Midterm Exam

Yuxi Wang

11/2/2020

Instruction

This is your midterm exam that you are expected to work on it alone. You may NOT discuss any of the content of your exam with anyone except your instructor. This includes text, chat, email and other online forums. We expect you to respect and follow the GRS Academic and Professional Conduct Code (http://www.bu.edu/cas/files/2017/02/GRS-Academic-Conduct-Code-Final.pdf).

Although you may NOT ask anyone directly, you are allowed to use external resources such as R codes on the Internet. If you do use someone's code, please make sure you clearly cite the origin of the code.

When you finish, please compile and submit the PDF file and the link to the GitHub repository that contains the entire analysis.

Introduction

In this exam, you will act as both the client and the consultant for the data that you collected in the data collection exercise (20pts). Please note that you are not allowed to change the data. The goal of this exam is to demonstrate your ability to perform the statistical analysis that you learned in this class so far. It is important to note that significance of the analysis is not the main goal of this exam but the focus is on the appropriateness of your approaches.

Data Description (10pts)

Please explain what your data is about and what the comparison of interest is. In the process, please make sure to demonstrate that you can load your data properly into R.

Objective: I want to understand the relationship between the price of textbooks currently used by graduate students in the Department of Statistics and some attributes of the books, such as the number of pages.

Approach: I first inquired about the relevant information about the textbook I was using, and then asked three MSSP students, and two other students studying applied statistics in other schools, about their textbook information.

Ideas of the survey: Inquire by myself, and ask five of my friends who are studying satistics about some simple questions, reducing the time cost of collecting data. In addition, since it is only investigating the relevant content of the textbook, it does not involve any privacy issues, and it is very convenient to get an answer when consulting other students' textbooks. For the prices of these books, I found information about new books from Amazon.com, and I checked the prices of new books in Chinese from Dangdang.com.

```r
# Load the data
raw_data <- read.csv("/Users/mac/Desktop/BU Mssp/MA678/Material/data collection/Raw_data.csv",head=TRUE)
head(raw_data[,-c(1)])
```

	Book.ID	Pages	Chapters	English.edition	Price
	<int>	<int>	<int>	<int>	<dbl>
1	1	534	22	1	52.20

	Book.ID <int>	Pages <int>	Chapters <int>	English.edition <int>	Price <dbl>
2	2	625	25	1	65.68
3	3	698	10	1	77.19
4	4	376	14	1	83.95
5	5	492	24	1	45.94
6	6	351	31	1	75.63

6 rows

Interpreting: ‘Name’: The name of the textbook that the surveyed graduate student reads. ‘Book ID’: The ID of the book being investigated. ‘Pages’: Number of pages in the book ‘Chapters’: Number of chapters in the book ‘English edition’: Whether the book under investigation is in English, 1 means yes, 0 means no. ‘Price’: I checked the price of the book from a website with a lot of purchases.

EDA (10pts)

Please create one (maybe two) figure(s) that highlights the contrast of interest. Make sure you think ahead and match your figure with the analysis. For example, if your model requires you to take a log, make sure you take log in the figure as well.

```
# Data cleaning
# Because the variable 'price' has different unit(USD or RMB), so I need to exchang t
he unit into dollar.
url<-'http://stockq.cn/market/currency.php'
web<-read_html(url)
res<-web%>%
  html_table(fill=T)%>%.[8]%>%as.data.frame
res<-res[c(16),c(1:3,5)]
res[,2:3]<-lapply(res[,2:3],as.numeric)
ChangeRate<-round(res$X3/res$X2,4)
currency<-data.frame(res[,1:3],ChangeRate,res[,4])
names(currency)<-c("Name","Price","ChangePrice","ChangeRate","Time")
currency<-currency[order(-currency$ChangeRate),]
print(paste("The exchange rate between RMB and USD is:", currency$Price))
```

```
## [1] "The exchange rate between RMB and USD is: 6.608"
```

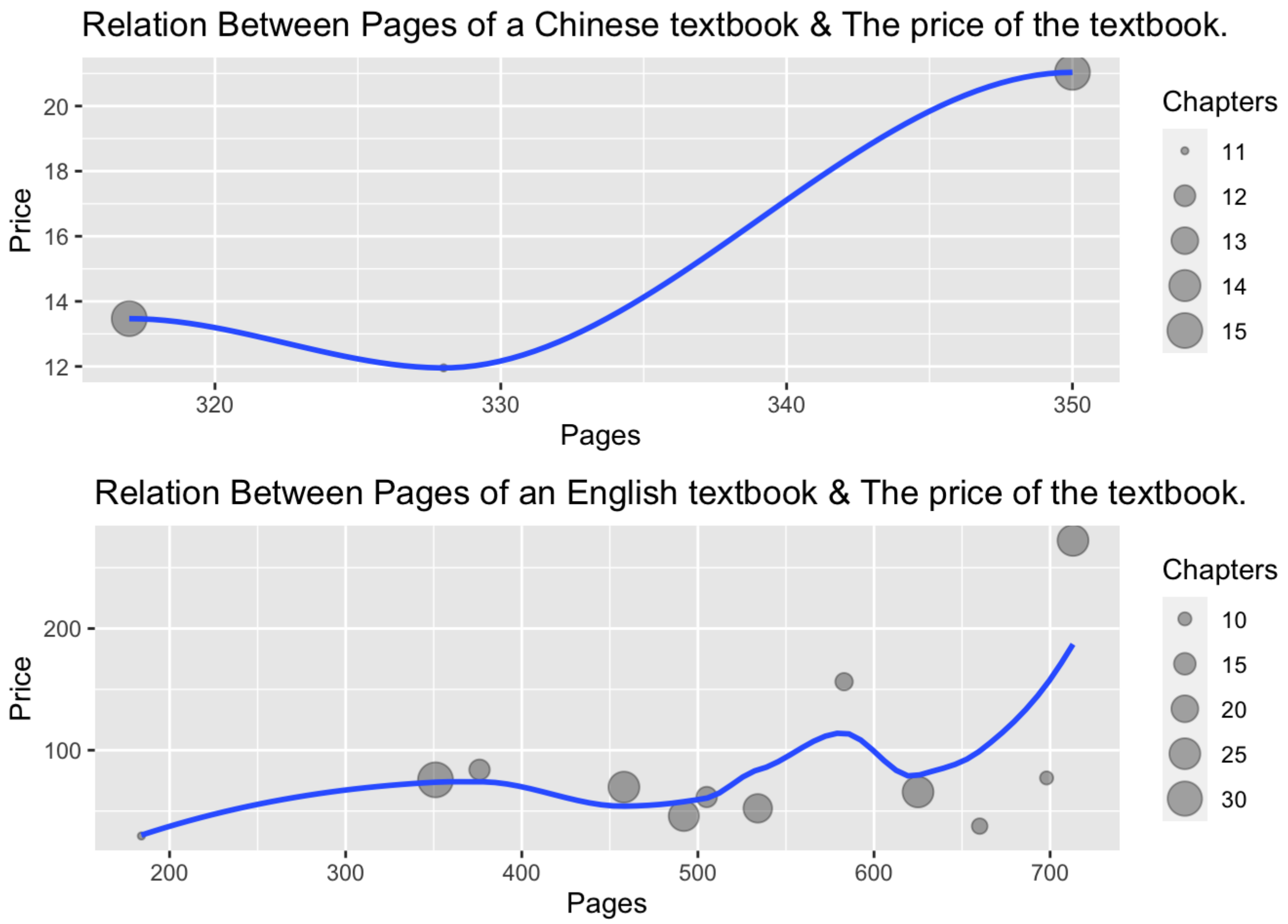
```
# Make the unit of price variable in U.S. dollars
data <-raw_data
data[10:12,6] <- raw_data[10:12,6]/currency$Price
view(data)

# Seprete the data with English edition and Chinese edition
data_C <- data[10:12,]
data_E <- data[-c(10:12),]
```

Through data cleaning, I unified the data units so that they can be directly brought into EDA and modeling later. At the same time, I seperate data

```
# EDA
p1 <- ggplot(data = data_C, mapping = aes(x = Pages, y = Price)) +
  geom_point(aes(size = Chapters ), alpha = 1/3) +
  geom_smooth(se = FALSE)+
  ggtitle("Relation Between Pages of a Chinese textbook & The price of the textbook."
)

p2 <- ggplot(data = data_E, mapping = aes(x = Pages, y = Price)) +
  geom_point(aes(size = Chapters ), alpha = 1/3) +
  geom_smooth(se = FALSE)+
  ggtitle("Relation Between Pages of an English textbook & The price of the textboo
k.")
plot_grid(p1,p2,label_x = 0.2,nrow = 2)
```



By observing this figure, the first thing we can see is the relationship between the number of pages of the book and the price of the book. Through the trend curve, we can clearly understand that as the number of pages of a book increases, the price of a book generally increases. In the middle of the curve, the price of books has fallen. This may because the number of data sets is not large enough, and the book types are not exactly the same as the publishers, resulting in certain differences in pricing.

In addition, I define the size of each data point as the number of chapters in the book. From this figure alone, it is difficult to see the relationship between the number of chapters of a book and the number of pages of a book or the price of the book. This situation may due to the different chapter methods and different types of books. For example, the pages of each chapter of a theoretical textbook tends to be bigger, because such a book requires a lot of space to prove each definition or theorem and derive each formula. The application book has relatively fewer pages per chapter.

Finally, I also made a comparison between Chinese and English books. Judging from the collected data alone, the price and number of pages of Chinese books are usually less than those of English books.

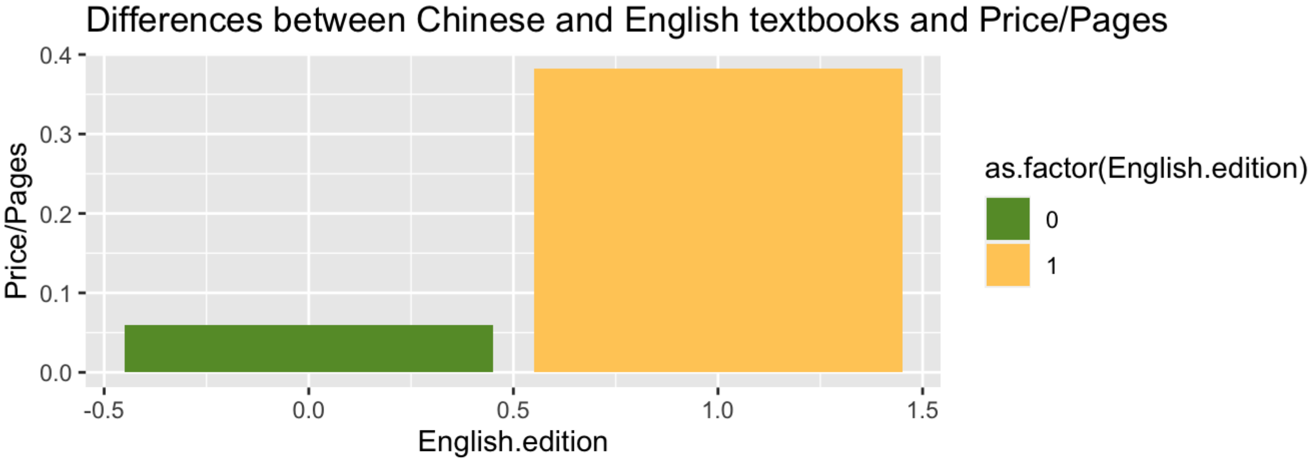
Therefore, I want to further explore other differences between Chinese and English books.

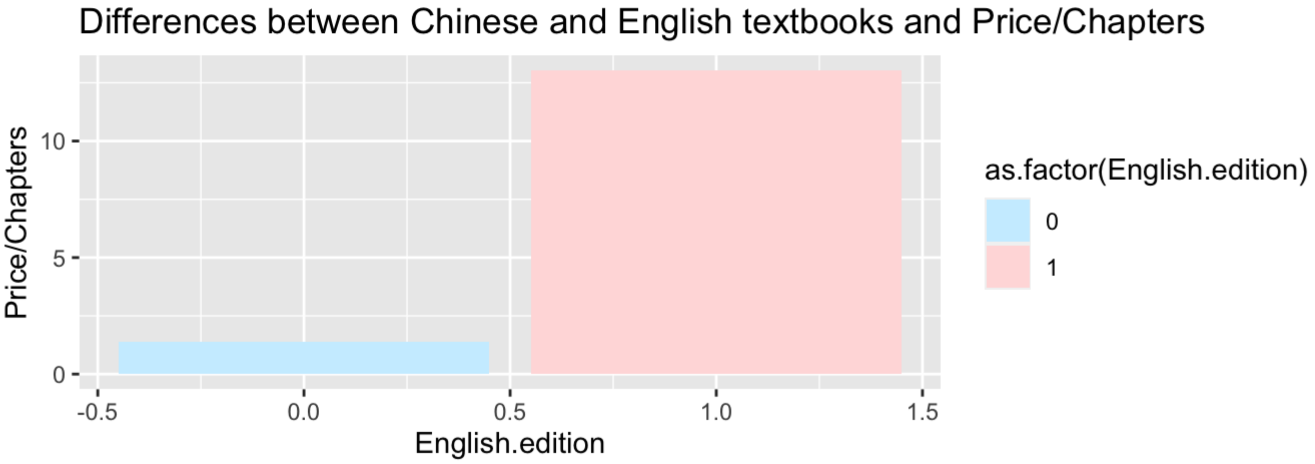
```
data <- select(data,
       Book.ID:Price
      )
mutate(data,
       price_page <- Price/Pages,
       price_chapter <- Price/Chapters
)
```

Book.ID <int>	Pa... <int>	Chapters <int>	English.edition <int>	Price <dbl>	price_page <- Price/Pages <dbl>
1	534	22	1	52.20000	0.09775281
2	625	25	1	65.68000	0.10508800
3	698	10	1	77.19000	0.11058739
4	376	14	1	83.95000	0.22327128
5	492	24	1	45.94000	0.09337398
6	351	31	1	75.63000	0.21547009
7	184	9	1	29.49000	0.16027174
8	660	11	1	37.57000	0.05692424
9	505	14	1	61.52000	0.12182178
10	328	11	0	11.95521	0.03644880

1-10 of 15 rows | 1-6 of 7 columns Previous **1** 2 Next

```
p1 <- ggplot(data = data, aes(x = English.edition, y = Price/Pages, fill=as.factor(En
glish.edition)))+
  geom_bar(position = 'dodge',stat = 'identity')+
  scale_fill_manual(values = c("#669933","#FFCC66"))+
    ggtitle("Differences between Chinese and English textbooks and Price/Pages")
p2 <- ggplot(data = data, aes(x = English.edition, y = Price/Chapters, fill=as.factor
(English.edition)))+
  geom_bar(position = 'dodge',stat = 'identity')+
  scale_fill_manual(values = c("#CCEEFF","#FFDDDD"))+
  ggtitle('Differences between Chinese and English textbooks and Price/Chapters')
plot_grid(p1,p2,label_x = 0.2,nrow = 2)
```





From Chinese and British textbooks, we can know that the relative price of Chinese textbooks is much lower than that of American textbooks. In my opinion, there are two main reasons for this. First, there is a difference in exchange rate between China and the United States, so purchasing power is different. The second is that the agency rights of many Chinese textbooks are buyouts, and publishers will subsequently be subject to government pricing restrictions. Therefore, they no longer benefit from high prices, but through the huge number of sales.

Power Analysis (10pts)

Please perform power analysis on the project. Use 80% power, the sample size you used and infer the level of effect size you will be able to detect. Discuss whether your sample size was enough for the problem at hand. Please note that method of power analysis should match the analysis. Also, please clearly state why you should NOT use the effect size from the fitted model.

The power that I use is 80%, the sig.level I use is 5%, and the simple size is known, we may can use `pwr.t.test()` from the `pwr` package to calulate the power of a t-test.

```
pwr.t2n.test(n1=12 , n2=3 , d =NULL , sig.level =0.05, power =0.8)
```

```
##
##      t test power calculation
##
##               n1 = 12
##               n2 = 3
##                d = 1.95669
##        sig.level = 0.05
##            power = 0.8
##      alternative = two.sided
```

As we can see, the d=1.95669. After looking up some related articals, I found that this number may not make sense.

The d is defined as(only for two sample t test):

$$d = \frac{\mu_1 - \mu_2}{\sigma_{pooled}}$$

So, I believe that the best way for me to do the power analysis is to calculate the effect at first. Then, compare the sample size of calculation with the data set that I have, in order to find out the sample size is enough or not.

```
d <-( mean(data_E$Price)-mean(data_C$Price) )/sd(data$Price)
print(paste("The effect of the dataset is:", d))
```

```
## [1] "The effect of the dataset is: 1.06004981665297"
```

The result of calculating the equation is 1.06004981665297, which is also bigger than 1. I do not sure this result make sense, and from Cohen (1988, pages 24-27) I know:

– Small effect: 1% of the variance; d = 0.25 (too small to detect other than statistically; lower limit of what is clinically relevant) – Medium effect: 6% of the variance; d = 0.5 (apparent with careful observation) – Large effect: at least 15% of the variance; d = 0.8 (apparent with a superficial glance; unlikely to be the focus of research because it is too obvious)

So, maybe the data have a large effect.

Modeling (10pts)

Please pick a regression model that best fits your data and fit your model. Please make sure you describe why you decide to choose the model. Also, if you are using GLM, make sure you explain your choice of link function as well.

Since the data that I colleted is very simple, The outcome of data is a continuous price, not a binary outcome or counting model, etc., so I can just use the linear regression to fit it, and the language edition I will trated as a factor in the model.

```
data$Pages_scale <- scale(data$Pages, center = TRUE, scale = TRUE)
data$Chapters_scale <- scale(data$Chapters, center = TRUE, scale = TRUE)
fit1 <- lm(Price~Pages_scale+Chapters_scale+as.factor(English.edition), data=data)
fit <- fit1%>%summary(digits=2)
pander(fit)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	50.66	38.31	1.322	0.2129
Pages_scale	30.78	17.7	1.739	0.1098
Chapters_scale	12.2	16.21	0.7526	0.4675
as.factor(English.edition)1	26.17	44.06	0.594	0.5646

Fitting linear model: Price ~ Pages_scale + Chapters_scale + as.factor(English.edition)

Observations	Residual Std. Error	R^2	Adjusted R^2

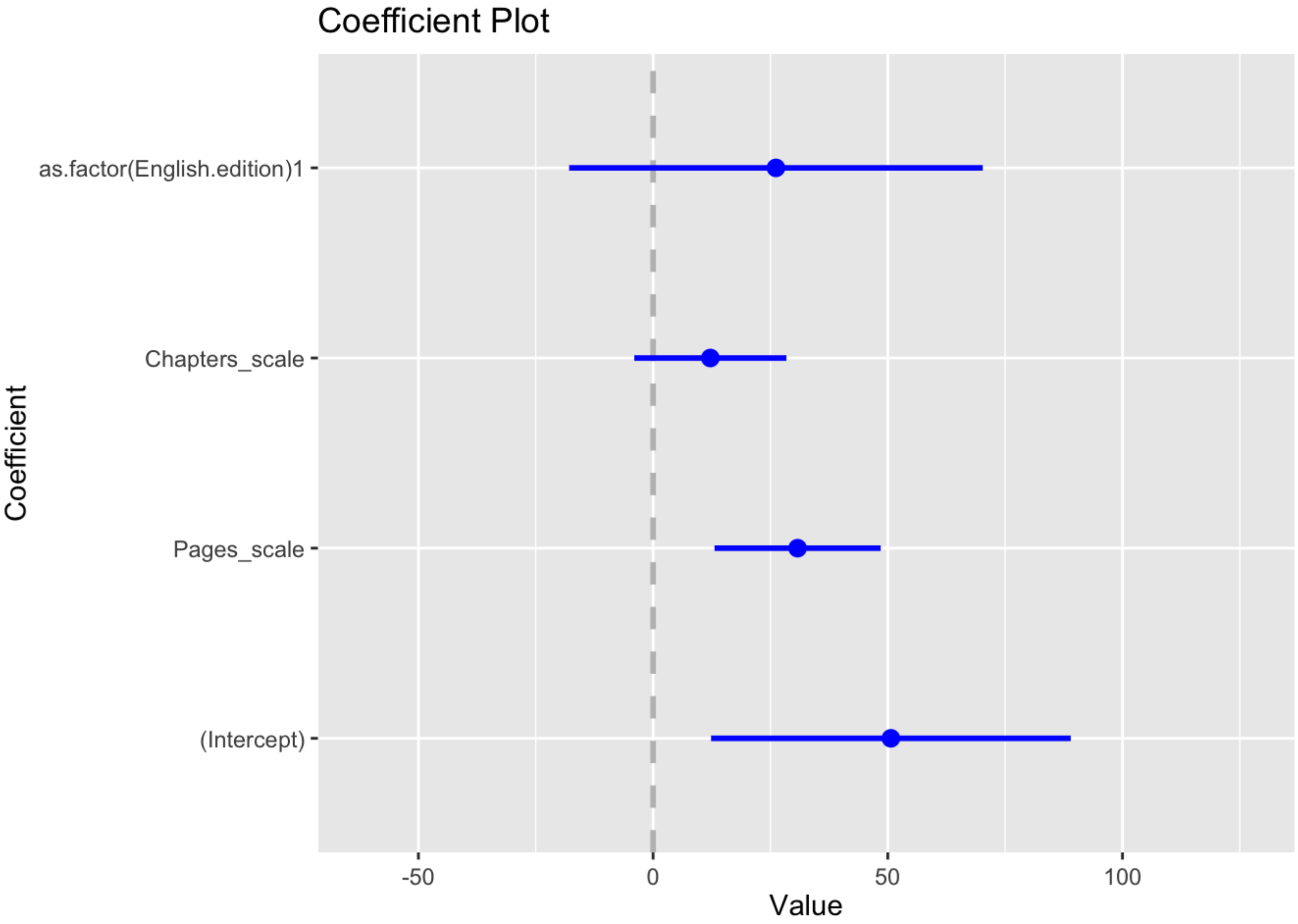
Observations	Residual Std. Error	R^2	Adjusted R^2
15	58.13	0.3936	0.2283

```
ci<-confint(fit1)
pander(exp(-coef(fit1)))
```

(Intercept)	Pages_scale	Chapters_scale	as.factor(English.edition)1
9.939e-23	4.272e-14	5.043e-06	4.313e-12

After running my code, I can easily check the coefficients, there are no significant variable in this model. Also, the R^2 of the model is 0.3936, the Adjusted R^2 is 0.2283. Both of the result are not near 1. So, I have to say, the model is not fitted well. So, I want to figure out where the problems are. I make a coefplot:

```
coefplot(fit1,mar=c(1,10,5,2))
```



By watching this figure, I found that the confidence interval for both ‘English.edition’ and ‘Chapters’ cross 0, so we will interpret the ParentBeliefVariation and DomainScience. These two variable are not proporate in the model.

So, I drop these two variables, and only use pages as predictor and price as the ooutcome to fit a simple linear model.

```
fit2 <- lm(Price~Pages, data=data)
fit_2 <- fit2%>%summary(digits=2)
pander(fit_2)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-42.6	47.7	-0.8929	0.3881
Pages	0.2388	0.09499	2.514	0.02592

Fitting linear model: Price ~ Pages This time, the result makes sense. At least, the results show that the variable ‘pages’ is significant. The model can be written as:

$$Price = -42.6 + 0.2388x_(pages)$$

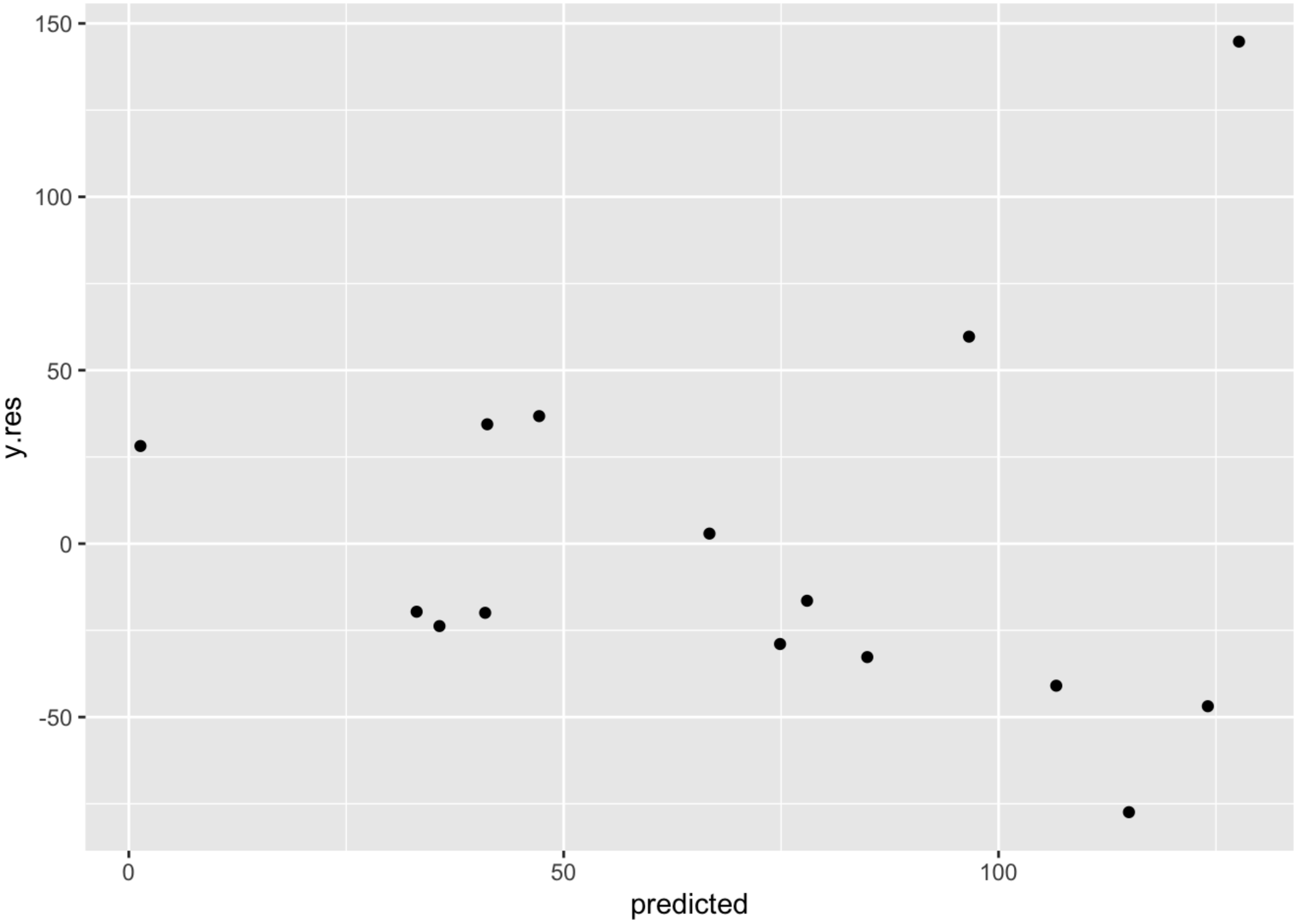
The intercept of the model is -42.6, which means that if we want to add some data in this model, we must have a textbook of $42.6/0.2388 = 178.392 near 179$ pages or more.

Observations	Residual Std. Error	R^2	Adjusted R^2
15	56.33	0.327	0.2753

The coefficient of ‘pages’ is 0.2388, which means If the number of pages in a book increases by one page, the price of the book will rise by $0.2388.

But the standard deviation of the model is still very large, and the value of $R^2 = 0.327 << 1$. This is caused by the fact that the observed value of the data is too small. Although there are still many problems, this simple model currently fits best The data I collected.

```
# The residual plot of the model
y.res <- resid(fit2)
predicted <- predict(fit2)
data_a <- data.frame(y.res,predicted)
ggplot(data=data_a, mapping= aes(x=predicted, y=y.res)) +
  geom_point()
```



Through the residual graph, we can temporarily think that the residual has no trend and spreads evenly, so there is no problem. But this may also be caused by too few observation points. Insufficient observations often cause us to fail to see what trend the residual graph has.

Validation (10pts)

Please perform a necessary validation and argue why your choice of the model is appropriate.

I use the `loo()` function in the rstanarm package to use LOO cross-validation and put it here.

```
fit3 <- stan_glm(Price~Pages, data=data,refresh=0)
loo(fit3)
```

```
## 
## Computed from 4000 by 15 log-likelihood matrix
## 
##          Estimate   SE
## elpd_loo    -86.0  6.1
## p_loo         5.2  3.8
## looic       172.1 12.2
## ------
## Monte Carlo SE of elpd_loo is NA.
## 
## Pareto k diagnostic values:
##                          Count Pct.    Min. n_eff
## (-Inf, 0.5]   (good)     14    93.3%   1707      
##  (0.5, 0.7]   (ok)        0     0.0%   <NA>      
##    (0.7, 1]   (bad)       0     0.0%   <NA>      
##    (1, Inf)   (very bad)  1     6.7%   6         
## See help('pareto-k-diagnostic') for details.
```

The `elpd_loo` of the model is -85.5, which is the “Estimated Log Predictive Denstiy” calculated through Loo. Since it is not close to 0, the model may have high test error. The `looic` of the model is 4.6, which is the “LOO Information Criterion”. In this model it is closeto 0, so the result is nice. The `p_loo` is the hypothetical number of parameters. In the model, it is 170.9.

```
kfold(fit3)
```

```
## 
## Based on 10-fold cross-validation
## 
##            Estimate   SE
## elpd_kfold    -85.1  5.8
## p_kfold          NA   NA
## kfoldic       170.2 11.6
```

From the results, we can see that the results obtained by him and loo are similar. Since the observed value of the data is too small, the model may have high test error.

Inference (10pts)

Based on the result so far please perform statistical inference to compare the comparison of interest. Calculating the confidence and plot the interval of it

```
confint(fit2)
```

```
##                    2.5 %     97.5 %
## (Intercept) -145.65236841 60.4607938
## Pages          0.03354688  0.4439875
```

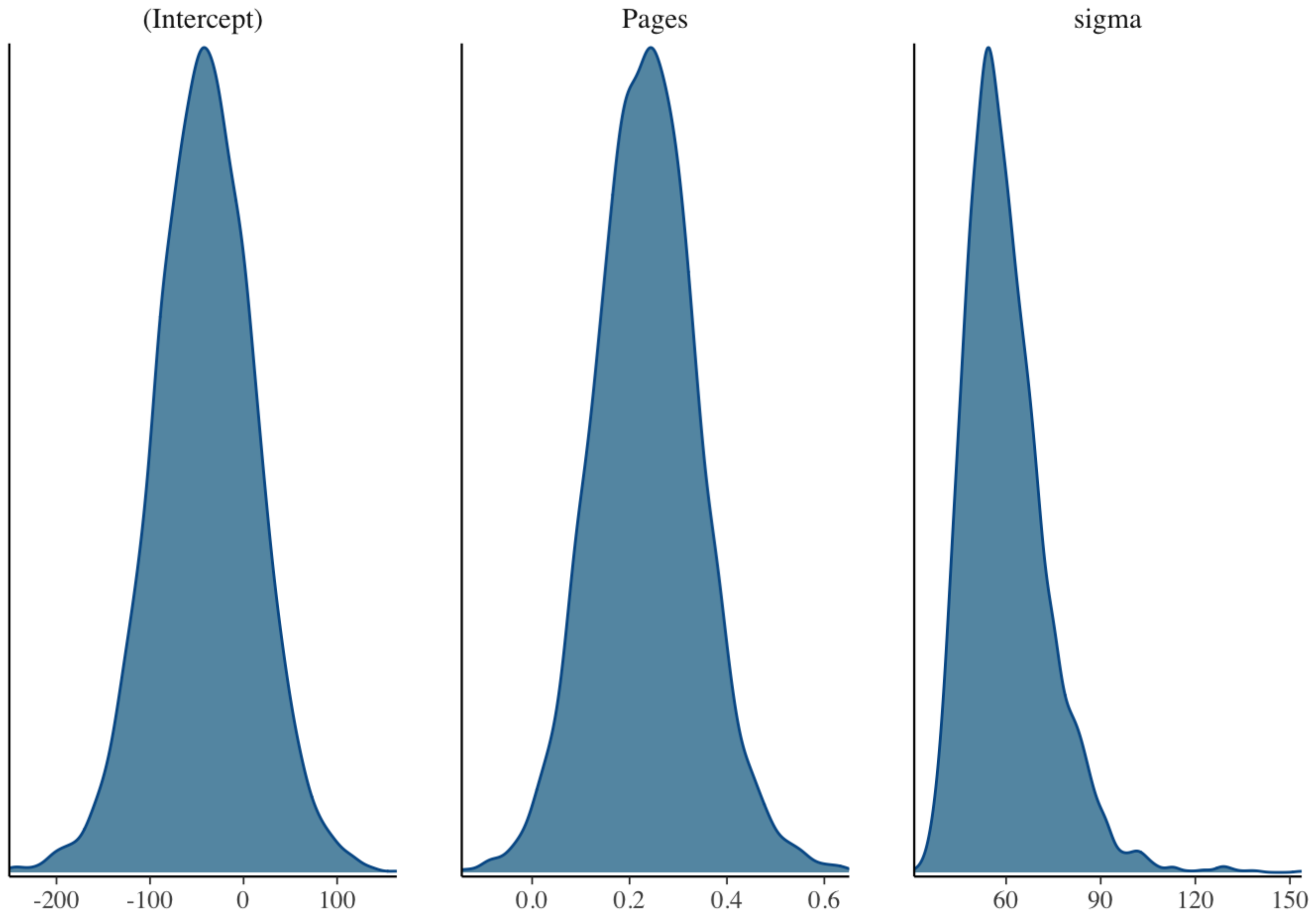
Using the matrix express uncertainty about a parameter estimate or function of parameter estimates.

```
sims <- as.matrix(fit3)
head(sims)
```

```
##           parameters
## iterations (Intercept)     Pages    sigma
##       [1,] -103.237204 0.4871390 74.38884
##       [2,] -178.223601 0.5988866 99.46918
##       [3,]   10.534861 0.1050643 48.35610
##       [4,]  -47.537600 0.2176843 50.18825
##       [5,]   -6.930375 0.1841786 67.01799
##       [6,]   -7.137908 0.1854072 62.83499
```

Looking at the distribution of the coefficients

```
mcmc_dens(sims)
```



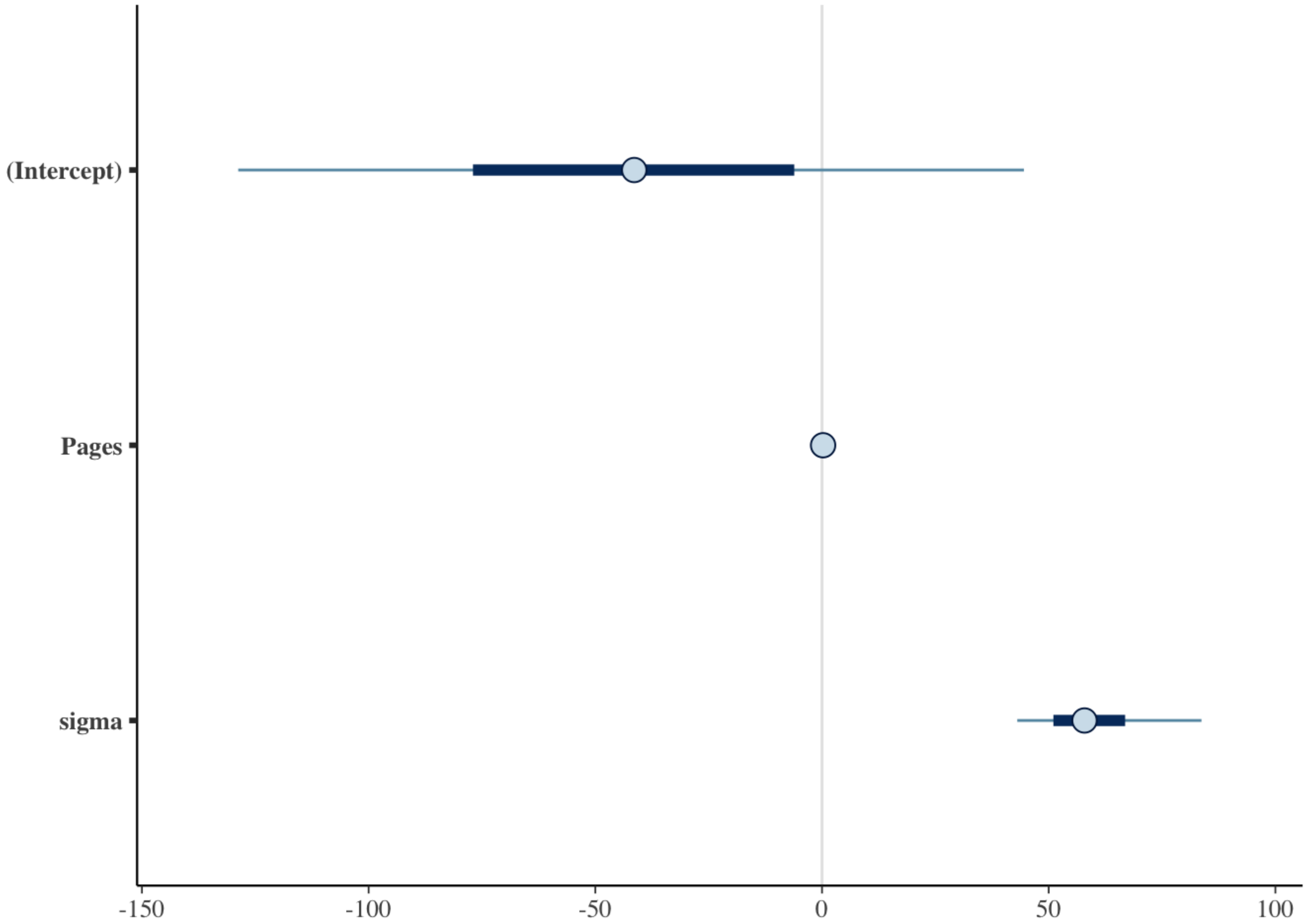
Computes Bayesian posterior uncertainty intervals. These intervals are often referred to as credible intervals.

```
posterior_interval(fit3)
```

```
##                        5%       95%
## (Intercept) -128.76159715 44.530754
## Pages          0.06791043  0.410296
## sigma         43.06764191 83.691949
```

And we can visualize the posterior intervals.

```
mcmc_intervals(sims)
```



By looking at the figure, the confidence interval of intercept has some problems, but this is caused by data, and there is no more suitable model.

Discussion (10pts)

Please clearly state your conclusion and the implication of the result.

This result makes sense to a certain extent. At least, the results show that the variable ‘pages’ is significant. The model can be written as:

$$Price = -42.6 + 0.2388x_(pages)$$

The intercept of the model is -42.6, which means that if we want to add some data in this model, we must have a textbook of $42.6/0.2388 = 178.392 near 179$ pages or more.

The coefficient of ‘pages’ is 0.2388, which means If the number of pages in a book increases by one page, the price of the book will rise by $0.2388.

But the standard deviation of the model is still very large, and the value of $R^2 = 0.327 << 1$. This is caused by the fact that the observed value of the data is too small. Although there are still many problems, this simple model currently fits best The data I collected.

Limitations and future opportunity. (10pts)

Please list concerns about your analysis. Also, please state how you might go about fixing the problem in your future study.

Firstly, the type of data and the sample size are too small. In the next time of data collection, I will collect as much and comprehensive data as possible.

Moreover, I don’t have a comprehensive understanding of regression models, I just studied the models in the textbook.. I clearly know that for this data set, the regression of the binomial class is not available. Categorical regressions are also not applicable. I also considered the random effect model, but I think it might not be applicable. Therefore, a simple linear regression is used according to the characteristics of the data. Regarding this point, I think we should read more papers and reports in the future to see how other statisticians solve problems when they encounter problems.

Finally, there are still some problems with my understanding of Power Analysis. I have searched some related documents, but still do not fully understand. After finishing this homework, I will ask my classmates for advice.

Comments or questions

If you have any comments or questions, please write them here.

I want to know if there are other regression model, other than linear regression, for this simple dataset.

Reference

[1]Regarding the procedure of importing currency exchange rates, I refer to the website: http://stockq.cn/market/currency.php (http://stockq.cn/market/currency.php).

[2]The power analysis part have a citation: https://stats.idre.ucla.edu/other/mult-pkg/seminars/intro-power/ (https://stats.idre.ucla.edu/other/mult-pkg/seminars/intro-power/)