COMM1190-Solutions-Week-5-10-and-Exam

Data, Insights and Decisions (University of New South Wales)

Question 1

Explain whether each scenario is a classification or regression problem and indicate whether we are most interested in inference or prediction. Finally, identify the predictors ﷺ and the target variable ﷺ.

a) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.
b) We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.
c) We are interested in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence, we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.
d) Most cameras have the ability to change their settings (ISO, shutter speed, white balance, etc) to improve the quality of the image based on the content (urban, landscape, night, indoors, etc). Unfortunately, this has to be done manually but you would like to create software which will automatically identify the type of scene and allow the camera to automatically adjust these settings. To begin with you'll only consider trying to determine if an image is indoors or not.
e) Given the current and past price movements for a stock, we are interested in determining whether the stock should be bought, held or sold. For this, we have information on the past movements of the stock with data coming very quickly, there being potentially thousands or more trades every second.
f) We have collected fertility data and socio-economic data for all the countries in the world. For each country we record GDP per capita, level of urbanisation, employment rate, average level of education and percentage of migrant population. We are interested in understanding which factors affect a countries fertility rate.
g) A company has a large number of documents that need to be sorted into one of three categories: Research & Development, Finance or Marketing. They have been able to identify a number of phrases which are commonly used in these documents and may help disambiguate them but there are a large number (thousands) of these phrases and each one only appears in a small number of documents. Thankfully someone has labeled a few hundred documents for you, but you need to build a method to automatically label the rest.

Solutions

a) ﷺ: Ceo Salary. This is a continuous variable, so this is a regression problem ﷺ: Record profit, number of employees, industry

Type: Inference as we are more interested in understanding the factors which affect the CEO salary.

b) ��: Success or Failure. This is a discrete variable, so this is a classification problem ��: marketing budget, price charged, competition price
Type: Prediction but we may also be interested in inference if we want to use the model to set prices or marketing budget.

c) ��: the response (% change in US dollar) is continuous. Regression problem ��: % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.
Type: Prediction. It is written in the statement!

d) ��: Indoor or not indoor. This is a discrete variable, so this is a classification problem ��: Content in the image
Type: Prediction. In this case as a phone user, we are just interested in the accuracy of the classification of the image and not on the reasons of the classification.

e) ��: bought, held or sold. Classification as this is a discrete variable. Note that it has 3 possible levels.
��: past movements of the stoc
Type: Prediction, specially if we interested in doing high frequency trading.

f) ��: Fertility rate. Regression as this is a continuous variable
��: GDP per capita, level of urbanisation, employment rate, average level of education and percentage of migrant population.
Type: Inference as we are interested in understanding.

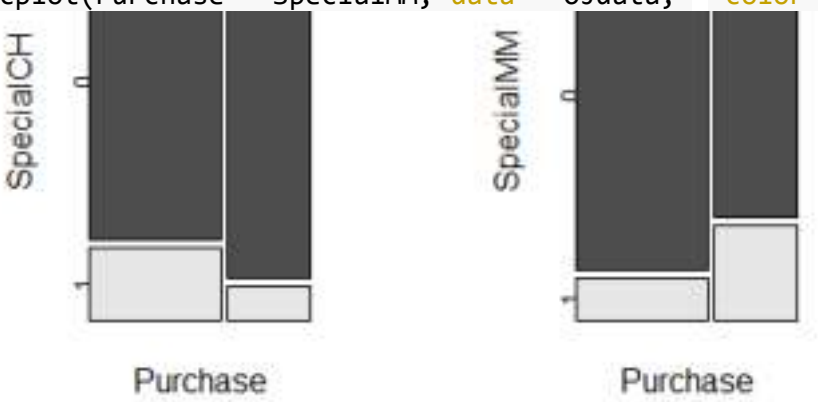
g) ��: Type of document (Research & Development, Finance or Marketing). Classification as this is a discrete variable. Note that it has 3 possible levels.
��: Phrases in the document.
Type: Mainly prediction. In this case we just on to get an accurate classification of the documents

UNSW SYDNEY

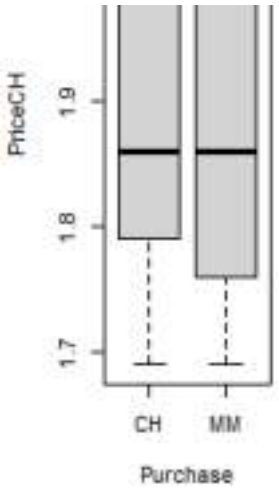
Activity_1.R

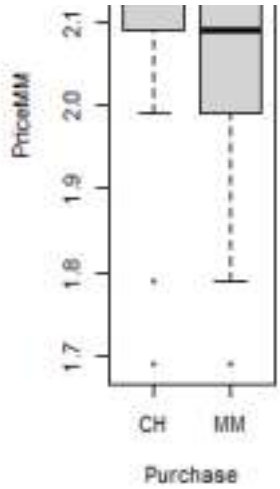
```
OJdata <- read.csv("OrangeJuice.csv")
#Install mosaic package and rplot package
#Purchase vs STORE
mosaicplot(Purchase ~ STORE, data = OJdata, color = TRUE)
```

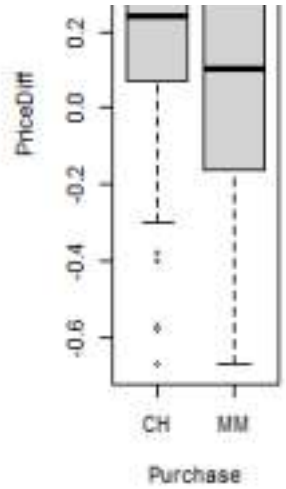

```
par(mfrow=c(1, 2))
#Purchase vs SpecialCH
mosaicplot(Purchase ~ SpecialCH, data = OJdata,  color = TRUE)
#Purchase vs SpecialMM
mosaicplot(Purchase ~ SpecialMM, data = OJdata,  color = TRUE)
```




```
par(mfrow=c(1, 3))
#Purchase vs Price
boxplot(PriceCH ~ Purchase, data = OJdata)
#Purchase vs PriceMM
boxplot(PriceMM ~ Purchase, data = OJdata)
#Purchase vs PriceDiff
boxplot(PriceDiff ~ Purchase, data = OJdata)
```








```
#2 Classification tree
set.seed(1) #Set the seed so that we always get the same results train <- sample(1070, 800)
OJTrain <- OJdata[train, ]
OJTest <- OJdata[-train, ]
dim(OJTrain)

## [1] 800 16

dim(OJTest)

## [1] 270 16

#load rpart
#3 fitting a tree
library(rpart) #Needs to be loaded to be able to fit the tree treeOJ <- rpart(Purchase ~ ., data
= OJTrain)

#load rpart.plot
#4 plotting the tree
library(rpart.plot) #Needs to be loaded to have access to plotting function ## Warning: package

'rpart.plot' was built under R version 4.1.3 rpart.plot(treeOJ)
```

```r
#5 make predictions
OJPredTree <- predict(treeOJ, OJTrain, type = "class")

#Compute the confusion matrix
table(OJPredTree, OJTrain$Purchase)
```

```
##
## OJPredTree CH MM
## CH 439 70
## MM 46 245
```

```r
# Accuracy score of the model
(439+245)/800
```

```
## [1] 0.855
```

```r
#Accuracy in predicting CH
439/(439+46)
```

```
## [1] 0.9051546
```

```r
#Accuracy in predicting MM
245/(70+245)
```

```
## [1] 0.7777778
```

```r
#predict the response on test data
#Make predictions
testOJPredTree <- predict(treeOJ, OJTest, type = "class")

#Compute the confusion matrix
table(testOJPredTree, OJTest$Purchase)
```

```
##
## testOJPredTree CH MM
## CH 154 35
## MM 14 67
```

```r
#Accuracy of the model using test data
(154 + 67)/270
```

```
## [1] 0.8185185
```

```r
#accuracy of predicting CH using test data
154/(154+14)
```

```
## [1] 0.9166667
```

```r
#accuracy of predicting MM using test data 67/(35+67)
```

```
## [1] 0.6568627
```

Activity_2.R

```r
#1. Exploring data
#Install ISLR and Leap packages
library(ISLR) #To have access to the Smarket data
```

```
## Warning: package 'ISLR' was built under R version 4.1.3
```

```r
#Scatter plot
pairs(Smarket)
```

```r
#Correlations
cor(Smarket[,-9]) #We reomve the 9th column as it is not numerical
```

```
## Year Lag1 Lag2 Lag3 Lag4 ## Year 1.00000000 0.029699649 0.030596422 0.033194581 0.035688718 ##
Lag1 0.02969965 1.000000000 -0.026294328 -0.010803402 -0.002985911 ## Lag2 0.03059642
-0.026294328 1.000000000 -0.025896670 -0.010853533 ## Lag3 0.03319458 -0.010803402 -0.025896670
1.000000000 -0.024051036 ## Lag4 0.03568872 -0.002985911 -0.010853533 -0.024051036 1.000000000 ##
Lag5 0.02978799 -0.005674606 -0.003557949 -0.018808338 -0.027083641 ## Volume 0.53900647
0.040909908 -0.043383215 -0.041823686 -0.048414246 ## Today 0.03009523 -0.026155045 -0.010250033
-0.002447647 -0.006899527 ## Lag5 Volume Today
## Year 0.029787995 0.53900647 0.030095229
## Lag1 -0.005674606 0.04090991 -0.026155045
## Lag2 -0.003557949 -0.04338321 -0.010250033
## Lag3 -0.018808338 -0.04182369 -0.002447647
## Lag4 -0.027083641 -0.04841425 -0.006899527
## Lag5 1.000000000 -0.02200231 -0.034860083
## Volume -0.022002315 1.00000000 0.014591823
## Today -0.034860083 0.01459182 1.000000000
plot(Smarket$Volume)
```

```
#Do any other plots/calculations that you find useful par(mfrow = c(1,3))
boxplot(Lag1 ~ Direction, data = Smarket)
boxplot(Lag2 ~ Direction, data = Smarket)
boxplot(Lag3 ~ Direction, data = Smarket)
```

```
boxplot(Lag4 ~ Direction, data = Smarket)
boxplot(Lag5 ~ Direction, data = Smarket)
```

```r
#2 data preparation
SP500data <- Smarket
SP500data$IsUp <- ifelse(SP500data$Direction == "Up", 1, 0)

#3 data split
SP500Train <- SP500data[SP500data$Year != 2005, ]
SP500Test <- SP500data[SP500data$Year == 2005, ]

#4. logistic regression
logisticSP500 <- glm(IsUp ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + Volume,  family = binomial(),
data = SP500Train)  summary(logisticSP500)

##
## Call:
## glm(formula = IsUp ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + Volume,  ## family = binomial(), data
= SP500Train)
##
## Deviance Residuals:
## Min 1Q Median 3Q Max
## -1.302 -1.190 1.079 1.160 1.350
##
## Coefficients:
## Estimate Std. Error z value Pr(>|z|)
## (Intercept) 0.191213 0.333690 0.573 0.567
## Lag1 -0.054178 0.051785 -1.046 0.295
## Lag2 -0.045805 0.051797 -0.884 0.377
## Lag3 0.007200 0.051644 0.139 0.889
## Lag4 0.006441 0.051706 0.125 0.901
## Lag5 -0.004223 0.051138 -0.083 0.934
## Volume -0.116257 0.239618 -0.485 0.628
##
## (Dispersion parameter for binomial family taken to be 1) ##
## Null deviance: 1383.3 on 997 degrees of freedom
## Residual deviance: 1381.1 on 991 degrees of freedom
## AIC: 1395.1
##
## Number of Fisher Scoring iterations: 3

#5. Make predictions
probsSP500 <- predict(logisticSP500, newdata = SP500Train, type = "response")  dim(SP500Train)

## [1] 998 10

probsSPPredLogistic <- rep("Down", 998)
probsSPPredLogistic[probsSP500 > 0.5] <- "Up"
#Confusion matrix
table(probsSPPredLogistic, SP500Train$Direction)

##
## probsSPPredLogistic Down Up
## Down 175 156
## Up 316 351

(175+351)/998

## [1] 0.5270541

175/(175+316)
```

```
## [1] 0.3564155

351/(156+351)

## [1] 0.6923077

#6. Make predictions
probsSP500Test <- predict(logisticSP500, newdata = SP500Test, type = "respons e")
dim(SP500Test)

## [1] 252 10

probsSPredLogisticTest <- rep("Down", 252)
probsSPredLogisticTest[probsSP500Test > 0.5] <- "Up"

#Confusion matrix
table(probsSPredLogisticTest, SP500Test$Direction)

##
## probsSPredLogisticTest Down Up
## Down 77 97
## Up 34 44

(77+44)/252

## [1] 0.4801587

77/(77+34)

## [1] 0.6936937

44/(97+44)

## [1] 0.3120567

#7. Fit model: IsUp ~ Lag1 + Lag2
logisticLag12 <- glm(IsUp ~ Lag1 + Lag2,
 family = binomial(), data = SP500Train)  summary(logisticLag12)
##
## Call:
## glm(formula = IsUp ~ Lag1 + Lag2, family = binomial(), data = SP500Train) ##
## Deviance Residuals:
## Min 1O Median 3O Max
## -1.345 -1.188 1.074 1.164 1.326
##
## Coefficients:
## Estimate Std. Error z value Pr(>|z|)
## (Intercept) 0.03222 0.06338 0.508 0.611
## Lag1 -0.05562 0.05171 -1.076 0.282
## Lag2 -0.04449 0.05166 -0.861 0.389
##
## (Dispersion parameter for binomial family taken to be 1) ##
## Null deviance: 1383.3 on 997 degrees of freedom
## Residual deviance: 1381.4 on 995 degrees of freedom
## AIC: 1387.4
##
## Number of Fisher Scoring iterations: 3

# Confusion matrix in train data
#Make predictions

probsLag12 <- predict(logisticLag12, newdata = SP500Train, type = "response")
probsLag12PredLogistic <- rep("Down", 998)
probsLag12PredLogistic[probsLag12 > 0.5] <- "Up"

#Confusion matrix
table(probsLag12PredLogistic, SP500Train$Direction)

##
## probsLag12PredLogistic Down Up
## Down 168 160
## Up 323 347

(168+347)/998

## [1] 0.5160321
```

```
168/(168+323)

## [1] 0.3421589

347/(160+347)

## [1] 0.6844181

# Confusion matrix in test data
#Make predictions
probsLag12Test <- predict(logisticLag12, newdata = SP500Test, type = "respons
e")
probsLag12PredLogisticTest <- rep("Down", 252) probsLag12PredLogisticTest[probsLag12Test > 0.5]
<- "Up"

#Confusion matrix
table(probsLag12PredLogisticTest, SP500Test$Direction)

##
## probsLag12PredLogisticTest Down Up
## Down 35 35
## Up 76 106

(35+106)/252

## [1] 0.5595238

35/(35+76)

## [1] 0.3153153

106/(35+106)

## [1] 0.751773
```

Week 7 Workshop: Research Design and Experiments I

• *Read over the lecture notes thoroughly.*

R

• One problem requires R using data file *women.csv*.

Problem Set (these will be discussed in tutorial classes)

Q1. *Critically evaluate the following experiment conducted by a bank that wanted to investigate the preferences of customers towards two features of their credit cards: the annual fee charged, and the annual percentage rate charged. The bank selected a large sample of people at random from a mailing list and randomly sent half the sample offers with a low rate and no card fee. While the other half of the sample received offers with a higher rate and a $50 annual fee. Identify the key problem with this experiment and explain how you would avoid it by redesigning the experiment?*

• *There are two attributes of interest here that potentially impact credit card choice and two "levels" (high/low) have been chosen for each attribute. So, there are 4 possible credit cards that could be offered by manipulating these attributes.*

• *Here the experimental design asks customers to choose between 2 of these hypothetical credit cards, one with a low rate and low fee and the other a high rate and high fee. It would be no surprise to find customers choosing the first option but then we wouldn't know whether it's the low fee and/or the low rate that is more important.*

• *A redesigned experiment needs to control for the confoundment problem. You would randomly assign customers to one of the four possible cards. Now we could compare acceptance rates for the low-low versus low-high card with the acceptance rates for the low-low versus high-low.*

Q2. *An energy company is considering a change to its pricing policy involving the introduction of time-of-day pricing whereby customers are charged more for electricity use at peak times of the day (e.g. 6-8pm) in an attempt to smooth out the high demand periods. Before they implement the policy, they decide to conduct an experiment to determine the impact of the change on electricity demand. They chose several regions of the state to be treated (charged higher prices at peak times and off-peak prices remaining unchanged) while in other areas all prices remained constant. Because management was concerned about possible customer complaints, they decided not to reveal to customers that the changes were being made.*

a) What are the problems with this as an experimental design?

• *First you worry about the choice of regions and whether there was a good matching of treatment and control regions. Think of NSW where the climate (main determinant of electricity demand) can vary quite a lot. You do not want all treatment regions concentrated on the North Coast say.*

• *While we have discussed experiments where the participants didn't know they were in an experiment e.g. in Q1 customers didn't know that their offer is different from others but they do know the form of their offer. Here the treatment is a price change, but customers don't know the price has changed (well not until they get their next bill by which time it is too late to change behaviour.)*

b) How might you overcome these problems using a different design?

• *Clearly make sure you have treatment and control regions that on average are similar.* • *One way is to enlist customers into the treatment group and say that based on your bills for the experimental period in the past, if you do not change behaviour over the course of the experiment then your overall bill would stay the same (you would need lower off-peak rates to compensate for higher peak rates). Thus, they can only benefit from the change by shifting when they use their electrical appliances.*

Q3. *Lewis & Reiley (LR) (2014) contrast their experimental approach to estimating the impact of online ads to that proposed by Abraham (2008) described as:*

> *"Measuring the online sales impact of an online ad … in which a company pays to have its link appear at the top of the page of search results – is straightforward: We determine who has viewed the ad, then compare online purchases made by those who have not seen it."*

a) What is the problem with the alternative research design of Abraham (2008) and how do LR avoid the problem?

• *There is a clear self-selection problem. Those who click on and view an ad are more likely to purchase the product.*

• *In the experimental approach customers are randomly assigned to whether they see the ad or not. Treatment is randomly assigned and not determined by the customer. Thus, any difference in sales between the treatment and control group can reasonably be attributed to the ad as all other contributing factors are not related to whether anyone is treated or not.*

b) LR present the following Table of summary statistics comparing the control and treatment groups of customers. What does this show and why is this important in terms of reporting results from a field experiment?

• *In terms of respondent characteristics (female) and internet activity (Yahoo! page views) there is almost no difference between the control and treatment groups. Such comparisons confirm the success of the random assignment.*

• *The other measures describe the exposure of the treatment group to the ads and that the control group had zero exposure.*

• *Notice that the treatment did not ensure everyone in the treatment group saw the ads just that if they did visit Yahoo! the ad would be there.*

c) LR note that the difference in mean number of Yahoo! page views, 363 versus 358, was statistically significant. In other words, the 95% confidence interval for the difference does not include zero. Does this have any impact on your answer to (b)?

• *It remains the case that the actual size of the difference is small. We now know that because of the large sample size this difference is precisely estimated and turns out to be statistically significant.* • *In LR they do note that there were a few customers with excessively large numbers of views and that these happened to be in the treatment group. After accounting for these outliers, the significance of the difference disappears.*

d) The initial baseline estimate provided by LR is the difference in mean sales of the treatment and control groups during the experiment. Because only 63.7% of the treatment group actual viewed the experimentally allocated ads, do you agree that this is a conservative estimate of the effect of advertising on sales.

• *Yes, it is conservative in that the difference could have been larger (it could not have been smaller) depending on how those in the treatment group would have responded if they had viewed the ads.*

e) In one of their additional analyses, LR redefine the treatment group to be only those in the treatment group who viewed the ads, while the rest of the treatment group who did not view the ads were added to the control group. What are the advantages and disadvantages of this modelling decision?

• *The advantage is that all of those in the treatment group who were not exposed are treated as controls so that we isolate the effect of treatment on those who received the treatment (in causal inference terminology this is the* **treatment effect on the treated***). Potentially this is a more interesting treatment effect than the initial estimate (***intention to treat effect***).*

• *The problem is that people choose not to be exposed and you worry that they are systematically different from others in the control group with whom they are pooled. LR check this and it seems not to be a worry.*

The following problem involves the use of R.

Q4. *Chattopadhyay and Duflo (2004)* is a highly cited field experiment that investigated the casual effects on policy outcomes of having female politicians in government. In India there was a period when one-third of village council heads were randomly reserved for female politicians. Part of the data from the study are provided as women.csv. The policy was implemented at the GP level of government, so some GPs were treated and others not. In the data the GP was said to be reserved or not and this indicator is in the data as the variable reserved. In the data there*

are 161 GPs and for each GP the study included two randomly selected villages giving a total of 322 village level observations. The variable female indicated whether the village in the GP in fact had a female leader or not

The outcomes of interest represent the number of new or repaired facilities of two types, irrigation, and water. The hypothesis being investigated is that women will support policies that women voters care about more. Previous research had indicated that women tended to complain about water quality while men tended to be more concerned with irrigation problems.

(a) Why would an observational study be problematic in determining whether women politicians promote different policies when in government? (Q7.4 in lectures)

- *In an observational study the presence of female heads would not be randomly assigned. The women would need to choose to run and then be elected. The successful woman candidate could be very different from one willing to be assigned to a GP.*

- *The villages willing to elect a woman may be very different from those not willing.* - *In each of these cases you may be able to find appropriate variables that could explain these differences, but it is not obvious. Some of the key variables maybe things that are difficult to measure.*

- *An even more pragmatic answer is that there may not be enough villages with female heads to draw any meaningful comparisons between the two groups. In the next question we see that in the control group villages that were not subject to the policy intervention only 7.5% had female heads.*

(b) Confirm the analysis provided in lectures that checks whether the policy had been successfully applied by determining whether reserved GPs have female leaders?

- *R output below in the form of a cross-tab confirms that all reserved GPs did have a female head and that of those villages in the control group 7.5% had female heads.*

```
> res.fem.tab <- table(res = women$reserved, fem = women$female) > addmargins(res.fem.tab)

fem
res 0 1 Sum
0 198 16 214
1 0 108 108
Sum 198 124 322

> res.fem.tab[1, 2] / sum(res.fem.tab[1, ])
[1] 0.07476636
```

(c) Confirm the estimates of the ATE of the reservation policy for both irrigation, and water that were provided in lectures. Does your analysis support the hypothesis that women tend to support the preferences of women voters?

- *R output below indicates support for the hypothesis – the estimated increase in projects attributable to the reservation policy was 9.25 for water but for irrigation there was a slight decrease of -0.37.*

```
Call:
lm(formula = water ~ reserved, data = women)

Residuals:
Min 1Q Median 3Q Max
-23.991 -14.738 -7.865 2.262 316.009

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.738 2.286 6.446 4.22e-10 ***
```

reserved 9.252 3.948 2.344 0.0197 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 33.45 on 320 degrees of freedom
Multiple R-squared: 0.01688, Adjusted R-squared: 0.0138
F-statistic: 5.493 on 1 and 320 DF, p-value: 0.0197

> confint(water.women)
2.5 % 97.5 %
(Intercept) 10.240240 19.23640
reserved 1.485608 17.01924
>
> irrigation.women <- lm(irrigation ~ reserved, data=women)
> summary(irrigation.women)

Call:
lm(formula = irrigation ~ reserved, data = women)

Residuals:
Min 1Q Median 3Q Max

-3.388 -3.388 -3.019 -1.019 86.612

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.3879 0.6498 5.214 3.33e-07 ***
reserved -0.3693 1.1220 -0.329 0.742

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.506 on 320 degrees of freedom
Multiple R-squared: 0.0003385, Adjusted R-squared: -0.002785 F-statistic: 0.1084 on 1 and 320 DF, p-value: 0.7422

> confint(irrigation.women)
2.5 % 97.5 %
(Intercept) 2.109436 4.666265
reserved -2.576766 1.838103

* *Chattopadhyay, R. and Duflo, E. (2004), "Women as policy makers: Evidence from a randomized policy experiment in India", Econometrica 72, 1409-1443.*

COMM1190 Data, Insights, and Decisions

Week 8 Tutorials

• *Read over the lecture notes thoroughly.*

R

• One problem requires R to use data file *housing.xls*.

Problem Set (these will be discussed in tutorial classes)

Q1. *A certain school typically has two, year 7 classes that occupy one of two rooms. These rooms have different capacities (20 or 25 students) presenting the opportunity to exploit a natural experiment to investigate the impact of class size on student performance. Suppose you have data from this school's year 7 cohort collected over many years and comprising student grades and whether they were in the small or large class.*

Under what conditions would you expect the difference in year 7 grades between those in the large class relative to those in the small class to best reflect the causal effect of class size on student performance?

• Allocation to classes needs to be random & not by saying: date of birth with the split being into younger & older students; or by year 6 performance into better & poorer performers. • The allocation of teachers needs to be random.

• The performance should be measured on a standardized test that is comparable over time. • The population of students needs to be relatively stable over time again to ensure comparability in results.

Q2. *Jenny Craig is a weight-loss intervention. Their commercials show a photo of some celebrity before and after joining Jenny Craig.*

a) What are the control and treatment in this experiment?

- *This is a (within the person) before and after design. It is the weight of the same person being compared before (control) and after (treatment).*

b) Are you worried about any selection problems? Explain.

- *Jenny Craig has control over who they use in their ads and so are likely to choose only success stories and presumably the more successful the better.*
- *This is a* **sample selection** *problem. We need to distinguish this from the* **endogenous selection** *issue, where people choose whether to be treated or not. This latter selection issue is also relevant as people choose to join Jenny Craig and those that do will typically be motivated to join and to be successful in losing weight.*

TN: *Some students may worry about the sample size, n=1. But there is a history of self experimentation, especially in medicine. Researchers trying out their ideas on themselves. Often with success. So that's not really the main problem, although in general, we do prefer more observations to increase the precision of our estimates. Also in an RCT version of testing whether Jenny Craig works or not, you would want a large sample to ensure that randomization has worked.*

Q3. *Your statistically naïve friend, Denzil, is very interested in the impact on housing prices of being located under the flight path. Given data on sales over a month, the regression of housing price on flightpath (a dummy variable indicating whether the house was under the Kingsford-Smith airport flightpath) provides a difference in means estimate.*

(a) Explain why this estimate would be a poor indication of the impact on sales of being under the flight path.

- *This is a "difference-in-means" regression as the flight path is a binary variable that divides sales according to whether they were under the flight path or not. This is fine as a descriptive tool but not if you want to infer causality.*
- *There are likely to be many other confounding factors here that would be correlated with flightpath and would help explain the price difference.*

(b) After the Western Sydney International Airport at Badgerys Creek Sydney opens, suppose the decision is made to close the east-west runway at the Kingsford-Smith airport leaving just the two north-south runways in operation. This is purely hypothetical and highly unlikely to

happen but suppose it does. Denzil now revisits his original research question, but this time uses two samples of houses that were located under the east-west flight path. But one sample represented sales in a year before the closure of the runway while the second sample represented sales in a year after the runway was closed. Now he runs a regression of the following form:

$$\text{price}_{i} = \beta_0 + \beta_1 \text{after}_{i} + \varepsilon_{i}$$

where price = *housing sale price;* after = 1 *if the house was sold after the runway closure and zero otherwise.*

How do you interpret the regression parameters for this model?

- *This is a "difference-in-means" regression as after is a binary variable that divides sales according to whether they were sold after the closure of the runway or not.*
- *The estimated intercept will be the mean of the sales price for those houses sold in suburbs under the runway and before it was closed. The estimated after parameter will be the difference in means associated with being sold after the closure.*
- *Because the negative features of being under the flightpath (notably aircraft noise) have been removed you would expect house prices of those previously under the flightpath would increase and hence the estimate of* β_1 *to be positive.*

(c) Do you think this is a good research design to address this problem? Explain.

- *As always, the interpretation of* β_1 *relies on the likely unobservables captured in the disturbance and whether they are likely correlated with after.*
- *Here you would worry about what is happening in the housing market in general. You may find a positive estimate of* β_1 *but is that attributable to the reduced aircraft noise or simply that the housing market is active, and all houses are increasing in value?*
- *This is an example of a* **natural experiment.** *In such cases, you worry that the treatment is truly exogenous. For example, if the closure has been mooted for a while, then developers may buy up houses in the affected area and renovate them in anticipation of the expected post-closure appreciation in value. This again would distort the before and after estimate as a reflection of just being under the flight path or not.*

TN: *Explore DiD as a solution. Ask students if they also had two samples of sales (before and after) for houses in control suburbs (i.e. close to the treated suburbs and hence similar but unaffected by aircraft noise either before or after). What could they do?*

If you take the difference in mean sales for the control and treated groups before the closure and compare this to the same difference after, then this difference-in-difference estimate is a better estimate because any autonomous trends in housing prices are accounted for. It does rely on the assumption that these trends are the same in both the control and treated suburbs. The Varian (2016) paper is a reference for this DiD approach.

Q4. *The Australian Federal Government introduced a policy designed to provide an incentive for women to have more children: the so-called baby bonus. In the budget released in May 2004, the Government announced that families whose babies were born on or after July 1, 2004, would receive a sum of $3,000. No bonus was paid for births before this date. The amount of the bonus continued to rise and on July 1, 2008, rose to $5,000.*

Here we only consider an analysis of Australian Bureau of Statistics birth data just before and immediately after the introduction of the new policy on July 1, 2004. In particular, the outcome

variable of interest is the number of daily births in Australia for the period January 1

to July 31 of 2004, which is 6 months before the introduction and one month after. The time series graph below depicts the birth data with $t = 183$ *corresponding to July 1, 2004.*

Suppose the primary objective of the analysis is to determine the impact of the introduction of the baby bonus. Did Australian women (and doctors) respond to the financial incentives and if so by how much? To address these issues, consider the following model:

$$\text{number of births}_t = \beta_1 + \beta_2 \text{weekend}_t + \beta_3 A_t + \varepsilon_t$$

where

$\text{number of births}_t$ *= the number of births in Australia on day t ;*

$\text{weekend}_t = 1$ *if day t is a weekend or a public holiday (=0 otherwise);*

$A_t = 1$ *if day t is after June 30 (=0 otherwise).*

(a) How do you interpret the regression parameters for this model?

- *Again, this is a "difference-in-means" regression as A is a binary variable that divides births according to whether the time is after the introduction of the Baby Bonus (BB) or not. But here there is an added control for weekends and public holidays; a variable that is justified by observing the obvious weekly cycle in the data (doctors prefer not to work on weekends).*
- *Because there are 4 possible regimes, this complicates the interpretation of the parameters:*

$$E(\text{number of births} \mid \text{weekend} = A = 0) = \beta_1$$
$$E(\text{number of births} \mid \text{weekend} = 1, A = 0) = \beta_1 + \beta_2$$
$$E(\text{number of births} \mid \text{weekend} = 0, A = 1) = \beta_1 + \beta_3$$
$$E(\text{number of births} \mid \text{weekend} = 1, A = 1) = \beta_1 + \beta_2 + \beta_3$$

- *Thus,* β_1 *is the mean births on weekdays before the BB.*
- β_2 *is the difference in mean births between weekdays & weekends irrespective of whether it is before or after the BB.*
- β_3 *is the difference in mean births before & after the BB irrespective of whether it is a weekday or weekend.*

(b) Do you think this is a good research design and associated model specification to address this problem? Explain.

- *There are some problems with this design including the likelihood of an interaction effect between the BB and whether the day is a weekday or not. But more fundamentally it "imposes" a once-off permanent change (presumably increase) in births after the BB.*
- *Because of the limited time interval here, every woman who gave birth in this period was already pregnant at the start of the period. The model is inconsistent with the data chosen.* • *With these data what can be identified is a shift in the birthdate of babies that would be born during this period. See Gans & Leigh (2009) for an extensive analysis of these data. They estimate a large "shift" in the* **timing of births** *of over 1000 births that can be attributable to the policy change.*

TN: *Salient message is that the model to be estimated needs to make sense! A useful general modelling lesson is to describe the model in words before doing any estimation and ask is it sensible/reasonable? It can never match reality exactly, but it should capture the main features of the process being modelled – think of models as maps.*

The following problem involves the use of R.

Q5: *Upgrading infrastructure, such as freeways, has the potential to increase property values for impacted dwellings. Here the claim is that a major freeway upgrade has led to an increase in the value of residential houses.*

(a) Explain how you would design an experiment to test this causal claim. Is the experiment you have proposed feasible to conduct?

- *Randomly allocate houses to whether they benefit from the upgrade and compare average selling prices between the two groups.*
- *Upgrade the freeway and observe the selling price of a sample of houses. Now rewind the clock and don't upgrade the freeway and observe the selling prices of the same houses. Attribute any differences in average selling prices to the upgrade.*
- *Clearly, neither is feasible.*

(b)Instead of the experiment outlined in (a), consider investigating the claim by using available observational data contained in housing.xls. This is a sample of house sales over 12 months covering 9 months before and 3 months after the completion of the freeway upgrade. The data includes information on dwelling characteristics including a dummy variable, ⬚, *that indicates whether the house was in an area with ready access to the freeway. Using only the data where* ⬚ $= 1$, *estimate the regression,* ⬚ $= \beta_0 + \beta_1$ ⬚ $+ u$, *where* ⬚ *is a dummy variable that is equal to one when the sale is made in the three months after the completion of the upgrade and zero otherwise. Interpret these results and discuss whether they are consistent with the claim of improved property values.*

⬚$\widehat{\ }$ $= 1212.8$ (24.4) $+122.4$ ⬚ (68.3)

$n = 282$, $R^2 = .011$, ⬚

⬚ ⬚ $(\ .\ )$

- *Houses sold after attracted a selling price that was $122,400 higher and you would reject the null*

hypothesis of no effect at the 5% level with a one-tail test.

- *The results are consistent with the claim of improved property values BUT see next part.*

(c) Identify the potential problems associated with the approach in (b) by contrasting it with the experiment you described in (a).

- *You worry that this estimate reflects, at least in part, other factors. A confoundment problem.* •

Maybe it is just an overall increase in property values irrespective of the upgrade. • *Such problems were avoided by either winding back the clock or random assignments.*

(d) Consider using a difference-in-difference approach using the entire sample to estimate following the regression model. What do you conclude from these results?

$$\text{Price} = \beta_0 + \beta_1 \text{Upgrade} + \beta_2 \text{Distance} + \beta_3 \text{Upgrade} \times \text{Distance} + \varepsilon$$

$$\widehat{\text{Price}} = 1300.7\ (11.1) - 87.9\ \text{Distance}\ (44.9) + 80.3\ \text{Upgrade}\ (29.5) + 42.1\ \text{Upgrade} \times \text{Distance}\ (125.3)$$

$n = 4{,}663$, $R^2 = .003$, (standard errors in parentheses)

- *The DiD estimate of $42,100 is still positive but much smaller than in (b).* • *Moreover, the estimate is very imprecise (CI is wide) and formally you would not reject the hypothesis of no effect at say 5%.* -→ *not statistically significant*

(e) Someone evaluating your results notes that the freeway extension only impacts people living in the outer suburbs where prices are lower. They are concerned that the results in (d) may be biased and your conclusions problematic. Run the following regression to explore this concern. What do you conclude?

$$\text{Price} = \beta_0 + \beta_1 \text{Upgrade} + \beta_2 \text{Distance} + \beta_3 \text{Upgrade} \times \text{Distance} + \beta_4 \text{Outer} + \varepsilon$$

The results from the extended regression model are given below:

$$\widehat{\text{Price}} = 1768.5\ (28.5) + 137.7\ \text{Upgrade}\ (45.3) + 74.6\ \text{Distance}\ (28.5) + 63.5\ \text{Upgrade} \times \text{Distance}\ (121.3) - 52.7\ \text{Outer}\ (3.0)$$

$n = 4{,}663$, $R^2 = .066$, (standard errors in parentheses).

- *It is true that distance from the CBD does imply lower prices (on average $52,700 less for each km ceteris paribus).*

- *This impacts the estimated coefficient on the upgrade, which is now positive and statistically significant, but this is not the parameter of primary interest.*
- *Controlling for distance does not change the qualitative conclusion, the DiD estimate of $63,500 is positive and imprecise.*

TN: *As well as being statistically insignificant, the DiD estimates are relatively small although maybe not for a student. But relative to prices in the millions (see the first regression) these are small differences. It is good to remind students that the magnitude of an effect is often more important than its statistical significance.*

Week 9 Tutorial: Data Ethics (discussion points)

1. **HBR Case: Algorithm bias in marketing**
   https://hbsp.harvard.edu/tu/53b7b99b

   Read the above case study and answer the following questions:

   (a) Analysing the examples:

   Place an example on the 2x2 diagram from the case. Explain why it belongs in that quadrant and would you recommend using algorithms for this decision? Explain why this type of bias prevalent in the case study is a concern.

Data Characteristics		**Incomplete or unrepresentative data?**	
Contain pre-existing biased pattern or correlation between group preferences and outcome		No	Yes
	No	N/A	Facial recognition
	Yes	**Perpetuate bias:** Lending access (yes/no) Ridesharing prices Amazon Prime **Correlations:** STEM Ads on Facebook Insurance prices Recommendations systems	**Perpetuate bias:** Lending specific terms

Benefits of placing the examples in the diagram:

- to reinforce the understanding of the causes of algorithm bias
- to fix discussions around different quadrants

(b) Targeting and personalisation in marketing:

Do targeting and targeting personalisation practices always create bias? What is the marketer's problem in this case study?

- A new store opens and decides to target all customers within a 1-5 miles radius. However, it may have ethical implications if only a particular class of people received the coupons by their proximity.
- Targeting customers by their gender by online department stores imply that if the algorithm indicates that the potential customer is a woman, she will be sold only items such as make-up and dresses. She may not be offered other items.
- Discrimination may be considered an ethical issue in the case study

(c) Detection and prevention:

How would design a customised and personalised offering to mitigate the algorithm bias?

Should algorithms be designed with fairness in mind? If the consumers cannot compare their offers, it should be fine. However, with social media and increased public awareness, consumers may quickly know about the algorithm bias.
Possible measures

- Simulations
- Tests
- Audit
- Team diversity

NB: The above discussion points are based on Harvard Business Review's teaching note.

2. Facebook has been collecting big data on its clients and even generates their respective social graphs using its sophisticated algorithm. Meanwhile, Amazon tracks your purchase history and the ads clicked. The business model of the gigantic social media is to sell targeted ads to commercial businesses. However, Facebook has been giving access to its customers' private data to other companies. It has been also reported false data on ads.
In a 2012 New York Times article, there was a story of a father learning of the pregnancy of his teenage daughter when Target Inc sent coupons for baby items. The question remains whether it was the access to private data or pure coincidence that can lead to customer targeting by Target Inc.
https://www.cmswire.com/digital-marketing/facebook-a-case-study-in-ethics/
(i) What is the ethical dilemma facing the employees at Facebook?

- Employees at Facebook have a duty of care to their employer to generate revenue for the business by selling targeted ads and generating social graphs.
- However, giving access to their customers' private data without their explicit consent is an ethical issue to be concerned with.

(ii) To what extent the ethical implications of Amazon and Facebook's practices are different?

- Amazon collects data but does not disclose your private data to the business. Since their business model includes tracking purchase history and selling targeted ads, they do not reveal their customers' deepest secrets.
- Meanwhile, the Facebook malpractice of revealing private data to other businesses is gross negligence of ethical values.
- Besides, reporting false data is unethical on ads and can be subject to litigation.

(iii) Which of the following ethical approaches -deontological or utilitarianism, best apply to the Facebook case study?

- Deontological ethical approaches set the rules which bind each employee to a duty of care in their task.
- Although implementing detailed rules may help minimise ethical issues, the utilitarian approach works better as employees would ask the right question whether the company's action of false data and disclosing private data can be more harmful to society at large as it can adversely many people.

(iv) If Target Inc's practice is not breaching any laws, why does a business need to care about ethical principles?

- Target Inc has not infringed the laws by targeting customers with promotions and ads. Yet, it may be an ethical case study if the company has used private data in its predictive analytics.

3. A married couple, John and Jane applied for a mortgage loan to buy their first home. They both

earned six-digit salaries with John working as an aircraft engineer and Jane as a dentist. The bank manager feeds their applications into an algorithm, which sourced data from multiple sources beyond what they have disclosed in their application. The data comprise the timeliness in settling utility bills, medical histories, genetic reports, internet sites visited, movies watched, books read, and other information on the ethnic and family background. After a couple of days,
the bank manager informed them of the outcome of their applications as rejected. Upon inquiry, they were told that there is no legal infringement by the bank as the same software is used on all loan applications and that even the software designer does not know the full details of how the data is being processed.

(i) What are the ethical issues adversely impacting the couple?

- Despite their credit score is good, they were penalised for lack of transparency on how the algorithm makes the decision.
- There was also an invasion of data privacy without their consent

(ii) What are the wider societal implications of using a similar algorithm for evaluating loans?

- Although there are no legal ramifications for using the software, there is a violation of the ethical principle of utilitarianism as many successful applicants could find their loans being rejected, possibly causing broader harmful effects on class mobility.

(iii) Outline the main measures that the bank can implement to minimise the harms caused by the loan approval process

- More transparency in how the software processes the application
- The manager can intervene by tweaking the algorithm if it is arbitrary.
  [Obviously, this management's discretion must be approved by an ethical board]
- The applications' consent can be sought before using their private data

Week 10 Workshop: Data Communication

1. Choosing Charts

(a) Pitching to a VC

Imagine that you are pitching an idea to a VC. You want to show what you call “a huge chasm” in the market between products and customers’ access to them. Your solution is “the bridge” connecting customers to products.

i. What kind of chart address your goal? Conceptual? Declarative? How would you classify the chart amongst the four types?

- We are looking for a conceptual chart where you are pitching an idea.

ii. Which of the following sketches might be a good start for visualizing your value proposition? The middle chart is the best as it clearly shows a connector between the two domains. The first figure is not great because it mixes metaphors. While the goal is to convey the idea of a bridge or a connector, a Venn Diagram implies overlap or commonality, which is not the same. The last figure is too creative and over-designed; the detailed decoration distracts from the message.

(b) Forecasts

Your boss has asked you to present the year-to-year trends and forecast for three product groups in your next meeting. You create two-line charts to present.

Which chart do you choose to present to your boss?
You should choose the first chart. Although the second chart is simpler and perhaps more elegant, it is not fit for purpose. It could be useful for a five-year forecast. However, the lack of units on the y-axis, years on the x-axis, and without the grid lines, it is difficult to understand the slopes of the lines (i.e., the trends).

(c) Redundancy

Which features of this graph are redundant or irrelevant?

There are four places where colour is redundant or not useful:
- The category labels. Why are they rainbow colours?
- There is redundancy in the blocks and the text. If the blocks are already in colour the chart could be simpler with non-coloured text.
- The headline. Why are there different colours in the headline? It does not help convey what is important regarding buying a drone?
- Why are the x-axis labels coloured? They do not need to be coloured, and at 60, which is coloured not at all important has all bars above in all three colours.

2. **The Fallen**

In your group, discuss the data viz epic the Fallen. Specifically, focus on

i. What was the purpose of the video?

- The video's primary purpose is to show the magnitude of death that occurred in World War II.
-Towards the end, there is also a message of hopefulness, by showing that the soldier who died in World War II has been dramatically higher than in all combined wars since.

ii. What were the potential challenges that the filmmakers faced when creating the video?

- Conveying a human tragedy of a magnitude of 70 million is difficult to comprehend
- The farther away events are, the more we treat them as statistics.
- The challenge is presenting the deaths as tragedies rather than reporting them as statistics.

iii. What visual strategies did the visualisation employ to connect data with individual lives?

- Deconstruction and reconstruction of data deepen engagement with a complex topic - Uses the soldier as a concrete representation of 1000 people; the soldier is more effective than a dot… a dot would make the visuals feel more like a statistic than the loss of life - The most powerful moment is the 45 seconds tallying the death of Russian soldiers. The time creates tension (any longer and it would pass the threshold of engaging to annoyance).
- Almost as powerful is what happens afterwards, where the zoom out contrasts the 9 million compared to the other countries. It shows how much of an outlier, and the small countries provide a frame of reference.

3. BCOM Majors

Imagine that you work for the Business School's Educational Portfolio team, and your colleagues are interested in understanding the compositions of the various BCOM majors. They also want to understand how the makeup of the program changes over time since shifting proportions can help plan future course development. So, you decide to make a visual to help them out.

i. You have five minutes to present to your boss on the Educational Portfolio team. You have been asked to show how students' majors have shifted from 5 years ago to today. You have sketched several options including two pie charts and two stacked bar charts. However, you are also thinking of using an alluvial diagram to impress the directors.

i. What kind of chart addresses your goal? How would you classify these charts amongst the four types?

Although a conceptual chart could accomplish the goal of providing information, your boss likely wants actual figures, in which case we want a declarative chart. The first two charts are everyday data viz. The Alluvial could be classified as since there is more information that could be used to explore the changes in more detail.

ii. Which of the three options should you use?

The answer to which option is best is "it depends". If your boss understands the alluvial chart, then this could be a good choice, particularly if you want to highlight specific paths, as this information is not available in the pie charts or stacked bar charts. However, if they are unfamiliar with the diagram, too much time would be spent explaining the chart. So, in this case, the stacked bar chart is more appropriate, as it does not require explanation and is generally more useful than pie charts when there are more than 3 categories.

iii. Now that you have discussed which of the chart you want to pursue to present to your boss on the Educational Portfolio team, download the dataset majors.xlsx to create two charts. Use the software infogram to create the charts. One chart should be based on one of the three options above. As you create the chart discuss what you are trying to say or show or prove or learn.

A good starting point is to take the stacked bar chart from Question 3 and remove the years 2018-2020 to highlight the changes between 2017 and 2021.

It is difficult to compare the different categories, so perhaps it would be better to make the graph horizontal bars. This can be done easily by changing the chart type.

Now the benefit of this graph is that is clearer that finance and information systems have proportionally increased their student numbers, largely at the expense of management. However, the years going from 2021 on the bottom to 2017 on top is strange. We typically expect to see years on the x-axis. So, let's try a line graph since the x-axis is time.

This will require adding in the actual proportions of each group as the data. Once, the correct

data has been added in, we can change the colours to highlight the insights that the Finance and Information Systems majors have substantially increased, and that the Management major has decreased substantially.

4. Supplementary Question: Happy Cows Revisited

In week 2, you discussed data visualisation and exploration of Happy Cow ice cream. In this case, both Mary and Prem, the senior sales assistant, believed that different flavours sell better. To address this conflict, you were required to use R to visualise the consumption of individual flavours, and to group them to generate insights into the ice cream sales by flavour or a category. For staff, the bar chart and pie charts were as follows:

i. Identify issues with these graphs.

There are many issues with these charts. They could be considered prototypes just to get a sense of the data. However, even then it is hard to extract valuable information from these charts. Some issues with the chart include:

- Titles not aligned
- Graphs are just plotting data, not telling a story
- Eye travel is not minimised and it is unclear where to focus
- Too much information is presented
- Nothing stands out

ii. Create a good chart that contrasts the difference in sales between students and staff. To refine your graphics, decide which extraneous element to remove. Use the remaining elements to highlight the idea, rather than describe the chart's structure. Make sure to minimize colours and limit eye travel. To do this, download the dataset "ice_cream.csv", and use a software tool like infogram, excel, or R.

Previously we grouped flavours based on fruits, chocolate, caramel, nuts, and tea. This is still a lot of information. Looking at the data, I started noticing that some of the flavours had very low sales, so this is probably less relevant information. I also looked at the fruit and noticed that mango, coconut, and banana, were by far the most popular fruits. So, I grouped those and clumped the non-chocolate and non-caramel into another category. From here I noticed that the proportion of

flavours was largely the same. So, I decided that my story would be that where it counts, students and staff are not so different. Given that I was considering the percentage of sales for a given set of flavours, I thought that the stacked bar was appropriate. I also tried to make the colours match the flavours to ease interpretation.

More refinement….
Then to facilitate eye movement, I narrowed the bars. This also uses less ink, which increases the "ink to impact ratio". I also emphasized the message in the title of the graph.
Next, I moved the flavours back to the top, as this is where the eye is naturally drawn to the emphasized title. Finally, I copied and pasted the figure into the paint, and minimized the white space between the title and subtitle and figure.

COMM1190: DATA, INSIGHTS, AND DECISIONS TERM 1 2022

SAMPLE EXAMINATION

QUESTION 1 30 MARKS

You have been brought in as a Data Science consultant on a court case. A chemical company has been found negligent after a chemical spill at one of their plants. All that remains in the court case is to decide on the extent of the damages for which the company is liable. One way

the court has been deciding on this amount is to look at the impact the spill has had on the value of houses located near to the chemical plant where the spill occurred.

As the expert witness, you have been asked to evaluate some alternative strategies to estimate the impact on housing prices (*price*). Strategy A involves taking a sample of sales that occurred after the spill where the houses are classified as either being close to the plant or not. This feature was designated by a variable *near* that was equal to 1 if the house was deemed to be close to the chemical plant and zero otherwise. Then a regression analysis is performed using the following model (*MA*):

$$\text{price}: \text{price}_i = \beta_0 + \beta_1 \text{near}_i + u_i.$$

Strategy B involves taking a sample of sales for houses near to the plant but where some sales occurred before the spill and some after. The variable *after* is equal to 1 if the house was sold after the spill and zero if the sale was before. Then a regression analysis is performed using the following model (*MB*):

$$\text{price}: \text{price}_i = \beta_0 + \beta_1 \text{after}_i + u_i.$$

a) Explain A and B as strategies to estimate the impact of the chemical spill and critically evaluate each of them. Is either preferable to the other?

[max 200 words] (10 marks)

b) Suggest an alternative regression model that is preferable to price given that you only have data from after the spill. Does this address all your criticisms of Strategy A that you outlined in Q1a)?

[max 150 words] (5 marks)

c) Using housing data models *MA* and *MB* are estimated, and the results given below.

How do you interpret these results? (Note that price is expressed in $1000)

$$\widehat{\text{price}} = 131.9 - 40.0\,\text{near} \quad (7.6)$$
$$(4.0)$$

$n = 142,\ R^2 = .165,$ root mse $= 63.7$

$$(5.9)$$

$$\widehat{\text{price}} = 63.7 + 28.3\,\text{after} \quad (9.1)$$

$\square\square\square\square = 96, \square\square\square\square^2 = .094,$ (.)

[max 200 words] (10 marks)

d) Suppose you have sales both near and not near to the plant as well as sales before and after the spill. Suggest an alternative strategy to estimate the effect of the oil spill on housing prices that is preferable to both *MA* and *MB*?

[max 150 words] (5 marks)

QUESTION 2 30 MARKS

Imagine you work for a large department store, which highly values customer service. The following chart shows how customers contact the customer service centres.

You begin to discuss the chart with your manager. Immediately, she has the following queries: “I want to see the overall trends, but it is difficult to see with all the seasonal spikes in the time series. I’d like a simpler view into the trend.” You decide to create some charts to address your manager’s queries.

a) Identify the type of chart you would use to address the query and explain why. [max 100 words] (5 marks)

b) Sketch two alternative charts for the query. For each chart, provide a brief explanation of your design choices. To sketch the chart, you can use any tool you want (e.g., you can use a software tool like infogram, excel, or R). Alternatively, you can sketch the chart using pencils, pens or markers on paper, then take a picture of the charts and paste them into your solutions document. You can access the underlying data “customer_service.xlsx” on Ed.

[max 300 words] (15 marks)

c) Evaluate your two charts and explain which you would select to further develop to

present to your manager.

[max 200 words] (10 marks)

QUESTION 3 20 MARKS

Recall the Data Analytics Simulation: Strategic Decision Making that you played in Week 1's Workshop. The simulation provided first-hand experience in the benefits and challenges of making data-driven decisions.

1. Reflect on how the data visualisations, dashboards, and filters helped you make decisions and the challenges you experienced while playing the game.

2. Based on your experience, identify organisational benefits and challenges from using data and modern visualization to make decisions.

The critical reflection should include lessons learned during the game, with a focus on the value and challenges of working with data.

[max 400 words] (20 marks)

QUESTION 4 20 MARKS

You are a data analyst for AppCo. AppCo produces a smartphone app that allows users to virtually try on clothes. It is funded by having sponsored links to online clothing retailers. AppCo tells its users, "top brands, all sizes, best prices". AppCo has a dashboard that is used by Board members that shows sales, revenues and a comparison of sales by retailer. You have been asked to consider whether having a selection of the most popular sizes at a slightly lower price would increase revenue. The CEO has said, "well just use some of your magic A/B testing …".

a) Identify potential legal issues that may arise from A/B testing if AppCo users are unaware the experiment is taking place.

[max 100 words] (5 marks)

b) Evaluate whether A/B testing would lead to any legal consumer issues? [max 200 words] (10 marks)

c) Recommend steps to ensure that your organisation maintains good governance when developing analytics at AppCo.

[max 100 words] (5 marks) — END OF EXAMINATION PAPER —

Question 1

Marks: 30

Word Limit: 700 words

Your word count: *** words

Your answer:

Q1a)

Both regressions have a design with a structure like an experiment with a treatment and control group. This results in estimates that can be interpreted as differences in means.

Strategy A estimates the difference in means for housing near compared to houses away from the spill. • The **primary flaw** in that we cannot determine if other factors contribute to price differences based on distance to the plant.

- If the chemical plant was built in a commercial area that is unattractive to homeowners, differences found using this strategy may reflect differences in the attractiveness of the near/away areas that preceded the oil spill and hence are not directly attributable to the spill.
- There is an additional concern about the representativeness of houses for sale after the spill, a sample selection problem.

Strategy B is a before and after comparison but confined to houses near the chemical plant. • We avoid the previous problem of comparing near and away houses by using houses that were near, making them more comparable.

- The **primary flaw** in this strategy is a confoundment problem, which now happens over time. For example, housing market dynamics may create broad changes in the price of all houses.
- There again is a selection problem relating to the representativeness of houses put up for sale before the spill compared to those after the spill. For example, if those before were broadly representative but only poorer houses were put on the market after the spill then this would induce a downward bias of the estimated impact of the spill.

You would have to argue about the relative size of these biases so in general it is not obvious that one strategy is better than the other.

Q1b)

Adding covariates to the regression model reflecting house characteristics would help. This would control for observable differences between near and away houses that were put up for sale.

This could also help with the selection problem to the extent that observable differences explain whether houses were put up for sale or not.

Conversely if the reason the houses were put up for sale are unobservable to the analyst, then the omitted variable and selection problems remain.

Q1c)

Regression 1:

- The first regression implements Strategy A.
- As expected, houses near the chemical plant sell for prices less than those not near the plant. The estimated difference in the mean price of houses is $40,000.
- This is a large difference, implying houses in the affected areas sell for 30% less than those houses sold in the unaffected area. [$30 \cong (40/131.9) \times 100$]
- This difference is precisely estimated as the 95% CI of $40 \pm 1.96 \times 7.6$ or [25.1, 54.9] does not include zero.

Regression 2:

- The second regression implements Strategy B. Here houses in the affected area sell for prices $28,300 more than those sold before the spill.
- This result is not consistent with a detrimental impact of the spill, but it is consistent with the concerns expressed in (a) that there may be biases here due to general upward movement in house prices. • This is a large difference, implying houses sold after the spill sell for 44% more than those sold before. [$44 \cong (28.3/63.7) \times 100$]
- This difference is precisely estimated as the 95% CI of $28.3 \pm 1.96 \times 9.1$ or [10.5, 46.1] does not include zero.

Q1d)

A difference in differences (DiD) approach is an appropriate way to proceed when data are available on houses sold before and after the spill and near and away.

Essentially calculate the difference in average near and away house sale prices before the spill and compare it to the difference in average near and away house prices after the spill. The resulting difference in difference estimate then helps control for many of the concerns in Strategy A and B.

Controlling for time effects and household characteristics would also help.

Question 2

Marks: 30

Word Limit: 600 words

Your word count: *** words

Your answer:

Q2a)

This is a declarative and data-driven chart. The idea being presented is the trends, so this is data-driven. In addition, you are not exploring trends, you are just presenting a simple view of the trends, so this is declarative. Thus, the type of chart needed is an everyday data viz.

Q2b)

Q2c)

The line chart seems more appropriate as it better illustrates the upwards trend of the virtual agents and the decrease of the phone (whereas e-mails and in persons are fairly consistent).

The stacked bar chart shows that virtual agents are a growing aspect of the business, proportionally, but it is more difficult to see the clear trends that are visible on the line graph.

To further develop the line chart, I would explore adding future trend line, potentially with scenarios, since this would even better highlight the trends of growth in virtual agents. In addition, I would consider moving the channels to be the graph next to the lines to limit eye travel.

When presenting the data, I would consider presenting two versions of the graph. The first would be without virtual agents. I would make the title “modest changes in customer service…” and make the comment that this does not include new technologies for handling customer service. After allowing the manager to digest that information, I would then present a version of the graph above with virtual agents accentuated in blue (greying out the other channels).

Question 3

Marks: 20

Word Limit: 400 words

Your word count: *** words

Your answer:

The dashboards, visualisations, and filters enable data driven decisions. For example, line charts quickly

facilitate identifying trends (e.g., product popularity and competitor pricing patterns), which in turn facilitate discussion regarding strategy. The filters provided further functionality by providing answers for questions regarding specific market segments, which helped create a strategy for Blue. The filters also had the benefit of streamline the overload of data available through the various dashboards. After selecting a strategy, as the data was aggregated to an annual level, it was difficult to determine if increased profits resulted by chance or good decision making. Also, positive performance reinforced the acceptance and use of our strategy, which means that we would miss potential new opportunities to do even better resulting from new data and other market trends. Compounding this issue is that using our filters, it was easy to miss new developments occurring outside our area of focus. Unfortunately, we did not notice these omissions until after performance would fall. Thus, our responses were often reactionary rather than proactive, undermining a key advantage of data driven decisions.

For organisations, data driven decisions enables greater confidence in the actions, since logic and evidence underpin critical decisions. Using data as evidence can also help make arguments (e.g., for financing or resources) more compelling to decision makers. Moreover, being data-driven can help overcome biases and preconceptions. I personally know a lot of people who prefer liquid detergent to pods, and initially gravitated towards that product form. However, a quick inspection of the data showed that this hypothesis was false: Pods were more appealing to our target market. Our group found the volume of data and array of filters challenging; however, this challenge is likely to magnified for organisations. For example, each of our team members initially explored the data separately and came to different conclusions on the best strategies. For large teams in organisations, with even more data and analysis approaches could lead to debate and opinions over the best way forward. Although this is a substantial challenge, with proper structures, this type of debate is extremely valuable, as it can ultimately lead to better decision making.

Question 4

Marks: 20

Word Limit: 500 words

Your word count: *** words

Your answer:

Q4a)
Poor design of an AB testing experiment could lead to issues under consumer law in Australia. For example, if the subjects who receive a service in group B do not receive a service which is fit for purpose, this could expose the business to legal action. There is also an important consideration in respect of AB testing to ensure that it does not breach Australia's antidiscrimination laws.

Q4b)
The "all sizes" and "best prices" might be misleading or a misrepresentation if there is no basis for the statement for some AppCo users. A business will be liable for engaging in misleading or deceptive conduct if it makes a statement about the future that later proves to be incorrect, unless the business had reasonable grounds for making the statement. Misrepresentation is associated with advertising where a business claims something about a good or a service which is untrue or misleading. However, the legislation about misrepresentation does not specify that any misrepresentation needs to be in an advertisement. As a result, it is possible for a business to mislead others by making claims which are false or misleading.

Q4c)
One approach is to have a "red team" or "devil's advocate" when decisions are made within a business. One person or one team advocating for one approach and another team or person advocating for an alternative approach. It means that a great solution to a problem can be synthesised from the advocacy of each of the two groups.