Solution HW 2

Martin Waibel

March 14, 2022

Exercise 1

For this exercise, use the data in "pension.dta". The equation of interest is a linear probability model

$$pira = \beta_0 + \beta_1\, p401k + \beta_2\, inc + \beta_3\, inc^2 + \beta_4\, age + \beta_3\, age^2 + u \qquad (1)$$

where pira is a binary (e.g., 0/1) variable equal to 1 if the person has an IRA account. The goal is to test whether there is a tradeoff between participating in a 401(k) plan and having an individual retirement account (IRA). Therefore, we want to estimate β_1.

a) Estimate the equation (1) by OLS and discuss the estimated effect of *p401k*.

   - **Solution:** The first column in Table (1) lists the OLS estimation result for model (1). The estimated effect of p401k is about 0.054 and is significant at the 1% significance level. It implies that if an individual participates in a 401(k) plan the probability of participating in an IRA plan will increase by around 5.4 percentage point. Hence, the participation rates of the two retirement plans are positively correlated.

b) For the purposes of estimating the ceteris paribus tradeoff between participation in two different types of retirement savings plans, what might be a problem with ordinary least squares (OLS)?

   - **Solution:** One problem might be a classic omitted variable bias problem. For instance, high income people might save in both 401(k) and IRA plans, i.e. income is positively correlated with both plan dummies. As we omit the income variable from the model it might well be that β_1 turns insignificant once we control for the omitted factor (just a hypothesis, not acutally tested). Another problem, inherent to the linear probability model is that the the predicted probability, $\hat{Y}$, might be outside the unit interval. In this case the interpretation would not be aligned with a (sensible) percent interpretation anymore. In contrast, probit and logit models are by construction restricted to yield predictions inside the unit interval.

c) The variable $e401k$ is a binary variable equal to one if a worker is elegible to participate in a 401(k) plan. Explain what is required for $e401k$ to be a valid IV for p401k. Do these assumptions seem reasonable?

- **Solution:** Consider the general system $Y = \beta_1 \mathbf{X_1} + \beta_2 \mathbf{X_2} + \epsilon$. We suspect $\mathbf{X_1}$ to be endogenous and therefore want to apply an IV approach using the instrument $\mathbf{Z}$. $\mathbf{X_2}$ is assumed to be a matrix of exogenous regressors. Hence, in our case the instrument $\mathbf{Z}$ is $e401k$.
- Condition 1: **Relevance condition** (mathematically: $plim \frac{1}{N}(\mathbf{Z}'\mathbf{X_1}) \neq 0$). Intuitively, this implies that the instrument needs to be significantly correlated with the endogenous regressor (conditional on all exogenous regressors). Conceptually, we want to "*only exploit variation in* X_1 *driven by the instrument*."
- Condition 2: **Exclusion condition** (mathematically: $plim \frac{1}{N}(\mathbf{Z}'\epsilon) \neq 0$). The intuition of this condition is that the instrument is **uncorrelated** with the error term. In particular, the exclusion restriction consists of the following two conditions.
  - Condition 2.1: **Externality condition:** Intuitively, this implies that Z is determined like in an experiment. This guarantees that the instrument is uncorrelated to the error term. Note that this condition is fundamentally untestable and hence needs to be defended argumentatively.
  - Condition 2.2: **Excludability condition:** Intuitively, this implies that the instrument affects the dependent variable **only** through X_1, i.e. through the endogenous regressor. This is very important and requires you to carefully think about channels through which $\mathbf{Z}$ impacts Y **other than** through the endogenous regressor $\mathbf{X_1}$.
  - Condition 2.1 and Condition 2.2 together form the exclusion restriction.
- **Justifiability:**
  - Relevance condition: See 1d). Note we require significant correlation between the instrument and the endogenous regressor **conditional** on the independent regressors.
  - Exclusion restriction:
    * Externality: Fundamentally untestable ! Eligibility to a 401(k) might not be entirely random -> Here we need more institutional background knowledge how this is determined. Likely, not random. Maybe, workers are more likely to be eligible based on sociodemographic factors (age, income, location, etc.)
    * Excludability: It might be that your eligibility status for 401(k) also influences a variable that determines your credit worthiness (e.g. signal effect for better credit scores). This in turn might affect your decision to sign up for an IRA plan as you have more access to funding.

d) Estimate the reduced form for p401k and verify that e401k has significant partial correlation with p401k. Since the reduced form is also a linear probability model, use a heteroskedasticity-robust standard errors (e.g., White S.Es).

- **Solution:** Conceptually, we run the reduced form regression to test if Condition 1 is satisfied. In our exercise, only $p401k$ is subject to the endogeneity issue. All other regressors in equation (1) plus the IV e401k will serve as the RHS variables. Importantly, the reduced form regression includes **both** the exogenous regressors ($\mathbf{X_2}$) and the instrument ($\mathbf{Z}$). The reduced-form regression is:

$$p401k = \underbrace{\pi_0 + \pi_1 e401k}_{\pi_1 \mathbf{Z}} + \underbrace{\pi_2 inc + \pi_3 inc^2 + \pi_4 age + \pi_5 age^2}_{\pi_2 \mathbf{X_2}} + v \quad (2)$$

The second column in Table (1) presents the result for our reduced form regression. π_1 is the estimator of interest. $\pi_1 = 0.689$ and is significantly different from 0 with a white-corrected standard error of 0.008. Note, the rule of thumb is that with a first stage F-statistics < 10, instruments were considered "weak" (Staiger and Stock, 1997). However, Young (2020) provides a more recent discussion which you should consult in case you work with IV yourself. In our case both the coefficient and the F-statistic clearly indicate that the instrument fulfills the relevance condition.

e) Now, estimate the structural equation by instrumental variable (IV) and compare the estimate of π_1 with the OLS estimate.

- **Solution:** Replacing p401k by the fitted value from the reduced regression, the structural equation becomes

$$pira = \beta_0 + \beta_1 \widehat{401k} + \beta_2 inc + \beta_3 inc^2 + \beta_4 age + \beta_5 age^2 + u \quad (3)$$

The IV estimate for p401k is less than half of the OLS estimate, but the standard error of the IV estimate is slightly larger than the OLS standard error. The 95% confidence interval for the IV estimate is between -0.004 and 0.046, which includes zero. Hence, the participation in 401(k) does not significantly affect the participation in IRA, after accounting for the endogeneity problem. The presence of a wider confidence interval is a price we must pay to get a consistent estimator of β_{p401k} when we think p401k is endogenous. Importantly, the R^2 in the second stage is not meaningful. The point of the second stage is not to fit y to X but solely to estimate the coefficient β_1.

Table 1: Regression output Exercise 1

	Dependent variable:		
	pira default	p401k robust	pira default
	(1)	(2)	(3)
p401k	0.054***		
	(0.010)		
e401k		0.689***	
		(0.008)	
hatp401k			0.021
			(0.013)
inc	0.009***	0.001***	0.009***
	(0.001)	(0.0003)	(0.001)
incsq	−0.00002***	0.00000	−0.00002***
	(0.00000)	(0.00000)	(0.00000)
age	−0.002	−0.005**	−0.001
	(0.003)	(0.002)	(0.003)
agesq	0.0001***	0.0001**	0.0001***
	(0.00004)	(0.00003)	(0.00004)
Constant	−0.198***	0.059	−0.207***
	(0.069)	(0.046)	(0.069)
Observations	9,275	9,275	9,275
R^2	0.180	0.596	0.177
Adjusted R^2	0.180	0.596	0.177
Residual Std. Error (df = 9269)	0.394	0.284	0.395
F Statistic (df = 5; 9269)	406.851***	2,738.415***	399.846***
Note:			*p<0.1; **p<0.05; ***p<0.01

Exercise 2

a) Estimate a probit model of approve on white. Find the estimated probability of loan approval for both whites and nonwhites.

- **Solution:** Estimate the following regression via a Probit model:

$$approval = \beta_0 + \beta_1 white + u \tag{4}$$

where both *approval* and *white* are dummy variables.

The first column in Table (2) lists the probit estimates of approval on white. The coefficient on white is significantly positive, which implies a significant difference in the loan approval rates among the white and non-white group. Denote $\hat{y_1} = \hat{\beta_0} + \hat{\beta_1} * 1 = 0.547 + 0.784 = 1.331$. The probit model assumes that $\hat{y_1}$ and $\hat{y_0}$ are normally distributed and maps them to the corresponding probabilities. The probability of a loan approval for whites therefore is $\Phi(\hat{y_1}) = \Phi(1.331) = 0.91$ and for non-white is $\Phi(\hat{y_0}) = \Phi(0.547) = 0.71$. The partial effect of white on loan approval is therefore 20%, i.e. it is 20 percentage points more likely for whites to get a loan than for non-whites. Clearly, it has to be considered that at this stage we do NOT control for any other factors that might influence loan approval rates.

b) Estimate a simple linear probability model of approve on white. How do these estimates compare to the ones in point (a)?

- **Solution:** The second column in Table (2) presents the OLS estimates for the same regression. The estimate on white is 0.201, which means whites on average and ceteris paribus have a 20 percentage points higher loan approval rate than non-whites. The constant term means for the non-white group, only 71% of their total loan applications are approved on average. Interestingly, as is shown in Table (3), the estimated partial effects for white on loan approval are the same, regardless of the model we are using. The convenient thing about linear probability models is that we can directly interpret the coefficients as probabilities, rather than going through the hassle of evaluating the coefficients at the CDF of the normal distribution.

c) Now, add the variables *hrat, obrat, loanprc, unem, male, married, dep, sch, cosign, chist, pubrec, mortlat1, mortlat2,* and *vr* to the probit model. Re-estimate the model. Is there statistically significant evidence of discrimination against nonwhites now?

- **Solution:** The first column in Table 4 shows the estimates of the probit regression. The estimates on white is 0.520 and is still statistically

significantly different from zero. In other words, the evidence of discrimination against nonwhites still exists after controlling for several covariates.

d) Estimate the model in part (c) by logit. Compare the coefficient on white to the probit estimate.

- **Solution:** The second column in Table 4 shows the estimates of the logit regression. The evidence of discrimination against nonwhites is also significantly different from 0 if we assume that loan approval rates follow a logistic distribution. Note, however, that the estimate on *white* does not directly reflect the partial effect of white on loan approval. As white is a binary variable, we hold all other variables fixed at their respective mean level and compare the estimated partial effect of changing white from 0 to 1. Table (5) provides the estimated partial effects for the different models. Holding the other variables at the sample average level, an average person has a 11 percentage point (Probit), 10 percentage point (Logit) higher approval rate if he/she is white under the Probit (Logit) model. The estimated partial effect from the linear probability model, is at 13 percentage point higher than the Probit and Logit estimates.

Table 2

	Dependent variable:	
	approve	
	probit	*OLS*
	(1)	(2)
white	0.784^{***}	0.201^{***}
	(0.087)	(0.020)
Constant	0.547^{***}	0.708^{***}
	(0.075)	(0.018)
Observations	1,989	1,989
R^2		0.049
Adjusted R^2		0.048
Log Likelihood	−700.877	
Akaike Inf. Crit.	1,405.755	
Residual Std. Error		0.320 (df = 1987)
F Statistic		102.226^{***} (df = 1; 1987)
Note:		*p<0.1; **p<0.05; ***p<0.01

	nonwhites	whites
Probit	0.71	0.91
OLS	0.71	0.91

Table 3: Estimated Loan Approval Probability

Table 4: Regression output Probit/Logit with additional covariates

	Dependent variable:	
	approve	
	probit	*logistic*
	(1)	(2)
white	0.520^{***} (0.097)	0.938^{***} (0.173)
hrat	0.008 (0.007)	0.013 (0.013)
obrat	-0.028^{***} (0.006)	-0.053^{***} (0.011)
loanprc	-1.012^{***} (0.240)	-1.905^{***} (0.460)
unem	-0.037^{**} (0.018)	-0.067^{**} (0.033)
male	-0.037 (0.110)	-0.066 (0.206)
married	0.266^{***} (0.095)	0.503^{***} (0.178)
dep	-0.050 (0.039)	-0.091 (0.073)
sch	0.015 (0.095)	0.041 (0.178)
cosign	0.086 (0.241)	0.132 (0.446)
chist	0.585^{***} (0.096)	1.067^{***} (0.171)
pubrec	-0.779^{***} (0.127)	-1.341^{***} (0.217)
mortlat1	-0.188 (0.257)	-0.310 (0.464)
mortlat2	-0.494 (0.326)	-0.895 (0.569)
vr	-0.201^{**} (0.081)	-0.350^{**} (0.154)
Constant	2.062^{***} (0.316)	3.802^{***} (0.595)
Observations	1,971	1,971
Log Likelihood	-600.271	-600.496
Akaike Inf. Crit.	1,232.542	1,232.992
Note:		*p$<$0.1; **p$<$0.05; ***p$<$0.01

	Probit	Logit	OLS
white	0.11	0.10	0.13

Table 5: Partial effects

Exercise 3

Two researchers are asked to perform a time series analysis of the equal-weighted index return series (EWRET) from January 2, 1997 to November 30, 2011, denoted x_t. They determine the appropriate model by using an estimation of the information criteria for each ARMA model order from (0,0) to (5,5). Researcher A uses Akaike's information criterion (AIC), Researcher B uses Schwarz's Bayesian information criterion (BIC). Obtain the data from the sheet "stock_daily" of the file "data_assignment2.xlsx" and compute the log (or continuously compounded) return series.

a) Plot the sample ACF and the sample PACF, up to 22 lags, of the log return serie and comment on their patterns.

   - As is shown in Figure (1), both ACF and PACF die out gradually from the first to the fourth lag. It is highly possible that both AR and MA components exist in the true model, with an order less than 4. Note that in case we had a pure AR(p) process we would expect the PACF to to be insignificant from 0 for all lags j where $j \geq p + 1$. This, however doesn't seem to be the case. Similarly, for a pure MA(q) process we would want to see the ACF to be cut off (indifferent from 0) for all lags j where $j \geq q + 1$. As neither seems to be the case, it is most likely to have an ARMA(p,q) model.

b) Find the optimal model estimated by each researcher, produce parameter estimates with their standard errors and check the fitted model.

   - Both researchers will chose ARMA(3,3) as their optimal model. Please see the attached code to learn how to select the best order. Table (6) lists the estimates with corresponding standard errors.

     In order to verify the obtained model, one could fit a Ljung-Box test in order to determine whether the null hypothesis of independence in the residuals can or cannot be rejected at a 1% significance level.

c) Use the fitted ARMA models to compute 1-step to 21-step ahead forecasts at the forecast origin November 30, 2011. Show the prediction plots including the two standard error limits of the forecasts and the actual observed values of the (EWRET) return series. Compare and comment on the forecasting accuracy of the models using the MSE and the MAE of the forecasts.

   - Figure 2 shows the prediction plots as required. Denote A_t and F_t as the true value and the forecasting values respectively. Then the accuracy statistics can be computed via the formulas listed below. Table 6 lists the forecasting accuracy of ARMA(3,3) using MSE, MAE, MAPE and SMAPE of the forecasts. Note that for MSE and MAE only the relative ordering matters. This is because the statistics are unbounded

from above and hence little can be learned from their absolute size. MAPE is the mean absolute percentage deviation, and is as high as 111.32% in the given case. This implies quite a high forecasting error. SMAPE, in comparison, has both a lower bound and an upper bound (between 0% and 100 %). SMAPE approaches 100% when F_t, i.e. the forecasting values, are too high or too low. SMAPE = 80.75% therefore implies a large forecasting error. Obviously, given that both researcher choose exactly the same model (ARMA(3,3)) the forecasting accuracy of the models is identical.

$$MSE = \frac{1}{n}\sum_{t=1}^{n}|A_t - F_t|^2 \tag{5}$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|A_t - F_t| \tag{6}$$

$$MAPE = \frac{100}{n}\sum_{t=1}^{n}|\frac{A_t - F_t}{A_t}|^2 SMPE \qquad = \frac{100}{n}\sum_{t=1}^{n}\frac{|A_t - F_t|}{|A_t| + |F_t|} \tag{7}$$

Table 6: Forecast accuracy metrics

	MSE	MAE	MAPE	SMAPE
1	0.0001	0.01	111.32	80.75

x

d) Keeping the residuals series from the estimations performed in (b) by Researcher B, form the absolute residuals and squared residuals series.

i) Compute the variance, skewness, excess kurtosis, minimum and maximum of the absolute residuals and squared residuals series.

- For output refer to Table 6

Table 7: Residual Statistics of absolute and squared residuals

	resAbs	resSq
Variance	0.00	0.00
Skewness	3.14	11.26
Kurtosis	16.57	209.43
Minimum	0.00	0.00
Maximum	0.11	0.01

ii) Plot the autocorrelogram (ACF) and the partial autocorrelogram (PACF) of the absolute residuals and squared residuals series.

- For output, please refer to Figure 3

iii) Test whether the squared residuals are stationary or have a unit root.

- Use adf.test() to test if the squared residuals are stationary. Alternatively, use the ur.df() function which allows for slightly more flexibility. The test result can be seen in Table 9. The ADF test rejects the null hypothesis of non-stationarity at the 1% significance level.

iv) Build an ARMA model for the squared residuals series and check the fitted model (for this part, use the BIC criterion).

- Auto.arima() finds an optimal order (2,1,10) to fit the squared residuals, under the BIC criterion (compare Table 10). **Note: I realized that this depends on your specifications of the maximum lag sizes. In alternative specifications, I find the optimal model to be ARMA(5,5).** Figure 4 is a diagnostic graph for the fitted model. The residual of the fitted model passes the Ljung-Box test (recall: H_0 is that the data is independently distributed. Hence, the high p-values indicate that we cannot reject H_0 at any standard significance level). However, the squared residuals do not seem to be white noise. Rather, the time series exhibits relatively strong evidence of heteroskedasticity across time (compare the first subplot in Figure 4). So you may further want to test your model for this hypothesis and potentially modify it accordingly, for instance by introducing a GARCH term.

Table 8: Regression output of optimal ARMA(3,3) model

	Dependent variable:
	r
ar1	-0.029
	(0.103)
ar2	-0.468^{***}
	(0.078)
ar3	0.689^{***}
	(0.102)
ma1	0.113
	(0.114)
ma2	0.513^{***}
	(0.088)
ma3	-0.595^{***}
	(0.113)
intercept	0.001^{***}
	(0.0002)
Observations	3,755
Log Likelihood	11,508.310
σ^2	0.0001
Akaike Inf. Crit.	$-23{,}000.620$
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 1: ACF and PACF of the log return series

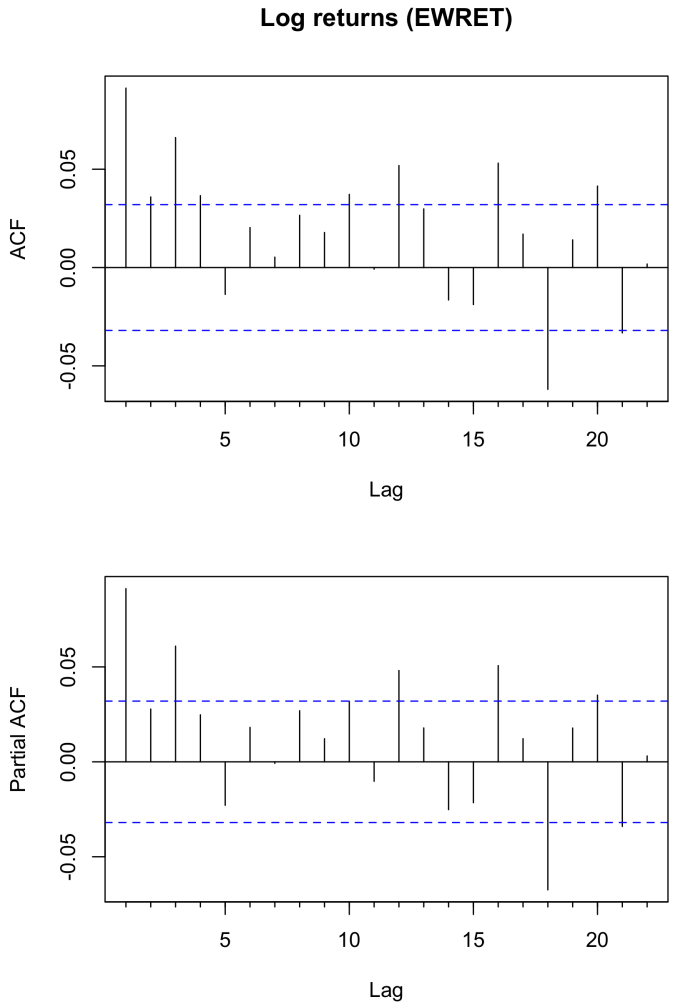


Table 9: t-statistics of Dickey-Fuller Test from different functions

	t_adf.test	t_ur.df
Dickey-Fuller	-8.10	-8.10

Figure 2: Prediction Plot (ARMA(3,3) - 21-step ahead forecast)

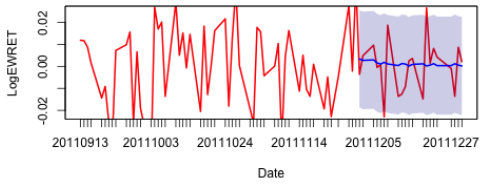


Figure 3: ACF and PACF of the absolute residuals and squared residuals

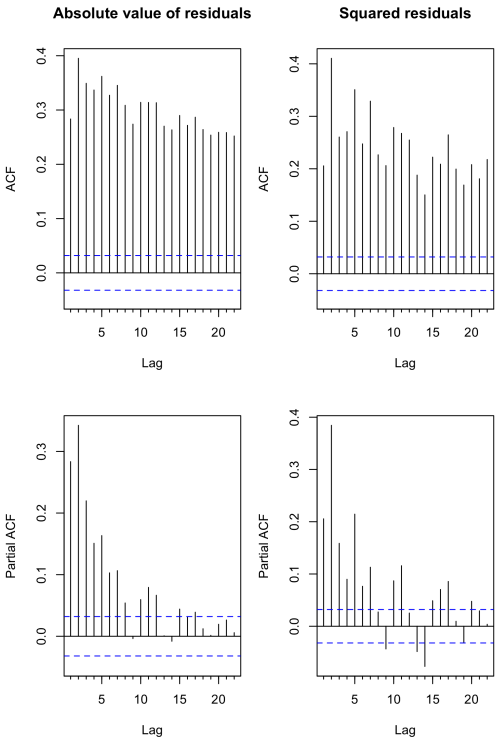


Figure 4: Diagnostic Graph

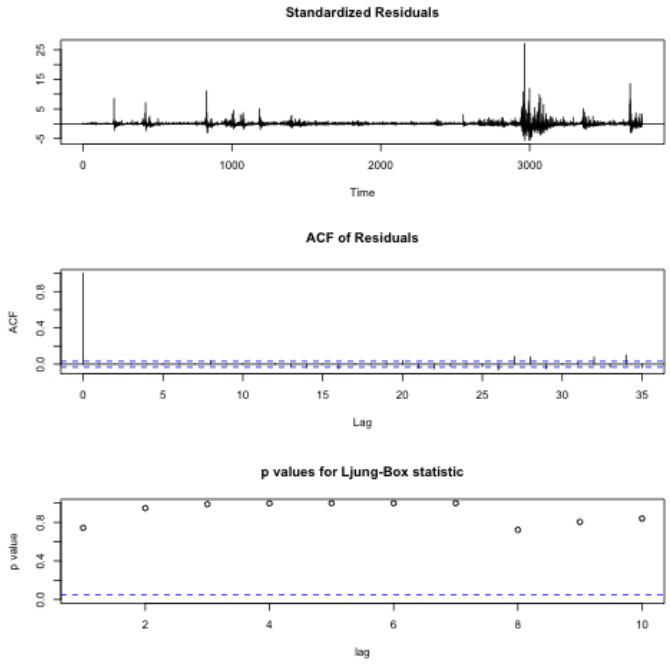


Table 10: Regression output of optimal ARMA(2,10) model

	Dependent variable:
	resSq
ar1	0.869^{***}
	(0.008)
ar2	-0.968^{***}
	(0.007)
ma1	-1.893^{***}
	(0.018)
ma2	2.178^{***}
	(0.036)
ma3	-1.482^{***}
	(0.050)
ma4	0.526^{***}
	(0.054)
ma5	-0.188^{***}
	(0.054)
ma6	-0.111^{**}
	(0.054)
ma7	0.287^{***}
	(0.054)
ma8	-0.407^{***}
	(0.048)
ma9	0.328^{***}
	(0.035)
ma10	-0.144^{***}
	(0.017)
Observations	3,754
Log Likelihood	24,691.260
σ^2	0.00000
Akaike Inf. Crit.	$-49,356.520$
Note:	*p<0.1; **p<0.05; ***p<0.01

Exercise 4

Consider the monthly U.S. 1-year and 5-year Treasury rates from January 1980 to December 2016. Data can be found in the "rates" sheet of the file "data_assignment2.xlsx". Define the vector $x_t = \begin{pmatrix} x_{1,t} \\ x_{2,t} \end{pmatrix}$ where $x_{1,t}$ is the 1-year Treasury rate and $x_{2,t}$ is the 5-year Treasury rate.

1. Estimate a VAR(2) model for the bivariate interest rate series. Write down the fitted model.

    - **Solution:** Table 11 summarizes the VAR(2) estimates for the bivariate interest rate series. Only the lagged 1-year Treasury rate has a significant estimate, yet only at the 10 % significance level. Hence, this indicates that neither within a series nor across a series there seems to be serial dependence. **Note: I stationarized the series first by first-differencing. If you didn't do this, check my code for the results with not stationarized series to compare it to your results.**

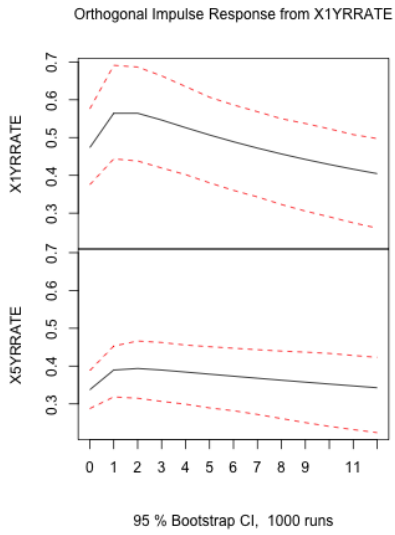
2. Compute and plot the impulse response function of the fitted VAR model. You should compute 12-periods impulse responses.

    - Recall that Impulse responses trace out the responsiveness of the dependent variables in the VAR to shocks to the error term. A unit shock is applied to each variable and its effects are noted
    - **Solution:** The left panel shows the response of an one unit shock in the 1-year treasury rate. The shock is directly reflected in the contemporaneous period of the 1-year rate. This positive shock affects the future 1-year rate for only up to 1 month. There is no significant effect on the 5-year treasury rate. When there is a one-unit shock in the 5-year rate (as shown in the right panel), the 1-year rate will temporarily increase by a small amount, yet not significantly different from zero. Meanwhile, the positive shock in the 5-year rate will persist for only one period and then dies quickly out.

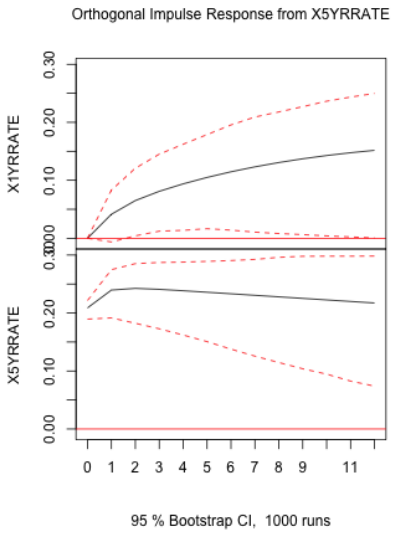
Table 11: Regression model of the VAR(2) model

	Dependent variable:	
	y	
	(1)	(2)
X1YRRATE.l1	0.158*	0.051
	(0.092)	(0.077)
X5YRRATE.l1	0.134	0.149
	(0.111)	(0.092)
X1YRRATE.l2	−0.138	−0.033
	(0.091)	(0.075)
X5YRRATE.l2	−0.019	−0.088
	(0.111)	(0.092)
const	−0.022	−0.018
	(0.024)	(0.020)
Observations	442	442
R^2	0.069	0.045
Adjusted R^2	0.061	0.036
Residual Std. Error (df = 437)	0.496	0.411
F Statistic (df = 4; 437)	8.115***	5.155***
Note:	*p<0.1; **p<0.05; ***p<0.01	

Figure 5: Impulse Response Functions



(a) 1-year Treasury rate



(b) 5-year Treasury rate