Multiple Regression Analysis of Taxable Income

CS – 593

Xiaoqin, Chen

Lan, Chang

May 05, 2016

Variables

Taxable income for household	Federal income taxes	husband's number of siblings	husband's father's education level	husband's mothers's education level	Wife's number of siblings	dummy variable = 1 if woman worked in 1975, else 0	Wife's hours of work in 1975	Number of children less than 6 years old in household	Number of children between ages 6 and 18 in household
Wife's age	Wife's educational attainment, in years	Wife's 1975 average hourly earnings, in 1975 dollars	Wife's wage reported at 1976 interview, for 1976	Husband's hours worked in 1975	Husband's age	Husband's educational attainment, in years	Husband's wage, in 1975 dollars	Family income, in 1975 dollars	marginal tax rate facing the wife, includes Soc Sec taxes
wife's mother's education level	wife's father's education level	Unemployment rate in county of residence	Dummy variable = 1 if live in large city (SMSA), else 0	Actual years of wife's previous labor market experience					
12	7	5	0	14	34	12	4.0288000107	16310	0.7214999795
7	7	11	1	5	30	9	8.4415998459	21800	0.6614999771
12	7	5	0	15	40	12	3.5806999207	21040	0.6915000081
7	7	5	0	6	53	10	3.5416998863	7300	0.7814999819
12	14	9.5	1	7	32	12	10	27300	0.6215000153
14	7	7.5	1	33	57	11	6.7105998993	19495	0.6915000081
14	7	5	0	11	37	12	3.4277000427	21152	0.6915000081
3	3	5	0	35	53	8	2.548500061	18900	0.6915000081
7	7	3	0	24	52	4	4.2206001282	20405	0.7515000105
7	7	5	0	21	43	12	5.7143001556	20425	0.6915000081

Dataset resource: http://www.principlesofeconometrics.com/sas.htm

Correlation

	HSIBLINGS	HFATHEREDUC	HMOTHEREDUC	SIBLINGS	LFP
TAXABLEINC Taxable income for household	−0.17687 <.0001	0.26064 <.0001	0.22113 <.0001	−0.20938 <.0001	−0.14872 <.0001

Pearson Correlation Coefficients, N = 753
Prob > |r| under H0: Rho=0

HOURS	KIDSL6	KIDS618	AGE	EDUC	WAGE	WAGE76	HHOURS	HAGE
−0.05729 0.1162	0.08050 0.0272	−0.02481 0.4966	0.00194 0.9576	−0.08065 0.0269	−0.11299 0.0019	−0.03837 0.2930	−0.05729 0.1162	0.01277 0.7265

HEDUC	HWAGE	FAMINC	MTR	MOTHEREDUC	FATHEREDUC	UNEMPLOYMENT	LARGECITY	EXPER
−0.04090 0.2623	−0.00135 0.9704	−0.04259 0.2431	0.04870 0.1819	−0.03691 0.3118	−0.02728 0.4548	0.00425 0.9074	−0.02097 0.5656	−0.03331 0.3613

Multiple Regression using backward selection

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: TAXABLEINC Taxable income for household

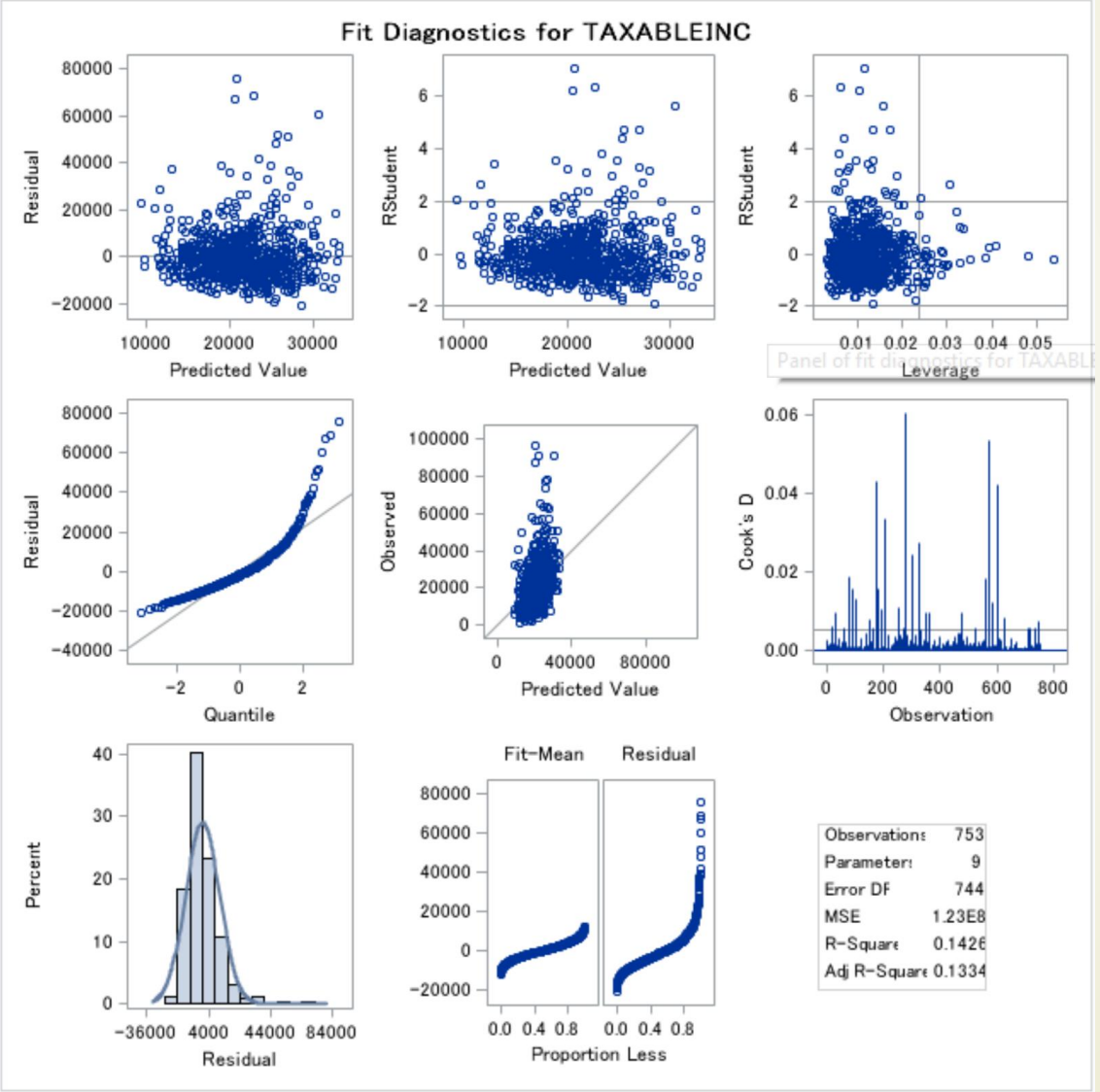
Number of Observations Read	753
Number of Observations Used	753

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	15166732684	1895841585	15.47	<.0001
Error	744	91169844945	122540114		
Corrected Total	752	1.063366E11			

Root MSE	11070	R-Square	0.1426
Dependent Mean	21152	Adj R-Sq	0.1334
Coeff Var	52.33471		

Parameter Estimates								
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Squared Partial Corr Type I	Variance Inflation
Intercept	Intercept	1	21947	2926.46901	7.50	<.0001	.	0
HSIBLINGS	husband's number of siblings	1	-328.45814	180.94406	-1.82	0.0699	0.03128	1.16449
HFATHEREDUC	husband's father's education level	1	595.33288	143.39536	4.15	<.0001	0.04988	1.30612
HMOTHEREDUC	husband's mothers's education level	1	373.23815	138.83562	2.69	0.0073	0.00979	1.33169
SIBLINGS	Wife's number of siblings	1	-749.77182	180.64282	-4.15	<.0001	0.02171	1.07119
LFP	dummy variable = 1 if woman worked in 1975, else 0	1	-4530.14807	1072.05078	-4.23	<.0001	0.02452	1.73255
KIDSL6	Number of children less than 6 years old in household	1	1532.08248	799.93951	1.92	0.0558	0.00351	1.07807
EDUC	Wife's educational attainment, in years	1	-367.44234	186.26350	-1.97	0.0489	0.00309	1.10703
WAGE76	Wife's wage reported at 1976 interview, for 1976	1	529.14222	220.54980	2.40	0.0167	0.00768	1.74801

Residual Analysis



Transformation to the variable

```
data mroz;
set mroz;
LOG_TAXABLEINC = log(TAXABLEINC);
run;
```

Taxable income for household	LOG_TAXABLEINC
12200	9.4091912307
18000	9.7981270369
24000	10.085809109
16400	9.7050366138
10000	9.210340372
6295	8.7475109465
9952	9.205528815
18900	9.846917201
1500	7.3132203871
22000	9.9987977323
30000	10.308952661
21950	9.9965224185
22000	9.9987977323
9296	9.1373394791
12600	9.4414520929
25000	10.126631104
5878	8.678971847
11100	9.3147003873
25320	10.139349876
56100	10.934891092

Map Reduce

Divide

```
*divide the data into two datasets by LFP;
data income_1 income_2;
set mroz;
if mod(LFP, 2)=0 then output income_1;
else if mod(LFP, 2)=1 then output income_2;
run;
```

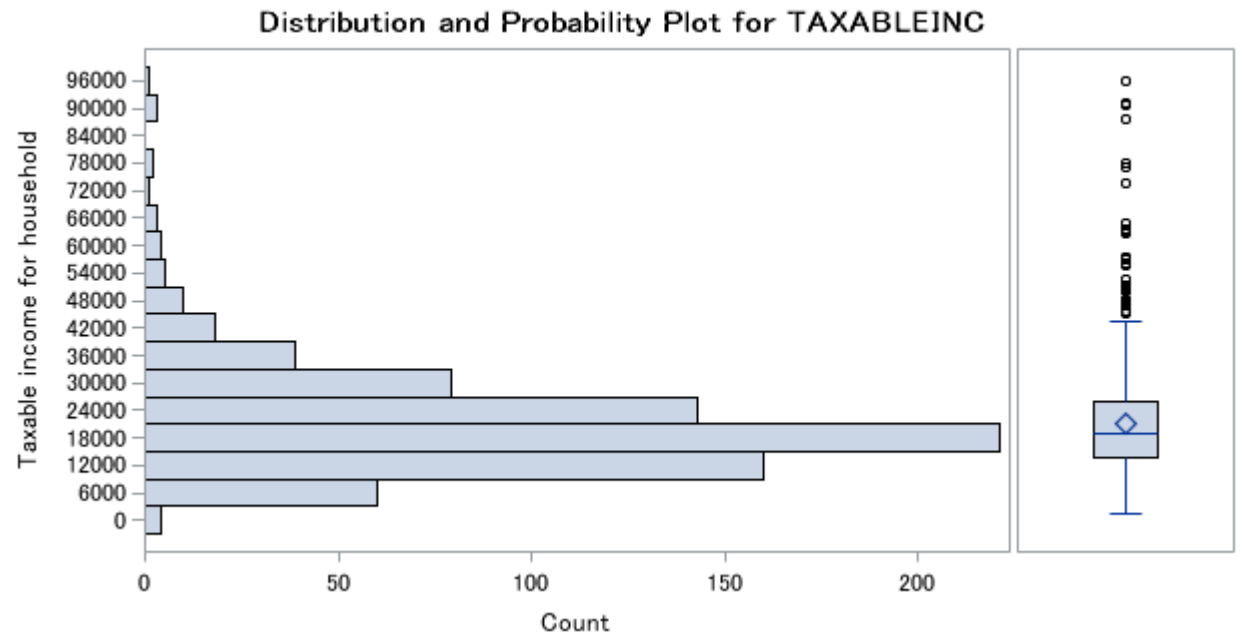
Map task

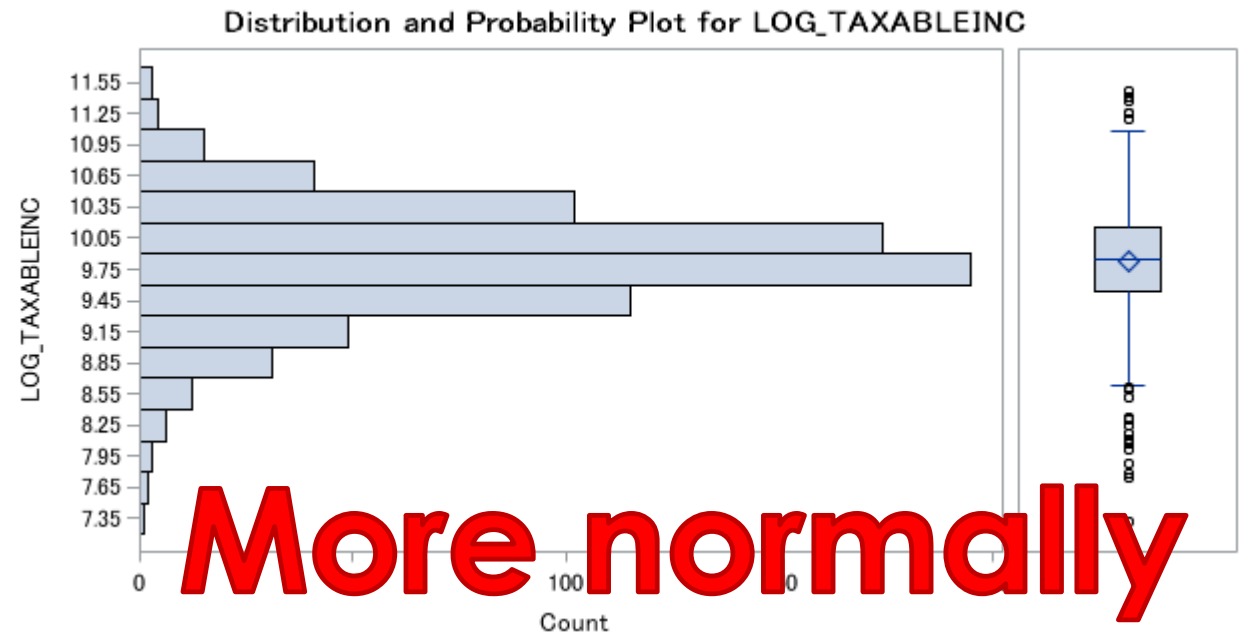
```
data sasdata1.log_income_1;
set sasdata1.income_1;
LOG_TAXABLEINC=log(TAXABLEINC);
run;
```

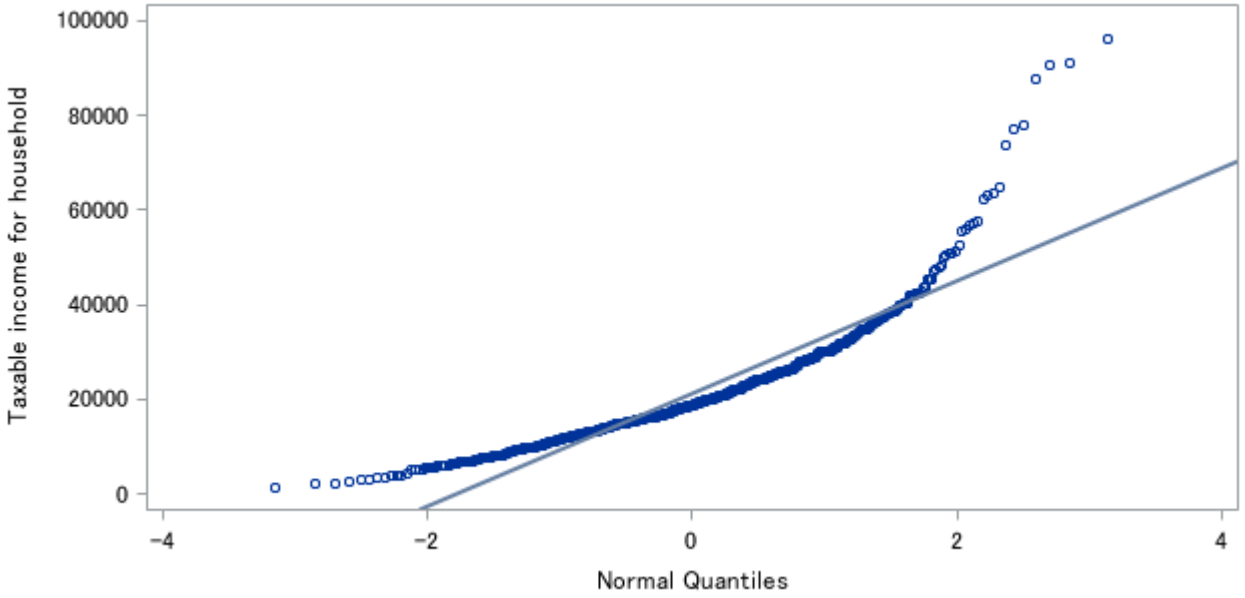
```
proc sql;
create table reduce1.income_1 as
select *
from sasdata1.income_1
;
run;
```

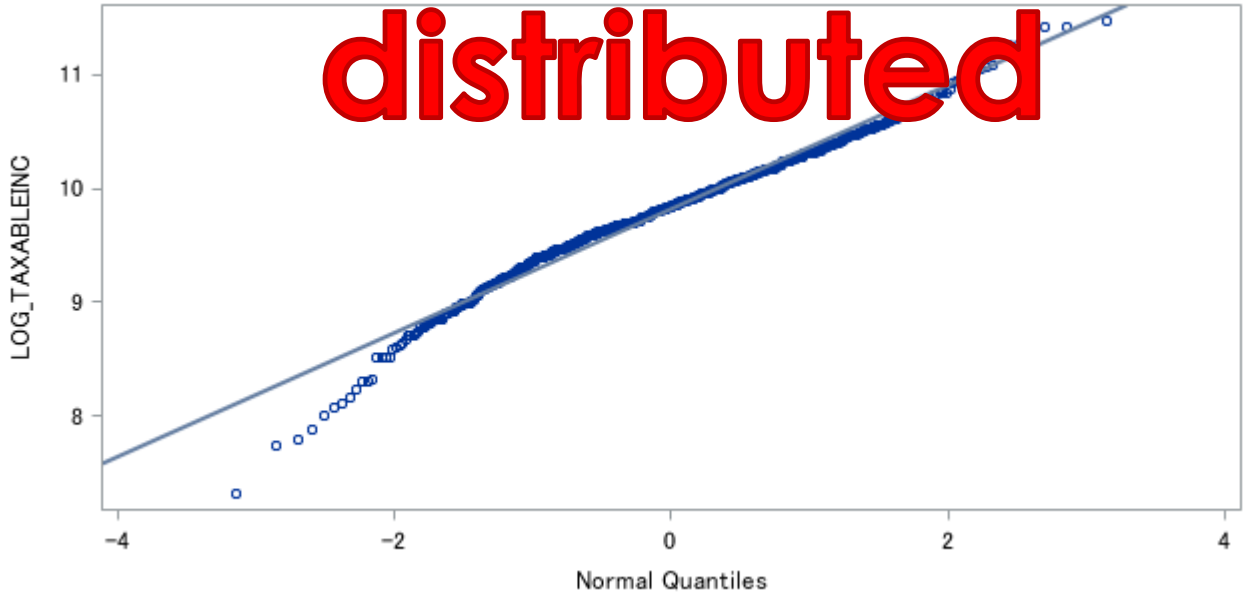
Reduce task

```
proc sql;
insert into main.income_1 select * from main.income_2;
run;
```









More normally distributed

Correlation

	HSIBLINGS	HFATHEREDUC	HMOTHEREDUC	SIBLINGS	LFP
LOG_TAXABLEINC	-0.20029 <.0001	0.26687 <.0001	0.21509 <.0001	-0.23683 <.0001	-0.20842 <.0001

Pearson Correlation Coefficients, N = 753
Prob > |r| under H0: Rho=0

HOURS	KIDSL6	KIDS618	AGE	EDUC	WAGE	WAGE76	HHOURS
-0.09456 0.0094	0.10329 0.0046	-0.00868 0.8120	-0.00688 0.8506	-0.07906 0.0301	-0.13461 0.0002	-0.04118 0.2591	-0.05435 0.1362

HAGE	HEDUC	HWAGE	FAMINC	MTR	MOTHEREDUC	FATHEREDUC	UNEMPLOYMENT	LARGECITY	EXPER
0.00468 0.8980	-0.02948 0.4192	0.00320 0.9300	-0.06365 0.0809	0.06005 0.0996	-0.05003 0.1702	-0.01134 0.7561	0.01557 0.6696	0.01727 0.6362	-0.05240 0.1509

Multiple Regression using backward selection

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: TAXABLEINC Taxable income for household

Number of Observations Read	753
Number of Observations Used	753

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	15166732684	1895841585	15.47	<.0001
Error	744	91169844945	122540114		
Corrected Total	752	1.063366E11			

Root MSE	11070	R-Square	0.1426
Dependent Mean	21152	Adj R-Sq	0.1334
Coeff Var	52.33471		

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: LOG_TAXABLEINC

Number of Observations Read	753
Number of Observations Used	753

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	41.43578	5.17947	21.30	<.0001
Error	744	180.93228	0.24319		
Corrected Total	752	222.36806			

Root MSE	0.49314	R-Square	0.1863
Dependent Mean	9.81989	Adj R-Sq	0.1776
Coeff Var	5.02186		

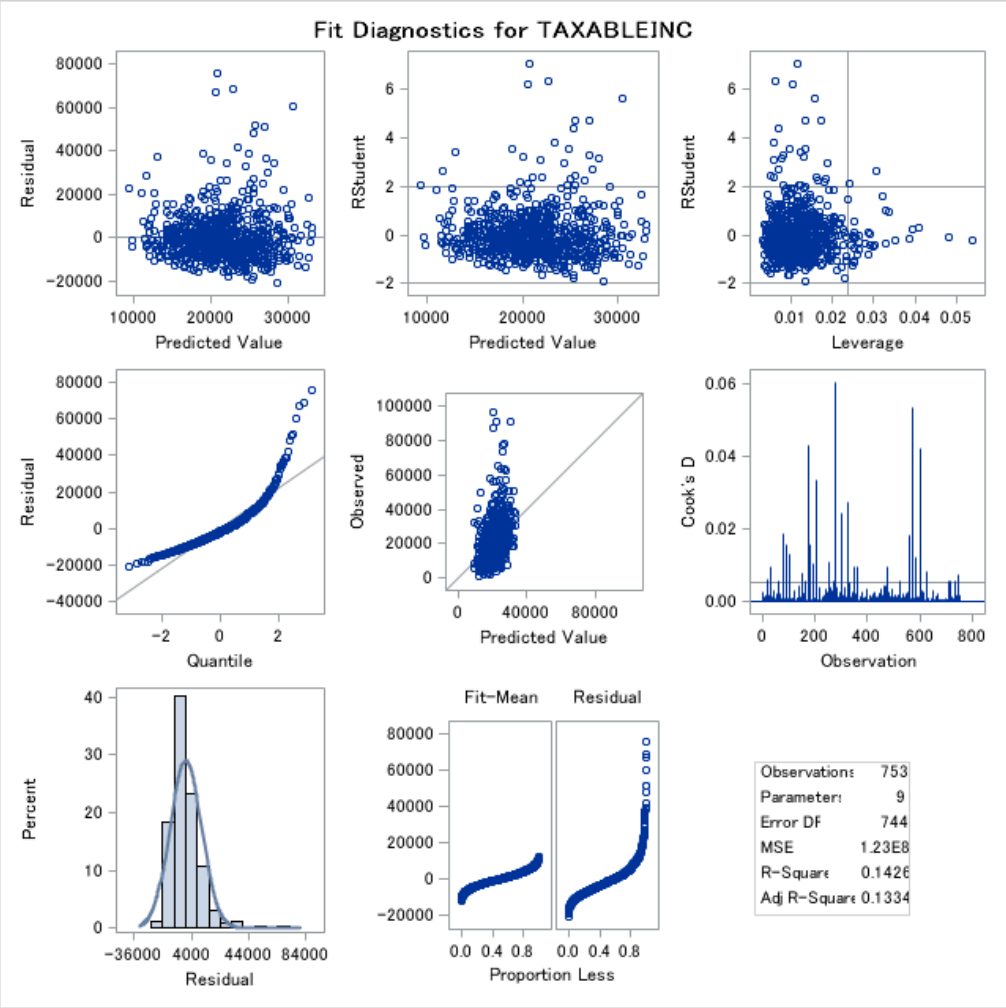
Not good enough

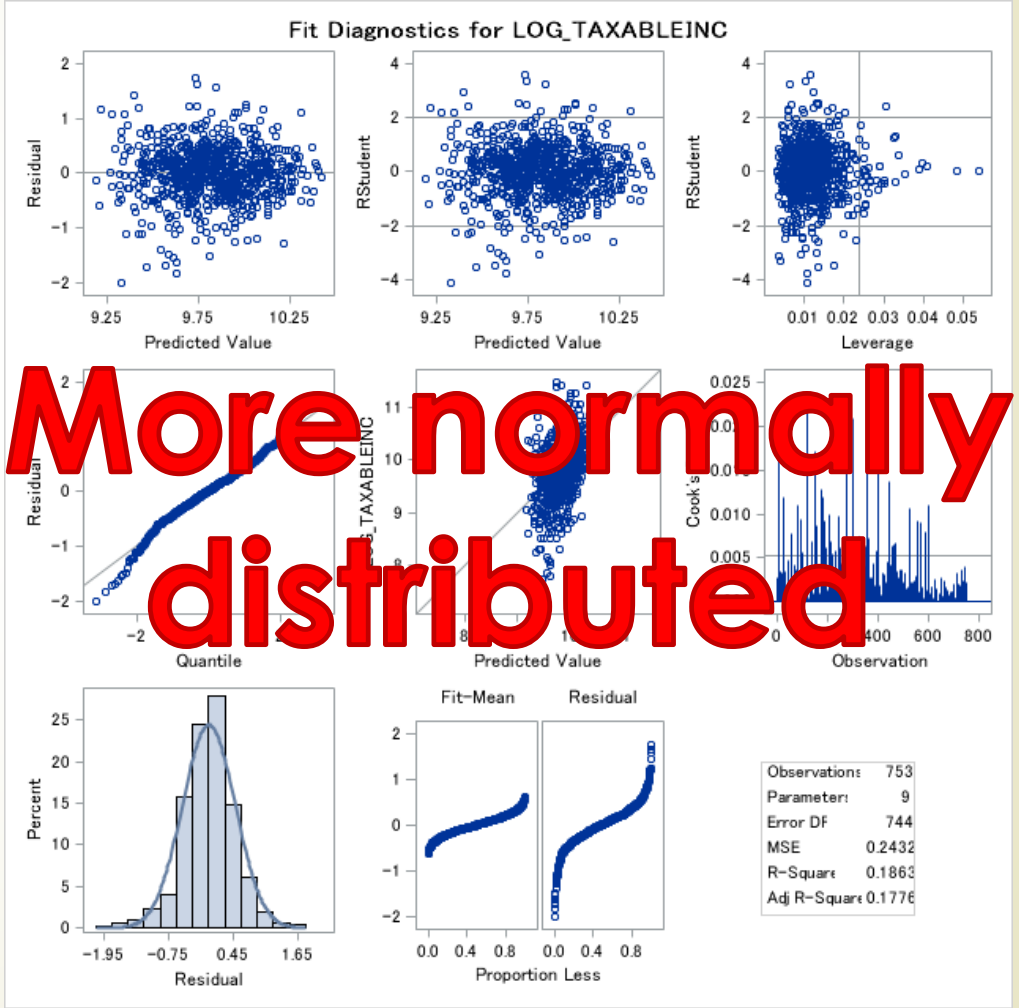
Parameter Estimates								
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Squared Partial Corr Type I	Variance Inflation
Intercept	Intercept	1	9.94117	0.13037	76.25	<.0001	.	0
HSIBLINGS	husband's number of siblings	1	−0.02020	0.00806	−2.51	0.0124	0.04012	1.16449
HFATHEREDUC	husband's father's education level	1	0.02711	0.00639	4.24	<.0001	0.05017	1.30612
HMOTHEREDUC	husband's mothers's education level	1	0.01454	0.00618	2.35	0.0190	0.00696	1.33169
SIBLINGS	Wife's number of siblings	1	−0.04057	[illegible]	[illegible]	[illegible]	[illegible]	1.07119
LFP	dummy variable = 1 if woman worked in 1975, else 0	1	−0.31098	0.04776	−6.51	<.0001	0.04875	1.73255
KIDSL6	Number of children less than 6 years old in household	1	0.07958	0.03564	2.23	0.0258	0.00502	1.07807
EDUC	Wife's educational attainment, in years	1	−0.01596	0.00830	−1.92	0.0548	0.00187	1.10703
WAGE76	Wife's wage reported at 1976 interview, for 1976	1	0.03723	0.00983	3.79	0.0002	0.01894	1.74801

Significant

The regression equation is:
LOG_TAXABLEINC = 9.941 – 0.020 HSIBLINGS + 0.027 HFATHEREDUC + 0.015 HMOTHEREDUC – 0.041 SIBLINGS – 0.311 LFP + 0.080 KIDSL6 – 0.016 EDUC + 0.037 WAGE76

Residual Analysis





More normally distributed

High Leverage and Influential Observation

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0.00309440	224	0.0386585	423
0.00332556	208	0.0395596	484
0.00343871	105	0.0409973	111
0.00359925	411	0.0482092	715
0.00361745	320	0.0539029	605

Lev

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
1.57563E-09	415	0.0171683	151
1.10555E-08	365	0.0185696	358
3.25095E-08	622	0.0203132	9
3.43374E-08	276	0.0208717	300
1.02093E-07	188	0.0235456	119

Cookd

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
-4.64E-01	119	0.315902	601
-4.32E-01	9	0.327814	184
-4.12E-01	358	0.342666	174
-3.95E-01	151	0.385341	275
-3.94E-01	399	0.434860	300

Dffits

Multicollinearity Analysis

Parameter Estimates								
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	Intercept	1	9.94117	0.13037	76.25	<.0001	.	0
HSIBLINGS	husband's number of siblings	1	-0.02020	0.00806	-2.51	0.0124	0.85875	1.16449
HFATHEREDUC	husband's father's education level	1	0.02711	0.00639	4.24	<.0001	0.76563	1.30612
HMOTHEREDUC	husband's mothers's education level	1	0.01454	0.00618	2.35	0.0190	0.75093	1.33169
SIBLINGS	Wife's number of siblings	1	-0.04057	0.00805	-5.04	<.0001	0.93354	1.07119
LFP	dummy variable = 1 if woman worked in 1975, else 0	1	-0.31098	0.04776	-6.51	<.0001	0.57719	1.73255
KIDSL6	Number of children less than 6 years old in household	1	0.07958	0.03564	2.23	0.0258	0.92758	1.07807
EDUC	Wife's educational attainment, in years	1	-0.01596	0.00830	-1.92	0.0548	0.90332	1.10703
WAGE76	Wife's wage reported at 1976 interview, for 1976	1	0.03723	0.00983	3.79	0.0002	0.57208	1.74801

X multicollinearity

Conclusion

LOG_TAXABLEINC	---------- taxable income for household
= 9.941	
– 0.020 HSIBLINGS	---------- husband's number of siblings
+ 0.027 HFATHEREDUC	---------- husband's father's education level
+ 0.015 HMOTHEREDUC	---------- husband's mother's education level
– 0.041 SIBLINGS	---------- wife's number of siblings
– 0.311 LFP	---------- dummy variable = 1 if woman worked in 1975, else 0
+ 0.080 KIDSL6	---------- number of children less than 6 years old in household
– 0.016 EDUC	---------- wife's educational attainment, in years
+ 0.037 WAGE76	---------- wife's wage reported at 1976 interview, for 1976

Thank you!