R Notebook

```
library(readxl)
pcancerdf <- read_xlsx("../823data/pcancer.xlsx")
#change score and seminal to factors
#the levels keep the default names for now...
pcancerdf$score <- factor(pcancerdf$score)
pcancerdf$seminal <- factor(pcancerdf$seminal)
```

```
str(pcancerdf)
```

```
## tibble [97 x 9] (S3: tbl_df/tbl/data.frame)
##  $ idnum      : num [1:97] 1 2 3 4 5 6 7 8 9 10 ...
##  $ psa        : num [1:97] 0.651 0.852 0.852 0.852 1.448 ...
##  $ cancerv    : num [1:97] 0.56 0.372 0.601 0.301 2.117 ...
##  $ weight     : num [1:97] 16 27.7 14.7 26.6 30.9 ...
##  $ age        : num [1:97] 50 58 74 58 62 50 64 58 47 63 ...
##  $ hyperplasia: num [1:97] 0 0 0 0 0 ...
##  $ seminal    : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ capsular   : num [1:97] 0 0 0 0 0 0 0 0 0 0 ...
##  $ score      : Factor w/ 3 levels "6","7","8": 1 2 2 1 1 1 1 1 2 1 ...
```

```
print("factor levels for score")
```

```
## [1] "factor levels for score"
```

```
levels(pcancerdf$score)
```

```
## [1] "6" "7" "8"
```

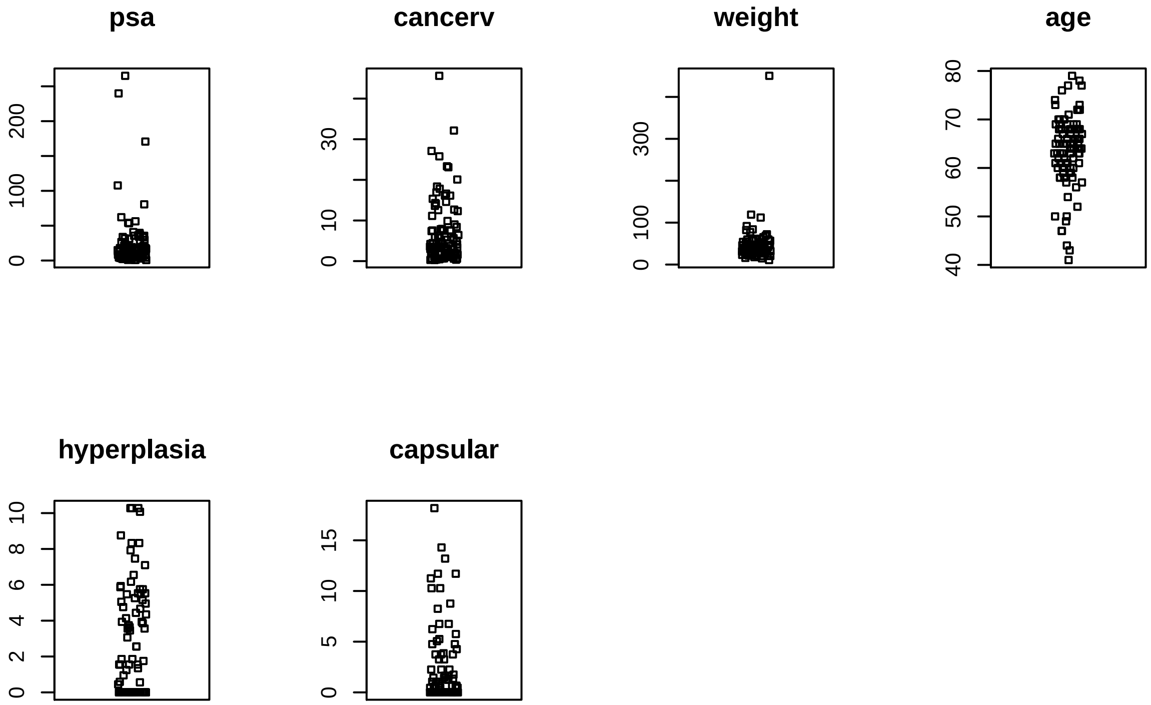
```
print("factor levels for seminal")
```

```
## [1] "factor levels for seminal"
```

```
levels(pcancerdf$seminal)
```

```
## [1] "0" "1"
```

```
par(mfrow = c(2,4))
for(i in c(2,3,4,5,6,8)){ #7 and 9 are now factors - so skip
  stripchart(pcancerdf[,i], main = names(pcancerdf[i]),
             vertical = TRUE,method = "jitter")
}
```



```
#library(epiDisplay)
#library(car)
#summ(pcancerdf)
summary(pcancerdf)
```

```
##      idnum          psa            cancerv            weight
##  Min.   : 1   Min.   :  0.651   Min.   : 0.2592   Min.   : 10.70
##  1st Qu.:25   1st Qu.:  5.641   1st Qu.: 1.6653   1st Qu.: 29.37
##  Median :49   Median : 13.330   Median : 4.2631   Median : 37.34
##  Mean   :49   Mean   : 23.730   Mean   : 6.9987   Mean   : 45.49
##  3rd Qu.:73   3rd Qu.: 21.328   3rd Qu.: 8.4149   3rd Qu.: 48.42
##  Max.   :97   Max.   :265.072   Max.   :45.6042   Max.   :450.34
##       age         hyperplasia      seminal    capsular        score
##  Min.   :41.00   Min.   : 0.000   0:76    Min.   : 0.0000   6:33
##  1st Qu.:60.00   1st Qu.: 0.000   1:21    1st Qu.: 0.0000   7:43
##  Median :65.00   Median : 1.350           Median : 0.4493   8:21
##  Mean   :63.87   Mean   : 2.535           Mean   : 2.2454
##  3rd Qu.:68.00   3rd Qu.: 4.759           3rd Qu.: 3.2544
##  Max.   :79.00   Max.   :10.278           Max.   :18.1741
```

```
library(leaps)
```

```
## Warning: package 'leaps' was built under R version 4.0.4
```

```
regsubsets.model <- regsubsets(psa ~ cancerv + weight + age + hyperplasia + capsular ,data = pcancerdf)

regsubsets.model.summ <- summary(regsubsets.model)
```

```r
regsubsets.model.summ
```

```
## Subset selection object
## Call: regsubsets.formula(psa ~ cancerv + weight + age + hyperplasia + 
##     capsular, data = pcancerdf)
## 5 Variables  (and intercept)
##             Forced in Forced out
## cancerv         FALSE      FALSE
## weight          FALSE      FALSE
## age             FALSE      FALSE
## hyperplasia     FALSE      FALSE
## capsular        FALSE      FALSE
## 1 subsets of each size up to 5
## Selection Algorithm: exhaustive
##          cancerv weight age hyperplasia capsular
## 1  ( 1 ) "*"     " "    " " " "         " "     
## 2  ( 1 ) "*"     " "    " " " "         "*"     
## 3  ( 1 ) "*"     " "    " " "*"         "*"     
## 4  ( 1 ) "*"     " "    "*" "*"         "*"     
## 5  ( 1 ) "*"     "*"    "*" "*"         "*"     
```

```r
print("adjusted $R^{2}$")
```

```
## [1] "adjusted $R^{2}$"
```

```r
regsubsets.model.summ$adjr2
```

```
## [1] 0.3831383 0.4040770 0.4020477 0.3983838 0.3917987
```

```r
print("cp")
```

```
## [1] "cp"
```

```r
regsubsets.model.summ$cp
```

```
## [1] 3.352737 1.102335 2.432833 4.003900 6.000000
```

```r
print("bic")
```

```
## [1] "bic"
```

```r
regsubsets.model.summ$bic
```

```
## [1] -38.72801 -38.52950 -34.66247 -30.54388 -25.97332
```

Look at cancerv and capsular together

```r
lm.psa.1 <- lm(data = pcancerdf, formula = psa ~ cancerv  + capsular)
summary(lm.psa.1)
```

```
## 
## Call:
## lm(formula = psa ~ cancerv + capsular, data = pcancerdf)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -60.346  -8.324  -1.205   4.159 183.843 
## 
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.3276     4.2861   0.310    0.757
## cancerv       2.4139     0.5655   4.269 4.69e-05 ***
## capsular      2.4533     1.1779   2.083    0.040 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31.48 on 94 degrees of freedom
## Multiple R-squared:  0.4165, Adjusted R-squared:  0.4041
## F-statistic: 33.55 on 2 and 94 DF,  p-value: 1.01e-11
```

```
#plot(psa ~ cancerv  + capsular, data = pcancerdf)
```

cancerv and capsular are both significant.

look at seminal, cancerv

```
lm.psa.2 <- lm(data = pcancerdf, formula = psa ~ seminal  + cancerv)
summary(lm.psa.2)
```

```
##
## Call:
## lm(formula = psa ~ seminal + cancerv, data = pcancerdf)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -55.145  -7.535  -1.129   4.256 170.018
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.060      4.231   0.251   0.8027
## seminal1      24.647      9.423   2.616   0.0104 *
## cancerv        2.477      0.495   5.003 2.62e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31.09 on 94 degrees of freedom
## Multiple R-squared:  0.431,  Adjusted R-squared:  0.4189
## F-statistic:  35.6 on 2 and 94 DF,  p-value: 3.098e-12
```

```
#plot(psa ~ seminal + cancerv, data = pcancerdf)
```

seminal and cancerv are both statistically significant in this regression

look at seminal alone and try regression again

```
lm.psa.2 <- lm(data = pcancerdf, formula = psa ~ seminal)
summary(lm.psa.2)
```

```
##
## Call:
## lm(formula = psa ~ seminal, data = pcancerdf)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -55.415  -9.598  -4.918   4.154 200.541
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   12.456      3.992    3.12  0.00239 ** 
## seminal1      52.075      8.579    6.07 2.61e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 34.8 on 95 degrees of freedom
## Multiple R-squared:  0.2794, Adjusted R-squared:  0.2719 
## F-statistic: 36.84 on 1 and 95 DF,  p-value: 2.614e-08
```

seminal alone is statistically significant

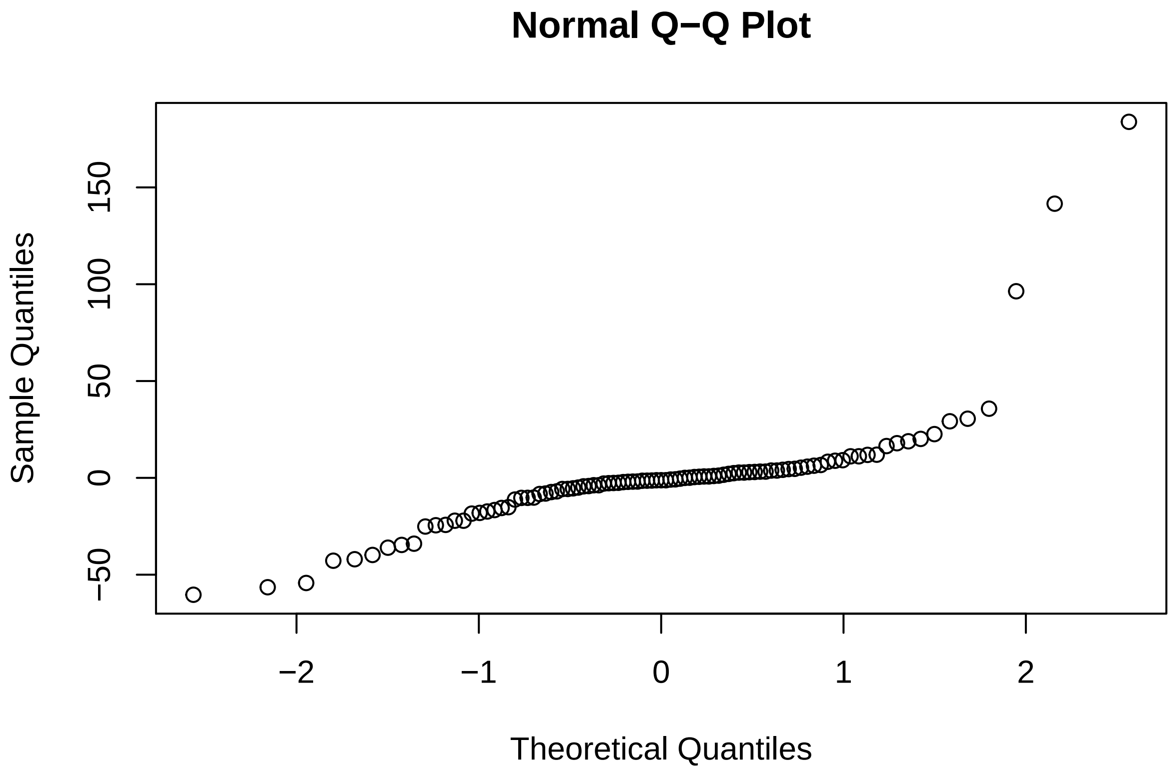
Look at seminal and score

```
summary(lm(data = pcancerdf, formula = psa ~ seminal + score))
```

```
## 
## Call:
## lm(formula = psa ~ seminal + score, data = pcancerdf)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -67.800  -7.813  -2.581   6.136 184.206 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)    9.261      5.689   1.628    0.107    
## seminal1      39.398      8.943   4.406 2.82e-05 ***
## score7        -2.331      7.722  -0.302    0.763    
## score8        32.206     10.125   3.181    0.002 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 32.64 on 93 degrees of freedom
## Multiple R-squared:  0.3793, Adjusted R-squared:  0.3593 
## F-statistic: 18.95 on 3 and 93 DF,  p-value: 1.133e-09
```

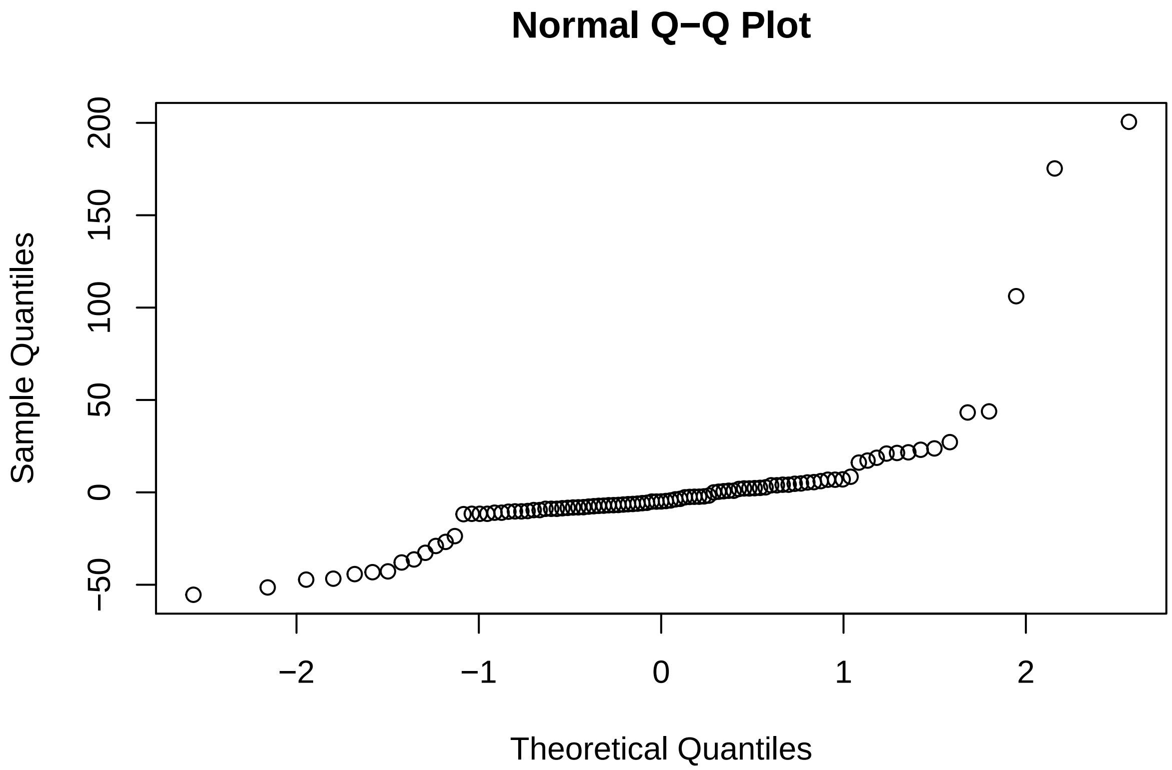
not all factors are statistically significant

```
qqnorm(residuals(lm.psa.1))
```



The residuals for cancerv and capsular do not look normally distributed

```
qqnorm(residuals(lm.psa.2))
```



the residuals for seminal doe not look normally distributed

```
pcancerdf[pcancerdf$capsular == 0,]
```

```
## # A tibble: 45 x 9
##    idnum   psa cancerv weight   age hyperplasia seminal capsular score
##    <dbl> <dbl>   <dbl>  <dbl> <dbl>       <dbl> <fct>      <dbl> <fct>
##  1     1 0.651   0.560   16.0    50        0    0              0 6
##  2     2 0.852   0.372   27.7    58        0    0              0 7
##  3     3 0.852   0.600   14.7    74        0    0              0 7
##  4     4 0.852   0.301   26.6    58        0    0              0 6
##  5     5 1.45    2.12    30.9    62        0    0              0 6
##  6     6 2.16    0.350   25.3    50        0    0              0 6
##  7     7 2.16    2.10    32.1    64        1.86 0              0 6
##  8     8 2.34    1.99    34.5    58        4.66 0              0 6
##  9     9 2.86    0.458   34.5    47        0    0              0 7
## 10    10 2.86    1.25    25.5    63        0    0              0 6
## # ... with 35 more rows
```

```
pcancerdf[pcancerdf$weight > 100,]
```

```
## # A tibble: 3 x 9
##   idnum   psa cancerv weight   age hyperplasia seminal capsular score
##   <dbl> <dbl>   <dbl>  <dbl> <dbl>       <dbl> <fct>      <dbl> <fct>
## 1    32  7.46    1.20   450.    65        5.47 0          0     6
## 2    70 19.5     3.29   119.    72       10.3  0          0.449 7
## 3    89 53.5    16.6    112.    65        0    1         11.7   8
```

```
pcancerdf[pcancerdf$hyperplasia == 0,]
```

```
## # A tibble: 43 x 9
##    idnum   psa cancerv weight   age hyperplasia seminal capsular score
##    <dbl> <dbl>   <dbl>  <dbl> <dbl>       <dbl> <fct>      <dbl> <fct>
##  1     1 0.651   0.560   16.0    50           0 0          0     6
##  2     2 0.852   0.372   27.7    58           0 0          0     7
##  3     3 0.852   0.600   14.7    74           0 0          0     7
##  4     4 0.852   0.301   26.6    58           0 0          0     6
##  5     5 1.45    2.12    30.9    62           0 0          0     6
##  6     6 2.16    0.350   25.3    50           0 0          0     6
##  7     9 2.86    0.458   34.5    47           0 0          0     7
##  8    10 2.86    1.25    25.5    63           0 0          0     6
##  9    11 3.56    1.28    36.6    65           0 0          0     6
## 10    13 3.56    5.00    20.5    63           0 0          0.549 7
## # ... with 33 more rows
```

creating new columns with 0's removed and a new data frame with the highest weights removed

```
pcancerdf$hyperplasia.na <- sapply(pcancerdf$hyperplasia,function(x){if (x == 0){return (NA)} else {retu
pcancerdf$capsular.na <- sapply(pcancerdf$capsular,function(x){if (x == 0){return (NA)} else {return (x
pcancerdf.100 <- pcancerdf[pcancerdf$weight < 100,]
```

Check the coefficients for cancerv and capsular with capsular 0's replaced with na

```
lm.psa.1.1 <- lm(data = pcancerdf, formula = psa ~ cancerv  + capsular.na)
print("coefficients for cancerv and capsular - zeros replaced")
```

```
## [1] "coefficients for cancerv and capsular - zeros replaced"
```

```
round(summary(lm.psa.1.1)$coefficient,5)
```

```
##             Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -2.15983    8.95937 -0.24107  0.81051
## cancerv      2.87308    0.98824  2.90726  0.00546
## capsular.na  1.99648    1.99710  0.99969  0.32237
```

```
print("coefficients for cancerv and capsular - raw data")
```

```
## [1] "coefficients for cancerv and capsular - raw data"
```

```
round(summary(lm.psa.1)$coefficient,5)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.32760    4.28605 0.30975  0.75744
## cancerv      2.41388    0.56547 4.26884  0.00005
## capsular     2.45330    1.17789 2.08278  0.03999
```

check hyperplasia again to see if the attribute is more useful with zeros removed

```
lm.psa.3 <- lm(data = pcancerdf, formula = psa ~ hyperplasia)
lm.psa.3.1 <- lm(data = pcancerdf, formula = psa ~ hyperplasia.na)
print("coefficients for seminal - zeros replaced")
```

```
## [1] "coefficients for seminal - zeros replaced"
```

```
round(summary(lm.psa.3.1)$coefficient,5)
```

```
##                Estimate Std. Error  t value Pr(>|t|)
## (Intercept)    34.16409   12.60306  2.71078  0.00907
## hyperplasia.na -1.83426    2.38708 -0.76841  0.44572
```

```
print("coefficients for seminal - raw data")
```

```
## [1] "coefficients for seminal - raw data"
```

```
round(summary(lm.psa.3)$coefficient,5)
```

```
##             Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 24.29238    5.43709  4.46790  0.00002
## hyperplasia -0.22182    1.38021 -0.16071  0.87266
```

```
lm.psa.3 <- lm(data = pcancerdf, formula = psa ~ weight)
lm.psa.3.1 <- lm(data = pcancerdf.100, formula = psa ~ weight)
print("coefficients for weight  - weight > 100 removed")
```

```
## [1] "coefficients for weight  - weight > 100 removed"
```

```
round(summary(lm.psa.3.1)$coefficient,5)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.58013   11.40328 0.66473  0.50788
## weight       0.40440    0.26681 1.51567  0.13303
```

```
print("coefficients for weight  - original data")
```

```
## [1] "coefficients for weight  - original data"
```

```
round(summary(lm.psa.3)$coefficient,5)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 22.66607    5.88626 3.85067  0.00021
## weight       0.02339    0.09152 0.25558  0.79882
```