



Линейная модель



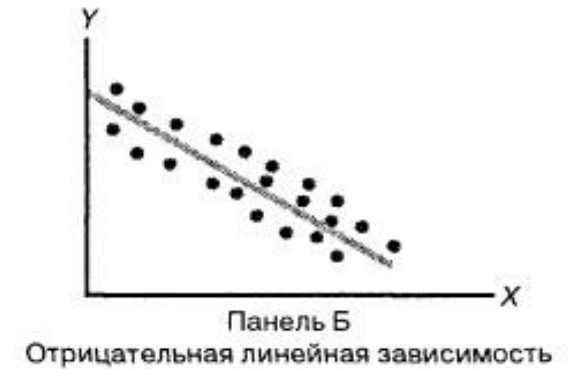
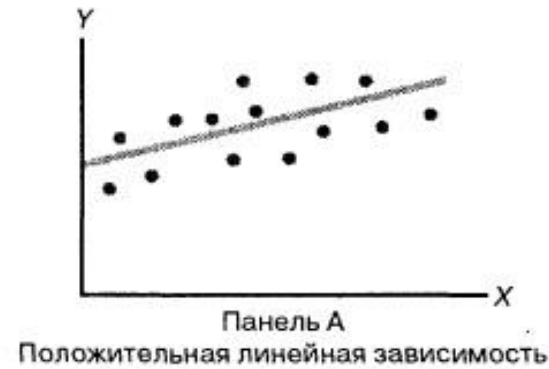
Российский университет
дружбы народов

Яссин Мохмад Аламин,
НКНбд-01-20

Линейные модели

$$Y = a + bX + \epsilon$$

Принцип работы линейной регрессии



Формула для линейной регрессии

Y_i = зависимую переменную
 β_0, β_1 = коэффициенты регрессии
 ε_i = остаточная ошибка

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

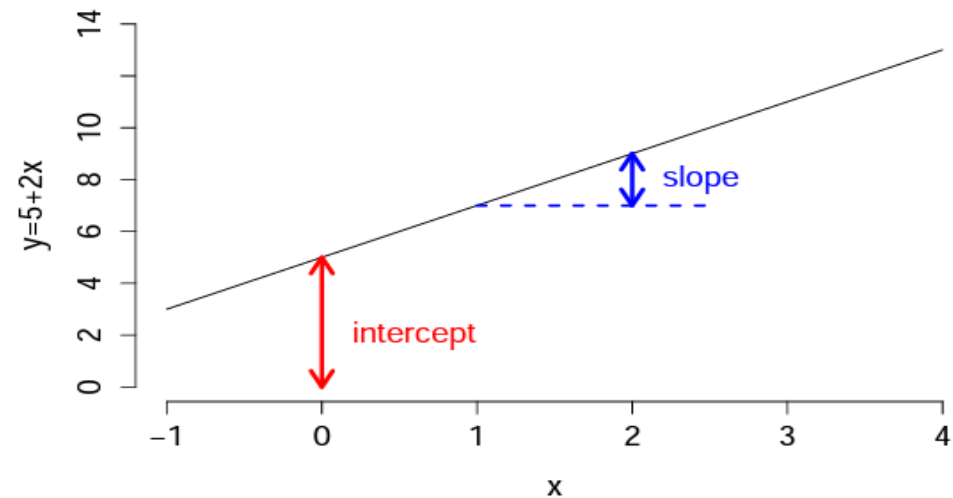
Diagram labels for the equation:

- Dependent Variable: Y_i
- Population Y intercept: β_0
- Population Slope Coefficient: β_1
- Independent Variable: X_i
- Random Error term: ε_i

Component labels:

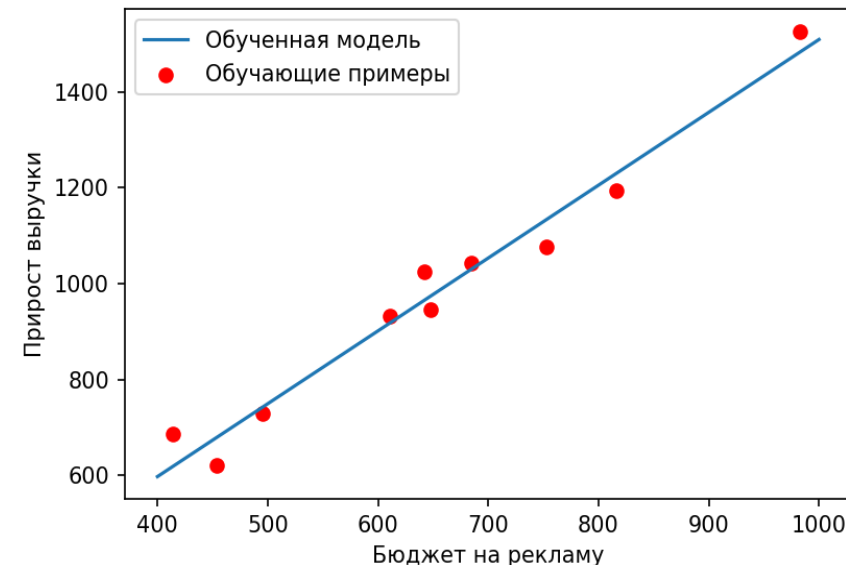
- Linear component: $\beta_0 + \beta_1 X_i$
- Random Error component: ε_i

$$y = 5 + 2x$$

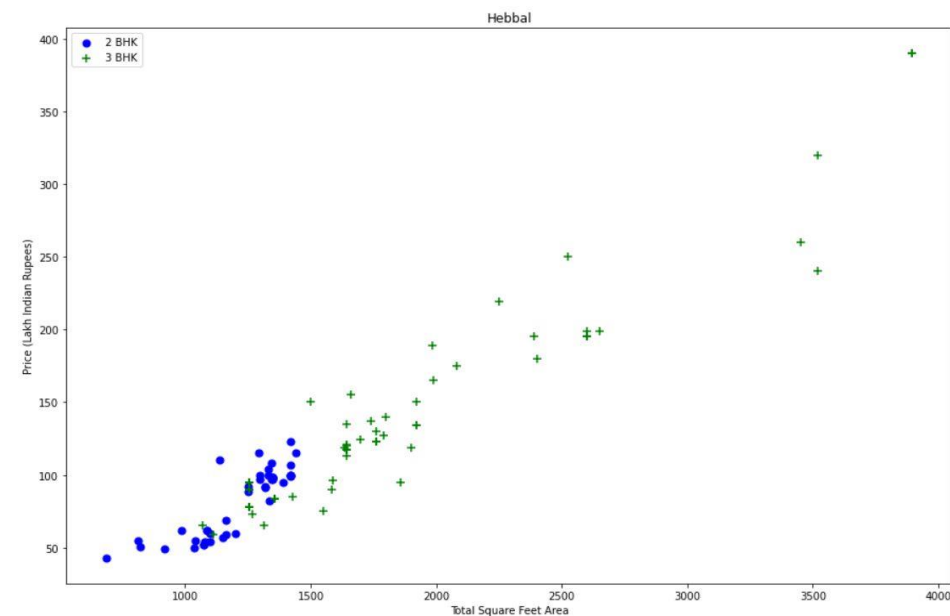


Примеры

- Предсказание продаж в зависимости от рекламы
- определение влияния года выпуска на стоимость автомобиля
- предсказание температуры воздуха на основе времени года
- предсказание цены на аренду квартир



<https://sysblok.ru/glossary/chto-takoe-linejnaja-regressija/>



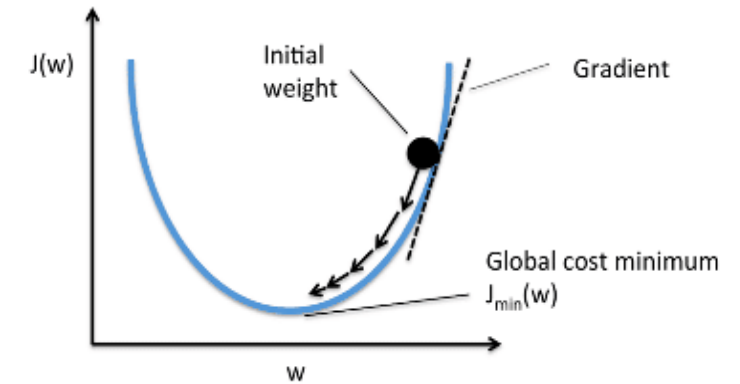
<https://github.com/Strikoder/DS-ML-DL/tree/main/Projects>

Функция потерь

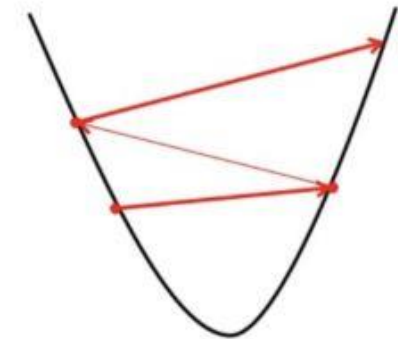
$$\text{minimize } \frac{1}{n} \sum_{i=1}^n (\text{pred}_i - y_i)^2$$

- функция потерь - измеряет соответствие модели данным. В линейной регрессии используется MSE. Цель - минимизировать функцию, настраивая параметры.
- Градиентный спуск: Метод минимизации ошибки линейной регрессии, итеративно обновляющий коэффициенты.

Градиентный спуск



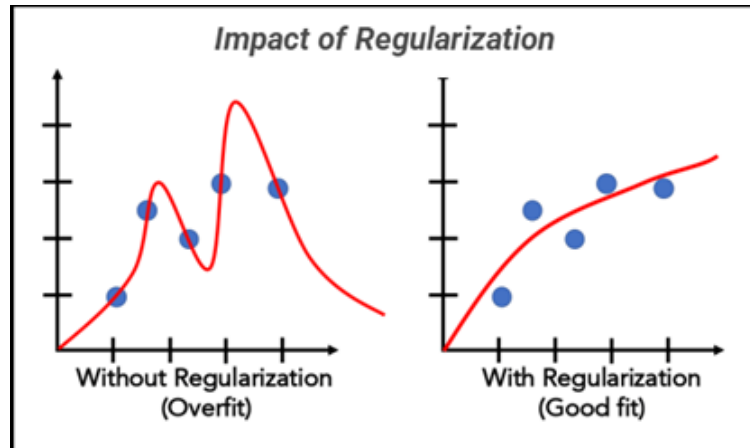
выбор больших шагов



Регуляризация: Ridge - Lasso

Regularized linear regression

$$\min_{\vec{w}, b} J(\vec{w}, b) = \min_{\vec{w}, b} \left[\frac{1}{2m} \sum_{i=1}^m (f_{\vec{w}, b}(\vec{x}^{(i)}) - y^{(i)})^2 + \frac{\lambda}{2m} \sum_{j=1}^n w_j^2 \right]$$

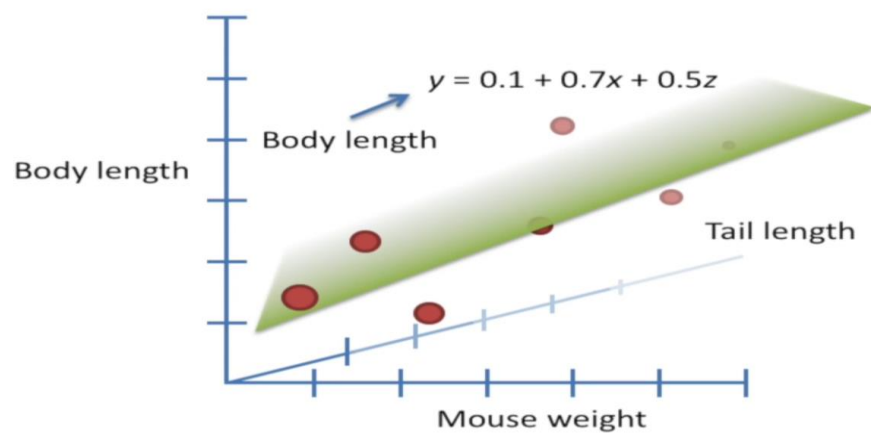


$$L1 \text{ Regularization} = (\text{loss function}) + \alpha \sum_{j=1}^p |b_j|$$

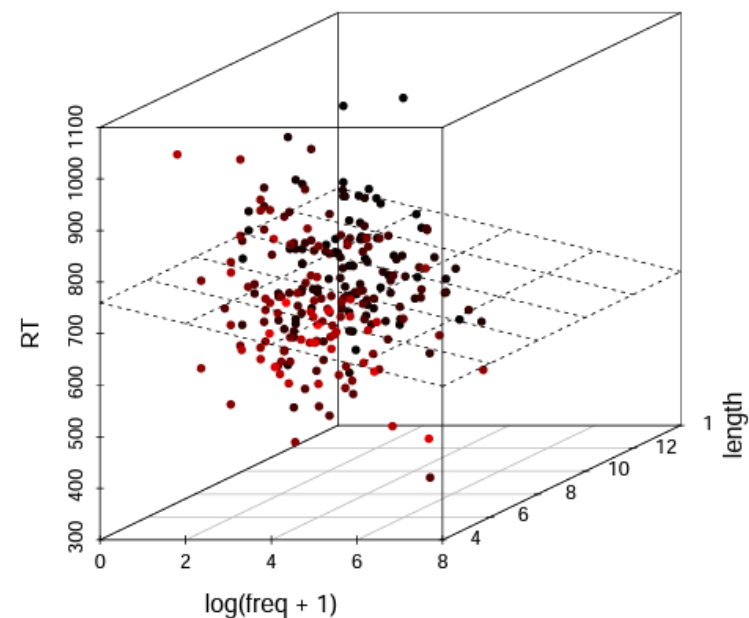
$$L2 \text{ Regularization} = (\text{loss function}) + \alpha \sum_{j=1}^p b_j^2$$

Модель множественной линейной регрессии

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$$

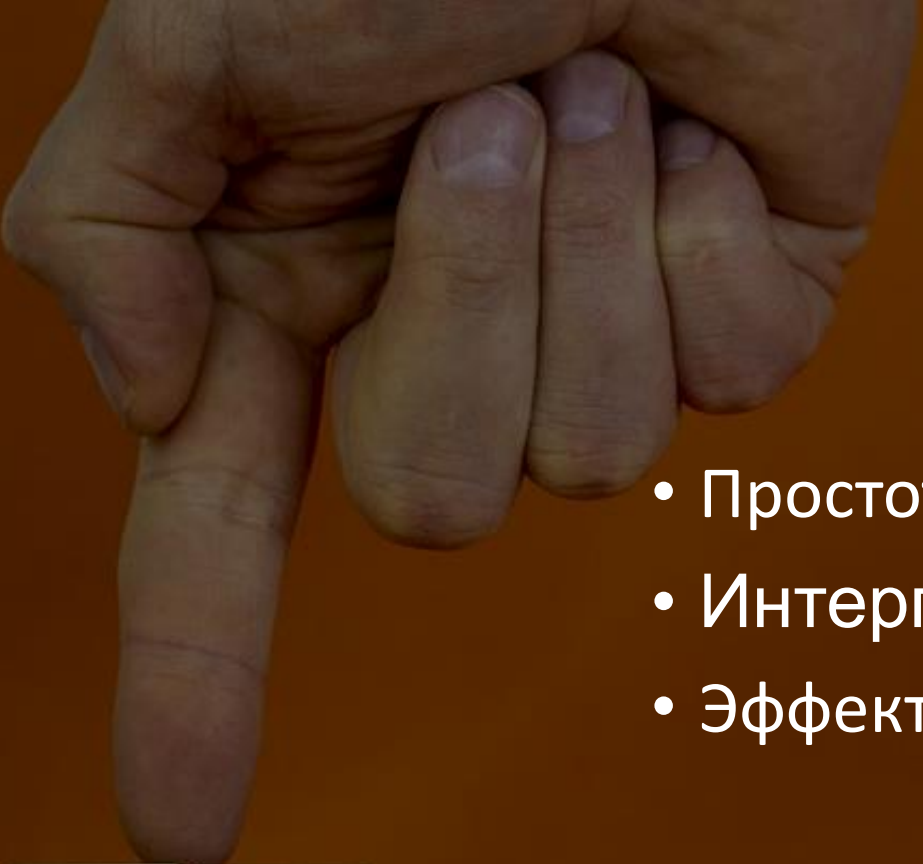


Naming Time by Log Frequency and Word Length



METRICS

	Линейная регрессия (оценка МНК)	Линейная регрессия (метод инструментальных переменных)	Модель бинарного выбора	Нелинейная регрессия
Проверка значимости каждого фактора	t-статистика и её вероятность	t-статистика и её вероятность	z-статистика и её вероятность	t-статистика и её вероятность
Проверка корреляции факторов	матрица корреляции	матрица корреляции	матрица корреляции	-
Проверка критериев качества	R ² , Adj R ²	R ² , Adj R ²	McFadden R ²	R ² , Adj R ²
Проверка значимости всех факторов	F-статистика и её вероятность	F-статистика и её вероятность, J-статистика и её вероятность	LR-статистика и её вероятность	F-статистика и её вероятность
Сопоставление сравнительных критериев	SSR, LogL	SSR, LogL	LogL	SSR
	AIC, SC, HQ	AIC, SC, HQ	AIC, SC, HQ	AIC, SC, HQ
Анализ остатков	DW	DW	DW	DW

- 
- Простота
 - Интерпретируемость
 - Эффективность

- Линейность
- Предположения о распределении
- Выбросы
- Мультиколлинеарность
- Ограниченность в предсказаниях

DISADVANTAGES

Заключение

Linear regression – это один из самых хороших алгоритмов для анализа данных и предсказания будущих значений. Однако, она имеет ограничения, и для ее использования необходимы чистые и обработанные данные.

```
*****
*****
*****
*****
```

```
# Find the best algorithm
best_model_name = max(best_scores, key=lambda x: x['best_score'])['model']
print("Best algorithm:", best_model_name)

# Get the best model object and its parameters
best_model = model_params[best_model_name]['model']
best_params = next(item for item in best_scores if item['model'] == best_model_name)['best_params']

# Fit the best algorithm into the pipeline
steps = [('imputer', SimpleImputer()),
         ('scaler', StandardScaler()),
         (best_model_name, best_model.set_params(**best_params))]

pipeline = Pipeline(steps)

pipeline.fit(X_train, y_train)

# Fit the pipeline and compute its evaluation metric
if best_model_name == 'linear_regression':
    predictions = pipeline.predict(X_test)
    mse = mean_squared_error(y_test, predictions)
    print("MSE:", mse)
else:
    score = pipeline.score(X_test, y_test)
    print("Score:", score)
```

```
Best algorithm: linear_regression
MSE: 1111.875328640399
```

<https://www.kaggle.com/code/strikoder/randomizedsearchcv-pipeline-ann/notebook>

ИСТОЧНИКИ

- https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html
- <https://habr.com/ru/post/278513/>
- Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to linear regression analysis (5th ed.). John Wiley & Sons. (p. 38)
- Kumari, Khushbu & Yadav, Suniti. (2018). Linear regression analysis study. Journal of the Practice of Cardiovascular Sciences. 4. 33. 10.4103/jpcs.jpcs_8_18.
- Gujarati, D. N. (2018). Linear Regression: A Mathematical Introduction.



Спасибо за внимание!