Статистика и анализ данных в R

Лекция 8. Линейные модели. Часть 2

(5.11.2022)

Даниил Литвинов

Что мы уже знаем?

Вопросы

$y = w_0 + w_1 x_1 + \varepsilon$

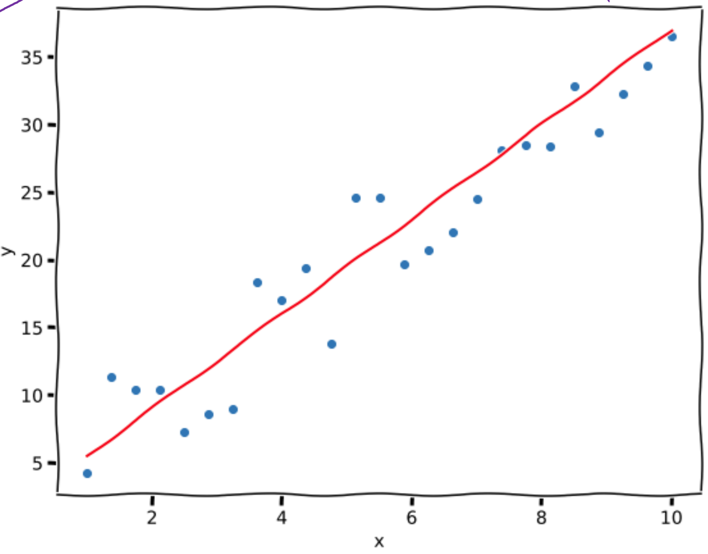
y - знаем

$\hat{y}$ - предсказани

количеств

SSE / MSE $= \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n}$

MAE $= \frac{\sum |y_i-\hat{y}_i|}{n}$



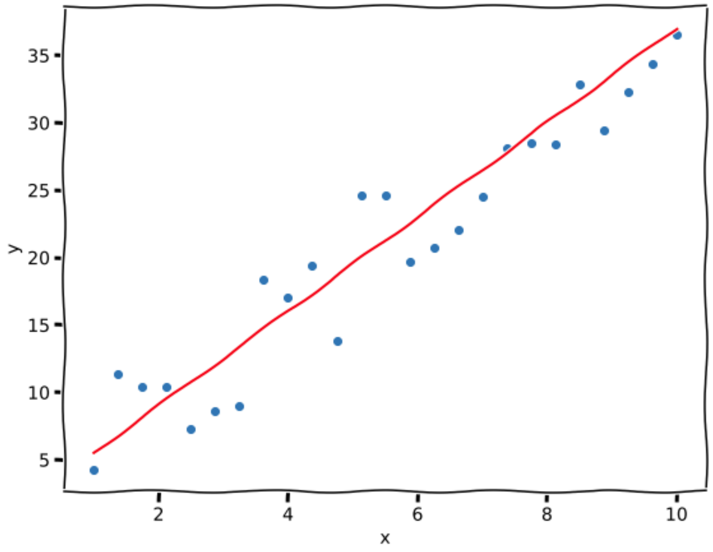
1. Какой формулой задается линейная регрессия?
2. Какие переменные мы можем предсказывать при помощи ЛР?
3. Что мы используем для оценки качества модели?
4. Какие задачи решают ЛМ?
5. А если у нас квадратичная зависимость?
6. Что такое МНК? $\sum_{i}^{n}(y_i-\hat{y}_i)^2 \to \min$ | $w=(X^TX)^{-1}X^Ty$
7. Что такое bias trick? Xw $(Xw+w_0)$
8. Что такое скалярное произведение? $\vec{a}\cdot\vec{b} = \sum a_i b_i$
9. Что такое градиент? $\nabla Loss(w) = \left[\frac{\partial Loss}{\partial w_i}\right]$
10. Как интерпретировать коэффициенты в множественной ЛР?

Что это такое ЛМ?

x — баллы за экзамен по английскому 1

y — баллы за экзамен по английскому 2

x	y
1	5
3	11
9	35
10	33



Что это такое ЛМ?

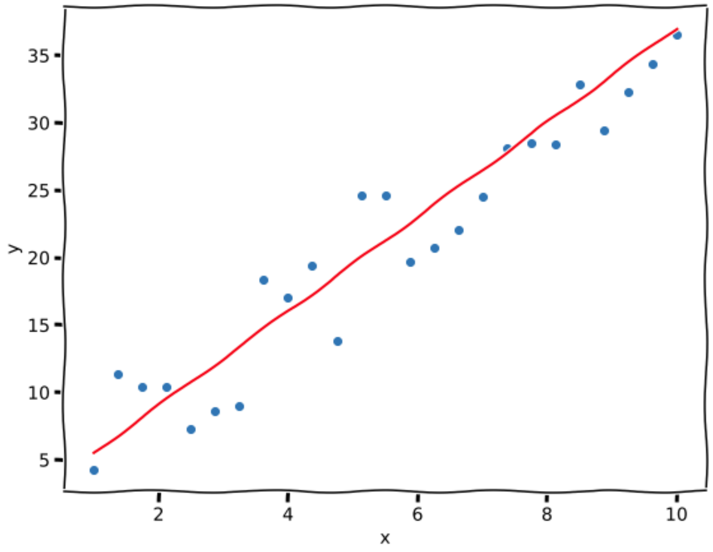
свободный член
(intercept)

случайная ошибка
(error)

$$y = w_0 + w_1 x + \epsilon$$

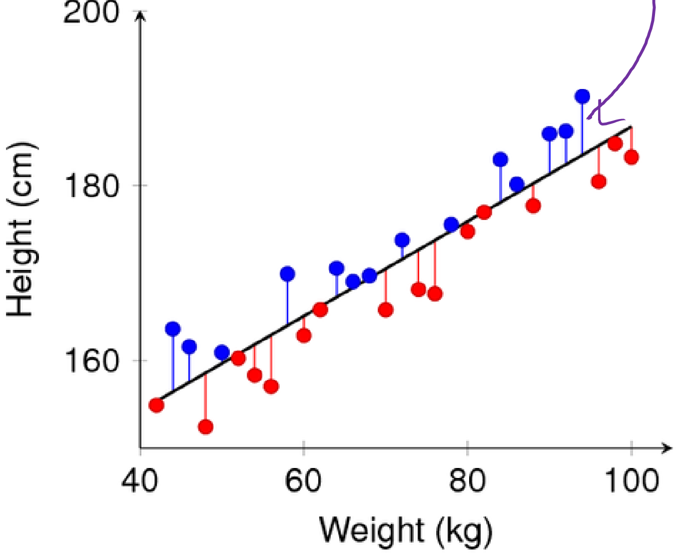
целевая переменная
(aka target)

независимая переменная
(slope, predictor)



Остатки модели

$\varepsilon \sim N(0, 6)$



Формулы

- $y = w_0 + w_1 x + \epsilon$
- $y = Xw$

w_0

$$\frac{dLoss}{dw} = \nabla Loss = 2X^T(Xw - y)$$

$$\underline{Loss = (y - Xw)^T(y - Xw)}$$

$$\boxed{w = (X^T X)^{-1} X^T y}$$

Как проверить, что модель хорошая? (метрики)

$\hat{y} = w_0 \xrightarrow{\text{МНК}} w_0 = \bar{y}$

1) $$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

ошибка модели

ошибка модели $\hat{y} = \bar{y}$

$\bar{y} = \frac{\sum y_i}{n}$

$R^2 : (-\infty; 1]$

$H_0 : R^2 = 0$

$H_1 : R^2 \neq 0$

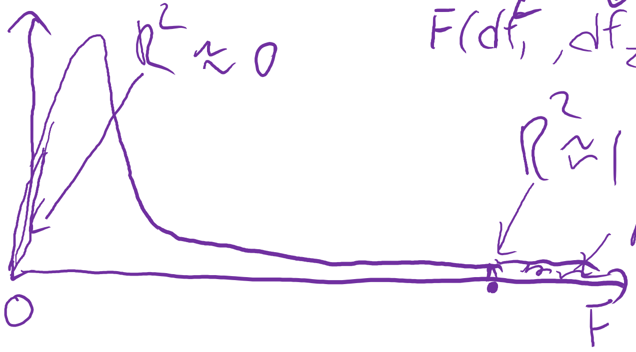
2) $$R^2_{adj} = 1 - (1 - R^2)\frac{(n-1)}{(n-k)}$$

число набл.

число признаков в модели

3) $$F = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}$$

$F(df_1, df_2)$



p-value для коэффициентов

```
> summary(m)

Call:
lm(formula = y ~ u + v + w)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3965 -0.9472 -0.4708  1.3730  3.1283

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.4222     1.4036   1.013  0.32029
u             1.0359     0.2811   3.685  0.00106 **
v             0.9217     0.3787   2.434  0.02211 *
w             0.7261     0.3652   1.988  0.05744 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.625 on 26 degrees of freedom
Multiple R-squared: 0.4981,     Adjusted R-squared: 0.4402
F-statistic: 8.603 on 3 and 26 DF,  p-value: 0.0003915
```

$1.0359 = w_1$

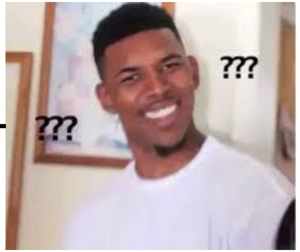
X, μ

$$t = \frac{\bar{X} - \mu}{\frac{\sigma(X)}{\sqrt{n}}}$$

$H_0: \bar{X} = \mu$

$H_1: \bar{X} \neq \mu$

$$t = \frac{1.0359 - 0}{\frac{\sigma(w_1)}{\cancel{\sqrt{n}}}}$$



$H_0: w_1 = 0$

$H_1: w_1 \neq 0$

$w_1 = 10$

$\sigma(w_1) = 100$

p-value для коэффициентов

$D(y_i - \hat{y}_i)$

$$\boxed{w = (X^T X)^{-1} X^T y}$$

$$\boxed{\hat{\sigma}^2(w_i) = \frac{s^2}{(n-k)\hat{\sigma}^2(X_i)} \cdot \frac{1}{1-R_i^2}}$$

$$t = \frac{w_i - \overset{0}{\overset{\shortparallel}{w_{i\_real}}}}{\hat{\sigma}(w_i)}$$

t-распределение с (n - k) степенями свободы

- ✓ $\hat{\sigma}^2(w_i)$ – дисперсия i – го коэффициента
- ✓ s^2 – дисперсия остатков модели
- ✓ n – число наблюдений
- ✓ k – число параметров модели
- ✓ $\hat{\sigma}^2(X_i)$ – дисперсия i – го признака
- ✓ R_i^2 – R^2 модели, где мы предсказываем признак i по всем остальным переменным

$\hat{y} = w_0 + w_1 x_1 + w_2 x_2 + \dots + x_n \;\Big|\; x_1 \sim x_2 + \dots x_n$

Теперь мы знаем все

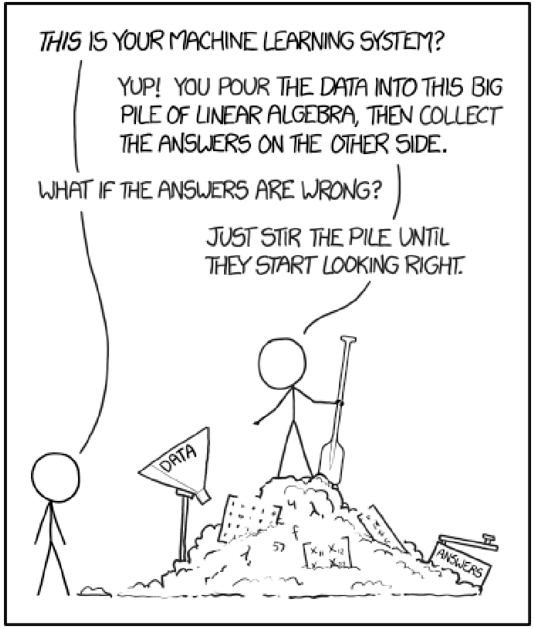
```
> summary(m)

Call:
lm(formula = y ~ u + v + w)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3965 -0.9472 -0.4708  1.3730  3.1283

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.4222     1.4036   1.013  0.32029
u             1.0359     0.2811   3.685  0.00106 **
v             0.9217     0.3787   2.434  0.02211 *
w             0.7261     0.3652   1.988  0.05744 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.625 on 26 degrees of freedom
Multiple R-squared: 0.4981,     Adjusted R-squared: 0.4402
F-statistic: 8.603 on 3 and 26 DF,  p-value: 0.0003915
```

Отдых

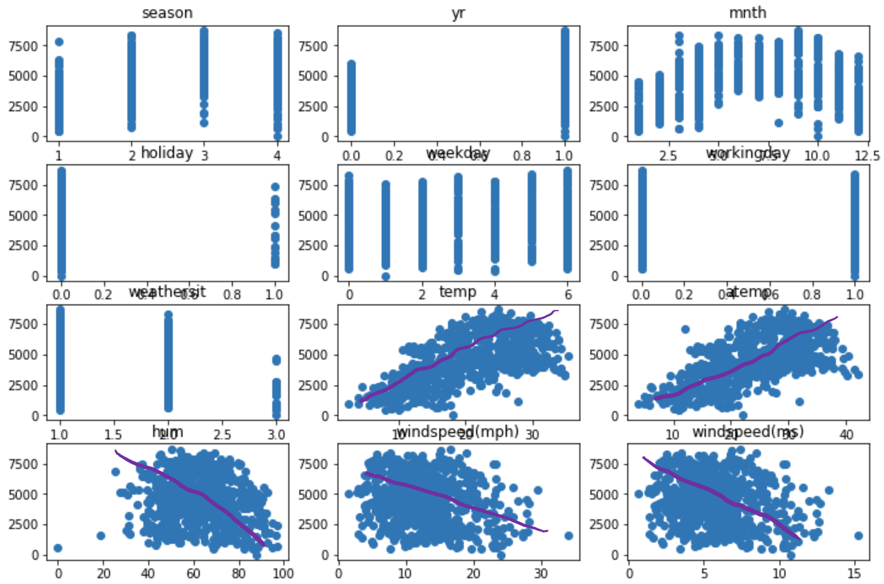
Условия применимости ЛМ

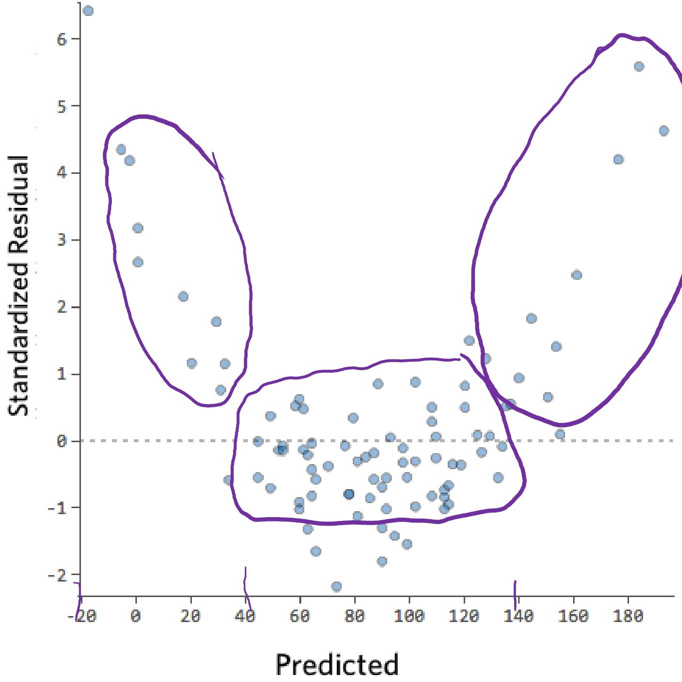
1. линейность зависимости
2. отсутствие влиятельных наблюдений
3. независимость наблюдений
4. нормальное распределение остатков
5. постоянство дисперсии остатков (a.k.a отсутствие гетероскедастичности)
6. отсутствие коллинеарности предикторов

$$y = w_0 + w_1 x + \epsilon$$



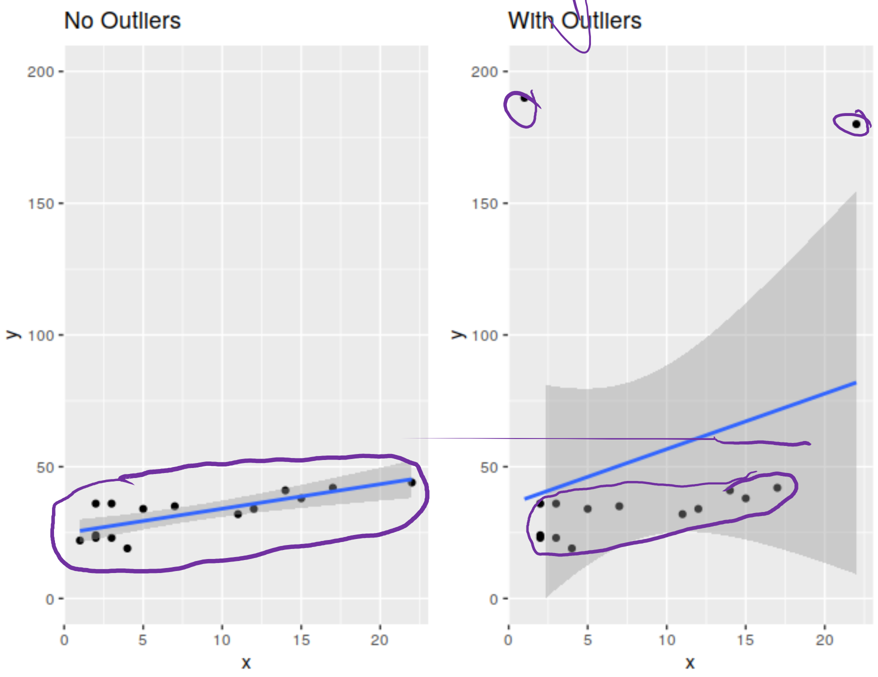
Линейность зависимости





Влиятельные наблюдения (выбросы)

1. Берем k подвыборок наших данных
2. Для каждой строим свою модель
3. В итоге берем для каждого коэффициента медианный



Выбросы (расстояния Кука)

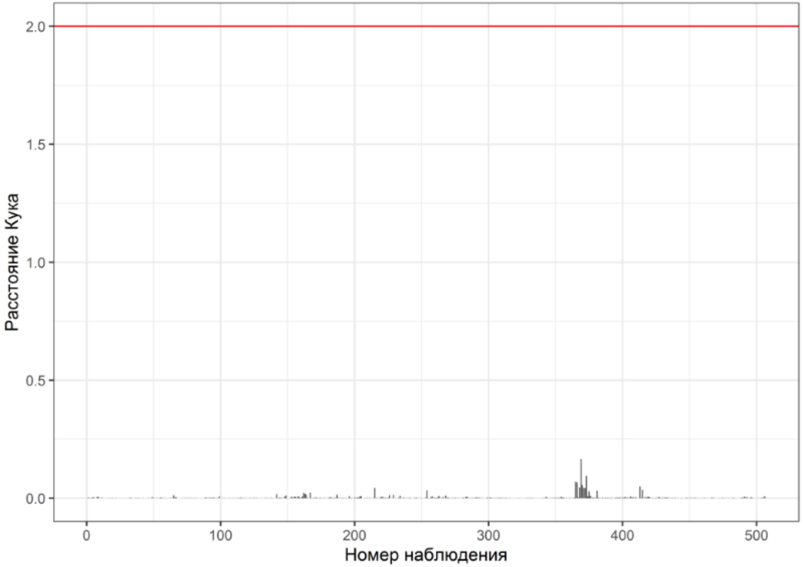
(n-1)

$$D_i = \frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}_{j(i)})^2 / k}{\sum_{j=1}^{n} (y_j - \hat{y}_j)^2 / n}$$

$\hat{y}_{j(i)}$ — значение, предсказанное моделью, построенной без учета i — го наблюдения

1. $D_i > \frac{4}{n}$
2. $D_i > 1$

thresholds



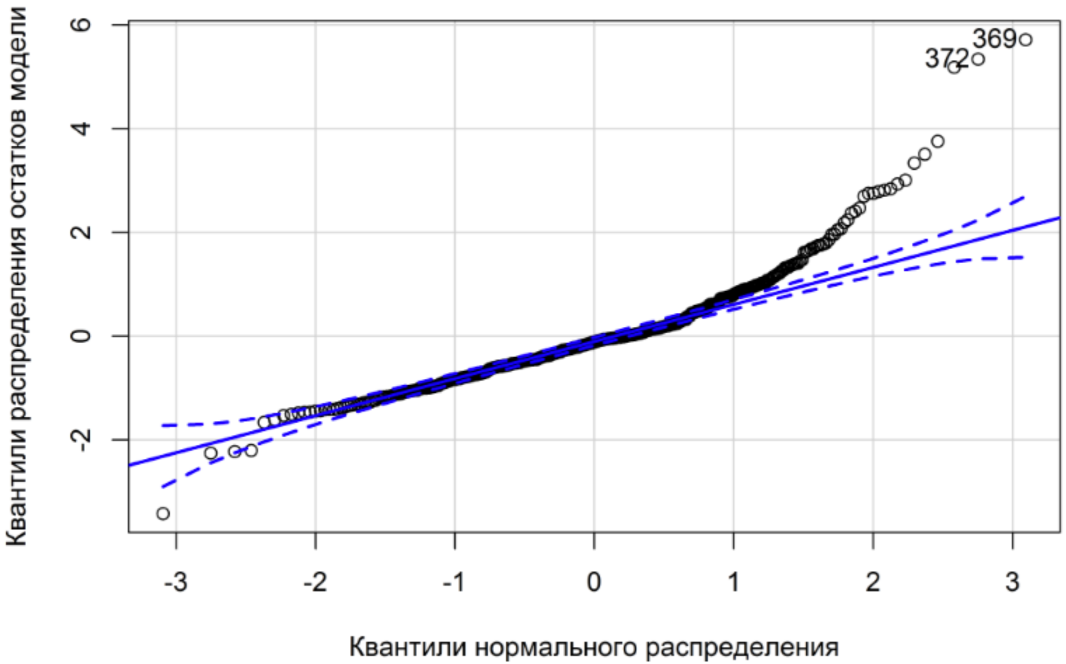
Независимость наблюдений

В линейной регрессии наблюдения должны быть независимы.

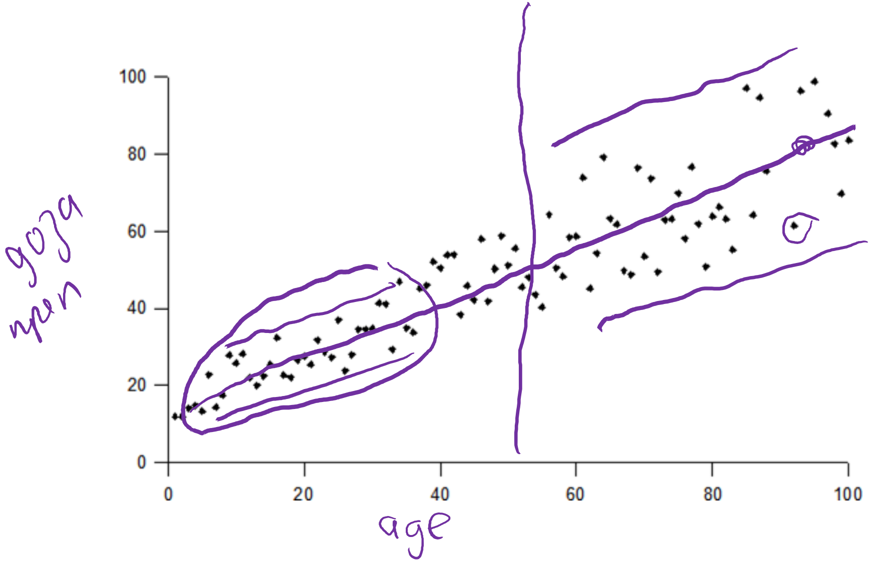
$$X^{-1} = \frac{A}{|X|}$$

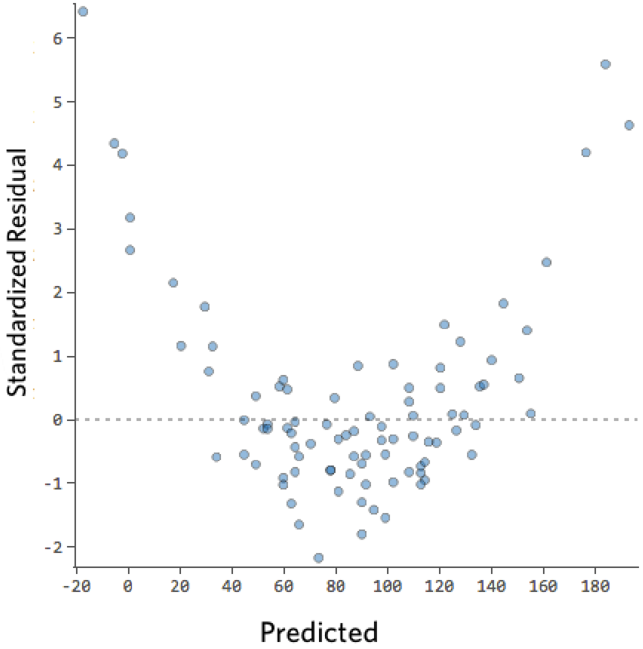
$$B = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 5.9 \\ -1 & 0 & 2 \end{pmatrix} \quad |B| = 0$$

Распределение остатков



Постоянство дисперсии





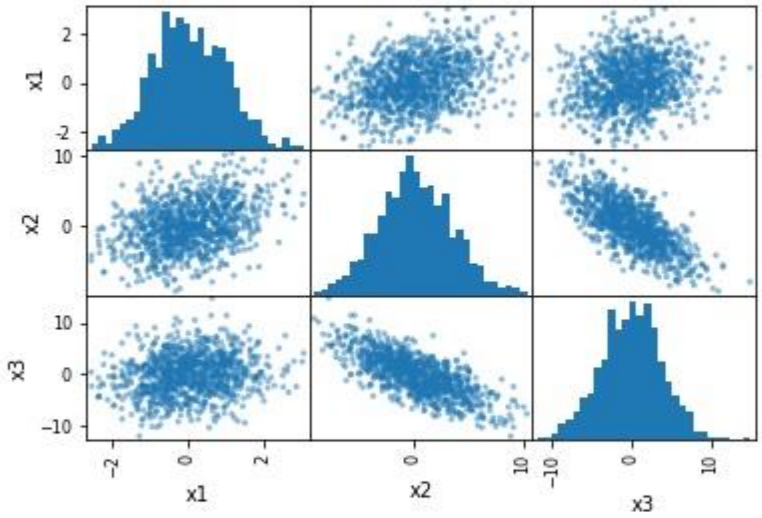
Мультиколлинеарность

$$y = w_0 + w_1 x_1 + w_2 x_2$$

$$x_1 = \frac{1}{3} x_2$$

$$y = w_0 + w_1 x_1 + 3 w_2 x_1$$

$$y = w_0 + x_1 (w_1 + 3 w_2)$$



$w_1 + 3w_2 = 100$

Следовательно, получим оценку (w1 + 3w2), а из нее не очень понятно, чему равны коэффициенты в отдельности

Мультиколлинеарность

$$\hat{\sigma}^2(w_i) = \frac{s^2}{(n-k)\hat{\sigma}^2(X_i)} \cdot \boxed{\frac{1}{1-R_i^2}}$$

$\hat{\sigma}^2(w_i)$ – дисперсия i – го коэффициента

s^2 – дисперсия остатков модели

n – число наблюдений

k – число параметров модели

$\hat{\sigma}^2(X_i)$ – дисперсия i – го признака

R_i^2 – R^2 модели, где мы предсказываем признак i по всем остальным переменным

Variance inflation factor (VIF)

$$\boxed{VIF_i = \frac{1}{1-R_i^2}}$$

Квадратный корень из VIF показывает насколько возрастает стандартная ошибка измерения коэффициента в сравнении со случаем, если бы предиктор не коррелировал с другими.

$VIF > 5$ – не очень хорошо

Отбор по VIF

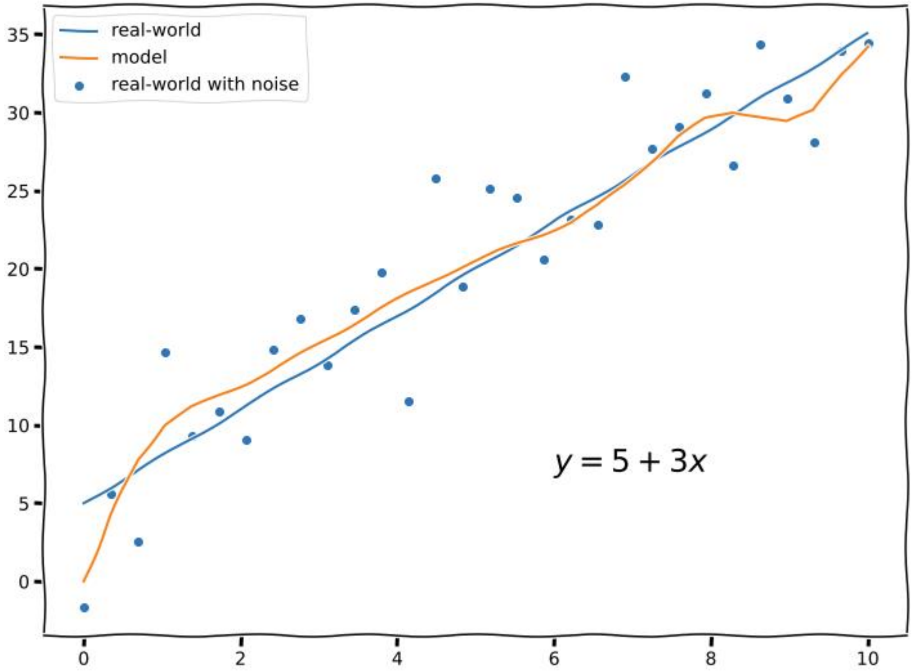
```
| vif(mod_1) # исходная модель со всеми возможными предикторами
| mod_2 <- update(mod_1, .~. - gs) # удалили предиктор gs

| vif(mod_2)
| mod_3 <- update(mod_2, .~. - hb) # удалили предиктор hb

| vif(mod_3)
| mod_4 <- update(mod_3, .~. - plag) # удалили предиктор plag

| vif(mod_4) # в модели не осталось мультиколлинеарности
```

Итак, что же делать?

площадь	число комнат	школа близко	цена квартиры
50	2	нет	5000
1000	7	да	11000
30	1	нет	3500
100	4	нет	33333

$$\Rightarrow \quad R^2 = 0.99$$

Зачем делить на test и train?



Оптимальная модель

1. Если цель в том, чтобы добиться максимально точных предсказаний, то большее число предикторов более полно опишет данные

2. Если цель состоит в выявлении закономерностей в данных и интерпретации полученной модели (такая задача в биологии стоит чаще), то имеет смысл оставлять только значимые предикторы

$$t = \frac{w_i - w_{i\_real}}{\hat{\sigma}(w_i)}$$

Частный F-тест

$$F = \frac{(SE_{reduced} - SE_{full})/(df_{reduced} - df_{full})}{SE_{full}/df_{full}}$$

$$full = w_0 + w_1 x_1 + w_2 x_2$$

$$reduced = w_0 + w_1 x_1$$

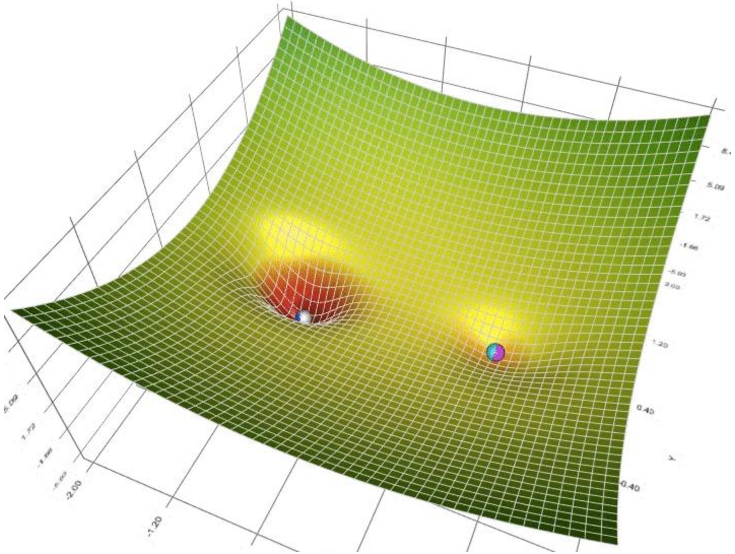
$$SE = \sum (y_i - \hat{y}_i)^2$$

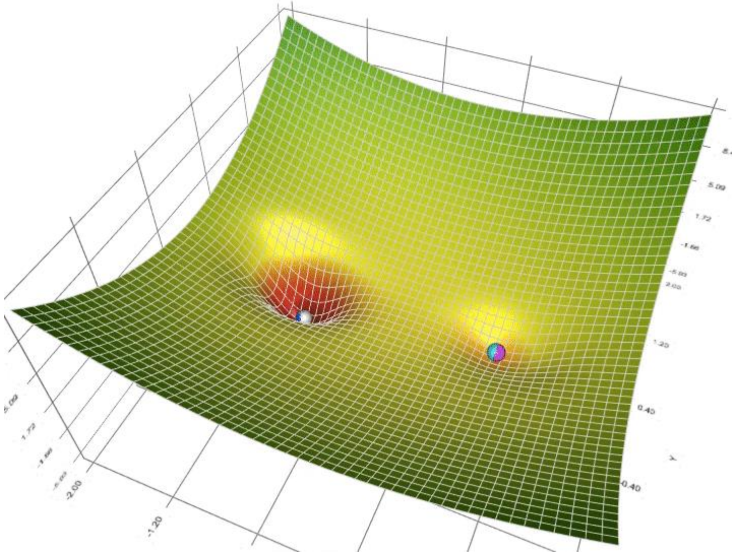
$$df = n - k$$

Частный F-тест сравнивает объясненную изменчивость между полной и вложенной моделью. Если модель после удаления предиктора значимо ухудшается, то такой предиктор важен.

Отдых

Градиентный спуск





Градиентный спуск

$$Loss = (y - Xw)^T(y - Xw) \quad \frac{dLoss}{dw} = \nabla Loss = 2X^T(Xw - y)$$

```
w = np.random.randn(m + 1)
Пока grad(Loss) != 0:
        w -= η * grad(Loss)
```

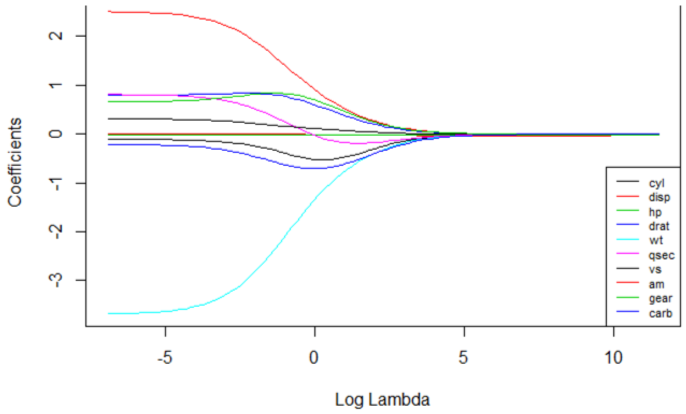
Регуляризация

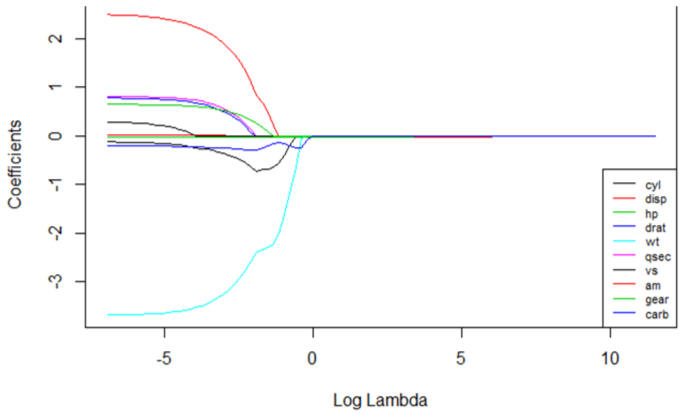
$$L1\ (Lasso) : Loss = (y - Xw)^T (y - Xw) + \lambda \sum |w_i|$$

$$L2\ (Ridge) : Loss = (y - Xw)^T (y - Xw) + \lambda \sum w_i^2$$

$$Elastic\, net : Loss = (y - Xw)^T (y - Xw) + \lambda_{l1} \sum |w_i| + \lambda_{l2} \sum w_i^2$$

Регуляризация





Регуляризация