Классическая линейная регрессия

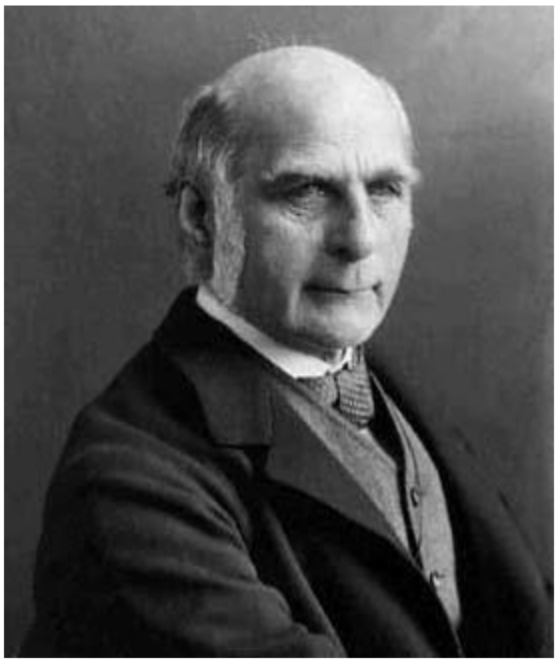
План лекции

- Понятие регрессии
- Классическая линейная регрессионная модель
- Метод наименьших квадратов (МНК)
- Критерии качества подгонки регрессии
- Свойства оценок МНК
- Статистический анализ результатов
- Прогнозирование по регрессионной модели.

Происхождение термина «регрессия»

- По смысловой нагрузке слово «регрессия» не имеет отношения к существу стохастических связей, для описания которых оно используется.
- Термин был введён Фрэнсисом Гальтоном в конце 19-го века.

Происхождение термина «регрессия»



- Френсис Гальтон
- (16 февраля 1822 — 17 января 1911)
- английский исследователь, географ, антрополог и психолог; основатель дифференциальной психологии и психометрики, статистик.

Происхождение термина «регрессия»

- Занимаясь антропологическими исследованиями, Гальтон обнаружил, что сыновья отцов с высоким или низким ростом обычно не наследуют выдающийся рост и назвал этот феномен "регрессия к посредственности".
- Сначала этот термин использовался исключительно в биологическом смысле.
- После работ ученика Гальтона, Карла Пирсона, этот термин стали использовать и в статистике.

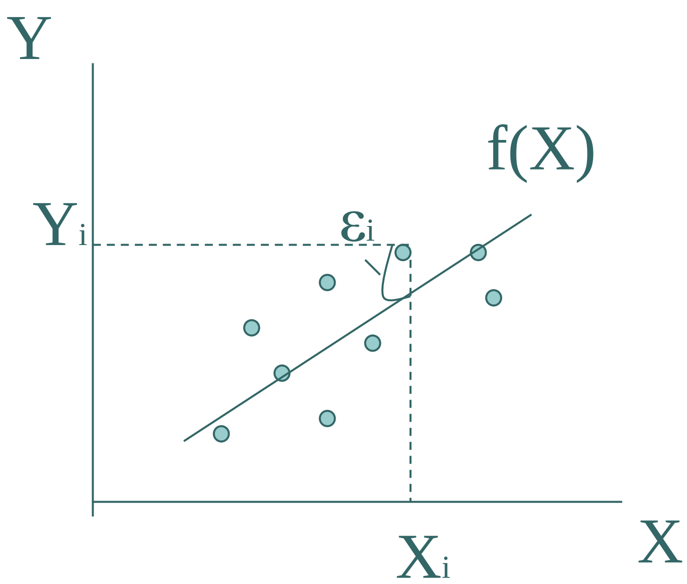
Возможности регрессионного подхода

Он позволяет

- выявить, влияют ли управляемые показатели, факторы внешней среды, статусные факторы (теперь для удобства мы будем обозначать совокупность этих показателей буквой X) на результирующий показатель Y
- построить приближенную функциональную зависимость Y от X, которую можно использовать для прогнозирования поведения Y при известных значениях X

Постановка задачи подгонки зависимости

Пусть нас интересует некоторое экономическое явление, например, потребление домохозяйствами продуктов питания.



У нас есть данные о расходах на продукты (У) и доходах (Х) домохозяйств.
Мы хотим построить по этим данным зависимость У= f(X), например, линейную: $f(X) = \beta_0 + \beta_1 X$.
Наша задача: подобрать параметры β_0 и β_1 так, чтобы линия, изображающая эту зависимость прошла через основную массу точек

Какими способами можно это осуществить?

Нужно найти такой способ подбора параметров функции *f(X)*,

при котором различия между фактически наблюдаемыми значениями Y_i и значениями функции $f(X_i)$ были как можно меньше

$$\varepsilon_i = Y_i - f\left(X_i\right) = Y_i - \beta_0 - \beta_1 X_i$$

(эту разницу называют невязкой или ошибкой)

Наилучший прогноз

- Задача подбора параметров функции $f(X_i)$ – задача поиска наилучшего прогноза Y_i по X_i
- Это оптимизационная задача
- Для ее решения надо определить целевую функцию – «функцию потерь»

$$\sum_i \rho(\varepsilon_i;\beta) \to \min_\beta$$

Метод наименьших квадратов (МНК)

$$\sum_i \rho(\varepsilon_i, \beta) = \sum_i \varepsilon_i^2 \to \min_\beta$$

$$\sum_i \varepsilon_i^2 = \sum_i (Y_i - \beta_0 - \beta_1 X_i)^2 =$$

$$= \sum_i (Y_i - f(X_i))^2 = nE(Y - f(X))^2 \to \min$$

Теорема. $E(Y - f(X))^2 \geq E(Y - E(Y \mid X))^2$

Таким образом , решение будет соответствовать оценке условного по X_i среднего значения Y_i

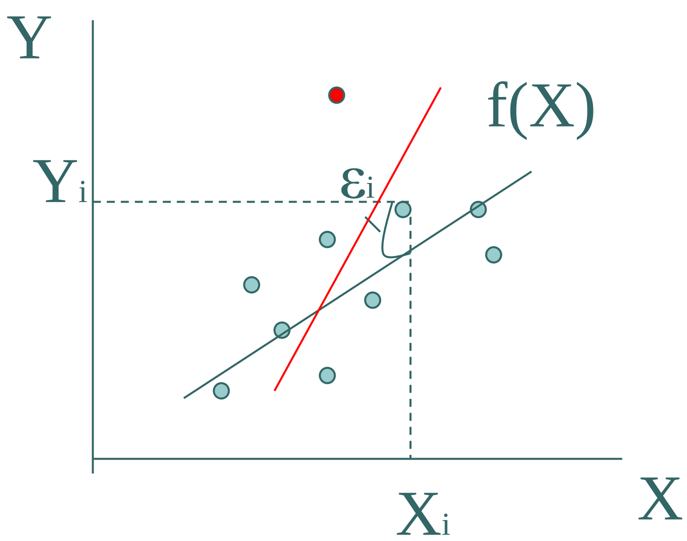
$$\hat{f}(X_i) = \hat{\beta}_0 - \hat{\beta}_1 X_i = \hat{E}(Y_i \mid X_i)$$

Метод наименьших квадратов (МНК)

- Достоинства:
  - дифференцируемость функции потерь,
  - вычислительная простота,
  - единственность решения
- Недостатки:
  - неробастность

Неробастность МНК

Нетипичные значения (выбросы) приводят к существенному ухудшению прогностических свойств функции $f(X) = \beta_0 + \beta_1 X$.



Робастные методы подгонки зависимости (М-оценки)

$$\sum_i \rho(\varepsilon_i;\beta) \to \min_\beta$$

функция ρ(.) растет по ε медленнее, чем само ε.

- Например: $\rho(\varepsilon,\beta) = |\,\varepsilon\,|$
- Полученная регрессия называется медианной, поскольку соответствует условной медиане Y_i

$$\hat{f}(X_i) = \hat{\beta}_0 - \hat{\beta}_1 X_i = \widehat{med}(Y_i \mid X_i)$$

Медианная регрессия

- Достоинства:
  - робастность
- Недостатки:
  - недифференцируемость функции потерь,
  - вычислительная сложность (симплекс-метод, методы линейного программирования)
  - неединственность решения

Квантильная регрессия

- Используется, когда предметом исследования служат не средние значения зависимой переменной при фиксированных объясняющих, а определенные квантили распределения

$$\Pr\left(Y < f\left(X\right) | X\right) = q$$

- При q=0.5 превращается в медианную регрессию
- Хорошо работает для асимметричных распределений, например, при исследованиях
  - финансового рынка (доли аутсайдеров среди акционеров),
  - доли расходов на питание домохозяйств,
  - данных о предприятиях, сильно различающихся размером

Непараметрическая регрессия

- Является интуитивной формализацией идеи сглаживания «на глаз», когда линия проводится с учетом локальных особенностей поведения У вблизи интересующих исследователя Х

$$\frac{1}{n}\sum W_{ni}(X_i)\left(Y_i - \widehat{f}(X_i)\right)^2 \to \min_{f(X)}$$

- Ее можно интерпретировать, как локально взвешенный МНК с весами

$$W_{ni}(X) = K_{h_n}(X - X_i)/\bar{K}_{h_n}(X)$$

$$\bar{K}_{h_n}(X) = \frac{1}{n}\sum_i K_{h_n}(X - X_i)$$

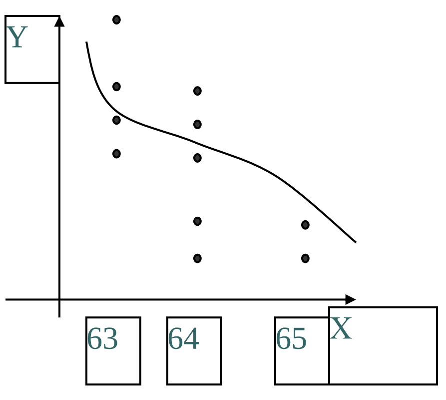
$$K_{h_n}(u) = \frac{1}{h_n}K(u/h_n), \quad где \quad \int K(u)du = 1$$

Непараметрическая регрессия

- h_n – окно сглаживания
- K(u) – ядерная функция, может быть выбрана в виде плотности стандартного нормального распределения
- Достоинства: нет необходимости в строгой спецификации модели
- Недостатки: одномерность
- Полезна для проверки точности подгонки

Графическое представление данных – диаграмма рассеяния

Определим понятие теоретической регрессии величины Y на величину X. Это будет означать, что линия регрессии строится по всей генеральной совокупности (в нашем примере – по всем домохозяйствам России).



Терминология:

X – независимая, объясняющая, экзогенная переменная, регрессор,

У – зависимая, объясняемая, эндогенная величина, регрессант.

Уравнение теоретической регрессии

- Расходы на продукты (Y) в разных домохозяйствах при одном и том же доходе (X) могут различаться

  (на рисунке показано, что при одном и том же значении X могут быть разные Y)
- Из Y можно выделить некоторую часть, определяемую X – ожидаемое значение расходов при данном доходе: $f(X) = E(Y \mid X)$
- Ту часть Y, что не укладывается в f(X), обозначают ε_i и называют случайной ошибкой
- Уравнением теоретической регрессии называют зависимость вида: $Y_i = E(Y_i \mid X_i) + \varepsilon_i$

Эконометрическая модель

Эконометрическая модель – это совокупность уравнения теоретической регрессии

$Y_i = E(Y_i \mid X_i) + \varepsilon_i$

и предположений о природе ε_i.

Какова природа ε_i, причина появления?

- Пропуск в модели ряда существенных переменных, влияющих на поведение Y
- Врожденная неопределенность поведения экономических агентов
- Использование в уравнении тех величин, которые можно измерить, а не тех, которые хотелось бы иметь теоретически
- Наличие ошибок измерения

Линейность модели

- Уравнение теоретической регрессии $Y_i = f(X_i) + \varepsilon_i$ в зависимости от $f(X_i)$ может быть линейным, квадратичным, логарифмическим и т.п.
- Мы будем рассматривать (для начала) полностью линейную модель: $f(x)=a+b*x$ – линейна по X и по параметрам
- Впоследствии станет ясно, что важна лишь линейность по параметрам (модели $f(x)=a+b*\ln(x)$, $f(x)=a+b*(1/x)$ – линейны по параметрам a и b)

Выборочная регрессия

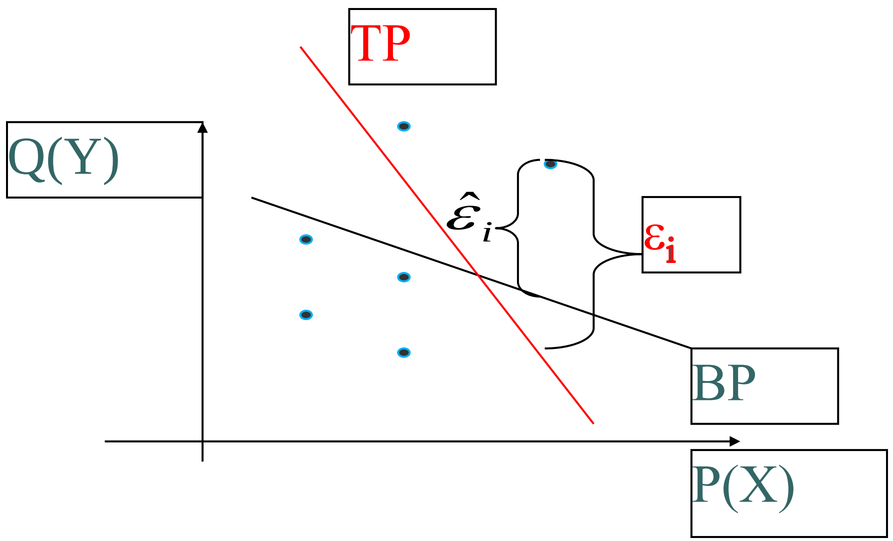
- Как правило, теоретическую регрессию построить невозможно из-за недоступности полной информации о генеральной совокупности.
- Обычно нам бывает доступна только выборка.
- Пусть теперь в нашем примере выборка из 100 домохозяйств. При использовании выборки, мы не можем построить условное ожидание – теоретическую регрессию, но мы можем оценить ее.
- Выборочной оценкой теоретической регрессии (ТР)

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

- является выборочная регрессия (ВР) $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$
- Разницу $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$ называют остатком.

Выборочная регрессия

Графическая интерпретация



Метод наименьших квадратов (МНК)

Как оценить выборочную линию регрессии?

Естественно потребовать, чтобы остатки $\hat{\varepsilon}_i \to$min.

- $\min \sum_i \hat{\varepsilon}_i$ - плохо т.к. разные знаки компенсируют друг друга, и сумма равна 0
- $\min \sum_i |\hat{\varepsilon}_i|$ - тоже плохо, т.к. эта функция не дифференцируема
- $\min \sum_i \hat{\varepsilon}^2{}_i = \min \sum_i (Y_i - \hat{Y}_i)^2$ - лучший вариант

В этом и заключается МНК

(OLS – ordinary least squares).

Как найти

$$\min \sum \hat{\varepsilon}_i^2 = \min_{\alpha,\beta} \sum (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2$$

- Обозначим $\sum (Y_i - \alpha - \beta X_i)^2 = S$
- Чтобы найти минимум этой функции необходимо приравнять к нулю частные производные

$$\begin{cases} \dfrac{\partial S}{\partial \alpha} = -2\sum (Y_i - \alpha - \beta X_i) = 0 \\ \dfrac{\partial S}{\partial \beta} = -2\sum (Y_i - \alpha - \beta X_i) X_i = 0 \end{cases} \qquad \begin{cases} \sum (Y_i - \alpha - \beta X_i) = 0 \\ \sum (Y_i - \alpha - \beta X_i) X_i = 0 \end{cases}$$

Система нормальных уравнений

$$\begin{cases} \sum_i Y_i - \alpha \sum_i 1 - \beta \sum_i X_i = 0 \\ \sum_i Y_i X_i - \alpha \sum_i X_i - \beta \sum_i X_i X_i = 0 \end{cases}$$

$$\begin{cases} \sum_i Y_i - \alpha \cdot n - \beta \sum_i X_i = 0 \\ \sum_i Y_i X_i - \alpha \sum_i X_i - \beta \sum_i X_i^2 = 0 \end{cases}$$

$$\begin{cases} \overline{Y} = \alpha + \beta \cdot \overline{X} \\ \sum_i Y_i X_i - \alpha \sum_i X_i - \beta \sum_i X_i^2 = 0 \end{cases}$$

Решение системы:

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{\sum X_i Y_i - n\overline{X}\overline{Y}}{\sum X_i^2 - n\overline{X}^2}$$

$$\hat{\alpha} = \overline{Y} - \hat{\beta} \cdot \overline{X}$$

где –

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \qquad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$x_i = X_i - \overline{X},\ y_i = Y_i - \overline{Y}$$

$\sum x_i^2 \neq 0$ не все X равны между собой

Проверка соответствия решения системы условию минимума

$$H=\begin{pmatrix} \dfrac{\partial^2 S}{\partial\alpha^2} & \dfrac{\partial^2 S}{\partial\alpha\partial\beta} \\ \dfrac{\partial^2 S}{\partial\alpha\partial\beta} & \dfrac{\partial^2 S}{\partial\beta^2} \end{pmatrix}=\begin{pmatrix} 2n & 2\sum X_i \\ 2\sum X_i & 2\sum X_i^2 \end{pmatrix}$$

- главные угловые миноры должны быть >0 – тогда это будет минимум.
- Это так: 2n>0; $2n\sum x_i^2-4(\sum x_i)^2>0$

Множественная регрессия

Обозначения

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \varepsilon_i$$

- X_{ij} - i-ое выборочное значение объясняющей переменной X_j
- Y_i - i-ое выборочное значение объясняемой переменной Y
- β_j значение коэффициента при регрессоре X_j
- ε_i - случайная ошибка

Множественная регрессия

Теоретическая регрессия

$$E(Y_i \mid X_i) = \beta_0 + \beta_1 X_{i1} + \cdots \beta_k X_{ik}$$

Дисперсия теоретической регрессии

$$V(Y_i \mid X_i) = V(\varepsilon_i) = \sigma_\varepsilon^2$$

Регрессия в матричных обозначениях

$$Y = X\beta + \varepsilon$$

где

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \qquad X = \begin{pmatrix} 1 & X_{11} & \dots & X_{1k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & \dots & X_{nk} \end{pmatrix}, \qquad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix}, \qquad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Метод наименьших квадратов

Позволяет найти минимум функции

$$f(\beta_{0,}\beta_1,\ldots,\beta_k) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \ldots - \beta_k X_{ik})^2$$

В матричных обозначениях эта задача может быть записать так

$$\sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon'\varepsilon = (Y - X\beta)'(Y - X\beta) \to \min$$

МНК

Условие 1-го порядка $$\frac{\partial \varepsilon'\varepsilon}{\partial \beta} = -2X'Y + 2X'X\beta$$

система нормальных уравнений $$X'Y = X'X\beta$$

вектор оценок коэффициентов регрессии

$$\hat{\beta} = (X'X)^{-1}X'Y$$

вектор оцененных (предсказанных моделью) значений Y $$\widehat{Y} = X\hat{\beta}$$

вектор остатков $$\widehat{\varepsilon} = Y - \widehat{Y}$$

Алгоритм МНК

Рассмотрим конкретный численный пример:

$$Y = \begin{bmatrix} 3 \\ 1 \\ 8 \\ 3 \\ 5 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & 3 & 5 \\ 1 & 1 & 4 \\ 1 & 5 & 6 \\ 1 & 2 & 4 \\ 1 & 4 & 6 \end{bmatrix}$$

Задача поиска $\min\limits_{\beta} \sum\limits_{i} \varepsilon_i^2 = \min\limits_{\beta} (\varepsilon' \varepsilon) = \min\limits_{\beta} (Y - X\beta)' (Y - X\beta)$ приводит к системе нормальных уравнений

$$X'X\beta = X'Y$$

Алгоритм МНК

В системе нормальных уравнений используются следующие конструкты:

$$X'X=\begin{bmatrix}5 & 15 & 25\\ 15 & 55 & 81\\ 25 & 81 & 129\end{bmatrix}=\begin{bmatrix}n & \sum X_1 & \sum X_2\\ \sum X_1 & \sum X_1^2 & \sum X_1X_2\\ \sum X_2 & \sum X_2X_1 & \sum X_2^2\end{bmatrix}$$

$$X'Y=\begin{bmatrix}20\\ 76\\ 109\end{bmatrix}=\begin{bmatrix}\sum Y\\ \sum X_1Y\\ \sum X_2Y\end{bmatrix}$$

Алгоритм МНК

Конкретный вид системы нормальных уравнений

$$\begin{bmatrix} 5 & 15 & 25 \\ 15 & 55 & 81 \\ 25 & 81 & 129 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 20 \\ 76 \\ 109 \end{bmatrix}$$

Решение системы нормальных уравнений –
оценки МНК для коэффициентов регрессии $Y = X\beta + \varepsilon$

$$\hat{\beta} = \left[X'X\right]^{-1} X'Y \qquad \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 4 \\ 2.5 \\ -1.5 \end{bmatrix}$$

Геометрическая суть МНК для регрессии со свободным членом (β_0)

Имеется плоскость $\pi = (i, X)$,

образованная единичным вектором *i* и векторами регрессоров X.

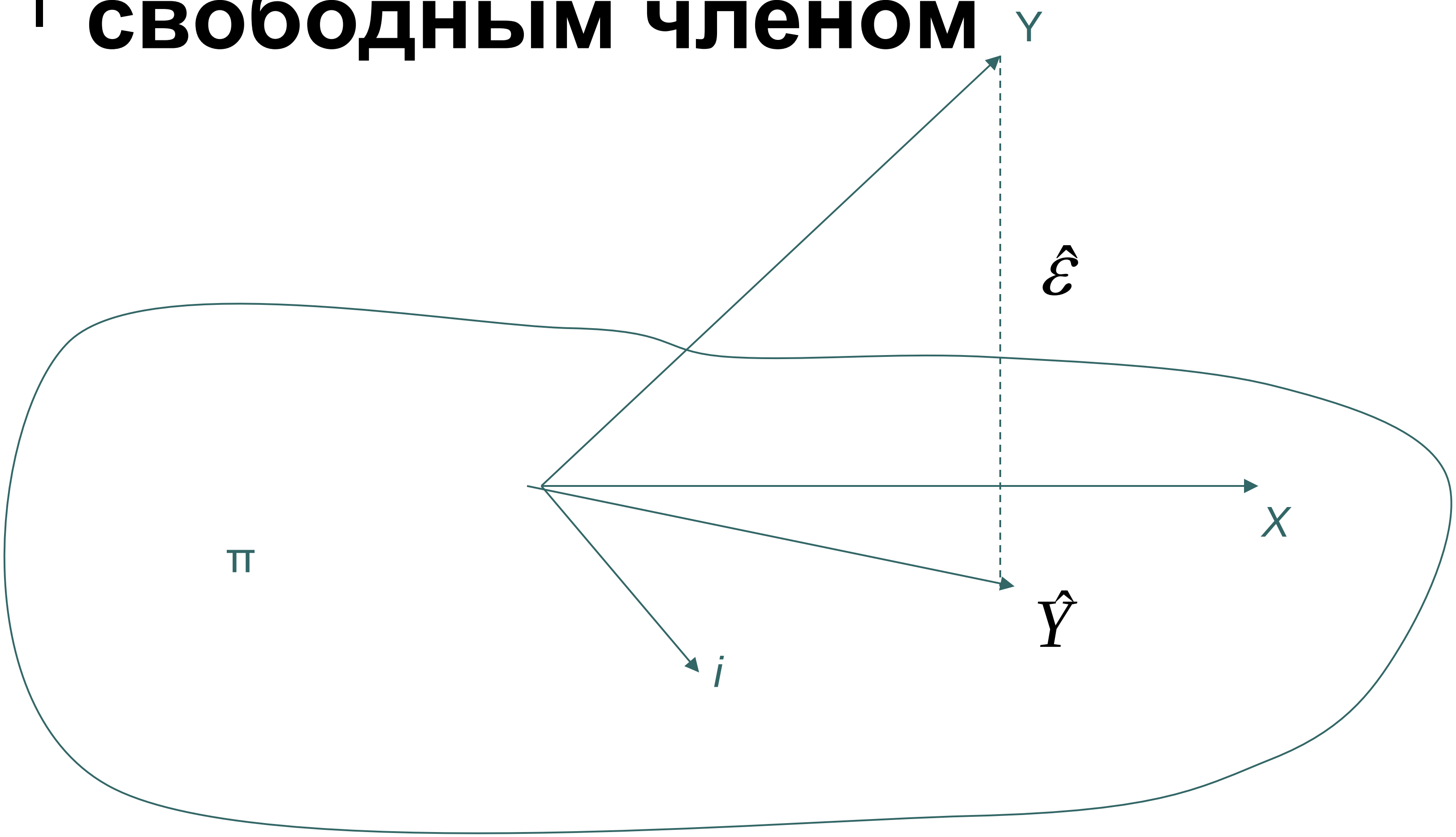
Имеется вектор значений зависимой переменной Y.

Мы ищем проекцию Y на π так, чтобы расстояние от конца Y до плоскости было минимальным. Такое возможно, если

$$\widehat{\varepsilon} \perp \pi \quad \Rightarrow \quad \widehat{\varepsilon} \perp X \quad \Rightarrow \quad X'\widehat{\varepsilon} = 0, i'\widehat{\varepsilon} = 0$$

$$\widehat{\varepsilon} \perp \widehat{Y} \quad \Rightarrow \quad \widehat{Y}'\widehat{\varepsilon} = 0$$

Геометрическая суть МНК для регрессии со свободным членом



Дисперсионный анализ результатов регрессии

$$TSS = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n} y_i^2 = y'y$$

$$ESS = \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2 = \sum_{i=1}^{n} \widehat{y}_i^2 = \widehat{y}'\widehat{y}$$

$$RSS = \sum_{i=1}^{n}(Y_i - \widehat{Y})^2 = \sum_{i=1}^{n} \widehat{\varepsilon}_i^2 = \widehat{\varepsilon}'\widehat{\varepsilon}$$

Дисперсионный анализ результатов регрессии

- TSS – общая сумма квадратов отклонения наблюдаемых значений Y от среднего значения
- ESS – сумма квадратов отклонения от среднего значения объясненных с помощью регрессии значений
- RSS – остаточная сумма квадратов отклонения наблюдаемых значений Y от объясненных с помощью регрессии значений

- TSS – total sum of squares
- ESS – explained sum of squares
- RSS – residual sum of squares

Критерии качества подгонки регрессии

Очевидно, что регрессия тем лучше, чем меньше RSS и чем больше ESS.

Однако более удобным критерием качества является относительный показатель - коэффициент детерминации:

$$R^2 = \frac{ESS}{TSS}$$

- доля объясненного разброса наблюдений Y

$$0 \le R^2 \le 1 \qquad R^2 = r^2_{Y\hat{Y}}$$

Модифицированный коэффициент детерминации регрессии

Чем ближе R^2 к 1, тем лучше качество подгонки, хотя надо помнить, что этот показатель всегда механически увеличивается при добавлении нового регрессора, даже если он никак не связан с Y.

Более чувствителен к качеству регрессии модифицированный R^2, нормированный на степени свободы :

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

.

Свойства оценок МНК, обязанные наличию в регрессии свободного члена

1. Сумма остатков равна 0: $\sum_i \hat{\varepsilon}_i = \vec{i}\,'\hat{\varepsilon} = 0$
2. Среднее значение наблюдаемых Y равно среднему значению оцененных Y $\overline{Y} = \overline{\hat{Y}}$
3. Точка $\left(\overline{X}, \overline{Y}\right)$ лежит на линии регрессии
4. Выполняется теорема Пифагора TSS=ESS+RSS
5. Эквивалентны два определения коэффициента детерминации

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

Регрессия без свободного члена

1. Сумма остатков **не** равна 0
2. Среднее значение наблюдаемых Y **не** равно среднему значению оцененных Y
3. Точка $(\overline{X}, \overline{Y})$ **не** лежит на линии регрессии
4. **Не** выполняется теорема Пифагора TSS≠ESS+RSS
5. **Не** эквивалентны два определения коэффициента детерминации

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

Статистические свойства оценок

- Оценки считаются «хорошими», если они обладают определенными свойствами:
- ***несмещенностью*** (в этом случае математическое ожидание оценки совпадает с оцениваемым теоретическим параметром);
- ***состоятельностью*** (это означает, что для больших выборок вероятность значимых отклонений величины оценки от значения оцениваемого теоретического параметра равна нулю);
- ***эффективностью*** (чем меньше дисперсия оценки, тем она считается эффективнее).
- Исследование свойств оценок – это важная теоретическая задача.

Теорема Гаусса-Маркова

Иога́нн Карл Фри́дрих Га́усс
(1777- 1855) — немецкий математик, механик, физик, астроном, геодезист Иностранный член Шведской (1821) и Российской (1824) Академий наук, английского Королевского общества. Создатель МНК

Андре́й Андре́евич Ма́рков
(1856 -1922) — русский математик, академик. Создатель теории стохастических процессов, цепей Маркова

Свойства оценок МНК (теорема Гаусса-Маркова)

Если выполнены следующие условия:

1. Модель $Y = X\beta + \varepsilon$ верно специфицирована
2. Матрица X – детерминирована и имеет ранг k+1
3. Ошибка – случайный вектор с математическим ожиданием и ковариационной матрицей

$$E(\varepsilon) = 0, \quad V(\varepsilon) = E\left[(\varepsilon - E(\varepsilon))(\varepsilon - E(\varepsilon))'\right] = \sigma_{\varepsilon}^{2} I$$

Свойства оценок МНК (теорема Гаусса-Маркова)

тогда оценка МНК

$$\hat{\beta} = (X'X)^{-1} X'Y$$

является наилучшей (наиболее эффективной) в классе линейных несмещенных оценок, т.е.

она линейна по Y и по ε,

$$E(\hat{\beta}) = \beta$$

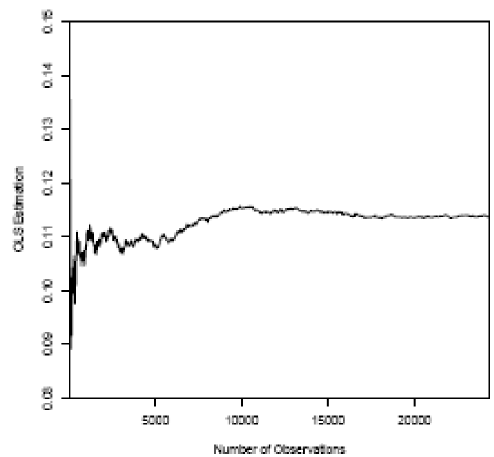
и обладает наименьшей дисперсией в классе линейных несмещенных оценок.

Асимптотические свойства оценок МНК

Для больших выборок для оценок МНК выполняется свойство состоятельности.

Слишком жесткое требование детерминированности матрицы регрессоров X заменяется на условие:

$$\lim_{n\to\infty} P\left[\left|\frac{X'\varepsilon}{n}\right| > \eta\right] = 0 \qquad \underset{n\to\infty}{p\lim} \frac{X'\varepsilon}{n} = 0$$



тогда

$$\lim_{n\to\infty} P\left[\left|\hat{\beta} - \beta\right| > \eta\right] = 0$$

$$\underset{n\to\infty}{p\lim} \widehat{\beta} = \beta + \underset{n\to\infty}{p\lim}\left[\left(\frac{X'X}{n}\right)^{-1} \frac{X'\varepsilon}{n}\right] = \beta + \left(\underset{n\to\infty}{p\lim} \frac{1}{n} X'X\right)^{-1} \cdot \underset{n\to\infty}{p\lim} \frac{X'\varepsilon}{n} = \beta$$

Статистический анализ результатов

- Следующий вопрос: насколько достоверны полученные оценки, ведь есть проблема выборочного смещения?
- Кроме того, у нас могут иметься различные гипотезы о влиянии тех или иных показателей на Y, и мы хотели бы их проверить, пользуясь построенной моделью.
- Для этого надо знать, каким вероятностным распределениям подчиняются полученные оценки
- Распределение оценок зависит от распределения ошибок
- В КЛРМ делается следующее предположение:

$$\varepsilon \sim N(0, \sigma_\varepsilon^2 I)$$

- это предположение о нормальности случайной ошибки.

Статистический анализ результатов

Для построения необходимых тестовых статистик важно знать, как распределены показатели теоретической и выборочной регрессии.

В силу линейности модели линейные комбинации нормальных случайных векторов будут тоже нормальными векторами:

$$Y \sim N(X\beta, \sigma_\varepsilon^2 I), \quad \widehat{Y} \sim N(X\beta, \sigma_\varepsilon^2 X(X'X)^{-1}X = \sigma_\varepsilon^2 P),$$

$$\widehat{a} \sim N(\beta, \sigma_\varepsilon^2 (X'X)^{-1}), \quad \widehat{\varepsilon} \sim N(0, \sigma_\varepsilon^2 (I-P))$$

Статистический анализ результатов

А что можно сказать о нелинейных комбинациях?

$$\frac{(n-k-1)\hat{\sigma}_{\varepsilon}^{2}}{\sigma_{\varepsilon}^{2}}=\frac{RSS}{\sigma_{\varepsilon}^{2}}=\frac{\hat{\varepsilon}'\hat{\varepsilon}}{\sigma_{\varepsilon}^{2}}\sim\chi^{2}(n-k-1)$$

$$\frac{ESS}{\sigma_{\varepsilon}^{2}}\sim\chi^{2}(k),\quad\frac{TSS}{\sigma_{\varepsilon}^{2}}\sim\chi^{2}(n-1)$$

Можно показать, что оценки $\hat{\beta}$ и $\hat{\sigma}_{\varepsilon}^{2}$ статистически независимы, и тогда

$$\frac{\hat{\beta}_{j}}{\hat{\sigma}_{\varepsilon}\sqrt{\left[(X'X)^{-1}\right]_{jj}}}\sim t(n-k-1)$$

Проверка гипотез

Статистический анализ оценок сводится в стандартном случае к проверке следующих статистических гипотез:

1) $H_0: \quad \beta_j = 0$ - проверка значимости отдельного коэффициента регрессии,
при альтернативной гипотезе $H_A: \beta_j \neq 0$;
осуществляется на основании t-статистики,

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}} = \frac{\hat{\beta}_j}{\hat{\sigma}_{\varepsilon}\sqrt{\left[(X'X)^{-1}\right]_{jj}}} \overset{H_0:\beta_j=0}{\sim} t(n-k-1)$$

Пример 1.
Оценивание множественной регрессии для анализа капитализации банковской системы РФ за период 2004-2009 г.г.

```
      Source |       SS       df       MS              Number of obs =      64
-------------+------------------------------           F(  7,    56) =  621.86
       Model |  3.3375e+18     7  4.7678e+17           Prob > F      =  0.0000
    Residual |  4.2935e+16    56  7.6670e+14           R-squared     =  0.9873
-------------+------------------------------           Adj R-squared =  0.9857
       Total |  3.3804e+18    63  5.3657e+16           Root MSE      =  2.8e+07

------------------------------------------------------------------------------
           a |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          nf |  -.0899685   .2113737    -0.43   0.672    -.5134008    .3334639
          na |   .0844604   .0227928     3.71   0.000     .0388009    .1301199
          nh |   -.185508   .1047839    -1.77   0.082    -.3954154    .0243994
          db |   .0414967   .0641912     0.65   0.521    -.0870938    .1700871
          df |   .1308707   .0248723     5.26   0.000     .0810455    .1806958
          da |  -.0087314   .0393488    -0.22   0.825    -.0875565    .0700938
          dh |   .0277084   .0299776     0.92   0.359    -.0323439    .0877608
       _cons |   1.07e+08   2.79e+07     3.84   0.000     5.13e+07    1.63e+08
```

Пример 1

В примере с моделированием капитализации значимость влияния, скажем, депозитов фирм (da) можно проверить так:

$$t_{\hat{\beta}_{da}} = \frac{\hat{\beta}_{da}}{\widehat{\sigma}_{\hat{\beta}_{da}}} = \frac{-0.0087}{0.0394} = -0.22 \qquad P(t > |-0.22|) = 0.825$$

поскольку вероятность оказалась велика – 82.5% (например, по сравнению с 5%-ым уровнем значимости), нет оснований отбрасывать основную гипотезу . Это означает, что объем депозитов фирм не оказывает значимого влияния на капитализацию банковской системы РФ в анализируемом периоде.

Пример 2. Оценивание множественной регрессии для анализа детерминант заработной платы жителей Москвы в 2000 году.

```
      Source |       SS       df       MS              Number of obs =     157
-------------+------------------------------           F(  6,   150) =    8.86
       Model |  37.0211059     6  6.17018432           Prob > F      =  0.0000
    Residual |  104.515396   150  .696769304           R-squared     =  0.2616
-------------+------------------------------           Adj R-squared =  0.2320
       Total |  141.536502   156  .907285266
------------------------------------------------------------------------------
 logrealwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |  -.6079527   .1378764    -4.41   0.000    -.8803834   -.3355219
         age |   .1590116   .0309307     5.14   0.000     .0978954    .2201277
        age2 |  -.0018494   .0003508    -5.27   0.000    -.0025425   -.0011562
   education |  -.1191102   .0380372    -3.13   0.002    -.1942681   -.0439524
      stagna |  -.3624113   .1892399    -1.92   0.057    -.7363315    .0115089
     stagna2 |   .0496421   .0267672     1.85   0.066    -.0032472    .1025315
       _cons |   2.627421   .6371025     4.12   0.000     1.368566    3.886275
------------------------------------------------------------------------------
```

Пример 2

В примере с уравнением заработной платы значимость влияния, скажем, возраста (age) можно проверить так:

$$t_{\hat{\beta}_{age}} = \frac{\hat{\beta}_{age}}{\hat{\sigma}_{\hat{\beta}_{agej}}} = \frac{0.159}{0.031} = 5.14 \qquad P(t > |5.14|) = 0.000$$

поскольку вероятность оказалась мала (например, по сравнению с 5%-ым уровнем значимости), основную гипотезу следует отбросить. Это означает, что возраст оказывает значимое влияние на заработную плату.

Проверка гипотез

2) проверка адекватности регрессии

$$H_0: \quad \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

(при этом $R^2 = 0$)

при альтернативной гипотезе

$$H_A: a_1^2 + a_2^2 + \ldots + a_k^2 > 0$$

(при этом $R^2 > 0$)

Проверка гипотезы об адекватности регрессии

осуществляется на основании F-статистики, которая в условиях справедливости основной гипотезы, т.е. гипотезы о неадекватности регрессии, подчиняется F-распределению с k и n-k-1 степенями свободы:

$$F = \frac{ESS / k}{RSS / (n-k-1)} =$$

$$= \frac{R^2 / k}{(1-R^2)/(n-k-1)} \overset{H_0:\ \beta_1=\beta_2=\ldots=\beta_k=0}{\sim} F(k, n-k-1)$$

Пример 1

В примере с капитализацией адекватность регрессии в целом можно проверить так:

$$F = \frac{3.34*10^{18}/7}{4.29*10^{16}/56} = 621.86 \qquad P(F > 621.86) = 0.000$$

поскольку вероятность оказалась мала (например, по сравнению с 5%-ым уровнем значимости), основную гипотезу следует отбросить. Это означает, что оцененная регрессия оказалась адекватна данным.

Об этом же свидетельствует величина коэффициента детерминации R^2 =0.9873 и его модифицированного с учетом степеней свободы аналога R^2_{adj} =0.9857.

Однако для регрессии на основе временных рядов высокие показатели коэффициентов детерминации – явление типичное, связанное с наличием общих временных тенденций в анализируемых показателях.

Пример 2

В примере с уравнением заработной платы адекватность регрессии в целом можно проверить так:

$$F=\frac{37.021/6}{104.515/150}=8.86 \quad P(F>8.86)=0.000$$

поскольку вероятность тоже оказалась мала (например, по сравнению с 5%-ым уровнем значимости), основную гипотезу следует отбросить.

Это означает, что оцененная регрессия оказалась адекватна данным, несмотря на то, что коэффициент детерминации R^2 =0.26, а его модифицированный с учетом степеней свободы аналог R^2_{adj}=0.23.

Следует отметить, что такие маленькие значения коэффициентов детерминации - довольно типичное явление для данных опросов домохозяйств из-за сильной неоднородности объектов выборки.

Проверка гипотез

3) $H_0: \quad Q\beta = q$

- проверка линейного ограничения на коэффициенты, при альтернативной гипотезе

$$H_A: Q\beta \neq q$$

Проверка линейного ограничения

- Можно проверить **гипотезу о не значимости группы переменных**.
- В нашем примере с капитализацией есть целый ряд показателей, которые по отдельности не оказывают значимого влияния на капитализацию. Это - расчетные счета нерезидентов (nf), МБК (db), депозиты фирм (da), срочные депозиты населения (dh). Можно проверить гипотезу о том, что они не оказывают влияния и в совокупности:

$$H_0: \quad \beta_1 = \beta_4 = \beta_6 = \beta_7 = 0$$

$$H_A: \beta_1^2 + \beta_4^2 + \beta_6^2 + \beta_7^2 > 0$$

Проверка линейного ограничения

- В таких случаях необходимо строить дополнительную регрессию, в которую не будут включены соответствующие регрессоры. Для каждой регрессии вычисляется сумма квадратов остатков: RSS (RSSд для исходной регрессии и RSSк для дополнительной).
- Затем, с помощью F-статистики производится их сравнение

$$F = \frac{(RSS_{к} - RSS_{Д})/r}{RSS_{Д}/n-k-1} \sim F_{r,(n-k-1)}$$

Проверка линейного ограничения

Для наших данных оценка короткой регрессии выглядит следующим образом:

```
      Source |       SS       df       MS              Number of obs =      64
-------------+------------------------------           F(  3,    60) = 1499.02
       Model |  3.3359e+18     3  1.1120e+18           Prob > F      =  0.0000
    Residual |  4.4508e+16    60  7.4179e+14           R-squared     =  0.9868
-------------+------------------------------           Adj R-squared =  0.9862
       Total |  3.3804e+18    63  5.3657e+16           Root MSE      =  2.7e+07
------------------------------------------------------------------------------
           a |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          na |    .083186   .0166485     5.00   0.000      .049884     .116488
          nh |  -.1715743   .0828866    -2.07   0.043    -.3373722   -.0057765
          df |    .153684   .0140171    10.96   0.000     .1256457    .1817224
       _cons |   1.32e+08   1.40e+07     9.46   0.000     1.04e+08    1.60e+08
------------------------------------------------------------------------------
```

$$F = \frac{(4.4508*10^{16} - 4.2935*10^{16})/4}{4.2935*10^{16}/56} = 0.51 \qquad P(F > 0.51) = 0.73$$

Проверка линейного ограничения

- Этот результат интерпретируется следующим образом: при любом разумном уровне значимости основная гипотеза не может быть отвергнута, т.е. можно исключить из регрессии группу незначимых показателей.
- Об этом так же свидетельствует несколько возросшее в короткой регрессии значение

$$R^2_{adj} = 0.9862$$

Проверка линейного ограничения

В примере с заработной платой есть 2 переменные, stagna и stagna2 – стаж работы на данном предприятии и его квадрат, которые по отдельности не оказывают значимого влияния на заработную плату.

Проверим гипотезу о том, что и в совокупности эти переменные не значимы:

$$H_0: \quad \beta_5 = \beta_6 = 0 \quad H_A: \beta_5^2 + \beta_6^2 > 0$$

Проверка линейного ограничения

Для наших данных оценка короткой регрессии выглядит следующим образом:

```
      Source |       SS       df       MS              Number of obs =     158
-------------+------------------------------           F(  4,   153) =   12.23
       Model |  34.8600216     4  8.7150054           Prob > F      =  0.0000
    Residual |  109.023637   153  .71257279           R-squared     =  0.2423
-------------+------------------------------           Adj R-squared =  0.2225
       Total |  143.883659   157  .916456424          Root MSE      =  .84414
------------------------------------------------------------------------------
 logrealwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |  -.6106794   .1378193    -4.43   0.000    -.8829538    -.338405
         age |   .1463342   .0299831     4.88   0.000     .0870999    .2055684
        age2 |  -.0017299   .0003461    -5.00   0.000    -.0024136   -.0010462
   education |  -.1071937   .0381301    -2.81   0.006    -.1825232   -.0318642
       _cons |   2.385354   .6251253     3.82   0.000     1.150363    3.620346
```

$$F = \frac{(109.024 - 104.515)/2}{104.515/150} = 3.24 \qquad P(F > 8.86) = 0{,}0419$$

Проверка линейного ограничения

- Этот результат интерпретируется следующим образом: при уровне значимости 5% основная гипотеза должна быть отвергнута, т.е. нежелательно исключать из регрессии переменные, отвечающие за стаж. Об этом же свидетельствует упавшее в короткой регрессии значение R^2_{adj} =0.22.
- Аналогичным образом могут быть проверены любые линейные гипотезы относительно регрессионных коэффициентов.

Доверительные интервалы для коэффициентов

- В последних двух столбцах таблицы результатов оценивания регрессии в некоторых статистических пакетах выдаются интервальные оценки - доверительные интервалы - для коэффициентов.
- Они строятся на основании t-статистик для указанной (обычно 95%) доверительной вероятности:

$$\hat{\beta}_j - t_{2.5\%}(n-k-1)\cdot\widehat{\sigma}_{\hat{\beta}_j} < \beta_j < \hat{\beta}_j + t_{2.5\%}(n-k-1)\cdot\widehat{\sigma}_{\hat{\beta}_j}$$

Прогнозирование по регрессионной модели

Более интересно и целесообразно строить интервальные оценки для прогноза зависимой переменной:

$$X_0\hat{\beta} - t_{\alpha/2}(n-k-1)\cdot\hat{\sigma}\sqrt{1+X_0'(X'X)^{-1}X_0} < Y_0 <$$

$$< X_0\hat{\beta} - t_{\alpha/2}(n-k-1)\cdot\hat{\sigma}\sqrt{1+X_0'(X'X)^{-1}X_0}$$

здесь X_0 - набор значений регрессоров, для которого мы намереваемся вычислить прогноз Y_0 .

Прогнозирование по регрессионной модели

Пусть в нашем примере мы хотим оценить заработную плату жителя Москвы в 2000 году, при условии, что это 30-ти летний мужчина с аспирантурой и 2-х летним стажем работы на некоем предприятии.

Согласно оцененному уравнению регрессии:

$$\widehat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 sex + \hat{\beta}_2 age + \hat{\beta}_3 age2 + \hat{\beta}_4 education + \hat{\beta}_5 stagna + \hat{\beta}_6 stagna2$$

$$= 2.63 - 0.61*0 + 0.16*30 - 0.002*900 -$$
$$- 0.12*2 - 0.36*2 + 0.05*4 = 5.088$$

Прогнозирование по регрессионной модели

Мы предсказали логарифм заработной платы

Это соответствует оценке величины самой заработной платы 162 условных единиц

Можно вычислить доверительный интервал для логарифма заработной платы

$$s.e.(Y_0) = \hat{\sigma}\sqrt{1 + X_0'(X'X)^{-1}X_0} = 0.847$$

$$t_{\alpha/2}(n-k-1) = 1.645,$$

$$3.695 < Y_0 < 6.481.$$

Это означает, что в 2000 году сама заработная плата такого индивида могла лежать в интервале

от 40 до 653-х условных единиц.

Спасибо за внимание!