Suppose that the sales manager of a large automotive parts distributor wants to estimate as early as April the total annual sales of a region. On the basis of regional sales, the total sales for the company can also be estimated. If, based on past experience, it is found that April estimates of annual sales are reasonably accurate, then in future years the April fore cast could be used to revise production schedules and maintain the correct inventory at the retail outlets.

Several factors appear to be related to sales, including the number of retail outlets in the region stocking the company’s parts, the number of automobiles in the region registered as of April 1, and total personal income for the first quarter of the year. Five independent variables were finally selected as being the most important (according to the sales manager). Then the data were gathered for a recent year. The total annual sales for that year for each region were also recorded. Note in the following table that for region 1 there were 1,739 retail outlets stocking the company’s automotive parts, there were 9,270,000 registered automobiles in the region as of April 1 and so on. The sales for that year were $37,702,000.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Annual Sales ($millions),  Y | Number of Retail Outlets,  X1 | Number of Automobiles Registered (millions), X2 | Personal Income ($ billions),  X3 | Average Age of Automobiles (years),  X4 | Number of Supervisors,  X5 |
| 37.702 | 1,793 | 9.27 | 85.4 | 3.5 | 9.0 |
| 24.196 | 1,221 | 5.86 | 60.7 | 5.0 | 5.0 |
| 32.055 | 1,846 | 8.81 | 68.1 | 4.4 | 7.0 |
| 3.611 | 120 | 3.81 | 20.2 | 4.0 | 5.0 |
| 17.625 | 1,096 | 10.31 | 33.8 | 3.5 | 7.0 |
| 45.919 | 2,290 | 11.62 | 95.1 | 4.1 | 13.0 |
| 29.600 | 1,687 | 8.96 | 69.3 | 4.1 | 15.0 |
| 8.114 | 241 | 6.28 | 16.3 | 5.9 | 11.0 |
| 20.116 | 649 | 7.77 | 34.9 | 5.5 | 16.0 |
| 12.994 | 1,427 | 10.92 | 15.1 | 4.1 | 10.0 |

1. Consider the following correlation matrix. Which single variable has the strongest correlation with the dependent variable? The correlations between the independent variables outlets and income and between cars and outlets are fairly strong. Could this be a problem? What is this condition called?

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Sales | Outlets | Cars | Income | Age |
| Outlets | 0.899 |  |  |  |  |
| Cars | 0.605 | 0.775 |  |  |  |
| Income | 0.964 | 0.825 | 0.409 |  |  |
| Age | -0.323 | -0.489 | -0.447 | -0.349 |  |
| Bosses | 0.286 | 0.183 | 0.395 | 0.155 | 0.291 |

1. The output for all five variables is on the following page. What percent of the variation is explained by the regression equation?

The Regression equation is:

Sales= -19.7 - 0.00063 outlets + 1.74 cars + 0.410 income + 2.04 age - 0.034 bosses

|  |  |  |  |
| --- | --- | --- | --- |
| Predictor | Coef | StDev | t-ratio |
| Constant | -19.672 | 5.422 | -3.63 |
| Outlets | -0.000629 | 0.002638 | -0.24 |
| Cars | 1.7399 | 0.5530 | 3.15 |
| Income | 0.40994 | 0.04385 | 9.35 |
| Age | 2.0357 | 0.8779 | 2.32 |
| Bosses | -0.0344 | 0.1880 | -0.18 |

Analysis of Variance

|  |  |  |  |
| --- | --- | --- | --- |
| Source | DF | SS | MS |
| Regression | 5 | 1593.81 | 318.76 |
| Error | 4 | 9.08 | 2.27 |
| Total | 9 | 1602.89 |  |

1. Conduct a global test of hypothesis to determine whether any of the regression coefficients are not zero. Use the .05 significance level.
2. Conduct a test of hypothesis on each of the independent variables. Would you consider eliminating “outlets” and “bosses”? Use the .05 significance level.
3. The regression has been rerun below with “outlets” and “bosses” eliminated. Compute the coefficient of determination. How much has R^2 changed from the previous analysis?

The Regression equation is:

Sales= -18.9 + 1.61 cars +0.400 income +1.96 age

|  |  |  |  |
| --- | --- | --- | --- |
| Predictor | Coef | StDev | t-ratio |
| Constant | -18.924 | 3.636 | -5.20 |
| Cars | 1.6129 | 0.1979 | 8.15 |
| Income | 0.40031 | 0.01569 | 25.52 |
| Age | 1.9637 | 0.5846 | 3.36 |

Analysis of Variance

|  |  |  |  |
| --- | --- | --- | --- |
| Source | DF | SS | MS |
| Regression | 3 | 1593.66 | 531.22 |
| Error | 6 | 9.23 | 1.54 |
| Total | 9 | 1602.89 |  |

1. Following is a histogram of the residuals. Does the normality assumption appear reasonable?

Histogram of residual N=10 Stem-and-leaf of residual N=10

Leaf Unit=0.10

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Midpoint Count | | | | | | | |
| -1.5 | 1 | \* |  |  | 1 | -1 | 7 |
| -1.0 | 1 | \* |  |  | 2 | -1 | 2 |
| -0.5 | 2 | \*\* |  |  | 2 | -0 |  |
| -0.0 | 2 | \*\* |  |  | 5 | -0 | 440 |
| 0.5 | 2 | \*\* |  |  | 5 | 0 | 24 |
| 1.0 | 1 | \* |  |  | 3 | 0 | 68 |
| 1.5 | 1 | \* |  |  | 1 | 1 |  |
|  |  |  |  |  | 1 | 1 | 7 |

1. Following is a plot of the fitted values of Y (i.e., Y [^ is over the Y]) and the residuals. Do you see any violations of the assumptions?

