Домашнее задание 3

2022-10-23

Введение

В данном мини-исследовании проводится анализ зависимости рейтингов онлайн-курсов на платформе Udemy от различных факторов.

Гипотезы:

1. Чем больше подписчиков у курса, тем выше его рейтинг, так как количество подписчиков определенно является показателем привлекательности курса для аудитории данной платформы.
2. Люди ценят курсы, на которых много практики, поэтому количество практических тестов положительно влияет на их рейтинг.
3. Рейтинг более новых курсов выше вследствие того, что аудитория считает их более актуальными на данный момент.

```
courses <- read.csv("courses2.csv")
courses$log_num_subscribers <- log(courses$num_subscribers)
years_in_2020 <- 2020 - as.integer(format(as.Date(courses$published_time),"%Y"))
courses$is_new <- ifelse(years_in_2020 < 5, 1, 0)
```

Построение множественной линейной регрессии

Для проверки сформулированных гипотез построим модель. Целевой (зависимой) переменной будет “avg_rating” – средний рейтинг по курсу. Независимыми переменнами являются: “num_subscribers” – дискретная переменная, количество людей, подписанных на курс; “num_ published_ practice_tests”: - дискретная переменная, количество практических тестов на курсе; “is_new” – dummy переменная , прошло ли больше 5 лет с даты публикации.

Перед построением модели необходимо преобразовать столбец “num_subscribers”, заменим на его логарифм, столбец “log_num_subscribers”, чтобы уменьшить эффект масштаба и разброс значений числовой переменной. Столбец “published_time” будет использован для создания dummy переменной таким образом: вычтем из 2020 год выпуска курса и присвоим значения 1 и 0 в зависимости от того, является ли курс новым или нет (меньше 5 лет от даты публикации).

```
lin_mod <- lm(avg_rating ~ log_num_subscribers +  num_published_practice_tests + is_new, data=courses)
summary(lin_mod)
```

```
##
## Call:
## lm(formula = avg_rating ~ log_num_subscribers + num_published_practice_tests +
##     is_new, data = courses)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.14289 -0.24730  0.07365  0.30422  1.14400
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   4.113174   0.019053 215.879  < 2e-16 ***
## log_num_subscribers           0.001643   0.002351   0.699   0.4847
## num_published_practice_tests -0.048132   0.008091  -5.949  2.8e-09 ***
## is_new                        0.025777   0.011049   2.333   0.0197 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4527 on 9443 degrees of freedom
## Multiple R-squared:  0.004176,   Adjusted R-squared:  0.00386
## F-statistic:  13.2 on 3 and 9443 DF,  p-value: 1.345e-08
```

Уравнение модели:

avg_price = 4.113174 + 0.001643 * log_num_subscribers - 0.048132 * num_published_practice_tests + 0.025777 * is_new

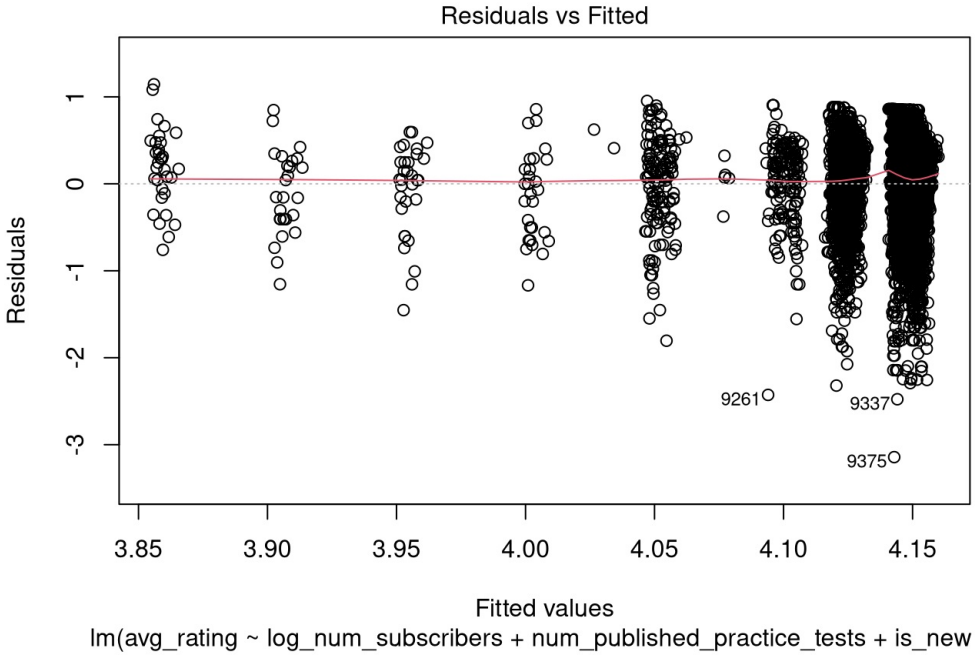
Проинтерпретируем полученные результаты. Независимый коэффициент равен 4.113174, т.е. при нулевых значениях предикторов, курс будет являться хорошим, и так как p-value данного признак практически нулевое, то можно утверждать о его значимости. Коэффициент при логарифме количества подписчиков равен -0.048132, то есть околонулевое значение, но он также и незначим, то есть рейтинг курса не зависит от того, является ли создать курса популярным. Коэффициент при количестве тестов на курсе является значимым и имеет близкое к нулю отрицательное, однако отрицательное значение. Можно сказать, что рейтинг курса практически не зависит от наличия практических заданий. И после коэффициент при dummy переменной тоже оказался низким хоть и значимым, всего 0.025777. То есть новые курсы имеют средний рейтинг немного выше курсов с возрастом от 5 и более лет.

Оценка качества построенной модели

Предсказательная сила модели равна 0.4176 %.

Чтобы понять, выполняется ли условие гомоскедастичности, построим диаграмму рассеивания Fitted values vs Residuals, то есть предсказанные значения среднего рейтинга против остатков модели. Этот график позволит понять, одинаковы дисперсии во все моменты измерения или нет.

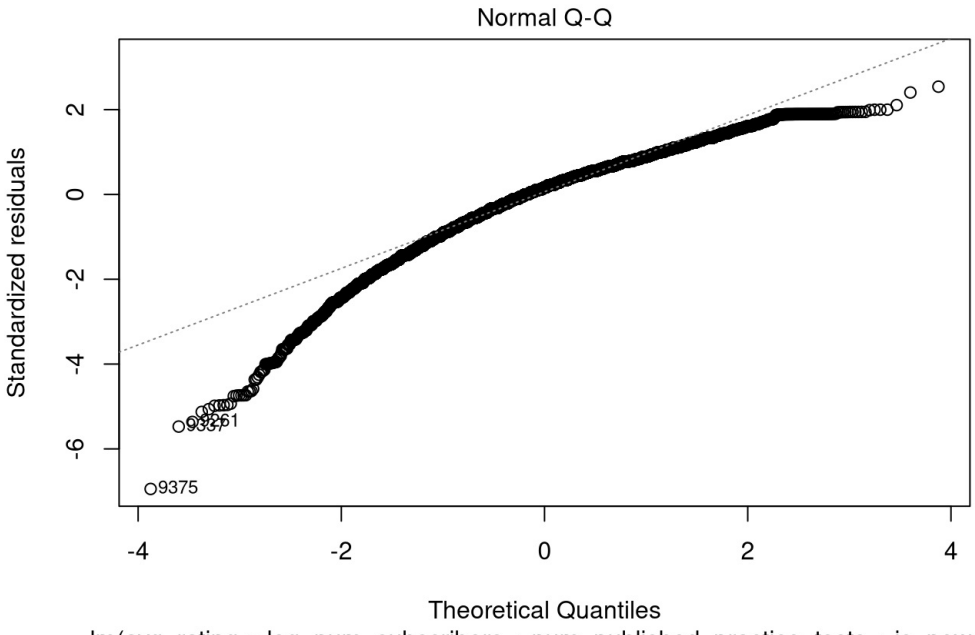
```
plot(lin_mod, which=1)
```



При увеличении значений зависимой переменной разброс точек становится все больше и больше, следовательно, условие гомоскедастичности нарушается.

Теперь построим построим Q-Q plot, чтобы проверить распределение остатков модели на нормальность:

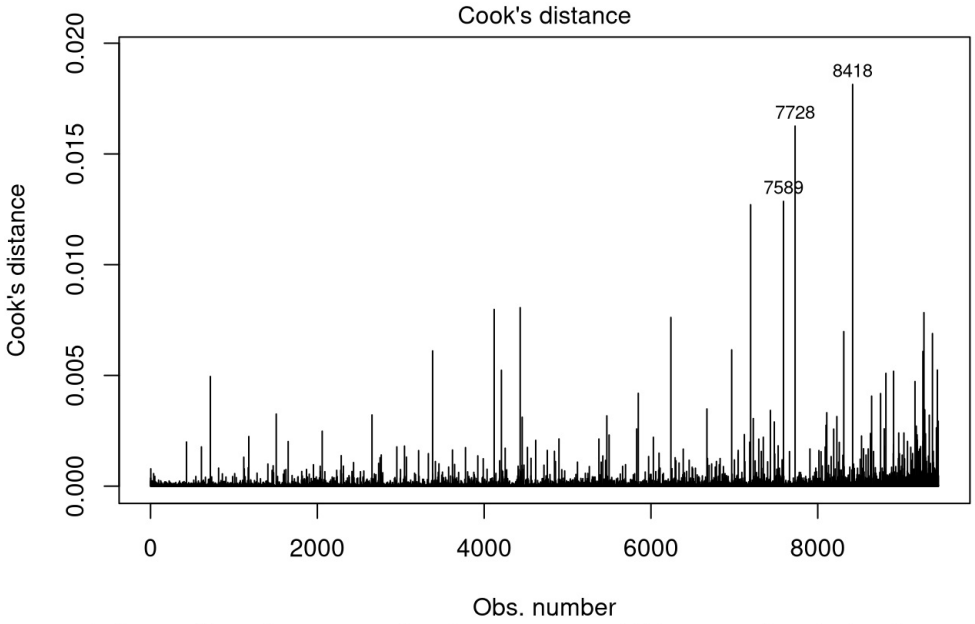
```
plot(lin_mod, which=2)
```



На основе данного графика можно предположить, что распределение остатков модели является нормальным со скосом влево, это можно понять по количеству отклоняющихся от прямой точек слева.

Построим ещё один график, который позволит понять, есть ли в данных влиятельные наблюдения, которые могут существенно искажать оценки коэффициентов в модели:

```
plot(lin_mod, which = 4)
```



На основе данного графика можно утверждать, что нетипичных данных нет.

Проверка условий теоремы Гаусса-Маркова и наличия влиятельных наблюдений: более глубокий анализ. Для удобства сохраним остатки модели и предсказанные значения в столбцы residuals и fitted соответсвенно:

```
courses$residuals <- lin_mod$residuals
courses$fitted <- lin_mod$fitted.values
```

Проверим самое первое условие — условие о равенстве математического ожидания остатков модели нулю. Посмотрим на описательные статистики для остатков:

```
summary(courses$residuals)
```

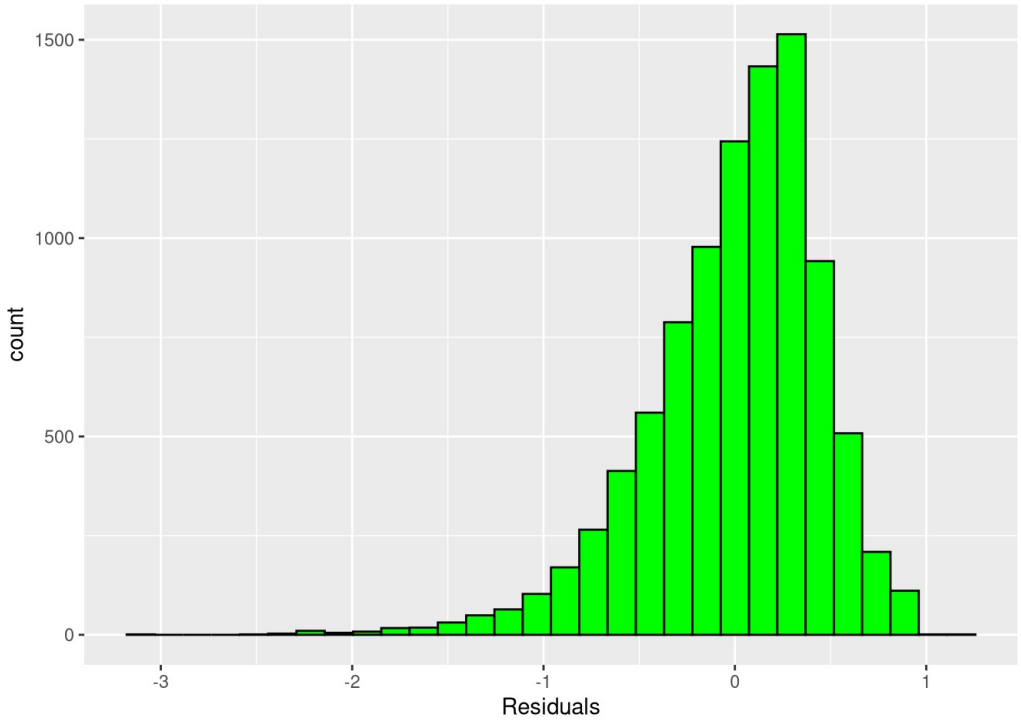
```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -3.14289 -0.24730  0.07365  0.00000  0.30422  1.14400
```

С учетом того, что данном случае среднее равно нулю, а медиана почти равна нулю, то можно предположить, что мат ожидание теоретического распределения также равно нулю, так как все центральные метрики для нормального распределение совпадают.

Чтобы проверить распределение остатков, построит гистограмму:

```
library(ggplot2)
ggplot(data = courses, aes(x = residuals)) +
geom_histogram(fill = "green", color = "black") +
xlab("Residuals")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Наше предположение о том, что распределение остатков является нормальным со скосом влево все еще нельзя отвергнуть.

Проведём тест Шапиро-Уилка на первых 5-ти тысячах наблюдений (для бОльших выборок данный критерий не подходит), чтобы проверить распределение на нормальность:

```
shapiro.test(courses$residuals[5000:10000])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  courses$residuals[5000:10000]
## W = 0.96667, p-value < 2.2e-16
```

Так как p-value ~ 0, гипотезу о нормальном распределении остатков можно отвергнуть.

Для полноты проверим данную гипотезу еще раз, но уже с помощью теста Колмогорова-Смирнова на полной выборке:

```
ks_test_sample <- order(unique(courses$residuals))
ks.test(ks_test_sample, 'pnorm')
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  ks_test_sample
## D = 0.99964, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

Выводы те же - распределение остатков не является нормальным.

Проведем тест Бройша-Пагана для проверки гомоскедастичности

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
bptest(lin_mod)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  lin_mod
## BP = 106.12, df = 3, p-value < 2.2e-16
```

p-value ~ 0, следовательно, гипотеза о постонстве дисперсии отвергается, модель гетероскедостична.

Выводы:

1. Влиятельные наблюдения остутсвуют, то есть значения коэффициентов при модели имеют смысл, то есть они основываются на распределении данных в целом, а не отдельных точках.
2. Ни одно из условий Гаусса-Маркова не выполняется, возможны высокие значения стандартных ошибок вследствие гетероскедастичности.

Итоги исследования:

В ходе данного исследования были сделаны следующие выводы. Гипотеза о том, что средний рейтинг курса на платформе Udemy зависит от количества людей, подписанного на него, отвергнута. Также наличие большого количества практических тестов не играет важной роли, а даже может незначительно снизить средний рейтинг курса. И последняя гипотеза о том, что чем больше курсу лет, тем более высокий у него рейтинг принята, однако влияние возраста также совсем незначительно.