INFSCI 2725: Data Analytics

Assignment 2: Fundamental concepts from statistics

Yumeng Lu
Zhaoxuan Ren
Tao Li

In this assignment 2, we use SPSS to analyze the given data.

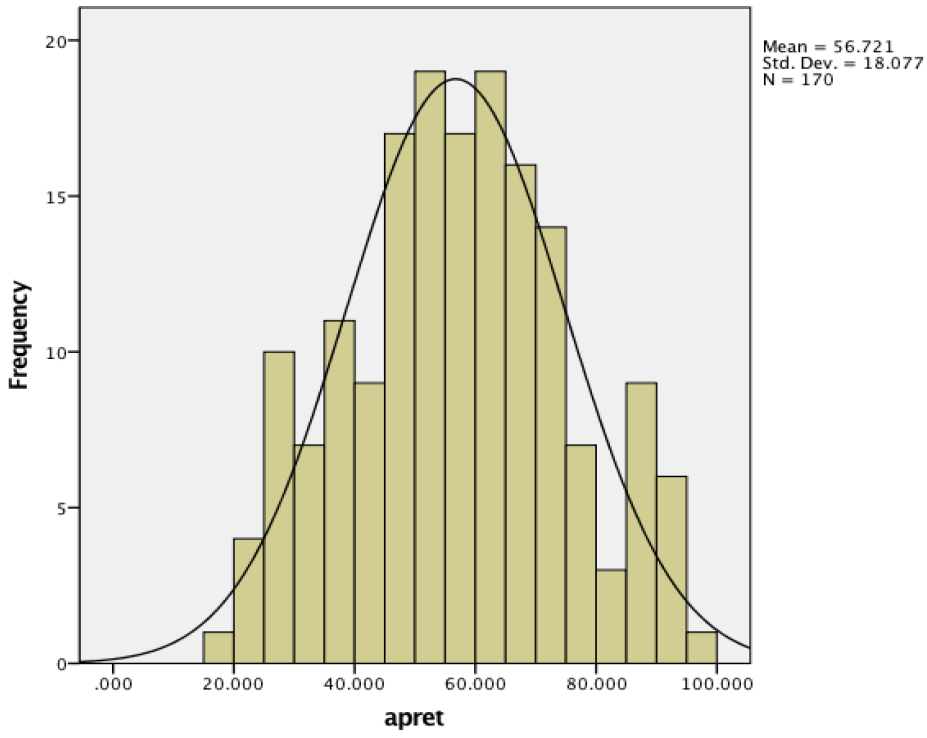
(1) Generate descriptive statistics and histograms for apret, tstsc, and salar.

apret:

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean		Std. Deviation	Variance
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic
apret	170	76.500	18.750	95.250	56.72108	1.386450	18.077097	326.781
Valid N (listwise)	170							

Histogram for apret

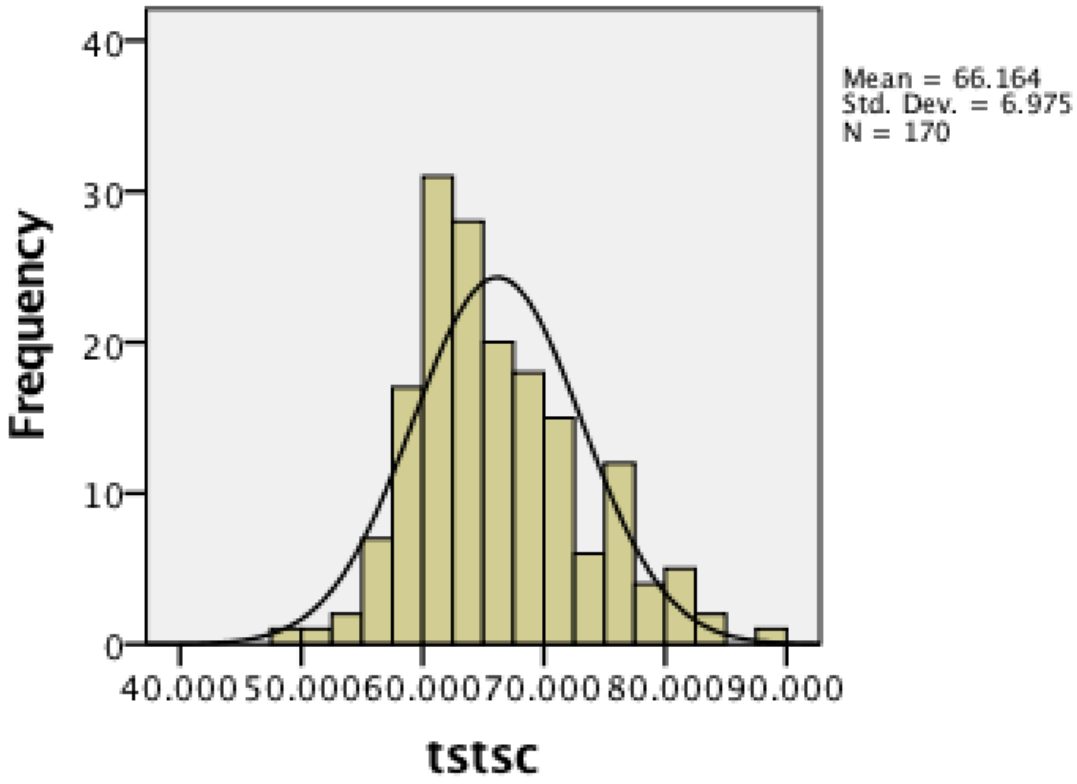


tstsc:

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean		Std. Deviation	Variance
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic
tstsc	170	39.375	48.125	87.500	66.16416	.534982	6.975306	48.655
Valid N (listwise)	170							

Histogram for tstsc

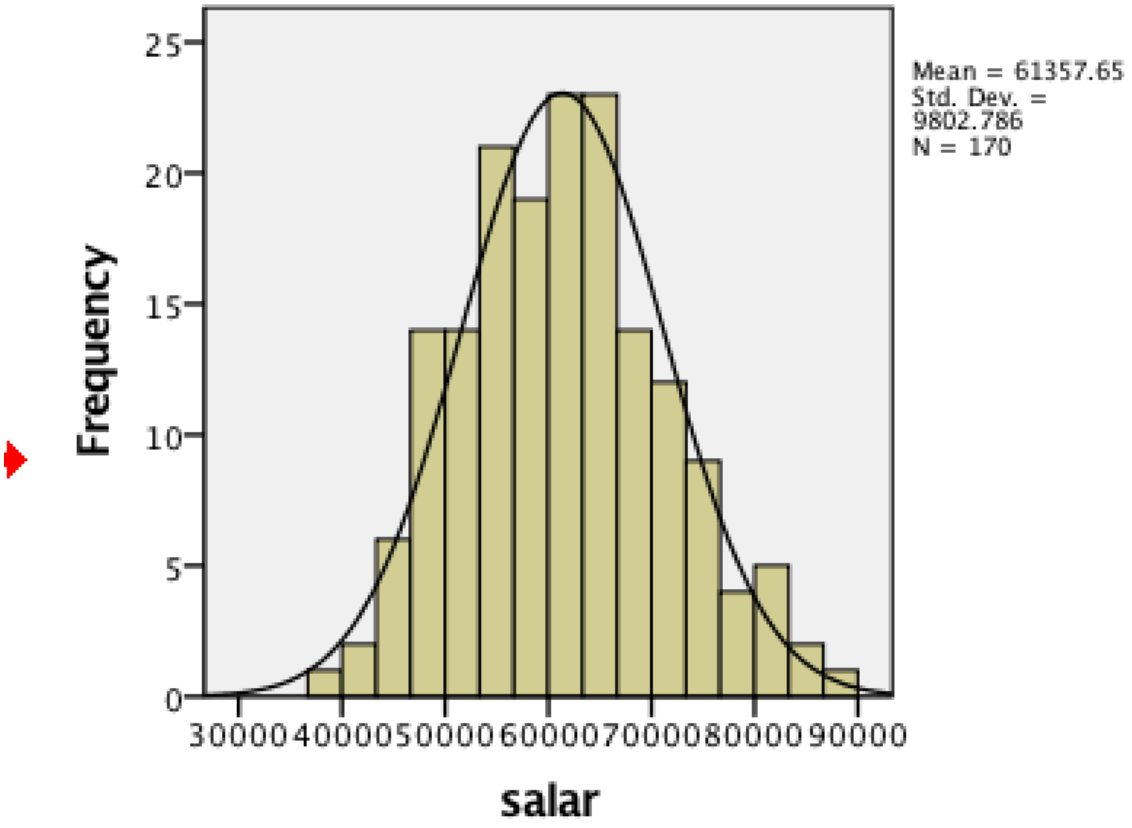


salar:

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean		Std. Deviation	Variance
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic
salar	170	49260	38640	87900	61357.65	751.839	9802.786	96094622.3
Valid N (listwise)	170							

Histogram for tstsc



（2）perform linear regression

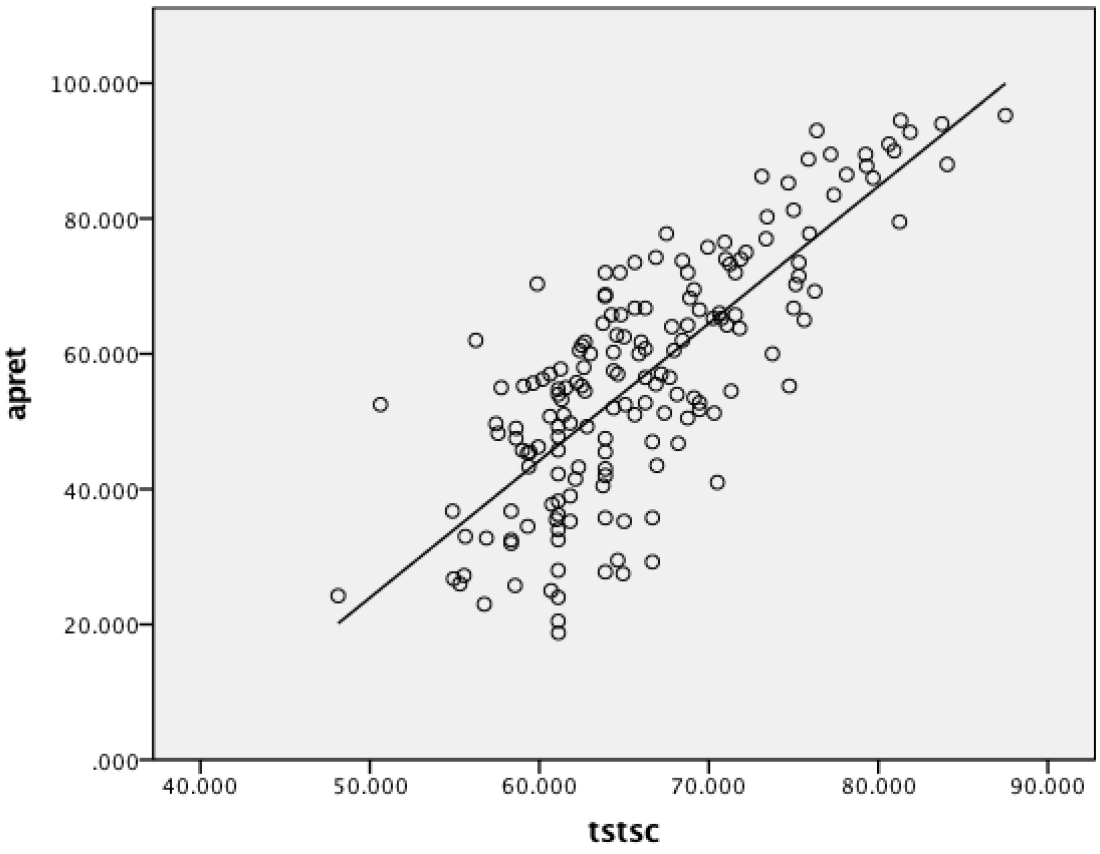
(a) linear regression of apret on tstsc

Dependent Variable: apret

Variables Entered: tstsc

Linear model:Y=AX+B

95% confidence intervals



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.782[a]	.612	.609	11.296381

R^2=0.612 and adjusted R^2=0.609, so the fitting degree of model and data is good

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	33787.880	1	33787.880	264.778	.000[b]
	Residual	21438.181	168	127.608		
	Total	55226.061	169			

a. Dependent Variable: apret

The sig=0.000<0.05, so the regression equation is valid

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-77.400	8.288		-9.339	.000	-93.762	-61.038
	tstsc	2.027	.125	.782	16.272	.000	1.781	2.273

a. Dependent Variable: apret

Constant=-77.400

tstsc(B)=2.027

apret=2.027tstsc-77.4000

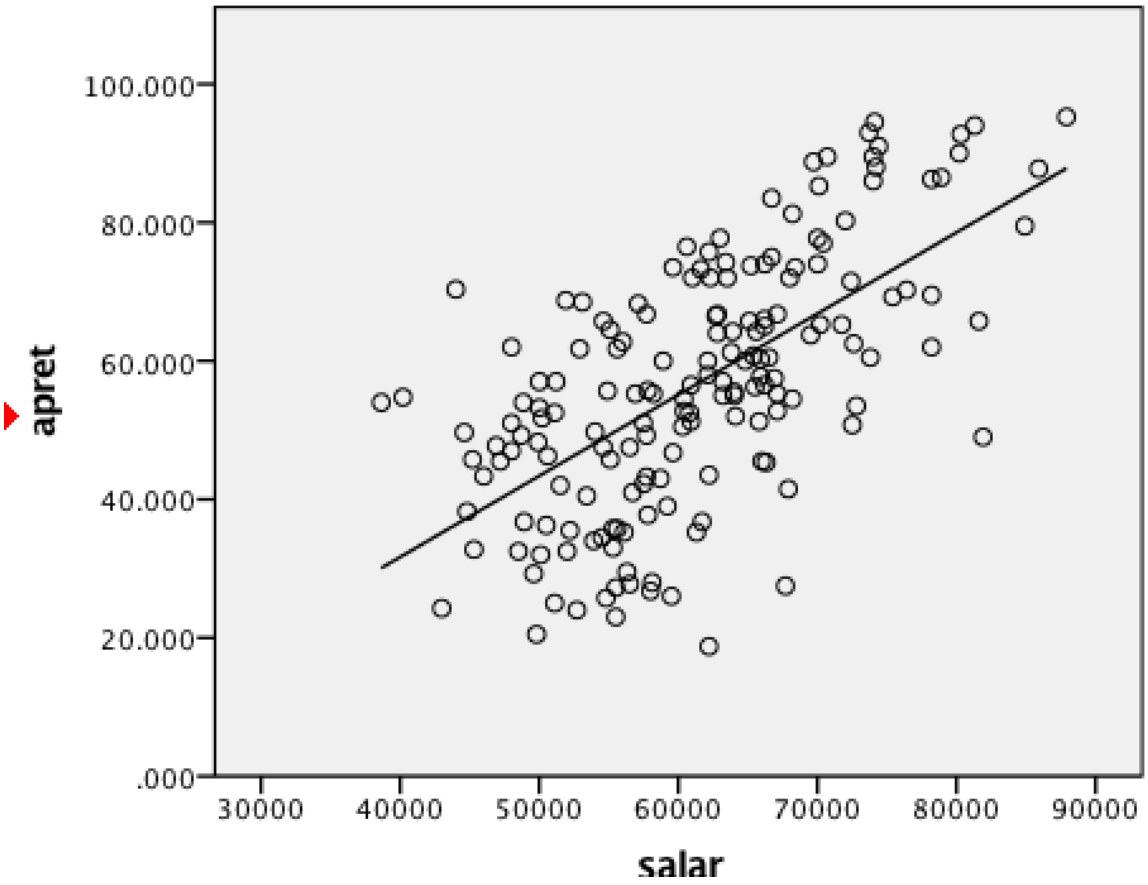
（b）linear regression of apret on salar

Dependent Variable: apret

Variables Entered: salar

Linear model:Y=AX+B

95% confidence intervals



Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.636[a]	.404	.401	13.993569

a. Predictors: (Constant), salar

b. Dependent Variable: apret

R^2=0.404 and adjusted R^2=0.404, so the fitting degree of model and data is good

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	22328.307	1	22328.307	114.025	.000[b]
	Residual	32897.755	168	195.820		
	Total	55226.061	169			

a. Dependent Variable: apret

b. Predictors: (Constant), salar

Sig=0.000<0.05 , so the regression equation is valid

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-15.224	6.823		-2.231	.027	-28.693	-1.755
	salar	.001	.000	.636	10.678	.000	.001	.001

a. Dependent Variable: apret

Constant=-15.224

salar(B)=0.001

apret=0.001salar-15.224

(c) linear regression of apret on tstsc and salar

Dependent Variable: apret
Variables Entered: salar， tstsc
R^2=0.624
Adjusted R^2=0.619
Sig=0.000
Constant=-75.911
tstsc(B)=1.738
salar(B)=0.000

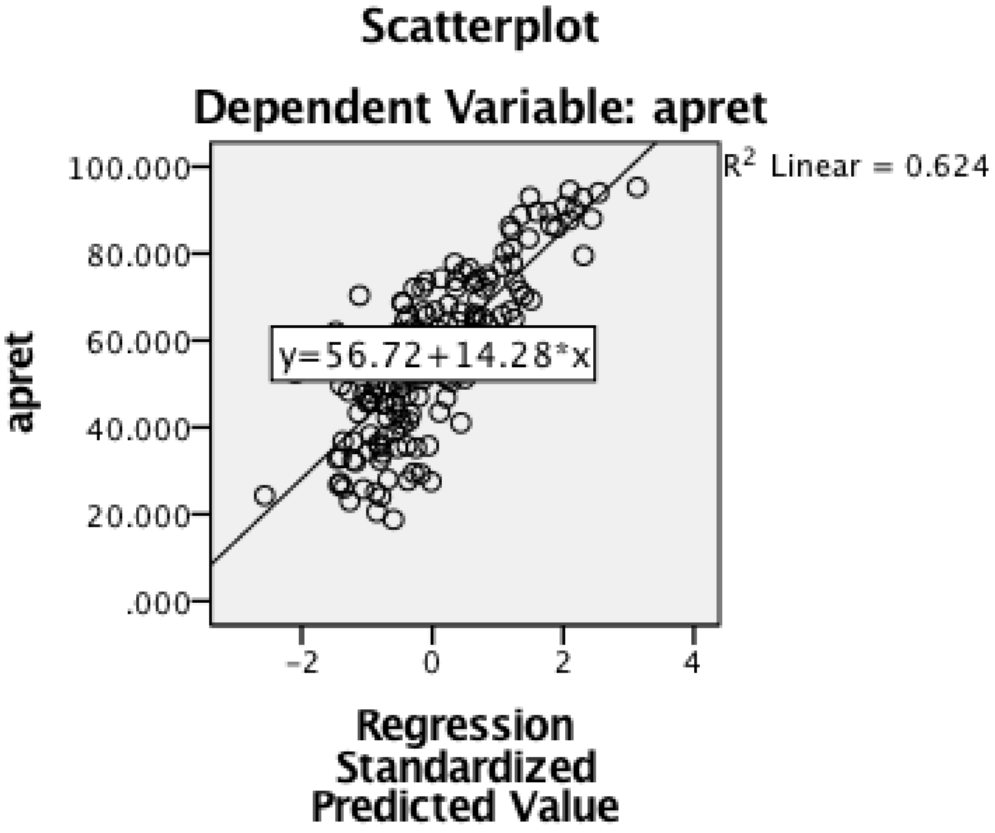
ANOVA[a]

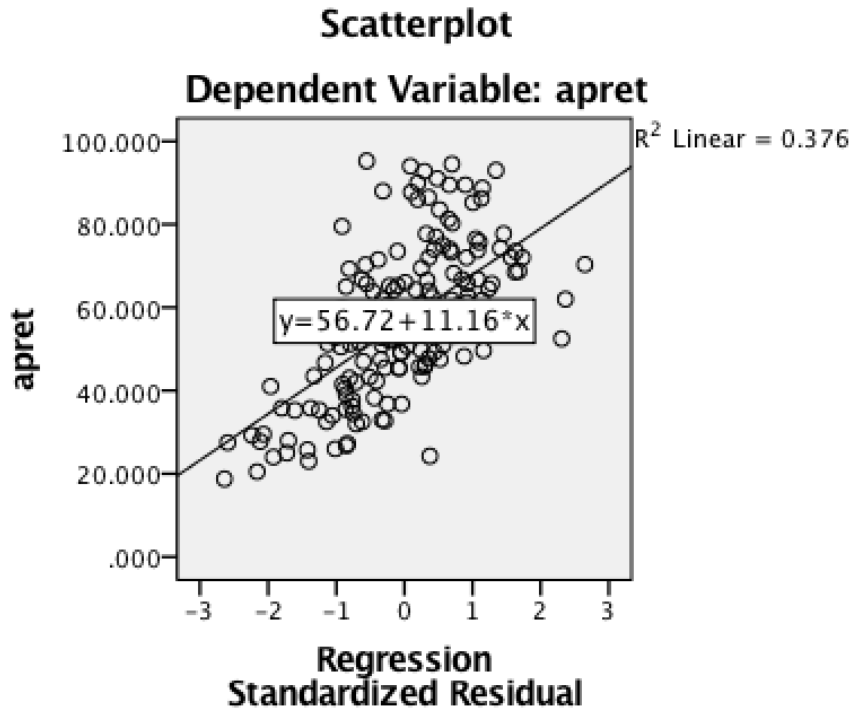
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	34445.228	2	17222.614	138.405	.000[b]
	Residual	20780.833	167	124.436		
	Total	55226.061	169			

a. Dependent Variable: apret
b. Predictors: (Constant), salar, tstsc

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-75.911	8.210		-9.246	.000	-92.119	-59.703
	tstsc	1.738	.176	.670	9.868	.000	1.390	2.085
	salar	.000	.000	.156	2.298	.023	.000	.001


Scatterplot
Dependent Variable: apret
R² Linear = 0.624
y=56.72+14.28*x
apret
100.000
80.000
60.000
40.000
20.000
.000
-2
0
2
4
Regression
Standardized
Predicted Value

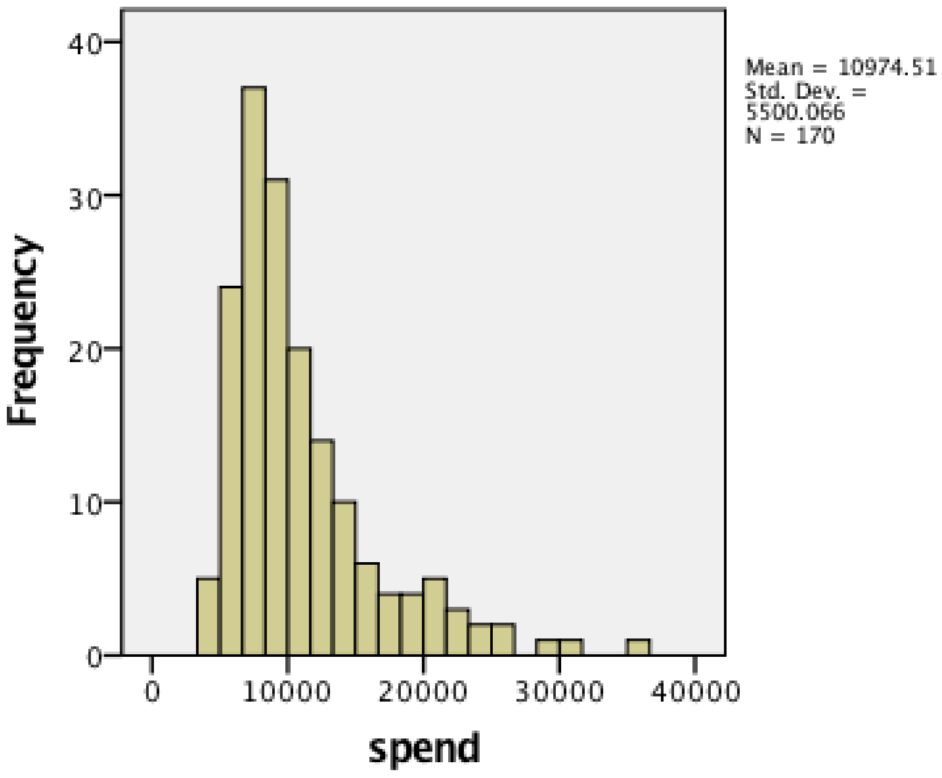

Scatterplot
Dependent Variable: apret
R² Linear = 0.376
y=56.72+11.16*x
apret
100.000
80.000
60.000
40.000
20.000
.000
-3
-2
-1
0
1
2
3
Regression
Standardized Residual

（3）Generate descriptive statistics and histograms for spend，top10，rejr，pacc，strat

(a)spend

Descriptive Statistics

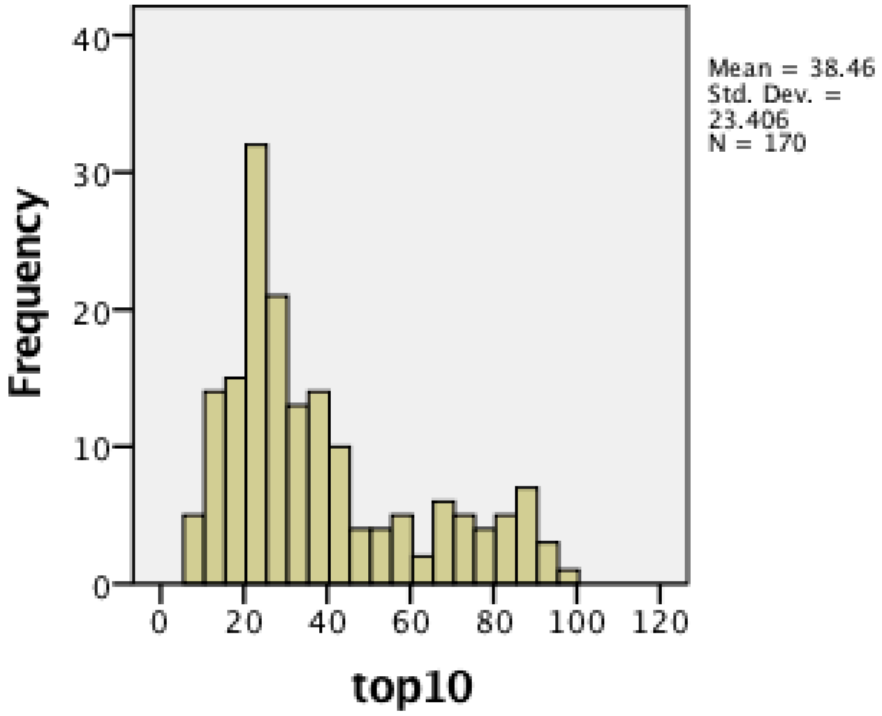
	N	Range	Minimum	Maximum	Mean		Std. Deviation	Variance
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic
spend	170	31738	4125	35863	10974.51	421.836	5500.066	30250721.4
Valid N (listwise)	170							



(b) top 10

Descriptive Statistics

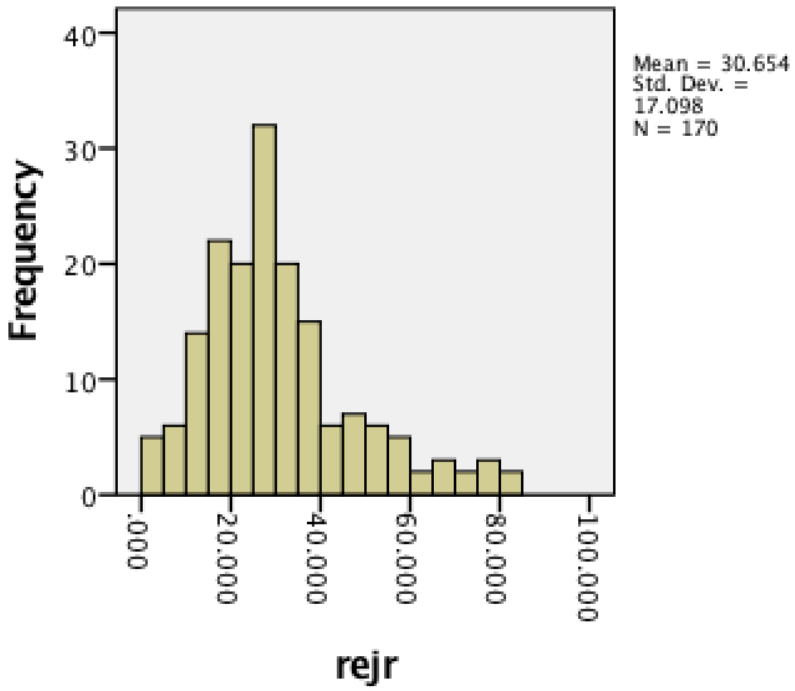
	N	Range	Minimum	Maximum	Mean		Std. Deviation	Variance
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic
top10	170	90	8	98	38.46	1.795	23.406	547.859
Valid N (listwise)	170							



(c) rejr

Descriptive Statistics

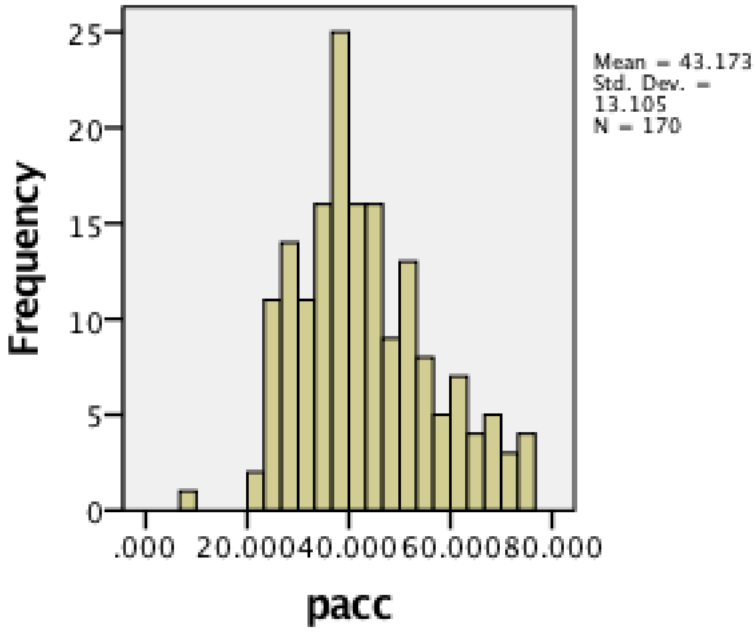
	N	Range	Minimum	Maximum	Mean		Std. Deviation	Variance
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic
rejr	170	84.067	.000	84.067	30.65422	1.311365	17.098104	292.345
Valid N (listwise)	170							



(d) pacc

Descriptive Statistics

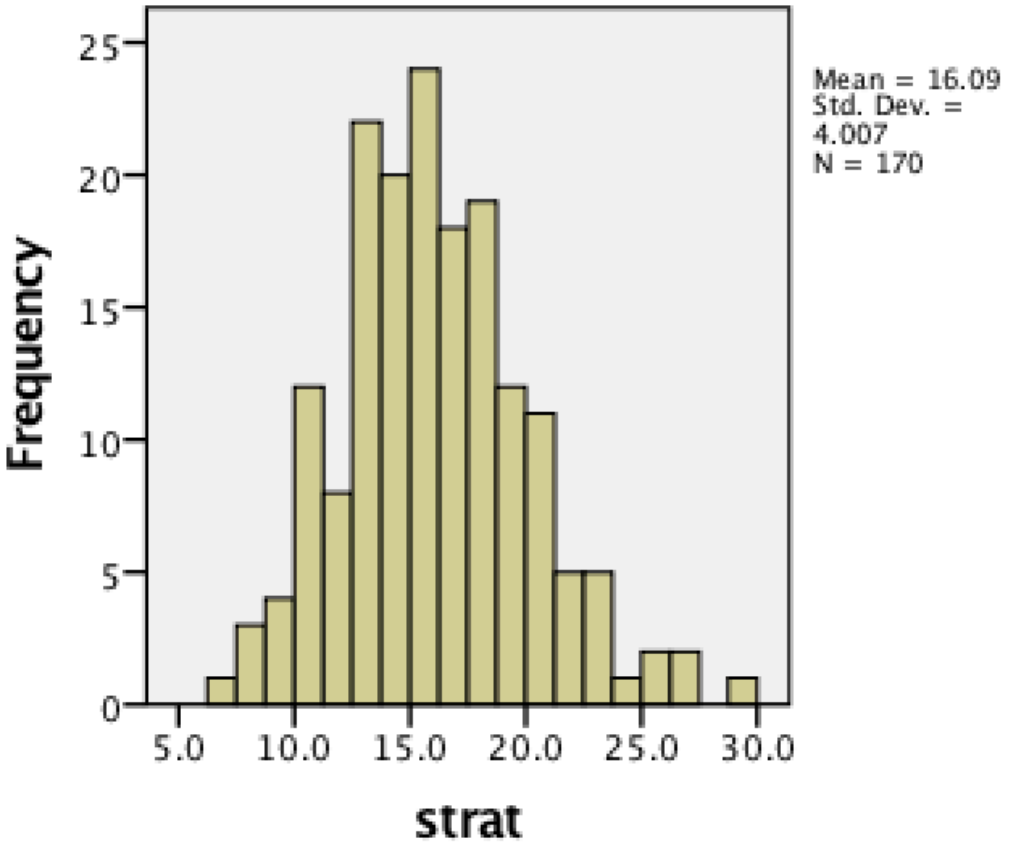
	N	Range	Minimum	Maximum	Mean		Std. Deviation	Variance
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic
pacc	170	67.289	8.964	76.253	43.17311	1.005123	13.105195	171.746
Valid N (listwise)	170							



(e) strat

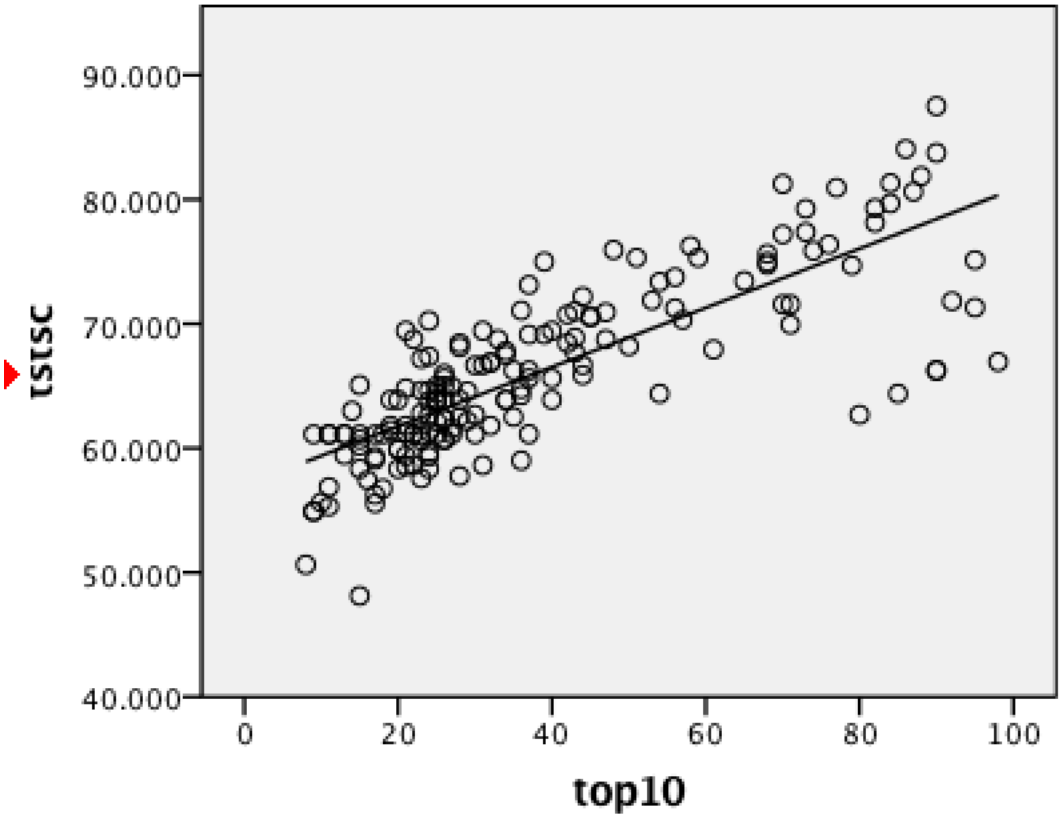
Descriptive Statistics

	N	Range	Minimum	Maximum	Mean		Std. Deviation	Variance
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic
strat	170	22.0	7.2	29.2	16.086	.3073	4.0065	16.052
Valid N (listwise)	170							



(4) Analyze the tstsc and top10

Dependent Variable: tstsc
Variables Entered: top10



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.799[a]	.638	.636	4.208723

a. Predictors: (Constant), top10

R^2=0.638 and Adjusted R^2=0.636 so the fitting degree of model and data is good

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	57.009	.622		91.621	.000
	top10	.238	.014	.799	17.211	.000

a. Dependent Variable: tstsc

Constant=57.009
Top10(B)=0.238
So tstsc=0.238top10+57.009