Homework 3

Q1).

Ans:

Given linear equation

```
   Y = alpha+beta1*X1+beta2*X2+noise;
   X1 = rand("normal",1,4);        /*Given X1 ~ (1,4)   */
   X2 = rand("normal",3,1);       /*Given X2 ~ (3,1)    */
   noise=rand("normal",0,1              /*Given noise ~ N(0,1)*/
```

for model diagnostics, we perform the following

1.Independent:

Residuals are taken into new data set named diagnostic using

```
proc reg data=LinearSimulation;
model y=x1 x2 / r;
output out=diagnostics r=residual;
```

for checking independence, we plot residuals vs x1 and residuals vs x2 using

```
proc plot data=diagnostics;
plot residual*x1;
plot residual*x2;
run;
```

2. Normally distributed.

We plot histogram for residuals using proc chart

```
proc chart data=diagnostics;
vbar residual;
RUN;
proc chart data=diagnostics;
hbar residual;
RUN;
```

To test normality, we use univariate for the variable residual

```
proc UNIVARIATE data=diagnostics NORMAL PLOT;
VAR residual;
RUN;
```

3)Mean 0.

We can say mean is 0 after checking the plots drawn in previous step

To perform hypothesis test, we use proc means

```
PROC MEANS data=diagnostics T PRT;
 VAR residual;
```

4. Constant variance

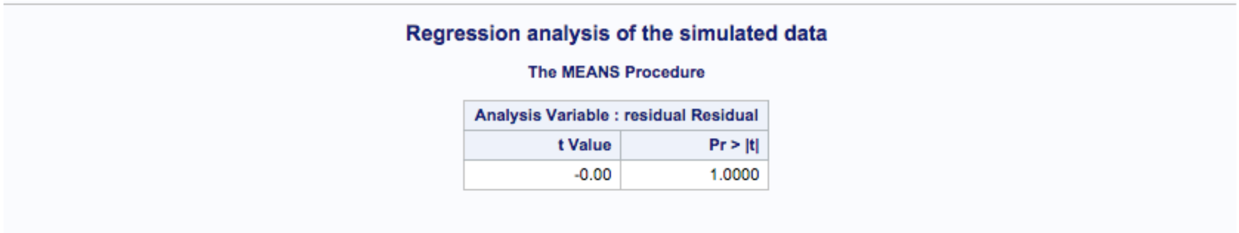
For variance we check the plots generated in Univariate plot step.

Program:

```
DATA LinearSimulation(keep=X1 X2 Y y_bin);
alpha=1;
beta1=2;
beta2=-1.5;
DO i = 1 TO 200;                          /*200 observations */
 X1 = rand("normal",1,4);       /*Given X1 ~ (1,4)   */
   X2 = rand("normal",3,1);       /*Given X2 ~ (3,1)    */
   noise=rand("normal",0,1);               /*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 reg data=LinearSimulation;
model y=x1 x2 / r;
output out=diagnostics r=residual;
title Regression analysis of the simulated data;
RUN;
proc plot data=diagnostics;
plot residual*x1;
plot residual*x2;
run;
proc chart data=diagnostics;
vbar residual;
RUN;
proc chart data=diagnostics;
hbar residual;
RUN;
proc UNIVARIATE data=diagnostics NORMAL PLOT;
```

```
VAR residual;
RUN;
PROC MEANS data=diagnostics T PRT;
 VAR residual;
run;
```

Output:



Regression analysis of the simulated data

The MEANS Procedure

Analysis Variable : residual Residual	
t Value	Pr > |t|
-0.00	1.0000

Pr value is 1. So that, hypothesis is strong pr value. We can not reject the hypothesis

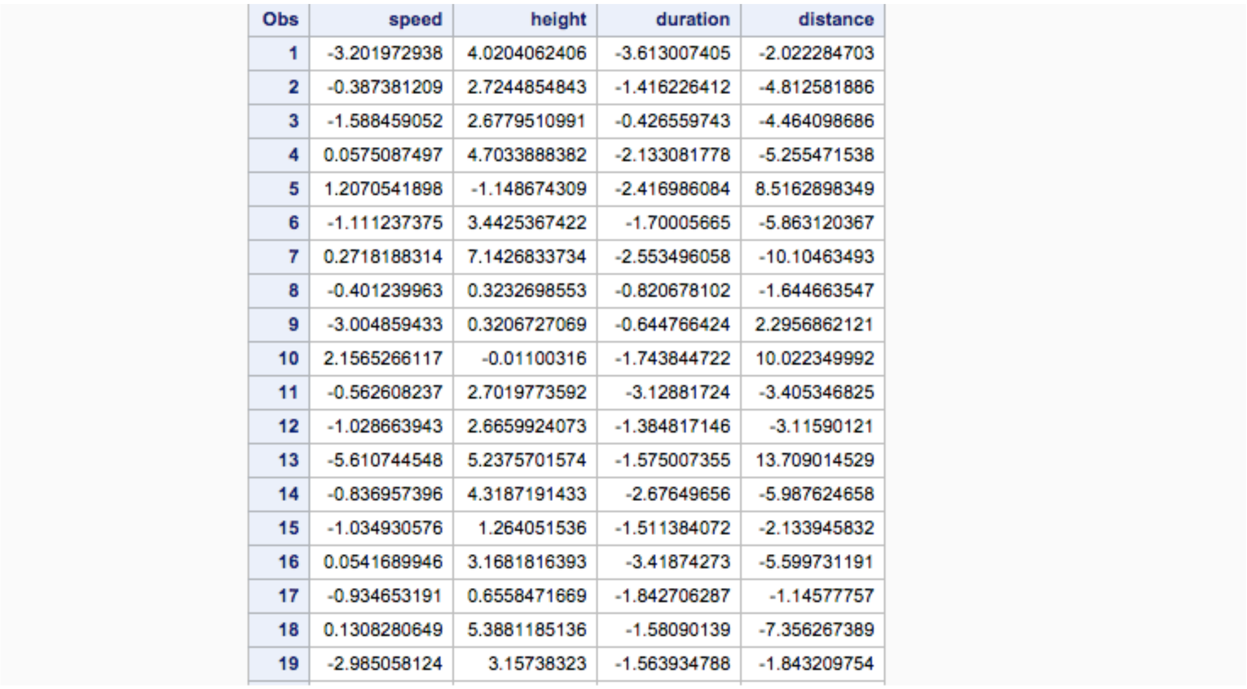
Q2).

1.
Ans: we read data from xls file and print the data using Proc statement

Program:

```
proc import out=FAA
datafile='/folders/myfolders/FAA.xls'
dbms=xls replace;
getnames=yes;
run;
proc print data=faa;
run;
proc univariate data=faa;
run;
```

Output:



Obs	speed	height	duration	distance
1	-3.201972938	4.0204062406	-3.613007405	-2.022284703
2	-0.387381209	2.7244854843	-1.416226412	-4.812581886
3	-1.588459052	2.6779510991	-0.426559743	-4.464098686
4	0.0575087497	4.7033888382	-2.133081778	-5.255471538
5	1.2070541898	-1.148674309	-2.416986084	8.5162898349
6	-1.111237375	3.4425367422	-1.70005665	-5.863120367
7	0.2718188314	7.1426833734	-2.553496058	-10.10463493
8	-0.401239963	0.3232698553	-0.820678102	-1.644663547
9	-3.004859433	0.3206727069	-0.644766424	2.2956862121
10	2.1565266117	-0.01100316	-1.743844722	10.022349992
11	-0.562608237	2.7019773592	-3.12881724	-3.405346825
12	-1.028663943	2.6659924073	-1.384817146	-3.11590121
13	-5.610744548	5.2375701574	-1.575007355	13.709014529
14	-0.836957396	4.3187191433	-2.67649656	-5.987624658
15	-1.034930576	1.264051536	-1.511384072	-2.133945832
16	0.0541689946	3.1681816393	-3.41874273	-5.599731191
17	-0.934653191	0.6558471669	-1.842706287	-1.14577757
18	0.1308280649	5.3881185136	-1.58090139	-7.356267389
19	-2.985058124	3.15738323	-1.563934788	-1.843209754

2.

To study the relation among variables speed, height, duration and distance

i) We do plots for distance vs speed, height and duration

```
proc plot data=faa;
plot distance*speed;
plot distance*height;
plot distance*duration;
```

ii) calculate the correlation matrix for distance vs speed, height and duration

```
PROC CORR data = faa;
VAR distance;
WITH speed height duration;
RUN;
proc corr data=faa;
var speed height duration;
with distance;
title Correlaiton coefficients with Y;
run;
```

iii) we do regression analysis using

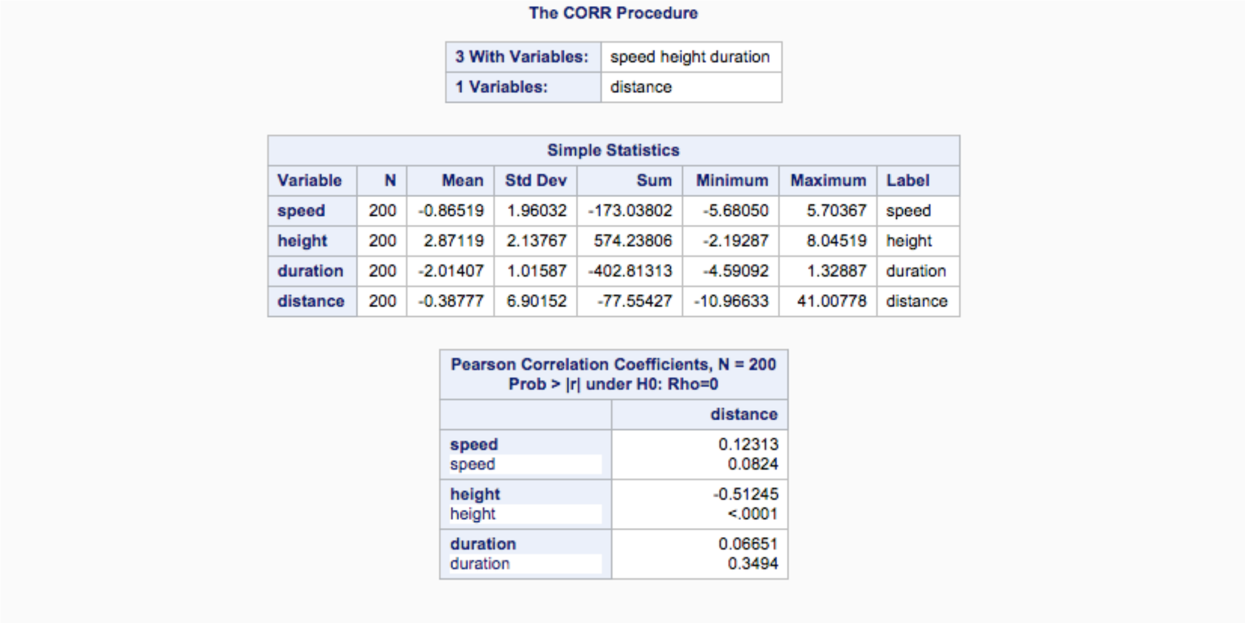
```
proc reg data=faa;
model distance=speed height duration;
title Regression analysis of the simulated data set;
```

run;

iv)model checking

1. Independent. 2. Normally distributed. 3. Mean 0. 4. Constant variance are checked for regression model

output :



The CORR Procedure

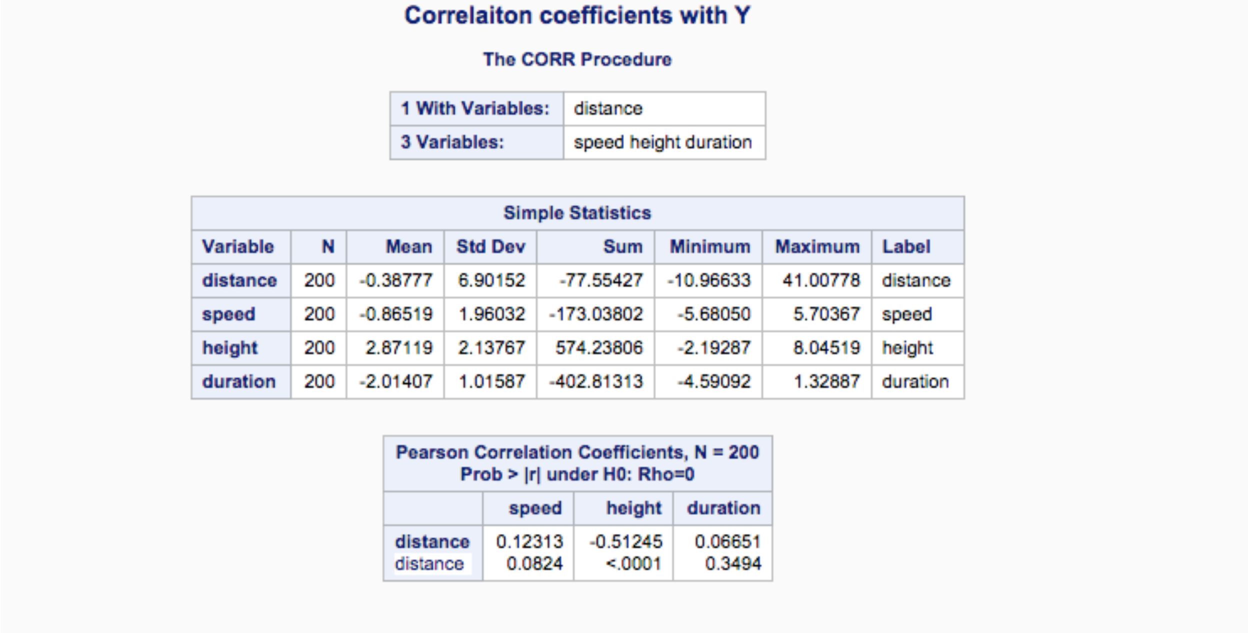
3 With Variables:	speed height duration
1 Variables:	distance

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
speed	200	-0.86519	1.96032	-173.03802	-5.68050	5.70367	speed
height	200	2.87119	2.13767	574.23806	-2.19287	8.04519	height
duration	200	-2.01407	1.01587	-402.81313	-4.59092	1.32887	duration
distance	200	-0.38777	6.90152	-77.55427	-10.96633	41.00778	distance

Pearson Correlation Coefficients, N = 200
Prob > |r| under H0: Rho=0

	distance
speed speed	0.12313 0.0824
height height	-0.51245 <.0001
duration duration	0.06651 0.3494



Correlaiton coefficients with Y

The CORR Procedure

1 With Variables:	distance
3 Variables:	speed height duration

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
distance	200	-0.38777	6.90152	-77.55427	-10.96633	41.00778	distance
speed	200	-0.86519	1.96032	-173.03802	-5.68050	5.70367	speed
height	200	2.87119	2.13767	574.23806	-2.19287	8.04519	height
duration	200	-2.01407	1.01587	-402.81313	-4.59092	1.32887	duration

Pearson Correlation Coefficients, N = 200
Prob > |r| under H0: Rho=0

	speed	height	duration
distance distance	0.12313 0.0824	-0.51245 <.0001	0.06651 0.3494

```
3) data FAA;
 do i=1 to 200; speed=-1+2*rannor(12); height=3+2*rannor(12);
duration=-2+rannor(12); error=rannor(12);
 alpha=1;
 beta_1=2;
 beta_2=-1.5;
 beta_3=1; distance=alpha+beta_1*speed+beta_2*height+beta_3*speed*speed+error; output;
end; keep speed height duration distance; run; proc export data=FAA dbms=excel2002
outfile='C:\Users\...\FAA.xls' rep
```

after observing the ouput, the given linear equation satisfies the all conditions we checked in 4 steps i.e Plots, correlation matric, regression analysis and model checking as the data from the output is similar to give input file.