# project 1

*Zijiang Yang, Dylan Sun*

*2017-01-08*

### Load in the data

### Load packages

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

```r
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)]
prostate[, centered_age := Age - mean(Age)]
```

## This is what our data looks like

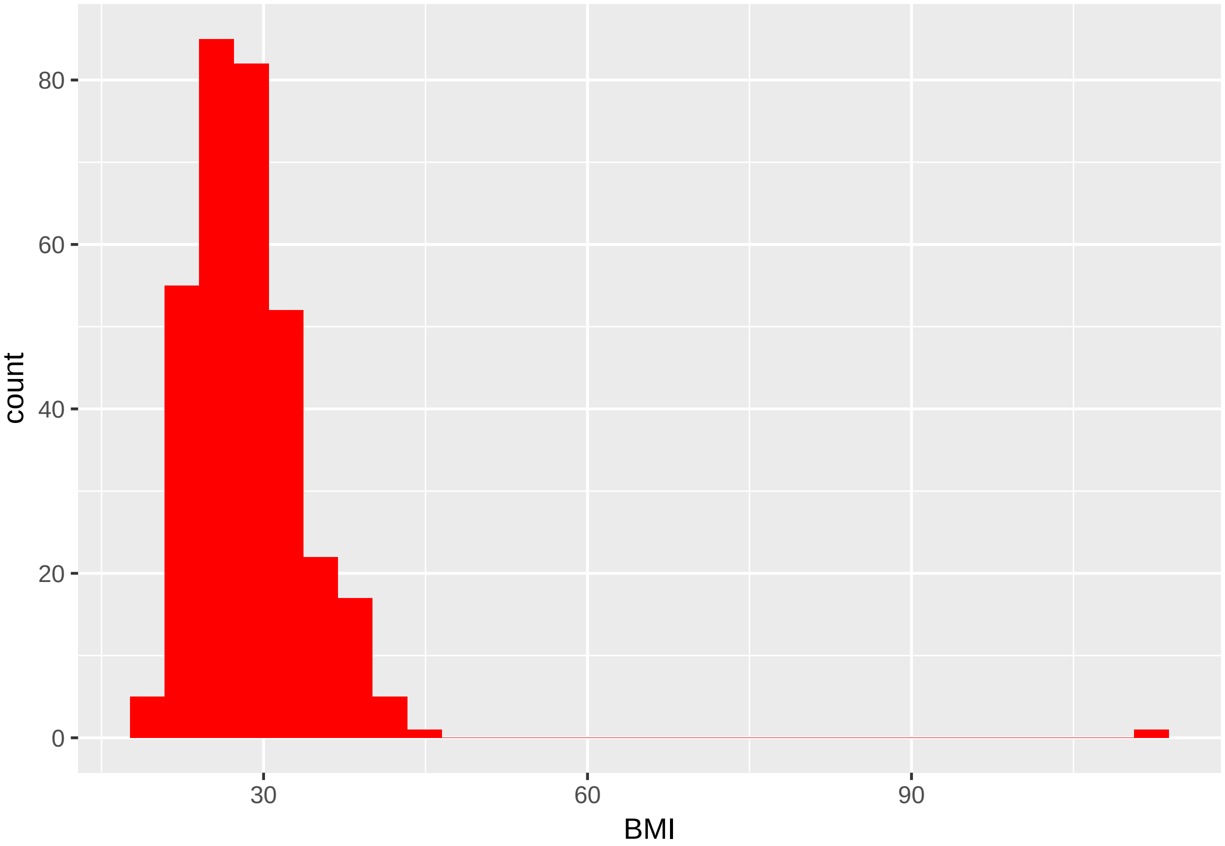
```r
kable(head(prostate))
```

| 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 | |

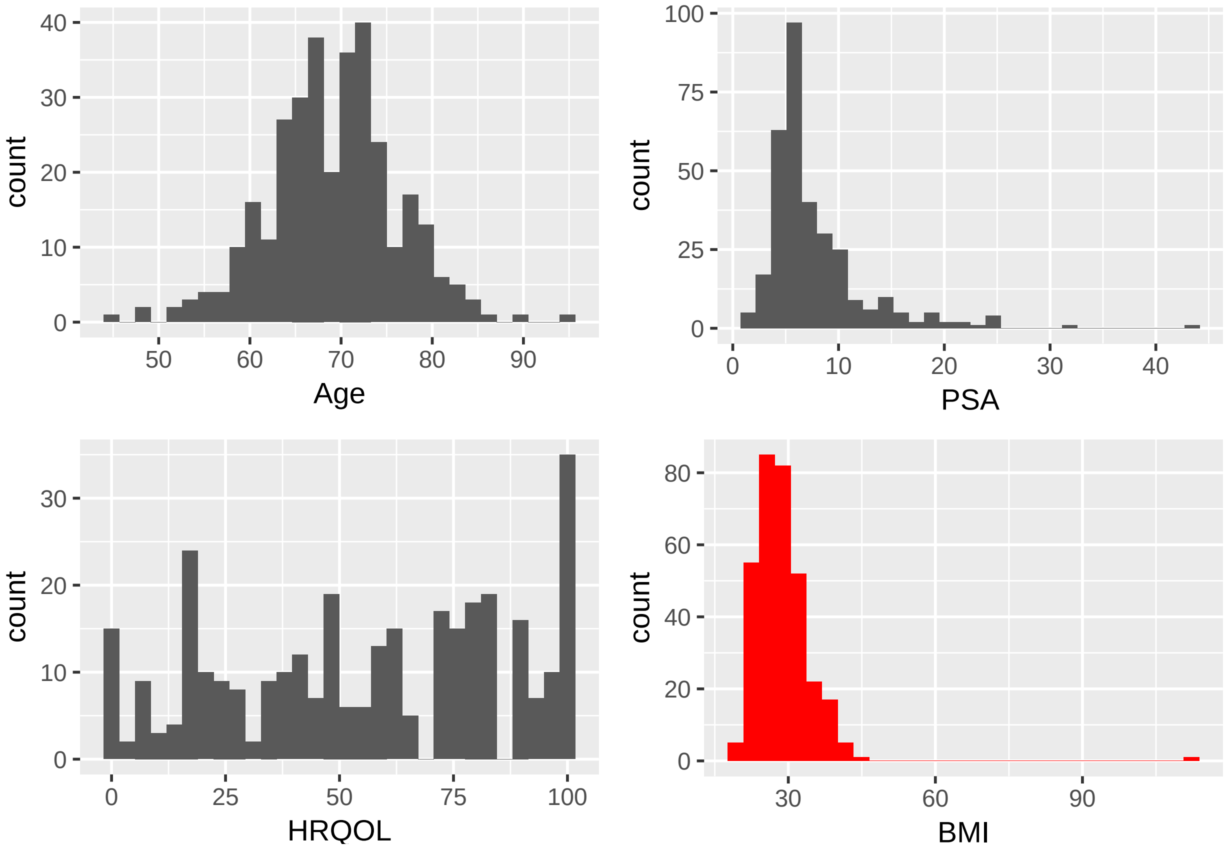
## Exploratory Data analysis

### Histograms

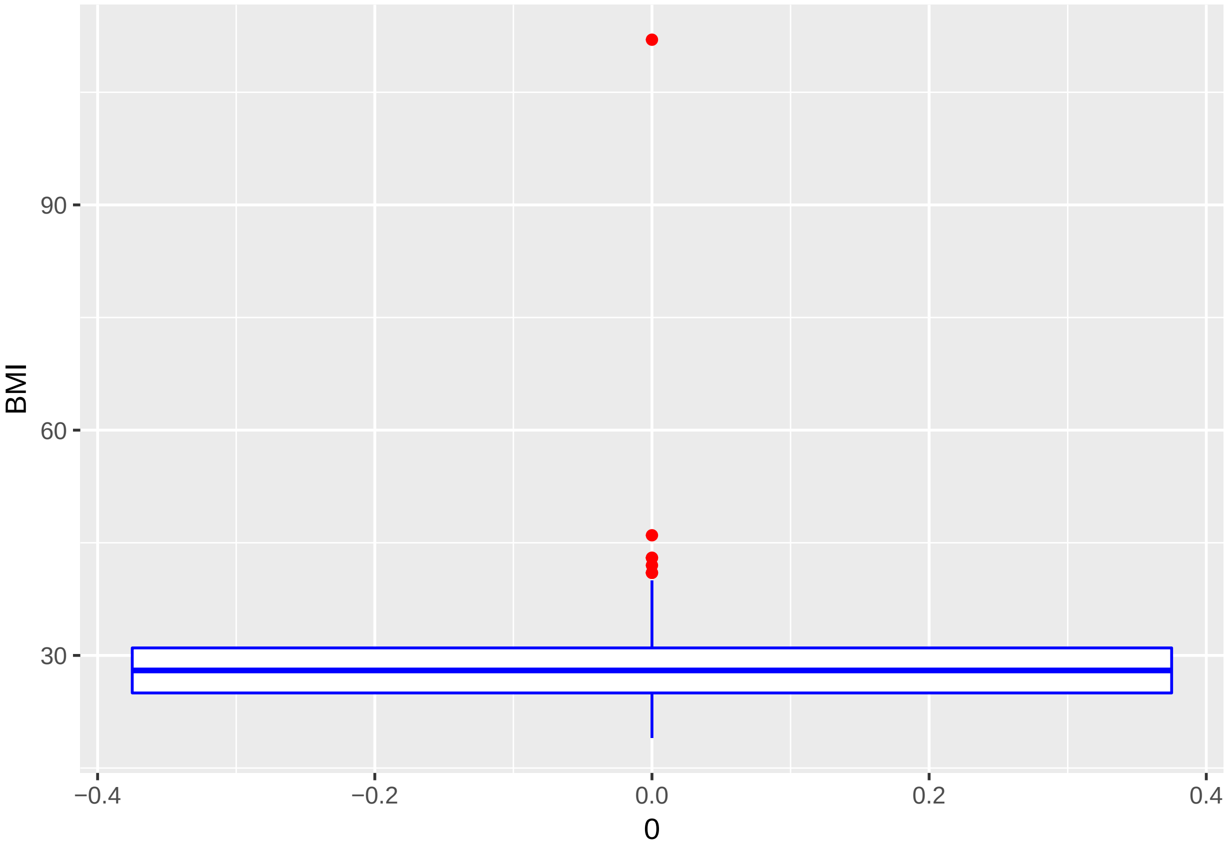
```r
age_hist <- ggplot(data = prostate, aes(x = Age)) +
  geom_histogram()
psa_hist <- ggplot(data = prostate, aes(x = PSA)) +
  geom_histogram()
hrqol_hist <- ggplot(data = prostate, aes(x = HRQOL)) +
  geom_histogram()
bmi_hist <- ggplot(data = prostate, aes(x = BMI)) +
  geom_histogram(fill = "red")
bmi_hist
```



```
grid.arrange(age_hist, psa_hist, hrqol_hist, bmi_hist, ncol = 2)
```



```
bmi_boxplot <- ggplot(data = prostate, aes(x = 0, y = BMI)) +
  geom_boxplot(color = "blue", outlier.color = "red")
bmi_boxplot
```

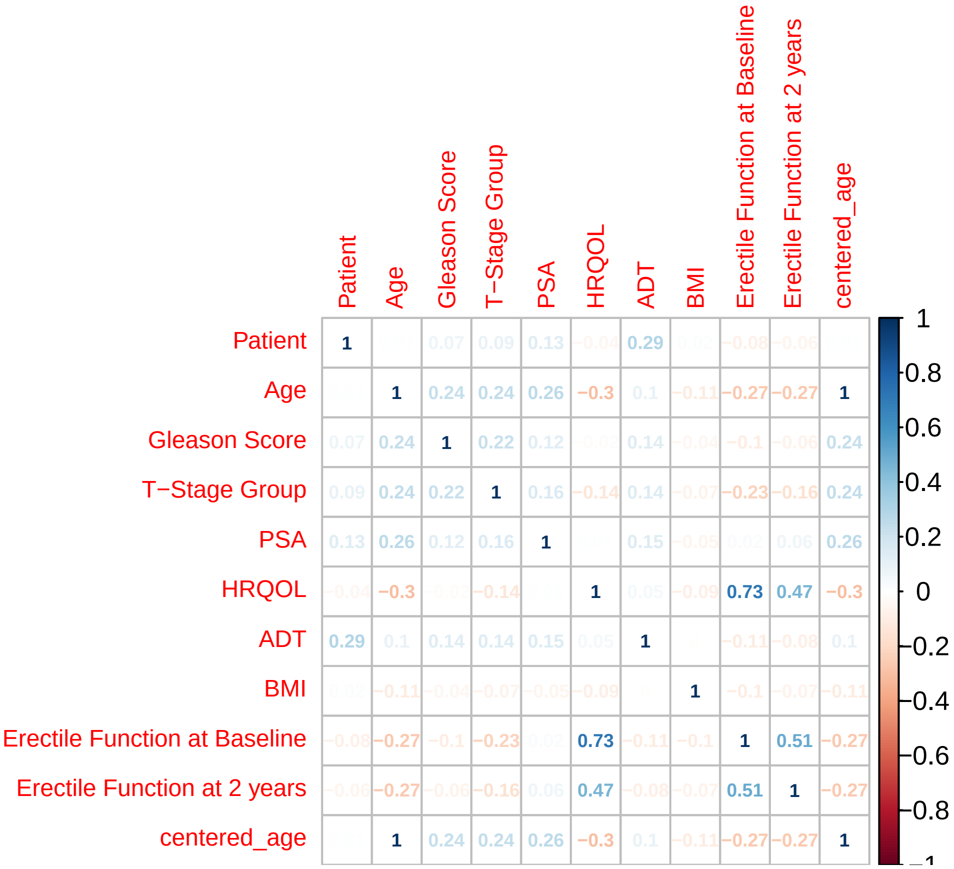


There appears to be an outlier for BMI: subject 67 has a BMI of 112. However, based on the Wikipedia page of the world's heaviest people, a BMI of 112 is realistically possible and so we decided against removing this row.

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_func
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 proportio |

### Logistic Regression

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

```
## 
## Call:
## glm(formula = `Erectile Function at 2 years` ~ Age * PSA + `Gleason Score` + 
##     `T-Stage Group` + PSA + HRQOL + ADT + BMI + `Erectile Function at Baseline`, 
##     family = binomial(link = "logit"), data = prostate)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.8035  -0.5877  -0.3165   0.8392   2.6907  
## 
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)    
## (Intercept)                      1.047147   3.551952   0.295 0.768140    
## Age                             -0.065057   0.043962  -1.480 0.138915    
## PSA                              0.026806   0.318300   0.084 0.932885    
## `Gleason Score`                  0.080093   0.249750   0.321 0.748442    
## `T-Stage Group`                 -0.578464   0.531499  -1.088 0.276435    
## HRQOL                            0.022820   0.008153   2.799 0.005124 ** 
## ADT                             -0.754107   0.606307  -1.244 0.213584    
## BMI                             -0.021589   0.030213  -0.715 0.474875    
## `Erectile Function at Baseline`  1.475934   0.439846   3.356 0.000792 ***
## Age:PSA                          0.000492   0.004483   0.110 0.912615    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 404.39  on 324  degrees of freedom
## Residual deviance: 289.85  on 315  degrees of freedom
## AIC: 309.85
## 
## Number of Fisher Scoring iterations: 5
```

```
# edit results table
prostate_logreg_results <- data.table(tidy(prostate_logreg))
prostate_logreg_results[, lower := exp(estimate - std.error)]
```

```
##                                    term      estimate   std.error   statistic
##  1:                         (Intercept)  1.0471466523 3.551951852  0.29480880
##  2:                                 Age -0.0650571085 0.043962191 -1.47984226
##  3:                                 PSA  0.0268056565 0.318299762  0.08421513
##  4:                     `Gleason Score`  0.0800932541 0.249749913  0.32069382
##  5:                     `T-Stage Group` -0.5784640933 0.531498799 -1.08836388
##  6:                               HRQOL  0.0228200589 0.008152590  2.79911769
##  7:                                 ADT -0.7541073047 0.606307126 -1.24377114
##  8:                             BMI -0.0215892696 0.030212938 -0.71457034
##  9: `Erectile Function at Baseline`  1.4759345141 0.439845694  3.35557341
## 10:                         Age:PSA  0.0004919726 0.004483063  0.10974028
##          p.value      lower
##  1: 0.7681399459 0.08169151
##  2: 0.1389153470 0.89671311
##  3: 0.9328853873 0.74714642
##  4: 0.7484424300 0.84395453
##  5: 0.2764345019 0.32957119
##  6: 0.0051242457 1.01477556
##  7: 0.2135838056 0.25655443
##  8: 0.4748745782 0.94951666
##  9: 0.0007920063 2.81817305
## 10: 0.9126153531 0.99601686
```

```
prostate_logreg_results[, OR := exp(estimate)]
```

```
##                                term      estimate   std.error   statistic
##  1:                     (Intercept)  1.0471466523 3.551951852  0.29480880
##  2:                             Age -0.0650571085 0.043962191 -1.47984226
##  3:                             PSA  0.0268056565 0.318299762  0.08421513
##  4:                 `Gleason Score`  0.0800932541 0.249749913  0.32069382
##  5:                 `T-Stage Group` -0.5784640933 0.531498799 -1.08836388
##  6:                           HRQOL  0.0228200589 0.008152590  2.79911769
##  7:                             ADT -0.7541073047 0.606307126 -1.24377114
##  8:                             BMI -0.0215892696 0.030212938 -0.71457034
##  9: `Erectile Function at Baseline`  1.4759345141 0.439845694  3.35557341
## 10:                         Age:PSA  0.0004919726 0.004483063  0.10974028
##          p.value      lower        OR
##  1: 0.7681399459 0.08169151 2.8495089
##  2: 0.1389153470 0.89671311 0.9370140
##  3: 0.9328853873 0.74714642 1.0271682
##  4: 0.7484424300 0.84395453 1.0833881
##  5: 0.2764345019 0.32957119 0.5607590
##  6: 0.0051242457 1.01477556 1.0230824
##  7: 0.2135838056 0.25655443 0.4704304
##  8: 0.4748745782 0.94951666 0.9786421
##  9: 0.0007920063 2.81817305 4.3751225
## 10: 0.9126153531 0.99601686 1.0004921
```

```
prostate_logreg_results[, upper := exp(estimate + std.error)]
```

```
##                                term      estimate   std.error   statistic
##  1:                     (Intercept)  1.0471466523 3.551951852  0.29480880
##  2:                             Age -0.0650571085 0.043962191 -1.47984226
##  3:                             PSA  0.0268056565 0.318299762  0.08421513
##  4:                 `Gleason Score`  0.0800932541 0.249749913  0.32069382
##  5:                 `T-Stage Group` -0.5784640933 0.531498799 -1.08836388
##  6:                           HRQOL  0.0228200589 0.008152590  2.79911769
##  7:                             ADT -0.7541073047 0.606307126 -1.24377114
##  8:                             BMI -0.0215892696 0.030212938 -0.71457034
##  9: `Erectile Function at Baseline`  1.4759345141 0.439845694  3.35557341
## 10:                         Age:PSA  0.0004919726 0.004483063  0.10974028
##          p.value      lower        OR      upper
##  1: 0.7681399459 0.08169151 2.8495089 99.3946713
##  2: 0.1389153470 0.89671311 0.9370140  0.9791260
##  3: 0.9328853873 0.74714642 1.0271682  1.4121388
##  4: 0.7484424300 0.84395453 1.0833881  1.3907500
##  5: 0.2764345019 0.32957119 0.5607590  0.9541205
##  6: 0.0051242457 1.01477556 1.0230824  1.0314573
##  7: 0.2135838056 0.25655443 0.4704304  0.8626035
##  8: 0.4748745782 0.94951666 0.9786421  1.0086610
##  9: 0.0007920063 2.81817305 4.3751225  6.7922361
## 10: 0.9126153531 0.99601686 1.0004921  1.0049874
```

```
kable(prostate_logreg_results[, .(term, lower, OR, upper, p.value)], digits = 2)
```

| term | lower | OR | upper | p.value |
|---|---|---|---|---|
| (Intercept) | 0.08 | 2.85 | 99.39 | 0.77 |
| Age | 0.90 | 0.94 | 0.98 | 0.14 |
| PSA | 0.75 | 1.03 | 1.41 | 0.93 |
| Gleason Score | 0.84 | 1.08 | 1.39 | 0.75 |
| T-Stage Group | 0.33 | 0.56 | 0.95 | 0.28 |
| HRQOL | 1.01 | 1.02 | 1.03 | 0.01 |
| ADT | 0.26 | 0.47 | 0.86 | 0.21 |
| BMI | 0.95 | 0.98 | 1.01 | 0.47 |
| Erectile Function at Baseline | 2.82 | 4.38 | 6.79 | 0.00 |
| Age:PSA | 1.00 | 1.00 | 1.00 | 0.91 |

```
# calculate error rate
prostate <- prostate[, predicted_prob := predict(prostate_logreg, type = "response")]
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] 0.2246154
```

## Manually create small prostate table for example predictions

```
example_prostate <- data.table(Age = 69, "Gleason Score" =  7, "T-Stage Group" = 0, "PSA" = 7.7,
                               HRQOL = 56, ADT = 0, BMI = 29, "Erectile Function at Baseline" =  1)
example_prostate <- example_prostate[, probability_of_function := predict(prostate_logreg, example_pros
example_prostate <- example_prostate[, predicted_function := ifelse(probability_of_function >= 0.5, 1,
```

```
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))
```

Table 5:

| | *Dependent variable:* |
|---|---|
| | ‘Erectile Function at 2 years‘ |
| Age | −0.065 |
| | (0.044) |
| PSA | 0.027 |
| | (0.318) |
| ‘Gleason Score‘ | 0.080 |
| | (0.250) |
| ‘T-Stage Group‘ | −0.578 |
| | (0.531) |
| HRQOL | 0.023$^{***}$ |
| | (0.008) |
| ADT | −0.754 |
| | (0.606) |
| BMI | −0.022 |
| | (0.030) |
| ‘Erectile Function at Baseline‘ | 1.476$^{***}$ |
| | (0.440) |
| Age:PSA | 0.0005 |
| | (0.004) |
| Constant | 1.047 |
| | (3.552) |
| Observations | 325 |
| Log Likelihood | −144.926 |
| Akaike Inf. Crit. | 309.852 |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| (Intercept) | 1.6092121 | 3.3021126 | 0.4873281 | 0.6260259 |
| Age | -0.0487965 | 0.0297098 | -1.6424369 | 0.1004995 |
| Gleason Score | 0.0436488 | 0.2908530 | 0.1500718 | 0.8807080 |
| T-Stage Group | -0.6237373 | 0.6979983 | -0.8936086 | 0.3715313 |
| PSA | 0.0520421 | 0.0378554 | 1.3747617 | 0.1692053 |
| HRQOL | 0.0307383 | 0.0110775 | 2.7748453 | 0.0055228 |
| ADT | -0.6576314 | 0.7433021 | -0.8847432 | 0.3762952 |
| BMI | -0.0396373 | 0.0396590 | -0.9994534 | 0.3175751 |

```
func_at_baseline <- func_at_baseline[, predicted_prob := predict(sub_logreg)]
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] 0.1846154
```

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.

## Base guessing

```
sub_logreg <- glm(`Erectile Function at 2 years` ~ PSA + HRQOL + Age + `Erectile Function at Baseline`,
kable(tidy(sub_logreg))
```

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| (Intercept) | 0.7450689 | 1.6581609 | 0.4493345 | 0.6531904 |
| PSA | 0.0483453 | 0.0313912 | 1.5400908 | 0.1235382 |
| HRQOL | 0.0203683 | 0.0074747 | 2.7249643 | 0.0064308 |
| Age | -0.0619940 | 0.0231581 | -2.6769936 | 0.0074286 |
| Erectile Function at Baseline | 1.6815924 | 0.4113486 | 4.0879981 | 0.0000435 |

```
# calculate error rate
prostate <- prostate[, predicted_prob := predict(sub_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] 0.2553846
```