Assignment 2

108048110

2022-10-12

Assignment 2

1. Yuan data set

Table 1: Chinese yuan data set for 17 factories in Shanhai.

Item	Variable	Description
1	Output	Per capita output in Chinese yuan
2	SI	Number of workers in the factory
3	SP	land area of the factory in square meters per worker
4	I	Investment in yuans per worker

Missing values check

Output	SI	SP	I
Mode :logical	Mode :logical	Mode :logical	Mode :logical
FALSE:17	FALSE:17	FALSE:17	FALSE:17

A quick look at the data

Output	SI	SP	I
Min. :11360	Min. : 56.0	Min. : 840	Min. :10.54
1st Qu.:12930	1st Qu.: 408.0	1st Qu.:2480	1st Qu.:12.45
Median :16680	Median : 805.0	Median :2840	Median :14.74
Mean :18348	Mean : 856.9	Mean :2859	Mean :16.94
3rd Qu.:20030	3rd Qu.:1217.0	3rd Qu.:3200	3rd Qu.:19.52
Max. :30750	Max. :1754.0	Max. :4240	Max. :29.19

Conjectures and Suppositions

- Output: Quantitative continuous; SI: Quantitative Discrete; SP: Quantitative continuous; I: Quantitative continuous
- Consider the relationship between the response and other explanatory variables.
  - The more workers are in a factory, the higher output per capita in yuan.
- Consider correlations between explanatory variables.
  - The more workers there are in a factory, the less place workers can have.

a. Fit a model using least squares, express "output" in terms of other variables

$$Output = \beta_0 + \beta_1 * SI + \beta_2 * SP + \beta_3 * I$$

```
##
## Call:
## lm(formula = Output ~ SI + SP + I, data = Yuan_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6638.7 -3578.0  -558.5  4011.6  9637.4
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6026.061   5245.659   1.149   0.2713
## SI             1.742      3.777   0.461   0.6523
## SP             5.302      2.188   2.423   0.0307 *
## I           -255.506    333.194  -0.767   0.4569
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5262 on 13 degrees of freedom
## Multiple R-squared:  0.4174, Adjusted R-squared:  0.2829
## F-statistic: 3.104 on 3 and 13 DF,  p-value: 0.06371
```

b. Add SI^2 and $SP * I$ to obtain another model

$$Output = \beta_0 + \beta_1 * SI + \beta_2 * SP + \beta_3 * I + \beta_4 * SI^2 + \beta_3 * SP * I$$

```
##
## Call:
## lm(formula = Output ~ SI + SP + I + SI_square + SP_and_I, data = Yuan_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4821.1 -1766.4  -316.4  1032.9  5638.4
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.240e+04  1.354e+04   3.869  0.00261 **
## SI           3.513e+01  1.029e+01   3.414  0.00579 **
## SP          -1.372e+01  4.978e+00  -2.755  0.01871 *
## I           -3.716e+03  1.001e+03  -3.710  0.00344 **
## SI_square   -1.454e-02  4.894e-03  -2.971  0.01273 *
## SP_and_I     1.022e+00  2.841e-01   3.597  0.00419 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3560 on 11 degrees of freedom
## Multiple R-squared:  0.7743, Adjusted R-squared:  0.6717
## F-statistic: 7.548 on 5 and 11 DF,  p-value: 0.002667
```

- Adding more features does improve the performance of the model regardless of the feature's effect on the response.

c. Use the model acquired from *b.*, find out the value of SP, SI, I that maximize per capita *output*.

- SP, I, SI^2 these variables have a large negative effect on the response.

	x
12	28566.52
11	27307.13
13	24551.52
9	22761.08
14	22432.81
10	21352.74
8	20843.11
15	17729.12
6	17451.55
5	16996.37
17	14909.07
3	13950.52
16	13726.44
4	13060.41
7	12860.51
1	12689.03
2	10722.05

- Among 17 factories, factory number 12 has the maximum output. So I dig in to find its SP, SI and I.

	Output	SI	SP	I	SI_square	SP_and_I
12	30750	1181	4240	21.21	1394761	89930.4

2. Prostate data set

Table 6: Prostate dataset

Item	Variable	Description
1	lcavol	$log(size_{tumor})$
2	lweight	$log(weight_{tumor})$
3	age	patient age
4	lbph	$log(Hyperplasia_{prostatic})$
5	svi	seminal vesicle invasion
6	lcp	$log(increment_{penetration})$
7	gleason	Gleason score
8	pgg45	percentage Gleason score
9	lpsa	$log(PSA)$

- SVI - seminal vesicle invasion:

  the presence of prostate cancer in the areolar connective tissue around the seminal vesicles and outside the prostate.

- LCP - log capsular penetration:

  cancer that has reached the outer wall of the prostate.

- Gleason - gleason score

  The most common system doctors use to grade prostate cancer, the grade of a cancer tells you how much the cancer cells look like normal cells. (10: very abnormal)

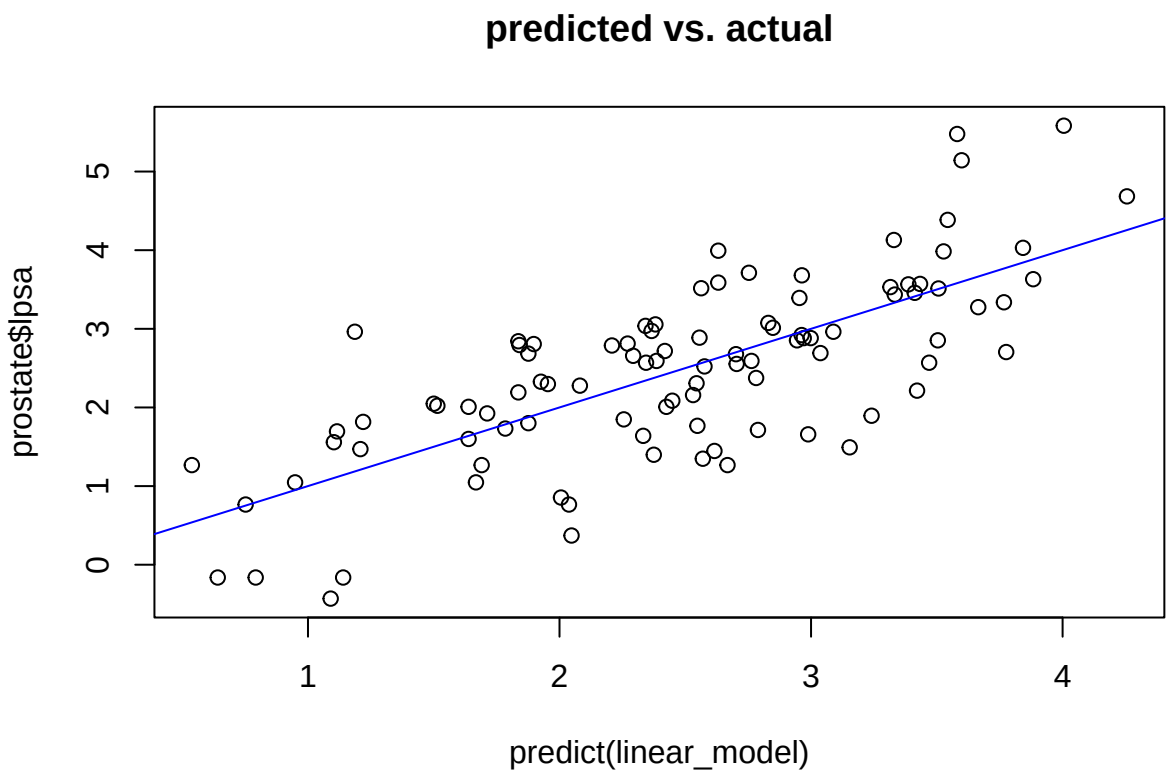
  Grades: low= 6; medium= 7; high= 8~10

- lpsa - log prostate specific antigen

  The PSA test is used to monitor men after surgery or radiation therapy for prostate cancer to see if their cancer has recurred (come back).

a. Fit a model with *lpsa* as the response and *lcavol* as the predictor. Report the residual standard error and the R^2.

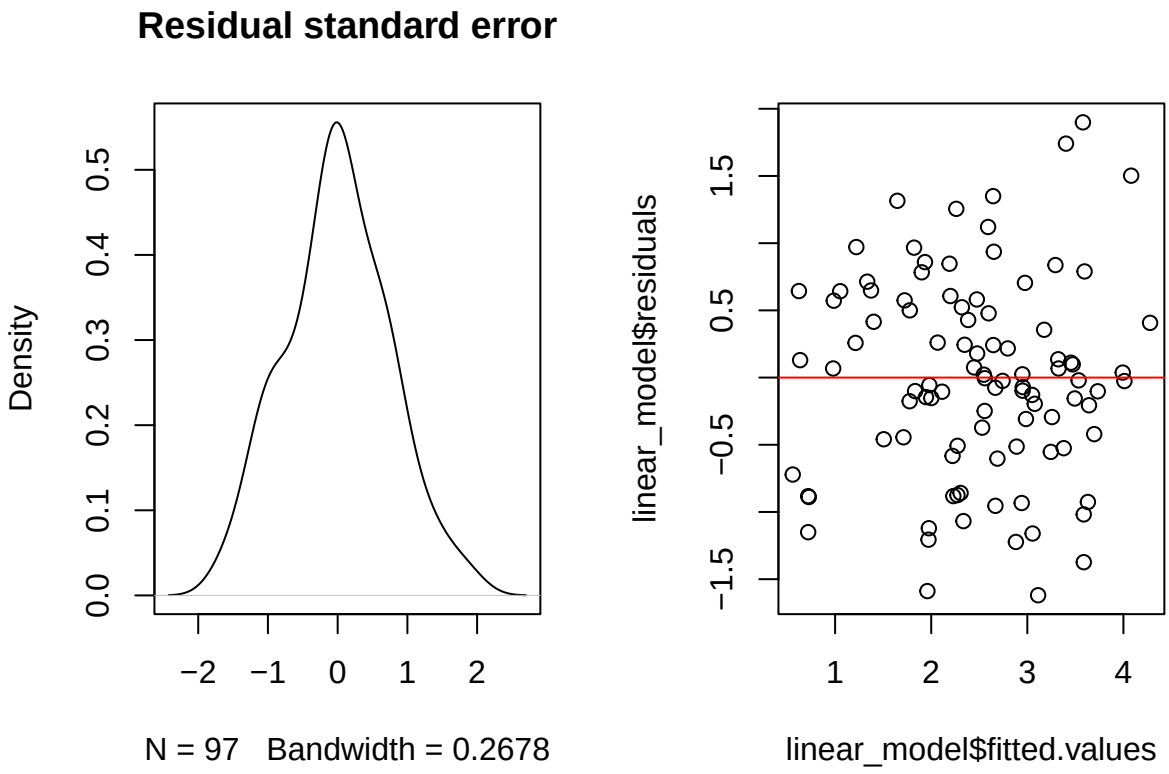
- Residual standard error: 78.75%
- $R^2 = 53.94\%$



b. Now add *lweight*, *svi*, *lbph*, *age*, *lcp*, *pgg45*, and *gleason* to the model *one at a time*. For each model record the residual standard error and the R^2. Plot the trends in these two statistics and comment on any features that you find interesting.

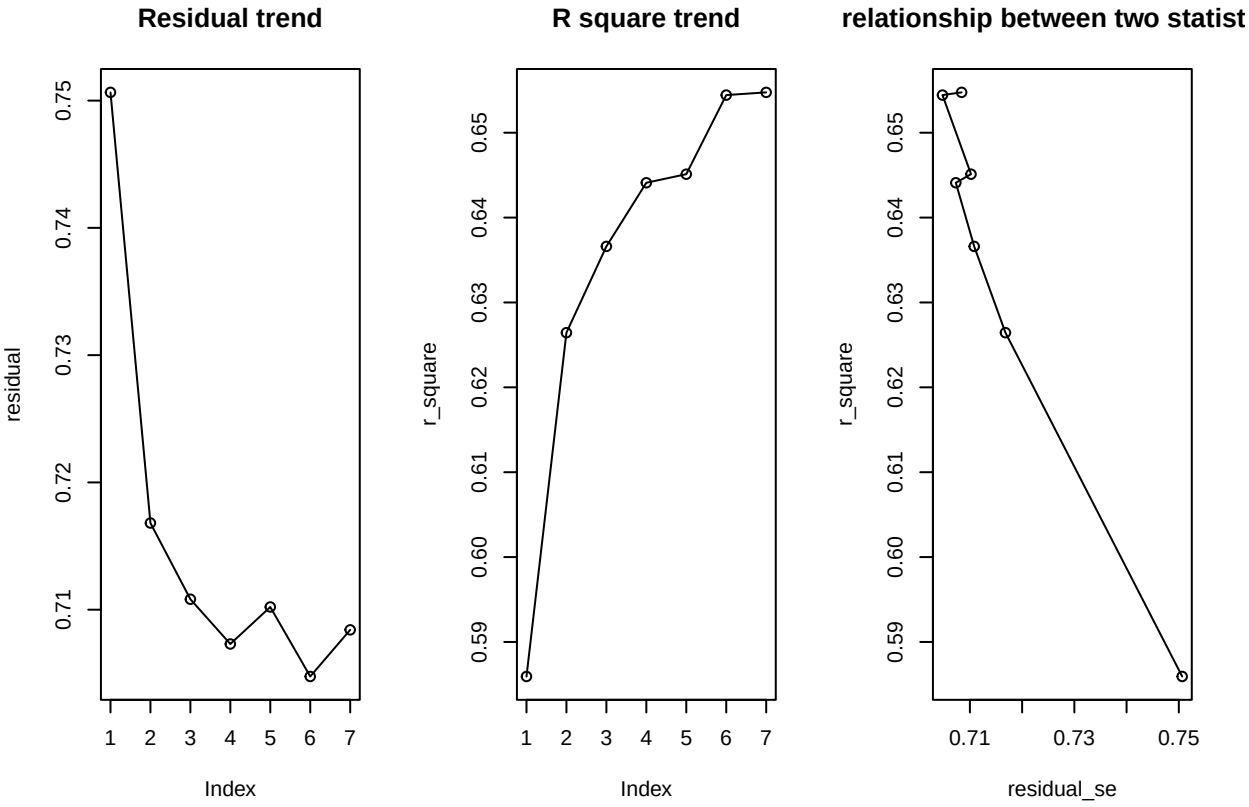
- Adding *lweight*

  - Residual standard error: 75.06%
  - $R^2 = 58.59\%$
- Adding SVI
  - Residual standard error: 71.68%
  - $R^2 = 62.64\%$
- Adding $lbph$
  - Residual standard error: 71.08%
  - $R^2 = 63.66\%$
- Adding age
  - Residual standard error: 70.73%
  - $R^2 = 64.41\%$
- Adding lcp
  - Residual standard error: 71.02%
  - $R^2 = 64.51\%$
- Adding $pgg45$
  - Residual standard error: 70.48%
  - $R^2 = 65.44\%$



- Adding *gleason*
  - Residual standard error: 70.84%
  - $R^2 = 65.48\%$

Residual standard error	R square
0.7506469	0.5859345
0.7168094	0.6264403
0.7108232	0.6366035
0.7073054	0.6441024
0.7102135	0.6451130
0.7047533	0.6544317
0.7084155	0.6547541



- As we can observe from the plot, residual standard error had a huge dump after adding *lweight* to the model, and the standard error kept decreasing afterwards until I added variable *lcp*.
- Additionally, according to the professor both R^2 and residual standard error are two key goodness-of fit measures for regression analysis. Yet, I found that the increment of R^2 does not necessarily represents the reduction in residual standard error.
  - While residual standard error is defined as the standard deviation of the residuals, intuitively, I thought the smaller residual standard error was, the closer predicted values are to actual values, the better the model fits a dataset.

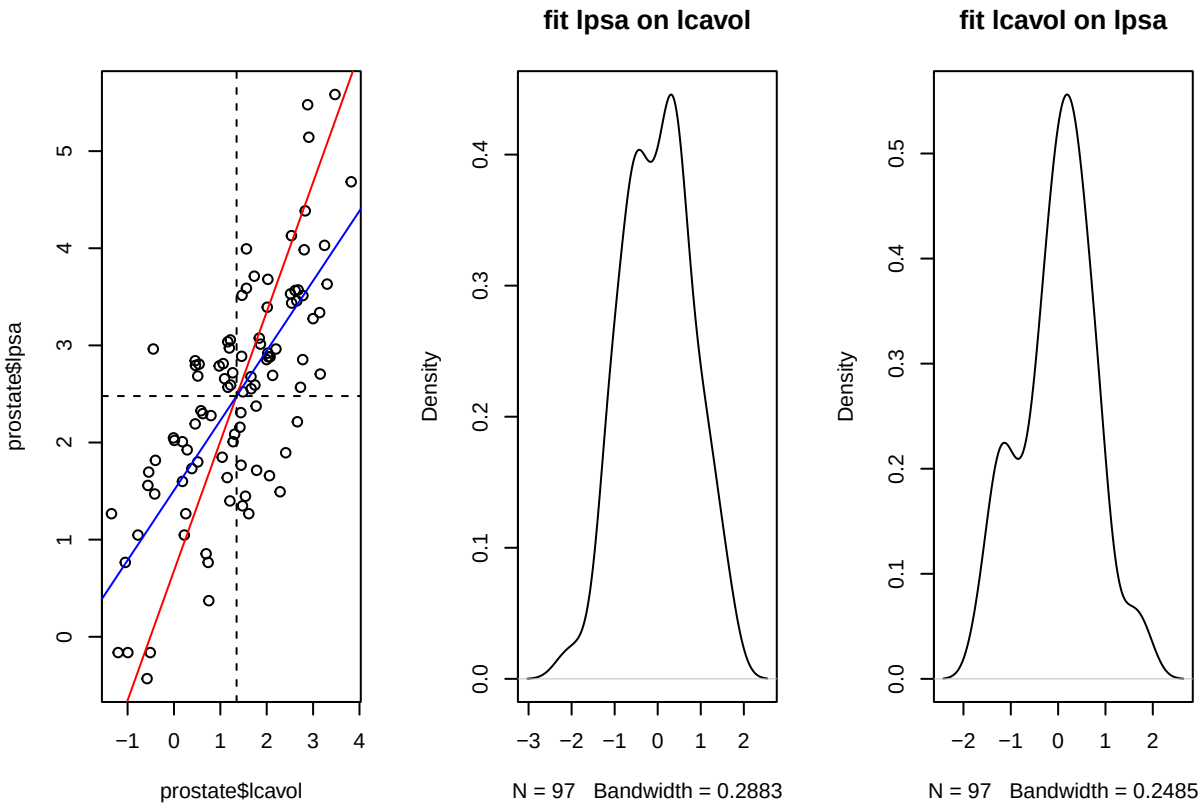
$$e_i = y_i - \hat{y}_i \sigma = sd(\frac{\sum e_i^2}{n-p})$$

- As for R^2 this value can be interpreted as proportion of variation in the response variable *lpsa* , accounted for by the model. If the data collection process does not involved any manipulation, then larger value of R^2 usually means a better model.

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum(\hat{y}_i - y_i)^2}{\sum(y_i - \bar{y_i})^2}$$

- After given some thoughts, for R^2 , it measures the percentage the regression model explain of the variance, higher R^2 values indicate the predicted data points are closer to the actual points. While you can simply increase the value of R^2 even if the added predictors are irrelevant to the response, R^2 values do not tell you how far exactly the data points are from the regression line. On the other hand, residual standard error signifies the distances between fitted values and actual data, these values could tell you how precise the model predictions are using the units of the predictors.
- In conclusion, residual standard error is in the units of the predictors, as a result, it can provide a more concrete insight about your prediction, while R^2 does not have any units, it only measures the how much variance is explained by the model over total variance. Hence, an increase in R^2 values do not necessarily imply a better fit unless you have further information about the residuals.

c. Plot *lpsa* against *lcavol*. Fit the simple regressions of *lpsa* on *lcavol* and *lcavol* on *lpsa*.



- Two lines intersect at the mean point of each variables as we can verify by the plot above.

- I found that two regression lines have the same R^2 values, they are both 78.75%. However, two residual standard errors are not the same, using *lpsa* to fit response *lcavol* has a higher standard error value.
- If we analyze the goodness of a model simply by observing , we would conclude that the two regressions are as good. However, they weren't. This discovery not only verifies my previous observations but also reminds us that R^2 is not an objective indicator, the number of variables should also taken into account when measuring performance of regression models.

3. Economic Data

This dataset gives us information about features of a production function, variables including capital, labor, and value added for 3 economic sectors.

The production function states the amount of product that can be obtained from every combination of factors, assuming that the most efficient available methods of production are used. While a Cobb-Douglas production function models the relationship between production output(`Value added`) and production inputs (`Capital and Labor`) it is used to calculate ratios of inputs to one another for efficient production and to estimate technological change in production methods.

Table 8: Economic data for each sectors.

Item	Variable	Description
1	**Year**	Range from 1972 to 1986
2	**Capital (20, 36, 37)** denoted as K	Usually represents the amount of physical capital input.
3	**Labor (20, 36, 37)** denoted as L	Usually represents the amount of labor expended, which is typically expressed in hours.
4	**Value added (20, 36, 37)** denoted as V	The amount of output produced from the inputs K and L.

a. Consider the below model, assuming that the errors are independent, and taking logs of both sides of the above model, estimate β_1, β_2.

$$V_t = \alpha K_t^{\beta_1} L_t^{\beta_2} \epsilon_t log(V_t) = log(\alpha) + \beta_1 log(K_t) + \beta_2 log(L_t) + log(\epsilon_t) = \beta_0 + \beta_1 log(K_t) + \beta_2 log(L_t) + \epsilon_t$$

```
##       (Intercept) log(`Capital 20`)   log(`Labor 20`)
##        25.4928845         0.2268538        -1.4584782
```

- For food and kindred products sector, the estimated

$$\beta_1 = 0.2268538$$

$$\beta_2 = -1.4584782$$

```
##       (Intercept) log(`Capital 36`)   log(`Labor 36`)
##        -1.2332115         0.5260689         0.2543206
```

- For electrical and electronic machinery, equipment and supplies sector, the estimated

$$\beta_1 = 0.5260689$$

$$\beta_2 = 0.2543206$$

```
##       (Intercept) log(`Capital 37`)   log(`Labor 37`)
##        -9.6259339         0.5056509         0.8454644
```

- For transportation equipment sector, the estimated

$$\beta_1 = 0.5056509$$

$$\beta_2 = 0.8454644$$

b. Estimate β_1, β_2 under where $\beta_1 + \beta_2 = 1$.

-

$$\beta_2 = 1 - \beta_1$$

$$log(V_t) = log(\alpha) + \beta_1 log(K_t) + \beta_2 log(L_t) + log(\epsilon_t)$$

$$log(V_t) - log(L_t) = log(\alpha) + \beta_1 (log(K_t) - log(L_t)) + log(\epsilon_t)$$

$$Y_t = log(\alpha) + \beta_1 X_t + log(\epsilon_t)$$

```
##
## Call:
## lm(formula = response ~ predictor)
##
## Coefficients:
## (Intercept)    predictor
##      -3.484        1.290
```

- For food and kindred products sector, under the condition, the estimated

$$\beta_1 = 1.29$$

$$\beta_2 = -0.29$$

```
##
## Call:
## lm(formula = response ~ predictor)
##
## Coefficients:
## (Intercept)    predictor
##     -3.8171       0.9001
```

- For electrical and electronic machinery, equipment and supplies sector, under the condition, the estimated

$$\beta_1 = 0.9001$$

$$\beta_2 = 0.0999$$

```
##
## Call:
## lm(formula = response ~ predictor)
##
## Coefficients:
## (Intercept)    predictor
##   -4.712885     0.009609
```

- For transportation equipment sector, under the condition, the estimated

$$\beta_1 = 0.009609$$

$$\beta_2 = 0.990391$$

c. Sometimes the model

$$V_t = \alpha\gamma^t K_t^{\beta_1} L_t^{\beta_2} \epsilon_t$$

is considered, where γ_t is assumed to account for technological development. Estimate

$$\beta_1, \beta_2$$

for this model.

$$log(V_t) = log(\alpha) + t * log(\gamma) + \beta_1 * log(K_t) + \beta_2 * log(L_t) + log(\epsilon_t)$$

- Since γ is a parameter, the same $log(\gamma)$ is also a parameter, and t is a variable we know, so the model should be estimated by γ, K_t, L_t.
- For food and kindred products sector, under the condition, the estimated

$$\beta_1 = 0.69852$$

$$\beta_2 = -0.63414$$

- For electrical and electronic machinery, equipment and supplies sector, under the condition, the estimated

$$\beta_1 = 0.5403$$

$$\beta_2 = 0.01317$$

- For transportation equipment sector, under the condition, the estimated

$$\beta_1 = 7.4407$$

$$\beta_2 = 6.3788$$

d. Estimate β_1, β_2 in the model in part c, under the constraint $\beta_1 + \beta_2$.

- Now the model become

$$log(V_t) = log(\alpha) + log(\gamma) * t + \beta_1 * log(K_t) + \beta_2 * log(L_t) + log(\epsilon_t)$$

$$log(V_t) - log(L_t) = log(\alpha) + \beta_1(log(K_t) - log(L_t)) + log(\gamma)t + log(\epsilon_t)$$

$$Y_t = \beta_0 + \beta_1 X_1 + log(\gamma)t + \epsilon_t$$

- For food and kindred products sector, under the condition, the estimated

$$\beta_1 = 1.192886$$

$$\beta_2 = -0.192886$$

- For electrical and electronic machinery, equipment and supplies sector, under the condition, the estimated

$$\beta_1 = 1.592027$$

$$\beta_2 = -0.592027$$

- For transportation equipment sector, under the condition, the estimated

$$\beta_1 = 0.33476$$

$$\beta_2 = 0.66524$$