Homework2

Step 1:
Simulate 200 observations from the following linear model: Y = alpha + beta1 * X1 + beta2 * X2 + noise where • alpha=1, beta1=2, beta2=-1.5 • X1 ~ N(1, 4), X2 ~ N(3,1), noise ~ N(0,1)

Program:

```
DATA LinearSimulation(keep=X1 X2 Y);
TITLE Simulation;
alpha=1;
beta1=2;
beta2=-1.5;

DO i = 1 TO 200;                                  /*200 observations */
  UnifVals = rand("Uniform");     /*U(0,1)*/
  X1 = 1 + (4-1)*UnifVals;        /*Given X1 ~ (1,4)   */
  X2 = 3 + (1-3)*UnifVals;        /*Given X2 ~ (3,1)    */
  noise=UnifVals;                 /*Given noise ~ N(0,1)*/
  Y = alpha+beta1*X1+beta2*X2+noise;
OUTPUT;
END;
RUN;
PROC PRINT DATA=LinearSimulation LABEL;
RUN;
```

Output:

Simulation

Obs	X1	X2	Y
1	1.33466	2.77689	-0.38445
2	2.58702	1.94198	3.79008
3	1.94835	2.36777	1.66115
4	1.14694	2.90204	-1.01021
5	2.19816	2.20122	2.49388
6	2.10061	2.26626	2.16870
7	3.71892	1.18739	7.56306
8	3.73684	1.17544	7.62281
9	2.55083	1.96611	3.66945
10	2.36519	2.08987	3.05064
11	3.13706	1.57529	5.62354
12	1.40212	2.73192	-0.15961
13	3.05912	1.62725	5.36373
14	1.96318	2.35788	1.71060
15	2.92950	1.71366	4.93168
16	1.75655	2.49563	1.02184
17	1.68123	2.54585	0.77076
18	2.95793	1.69471	5.02645
19	3.01576	1.65616	5.21921
20	2.39373	2.07084	3.14578
21	2.51180	1.99214	3.53932
22	1.51558	2.65628	0.21859
23	1.43915	2.70724	-0.03618
24	3.17157	1.55229	5.73856
25	2.88624	1.74251	4.78745
26	2.64572	1.90285	3.98573
27	1.61546	2.58969	0.55153
28	3.46299	1.35801	6.70997

Step 2 & 3:
Define a new binary variable Y_bin such that Y_bin=1 if Y>0 and Y_bin=0 otherwise

Make the final data contain only 4 variables: X1, X2, Y and Y_bin.

Program:

```
DATA LinearSimulation(keep=X1 X2 Y Y_bin);
TITLE Simulation;
alpha=1;
beta1=2;
beta2=-1.5;

DO i = 1 TO 200;                                  /*200 observations */
  UnifVals = rand("Uniform");      /*U(0,1)*/
  X1 = 1 + (4-1)*UnifVals;        /*Given X1 ~ (1,4)   */
  X2 = 3 + (1-3)*UnifVals;        /*Given X2 ~ (3,1)    */
  noise=UnifVals;                 /*Given noise ~ N(0,1)*/
  Y = alpha+beta1*X1+beta2*X2+noise;
  if(Y>0)then Y_bin=1;
  ELSE Y_bin=0;
OUTPUT;
END;
RUN;
PROC PRINT DATA=LinearSimulation LABEL;
RUN;
```

Output:

Simulation

Obs	X1	X2	Y	Y_bin
1	1.79199	2.47201	1.13996	1
2	2.89027	1.73982	4.80091	1
3	2.28982	2.14012	2.79941	1
4	1.39180	2.73880	-0.19399	0
5	1.14107	2.90595	-1.02976	0
6	2.63957	1.90695	3.96523	1
7	3.25413	1.49725	6.01377	1
8	3.34432	1.43712	6.31440	1
9	1.75328	2.49781	1.01093	1
10	2.47007	2.01996	3.40022	1
11	1.81895	2.45403	1.22984	1
12	1.96404	2.35730	1.71348	1
13	3.42343	1.38438	6.57809	1
14	3.73203	1.17864	7.60678	1
15	1.35606	2.76263	-0.31314	0

Step 4:

Calculate the range of X1 and X2 (only range and no other statistics).

Program:

```
DATA LinearSimulation(keep=X1 X2 Y Y_bin);
TITLE Simulation;
alpha=1;
beta1=2;
beta2=-1.5;

DO i = 1 TO 200;                                  /*200 observations */
  UnifVals = rand("Uniform");     /*U(0,1)*/
  X1 = 1 + (4-1)*UnifVals;        /*Given X1 ~ (1,4)   */
  X2 = 3 + (1-3)*UnifVals;        /*Given X2 ~ (3,1)    */
  noise=UnifVals;                 /*Given noise ~ N(0,1)*/
  Y = alpha+beta1*X1+beta2*X2+noise;
  if(Y>0)then Y_bin=1;
  ELSE Y_bin=0;
OUTPUT;
END;
RUN;
PROC MEANS DATA=LinearSimulation RANGE;
VAR X1 X2;
PROC PRINT DATA=LinearSimulation LABEL;
RUN;
```

Simulation

The MEANS Procedure

Variable	Range
X1	2.9566656
X2	1.9711104

Simulation

Obs	X1	X2	Y	Y_bin
1	2.43561	2.04293	3.28535	1
2	2.42528	2.04981	3.25095	1
3	1.49614	2.66924	0.15380	1
4	1.27544	2.81638	-0.58188	0
5	1.43157	2.71229	-0.06143	0
6	2.01630	2.32246	1.88768	1
7	2.64278	1.90482	3.97592	1
8	2.48138	2.01241	3.43794	1
9	1.90904	2.39397	1.53015	1

Step 5:

Check if Y follows a normal distribution. What graphics and statistics would you look into?

ANS:

After making histogram data, if the shape of the distribution resembles bell curve the data is likely normal.
Graphic: Plotting histogram and normal distribution curve

If the data meets the requirement of 68-95-99, that means 68% of the data should be in the range of one standard deviation, 95% data in the range of two standard deviations and 99% of the data in the range of 3 standard deviation

We look statistics of Mean, Mode and standard deviation. In normal distribution mean and mode are equal.

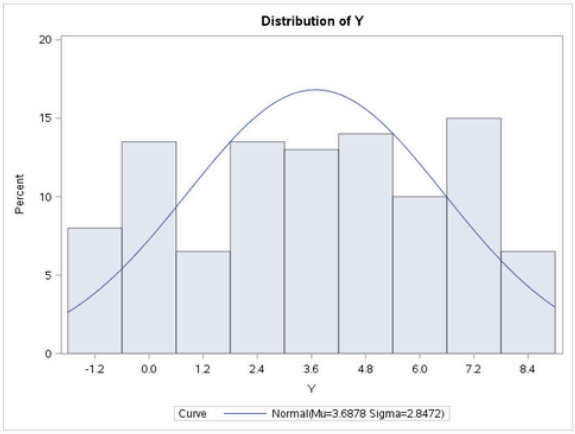
Program:

```
DATA LinearSimulation(keep=X1 X2 Y Y_bin);
TITLE Simulation;
alpha=1;
beta1=2;
beta2=-1.5;

DO i = 1 TO 200;                                    /*200 observations */
  UnifVals = rand("Uniform");      /*U(0,1)*/
  X1 = 1 + (4-1)*UnifVals;        /*Given X1 ~ (1,4)   */
  X2 = 3 + (1-3)*UnifVals;        /*Given X2 ~ (3,1)    */
  noise=UnifVals;                 /*Given noise ~ N(0,1)*/
  Y = alpha+beta1*X1+beta2*X2+noise;
  if(Y>0)then Y_bin=1;
  ELSE Y_bin=0;
OUTPUT;
END;
RUN;
PROC MEANS DATA=LinearSimulation RANGE;
VAR X1 X2;
proc univariate;
VAR Y;
HISTOGRAM Y / NORMAL;
PROBPLOT Y / NORMAL;
  run;
PROC PRINT DATA=LinearSimulation LABEL;
RUN;
```

Output:

The UNIVARIATE Procedure



Simulation

The UNIVARIATE Procedure
Fitted Normal Distribution for Y

Parameters for Normal Distribution		
Parameter	Symbol	Estimate
Mean	Mu	3.687807
Std Dev	Sigma	2.847246

Goodness-of-Fit Tests for Normal Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.07775958	Pr > D	<0.010
Cramer-von Mises	W-Sq	0.26372389	Pr > W-Sq	<0.005

Step 6:

Count how many ``1'' s you observed in Y_bin.

Program:

```
DATA YBinCount;
SET LinearSimulation;
PROC FREQ DATA=YBinCount;
TABLE Y_bin;
RUN;
```

Output:

Simulation

The FREQ Procedure

Y_bin	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	35	17.50	35	17.50
1	165	82.50	200	100.00

Step 7:

Create a new variable “sign” such that sign=“positive” when Y_bin=1 and sign=“negative” otherwise.

Program:

```
DATA CreateSign;
SET LinearSimulation;
if(Y_bin)then sign="Positive";
ELSE sign="Negative";
RUN;
PROC PRINT Data=CreateSign;
RUN;
```

Output:

Obs	X1	X2	Y	Y_bin	sign
1	3.13551	1.57632	5.61838	1	Positive
2	1.87279	2.41814	1.40931	1	Positive
3	3.11516	1.58990	5.55052	1	Positive
4	3.88929	1.07381	8.13097	1	Positive
5	1.01411	2.99059	-1.45295	0	Negative
6	2.77378	1.81748	4.41258	1	Positive
7	1.78332	2.47779	1.11107	1	Positive
8	1.94119	2.37254	1.63731	1	Positive
9	1.13823	2.90785	-1.03923	0	Negative
10	1.94575	2.36950	1.65248	1	Positive
11	2.97359	1.68427	5.07864	1	Positive
12	3.64853	1.23432	7.32842	1	Positive
13	1.15902	2.89399	-0.96994	0	Negative
14	1.27000	2.82000	-0.59999	0	Negative
15	3.24061	1.50626	5.96868	1	Positive
16	3.60297	1.26469	7.17655	1	Positive
17	2.44252	2.03832	3.30841	1	Positive
18	3.95827	1.02782	8.36089	1	Positive
19	2.62818	1.91454	3.92728	1	Positive
20	1.89529	2.40314	1.48430	1	Positive
21	3.67989	1.21340	7.43298	1	Positive
22	3.05249	1.63168	5.34162	1	Positive
23	1.21414	2.85724	-0.78619	0	Negative
24	3.39709	1.40194	6.49028	1	Positive
25	3.78112	1.14592	7.77041	1	Positive
26	2.70768	1.86155	4.19226	1	Positive
27	3.02774	1.64817	5.25915	1	Positive
28	1.67039	2.55308	0.73462	1	Positive

Step 8:

Compare the distribution of X1 in the "positive" group and the "negative" group. What statistics would you look into?

Ans:

Mean, Median and standard deviation by running proc univariate command sorted by sign (either positive or negative)

Program:

```
PROC SORT DATA=CreateSign;
BY sign;
RUN; /*in PROC UNIVARIATE*/
PROC UNIVARIATE DATA=CreateSign;
BY sign; /* tells SAS to sort data by SIGN*/
VAR X1; /* tells SAS to produce statistics of X1*/
RUN;
```

Output:

The UNIVARIATE Procedure
Variable: X1

sign=Negative

Moments			
N	35	Sum Weights	35
Mean	1.24432809	Sum Observations	43.551483
Std Deviation	0.12599533	Variance	0.01587482
Skewness	0.00642775	Kurtosis	-1.2910577
Uncorrected SS	54.7320776	Corrected SS	0.53974398
Coeff Variation	10.1255714	Std Error Mean	0.0212971

Basic Statistical Measures			
Location		Variability	
Mean	1.244328	Std Deviation	0.12600
Median	1.214144	Variance	0.01587
Mode	.	Range	0.42857
		Interquartile Range	0.22486

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	58.42712	Pr > |t|	<.0001
Sign	M	17.5	Pr >= |M|	<.0001
Signed Rank	S	315	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	1.44268
99%	1.44268
95%	1.43398
90%	1.41704
75% Q3	1.36309
50% Median	1.21414
25% Q1	1.13823
10%	1.07103
5%	1.05606
1%	1.01411
0% Min	1.01411

The UNIVARIATE Procedure
Variable: X1

sign=Positive

Moments			
N	165	Sum Weights	165
Mean	2.78431786	Sum Observations	459.412447
Std Deviation	0.71951068	Variance	0.51769562
Skewness	-0.1326752	Kurtosis	-1.1335859
Uncorrected SS	1364.05236	Corrected SS	84.9020814
Coeff Variation	25.8415424	Std Error Mean	0.05601383

Basic Statistical Measures			
Location		Variability	
Mean	2.784318	Std Deviation	0.71951
Median	2.864620	Variance	0.51770
Mode	.	Range	2.54195
		Interquartile Range	1.18542

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	49.70769	Pr > |t|	<.0001
Sign	M	82.5	Pr >= |M|	<.0001
Signed Rank	S	6847.5	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	3.99601
99%	3.95827
95%	3.88070
90%	3.75152
75% Q3	3.35138
50% Median	2.86462
25% Q1	2.16596
10%	1.78648
5%	1.63010
1%	1.47301
0% Min	1.45406