Statistical Modeling for Business Analytics – MBA652A – Project 2

PANEL DATA ANALYSIS : Grunfeld Investment Data

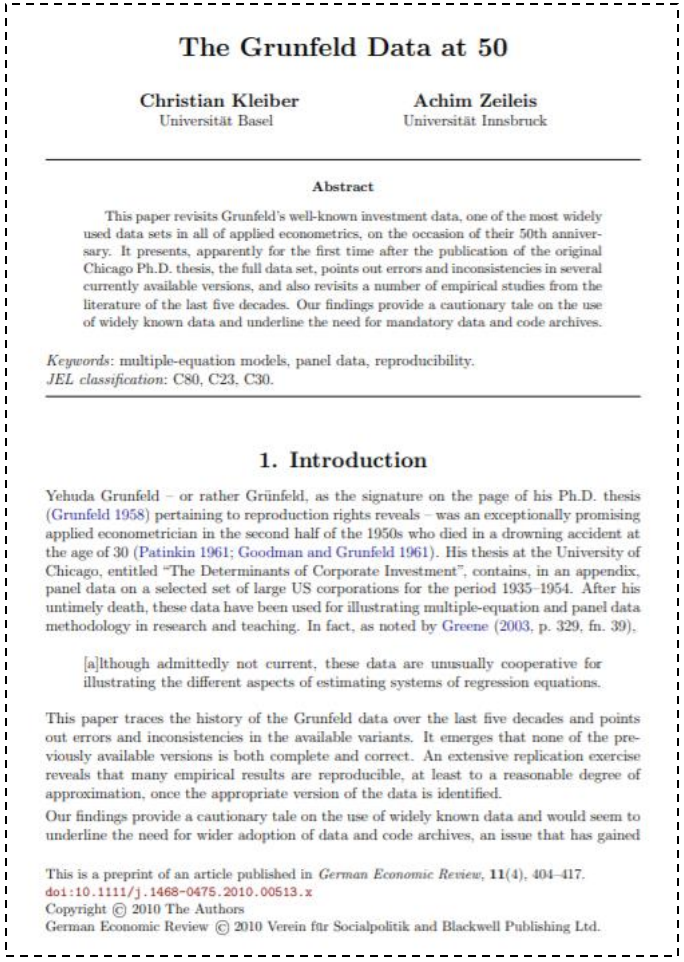
Submitted To:
Prof. (Dr.) Devlina Chatterjee

Submitted By: Group 5

1. Ashish Tiwari (21129004)
2. Jyoti Sharma (21129265)
3. Shiv Shakti Singh (21129024)
4. Shreeyash Nitin Malode (20214271)

Outline of the Presentation

- Introduction
- Objective
- Descriptive Analysis
- Scatter plot & Correlation Matrix
- Models
- Interpretation of Results
- Inference & Conclusion



The Grunfeld Data at 50

Christian Kleiber
Universität Basel

Achim Zeileis
Universität Innsbruck

Abstract

This paper revisits Grunfeld's well-known investment data, one of the most widely used data sets in all of applied econometrics, on the occasion of their 50th anniversary. It presents, apparently for the first time after the publication of the original Chicago Ph.D. thesis, the full data set, points out errors and inconsistencies in several currently available versions, and also revisits a number of empirical studies from the literature of the last five decades. Our findings provide a cautionary tale on the use of widely known data and underline the need for mandatory data and code archives.

Keywords: multiple-equation models, panel data, reproducibility.
JEL classification: C80, C23, C30.

1. Introduction

Yehuda Grunfeld – or rather Grünfeld, as the signature on the page of his Ph.D. thesis (Grunfeld 1958) pertaining to reproduction rights reveals – was an exceptionally promising applied econometrician in the second half of the 1950s who died in a drowning accident at the age of 30 (Patinkin 1961; Goodman and Grunfeld 1961). His thesis at the University of Chicago, entitled "The Determinants of Corporate Investment", contains, in an appendix, panel data on a selected set of large US corporations for the period 1935–1954. After his untimely death, these data have been used for illustrating multiple-equation and panel data methodology in research and teaching. In fact, as noted by Greene (2003, p. 329, fn. 39),

> [a]lthough admittedly not current, these data are unusually cooperative for illustrating the different aspects of estimating systems of regression equations.

This paper traces the history of the Grunfeld data over the last five decades and points out errors and inconsistencies in the available variants. It emerges that none of the previously available versions is both complete and correct. An extensive replication exercise reveals that many empirical results are reproducible, at least to a reasonable degree of approximation, once the appropriate version of the data is identified.

Our findings provide a cautionary tale on the use of widely known data and would seem to underline the need for wider adoption of data and code archives, an issue that has gained

This is a preprint of an article published in *German Economic Review*, **11**(4), 404–417.
doi:10.1111/j.1468-0475.2010.00513.x

Econometric Analysis of Panel Data
Professor William Greene

Main Reference - Kleiber, C., and Zeileis, A. (2010). "The Grunfeld Data at 50." *German Economic Review*, **11**(4), 404–417.

http://dx.doi.org/10.1111/j.1468-0475.2010.00513.x

Dataset Source - Econometric Analysis of Panel Data by NYU

Software used - R & Excel

Image Source – Screenshot of research paper & dataset website - http://people.stern.nyu.edu/wgreene/

Introduction

- A panel data on 10 large US manufacturing firms
- Total time span of data over 20 years, for the years 1935–1954.
- Total number of observations : 200
- Number of different variables: 5 of which two are identifiers
- Variables
  - **FIRM:** company under consideration
  - **YEAR:** the year of observation
  - **INV (I):** gross investment (USD in Mn.)
  - **VALUE(V):** market value of the firm (USD in Mn.)
  - **CAPITAL(C):** value of stock of plant and equipment (USD in Mn.)

Image Source – Computed R Output

```
> str(Gdata)
'data.frame':   200 obs. of  5 variables:
 $ FIRM   : int  1 1 1 1 1 1 1 1 1 1 ...
 $ YEAR   : int  1935 1936 1937 1938 1939 1940
 $ INV    : num  318 392 411 258 331 ...
 $ VALUE  : num  3078 4662 5387 2792 4313 ...
 $ CAPITAL: num  2.8 52.6 156.9 209.2 203.4 ...
```

```
      FIRM
YEAR   1 2 3 4 5 6 7 8 9 10
  1935 1 1 1 1 1 1 1 1 1  1
  1936 1 1 1 1 1 1 1 1 1  1
  1937 1 1 1 1 1 1 1 1 1  1
  1938 1 1 1 1 1 1 1 1 1  1
  1939 1 1 1 1 1 1 1 1 1  1
  1940 1 1 1 1 1 1 1 1 1  1
  1941 1 1 1 1 1 1 1 1 1  1
  1942 1 1 1 1 1 1 1 1 1  1
  1943 1 1 1 1 1 1 1 1 1  1
  1944 1 1 1 1 1 1 1 1 1  1
  1945 1 1 1 1 1 1 1 1 1  1
  1946 1 1 1 1 1 1 1 1 1  1
  1947 1 1 1 1 1 1 1 1 1  1
  1948 1 1 1 1 1 1 1 1 1  1
  1949 1 1 1 1 1 1 1 1 1  1
  1950 1 1 1 1 1 1 1 1 1  1
  1951 1 1 1 1 1 1 1 1 1  1
  1952 1 1 1 1 1 1 1 1 1  1
  1953 1 1 1 1 1 1 1 1 1  1
  1954 1 1 1 1 1 1 1 1 1  1
```

Varying across both entity & time – cross-sectional & time series data

Checking whether data is balanced or not ? **Balanced data**

```
> Gdata %>%
+   is.pbalanced()
[1] TRUE
```

Objective

- The objective is to investigate two determinants of gross investment for a firm viz., market value of the firm and stock of plant and equipment by using different type of panel data regression model on the Grunfeld Investment data.
- **Null Hypothesis (H_0) :** No relationship exist between gross investment (INV) for a firm, and market value (VALUE) of the firm and stock of plant and equipment (CAPITAL) of the firm.
- **Alternate Hypothesis (H_1)** : Relationship exist between gross investment (INV) for a firm, and market value (VALUE) of the firm and stock of plant and equipment (CAPITAL) of the firm.

Image Source – Computed R Output

Mathematical Representation

$$I_{it} = f(V_{it}, C_{it})$$

Where, I_{it} = Gross investment

V_{it} = Market value of the firm= X_{1it}

C_{it} = Capital Stock of Plant and Equipment = X_{2it}

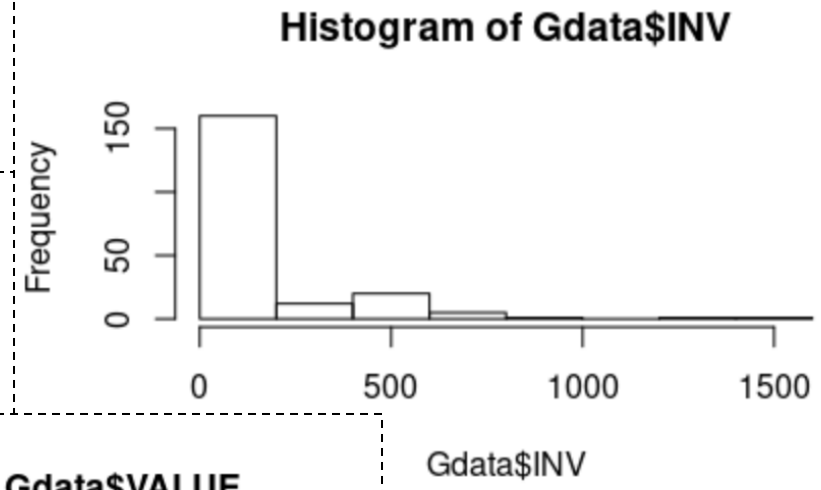
The equation between dependent and independent variable can be written as:-

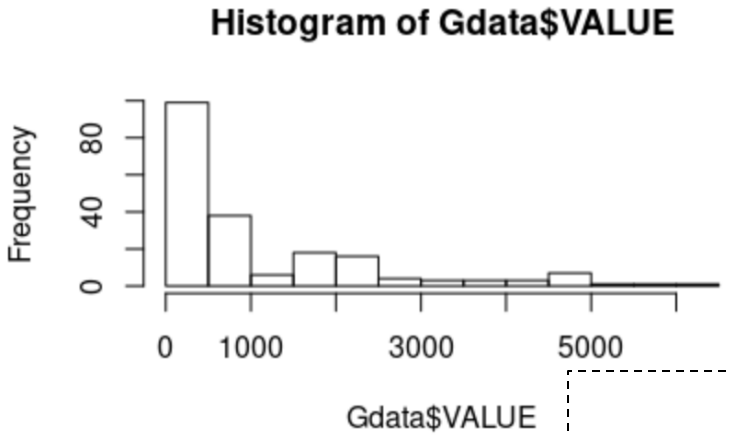
$$Y_{it} = \beta_0 + \beta_{1it} X_{1it} + \beta_{2it} X_{2it} + U_{it}$$

Image Source – Computed R Output

Descriptive Analysis

	INV	VALUE	CAPITAL
Mean	145.96	1081.68	276.02
Standard Error	15.34	92.95	21.29
Mode	361.60	156.70	67.10
Median	57.49	517.95	205.60
First Quartile	33.56	199.98	79.18
Third Quartile	138.04	1679.85	358.10
Variance	47034.89	1727830.58	90663.56
Standard Deviation	216.88	1314.47	301.10
Kurtosis	11.21	2.83	12.05
Skewness	2.92	1.81	2.79
Range	1485.77	6183.58	2225.50
Minimum	0.93	58.12	0.80
Maximum	1486.70	6241.70	2226.30
Sum	29191.65	216336.22	55203.43
Count	200.00	200.00	200.00





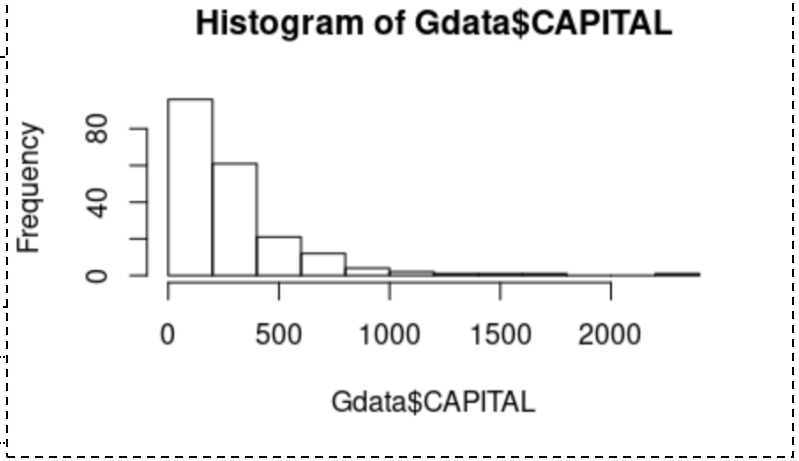
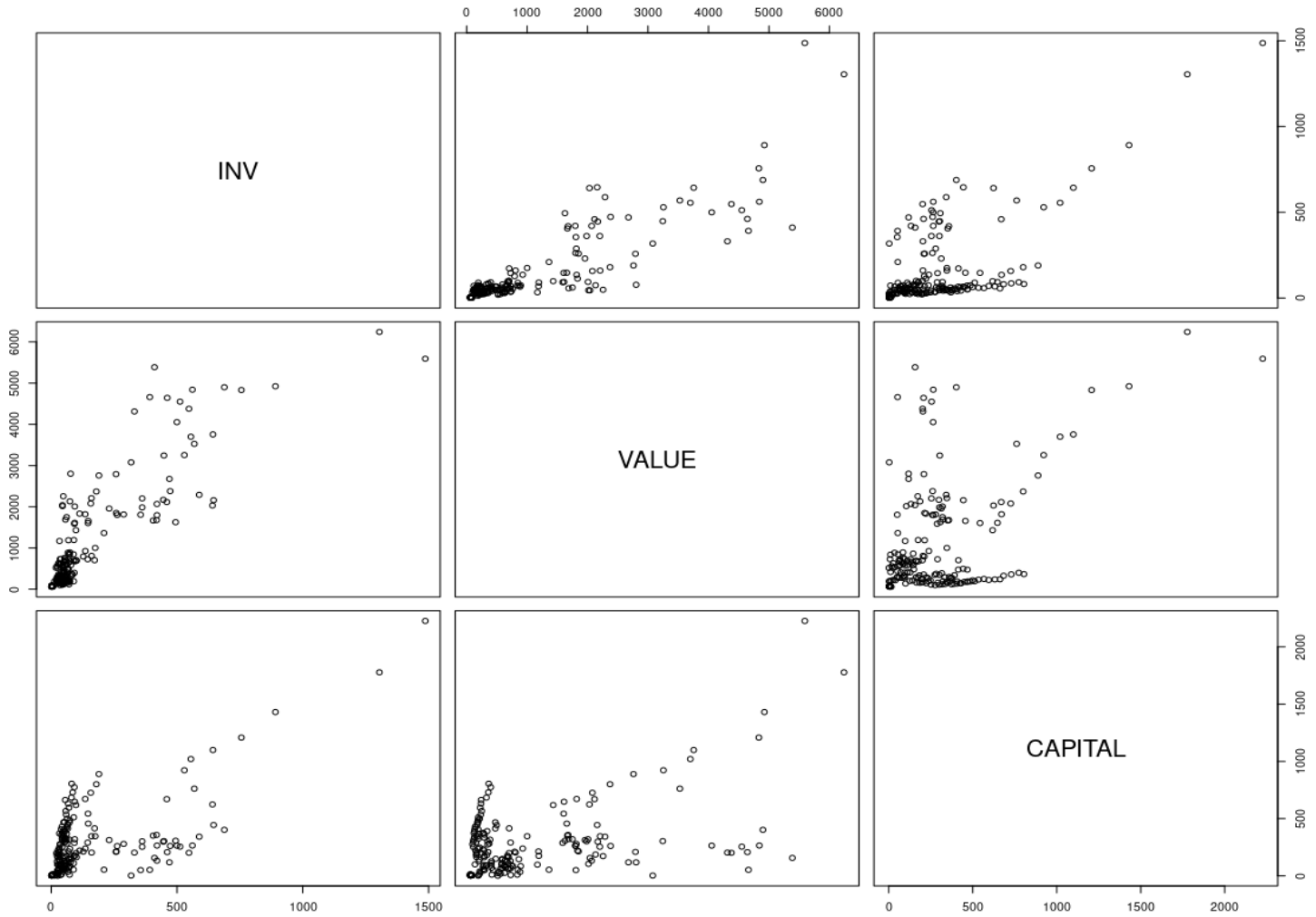


Image Source – Computed R & Excel Output

Scatter Plot and Correlation Matrix



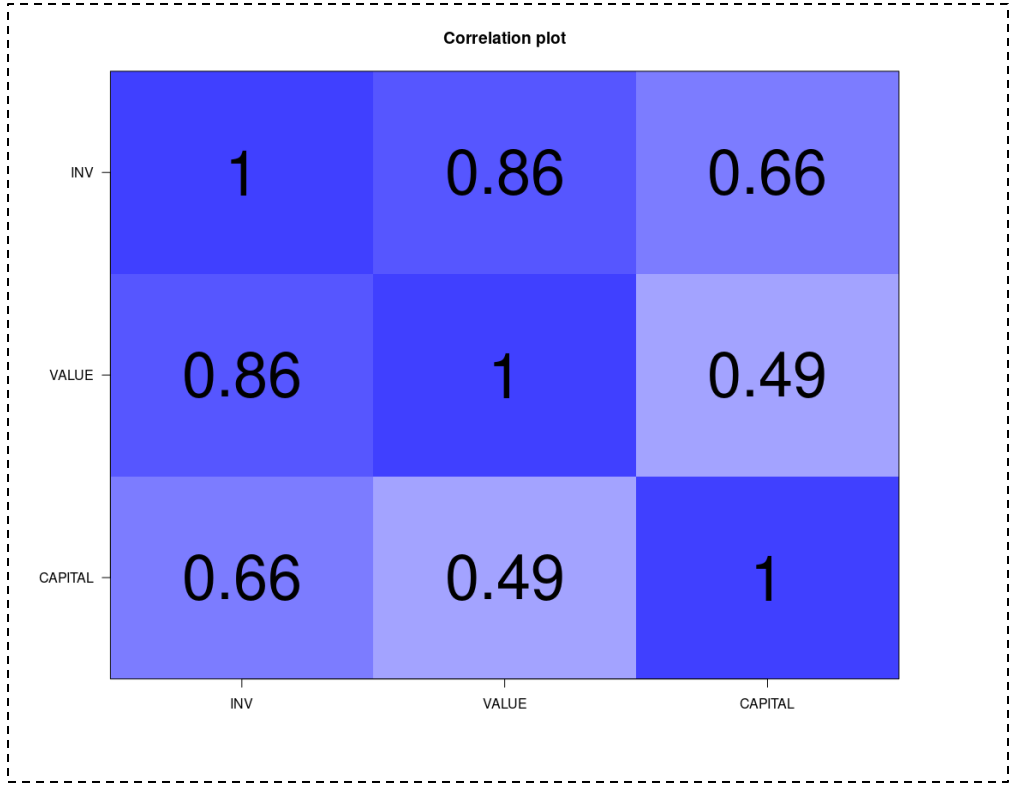
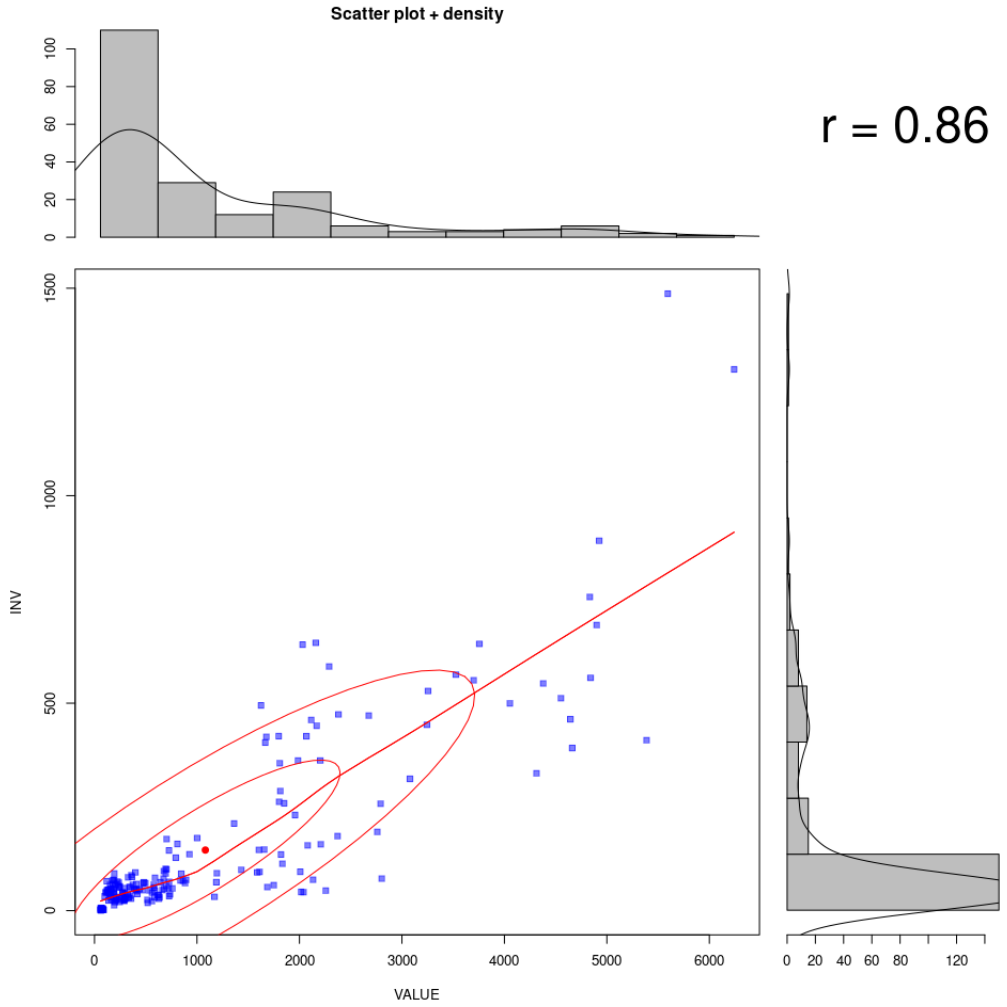


Image Source – Computed R Output

Scatter Plot and Correlation Matrix (Cond.)



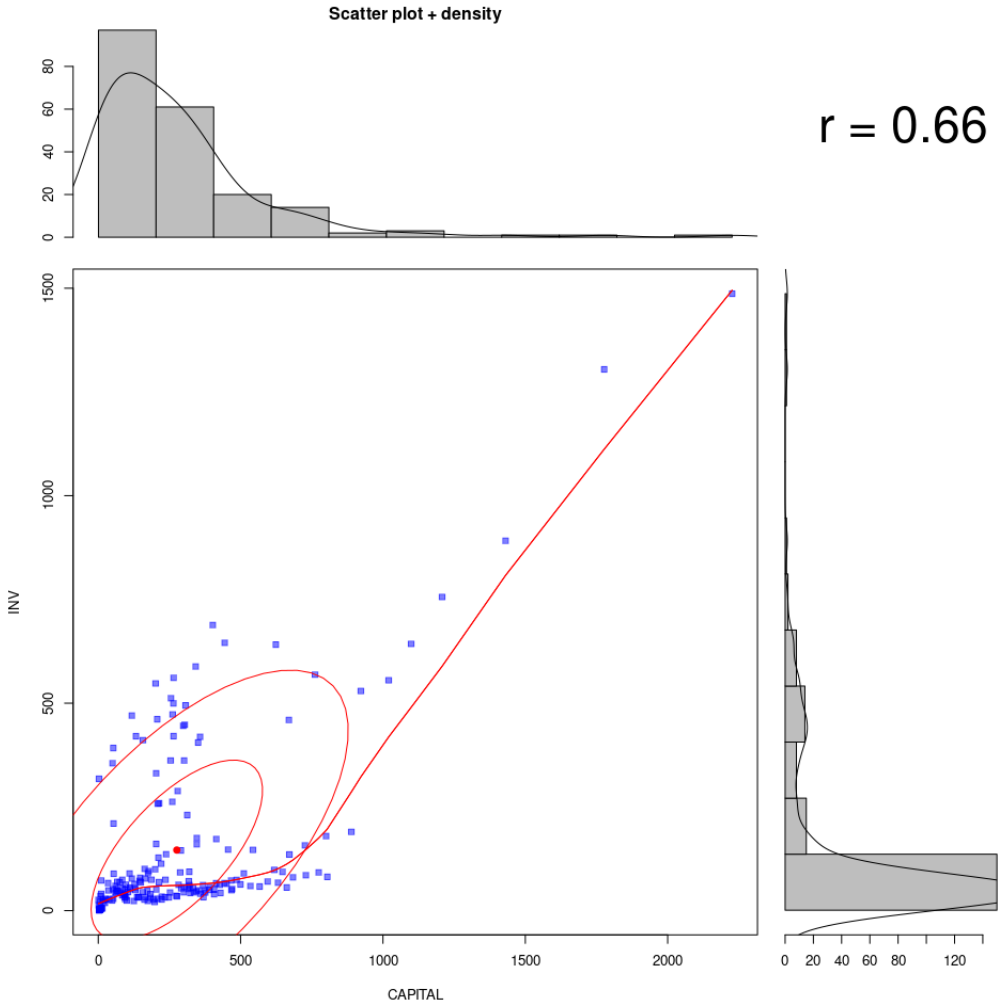
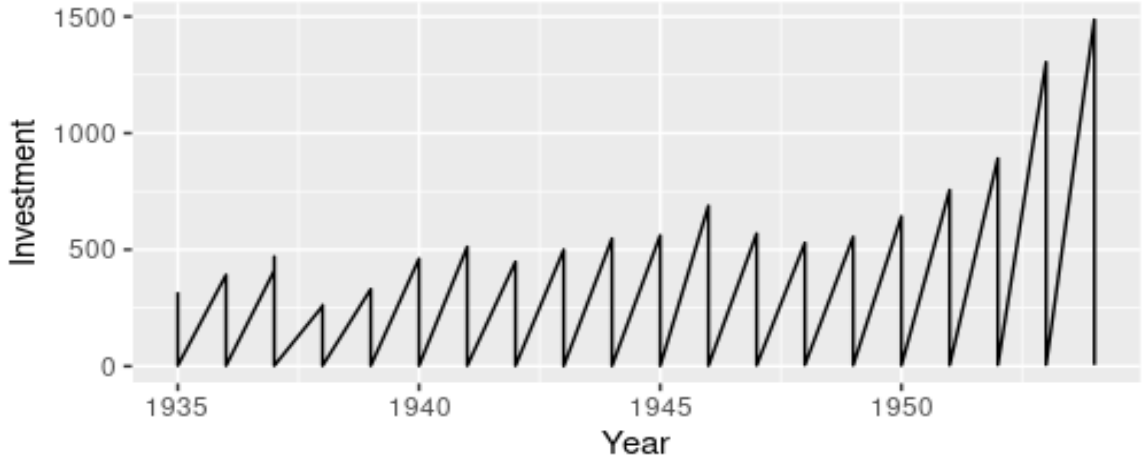
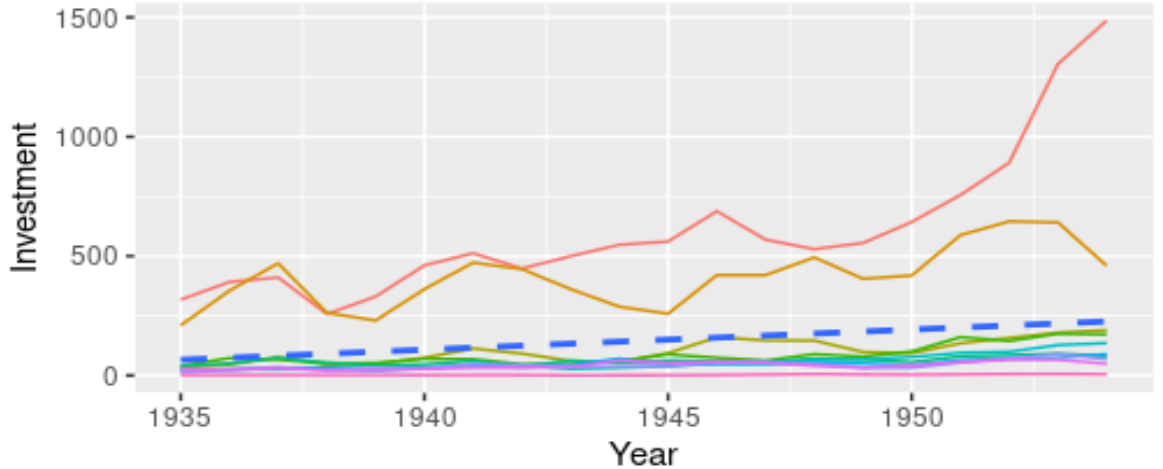
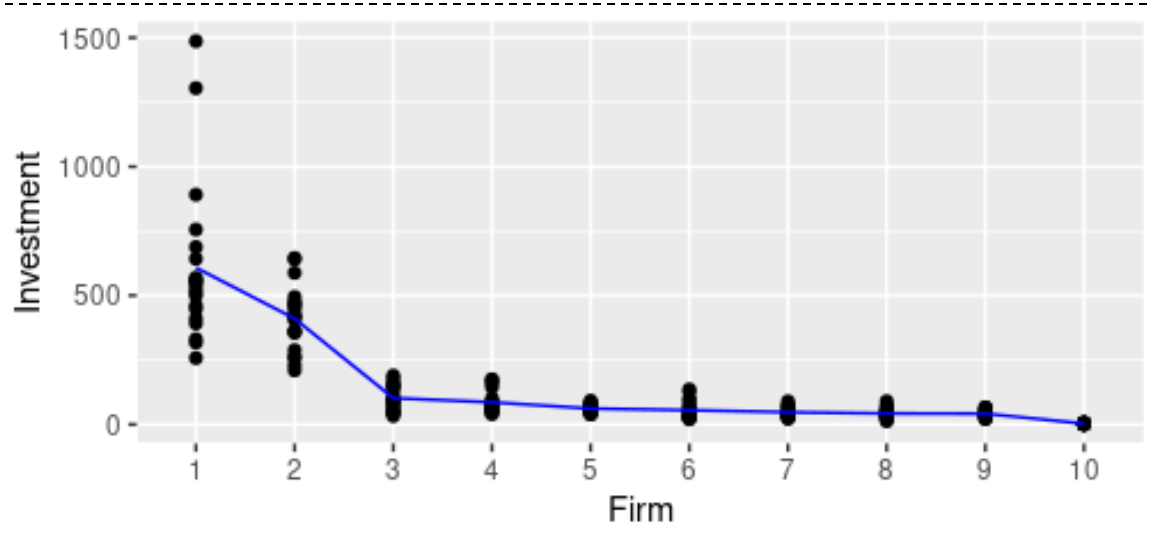


Image Source – Computed R Output

Entity & Time Heterogeneity







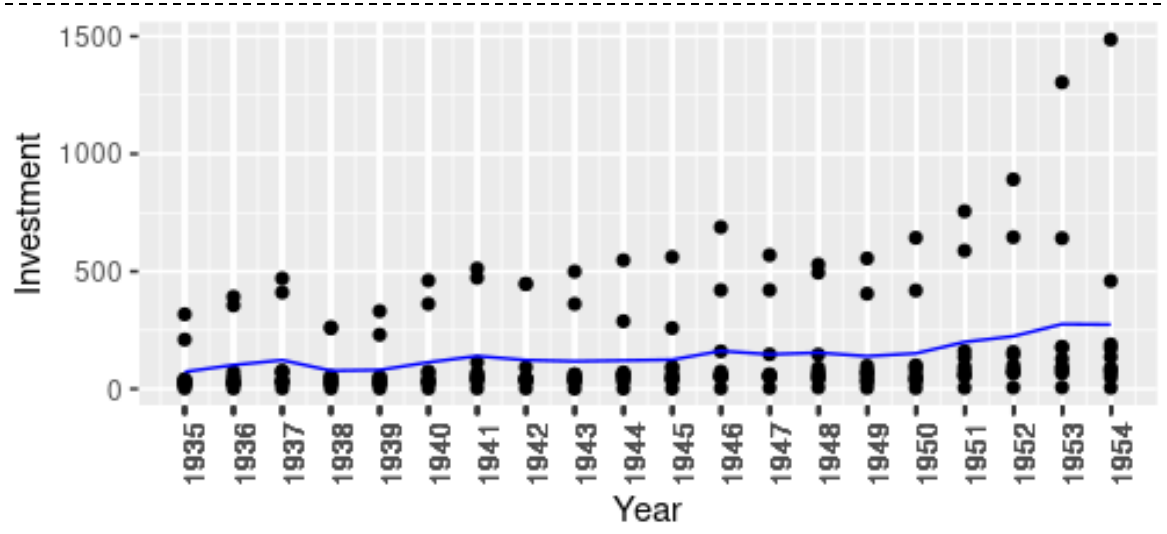


Image Source – Computed R Output

OLS Regression

Model	Dependent Variable	Independent Variable	Multiple R-squared
OLS 1	INV	CAPITAL	0.439
OLS 2	INV	VALUE	0.734
OLS 3	INV	VALUE + CAPITAL	0.812

```
Call:
lm(formula = INV ~ VALUE + CAPITAL, data = Gdata)

Residuals:
    Min      1Q  Median      3Q     Max
-291.68  -30.01    5.30   34.83  369.45

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -42.714369   9.511676  -4.491 1.21e-05 ***
VALUE         0.115562   0.005836  19.803  < 2e-16 ***
CAPITAL       0.230678   0.025476   9.055  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 94.41 on 197 degrees of freedom
Multiple R-squared:  0.8124,    Adjusted R-squared:  0.8105
F-statistic: 426.6 on 2 and 197 DF,  p-value: < 2.2e-16
```

Image Source – Computed R Output

Entity Fixed Effect Models

Fixed Effect Regression (Entity)				
Entity	**Intercept**	**Co-efficient CAPITAL**	**Co-efficient VALUE**	**Multiple R-squared**
FIRM 1	-149.78	0.37	0.11	0.92
FIRM 2	-49.19	0.38	0.17	0.47
FIRM 3	-9.95	0.15	0.02	0.70

N – 1 Binary Regression | **Multiple R-squared = 0.96**

```
Call:
lm(formula = inv ~ value + capital + factor(firm) - 1, data = Grunfeld)

Residuals:
     Min       1Q   Median       3Q      Max
-184.009  -17.643    0.563   19.192  250.710

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
value            0.11012    0.01186   9.288  < 2e-16 ***
capital          0.31007    0.01735  17.867  < 2e-16 ***
factor(firm)1  -70.29672   49.70796  -1.414   0.1590
factor(firm)2  101.90581   24.93832   4.086 6.49e-05 ***
factor(firm)3 -235.57184   24.43162  -9.642  < 2e-16 ***
factor(firm)4  -27.80929   14.07775  -1.975   0.0497 *
factor(firm)5 -114.61681   14.16543  -8.091 7.14e-14 ***
factor(firm)6  -23.16130   12.66874  -1.828   0.0691 .
factor(firm)7  -66.55347   12.84297  -5.182 5.63e-07 ***
factor(firm)8  -57.54566   13.99315  -4.112 5.85e-05 ***
factor(firm)9  -87.22227   12.89189  -6.766 1.63e-10 ***
factor(firm)10  -6.56784   11.82689  -0.555   0.5793
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 52.77 on 188 degrees of freedom
Multiple R-squared:  0.9616,    Adjusted R-squared:  0.9591
F-statistic:   392 on 12 and 188 DF,  p-value: < 2.2e-16
```

Image Source – Computed R Output

Time Fixed Effect Models

Fixed Effect Regression (Time)				
YEAR	Intercept	Co-efficient CAPITAL	Co-efficient VALUE	Multiple R-squared
1935	0.35	0.00	0.10	0.86
1936	15.21	-0.05	0.08	0.69
1937	-3.38	0.21	0.07	0.66
P value large than 0.05				

T– 1 Binary Regression	Multiple R-squared = 0.86

```
Call:
lm(formula = INV ~ VALUE + CAPITAL + factor(YEAR) - 1, data = Gdata)

Residuals:
    Min      1Q  Median      3Q     Max
-292.16  -26.37    8.37   31.42  380.14

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
VALUE              0.116798   0.006331  18.448  < 2e-16 ***
CAPITAL            0.219707   0.032296   6.803  1.5e-10 ***
factor(YEAR)1935 -23.574968  31.254082  -0.754   0.4517
factor(YEAR)1936 -40.787307  31.579889  -1.292   0.1982
factor(YEAR)1937 -58.066240  31.878940  -1.821   0.0702 .
factor(YEAR)1938 -52.017730  31.393606  -1.657   0.0993 .
factor(YEAR)1939 -79.818004  31.585049  -2.527   0.0124 *
factor(YEAR)1940 -54.079700  31.635759  -1.709   0.0891 .
factor(YEAR)1941 -26.202078  31.628133  -0.828   0.4085
factor(YEAR)1942 -24.997122  31.583540  -0.791   0.4297
factor(YEAR)1943 -45.376238  31.672627  -1.433   0.1537
factor(YEAR)1944 -45.692318  31.684692  -1.442   0.1510
factor(YEAR)1945 -57.171437  31.779365  -1.799   0.0737 .
factor(YEAR)1946 -30.603029  31.872994  -0.960   0.3383
factor(YEAR)1947 -28.821095  32.144500  -0.897   0.3711
factor(YEAR)1948 -27.494440  32.452843  -0.847   0.3980
factor(YEAR)1949 -52.368285  32.757614  -1.599   0.1117
factor(YEAR)1950 -51.929057  32.937171  -1.577   0.1167
factor(YEAR)1951 -35.246908  33.135802  -1.064   0.2889
factor(YEAR)1952 -29.188192  33.738078  -0.865   0.3881
factor(YEAR)1953 -21.125971  34.643537  -0.610   0.5428
factor(YEAR)1954 -35.889838  35.726906  -1.005   0.3165
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 98.1 on 178 degrees of freedom
Multiple R-squared:  0.8742,    Adjusted R-squared:  0.8587
F-statistic: 56.24 on 22 and 178 DF,  p-value: < 2.2e-16
```

Image Source – Computed R Output

Entity & Time Fixed Effect Models

```
Call:
lm(formula = INV ~ VALUE + CAPITAL + factor(YEAR) + factor(FIRM) -
    1, data = Gdata)

Residuals:
     Min       1Q   Median       3Q      Max
-162.609  -19.471   -1.267   19.128  211.842
```

```
Residual standard error: 51.72 on 169 degrees of freedom
Multiple R-squared:  0.9668,    Adjusted R-squared:  0.9607
F-statistic: 158.8 on 31 and 169 DF,  p-value: < 2.2e-16
```

	Estimate	Std. Error	t value	Pr(>|t|)	
VALUE	0.11772	0.01375	8.560	6.65e-15	***
CAPITAL	0.35792	0.02272	15.754	< 2e-16	***
factor(YEAR)1935	-86.90023	56.04663	-1.550	0.122893	
factor(YEAR)1936	-106.09764	60.75981	-1.746	0.082597	.
factor(YEAR)1937	-127.59024	64.19239	-1.988	0.048469	*
factor(YEAR)1938	-126.12663	57.65495	-2.188	0.030071	*
factor(YEAR)1939	-156.37052	60.58431	-2.581	0.010699	*
factor(YEAR)1940	-131.13531	61.26344	-2.141	0.033747	*
factor(YEAR)1941	-105.70469	60.64316	-1.743	0.083142	.
factor(YEAR)1942	-108.04002	57.98137	-1.863	0.064147	.
factor(YEAR)1943	-129.87785	59.41025	-2.186	0.030182	*
factor(YEAR)1944	-129.99900	59.82469	-2.173	0.031173	*
factor(YEAR)1945	-142.58327	61.23810	-2.328	0.021078	*
factor(YEAR)1946	-118.06951	62.02857	-1.903	0.058679	.
factor(YEAR)1947	-126.29247	58.47605	-2.160	0.032203	*
factor(YEAR)1948	-130.61674	58.09065	-2.248	0.025837	*
factor(YEAR)1949	-160.39533	58.35005	-2.749	0.006631	**
factor(YEAR)1950	-162.79634	59.09141	-2.755	0.006513	**
factor(YEAR)1951	-149.38114	61.96340	-2.411	0.016990	*
factor(YEAR)1952	-151.53257	62.56240	-2.422	0.016488	*
factor(YEAR)1953	-154.61820	65.39079	-2.365	0.019188	*
factor(YEAR)1954	-180.42645	65.00056	-2.776	0.006128	**
factor(FIRM)2	207.05424	35.17275	5.887	2.07e-08	***
factor(FIRM)3	-135.23080	35.70897	-3.787	0.000212	***
factor(FIRM)4	95.35384	50.72212	1.880	0.061839	.
factor(FIRM)5	-5.43860	57.83052	-0.094	0.925186	
factor(FIRM)6	102.88864	54.17388	1.899	0.059238	.
factor(FIRM)7	51.46661	58.17922	0.885	0.377617	
factor(FIRM)8	67.49051	50.97093	1.324	0.187258	
factor(FIRM)9	30.21756	55.72307	0.542	0.588339	
factor(FIRM)10	126.83712	58.52545	2.167	0.031618	*

Image Source – Computed R Output

Random Effect Models

```
Call:
plm(formula = INV ~ VALUE + CAPITAL, data = Gdata, effect = "individual",
    model = "random", index = c("FIRM", "YEAR"))

Balanced Panel: n = 10, T = 20, N = 200

Effects:
                  var std.dev share
idiosyncratic 2784.46   52.77 0.282
individual    7089.80   84.20 0.718
theta: 0.8612

Residuals:
     Min.   1st Qu.    Median   3rd Qu.      Max.
-177.6063  -19.7350    4.6851   19.5105  252.8743

Coefficients:
             Estimate Std. Error z-value Pr(>|z|)
(Intercept) -57.834415  28.898935 -2.0013  0.04536 *
VALUE         0.109781   0.010493 10.4627  < 2e-16 ***
CAPITAL       0.308113   0.017180 17.9339  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    2381400
Residual Sum of Squares: 548900
R-Squared:      0.7695
Adj. R-Squared: 0.76716
Chisq: 657.674 on 2 DF, p-value: < 2.22e-16
```

Image Source – Computed R Output

Hausman Test

Hausman test is done to check between fixed and random effects model. It checks whether the individual error terms are correlated with variable. The null hypothesis states that there is no such correlation (RE). The alternative hypothesis is that a correlation exists (FE).
RE can be rejected.

```
        Hausman Test

data:  INV ~ VALUE + CAPITAL
chisq = 10.4, df = 2, p-value = 0.005517
alternative hypothesis: one model is inconsistent
```

P value is significant (< 0.05)

Image Source – Computed R Output

Conclusion

- There is causal relationship between gross investment of a firm to the value and capital of the firm which is variable cost and fixed cost.
- Entity & time heterogeneity exist in the data. Thus, OLS regression is not correct model to predict investment.
- N – 1 / T –1 Binary Regression showed better result than OLS regression.
- Time & Entity Fixed Effect showcased best result.
- Hausman test indicate that Random Effect model can be ignored, and there is no correlation between individual error & variables.

Image Source – Computed R Output

References

1) Kleiber, Christian, and Achim Zeileis. "The Grunfeld data at 50." *German Economic Review* 11.4 (2010): 404-417.
2) Grunfeld, Yehuda. *The determinants of corporate investment*. Diss. The University of Chicago, 1958.
3) Croissant, Yves, and Giovanni Millo. "Panel data econometrics in R: The plm package." *Journal of statistical software* 27.2 (2008).
4) https://rpubs.com/phle/r_tutorial_panel_data_analysis
5) https://rpubs.com/bambangpe/778639
6) https://www.zeileis.org/grunfeld/
7) https://www.princeton.edu/~otorres/Panel101R.pdf
8) Croissant, Yves, and Giovanni Millo. *Panel data econometrics with R*. John Wiley & Sons, 2018.
9) https://rstudio-pubs-static.s3.amazonaws.com/79177_5a4eca630d2942bc927b6da7b94f6ab4.html

Thank you