Итоговый проект

«Анализ выборочных совокупностей»

по курсу «Математическая статистика»

Выполнила: Грошева Валерия

Содержание:

1. Введение 2

2. Выявление выбросов в наборе данных 3

3. Статистика вывода 4

4. Построение парной и множественной линейной регрессии 12

5. Применимые управленческие решения 16

1. ВВЕДЕНИЕ

В современном мире розничной торговли компаниям необходимо постоянно адаптироваться к меняющимся условиям рынка и предпочтениям потребителей. Одним из ключевых факторов, влияющих на успех продаж, является эффективность маркетинговых кампаний.

Цели исследования:

Цель данного исследования заключается в выявлении и анализе переменных, которые влияют на количество продаж в магазинах торговой сети, а также определении, какая маркетинговая кампания является самой эффективной.

Выдвигаемые гипотезы:

1) Распределение количества продаж в магазинах, где была проведена маркетинговая кампания первого, второго и третьего типа не одинаково.
2) Первая маркетинговая кампания является наиболее эффективной.
3) Размер магазина (большой, средний, маленький) не влияет на количество продаж.
4) Количество продаж магазина зависит от того, сколько лет уже функционирует этот магазин.

Набор данных:

Для проведения данного исследования я использую датасет, содержащий данные касательно трех маркетинговых кампаний, проведенных сетью магазинов. Он содержит следующие столбцы:

MarketID	Уникальный идентификатор рынка
MarketSize	Размер рыночной площади по объемам продаж
Location	Уникальный идентификатор местоположения магазина
AgeOfStore	Возраст магазина в годах
Promotion	Одна из трех протестированных маркетинговых кампаний
Week	Одна из четырех недель, в которую проводилась акции
SalesInThousands	Количество продаж в тысячах

2. ВЫЯВЛЕНИЕ ВЫБРОСОВ В НАБОРЕ ДАННЫХ

Построим ящик с усами по изучаемой переменной SalesInThousands для визуальной оценки наличия выбросов в ней.

```
> sales <- my_dataset$SalesInThousands
> boxplot(sales)
```

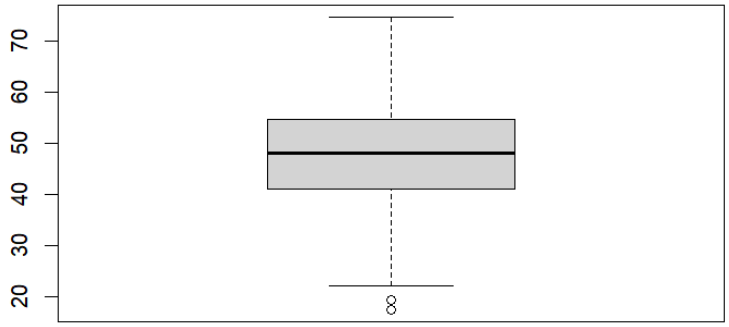

Рис.1. *Ящик с усами*

Можно наглядно увидеть, что среди данных по продажам присутствуют выбросы. Найдем их двумя известными нам методами: методом Тьюки и методом z-оценок.

Идентификация выбросов по методу Тьюки:

Находим $Q_1 - 1{,}5\ IQR$ и $Q_3 + 1{,}5\ IQR$.

```
> lb <- quantile(sales,0.25, na.rm = TRUE) - 1.5*IQR(sales, na.rm = TRUE)
> lb
     25%
20.86125
> rb <- quantile(sales,0.75, na.rm = TRUE) + 1.5*IQR(sales, na.rm = TRUE)
> rb
     75%
74.85125
```

Элементы меньшие lb или большие rb считаются выбросами. Таких элементов в данной переменной два: 19,26 и 17,34.

```
> vybrosy <- subset(sales, sales < lb | sales > rb)
> vybrosy
[1] 19.26 17.34
```

Идентификация выбросов по методу z-оценок:

Для проверки наличия выбросов методом z-оценок, данные должны быть извлечены из выборки с распределением близким к нормальному.

```
> shapiro.test(sales)

        Shapiro-Wilk normality test

data:  sales
W = 0.99619, p-value = 0.332
```

Критерий Шапиро–Уилка показал, что выборка действительно имеет распределение близкое к нормальному, так как p-value > 0,05. Следовательно, мы можем применить данный метод.

```
> vybrosy_z <- subset(sales, abs((sales-mean(sales))/sd(sales))>3)
> vybrosy_z
numeric(0)
```

Метод z-оценок не выявил ни одного выброса.

3. СТАТИСТИКА ВЫВОДА

3.1 Проверка на нормальность распределения

Проверим на нормальность распределения каждую количественную переменную в общей выборке, используя критерии Шапиро–Уилка и Крамера–фон Мизеса.

```
> # Используем критерий Шапиро-Уилка
> shapiro.test(sales)

        Shapiro-Wilk normality test

data:  sales
W = 0.99619, p-value = 0.332

> shapiro.test(my_dataset$AgeOfStore)

        Shapiro-Wilk normality test

data:  my_dataset$AgeOfStore
W = 0.89545, p-value < 2.2e-16

> # Используем критерий Крамера-фон Мизеса
> cvm.test(sales)

        Cramer-von Mises normality test

data:  sales
W = 0.055492, p-value = 0.433

> cvm.test(my_dataset$AgeOfStore)

        Cramer-von Mises normality test

data:  my_dataset$AgeOfStore
W = 2.1751, p-value = 7.37e-10
```

Оба критерия показали, что переменная SalesInThousands имеет нормальное распределение (p-value > 0,05), а распределение переменной AgeOfStore значительно отличается от нормального (p-value < 0,05). В подтверждение гипотезы о нормальности распределения изучаемой переменной sales дополнительно построим гистограмму и график Q-Q Plot.

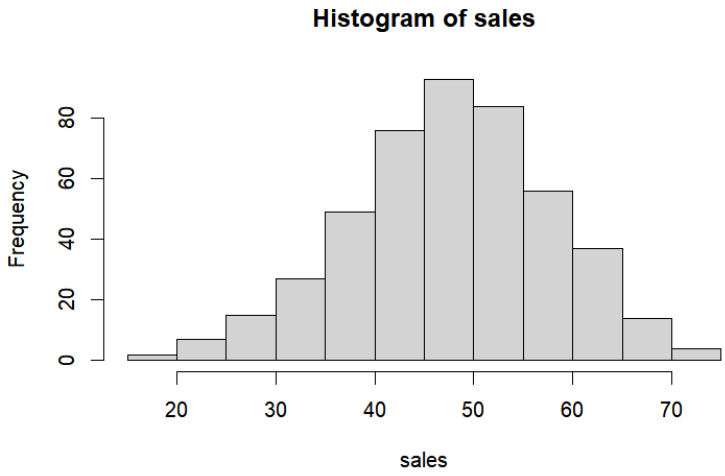

Рис.2. *Гистограмма выборки SalesInThousands*

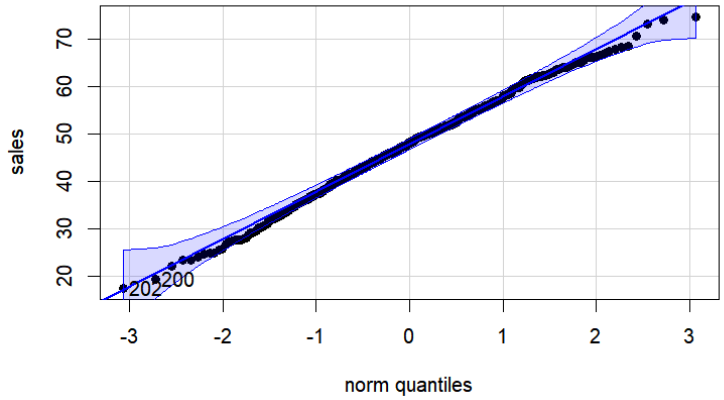

Рис.3. *График Q-Q Plot выборки SalesInThousands*

Гистограмма и график Q-Q Plot подтверждают гипотезу о том, что распределение изучаемой переменной sales несущественно отличается от нормального.

Мы также можем построить гистограмму и график Q-Q Plot для переменной AgeOfStore.

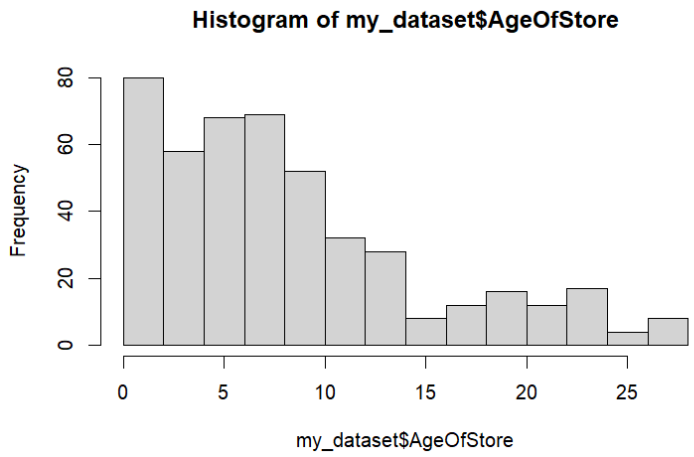

Рис.4. *Гистограмма выборки AgeOfStore*

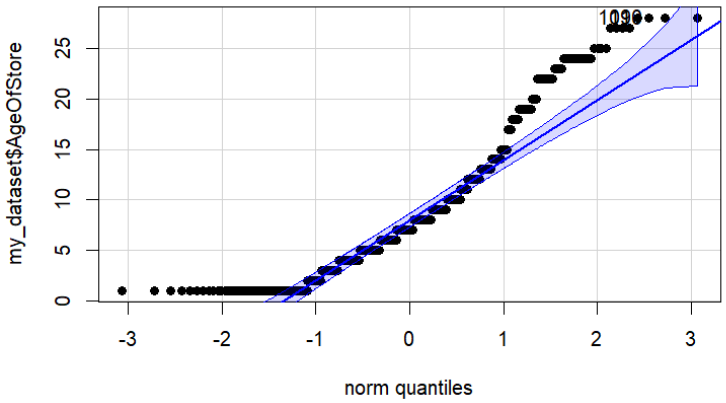

Рис.5. *График Q-Q Plot выборки AgeOfStore*

Здесь мы можем наглядно увидеть, что распределение переменной AgeOfStore значительно отличается от нормального, что и показали критерии Шапиро–Уилка и Крамера–фон Мизеса.

3.2 Оценивание параметров

3.2.1 Точечные оценки параметров распределений.

Рассчитаем некоторые числовые характеристики для количества продаж (переменная SalesInThousands).

- Выборочное среднее: 47,65127;
- Выборочная дисперсия: 104,0667;
- Выборочное среднеквадратическое отклонение: 10,20131;
- Выборочная медиана: 48,15;

- Выборочный коэффициент асимметрии A: -0,1672835 (< 0, следовательно левый хвост распределения «тяжелее»);
- Выборочный коэффициент эксцесса E: -0,171872 (< 0, следовательно хвосты распределения «тяжелее», а пик более «приплюснутый», чем у нормального распределения).

Рассчитаем те же числовые характеристики для возраста магазина (переменная AgeOfStore).

- Выборочное среднее: 8,75431;
- Выборочная дисперсия: 44,82504;
- Выборочное среднеквадратическое отклонение: 6,69515;
- Выборочная медиана: 7;
- Выборочный коэффициент асимметрии: 1,019697 (> 0, следовательно правый хвост распределения «тяжелее»);
- Выборочный коэффициент эксцесса E: 0,3403065 (> 0, следовательно хвосты распределения «легче», а пик острее, чем у нормального распределения).

3.2.2 Интервальные оценки неизвестных параметров нормального распределения.

Построим доверительный интервал для количества продаж и возраста магазинов:
С вероятностью 95% среднее количество продаж будет находиться в интервале между 46,7206 и 48,5819.
С вероятностью 95% средний возраст магазинов будет находиться в интервале между 8,143528 и 9,365092.

3.3 Проверка гипотез

3.3.1

Проверим гипотезу о том, что средние значения количества продаж в магазинах большего размера (Large) равны средним значениям количества продаж в магазинах маленького размера (Small), а не больше, как можно было предположить.
Сначала проверим гипотезу о нормальности распределения подвыборок:
H0: распределение нормальное

H1: распределение отлично от нормального

```
> small <- subset(my_dataset, my_dataset$MarketSize=='Small')
> shapiro.test(small$SalesInThousand)

        Shapiro-Wilk normality test

data:  small$SalesInThousand
W = 0.97775, p-value = 0.341

> large <- subset(my_dataset, my_dataset$MarketSize=='Large')
> shapiro.test(large$SalesInThousand)

        Shapiro-Wilk normality test

data:  large$SalesInThousand
W = 0.96749, p-value = 0.03205
```

В обоих случаях p-value > 0.05, следовательно нет оснований отвергнуть H0 о том, что распределение подвыборок нормальное. Значит, мы можем применять параметрические критерии.

Проверим гипотезу о равенстве двух математических ожиданий:

H0: μ(large) = μ(small)

H1: μ(large) > μ(small)

Для выбора одного из двух t-критериев сформулируем и протестируем вспомогательную гипотезу о равенстве двух дисперсий:

H0: var(large) = var(small)

H1: var(large) ≠ var(small)

```
> var.test(small$SalesInThousands, large$SalesInThousands, conf.level = 0.95, alternative = "two.sided")

        F test to compare two variances

data:  small$SalesInThousands and large$SalesInThousands
F = 0.68542, num df = 59, denom df = 83, p-value = 0.1264
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.4299672 1.1140086
sample estimates:
ratio of variances
         0.6854227
```

p-value > 0.05, следовательно не отвергаем H0 о том, что дисперсии несущественно отличаются друг от друга. Для проверки исходной гипотезы о равенстве двух математических ожиданий применим t-test с одинаковыми дисперсиями:

```
> t.test(large$SalesInThousands, small$SalesInThousands, mu=0, alternative = "greater", var.equal = TRUE)

        Two Sample t-test

data:  large$SalesInThousands and small$SalesInThousands
t = -2.188, df = 142, p-value = 0.9848
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -4.853199       Inf
sample estimates:
mean of x mean of y
 54.64667  57.40933
```

p-value > 0.05, следовательно нет оснований отвергнуть H0 о том, что средние значения количества продаж в магазинах большего размера (Large) и маленького размера (Small) равны.

3.3.2

Проверим гипотезу об одинаковом распределении количества продаж в магазинах, где была проведена маркетинговая кампания (столбец Promotion) первого, второго и третьего типа.

H0: распределение количества продаж в магазинах, где была проведена маркетинговая кампания первого, второго и третьего типа одинаково

H1: распределение количества продаж в магазинах, где была проведена маркетинговая кампания первого, второго и третьего типа хотя бы при одном из этих типов маркетинговых кампаний отличается от остальных

Для этого используем непараметрический критерий для нескольких совокупностей Краскела–Уоллиса:

```
> kruskal.test(my_dataset, my_dataset$Promotion)

        Kruskal-Wallis rank sum test

data:  my_dataset
Kruskal-Wallis chi-squared = 2010.6, df = 7, p-value < 2.2e-16
```

p-value < 0.05, следовательно отвергаем гипотезу H0 и принимаем гипотезу H1: распределение количества продаж в магазинах, где была проведена маркетинговая кампания первого, второго и третьего типа хотя бы при одном из этих типов маркетинговых кампаний отличается от остальных.

3.3.3

Проверим гипотезу о том, что средние значения количества продаж в магазинах, где была проведена первая маркетинговая кампания равны средним значениям количества продаж в магазинах, где была проведена вторая маркетинговая кампания, а не больше, как утверждает эксперт.

Сначала проверим гипотезу о нормальности распределения подвыборок:

H0: распределение нормальное

H1: распределение отлично от нормального

```
> promo1 <- subset(my_dataset, my_dataset$Promotion==1)
> shapiro.test(promo1$SalesInThousand)

        Shapiro-Wilk normality test

data:  promo1$SalesInThousand
W = 0.96869, p-value = 0.002215

> promo2 <- subset(my_dataset, my_dataset$Promotion==2)
> shapiro.test(promo2$SalesInThousand)

        Shapiro-Wilk normality test

data:  promo2$SalesInThousand
W = 0.98126, p-value = 0.02343
```

У обеих подвыборок p-value < 0.05, следовательно отвергаем H0 и принимаем альтернативную гипотезу о том что распределение выборок отлично от нормального. Значит, для проверки равенства средних значений будем использовать непараметрические критерии (применение параметрических нежелательно).

Проверим гипотезу о равенстве двух медиан:

H0: median(promo1) = median(promo2)

H1: median(promo1) > median(promo2)

```
> wilcox.test(promo1$SalesInThousands, promo2$SalesInThousands, mu=0, conf.level = 0.95, alternative = "great
er")

        Wilcoxon rank sum test with continuity correction

data:  promo1$SalesInThousands and promo2$SalesInThousands
W = 17753, p-value = 2.168e-13
alternative hypothesis: true location shift is greater than 0
```

p-value < 0.05, следовательно отвергаем гипотезу H0 и принимаем гипотезу H1 о том, что медианные значения количества продаж в магазинах, где была проведена первая маркетинговая кампания больше, чем средние значения количества продаж в магазинах, где была проведена вторая маркетинговая кампания.

3.3.4

Проверим гипотезу о том, что доля магазинов сети, существующих уже более 10 лет, меньше 0,5.

Сначала проверим гипотезу о нормальности распределения возраста магазинов.

H0: распределение нормальное

H1: распределение отлично от нормального

```
> shapiro.test(my_dataset$AgeOfStore)

        Shapiro-Wilk normality test

data:  my_dataset$AgeOfStore
W = 0.89545, p-value < 2.2e-16
```

Так как p-value < 0.05, отвергаем основную гипотезу H0 и принимаем альтернативную гипотезу H1 о том, что распределение возраста магазинов отлично от нормального. Несмотря на это, мы все равно используем параметрический критерий проверки гипотезы о величине генеральной доли. Однако отметим, что полученный результат может быть не совсем достоверным из-за того, что распределение отличается от нормального.

H0: p = 0,5

H1: p < 0,5

```
> prop.test(length(subset(my_dataset$AgeOfStore, my_dataset$AgeOfStore>10)), length(my_dataset$AgeOfStore), p
=0.5, conf.level = 0.95, alternative = "less")

        1-sample proportions test with continuity correction

data:  length(subset(my_dataset$AgeOfStore, my_dataset$AgeOfStore > 10)) out of length(my_dataset$AgeOfStor
e), null probability 0.5
X-squared = 76.985, df = 1, p-value < 2.2e-16
alternative hypothesis: true p is less than 0.5
95 percent confidence interval:
 0.000000 0.332305
sample estimates:
        p
0.2952586
```

p-value < 0.05, следовательно отвергаем H0 и принимаем альтернативную гипотезу о том, что доля магазинов сети, существующих уже более 10 лет, меньше 0,5.

3.3.5

Проверим гипотезу об одинаковом распределении количества продаж в магазинах на разных неделях промокампании.

H0: распределение количества продаж в магазинах на разных неделях промокампании одинаково

H1: распределение уровня количества продаж в магазинах на разных неделях промокампании хотя бы на одной из этих недель отличается от остальных

Для этого используем непараметрический критерий для нескольких совокупностей Краскела–Уоллиса:

```
> kruskal.test(my_dataset, my_dataset$week)

        Kruskal-Wallis rank sum test

data:  my_dataset
Kruskal-Wallis chi-squared = 2010.6, df = 7, p-value < 2.2e-16
```

p-value < 0.05, следовательно отвергаем гипотезу H0 и принимаем гипотезу H1: распределение уровня количества продаж в магазинах на разных неделях промокампании хотя бы на одной из этих недель значительно отличается от остальных.

4. ПОСТРОЕНИЕ ПАРНОЙ И МНОЖЕСТВЕННОЙ ЛИНЕЙНОЙ РЕГРЕССИИ

4.1 Парная линейная регрессия

Вычислив коэффициенты корреляции, я сделала такие выводы:

SalesInThousands и AgeOfStore, Location – прямая линейная зависимость;

SalesInThousands и week, Promotion – обратная линейная зависимость.

Коэффициенты корреляции Пирсона:

- 0.1022755 для SalesInThousands и AgeOfStore
- -0.02801419 для SalesInThousands и week
- 0.2649495 для SalesInThousands и Location
- -0.1284265 для SalesInThousands и Promotion

Далее построим модели линейной регрессии для переменных, связанных линейной зависимостью.

Построим попарную диаграмму рассеяния:

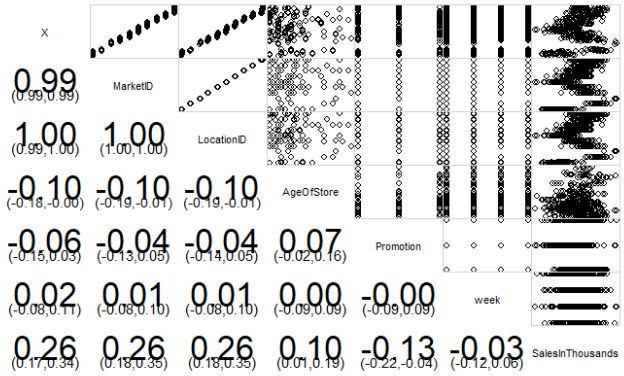

Рис.6. *Попарная диаграмма рассеяния*

Попарная диаграмма рассеяния подтверждает выводы, сделанные исходя из расчетов коэффициентов.

Модель парной линейной регрессии для количества продаж и возраста магазинов (SalesInThousands и AgeOfStore):

Модель зависимости количества продаж от возраста магазинов:

```
> summary(lm(SalesInThousands~AgeOfStore, my_dataset))

Call:
lm(formula = SalesInThousands ~ AgeOfStore, data = my_dataset)

Residuals:
     Min       1Q   Median       3Q      Max
-29.7262  -6.4171   0.4184   6.8879  27.8396

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 46.28704    0.77686   59.58   <2e-16 ***
AgeOfStore   0.15584    0.07052    2.21   0.0276 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.16 on 462 degrees of freedom
Multiple R-squared:  0.01046,   Adjusted R-squared:  0.008318
F-statistic: 4.884 on 1 and 462 DF,  p-value: 0.0276
```

Уравнение парной регрессии: SalesInThousands = 46.2870 + 0.1558*AgeOfStore

Это означает, что при увеличении возраста магазина на 1 год, количество продаж в нем увеличивается на 2.272 тысячи долларов.

Pr(>|t|) < 0.05, следовательно модель и коэффициенты значимы.

Коэффициент детерминации 0.01046, то есть 1,04% изменения количества продаж объясняется тем, как давно магазин уже функционирует.

Отсутствует гомоскедастичность.

Есть автокорреляция.

Все необходимые предпосылки выполняются.

Средняя ошибка аппроксимации = 10.16% >10%, не очень хороший подбор модели, но приемлемый.

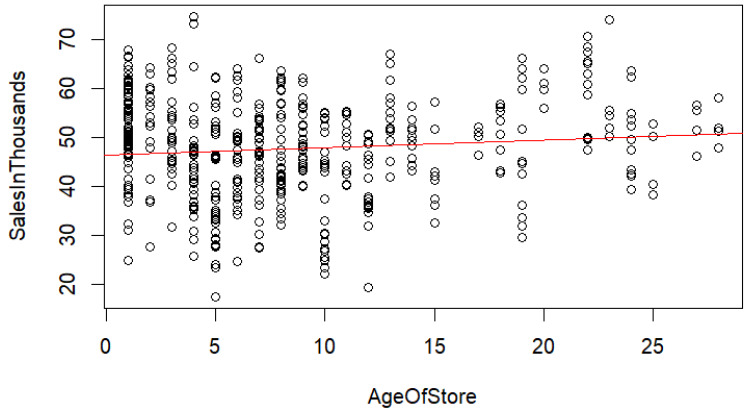

Рис.7. *Модель зависимости количества продаж от возраста магазина*

Модель парной линейной регрессии для количества продаж и недели, на которой проводилась маркетинговая кампания (SalesInThousands и week):

Модель зависимости количества продаж от недели, на которой проводилась маркетинговая кампания:

```
> summary(lm(SalesInThousands~week, my_dataset))

Call:
lm(formula = SalesInThousands ~ week, data = my_dataset)

Residuals:
     Min       1Q   Median       3Q      Max
-30.4367  -6.4477   0.3263   6.8659  26.9733

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  48.2874     1.1575  41.716   <2e-16 ***
week         -0.2553     0.4239  -0.602    0.547
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.21 on 462 degrees of freedom
Multiple R-squared:  0.0007848, Adjusted R-squared:  -0.001378
F-statistic: 0.3629 on 1 and 462 DF,  p-value: 0.5472
```

Уравнение парной регрессии: SalesInThousands = 48.2874 – 0.2553*week

Это означает, что при увеличении возраста магазина на 1 год, количество продаж в нем увеличивается на 2.272 тысячи долларов.

Pr(>|t|) < 0.05, следовательно модель и коэффициенты значимы.

Коэффициент детерминации 0.01046, то есть 0,08% изменения количества продаж объясняется тем, на какой неделе проводилась маркетинговая кампания.

Гомоскедастичность присутствует.

Есть автокорреляция.

Все необходимые предпосылки выполняются.

Средняя ошибка аппроксимации = 10.21% >10%, не очень хороший подбор модели, но приемлемый.

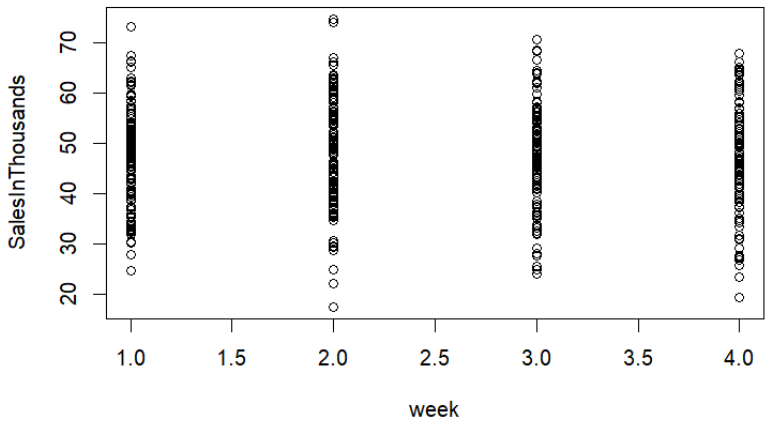

Рис.8. *Модель зависимости количества продаж от возраста магазина*

Дисперсионный анализ для количества продаж и маркетинговой кампании (SalesInThousands и Promotion):

Данные распределены ненормально, поэтому результаты анализа не корректны.

4.2 Множественная линейная регрессия

Для того, чтобы лучше объяснить изучаемую переменную, перейдем к множественному регрессионному анализу.

```
> summary(lm(SalesInThousands~Promotion+AgeOfStore+week, my_dataset))

Call:
lm(formula = SalesInThousands ~ Promotion + AgeOfStore + week,
    data = my_dataset)

Residuals:
     Min       1Q   Median       3Q      Max
-29.8321  -6.7268   0.7525   6.7466  29.4844

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 50.30840    1.72808  29.112  < 2e-16 ***
Promotion   -1.73605    0.58658  -2.960  0.00324 **
AgeOfStore   0.17039    0.07015   2.429  0.01553 *
week        -0.25808    0.41861  -0.617  0.53786
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.08 on 460 degrees of freedom
Multiple R-squared:  0.02973,   Adjusted R-squared:  0.0234
F-statistic: 4.698 on 3 and 460 DF,  p-value: 0.003047
```

Больше всего влияния имеет тип маркетинговой кампании.

Коэффициент детерминации (Adjusted R-squared): 0.0234. То есть, изменения количества продаж на 2% объясняется совокупным влиянием всех исследуемых факторов.

Гомоскедастичности нет.

Автокорреляция есть.

Предпосылки регрессионного анализа не выполнены.

Мультиколлинеарность: значения vif слегка больше 1 (<10). То есть, независимые переменные не связаны линейной зависимостью.

5. ПРИМЕНИМЫЕ УПРАВЛЕНЧЕСКИЕ РЕШЕНИЯ

Результаты проведённого анализа могут быть полезны для менеджмента, владельцев и маркетологов сети магазинов, по которому проводилось исследование.

Исследование показало, что:

Маркетинговые кампании, проводимые в магазинах, дали разный результат. Наиболее эффективная кампания – первая. Наименее эффективная – вторая. В дальнейшей своей деятельности управляющим компании рекомендуется использовать первую маркетинговую кампанию для максимизации количества продаж.

На количество продаж также влияют и другие факторы, например, сколько лет уже существует магазин: в более старых магазинах количество продаж выше. При этом оно не зависит от территории магазина: в маленьких и больших магазинах этот показатель отличается незначительно. Учитывая эти данные, руководство сможет более эффективно вести маркетинговые кампании и распределять их бюджет.