project 1

Zijiang Yang, Dylan Sun

2017-01-08

Load in the data

Load packages

```
library(readxl)
library(knitr)
library(data.table)
library(corrplot)
library(broom)
library(stargazer)
library(ggplot2)
library(gridExtra)
```

```
prostate <- read_excel("~/Downloads/Prostate SBRT Sexual Function Data.xlsx", skip = 7)
prostate <- data.table(prostate)
prostate <- prostate[complete.cases(prostate)] # remove last row which is empty
# Code ADT as numeric; Y = 1, N = 0
prostate[ADT == "Y", ADT := 1]
prostate[ADT == "N", ADT := 0]
prostate[, ADT := as.numeric(ADT)]
```

This is what our data looks like

```
kable(head(prostate))
```

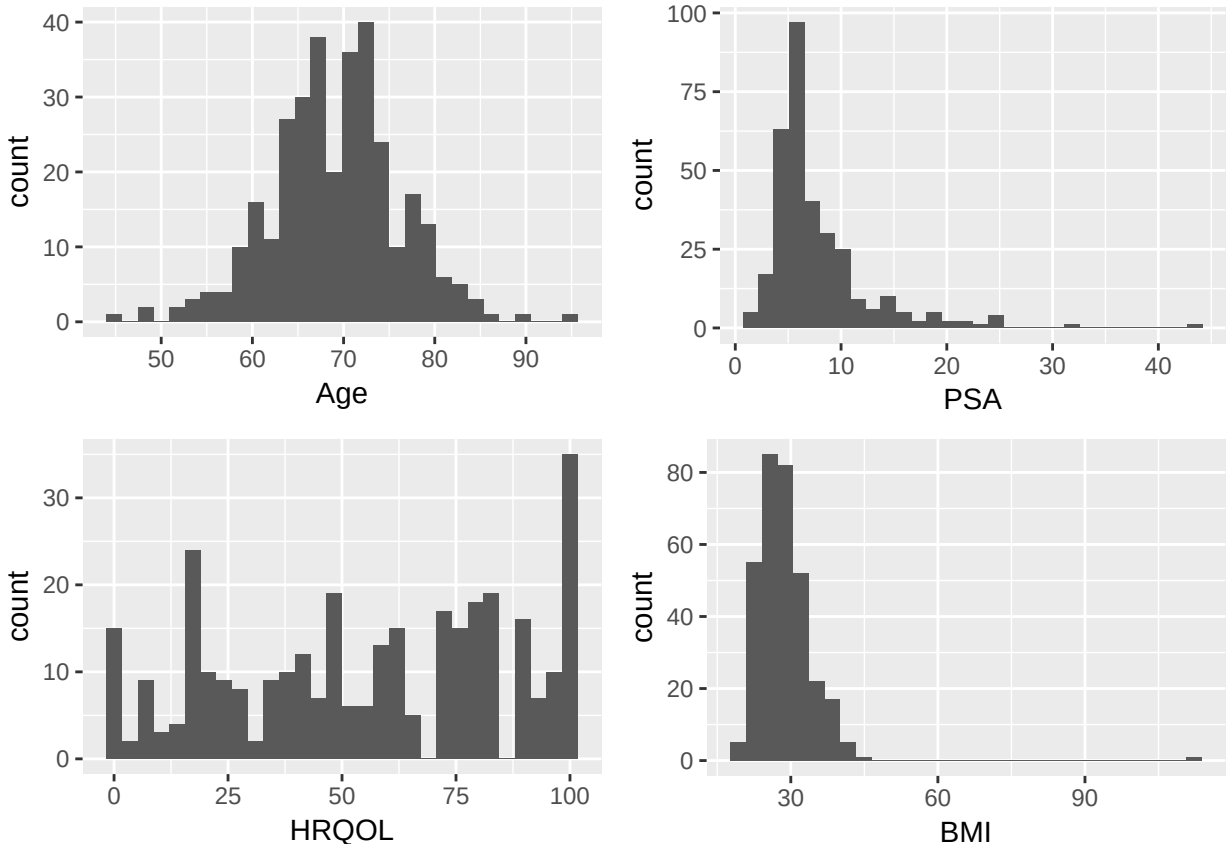
Patient	Age	Gleason Score	T-Stage Group	PSA	HRQOL	ADT	BMI	Erectile Function at Baseline	Erectil
1	82	6	0	16.70	17	0	21	0	
2	73	7	0	6.90	83	0	24	1	
3	70	7	0	7.50	71	0	30	1	
4	69	7	0	4.60	75	0	26	1	
5	69	6	0	5.60	96	0	25	1	
6	72	7	0	7.54	83	0	27	1	

Exploratory Data analysis

Histograms

```
age_hist <- ggplot(data = prostate, aes(x = Age)) +
  geom_histogram()
psa_hist <- ggplot(data = prostate, aes(x = PSA)) +
```

```
  geom_histogram()
hrqol_hist <- ggplot(data = prostate, aes(x = HRQOL)) +
  geom_histogram()
bmi_hist <- ggplot(data = prostate, aes(x = BMI)) +
  geom_histogram()
grid.arrange(age_hist, psa_hist, hrqol_hist, bmi_hist, ncol = 2)
```

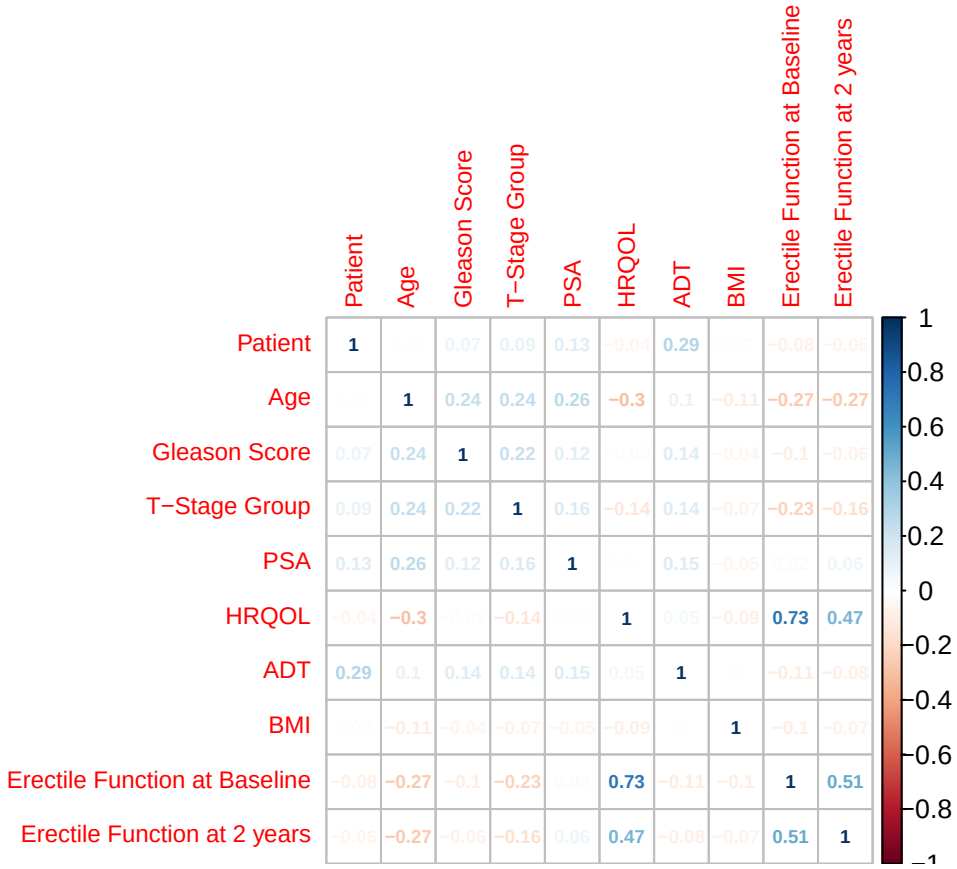


There appears to be an outlier for BMI: subject 67 has a BMI of 112. However, based on the Wikipedia page of the world's heaviest people, a BMI of 112 is realistically possible and so we decided against removing this row.

The PSA values also appear to be reasonable.

Correlation matrix

```
correlations <- cor(prostate)
corrplot(correlations, method = "number", tl.cex = 0.75, number.cex = 0.6)
```



Count the number of people in each category

We should count the number of people in each category: Has function -> no function no function -> no function Has function -> has function no function -> has function

```
nf_nf <- dim(prostate[`Erectile Function at Baseline` == 0 & `Erectile Function at 2 years` == 0])[1]
nf_hf <- dim(prostate[`Erectile Function at Baseline` == 0 & `Erectile Function at 2 years` == 1])[1]
hf_nf <- dim(prostate[`Erectile Function at Baseline` == 1 & `Erectile Function at 2 years` == 0])[1]
hf_hf <- dim(prostate[`Erectile Function at Baseline` == 1 & `Erectile Function at 2 years` == 1])[1]
counts <- data.table(never_functional = nf_nf, gain_function = nf_hf, loss_function = hf_nf, retain_fun
kable(counts)
```

never_functional	gain_function	loss_function	retain_function
153	14	70	88

Test of proportions

```
successes <- prostate[, c(sum(`Erectile Function at Baseline`), sum(`Erectile Function at 2 years`))]
failures <- dim(prostate)[1] - successes
prop_table <- data.table(successes, failures)

kable(tidy(prop.test(as.matrix(prop_table))))
```

estimate1	estimate2	statistic	p.value	parameter	conf.low	conf.high	method
0.4861538	0.3138462	19.39103	1.07e-05	1	0.0950818	0.2495336	2-sample test for equality of proportion

Logistic Regression

```
prostate_logreg <- glm(`Erectile Function at 2 years` ~ Age + `Gleason Score` + `T-Stage Group` + PSA +
# summary(prostate_logreg)
kable(tidy(prostate_logreg))
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.6215224	0.3389228	1.8338172	0.0676215
Age	-0.0094434	0.0034752	-2.7173835	0.0069428
`Gleason Score`	0.0078768	0.0372337	0.2115504	0.8325942
`T-Stage Group`	-0.0511802	0.0628191	-0.8147237	0.4158440
PSA	0.0099910	0.0047340	2.1104646	0.0356039
HRQOL	0.0029451	0.0010672	2.7596882	0.0061232
ADT	-0.0909474	0.0772114	-1.1779012	0.2397223
BMI	-0.0024765	0.0033832	-0.7319945	0.4647145
`Erectile Function at Baseline`	0.2867081	0.0659157	4.3496193	0.0000184

```
prostate <- prostate[, predicted_prob := predict(prostate_logreg)]
prostate <- prostate[, predicted := ifelse(predicted_prob >= 0.5, 1, 0)]
error_rate <- prostate[, sum(abs(predicted - `Erectile Function at 2 years`))]/dim(prostate)[1]
error_rate
```

```
## [1] 0.2123077
```

```
stargazer(prostate_logreg, header = F, type = "latex")
```

Despite HRQOL and Erectile Function at Baseline being highly correlated (erectile function of baseline in included in the HRQOL score), we do not feel that removing either one of them is justified. Both are highly predictive in the model and removing one or the other would lose information.

Logistic regression on subsets of the data

```
func_at_baseline <- prostate[`Erectile Function at Baseline` == 1] # subset only rows in which the men
sub_logreg <- glm(`Erectile Function at 2 years` ~ Age + `Gleason Score` + `T-Stage Group` + PSA + HRQO
kable(tidy(sub_logreg))
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.8281947	0.7375664	1.1228747	0.2632847
Age	-0.0105925	0.0064690	-1.6374311	0.1036366
`Gleason Score`	0.0128593	0.0665606	0.1931966	0.8470664

term	estimate	std.error	statistic	p.value
T-Stage Group	-0.1257358	0.1574006	-0.7988269	0.4256538
PSA	0.0105707	0.0074877	1.4117426	0.1600968
HRQOL	0.0067983	0.0023741	2.8634978	0.0047904
ADT	-0.1211836	0.1633098	-0.7420475	0.4592189
BMI	-0.0087521	0.0089909	-0.9734367	0.3319035

```
func_at_baseline <- func_at_baseline[, predicted_prob := predict(prostate_logreg)]
```

```
## Warning in `[.data.table`(func_at_baseline, , `:=`(predicted_prob,
## predict(prostate_logreg))): Supplied 325 items to be assigned to 158 items
## of column 'predicted_prob' (167 unused)
```

```
func_at_baseline <- func_at_baseline[, predicted := ifelse(predicted_prob >= 0.5, 1, 0)]
error_rate <- func_at_baseline[, sum(abs(predicted - `Erectile Function at 2 years`))]/dim(prostate)[1]
error_rate
```

```
## [1] 0.2861538
```

Other stuff to consider adding to the paper: We ran models removing one of either quality of life or erectile function at baseline. Those models are not included by we can incldue them if we want to.

Table 5:

	Dependent variable:
	'Erectile Function at 2 years'
Age	-0.009^{***}
	(0.003)
'Gleason Score'	0.008
	(0.037)
'T-Stage Group'	-0.051
	(0.063)
PSA	0.010^{**}
	(0.005)
HRQOL	0.003^{***}
	(0.001)
ADT	-0.091
	(0.077)
BMI	-0.002
	(0.003)
'Erectile Function at Baseline'	0.287^{***}
	(0.066)
Constant	0.622^{*}
	(0.339)
Observations	325
Log Likelihood	-152.830
Akaike Inf. Crit.	323.660
Note:	*p<0.1; **p<0.05; ***p<0.01