STA457/STA2202 - Assignment 3

Submission instructions:

Submit *two separate files* to A3 on Quercus - the deadline is 11:59PM on Monday, June 15.
- A PDF file with your Theory part answers.
- A PDF file with your Practice part report (w/ code in RMarkdown chunks or in Appendix).

Theory

1. Consider the causal representation of a VAR(p) model

$$\boldsymbol{X}_t = \overbrace{\sum_{j=1}^{p} \boldsymbol{\Phi}_j \boldsymbol{X}_{t-j} + \boldsymbol{W}_t}^{\text{VAR}(p)} = \overbrace{\sum_{j\geq 0} \boldsymbol{\Psi}_j \boldsymbol{W}_{t-j}}^{\text{causal repr.}}$$

for causal weight matrices $\{\boldsymbol{\Psi}_j\}$ and $\boldsymbol{W}_t \sim \text{WN}(\boldsymbol{0}, \boldsymbol{\Sigma}_w)$.

a. [3 marks] Prove equation (5.95) on SS p.280, which gives the series auto-covariance matrix at lag $h > 0$ as

$$\boldsymbol{\Gamma}(h) = \text{Cov}(\boldsymbol{X}_{t+h}, \boldsymbol{X}_t) = \sum_{j\geq 0} \boldsymbol{\Psi}_{j+h} \boldsymbol{\Sigma}_w \boldsymbol{\Psi}_j'$$

b. [2 marks] Show that $\boldsymbol{\Gamma}(h) = \boldsymbol{\Gamma}'(-h)$ for $h \geq 0$.

2. Consider the following bivariate time series model:

$$\begin{cases} X_{1,t} = .5X_{1,t-1} + U_t \\ X_{2,t} = .5X_{2,t-2} + U_t + V_t \end{cases}$$

where U_t, V_t are *independent* WN$(0, 1)$ sequences. Note that X_1 marginally follows AR(1) and X_2 marginally follows AR(2).

a. [2 marks] Write the model as a bivariate VAR(p) model of the form

$$\boldsymbol{X}_t = \sum_{j=1}^{p} \boldsymbol{\Phi}_j \boldsymbol{X}_{t-j} + \boldsymbol{W}_t$$

where $\boldsymbol{W}_t \sim \text{WN}(\boldsymbol{0}, \boldsymbol{\Sigma}_w)$ for some variance-covariance matrix $\boldsymbol{\Sigma}_w$. Specify the values of the parameters $(\{\boldsymbol{\Phi}_j\}, \boldsymbol{\Sigma}_w)$

b. [4 marks] Find a closed form expression for the causal weight matrices $\{\boldsymbol{\Psi}_j\}_{j\geq 1}$, from the model's causal representation $\boldsymbol{X}_t = \sum_{j\geq 0} \boldsymbol{\Psi}_j \boldsymbol{W}_{t-j}$.
(*Hint*: you can use the recurrence equation $\boldsymbol{\Psi}_k = \sum_{j=1}^{\min\{p,k\}} \boldsymbol{\Psi}_{k-j} \boldsymbol{\Phi}_j$, where $\boldsymbol{\Psi}_0 = \boldsymbol{I}$)

c. [4 marks] Find the cross-covariance function $\gamma_{12}(h) = \text{Cov}(X_{1,t+h}, X_{2,t})$ for any h.

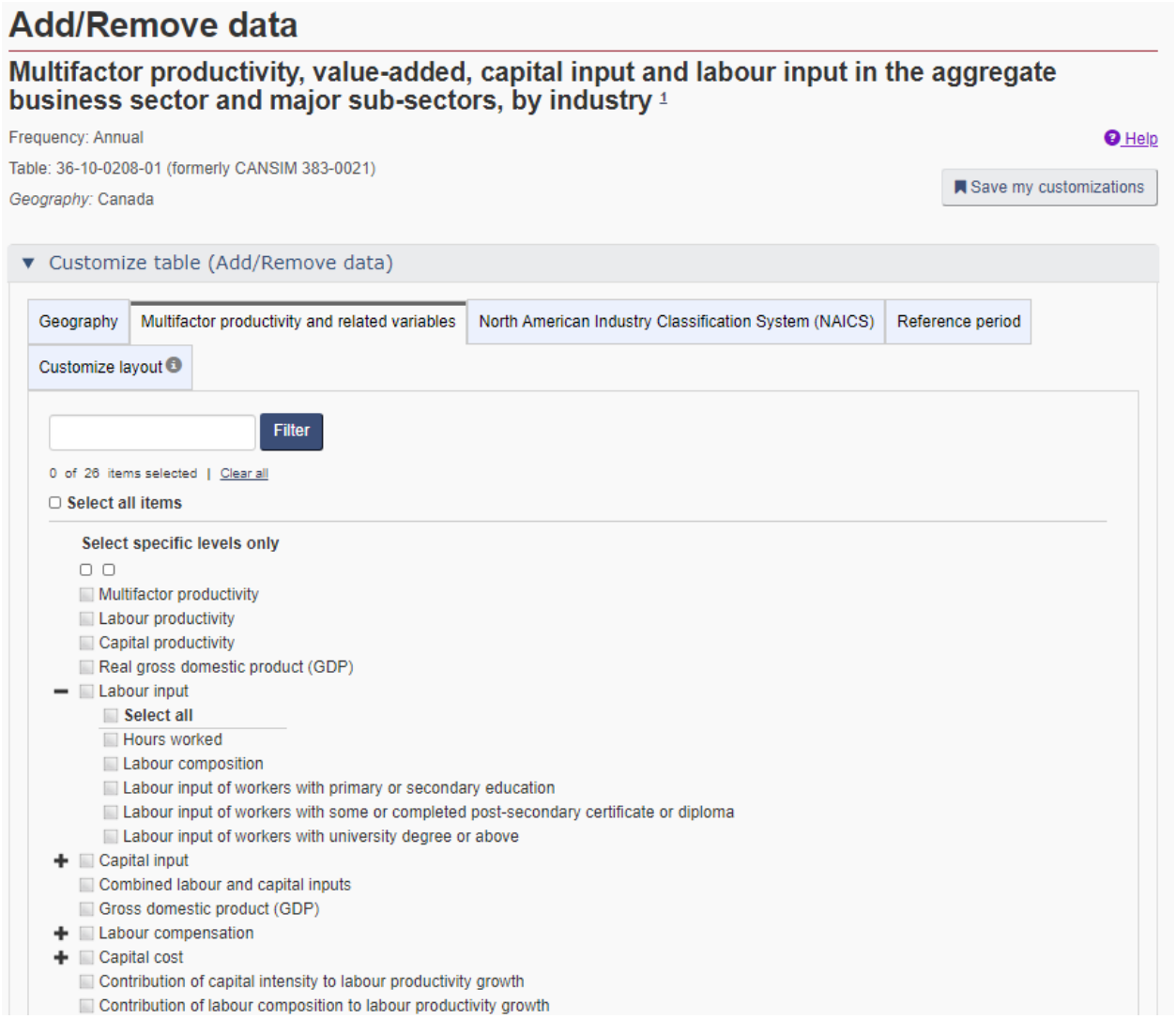
Practice

For this part you will work on this year's Statistics Canada: Business Data Scientist Challenge. The goal of this challenge is to create timely estimates of current GDP based on other, more readily available information; this problem is referred to as *nowcasting*. Each student will work on one of 20 different industry/sector groups as follows:

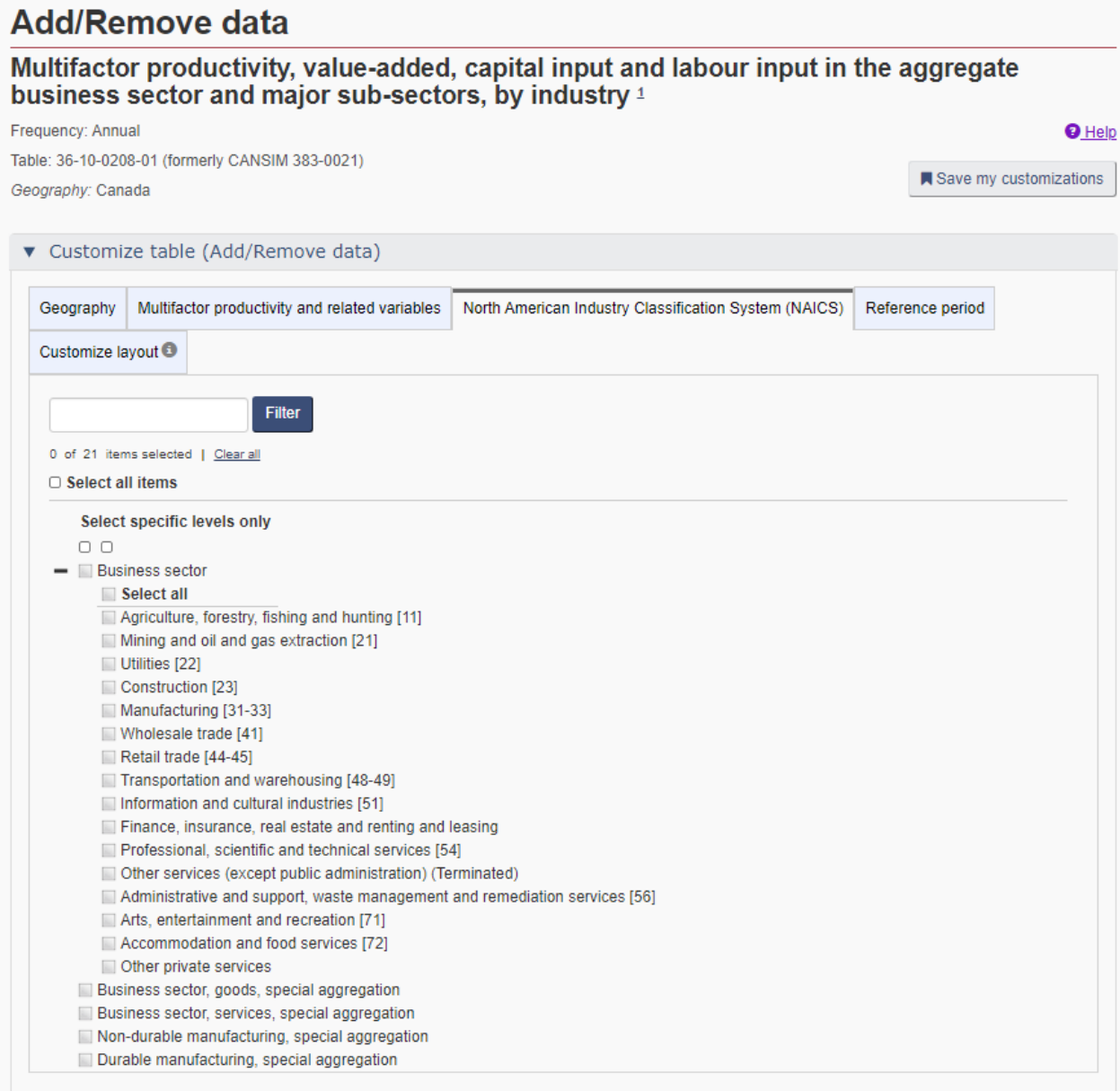
Sector/Industry Group	Last 2 digits of student #
Agriculture, forestry, fishing and hunting	00-04
Mining and oil and gas extraction	05-09
Utilities	10-14
Construction	15-19
Manufacturing	20-24
Wholesale trade	25-29
Retail trade	30-34
Transportation and warehousing	35-39
Information and cultural industries	40-44
Finance, insurance, real estate and renting and leasing	45-49
Professional, scientific and technical services	50-54
Other services (except public administration) (Terminated)	55-59
Administrative and support, waste management and remediation services	60-64
Arts, entertainment and recreation	65-69
Accommodation and food services	70-74
Other private services	75-79
Business sector, goods, special aggregation	80-84
Business sector, services, special aggregation	85-89
Non-durable manufacturing, special aggregation	90-94
Durable manufacturing, special aggregation	95-99

Data

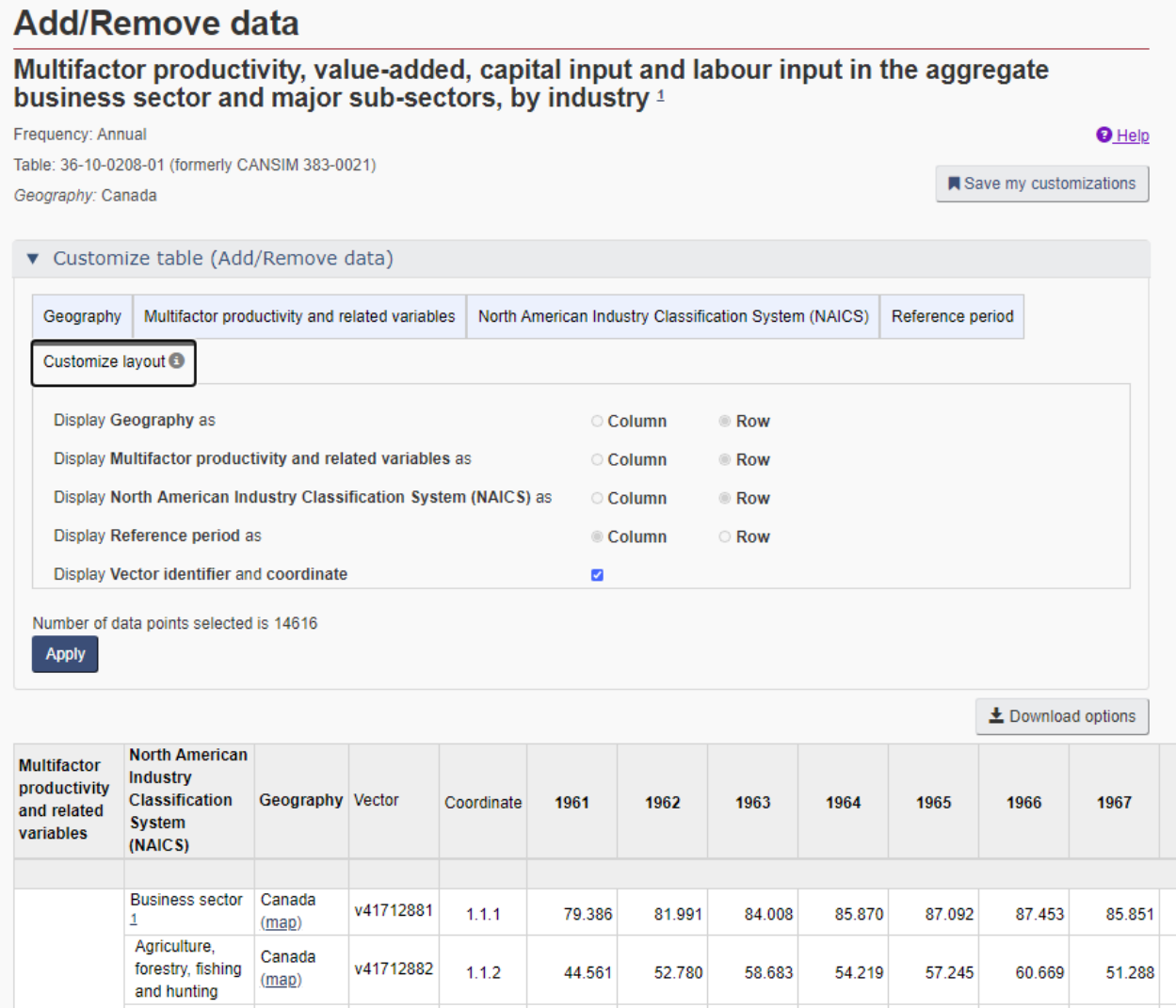
The data are given in StatCan Table: 36-10-0208-01 called *"Multifactor productivity, value-added, capital input and labour input in the aggregate business sector and major sub-sectors, by industry"*. This table contains annual data from 1961-2018 for a range of economic variables listed under the *Add/Remove data* option , as shown below:



You only need to work with data from your own Business Sector/Industry group, selected under the *North American Industry Classification System (NAICS)* tab:



You will notice that the range of values for the variable *Gross Domestic Product (GDP)* is two years **shorter** (ends in 2016) than the other valriables (end in 2018). You can extract table data using R's `cansim` library, as in Assignment 1; you can find each series' *vector* identifier using the *Customize Layout* tab:



Add/Remove data

Multifactor productivity, value-added, capital input and labour input in the aggregate business sector and major sub-sectors, by industry [1]

Frequency: Annual

Table: 36-10-0208-01 (formerly CANSIM 383-0021)

Geography: Canada

Help

Save my customizations

▼ Customize table (Add/Remove data)

Geography | Multifactor productivity and related variables | North American Industry Classification System (NAICS) | Reference period

Customize layout

Display **Geography** as	Column	Row
Display **Multifactor productivity and related variables** as	Column	Row
Display **North American Industry Classification System (NAICS)** as	Column	Row
Display **Reference period** as	Column	Row
Display **Vector identifier** and **coordinate**	☑	

Number of data points selected is 14616

Apply

Download options

Multifactor productivity and related variables	North American Industry Classification System (NAICS)	Geography	Vector	Coordinate	1961	1962	1963	1964	1965	1966	1967
	Business sector [1]	Canada (map)	v41712881	1.1.1	79.386	81.991	84.008	85.870	87.092	87.453	85.851
	Agriculture, forestry, fishing and hunting	Canada (map)	v41712882	1.1.2	44.561	52.780	58.683	54.219	57.245	60.669	51.288

Description

The goal is to fit a model for predicting "current" GDP, call it Y_t, based on current and lagged values of the other variables (e.g. $X_{1,t}, X_{1,t-1}, X_{2,t}$) and possibly lagged values of GDP (Y_{t-1}). For this, you will use VAR and regression with ARMA error models.

Note: Most economic time-series are integrated of order 1, so you might need to difference the data

1. [2 marks] Plot of the (nominal) GDP series and perform an `adf.test` for stationarity. Report the p-value and the conclusion for your series (integrated or stationary).
2. [3 marks] Fit a bivariate VAR(1) model on (nominal) GDP and Real GDP. Do not transform the series, but include both constant and trend term in your model. Report the coefficient matrix and check whether the model is stationary, i.e. its eigen-values are within the unit disk (use functions `eigen` and `Mod`).
3. [2 marks] Plot the residuals and their ACF/CCF from the previous VAR(1) model, and comment on its fit. Report the residual MAPE for (nominal) GDP only.
4. [3 marks] Now fit an ARMA-error regression model for (nominal) GDP (Y_t) with simultaneous Real

GDP (X_t) as the external regressor. Use `forecast::auto.arima` to select the order of the model (including differencing) and report the final model, its AIC and MAPE.

5. [5 marks] Finally, fit an ARMA-error regression model for (nominal) GDP with any of the other variables (Real GDP, Labour/Capital productivity/input/cost, etc.) as external regressors, simultaneous or lagged. Find a model that gives a better AIC than the previous part, or report three different models that you tried with worse AIC. Report the best-AIC model's MAPE and plot its diagnostics, commenting briefly on its fit.

6. [10 marks; **STA2202 (grad) students ONLY**] The in-sample MAPE used above is a biased measure of predictive performance. A better measure is given by using time series cross-validation, as described in chapter 3.4 of fpp2. For this part, you have to evaluate the predictive performance of your previous model using TS cross-validation on the last 10 available GDP values. More specifically, create a loop for $i = 1, \ldots, 10$ and do the following:

- Fit the model specification you chose in the previous part to the data from 1961 to $2006 + i = n_i$.
- Use the model to create a 1-step-ahead forecast for (nominal) GDP, call it $Y^{n_i}_{n_i+1}$; make sure to use the appropriate regressor values for *newxreg*.
- Calculate the percentage error: $|Y_{n_i+1} - Y^{n_i}_{n_i+1}|/Y_{n_i+1}$
  In the end, average the percentage errors over all i and report the resulting MAPE value.
  (Note: this will give you a more objective measure of predictive performance, because you are only using *out-of-sample* 1-step-ahead forecasts.)