Линейный регрессионный анализ

Примеры

Цены на женские кольца с бриллиантами

Построить модель цены в зависимости от веса бриллианта

Наблюдения `diamond.dat` — цены на женские золотые кольца с бриллиантами (2 столбец) и вес бриллиантов в каратах (1 столбец). Все кольца сделаны из золота пробы 20 каратов, на каждом один бриллиант (1 карат = 1/24 часть чистого золота, то есть чистое золото имеет пробу 24 карата).

Варианты моделей

$$price = a + b * weight,$$
$$price = a + b * weight^2,$$
$$price = a + b * weight + c * weight^2.$$

Какая из моделей лучше?

Коэффициент детерминации

```
# Простейшая модель
itog1 <- lm(PRICE ~ WEIGHT, x.1)
# Сделаем так как предлагают ювелиры
itog2 <- lm(PRICE ~ WEIGHT2, x.1)
# Модель, зависящая ото всех переменных
itog3 <- lm(PRICE ~ WEIGHT + WEIGHT2, x.1)
```

Модель	R^2
1	0.9783
2	0.9703
3	0.9789

Так выбрать модель не получится.

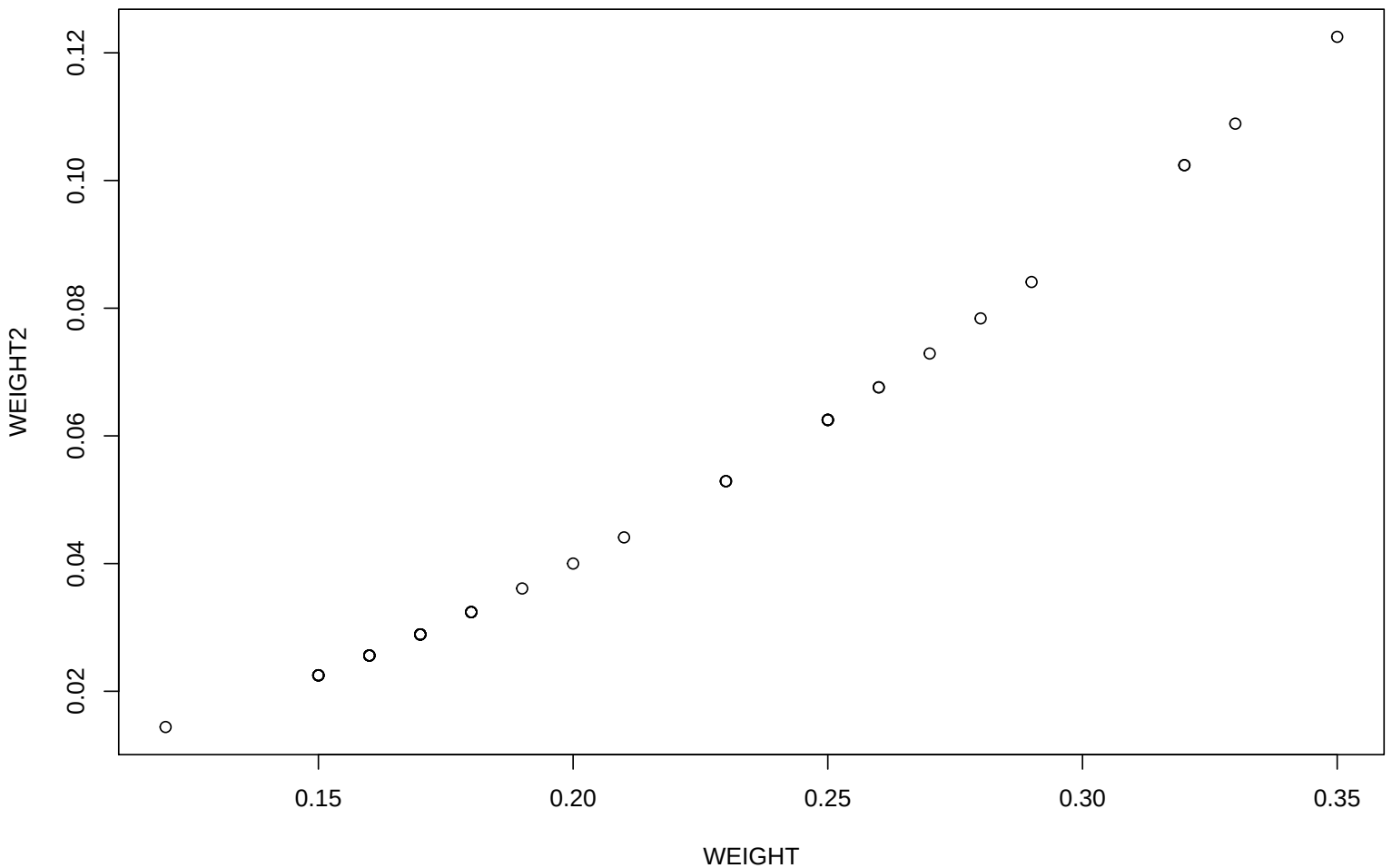
Содержательная интерпретация

```
# 1)
itog1 <- lm(PRICE ~ WEIGHT, x.1)
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept)  -259.63      17.32  -14.99   <2e-16 ***
# WEIGHT       3721.02      81.79   45.50   <2e-16 ***
# 2)
itog2 <- lm(PRICE ~ WEIGHT2, x.1)
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept)   141.66      10.69   13.25   <2e-16 ***
# WEIGHT2      7993.11     206.15   38.77   <2e-16 ***
# 3)
itog3 <- lm(PRICE ~ WEIGHT + WEIGHT2, x.1)
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept)  -174.13      74.24  -2.346   0.0235 *
# WEIGHT       2920.13     681.30   4.286 9.47e-05 ***
# WEIGHT2      1739.90    1469.47   1.184   0.2426
```

1. Модель проста, но цена кольца без бриллианта -259.63 сингапурских доллара, что не логично.
2. Здесь свободный член $a > 0$, что логично.
3. За кольцо без бриллианта снова приплачивают. Кроме того получается, что коэффициент при квадрате веса можно считать равным нулю. Хотя в модели 2 вес в квадрате был информативной переменной.

Противоречие: WEIGHT2 одновременно информативная и неинформативная переменная

Безумная гипотеза: что если WEIGHT и WEIGHT2 линейно зависимы?



Вывод

В первой и третьей моделях значение свободного члена нелогично.

В третьей модели, благодаря коллинеарности WEIGHT и WEIGHT2, можно удалить переменную WEIGHT.

Таким образом, следует предпочесть вторую модель.

Дополнительная информация

- Multicollinearity - Wikipedia (https://en.wikipedia.org/wiki/Multicollinearity) — о диагностике и лекарствах от коллинеарности и мультиколлинеарности.