Проверка гипотез в классической линейной нормальной регрессионной модели

Краткое содержание предыдущей лекции.
Мы рассмотрели классическую линейную нормальную регрессионную модель (КЛНРМ) $Y_i=\beta_1+\beta_2 X_{2,i}+...+\beta_k X_{k,i}+\epsilon_i$ (или, в матричной записи, $y=X\beta+\epsilon$) и выяснили, как рассчитать доверительный интервал для любого коэффициента регрессии β_j, $j=1,\dots,k$.

Сегодня мы рассмотрим гипотезы, затрагивающие эти коэффициенты:
► о значении отдельного коэффициента;
► о (не)значимости регрессии в целом;
► о линейном ограничении на коэффициенты.

Критерий для проверки гипотезы о значении коэффициента опирается на те же теоретические результаты, что и рассмотренный в прошлый раз доверительный интервал.
Теоретическое обоснование критериев для проверки остальных гипотез — штука непростая, расписывать её здесь не буду. Теория есть в книжке, которую я всегда рекомендую: Я.Р. Магнус, П.К. Катышев, А.А. Пересецкий, «Эконометрика. Начальный курс». Не скажу, что там всё сразу понятно, но вариант удачнее предложить не могу.

Проверка гипотезы о значении коэффициента регрессии

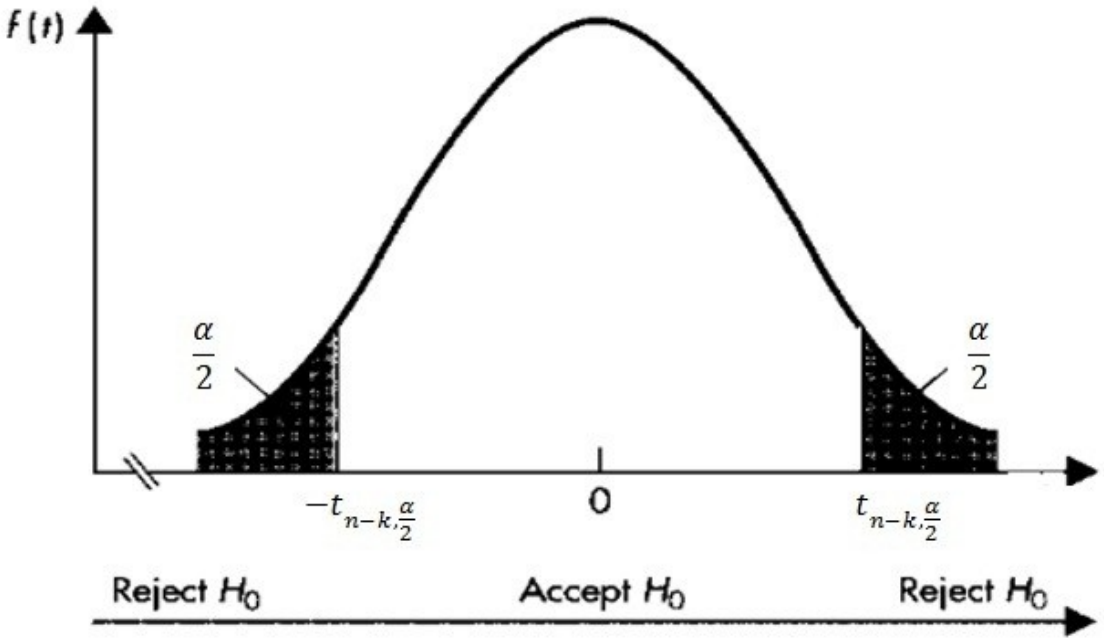
Нужно проверить основную гипотезу $H_0:\beta_j=\beta_j^0$, где β_j^0 — предполагаемое значение коэффициента (верхний индекс «0» здесь означает, что таково значение коэффициента согласно H_0, а не степень).
Обычно на автомате выбирается двусторонняя альтернатива $H_A:\beta_j\neq\beta_j^0$, хотя односторонние гипотезы тоже можно использовать, иногда они более уместны. Дальше речь пойдёт именно о двусторонней альтернативной гипотезе.

Снова используем следствие 2 из теоремы Фишера: $\dfrac{\hat\beta_j-\beta_j}{\hat\sigma(\hat\beta_j)}\sim t_{n-k}$. Если в этом утверждении заменить неизвестное истинное значение β_j на предполагаемое β_j^0, получится *t-статистика для проверки гипотезы о значении коэффициента регрессии*:

$$t=\frac{\hat\beta_j-\beta_j^0}{\hat\sigma(\hat\beta_j)}\overset{H_0}{\sim}t_{n-k}.$$

Если оценка $\hat\beta_j$ близка к предполагаемому значению β_j^0, то *t*-статистика близка к нулю. В этом случае основную гипотезу отклонять, конечно, не следует. Если же расхождение между оценкой и предполагаемым значением велико по сравнению со стандартной ошибкой $\hat\sigma(\hat\beta_j)$, то данные плохо согласуются с основной гипотезой и её стоит отвергнуть.

Критерий: H_0 отвергается, если $|t|>t_{n-k,\alpha/2}$.
На картинке этот критерий можно показать так:



p-значение рассчитывается как обычно — ищем уровень значимости, для которого рассчитанная *t*-статистика совпадает с критическим значением. Можно записать это и в виде формулы:

$$p-значение = P\left(|U| > |t|\right),\ U \sim t_{n-k}\ .$$

Здесь U — случайная величина, распределённая по закону Стьюдента с (n – k) степенями свободы, t — *рассчитанное* значение *t*-статистики (т. е. *t*-статистика как число, а не как случайная величина). Обдумайте эту формулу. Ответьте на вопрос: почему именно так?

Пример (и обещанная байка). Как-то на рубеже XX и XXI веков случился торговый конфликт между двумя членами Всемирной Торговой Организации (ВТО): Чили и Евросоюзом. В Чили была установлена система налогообложения, выгодная для чилийских производителей писко (крепкий алкогольный напиток, который мне до сих пор не доводилось отведать). Это расстроило европейских производителей виски, которые потребовали разбирательства ВТО. Согласно жалобе европейцев, щадящие налоги позволяли производителям писко назначать низкие цены на свой товар, что приводило к вытеснению европейского виски с чилийского рынка (производители из ЕС не могли снизить свои цены — их чилийские налоги не щадили). В ответ на жалобу сторона Чили представила результаты оценивания регрессионной модели

$$Спрос\ на\ писко_t = \beta_1 + \beta_2 Доход_t + \beta_3 Цена\ писко_t + \beta_4 Цена\ виски_t + \beta_5 Цена\ пива_t + \beta_6 Цена\ вина_t + \epsilon_t .$$

Индекс *t* здесь обозначает год наблюдения (данные относились ко всей стране в целом). Возможно, спрос на писко измерялся объёмом продаж, а доход — чилийским ВВП, точно не скажу. Отчёт о деле «Chile – EU: alcoholic beverages» доступен в интернете, хотя читать его скучно. Скучнее чем эту лекцию — он куда дольше.

В центре внимания был коэффициент β_4 при цене виски. Жалоба ЕС опиралась на предположение, что чилийцы рассматривают писко как заменитель виски, поэтому при падении цены на писко переключаются с дорогого европейского виски на дешёвый отечественный аналог. Чилийцы проверили гипотезу $H_0: \beta_4 = 0$ — спрос на писко не зависит от цены виски, так что эти два товара не заменители друг другу. Альтернативная гипотеза была $H_A: \beta_4 \neq 0$, хотя по смыслу больше подошла бы односторонняя альтернатива $H_A: \beta_4 > 0$ — при росте цены виски растёт спрос на писко, как и должно быть в случае замещения). Как я уже писал, исследователи берут двустороннюю альтернативную гипотезу на автомате, не задумываясь (и все статистические программы автоматически проверяют именно так).

Вот уравнение, оценённое чилийцами по данным за 15 лет:

$$\widehat{\textit{Спрос на писко}} = \underset{(3.66)}{3.58} - \underset{(1.21)}{0.007}\,\textit{Доход} - \underset{(0.46)}{1.31}\,\textit{Цена писко} + \underset{(0.52)}{0.12}\,\textit{Цена виски} + \underset{(0.40)}{0.60}\,\textit{Цена пива} + \underset{(1.21)}{0.36}\,\textit{Цена вина}.$$

Судя по точечной оценке коэффициента β_4 , спрос на писко положительно связан с ценой виски. Рассчитаем *t*-статистику:

$$t = \frac{\hat{\beta}_4 - \beta_4^0}{\hat{\sigma}(\hat{\beta}_4)} = \frac{0.12 - 0}{0.52} = 0.23.$$

Критическое значение: $t_{n-k,\alpha/2} = t_{15-6,0.05/2} = 2.262$.

Получили $|t| < 2.262$, так что нет оснований отвергнуть основную гипотезу. Данные не позволяют утверждать наличие связи между спросом на писко и ценой виски — ни положительной, ни отрицательной.
Попробуйте рассчитать *p*-значение. Должно получиться 0.82. Подумайте, чему было бы равно *p*-значение, если бы альтернатива была односторонней.

В общем, связь продаж писко с ценой виски не обнаружилась. Таким образом у Чили появляется аргумент против ЕС: ваши обвинения голословны, у вас нет надёжного свидетельства.

Регрессия не помогла чилийцам, но это не наша забота.

Проверка значимости регрессии в целом

Основная гипотеза, $H_0: \beta_2 = \beta_3 = ... = \beta_k = 0$ (ни одна из объясняющих переменных не связана с объясняемой).
Альтернативная гипотеза, $H_A: \beta_2^2 + \beta_3^2 + ... + \beta_k^2 > 0$ (хотя бы один из коэффициентов при регрессорах отличен от нуля — есть связь хотя бы с одной объясняющей переменной).

Внимание: свободный член β_1 не участвует в основной и альтернативной гипотезах, он не отражает связь между регрессантом и регрессорами.

Зачем это? Почему бы не проверить каждый коэффициент *t*-статистикой?
А вот не стоит так делать. Если у вас много переменных в модели и вы проверяете значимость коэффициента при каждой, то получается неприятность. Каждый раз вы допускаете ошибку первого рода с выбранной вами вероятностью. Если вы используете уровень значимости 5%, то при отсутствии связи объясняемой переменной с регрессором вы ошибочно обнаружите эту связь с вероятностью 5%. Если у вас десять регрессоров, то очень вероятно, что хотя бы один из них окажется значимым, даже если на самом деле связи никакой нет. Поэтому при *оценивании принято сразу делать проверку значимости регрессии в целом и придавать значение отдельным коэффициентам и их значимости, только если регрессия в целом значима*.

***F*-статистика** для проверки гипотезы о значимости регрессии в целом и её распределение при верной основной гипотезе:

$$F = \frac{ESS/(k-1)}{RSS/(n-k)} = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} \overset{H_0}{\sim} F_{k-1,n-k}.$$

Предложение. Сравните эту формулу с формулой *t*-статистики для коэффициента корреляции. Они похожи. И это неспроста.

F-статистика тем больше, чем большую долю вариации объясняемой переменной может объяснить уравнение (чем больше ESS по отношению к RSS, чем больше $R^2 = ESS/TSS$). Вспомните: коэффициент детерминации равен нулю, когда $\hat{\beta}_2 = \hat{\beta}_3 = ... = \hat{\beta}_k = 0$ — естественно с его помощью проверить гипотезу, что $\beta_2 = \beta_3 = ... = \beta_k = 0$.
Когда $ESS = 0$ и $R^2 = 0$, то и $F = 0$. В этом случае ни малейшего повода признавать регрессию значимой нет. А с ростом коэффициента R^2 растёт и *F*-статистика. Значит, при маленьких значенияъ статистики основную гипотезу не отвергаем, а при больших — отвергаем.

Критерий: отвергнуть H_0, если $F > F_{k-1,n-k,\alpha}$ где α — выбранный уровень значимости.

$p-$значение $= P(W > F)$, где $W \sim F_{k-1,n-k}$, а F — *рассчитанное* значение *F*-статистики (в том смысле, что это число, а не случайная величина).

Как обычно советую подумать, почему это так, и вспомнить, что вообще такое *p*-значение.

Пример. Вернёмся к уравнению спроса на писко:

$$\underset{}{\hat{Спрос}\ на\ писко} = \underset{(3.66)}{3.58} - \underset{(1.21)}{0.007}\ Доход - \underset{(0.46)}{1.31}\ Цена\ писко + \underset{(0.52)}{0.12}\ Цена\ виски + \underset{(0.40)}{0.60}\ Цена\ пива + \underset{(1.21)}{0.36}\ Цена\ вина,$$
$$R^2 = 0.976,\ n = 15.$$

На этот раз я добавил коэффициент детерминации — как видите, он у чилийцев оказался очень высоким.

Проверим значимость всей регрессии, то есть гипотезу $H_0: \beta_2 = ... = \beta_6 = 0$ против $H_A: \beta_2^2 + ... + \beta_6^2 > 0$, на уровне 5%.

Рассчитаем статистику:

$$F = \frac{0.976/(6-1)}{(1-0.976)/(15-6)} = 72.68.$$

Критическое значение: $F_{k-1,n-k,\alpha} = F_{6-1,15-5,0.01} = F_{5,9,0.01} = 6.06.$

Вывод: основная гипотеза отвергается, потому что $F > 6.06$, регрессия в целом значима. Какие-то из объясняющих переменных связаны с продажами писко.

$p-$значение $= P(W > 72.68),\ W \sim F_{5,9}.$

Посмотрев в таблицы или спросив у компьютера, узнаем, что $p-$значение ≈ 0. Уравнение чилийцев значимо практически на любом уровне.

Интересные случаи, с которыми вас может столкнуть жизнь.
1) *В целом регрессия незначима, а отдельные коэффициенты значимы.* Об этом уже говорилось. Отдельные коэффициенты могут быть значимыми из-за ошибок первого рода. Особенно эти ошибки вероятны, если в модели много переменных. Может быть и другой вариант — ошибка второго рода («правильная» регрессия случайно оказалась незначимой). Так или иначе, данные не дают веского свидетельства связи между объясняемой переменной и объясняющими.
2) *В целом регрессия значима, а отдельные коэффициенты незначимы.* Такое случается при нестрогой мультиколлинеарности — тесной зависимости между объясняющими переменными. Интуитивное объяснение примерно такое: если несколько регрессоров ведут себя почти одинаково, каждый по отдельности можно выкинуть из модели и она почти не ухудшится. Но если выкинуть все регрессоры, модель перестаёт отражать реально существующую связь (это «видит» *F*-статистика и даёт сигнал признать регрессию значимой). То есть объясняемая величина связана с какими-то регрессорами, а с какими — непонятно.

Конечно, есть и другой вариант — при проверке значимости в целом произошла ошибка первого рода. Никакой связи нет, а вам показалось, что есть. Но над ошибкой первого рода у вас больше власти, чем над ошибкой второго рода — вы ведь сами выбираете уровень значимости.

Пример вывода результатов регрессионного анализа

Естественно, все процедуры оценивания и проверки гипотез сейчас делаются специализированными программами. Программ много, но формат вывода у них похож. Вернёмся к регрессии цен коттеджных участков (см. предыдущую лекцию):

$$\ln Price_i = \beta_1 + \beta_2 \ln Area_i + \beta_3 \ln House_i + \beta_4 \ln Dist_i + \beta_5 Eco_i + \epsilon_i .$$

Ниже пример команды оценивания вместе с выводом программы Stata.

```
. regress ln_price ln_area ln_house ln_dist eco

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  4,    45) =   75.20
       Model |  55.8749116     4  13.9687279            Prob > F      =  0.0000
    Residual |  8.35841853    45  .185742634            R-squared     =  0.8699
-------------+------------------------------           Adj R-squared =  0.8583
       Total |  64.2333302    49  1.31088429            Root MSE      =  .43098

------------------------------------------------------------------------------
    ln_price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     ln_area |   .3134958    .164324     1.91   0.063    -.0174698    .6444613
    ln_house |   .8190779   .1013727     8.08   0.000     .6149029    1.023253
     ln_dist |  -.2688648   .0875249    -3.07   0.004    -.4451489   -.0925807
         eco |   .5300659   .1296547     4.09   0.000      .268928    .7912038
       _cons |  -.3395755   .6107133    -0.56   0.581    -1.569615    .8904643
------------------------------------------------------------------------------
```

Слева вверху бирюзовым я выделил *ESS* (Стата называет эту величину Model Sum of Squares вместо Explained Sum of Squares), *RSS* и *TSS*.

Справа вверху светло-оранжевым я подсветил:
> обычный и скорректированный R^2 ;
> оценку стандартного отклонения случайной ошибки $\hat{\sigma}_\epsilon = \sqrt{\hat{\sigma}_\epsilon^2}$ — Стата называет её Root MSE;
> результаты проверки значимости регрессии в целом: *F*-статистику (в скобках указано числа степеней свободы: 4 и 45) и *p*-значение (`Prob > F`).
> число наблюдений (на него надо обращать внимание — наблюдения, в которых неизвестно значение хотя бы одной из переменных модели, не используются при оценивании, поэтому число наблюдений, по которым оценивалась регрессия, может сильно отличаться от всего объёма выборки).

Нижняя часть посвящена оценкам коэффициентов и их значимости:
`Coef.` — оценённые коэффициенты;
`Std. Err.` — стандартные ошибки;
`t` — *t*-статистики для проверки значимости коэффициентов;
`P>|t|` — соответствующие *p*-значения;
`[95% Conf. Interval]` — 95% доверительные интервалы для коэффициентов.

Я выбрал Стату, потому что пользуюсь ей сам. Вывод другой программы будет похож. Для сравнения я оценил то же уравнение в Gretl — вывод можете посмотреть в Приложении. У меня стоит русифицированная версия, но я не советую вам ставить русифицированные версии ни Gretl, ни других статистических пакетов. В отличие от Статы, Gretl не приводит доверительные интервалы, зато отмечает значимость коэффициентов звёздочками.

Проверка гипотезы о линейном ограничении

В подавляющем большинстве случаев гипотезы о коэффициентах регрессии имеют линейный вид (линейная комбинация коэффициентов чему-то равна или несколько линейных комбинаций чему-то равны). Сейчас мы разберёмся, как проверить любую линейную гипотезу, а нелинейные рассматривать не будем.

Сначала на языке матриц (не унывайте, ниже будут примеры на языке людей). В общем виде гипотезу о линейном ограничении можно сформулировать так. $H_0: R\beta = r$, где R — какая-то матрица, а r — какой-то вектор подходящего размера. Например, вот матричная запись гипотезы о значимости регрессии в целом $H_0: \beta_2 = ... = \beta_k = 0$.

$$H_0: \begin{pmatrix} 0 & 0 & 0 & ... & 0 \\ 0 & 1 & 0 & ... & 0 \\ 0 & 0 & 1 & ... & 0 \\ ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ ... \\ \beta_k \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ ... \\ 0 \end{pmatrix}.$$

Альтернативная гипотеза всегда будет $H_A: R\beta \neq r$ — хотя бы одно из накладываемых ограничений не выполняется (в случае проверки значимости — хотя бы один из коэффициентов $\beta_2, ..., \beta_k$ отличен от нуля).

Пусть $q = rank(R)$. То есть q — число линейно независимых ограничений, накладываемых на коэффициенты. При проверке значимости в целом $q = k - 1$ — в матрице R всего $k - 1$ независимых строк, соответствующих $k - 1$ ограничению.

Проверка гипотезы опирается на сравнение качества подгонки двух моделей:
— регрессии без ограничения (просто обычной регрессии, коэффициенты которой вы изучаете);
— регрессии с ограничением (здесь при оценивании на коэффициенты накладывается ограничение $R\beta = r$).

Качество подгонки будем измерять суммой квадратов остатков RSS . Для неограниченной регрессии обозначим её RSS_{UR} (UR — UnRestricted), для ограниченной — RSS_R (Restricted). Обратите внимание: $RSS_R \geq RSS_{UR}$. Накладывая ограничение на коэффициенты мы не можем улучшить качество подгонки, от этого RSS может только вырасти.

Как оценить регрессию с ограничением?
Формально нужно оценить регрессию условным методом наименьших квадратов, то есть решить задачу условной минимизации:

$$(y - X\beta)'(y - X\beta) \to \min_{\beta},$$
$$R\beta = r.$$

У этой задачи есть аналитическое решение, но мы обойдёмся без него. Дело в том, что её можно свести к обычному оцениванию регрессии простым методом наименьших квадратов. Чуть позже я опишу, как это сделать, а пока что просто поверим, что мы можем оценить регрессию с ограничением и рассчитать RSS_R .

***F*-статистика для проверки гипотезы о линейном ограничении:**

$$F = \frac{(RSS_R - RSS_{UR})/q}{RSS_{UR}/(n-k)} \overset{H_0}{\sim} F_{q,n-k}.$$

Смысл таков. Если наложенное условие $R\beta=r$ почти не ухудшает качество подгонки, $RSS_R \approx RSS_{UR}$, то данные согласуются с основной гипотезой и значение статистики близко к нулю. Если же у ограниченной регрессии качество много хуже, то основную гипотезу наверное стоит отвергнуть, при этом *F*-статистика будет принимать большие значения.

Критерий: отвергнуть H_0, если $F>F_{q,n-k,\alpha}$, где α — выбранный уровень значимости.

$p-$значение $=P(W>F)$, где $W\sim F_{q,n-k}$, а F — рассчитанное значение *F*-статистики.

Пример 1: проверка совместной значимости коэффициентов.

По данным переписи США 1980 года оценивается зависимость

$$drate_i=\beta_1+\beta_2 medage_i+\beta_3 medage_i^2+\beta_4 pcturban_i+\epsilon_i,$$

где $drate_i$ — коэффициент смертности в штате *i* (число умерших на 1000 населения),
$medage_i$ — медианный возраст в штате *i*,
$pcturban_i$ — процент населения штата *i*, проживающего в городской местности.

Всего штатов $n=50$. Данные находятся в файлах «census.ods» и «census.dta».

Вопрос: есть ли основания считать, что смертность зависит от медианного возраста? Понятно, что есть, но мы ответим на этот вопрос статистически, выбрав уровень значимости 5%.

Нужно проверить гипотезу $H_0:\beta_2=\beta_3=0$ против альтернативы $H_A:\beta_2^2+\beta_3^2>0$. Заметьте, что по отдельности гипотезы $\beta_2=0$ и $\beta_3=0$ не подходят по смыслу: про первую вообще трудно сказать, что она означает, а вторая говорит, что зависимость смертности от возраста не квадратична — это нам не нужно.

Число линейно независимых ограничений $q=2$, вот эти ограничения: (1) $\beta_2=0$ и (2) $\beta_3=0$.

Как подсчитать число ограничений. Для этого не нужно задумываться над матрицами и рангами. Просто посчитайте число знаков “=” в записи основной гипотезы.

Это не сработает, если пользоваться извращёнными способами формулировки гипотез. В нашем случае извращённо можно записать так: $H_0:\beta_2^2+\beta_3^2=0$ (нелинейная запись). А ещё этак: $H_0:\beta_2=\beta_3=0;\ 2\beta_2=2\beta_3=0$ (линейно зависимые ограничения). Я обещаю так не делать. Давайте записывать основную гипотезу в виде набора независимых и линейных ограничений.

Вернёмся к регрессии. Оценив уравнение, получаем (ниже `medage_sq` — это $medage^2$):

```
. reg drate medage medage_sq pcturban
      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  3,    46) =   31.47
       Model |  55.9301763     3  18.6433921            Prob > F      =  0.0000
    Residual |  27.2486414    46  .592361771            R-squared     =  0.6724
-------------+------------------------------           Adj R-squared =  0.6510
       Total |  83.1788177    49  1.69752689            Root MSE      =  .76965
------------------------------------------------------------------------------
       drate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      medage |    .485067   1.207049     0.40   0.690    -1.944596     2.91473
   medage_sq |   .0023664   .0205546     0.12   0.909    -.0390079    .0437407
    pcturban |  -.0353483   .0082932    -4.26   0.000    -.0520416   -.0186549
       _cons |  -5.597969    17.8979    -0.31   0.756    -41.62457    30.42863
------------------------------------------------------------------------------
```

Выделенное число 27.25 — это сумма квадратов остатков, RSS_{UR} .

Как оценить регрессию с ограничением? Условие $\beta_2=\beta_3=0$ очень легко учесть в модели — нужно просто выкинуть из уравнения переменные *medage* и $medage^2$, это тоже самое, что приравнять у нулю коэффициенты при них. Значит, уравнение с ограничением выглядит так:

$$drate_i=\beta_1+\beta_4\, pcturban_i+\epsilon_i.$$

Оценим его.

```
. reg drate pcturban

      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  1,    48) =    2.28
       Model |  3.77088896     1  3.77088896           Prob > F      =  0.1377
    Residual |  79.4079287    48  1.65433185           R-squared     =  0.0453
-------------+------------------------------           Adj R-squared =  0.0254
       Total |  83.1788177    49  1.69752689           Root MSE      =  1.2862

------------------------------------------------------------------------------
       drate |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    pcturban |  -.0192519   .0127515    -1.51   0.138    -.0448906    .0063868
       _cons |    9.72488   .8728678    11.14   0.000     7.969861     11.4799
------------------------------------------------------------------------------
```

Снова я выделил сумму квадратов остатков, $RSS_R=79.41$.

Рассчитываем *F*-статистику:

$$F=\frac{(RSS_R-RSS_{UR})/q}{RSS_{UR}/(n-k)}=\frac{(79.41-27.25)/2}{27.25/(50-4)}=44.02.$$

Обратите внимание: в знаменателе фигурирует $k=4$ — число коэффициентов в исходном, не ограниченном, уравнении.

Критическое значение: $F_{q,n-k,\alpha}=F_{2,46,0.05}=3.20$.

Вывод: $F>F_{q,n-k,\alpha}$, поэтому основная гипотеза отвергается. Хотя бы из коэффициентов β_2 и β_3 отличается от нуля, а это значит, что есть связь смертности с медианным возрастом.

Stata согласна с моим выводом. Если воспользоваться готовой командой для проверки гипотез, получается так:

```
. reg drate medage medage_sq pcturban
. test medage medage_sq

 ( 1)  medage = 0
 ( 2)  medage_sq = 0

       F(  2,    46) =   44.03
            Prob > F =    0.0000
```

Цветом выделены *F*-статистика и *p*-значение. Как видно, связь прослеживается на любом уровне значимости.

Заметьте. В нашем уравнении коэффициенты при переменных *medage* и $medage^2$ по отдельности незначимы (*p*-значения равны 0.690 и 0.909 соответственно), а вместе два коэффициента

значимы. Причина — тесная связь между переменными. Она не линейная, но близкая к линейной (выборочный коэффициент корреляции между *medage* и $medage^2$ равен ?????) Поэтому модель мало теряет, если выкинуть любую одну из этих переменных, а вот если выкинуть обе, то качество подгонки снижается существенно, что и замечает *F*-статистика. Это пример нестрогой мультиколлинеарности, о которой я уже упоминал повыше, говоря о проверке значимости регрессии в целом.

Когда проверяется совместная значимость? Чаще всего — когда один статистический признак отражается несколькими переменными. В рассмотренном примере возраст учитывается двумя переменными *medage* и $medage^2$. Можно вспомнить ещё уравнение заработной платы из лекции про дамми-переменные — там образование работника учитывалось четырьмя дамми. Качественные объясняющие признаки с более чем двумя категориями включаются в модель как набор переменных, коэффициенты при которых проверяются совместно.

Почти философские размышления. В разобранном примере есть неясное место. Какой вообще смысл проверять гипотезы об истинных коэффициентах регрессии? Обычно параметр — это то, что относится к генеральной совокупности (иногда воображаемой), а оценка — то, что относится к случайной выборке. Расчёт доверительных интервалов и проверка гипотез — это распространение результатов выборочного обследования на генеральную совокупность. В нашем случае роль выборки играют все штаты США, а что такое генеральная совокупность? Все штаты, которые могли бы быть? Штаты из параллельных вселенных? Зачем нам вообще знать о них что-то, почему бы не довольствоваться просто выборочными характеристиками? Я на этот вопрос отвечать не буду.

Пример 2: проверка бессмысленных ограничений.

Равенство нескольких коэффициентов нулю — самый частый вид линейных ограничений, но не единственный. Рассмотрим отвлечённое уравнение

$$Y_i=\beta_1+\beta_2 X_{2,i}+\beta_3 X_{3,i}+\beta_4 X_{4,i}+\epsilon_i.$$

Гипотезы:

$$H_0: \beta_2=-\beta_3;\ \beta_4=0.$$
$$H_A: \beta_2\neq-\beta_3 \text{ или } \beta_4\neq0.$$

Число ограничений $q=2$.
Для регрессии без ограничений получили

$$\hat{Y}=1.3+0.2\,X_2-0.4\,X_3-0.5\,X_4,\ \ RSS=45.2,\ n=20.$$

Как оценить регрессию с ограничением? Равенство $\beta_4=0$ учесть легко — нужно выкинуть из уравнения переменную X_4 . Чтобы учесть равенство $\beta_2=-\beta_3$ можно вместо отдельных переменных X_2 и X_3 включить в уравнение их разность X_2-X_3 . Получим уравнение с ограничением:

$$Y_i=\beta_1+\beta_2(X_{2,i}-X_{3,i})+\epsilon_i.$$

Коэффициент при новом регрессоре X_2-X_3 — это сразу и β_2, и $-\beta_3$, так что нужное ограничение выполнено.
Пусть оценённое уравнение выглядит так:

$$\hat{Y}=-10.2+0.3(X_2-X_3),\ \ RSS=47.0,\ n=20.$$

Теперь можно рассчитать *F*-статистику: $F=\dfrac{(47.0-45.2)/2}{45.2/(20-4)}=\dfrac{0.9}{2.825}=0.32.$

Критическое значение для уровня значимости 5%: $F_{q,n-k,\alpha} = F_{2,16,0.05} = 3.63$.

Вывод: не отвергаем основную гипотезу, потому что F < 3.63. Нет оснований считать, что ограничение не выполняется.

На этом всё.

Подытожим

▶ Гипотеза $H_0: \beta_j = \beta_j^0$ проверяется статистикой $t = \frac{\hat{\beta}_j - \beta_j^0}{\hat{\sigma}(\hat{\beta}_j)} \overset{H_0}{\sim} t_{n-k}$.

▶ Критерий при проверке $H_0: \beta_j = \beta_j^0$ против $H_A: \beta_j \neq \beta_j^0$ таков: основная гипотеза отвергается на уровне значимости α, если $|t| > t_{n-k,\alpha/2}$.

▶ Проверка гипотезы $H_0: \beta_j = 0$ против $H_A: \beta_j \neq 0$ называется проверкой значимости (оценённого) коэффициента регрессии, а если основная гипотеза отвергается, говорят, что коэффициент $\hat{\beta}_j$ значим. В этом случае есть основания считать, что изменения объясняющей переменной X_j статистически связаны с изменениями объясняемой переменной Y.

▶ Коэффициент $\hat{\beta}_j$ незначим, если 1) оцениваемой статистической связи нет, то есть $\beta_j = 0$, или 2) оценка $\hat{\beta}_j$ имеет большую стандартную ошибку — коэффициент β_j не удаётся точно оценить.

▶ Проверка значимости регрессии в целом — это проверка гипотезы $H_0: \beta_2 = \beta_3 = ... = \beta_k = 0$ против $H_A: \beta_2^2 + \beta_3^2 + ... + \beta_k^2 > 0$.

▶ Статистика для проверки значимости в целом:

$$F = \frac{ESS/(k-1)}{RSS/(n-k)} = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} \overset{H_0}{\sim} F_{k-1,n-k}.$$

▶ Проверка гипотезы о линейном ограничении на коэффициенты регрессии $H_0: R\beta = r$ против альтернативы $H_A: R\beta \neq r$ проводится с помощью статистики

$$F = \frac{(RSS_R - RSS_{UR})/q}{RSS_{UR}/(n-k)} \overset{H_0}{\sim} F_{q,n-k},$$

где $q = rank(R)$ — число линейно независимых ограничений в основной гипотезе,
RSS_R и RSS_{UR} — суммы квадратов остатков у регрессии с ограничением (**R**estricted) и без ограничения (**U**n**R**estricted),
k — число коэффициентов регрессии без ограничения.

▶ Как правило, число ограничений q равно числу знаков “=” в записи основной гипотезы.

▶ Частный случай линейного ограничения — гипотеза о равенстве нулю нескольких коэффициентов регрессии. Проверку этой гипотезы называют проверкой совместной значимости коэффициентов.

▶ Проверка совместной значимости важна, когда один объясняющий признак учитывается несколькими переменными. Тогда гипотеза о том, что этот признак не связан с объясняемой переменной, означает равенство нулю сразу нескольких коэффициентов.

В следующий раз

▶ Метод главных компонент.

Приложение. Вывод результатов регрессионого анализа программой gretl.

```
? ols ln_price const ln_area ln_house ln_dist eco

Модель 1: МНК, использованы наблюдения 1-50
Зависимая переменная: ln_price

             Коэффициент   Ст. ошибка   t-статистика   P-значение
  ------------------------------------------------------------------
  const       -0,339576    0,610713       -0,5560      0,5809
  ln_area      0,313496    0,164324        1,908       0,0628     *
  ln_house     0,819078    0,101373        8,080       2,64e-010  ***
  ln_dist     -0,268865    0,0875249      -3,072       0,0036     ***
  eco          0,530066    0,129655        4,088       0,0002     ***

Среднее зав. перемен    3,784721   Ст. откл. зав. перемен  1,144939
Сумма кв. остатков      8,358419   Ст. ошибка модели       0,430979
R-квадрат               0,869874   Испр. R-квадрат         0,858307
F(4, 45)                75,20475   P-значение (F)          2,43e-19
Лог. правдоподобие     -26,22808   Крит. Акаике            62,45616
Крит. Шварца            72,01628   Крит. Хеннана-Куинна    66,09671
```