Results

Pochodzenie Etniczne vs Występowanie Choroba

Descriptive Statistics

	edit_scale						
	African American or Black	American Indian or Alaskan Native	Asian	Caucasian or White	Multiple	Native Hawaiian or Other Pacific Islander	Unknown, Unavailable or Unreported
Valid	2432	14	232	1844	7	41	210
Missing	0	0	0	0	0	0	0
Mean	0.212	0.000	0.228	0.286	0.571	0.146	0.143
Std. Deviation	0.409	0.000	0.421	0.452	0.535	0.358	0.351
95% CI Std. Dev. Upper	0.421	0.000	0.463	0.467	1.177	0.458	0.388
95% CI Std. Dev. Lower	0.398	0.000	0.386	0.438	0.344	0.294	0.320
Minimum	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Maximum	1.000	0.000	1.000	1.000	1.000	1.000	1.000
25th percentile	0.000	0.000	0.000	0.000	0.000	0.000	0.000
50th percentile	0.000	0.000	0.000	0.000	1.000	0.000	0.000
75th percentile	0.000	0.000	0.000	1.000	1.000	0.000	0.000

Note. Excluded 8 rows from the analysis that correspond to the missing values of the split-by variable ETHNICITY_DESC

1. Analizowane są zależności między Wiekie, Pochodzeniem Etnicznym i Występowaniem choroby

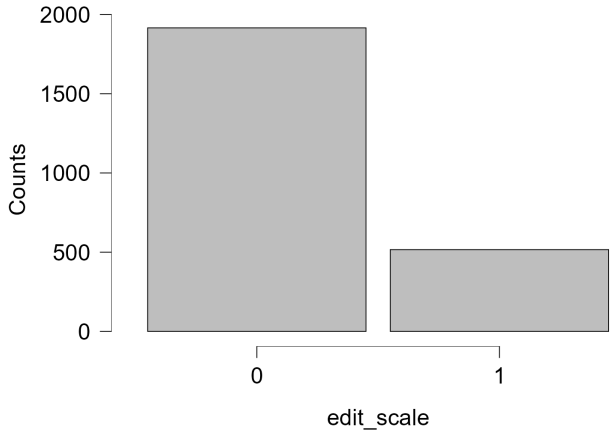
2. Występowanie choroby jest na podstawie path_severity - 0 uznaje za chorobe 1-6 za brak (wiem niezbyt dobre ale to na razie wymyśliłem idk)

3. Brane są tylko dane po biopsi bo tylko te mają path_severity

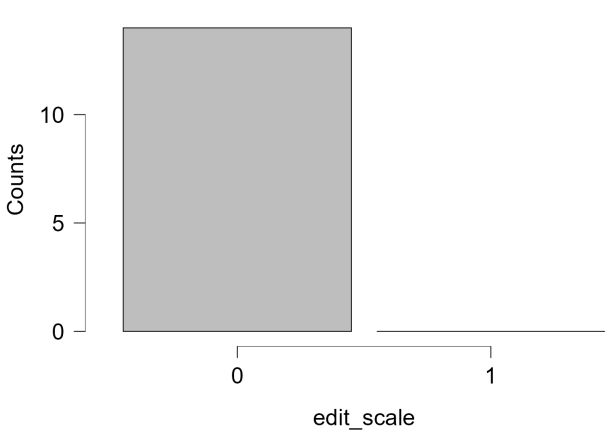
Distribution Plots

edit_scale

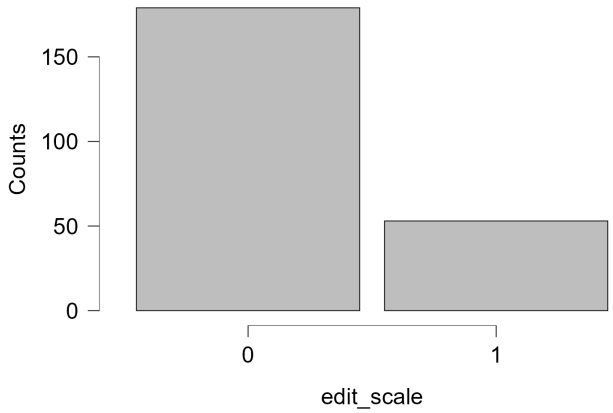
African American or Black



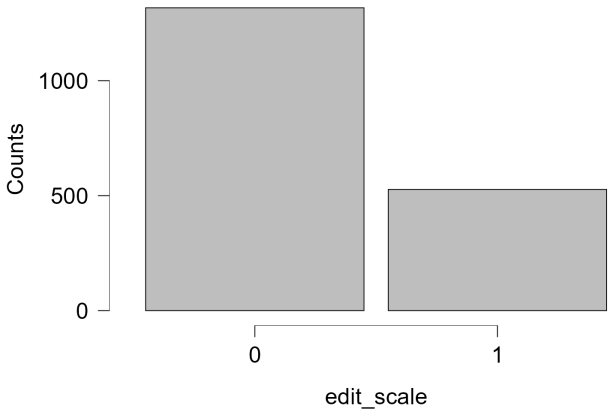
American Indian or Alaskan Native



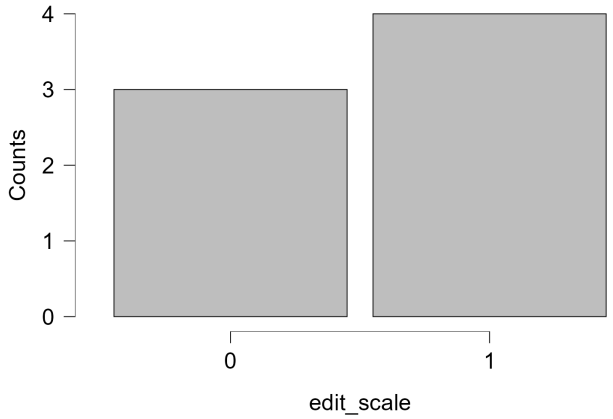
Asian



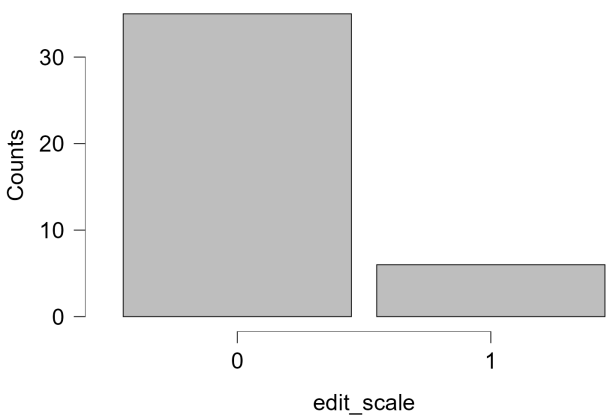
Caucasian or White



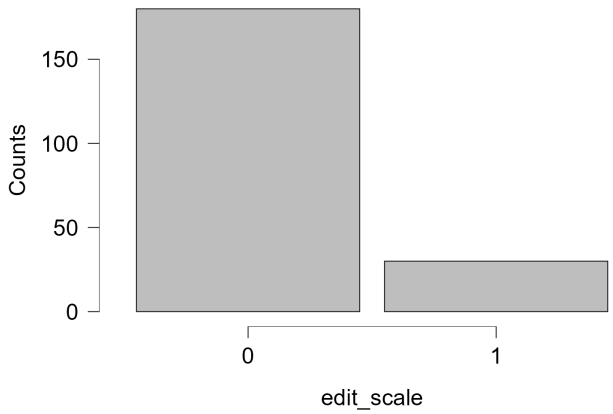
Multiple



Native Hawaiian or Other Pacific Islander



Unknown, Unavailable or Unreported



Wiek vs Występowanie Choroba

Descriptive Statistics

	age_at_study	
	0	1
Valid	3644	1136
Missing	6	2
Mean	55.939	62.230
Std. Deviation	12.374	12.042
95% CI Std. Dev. Upper	12.665	12.558
95% CI Std. Dev. Lower	12.096	11.566
Minimum	25.438	28.028
Maximum	89.000	89.000
25th percentile	46.344	53.428
50th percentile	54.870	62.498
75th percentile	65.382	71.730

1. Widocznie że średnia wieku osób chorych i zdrowych, się różnia

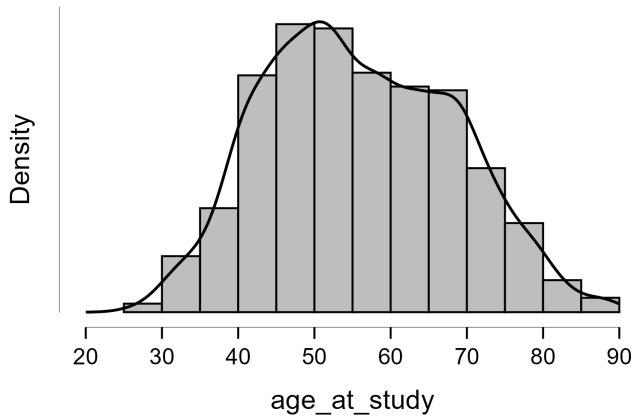
2. Obie dane mają podobna st dev - czyli wartości są w podoby sposób porozrzucane w okół średnich

3. Oba wykresy mają podobne wykresy do normalnych - gaussowskich

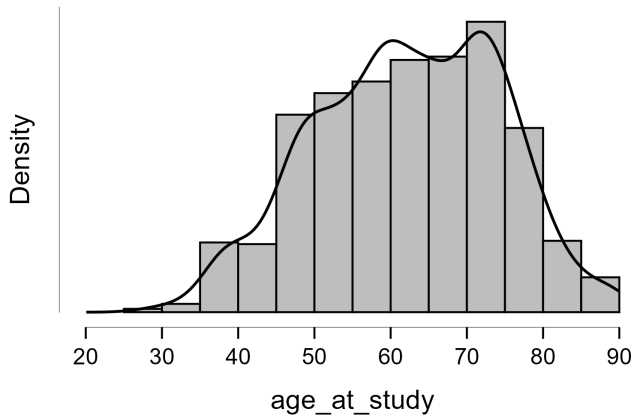
Distribution Plots

age_at_study

0



1



Chi kwadrat - Test zależności między etnicznością a chorobą

Contingency Tables

ETHNICITY_DESC		edit_scale 0	edit_scale 1	Total
African American or Black	Count	1916.000	516.000	2432.000
	Unstandardized residuals	61.982	−61.982	
	Pearson residuals	1.439	−2.578	
	Standardized residuals	4.213	−4.213	
American Indian or Alaskan Native	Count	14.000	0.000	14.000
	Unstandardized residuals	3.327	−3.327	
	Pearson residuals	1.018	−1.824	
	Standardized residuals	2.092	−2.092	
Asian	Count	179.000	53.000	232.000
	Unstandardized residuals	2.136	−2.136	
	Pearson residuals	0.161	−0.288	
	Standardized residuals	0.338	−0.338	
Caucasian or White	Count	1317.000	527.000	1844.000
	Unstandardized residuals	−88.761	88.761	
	Pearson residuals	−2.367	4.240	
	Standardized residuals	−6.196	6.196	
Multiple	Count	3.000	4.000	7.000
	Unstandardized residuals	−2.336	2.336	
	Pearson residuals	−1.011	1.811	
	Standardized residuals	−2.076	2.076	
Native Hawaiian or Other Pacific Islander	Count	35.000	6.000	41.000
	Unstandardized residuals	3.744	−3.744	
	Pearson residuals	0.670	−1.199	
	Standardized residuals	1.380	−1.380	
Unknown, Unavailable or Unreported	Count	180.000	30.000	210.000
	Unstandardized residuals	19.908	−19.908	
	Pearson residuals	1.573	−2.818	
	Standardized residuals	3.301	−3.301	
Total	Count	3644.000	1136.000	4780.000

Residuals to porównanie między ile osób powinno być w kategori (1-chory, 0 zdrowy) na pdostawie rozkladu losowego a ile jest, tz od -1 do 1 to nie sa duze roznice ale 4 oznacza ze w tej grupie jest wiece osob chorych/zdrowych niz przewidywano a -4 ze mniej.Odchylenie od normy przy braku zwiazku miedzy pcohodeniem a choroba

Chi-Squared Tests

	Value	df	p
X^2	53.382	6	< .001
X^2 continuity correction	.		
N	4780		

Note. Continuity correction is available only for 2x2 tables.

p<0.001 czyli zależność statystycznie istotna (<0.05) czyli możemy odrzucić hipotezę zerową - jest naprawdę mała szansa, że różnice były by przypadkiem

Nominal

	Value[a]
Phi-coefficient	NaN
Cramer's V	0.106

[a] Phi coefficient is only available for 2 by 2 contingency Tables

Współczynnik "siły" zależnosci na podstawie chi kwadrat (wartosci 0,1) jest bardzo blisko 0 więc gdzie zależność jest jest ona słaba (ma mał wpływ)

Korelacja Wiek - Wystepowanie choroby

Independent Samples T-Test

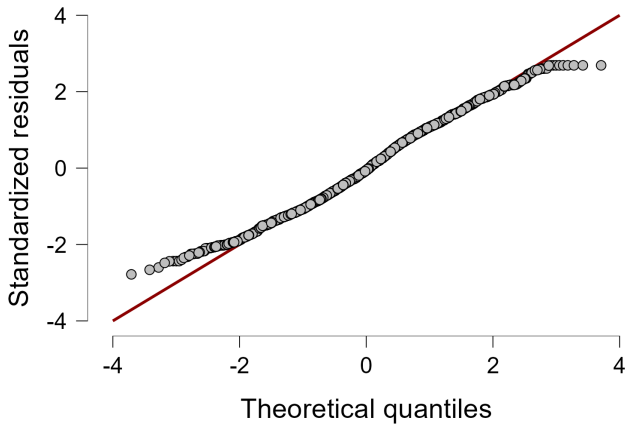
	Test	Statistic	df	p
age_at_study	Student	−15.056	4778.000	< .001
	Welch	−15.273	1939.722	< .001
	Mann-Whitney	$1.472 \times 10^{+6}$		< .001

Znowu p małe więc możmey odrzucić Hipoteze zerową co ocznacza ze wartosci nie są losowe prawdopodobie

Assumption Checks

Q-Q Plots

age_at_study



porównanie kwatyli teoretycznych z rozkładu normalnego do kwantyli rzechywistych oznacza to ze kwantyleśrodek sa zblizowe ale ich ogony nie sa idelanie normalne (mozna policzyc kurtosi i skewness ale to by nam tylko poeiwedzilo jak normalny jest rozklad wieku)

Logiczna Regresja

niebieski - jak bardzo sie zmienia model z dodaniem kolejnej źmienj (M1-M0, M2-M1)
znowu duzy skok z M0 na M1 ale maly z M1 na M2

Model Summary - edit_scale

Model	Deviance	AIC	BIC	df	ΔX²	p	McFadden R²	Nagelkerke R²	Tjur R²	Cox & Snell R²
M_0	5242.360	5244.360	5250.832	4779			0.000		0.000	
M_1	5022.903	5026.903	5039.847	4778	219.457	< .001	0.042	0.067	0.046	0.045
M_2	4997.920	5013.920	5065.697	4772	24.983	< .001	0.047	0.075	0.051	0.050

Note. M_1 includes age_at_study
Note. M_2 includes age_at_study, ETHNICITY_DESC

Jak dobrze model wyjaśnia wystepowanie choroby (wiecej lepiej)

Bla bla to samo odżucenie H0 nie lsoowe

czerwony - miary dopasowania modelu (mniejszy bardziej dopasowany) co ważne dodanie wieku bardzo zmiejsza ale pochodzenie troche a w BIC zwieksza

Coefficients

					Wald Test		
Model		Estimate	Standard Error	z	Wald Statistic	df	p
M_0	(Intercept)	−1.166	0.034	−34.301	1176.536	1	< .001
M_1	(Intercept)	−3.589	0.176	−20.351	414.169	1	< .001
	age_at_study	0.041	0.003	14.387	206.985	1	< .001
M_2	(Intercept)	−3.583	0.182	−19.707	388.376	1	< .001
	age_at_study	0.039	0.003	13.343	178.033	1	< .001
	ETHNICITY_DESC (American Indian or Alaskan Native)	−12.897	231.273	−0.056	0.003	1	0.956
	ETHNICITY_DESC (Asian)	0.219	0.167	1.308	1.712	1	0.191
	ETHNICITY_DESC (Caucasian or White)	0.270	0.074	3.672	13.485	1	< .001
	ETHNICITY_DESC (Multiple)	1.633	0.798	2.046	4.184	1	0.041
	ETHNICITY_DESC (Native Hawaiian or Other Pacific Islander)	−0.087	0.451	−0.192	0.037	1	0.848
	ETHNICITY_DESC (Unknown, Unavailable or Unreported)	−0.164	0.207	−0.789	0.622	1	0.430

Note. edit_scale level '1' coded as class 1.

Estimate logarytmu szns na poziomie 0,039 ożnacza e^(0,039)=1,0397,czyli kazdy rok to więcej o 3,9 77% wieksza sznsa na chorobe

To samo co wyżej ale e^(0,270)=1.3099 wiec, pochodzenie caukazwie zwieksza o 30.99% na raka w porownaniu z kategrią bazową (black/african)

Bardzo statycznie istotne czyli osoba czarnosóra i biała które mają tyle samo lat to biała ma większą szanse na chorobe bo czarnoskóra to bazowa

Wieksze p więc nie ma pewnosci czy bycie innym pochodzeniem, niż pochodzenie bazowe zwieksza zachorowalnosc

Bardzo niskie "z" spowodowane wysokim bledem wzgledem estimat co powoóje, że ilosc przypadkó prawdopodobnie byla za mala

Performance Diagnostics

Performance metrics

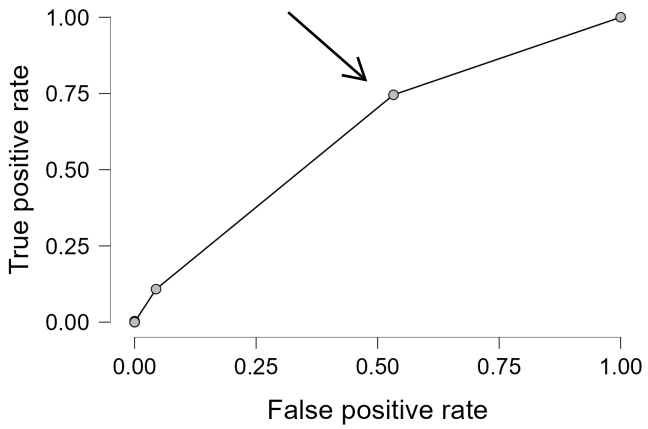
	Value
AUC	0.649

AUC to pole pod wykresem krzywęj roc - od 0.5 do 1
65% oznacza ze jesli mamy osobe chorą i zdrową to w 65% przypadków osoba chora osiągnie wyższy wynik niż zdrowa w tescie z modelu

Performance plots

ROC plot

Krzywa ROC służąca do analizowania skuteczności modeli do klasyfikacji binarnej (chory nie chory) kolejne punkty oznaczają kolejne tresholdy jakie uznamy dla naszeg modelu i jakie poziomy True i FFalse positive rate osiagniemy z wykresu mozna zobaczy ze dla 0.2 czułości osiągamy najlepszy wynik



I dlaczego wykorzystuję tutaj AUC a nie accurecy czy coś innego?

Bo ten model osiąga accurecy na poziomie najwyższym 76% jednak robi to bo przypisuje on 99% przypadków za chore (nie zbalansowana baza) i treshhold vaulena 50% model ma lepsze true positive and false positive rate jesli zmiejszymy treshold vaule

Korelacja pochodzenie etniczne a wystepowanie z podzialem na kohorty wiekowe

Dekady	Racism Reduce	edit_scale 0	edit_scale 1	Total
3	African American or Black	60	3	63
	Cauciasian or White	32	4	36
	Other	29	0	29
	Total	121	7	128
4	African American or Black	336	56	392
	Cauciasian or White	181	24	205
	Other	122	3	125
	Total	639	83	722
5	African American or Black	551	118	669
	Cauciasian or White	341	102	443
	Other	180	31	211
	Total	1072	251	1323
6	African American or Black	492	157	649
	Cauciasian or White	327	107	434
	Other	53	27	80
	Total	872	291	1163
7	African American or Black	362	119	481
	Cauciasian or White	304	192	496
	Other	20	18	38
	Total	686	329	1015
8	African American or Black	107	56	163
	Cauciasian or White	113	84	197
	Other	7	14	21
	Total	227	154	381
9	African American or Black	8	7	15
	Cauciasian or White	19	14	33
	Other	0	0	0
	Total	27	21	48
Total	African American or Black	1916	516	2432
	Cauciasian or White	1317	527	1844
	Other	411	93	504
	Total	3644	1136	4780

Note. Each cell displays the observed counts

Chciałem sprawdzić jak dokładnie wygląda zależność pochodzenie etniczne - choroba i jak bardzo zależy to od wieku (bo niektóre grupy mają większy wiek średni) a jak bardzo od pochodzenia etnicznego

Przydzieliłem każdą osobę do kohorty (ile ma lat w dekadach minus reszta) i w każdej kohorcie sprawdzam test chi kwadrat

z powodu małej ilości Podzieliłem je na caucasian/white, african/black i other

Chi-Squared Tests

Dekady		Value	df	p
3	X^2	3.956	2	0.138
	X^2 continuity correction	.		
	N	128		
4	X^2	13.172	2	0.001
	X^2 continuity correction	.		
	N	722		
5	X^2	8.022	2	0.018
	X^2 continuity correction	.		
	N	1323		
6	X^2	3.519	2	0.172
	X^2 continuity correction	.		
	N	1163		
7	X^2	25.783	2	< .001
	X^2 continuity correction	.		
	N	1015		
8	X^2	8.899	2	0.012
	X^2 continuity correction	.		
	N	381		
9	X^2	NaN		
	X^2 continuity correction	.		
	N	48		
Total	X^2	40.154	2	< .001
	X^2 continuity correction	.		
	N	4780		

Note. Continuity correction is available only for 2x2 tables.
[a] X^2 could not be calculated - At least one row or column contains all zeros

W dekdzie 3 i 9 liczba osób jest bardzo mała, więc można je teoretycznie pominac
co ciekawe według wyników w dekadzie 6 liczba p wyniosła aż 0.172 co nie pozwala
nam odrzucić hipoteze zerową ale w dekadzie 7 p już ma wartość statycznie znacząc

Ogólne podsumowanie, wiek ma bardzo duży związek z chorobą i
jest najważniejszy czynnikiem, pochodzenie etniczne też może być czynnikiem ale slabszym niż wiek
pochodzenie biale/kaukaskie wplywa na szanse na zachorowanie ale w innych nie znaleziono tej
zależnozci?