

Статистический взгляд на линейную регрессию: прогнозы и интерпретация

Парная регрессия

- Есть данные про две переменные: x, y
- Надо оценить как одна переменная зависит от другой

Задача: понять как меняются результаты школьников (баллы за тест) в зависимости от размера класса, есть ли между этими переменными значимая связь?

- Один из самых простых подходов: оценить линейную регрессию

Парная регрессия

Задача: понять как меняются результаты школьников (баллы за тест) в зависимости от размера класса, есть ли между этими переменными значимая связь?

Модель: $y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i$

x_i — среднее значение размера класса в i -ом округе

y_i — среднее значение за тест в i -ом округе

ε_i — прочие факторы, влияющие на результаты обучения в i -ом округе

β_0, β_1 — коэффициенты линейной регрессии

Как обучаем

Чтобы понять как переменные зависят, надо обучить модель, в машинном обучении мы использовали для этого какую-нибудь функцию потерь:

$$L(\beta) = MSE(\beta) = \frac{1}{n} \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2 \rightarrow \min_{\beta_0, \beta_1}$$

- ❗ Можно минимизировать любую другую функцию потерь, но у MSE есть ряд хороших статистических свойств

Парная регрессия

Задача: понять как меняются результаты школьников (баллы за тест) в зависимости от размера класса, есть ли между этими переменными значимая связь?

Оценённая модель: $y = \hat{\beta}_0 + \hat{\beta}_1 x$

- Как понять, есть ли между переменными связь?
- Если коэффициент $\hat{\beta}_1$ оказался равен нулю, связи нет

Парная регрессия

Задача: понять как меняются результаты школьников (баллы за тест) в зависимости от размера класса, есть ли между этими переменными значимая связь?

Оценённая модель: $y = \hat{\beta}_0 + \hat{\beta}_1 x$

- Как понять, есть ли между переменными связь?
- Если коэффициент $\hat{\beta}_1$ оказался равен нулю, связи нет

❗ Что значит равен нулю? Эта оценка – случайная величина, нужно проверять гипотезу о равенстве нулю

Когда важна интерпретация

Дискриминация на рынке труда

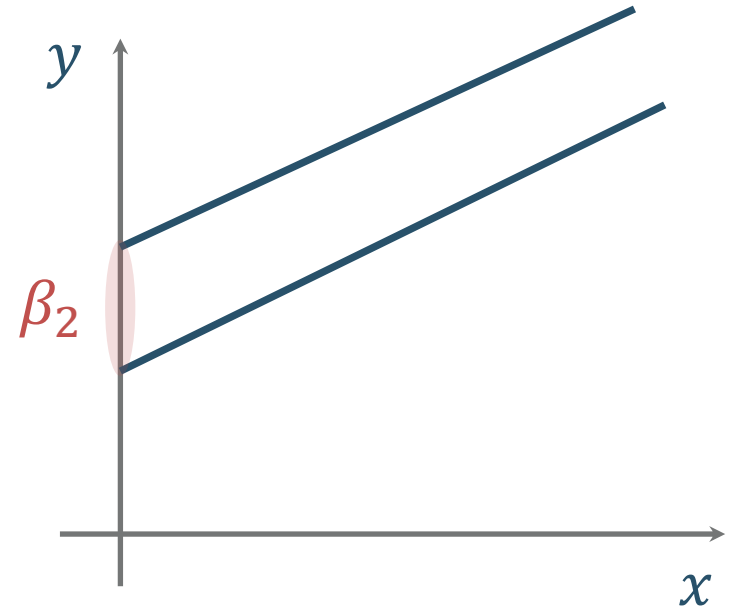
Задача: понять, есть ли на рынке труда дискриминация

Модель: $y_i = \beta_0 + \beta_1 \cdot x_i + \beta_2 \cdot d_i + \varepsilon_i$

y_i — зарплата i -ого работника в долларах в час

x_i — стаж i -го работника в годах

d_i — фиктивная переменная, которая равна единице, если работник мужчина и нулю, если женщина



$H_0: \beta_2 = 0$ Нет дискриминации

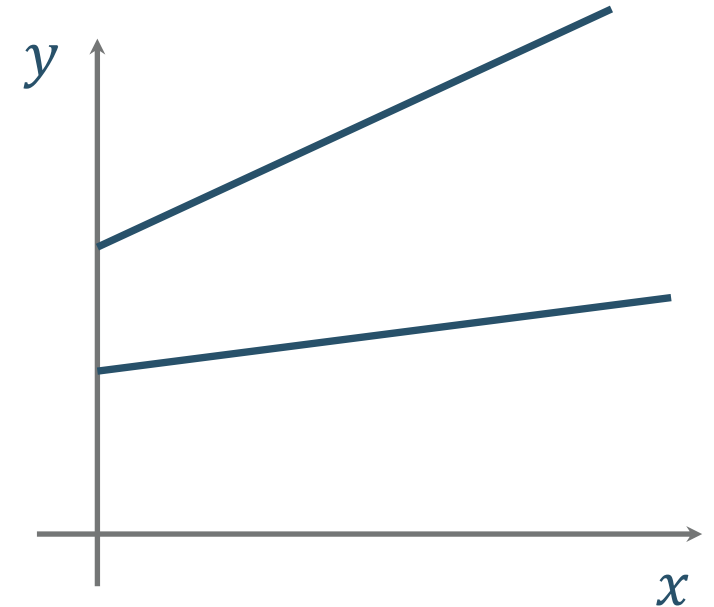
$H_a: \beta_2 \neq 0$ Есть

Дискриминация на рынке труда

Задача: понять, есть ли на рынке труда дискриминация

Модель: $y_i = \beta_0 + \beta_1 \cdot x_i + \beta_2 \cdot d_i + \beta_3 \cdot d_i \cdot x_i + \varepsilon_i$

! Этого недостаточно,
разрыв между
зарплатами может
расти вместе со стажем



$H_0: \beta_2 = 0, \beta_3 = 0$ Нет дискриминации

H_a : хотя бы одно неравенство **Есть**

Дискриминация на рынке труда

Чтобы проверить гипотезу о дискриминации, надо:

- Получить оценки с хорошими статистическими свойствами
 - Аккуратно выбрать союзника для проверки гипотезы
- ❗ Тест отношения правдоподобий может в этом помочь, если с данными нет проблем и его предпосылки не нарушаются

Центральный банк и ставка процента

- Высокая инфляция (рост цен) – плохо влияет на экономику
- ЦБ должен с ней бороться и держать её на стабильно низком уровне
- Рост процентных ставок помогает уменьшить инфляцию

Задача: понять насколько ЦБ должен поднять ставку, чтобы инфляция уменьшилась до целевого уровня

Центральный банк и ставка процента

Задача: понять насколько ЦБ должен поднять ставку, чтобы инфляция уменьшилась до целевого уровня

Модель:

$$y_t = \beta_0 + \beta_1 \cdot x_t + \varepsilon_t$$

y_t — инфляция

x_t — ставка процента

Отражает, насколько изменится инфляция при изменении ставки

- ❗ Если мы получим для коэффициента смещённую оценку, мы неправильно оценим насколько надо поднять ставку \Rightarrow инфляция изменится на непредсказуемую величину

Спрос на овощи

Задача: предсказать сколько овощей купят в разных магазинах большой торговой сети, чтобы не привезти туда лишних и они не испортились

- При решении такой задачи нас волнуют точечный прогнозы для каждого типа овощей и каждого магазина
- Нас не очень интересует, какие именно коэффициенты получатся и какими свойствами они будут обладать

Два великих вопроса

- Как устроен мир?
- Как переменная y зависит от переменной x ?
- Что будет завтра?
- Как удачно спрогнозировать переменную y ?

! Удивительно, но ответы на эти вопросы ищутся по-разному

Два великих вопроса

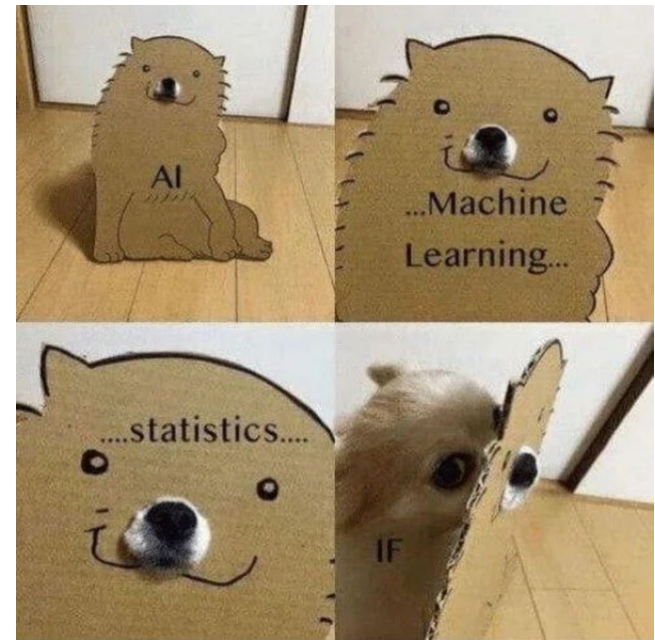
Как устроен мир?

- Важна интерпретация
- Поиск верёвочки, за которую можно дёрнуть

❗ Обе задачи можно решать одними и теми же моделями, например, линейной регрессией, но интересоваться нас будут её разные аспекты

Что будет завтра?

- Интерпретация часто приносится в жертву
- Поиск хорошего прогноза



Эконометрика

- Если нас интересует ответ на вопрос “Как устроен мир?”, найти ответ на него нам может помочь эконометрика
- Она концентрируется на поиске интерпретируемых оценок и проверке гипотез
- Из-за этого в ней идёт яростная борьба за статистические предпосылки
- Концентрируется на MSE из-за удобных статистических свойств, однако ничто не мешает переключиться на любые другие потери и использовать бутстрап

► <https://www.coursera.org/learn/ekonometrika>

Интерпретируемое машинное обучение

- С моделями машинного обучения, которые нельзя проинтерпретировать, возникают проблемы
- Например, дискриминация в банковском секторе
- Законодательное регулирование моделирования
- Тренд на интерпретируемое машинное обучение, разработка алгоритмов, которые могли бы объяснить, что именно происходит внутри чёрного ящика

➤ <https://arxiv.org/abs/1606.08813>

➤ <https://christophm.github.io/interpretable-ml-book/>

Линейная регрессия: статистический подход

Линейная регрессия: предпосылки

Модель:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + \varepsilon_i$$

случайная ошибка
(прочие факторы)

↑
цена
квартиры

↑
площадь

↑
центральный
район (дамми)

↑
расстояние
до метро

Линейная регрессия: предпосылки

Модель:

$$y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{21} + \cdots + \beta_k x_{k1} + \varepsilon_1$$

$$y_2 = \beta_0 + \beta_1 x_{12} + \beta_2 x_{22} + \cdots + \beta_k x_{k2} + \varepsilon_2$$

...

$$y_n = \beta_0 + \beta_1 x_{1n} + \beta_2 x_{2n} + \cdots + \beta_k x_{kn} + \varepsilon_n$$

Линейная регрессия: предпосылки

Модель:

$$y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{21} + \cdots + \beta_k x_{k1} + \varepsilon_1$$

$$y_2 = \beta_0 + \beta_1 x_{12} + \beta_2 x_{22} + \cdots + \beta_k x_{k2} + \varepsilon_2$$

...

$$y_n = \beta_0 + \beta_1 x_{1n} + \beta_2 x_{2n} + \cdots + \beta_k x_{kn} + \varepsilon_n$$

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix} \quad X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_k \end{pmatrix} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{pmatrix}$$

$$y = X \beta + \varepsilon$$

Линейная регрессия: предпосылки

Модель:

$$y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{21} + \cdots + \beta_k x_{k1} + \varepsilon_1$$

$$y_2 = \beta_0 + \beta_1 x_{12} + \beta_2 x_{22} + \cdots + \beta_k x_{k2} + \varepsilon_2$$

...

$$y_n = \beta_0 + \beta_1 x_{1n} + \beta_2 x_{2n} + \cdots + \beta_k x_{kn} + \varepsilon_n$$

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix} \quad X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_k \end{pmatrix} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{pmatrix}$$

$$y = X \beta + \varepsilon$$

Линейная регрессия: предпосылки

Модель:

$$y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{21} + \cdots + \beta_k x_{k1} + \varepsilon_1$$

$$y_2 = \beta_0 + \beta_1 x_{12} + \beta_2 x_{22} + \cdots + \beta_k x_{k2} + \varepsilon_2$$

...

$$y_n = \beta_0 + \beta_1 x_{1n} + \beta_2 x_{2n} + \cdots + \beta_k x_{kn} + \varepsilon_n$$

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix} \quad X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_k \end{pmatrix} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{pmatrix}$$

$$y = X \beta + \varepsilon$$

Метод наименьших квадратов

$$\begin{aligned} L(\beta) &= MSE(y, \hat{y}) = \frac{1}{n} \|y - X\beta\|^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \\ &= \frac{1}{n} \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}))^2 \rightarrow \min_{\beta} \end{aligned}$$

❗ Оценку, полученную минимизацией MSE, обычно называют оценкой наименьших квадратов (ordinary least squares, OLS)

- Легко взять производную, можно получить аналитическое решение:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Предпосылки классической линейной модели:

1. Модель линейна по параметрам и корректно специфицирована: $y = X\beta + \varepsilon$
 - Мы учли все важные переменные
 - Взаимосвязь между переменными правда линейная
- ❗ Если мы пропустили важную переменную либо неверно специфицировали модель, оценки коэффициентов будут неоптимальными

Предпосылки классической линейной модели:

1. Модель линейна по параметрам и корректно специфицирована: $y = X\beta + \varepsilon$
2. Объясняющие переменные x_{ik} детерминированы и линейно независимы
 - В данных нет мультиколлинеарности
 - Наши объясняющие переменные (регрессоры) – неслучайные величины, это упрощает статистические выкладки

Пример: выборка из квартир, разные площади, x . Собираем заново с такими же площадями, x не изменится, а значения y (например, цен) может поменяться

Предпосылки классической линейной модели:

1. Модель линейна по параметрам и корректно специфицирована: $y = X\beta + \varepsilon$
 2. Объясняющие переменные x_{ik} детерминированы и линейно независимы
 3. Математическое ожидание случайных ошибок равно нулю $E(\varepsilon) = 0$
- Прочие факторы могут приводить к отклонению y в любую сторону, но в среднем эти отклонения компенсируют друг-друга

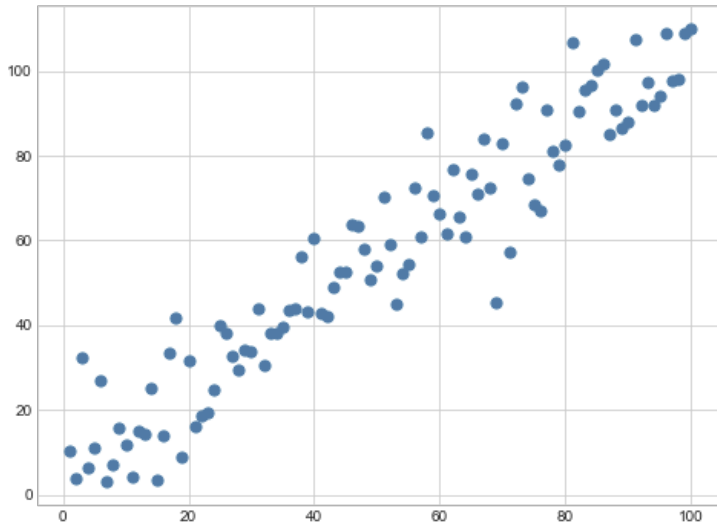
Предпосылки классической линейной модели:

1. Модель линейна по параметрам и корректно специфицирована: $y = X\beta + \varepsilon$
2. Объясняющие переменные x_{ik} детерминированы и линейно независимы
3. Математическое ожидание случайных ошибок равно нулю $E(\varepsilon) = 0$
4. Случайные ошибки, относящиеся к разным наблюдениям независимы и обладают равной дисперсией

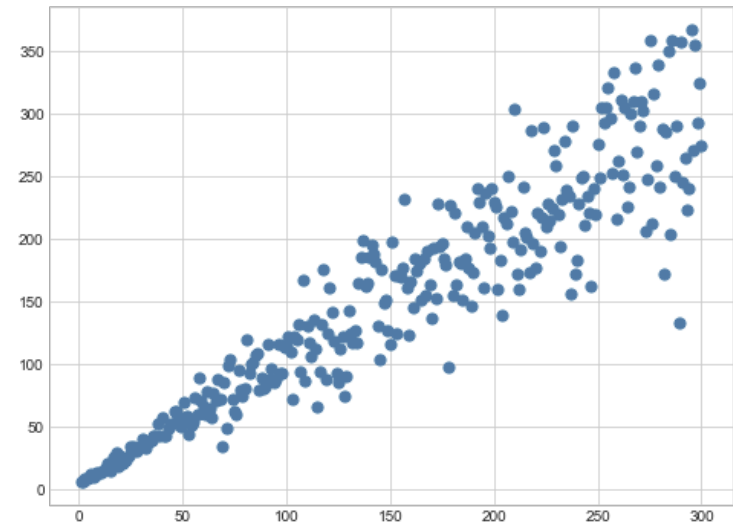
$$Var(\varepsilon) = \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix}$$

Предпосылки классической линейной модели:

- Разброс ошибок должен быть, в среднем, постоянным
- Ошибки не должны коррелировать друг с другом



**Гомоскедастичность
(одноразбросие)**



**Гетероскедастичность
(разноразбросие)**

Реалистичны ли эти предпосылки?



Не очень!

- Однако, если ввести их мы можем понять ряд простых идей и не погрязнуть в технических трудностях
- Дальше от этих предпосылок можно постепенно отказаться и усовершенствовать свой статистический аппарат

Теорема Гаусса-Маркова

Если выполнены предпосылки классической линейной модели, тогда оценка, получаемая минимизацией MSE (оценка наименьших квадратов):

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Несмещённая и эффективная в классе всех несмещённых и линейных по y оценок
(best linear unbiased estimate, *BLUE*)

Простым языком: доверительные интервалы для этой оценки самые узкие, её можно интерпретировать как то, насколько изменится переменная y при изменении x на единицу

Оценка дисперсии

Модель: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i$

- Кроме вектора β , в модели есть ещё один параметр, дисперсия σ^2
- Состоятельной и несмещённой оценкой дисперсии будет:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - k - 1} = \frac{RSS}{n - k - 1}$$

Residual Sum of Squares

Качество модели

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Residual Sum
of Squares

$$ESS = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

Explained Sum
of Squares

$$TSS = \sum_{i=1}^n (y_i - \bar{y})^2$$

Total Sum
of Squares

Качество модели

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

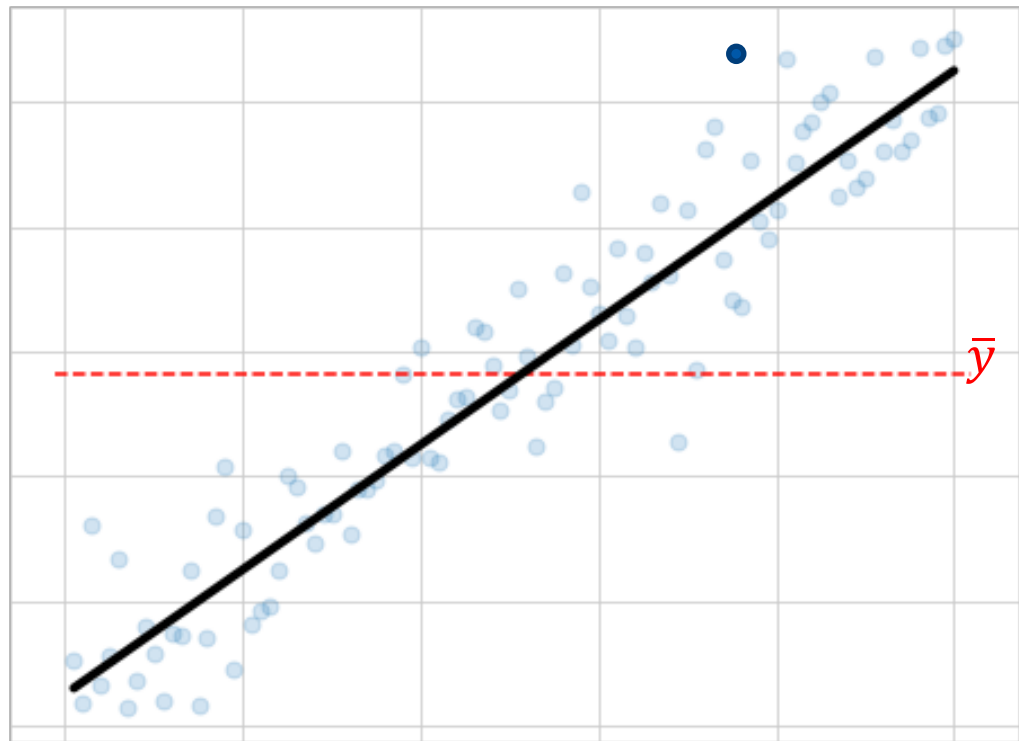
Residual Sum
of Squares

$$ESS = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

Explained Sum
of Squares

$$TSS = \sum_{i=1}^n (y_i - \bar{y})^2$$

Total Sum
of Squares



Качество модели

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

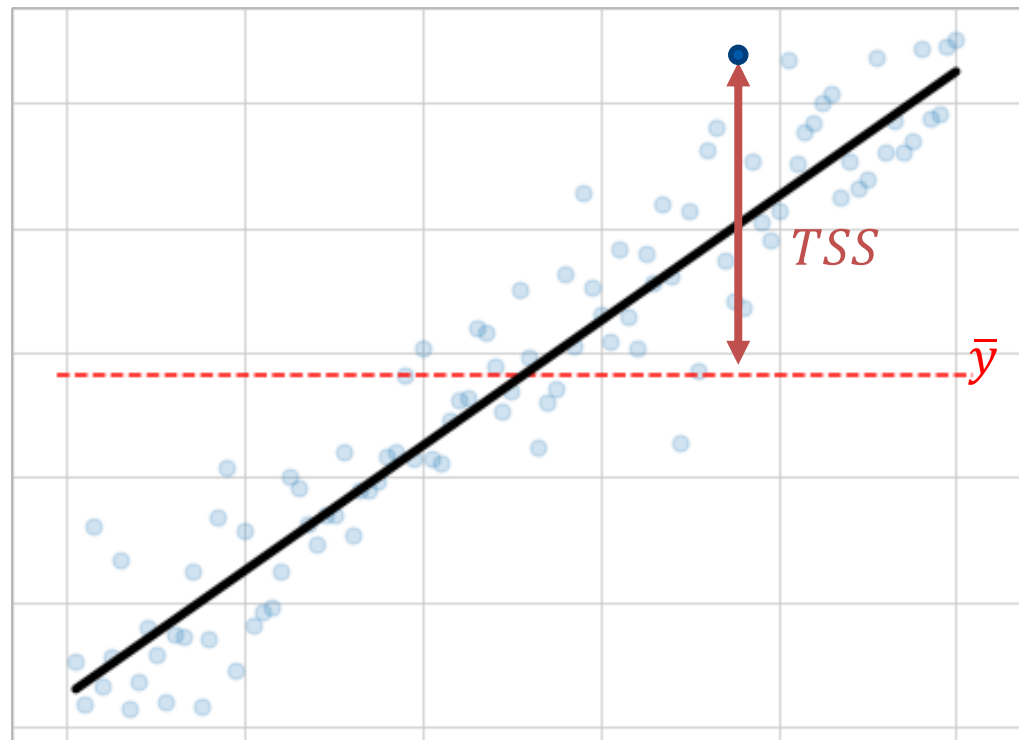
Residual Sum
of Squares

$$ESS = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

Explained Sum
of Squares

$$TSS = \sum_{i=1}^n (y_i - \bar{y})^2$$

Total Sum
of Squares



Качество модели

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

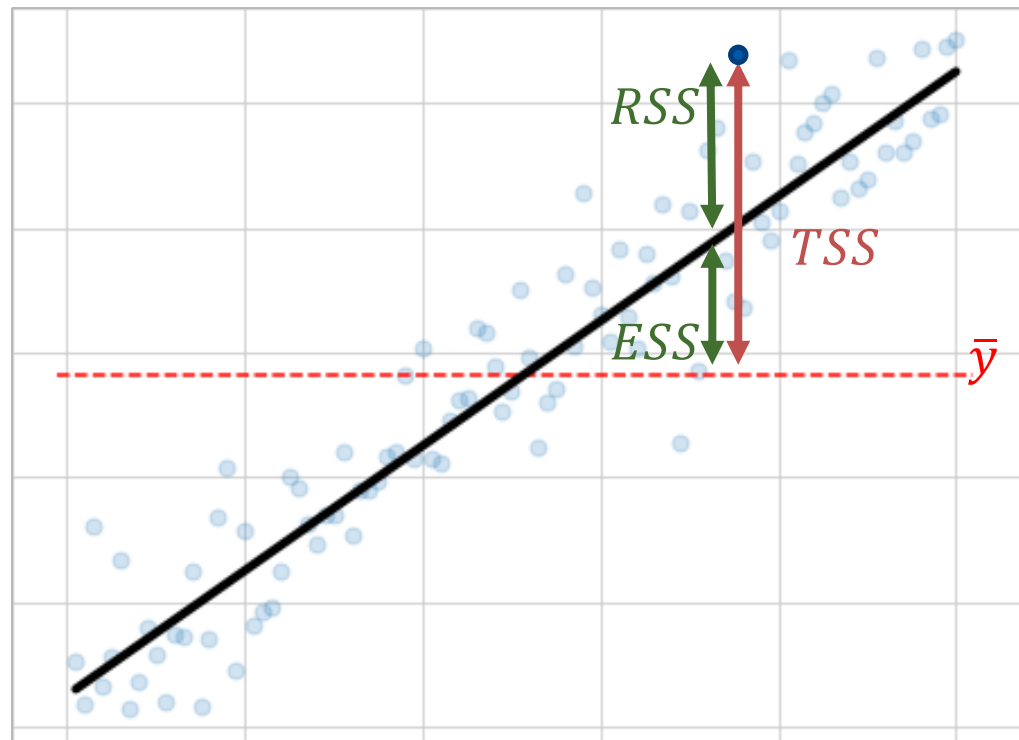
Residual Sum
of Squares

$$ESS = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

Explained Sum
of Squares

$$TSS = \sum_{i=1}^n (y_i - \bar{y})^2$$

Total Sum
of Squares



Качество модели

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Residual Sum
of Squares

$$ESS = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

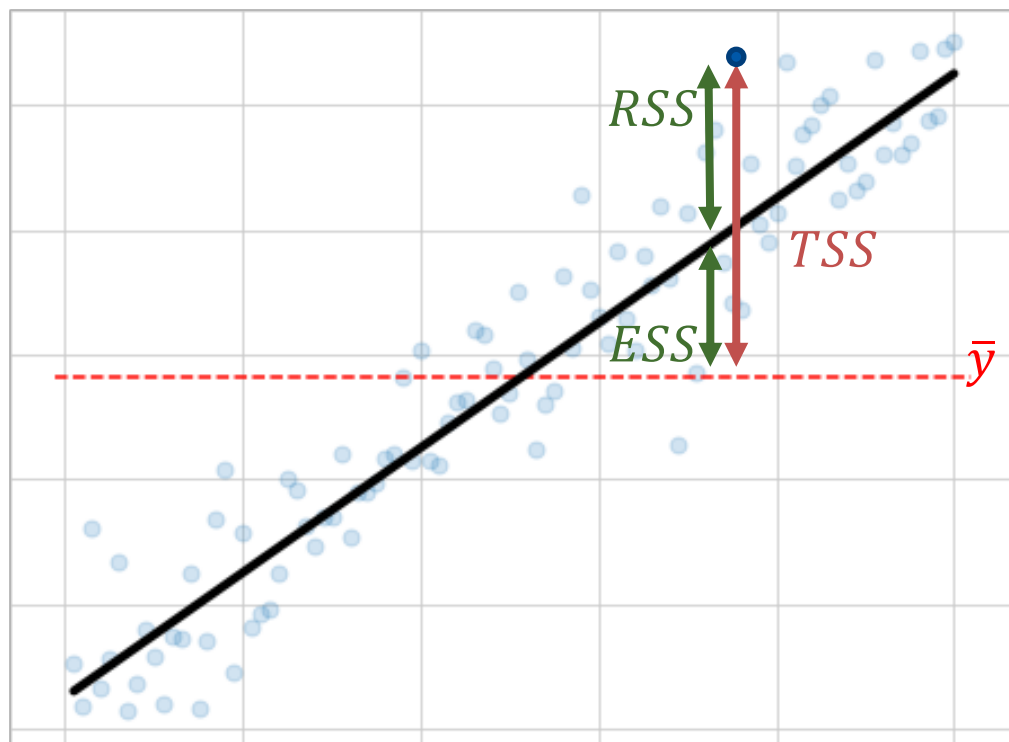
Explained Sum
of Squares

$$TSS = \sum_{i=1}^n (y_i - \bar{y})^2$$

Total Sum
of Squares

$$TSS = RSS + ESS$$

❗ Это равенство
выполнено
только для
моделей с
константой



Качество модели

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Residual Sum
of Squares

$$ESS = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

Explained Sum
of Squares

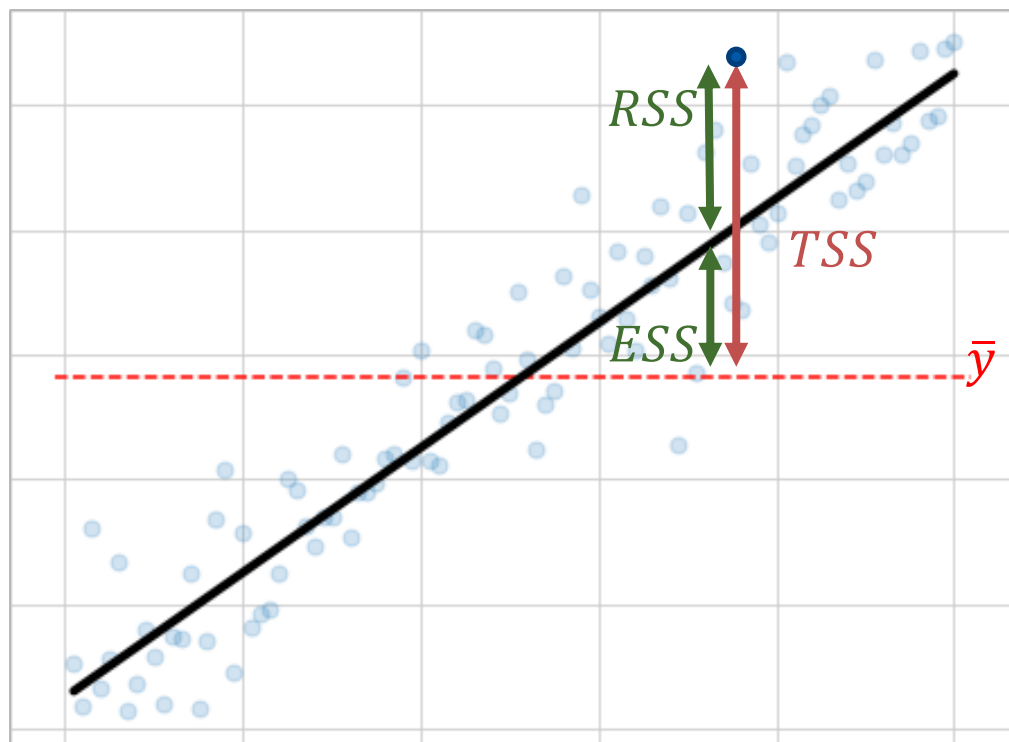
$$TSS = \sum_{i=1}^n (y_i - \bar{y})^2$$

Total Sum
of Squares

$$TSS = RSS + ESS$$

Коэффициент
детерминации:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$



Коэффициент детерминации

- Когда в модели есть константа, коэффициент детерминации лежит между нулём и единицей
- Это не лучшая метрика качества модели, она обладает рядом недостатков
- Например, при добавлении новых регрессоров в модель, она всегда увеличивается, из-за этого вводится “скорректированный” коэффициент детерминации
- Нас R^2 будет интересовать только в рамках проверки гипотез

Линейная регрессия: проверка гипотез

Распределение вектора оценок

- При выполнении предпосылок, минимизация MSE даёт нам несмещённую, эффективную оценку
- Нам этого мало, чтобы тестировать гипотезы, нам нужно знать распределение $\hat{\beta}$
- Гипотезы можно тестировать в предположении о нормальности остатков, либо руководствуясь асимптотикой

Нормальность остатков

Модель: детерминированная часть случайная часть

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i$$

Добавим предположение о нормальности остатков:

$$\varepsilon_i \sim N(0, \sigma^2)$$

$$y_i \sim N(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}, \sigma^2)$$

Нормальность остатков

Модель:

$$y_i = X \beta + \varepsilon$$

Добавим предположение о нормальности остатков:

$$\varepsilon \sim N(0, \sigma^2 \cdot I_n)$$

$$y \sim N(X \beta, \sigma^2 \cdot I_n)$$

Наша МНК-оценка – линейная комбинация нормальных случайных величин

$$\hat{\beta} = (X^T X)^{-1} X^T y \quad \hat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1})$$

Нормальность остатков

Модель:

$$y_i = X \beta + \varepsilon$$

❗ Если снять предположения о нормальности, дисперсии и математические ожидания не изменятся

Добавим предположение о нормальности остатков:

$$\varepsilon \sim N(0, \sigma^2 \cdot I_n)$$

$$y \sim N(X \beta, \sigma^2 \cdot I_n)$$

Наша МНК-оценка – линейная комбинация нормальных случайных величин

$$\hat{\beta} = (X^T X)^{-1} X^T y \quad \hat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1})$$

Гипотеза о незначимости коэффициентов

Модель:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + \varepsilon_i$$

$$H_0: \beta_j = 0$$

$$H_a: \beta_j \neq 0$$

$$t = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$

- Если коэффициент равен нулю, значит он незначим, т.е. между x_j и y нет статистически-значимой связи
- Обычно результат проверки такой гипотезы выводится всеми статистическими пакетами
- Если мы хотим проверить гипотезу о незначимости нескольких коэффициентов, мы сталкиваемся с проблемой множественного тестирования

Гипотеза о незначимости коэффициентов

Модель:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i$$

$$H_0: \beta_2 = 0, \beta_5 = 0, \beta_{15} = 0$$

H_a : хотя бы один коэффициент отличается от нуля

- **Выход 1:** ввести корректировку на множественное тестирование (например, методом Холма)
- **Выход 2:** использовать F-статистику

Гипотеза о незначимости коэффициентов

Модель:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + \varepsilon_i$$

$$H_0: \beta_2 = 0, \beta_5 = 0, \beta_{15} = 0$$

F-тест:

$$F = \frac{R_{UR}^2 - R_R^2}{1 - R_{UR}^2} \cdot \frac{n - k - 1}{q} \sim F_{q, n-k-1}$$

- R_{UR}^2 – коэффициент детерминации в модели без ограничений
- R_R^2 – коэффициент детерминации в модели с ограничениями, q – число ограничений

Гипотеза о незначимости коэффициентов

Модель:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + \varepsilon_i$$

$$H_0: \beta_2 = 0, \beta_5 = 0, \beta_{15} = 0$$

F-тест:

$$F = \frac{R_{UR}^2 - R_R^2}{1 - R_{UR}^2} \cdot \frac{n - k - 1}{q} \sim F_{q, n-k-1}$$

! Без нормальности остатков, если в данных нет выбросов: $q \cdot F \rightarrow \chi_q^2$

Гипотеза о незначимости модели в целом

Модель:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i$$

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

F-тест:

$$F = \frac{R_{UR}^2 - R_R^2}{1 - R_{UR}^2} \cdot \frac{n - k - 1}{q} \sim F_{q, n-k-1}$$

$$F = \frac{R^2}{1 - R^2} \cdot \frac{n - k - 1}{k} \sim F_{k, n-k-1}$$

Резюме

- Если выполнены предпосылки модели классической линейной регрессии, мы получаем ряд хороших свойств
- Это даёт нам возможность проверять гипотезы и находить величину, на которую в среднем изменится y при изменении x

Обзор проблем, возникающих при оценке регрессии

Предпосылки классической линейной модели:

1. Модель линейна по параметрам и корректно специфицирована: $y = X\beta + \varepsilon$
2. Объясняющие переменные x_{ik} детерминированы и линейно независимы
3. Математическое ожидание случайных ошибок равно нулю $\mathbb{E}(\varepsilon) = 0$
4. Случайные ошибки, относящиеся к разным наблюдениям независимы и обладают равной дисперсией

$$Var(\varepsilon) = \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix}$$

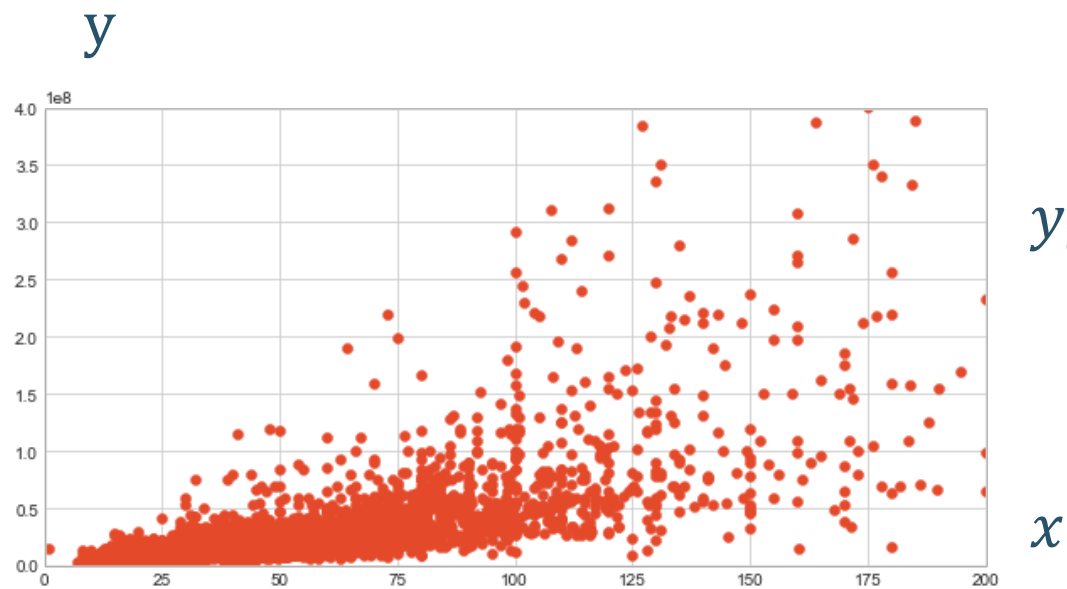
Нелинейность

1. Модель линейна по параметрам и корректно специфицирована: $y = X\beta + \varepsilon$

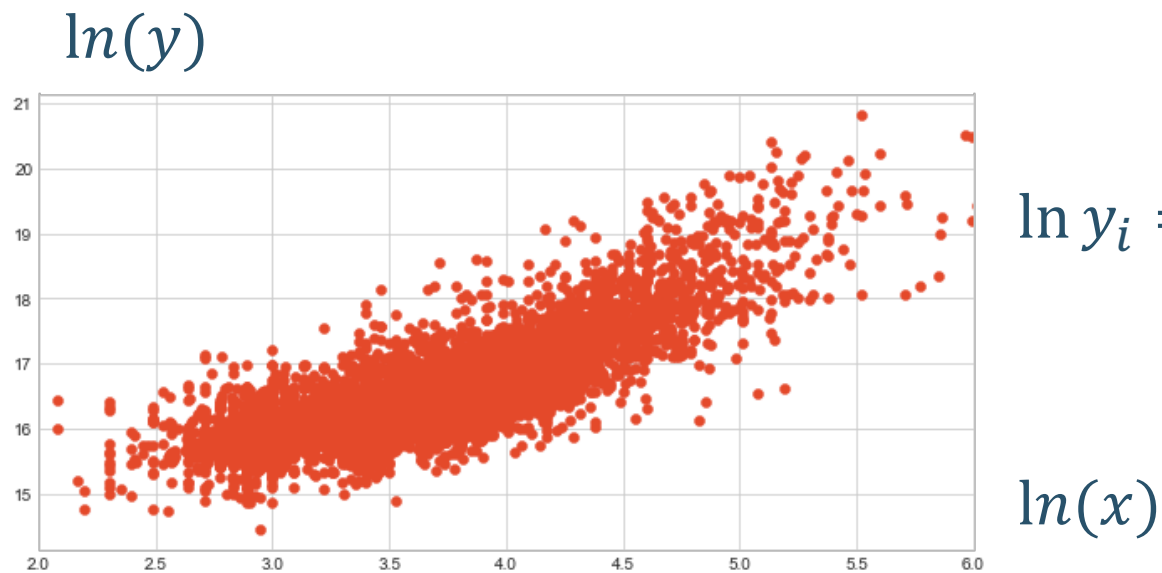
❗ В точности не выполняется никогда, все модели неверны. При сильных отклонениях от линейности оценки смещены и несостоятельны.

Решение: Графический анализ, различные тесты на спецификацию модели (тест Рамсея)

Линеаризация зависимости



$$y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_t$$



$$\ln y_i = \beta_0 + \beta_1 \cdot \ln x_i + \varepsilon_t$$

Мультиколлинеарность

2. Объясняющие переменные x_{ik} детерминированы и линейно независимы

! Если переменные зависимы, возникает проблема мультиколлинеарности, мы не можем найти МНК-оценку, так как определитель матрицы $X^T X$ оказывается близок к нулю

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Решение: следить, чтобы среди регрессоров не было переменных, связь между которыми близка к линейной

Случайные регрессоры

2. Объясняющие переменные x_{ik} детерминированы и линейно независимы

! От предпосылки, что x_{ik} детерминированы обычно отказываются и рассматривают модель со случайными регрессорами

- Доказательства теорем из-за этого становятся более сложными
- На вектор ошибок накладывается дополнительное ограничение $\text{Cov}(\varepsilon_i, x_j) = 0$ либо $\mathbb{E}(\varepsilon_i | x_j) = 0$
- Если эта предпосылка нарушена, говорят о **проблеме эндогенности**

Эндогенность

- Если $\text{Cov}(\varepsilon_i, x_j) \neq 0$, значит среди “прочих” факторов есть такие, которые связаны с x_j
- Можно показать, что это приводит к несостоятельным и смещённым оценкам коэффициентов
- Эндогенность может возникать из-за разных причин:
 1. наблюдаемая пропущенная переменная
 2. ненаблюдаемая пропущенная переменная
 3. ошибки измерения
 4. двухсторонняя причинно-следственная связь

Эндогенность

Решение: разработка более сложных статистических процедур, которые помогут получить состоятельные несмещённые оценки (или хотя бы просто состоятельные оценки):

- Метод инструментальных переменных
- Двухшаговый МНК
- Панельные данные

Математическое ожидание ошибок

3. Математическое ожидание случайных ошибок равно нулю $\mathbb{E}(\varepsilon) = 0$

Условное математическое ожидание случайных ошибок равно нулю $\mathbb{E}(\varepsilon_i | x_j) = 0$

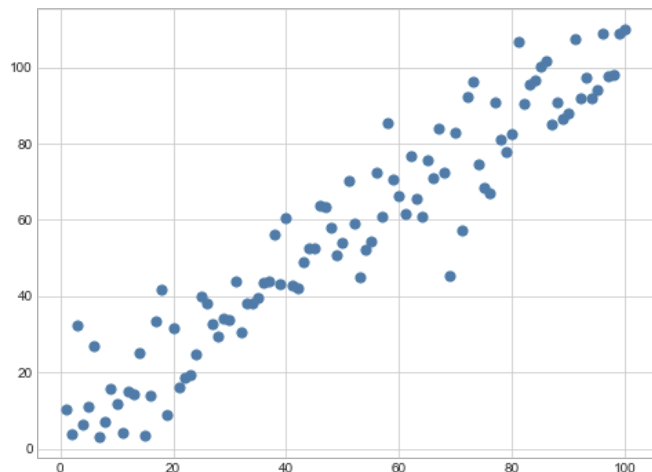


Мы не включили в модель какие-то важные факторы. В условиях стохастических регрессоров мы получаем несостоятельные смещённые оценки.

Гетероскедастичность

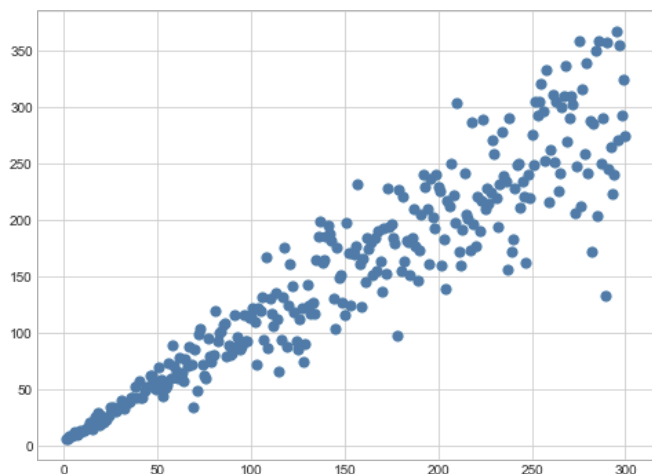
4. Случайные ошибки, относящиеся к разным наблюдениям независимы и обладают равной дисперсией

Гетероскедастичность



❗ Оценки коэффициентов останутся несмещёнными и состоятельными, но перестанут быть эффективными, это приведёт к искажению доверительных интервалов.

Гомоскедастичность



Гетероскедастичность

$$Var(\varepsilon) = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n^2 \end{pmatrix}$$

Автокоррелированность

4. Случайные ошибки, относящиеся к разным наблюдениям независимы и обладают равной дисперсией

! Если это не так, оценки коэффициентов останутся несмещёнными и состоятельными, но перестанут быть эффективными, это приведёт к искажению доверительных интервалов

$$Var(\varepsilon) = \begin{pmatrix} \sigma^2 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & \sigma^2 & \cdots & \rho_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \cdots & \sigma^2 \end{pmatrix}$$

Автокоррелированность и гетероскедастичность

Решение: разработка более сложных статистических процедур, которые скорректируют оценки дисперсий:

- Обобщённый метод наименьших квадратов
- Различные процедуры коррекции оценок дисперсии

Для поиска этих отклонений от стандартных предпосылок разработано довольно много статистических тестов.

Корреляция и причинность

- Наличие значимого коэффициента в модели вовсе не означает причинно-следственной связи между переменными
- Значимый коэффициент означает, что между переменными есть корреляция

Решение: разработка более сложных статистических процедур, которые помогут выявить причинно-следственные связи, а также опора на теорию и здравый смысл

Резюме

- Мы посмотрели на приблизительную схему того, как можно тестировать сложные гипотезы о взаимосвязях между переменными
- На самом деле в этой процедуре есть куча сложностей, проблем и нюансов
- Все они обычно освещаются в курсе эконометрики :)

► <https://www.coursera.org/learn/ekonometrika>

Разложение на смещение и разброс

Теорема Гаусса-Маркова

1. $y = X\beta + \varepsilon$
2. Матрица X детерминирована и её столбцы линейно независимы
3. $\varepsilon \sim (0, \sigma^2 \cdot I_n)$, где I_n – единичная матрица

Тогда оценка, получаемая минимизацией MSE (оценка наименьших квадратов):

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Несмещённая и эффективная в классе всех несмещённых и линейных по y оценок
(best linear unbiased estimate, *BLUE*)

Разложение ошибки на смещение и разброс

Пусть $\hat{\beta}$ – оценка параметра, тогда её квадратичную ошибку можно разложить на смещение и разброс:

$$MSE = \mathbb{E}(\hat{\beta} - \beta)^2 = \text{Var}(\hat{\beta}) + \text{bias}^2(\hat{\beta})$$

Ошибку прогноза тоже можно разложить подобным образом, но в уравнении появится новая компонента: неустраняемая ошибка σ^2 , которая берётся из-за наличия в модели шума ε

$$MSE = \mathbb{E}(\hat{y} - y)^2 = \text{Var}(\hat{y}) + \text{bias}^2(\hat{y}) + \sigma^2$$

Разложение ошибки на смещение и разброс

$$MSE = \mathbb{E}(y - \hat{y})^2 = \sigma^2 + Var(\hat{y}) + bias^2(\hat{y})$$

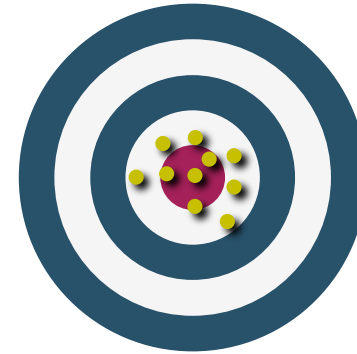
- σ^2 — неустраняемая ошибка
- $Var(\hat{y})$ — разброс, дисперсия прогноза
- $bias(\hat{y})$ — смещение, какой в среднем будет наша ошибка при регулярном использовании модели
- С неустраняемой ошибкой мы ничего не можем сделать, зато можем повлиять на смещение и разброс

Разложение на смещение и разброс

Низкий разброс

Высокий разброс

Низкое
смещение

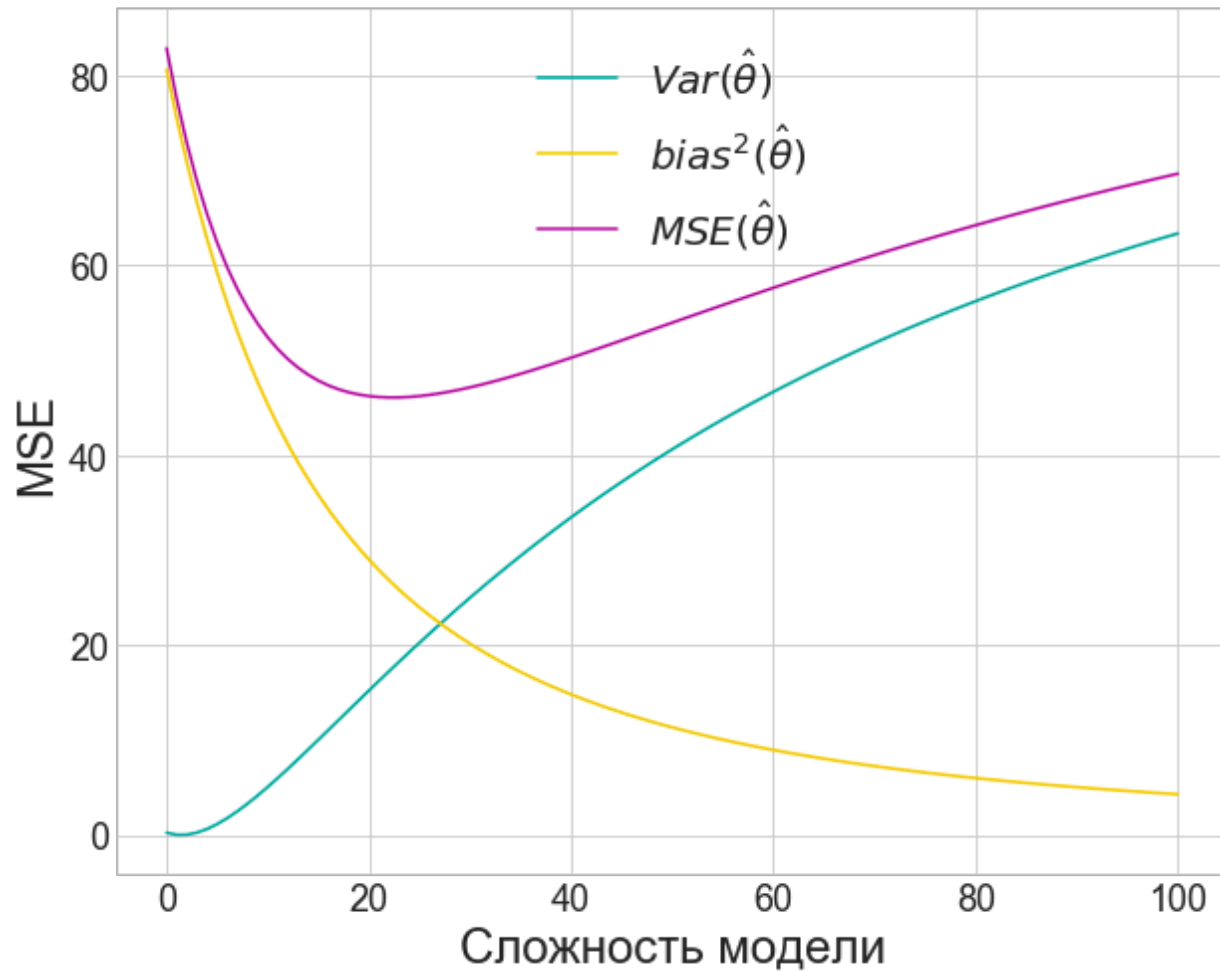


Высокое
смещение



$$MSE = \mathbb{E}(\hat{y} - y)^2 = \sigma^2 + Var(\hat{y}) + bias^2(\hat{y})$$

Разложение на смещение и разброс



$$MSE = \mathbb{E}(\hat{y} - y)^2 = \sigma^2 + Var(\hat{y}) + bias^2(\hat{y})$$

Разложение на смещение и разброс

- При увеличении сложности модели (рост числа оцениваемых параметров) смещение убывает, разброс растёт
- Теорема Гаусса-Маркова говорит, что \hat{y}^{OLS} обладает нулевым смещением и наименьшим разбросом
- То есть, для любого другого несмещённого прогноза \tilde{y} выполняется $Var(\hat{y}^{OLS}) \leq Var(\tilde{y})$
- То есть для такой оценки мы будем получать самые узкие доверительные интервалы из всех возможных несмещённых оценок

Разложение на смещение и разброс

- Иногда можно намеренно увеличивать смещение модели ради её стабильности (более низкого разброса)
- Тогда мы будем получать смещённые оценки коэффициентов
- Такой приём называется регуляризацией

Резюме

- Как устроен мир?
- Как переменная y зависит от переменной x ?
- Что будет завтра?
- Как удачно спрогнозировать переменную y ?

Оказывается, за счёт смещения можно уменьшить разброс точечных прогнозов и решить задачу лучше с точки зрения MSE, пожертвовав интерпретацией.

 Коэффициенты в Ridge и Lasso регрессиях неинтерпретируемы