Практикум 6. Множественная линейная регрессия. Проверка условий Гаусса-Маркова.

Алла Тамбовцева

Содержание

Загрузка библиотек и подготовка данных

Загрузим библиотеки `ggplot2` и `lmtest`, которые мы устанавливали на занятиях ранее:

```
library(ggplot2)
library(lmtest)
```

Если не запустить строки кода выше, R не будет понимать функции из библиотек `ggplot2`. В нашем случае это будет выражаться, например, в невозможности построить графики при помощи `ggplot()` из-за ошибки *не могу найти функцию "ggplot"*.

Если при загрузке библиотек R выдал ошибку *нет пакета под названием ...*, то соответствующую библиотеку нужно установить:

```
install.packages("ggplot2")
install.packages("lmtest")
```

Напоминание. При установке библиотеки в консоли будет появляться много текста, это нормально. R последовательно информирует нас о том, как он подключается к серверу, скачивает необходимые компоненты библиотеки, распаковывает архивы с ними и устанавливает на компьютер.

Если на Windows при установке R выдает ошибку и пишет что-то про отсутствие доступа, выдавая при этом путь к папке с вопросительными знаками, можно запустить RStudio от имени администратора (кликнуть правой клавишей на значок *RStudio*, и выбрать соответствующий пункт в *Дополнительно*) и попробовать установить библиотеку. Эта проблема обычно возникает, если в пути к папке, куда RStudio устанавливает библиотеки, есть слова на кириллице.

Загрузим данные из файла `Salaries.csv` и сохраним их в датафрейм `salaries`:

```
salaries <- read.csv(file.choose())
```

В этом файле хранятся данные о заработной плате сотрудников некоторого университета США и характеристики самих сотрудников. Переменные в файле:

- `rank`: должность;
- `discipline`: тип преподаваемой дисциплины (`A` – теоретическая, `B` – практическая);
- `yrs.since.phd`: число лет с момента получения степени PhD;
- `yrs.service`: число лет опыта работы;
- `sex`: пол;
- `salary`: заработная плата за 9 месяцев, в долларах.

В этот раз давайте не будем использовать в регрессионных моделях столбцы с текстовыми значениями, которые R автоматически кодирует с помощью чисел 0 и 1, а создадим такие бинарные столбцы самостоятельно. Для этого нам понадобится функция `ifelse()`.

Создадим столбец `male`, в котором значение 1 соответствует сотрудникам мужского пола, а значение 0 соответствует сотрудникам женского пола.

```
salaries$male <- ifelse(salaries$sex == "Male", 1, 0)
```

Код выше работает так: функция `ifelse()` проверяет для каждого значения в столбце `sex` выполнение условия равенства строке `"Male"` и, если условие выполняется, ставит 1, если нет — ставит 0. А затем полученный набор из 0 и 1 мы добавляем в датафрейм `salaries` в качестве нового столбца `male`.

Аналогичным образом создадим столбец `practice` в котором значение 1 соответствует практическим курсам, а значение 0 — теоретическим:

```
salaries$practice <- ifelse(salaries$discipline == "B", 1, 0)
```

Построение регрессионной модели

Построим модель линейной регрессии, которая покажет нам характер взаимосвязи между заработной платой, числом лет опыта работы, числом лет после получения степени PhD, пола и типа преподаваемых курсов:

```
model1 <- lm(data = salaries, salary ~ yrs.service +
               yrs.since.phd + male + practice)
summary(model1)
```

```
##
## Call:
## lm(formula = salary ~ yrs.service + yrs.since.phd + male + practice,
##     data = salaries)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -75974 -17094  -3799  16073  97055
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    71325.8     4981.8  14.317  < 2e-16 ***
## yrs.service     -770.1      244.1  -3.155  0.00173 **
## yrs.since.phd   1804.1      248.9   7.249 2.25e-12 ***
## male            7545.3     4462.6   1.691  0.09167 .
## practice       16325.4     2708.9   6.027 3.87e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26130 on 392 degrees of freedom
## Multiple R-squared:  0.2634, Adjusted R-squared:  0.2558
## F-statistic: 35.04 on 4 and 392 DF,  p-value: < 2.2e-16
```

Проинтерпретируем полученные коэффициенты:

- `Intercept`. Если все независимые переменные равны 0, то, согласно нашей модели, средняя заработная плата составляет примерно 71325.8 долларов. Ситуация в данном случае вполне реалистичная, это средняя заработная плата женщины (`male = 0`) без опыта работы (`yrs.service = 0`), только что получившей степень `phd` (`yrs.since.phd = 0`) и читающей теоретические курсы (`practice = 0`).
- `yrs.service`: при прочих равных, то есть если мы будем сравнивать сотрудников, которые отличаются только стажем, у человека с опытом работы на 1 год выше, заработная плата, в среднем, была ниже на 770 долларов;
- `yrs.since.phd`: при прочих равных, то есть если мы будем сравнивать сотрудников, которые отличаются только числом лет после получения степени, у человека, который получил степень на 1 год раньше, заработная плата, в среднем, будет выше на 1804 долларов;
- `male`: при прочих равных, то есть если мы будем сравнивать сотрудников, которые отличаются только по полу, заработная плата мужчин, в среднем, будет выше заработной платы женщин на 7545 долларов;
- `practice`: при прочих равных, то есть если мы будем сравнивать сотрудников, которые отличаются только типом читаемых курсов, заработная плата сотрудников, преподающих практические курсы, в среднем, будет выше на 16325 долларов.

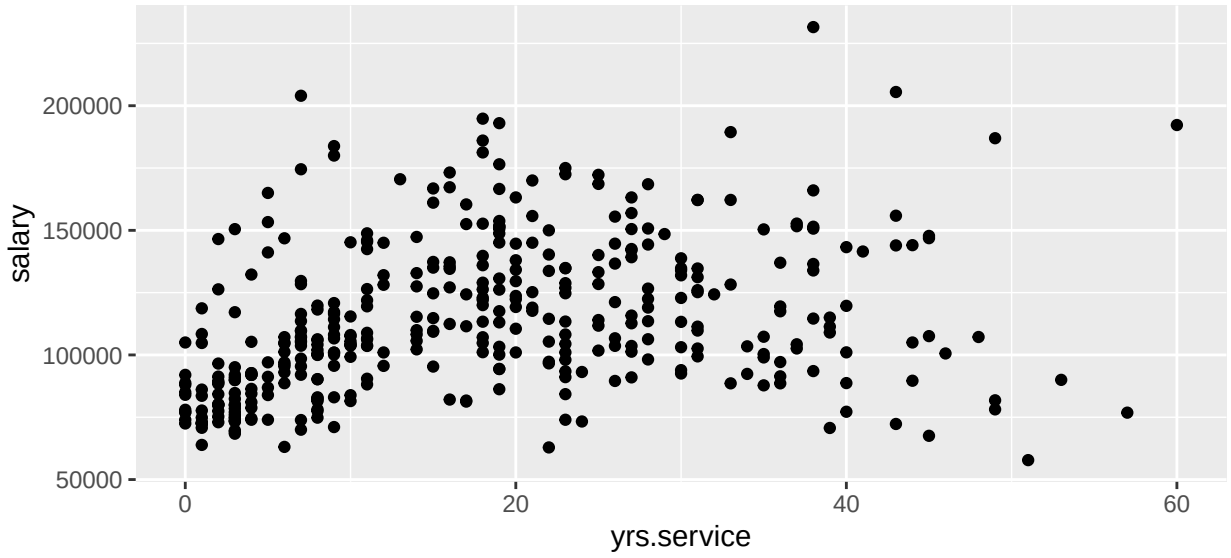
Если подумать, коэффициент при опыте работы немного странный — он отрицательный, получается, больший опыт работы сопровождается снижением заработной платы. При этом число лет после получения степени PhD, наоборот, сказывается на заработной плате положительно. Такая странность отчасти объяснима: оба показателя связаны с возрастом человека, а связь между возрастом и заработной платой не совсем линейная, поэтому наша линейная модель может нас немного вводить в заблуждение.

Посмотрим на последний столбец в выдаче и найдем p-value. Судя по звездочкам и самим значениям p-value, все оценки коэффициентов, кроме коэффициента при `male`, являются статистически значимыми на уровне значимости 5%. С полом ситуация немного другая — этот фактор значим только на 10% уровне значимости. Но, тем не менее, давайте его не выкидывать из модели, сложно отрицать, что пол не влияет на заработную плату (на самом деле, эти данные собирались как раз для того, чтобы выявить и изучить неравенство доходов мужчин и женщин в университете). Итак, истинные значения всех коэффициентов можно считать отличными от нуля, изменчивость заработной можно объяснить изменчивостью тех факторов, которые мы включили в модель.

Если мы попытаемся оценить предсказательную силу модели, она будет невысокой, R^2 у модели всего 26%. Это неслучайно. Ключевые показатели модели, `yrs.service` и `yrs.since.phd`, те, которые в силу большего разнообразия значений объясняют большую часть дисперсии заработной платы, связаны с ней не линейно, а квадратично. Все упирается в возраст. До какого-то возраста увеличение стажа приводит к увеличению заработной платы (больше опыта, больше навыков), а затем зависимость становится обратной (меньше сил, меньше желания обучаться новым навыкам, ограничения по здоровью). Поэтому линейная модель может быть логичной и удобной для интерпретации, но несовершенной.

Как увидеть нелинейную зависимость? Построим диаграмму рассеивания между `yrs.service` и `salary`:

```
ggplot(data = salaries, aes(x = yrs.service, y = salary)) +
  geom_point()
```



Пояснения к коду:

В аргумент `data` записываем название датафрейма, внутри `aes()` указываем, какие переменные идут по осям `x` и `y`. Далее через `+` добавляем слои, отвечающие за тип и внешний вид графика, `geom_point()` строит диаграмму рассеивания (точечный график).

На графике видно, что облако точек лучше описывается не прямой, а некоторой параболой с ветвями вниз: до 25 лет стажа связь между показателями прямая, а после — обратная (хотя, конечно, не у всех, есть сотрудники, над которыми время не властно).

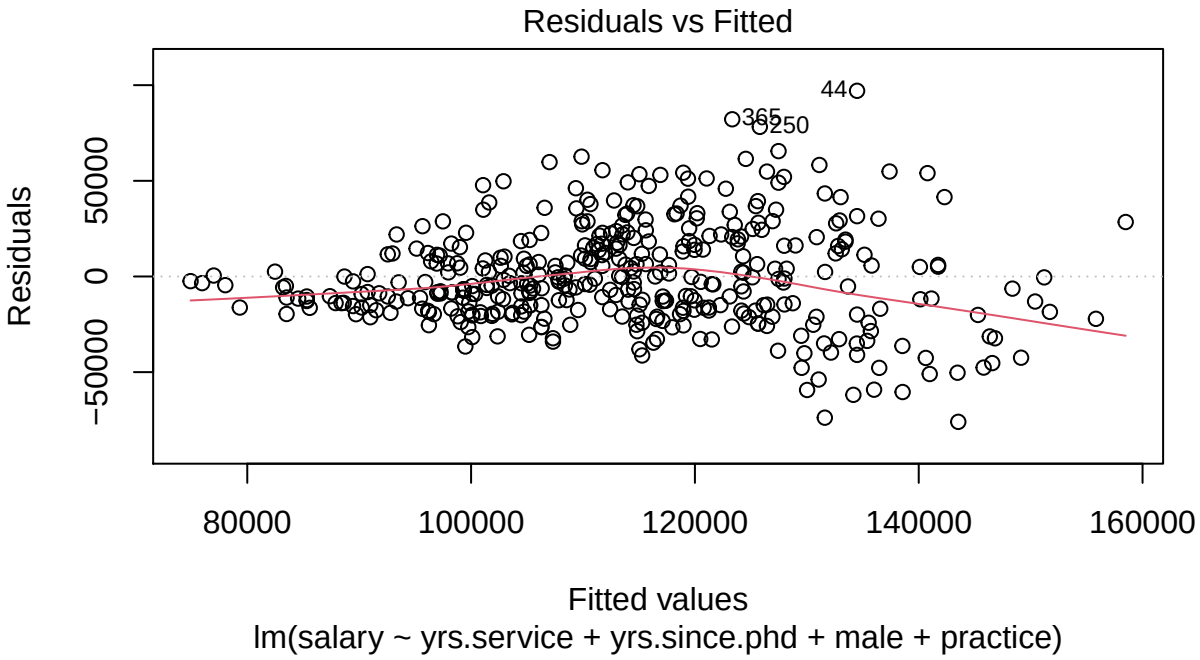
Проверка условий Гаусса-Маркова и наличия влиятельных наблюдений: предварительный анализ

Посмотрим, выполняются ли допущения линейной модели — условия Гаусса-Маркова. Проанализируем остатки (ошибки) модели.

Начнем с самого простого, но не очень подробного способа (и графики не очень симпатичные). Воспользуемся базовой функцией `plot()`, которая построит четыре основных графика для диагностики модели `model1`.

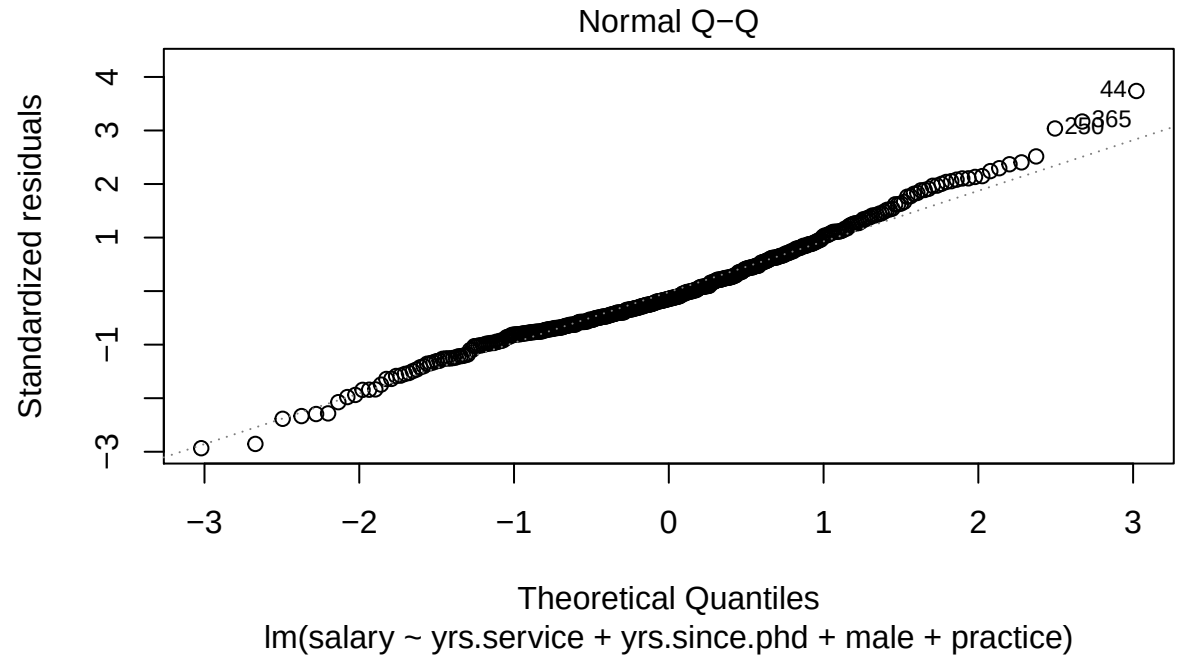
```
plot(model1)
```

```
plot(model1, which = 1)
```



Первый график представляет собой диаграмму рассеивания `Fitted values` vs `Residuals`, то есть предсказанные значения заработой платы против остатков модели. Этот график позволяет понять, можно ли считать дисперсию остатков постоянной, то есть выполняется ли условие гомоскедастичности. Хорошая модель может ошибаться, но она должна это делать примерно одинаково во всех случаях. Что мы видим на графике? Что модель ошибается всегда по-разному: при предсказании маленьких значений заработной платы модель ошибается мало (точки лежат близко к линии `Residuals = 0`, разброс остатков минимальный), при предсказании средних значений заработной платы — чуть больше (разброс точек более заметный), а при больших — совсем сильно. Визуально точки на графике образуют расширяющуюся «воронку», а значит, дисперсия остатков увеличивается при увеличении предсказанных значений зависимой переменной. Условие гомоскедастичности не выполнено, имеет место гетероскедастичность.

```
plot(model1, which = 2)
```



Второй график — график *Q-Q plot*, то есть *quantile-to-quantile plot*. Он строится для остатков модели и позволяет понять, можно ли считать их распределение нормальным. Идея такая: если мы стандартизуем остатки (вычтем из

каждого среднее и поделим на стандартное отклонение), мы сможем сравнить их распределение со стандартным нормальным.

Для этого для каждого значения стандартизованного остатка в нашей выборке вычисляется доля наблюдений, которые его не превышают — уровень квантиля, а затем в стандартном нормальном распределении находится квантиль этого уровня. По горизонтальной оси на графике отмечаются квантили стандартного нормального распределения, а по вертикальной — реальные значения стандартизованных остатков, точки с соответствующими координатами наносятся на график. Если значение по горизонтальной оси совпадает со значением по вертикальной оси, это означает, что отклонения от стандартного нормального распределения нет, визуально такая точка лежит на диагональной прямой, отмеченной на графике.

Для закрепления понимания рассмотрим пример. У нас есть остаток модели, который равен -75973.7. То есть, предсказывая заработную плату для некоторого человека, наша модель ошибается на 75973.7 долларов (не пугайтесь числам, заработная плата за 9 месяцев, значения там огромные сами по себе). После стандартизации этот остаток стал равным -2.92. Сколько наблюдений не превышают это значение? Ровно одно из 397, значит, доля таких наблюдений равна 0.0025. А чему было бы равно значение остатка, если бы оно было взято из стандартного нормального распределения? Найдем квантиль уровня 0.0025:

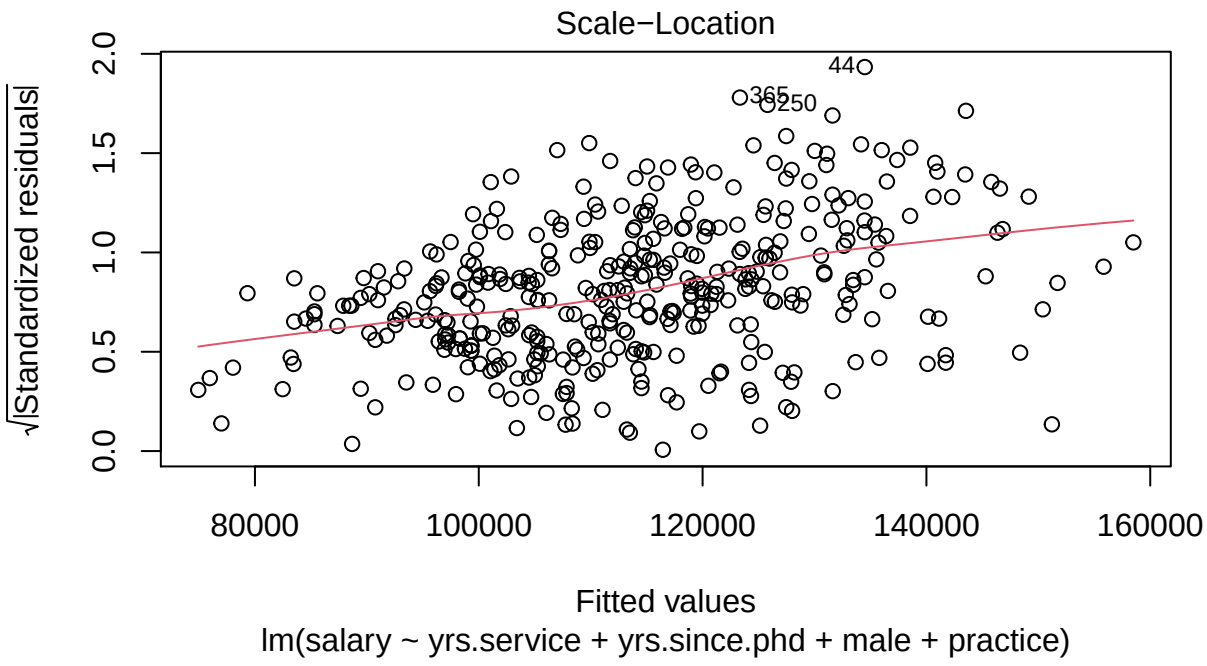
```
qnorm(p = 0.0025)
```

```
## [1] -2.807034
```

Оно было бы равно —2.81. Наносим на график точку с координатами —2.81 и —2.92 и видим, что точка почти лежит на прямой.

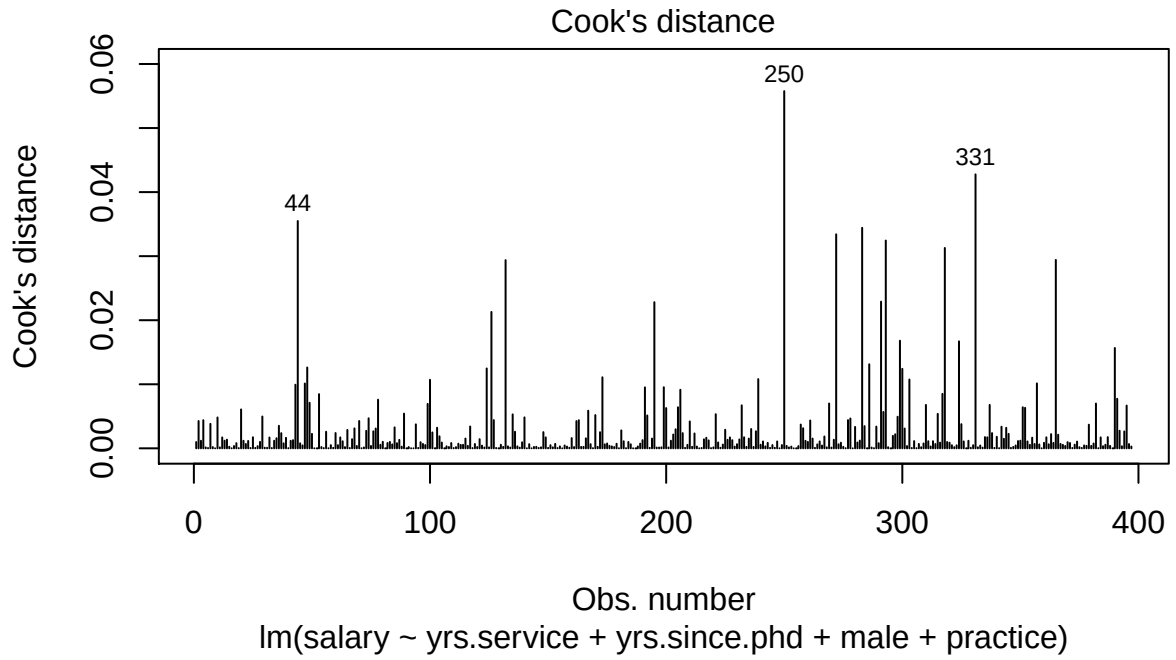
Итого: если на графике есть точки, которые сильно отклоняются от прямой (R подписывает номера таких наблюдений), говорить о нормальном распределении нельзя. В данном случае распределение остатков на нормальное не похоже.

```
plot(model1, which = 3)
```



Третий график можем пропустить — его смысл такой же, как у первого, только остатки здесь взяты по модулю, а затем из них извлечен корень. А вот четвертый график новый, он позволяет понять, есть ли в данных влиятельные наблюдения, те, которые могут существенно искажать оценки коэффициентов в модели.

```
plot(model1, which = 4)
```



Мер влиятельности существует множество, и их вычисление выглядит не всегда понятно. Поэтому давайте просто зафиксируем, что на этом графике отображены два показателя — потенциал влияния каждого наблюдения (*leverage*) и сами остатки. Влиятельными будут считаться те наблюдения, которые одновременно обладают высоким потенциалом влияния и являются нетипичными, то есть, чьи значения остатков лежат далеко от 0. В общем, на таком графике внимание стоит обращать на верхний правый угол и на нижний правый угол, если какие-то точки в этих углах есть, наблюдения являются нетипичными (R их подписывает и отчерчивает красной пунктирной линией).

Проверка условий Гаусса-Маркова и наличия влиятельных наблюдений: более глубокий анализ

Второй вариант проверки условий — более фундаментальный. Мы построим больше графиков и задействуем формальные статистические тесты.

Для начала сохраним остатки модели и предсказанные значения заработной платы в отдельные столбцы `salaries`:

```
salaries$res <- model1$residuals
salaries$fitted <- model1$fitted.values
```

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

```
summary(salaries$res)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -75974  -17094   -3799       0   16073   97055
```

Как и всегда, видим, что среднее арифметическое равно 0. Это обеспечивается по построению, так устроен метод наименьших квадратов. Поэтому среднее здесь не совсем информативно, стоит смотреть на медиану. А медиана здесь сильно меньше среднего, отрицательна, на 0 не похожа. Если медиана выборки не 0, можем предположить, что и медиана теоретического распределения остатков тоже не 0. А если так, то и математическое ожидание теоретического распределения остатков не 0, так как в случае нормального распределения эти параметры совпадают. Условие не выполняется.

Раз речь зашла о нормальности, давайте сразу, более красивым образом, проверим, можно ли считать остатки нормально распределенными. Для этого просто построим гистограмму:

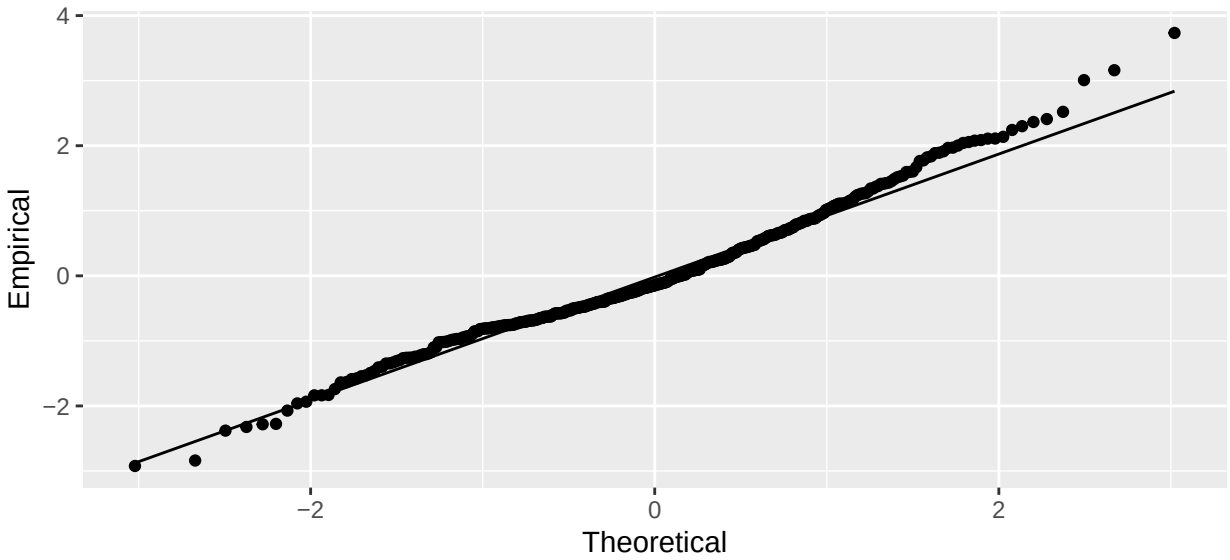
```
ggplot(data = salaries, aes(x = res)) +
  geom_histogram(fill = "cornflowerblue", color = "black") +
  xlab("Model residuals")
```

Пояснения к коду:

- так как R сам определяет высоты столбцов в гистограмме, в `aes()` достаточно указать только показатель по оси `x`, а здесь это остатки; слой `geom_histogram()` — слой для построения гистограммы.

Да, распределение не похоже на нормальное. В нем как будто бы два пика, большой и маленький, плюс, распределение скошено вправо. Проверим то же самое через красивый *Q-Q plot*:

```
ggplot(data = salaries, aes(sample = scale(res))) +
  stat_qq() + stat_qq_line() +
  xlab("Theoretical") + ylab("Empirical")
```



Пояснения к коду: `scale(res)` — не забываем стандартизовать остатки, `stat_qq()` — высчитывает квантили и добавляет точки, `stat_qq_line()` — добавляет диагональную линию для удобства сравнения.

Как мы уже знаем, нормальности нет. Финальный аккорд в этой части — формальный тест. Так как разобранный на лекции критерий Колмогорова-Смирнова в R реализуется не совсем удобно, реализуем другой критерий со схожей гипотезой.

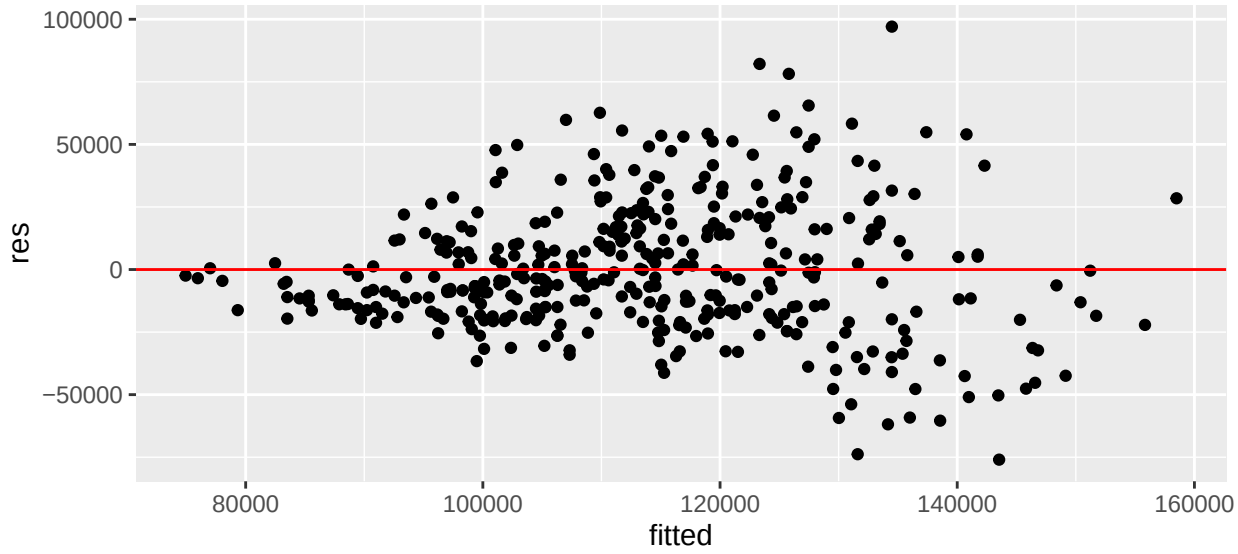
```
shapiro.test(salaries$res)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  salaries$res
## W = 0.98374, p-value = 0.0001931
```

Это критерий Шапиро-Уилка для проверки нормальности, нулевая гипотеза здесь такая: выборка (остатки модели в нашем случае) взята из нормального распределения с некоторыми параметрами. Судя по p-value, которое примерно равно 0.0002, эту гипотезу следует отвергнуть. Теперь даже формально доказали, что условие о нормальности остатков модели стоит отвергнуть.

Два условия проверили. Давайте теперь более обстоятельно посмотрим на условие гомоскедастичности. В самом начале мы выяснили, что это условие не выполняется, посмотрев на график *предсказанные значения* vs *остатки*. При желании можем построить такой же график, только с `ggplot()`:

```
ggplot(data = salaries, aes(x = fitted, y = res)) +
  geom_point() + geom_hline(yintercept = 0, color = "red")
```



Диаграмму рассеивания мы уже строили, стоит отметить только, что здесь мы добавили на нее горизонтальную линию `y = 0` красного цвета для удобства оценки разброса точек относительно 0.

Формальный критерий для проверки гомоскедастичности тоже есть, это критерий Бройша-Пагана. Применим его к нашей модели `model1`:

```
bptest(model1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model1
## BP = 60.249, df = 4, p-value = 2.572e-12
```

Гипотезы здесь такие:

$$H_0 : Var(\varepsilon) = \sigma^2, \text{ гомоскедастичность}$$

$$H_1 : Var(\varepsilon) \neq \sigma^2, \text{ гетероскедастичность}$$

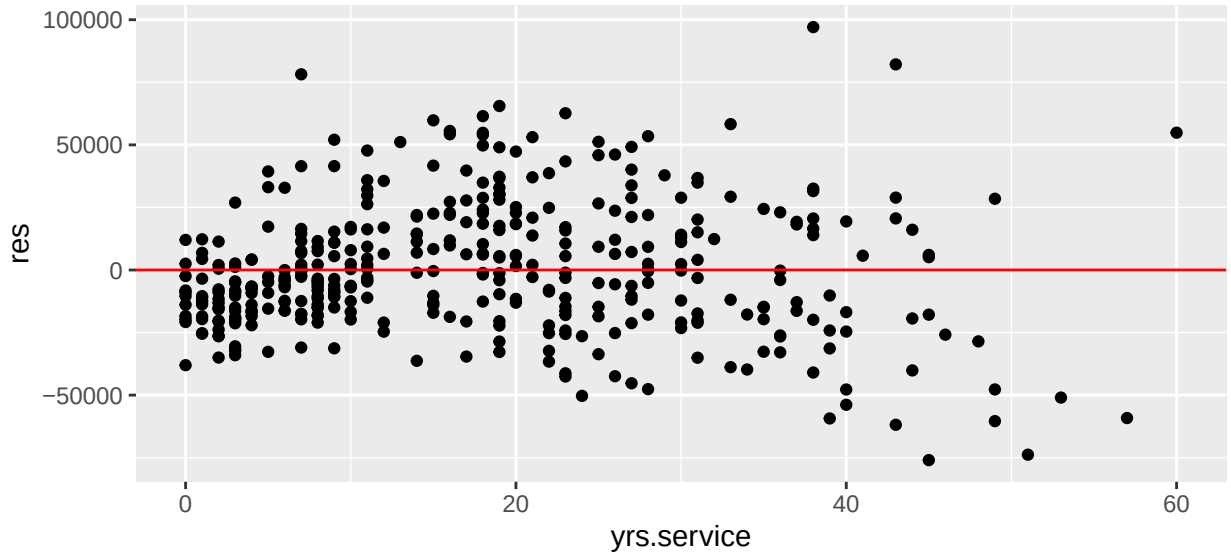
Значение p-value здесь очень маленькое, примерно 0, поэтому нулевую гипотезу мы отвергаем, условие гомоскедастичности, постоянства дисперсии остатков, не выполняется.

Теперь посмотрим на выполнение условия случайности остатков и проверим, есть ли связь между ошибками модели и независимыми переменными. По идее, если модель хорошая, связи между ними быть не должно, потому что модель должна ошибаться случайно, без всяких закономерностей и зависимостей.

Построим диаграммы рассеивания *независимая переменная* vs *остатки* для каждой независимой переменной. Их у нас четыре.

Итак, первая диаграмма:

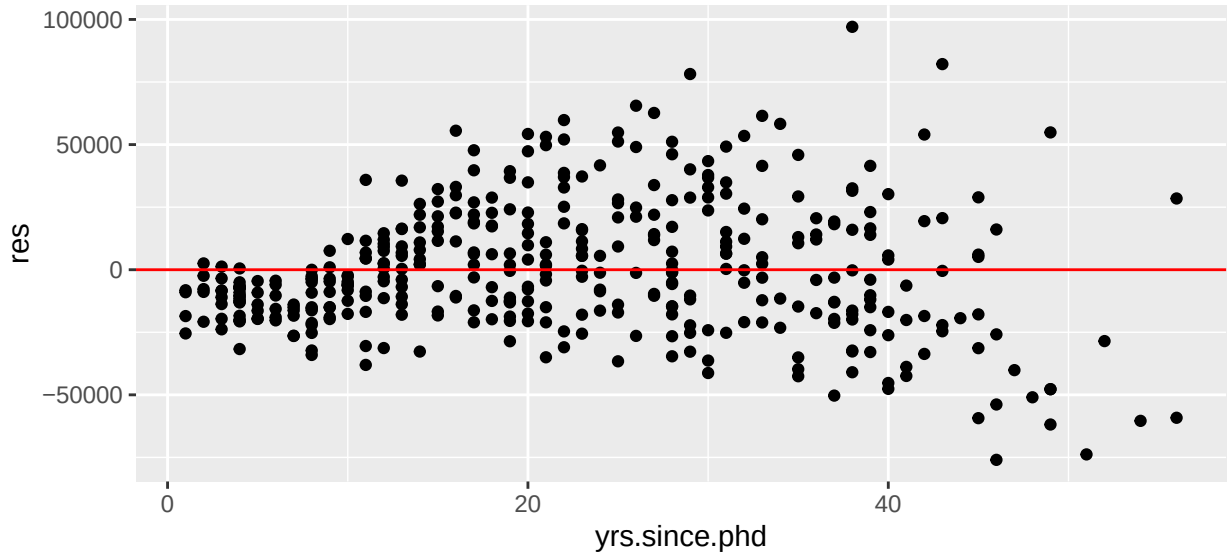
```
ggplot(data = salaries, aes(x = yrs.service, y = res)) +
  geom_point() + geom_hline(yintercept = 0, color = "red")
```



Сказать, что точки разбросаны случайно, нельзя. Явно видна некоторая зависимость между значениями лет опыта работы и остатками, причем эта зависимость нелинейная.

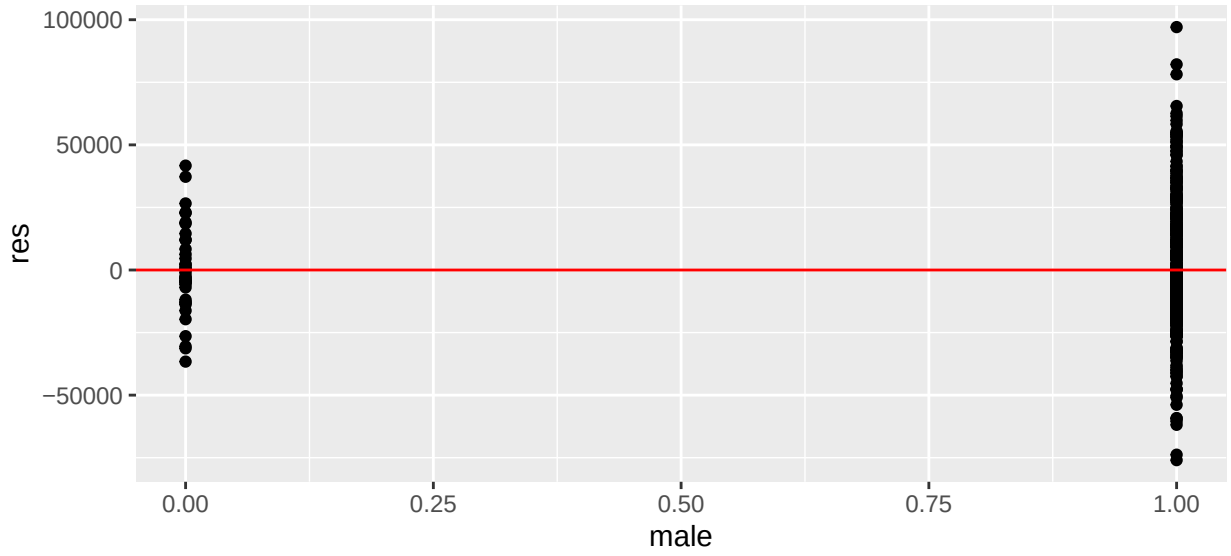
Со второй диаграммой та история:

```
ggplot(data = salaries, aes(x = yrs.since.phd, y = res)) +
  geom_point() + geom_hline(yintercept = 0, color = "red")
```



А вот с третьей диаграммой интереснее.

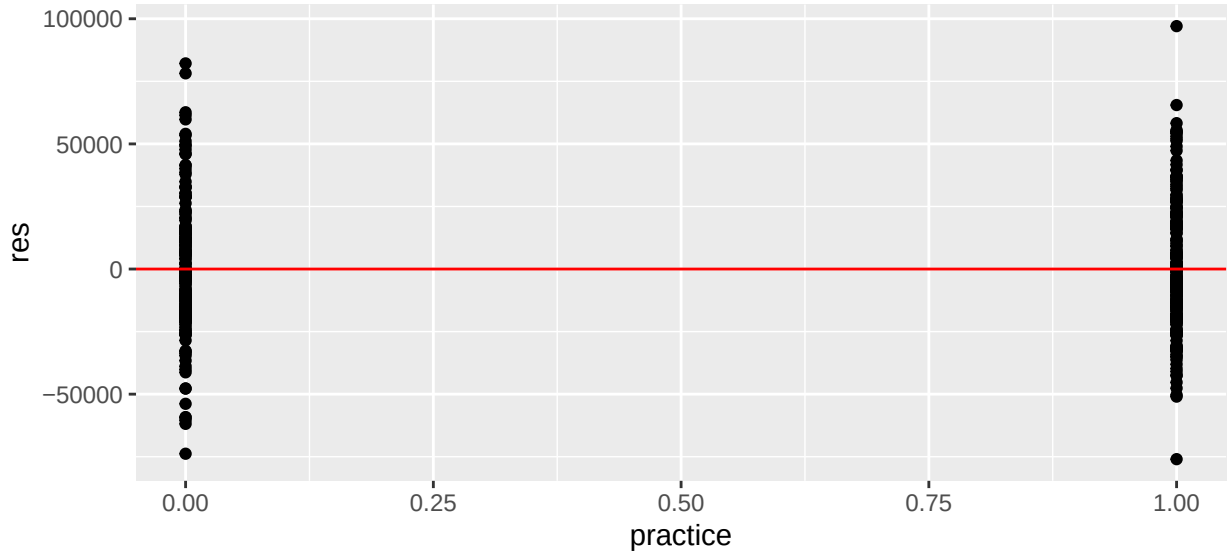
```
ggplot(data = salaries, aes(x = male, y = res)) +
  geom_point() + geom_hline(yintercept = 0, color = "red")
```



Она может показаться странной, но на самом деле, все в порядке: у переменной `male` только два значения, 0 и 1. Но и в таком случае разброс точек можно оценить. По графику видно, что при `male = 0` точки более скучены относительно прямой `res = 0`, а при `male = 1` точки более разбросаны. Отчасти это, конечно, связано с тем, что сотрудников мужского пола в выборке больше, но, тем не менее, отрицать, что модель ошибается по-разному в случае предсказания заработной платы мужчин и женщин, нельзя.

На четвертой диаграмме ситуация уже другая:

```
ggplot(data = salaries, aes(x = practice, y = res)) +
  geom_point() + geom_hline(yintercept = 0, color = "red")
```



Разброс точек примерно одинаков и для `practice = 0`, и для `practice = 1`. Но три случая неслучайности остатков против одного — о случайности остатков и их независимости от переменных в модели говорить не приходится.

Проверка условий Гаусса-Маркова и наличия влиятельных наблюдений: запасной вариант

Есть еще один способ быстро проверить выполнение некоторых условий Гаусса-Маркова и наличие нетипичных наблюдений — воспользоваться функцией `autoplot()` из библиотеки `ggfortify`. Это аналог базовой функции `plot()`,

которую мы использовали в начале, только строит она более продвинутые и симпатичные графики. Один минус — не всегда эта библиотека благополучно устанавливается (хотя, по сути, это просто некоторая надстройка над библиотекой `ggplot`).

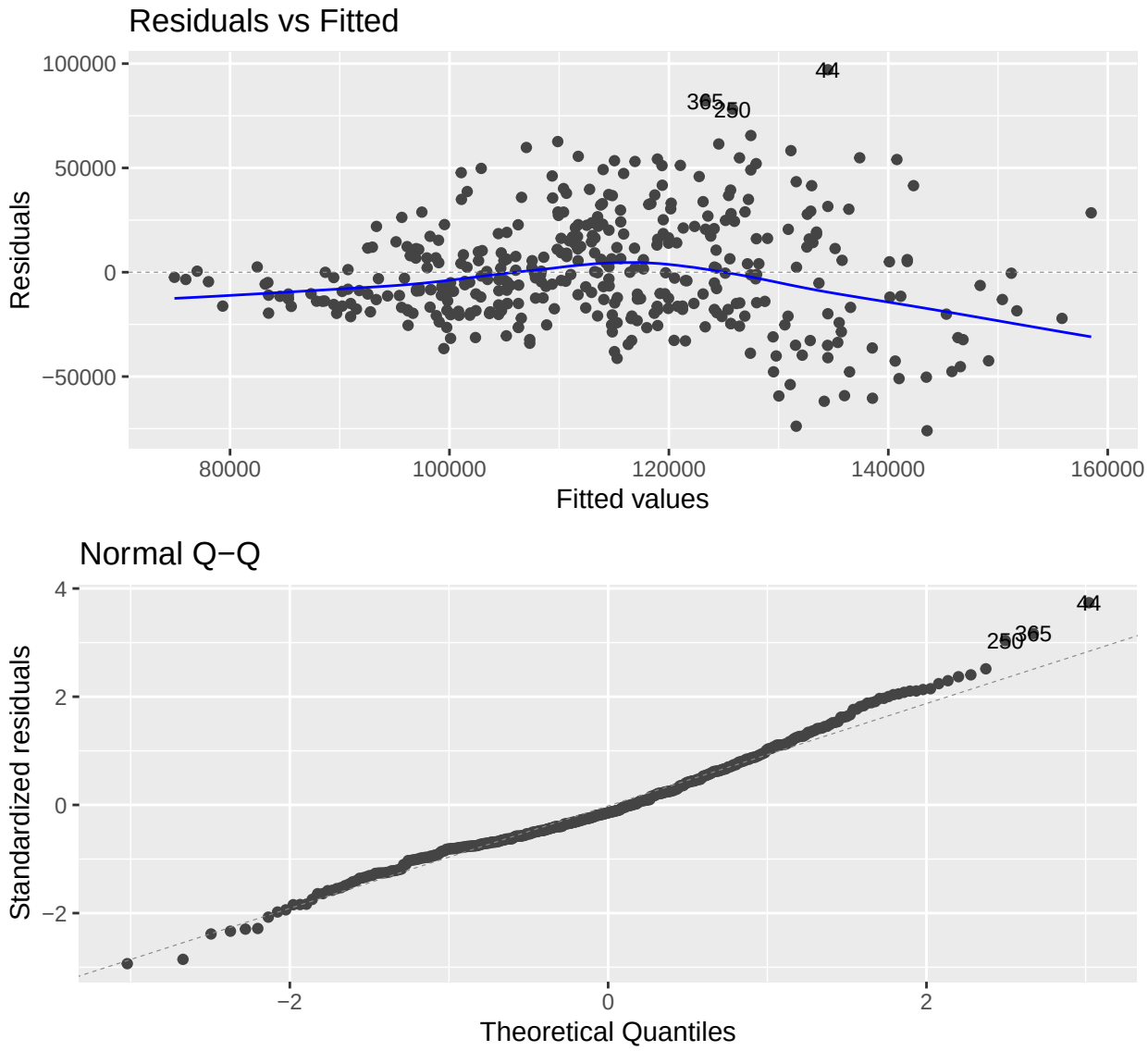
```
install.packages("ggfortify")
```

Если библиотека установилась, к ней можно обратиться:

```
library(ggfortify)
```

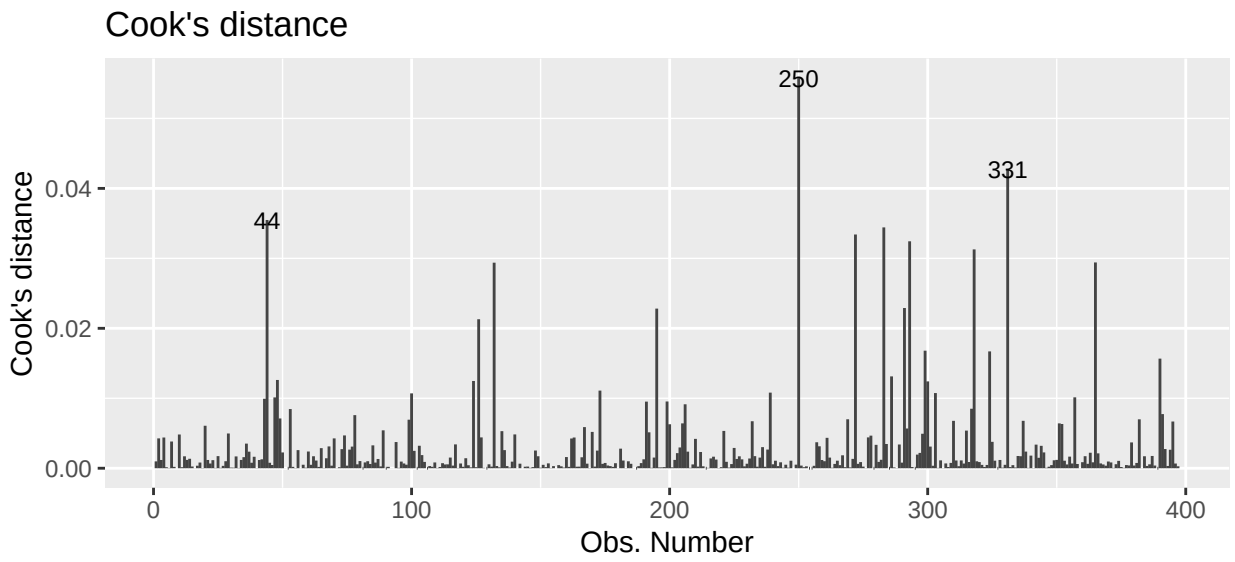
Одной функцией `autoplot()` можно построить много графиков для диагностики модели. Мы выберем лишь некоторые из них. Знакомые графики для выявления гетероскедастичности и проверки нормальности остатков:

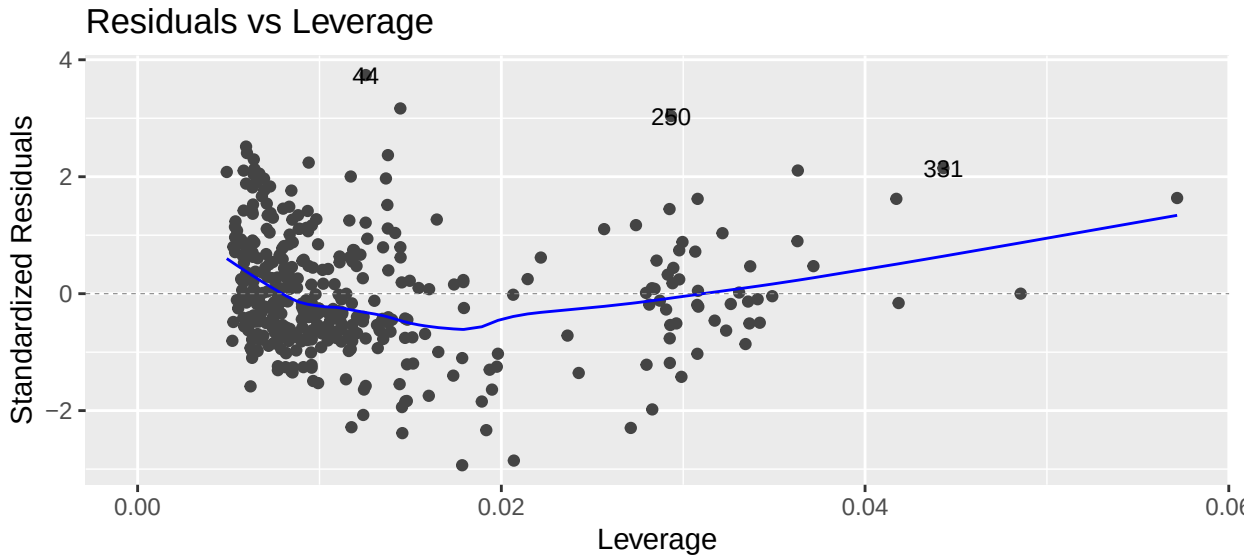
```
autoplot(model1, which = 1:2, nrow = 2, label.size = 3)
```



Графики для выявления влиятельных наблюдений:

```
autoplot(model1, which = 4:5, nrow = 2, label.size = 3)
```





На первом графике отмечены номера наблюдений (строк в таблице), и для каждого наблюдения вычислена мера влиятельности — мера Кука. Каких-то строгих критериев для того, чтобы решать, какие значения меры Кука являются достаточно большими, чтобы считать наблюдения влиятельными, нет, но обычно в качестве порогового значения выбирают 1. Второй график мы уже видели — потенциал влияния и стандартизованные остатки.

Подведем итоги

Итак, мы выяснили две вещи:

- влиятельных наблюдений в наших данных нет;
- ни одно из условий Гаусса-Маркова не выполняется.

Первый факт — обнадеживающий. На оценки коэффициентов, которые мы видели в выдаче, можно положиться; их значения обеспечиваются распределением данных и связям в них, а не наличием особых точек, которые «перетягивают» на себя регрессионную плоскость. Если бы влиятельные наблюдения были, соответствующие им строки в таблице нужно было бы удалить (но прежде, конечно, изучить, что это за наблюдения и почему они могут быть такими).

Второй факт — настораживающий. Основная проблема здесь — гетероскедастичность. Эта проблема приводит к тому, что значения стандартных ошибок могут получаться завышенными. Почему это плохо? Потому что если

стандартные ошибки высокие, то абсолютные значения t-статистик будут, наоборот, заниженными (вспомните, как считается `t value` в выдаче, это значение оценки коэффициента, деленное на стандартную ошибку). А если значения t-статистики по модулю близки к нулю, то p-value, наоборот, будут высокими. Если p-value высокое, то нулевая гипотеза о равенстве истинного коэффициента будет не отвергаться. В итоге мы будем получать незначимые коэффициенты, и они будут таковыми не потому что связи между показателями нет, а потому что гетероскедастичность исказила ошибки коэффициентов. Как поступают в таком случае? Выбирают тип стандартных ошибок, устойчивый к наличию гетероскедастичности и строят регрессионную модель заново, с выбранными ошибками. Но вопрос выбора этих стандартных ошибок мы пока оставим за скобками.