Математическая статистика

Линейная регрессия


МОЁ ХОББИ: ЭКСТРАПОЛИРОВАТЬ
КАК ВИДИМ, К КОНЦУ СЛЕДУЮЩЕГО МЕСЯЦА У ТЕБЯ БУДЕТ БОЛЕЕ ПЯТИ ДЕСЯТКОВ МУЖЕЙ.
ЛУЧШЕ ЗАКАЖИ СВАДЕБНЫЕ ТОРТЫ ОПТОМ.
КОЛИЧЕСТВО МУЖЕЙ
1
0
ВЧЕРА
СЕГОДНЯ

Пример

Простая зависимость:

$$y = \theta_0 + \theta_1 x + \varepsilon,$$

x — рост песика,

y — вес песика,

θ_0, θ_1 — неизвестные параметры,

ε — случайная составляющая с нулевым средним.

Зависимость **линейна по параметрам**, линейна по аргументу.

Пример

Более сложная зависимость:

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_2^2 + \varepsilon,$$

x_1 — рост песика,

x_2 — обхват туловища песика,

y — вес песика,

$\theta_0, \theta_1, \theta_2, \theta_3$ — неизвестные параметры,

ε — случайная составляющая с нулевым средним.

Зависимость **линейна по параметрам**, квадратична по аргументам.

Модель линейной регрессии

Рассматриваем функциональную зависимость вида

$$y = y(x) = \theta_1 x_1 + \ldots + \theta_d x_d$$

$x_1, \ldots, x_d$ — признаки,

$\theta = (\theta_1, \ldots, \theta_d)^T$ — вектор параметров.

Модель линейной регрессии

Рассматриваем функциональную зависимость вида

$$y = y(x) = \theta_1 x_1 + \ldots + \theta_d x_d$$

$x_1, \ldots, x_d$ — признаки,

$\theta = (\theta_1, \ldots, \theta_d)^T$ — вектор параметров.

Для оценки θ производится n испытаний вида

$$Y_i = \theta_1 x_{i1} + \ldots + \theta_d x_{id} + \varepsilon_i, \quad i = 1, \ldots, n,$$

$x_i = (x_{i1}, \ldots, x_{id})$ — признаковые описания объекта i

(обычно неслучайные),

ε_i — случайная ошибка измерений.

Модель линейной регрессии

Рассматриваем функциональную зависимость вида

$$y = y(x) = \theta_1 x_1 + \ldots + \theta_d x_d$$

$x_1, \ldots, x_d$ — признаки,

$\theta = (\theta_1, \ldots, \theta_d)^T$ — вектор параметров.

Для оценки θ производится n испытаний вида

$$Y_i = \theta_1 x_{i1} + \ldots + \theta_d x_{id} + \varepsilon_i, \quad i = 1, \ldots, n,$$

$x_i = (x_{i1}, \ldots, x_{id})$ — признаковые описания объекта i

(обычно неслучайные),

ε_i — случайная ошибка измерений.

Модель линейной регрессии

Введем обозначения

$$Y = \begin{pmatrix} Y_1 \\ \dots \\ Y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_{11} & \dots & x_{1d} \\ \dots & & \\ x_{n1} & \dots & x_{nd} \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \dots \\ \varepsilon_n \end{pmatrix}.$$

Модель линейной регрессии

Введем обозначения

$$Y = \begin{pmatrix} Y_1 \\ \dots \\ Y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_{11} & \dots & x_{1d} \\ \dots & & \\ x_{n1} & \dots & x_{nd} \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \dots \\ \varepsilon_n \end{pmatrix}.$$

Матричная форма записи проведенных испытаний

$$Y = X\theta + \varepsilon.$$

$X \in \mathbb{R}^{n \times d}$ — регрессоры (или матрица плана эксперимента),

$Y \in \mathbb{R}^n$ — отклик.

Матричный вид зависимости: $y(x) = x^T\theta$.

Модель линейной регрессии

Введем обозначения

$$Y = \begin{pmatrix} Y_1 \\ \dots \\ Y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_{11} & \dots & x_{1d} \\ \dots & & \\ x_{n1} & \dots & x_{nd} \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \dots \\ \varepsilon_n \end{pmatrix}.$$

Матричная форма записи проведенных испытаний

$$Y = X\theta + \varepsilon.$$

$X \in \mathbb{R}^{n \times d}$ — регрессоры (или матрица плана эксперимента),

$Y \in \mathbb{R}^n$ — отклик.

Матричный вид зависимости: $y(x) = x^T\theta$.

Категориальные переменные

x — id должности сотрудника (натуральное число),

y — его зарплата.

Предположим, что должности занумерованы следующим образом:

- $x = 1$ — простой рабочий;
- $x = 2$ — сисадмин, присваивающий id;
- $x = 3$ — директор.

Сисадмин предложит рассмотреть модель $y = \theta_0 + \theta_1 x$:))

Категориальные переменные

x — id должности сотрудника (натуральное число),

y — его зарплата.

Предположим, что должности занумерованы следующим образом:

- $x = 1$ — простой рабочий;
- $x = 2$ — сисадмин, присваивающий id;
- $x = 3$ — директор.

Сисадмин предложит рассмотреть модель $y = \theta_0 + \theta_1 x$:))

Если $x \in \{1, ..., k\}$, то рассматриваются **dummy-переменные**:

$x_j = I\{x = j\}, \quad j = 1, ..., k-1,$

модель $y = \theta_0 + \theta_1 x_1 + ... + \theta_{k-1} x_{k-1}$.

Категориальные переменные

x — id должности сотрудника (натуральное число),

y — его зарплата.

Предположим, что должности занумерованы следующим образом:

- $x = 1$ — простой рабочий;
- $x = 2$ — сисадмин, присваивающий id;
- $x = 3$ — директор.

Сисадмин предложит рассмотреть модель $y = \theta_0 + \theta_1 x$:))

Если $x \in \{1, ..., k\}$, то рассматриваются **dummy-переменные**:

$x_j = I\{x = j\}, \quad j = 1, ..., k-1,$

модель $y = \theta_0 + \theta_1 x_1 + ... + \theta_{k-1} x_{k-1}$.

Метод наименьших квадратов

Задача: $RSS(\theta) = \|Y - X\theta\|_2^2 \to \min\limits_{\theta \in \mathbb{R}^d}.$

Метод наименьших квадратов

Задача: $RSS(\theta) = \|Y - X\theta\|_2^2 \to \min\limits_{\theta \in \mathbb{R}^d}$.

Решение: $\widehat{\theta} = \left(X^T X\right)^{-1} X^T Y$ если $rank\ X = d$.

Метод наименьших квадратов

Задача: $RSS(\theta) = \|Y - X\theta\|_2^2 \to \min\limits_{\theta \in \mathbb{R}^d}$.

Решение: $\widehat{\theta} = \left(X^T X\right)^{-1} X^T Y$ если $rank\ X = d$.

Предположения и следствия:

1. $\mathsf{E}\varepsilon = 0$ — несмещенность:

Метод наименьших квадратов

Задача: $RSS(\theta) = \|Y - X\theta\|_2^2 \to \min\limits_{\theta \in \mathbb{R}^d}$.

Решение: $\widehat{\theta} = \left(X^T X\right)^{-1} X^T Y$ если $rank\ X = d$.

Предположения и следствия:

1. $\mathsf{E}\varepsilon = 0$ — несмещенность:
   - $\widehat{\theta}$ — несмещенная оценка θ.

Метод наименьших квадратов

Задача: $RSS(\theta) = \|Y - X\theta\|_2^2 \to \min\limits_{\theta \in \mathbb{R}^d}$.

Решение: $\widehat{\theta} = \left(X^T X\right)^{-1} X^T Y$ если $rank\ X = d$.

Предположения и следствия:

1. $\mathsf{E}\varepsilon = 0$ — несмещенность:
   - $\widehat{\theta}$ — несмещенная оценка θ.
2. $\mathsf{E}\varepsilon = 0$ и $\mathsf{D}\varepsilon = \sigma^2 I_n$ — несмещенность и гомоскедастичность:

Метод наименьших квадратов

Задача: $RSS\,(\theta) = \|Y - X\theta\|_2^2 \to \min\limits_{\theta \in \mathbb{R}^d}.$

Решение: $\widehat{\theta} = \left(X^T X\right)^{-1} X^T Y$ если $rank\ X = d$.

Предположения и следствия:

1. $\mathsf{E}\varepsilon = 0$ — несмещенность:
   - $\widehat{\theta}$ — несмещенная оценка θ.
2. $\mathsf{E}\varepsilon = 0$ и $\mathsf{D}\varepsilon = \sigma^2 I_n$ — несмещенность и гомоскедастичность:
   - $\mathsf{D}\widehat{\theta} = \sigma^2 \left(X^T X\right)^{-1}$;
   - $\widehat{\theta}$ — оптимальная оценка θ среди линейных по y;
   - $\widehat{\sigma}^2 = RSS(\widehat{\theta})/(n-d)$ — несмещенная оценка σ^2.

Метод наименьших квадратов

Задача: $RSS(\theta) = \|Y - X\theta\|_2^2 \to \min\limits_{\theta \in \mathbb{R}^d}$.

Решение: $\widehat{\theta} = \left(X^T X\right)^{-1} X^T Y$ если $rank\ X = d$.

Предположения и следствия:

1. $\mathsf{E}\varepsilon = 0$ — несмещенность:
   - $\widehat{\theta}$ — несмещенная оценка θ.
2. $\mathsf{E}\varepsilon = 0$ и $\mathsf{D}\varepsilon = \sigma^2 I_n$ — несмещенность и гомоскедастичность:
   - $\mathsf{D}\widehat{\theta} = \sigma^2 \left(X^T X\right)^{-1}$;
   - $\widehat{\theta}$ — оптимальная оценка θ среди линейных по y;
   - $\widehat{\sigma}^2 = RSS(\widehat{\theta})/(n-d)$ — несмещенная оценка σ^2.
3. $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ — гауссовская линейная модель:

Метод наименьших квадратов

Задача: $RSS(\theta) = \|Y - X\theta\|_2^2 \to \min\limits_{\theta \in \mathbb{R}^d}$.

Решение: $\widehat{\theta} = \left(X^T X\right)^{-1} X^T Y$ если $rank\ X = d$.

Предположения и следствия:

1. $\mathsf{E}\varepsilon = 0$ — несмещенность:
   - $\widehat{\theta}$ — несмещенная оценка θ.
2. $\mathsf{E}\varepsilon = 0$ и $\mathsf{D}\varepsilon = \sigma^2 I_n$ — несмещенность и гомоскедастичность:
   - $\mathsf{D}\widehat{\theta} = \sigma^2 \left(X^T X\right)^{-1}$;
   - $\widehat{\theta}$ — оптимальная оценка θ среди линейных по y;
   - $\widehat{\sigma}^2 = RSS(\widehat{\theta})/(n-d)$ — несмещенная оценка σ^2.
3. $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ — гауссовская линейная модель:
   - МНК совпадает с ОМП для θ;
   - $\widehat{\theta}$ — оптимальная оценка θ;
   - $RSS(\widehat{\theta})/\sigma^2 \sim \chi^2_{n-d}$.

Метод наименьших квадратов

Задача: $RSS(\theta) = \|Y - X\theta\|_2^2 \to \min\limits_{\theta \in \mathbb{R}^d}$.

Решение: $\widehat{\theta} = \left(X^T X\right)^{-1} X^T Y$ если $rank\ X = d$.

Предположения и следствия:

1. $\mathsf{E}\varepsilon = 0$ — несмещенность:
   - $\widehat{\theta}$ — несмещенная оценка θ.
2. $\mathsf{E}\varepsilon = 0$ и $\mathsf{D}\varepsilon = \sigma^2 I_n$ — несмещенность и гомоскедастичность:
   - $\mathsf{D}\widehat{\theta} = \sigma^2 \left(X^T X\right)^{-1}$;
   - $\widehat{\theta}$ — оптимальная оценка θ среди линейных по y;
   - $\widehat{\sigma}^2 = RSS(\widehat{\theta})/(n-d)$ — несмещенная оценка σ^2.
3. $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ — гауссовская линейная модель:
   - МНК совпадает с ОМП для θ;
   - $\widehat{\theta}$ — оптимальная оценка θ;
   - $RSS(\widehat{\theta})/\sigma^2 \sim \chi^2_{n-d}$.

Реализация в sklearn

```
m = sklearn.linear_model.LinearRegression(fit_intercept=True)
```

Обучение модели:

```
m.fit(X, Y)
```

Вектор коэффициентов:

```
m.coef_
```

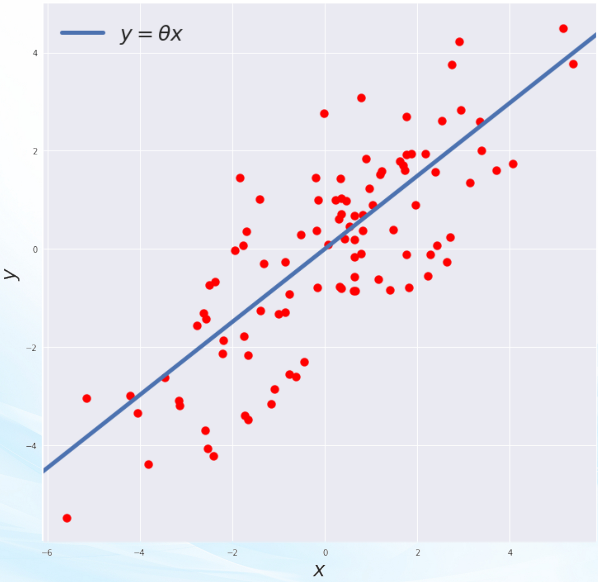
Свободный коэффициент:

```
m.intercept_
```

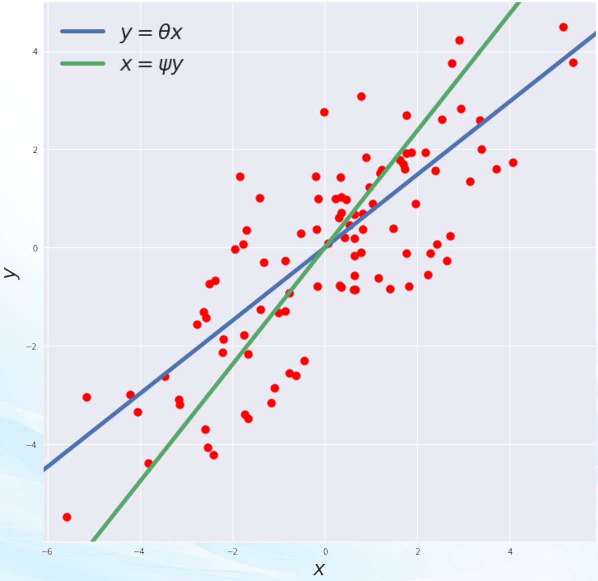
Предсказания:

```
m.predict(X)
```

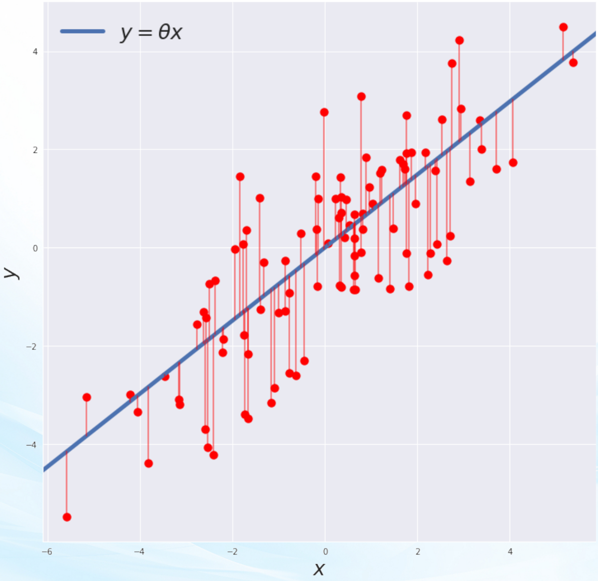
Инверсия



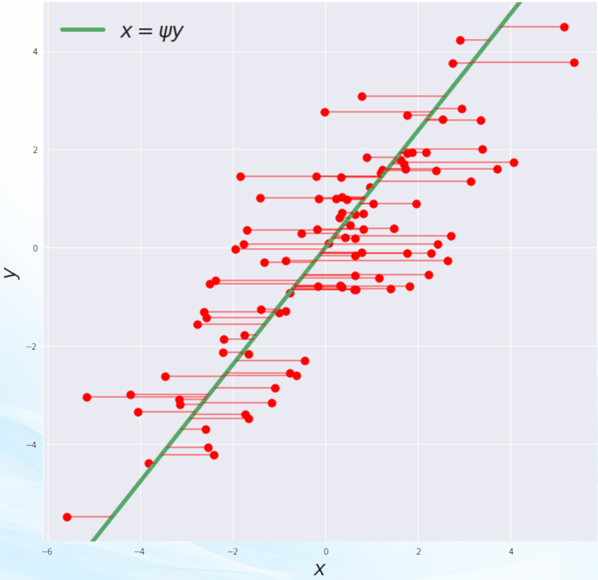
Инверсия



Инверсия



Инверсия



Качество модели

$$R^2 = 1 - \frac{RSS(\widehat{\theta})}{\left\| Y - \overline{Y} \right\|_2^2}$$ — коэффициент детерминации.

$R^2 \approx 1$ — хорошо, $R^2 \approx 0$ — плохо.

Качество модели

$$R^2 = 1 - \frac{RSS(\widehat{\theta})}{\left\| Y - \overline{Y} \right\|_2^2}$$ — коэффициент детерминации.

$R^2 \approx 1$ — хорошо, $R^2 \approx 0$ — плохо.

Проблема: чем больше признаков, тем больше значение R^2.

Качество модели

$$R^2 = 1 - \frac{RSS(\widehat{\theta})}{\left\| Y - \overline{Y} \right\|_2^2}$$ — коэффициент детерминации.

$R^2 \approx 1$ — хорошо, $R^2 \approx 0$ — плохо.

Проблема: чем больше признаков, тем больше значение R^2.

Для отбора признаков можно использовать подправленный R^2:

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

Качество модели

$$R^2 = 1 - \frac{RSS(\widehat{\theta})}{\|Y - \overline{Y}\|_2^2}$$ — коэффициент детерминации.

$R^2 \approx 1$ — хорошо, $R^2 \approx 0$ — плохо.

Проблема: чем больше признаков, тем больше значение R^2.

Для отбора признаков можно использовать подправленный R^2:

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

Информационные критерии: (*чем меньше, тем лучше*)

$$AIC = \frac{2d}{n} + \ln \frac{RSS(\widehat{\theta})}{n}$$ — Акаике, $$BIC = \frac{d \ln n}{n} + \ln \frac{RSS(\widehat{\theta})}{n}$$ — Шварца.

Качество модели

$$R^2 = 1 - \frac{RSS(\widehat{\theta})}{\left\| Y - \overline{Y} \right\|_2^2} \text{ — коэффициент детерминации.}$$

$R^2 \approx 1$ — хорошо, $R^2 \approx 0$ — плохо.

Проблема: чем больше признаков, тем больше значение R^2.

Для отбора признаков можно использовать подправленный R^2:

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

Информационные критерии: (*чем меньше, тем лучше*)

$$AIC = \frac{2d}{n} + \ln \frac{RSS(\widehat{\theta})}{n} \text{ — Акаике,} \quad BIC = \frac{d \ln n}{n} + \ln \frac{RSS(\widehat{\theta})}{n} \text{ — Шварца.}$$

Качество модели

Качество модели обычно считают на тестовом множестве наблюдений, которые не участвовали в обучении.

$X_{test} \in \mathbb{R}^{k \times d}$, $Y_{test} \in \mathbb{R}^{k}$ — тестовое множество

Качество модели

Качество модели обычно считают на тестовом множестве наблюдений, которые не участвовали в обучении.

$X_{test} \in \mathbb{R}^{k \times d}$, $Y_{test} \in \mathbb{R}^{k}$ — тестовое множество

$$MSE = \frac{1}{k}\sum_{i=1}^{k}\left(Y_i^{test} - \widehat{y}(X_i^{test})\right)^2$$

Качество модели

Качество модели обычно считают на тестовом множестве наблюдений, которые не участвовали в обучении.

$X_{test} \in \mathbb{R}^{k \times d}$, $Y_{test} \in \mathbb{R}^{k}$ — тестовое множество

$$MSE = \frac{1}{k} \sum_{i=1}^{k} \left(Y_i^{test} - \widehat{y}(X_i^{test}) \right)^2$$

$$MAE = \frac{1}{k} \sum_{i=1}^{k} \left| Y_i^{test} - \widehat{y}(X_i^{test}) \right|$$

Качество модели

Качество модели обычно считают на тестовом множестве наблюдений, которые не участвовали в обучении.

$X_{test} \in \mathbb{R}^{k \times d}$, $Y_{test} \in \mathbb{R}^{k}$ — тестовое множество

$$MSE = \frac{1}{k}\sum_{i=1}^{k}\left(Y_i^{test} - \widehat{y}(X_i^{test})\right)^2$$

$$MAE = \frac{1}{k}\sum_{i=1}^{k}\left|Y_i^{test} - \widehat{y}(X_i^{test})\right|$$

$$MAPE = \frac{100\%}{k}\sum_{i=1}^{k}\left|\frac{Y_i^{test} - \widehat{y}(X_i^{test})}{Y_i^{test}}\right|$$

Реализация метрик: `sklearn.metrics`

Качество модели

Качество модели обычно считают на тестовом множестве наблюдений, которые не участвовали в обучении.

$X_{test} \in \mathbb{R}^{k \times d}$, $Y_{test} \in \mathbb{R}^{k}$ — тестовое множество

$$MSE = \frac{1}{k} \sum_{i=1}^{k} \left(Y_i^{test} - \widehat{y}(X_i^{test}) \right)^2$$

$$MAE = \frac{1}{k} \sum_{i=1}^{k} \left| Y_i^{test} - \widehat{y}(X_i^{test}) \right|$$

$$MAPE = \frac{100\%}{k} \sum_{i=1}^{k} \left| \frac{Y_i^{test} - \widehat{y}(X_i^{test})}{Y_i^{test}} \right|$$

Реализация метрик: `sklearn.metrics`

Гауссовская линейная модель

Величина	Доверительный интервал
σ	$\left(\sqrt{RSS(\widehat{\theta})\Big/\chi^2_{n-d,1-\alpha/2}},\quad \sqrt{RSS(\widehat{\theta})\Big/\chi^2_{n-d,\alpha/2}}\right)$
Дов. интервал для размера шума в отклике	

Величина	**Доверительный интервал**
σ	$\left(\sqrt{RSS(\widehat{\theta}) \Big/ \chi^2_{n-d,1-\alpha/2}},\ \sqrt{RSS(\widehat{\theta}) \Big/ \chi^2_{n-d,\alpha/2}}\right)$
Дов. интервал для размера шума в отклике	
θ_j	$\left(\widehat{\theta}_j \ \pm\ T_{n-d,1-\alpha/2} \cdot \widehat{\sigma}\sqrt{(X^TX)^{-1}_{jj}}\right)$
Дов. интервал для коэффициента перед j-м признаком	

Величина	**Доверительный интервал**
σ	$\left(\sqrt{RSS(\widehat{\theta}) \Big/ \chi^2_{n-d,1-\alpha/2}},\ \sqrt{RSS(\widehat{\theta}) \Big/ \chi^2_{n-d,\alpha/2}}\right)$
Дов. интервал для размера шума в отклике	
θ_j	$\left(\widehat{\theta}_j \pm T_{n-d,1-\alpha/2} \cdot \widehat{\sigma}\sqrt{(X^TX)^{-1}_{jj}}\right)$
Дов. интервал для коэффициента перед j-м признаком	
$x_0^T\theta$	$\left(x_0^T\widehat{\theta} \pm T_{n-d,1-\alpha/2} \cdot \widehat{\sigma}\sqrt{x_0^T\,(X^TX)^{-1}\,x_0}\right)$
Дов. интервал для **среднего** отклика на объекте x_0	

Величина	Доверительный интервал
σ	$\left(\sqrt{RSS(\widehat{\theta})\Big/\chi^2_{n-d,1-\alpha/2}},\ \sqrt{RSS(\widehat{\theta})\Big/\chi^2_{n-d,\alpha/2}}\right)$
Дов. интервал для размера шума в отклике	
θ_j	$\left(\widehat{\theta}_j \pm T_{n-d,1-\alpha/2}\cdot\widehat{\sigma}\sqrt{(X^TX)^{-1}_{jj}}\right)$
Дов. интервал для коэффициента перед j-м признаком	
$x_0^T\theta$	$\left(x_0^T\widehat{\theta} \pm T_{n-d,1-\alpha/2}\cdot\widehat{\sigma}\sqrt{x_0^T\left(X^TX\right)^{-1}x_0}\right)$
Дов. интервал для **среднего** отклика на объекте x_0	
$x_0^T\theta+\varepsilon$	$\left(x_0^T\widehat{\theta} \pm T_{n-d,1-\alpha/2}\cdot\widehat{\sigma}\sqrt{1+x_0^T\left(X^TX\right)^{-1}x_0}\right)$
Предск. интервал для **наблюдаемого** отклика на объекте x_0	

Величина	Доверительный интервал
σ	$\left(\sqrt{RSS(\widehat{\theta}) \Big/ \chi^2_{n-d,1-\alpha/2}},\ \sqrt{RSS(\widehat{\theta}) \Big/ \chi^2_{n-d,\alpha/2}}\right)$
Дов. интервал для размера шума в отклике	
θ_j	$\left(\widehat{\theta}_j \pm T_{n-d,1-\alpha/2} \cdot \widehat{\sigma}\sqrt{(X^TX)^{-1}_{jj}}\right)$
Дов. интервал для коэффициента перед j-м признаком	
$x_0^T\theta$	$\left(x_0^T\widehat{\theta} \pm T_{n-d,1-\alpha/2} \cdot \widehat{\sigma}\sqrt{x_0^T\,(X^TX)^{-1}\,x_0}\right)$
Дов. интервал для **среднего** отклика на объекте x_0	
$x_0^T\theta + \varepsilon$	$\left(x_0^T\widehat{\theta} \pm T_{n-d,1-\alpha/2} \cdot \widehat{\sigma}\sqrt{1 + x_0^T\,(X^TX)^{-1}\,x_0}\right)$
Предск. интервал для **наблюдаемого** отклика на объекте x_0	

Примечание: если $\mathsf{P}_{a,b}$ — некоторое распределение с параметрами a и b, то $\mathsf{P}_{a,b,\alpha}$ — его α квантиль

Величина	Доверительный интервал
σ	$\left(\sqrt{RSS(\widehat{\theta})\Big/\chi^2_{n-d,1-\alpha/2}},\ \ \sqrt{RSS(\widehat{\theta})\Big/\chi^2_{n-d,\alpha/2}}\right)$
Дов. интервал для размера шума в отклике	
θ_j	$\left(\widehat{\theta}_j \ \pm\ T_{n-d,1-\alpha/2}\cdot\widehat{\sigma}\sqrt{(X^TX)^{-1}_{jj}}\right)$
Дов. интервал для коэффициента перед j-м признаком	
$x_0^T\theta$	$\left(x_0^T\widehat{\theta}\ \pm\ T_{n-d,1-\alpha/2}\cdot\widehat{\sigma}\sqrt{x_0^T\,(X^TX)^{-1}\,x_0}\right)$
Дов. интервал для **среднего** отклика на объекте x_0	
$x_0^T\theta+\varepsilon$	$\left(x_0^T\widehat{\theta}\ \pm\ T_{n-d,1-\alpha/2}\cdot\widehat{\sigma}\sqrt{1+\ x_0^T\,(X^TX)^{-1}\,x_0}\right)$
Предск. интервал для **наблюдаемого** отклика на объекте x_0	

Примечание: если $\mathsf{P}_{a,b}$ — некоторое распределение с параметрами a и b, то $\mathsf{P}_{a,b,\alpha}$ — его α квантиль

Типы гипотез

1. Значим ли признак j?

   $H_0 \colon \theta_j = 0$

Типы гипотез

1. Значим ли признак j?

   $H_0: \theta_j = 0$

2. Значима ли регрессия вообще?

   $H_0: \theta_2 = \ldots = \theta_d = 0$ если $x_1 = 1$

Типы гипотез

1. Значим ли признак j?

   $H_0: \theta_j = 0$

2. Значима ли регрессия вообще?

   $H_0: \theta_2 = ... = \theta_d = 0$ если $x_1 = 1$

1. Значим ли признак j?

Гипотеза о незначимости коэффициента θ_j

$\mathrm{H}_0: \theta_j = 0$ *vs.* $\mathrm{H}_1: \theta_j \; \{<, \neq, >\} \; 0$

1. Значим ли признак j?

Гипотеза о незначимости коэффициента θ_j

$\mathrm{H}_0 : \theta_j = 0$ *vs.* $\mathrm{H}_1 : \theta_j \; \{<, \neq, >\} \; 0$

Критерий Стьюдента

$$T_j^0(X, Y) = \frac{\widehat{\theta}_j}{\widehat{\sigma}\sqrt{(X^T X)^{-1}_{jj}}} \overset{\mathrm{H}_0}{\sim} T_{n-d}$$

$T_j^0(X, Y)$ — T-статистика критерия

1. Значим ли признак j?

Гипотеза о незначимости коэффициента θ_j

$\mathrm{H}_0: \theta_j = 0$ *vs.* $\mathrm{H}_1: \theta_j \ \{<, \neq, >\}\ 0$

Критерий Стьюдента

$$T_j^0(X, Y) = \frac{\widehat{\theta}_j}{\widehat{\sigma}\sqrt{(X^TX)^{-1}_{jj}}} \overset{\mathrm{H}_0}{\sim} T_{n-d}$$

$T_j^0(X, Y)$ — Т-статистика критерия

Для $\mathrm{H}_1: \theta_j \neq 0$ критерий $\left\{\left|T_j^0(X, y)\right| > T_{n-d, 1-\alpha/2}\right\}$,

где $T_{n-d,1-\alpha/2}$ — $(1-\alpha/2)$ квантиль распределения T_{n-d}

1. Значим ли признак j?

Гипотеза о незначимости коэффициента θ_j

$\mathrm{H}_0\colon \theta_j = 0$ *vs.* $\mathrm{H}_1\colon \theta_j \ \{<, \neq, >\}\ 0$

Критерий Стьюдента

$$T_j^0(X, Y) = \frac{\widehat{\theta}_j}{\widehat{\sigma}\sqrt{(X^T X)^{-1}_{jj}}} \overset{\mathrm{H}_0}{\sim} T_{n-d}$$

$T_j^0(X, Y)$ — T-статистика критерия

Для $\mathrm{H}_1\colon \theta_j \ \neq\ 0$ критерий $\left\{\left|T_j^0(X, y)\right| > T_{n-d,1-\alpha/2}\right\}$,

где $T_{n-d,1-\alpha/2}$ — $(1-\alpha/2)$ квантиль распределения T_{n-d}

Если H_0 не отвергается, то можно считать, что θ_j отклоняется от нуля статистически незначимо. Возможно, признак j стоит убрать.

1. Значим ли признак j?

Гипотеза о незначимости коэффициента θ_j

$\mathsf{H}_0 : \theta_j = 0$ *vs.* $\mathsf{H}_1 : \theta_j \ \{<, \neq, >\} \ 0$

Критерий Стьюдента

$$T_j^0(X, Y) = \frac{\widehat{\theta}_j}{\widehat{\sigma}\sqrt{(X^T X)^{-1}_{jj}}} \overset{\mathsf{H}_0}{\sim} T_{n-d}$$

$T_j^0(X, Y)$ — T-статистика критерия

Для $\mathsf{H}_1 : \theta_j \neq 0$ критерий $\left\{\left|T_j^0(X, y)\right| > T_{n-d, 1-\alpha/2}\right\}$,

где $T_{n-d, 1-\alpha/2}$ — $(1 - \alpha/2)$ квантиль распределения T_{n-d}

Если H_0 не отвергается, то можно считать, что θ_j отклоняется от нуля статистически незначимо. Возможно, признак j стоит убрать.

2. Значима ли регрессия вообще?

Гипотеза о незначимости регрессии

$\mathsf{H}_0\colon \theta_1 = \theta_2 = \dots = \theta_d = 0 \;\; vs. \;\; \mathsf{H}_1\colon \exists\, \theta_j \neq 0$

2. Значима ли регрессия вообще?

Гипотеза о незначимости регрессии

$\mathrm{H}_0: \theta_1 = \theta_2 = \cdots = \theta_d = 0 \;\; vs. \;\; \mathrm{H}_1: \exists\, \theta_j \neq 0$

Критерий Фишера

$$F(X, Y) = \frac{R^2/(d-1)}{(1-R^2)/(n-d)} \overset{\mathrm{H}_0}{\sim} F_{d-1,n-d},$$

где $R^2 = 1 - \frac{RSS(\widehat{\theta})}{\|Y - \overline{Y}\|_2^2}$ — коэффициент детерминации.

$F(X, Y)$ — F-статистика критерия

2. Значима ли регрессия вообще?

Гипотеза о незначимости регрессии

$\mathrm{H}_0: \theta_1 = \theta_2 = \dots = \theta_d = 0 \quad vs. \quad \mathrm{H}_1: \exists\, \theta_j \neq 0$

Критерий Фишера

$$F(X, Y) = \frac{R^2/(d-1)}{(1-R^2)/(n-d)} \overset{\mathrm{H}_0}{\sim} F_{d-1,n-d},$$

где $R^2 = 1 - \frac{RSS(\widehat{\theta})}{\|Y - \overline{Y}\|_2^2}$ — коэффициент детерминации.

$F(X, Y)$ — F-статистика критерия

Для $\mathrm{H}_1: \exists\, \theta_j \neq 0$ критерий $\left\{F_j^0(X, y) > F_{d-1,n-d,1-\alpha}\right\}$,

где $F_{d-1,n-d,1-\alpha}$ — $(1-\alpha)$ квантиль распределения $F_{d-1,n-d}$

2. Значима ли регрессия вообще?

Гипотеза о незначимости регрессии

$H_0: \theta_1 = \theta_2 = \dots = \theta_d = 0$ *vs.* $H_1: \exists\, \theta_j \neq 0$

Критерий Фишера

$$F(X, Y) = \frac{R^2/(d-1)}{(1-R^2)/(n-d)} \overset{H_0}{\sim} F_{d-1,n-d},$$

где $R^2 = 1 - \frac{RSS(\widehat{\theta})}{\left\|Y - \overline{Y}\right\|_2^2}$ — коэффициент детерминации.

$F(X, Y)$ — F-статистика критерия

Для $H_1: \exists\, \theta_j \neq 0$ критерий $\left\{F_j^0(X, y) > F_{d-1,n-d,1-\alpha}\right\}$,

где $F_{d-1,n-d,1-\alpha}$ — $(1-\alpha)$ квантиль распределения $F_{d-1,n-d}$

Если H_0 не отвергается, то можно считать, что регресия

статистически незначимо отличается от приближения константой.

2. Значима ли регрессия вообще?

Гипотеза о незначимости регрессии

$H_0: \theta_1 = \theta_2 = \dots = \theta_d = 0$ *vs.* $H_1: \exists\, \theta_j \neq 0$

Критерий Фишера

$$F(X, Y) = \frac{R^2/(d-1)}{(1-R^2)/(n-d)} \overset{H_0}{\sim} F_{d-1,n-d},$$

где $R^2 = 1 - \frac{RSS(\widehat{\theta})}{\left\|Y - \overline{Y}\right\|_2^2}$ — коэффициент детерминации.

$F(X, Y)$ — F-статистика критерия

Для $H_1: \exists\, \theta_j \neq 0$ критерий $\left\{F_j^0(X, y) > F_{d-1,n-d,1-\alpha}\right\}$,

где $F_{d-1,n-d,1-\alpha}$ — $(1-\alpha)$ квантиль распределения $F_{d-1,n-d}$

Если H_0 не отвергается, то можно считать, что регресия статистически незначимо отличается от приближения константой.

Таблица в statsmodels

```
# Load data
In [4]: dat = sm.datasets.get_rdataset("Guerry", "HistData").data

# Fit regression model (using the natural log of one of the regressors)
In [5]: results = smf.ols('Lottery ~ Literacy + np.log(Pop1831)', data=dat).fit()

# Inspect the results
In [6]: print(results.summary())
                            OLS Regression Results
==============================================================================
Dep. Variable:                Lottery   R-squared:                       0.348
Model:                            OLS   Adj. R-squared:                  0.333
Method:                 Least Squares   F-statistic:                     22.20
Date:                Mon, 14 May 2018   Prob (F-statistic):           1.90e-08
Time:                        21:48:09   Log-Likelihood:                -379.82
No. Observations:                  86   AIC:                             765.6
Df Residuals:                      83   BIC:                             773.0
Df Model:                           2
Covariance Type:            nonrobust
===================================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------
Intercept         246.4341     35.233      6.995      0.000     176.358     316.510
Literacy           -0.4889      0.128     -3.832      0.000      -0.743      -0.235
np.log(Pop1831)   -31.3114      5.977     -5.239      0.000     -43.199     -19.424
```


ВСЁ!