EXECUTIVE SUMMARY

Problem: Data for 18 people has been collected. We are trying to build a model that predicts a person's income (I) given their age (A), and the order (O) in which the persons age was collected. In addition, we have been asked to give a predicted income of a 35 and 50-year-old, and a 95% confidence interval for this prediction.

Variable	Mean	SD	Min	Max	Shape
Income (100s of dollars)	200	10.4	178	214.5	Unimodal
Age (years)	40	13.3	20	60	Unimodal
Order	9.5	5.3	1	18	Uniform

Recommended Model: I^=-51.8+20.55A-.53A^2+.0043A^3
I income in 100s of dollars. Age in years. Sia=6.32, R^2=.69

The predicted income for a person who is 35 years old = 202.56

- I^=-51.8+20.55(35)-.53*35^2+.0043*35^3
- The 95% CI= 202.56+/- (2.145)(6.32)(1+(1/18)+[(35-40)^2/((18-1)13.3^2)])^(1/2)
- =(188.56,216.56)

The predicted income for a person who is 50 years old is 188.2

- I^=-51.8+20.55(50)-.53*50^2+.0043*50^3
- The 95% CI = 188.2+/-(2.145)(6.32)(1+(1/18)+[(50-40)^2/((18-1)13.3^2)])^(1/2)
- =(174.2, 202.2)

ANALYSIS

Problem: Data for 18 people has been collected. We are trying to predict a person's income (I) given their age (A), and the order (O). In addition, we have been asked to give a predicted income of a 35 and 50-year-old, and a 95% confidence interval for this prediction.

Data: The data has been entered into the computer and printed out (p.1a). The data has been checked for accuracy and has been verified to be the same as the data provided to us.

Income: (In 100s of $), the dependent variable, has an average of 200, a standard deviation of 10.5, ranges from a minimum value of 178 and a maximum value of 214.5. The shape of the distribution appears unimodal and fairly symmetric, possibly a bit skewed to the right.

Age: An independent variable, has an average of 40, a standard deviation of 13.3, ranges from a minimum value of 20 and a maximum value of 60. The shape of the distribution appears uniform.

Order: An independent variable, has an average of 9.5, a standard deviation of 5.3, ranges from a minimum value of 1 and a maximum value of 18. The shape of the distribution is uniform.

Income (y) vs independent variables: We calculated the correlations of all pairs of variables including the order variables examining the correlation matrix (1d), we see the following significant results: The corr between:
Income vs. Age: R=.485, 23.5% of the variability in the y scores about y bar is explained by the simple regression between income and age. The std dev of the y scores about the simple regression model is 48.4% of the std dev of the y scores about y bar. As age increases, y-hat increases in this fitted simple regression model.

Income vs. Order: There is no significant relationship between Income and order

Scatterplots of Income vs Independent Variables: We calculated the scatterplots of all pairs of variables including the order variables (1E)
Income Vs. Age: There appears to be a positive linear relationship between these two variables.

Scatterplots between pair of independent variables:
Age vs. Order: There doesn't appear to be a linear relationship between these two variables.

No significant relationship found between Order and any other variables after examining correlation matrix and scatterplots.

Fit of first order model: E(I)=184.5 + .38A.
Examining the excel results for this model, we find Si.a= 9.44, Si=10.5, R^2=.23. For testing that B1=0 (Since A holds), we find a p value of .041., we find p for t = .041<.05=a. Thus the first order model is a significant improvement over the model E(y)=Bo

Fit of second order model: E(I) = 177.3+.78A-.005(A^2)
We see the standard deviation of the residuals about the second order model is 9.7, which we note is a tad larger/worse than the standard deviation of the y scores about the first order model. R^2 =.24. for testing Ho: B2=0, since A holds, we find p for t = .75>.05, so we accept Ho and thus the second order model is not a significant improvement over the first order model.

Residual plot for the highest order model fitted: Examining the residual plot of the second order model (2b-2), the expected value/ average of the residual seems to change as predicted y changes. Since there is a relationship between predicted y and the residual, assumption A is violated, we must try to fit a higher model.
Example: E(e^|y^=198)>0

Fit of the third order model: E(I)=-51.8+20.55A-.53A^2+.0043A^3
We see the standard deviation of the residuals about the third order model is 6.32, which we note is better than the standard deviation of the residuals about the first order model (current best) of 9.44. R^2=.699. For testing Ho: B3=0, since A holds, we find p for t = .00039<.05 so we reject Ho and conclude that we prefer the third order model to the first order model. We must always fit one higher so we move to the fourth.

Fit of the fourth order model: E(I)=
We see the standard deviation of the residuals about the fourth order model is 6.25 which we note is slightly better than the standard deviation of the residuals about the third order model (current best) of 6.32. R^2=.73. For testing Ho:B4=0 since A holds, we find p for t =.27>.05=a so we accept Ho and conclude that the fourth order model is not significantly better than the third order model.

Residual plot for the highest order model fitted: Examining the residual plot of the fourth order model (2d-2), the expected value/ average of the residual seems to remain at 0 regardless of predicted Y. Also the standard deviation appears to be constant with respect to predicted Y. So we stop here.

Since there is only one variable being added each time we don't need to test individual variables.

Residual plot for the FINAL (third order) model: Examining the residual plot for this model (2c-2), we note that:
(1) The mean of the residuals is 0 regardless of the predicted income, assumption A holds.
(2) The variance of the residuals is constant regardless of predicted income, assumption B holds.
Histogram of residuals for the FINAL (first order) model: Examining the Histogram of the residuals, we see that they appear to not be normal.

Independence of the residuals for the FINAL (third order) model: Given that the data has been ordered by the order it was collected, we examine the plot of residual t vs. residual t-1 (3a), there doesn't appear to be a line through this data which has a very positive or negative slope, so we conclude that the residuals are independent. If we run a regression model for e(t) = Bo + B1e(t-1) we get a correlation of r = .18 and p=.48 and conclude that B1 is not different than 0.

The prediction model is given by I^=-51.8+20.55A-.53A^2+.0043A^3
The predicted income for a person who is 35 years old = 202.56

- The 95% CI= (188.56, 216.56)

The predicted income for a person who is 50 years old is 188.20

- The 95% CI = (174.20, 202.20)

Conclusion: We recommend using the third order model given above to describe the relationship.

Table of contents:

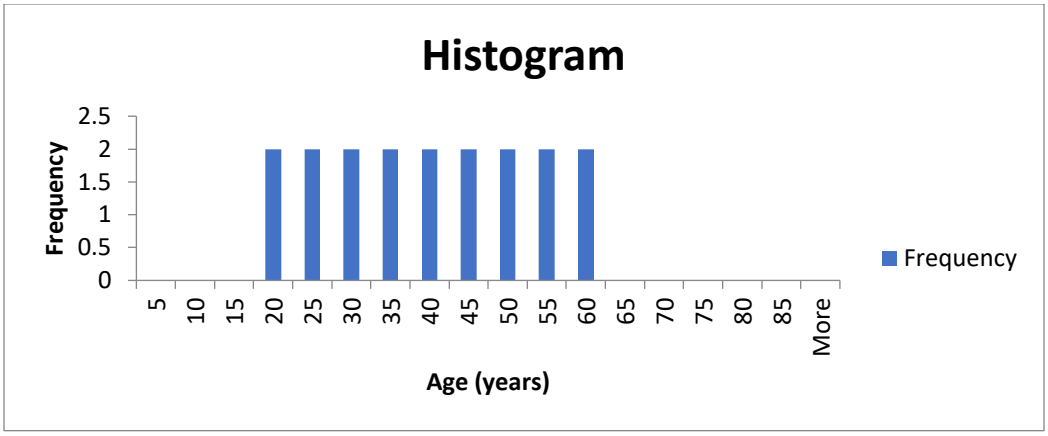
1a:

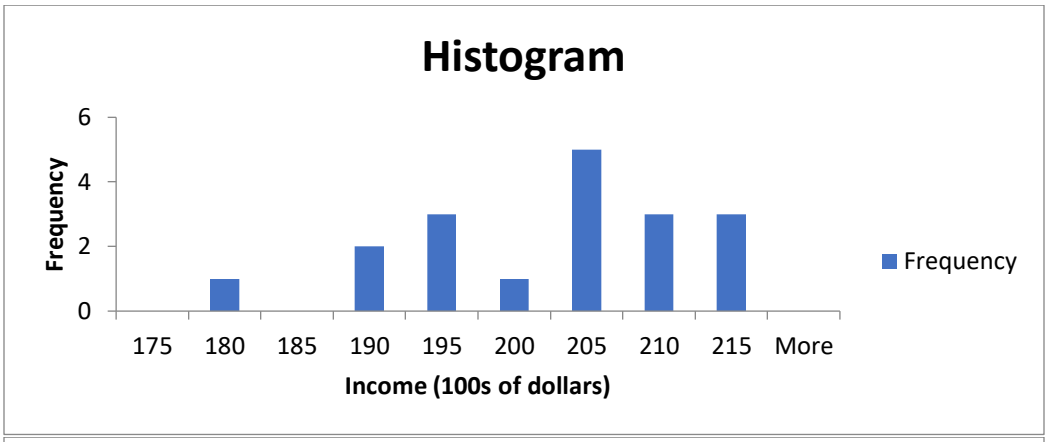
Order	Age	Age^2	Age^3	Age^4	Income
1	35	1225	42875	1500625	212.6
2	60	3600	216000	12960000	210.8
3	45	2025	91125	4100625	186
4	55	3025	166375	9150625	201.4
5	30	900	27000	810000	203.6
6	50	2500	125000	6250000	200.7
7	25	625	15625	390625	190.1
8	45	2025	91125	4100625	193.6
9	35	1225	42875	1500625	208.9
10	50	2500	125000	6250000	197
11	25	625	15625	390625	201.6
12	20	400	8000	160000	178
13	60	3600	216000	12960000	214.5
14	30	900	27000	810000	209.6
15	40	1600	64000	2560000	201.1
16	40	1600	64000	2560000	191.1
17	20	400	8000	160000	186
18	55	3025	166375	9150625	209.4

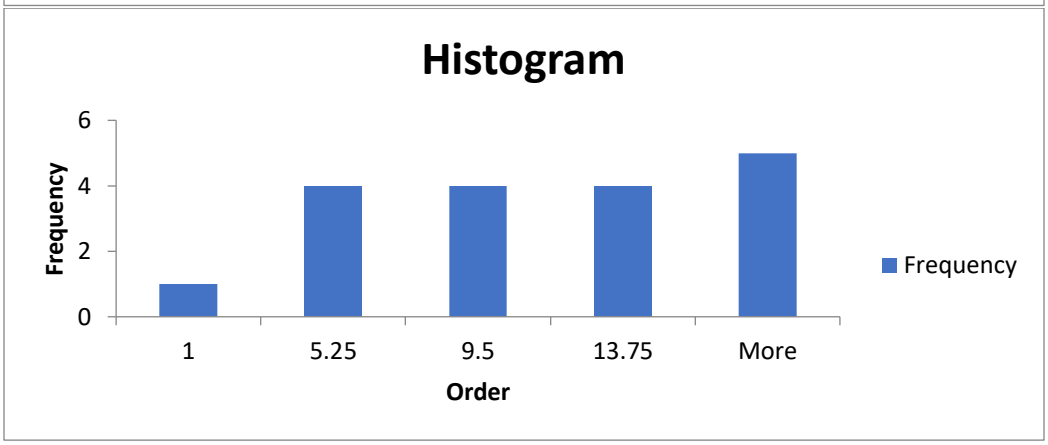
1b:

Order		*Age*		*Income*	
Mean	9.5	Mean	40	Mean	199.777778
Standard Deviation	5.33853913	Standard Deviation	13.2842233	Standard Deviation	10.4792543
Minimum	1	Minimum	20	Minimum	178
Maximum	18	Maximum	60	Maximum	214.5
Count	18	Count	18	Count	18

1c:


Histogram
Frequency
2.5
2
1.5
1
0.5
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
More
Age (years)
Frequency

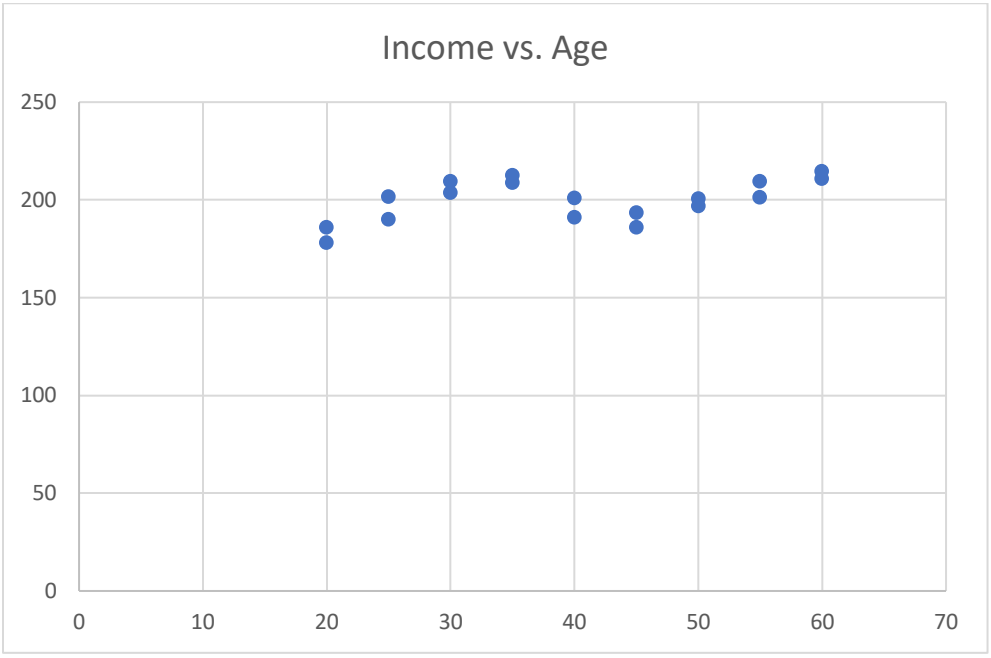

Histogram
Frequency
6
4
2
0
175
180
185
190
195
200
205
210
215
More
Income (100s of dollars)
Frequency

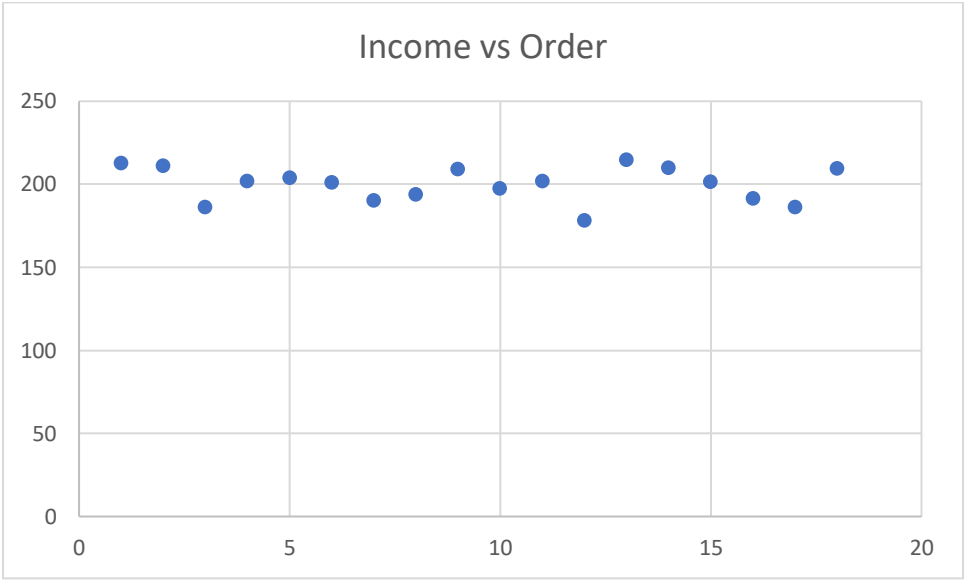

Histogram
Frequency
6
4
2
0
1
5.25
9.5
13.75
More
Order
Frequency

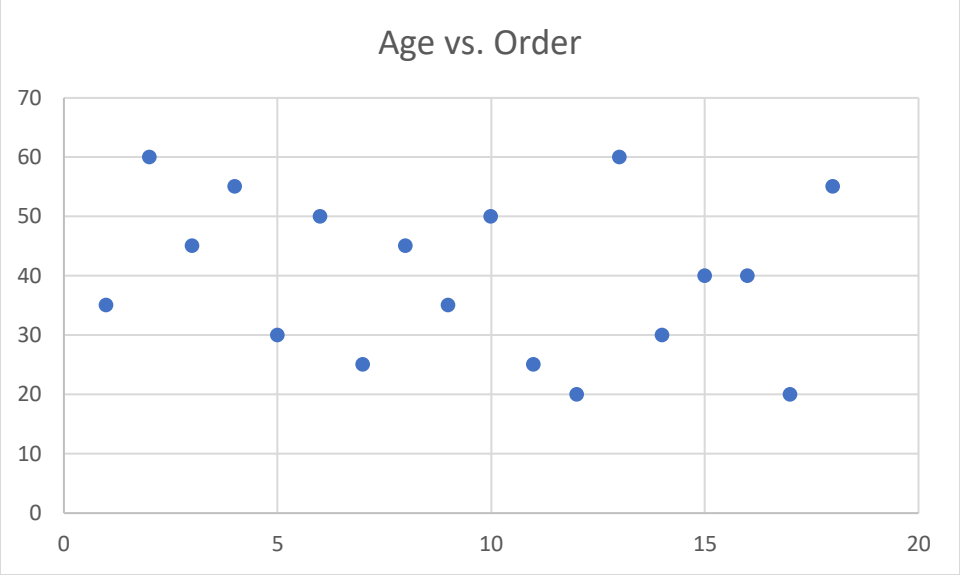
1D:

	Order	*Age*	*Income*
Order	1		
Age	-0.2032163	1	
Income	-0.1373224	0.48509458	1

1E:







2a-1:

SUMMARY OUTPUT

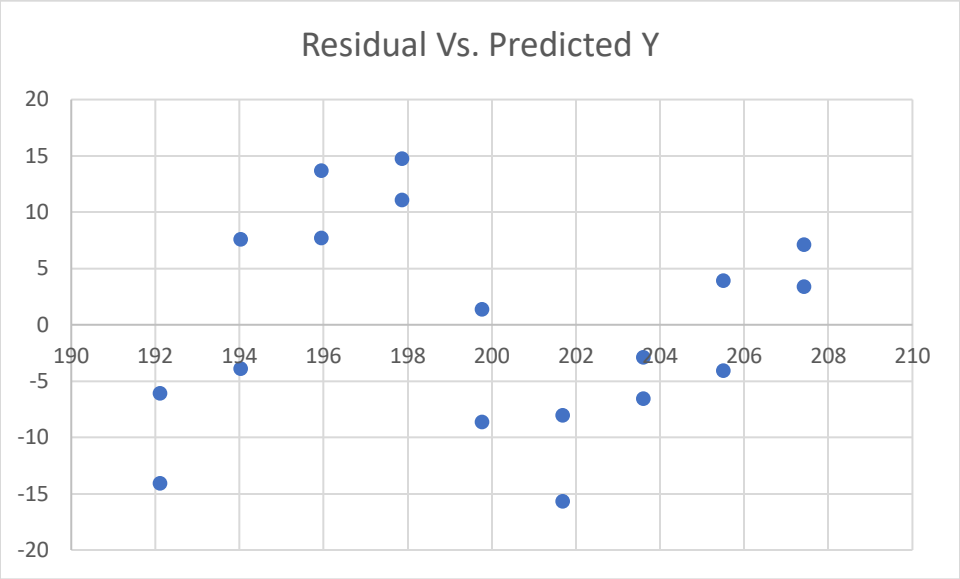
Regression Statistics	
Multiple R	0.48509458
R Square	0.23531675
Adjusted R Square	0.18752405
Standard Error	9.44573243
Observations	18

ANOVA

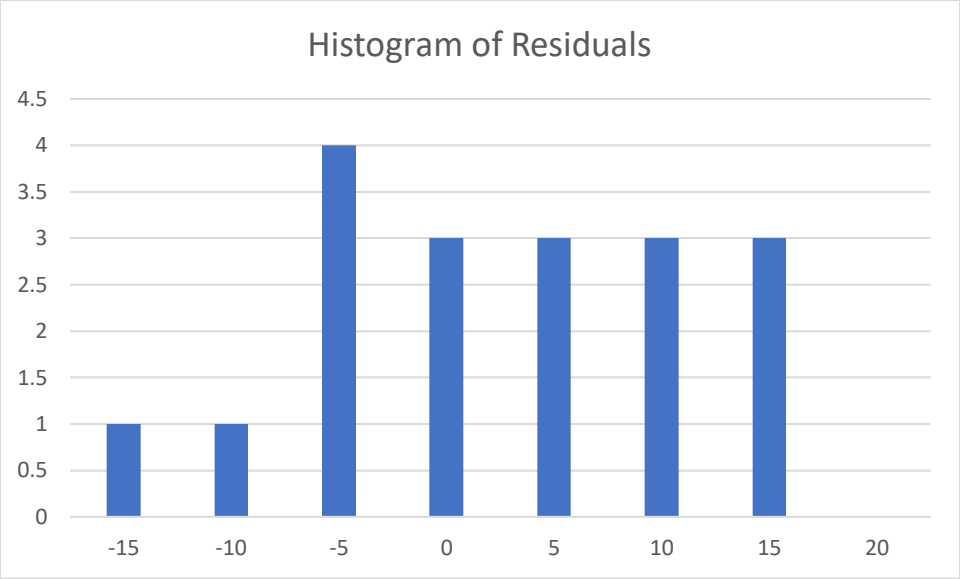
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	439.301333	439.30133	4.923696	0.041299
Residual	16	1427.54978	89.221861		
Total	17	1866.85111			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	184.471111	7.2485697	25.44931191	2.2655E-14	169.10483	199.837392	169.10483	199.837392
Age	0.38266667	0.17245469	2.21894032	0.04129914	0.01707905	0.74825428	0.01707905	0.74825428

2a-2:



2a-3:



2b-1:

SUMMARY OUTPUT

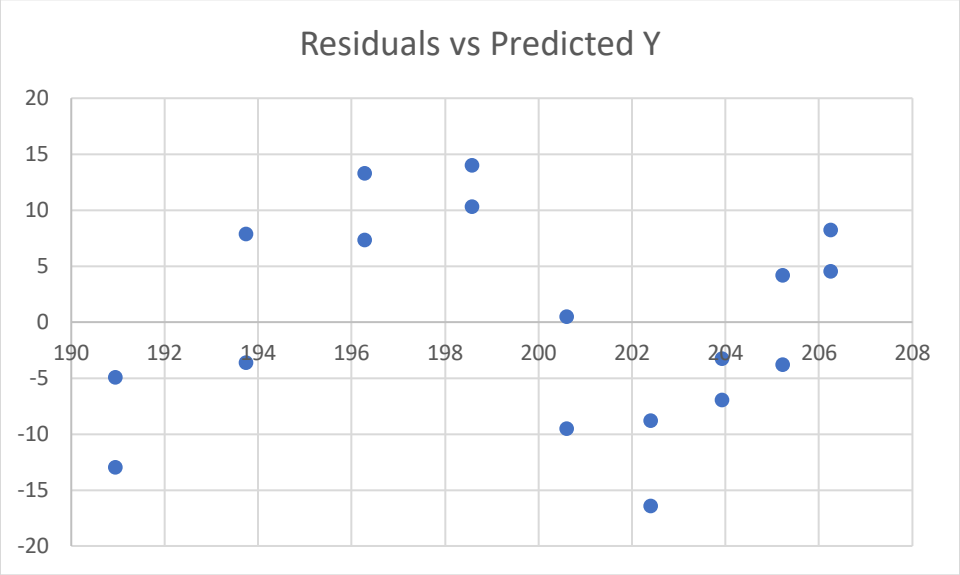
Regression Statistics	
Multiple R	0.490425 55
R Square	0.240517 22
Adjusted R Square	0.139252 85
Standard Error	9.722281 16
Observations	18

ANOVA

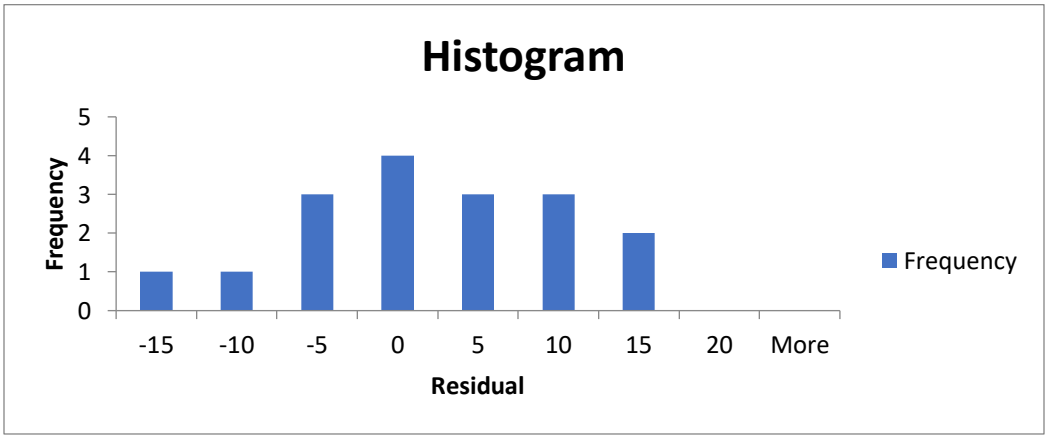
	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	449.009847	224.504924	2.375141	0.127023
Residual	15	1417.84126	94.5227509		
Total	17	1866.85111			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	177.27342	23.6655346	7.490784	1.917E-06	126.831527	227.715313	126.831527	227.715313
Age	0.7843982	1.2660153	0.619580	0.544831	-1.914049	3.4828462	-1.914049	3.4828462
Age^2	-0.005021	0.0156687	-0.320485	0.753021	-0.038419	0.0283757	-0.038419	0.0283757

2b-2:



2b-3:



2c-1:

SUMMARY OUTPUT

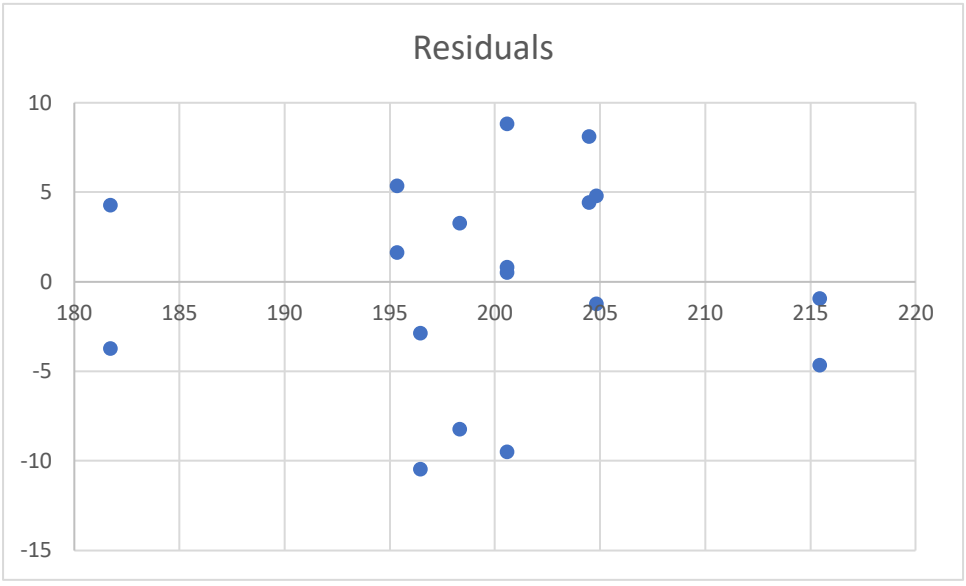
Regression Statistics	
Multiple R	0.836604
R Square	0.699906 25
Adjusted R Square	0.635600 45
Standard Error	6.325855 96
Observations	18

ANOVA

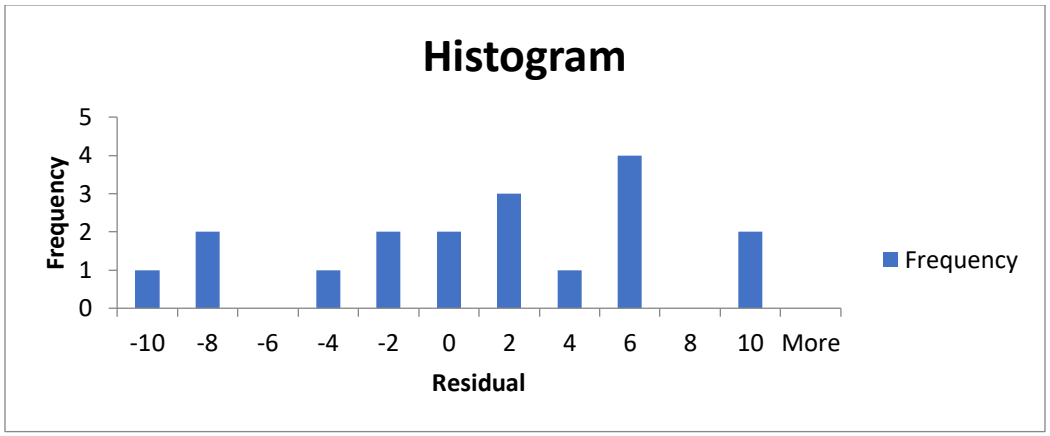
	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	1306.62076	435.540254	10.8840293	0.00059125
Residual	14	560.23035	40.0164536		
Total	17	1866.85111			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-51.756277	51.8136433	-0.998892 8	0.33479945	162.88549	59.372954	-162.88549	59.372954
Age	20.550275 4	4.34836608	4.725976 37	0.00032485	11.223957 7	29.876593 1	11.223957 7	29.876593 1
Age^2	-0.531526 7	0.11418647	-4.654900 8	0.00037161	-0.776432 3	-0.286621 1	-0.776432 3	-0.286621 1
Age^3	0.004387 54	0.00094775	4.629412 19	0.00039003	0.002354 81	0.006420 27	0.002354 81	0.006420 27

2c-2:



2c-3:



2d-1:

SUMMARY OUTPUT

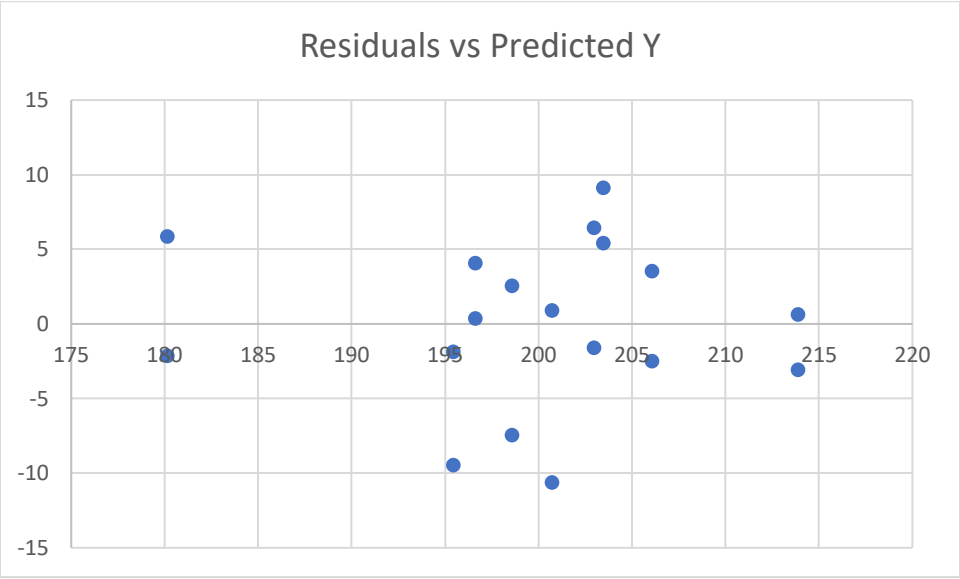
Regression Statistics	
Multiple R	0.852832 39
R Square	0.727323 08
Adjusted R Square	0.643422 49
Standard Error	6.257593 61
Observations	18

ANOVA

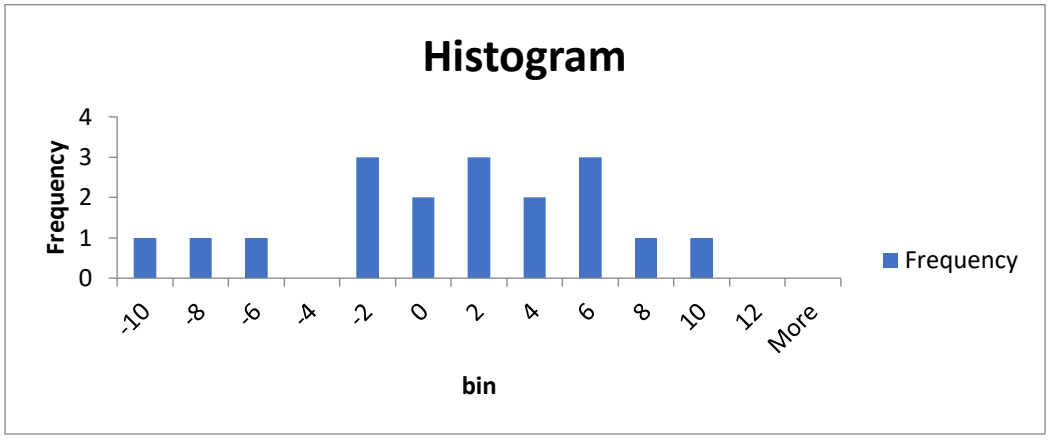
	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	1357.8039	339.450975	8.66886722	0.00122938
Residual	13	509.047211	39.1574777		
Total	17	1866.85111			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-254.589	184.667296	-1.37863	0.191268	-653.539	144.36043	-653.539	144.36043
Age	44.0973083	21.040251	2.09585467	0.056234	-1.35739	89.552072	-1.35739	89.552072
Age^2	-1.501221	0.8556503	-1.75448	0.102873	-3.34974	0.3472903	-3.34974	0.3472903
Age^3	0.0212714	0.0147957	1.43749645	0.174209	-0.01069	0.0532367	-0.01069	0.0532367
age**4	-0.000105	9.2299E-05	-1.14329	0.273536	-0.00031	9.3875E-05	-0.00031	9.3875E-05

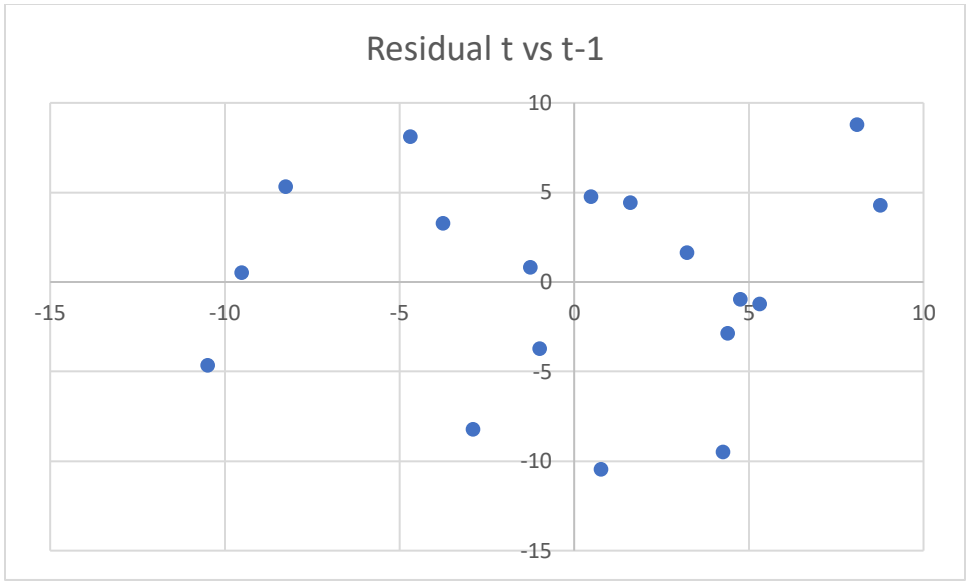
2d-2:



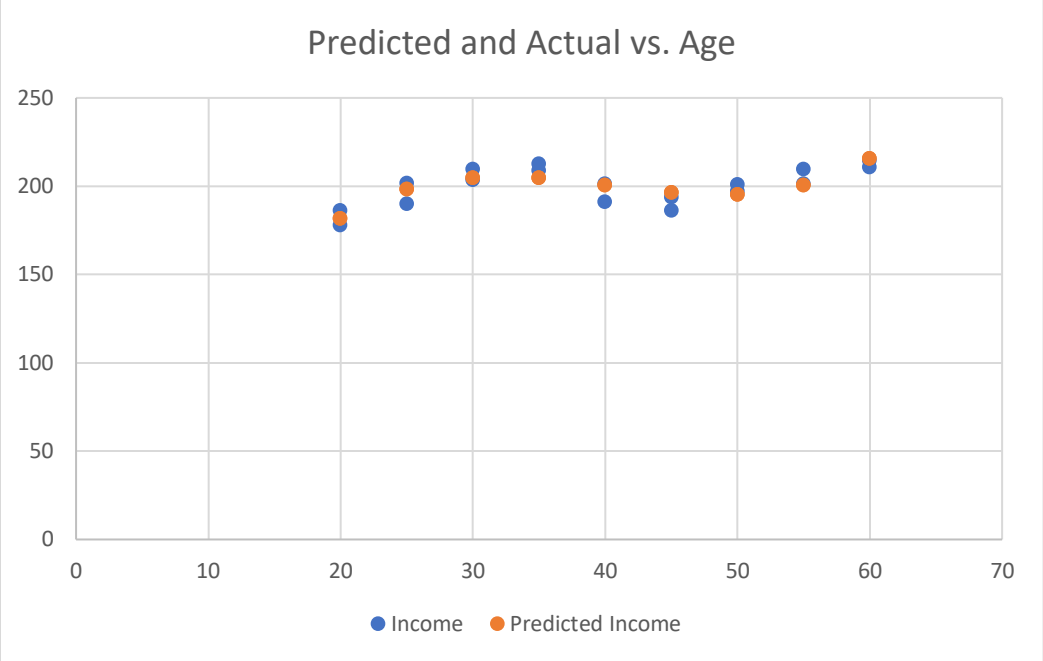
2d-3:


Histogram
Frequency
4
3
2
1
0
-10
-8
-6
-4
-2
0
2
4
6
8
10
12
More
bin
Frequency

3a


Residual t vs t-1
10
5
0
-5
-10
-15
-15
-10
-5
0
5
10

3b:


Predicted and Actual vs. Age
250
200
150
100
50
0
0
10
20
30
40
50
60
70
Income
Predicted Income