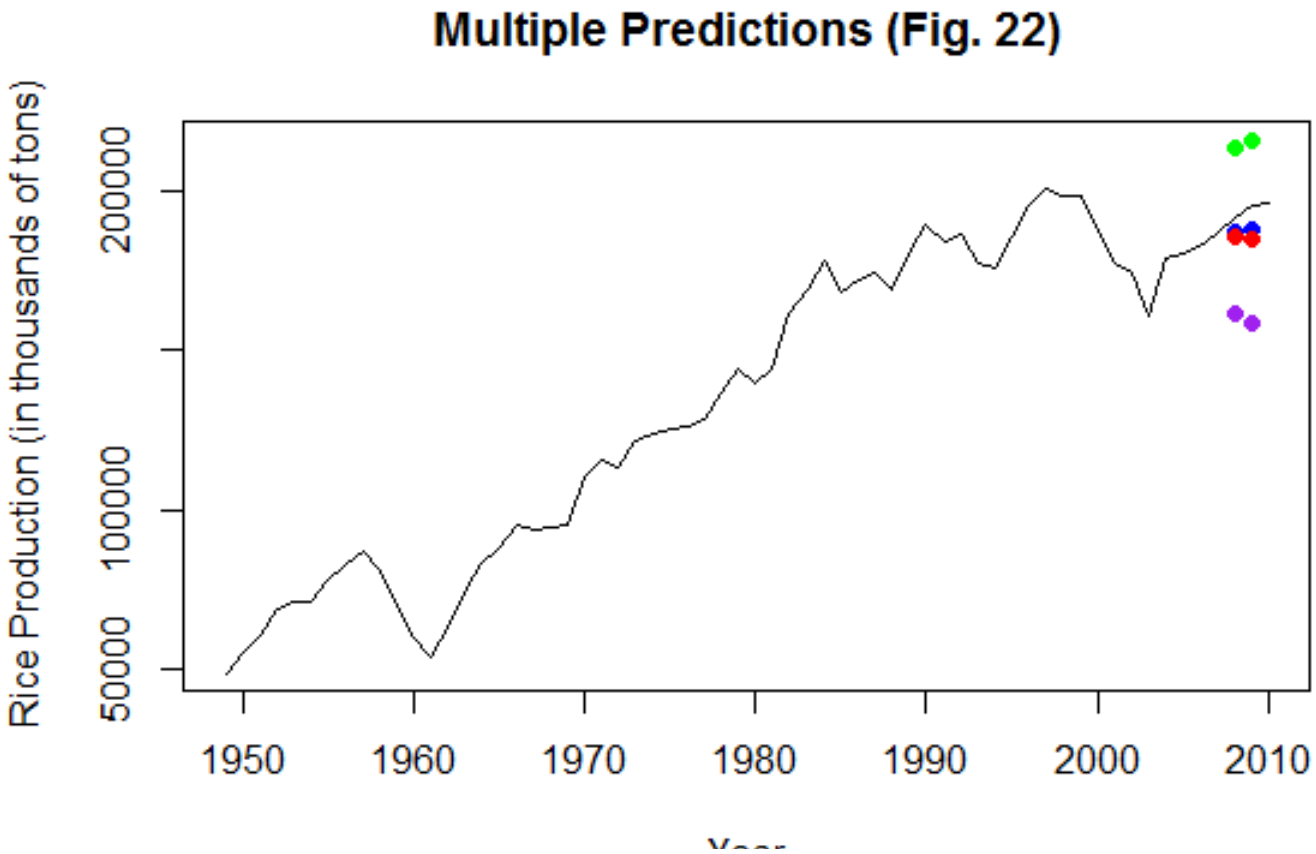
Fig. 20



From the cross-correlation (fig. 21), we see something quite interesting. It appears that rice production precedes GDP per capita increase! In fact, one may conjecture that rice production causes GDP per capita. Hence, rice may be able to predict GDP per capita instead of what we are doing now.



Back to our original question—which model has better predictions given: univariate or bivariate? Looking at the all 4 models juxtaposed, the prediction from the first difference of rice onto time (blue) is the best model out of the group with the smallest sum of squares of the residual being 53,819,632,768. The univariate AR(1) model (red) came next with sum of squares being 139,930,484. Using the linear model of rice onto time (green), the sum of squares is 880,378,938; for the bivariate rice onto time and GDP per capita (purple), it is 2,263,733,565



ANSWER: The answer to my question is that although we take account of GDP per capita in a bivariate model, just using looking at the first difference of rice data makes a much stronger prediction. However, the validity of our conclusion is limited by the fact that we only made 2 predictions into the future and in the long run, the bivariate model may do better. In addition, we built our model using highest R-squared for linear models and eacf for the ARMA. However, these tools generate the best model for given data, and if perhaps our assumption that the first difference of rice is simply a random walk, then they have weak predicting ability into the distant future. The irony is that our best model (which has the lowest R-squared of essentially 0) beats out the worst model that has the highest R-squared (.93)! Wow, statistics!

NOTES: The BIC value for my AR(1) of the rice data is 1233. This value is the lowest BIC value for ARMA process of different orders, justifying along with the eacf that AR(1) is an appropriate model.

For future deeper analysis, I noticed that the first difference may have seasonality, so one can construct a seasonal model that may have better prediction ability. For the GDP per capita, I did try using log as the growth seems exponential. Though the resulting model had a lower R-squared, one may try to lag GDP per capita forward as rice production seems to precede GDP per capita in order to generate a model with higher R-squared. For a better univariate model, one can try to find a polynomial regression onto time for a higher R-squared. For a better bivariate model, one may try to remove the effect of time from either rice and/or GDP per capita. Lastly, thank you Professor Brillinger for reading this far!

Appendix:

```
GDP=ts(rev(c(25575,23708,20169,16500,14185,12335,10541,9398,8621,7857,7158,6796,6420,
5845,5045,4044,2998,2311,1892,1644,1519,1365,1112,963,857,695,582,527,492,463,419,381,3
39,316,327,310,309,292,288,275,243,222,235,254,240,208,181,173,185,218,216,200,168,165,15
0,144,142,119)),start=1952,frequency=1)
rice=ts(rev(c(195760,195102,191900,186034,182571,180588,179087,160656,174539,177580,18
7908,198487,198713,200735,195103,185226,175933,177514,186222,183813,189331,180130,16
9107,174262,172224,168569,178255,168865,161595,143955,139910,143750,136930,128565,12
5810,125560,123905,121735,113355,115205,109990,95065,94530,93685,95390,87720,83000,7
3765,62985,53640,59730,69365,80850,86775,82480,78025,70850,71270,68425,60555,55100,48
645)),start=1949,frequency=1)
library(TSA)
library(MASS)
plot(rice,ylab="in thousands of tons", xlab="Year",main="Annual Rice Production in China
(Fig. 1)")
riceD=diff(rice)
plot(riceD,xlab="Year",ylab="in thousands of tons", main="First Difference of Rice Production
(Fig. 2)")
hist(riceD,xlab="Rice Production",main="Fig. 3",breaks=c(-17500,-12500,-7500,-
2500,2500,7500,12500,17500,22500))
library("car")
riceDModel=lm(riceD[1:58]~time(rice)[2:59])
plot(riceD,xlab="Year",ylab="in thousands of tons", main="First Difference of Rice Production
(Fig. 4)")
lines(time(rice)[2:59],riceDModel$fitted.values, col="blue")
points(time(rice)[60],riceDModel$coef[1]+riceDModel$coef[2]*time(rice)[60],col="blue",pch=
19)
points(time(rice)[61],riceDModel$coef[1]+riceDModel$coef[2]*time(rice)[61],col="blue",pch=
19)
acf(riceDModel$resid, main="Difference of Rice Time Series (Fig. 5)")
pacf(riceDModel$resid,main="Difference of Rice Time Series (Fig. 6)")
eacf(riceD)
Box.test(riceD,type="Ljung-Box")
hist(riceDModel$residual,main="Fig. 7", xlab="Linear Regression Residuals",breaks=c(-15000,-
9000,-3000,3000,9000,15000,21000))
qqPlot(riceDModel$residual,xlab=”Standardized Quantiles”,ylab=”Residual
Quantiles”,main=”Fig. 8”)
plot(rice,xlab="Year",ylab="Rice Production (in thousands of tons",main="Rice Production (Fig.
9)")
points(2008,riceDModel$coef[1]+riceDModel$coef[2]*time(rice)[60]+rice[59],col="blue",pch=
19)
points(2009,riceDModel$coef[1]+riceDModel$coef[2]*time(rice)[60]+
      riceDModel$coef[1]+riceDModel$coef[2]*time(rice)[61]+rice[59],col="blue",pch=19)

riceModel=lm(rice[1:59]~time(rice)[1:59])
hist(riceModel$resid,xlab="Residuals",main="Fig. 10", breaks=c(-48000,-30000,-18000,-
6000,6000,18000,30000),freq=T)
```

```
qqPlot(riceModel$residuals,xlab="Standardized Quantiles",ylab="Residual
Quantiles",main="Fig. 11")
plot(time(rice)[1:59],riceModel$fitted.values,type="l",xlab="Years",
    ylab="Rice Production (in thousands of tons)", main="Rice Production Linear Model (Fig.
12)", col="green",xlim=c(1949,2010))
lines(rice)
points(2008,riceModel$coef[1]+riceModel$coef[2]*2008,pch=19,col="green")
points(2009,riceModel$coef[1]+riceModel$coef[2]*2009,pch=19,col="green")
acf(riceModel$resid)
pacf(riceModel$resid)
eacf(rice)

riceAR=arima(rice[1:59],order=c(1,0,0))
plot(rice,xlab="Year",ylab="Rice Production (in thousands of tons",main="AR(1) Process (Fig
13)")
lines(time(rice)[1:59]-1,rice[1:59]-riceAR$resid, type="l",col="red",lty=2,lwd=3)
points(2008:2009,predict(riceAR,n.ahead=2)[[1]],col="red",pch=19)
Box.test(riceAR$resid,type="Ljung")
hist(riceAR$residuals,xlab="Residuals",main="Fig. 14",breaks=c(-21000,-15000,-9000,-
3000,3000,9000,15000,21000))
qqPlot(riceAR$residuals,xlab="Standardized Quantiles",ylab="Residual Quantiles",main="Fig.
15")
periodogram(rice,ylab="Amplitude",main="Periodogram (Fig. 16)")

plot(GDP,main="GDP per capita in China (Fig. 17)",xlab="Year",ylab="Yuan")
RvsG=lm(rice[4:59]~time(rice)[4:59]+GDP[1:56])
plot(time(rice)[4:59],RvsG$fitted.values,type="l",xlim=c(1949,2009),col="purple",xlab="Year",
    ylab="Rice Production (in thousands of tons)",main="Bivariate Analysis (Fig. 18)")
points(time(rice),rice,type="l")
points(2008,RvsG$coef[1]+2008*RvsG$coef[2]+GDP[57]*RvsG$coef[3],col="purple",pch=19)
points(2009,RvsG$coef[1]+2009*RvsG$coef[2]+GDP[58]*RvsG$coef[3],col="purple",pch=19)
hist(RvsG$resid,freq=T,xlab="Residuals",main="Fig. 19",breaks=c(-31000,-22500,-13500,-
4500,4500,13500,22500,31000))
qqPlot(riceModel,xlab="Standardized Quantiles",ylab="Residual Quantiles",main="Fig. 20")

ccf(rice[4:59],GDP[1:56],main="Rice vs GDP per capita (Fig. 21)",ylab="CCF")

plot(rice,ylim=c(50000,215000),xlab="Year",ylab="Rice Production (in thousands of
tons)",main="Multiple Predictions (Fig. 22)")
points(2008,riceDModel$coef[1]+riceDModel$coef[2]*time(rice)[60]+rice[59],col="blue",pch=
19)
points(2009,riceDModel$coef[1]+riceDModel$coef[2]*time(rice)[60]+
      riceDModel$coef[1]+riceDModel$coef[2]*time(rice)[61]+rice[59],col="blue",pch=19)
points(2008:2009,predict(riceAR,n.ahead=2)[[1]],col="red",pch=19)
points(2008,riceModel$coef[1]+riceModel$coef[2]*2008,pch=19,col="green")
points(2009,riceModel$coef[1]+riceModel$coef[2]*2009,pch=19,col="green")
points(2008,RvsG$coef[1]+2008*RvsG$coef[2]+GDP[57]*RvsG$coef[3],col="purple",pch=19)
points(2009,RvsG$coef[1]+2009*RvsG$coef[2]+GDP[58]*RvsG$coef[3],col="purple",pch=19)
```