Jacob Krol BIOS 7747 Sep 27, 2023

Study summary

The National Institutes of Health funded a study to analyze what factors may predict the survival time after diagnosis of a terminal type of liver cancer. 2,500 patients were enrolled in ten different U.S. hospital and each patient underwent a biopsy that provided a measurement of twenty quantitative cell measurements. The overall goal of this study is to identify which cell measurements may be predictors of the survival time (if any) in addition to basic patient demographic information.

Goal

1. **Build and evaluate a regression model that can predict the survival time using the available data. Provide:**
    1. **A description and justification of the pre-processing steps to use categorical features, solve errors in the dataset, explore feature correlations and tackle potential problem related to collinear features.**

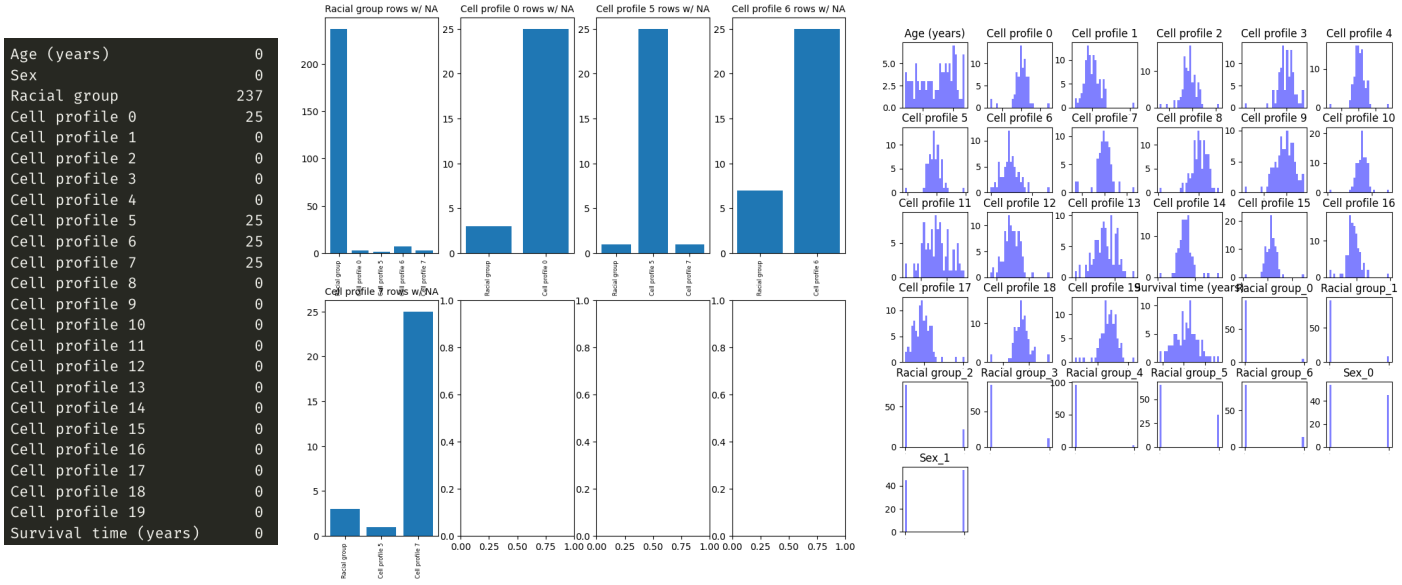


Fig 1 Missing Values: A (left) the number of missing values in each column. **B (middle)**: the overlap of missing values. Each subplot focuses on the NA rows of a particular column (e.g., top-left - the "Racial group" column), and a barplot shows how many NAs occur in other columns *given* the column of interest has a missing value (e.g., "Racial group"). **C (right)**: the distribution of values for rows that have a missing value in any column.

Fig 1 summarizes the amount of missing values (**fig 1A**), and **fig 1B** gives a qualitative look at the conditional association of NA occurrence column-wise to visualize a possible relationship between missing values of columns. **Fig 1c** shows there's a slight skew in some of the variables when a missing value is present.

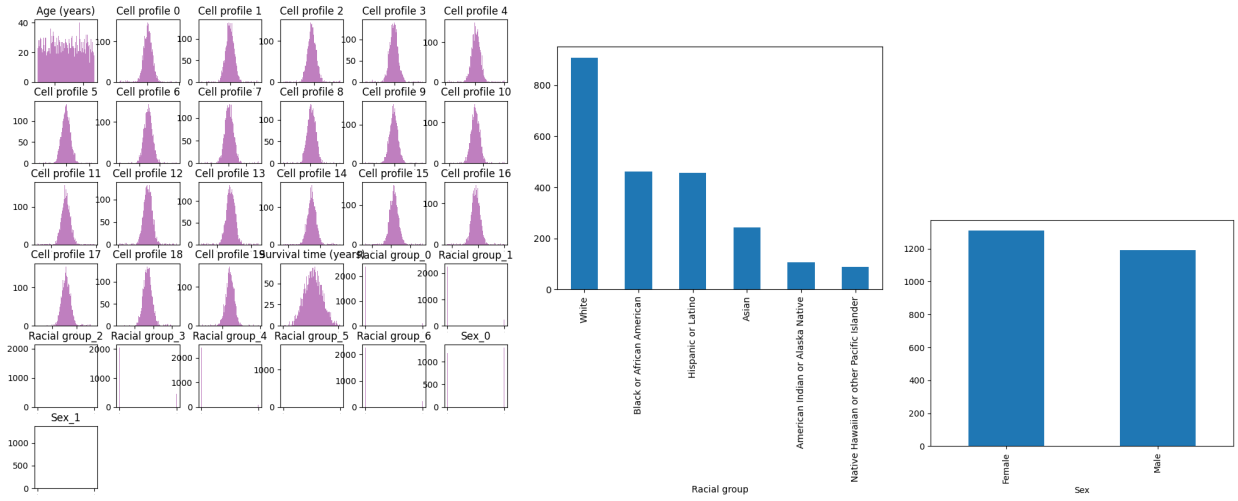


Fig 2 feature distributions: A (purple, left) the distribution of numerical variables before standardization; the encoded subplots can be ignored. **B**: (blue, middle) barplot of racial group before encoding. **C**: (blue, right) barplot of sex before encoding

	skew	pvalue
Age (years)	0.956336	9.601983e-27
Cell profile 0	0.912353	9.234053e-36
Cell profile 1	0.929665	9.155211e-33
Cell profile 2	0.915941	3.508366e-35
Cell profile 3	0.918976	1.124647e-34
Cell profile 4	0.922211	4.047334e-34
Cell profile 5	0.933181	4.368016e-32
Cell profile 6	0.930348	1.234378e-32
Cell profile 7	0.924330	9.582369e-34
Cell profile 8	0.920015	1.689694e-34
Cell profile 9	0.921239	2.742793e-34
Cell profile 10	0.927376	3.420822e-33
Cell profile 11	0.923157	5.932976e-34
Cell profile 12	0.928259	4.986142e-33
Cell profile 13	0.926288	2.161473e-33
Cell profile 14	0.934838	9.327582e-32
Cell profile 15	0.924373	9.750689e-34
Cell profile 16	0.924976	1.250945e-33
Cell profile 17	0.939004	6.735615e-31
Cell profile 18	0.939843	1.015698e-30
Cell profile 19	0.932597	3.354553e-32

	skew	pvalue
Age (years)	0.956336	9.601983e-27
Cell profile 0	0.912353	9.234053e-36
Cell profile 1	0.929665	9.155211e-33
Cell profile 2	0.915941	3.508366e-35
Cell profile 3	0.918976	1.124647e-34
Cell profile 4	0.922211	4.047334e-34
Cell profile 5	0.933181	4.368016e-32
Cell profile 6	0.930348	1.234378e-32
Cell profile 7	0.924330	9.582369e-34
Cell profile 8	0.920015	1.689694e-34
Cell profile 9	0.921239	2.742793e-34
Cell profile 10	0.927376	3.420822e-33
Cell profile 11	0.923157	5.932976e-34
Cell profile 12	0.928259	4.986142e-33
Cell profile 13	0.926288	2.161473e-33
Cell profile 14	0.934838	9.327582e-32
Cell profile 15	0.924373	9.750689e-34
Cell profile 16	0.924976	1.250945e-33
Cell profile 17	0.939004	6.735615e-31
Cell profile 18	0.939843	1.015698e-30
Cell profile 19	0.932597	3.354553e-32

Fig 3 normality test: **A** (left): A Shapiro-Wilk test was performed on the untransformed data (prior to standardization and imputation) where rows with missing values were excluded from the test. **B** (right): normality test after data transformations.

Imputation of missing values must take into consideration the type of data (e.g., categorical or numerical), and the distribution of data (e.g., normal, uniform, etc.). From **Fig 1A**, it is known that only racial group and four of the cell profile variables have missing values. To impute the values for the missing cell profile values, median imputation is used based on the qualitative normality seen by **fig 2A**; however, when testing for normality using the Shapiro-Wilk test, none of the apparent predictors are predicted to be normal (**fig 3**). **Fig 2B** shows the mode of racial group is a value of "White", and **fig 4** indicates there's minimal pairwise correlation between the racial group and other predictors. Imputing the mode ("White") to keep the 237 rows missing racial group in the final dataset seems appropriate.

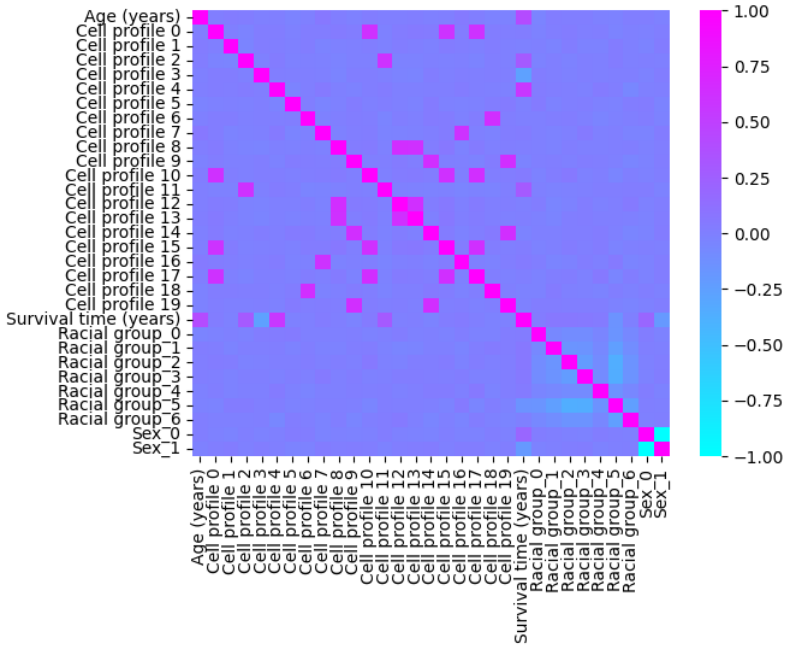


Fig 4: Pairwise correlation of variables: a correlation heatmap between variables after encoding and imputation.

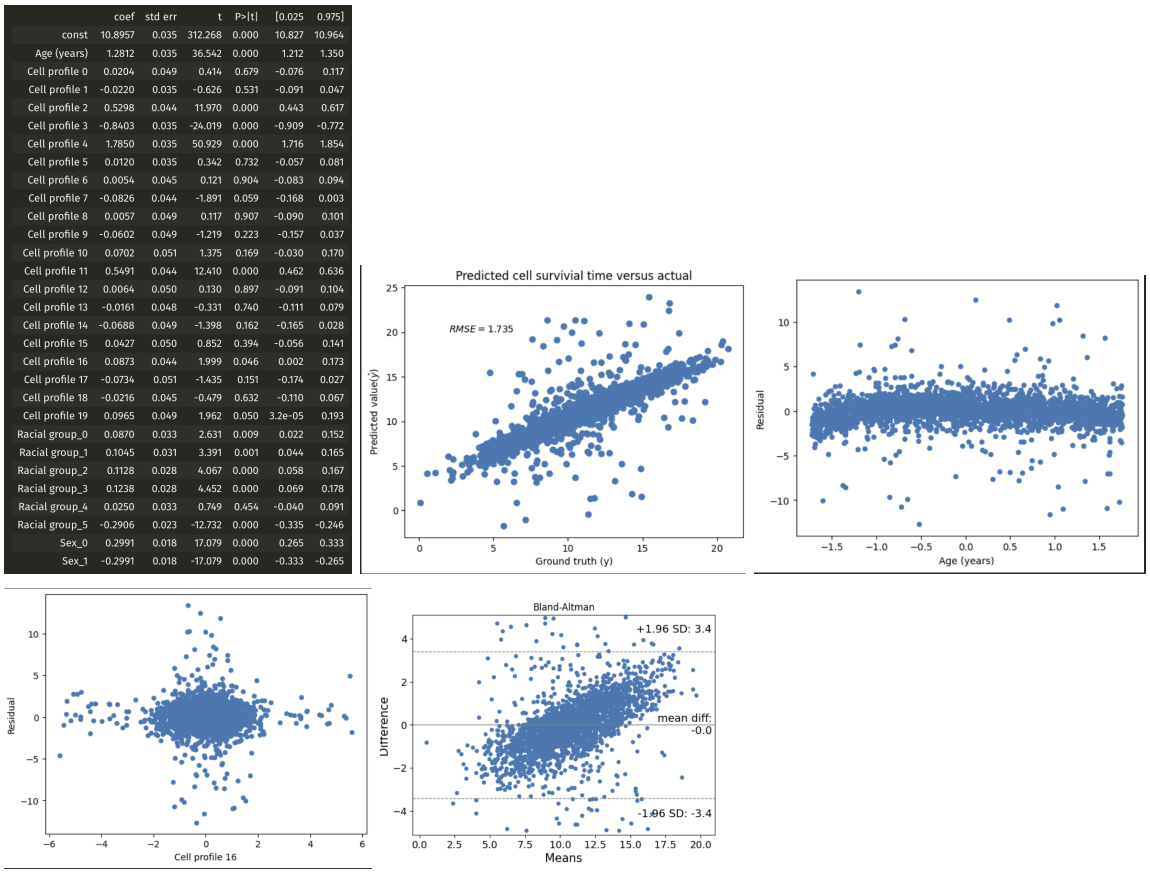
The pairwise correlation heatmap indicates only a few predictor variables will have a strong, direct linear relationship with the outcome, survival time. Some predictors do have non-zero correlation with the outcome: age, cell profiles 2, 3, 4, 11, racial groups, and sex. Also, there's multiple predictors with dependent relationships which will require a nonlinear model.

2. A mathematical equation that predicts the survival time.

$\hat{y}$ = ['bias*10.896', 'Sex_1*1.281', 'Age (years)*0.02', 'Cell profile 0*-0.022', 'Cell profile 1*0.53', 'Cell profile 2*-0.84', 'Cell profile 3*1.785', 'Cell profile 4*0.012', 'Cell profile 5*0.005', 'Cell profile 6*-0.083', 'Cell profile

```
7*0.006', 'Cell profile 8*-0.06', 'Cell profile 9*0.07', 'Cell
profile 10*0.549', 'Cell profile 11*0.006', 'Cell profile
12*-0.016', 'Cell profile 13*-0.069', 'Cell profile 14*0.043',
'Cell profile 15*0.087', 'Cell profile 16*-0.073', 'Cell profile
17*-0.022', 'Cell profile 18*0.096', 'Cell profile 19*131.371',
'Racial group_0*193.108', 'Racial group_1*253.206', 'Racial
group_2*251.728', 'Racial group_3*120.748', 'Racial
group_4*324.279', 'Racial group_5*-18.026', 'Sex_0*-18.624']
```

3. **A performance evaluation of the predictive model in the training dataset.**



	coef	std err	t	P>|t|	[0.025	0.975]
const	10.8957	0.035	312.268	0.000	10.827	10.964
Age (years)	1.2812	0.035	36.542	0.000	1.212	1.350
Cell profile 0	0.0204	0.049	0.414	0.679	-0.076	0.117
Cell profile 1	-0.0220	0.035	-0.626	0.531	-0.091	0.047
Cell profile 2	0.5298	0.044	11.970	0.000	0.443	0.617
Cell profile 3	-0.8403	0.035	-24.019	0.000	-0.909	-0.772
Cell profile 4	1.7850	0.035	50.929	0.000	1.716	1.854
Cell profile 5	0.0120	0.035	0.342	0.732	-0.057	0.081
Cell profile 6	0.0054	0.045	0.121	0.904	-0.083	0.094
Cell profile 7	-0.0826	0.044	-1.891	0.059	-0.168	0.003
Cell profile 8	0.0057	0.049	0.117	0.907	-0.090	0.101
Cell profile 9	-0.0602	0.049	-1.219	0.223	-0.157	0.037
Cell profile 10	0.0702	0.051	1.375	0.169	-0.030	0.170
Cell profile 11	0.5491	0.044	12.410	0.000	0.462	0.636
Cell profile 12	0.0064	0.050	0.130	0.897	-0.091	0.104
Cell profile 13	-0.0161	0.048	-0.331	0.740	-0.111	0.079
Cell profile 14	-0.0688	0.049	-1.398	0.162	-0.165	0.028
Cell profile 15	0.0427	0.050	0.852	0.394	-0.056	0.141
Cell profile 16	0.0873	0.044	1.999	0.046	0.002	0.173
Cell profile 17	-0.0734	0.051	-1.435	0.151	-0.174	0.027
Cell profile 18	-0.0216	0.045	-0.479	0.632	-0.110	0.067
Cell profile 19	0.0965	0.049	1.962	0.050	3.2e-05	0.193
Racial group_0	0.0870	0.033	2.631	0.009	0.022	0.152
Racial group_1	0.1045	0.031	3.391	0.001	0.044	0.165
Racial group_2	0.1128	0.028	4.067	0.000	0.058	0.167
Racial group_3	0.1238	0.028	4.452	0.000	0.069	0.178
Racial group_4	0.0250	0.033	0.749	0.454	-0.040	0.091
Racial group_5	-0.2906	0.023	-12.732	0.000	-0.335	-0.246
Sex_0	0.2991	0.018	17.079	0.000	0.265	0.333
Sex_1	-0.2991	0.018	-17.079	0.000	-0.333	-0.265

Fig 5: Ordinary least squares (OLS) regression summary table and residual plots:

A: Summary table of predictor-outcome significance from OLS. **B** (top, middle): residual plot of model performance (against original training data) with an RMSE label of 1.735 Survival time (years). **C** (top, right): residual plot for Age. **D** (bottom, left): residual plot for Cell profile 16. **E** (bottom, middle): bland-altman plot.

An ordinary least squares regression model was fit to the transformed data and the results were summarized to understand the relationship between the predictors and outcome (**fig 5**). **Fig 5a**'s p-values are low for numerous features which suggest they may not have a directly-significant

relationship with the outcome. The OLS model performance was predicted against the original training data with an RMSE of 1.735 Survival time (years) (**fig 5b)** and a bland-altman plot shows that the mean difference of the ground truth and predicted Survival time is not centered around 0. The directionality of the mean-difference plot indicates the model is under predicting for smaller ground truth values and over predicting for higher ground truth values.

For future directions, the multicollinearity between variables shown in **Fig 4** and the residual plots in **fig 5** indicate that the relationship between predictors needed to be investigated to build a model that could transform the data before model optimization.