Линейные модели — Сравнение моделей, ANCOVA Вариант № 01

Ф.И.О.: _____

1. (a) ☐ (b) ☐ (c) ☐ (d) ☐
2. ☐☐☐☐☐☐ . ☐☐☐
3. (a) ☐ (b) ☐ (c) ☐ (d) ☐
4. (a) ☐ (b) ☐ (c) ☐ (d) ☐
5. (a) ☐ (b) ☐ (c) ☐ (d) ☐
6. (a) ☐ (b) ☐ (c) ☐ (d) ☐
7. ☐☐☐☐☐☐ . ☐☐☐
8. ☐☐☐☐☐☐ . ☐☐☐
9. (a) ☐ (b) ☐ (c) ☐ (d) ☐
10. (a) ☐ (b) ☐ (c) ☐ (d) ☐
11. ☐☐☐☐☐☐ . ☐☐☐
12. (a) ☐ (b) ☐ (c) ☐ (d) ☐

1. Что такое доверительная вероятность (p-value)? Отметьте справедливые утверждения.

   (a) Высокое значение p показывает, что значение тестовой статистики вполне обычно, если H_0 была бы справедлива

   (b) Низкое значение p показывает, что значение тестовой статистики экстремально при условии справедливости H_O

   (c) p рассчитывается при условии того, что H_0 верна

   (d) $(1 - p)$ — это вероятность того, что удастся воспроизвести результат этого эксперимента

2. Выберете из предложенных операций те, которые, в общем виде, относятся к методам кросс-валидации регрессионной модели, и расставьте их в правильном порядке:

   1. Подбираем модель на тренировочном наборе данных
   2. Определяем точность предсказаний модели по тому, насколько предсказанные значения зависимой переменной отличаются от реальных
   3. Выполняем предсказание данных на тестовом наборе. При этом мы берем имеющиеся значения предиктора и предсказываем, с помощью модели, теоретические значения зависимой переменной
   4. Определяем точность предсказаний модели по тому, насколько теоретические значения предиктора отличаются от эмпирических и рассчитываем RMSE, аналогичным образом поступаем с зависимой переменной
   5. Сравниваем модели и выбираем ту, у которой расхождение предсказанных и эмпирических значения в тестовом наборе минимально
   6. Создаем тренировочный и тестовый наборы данных, случайным образом распределив исходные измерения на две группы без повторов
   7. Создаем тренировочный и тестовый наборы данных. При этом весь исходный набор становится тестовым, а тренировочный мы создаем случайным образом скопировав некоторые измерения из исходного набора
   8. Выполняем предсказание данных на тренировочном наборе. При этом мы предсказываем значения как предиктора, так и зависимой пременной с помощью модели
   9. Сравниваем модели и выбираем ту, у которой расхождение предсказанных и эмпирических значений в тренировочном наборе равно нулю

3. Отметьте верные утверждения, которые характеризуют переобученную модель (overfitted model)

   (a) В модель включено очень много предикторов

   (b) У переобученной модели *RMSE* на обучающей выборке будет выше, чем у хорошо (вмеру) обученной модели

   (c) В результатах анализа появляется *RMSE*

   (d) Сравнительно высокие значения ошибок, но высокая точность предсказаний на обучающей выборке

4. Перед вами графики, построенные по четырем моделям, описывающим одни и те же данные.

   Определите, на каком из графиков (рис. 1) приведена лучшая модель?

   (a) A

   (b) B

   (c) C

   (d) D

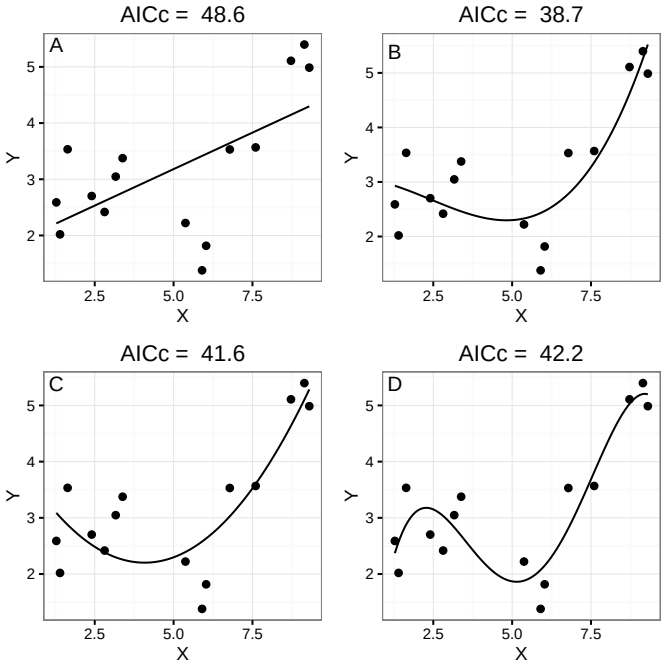

Рис. 1: Найдите на графиках лучшую модель.

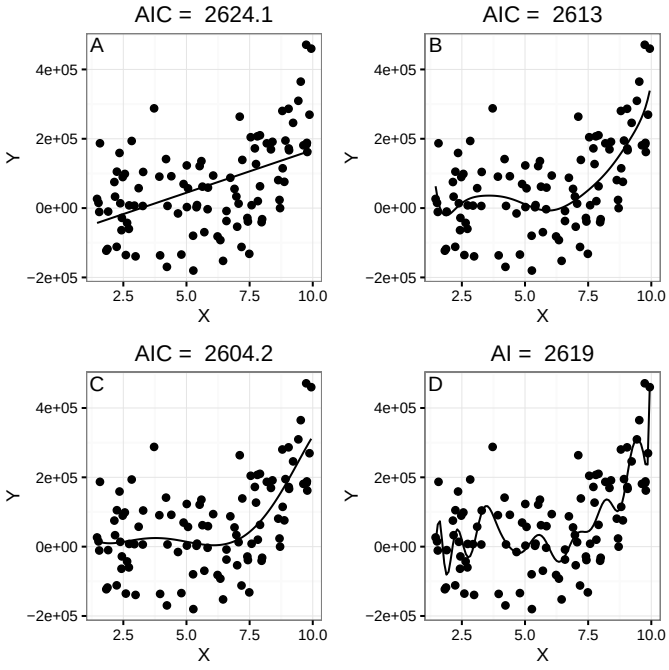

Рис. 2: Найдите на графиках самую переобученную модель.

5. Перед вами графики, построенные по четырем моделям, описывающим одни и те же данные. Определите, на каком из графиков (рис. 2) приведена самая переобученная модель?

(a) A
(b) B
(c) C
(d) D

6. Перед Вами несколько моделей

   - M1: Y X + Z + K + G
   - M2: Y X * Z * K * G
   - M3: Y X + Z + K + G + 1
   - M4: Y X + Z + K + G - 1
   - M5: Y X + Z + K + G + X:Z
   - M6: Y X + Z + K + G + X:K
   - M7: Y X + Z + K
   - M8: Y X * Z * K
   - M9: Y 1

   Найдите правильные утверждения

   (a) M5 вложена в M6
   (b) M9 вложена в M1
   (c) M6 вложена в M5
   (d) M1 вложена в M3

7. Перед вами модель:
   glm(formula = Y ~X + K + G + 1, data = table)
   Применение функции ‘logLik(model1)‘ дает следующие результаты.

   > logLik(model1)[1]

   [1] -429.14

   Вычислите значение AIC для данной модели, округлите до тысячных

8. Перед вами модель:
   glm(formula = Y ~X + K + G + 1, data = table)
   В этой модели переменная G - дискретный предиктор, имеющий 2 градаций. Остальные предикторы — непрерывные переменные. Применение функции ‘logLik(model1)‘ дает следующие результаты.

   > logLik(model1) [1]

   [1] -172.2649

   Вычислите значение AIC для данной модели, округлите до тысячных

9. Перед вами модель, в которой X - непрерывный предиктор, а Z - дискретный фактор

   - M1: Y ~X * Z

   Найдите правильные утверждения

(a) По результатам summary(M1) мы можем судить о значимости взаимодействий предикторов

(b) В этой модели будет выявлено взаимодействие предикторов

(c) Наклон кривой непрерывного предиктора, будет регулироваться значением дискретного фактора Z

(d) В summary(M1) уровень значимости при X:Z будет меньше 0.05

10. Датафрейм WBC содержит информацию о выбросах CO2 на душу населения в разных странах.

Предикторы: POP - непрерывный, численность населения, income - дискретный, ВВП на душу населения (градации rich и poor), demographic - дискретный, место проживания большинства населения (градации rural и urban).

Сравните три модели:

Модель 1

lm(formula = CO2 ~ POP + demographic + income, data = WBC)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	205.915	105.021	1.961	0.054
POP	409.755	69.364	5.907	0.000
demographicUrban	-109.293	137.817	-0.793	0.430
incomeRich	539.679	159.220	3.390	0.001

R^2 Adjusted R^2

0.4 0.375

Модель 2

lm(formula = CO2 ~ POP + income, data = WBC)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	150.036	77.675	1.932	0.057
POP	418.778	68.247	6.136	0.000
incomeRich	535.703	158.730	3.375	0.001

R^2 Adjusted R^2

0.395 0.378

Модель 3

lm(formula = CO2 ~ POP + demographic, data = WBC)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	327.789	105.617	3.104	0.003
POP	398.408	74.158	5.372	0.000
demographicUrban	-94.584	147.442	-0.642	0.523

R^2 Adjusted R^2

0.303 0.283

Определите, правомерно или нет исключение факторов, и почему. Выберете правильные утверждения.

(a) можно исключать предиктор POP, он не значим

(b) нельзя исключать предиктор income, без него модель хуже объясняет изменчивость

(c) нельзя исключать предиктор POP, он значим

(d) нельзя исключать предиктор demographic, без него модель хуже объясняет изменчивость

11. Австралийские фермеры изучали число семян в помидорах разных сортов. Всего было три сорта, или три градации дискретного предиктора: желтый, красный и черный. Были построены две модели, связывающие зависимую переменную с предиктором, с базовыми уровнями Yellow и Red:

	Names1	Base_Yellow
1	Intercept	22
2	red	1
3	black	7

	Names2	Base_Red
1	Intercept	23
2	yellow	-1
3	black	6

Рассчитайте (предскажите) значение интерсепта для модели, в которой в качестве базового уровня выбраны черные помидоры.

12. Перед вами результаты анализа ковариаций. Независимая переменная X дискретна (с тремя уровнями Large, Medium, Small), переменная Z непрерывна.

```
Call:
lm(formula = Y ~ X * Z, data = table)

Residuals:
    Min      1Q  Median      3Q     Max
-8.9880 -3.9469  0.3533  2.9633 14.8286

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  20.83425    1.90645  10.928 2.69e-15 ***
XMedium      -4.92996    4.06682  -1.212    0.231
XSmall      -17.69522    3.46359  -5.109 4.35e-06 ***
Z             7.58304    0.05401 140.388  < 2e-16 ***
XMedium:Z    -2.11463    0.10836 -19.514  < 2e-16 ***
XSmall:Z     -3.99427    0.08215 -48.619  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.066 on 54 degrees of freedom
Multiple R-squared:  0.9982,        Adjusted R-squared:  0.9981
F-statistic:  6107 on 5 and 54 DF,  p-value: < 2.2e-16
```

Из предложенных уравнений выберите то, которое соответствует уровню дискретного предиктора Large

(a) Y = (20.834 - 4.93) - 2.115 * Z

(b) Y = 20.834 - 17.695 - 3.994 * Z

(c) Y = - 4.93 - 2.115 * Z

(d) Y = 20.834 + 7.583 * Z