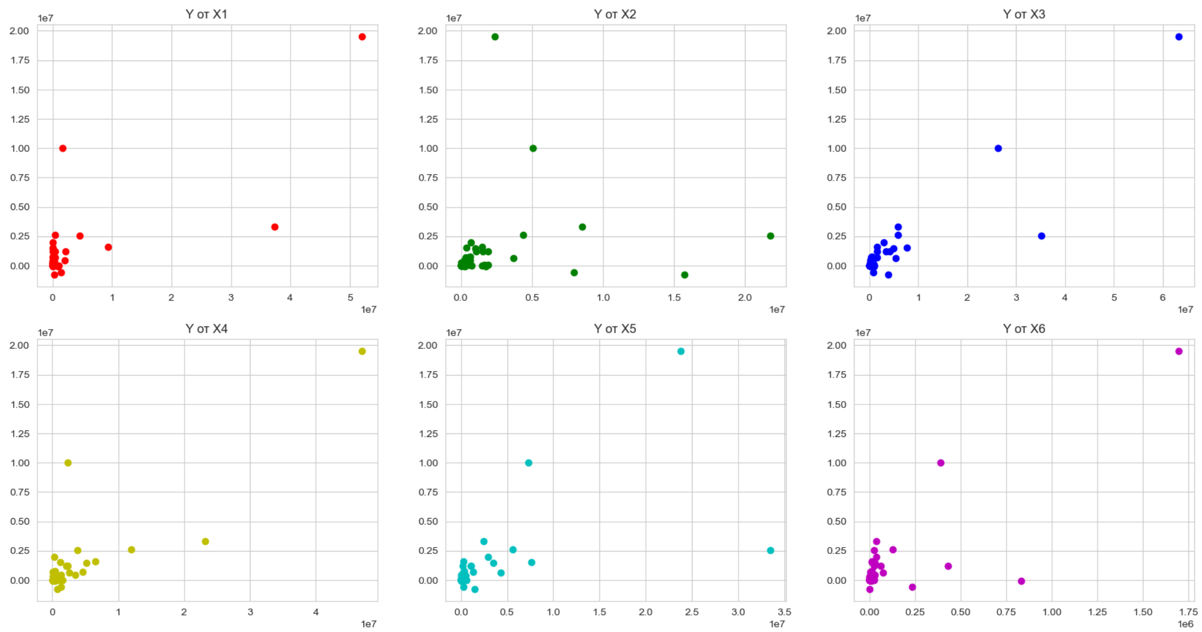
ДЗ №2 Множественная регрессия

Матусков Никита ПМ21-1

Анализ исходных данных

Представим графически исходные данные



Из графической интерпретации видно, что в данных присутствуют выбросы, которые значительно отличаются от других значений в выборке. Выбросы могут искажать статистические показатели и могут оказывать влияние на результаты статистических анализов. Поэтому, их необходимо заменить на медианное значение.

Работа с выбросами

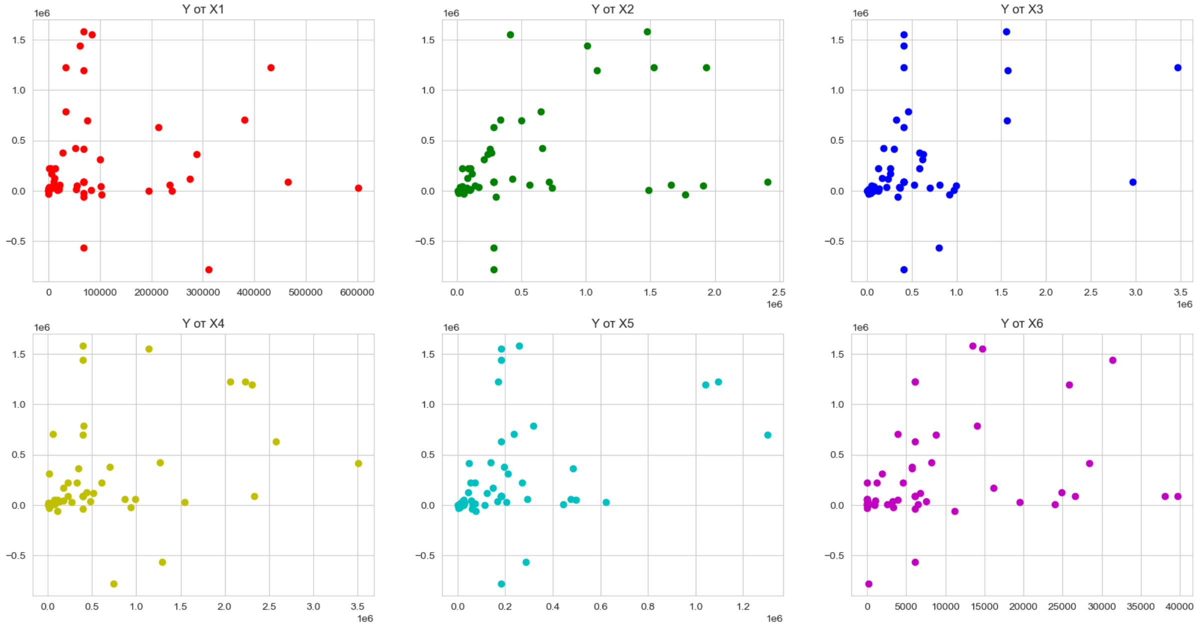
Так как данные не распределены нормально, определение выбросов будет производиться при помощи межквартильного диапазона (IQR). Это разность между значениями верхнего и нижнего квартилей в распределении данных.

Для поиска выбросов при помощи межквартильного диапазона можно использовать следующий алгоритм:

1. Найти значение первого квартиля (Q1) и третьего квартиля (Q3) для данных.
2. Вычислить межквартильный диапазон, вычитая значение Q1 из Q3.
3. Вычислить нижнюю границу выбросов, вычитая 1,5 раза значение IQR из Q1.
4. Вычислить верхнюю границу выбросов, добавляя 1,5 раза значение IQR к Q3.

Любое значение, которое будет меньше нижней границы или больше верхней границы будет являться выбросом. Заменим выбросы на медианы.

Для сравнения построим диаграммы рассеяния прибыли с регрессорами после преобразования

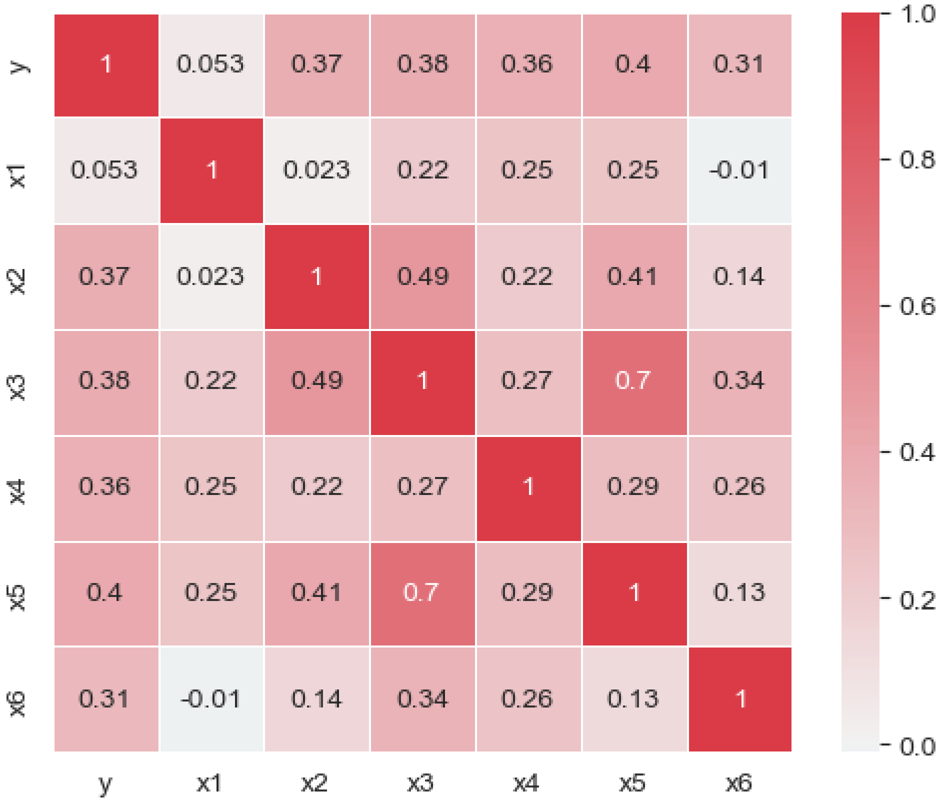


Из графика видно, что аномальных значений, значительно отличаются от других нет

Построение корреляционной матрицы

Для исследования связи между переменными построим корреляционную матрицу, отображающую коэффициенты корреляции между каждыми переменными

$$r = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \sigma_y}$$



Проверка значимости коэффициентов корреляции

Выдвигается гипотеза

$$H_0 : r_{x_i y} = 0$$
$$H_1 : r_{x_i y} \neq 0$$

Для проверки значимости используют t - распределение Стьюдента

$$t_{\text{расч } x_i} = \frac{r_{x_i y}}{\sqrt{1 - r^2_{x_i y}}} \cdot \sqrt{n-2}$$
$$t_{\text{табл}}(0,05;\ n-2) = 2,007$$

Получим значения t-критериев переменных

$$
\begin{aligned}
t_{\text{расч } x_1} &= 0,381 < t_{\text{табл}} \\
t_{\text{расч } x_2} &= 2,905 > t_{\text{табл}} \\
t_{\text{расч } x_3} &= 2,946 > t_{\text{табл}} \\
t_{\text{расч } x_4} &= 2,758 > t_{\text{табл}} \\
t_{\text{расч } x_5} &= 3,142 > t_{\text{табл}} \\
t_{\text{расч } x_6} &= 2,364 > t_{\text{табл}}
\end{aligned}
$$

Выводы

Корреляция между y и x1 не является статистически значимой

Корреляция между y и x2, x3, x4, x5, x6 является статистически значимой с вероятностью 0,95. Связь прямая слабая

Построение модели множественной линейной регрессии

Найдем коэффициенты уравнения регрессии при помощи матричного уравнения

$$B = (X^T X)^{-1} X^T Y$$

$$\hat{y} = -50578.67 - 0.22x_1 + 0.16x_2 - 0.01x_3 + 0.13x_4 + 0.45x_5 + 9.17x_6$$

Проверка значимости коэффициентов регрессии

Выдвигается гипотеза

$$
\begin{aligned}
H_0 &: b_j = 0 \\
H_1 &: b_j \neq 0
\end{aligned}
$$

Для проверки значимости используют t - распределение Стьюдента

$$
\begin{gathered}
t_{\text{расч}} = \frac{b_j}{S_{b_j}} \\
t_{\text{табл}}(0,05;\ n-2) = 2,007
\end{gathered}
$$

Стандартная ошибка коэффициента регрессии

$$
\begin{gathered}
S_{b_j} = S\sqrt{z_{jj}} \\
\text{где } z_{jj} - \text{диагональный элемент матрицы } (X^T \cdot X)^{-1}
\end{gathered}
$$

Стандартная ошибка отклонения

$$S = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n - m - 1}}$$

Получим значения t-критериев коэффициентов

$$t_{\text{расч } b_0} = -0,523$$
$$t_{\text{расч } b_1} = -0,475$$
$$t_{\text{расч } b_2} = 1,44$$
$$t_{\text{расч } b_3} = -0,079$$
$$t_{\text{расч } b_4} = 1,547$$
$$t_{\text{расч } b_5} = 1,446$$
$$t_{\text{расч } b_6} = 1,478$$

Определение значимых факторов

Будем использовать метод обратного пошагового отбора и последовательно исключать из модели незначимые переменные

Из полученных значений видно, что x_3 почти не влияет на y и его значение наименьшее и меньше $t_{\text{табл}}$, поэтому его можно исключить из модели

Пересчитаем t-критерии коэффициентов

$$t_{\text{расч } b_0} = -0,523$$
$$t_{\text{расч } b_1} = -0,493$$
$$t_{\text{расч } b_2} = 1,507$$
$$t_{\text{расч } b_4} = 1,567$$
$$t_{\text{расч } b_5} = 1,793$$
$$t_{\text{расч } b_6} = 1,563$$

Из полученных значений видно, что x_1 почти не влияет на y и его значение наименьшее и меньше $t_{\text{табл}}$, поэтому его можно исключить из модели

Пересчитаем t-критерии коэффициентов

$$t_{\text{расч } b_0} = -0,738$$
$$t_{\text{расч } b_2} = 1,581$$
$$t_{\text{расч } b_4} = 1,508$$
$$t_{\text{расч } b_5} = 1,74$$
$$t_{\text{расч } b_6} = 1,623$$

Исключим из модели x_4, так как его $t_{\text{расч}}$ меньше $t_{\text{табл}}$ и пересчитаем t-критерии

$$t_{\text{расч } b_0} = -0,385$$
$$t_{\text{расч } b_2} = 1,711$$
$$t_{\text{расч } b_5} = 2,078$$
$$t_{\text{расч } b_6} = 1,981$$

Исключим из модели незначимый x_2

$$t_{\text{расч } b_0} = 0,119$$
$$t_{\text{расч } b_5} = 2,938$$
$$t_{\text{расч } b_6} = 2,123$$

Коэффициент b_0 является незначимым, следовательно итоговая модель имеет вид:

$$\hat{y} = 0,655x_5 + 12,431x_6$$

Проверка значимости уравнения регрессии в целом

Выдвигается гипотеза

$$H_0 : \sigma^2_{\text{факт}} = \sigma^2_{\text{остат}}$$
$$H_1 : \sigma^2_{\text{факт}} \neq \sigma^2_{\text{остат}}$$

Для проверки значимости используют F - распределение Фишера

$$F_{\text{расч}} = \frac{r^2}{1-r^2} \cdot \frac{n-m-1}{m}$$
$$F_{\text{табл}}(0,05;\ m;\ n-m-1) = 3,18$$

$$r^2 = \frac{\sum(\hat{y}-\bar{y})^2}{\sum(y-\bar{y})^2}$$

$$F_{\text{расч}} = -33,399$$
$$|F_{\text{расч}}| > |F_{\text{табл}}|$$

Вывод

Уравнение регрессии является статистически значимым с вероятностью 0,95

Заключение

На основе проведенного анализа, наибольшее влияние на прибыль компании имеет количество дебиторских задолженностей и запасы готовой продукции.

При увеличении количества дебиторских задолженностей на 1 ед. прибыль увеличится на 0,655 ед.

При увеличении запасов готовой продукции на 1 ед. прибыль увеличится на 12,431 ед.