Team 6 RP#3 - Breast cancer Prediction

Daniel Miao, Rylan Keniston, Neha Bhattacharyya, Allyssa Weinbrecht

11/21/2020

1. Import Dataset

```
# Import dataset and install packages
library("tidyverse")

## -- Attaching packages ----------------------------------------------------
-----------------------------------------------------------------------------
---------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.1     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0

## -- Conflicts -------------------------------------------------------------
-----------------------------------------------------------------------------
---------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(readr)
library(leaps)

## Warning: package 'leaps' was built under R version 4.0.3

library(car)

## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

## The following object is masked from 'package:purrr':
##
##     some

library(Ecdat)

## Warning: package 'Ecdat' was built under R version 4.0.3

## Loading required package: Ecfun
## Warning: package 'Ecfun' was built under R version 4.0.3

##
## Attaching package: 'Ecfun'

## The following object is masked from 'package:base':
##
##     sign

##
## Attaching package: 'Ecdat'

## The following object is masked from 'package:carData':
##
##     Mroz

## The following object is masked from 'package:datasets':
##
##     Orange

library(lmtest)

## Warning: package 'lmtest' was built under R version 4.0.3

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 4.0.3

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

library(psych)

##
## Attaching package: 'psych'

## The following object is masked from 'package:car':
##
##     logit

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha

library(gridExtra)

## Warning: package 'gridExtra' was built under R version 4.0.3

##
## Attaching package: 'gridExtra'
## The following object is masked from 'package:dplyr':
##
##     combine
```

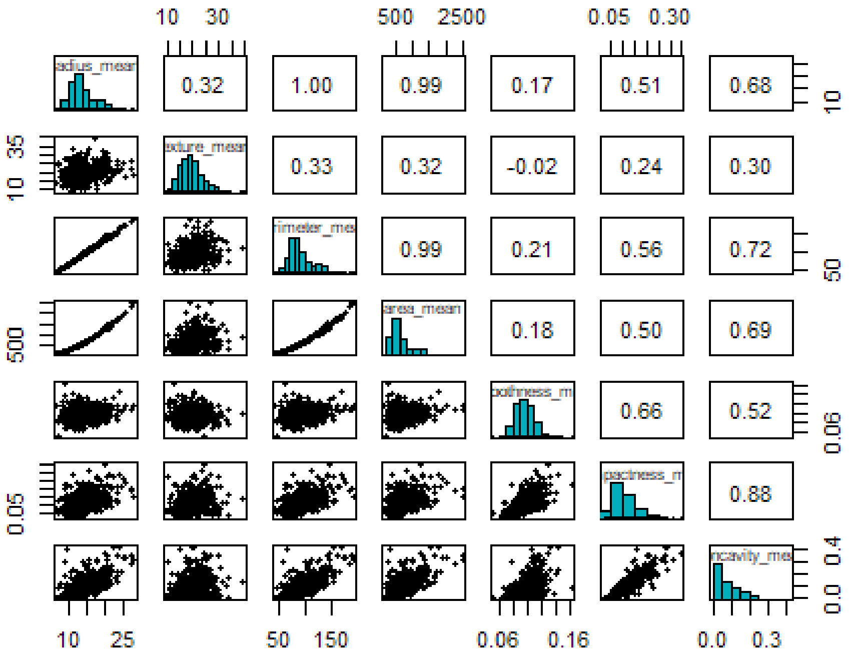
```
library(cowplot)
```

```
## Warning: package 'cowplot' was built under R version 4.0.3
```

```
source('https://tinyurl.com/y4krd9uy') # simple_anova function

setwd('D:/Documents/UT Austin/Classes/SDS 358/Project')
cancer<- read.csv("Cancer.csv")
cancer<-na.omit(cancer)

#clean the dataset, create a dummy variable for malignant=1, benign=0
response variable
cancer <- cancer %>%
  mutate(cancer, malignant=ifelse(diagnosis=='M',1,0))
```

2. Analysis of predictor relationships

```
#Predictor correlation matrix
pairs.panels((cancer)[c('radius_mean', 'texture_mean', 'perimeter_mean',
'area_mean', 'smoothness_mean', 'compactness_mean', 'concavity_mean' )],
 method = "pearson", # correlation method
 hist.col = "#00AFBB",
 smooth = FALSE, density = FALSE, ellipses = FALSE)
```



We see that the predictors radius_mean, perimeter_mean, and area_mean are all very highly correlated with each other. The predictor smoothness_mean seems to be the least correlated with the other predictors. Because almost all the predictors except for smoothness and texture are measuring mostly similar properties of the cell, it would be reasonable that most of the predictors would be moderately correlated with each other.

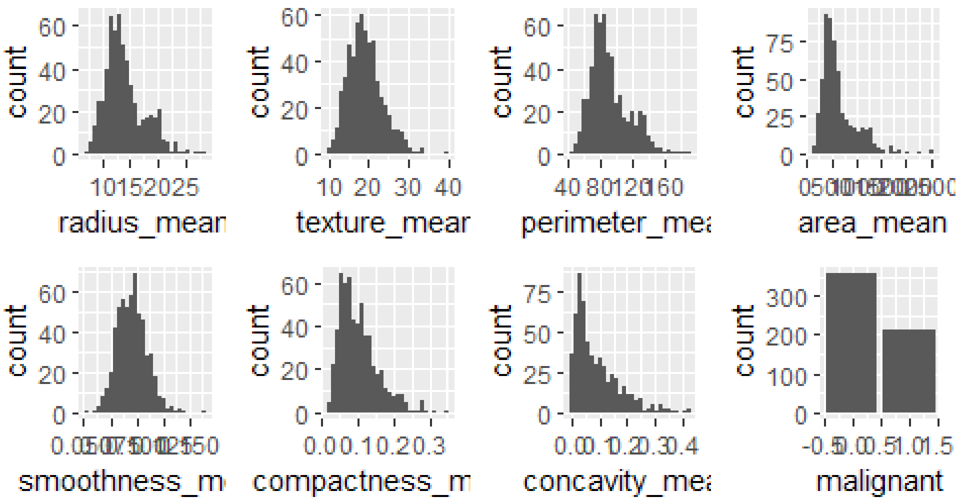
This should not be an issue when performing logistic regression, however.

2.1 Univariate Analysis

```
#Histograms to examine spread of predictors
hist1<- ggplot(data=cancer, aes(radius_mean)) + geom_histogram()
hist2<- ggplot(data=cancer, aes(texture_mean)) + geom_histogram()
hist3<- ggplot(data=cancer, aes(perimeter_mean)) + geom_histogram()
hist4<- ggplot(data=cancer, aes(area_mean)) + geom_histogram()
hist5<- ggplot(data=cancer, aes(smoothness_mean)) + geom_histogram()
hist6<- ggplot(data=cancer, aes(compactness_mean)) + geom_histogram()
hist7<- ggplot(data=cancer, aes(concavity_mean)) + geom_histogram()
bar8<-  ggplot(data=cancer, aes(x=malignant)) + geom_bar()
plot_grid(hist1,hist2,hist3,hist4,hist5,hist6,hist7,bar8,
nrows=7,ncol=4,labels=NULL)
```

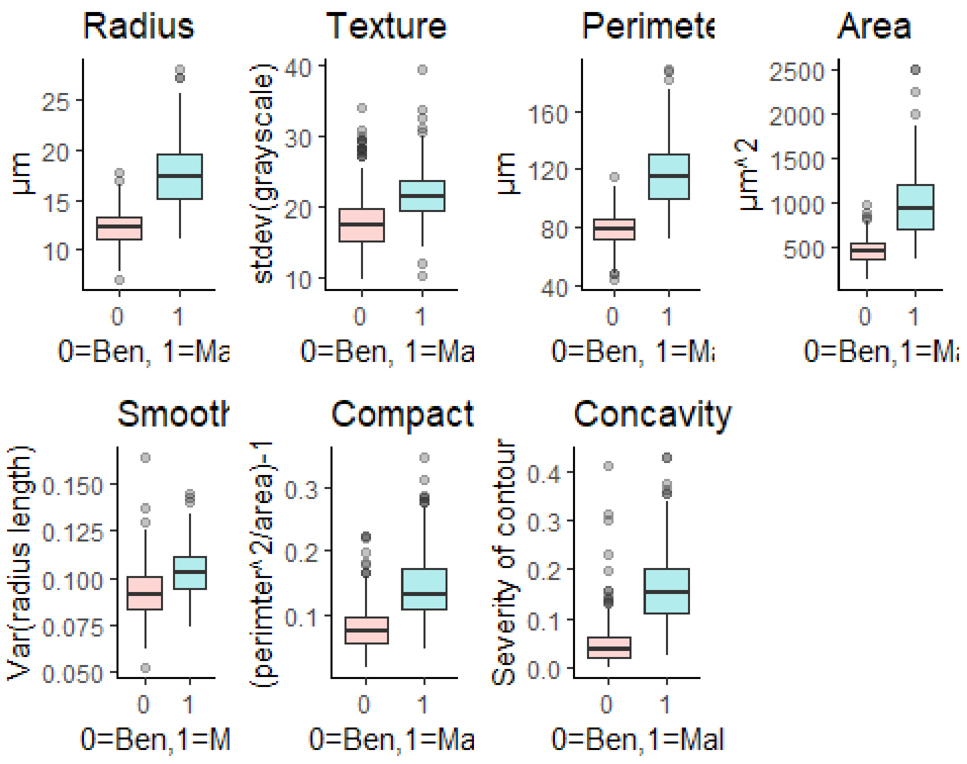
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning in as_grob.default(plot): Cannot convert object of class numeric
into a
## grob.
```



After briefly examining histograms of predictors by themselves, many of the predictors have medians closer to the left. The spread of the data is right skewed for almost all graphs besides smoothness, which stays symmetrical. For the response variable, there are around 50% more benign breast cells in the data sample than malignant. ### 2.2. Bivariate Analysis

```
# Comparing numerical variables with the response:
  boxplot1 <- ggplot(cancer, aes(x=as.factor(malignant), y=radius_mean,
fill=as.factor(malignant))) +
    geom_boxplot(alpha=0.3) +
    theme_classic() +
    theme(legend.position="none") +
    labs(title = "Radius",
     x = "0=Ben, 1=Mal", y= "µm")

  boxplot2 <- ggplot(cancer, aes(x=as.factor(malignant), y=texture_mean,
fill=as.factor(malignant))) +
    geom_boxplot(alpha=0.3) +
    theme_classic() +
    theme(legend.position="none") +
    labs(title = "Texture",
     x = "0=Ben, 1=Mal", y= "stdev(grayscale)")

  boxplot3 <- ggplot(cancer, aes(x=as.factor(malignant), y=perimeter_mean,
fill=as.factor(malignant))) +
    geom_boxplot(alpha=0.3) +
    theme_classic() +
    theme(legend.position="none") +
    labs(title = "Perimeter",
     x = "0=Ben, 1=Mal", y= "µm")

  boxplot4 <- ggplot(cancer, aes(x=as.factor(malignant), y=area_mean,
fill=as.factor(malignant))) +
    geom_boxplot(alpha=0.3) +
    theme_classic() +
    theme(legend.position="none") +
    labs(title = "Area",
     x = "0=Ben,1=Mal", y= "µm^2")

  boxplot5 <- ggplot(cancer, aes(x=as.factor(malignant), y=smoothness_mean,
fill=as.factor(malignant))) +
    geom_boxplot(alpha=0.3) +
    theme_classic() +
    theme(legend.position="none") +
    labs(title = "Smoothness",
     x = "0=Ben,1=Mal", y= "Var(radius length)")

  boxplot6 <- ggplot(cancer, aes(x=as.factor(malignant), y=compactness_mean,
fill=as.factor(malignant))) +
    geom_boxplot(alpha=0.3) +
    theme_classic() +
    theme(legend.position="none") +
    labs(title = "Compactness",
     x = "0=Ben,1=Mal", y= "(perimter^2/area)-1")

  boxplot7 <- ggplot(cancer, aes(x=as.factor(malignant), y=concavity_mean,
fill=as.factor(malignant))) +
    geom_boxplot(alpha=0.3) +
    theme_classic() +
    theme(legend.position="none") +
    labs(title = "Concavity",
     x = "0=Ben,1=Mal", y= "Severity of contour")

plot_grid(boxplot1,boxplot2,boxplot3,boxplot4,boxplot5,boxplot6,boxplot7,
nrow=2, ncol=4,labels=NULL)
```

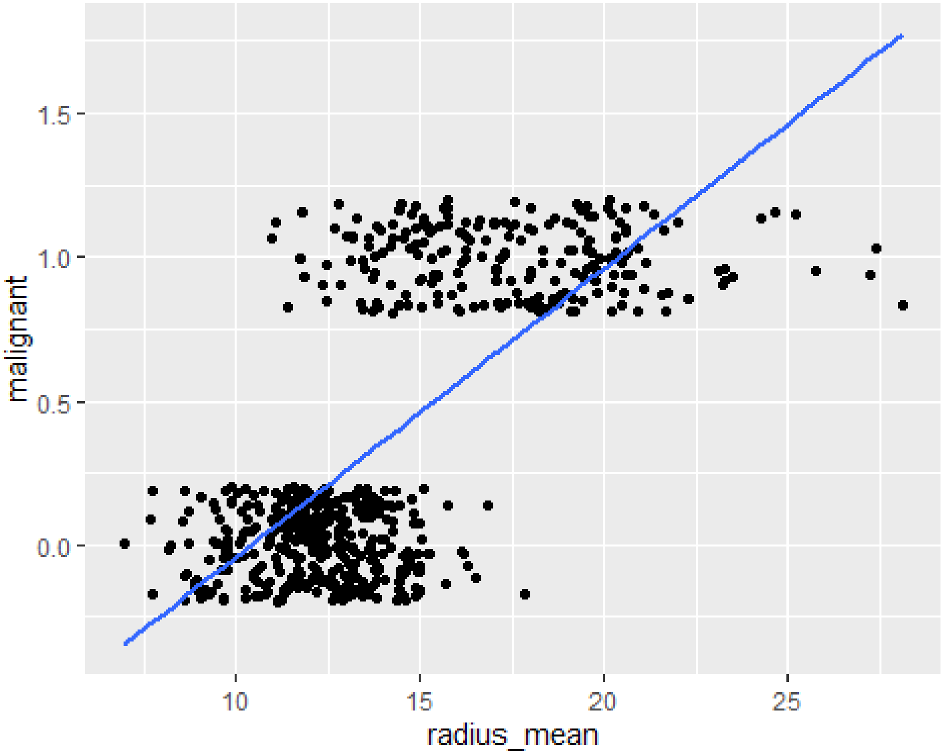


From the boxplots, we see that there are large differences in the physical properties of benign vs malignant (Ben, Mal) breast tissue cells. In addition, the IQR and top quartile spread in malignant cells seem to be significantly greater than in benign cells. The capacity for more extreme outliers seems to appear in malignant cells, but surprisingly texture and concavity have a large number of benign outliers as well.

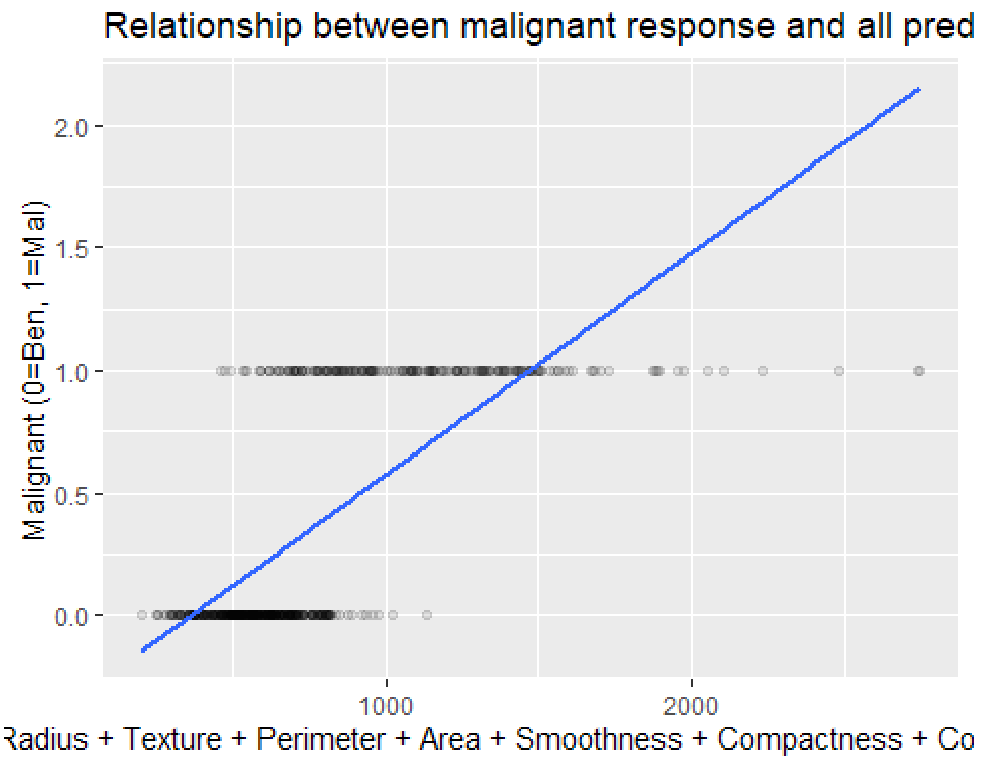
From the boxplots it appears that the predictors texture and smoothness are not as significant in differentiating between the two diagnoses. To further examine the data, we should perform a regression analysis of an initial full model.

3. Building an Initial Linear Model

```
#Create a scatter plot of response values for radius values
ggplot(cancer, aes(x=radius_mean, y=malignant, alpha=NA)) +
  geom_jitter(height=.2) +
  geom_smooth(method='lm', se = FALSE)

## Warning: Using alpha for a discrete variable is not advised.

## `geom_smooth()` using formula 'y ~ x'
```



```
lmlinear=lm(malignant~ radius_mean+
texture_mean+perimeter_mean+area_mean+smoothness_mean+compactness_mean+concav
ity_mean, data=cancer)

ggplot(cancer, aes(x=, radius_mean+
texture_mean+perimeter_mean+area_mean+smoothness_mean+compactness_mean+concav
ity_mean, y=malignant)) +
 geom_point(alpha=.1) +
 geom_smooth(method=lm, se=FALSE, fullrange=TRUE) +
 labs(title ="Relationship between malignant response and all predictors",
 x = "Radius + Texture + Perimeter + Area + Smoothness + Compactness +
Concavity", y = "Malignant (0=Ben, 1=Mal)")

## `geom_smooth()` using formula 'y ~ x'
```

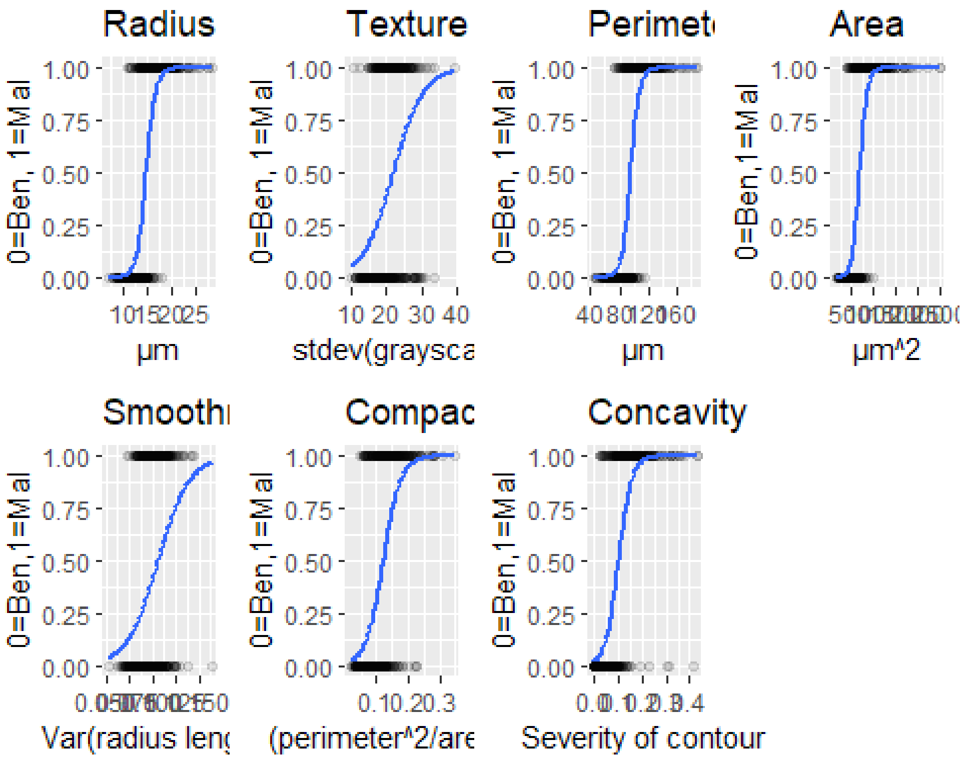


```
#Look at summary of data and R^2 improvement
summary(lmlinear)

##
## Call:
## lm(formula = malignant ~ radius_mean + texture_mean + perimeter_mean +
##     area_mean + smoothness_mean + compactness_mean + concavity_mean,
##     data = cancer)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.69625 -0.19852 -0.03572  0.18877  0.88995
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -2.4915932  0.2114522 -11.783  < 2e-16 ***
## radius_mean       0.4829980  0.1330684   3.630 0.000310 ***
## texture_mean      0.0218588  0.0029735   7.351 6.99e-13 ***
## perimeter_mean   -0.0500079  0.0208750  -2.396 0.016920 *
## area_mean        -0.0009070  0.0002321  -3.907 0.000105 ***
## smoothness_mean   5.7047685  1.1991245   4.757 2.50e-06 ***
## compactness_mean  0.6758191  0.8416452   0.803 0.422330
## concavity_mean    2.1624692  0.4145497   5.216 2.57e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2808 on 561 degrees of freedom
## Multiple R-squared:  0.6675, Adjusted R-squared:  0.6634
## F-statistic: 160.9 on 7 and 561 DF,  p-value: < 2.2e-16
```

Examining the linear regression graph, it is quite obvious that a logistic graph would be better for the categorical response variable. Howver, the greater the combined value of the predictors, the more likely it seems to be that the cell is malignant.

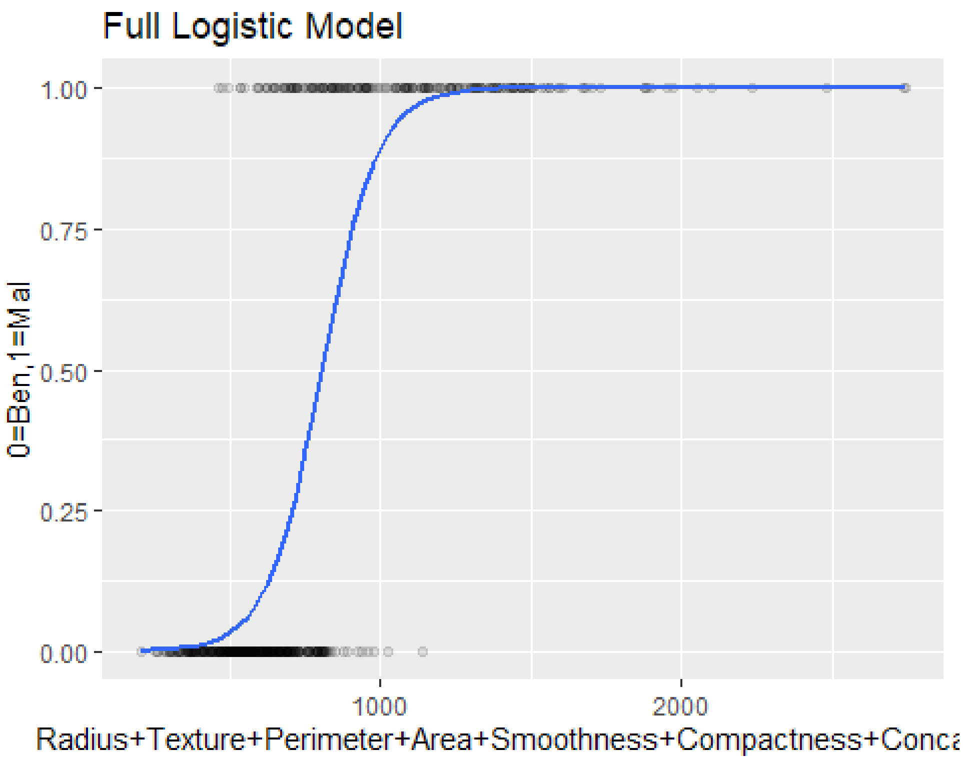
4. Building Simple Logistic Models

```
# Represent the logistic regression model
log1<-ggplot(cancer, aes(x=radius_mean, y=malignant)) +
 geom_point(alpha = 0.1) +
 geom_smooth(method = "glm", method.args = list(family = "binomial"), se =
FALSE) +
 labs(title = "Radius",y= "0=Ben, 1=Mal", x= "µm")

  log2 <- ggplot(cancer, aes(x=texture_mean, y=malignant)) +
    geom_point(alpha=.1) +
 geom_smooth(method = "glm", method.args = list(family = "binomial"), se =
FALSE) +
    labs(title = "Texture",y = "0=Ben, 1=Mal", x= "stdev(grayscale)")

  log3 <- ggplot(cancer, aes( x=perimeter_mean, y=malignant)) +
    geom_point(alpha=.1) +
 geom_smooth(method = "glm", method.args = list(family = "binomial"), se =
FALSE) +
    labs(title = "Perimeter",y = "0=Ben, 1=Mal", x= "µm")

  log4 <- ggplot(cancer, aes( x=area_mean,  y=malignant)) +
    geom_point(alpha=.1) +
 geom_smooth(method = "glm", method.args = list(family = "binomial"), se =
FALSE) +
    labs(title = "Area",y = "0=Ben,1=Mal", x= "µm^2")

  log5 <- ggplot(cancer, aes(x=smoothness_mean, y=malignant)) +
    geom_point(alpha=.1) +
 geom_smooth(method = "glm", method.args = list(family = "binomial"), se =
FALSE) +
    labs(title = "Smoothness",y = "0=Ben,1=Mal", x= "Var(radius length)")

  log6 <- ggplot(cancer, aes( x=compactness_mean,  y=malignant)) +
    geom_point(alpha=.1) +
 geom_smooth(method = "glm", method.args = list(family = "binomial"), se =
FALSE) +
    labs(title = "Compactness", y = "0=Ben,1=Mal", x= "(perimeter^2/area)-1")

  log7 <- ggplot(cancer, aes( x=concavity_mean,  y=malignant)) +
    geom_point(alpha=.1) +
 geom_smooth(method = "glm", method.args = list(family = "binomial"), se =
FALSE) +
    labs(title = "Concavity", y = "0=Ben,1=Mal", x= "Severity of contour")

plot_grid(log1,log2,log3,log4,log5,log6,log7, nrow=2, ncol=4,labels=NULL)

## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```



5. Building a Full Logistic Model

```
# Full model with all predictors:
log_full <- ggplot(cancer,
aes(x=radius_mean+texture_mean+perimeter_mean+area_mean+smoothness_mean+compa
ctness_mean+concavity_mean, y=malignant)) +
geom_point(alpha=.1) +
geom_smooth(method = "glm", method.args = list(family = "binomial"), se =
FALSE) +
labs(title = "Full Logistic Model", y = "0=Ben,1=Mal", x=
"Radius+Texture+Perimeter+Area+Smoothness+Compactness+Concavity")

log_full

## `geom_smooth()` using formula 'y ~ x'
```



```
reglog_full <- glm(malignant~
radius_mean+texture_mean+perimeter_mean+area_mean+smoothness_mean+compactness
_mean+concavity_mean, family=binomial, cancer)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

# Summary table of the full model
summary(reglog_full)

##
## Call:
## glm(formula = malignant ~ radius_mean + texture_mean + perimeter_mean +
##     area_mean + smoothness_mean + compactness_mean + concavity_mean,
##     family = binomial, data = cancer)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.96298  -0.16430  -0.03878   0.00740   3.04433
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -24.36955    9.72083  -2.507 0.012178 *
## radius_mean       -2.60480    3.62057  -0.719 0.471867
## texture_mean       0.38566    0.06330   6.093 1.11e-09 ***
## perimeter_mean     0.24624    0.47983   0.513 0.607829
## area_mean          0.02698    0.01485   1.817 0.069153 .
## smoothness_mean  136.11923   25.74295   5.288 1.24e-07 ***
## compactness_mean -14.44258   15.97534  -0.904 0.365967
## concavity_mean    21.17825    5.94281   3.564 0.000366 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 751.44  on 568  degrees of freedom
## Residual deviance: 155.96  on 561  degrees of freedom
## AIC: 171.96
##
## Number of Fisher Scoring iterations: 8
```

```
# Interpretation of slope coefficients in terms of odds
exp(coefficients(reglog_full))
```

```
##      (Intercept)      radius_mean     texture_mean   perimeter_mean
##     2.608779e-11     7.391788e-02     1.470580e+00     1.279202e+00
##        area_mean  smoothness_mean compactness_mean   concavity_mean
##     1.027348e+00     1.305663e+59     5.341569e-07     1.576141e+09
```

The full regression illustrates the comprehensive graph combining all the simple logistic regression models.

The equation of this model is: log(π/(1-π)) = -24.3696 - 2.6048(radius) + 0.3857(texture) + 0.2462(perimeter) + 0.0692(area) + 136.1192(smoothness) - 14.4426(compactness) + 21.1783 (concavity)

The most significant p values are from the predictors texture, smoothness, and concavity. The odds are most significantly impacted by changes in texture, area, smoothness, and concavity, using a significance level of a=.15 for significance in logistic regression. The p values for these predictors are 0, .0692, 0, and .0004, respectively. The exponent e raised to the slope of each predictor gives us the odds, which will enhance interpretability.

To interpret the odds, after holding all other predictors constant:

A 1 μm increase in mean radius of breast cancer cell tissue decreases the odds of the cell being malignant by a factor of 2.6048.

A 1 stdev(grayscale) increase in mean texture increases the odds of the cell being malignant by a factor of 0.3857.

A 1 μm increase in mean perimeter length increases the odds of the cell being malignant by a factor of 0.2462.

A 1 μm^2 increase in mean area of the cell increases the odds of the cell being malignant by a factor of 0.0692.

A 1 var(radius length) increase in mean smoothness variance increases the odds of the cell being malignant by a factor of 136.1192.

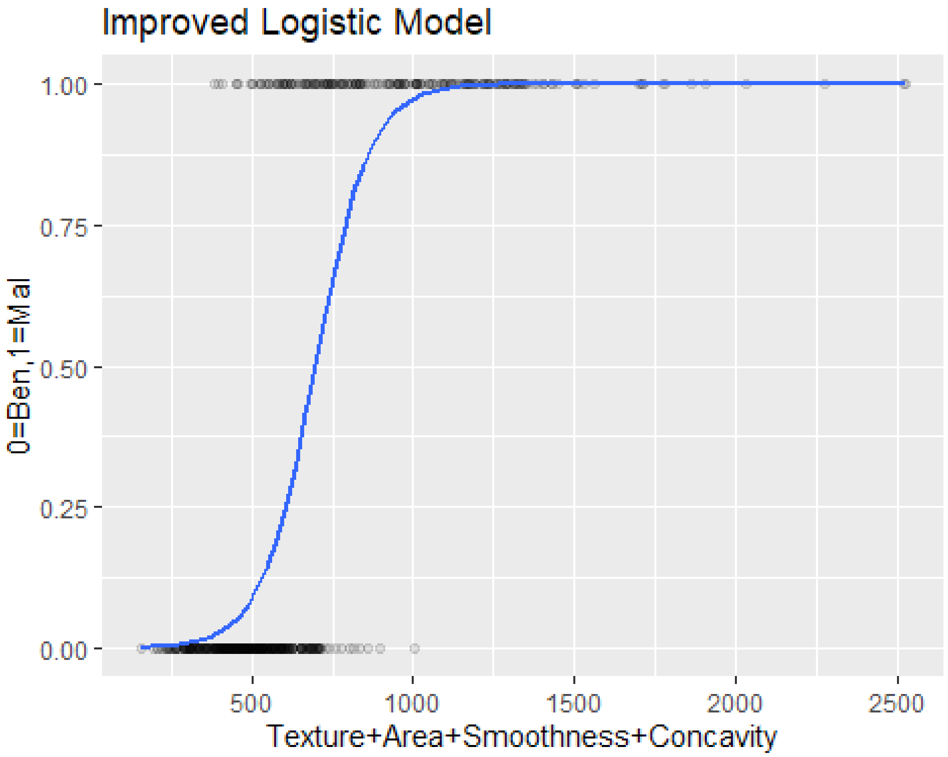
A 1 ((perimeter^2/area)-1) increase in mean compactness decreases the odds of the cell being malignant by a factor of 14.4426.

A 1 severity of contour increase in mean concavity increases the odds of the cell being malignant by a factor of 21.1783.

Now to examine an improved model.

6. Building an Improved Logistic Model

```
# Full model with all predictors:
log_improved <- ggplot(cancer,
aes(x=texture_mean+area_mean+smoothness_mean+concavity_mean, y=malignant)) +
geom_point(alpha=.1) +
geom_smooth(method = "glm", method.args = list(family = "binomial"), se =
FALSE) +
labs(title = "Improved Logistic Model", y = "0=Ben,1=Mal", x=
"Texture+Area+Smoothness+Concavity")

log_improved

## `geom_smooth()` using formula 'y ~ x'
```



```
reglog_improved <- glm(malignant~
texture_mean+area_mean+smoothness_mean+concavity_mean, family=binomial,
cancer)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
# Summary table of the full model
summary(reglog_improved)

##
## Call:
## glm(formula = malignant ~ texture_mean + area_mean + smoothness_mean +
##     concavity_mean, family = binomial, data = cancer)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.86649  -0.15162  -0.03710   0.01339   3.16165
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -30.887464   3.929528  -7.860 3.83e-15 ***
## texture_mean      0.381711   0.062645   6.093 1.11e-09 ***
## area_mean         0.015166   0.001932   7.849 4.19e-15 ***
## smoothness_mean 119.515647  21.141614   5.653 1.58e-08 ***
## concavity_mean   19.392935   3.876045   5.003 5.64e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 751.44  on 568  degrees of freedom
## Residual deviance: 158.11  on 564  degrees of freedom
## AIC: 168.11
##
## Number of Fisher Scoring iterations: 8

# Interpretation of slope coefficients in terms of odds
exp(coefficients(reglog_improved))

##     (Intercept)    texture_mean       area_mean smoothness_mean
concavity_mean
##    3.852519e-14    1.464788e+00    1.015281e+00    8.034999e+51
2.643898e+08
```

The equation of this model is: $\log(\pi/(1-\pi)) = -30.8875 + 0.3817(\text{texture}) + 0.0152(\text{area}) + 119.5156(\text{smoothness}) + 19.3929(\text{concavity})$

All the predictors are very significant at the a=.15 level.

To interpret the new odds, after holding all other predictors constant:

A 1 stdev(grayscale) increase in mean texture increases the odds of the cell being malignant by a factor of 1.4648.

A 1 µm^2 increase in mean area of the cell increases the odds of the cell being malignant by a factor of 1.0153.

A 1 var(radius length) increase in mean smoothness variance increases the odds of the cell being malignant by a factor of 8.035e+51.

A 1 severity of contour increase in mean concavity increases the odds of the cell being malignant by a factor of 2.644e+8.

Notably, the interpretation of odds for each individual have increased in magnitude significantly.

##7. Statistical Model Quality

```
reglog_full$deviance

## [1] 155.9614

reglog_improved$deviance

## [1] 158.1136

logLik(reglog_full)

## 'log Lik.' -77.98072 (df=8)

logLik(reglog_improved)

## 'log Lik.' -79.05678 (df=5)

AIC(reglog_full)

## [1] 171.9614

AIC(reglog_improved)

## [1] 168.1136

pseudoR2full <- 1 - reglog_full$deviance/reglog_full$null.deviance
pseudoR2full

## [1] 0.7924499

pseudoR2improved <- 1 -
reglog_improved$deviance/reglog_improved$null.deviance
pseudoR2improved

## [1] 0.7895859
```

Transitioning from the full model to the improved one, there seems to be a significant increase in AIC (171.9614 to 168.1136), at moderate costs to both deviance (155.9614 to 158.1136) and log likelihood (-77.9807 to -79.0568), with only minimal impact on the pseudo R^2 value (0.7924 to 0.7896).

Overall, the improved model appears to be a significant improvement as it factors in the most significant variables in determining the logistic regression model, reducing the risk of overfitting to the noise in the data. In addition, many of the highly correlated predictors have also been removed as a result of the high p values from before.