Универсиада по эконометрике 2024
Решения задач второго тура

Задача 1. (25 баллов)
Размышления о p-значении

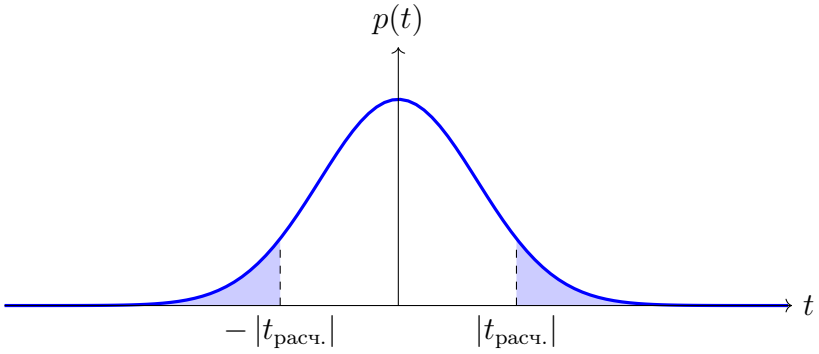
①

Если p-значение теста Стьюдента на значимость для параметра β_2 составило 0.007, то вероятность получить такое (или большее по модулю) значение t-статистики при $\beta_2 = 0$ оказывается меньше 0.7%. Это означает, что на уровне значимости 0.7% (и любом большем) гипотеза $H_0 \ : \ \beta_2 = 0$ должна быть отвергнута. Коэффициент статистически значимо отличается от нуля (2 балла).

②

Для (двустороннего) теста Стьюдента p-значение определяется как

$$p = \mathbb{P}\left(|t| \geq |t_{\text{расч.}}| \ \mid \ H_0 \text{ верна}\right) = 2\,\mathbb{P}\left(t \geq |t_{\text{расч.}}| \ \mid \ H_0 \text{ верна}\right), \qquad t \sim \mathrm{t}_{n-k}$$



Фактически p-значение теста – это минимальный уровень значимости, на котором гипотеза H_0 может быть отвергнута.

Пункт оценивался в 3 балла. Принимались также объяснения, что p-значение – это площадь «хвостов» в распределении t-статистики (см. рисунок) или вероятность отвергнуть H_0 при условии, что она верна.

③

Пункт оценивался в 6 баллов, по 2 балла за каждый подпункт. Точный ответ на этот вопрос можно дать только при условии, что верна нулевая гипотеза[1] (при отсутствии указания на это снимался 1 балл).

- $\mathbb{P}\left(p\text{-value} < 10\% \mid H_0 \text{ верна}, \ \varepsilon \sim \mathcal{N}(0, \sigma^2)\right) = 0.1$ (ясно из вопроса пункта ④)
- $\mathbb{P}\left(p\text{-value} < 10\% \mid H_0 \text{ верна}, \ \varepsilon \not\sim \mathcal{N}(0, \sigma^2)\right)$ не может быть однозначно определена и зависит от распределения ε (например, эта вероятность может существенно отличаться для дискретно распределенных ε)
- $\mathbb{P}\left(p\text{-value} < x \mid H_0 \text{ верна}, \ \varepsilon \sim \mathcal{N}(0, \sigma^2)\right) = \begin{cases} 0, & \text{при} \quad x < 0 \\ x, & \text{при} \quad 0 \le x \le 1 \\ 1, & \text{при} \quad x > 1 \end{cases}$ – снова ясно из формулировки ④ (и вновь, если распределение ε не является нормальным, то о распределении p-значений судить нельзя)

④

Рассуждение для (двустороннего!) теста Стьюдента.

$\mathbb{P}\left(p\text{-value} < x \mid H_0 \text{ верна}, \right) = \begin{cases} 0, & \text{при} \quad x < 0 \\ x, & \text{при} \quad 0 \le x \le 1 \\ 1, & \text{при} \quad x > 1 \end{cases}$ В выкладках использованы монотонность и непрерывность F_0 для распределения $t \sim \mathrm{t}_{n-2}$):

$$\begin{aligned}
\mathbb{P}\left(p\text{-value} < \alpha \mid H_0 \text{ верна}\right) &= \mathbb{P}\left(\, 2\,\mathbb{P}\left(t > |t_{\text{расч.}}|\right) < \alpha\right) = \mathbb{P}\left(\mathbb{P}\left(t > |t_{\text{расч.}}|\right) < \frac{\alpha}{2}\right) = \\
&= \mathbb{P}\left(1 - F_0\left(t_{\text{расч.}}\right) < \frac{\alpha}{2}\right) = \mathbb{P}\left(F_0\left(|t_{\text{расч.}}|\right) > 1 - \frac{\alpha}{2}\right) = \\
&= \mathbb{P}\left(|t_{\text{расч.}}| > F_0^{-1}\left(1 - \frac{\alpha}{2}\right)\right) = 2\,\mathbb{P}\left(t_{\text{расч.}} > F_0^{-1}\left(1 - \frac{\alpha}{2}\right)\right) = \\
&= 2\left(1 - \mathbb{P}\left(t_{\text{расч.}} < F_0^{-1}\left(1 - \frac{\alpha}{2}\right)\right)\right) = 2\left(1 - F_0\left(F_0^{-1}\left(1 - \frac{\alpha}{2}\right)\right)\right) = \\
&= 2\left(1 - \left(1 - \frac{\alpha}{2}\right)\right) = \alpha \qquad \text{(8 баллов)}
\end{aligned}$$

Пункт оценивался в 8 баллов. При рассмотрении одностороннего теста (без двойки, без модуля) баллы снижались. При рассмотрении только случая $\alpha \in [0,\, 1]$ баллы снижались.

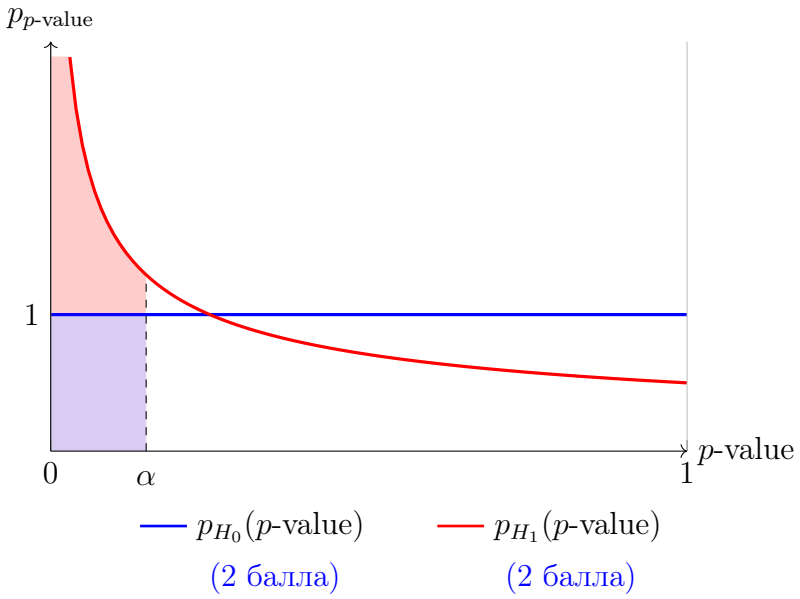
[1] Подробное обоснование здесь не требовалось (за его отсутствие балл не снижался), так как оно необходимо в пункте ④.

(5)–(6)

Сначала посмотрим, что меняется в рассуждении выше при верной альтернативной гипотезе H_1. Обозначим плотность X при верной H_0 как $f_0(x)$, а при верной H_1 – как $f_1(x)$. Из определения p-значения мы имеем $p\text{-value} = 1 - F_0\left(X_{\text{расч.}}\right)$. Тогда плотность p-значения имеет вид[2] (второй переход использует теорему о производной обратной функции)

$$p_{p\text{-value}}(\alpha) = f_1\left(F_0^{-1}(1-\alpha)\right)\left|\frac{d}{d\alpha}F_0^{-1}(1-\alpha)\right| = \frac{f_1\left(F_0^{-1}(1-\alpha)\right)}{f_0\left(F_0^{-1}(1-\alpha)\right)}$$

Видно, что распределение в общем случае уже не будет равномерным. При верной альтернативной гипотезе маленькие p-значения будут встречаться чаще, а большие – реже (то есть плотность p-значения будет убывать). Пример с некоторым заданным уровнем значимости α изображен на рисунке ниже.



(2 балла) (2 балла)

Уровень значимости α при этом является отсечкой по оси абсцисс и отсекает закрашенную площадь (синюю – для случая верной H_0 и красную – для H_1) (2 балла).

[2] Это получается из следующего факта. Пусть X и Y – случайные величины, причем $Y = g(X)$ и X имеет плотность $p_X(x)$. Тогда из соображений теоремы о замене переменной (в общей ситуации – в интеграле Лебега) при некоторых требованиях на функцию $g(x)$ плотность Y имеет вид

$$p_Y(y) = p_X\left(g^{-1}(y)\right)\left|\frac{d}{dy}g^{-1}(y)\right|$$

Дополнительный комментарий, не влияющий на баллы. В контексте примера из задания можно также упомянуть, что для случая теста Стьюдента распределение тестовой статистики при верной гипотезе H_1 имеет так называемое нецентрированное распределение Стьюдента, поэтому при большом желании точную форму плотности p-значения можно воспроизвести с его помощью. Соответствующий пример на сгенерированных данных можно посмотреть здесь.

Также отдельно можно было бы отметить работу [Murdoch, Tsai & Adcock, 2008], в которой идея задачи проиллюстрирована дополнительными симуляциями и несколько развита.

Задача 2. (25 баллов)
Boosted ridge or Ridged boosting?

(a)

При оценивании линейной регрессии с помощью МНК вектор y ортогонально проектируется на линейную оболочку столбцов X, поэтому вектор ошибок $e^{(0)} = y - \hat{y}^{(0)}$ будет лежать в ортогональном дополнении к этой линейной оболочке. Поэтому предсказанный линейной регрессией $\hat{e}^{(0)}$ окажется тождественно нулевым, и все остальные $\hat{y}^{(m)}$ в точности совпадут с МНК-оценкой нулевого шага[3]:

$$\hat{y}^{(m)} = X\left(X^\top X\right)^{-1} X^\top y \qquad \text{(3 балла)}$$

Выходит, никакого смысла в процедуре улучшения не было.

(b)

В векторно-матричном виде задачу Андрея по оцениванию ridge-модели можно записать в виде

$$(y - X\beta)^\top (y - X\beta) + \lambda\beta^\top\beta \to \min_\beta$$
$$\left(y^\top - \beta^\top X^\top\right)(y - X\beta) + \lambda\beta^\top\beta \to \min_\beta$$
$$y^\top y + \beta^\top X^\top X\beta - 2\beta^\top X^\top y + \lambda\beta^\top\beta \to \min_\beta$$

Оптимизируемая функция очевидно является выпуклой, поэтому для поиска оптимального β достаточно найти вектор, удовлетворяющий необходимому условию, которое записывается в матричном виде как

$$\nabla\left(y^\top y + \beta^\top X^\top X\beta - 2\beta^\top X^\top y\beta + \lambda\beta^\top\beta\right) = 0$$
$$2X^\top X\beta - 2X^\top y + 2\lambda\beta = 0 \implies \hat{\beta}_{\text{ridge}} = \left(X^\top X + \lambda I\right)^{-1} X^\top y$$

А тогда прогноз нулевого шага имеет вид

$$\hat{y}^{(0)} = X\hat{\beta}_{\text{ridge}} = X\left(X^\top X + \lambda I\right)^{-1} X^\top y \qquad \text{(4 балла)}$$

[3] То же самое можно получить напрямую, вычислив

$$\hat{e}^{(0)} = X\left(X^\top X\right)^{-1} X^\top e^{(1)} = X\left(X^\top X\right)^{-1} X^\top\left(y - \hat{y}^{(1)}\right) =$$
$$= X\left(X^\top X\right)^{-1} X^\top y - X\left(X^\top X\right)^{-1} \cancel{X^\top X\left(X^\top X\right)^{-1}} X^\top y = 0$$

(c)

Попробуем сначала вычислить $\hat{y}^{(1)}$ и $\hat{y}^{(2)}$, а затем обобщим результат. Вычисляем ошибки нулевого шага и предсказываем их тем же методом:

$$e^{(0)} = y - \hat{y}^{(0)} \implies \hat{e}^{(0)} = \underbrace{X\left(X^\top X + \lambda I\right)^{-1} X^\top}_{B}\left(y - \hat{y}^{(0)}\right)$$

$$\hat{y}^{(1)} = \hat{y}^{(0)} + e^{(0)} = By + B(y - By) = B\left(I + (I - B)\right)y$$

Теперь улучшим этот прогноз еще раз

$$e^{(1)} = y - \hat{y}^{(1)} \implies \hat{e}^{(1)} = B\left(y - \hat{y}^{(1)}\right)$$

$$\begin{aligned}\hat{y}^{(2)} = \hat{y}^{(1)} + \hat{e}^{(1)} &= B\left(I + (I - B)\right)y + B(y - B\left(I + (I - B)\right)y) \\ &= B\left(I + (I - B) + (I - B)^2\right)y\end{aligned}$$

Паттерн понятен и легко доказывается методом математической индукции (что, однако, не требовалось для получения полного балла за пункт). Заключаем, что

$$\hat{y}^{(m)} = B\sum_{k=0}^{m}(I - B)^k y \qquad \text{(3 балла)}$$

Сумма выше – частичная сумма матричной геометрической прогрессии (так называемого ряда Неймана), поэтому ее можно свернуть[4]

$$\hat{y}^{(m)} = B\left(B^{-1}\left(I - (I - B)^{m+1}\right)\right)y = \left(I - (I - B)^{m+1}\right)y$$

Подставляя сюда B, получаем ответ в виде[5]

$$\hat{y}^{(m)} = \left(I - (I - X\left(X^\top X + \lambda I\right)^{-1} X^\top)^{m+1}\right)y \qquad \text{(1 балл)}$$

(d)

Пусть мы считаем, что $y = X\beta + \varepsilon$ и ковариационная матрица ошибок имеет вид $\mathbb{V}\,\varepsilon = \Sigma$. Тогда ковариационная матрица y получается как

$$\mathbb{V}\,y = \mathbb{V}\left(X\beta + \varepsilon\right) = \mathbb{V}\,\varepsilon = \Sigma$$

---

[4] Для таких матриц A, что $I - A$ обратима, верно равенство $\sum_{k=0}^{m} A^k = (I - A)^{-1}(I - A^{m+1})$. При этом ниоткуда не следует, что этот ряд сходится при $m \to \infty$.

[5] То же можно было получить, раскрыв скобки в сумме и собрав результат по биному Ньютона.

Теперь, пользуясь свойством ковариационной матрицы[6], получаем

$$\mathbb{V}\,\hat{y}^{(m)} = \mathbb{V}\left(I-(I-B)^{m+1}\right)y = \left(I-(I-B)^{m+1}\right)\,\Sigma\,\left(I-(I-B)^{m+1}\right)^{\top}$$

Если дополнительно предположить, что $\Sigma = \sigma^2 I$, и заметить, что матрица B и, как следствие, матрица $I-(I-B)^{m+1}$ симметричны, то ответ упростится до вида (для полного балла такое упрощение в решении не требовалось)

$$\mathbb{V}\,\hat{y}^{(m)} = \sigma^2\left(I-(I-B)^{m+1}\right)^2 \qquad \text{(3 балла)}$$

(e)

Прогноз одношаговой ridge-регрессии, как известно, смещен, поэтому естественно ожидать смещения и от комбинации таких моделей. Рассмотрим вектор ошибок итогового прогноза

$$e^{(m)} = \hat{y}^{(m)} - y = -(I-B)^{m+1}y = -(I-B)^{m+1}(X\beta+\varepsilon)$$

Тогда ожидаемое смещение[7] прогноза имеет вид (в общем случае оно ненулевое)

$$\mathbb{E}\,e^{(m)} = -\mathbb{E}\left((I-B)^{m+1}(X\beta+\varepsilon)\right) = -(I-B)^{m+1}X\beta \qquad \text{(3 балла)}$$

(f)

Рассмотрим сингулярное разложение для матрицы X, то есть представим ее в виде $X = UDV^{\top}$, где U – ортогональная[8] матрица $n\times n$, V – ортогональная матрица $k\times k$, а D – прямоугольная диагональная матрица $n\times k$, на главной диагонали которой стоят корни из собственных чисел матрицы $X^{\top}X$, а за пределами главной диагонали – нули (если различных собственных чисел окажется меньше k, то следующие числа на диагонали также будут нулевыми). Тогда получаем

$$\begin{aligned}
B &= X\left(X^{\top}X+\lambda I\right)^{-1}X^{\top} = UDV^{\top}\left(VD^{\top}U^{\top}UDV^{\top}+\lambda I\right)^{-1}VD^{\top}U^{\top} = \\
&= UDV^{\top}\left(VD^{\top}DV^{\top}+\lambda VV^{\top}\right)^{-1}VD^{\top}U^{\top} = UDV^{\top}V\left(D^{\top}D+\lambda I\right)^{-1}V^{\top}VD^{\top}U^{\top} = \\
&= U\underbrace{D\left(D^{\top}D+\lambda I\right)^{-1}D^{\top}}_{\tilde{D}}U^{\top} = U\tilde{D}U^{\top}
\end{aligned}$$

[6] А именно тем, что $\mathbb{V}(A\boldsymbol{\xi}) = A\,\mathbb{V}(\boldsymbol{\xi})\,A^{\top}$.

[7] В некоторых источниках смещение может определяться как $\mathbb{E}\,(y-\hat{y})$, а не как $\mathbb{E}\,(\hat{y}-y)$, однако, поскольку сути это не меняет, оба варианта оценивались полным баллом.

[8] Ортогональными называются (вещественные) матрицы U, такие что $U^{\top}U = UU^{\top} = I$.

В записи выше матрица $\tilde{D}$ также диагональная, имеет размер $n \times n$, причем на ее диагонали стоят числа $\frac{d_i^2}{d_i^2+\lambda}$ вплоть до $i = k$, а затем – нули. Отсюда получаем

$$I - B = UU^\top - U\tilde{D}U^\top = U(I - \tilde{D})U^\top \implies (I - B)^{m+1} = U(I - \tilde{D})^{m+1}U^\top$$

Теперь рассмотрим вычисленное нами ранее смещение

$$-\mathbb{E}\, e^{(m)} = (I - B)^{m+1}X\beta = U(I - \tilde{D})^{m+1}U^\top UDV^\top\beta = U(I - \tilde{D})^{m+1}DV^\top\beta$$

На диагонали (при $i \leq k$) матрицы $(I-\tilde{D})^{m+1}$ стоят числа $\left(1 - \frac{d_i^2}{d_i^2+\lambda}\right)^{m+1}$ (которые бесконечно малы при $m \to \infty$, поскольку $1 - \frac{d_i^2}{d_i^2+\lambda} = \frac{\lambda}{\lambda+d_i^2} \in [0,\, 1)$ при всех $\lambda > 0$) и единицы при $i > k$. У матрицы D – наоборот – в этих строках нули.

$$(I - \tilde{D})^{m+1}D = \begin{pmatrix} \left(\frac{\lambda}{\lambda+d_1^2}\right)^{m+1} & & 0 & & \\ & \ddots & & 0 & \\ 0 & & \left(\frac{\lambda}{\lambda+d_k^2}\right)^{m+1} & & \\ \hdashline & 0 & & 1 & 0 \\ & & & & \ddots \\ & & & 0 & 1 \end{pmatrix} \begin{pmatrix} d_1 & & 0 & & \\ & \ddots & & 0 & \\ 0 & & d_k & & \\ \hdashline & 0 & & 0 & 0 \\ & & & & \ddots \\ & & & 0 & 0 \end{pmatrix}$$

Из этого представления ясно, что

$$(I - \tilde{D})^{m+1}D \xrightarrow{m\to\infty} 0 \implies \mathbb{E}\, e^{(m)} \xrightarrow{m\to\infty} 0 \qquad \text{(4 балла)}$$

(g)

При любом $\lambda > 0$ прогноз $\hat{y}^{(m)}$ линеен по y и притом его смещение при увеличении m стремится к нулю. В этот момент может возникнуть мысль о том, что прогноз $\hat{y}^{(m)}$ должен с ростом m стремиться к МНК-оценке $\hat{y}_{\text{OLS}}$. Попробуем это обосновать. Вычисленная ранее ковариационная матрица после замены через сингулярные числа примет вид

$$\begin{aligned} \mathbb{V}\,\hat{y}^{(m)} &= \sigma^2\left(I - (I - B)^{m+1}\right)^2 = \sigma^2\left(I - U(I - \tilde{D})^{m+1}U^\top\right)^2 = \\ &= \sigma^2\left(UU^\top - U(I - \tilde{D}\Big)^{m+1}U^\top)^2 = \sigma^2\, U\Big(I - (I - \tilde{D})^{m+1}\Big)^2U^\top \end{aligned}$$

Теперь выразим через U, D и V ковариационную матрицу МНК-оценок обычной линейной регрессии (предполагается, что матрица X имеет полный ранг, иначе МНК-оценка не будет существовать):

$$\begin{aligned} \mathbb{V}\,\hat{y}_{\text{OLS}} &= \mathbb{V}\,X\hat{\beta}_{\text{OLS}} = \mathbb{V}\,X(X^\top X)^{-1}X^\top(X\beta + \varepsilon) = \sigma^2 X(X^\top X)^{-1}X^\top = \\ &= \sigma^2 UDV^\top\left(VD^\top U^\top UDV^\top\right)^{-1}VD^\top U^\top = \sigma^2 UD\left(D^\top D\right)^{-1}D^\top U^\top \end{aligned}$$

Синяя матрица из записи выше состоит из k единиц, идущих по главной диагонали, а остальные ее элементы равны нулю. Что (не)удивительно, при $m \to \infty$ ровно такой же вид приобретает и синяя матрица из выражения для $\mathbb{V}\,\hat{y}^{(m)}$, что следует из последних рассуждений пункта f.

Получается, смещение прогноза $\hat{y}^{(m)}$ стремится к нулю, при этом он имеет линейный вид, а его ковариационная матрица сходится к матрице $\mathbb{V}\,\hat{y}_{\text{OLS}}$ при любом $\lambda > 0$. Тогда из соображений того, что МНК дает наиболее эффективную среди несмещенных оценок по теореме Гаусса–Маркова, в предпосылках этой теоремы имеет место сходимость

$$\hat{y}^{(m)} \xrightarrow{m\to\infty} \hat{y}_{\text{OLS}} = X\left(X^\top X\right)^{-1} X^\top y \qquad \text{(4 балла)}$$

В задаче не предполагалось настолько формальное обоснование. Для получения полного балла достаточно было сказать, что прогноз сходится к МНК-оценке из соображений асимптотической несмещенности, постепенного возрастания дисперсии и линейности.

Дополнительный комментарий. Задача, решение которой только что было представлено, фактически описывает процедуру градиентного бустинга над линейными ridge-регрессиями. Эта тема уже освещалась в литература, см., например, работы [Bühlmann & Yu, 2003] и [Tutz & Binder, 2007]. Алгоритм градиентного бустинга (хотя он и показал себя не с лучшей стороны в нашей задаче) является чрезвычайно мощным методом ансамблирования моделей машинного обучения, поэтому кажется достаточно интересным для рассмотрения.

Если участники не были знакомы с сингулярным разложением ранее, но заинтересовались им, автор может порекомендовать визуализированное объяснение и более подробное и формальное обоснование из открытого доступа.

Задача 3. (25 баллов)
Стать мракоборцем

---

(1)

Поскольку по условию переменные можно считать нормально распределенными, то пусть условные плотности $X(1,2)$ имеют следующий вид

$$p_1(X) = p(X \mid y=1) = \frac{1}{2\pi\sqrt{|\Sigma|}}\, e^{-\frac{1}{2}(X-\mu_1)^\top \Sigma^{-1}(X-\mu_1)}$$
$$p_2(X) = p(X \mid y=2) = \frac{1}{2\pi\sqrt{|\Sigma|}}\, e^{-\frac{1}{2}(X-\mu_2)^\top \Sigma^{-1}(X-\mu_2)}$$

Теперь преобразуем $\mathbb{P}(y=1 \mid X)$ – апостериорную вероятность обучения на мракоборца при условии на X – по формуле Байеса

$$\mathbb{P}(y=1 \mid X) = \frac{p(X \mid y=1)\overbrace{\mathbb{P}(y=1)}^{\pi}}{p(X)} = \frac{\pi\, p_1(X)}{\pi\, p_1(X) + (1-\pi)p_2(X)} = \frac{\frac{\pi}{1-\pi}\frac{p_1(X)}{p_2(X)}}{1+\frac{\pi}{1-\pi}\frac{p_1(X)}{p_2(X)}}$$

$$\frac{\pi}{1-\pi}\frac{p_1(X)}{p_2(X)} = e^{\ln\left(\frac{\pi}{1-\pi}\right)}\frac{\cancel{2\pi\sqrt{|\Sigma|}}}{\cancel{2\pi\sqrt{|\Sigma|}}}\, e^{-\frac{1}{2}\left((X-\mu_1)^\top\Sigma^{-1}(X-\mu_1)-(X-\mu_2)^\top\Sigma^{-1}(X-\mu_2)\right)} = e^{D(X)}$$

Попробуем вычислить это в нашем конкретном численном примере

$$X = \begin{pmatrix}60\\4.2\end{pmatrix},\quad \mu_1 = \begin{pmatrix}50\\4\end{pmatrix},\quad \mu_2 = \begin{pmatrix}40\\3\end{pmatrix},\quad \Sigma = \begin{pmatrix}100 & 3\\3 & 1\end{pmatrix} \implies \Sigma^{-1} = \frac{1}{91}\begin{pmatrix}1 & -3\\-3 & 100\end{pmatrix}$$

$$(X-\mu_1)^\top\Sigma^{-1}(X-\mu_1) = \frac{1}{91}(60-50)^2 - 2\times\frac{3}{91}(60-50)(4.2-4) + \frac{100}{91}(4.2-4)^2 = \frac{92}{91}$$
$$(X-\mu_2)^\top\Sigma^{-1}(X-\mu_2) = \frac{1}{91}(60-40)^2 - 2\times\frac{3}{91}(60-40)(4.2-3) + \frac{100}{91}(4.2-3)^2 = \frac{400}{91}$$

$$\implies D(X) = \ln\left(\frac{0.3}{1-0.3}\right) - \frac{1}{2}\left(\frac{92}{91}-\frac{400}{91}\right) = \ln\frac{3}{7} + \frac{22}{13} \approx 0.845 \implies e^{D(X)} \approx 2.328$$

Теперь, подставляя это в формулу условной вероятности, получаем

$$\mathbb{P}(y=1 \mid X) = \frac{e^{D(X)}}{1+e^{D(X)}} = \frac{2.328}{1+2.328} \approx 0.7$$

- Если участники давали (и разумно обосновывали) ответ только на основании средних значений, то за это ставилось 5 баллов
- Если участники развивали мысль о формуле Байеса или условной вероятности, то за это ставилось 5 баллов
- Если участники формализовали свои рассуждения в виде статистической модели, пробовали строить МНК-модель и получали ответ из $[0, 1]$, но никак не использовали ковариационные матрицы, то за это ставились 10 баллов
- Если студенты пробовали оценивать логит-модель, формируя и максимизируя соответствующее правдоподобие, то за это ставилось 15 баллов

②

Вернемся к виду условной вероятности, который мы получили в пункте ①

$$\mathbb{P}(y = 1 \mid X) = \frac{e^{D(X)}}{1 + e^{D(X)}} = \frac{1}{1 + e^{-D(X)}}$$

Видно, что это логистически преобразованное выражение $D(X)$. Покажем, что перед нами на самом деле выражение для вероятности из логистиечской регрессии. Еще раз рассмотрим

$$D(X) = \ln\left(\frac{\pi}{1-\pi}\right) - \frac{1}{2}\left((X-\mu_1)^\top \Sigma^{-1}(X-\mu_1) - (X-\mu_2)^\top \Sigma^{-1}(X-\mu_2)\right)$$

Поскольку ковариационные матрицы Σ для обоих классов используются одни и те же, то все элементы с квадратами и попарными произведениями взаимно уничтожаются, поэтому функция $D(X)$ оказывается **линейной** по X с точностью до константы, то есть имеет вид $D(X) = \beta_0 + X\beta$. А это означает, что перед нами в самом деле логит-модель (5 баллов)

$$\mathbb{P}(y = 1 \mid X) = \frac{e^{\beta_0 + X\beta}}{1 + e^{\beta_0 + X\beta}} = \frac{1}{1 + e^{-(\beta_0 + X\beta)}}$$

Задача 4. (25 баллов)
О взвешивании

Ⓐ

Как известно, МНК-оценка параметров линейной регрессии имеет вид $\hat{\beta}_{\text{OLS}} = \left(X^\top X\right)^{-1} X^\top y$. В нашем случае матрицы X и y имеют блочный вид, поэтому

$$\hat{\beta}_{\text{OLS}} = \left[\left(\begin{array}{c|c} X_1^\top & X_2^\top \end{array}\right)\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}\right]^{-1}\left(\begin{array}{c|c} X_1^\top & X_2^\top \end{array}\right)\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \left(X_1^\top X_1 + X_2^\top X_2\right)^{-1}\left(X_1^\top y_1 + X_2^\top y_2\right) =$$

$$= \left(X_1^\top X_1 + X_2^\top X_2\right)^{-1}\Big(\underbrace{X_1^\top X_1 \left(X_1^\top X_1\right)^{-1}}_{I} X_1^\top y_1 + \underbrace{X_2^\top X_2 \left(X_2^\top X_2\right)^{-1}}_{I} X_2^\top y_2\Big) =$$

$$= \left(X_1^\top X_1 + X_2^\top X_2\right)^{-1}\left(X_1^\top X_1\,\hat{\beta}_1 + X_2^\top X_2\,\hat{\beta}_2\right) = \underbrace{\left(X_1^\top X_1 + X_2^\top X_2\right)^{-1} X_1^\top X_1}_{A_1}\,\hat{\beta}_1 +$$

$$+ \underbrace{\left(X_1^\top X_1 + X_2^\top X_2\right)^{-1} X_2^\top X_2}_{A_2}\,\hat{\beta}_2$$

Осталось проверить, что $A_1 + A_2 = I$. Убедимся в этом непосредственно

$$A_1 + A_2 = \left(X_1^\top X_1 + X_2^\top X_2\right)^{-1} X_1^\top X_1 + \left(X_1^\top X_1 + X_2^\top X_2\right)^{-1} X_2^\top X_2 =$$

$$= \cancel{\left(X_1^\top X_1 + X_2^\top X_2\right)^{-1}}\cancel{\left(X_1^\top X_1 + X_2^\top X_2\right)} = I \qquad \text{(15 баллов)}$$

Ⓑ

Теперь исследуем, как себя ведет GLS-оценка модели. Нам известно, что истинная ковариационная матрица ошибок имеет блочный вид

$$\Omega = \left(\begin{array}{ccc|ccc} \sigma_1^2 & & 0 & & & \\ & \ddots & & & 0 & \\ 0 & & \sigma_1^2 & & & \\ \hline & & & \sigma_2^2 & & 0 \\ & 0 & & & \ddots & \\ & & & 0 & & \sigma_2^2 \end{array}\right) = \begin{pmatrix} \sigma_1^2 I & 0 \\ 0 & \sigma_2^2 I \end{pmatrix} \implies \Omega^{-1} = \begin{pmatrix} \frac{1}{\sigma_1^2} I & 0 \\ 0 & \frac{1}{\sigma_2^2} I \end{pmatrix}$$

Получим аналитический вид GLS-оценки (для получения полного балла за задачу этого не требовалось). Обобщенный метод наименьших квадратов решает задачу

$$e^\top \Omega^{-1} e = (y - X\beta)^\top \Omega^{-1} (y - X\beta) \to \min_\beta$$

Вновь пользуясь векторно-матричным дифференцированием, выписываем необходимое условие экстремума в этой задаче[9]

$$\nabla (y - X\beta)^\top \Omega^{-1} (y - X\beta) = 0 \iff \nabla \left(y^\top \Omega^{-1} y + \beta^\top X^\top \Omega^{-1} X\beta - 2\beta^\top X^\top \Omega^{-1} y \right) = 0$$
$$2X^\top \Omega^{-1} X\beta - 2X^\top \Omega^{-1} y = 0 \implies \hat{\beta}_{\text{GLS}} = \left(X^\top \Omega^{-1} X \right)^{-1} X^\top \Omega^{-1} y$$

Теперь еще раз воспользуемся блочной структурой матриц, данных в условии

$$\hat{\beta}_{\text{GLS}} = \left[\left(\begin{array}{c|c} X_1^\top & X_2^\top \end{array} \right) \begin{pmatrix} \frac{1}{\sigma_1^2} I & 0 \\ 0 & \frac{1}{\sigma_2^2} I \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \right]^{-1} \left(\begin{array}{c|c} X_1^\top & X_2^\top \end{array} \right) \begin{pmatrix} \frac{1}{\sigma_1^2} I & 0 \\ 0 & \frac{1}{\sigma_2^2} I \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} =$$
$$= \left(\frac{X_1^\top X_1}{\sigma_1^2} + \frac{X_2^\top X_2}{\sigma_2^2} \right)^{-1} \left(\frac{X_1^\top y_1}{\sigma_1^2} + \frac{X_2^\top y_2}{\sigma_2^2} \right) \qquad \text{(10 баллов)}$$

В случае, если участник записывал обратную матрицу в знаменателе дроби или обращался с матрицами и векторами, как с числами, то за соответствующий пункт выставлялось не более 5 баллов.

[9] И вновь оптимизируемая функция выпукла, поэтому необходимое условие оказывается одновременно и достаточным.