IT-Academy

КОЛЛИНЕАРНОСТЬ. ОБРАБОТКА КАТЕГОРИАЛЬНЫХ ПЕРЕМЕННЫХ

КУХАЛЬСКИЙ НИКОЛАЙ ГЕННАДЬЕВИЧ

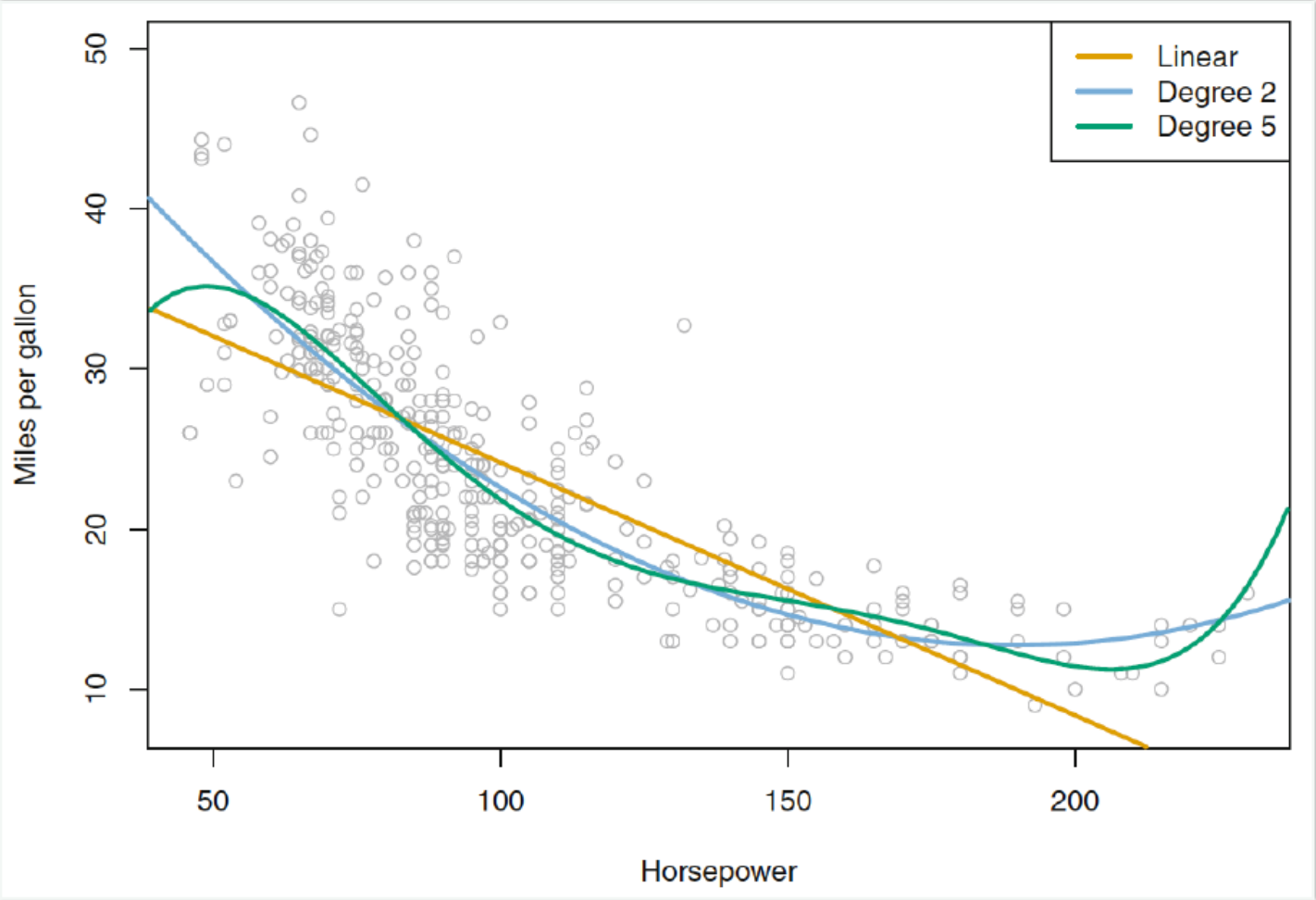
Вопросы занятия

1. Коллинеарность;

2. Обработка категориальных переменных:
- One hot encoding
- Counts
- Weights of evidence

НЕЛИНЕЙНЫЕ ДАННЫЕ

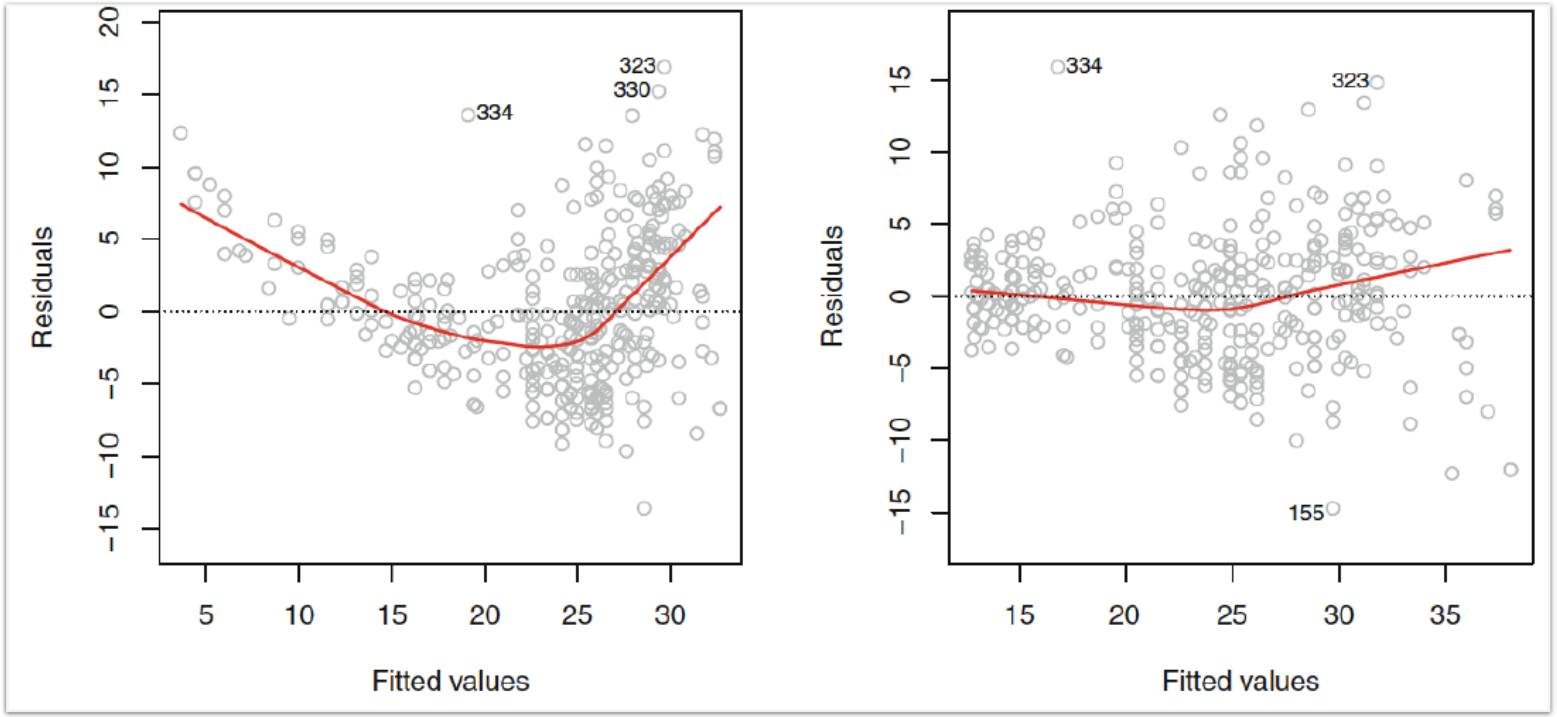
IT-Academy



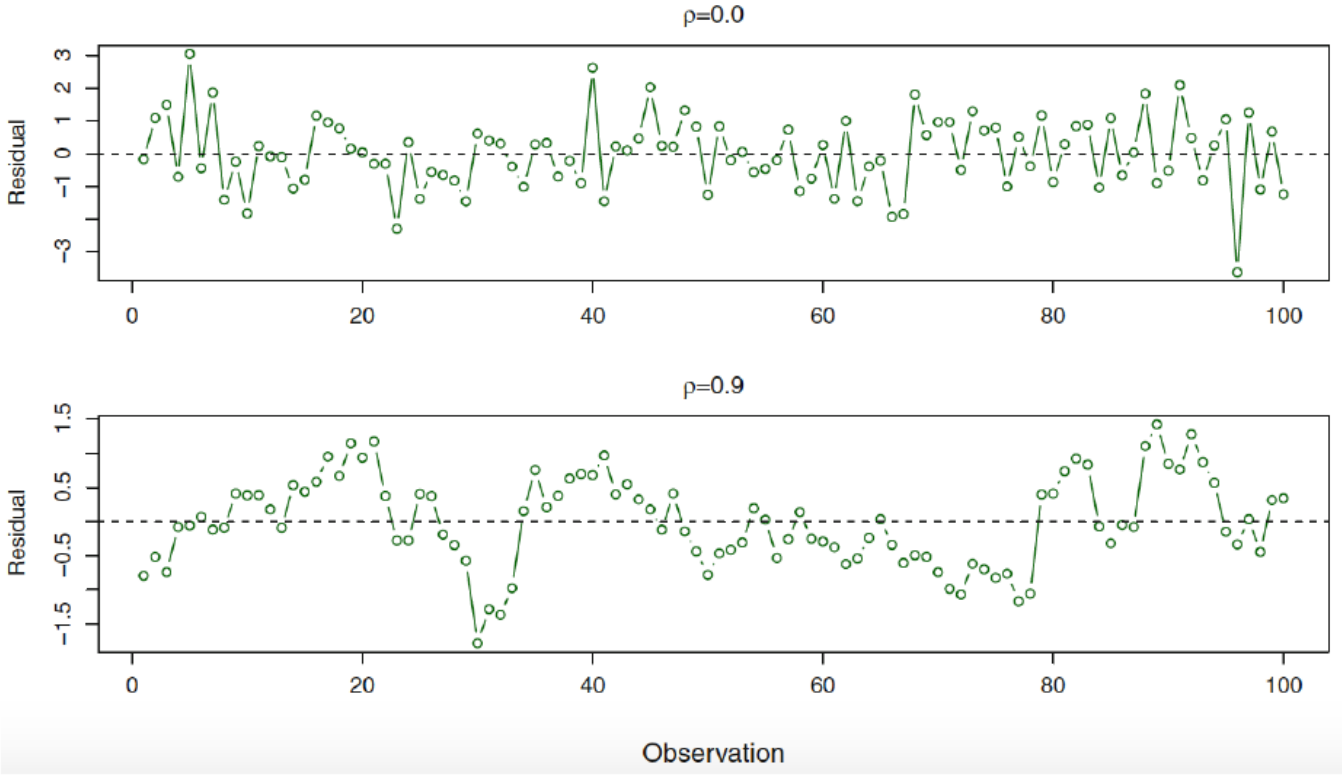
НЕЛИНЕЙНЫЕ ДАННЫЕ

IT-Academy

Графики ошибок для линейной и квадратичной регрессии



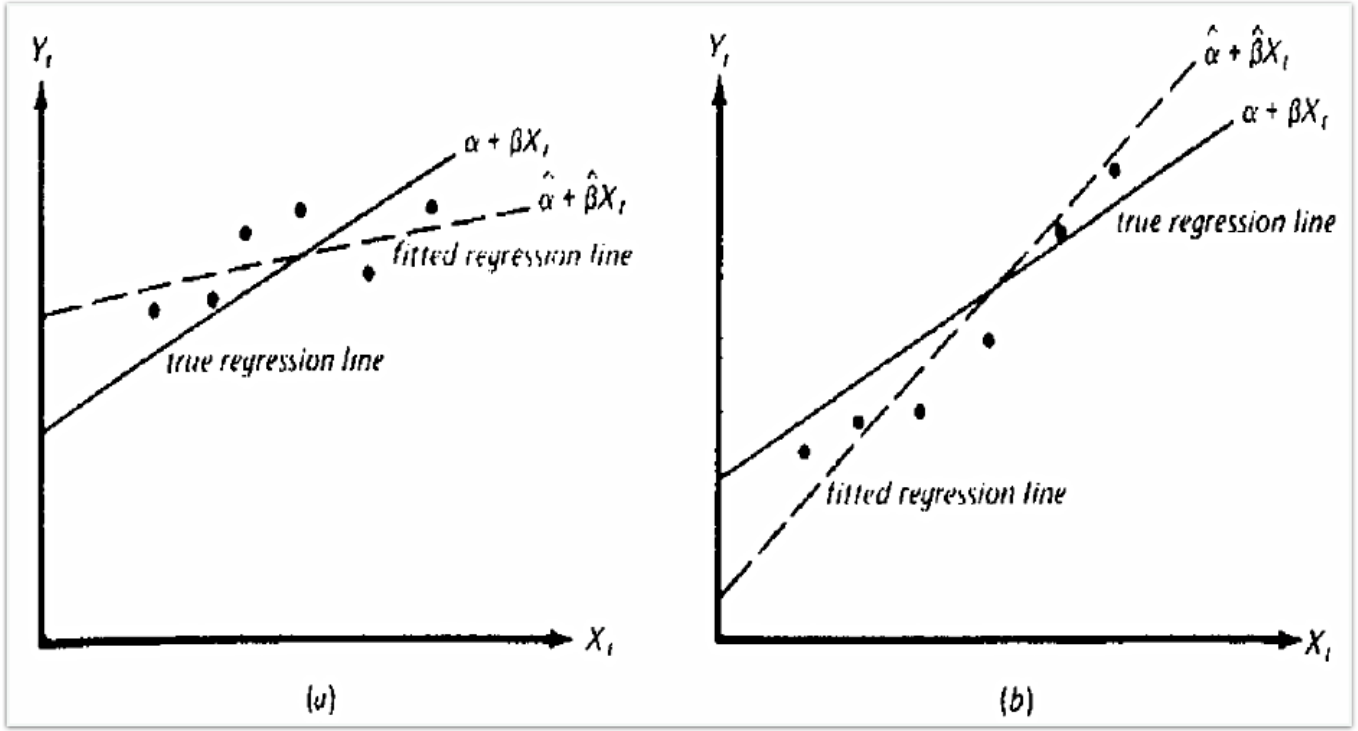
КОРРЕЛЯЦИЯ ОШИБОК



Ошибки, полученные при моделировании временной последовательности

ρ - уровень корреляции ошибок между смежными временными точками

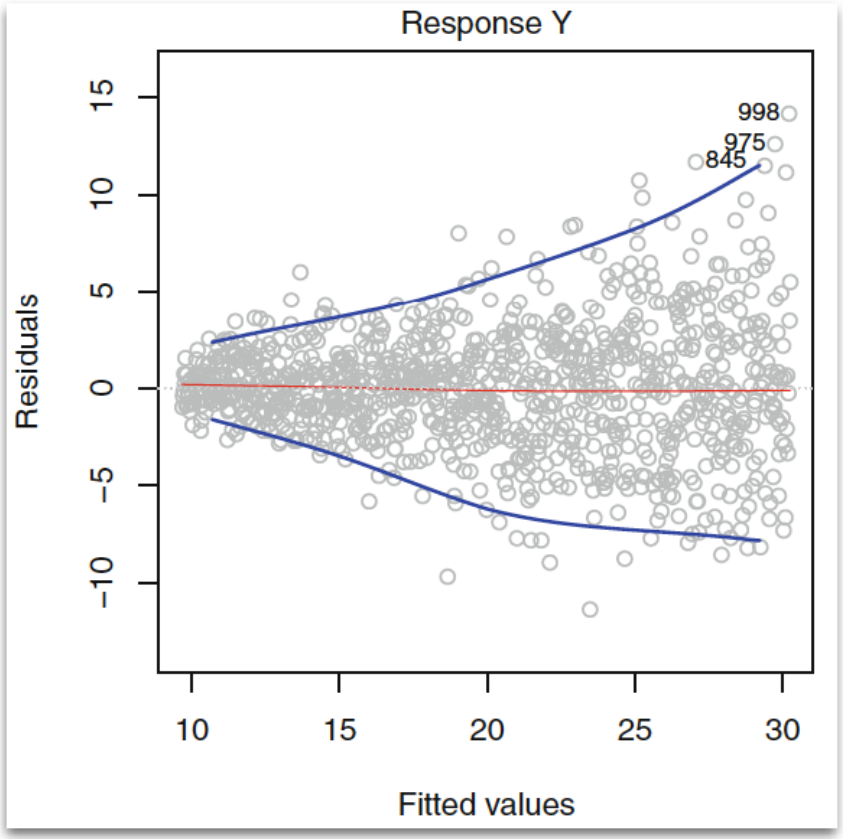
КОРРЕЛЯЦИЯ ОШИБОК



- Оцененная std. err. меньше чем реальная
- Параметры рассчитываются точнее, чем на самом деле
- Чаще отвергаем H_0
- Существуют тесты для определения корреляции ошибок

НЕПОСТОЯННОЕ ОТКЛОНЕНИЕ ОШИБКИ

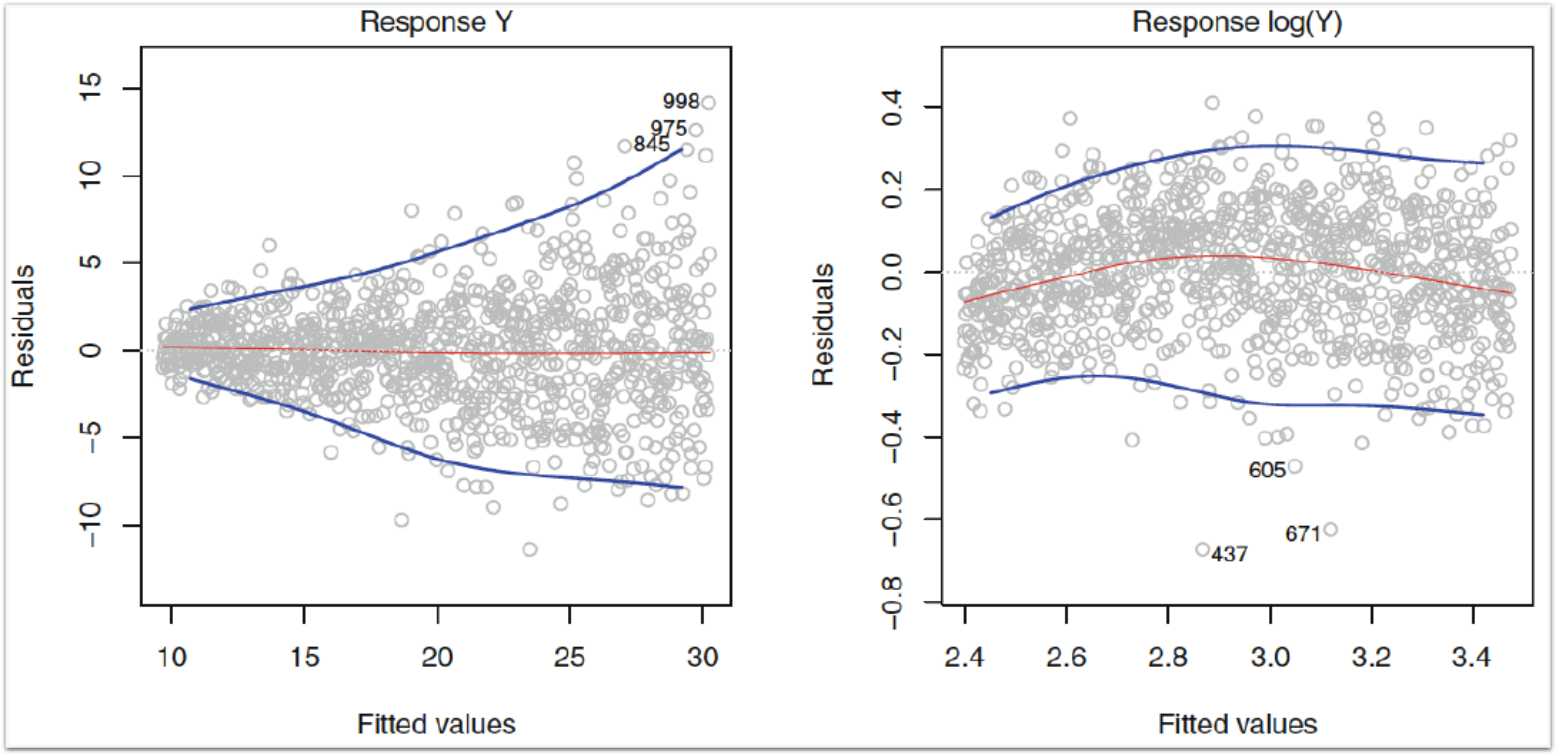
IT-Academy



Гетероскедастичность можно обнаружить по форме изображения ошибок (похожа на воронку)

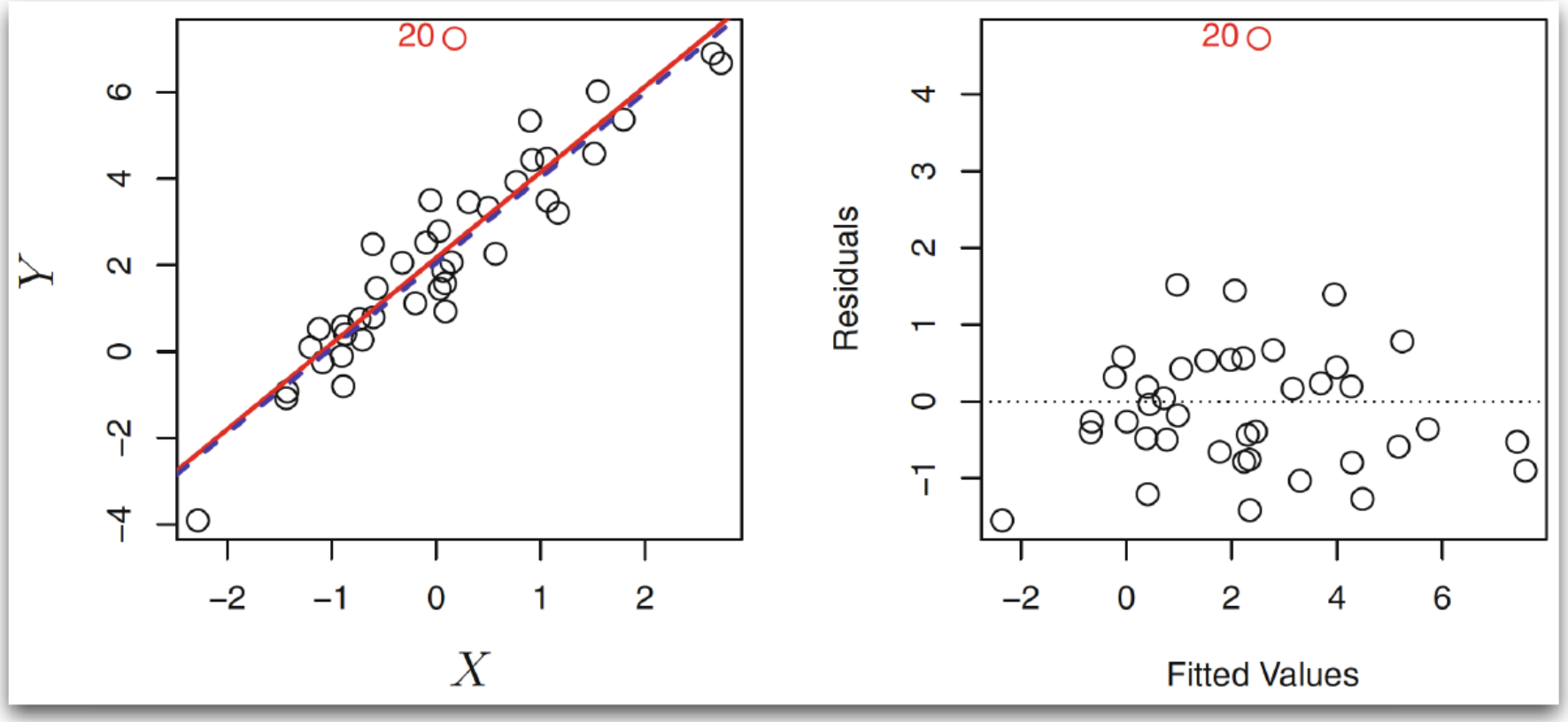
НЕПОСТОЯННОЕ ОТКЛОНЕНИЕ ОШИБКИ

IT-Academy



Лечится преобразованием предсказываемых данных

ВЫБРОСЫ



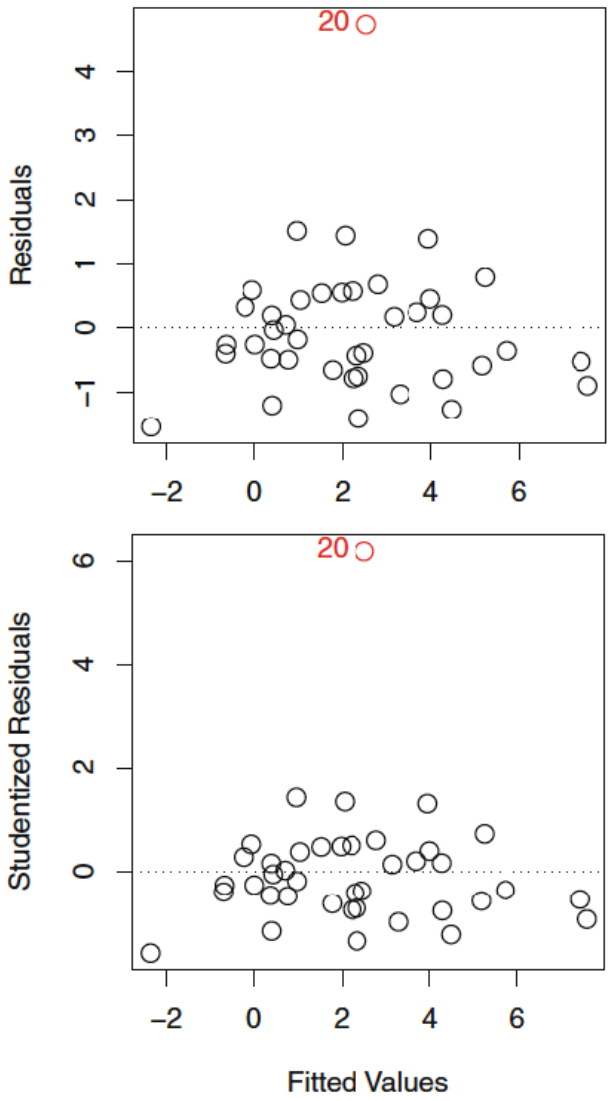
ВЫБРОСЫ

IT-Academy

- На графике Fitted value-Residual выброс не всегда очевиден
- Вместо Residual используют Studentize residual

$$e_i^{st} = \frac{e_i}{SE(\hat{y})}$$

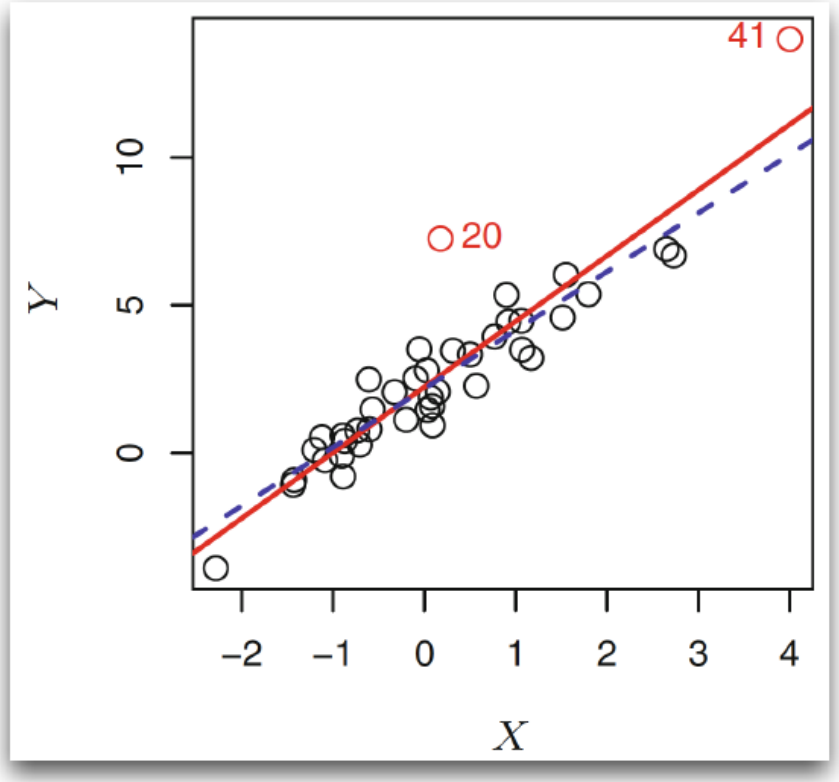
- Возможный выброс, если $e^{st}_i > 3$



HIGH-LEVERAGE POINTS

IT-Academy

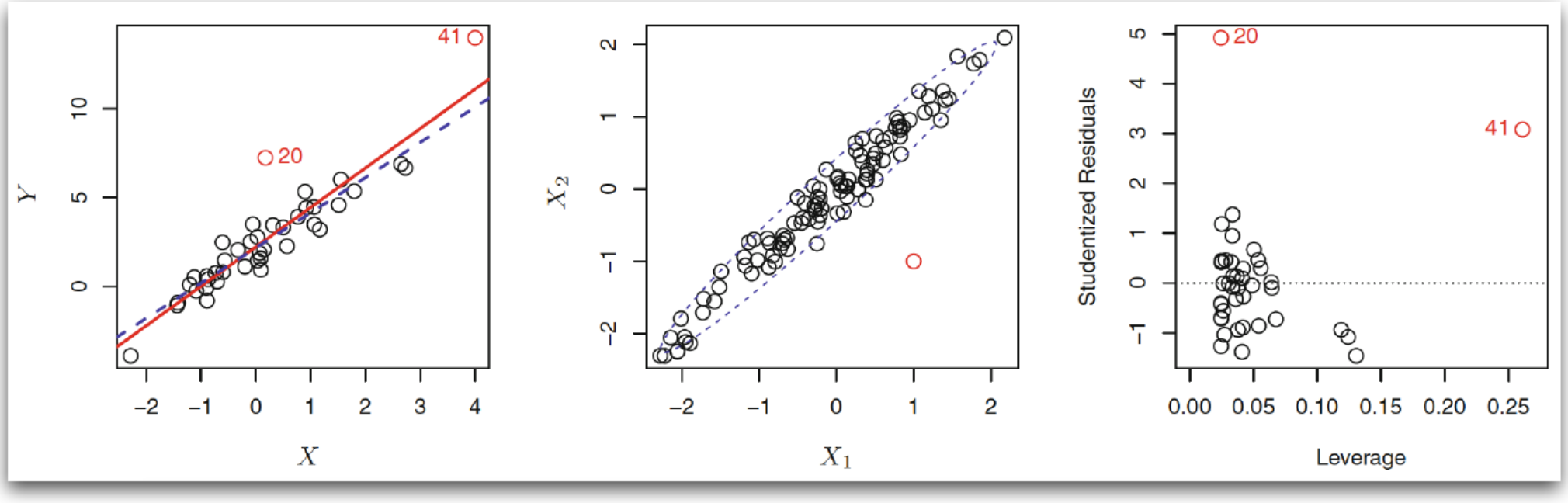
- Необычное значение y_i - выброс
- Необычное значение x_i - точка с высоким коэффициентом усиления
- HL-точка 41, влияет на результирующую модель намного сильнее, чем точка-выброс 20



HIGH-LEVERAGE POINTS

LEVERAGE STATISTICS

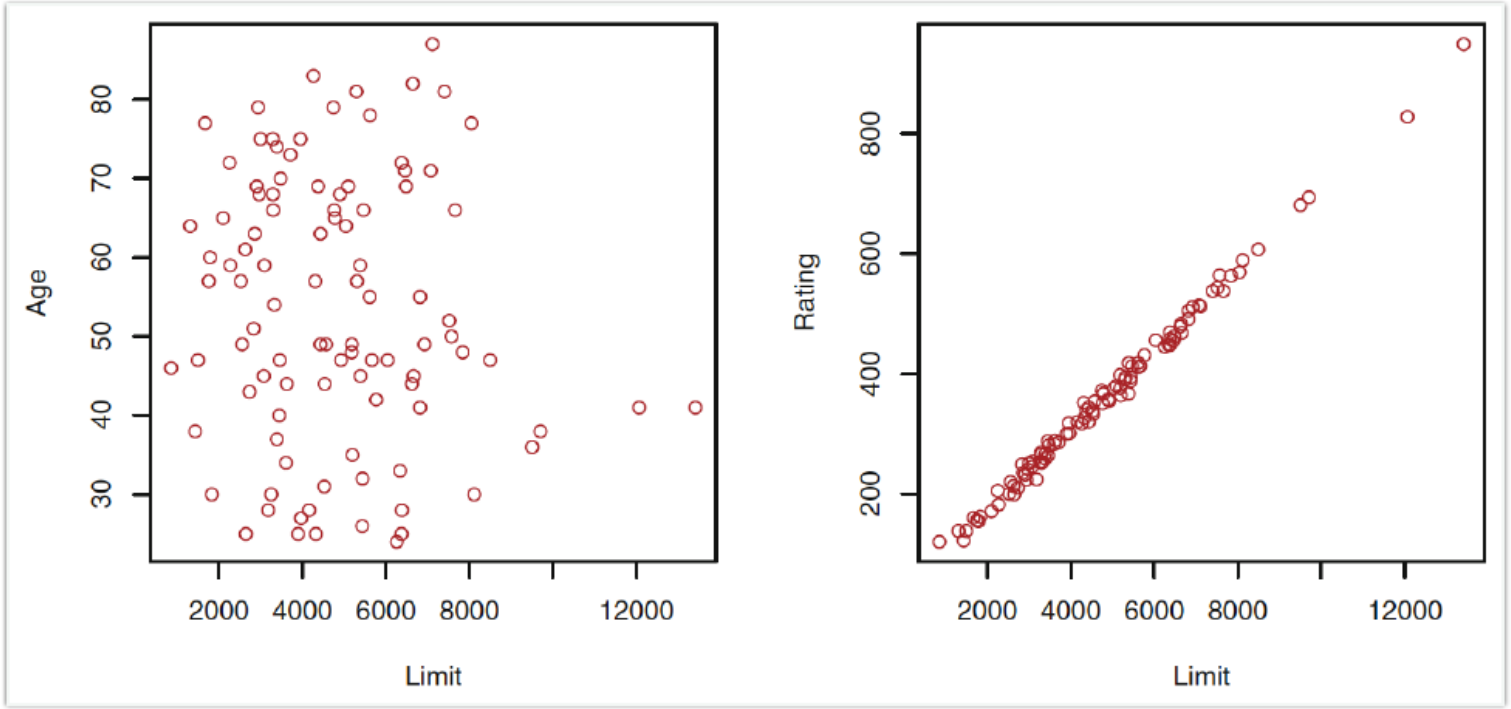
$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})}$$



КОЛЛИНЕАРНОСТЬ

IT-Academy

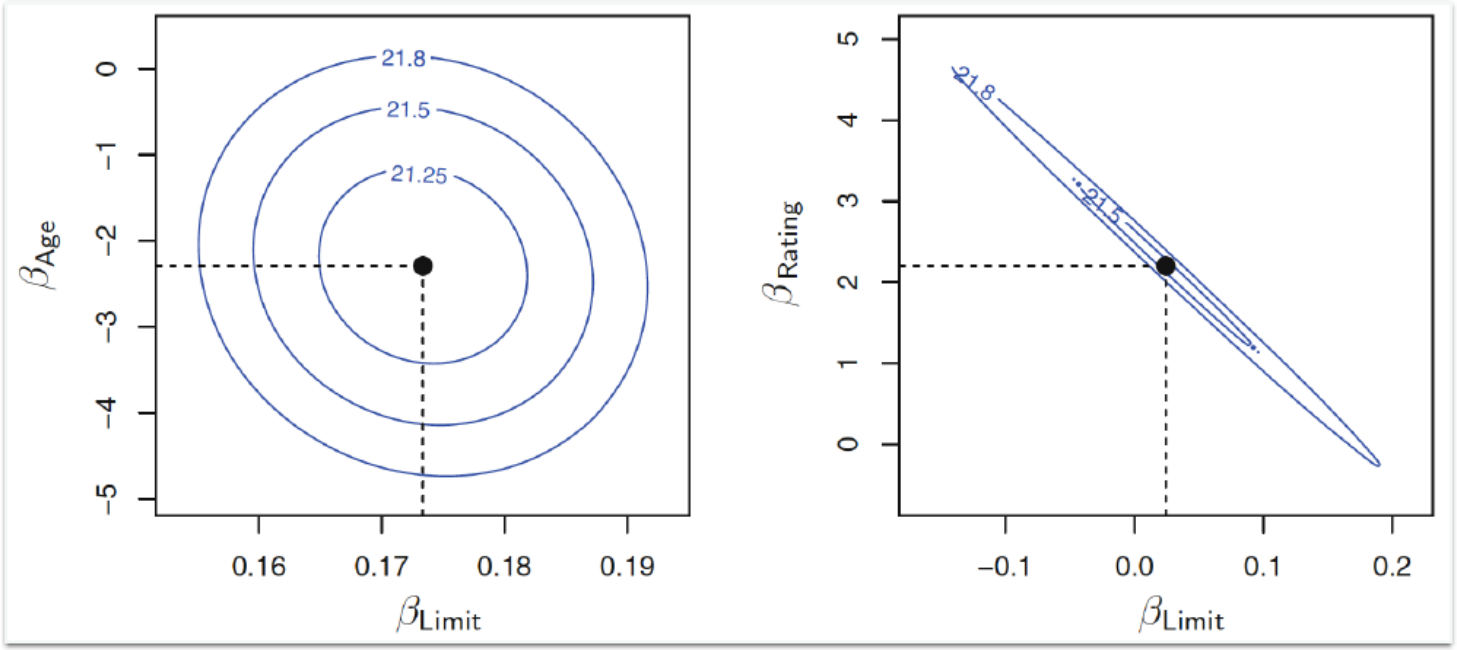
- Сильная корреляция между двумя и более предикторами



КОЛЛИНЕАРНОСТЬ

IT-Academy

Модель: предсказание кредитного баланса



Контуры RSS в зависимости от значений β для Age, Limit и Rating

КОЛЛИНЕАРНОСТЬ

IT-Academy

- Увеличивает std. err.
- Страдает t-statistics (не можем отвергнуть H_0)
- Простой способ определения - матрица корреляций (не работает с мультиколлинеарностью)
- Variance inflation factor: $$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

$R^2_{X_j|X_{-j}}$ - R^2 посчитанный для регрессии от X_j относительно всех остальных

- VIF = 1 - нет коллинеарности

КОЛЛИНЕАРНОСТЬ

Решение коллинеарности

- Удалить один из атрибутов
- Скомбинировать атрибуты

IT-Academy

ПРАКТИКА

cars.csv

ОБРАБОТКА КАТЕГОРИАЛЬНЫХ ПЕРЕМЕННЫХ

IT-Academy

Виды трансформации категориальных переменных

- Label encoding
- One hot encoding
- Counts (Likelihood encoding)
- Weights of evidence (WOE)

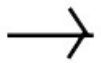
ОБРАБОТКА КАТЕГОРИАЛЬНЫХ ПЕРЕМЕННЫХ

IT-Academy

Encoding

Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50



One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

ОБРАБОТКА КАТЕГОРИАЛЬНЫХ ПЕРЕМЕННЫХ

IT-Academy

Counts (Likelihood encoding)

- Для каждой категории считаем среднюю долю целевого события
- С математической точки зрения – условная вероятность целевого события при известной категории соответствующей фичи
- Получается в какой-то степени смещенный результат
- Желательно делать KFold, исключая возможность переобучения

ОБРАБОТКА КАТЕГОРИАЛЬНЫХ ПЕРЕМЕННЫХ

IT-Academy

Weight of Evidence

$$Weight of Evidence = ln(\frac{DistributionGood_i}{DistributionBad_i})$$

где:
DistrGood – отношение числа хороших наблюдений, имевших значение атрибута из данного бина, к общему числу хороших наблюдений;

DistrBad – отношение числа плохих наблюдений, имевших значение атрибута из данного бина, к общему числу плохих наблюдений.

$$IV = \sum(DistributionGood_i - DistributionBad_i) \times WOE_i$$

- оценка информативности переменной.

На основе коэффициентов WoE вычисляется величина, определяющая значимость признака в модели бинарной классификации, называемая информационным индексом (IV)

IT-Academy

ПРАКТИКА

cars.csv

IT-Academy

ПРАКТИКА

Taxi_Moscow.ipynb

IT-Academy

ПРАКТИКА

Paribas.csv

IT-Academy

КОЛЛИНЕАРНОСТЬ. ОБРАБОТКА КАТЕГОРИАЛЬНЫХ ПЕРЕМЕННЫХ

КУХАЛЬСКИЙ НИКОЛАЙ ГЕННАДЬЕВИЧ