**Assignment1 : Solutions**

**SDGB 7840: EXPLANATORY MODELS**

**Minxia Ji**

**Question 1**

Population:All stores in a large chain of nationwide sporting goods stores

Sample:a random sample of 38 stores in a large chain of nationwide sporting goods stores.

Sample size:38

**Question 2**

Response variable:sales

Explanatory variables:age, growth, income, HS, college

|  |  |  |  |
| --- | --- | --- | --- |
| Histogram | Identify | type | discussion |
| 21 | Sales($)  Latest one-month sales total (dollars) | continuous | Skewness: right-skewed  Outlier: no outlier |
| 22 | Age(years)  Median age of customer base (years) | continuous | Skewness: almost normal  Outlier: no outlier |
| 23 | Growth(%)  Annual population growth rate of customer base over the past 10 years | continuous | Skewness: left-skewed  Outlier: an outlier with growth rate higher than 20% |
| 24 | Income($)  Median family income of customer base (dollars) | continuous | Skewness: almost normal  Outlier:no outlier |
| 25 | HS(%)  Percentage of customer base with a high school diploma | continuous | Skewness: left-skewed  Outlier: no outlier |
| 26 | College(%)  Percentage of customer base with a college diploma | continuous | Skewness: right-skewed  Outlier: no outlier |

Summary statistics

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Sales($) | Age(years) | Growth(%) | Income($) | HS(%) | College(%) |
| mean | 1638487 | 32.45 | 1.60 | 34174.6 | 77.24 | 23.67 |
| median | 1507047 | 32.67 | 0.76 | 3187.63 | 78.34 | 20.82 |
| Standard  deviation | 907758 | 1.72 | 3.80 | 8890.97 | 7.62 | 8.15 |
| minimum | 529215 | 28.52 | -1.04 | 18730.88 | 58.36 | 10.67 |
| maximum | 3860007 | 36.11 | 23.46 | 53063.79 | 93.50 | 41.71 |

**Question 3**

Correlations,r:

Sales vs age:0.041

Sales vs growth:0.112

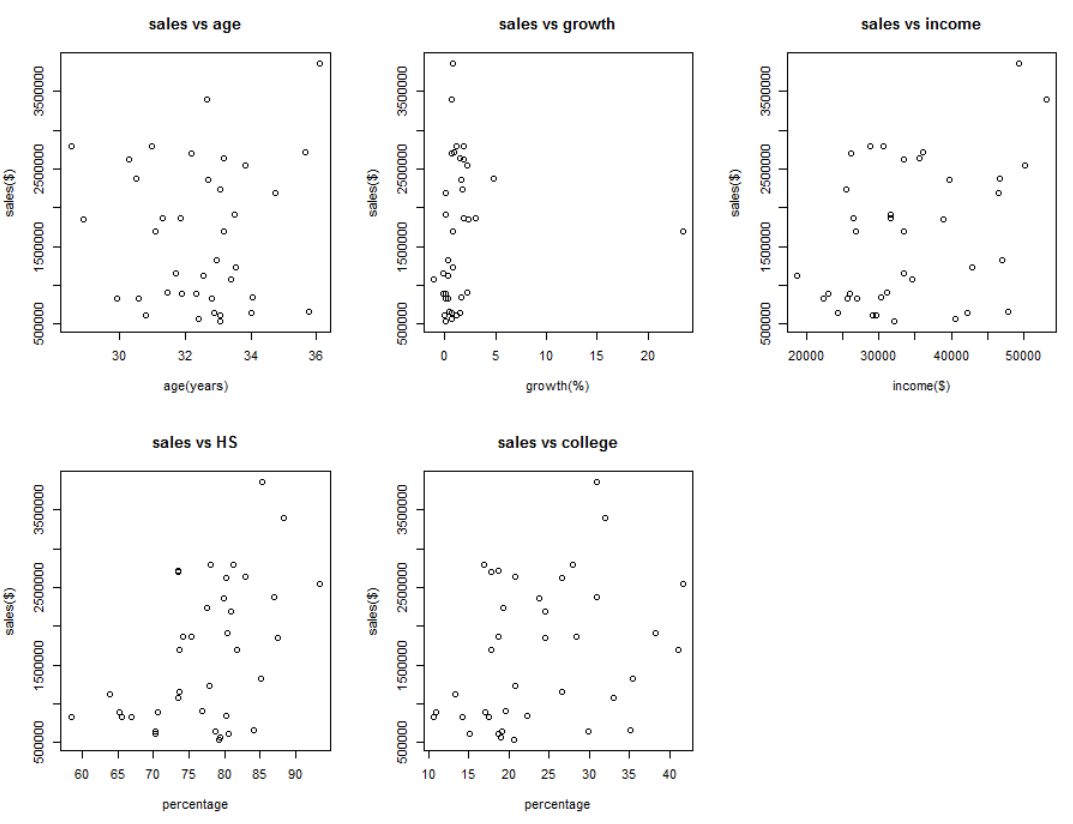
Sales vs income:0.384

Sales vs HS:0.490

Sales vs college:0.322

Correlation is an appropriate measure to describe relationships between sales vs HS.Since from the scatter plots we can see that the relationship between explanatory variables and response variable Sales can fairly be described by a line, the correlation is meaningful.

Correlation is not an appropriate measure to describe relationships between sales vs age,sales vs income, sales vs college and sales vs growth.Since from the scatter plots we can see that the relationship between explanatory variables and response variable cannot be described by a line, the correlation is meaningless.

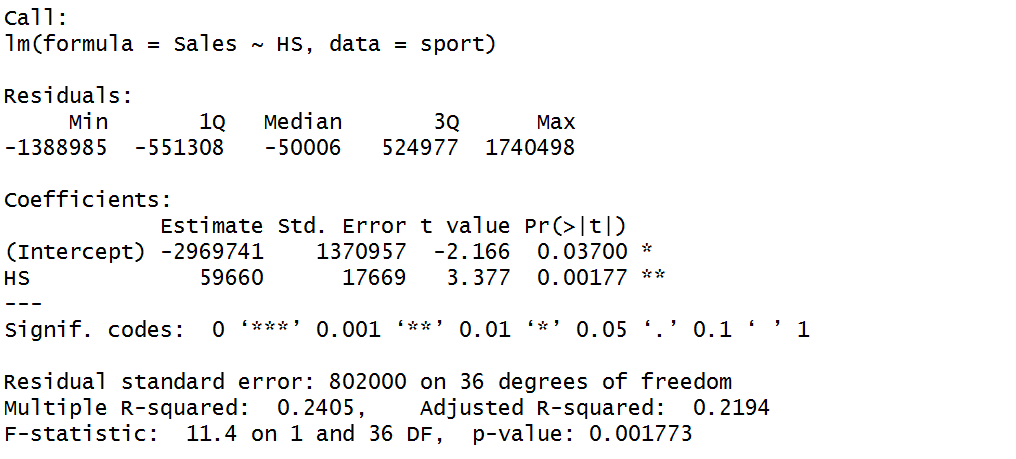


**Question 4**

Based on my analysis so far, the best explanatory variable should be HS,Percentage of customer base with a high school diploma.

Because the correlation coefficient r of sales and HS is the highest among others.r=0.49 shows that there is moderately correlation between sales and HS.

**Question 5**

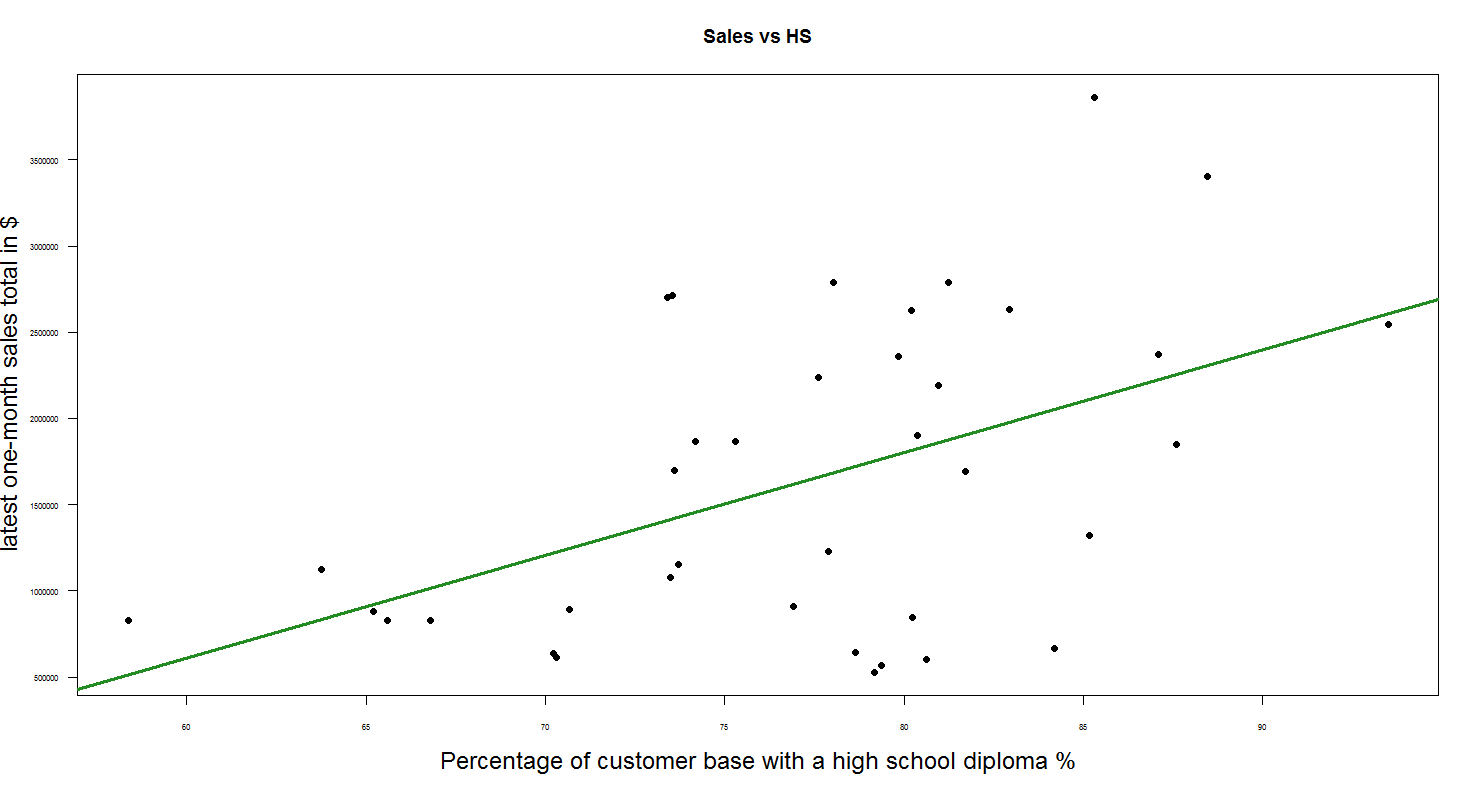


Model:

Sales($) = -2969741($) + 59660($/%)\*HS(%)

**Question 6**

The fit of my model is fairly good.It is a positive slope and the trend goes with the trend of the data.



**Question 7**

The slope of the regression line means that an additional one percent increase in customer base with a high school diploma is associated with a $59660 increase in the latest one-month sales total.

Over the range of 58.36% to 93.50% of HS is the interpretation meaningful.

**Question 8**

The y-intercept of the regression line means that when the percentage of customer base with a high school diploma is 0 % , we expect that the latest one-month sales total to be $-2969741.

It has no practical meaning for this application. Since 0 is not within the range of observed percentage of customer base with a high school diploma and there is no contextual interpretation for percentage of customer base with a high school diploma being 0 % .

**Question 9**

R square=0.2405

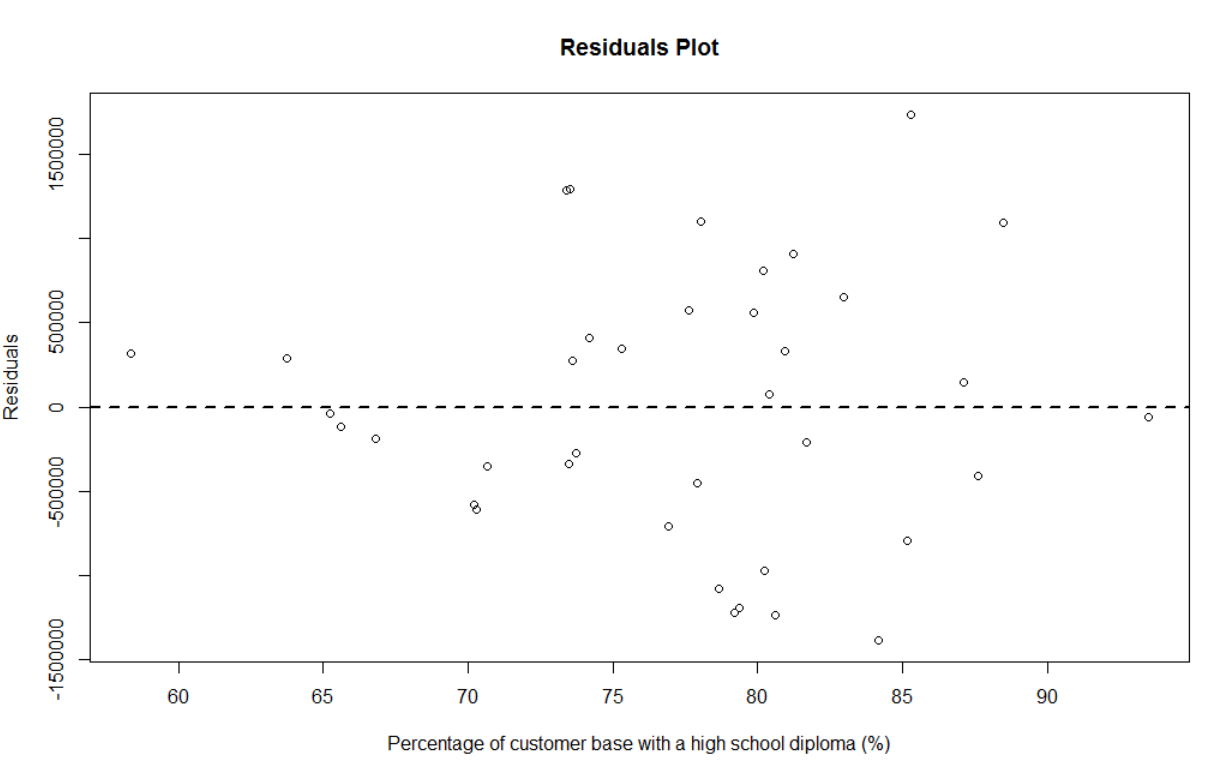
≈ 24% of the variability in the sales in latest one-month in $ can be explained by the percentage in customer base with a high school diploma % using the regression model.

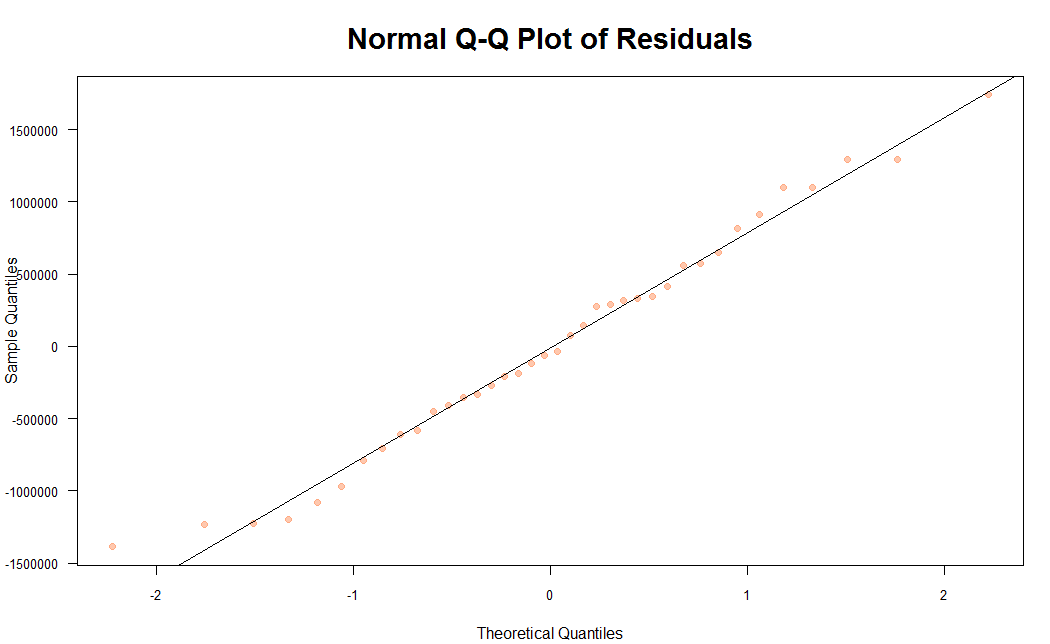
**Question 10**

Assumptions Check:

1. Measurement accuracy: Suppose no measurement error exist in data, assumption satisfied.
2. Linearity: No obvious trend or curve, assumption satisfied.
3. Constant variance: The variance of residuals seems get a bit of larger when x increasing, assumption is slightly violated.
4. Normality: The points almost form a straight line in the normal quantile plot of the residuals, assumption satisfied.
5. Independence: No time series data, assumption satisfied.

It is adequate for building the regression model.





**Question 11**

RMSE:$802003.8

If assumptions are satisfied, given the value of the percentage of customer base with a high school diploma, the sales has a standard deviation of ≈$802003.8.

It is high.Based on our range of y values, the RMSE is quite high.Since the R square is low, it represents that a lot of variability not explained by the model, therefore comes the high RMSE.

**Question 12**

Nrow(sport)=38

Let alpha=0.05

Hypothesis:

H0 : β1 = 0

H1 : β1 ≠ 0

T = (β1-β\*)/sβ1=3.38

Degree of freedom: n - 2 = 38 - 2 = 36

p-value = 2P(T >3.38) = 0.00177<α

Reject null hypothesis; slope is significantly different from 0.There is linear relationship between explanatory variable and response variable.

**Question 13**

Predict sales in A:$13263.4

Predict sales in B:$1504765.7

Predict sales in C:$1206465.2

The predictions in sales in B and C are interpolation while in A is an extrapolation.Because in the data set, for HS, maximum=93.49%,minimum=58.36%.For county A, the HS is 50%, beyond our range.

**Question 14**

It is better to compute a 95% prediction interval to learn some information about the quality of the predictions for the three counties. The prediction interval is used for examining value for a new or single y given x variable values, which is the case of the predictions for the three individual countries.

Prediction intervals for all three counties

|  |  |  |  |
| --- | --- | --- | --- |
|  | fit | Lower | Upper |
| County A | 13263.4 $ | -1901981.7 $ | 1928508 $ |
| County B | 1504765.7 $ | -144992.6 $ | 3154524 $ |
| County C | 1206465.2 $ | -461643.2 $ | 2874574 $ |

**Question 15**

I would recommend company to open a new store in County B.Because the model predicts a new store in County B would generate the highest sales revenues and it has best 95% prediction interval among all three counties ( largest lower and upper intervals ).