Линейный регрессионный анализ

10.05.2017

Продолжаем изучать зависимость между переменными

Коэффициент корреляции измеряет выраженность *линейной* связи между переменными.

В линейном регрессионном анализе строится линейное уравнение, описывающее статистическую зависимость переменной Y от переменной X.

В результате аналитик может прогнозировать значение переменной Y.

Более того, если он способен изменять значение переменной X, он может в некоторой степени управлять переменной Y.

Пытаемся управлять продажами

Если Y — уровень продаж, а X — затраты на рекламу, то можно управлять уровнем продаж, подбирая оптимальное значение переменной X.

Но не все так просто

Способность управлять ограничена, поскольку на уровень продаж влияют не только затраты на рекламу, но и многие другие показатели, которыми управлять труднее, достаточно упомянуть цену.

Кроме того, необходимо, чтобы была справедлива исходная гипотеза о виде зависимости переменных, а это часто не так.

Случай двух переменных X и Y

Дано

Наблюдения, то есть пары чисел (x_i, y_i).

Гипотеза, что имеется линейная статистическая зависимость между переменными X и Y

$$Y = a + bX. \tag{1}$$

Найти

Оценки коэффициентов a и b уравнения регрессии (1).

Решение

Геометрическая идея: уравнение регрессии определяет прямую, наиболее близко проходящую ко всем точкам с координатами (x_i, y_i).

Подбираем a и b так, чтобы сумма квадратов отклонений точек линии регрессии Y от наблюдаемых значений y_i была минимальной:

$$F(a,b) = \sum_{i=1}^{N} (y_i - f(x_i, a, b))^2 \to \min$$

Ответ нам уже известен

Коэффициент наклона b характеризует силу влияния X на Y:

$$b = \frac{\sum (x_i - \bar{X})(y_i - \bar{Y})}{\sum (x_i - \bar{X})^2}$$

Свободный член a:

$$a = \bar{Y} - b\bar{X}$$

Терминология

Широко используются термины:

- X — независимая переменная;
- Y — зависимая переменная.

Это неудачные термины! Часто приходится изучать зависимость независимых переменных.

Лучше:

- X — предикторы;
- Y — отклик.

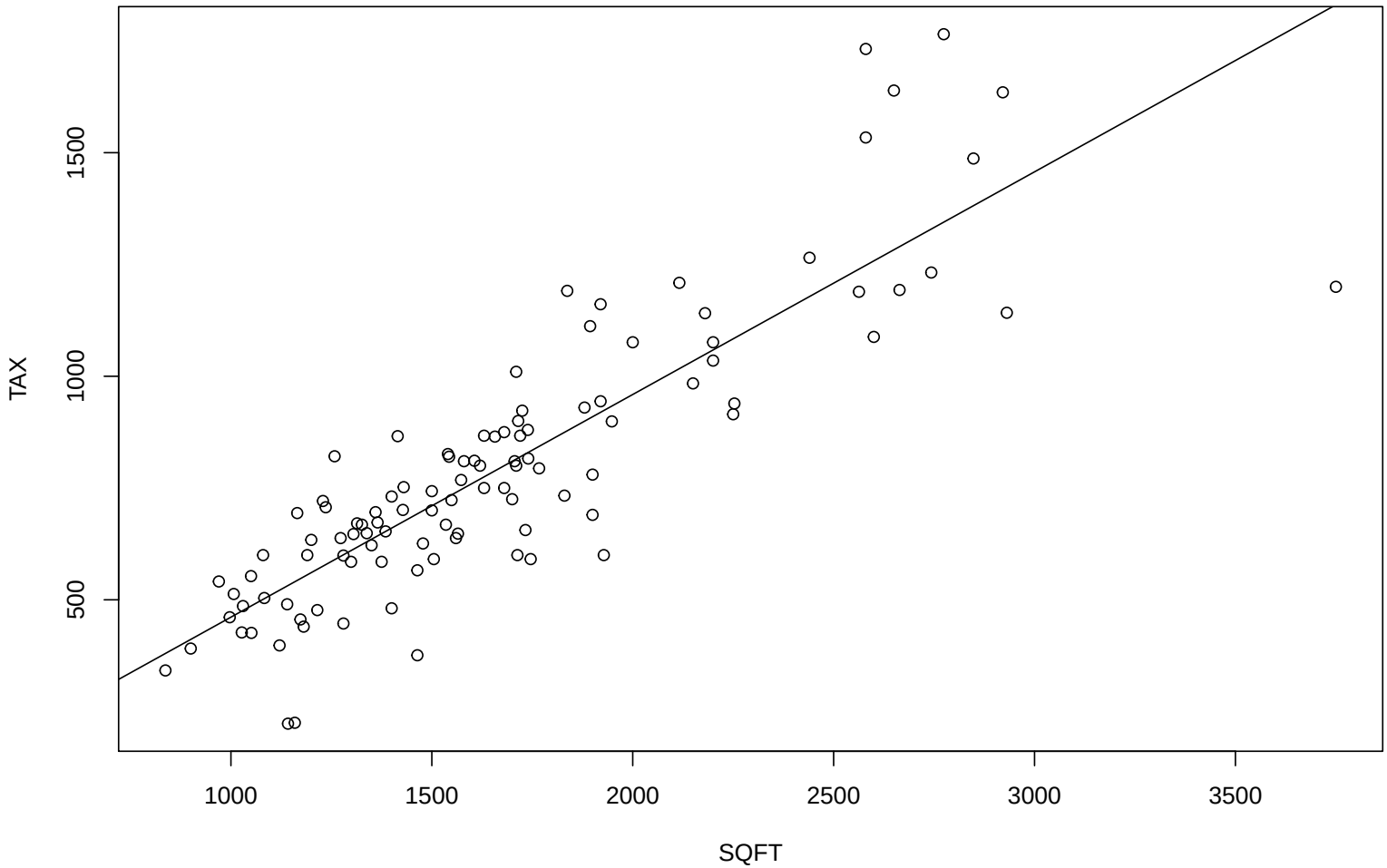
или

- X — входные переменные (входы);
- Y — выходная переменная (выход).

Альбукерке: зависимость налогов (TAX) от площади дома (SQFT)

```
x <- read.table("Albuquerque_Home_Prices_data.txt",
                header=T, na.strings="-9999")
# Чтобы не писать каждый раз 'x'
attach(x)
# Рассмотрим зависимость налогов от площади дома
plot(SQFT, TAX)
# Построим линейную регрессионную модель
reg <- lm(TAX ~ SQFT)
# Добавим к ней линию Y = a + bX
abline(a = reg$coefficients[1], b = reg$coefficients[2])
```

Зависимость налогов (TAX) от площади дома (SQFT)




```
## (Intercept)        SQFT
## -36.7857016   0.4979852
```

Измеряем качество регрессионной модели

Насколько построенная нами модель лучше описывает данные по сравнению с некой базовой характеристикой?

В качестве базовой характеристики выступает среднее арифметическое $\bar{Y}$.

- Измеряем сумму квадратов отклонений для регрессионной модели

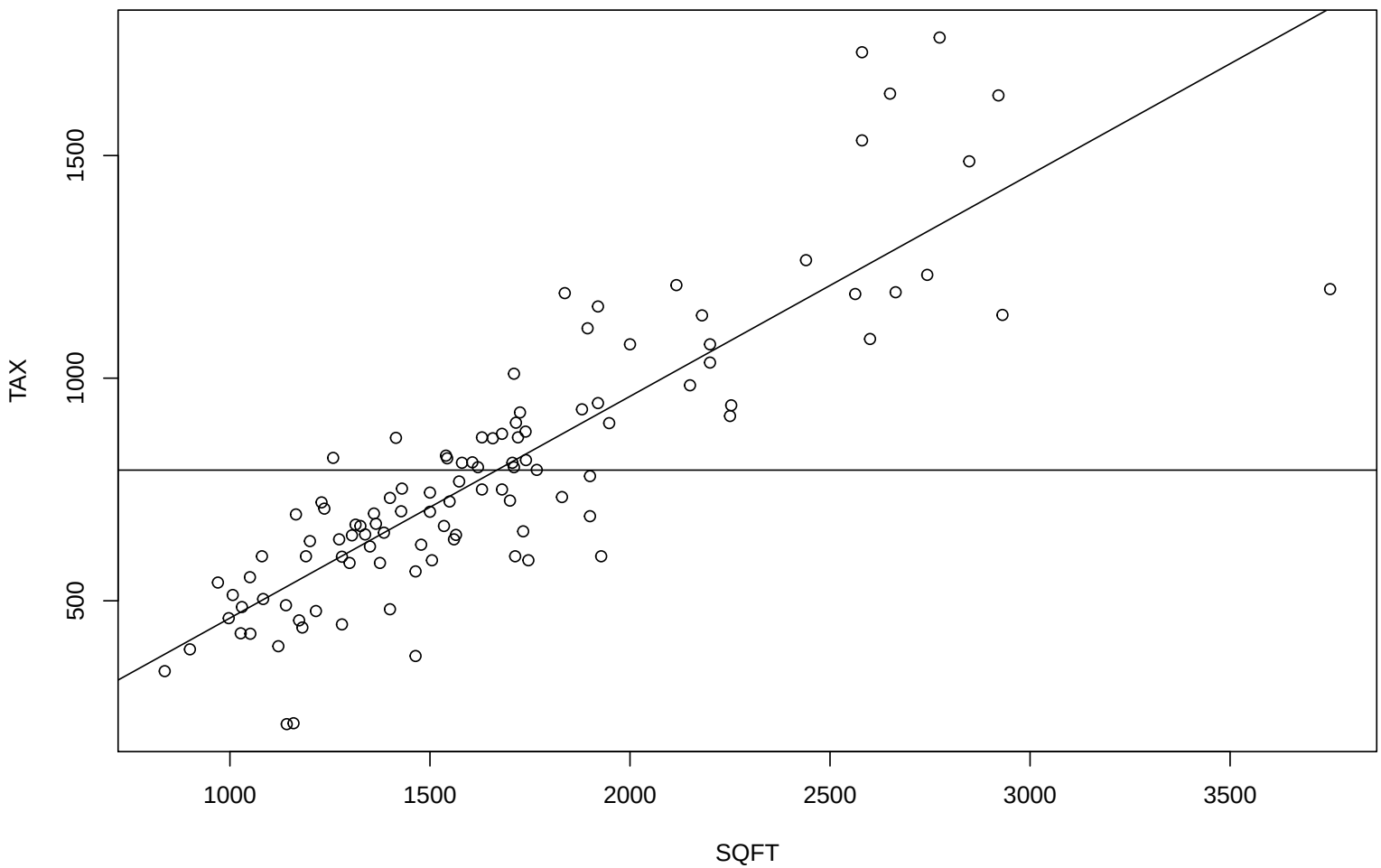
$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - (a + bX_i))^2$$

- Измеряем сумму квадратов отклонений для базовой модели

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

Коэффициент детерминации R^2

$$R^2 = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i - (a + bX_i))^2}{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}, \quad 0 \leq R^2 \leq 1$$



Альбукерке: коэффициент детерминации для зависимость налогов от площади дома

```
summary(reg)$r.squared
```

```
## [1] 0.7371644
```

Итак: наша модель хороша, если она дает большой выигрыш по сравнению с базовой моделью.

В данном случае это так.

Важно!

- Коэффициент детерминации, в отличие от MSE, — величина безразмерная. MSE (Mean Square Error): $(1/n)\sum_{i=1}^{n}(Y_i - (a + bX_i))^2$.
- Коэффициент детерминации работает для зависимостей вида

$$Y = a + b_1X_1 + b_2X_2 + ... + b_nX_n.$$

Интерпретации коэффициента детерминации

$$R^2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 - \frac{1}{n}\sum_{i=1}^{n}(Y_i - (a + bX_i))^2}{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \cdot 100\%$$

- На сколько процентов улучшилась модель по сравнению с базовой.
- Какой процент вариации Y объясняется влиянием всех независимых переменных (предикторов).

Недостатки коэффициента детерминации

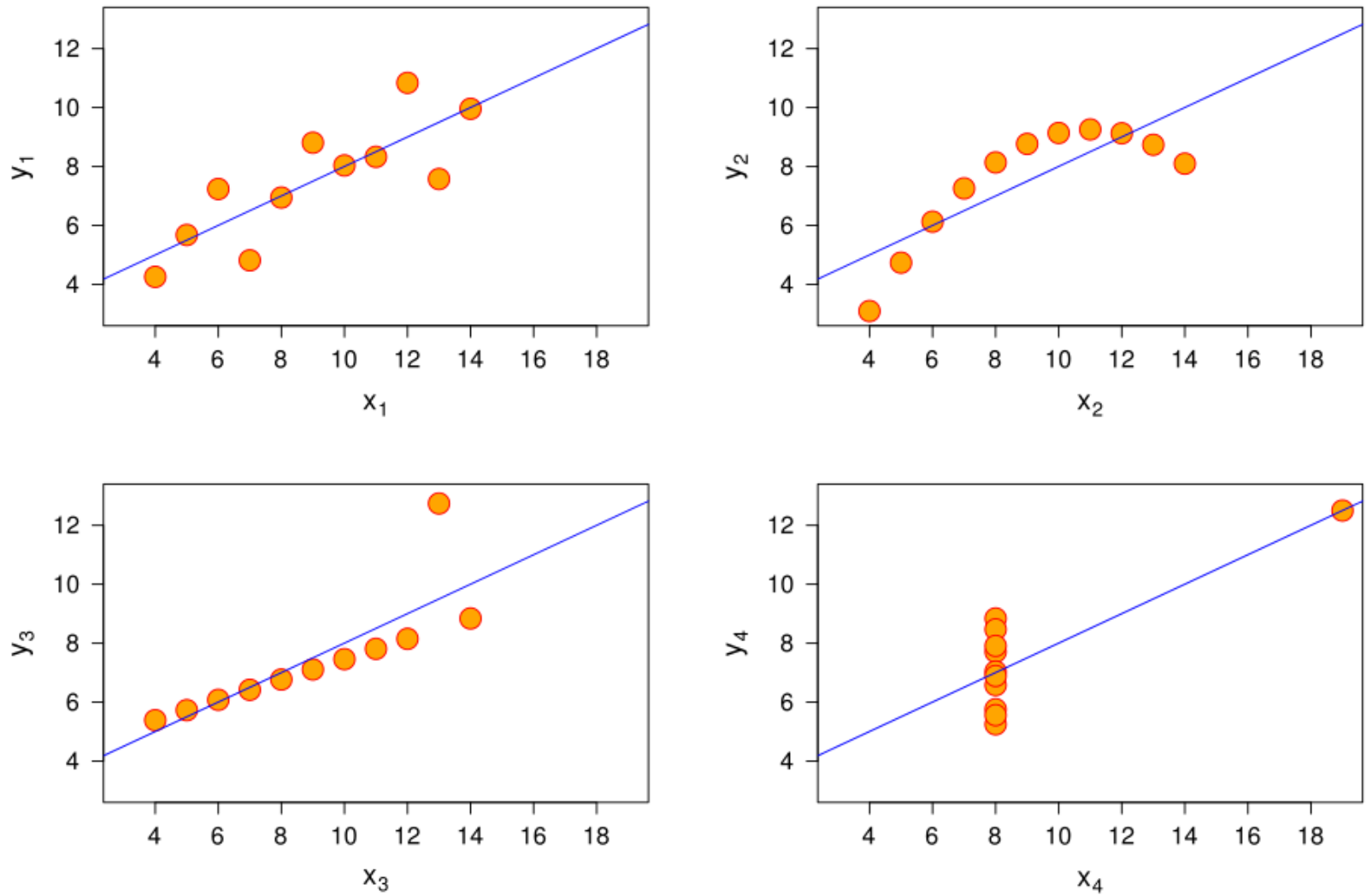


Figure 1: Квартет Анскомба: r везде равен 0.816

Недвижимость в г. Альбукерке, шт. Нью-Мексико, США

Данные (117 наблюдений) являются случайной выборкой из записей о перепродажах домов, совершенных между 15 февраля и 30 апреля 1993. Информация предоставлена Советом риэлтеров (Albuquerque Board of Realtors) Альбукерке.

Переменные:

- PRICE — продажная цена в сотнях долларов;
- SQFT — площадь в квадратных футах;
- AGE — возраст дома (количество лет);
- FEATS — количество дополнительных удобств из 11 возможных: dishwasher, refrigerator, microwave, disposer, washer, intercom, skylight(s), compactor, dryer, handicap fit, cable TV access;
- NE — дом расположен в престижном районе на северо-востоке города (1), или нет (0);
- CUST — тип постройки: был ли дом обычной постройки (0), или нет (1),
- COR — как расположен дом, на углу (1) или нет (0).
- TAX — величина налогов за владение домом (в долларах).

Дадим прогноз цен

Задача

Построить модель, позволяющую по имеющимся параметрам спрогнозировать цену дома.

Гипотеза

Зависимость цены от переменных — линейная.

Порядок решения

1. Построение модели цены
2. Интерпретация коэффициентов
3. Отбор переменных

С чего начать?

```
# Загрузка данных
setwd("week_07/data/")
x <- read.table("Albuquerque_Home_Prices_data.txt",
                    header=T, na.strings="-9999")
summary(x) # Проверка
# Построим модель, зависящую ото всех переменных
itog1 <- lm(PRICE ~ SQFT + AGE + FEATS + NE + CUST +
                COR + TAX, x)
summary(itog1)
```

Таблица x

	PRICE	SQFT	AGE	FEATS	NE	CUST	COR	TAX
1	2050	2650	13	7	1	1	0	1639
2	2080	2600	*NA*	4	1	1	0	1088
3	2150	2664	6	5	1	1	0	1193
4	2150	2921	3	6	1	1	0	1635
5	1999	2580	4	4	1	1	0	1732
6	1900	2580	4	4	1	0	0	1534
7	1800	2774	2	4	1	0	0	1765
8	1560	1920	1	5	1	1	0	1161
9	1450	2150	*NA*	4	1	0	0	*NA*

Figure 2:

Модель, зависящая ото всех переменных

```
Call:
lm(formula = PRICE ~ SQFT + AGE + FEATS + NE + CUST + COR + TAX,
    data = x)

Residuals:
    Min      1Q  Median      3Q     Max
-466.28  -82.29    6.75   78.70  484.84

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  92.74480  101.60704   0.913 0.365137
SQFT          0.35222    0.09575   3.679 0.000515 ***
AGE          -0.56508    2.00253  -0.282 0.778807
FEATS         4.38961   18.55499   0.237 0.813822
NE          -17.38534   47.27462  -0.368 0.714397
CUST        174.94108   53.72371   3.256 0.001887 **
COR         -73.58234   49.13007  -1.498 0.139633
TAX           0.49887    0.15849   3.148 0.002598 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 158.9 on 58 degrees of freedom
  (51 observations deleted due to missingness)
Multiple R-squared:  0.8623,    Adjusted R-squared:  0.8456
F-statistic: 51.86 on 7 and 58 DF,  p-value: < 2.2e-16
```

На что обращать внимание?

1. Множественный коэффициент детерминации `Multiple R-squared`. Если он мал, то нужно искать другую модель. Например, нелинейную.
2. Коэффициенты: столбец `Estimate`. Их можно интерпретировать.
3. Столбец `Pr(>|t|)` содержит результат проверки гипотезы о том, что коэффициент равен 0. Помогает понять, какие переменные нужно рассматривать.

Оставим в модели только переменные с "ненулевыми" коэффициентами

```
itog2 <- lm(PRICE ~ SQFT + CUST + TAX, x)
summary(itog2)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 164.7319                  91 0.004690 **
SQFT          0.1881                  39 0.002743 **
CUST        162.0664    45.1363   3.591 0.000507 ***
TAX           0.7090     0.1014   6.989 2.83e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 164.9 on 103 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared:  0.8207,    Adjusted R-squared:  0.8155
F-statistic: 157.2 on 3 and 103 DF,  p-value: < 2.2e-16
```

Коллинеарность

Допустим, что задана зависимость

$$Z = 6X + 8Y.$$

Одновременно известно, что

$$Y = 2X.$$

В таком случае к коэффициентам при X и Y нельзя относиться серьезно.

Возможно:

$$Z = 11Y,$$
$$Z = 4X + 90Y,$$
$$Z = 22X,$$
$$\ldots$$

Что делать?

1. Можно удалить одну из коллинеарных переменных.
2. В более сложных случаях — факторный анализ. Он преобразует исходные переменные в новые некоррелированные переменные. Новые переменные будут линейными комбинациями исходных. Их меньше. Но их труднее интерпретировать.

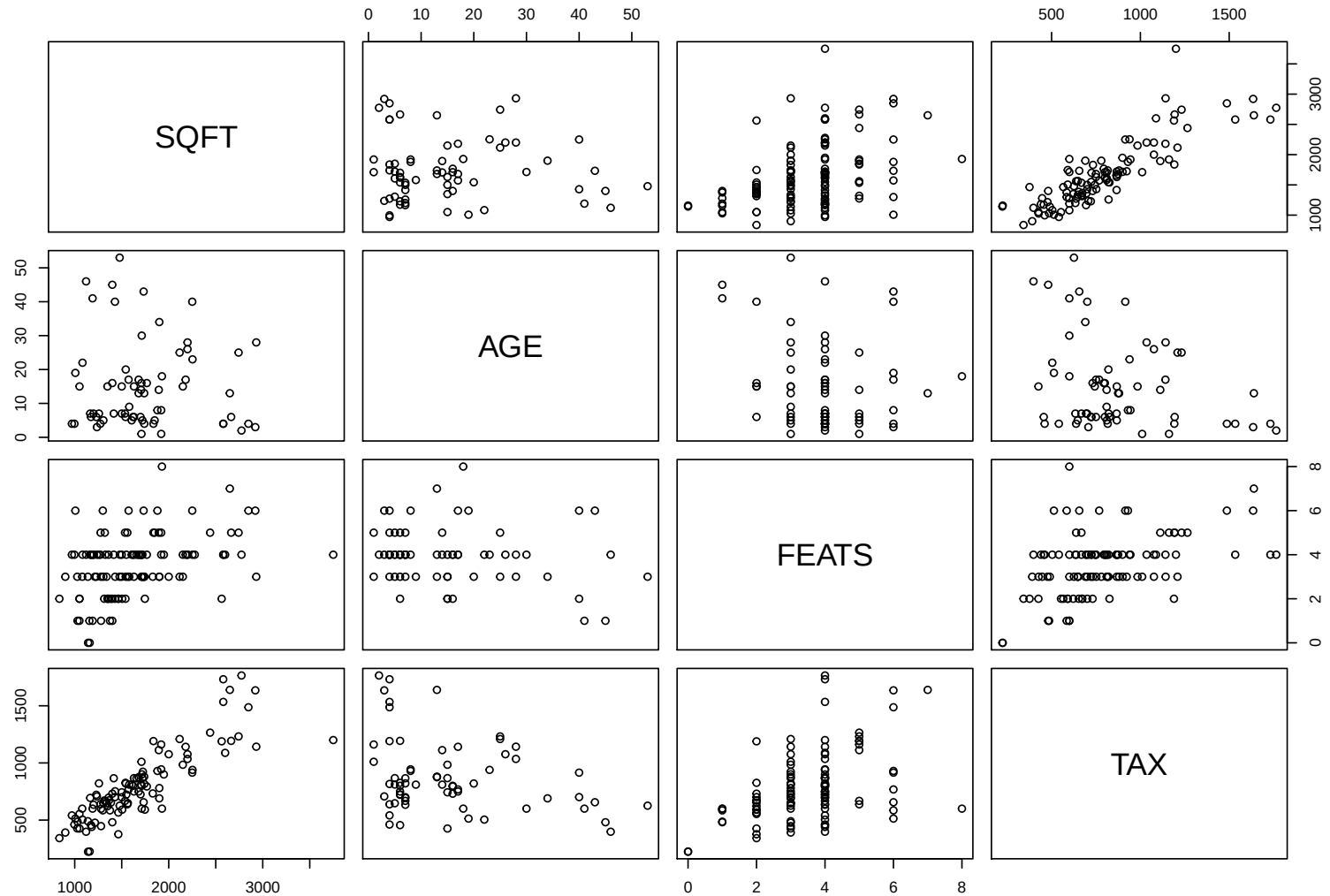
Посмотрим на корреляцию

```
cor(x, use = "complete.obs", method = "pearson")
```

```
##              PRICE        SQFT          AGE      FEATS          NE
## PRICE   1.0000000  0.88394183 -0.166662011  0.3663458  0.28916464
## SQFT    0.8839418  1.00000000 -0.037693593  0.3573967  0.36254721
## AGE    -0.1666620 -0.03769359  1.000000000 -0.1834804  0.21642412
## FEATS   0.3663458  0.35739666 -0.183480404  1.0000000  0.30963494
## NE      0.2891646  0.36254721  0.216424115  0.3096349  1.00000000
## CUST    0.5821164  0.49187084  0.008517219  0.3121949  0.15018688
## COR    -0.1875856 -0.07850150  0.162728128 -0.2491235 -0.02371519
## TAX     0.8775270  0.87524956 -0.291842247  0.3039824  0.30240397
##                CUST         COR        TAX
## PRICE   0.582116383 -0.18758563  0.8775270
## SQFT    0.491870845 -0.07850150  0.8752496
## AGE     0.008517219  0.16272813 -0.2918422
## FEATS   0.312194901 -0.24912353  0.3039824
## NE      0.150186875 -0.02371519  0.3024040
## CUST    1.000000000 -0.05368755  0.4370276
## COR    -0.053687549  1.00000000 -0.1531738
## TAX     0.437027555 -0.15317383  1.0000000
```

Или с помощью матрицы диаграмм рассеяния

```
pairs(~ SQFT + AGE + FEATS + TAX, data = x)
```



Исключаем налоги из модели

```
itog3 <- lm(PRICE ~ SQFT + AGE + FEATS + NE + CUST + COR, x)
summary(itog3)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  83.14037  102.73572   0.809  0.42151
SQFT          0.63719    0.05119  12.448  < 2e-16 ***
AGE          -3.72095    1.80540  -2.061  0.04357 *
FEATS         3.25714   18.93246   0.172  0.86398
NE          -14.32888   49.23057  -0.291  0.77200
CUST        148.47950   54.40590   2.729  0.00829 **
COR         -83.39862   51.26812  -1.627  0.10895
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 174.4 on 61 degrees of freedom
  (49 observations deleted due to missingness)
Multiple R-squared:  0.8295,    Adjusted R-squared:  0.8128
F-statistic: 49.47 on 6 and 61 DF,  p-value: < 2.2e-16
```

Делаем еще одну попытка: по новым “ненулевым” коэффициентам

```
itog4 <- lm(PRICE ~ SQFT + AGE + CUST, x)
summary(itog4)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  72.48077   84.52629   0.857  0.39437
SQFT          0.63914    0.04712  13.565  < 2e-16 ***
AGE          -4.28913    1.68111  -2.551  0.01313 *
CUST        149.31462   53.72717   2.779  0.00715 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 174.2 on 64 degrees of freedom
  (49 observations deleted due to missingness)
Multiple R-squared:  0.8216,    Adjusted R-squared:  0.8132
```

Идеологическое замечание

Разработка модели — процесс последовательных приближений к цели.