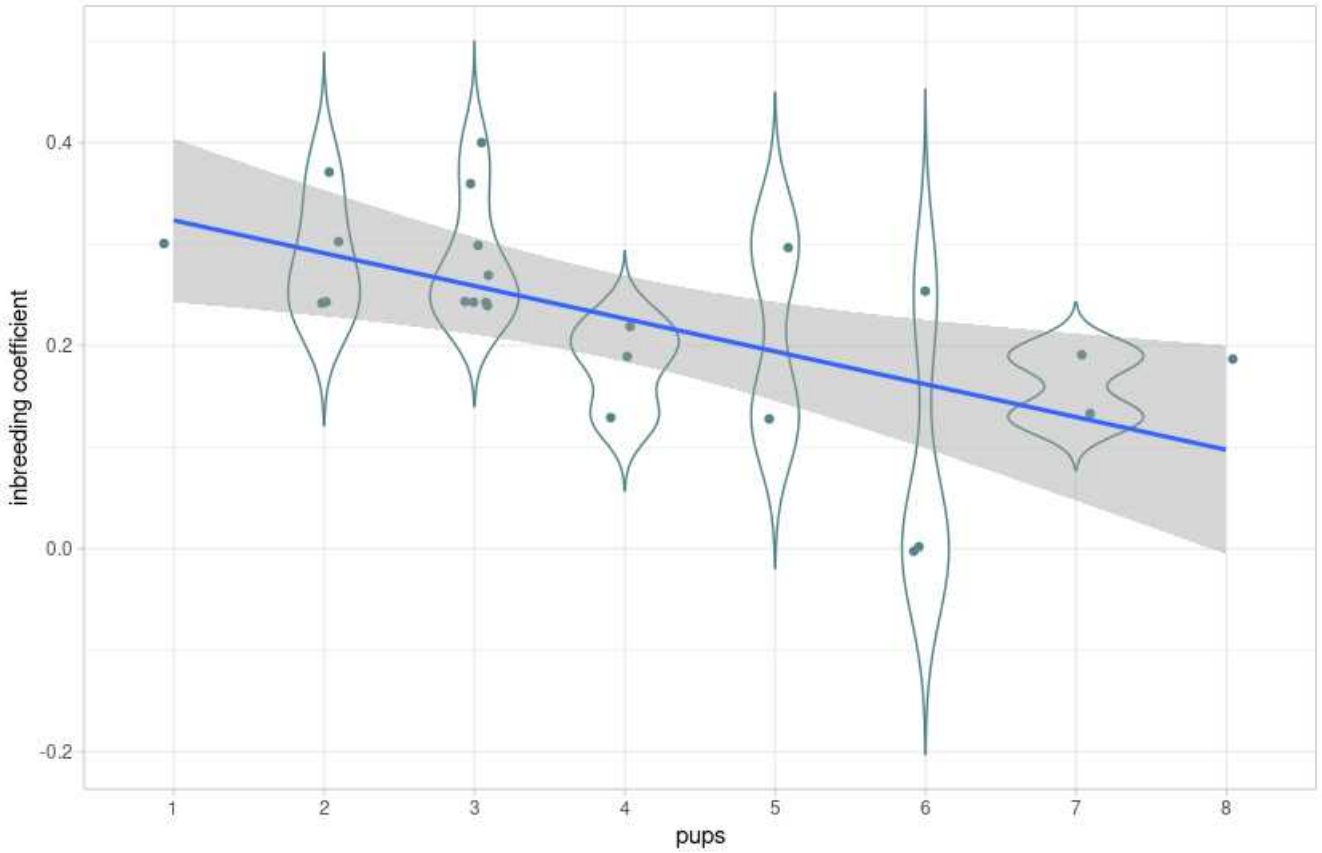
Github с кодом и графиками на R:
https://github.com/dsmutin/statistics_hw

1.



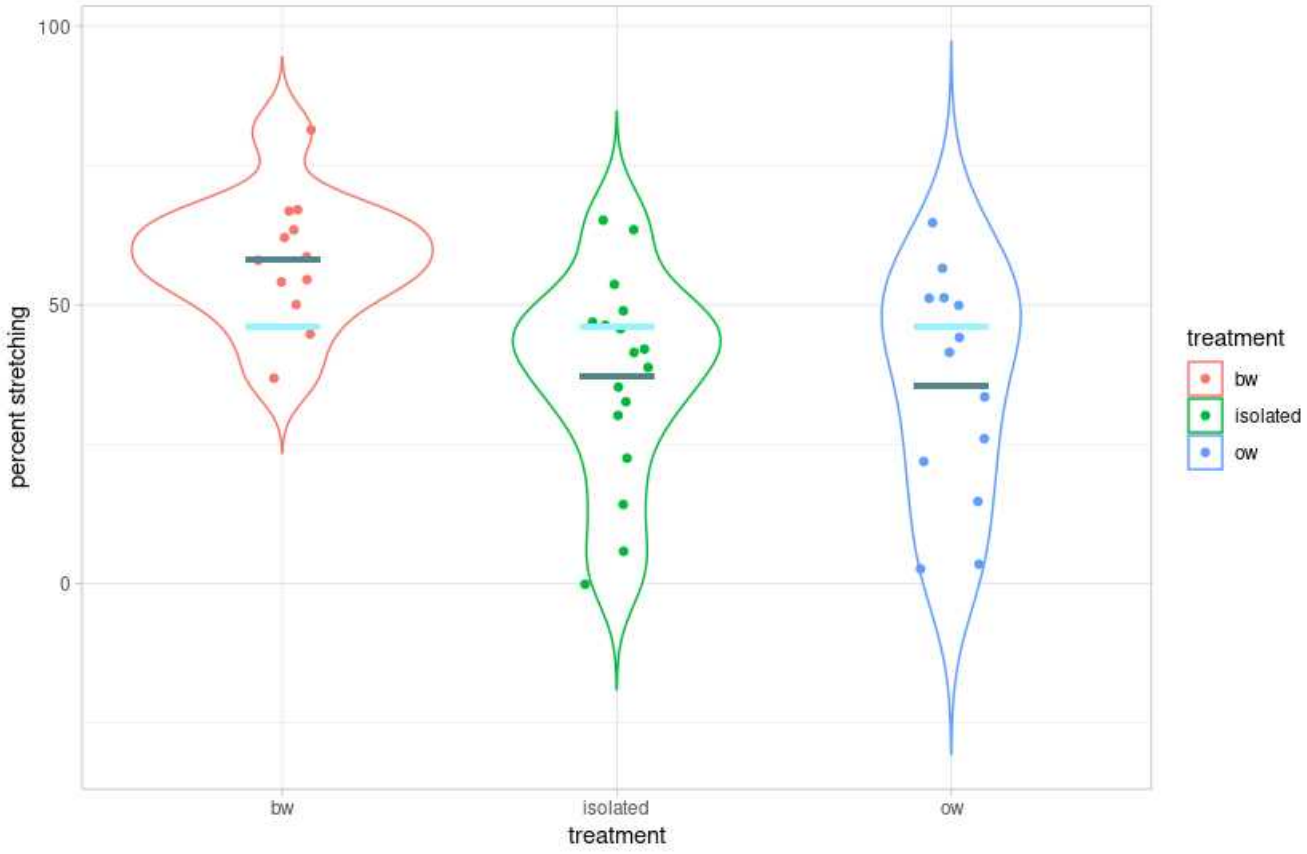
По графику можно предположить отрицательную корреляцию, но видно, что данных во всех категориях мало.
Графики плотности распределений по группам показывают неоднородность выборки, что связано с небольшим ее размером
Графики плотности распределений указывают на нормальность выборки

Посчитаем коэффициент корреляции (т. к. данные по количеству щенков дискретные и доступны в небольшом диапазоне, возьмем Спирмана):
cor = -0.6562775 , p = 0.0004966 .

Между коэффициентом инбридинга и количеством потомков есть достоверная отрицательная корреляция

2. Формула lm для альтернативной гипотезы:
percent.stretching = 58.050 - 20.856 (treatmentisolated) - 22.681 (treatmentow)

Для нулевой:
percent.stretching = mean(percent.stretching) = 42.59



Действительно, по glm bw > ow

Сравним модели ANOVA:

NULL	41	15848	
treatment	39	11807	deviance: 4040.9

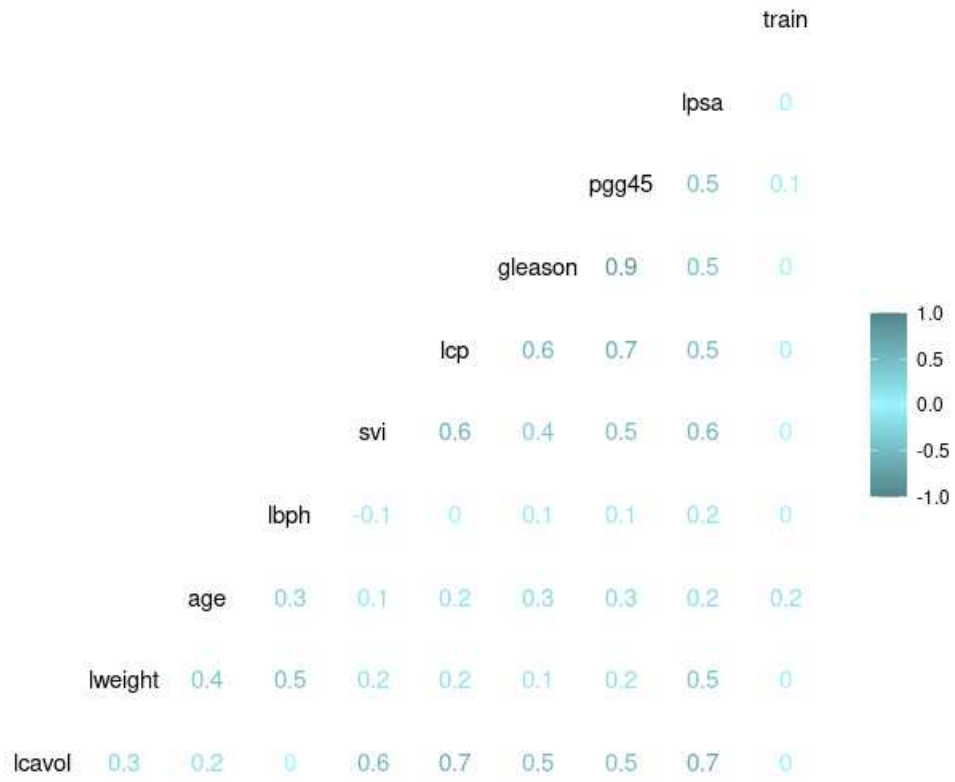
Различия достоверны

AIC(m1) = 364.02
AIC(m0) = 372.38

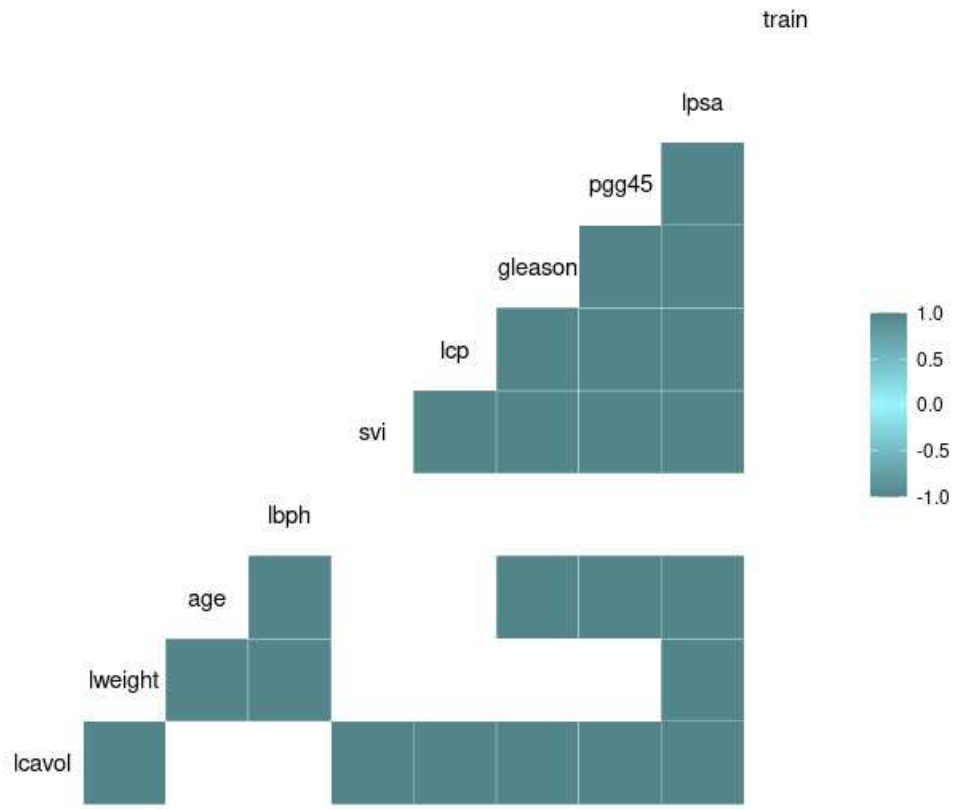
выбранная модель лучше, чем нулевая

3.
Посмотрим на данные
5, 7, 10 столбцы — категорийные
lcavol, lweight, lpsa обладают нормальным распределением

Чтобы не делать Z-приведения, воспользуемся коэффициентом корреляции Спирмана



Сверим p-значения. $p<0.05$:



Параметры, достоверно коррелирующие с lcavol:

- lweight
- svi
- lcp
- gleason/pgg45
- lpsa

Построим вложенные модели согласно значениям корреляции и сравним их ANOVA:

Model 1: lcavol ~ lcavol
Model 2: lcavol ~ lcp
Model 3: lcavol ~ lpsa
Model 4: lcavol ~ lcp * lpsa
Model 5: lcavol ~ lcp * lpsa * svi
Model 6: lcavol ~ lcp * lpsa * svi * gleason
Model 7: lcavol ~ lcp * lpsa * svi * gleason * lweight

	Resid. Df	Resid. Dev	Df	Deviance
1	96	133.36		
2	95	72.54	1	60.82
3	95	61.42	0	11.12
4	93	44.66	2	16.76
5	89	43.74	4	0.92
6	83	42.41	6	1.34
7	71	36.96	12	5.45

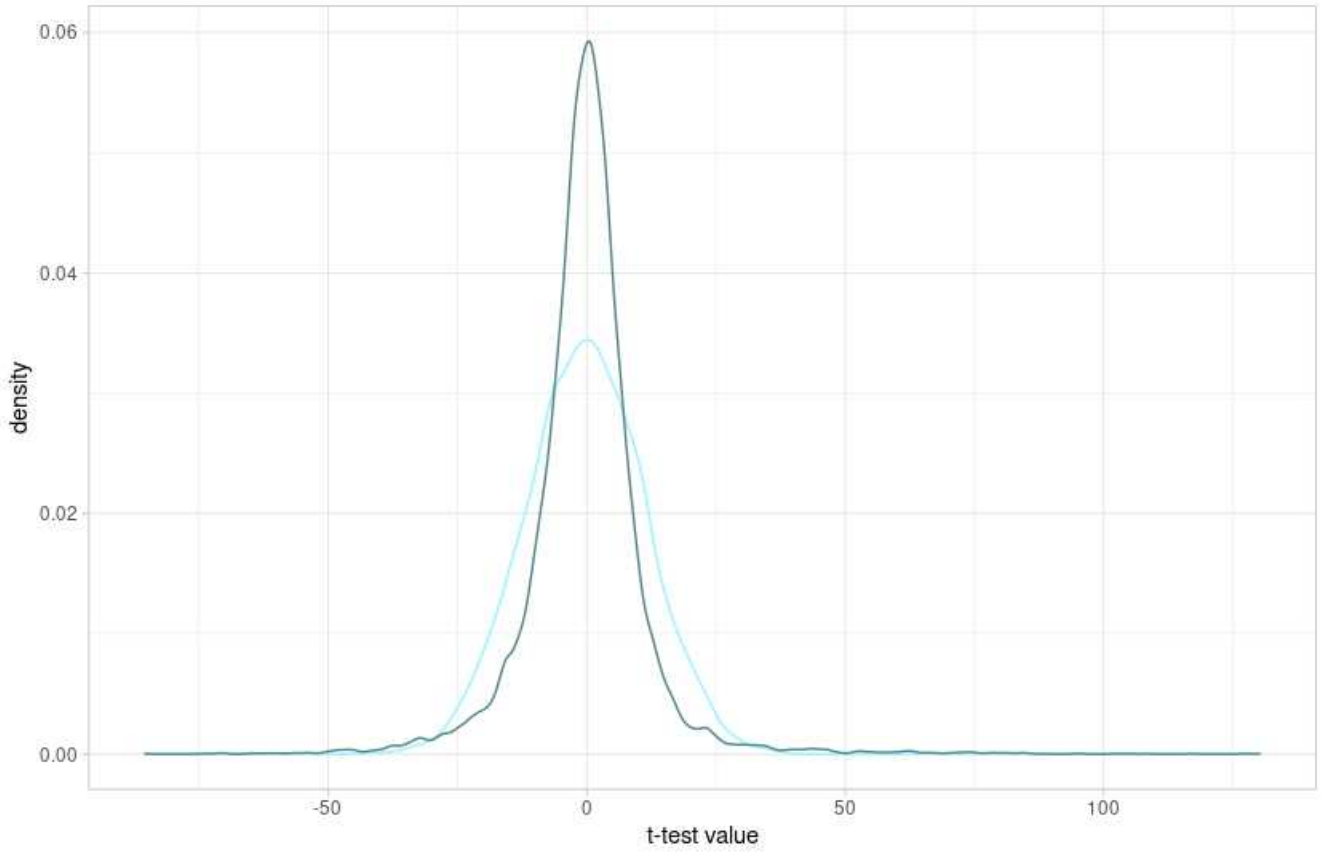
и AIC:

	df	AIC
1	2	310.15
2	3	253.09
3	3	236.95
4	5	210.05
5	9	216.03
6	15	225.01
7	27	235.67

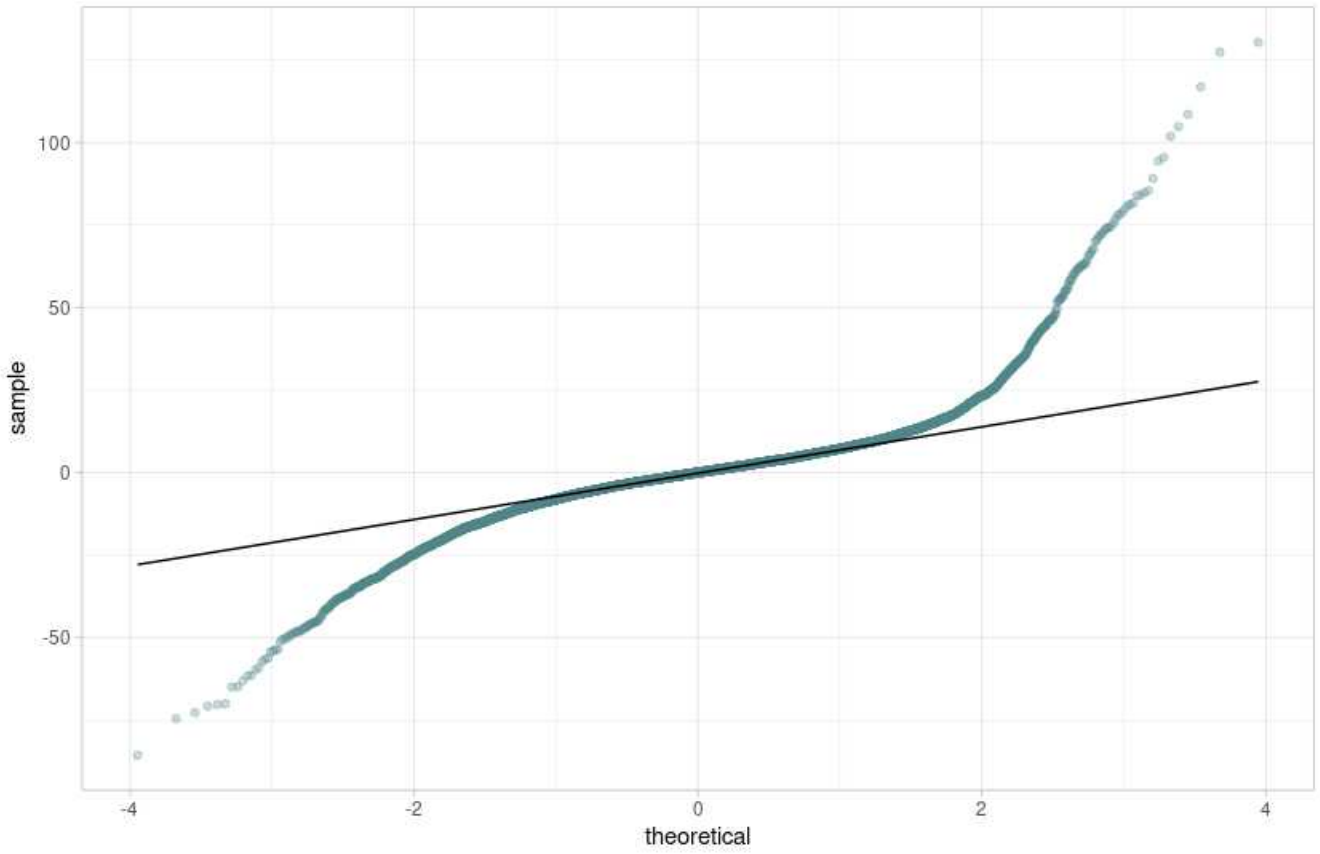
оптимальна модель 4. Ее коэффициенты:

(Intercept)	0.461
lcp	0.809
lpsa	0.42
svi	-1.843
lcp:lpsa	-0.183
lcp:svi	0.627
lpsa:svi	0.478
lcp:lpsa:svi	-0.1

4. Проверим результаты т-теста на сходство с распределением Гаусса:



Построим qqplot:



Видно, что разлеты не кластеризованы, генерального отклонения от нормального распределения нет

Чего мудрить, сравним t-тестом с нормальным распределением:
t = 1.1108, df = 25249, p-value = 0.2667

Различия от распределения Гаусса достоверные

Поправка Бонферрони для такого количества сравнений слишком маломощная (p/12627 — очень маленькое значение). Она дает 2957 генов. Имеет смысл использовать более подходящие подходы, например — поправку Шидака-Холма. Таким образом, нашли 2999 генов, достоверно различающихся между выборками. Многовато, конечно, но в целом — ожидаемая картина. Можно использовать более жесткие критерии или более низкое значение p

Наиболее перспективные гены:
14 15 3654 5401 9102 11552