

IE 5080 - EDA and Determination of Policy Form

```
# run variable selection
source(path) # run the variable selection function

## Loading required package: Matrix
## Loaded glmnet 3.0-1
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##   filter, lag
## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union

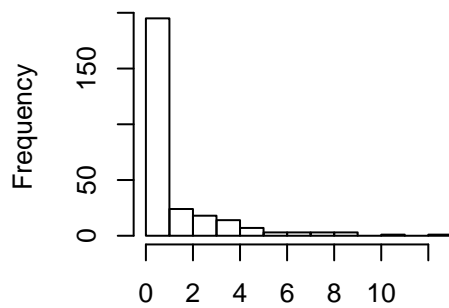
set.seed(1234) # this seed results in a dataset with 28 variables
df = variable_select()$newdata
```

For practical purposes we would like to construct our policy form using only the metadata. Since the imaging data has been normalized prior to obtaining possession, a policy form using abstracted information based on characteristics of imaging is less interpretable for clinicians. So, we have five candidate variables:

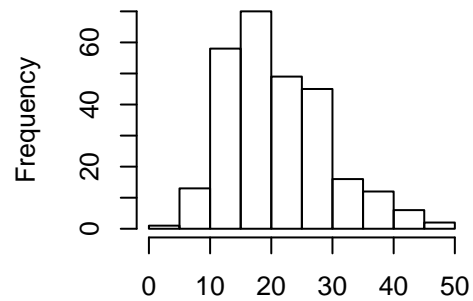
- posnodes (# positive lymph nodes)
- diam (diameter of primary tumor)
- hist (a measure of the severity of the histology?)
- age
- grade (histological grade)

```
# eda for metadata
par(mfrow=c(2,2))
hist(df$posnodes, main = "# Positive Lymph Nodes", xlab="")
hist(df$diam, main = "Diameter of Primary Tumor", xlab="")
hist(df$histtype, main = "Histological Severity?", xlab="")
hist(df$age, main = "Age", xlab="")
```

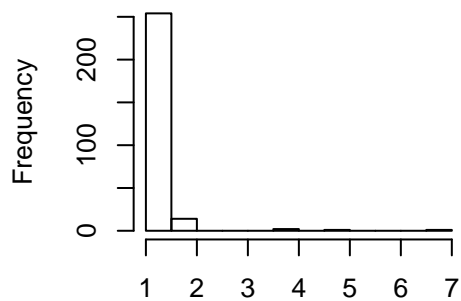
Positive Lymph Nodes



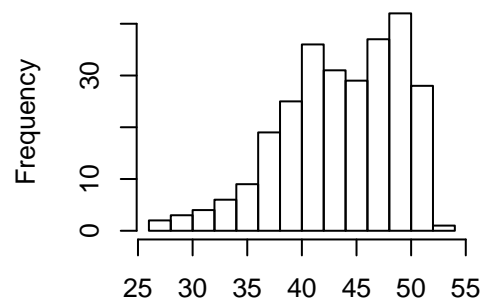
Diameter of Primary Tumor



Histological Severity?



Age



```
table(df$grade)
```

```
##
##    1    2    3
##  71   95  106
```

```
table(df$histtype)
```

```
##
##    1    2    4    5    7
## 254   14    2    1    1
```

```
table(df$posnodes)
```

```
##
##    0    1    2    3    4    5    6    7    8    9   11   13
## 137   58   24   18   14    7    3    3    3    3    1    1
```

```
table(df$angioinv)
```

```
##
##    1    2    3
## 169   30   73
```

```
table(df$lymphinfil)
```

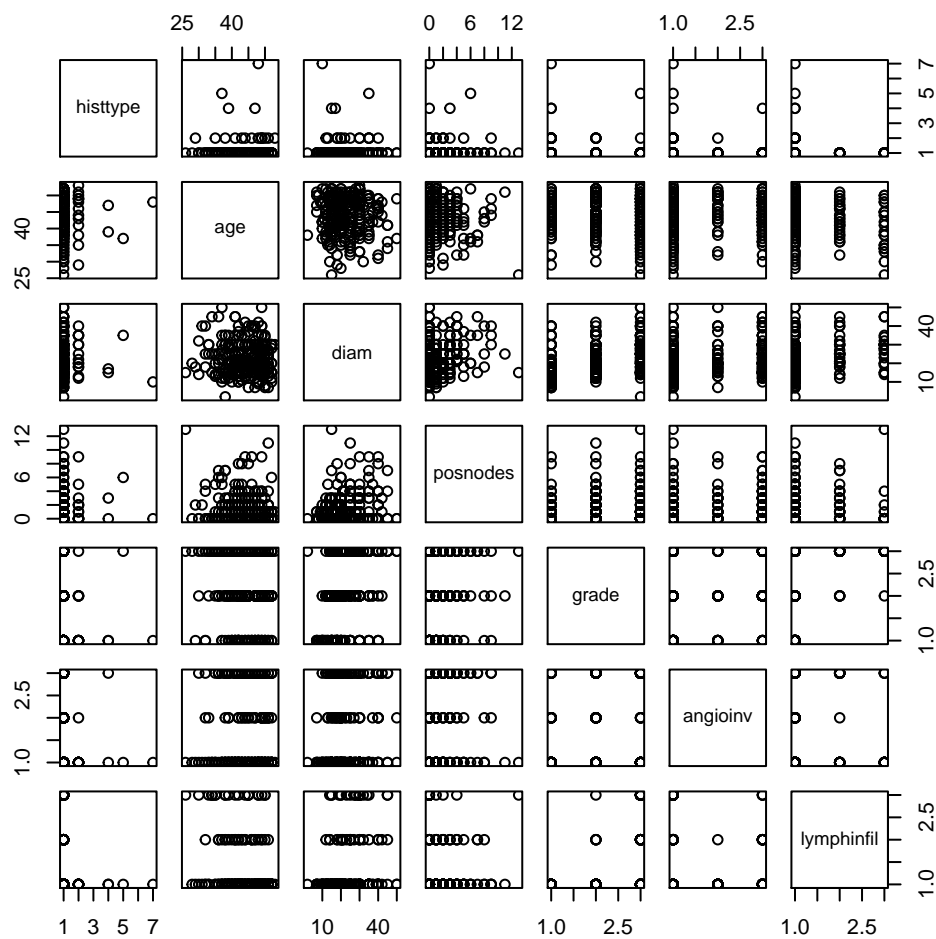
```
##
##    1    2    3
## 223   27   22
```

- posnodes contains a lot of zeros so any policy form with posnodes > 0 will automatically exclude 137 of the patients
- histtype contains 254 1s, so there is not enough variation to use it to determine treatment policy
- age, diameter, and grade seem like good candidates

```
metavars = df[,8:14]
cor = cor(metavars, method="spearman")
cor
```

```
##          histtype          age          diam          posnodes          grade
## histtype    1.00000000  0.019264075  0.033572695  0.05924833 -0.22221437
## age         0.01926407  1.000000000  0.006178775 -0.03597298 -0.09107382
## diam        0.03357269  0.006178775  1.000000000  0.15083359  0.35613078
## posnodes    0.05924833 -0.035972981  0.150833586  1.000000000 -0.01546974
## grade       -0.22221437 -0.091073825  0.356130779 -0.01546974  1.00000000
## angioinv    -0.14697620  0.044579330  0.112036374  0.23911023  0.13976685
## lymphinfil -0.12412800 -0.197331436  0.228940465 -0.00553728  0.46349511
##          angioinv lymphinfil
## histtype    -0.14697620 -0.12412800
## age         0.04457933 -0.19733144
## diam        0.11203637  0.22894047
## posnodes    0.23911023 -0.00553728
## grade       0.13976685  0.46349511
## angioinv    1.00000000 -0.07701179
## lymphinfil -0.07701179  1.00000000
```

```
plot(metavars)
```



Grade and age are moderately correlated ($r = 0.36$).

```
# run glm to determine which vars most influence death
# don't want to control for treatment
df_for_fit = df[, c(2, 8:14, 16:ncol(df))]
m1 = glm(eventdeath ~ ., data=df_for_fit, family=binomial)
summary(m1) # none of the metadata significant
```

```
##
## Call:
## glm(formula = eventdeath ~ ., family = binomial, data = df_for_fit)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.8668  -0.5517  -0.1994   0.2208   2.9002
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    2.3339205   2.2753820     1.026  0.305020
## histtype       -0.1630684   0.5897857    -0.276  0.782174
## age            -0.0898589   0.0375750    -2.391  0.016782 *
```

```
## diam          0.0393134  0.0230403   1.706 0.087955 .
## posnodes      -0.0114412  0.0865753  -0.132 0.894863
## grade         0.5992251  0.3382211   1.772 0.076445 .
## angioinv      0.1717580  0.2224016   0.772 0.439944
## lymphinfil    -1.6684194  0.4352339  -3.833 0.000126 ***
## NM_003258      -0.1379107  0.9337216  -0.148 0.882580
## NM_012067      -1.1060752  0.5642943  -1.960 0.049984 *
## NM_003430      -2.5261652  1.1783658  -2.144 0.032050 *
## AL117418       -0.5839900  0.9198213  -0.635 0.525497
## NM_006096       1.1455149  0.8556600   1.339 0.180652
## Contig23211_RC  3.3647197  1.5761202   2.135 0.032776 *
## NM_016109       1.2508821  0.7830600   1.597 0.110170
## AL049265       -0.0006939  0.7799368  -0.001 0.999290
## Contig55725_RC -0.0023075  0.7232520  -0.003 0.997454
## NM_016359       4.4894323  1.2946990   3.468 0.000525 ***
## Contig48913_RC  0.5169789  1.2654989   0.409 0.682893
## NM_001109       1.5655974  0.9619570   1.628 0.103628
## NM_001124      -0.6178550  0.7638710  -0.809 0.418603
## NM_001333       0.2912375  0.8078340   0.361 0.718461
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 324.14  on 271  degrees of freedom
## Residual deviance: 187.92  on 250  degrees of freedom
## AIC: 231.92
##
## Number of Fisher Scoring iterations: 6
```

Indeed, histtype and posnodes have the largest pvalues of the 5 metadata variables.

Next try a model with only the metadata:

```
m2 = glm(eventdeath ~ histtype + age + diam + posnodes + grade + angioinv + lymphinfil, data=df, family=
summary(m2))
```

```
##
## Call:
## glm(formula = eventdeath ~ histtype + age + diam + posnodes +
##      grade + angioinv + lymphinfil, family = binomial, data = df)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.6664  -0.7813  -0.4585   0.8599   2.5268
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.98579    1.48510  -1.337   0.1812
## histtype       0.28904    0.30870   0.936   0.3491
## age          -0.05926    0.02762  -2.146   0.0319 *
## diam          0.02834    0.01815   1.562   0.1184
## posnodes      0.02123    0.06821   0.311   0.7556
## grade         1.22355    0.24899   4.914 8.92e-07 ***
## angioinv      0.24476    0.17002   1.440   0.1500
## lymphinfil    -0.45202    0.26267  -1.721   0.0853 .
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 324.14  on 271  degrees of freedom
## Residual deviance: 272.40  on 264  degrees of freedom
## AIC: 288.4
##
## Number of Fisher Scoring iterations: 4
```

Once again, age, diam, and grade have the smallest p-values. Surprisingly, older ages correspond to higher death rates.

```
car::vif(m2)
```

```
##      histtype      age      diam  posnodes      grade  angioinv
##      1.068335  1.100621  1.125099  1.080445  1.293845  1.094403
## lymphinfil
##      1.335235
```

```
m3 = glm(eventdeath ~ age + diam + grade + lymphinfil, data=df, family=binomial)
summary(m3)
```

```
##
## Call:
## glm(formula = eventdeath ~ age + diam + grade + lymphinfil, family = binomial,
##      data = df)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.6123  -0.7712  -0.4590   0.9195   2.5206
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.17937    1.36504  -0.864   0.3876
## age          -0.06045    0.02720  -2.223   0.0262 *
## diam          0.03319    0.01779   1.866   0.0620 .
## grade         1.22738    0.24032   5.107 3.27e-07 ***
## lymphinfil   -0.53640    0.25671  -2.089   0.0367 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 324.14  on 271  degrees of freedom
## Residual deviance: 275.39  on 267  degrees of freedom
## AIC: 285.39
##
## Number of Fisher Scoring iterations: 4
```

```
car::vif(m3)
```

```
##      age      diam      grade lymphinfil
##      1.079144  1.080935  1.224555  1.292567
```

Based on these results, I think we should define the policy form based on three different sets of variables:

- age, diam, grade, lymphinfil
- age, diam, grade, lymphinfil, NM_003430, Contig23211_RC, NM_016359
- NM_003430, Contig23211_RC, NM_016359

The coefficient values for the variables were:

- positive for diam, grade, Contig23211_RC, and NM_016359
- negative for age (surprisingly), lymphinfil, and NM_003430

I next select a model for IPW and AIPW:

```
df_for_select_chemo = df[, c(5, 8:ncol(df))]
df_for_select_horm = df[, c(6, 8:ncol(df))]
df_for_select_amp = df[, c(7, 8:ncol(df))]

pi_chemo_full = glm(chemo ~., data=df_for_select_chemo, family=binomial)
pi_horm_full = glm(hormonal ~., data=df_for_select_horm, family=binomial)
pi_amp_full = glm(amputation ~., data=df_for_select_amp, family=binomial)

pi_chemo = (step(pi_chemo_full, direction="both", trace=0))
pi_horm = (step(pi_horm_full, direction="both", trace=0))
pi_amp = (step(pi_amp_full, direction="both", trace=0))
```

Look at the model selected for chemo:

```
pi_chemo

##
## Call:  glm(formula = chemo ~ posnodes + grade + NM_012067 + Contig55725_RC,
##        family = binomial, data = df_for_select_chemo)
##
## Coefficients:
##      (Intercept)      posnodes      grade      NM_012067
##      -0.2461      1.0613      -0.4685      0.7979
## Contig55725_RC
##      1.0172
##
## Degrees of Freedom: 271 Total (i.e. Null);  267 Residual
## Null Deviance:      364.6
## Residual Deviance: 252.7      AIC: 262.7
```

For the final models we will force the 7 variables that we are considering, plus the other variables that were selected by the procedure.

Chemo final IPW model:

(age, diam, grade, lymphinfil, NM_003430, Contig23211_RC, NM_016359) + posnodes, angioinv, Contig55725_RC, NM_001124

Look at the model selected for hormonal:

```
pi_horm

##
## Call:  glm(formula = hormonal ~ age + posnodes + lymphinfil + NM_012067 +
##        NM_003430 + AL117418 + Contig23211_RC, family = binomial,
##        data = df_for_select_horm)
##
## Coefficients:
##      (Intercept)      age      posnodes      lymphinfil
##      -14.4424      0.2522      0.2322      0.6361
```

```
##      NM_012067      NM_003430      AL117418  Contig23211_RC
##      1.3413      -1.5288      2.3039      1.9228
##
## Degrees of Freedom: 271 Total (i.e. Null); 264 Residual
## Null Deviance:      212.6
## Residual Deviance: 170.7      AIC: 186.7
```

For the final models we will force the 7 variables that we are considering, plus the other variables that were selected by the procedure.

Hormonal therapy final IPW model:

(age, diam, grade, lymphinfil, NM_003430, Contig23211_RC, NM_016359) + posnodes, NM_012067, AL117418

Look at the model selected for amputation:

```
pi_amp

##
## Call:  glm(formula = amputation ~ diam + posnodes + angioinv + Contig55725_RC +
##      NM_016359 + NM_001124, family = binomial, data = df_for_select_amp)
##
## Coefficients:
##      (Intercept)      diam      posnodes      angioinv
##      -1.9665      0.0816      0.1589      -0.3132
## Contig55725_RC      NM_016359      NM_001124
##      -1.6671      1.2846      0.8729
##
## Degrees of Freedom: 271 Total (i.e. Null); 265 Residual
## Null Deviance:      373.3
## Residual Deviance: 323.2      AIC: 337.2
```

For the final models we will force the 7 variables that we are considering, plus the other variables that were selected by the procedure.

Amputation final IPW model:

(age, diam, grade, lymphinfil, NM_003430, Contig23211_RC, NM_016359) + posnodes, angioinv, Contig55725_RC, NM_001124