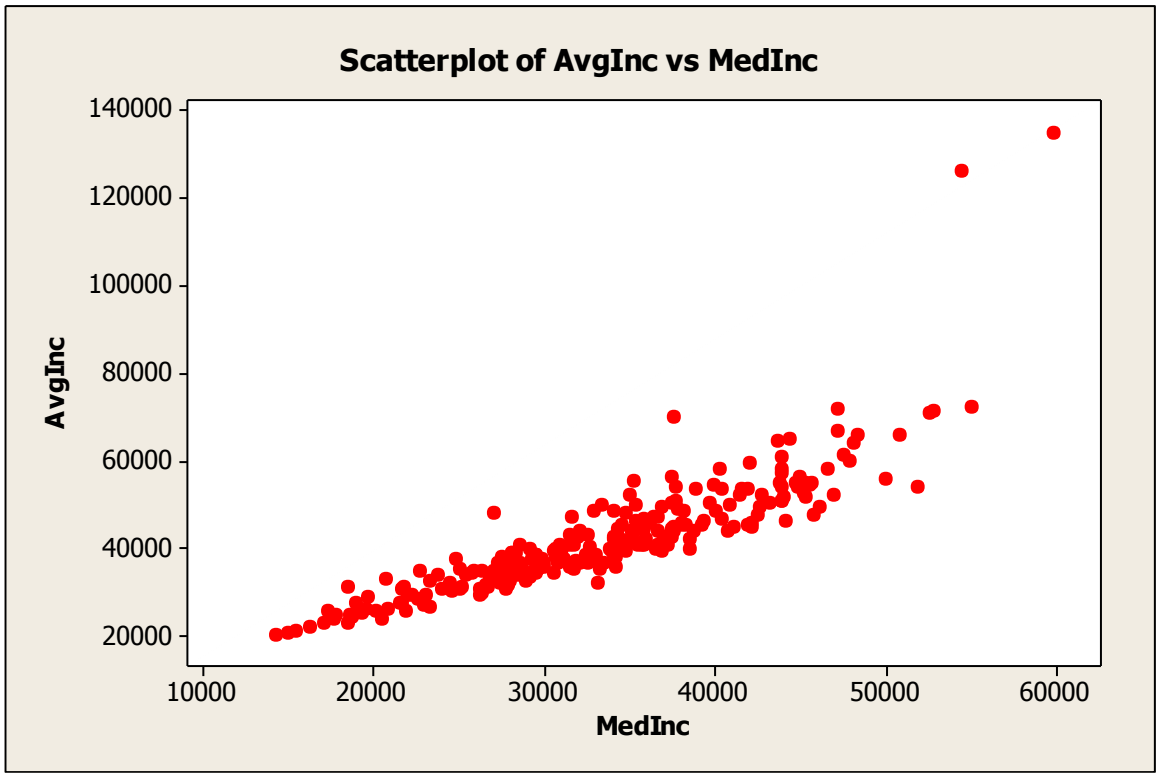
ADM 2304
Assignment 4 Solutions

Question 1. [35 marks]

a. [2 marks]



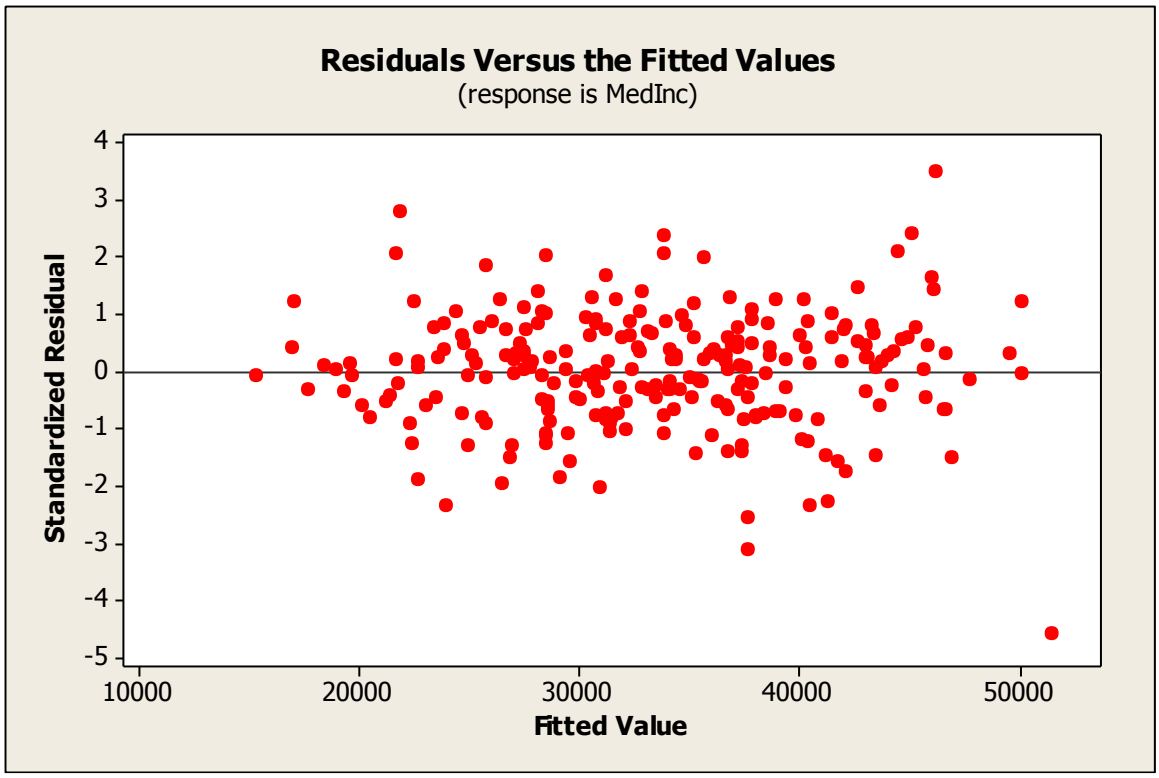
This graph shows how the average income can be so much larger than the median income. It suggests that income distributions are skewed.

1 mark for plot,
1 mark for indicating the skewness of the income distribution

b. [4 marks]

```
The regression equation is
MedInc = 14309 + 146517 P_mgmt + 30968 P_busfin + 34268 P_science
         + 123464 P_health + 63450 P_educgovt - 58923 P_cultsport
         - 47807 P_sales + 23221 P_trades - 19712 P_primaryind
         - 88377 P_manufact


S = 4143.68   R-Sq = 77.5%   R-Sq(adj) = 76.5%
```



1 mark for showing the regression equation and at least two summary statistics
1 mark for standardized residuals plotted against the fitted values
1 for comment that the extreme outliers suggest that the errors may not be normally distributed
1 for comment that the constant variance assumption is warranted given the relative constant vertical spread of the residuals

c. [1 mark]

The dropped observations are
Obs'n 24, Ctract 11.04
Obs'n 70, Ctract 54.00
Obs'n 85, Ctractd 110.00

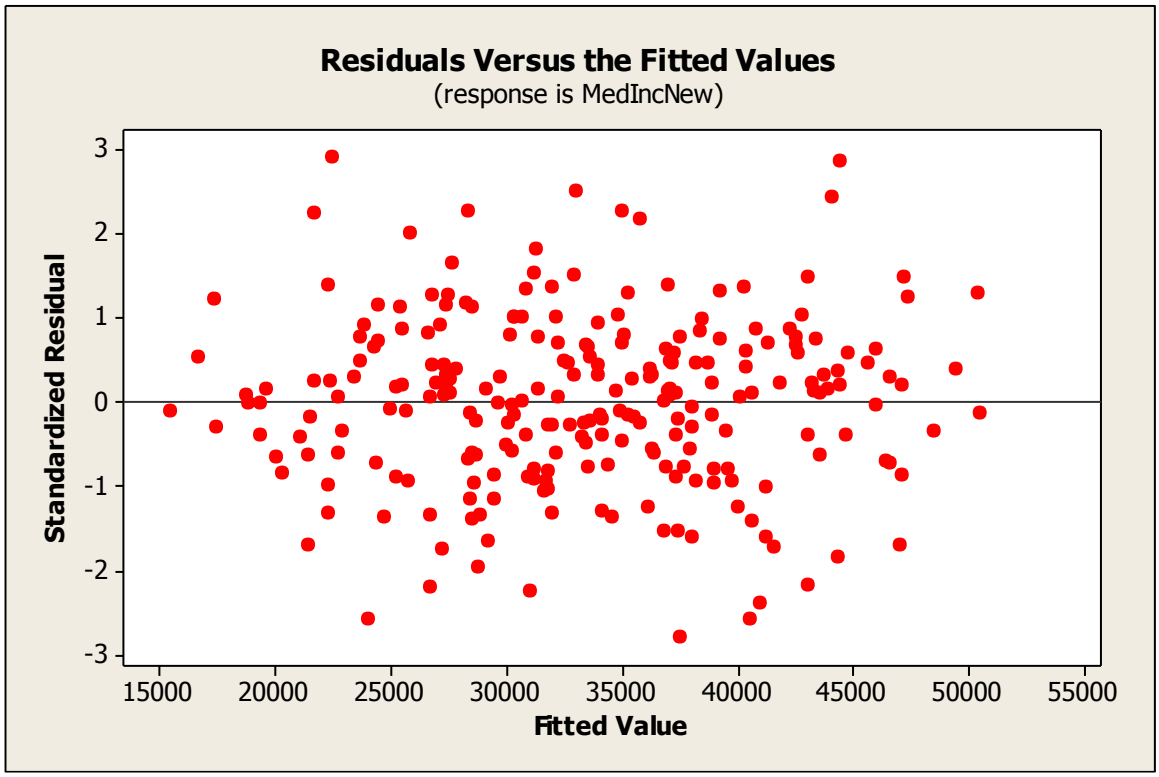
They were dropped because they are outliers (residual exceeded 3 in standard error units).

1 mark for identification and reason

d. [3 marks]

```
The regression equation is
MedIncNew = 13038 + 159862 P_mgmt + 30398 P_busfin + 34193 P_science
            + 108761 P_health + 59779 P_educgovt - 44271 P_cultsport
            - 43506 P_sales + 26940 P_trades - 33670 P_primaryind
            - 84643 P_manufact
```

```
S = 3782.55    R-Sq = 80.7%    R-Sq(adj) = 79.8%
```



1 mark for new regression equation
1 mark for new residual plot
1 mark for comment that the outlying residuals are gone and there are no particular problems with the residuals. (The number of outlying residuals is no more than 5%.)

e. [3 marks]

Ho: beta1=0,…beta10=0; Ha: at least one beta(j) is nonzero
F = 1.4 billion / 14 m = 98, with p-value = 0.000
Reject null H, conclude the model is useful

1 mark for hypotheses
1 mark for F-stat showing the ratio of MSE to MSE
1 mark for decision and conclusion

f. [3 marks]

```
Pearson correlation of MedInc and FITS1 = 0.879
```
Note that $.879^2$ = R-square = .807

1.5 mark for value of correlation coefficient
1.5 mark for noting relation to R-square

g. [1 mark]
There are no problems with multicollinearity since the VIF values are all small.

1 mark

h. [5 marks]

Ho: beta(mgmt) = 0; Ha: beta(mgmt) ≠ 0

T = 13.54 with p-value = 0.000

Reject null H and conclude the P_mgmt variable is useful, *given the other variables in the model.*

If the proportion increases by +0.01, the average increase in median incomes would be

($1599 +/- 2.58 * $118), (assuming the other variables remained constant)

1 mark for hypothesis
1 mark for t-statistic and p-value or critical value
1 mark for decision and conclusion
1 mark for CI above
1 mark for any comment about the effect of more people in management occupations.

i. [2 marks]

```
The regression equation is
MedIncNew = 15385 + 239826 P_mgmt

S = 4619.63   R-Sq = 70.1%   R-Sq(adj) = 69.9%

In the simple regression model, the coefficient is 239826, with std error of
10039.
In the multiple regression model, the coefficient is 159862, with std error of
11809.
```

They are different because of P_mgmt is somewhat correlated with a linear combination of the other variables.

1 mark for comparing the coefficients
1 mark for explanation

j. [4 marks]

```
Predicted Values for New Observations

New
Obs    Fit   SE Fit      99% CI          99% PI
  1  45012    1476  (41180, 48845)  (34468, 55556)X
```

```
X denotes a point that is an outlier in the predictors.


Values of Predictors for New Observations

New
Obs  P_mgmt  P_busfin  P_science  P_health  P_educgovt  P_cultsport  P_sales
  1  0.0881     0.113     0.0535    0.0755       0.148       0.0314   0.0597

New
Obs  P_trades  P_primaryind  P_manufact
  1  0.000000        0.0126    0.000000
```

The 99% PI is ($34468, $55556).

The standard error for the PI is sqrt(1476^2 + 3782.55^2) = $4060.

It does not contain the value $59831 because that value is an outlier.

1 mark for PI
1 mark for showing manual calculation of the standard error of 4060
1 mark for commenting the PI does not cover the value $59831.
1 mark for commenting that the value 59831 was an outlier in an earlier regression.

k. [4 marks]

Variable to consider dropping is P_primaryind because the coefficient is not statistically significantly nonzero.

Without this variable, the regression model is:

```
The regression equation is
MedIncNew = 12654 + 157591 P_mgmt + 32447 P_busfin + 32837 P_science
            + 111089 P_health + 61398 P_educgovt - 45435 P_cultsport
            - 41522 P_sales + 20748 P_trades - 84615 P_manufact


S = 3785.84   R-Sq = 80.5%   R-Sq(adj) = 79.8%
```

The original model had summary statistics

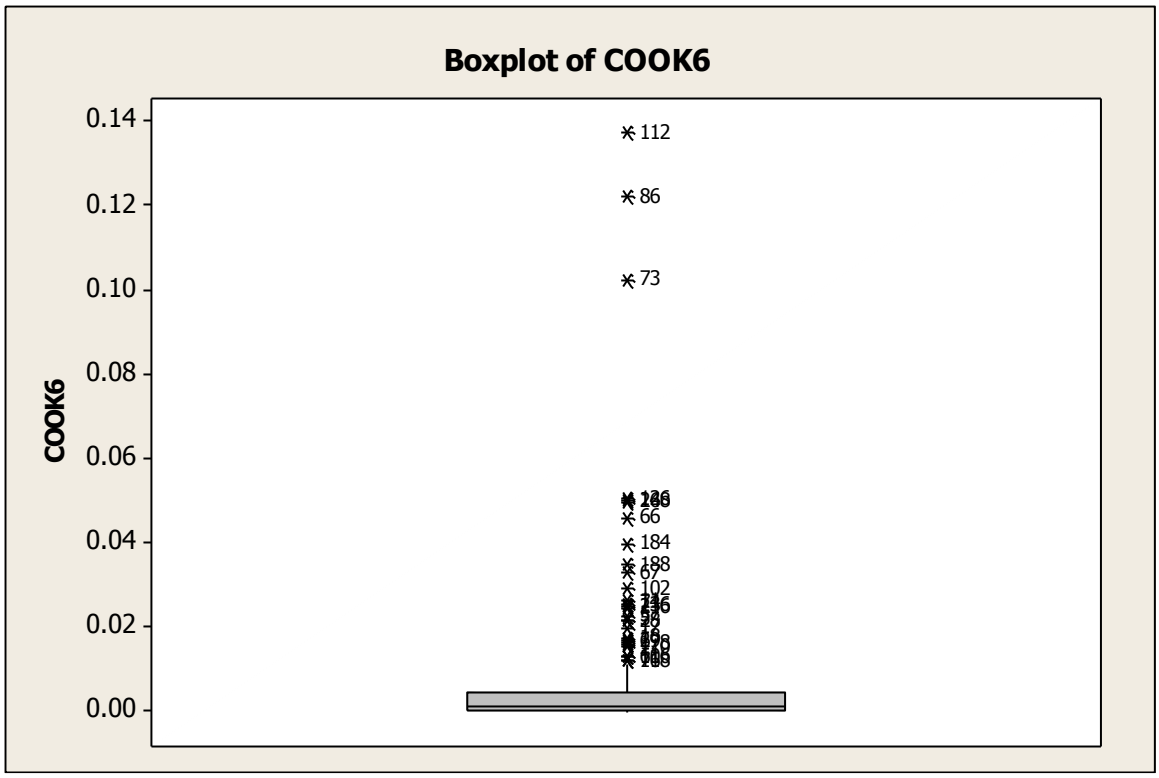
S = 3782.55 R-Sq = 80.7% R-Sq(adj) = 79.8%

The reduced model does not have better summary statistics, except the standard error is only slightly larger. The only advantage is that the model is smaller and all the coefficients are statistically significant.

1 mark for suggesting a variable to drop
1 mark for summary statistics for a reduced model
1 mark for comparing the summary statistics with the full model
1 mark for commenting the model may not have better summary statistics

l. [3 marks]

Boxplot with outliers labelled by observation number.



The most influential observations are 112, 86, 73. These are the unusual observations that are marked as having extreme residual values and as having potentially large influence.

The original coefficients are listed below:

Predictor	Coef	SE Coef	T	P	VIF
Constant	13038	2419	5.39	0.000	
P_mgmt	159862	11809	13.54	0.000	2.1
P_busfin	30398	8692	3.50	0.001	1.2
P_science	34193	9004	3.80	0.000	1.8
P_health	108761	24352	4.47	0.000	1.3
P_educgovt	59779	12146	4.92	0.000	2.6
P_cultsport	-44271	23098	-1.92	0.056	2.5
P_sales	-43506	8411	-5.17	0.000	1.6
P_trades	26940	11093	2.43	0.016	2.7
P_primaryind	-33670	28340	-1.19	0.236	1.6
P_manufact	-84643	41631	-2.03	0.043	1.4

Without the three unusual case, the new coefficients are:

Predictor	Coef	SE Coef	T	P	VIF
Constant	13642	2370	5.76	0.000	
P_mgmt	156082	11413	13.68	0.000	2.1

P_busfin	36477	8464	4.31	0.000	1.2
P_science	38228	8781	4.35	0.000	1.8
P_health	89519	24375	3.67	0.000	1.3
P_educgovt	63023	11783	5.35	0.000	2.6
P_cultsport	-58446	23645	-2.47	0.014	2.5
P_sales	-50021	8734	-5.73	0.000	1.6
P_trades	28485	10803	2.64	0.009	2.7
P_primaryind	-33631	27399	-1.23	0.221	1.6
P_manufact	-73739	40443	-1.82	0.070	1.4

The coefficients have all changed; the coefficient of the P_manufact variable is most affected (the p-value is now .043, which has changed from .070).

1 mark for graph showing labelled outliers
1 mark for identifying these observations as unusual observations as per comment above.
1 mark for comparing the coefficients between the two models