Исследование и улучшение модели

При выводе явной формулы для вектора Θ коэффициентов линейной регрессионной модели:

$$\hat{\Theta} = \left(X^T X\right)^{-1} X^T Y.$$

мы (неявно) предполагали, что матрица $\mathbb{X}^{\mathrm{T}}\mathbb{X}$ обратима, т.е. её столбцы линейно независимы. Однако на практике может оказаться, что *факторы* являются *линейно зависимыми*. Заметим, что вероятность того, что среди факторов есть линейно зависимые тем больше, чем большее количество факторов включается в модель. Хотя на первый взгляд, может показаться, что чем больше в модели факторов (предикторов), тем точнее эта модель, на само деле это не так. Линейная зависимость факторов носит название *мультиколлинеарности*.

Мультиколлинеарность - это (явная или стохастическая) линейная зависимость двух или нескольких факторов (объясняющих переменных).

Следствием мультиколлинеарности является вырожденность матрицы $\mathbb{X}^{\mathrm{T}}\mathbb{X}$ (т.е. равенство: $\det(\mathbb{X}^{\mathrm{T}}\mathbb{X}) = 0$), где

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1m} \\ 1 & x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{im} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nj} & \dots & x_{nm} \end{pmatrix}$$

а значит, невозможность найти решение (т.е. вектор коэффициентов регрессионной модели) по формуле:

$$\hat{\Theta} = \left(X^T X\right)^{-1} X^T Y.$$

Здесь Y – вектор (известных) значений переменной отклика.

При применении итерационных методов решения задачи, которая формулируется в Методе наименьших квадратов, построенная итерационная последовательность может не иметь предельных точек.

В случае, когда $\det(X^T X)$ близок к 0 (зависимость между факторами стохастическая), следствиями мультиколлинеарности являются следующие обстоятельства:

1) Оценки параметров становятся ненадежными (обнаруживают большие ошибки), причём по мере увеличения числа наблюдений (размера обучающей выборки) оценки сильно изменяются (даже могут поменять знак), что делает прогнозирование на их основе невозможным.
2) Параметры линейной регрессии теряют экономический (или иной) смысл, они не поддаются интерпретации.
3) Становится невозможным определить изолированное влияние каждого фактора на значение переменной отклика.

Рассмотрим некоторые *характеристики качества регрессионной модели*.

Заметим, что при выводе сводной информации о модели наряду с её коэффициентами выводятся ещё некоторые характеристики, в частности, величины :

- ***Коэффициент детерминации*** (**`Multiple R-squared`**) и
- ***Исправленный коэффициент детерминации*** (**`Adjused R-squared)`**

summary(mymodel)	

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.6560 -1.7738 -0.7540  0.6493  9.3139

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
totsquare   0.0519874  0.1177009   0.442    0.665
livesquare  0.1386667  0.1465748   0.946    0.360
floor      -0.0089039  0.2659979  -0.033    0.974
height     -0.2524347  0.2125690  -1.188    0.255
distcenter  0.1461910  0.1682269   0.869    0.399
distmetro  -0.0002076  0.0009206  -0.225    0.825

Residual standard error: 3.273 on 14 degrees of freedom
Multiple R-squared:  0.9395,    Adjusted R-squared:  0.9136
F-statistic: 36.25 on 6 and 14 DF,  p-value: 9.547e-08
```

Коэффициент детерминации (**`Multiple R-squared`**):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \quad (1)$$

где $\hat{y}_i$ – спрогнозированное значение переменной отклика для -ой строки обучающей выборки; y_i – истинное значение переменной отклика (в i-ой строке обучающей выборки); $\bar{y}$ – среднее значение переменной отклика.

Очевидно, что чем ***коэффициент детерминации*** ближе к 1 (т.е., чем числитель ближе к 0), тем точнее прогноз, полученный с помощью линейной регрессионной модели. (Действительно, величина в числителе дроби характеризует различие между прогнозом и фактическими значениями переменной отклика, а значит, чем эта величина меньше, тем точнее прогноз; знаменатель дроби – есть дисперсия переменной отклика (с точностью до множителя)). В приведённом выше примере $R^2 \approx 0{,}94$, что означает достаточно высокую точность прогноза.

Вместе с тем, с увеличением числа предикторов коэффициент детерминации возрастает, а значит, необоснованно «завышает оценку качества» прогноза, полученного с помощью линейной регрессионной модели. Для того, чтобы получить более адекватную оценку качества прогноза, используют так называемый «исправленный» коэффициент детерминации (**`Adjused R-squared`**):

$$R^2_{adj} = 1 - (1 - R^2)\,\frac{n-1}{n-m-1}, \qquad (2)$$

(здесь, как и ранее, n – число наблюдений (число строк в обучающей выборке), m – число факторов (предикторов) (число в обучающей выборке минус 1). Введённый в (2) коэффициент представляет собой штраф за большое число предикторов.

Подробнее о характеристиках линейной регрессионной модели см. в статье «Оценка результатов линейной регрессии»

Вычислим попарные коэффициенты корреляции между предикторами (а также между каждым из предикторов и переменной отклика) для рассмотренного выше примера:

```
myData = read.csv("myflats.txt", sep = '\t',header=TRUE)
myData
rm = cor(myData)
rm
```

	totsquare	livesquare	floor	height	distcenter	distmetro	price
totsquare	1.000000000	0.9510794	-0.3054832	0.003209003	-0.18622246	-0.5460350	0.78791852
livesquare	0.951079355	1.0000000	-0.1914513	0.105225446	-0.12384035	-0.5127576	0.77744783
floor	-0.305483161	-0.1914513	1.0000000	0.514470424	0.10466631	0.5249853	-0.25233447
height	0.003209003	0.1052254	0.5144704	1.000000000	0.04170783	0.2044663	-0.11189253
distcenter	-0.186222455	-0.1238404	0.1046663	0.041707831	1.00000000	0.2866833	0.05890125
distmetro	-0.546034979	-0.5127576	0.5249853	0.204466321	0.28668328	1.0000000	-0.39274420
price	0.787918520	0.7774478	-0.2523345	-0.111892529	0.05890125	-0.3927442	1.00000000

Естественно, диагональные элементы полученной корреляционной матрицы равны 1. Мы видим, что имеется сильная положительная корреляция между общей площадью квартиры (переменная **totsquare**) и жилой площадью (**livesquare**), поэтому имеет смысл исключить один из этих признаков (хотя, конечно, зависимость между ними нелинейная). Заметим, что влияние остальных признаков на цену квартиру существенно меньше, чем влияние фактора площади, причём факторы: «этаж», «количество этажей в доме» и «расстояние до метро» играют на понижение цены, остальные факторы – на повышение.

```
# Исключим 2-й фактор (livesquare)
myData = myData[-2]

# Пересчитаем корреляционную матрицу
rm = cor(myData)
rm
```

```
             totsquare      floor       height  distcenter  distmetro       price
totsquare   1.000000000 -0.3054832  0.003209003 -0.18622246 -0.5460350  0.78791852
floor      -0.305483161  1.0000000  0.514470424  0.10466631  0.5249853 -0.25233447
height      0.003209003  0.5144704  1.000000000  0.04170783  0.2044663 -0.11189253
distcenter -0.186222455  0.1046663  0.041707831  1.00000000  0.2866833  0.05890125
distmetro  -0.546034979  0.5249853  0.204466321  0.28668328  1.0000000 -0.39274420
price       0.787918520 -0.2523345 -0.111892529  0.05890125 -0.3927442  1.00000000
```

```
# Получим новую регрессионную модель (без livesquare)
mymodel = lm(price ~ -1+ totsquare + floor + height +
distcenter + distmetro, data = myData)
summary(mymodel)
```

```
    Min      1Q  Median      3Q     Max
-2.9597 -1.9119 -0.9866  0.8043  9.7132

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
totsquare   0.1609714  0.0240531   6.692  7.2e-06 ***
floor       0.0488326  0.2579945   0.189    0.852
height     -0.2233471  0.2095962  -1.066    0.303
distcenter  0.1668685  0.1662162   1.004    0.331
distmetro  -0.0005034  0.0008628  -0.583    0.568
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.262 on 15 degrees of freedom
Multiple R-squared:  0.9357,    Adjusted R-squared:  0.9142
F-statistic: 43.63 on 5 and 15 DF,  p-value: 2.062e-08
```

Обратите внимание, как изменились вероятности превышения (по модулю) абсолютного значения критерия Стьюдента! Кроме того, мы наблюдаем улучшение (хотя и незначительное) исправленного коэффициента детерминации, что говорит об улучшении качества модели.