Классическая линейная нормальная регрессионная модель
и оценивание её параметров, часть II

Краткое содержание предыдущей части.
Предпосылки классической линейной нормальной регрессионной модели (КЛНРМ):

1. $y = X\beta + \epsilon$.
2. Матрица регрессоров X детерминирована, $rank(X) = k$, где k — число столбцов матрицы X.
3. ϵ — случайный вектор, такой что:
3a. $E(\epsilon) = 0$;
3b. $V(\epsilon) = \sigma_\epsilon^2 I_n = \begin{pmatrix} \sigma_\epsilon^2 & 0 & \dots & 0 \\ 0 & \sigma_\epsilon^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_\epsilon^2 \end{pmatrix}$.
3c. $\epsilon \sim N(0, \sigma_\epsilon^2 I_n)$.

Мы рассмотрели оценку МНК для вектора коэффициентов регрессии β:

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

По теореме Гаусса–Маркова эта оценка несмещённая и эффективная в классе линейных несмещённых оценок. Свойства замечательные, но, как обычно, точечных оценок мало для надёжных выводов об истинных значениях коэффициентов β.
Хорошо бы знать, насколько могут оценки могут отклоняться от истинных коэффициентов. Для этого желательно знать их дисперсии, а лучше — всё распределение вектора оценок. А ещё лучше — построить доверительные интервалы для истинных коэффициентов регрессии. Этим мы теперь и будем заниматься.

На самом деле, ковариационную матрицу мы уже знаем из той же теоремы Гаусса–Маркова:

$$V(\hat{\beta}) = \sigma_\epsilon^2 (X'X)^{-1}.$$

Беда в том, что в этой формуле участвует неизвестный параметр σ_ϵ^2 — дисперсия случайной ошибки регрессии. Первое, что мы сделаем — научимся его оценивать.

Оценка дисперсии случайной ошибки.
Может показаться, что с этой задачей вы уже сталкивались — дисперсию к настоящему моменту стоило хоть раз оценить каждому. Кто ещё не оценивал, вряд ли читает этот текст. Однако изюминка тут есть: до сих пор вы оценивали дисперсии случайных величин по наблюдениям за ними, а ошибки регрессии $\epsilon_1, \dots, \epsilon_n$ ненаблюдаемы. В данных их нет. Это отклонения наблюдаемых значений $Y_1, \dots, Y_n$ от истинной линии (гиперплоскости) регрессии: $\epsilon = y - X\beta$. Но истинные значения β неизвестны, поэтому и ошибок мы не видим.
Конечно, есть хорошая оценка для β, которую даёт МНК, и мы можем рассчитать остатки — отклонения значений $Y_1, \dots, Y_n$ от оценённой зависимости:

$$e = y - \hat{y} = y - X\hat{\beta}.$$

Можно считать остатки как бы оценками для ненаблюдаемых случайных ошибок и попытаться по ним оценить σ_ϵ^2. Так и сделаем.

Утверждение. $E(RSS) = (n-k)\sigma_\epsilon^2$, где $RSS = \sum_{i=1}^{n} e_i^2$ — сумма квадратов остатков (residual sum of squares).

Напомню, что n — число наблюдений, а k — число коэффициентов в векторе β. Доказательство я привёл в Приложении 1 в конце этого документа.

Отсюда получаем несмещённую оценку дисперсии случайной составляющей:

$$\hat{\sigma}_\epsilon^2 = \frac{RSS}{n-k}.$$

Как и теорема Гаусса–Маркова, эта оценка не опирается на предпосылку **3c** о нормальности случайной составляющей, достаточно выполнения предпосылок (**1**)–(**3b**) КЛНРМ.

Скорректированный коэффициент детерминации. При добавлении объясняющей переменной в уравнение регрессии сумма квадратов остатков RSS, как правило, падает (и никогда не возрастает), а коэффициент детерминации R^2 не может упасть и обычно увеличивается, поэтому его не стоит использовать для сравнения моделей с разным числом объясняющих переменных — R^2 отдаёт предпочтение «длинным» моделям, где много регрессоров. А вот выведенная оценка дисперсии ошибки может увеличиться, если добавленная переменная почти не улучшает качество подгонки (т. е. почти не уменьшает RSS). Этим можно воспользоваться и предложить альтернативу R^2 — скорректированный (поправленный, нормированный) коэффициент детерминации:

$$R_{adj}^2 = 1 - \frac{\hat{\sigma}_\epsilon^2}{TSS/(n-1)} = 1 - \frac{RSS/(n-k)}{TSS/(n-1)}.$$

Индекс *adj* — сокращение от adjusted (поправленный). Когда оценка дисперсии падает, R_{adj}^2 увеличивается, и наоборот.

Обратите внимание на знаменатель: $TSS/(n-1) = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$. Он выглядит как несмещённая оценка дисперсии Y. Но будьте осторожны: на самом деле по предпосылкам классической модели величины Y_i в разных наблюдениях имеют разные математические ожидания: $E(Y_i) = E(\beta_1 + \beta_2 X_{2,i} + \ldots + \beta_k X_{k,i} + \epsilon_i) = \beta_1 + \beta_2 X_{2,i} + \ldots + \beta_k X_{k,i}$ — поэтому обычная оценка дисперсии для Y не работает. Более того, дисперсия Y_i совпадает с дисперсией ошибки: $D(Y_i) = D(\beta_1 + \beta_2 X_{2,i} + \ldots + \beta_k X_{k,i} + \epsilon_i) = D(\epsilon_i)$, потому что объясняющие переменные детерминированы.

Можно сказать, что в скорректированный коэффициент детерминации внесён штраф за число коэффициентов k, поэтому он больше пригоден для сравнения моделей с разным числом регрессоров. Правда, скорректированный коэффициент может принимать отрицательные значения и не имеет столь ясной интерпретации как обычный R^2.

Оценка ковариационной матрицы оценок МНК получается просто заменой неизвестной дисперсии σ_ϵ^2 в выражении $V(\hat{\beta}) = \sigma_\epsilon^2 (X'X)^{-1}$ на оценку $\hat{\sigma}_\epsilon^2$:

$$\hat{V}(\hat{\beta}) = \hat{\sigma}_\epsilon^2 (X'X)^{-1}.$$

Так как оценка $\hat{\sigma}_\epsilon^2$ несмещённая, а $(X'X)^{-1}$ — детерминированный множитель, то оценка ковариационной матрицы получается также несмещённой:

$$E[\hat{V}(\hat{\beta})] = E[\hat{\sigma}_\epsilon^2 (X'X)^{-1}] = E(\hat{\sigma}_\epsilon^2)(X'X)^{-1} = \sigma_\epsilon^2 (X'X)^{-1} = V(\hat{\beta}).$$

Пример. В таблице ниже приведены рост и вес пяти человек. Оценим регрессию веса на рост, рассчитаем обычный и скорректированный коэффициент детерминации и оценку ковариационной матрицы оценок МНК.

№	**Вес**	**Рост**
1	90	180
2	60	160
3	75	175
4	80	185
5	70	175

Представим данные в виде вектора объясняемой переменной *y* и матрицы регрессоров *X*:

$$y=\begin{pmatrix}90\\60\\75\\80\\70\end{pmatrix};\quad X=\begin{pmatrix}1 & 180\\1 & 160\\1 & 175\\1 & 185\\1 & 175\end{pmatrix}.$$

Для оценок коэффициентов и их ковариаций нам пригодятся вот эти матрицы:

$$(X'X)^{-1}=\begin{pmatrix}87.7 & -0.5\\-0.5 & 0.0029\end{pmatrix};\quad X'y=\begin{pmatrix}375\\65975\end{pmatrix}.$$

Оценённые коэффициенты:

$$\hat{\beta}=(X'X)^{-1}X'y=\begin{pmatrix}-100\\1\end{pmatrix}.$$

Оценённое уравнение регрессии: $\hat{Y}=-100+X$.

Считаем прогнозы и остатки:

№	**Вес**	**Рост**	**Прогноз веса (Рост - 100)**	**Остаток**
1	90	180	80	10
2	60	160	60	0
3	75	175	75	0
4	80	185	85	–5
5	70	175	75	–5

Получаем $RSS=\sum_{i=1}^{5} e_i^2=10^2+0^2+0^2+(-5)^2+(-5)^2=150;$

$$TSS=\sum_{i=1}^{5}(Y_i-\bar{Y})^2=(90-75)^2+(60-75)^2+(75-75)^2+(80-75)^2+(70-75)^2=350.$$

Оценка дисперсии случайной ошибки: $\hat{\sigma}_\epsilon^2=\frac{RSS}{n-k}=\frac{150}{5-2}=50.$

Коэффициент детерминации: $R^2=1-\frac{RSS}{TSS}=1-\frac{150}{350}=0.7.$

Скорректированный коэффициент детерминации: $R_{adj}^2=1-\frac{RSS/(n-k)}{TSS/(n-1)}=1-\frac{150/(5-2)}{350/(5-1)}=0.6.$

Оценка ковариационной матрицы:

$$\hat{V}(\hat{\beta})=\hat{\sigma}_{\epsilon}^{2}(X'X)^{-1}=50\times\begin{pmatrix}87.7 & -0.5\\ -0.5 & 0.0029\end{pmatrix}=\begin{pmatrix}4385 & -25\\ -25 & 0.1429\end{pmatrix}.$$

Стандартные ошибки. На главной диагонали ковариационной матрицы стоят оценки дисперсий оценок коэффициентов, обозначу их $\hat{D}(\hat{\beta}_j)$. Корни из элементов главной диагонали — оценки стандартных отклонений оценок коэффициентов — принято кратко называть стандартными ошибками (оценок) коэффициентов: $\hat{\sigma}(\hat{\beta}_j)=\sqrt{\hat{D}(\hat{\beta}_j)}$.

Привычнее было бы обозначение $\hat{\sigma}_{\hat{\beta}_j}$, но индексы в два уровня трудночитаемы, поэтому буду использовать скобки. И точнее было бы говорить, что $\hat{\sigma}(\hat{\beta}_j)$ — это оценка стандартной ошибки оценки коэффициента регрессии, но это тяжеловесно. В обыденной речи говорят просто «стандартная ошибка коэффициента».

В рассмотренном примере стандартная ошибка свободного члена $\hat{\sigma}(\hat{\beta}_1)=\sqrt{4385}=66.22$, стандартная ошибка коэффициента при росте $\hat{\sigma}(\hat{\beta}_2)=\sqrt{0.1429}=0.38$.

Стандартные ошибки — традиционная мера неточности оценок, их принято указывать вместе под оценками МНК в скобках. Уравнение из разобранного пример можно прилично представить так:

$$\hat{Y}=\underset{(66.22)}{-100.0}+\underset{(0.38)}{1.0}\,X,\quad R^2=0.7.$$

При взгляде на оценённое уравнение опытный читатель представляет величину возможных отклонений оценок от истинных коэффициентов регрессии. Как он это делает, сейчас разберём.

Теорема Фишера для регрессии

Оценённые коэффициенты $\hat{\beta}_j$ и стандартные ошибки $\hat{\sigma}(\hat{\beta}_j)$ позволяют понять, какие значения могут иметь истинные коэффициенты регрессии β_j (сделать *статистический вывод* о коэффициентах) . Сейчас мы и переходим к статистическому выводу: доверительным интервалам для коэффициентов регрессии и проверке гипотез об этих коэффициентах. Чтобы понять, откуда берутся эти интервалы и критерии, рассмотрим сначала обобщение теоремы Фишера для регрессии.

Теорема. Пусть выполнены все предпосылки КЛНРМ (теперь требуется и нормальность случайной составляющей). Тогда:
1) оценка МНК $\hat{\beta}=(X'X)^{-1}X'y$ имеет многомерное нормальное распределение, $\hat{\beta}\sim N(\beta,\sigma_{\epsilon}^{2}(X'X)^{-1})$;
2) распределение оценки дисперсии $\hat{\sigma}_{\epsilon}^{2}$ задаётся выражением $\dfrac{(n-k)\hat{\sigma}_{\epsilon}^{2}}{\sigma_{\epsilon}^{2}}\sim\chi_{n-k}^{2}$;
3) $\hat{\beta}$ и $\hat{\sigma}_{\epsilon}^{2}$ независимы.

Доказательство теоремы можно найти в третьей главе книги Я.Р. Магнуса, П.К. Катышева и А.А. Пересецкого «Эконометрика. Начальный курс». Она там не называется и не выделяется никак, но все три утверждения доказываются.

Дальше нам пригодится величина $\dfrac{\hat{\beta}_j-\beta_j}{\hat{\sigma}(\hat{\beta}_j)}$ - центрированная и нормированная (на оценку стандартного отклонения) оценка коэффициента регрессии β_j (здесь j — номер коэффициента, любое число от 1 до k). На неё опирается доверительный интервал для этого коэффициента.

Следствие 1 из теоремы Фишера: $\frac{\hat{\beta}_j-\beta_j}{\sigma(\hat{\beta}_j)} \sim N(0,1)$.

Это очевидно. Из утверждения 1 теоремы следует, что оценка $\hat{\beta}_j$ нормально распределена. Мы вычли её математическое ожидание, поделили на стандартное отклонение, получили стандартную нормальную величину.

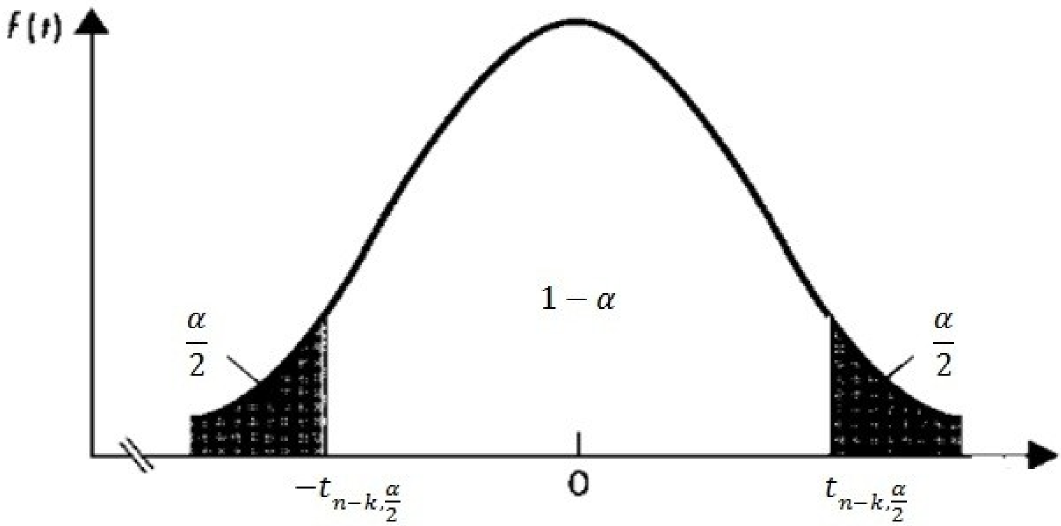
Следствие 2: $\frac{\hat{\beta}_j-\beta_j}{\hat{\sigma}(\hat{\beta}_j)} \sim t_{n-k}$.

Если вы разобрались с выводом доверительного интервала для среднего при неизвестной дисперсии, то сможете доказать следствие 2 самостоятельно. Будут трудности — посмотрите презентацию к лекции 3.

Доверительный интервал для коэффициента регрессии $\beta_j,\ j=1,\dots,k$.

Этот интервал выводится из следствия 2 аналогично доверительному интервалу для среднего.

Выбираем уровень доверия $1-\alpha$, так что α — вероятность ошибки, и делим график плотности распределения Стьюдента с $n-k$ степенями свободы на три части:



Как обычно, здесь $t_{n-k,\frac{\alpha}{2}}$ — такое число, которое отрезает вероятность $\frac{\alpha}{2}$ с правого хвоста распределения. Если мы возьмём случайную величину $U \sim t_{n-k}$, то $P(U>t_{n-k,\frac{\alpha}{2}})=\frac{\alpha}{2}$ и

$$P(-t_{n-k,\frac{\alpha}{2}}<U<t_{n-k,\frac{\alpha}{2}})=1-\alpha.$$

Вместо *U* тут можно подставить любую величину с тем же распределением, например так:

$$1-\alpha=P\left(-t_{n-k,\frac{\alpha}{2}}<\frac{\hat{\beta}_j-\beta_j}{\hat{\sigma}(\hat{\beta}_j)}<t_{n-k,\frac{\alpha}{2}}\right).$$

Остаётся преобразовать неравенство в скобках так, чтобы оцениваемый коэффициент остался в середине, а всё остальное разбрелось по бокам:

$$1-\alpha=P\left(-t_{n-k,\frac{\alpha}{2}}\hat{\sigma}(\hat{\beta}_j)<\hat{\beta}_j-\beta_j<t_{n-k,\frac{\alpha}{2}}\hat{\sigma}(\hat{\beta}_j)\right)=P\left(\hat{\beta}_j-t_{n-k,\frac{\alpha}{2}}\hat{\sigma}(\hat{\beta}_j)<\beta_j<\hat{\beta}_j+t_{n-k,\frac{\alpha}{2}}\hat{\sigma}(\hat{\beta}_j)\right).$$

Вот и получился доверительный интервал для коэффициента β_j с уровнем доверия $1-\alpha$:

$$\hat{\beta}_j - t_{n-k,\frac{\alpha}{2}}\hat{\sigma}(\hat{\beta}_j) < \beta_j < \hat{\beta}_j + t_{n-k,\frac{\alpha}{2}}\hat{\sigma}(\hat{\beta}_j).$$

Пример. Файлы «коттеджи_полный.ods», «cottages.dta» и «cottages.gdt» содержат сведения о 50 коттеджных участках. Выборка одна и та же, только записана в разных форматах (Open Office, Stata и Gretl). Частично с этими данными вы могли познакомиться, изучая корреляционный анализ.
По этим данным я оценил модель

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

где $Price_i$ — цена участка *i*, тыс. долл.,
$Area_i$ — площадь участка, сотки,
$House_i$ — площадь дома, м^2,
$Dist_i$ — расстояние до МКАД, км,
$Eco_i = 1$, если участок расположен рядом с водоёмом (речкой или озером), и 0 иначе.

Вот вектор оценок МНК и оценка его ковариационной матрицы:

$$\hat{\beta} = \begin{pmatrix} -0.34 \\ 0.31 \\ 0.82 \\ -0.27 \\ 0.53 \end{pmatrix}; \quad \hat{V}(\hat{\beta}) = \begin{pmatrix} 0.373 & -0.0003 & -0.044 & -0.042 & -0.012 \\ -0.0003 & 0.027 & -0.010 & -0.005 & -0.004 \\ -0.044 & -0.010 & 0.010 & 0.005 & 0.0009 \\ -0.042 & -0.005 & 0.005 & 0.008 & 0.003 \\ -0.012 & -0.004 & 0.0009 & 0.003 & 0.017 \end{pmatrix}.$$

Взяв корни диагональных элементов ковариационной матрицы, я получаю стандартные ошибки оценок коэффициентов:

$$\hat{\sigma}(\hat{\beta}_1) = \sqrt{0.373} = 0.61; \quad \hat{\sigma}(\hat{\beta}_2) = \sqrt{0.027} = 0.16; \quad \hat{\sigma}(\hat{\beta}_3) = 0.10; \quad \hat{\sigma}(\hat{\beta}_4) = 0.09; \quad \hat{\sigma}(\hat{\beta}_5) = 0.13.$$

Значит, оценённое уравнение выглядит так:

$$\ln \widehat{Price}_i = -\underset{(0.61)}{0.34} + \underset{(0.16)}{0.31} \ln Area_i + \underset{(0.10)}{0.82} \ln House_i - \underset{(0.09)}{0.27} \ln Dist_i + \underset{(0.61)}{0.53} Eco_i.$$

Теперь посмотрим, какую прибавку к цене участка даёт наличие водоёма поблизости. Начнём с интерпретации точечной оценки. Коэффициент 0.53 при переменной *Eco* означает, что прогнозируемый логарифм цены для участка рядом с водоёмом на 0.53 больше, чем у участка с теми же характеристиками, но без водоёма. То есть прогнозируемая цена (без логарифма) больше в $\exp(0.53) = 1.70$ раза, или на 70%.

Теперь рассчитаем 95% доверительный интервал для β_5 (коэффициента при *Eco*).

$$\hat{\beta}_5 - t_{n-k,\frac{\alpha}{2}}\hat{\sigma}(\hat{\beta}_5) < \beta_5 < \hat{\beta}_5 + t_{n-k,\frac{\alpha}{2}}\hat{\sigma}(\hat{\beta}_5).$$

В данном случае $\hat{\beta}_5 = 0.53$, $\hat{\sigma}(\hat{\beta}_5) = 0.13$, $t_{n-k,\frac{\alpha}{2}} = t_{50-5,\frac{0.05}{2}} = 2.014$, так что

$$0.53 - 2.014 \times 0.13 < \beta_5 < 0.53 + 2.014 \times 0.13;$$
$$0.27 < \beta_5 < 0.79.$$

Итак, наличие водоёма означает в среднем прибавку к логарифму цены в размере от 0.27 до 0.79. Потенцируем доверительный интервал:

$$e^{0.27} < e^{\beta_5} < e^{0.79};$$
$$1.31 < e^{\beta_5} < 2.20.$$

Получается, что средняя прибавка к цене участка от наличия водоёма поблизости находится в пределах от 31% до 120%. Не ахти какая точность. Чтобы получить доверительный интервал поуже, нужно иметь больше данных.

***Как прикинуть 95% доверительный интервал «на глаз»*. Если число степеней свободы $n-k$ не очень мало, то $t_{n-k,0.05/2}\approx 2$, поэтому границы 95% доверительного интервала отстоят от точечной оценки примерно на две стандартных ошибки: $\hat{\beta}_j \pm 2\hat{\sigma}(\hat{\beta}_j)$.**

Подытожим

▶ После того, как коэффициенты регрессии $y = X\beta + \epsilon$ оценены и рассчитаны остатки e , можно оценить дисперсию случайной ошибки:

$$\hat{\sigma}_\epsilon^2 = \frac{RSS}{n-k},$$

где $RSS = e'e = \sum_{i=1}^{n} e_i^2$ — сумма квадратов остатков.

▶ Оценка для ковариационной матрицы оценок МНК: $\hat{V}(\hat{\beta}) = \hat{\sigma}_\epsilon^2 (X'X)^{-1}$. На её главной диагонали стоят оценённые дисперсии оценок коэффициентов, $\hat{D}(\hat{\beta}_j),\ j=1,\ldots,k$.

▶ Корни из оценённых дисперсий называют стандартными ошибками (оценок) коэффициентов: $\hat{\sigma}(\hat{\beta}_j) = \sqrt{\hat{D}(\hat{\beta}_j)}$. Это традиционно используемая характеристика точности оценок. Их принято выписывать вместе с оценёнными коэффициентами. Например, так:

$$\hat{Y} = \underset{(5.4)}{89.1} + \underset{(0.2)}{0.7}\, X_2 - \underset{(1.3)}{0.9}\, X_3.$$

▶ Кроме обычного коэффициента детерминации в качестве меры объясняющей способности регрессии можно использовать скорректированный коэффициент детерминации:

$$R_{adj}^2 = 1 - \frac{RSS/(n-k)}{TSS/(n-1)}.$$

Он учитывает число оцениваемых коэффициентов и в отличие от простого R^2 может падать при добавлении объясняющей переменной, если она не позволяет заметно улучшить качество подгонки (он «штрафует» модель за лишние переменные). К сожалению, R_{adj}^2 не имеет столь ясной интерпретации как обычный коэффициент детерминации.

▶ Теорема Фишера распространяется на классическую линейную нормальную регрессионную модель, она говорит, какое распределение имеют оценки коэффициентов и дисперсии случайной ошибки.

▶ Доверительный интервал с уровнем доверия $1-\alpha$ для коэффициента регрессии:
$\hat{\beta}_j - t_{n-k,\frac{\alpha}{2}}\hat{\sigma}(\hat{\beta}_j) < \beta_j < \hat{\beta}_j + t_{n-k,\frac{\alpha}{2}}\hat{\sigma}(\hat{\beta}_j).$

В следующий раз:

- проверка гипотез о коэффициентах регрессии;
- байка про регрессию, политику и алкоголь.

ПРИЛОЖЕНИЕ. Доказательство утверждения $E(RSS)=(n-k)\sigma_\epsilon^2$.

Сначала получим полезное выражение для вектора остатков: $e=(I-P)\epsilon$, где $P=X(X'X)^{-1}X'$ — матрица-проектор на пространство регрессоров.

Выводим:

$$e=y-\hat{y}=y-Py=(I-P)y=(I-P)(X\beta+\epsilon)=IX\beta-PX\beta+I\epsilon-P\epsilon=$$
$$=X\beta-X(X'X)^{-1}X'X\beta+(I-P)\epsilon=X\beta-X\beta+(I-P)\epsilon=(I-P)\epsilon.$$

$(I-P)$ — матрица оператора ортогонального проецирования на ортогональное дополнение к пространству регрессоров. Как и все матрицы-проекторы она симметрична и идемпотентна (см. задачу 7 к лекции 13):

$$(I-P)'=I-P,\ (I-P)^2=I-P.$$

Дальше нам пригодятся математическое ожидание и ковариационное матрица остатков.

$E(e)=E[(I-P)\epsilon]=(I-P)E(\epsilon)=0.$

$V(e)=V[(I-P)\epsilon]=(I-P)V(\epsilon)(I-P)'=(I-P)\sigma_\epsilon^2 I(I-P)'=\sigma_\epsilon^2(I-P)(I-P)'=\sigma_\epsilon^2(I-P).$

Наконец перейдём к сумме квадратов остатков $RSS=\sum_{i=1}^{n}e_i^2$.

Обратите внимание: $E(e_i^2)=D(e_i)$, потому что $E(e_i)=0$. Значит,

$$E(RSS)=\sum_{i=1}^{n}E(e_i^2)=\sum_{i=1}^{n}D(e_i).$$

Дисперсии остатков стоят на главной диагонали ковариационной матрицы, поэтому сумма дисперсий — это след ковариационной матрицы остатков:

$$E(RSS)=tr(V(e))=tr[\sigma_\epsilon^2(I-P)].$$

Выношу общий множитель за знак следа:

$$E(RSS)=\sigma_\epsilon^2 tr(I-P).$$

Тут продвинутый читатель может заметить, что след матрицы-проектора обязательно равен её рангу и размерности пространства, на которое осуществляется проецирование. В нашем случае, *n*-мерный вектор y разделяется на две проекции: матрица *P* его проецирует на *k*-мерное пространство, а матрица — на $n-k$ -мерное, так что $tr(I-P)=n-k$.
Столь продвинутый читатель — редкая диковинка, поэтому дальше я привожу доказательство в лоб.

След разности — это разность следов, так что

$$E(RSS)=\sigma_\epsilon^2(tr\,I-tr\,P).$$

Матрицы *I* и *P* имеют размерность $n\times n$. На главной диагонали *I* стоят *n* единиц, которые в сумме дают *n*: $tr\,I=n$. С матрицей *P* чуть сложнее, мне пригодится свойство следа $tr(AB)=tr(BA)$ —

след не меняется при перемене мест множителей, если такая перемена допустима. Пользуясь этим свойством, получаю:

$$tr\,P = tr\left[X\left(X'X\right)^{-1}X'\right] = tr\left[X'X\left(X'X\right)^{-1}\right] = tr\,I_k = k.$$

Здесь I_k — единичная матрица $k \times k$ (та же размерность, что у $X'X$). Индекс k я добавил, чтобы отличить её от единичной матрицы $n \times n$, которая была раньше.

Вот и получается нужное выражение:

$$E\left(RSS\right) = \sigma_\epsilon^2\left(tr\,I - tr\,P\right) = \sigma_\epsilon^2\left(n-k\right).$$