HIGHER SCHOOL OF ECONOMICS
NATIONAL RESEARCH UNIVERSITY

Team 4 Report

Which Type of TV will Consumers Give Higher Ratings?

Empirical Evidence from Indian E-commerce Website Flipkart

Surname Name Program, Group

1. Introduction

Yu: notes marked as following, del before submission

a) describe the chosen economic problem and the goal of your research (no need to change much but make it briefly)

From black-and-white to colorful to modern technology, with intelligent software and hardware development, until now, TV is still popular and plays a vital role in both developed and developing countries. In the UK, 94% of children watched TV for an average of 13.25 h per week in 2018 (Nieto & Suhrcke, 2021). Malaysia's largest pay-TV provider has more than 3 million subscribers with a household penetration rate of 72% as of December 2017 and will continue to dominate the sector (Dawi, Jusoh, Streimikis & Mardani, 2018). In India, a 2008 Media Partners Asia (MPA) report revealed that 57% of homes in India own a TV set, out of which 61% (81 million households) subscribe to cable service. This indicates an increase of more than 25 million households in three years (Taylor, 2008).

In this research, we choose a market in developing countries; compared with developed countries, the market still has significant room for improvement and innovation, and the demand for TV is still relatively abundant, especially for digital ones. Because of the flourishing Indian economy and the growing demand for innovative technologies, the Indian market has excellent prospects for development. As a representative of developing economies, India has attracted many businesses to invest. Therefore, we focus on the Indian TV market. If we want to create an e-commerce store in India that specializes in selling TV, how can we ensure our TV's higher ratings to attract more customers? What factors will influence customers' ratings of TV?

b) write down stated hypotheses

Hypothesis 1: A better-selling/known TV brand indicates a higher average customer rating, ceteris paribus;

Hypothesis 2: A larger size of TV indicates a higher average customer rating, ceteris paribus;

Hypothesis 3: Selling price and original price contribute same to the average customer rating, ceteris paribus;

make hypothesis more, keep reasonable

Assumption:
1. Assuming dataset is true and reliable;
2. Assuming consumers are free to score by their own will.

c) describe your data (year of collecting the sample, the number of observations, country or region, names of variables used in the text).

Raw dataset is collected from an e-commerce website Flipkart of Television Brands available in the Indian Market 2021 using web scraping technique, contains 912 samples with 7 attributes, with some missing values. Our team find it in Kaggle, see:
https://www.kaggle.com/datasets/devsubhash/television-brands-ecommerce-dataset.
After data preprocessing, we got valid observations of 686 ($N = 686$), collected from e-commerce website Flipkart market in Indian, 2021, see:
https://github.com/Lecter314/Econometrics-HW1/blob/main/TV_dataset_preprocessing1.1.csv.

Names of variables used are as follows:

Table 1 Variables

Notation	Explanations
rate	Rating: average customer ratings on a scale of 5
brand	Brand: the manufacturer of the product "Others": 0, "SAMSUNG": 1, "LG": 2, "SONY": 3, "TCL": 4
reso	Resolution: the type of display, assign higher numbers for higher resolution "HD LED": 0, "Full HD LED": 1, "Ultra HD LED": 2
os	Operating System: the type of OS like Android, Linux, etc. "not Android": 0, "Android": 1
size	Size: the screen size in inches.
sellp	The Selling Price of the product in INR (Indian Rupee)
origp	The Original Price of the product from the manufacturer in INR (Indian Rupee)

data.describe

	brand	reso	size	sellp	origp	os	rate
count	686.0000	686.0000	686.0000	686.0000	686.0000	686.0000	686.0000
mean	0.8367	1.2114	44.6195	48256.4534	68297.5875	0.5656	4.2360
std	1.2165	0.8322	11.6417	47685.4602	64000.9571	0.4960	0.3643
min	0.0000	0.0000	20.0000	4849.0000	7999.0000	0.0000	2.0000
25%	0.0000	0.0000	32.0000	18460.0000	27000.0000	0.0000	4.1000
50%	0.0000	1.0000	43.0000	32999.5000	47994.5000	1.0000	4.3000
75%	2.0000	2.0000	55.0000	56855.0000	79990.0000	1.0000	4.4000
max	4.0000	2.0000	85.0000	324990.0000	409990.0000	1.0000	5.0000

data.info

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 686 entries, 0 to 685
Data columns (total 7 columns):

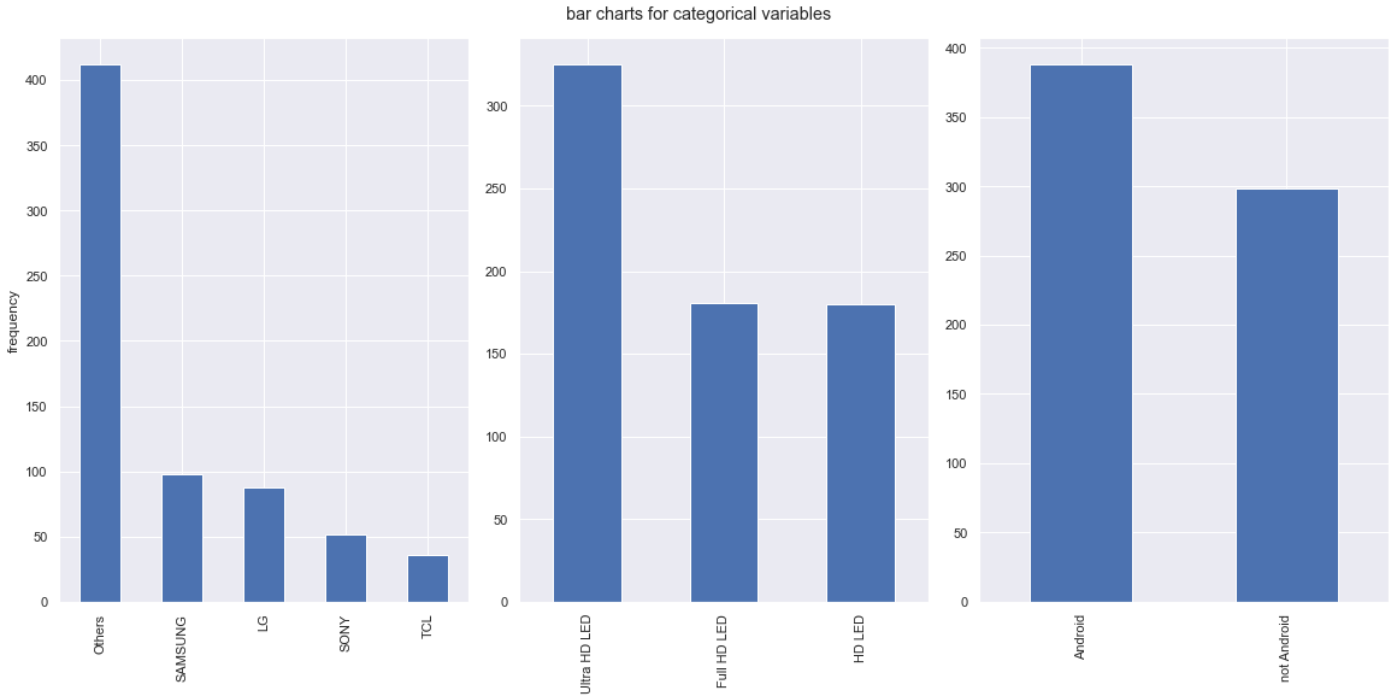
#	Column	Non-Null Count	Dtype
0	brand	686 non-null	int64
1	reso	686 non-null	int64
2	size	686 non-null	int64
3	sellp	686 non-null	int64
4	origp	686 non-null	int64

5	os	686 non-null	int64
6	rate	686 non-null	float64

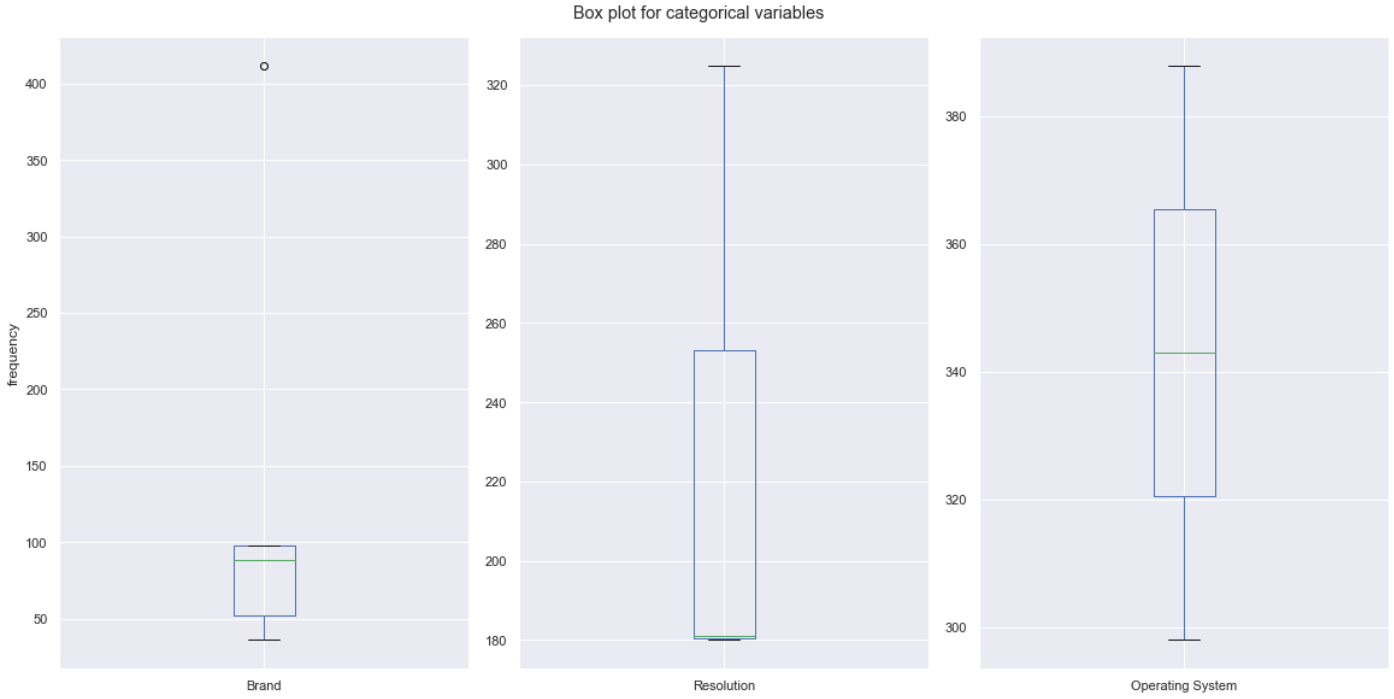
dtypes: float64(1), int64(6)
memory usage: 37.6 KB

d) Provide plots and graphs to illustrate the variables. Use bar charts and box plots to illustrate categorical variables, scatter plots to illustrate the relationships between the variables.

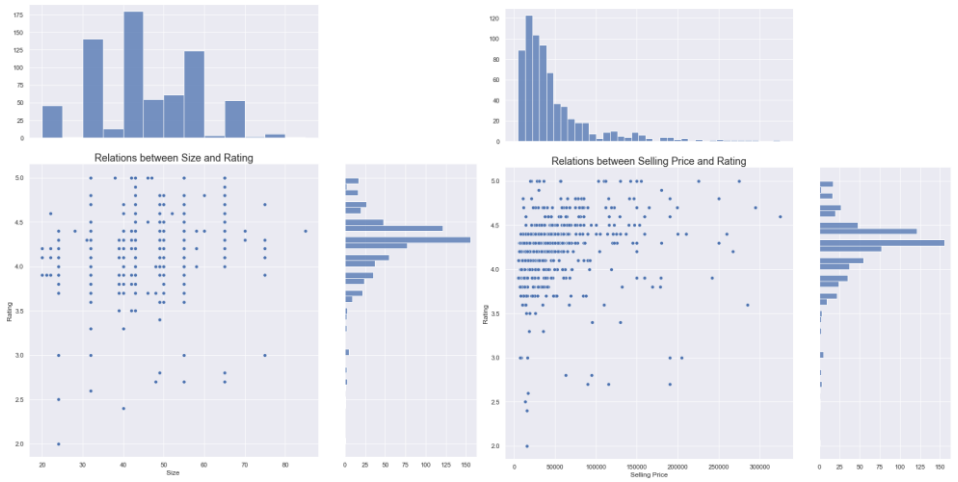
for categorical variables (plz number plots)

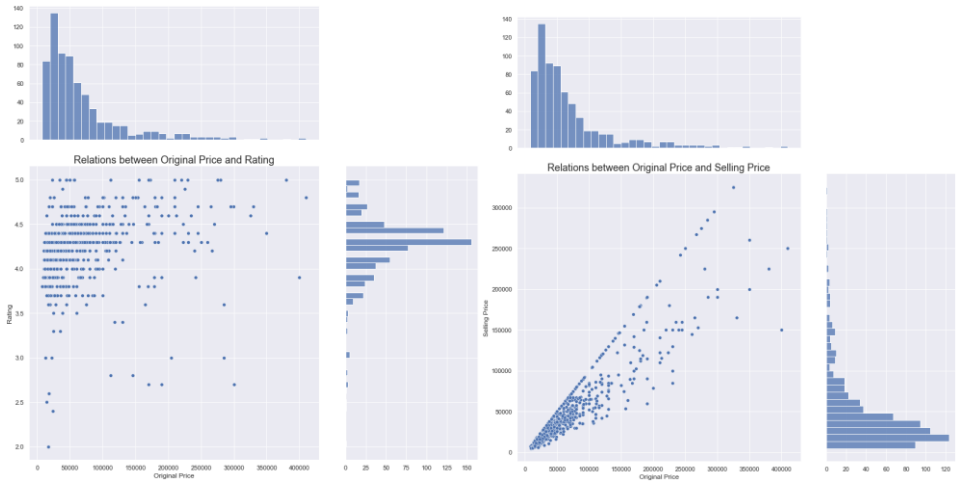


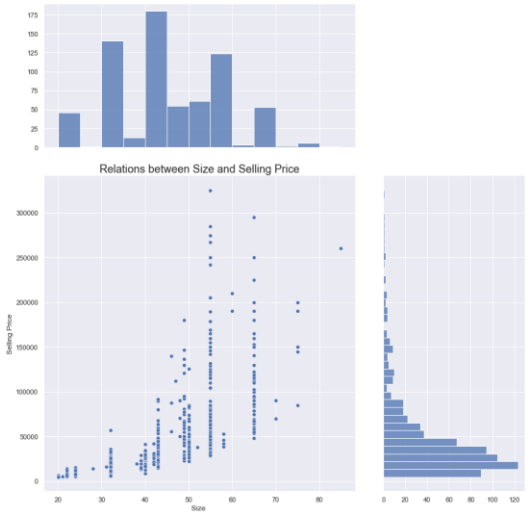
lots of observations are from Others, which cause the outlier in box plot at Brand ($brand$)



illustrate the relationships between the variables (adjust the graph size, do make a short comment here)







Multicollinearity between original price and selling price; size and selling price, the latter looks quadratic

Part 1a. Classical regression analysis (40 points)

a) Write an analytical model and state the assumptions about the nature of explanatory variables and disturbances. The model should include categorical and continuous explanatory variables as discussed

before.

To examine how ratings ($rate$) can be influenced by six explanatory variables (3 qualitative and 3 quantitative), first constructing a very fundamental model as baseline, illustrated as follows:

$$rate_i = \beta_0 + \beta_1 size_i + \beta_2\, sellp_i + \beta_3\, origp_i + \beta_4 brand_i + \beta_5 reso_i + \beta_6 os_i + \varepsilon_i \quad (1)$$

Naively, we hope variables and error term can satisfy Gauss-Markov assumptions (A1-A4) so as to adapt simple OLS method and keep the estimator BLUE. Although the actual dataset is rare likely to strictly satisfy Gauss-Markov assumptions, it's not so bad to assume these as a baseline.

Assumptions as follows (copied form the slides):

$$E\{\varepsilon_i\} = 0, \quad i = 1, \ldots, N \quad \text{(A1)}$$

$$\{\varepsilon_1, \ldots, \varepsilon_N\} \text{ and } \{x_1, \ldots, x_N\} \text{ are independent} \quad \text{(A2)}$$

$$V\{\varepsilon_i\} = \sigma^2, \quad i = 1, \ldots, N \quad \text{(A3)}$$

$$\text{cov}\{\varepsilon_i, \varepsilon_j\} = 0, \quad i, j = 1, \ldots, N, \; i \neq j. \quad \text{(A4)}$$

Try to make a linear regression, well, basically explained nothing in a R2 and each variable significance sense, thankfully the overall model has a valid F score.

```
> summary(model_baseline)

Call:
lm(formula = rate ~ size + sellp + origp + brand + reso + os,
    data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-2.15567 -0.11105  0.04639  0.18231  0.89356

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.969e+00  7.696e-02  51.574  < 2e-16 ***
size         3.652e-03  2.441e-03   1.496 0.135059
sellp        7.572e-07  7.578e-07   0.999 0.318015
origp       -6.053e-07  6.249e-07  -0.969 0.333089
brand        4.874e-02  1.254e-02   3.888 0.000111 ***
reso         5.404e-02  2.678e-02   2.018 0.044004 *
os           4.625e-03  3.114e-02   0.149 0.881982
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3502 on 679 degrees of freedom
Multiple R-squared:  0.0839,    Adjusted R-squared:  0.07581
F-statistic: 10.36 on 6 and 679 DF,  p-value: 5.233e-11
```

b) Propose a linear hypothesis and test it.

Hypothesis: TV $size$ and $brand$ have joint significance at 5% level.
H_0: $\beta_1 = \beta_4 = 0$
H_1: $\beta_1 \neq 0$ *or* $\beta_4 \neq 0$ (in later hypothesis I'll use the expression of *otherwise* for H1 unless it might leads to confusion)

```
> # F-test for testing brand=size=0
> linearHypothesis(model_baseline, c("brand=0", "size=0"))
Linear hypothesis test

Hypothesis:
brand = 0
size = 0

Model 1: restricted model
Model 2: rate ~ size + sellp + origp + brand + reso + os

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    681 85.397
2    679 83.292  2    2.1051 8.5806 0.0002088 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

F_obs = 8.5806 > F_cri, Pr(>F) = 0.0000 < 0.05, reject H0, β_1 and β_4 is significant from 0, which indicates TV $size$ and $brand$ do have joint significance at 5% level.

c) Propose an alternative model which can be considered as an extended version of the baseline model. Conduct F-tests to choose between the models.

Here we denote a new variable Discount Rate, which is (Original Price - Selling Price) / Original Price,

$$disr = (origp - sellp)/origp$$

What's more, we convert brand into 5 dummy variables because this category variable doesn't specifically contain orders as resolution or os, which means this approach might be more appropriate, dummy variable as:

Table 2 Dummy Variables for brand

Notation	Explanations
brand_0	"Others": 1, not "Others": 0
brand_1	"SAMSUNG": 1, not "SAMSUNG": 0
brand_2	"LG": 1, not "LG": 0
brand_3	"SONY": 1, not "SONY": 0
brand_4	"TCL": 1, not "TCL":

Introduce only 3 (drop brand_0 for "Others" and brand_4 for "TCL") to avoid perfect multicollinearity (because brand still in the regression!).

Consider these two together, we get the alternative model:

$$rate_i = \beta_0 + \beta_1 size_i + \beta_2\, sellp_i + \beta_3\, origp_i + \beta_4 brand_i + \beta_5 reso_i + \beta_6 os_i + \beta_7\, \text{disr}_i + \beta_8 brand\_1_i + \beta_9 brand\_2_i + \beta_{10}\, brand\_3_i + \varepsilon_i \quad (2)$$

```
> # add all of these above into the baseline model
> model_1a_c <- lm(rate ~ size + sellp + origp + brand + reso + os +
+                    + disr + brand_1 + brand_2 + brand_3, data = df)
> summary(model_1a_c)

Call:
lm(formula = rate ~ size + sellp + origp + brand + reso + os +
    +disr + brand_1 + brand_2 + brand_3, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-2.15349 -0.09642  0.04903  0.16510  0.89947

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.963e+00  9.334e-02  42.463  < 2e-16 ***
size         4.058e-03  2.441e-03   1.663   0.0969 .
sellp       -5.563e-07  1.104e-06  -0.504   0.6145
origp        1.800e-07  8.656e-07   0.208   0.8353
brand        6.840e-03  1.550e-02   0.441   0.6591
reso         5.113e-02  2.651e-02   1.928   0.0542 .
os           3.150e-02  3.931e-02   0.801   0.4232
disr        -1.025e-01  1.145e-01  -0.895   0.3710
brand_1      9.436e-02  5.457e-02   1.729   0.0842 .
brand_2      9.274e-02  5.966e-02   1.554   0.1205
brand_3      3.246e-01  6.812e-02   4.765 2.32e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3446 on 675 degrees of freedom
Multiple R-squared:  0.1186,    Adjusted R-squared:  0.1056
F-statistic: 9.086 on 10 and 675 DF,  p-value: 3.723e-14
```

Let's compare extended model with the baseline, using F-fest, which can construct hypothesis as:

$H_0: \beta_7 = \beta_8 = \beta_9 = \beta_{10} = 0$

$H_1: otherwise$

```
> # test which model to choose using F-test
> linearHypothesis(model_1a_c, c("disr=0", "brand_1=0", "brand_2=0", "brand_3=0"))
Linear hypothesis test

Hypothesis:
disr = 0
brand_1 = 0
brand_2 = 0
brand_3 = 0

Model 1: restricted model
Model 2: rate ~ size + sellp + origp + brand + reso + os + +disr + brand_1 +
    brand_2 + brand_3

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    679 83.292
2    675 80.134  4    3.1582 6.6506 2.979e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

F_obs = 6.6506 > F_cri, Pr(>F) = 0.0000 < 0.05, reject H0, $\beta_7, \beta_8, \beta_9, \beta_{10}$ are significant from 0, which indicates $disr$ and 3 dummy brand variables have joint significance at 5% level, so these factors do contain information that can explain rank which baseline model omits. Ergo, the extended model is better, choose the new model.

Part 1b.

d) Explain why the model can be misspecified. What are the consequences of misspecification on the estimates of the model?

Well…there're dozen, especially when analyzing real-life dataset. From the above we assume all of our variables satisfy Gauss-Markov assumptions, however, it's highly unlikely.

For example, it's hard to unearth all the influential variables for rate, which would make some of them hide

in the error term, cause problems of heteroscedasticity, autoregression, endogeneity, which would make our estimator not BLUE but biased, not efficiency or inconsistency.

What's more, some of the dependent variables are highly correlated, specifically, sellp and origp are highly related, corr = 0.9233, and if we test the VIF in baseline model, it stands our suspicion, which in this case cause problems of multicollinearity. Then lots of bad things would happen like the sign of estimators, poor robustness of the model etc. (check the lecture slides for more)

```
> vif(model_baseline)  # calculate VIFs, sellp = 7.235743, origp = 8.891786, suspected multicollinearity
    size    sellp    origp    brand    reso       os
4.508262 7.291711 8.931353 1.298791 2.773798 1.332368
```

And also, we denote two categorical variables (brand, reso) with linear change other than lots of dummy variables, which may lead to distorted identification of their true relationship. Basically, all of the above would make our model unreliable.
You can have more comments here

e) Explore specification of the model
From here we move to area of CLASS 6. SELECTING THE SET OF REGRESSORS and CLASS 7. CHOOSING THE FUNCTIONAL FORM
i. various sets of explanatory variables;

Then we check the proper regressors via J-test, which means hypothesis as:

M2 + fitted(M1)
H_0: $\delta = 0$, model 1 is nested in model 2
H_1: $\delta \neq 0$, model 1 is not nested in model 2

And do the other side:
M1 + fitted(M2)
H_0: $\delta = 0$, model 2 is nested in model 1
H_1: $\delta \neq 0$, model 2 is not nested in model 1

Actually, one can write the specific regression, but I write in a lazy way, hope it's understandable.

Start with baseline model as Model 1,

```
> # J-test
> jtest(model_1b_e_i1, model_1b_e_i2) # i1 in i2
J test

Model 1: rate ~ size + sellp + origp + brand + reso + os
Model 2: rate ~ size + sellp + reso + os + disr + brand_1 + brand_2 +
    brand_3 + brand_4
                Estimate Std. Error t value Pr(>|t|)
M1 + fitted(M2)  1.00637    0.19476  5.1672 3.13e-07 ***
M2 + fitted(M1) -0.29736    1.43006 -0.2079   0.8353
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Indicates Model 1 is nested in Model 2 and Model 2 is not nested in Model 1, which means switch brand to dummy variables is reasonable, and drop origp and add disr will not loss information about the regression.

Maybe try another one, drop os, the hypothesis stays the same as above,

```
> model_1b_e_i3 <- lm(rate ~ size + sellp + reso
+                     + disr + brand_1 + brand_2 + brand_3 + brand_4, data = df) # a good model we find
> jtest(model_1b_e_i1, model_1b_e_i3) # i1 in 13
J test

Model 1: rate ~ size + sellp + origp + brand + reso + os
Model 2: rate ~ size + sellp + reso + disr + brand_1 + brand_2 + brand_3 +
    brand_4
                Estimate Std. Error t value  Pr(>|t|)
M1 + fitted(M2)  1.02022    0.19845  5.1408 3.583e-07 ***
M2 + fitted(M1) -0.18136    1.42297 -0.1275    0.8986
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

Also, model 1 nested in model 2 and model 2 do not nested in model1, which indicates os is not that important for our regression. Which makes sense, few customer would compare Android or other os like Linux when shopping.

Then why not dropping more?

```
> model_1b_e_i4 <- lm(rate ~ size + sellp
+                     + disr + brand_1 + brand_2 + brand_3 + brand_4, data = df) # drop too much
> jtest(model_1b_e_i1, model_1b_e_i4) # inconclusive
J test

Model 1: rate ~ size + sellp + origp + brand + reso + os
Model 2: rate ~ size + sellp + disr + brand_1 + brand_2 + brand_3 + brand_4
                Estimate Std. Error t value  Pr(>|t|)
M1 + fitted(M2)   1.0054    0.19523  5.1498 3.421e-07 ***
M2 + fitted(M1)   0.7770    0.44694  1.7385   0.08258 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Well from here the result is inconclusive, both models are not nested in each other.

Non-nested F test cannot be adapted here consider the perfect multicollinearity between brand and brand dummies.

J-test indicates dependent variables like size, sellp, reso, disr, brand_1, brand_2, brand_3, brand_4 contains enough information from the baseline model, which means we find the variables for regression more properly.

ii. linear vs. non-linear form of the model;

Using PE test to specific whether to use linear or non-linear form of model,

```
> # PE test
> petest(model_1b_e_ii1, model_1b_e_ii2)      # conduct a PE test
PE test

Model 1: rate ~ size + sellp + reso + disr + brand_1 + brand_2 + brand_3 +
    brand_4
Model 2: ln_rate ~ ln_size + ln_sellp + reso + disr + brand_1 + brand_2 +
    brand_3 + brand_4
                          Estimate Std. Error t value Pr(>|t|)
M1 + log(fit(M1))-fit(M2) -2.26065     5.7359 -0.3941   0.6936
M2 + fit(M1)-exp(fit(M2)) -0.02275     0.4466 -0.0509   0.9594
>
```

Inconclusive for PE test, we turn to economic institution, the value of rate is more straightforward than the percentage one, so naively we stay with the original model. Add more here

iii. Ramsey test;

using the model rate ~ size + sellp + reso + disr + brand_1 + brand_2 + brand_3 + brand_4, data = df #

Convert it to regression form

```
> # iii. Ramsey test;
> model_1b_e_ii3 <- lm(rate ~ size + sellp + reso
+                      + disr + brand_1 + brand_2 + brand_3 + brand_4, data = df) # linear rate
> resettest(model_1b_e_i3, power=2:2) # conduct a RESET test

        RESET test

data:  model_1b_e_i3
RESET = 0.72632, df1 = 1, df2 = 676, p-value = 0.3944
```

p-value = 0.3944 >0.05, cannot reject H0, $\widehat{rate}^2$ is not significant from 0 in 5% level, the linear form of the model is valid.

iv. Chow test.

Write something about chow test, like hypothesis here

Test whether there's a structure break within os, p_value = 0.7308,

Inference: We cannot reject the null hypothesis ->> There is not a structural break

f) Conduct tests or procedures not listed above if you find them relevant;

add something here

g) Using steps of Part 1b choose the final model. Write down an equation with estimated parameters that respects to the chosen model. Interpret the coefficients.

Final model:

$$rate_i = \beta_0 + \beta_1 size_i + \beta_2 sellp_i + \beta_3 reso_i + \beta_4 disr_i + \beta_5 brand\_1_i + \beta_6 brand\_2_i + \beta_7 brand\_3_i + \beta_8 brand\_4_i + \varepsilon_i \quad (3)$$

```
> model_final <- lm(rate ~ size + sellp + reso
+                   + disr + brand_1 + brand_2 + brand_3 + brand_4, data = df)
> summary(model_final)

Call:
lm(formula = rate ~ size + sellp + reso + disr + brand_1 + brand_2 +
    brand_3 + brand_4, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-2.14871 -0.09677  0.05072  0.16802  0.90273

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.973e+00  7.385e-02  53.801  < 2e-16 ***
size         4.415e-03  2.279e-03   1.937   0.0532 .
sellp       -3.531e-07  4.481e-07  -0.788   0.4310
reso         4.928e-02  2.630e-02   1.874   0.0613 .
disr        -8.631e-02  8.164e-02  -1.057   0.2908
brand_1      7.509e-02  4.349e-02   1.727   0.0847 .
brand_2      8.171e-02  4.415e-02   1.851   0.0646 .
brand_3      3.332e-01  5.241e-02   6.358 3.75e-10 ***
brand_4      2.376e-02  6.129e-02   0.388   0.6983
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3442 on 677 degrees of freedom
Multiple R-squared:  0.1177,    Adjusted R-squared:  0.1073
F-statistic: 11.29 on 8 and 677 DF,  p-value: 4.459e-15
```

Interpret the coefficients here

Part 2. Violations of other assumptions (30 points)

a) Explain why multicollinearity can be a problem in general and, in particular, in the final model you have chosen. Test your model for multicollinearity. If there is a problem of multicollinearity, try to solve it.

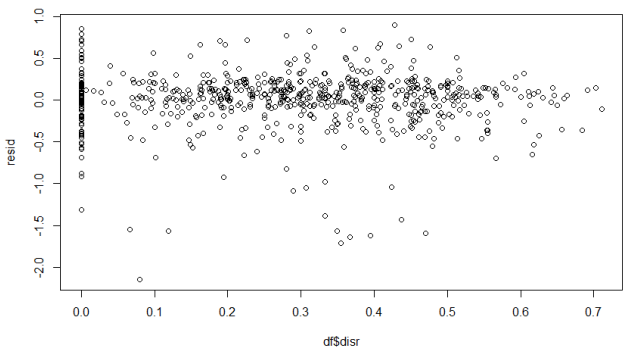
Make it specific, copy paste some theory

Quite suspect the relationship between sellp and origp and disr, our previous work has tried to reduce such effect with several methods, so in the final model there's not suspicious multicollinearity.

```
> # a) multicollinearity
> vif(model_final)
    size    sellp     reso     disr  brand_1  brand_2  brand_3  brand_4
4.070377 2.639156 2.768099 1.216995 1.340594 1.261689 1.113973 1.081332
```

b) Explain what the consequences of heteroscedasticity of residuals for OLS estimates are. Test for heteroscedasticity and make a conclusion. If there is heteroscedasticity describe approaches to deal with it. Run an estimation that provides better than OLS estimates if needed. Provide comments on the resulting estimates of the parameters and standard errors.

Make it specific, copy paste some theory



Goldfeld Quandt Test passed

```
> # Goldfeld Quandt Test
> gqtest(model_2_a,                  # p-value = 0.9875
+        order.by = df$sellp,  # for which variable do we sort our observations
+        fraction = 0.25)       # which fraction from the center we ELIMINATE

        Goldfeld-Quandt test

data:  model_2_a
GQ = 1.2354, df1 = 249, df2 = 248, p-value = 0.04819
alternative hypothesis: variance increases from segment 1 to 2
```

Breusch Pagan Test passed

```
> # Breusch Pagan Test (include only linear dependence by default)
> bptest(model_2_a) # p-value = 0.0000

        studentized Breusch-Pagan test

data:  model_2_a
BP = 37.229, df = 8, p-value = 1.044e-05
```

White Test, passed

```
> # White Test
> bptest(model_2_a, # p-value = 0.0032
+         varformula = ~ I(size^2) + I(sellp^2) + reso + disr + brand_1 + brand_2 + brand_3 + brand_4,
+         data = df)

        studentized Breusch-Pagan test

data:  model_2_a
BP = 31.047, df = 8, p-value = 0.0001378
```

c) Write the best model after all the steps of the analysis. Provide comments on the interpretation of the coefficients and goodness-of-fit. Are the stated in Introduction hypotheses confirmed by empirical evidence?

You know what to do next

Conclusion (15 points)

Provide a conclusion and a discussion to the analysis performed. Namely:

a) State the problem and the hypotheses you have investigated;

b) Briefly describe the features of data you have used;

c) Make salient the features of the model you have built;

d) Describe the results you have found. Do they support or contradict the stated hypothesis?

e) Could your results be applied for policy making?

Appendix (5 points)
With a .ipynb file and a .r file

I believe you guys can make this, good luck!