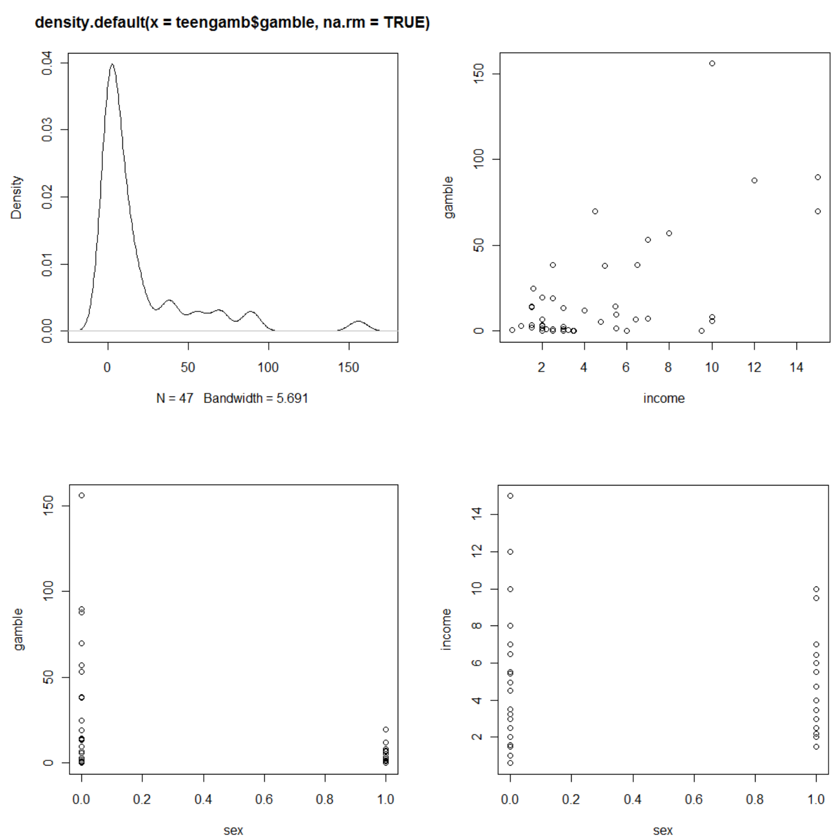
STAT 425 - Homework #1

PROBLEM 4

a) Numerical Summary

```
      sex             status          income           verbal
 Min.   :0.0000   Min.   :18.00   Min.   : 0.600   Min.   : 1.00
 1st Qu.:0.0000   1st Qu.:28.00   1st Qu.: 2.000   1st Qu.: 6.00
 Median :0.0000   Median :43.00   Median : 3.250   Median : 7.00
 Mean   :0.4043   Mean   :45.23   Mean   : 4.642   Mean   : 6.66
 3rd Qu.:1.0000   3rd Qu.:61.50   3rd Qu.: 6.210   3rd Qu.: 8.00
 Max.   :1.0000   Max.   :75.00   Max.   :15.000   Max.   :10.00
     gamble
 Min.   :  0.0
 1st Qu.:  1.1
 Median :  6.0
 Mean   : 19.3
 3rd Qu.: 19.4
 Max.   :156.0
```

b) Graphical Summary



c) Comments

- Looking at the numerical summary, the feature that stands out is the max in gamble, which is much higher than the mean and median
- The graphical summary corroborates this. Most of the expenditure in gamble is centered in the lower end, in which the frequency of observations decreases as the expenditure increases.
- Looking at the gamble vs income plot, the data might suggest a positive correlation as one would expect.
- The plot gamble vs sex shows an interesting feature: the expenditure in gamble is most dispersed (wider range) for men, and the high extremes values are also from men.
- This can be partially explained by the income vs sex plot, in which the range in income is greater for men than for women.

PROBLEM 5

a) Percentage of variation in the response explained by these predictors

```
Call:
lm(formula = gamble ~ ., data = teengamb)

Residuals:
    Min      1Q  Median      3Q     Max 
-51.082 -11.320  -1.451   9.452  94.252 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  22.55565   17.19680   1.312   0.1968    
sex         -22.11833    8.21111  -2.694   0.0101 *  
status        0.05223    0.28111   0.186   0.8535    
income        4.96198    1.02539   4.839 1.79e-05 ***
verbal       -2.95949    2.17215  -1.362   0.1803    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 22.69 on 42 degrees of freedom
Multiple R-squared: 0.5267,	Adjusted R-squared: 0.4816 
F-statistic: 11.69 on 4 and 42 DF,  p-value: 1.815e-06
```

- From the summary output: $R^2 = 0.5267$

b) Case number that corresponds to the highest positive residual

```
> sort(residual)
         39          18          23          27          17          30 
-51.0824078 -27.7998544 -27.2711657 -25.8747696 -25.2627227 -19.8090866 
          4          21          20          44          35           8 
-17.4957487 -16.0041386 -15.9510624 -14.8940753 -14.4016736 -12.3060734 
         10          22          26          28           7          29 
-10.3329505  -9.5801478  -9.1670510  -8.7455549  -7.0242994  -6.8803097 
         43          42          34          12           6          41 
 -4.3831786  -3.8361619  -3.5932770  -3.0958161  -2.9846919  -1.4513921 
         13          25          46          11          15          45 
  0.1172839   0.6993361   1.4092321   1.5934936   2.8488167   5.4506347 
          3           9          47          40           2          14 
  5.4630298   6.8496267   7.1662399   8.8669438   9.3711318   9.5331344 
          1          31          38          33          19          32 
 10.6507430  10.8793766  11.2429290  11.7462296  13.1446553  15.0599340 
         16          37           5          36          24 
 17.2107726  20.5472529  29.5194692  45.6051264  94.2522174 
> order(residual)
 [1] 39 18 23 27 17 30  4 21 20 44 35  8 10 22 26 28  7 29 43 42 34 12  6 4
[25] 13 25 46 11 15 45  3  9 47 40  2 14  1 31 38 33 19 32 16 37  5 36 24
```

- From the output: CASE NUMBER = 24 (Residual value = 94.252)

c) Mean and Median of the residuals

```
> mean(residual)
[1] -2.485822e-17
> median(residual)
[1] -1.451392
```

- $Mean(\varepsilon) \approx 0$
- $Median(\varepsilon) \approx -1.451$

d) Correlation of the residuals with the fitted values.

```
> cor(residual, gamble.full.lm$fit)
[1] 2.586181e-17
```

- $Correlation(residuals, fitted\ values) \approx 0.$

e) Correlation of the residuals with income.

```
> cor(residual, teengamb)
                 sex        status        income        verbal    gamble
[1,] -1.622003e-17 -1.496831e-18 -5.02741e-17 -1.067558e-17 0.687951
```

- $Correlation(residuals, income) \approx 0.$

f) Predicted expenditure on gamble for a male compared to a female when all other predictors are held constant

```
> gamble.full.lm$coeff
 (Intercept)          sex       status       income       verbal
 22.55565063 -22.11833009   0.05223384   4.96197922  -2.95949350
```

- $Coefficient(sex) = -22.118$
- All other things being equal, this is expected difference in response (the predicted expenditure on gambling) per unit difference in the sex predictor.
- Thus, the male is predicted to spend 22.118 pounds per year more than a female (unit change from 0=male to 1=female changes the response by -22.118)

g) Variables statistically significant at the 0.05 level

- Statistically significant at 0.05 level : p-value <0.05 (rejects the null hypothesis)
- Looking at the summary output (a), only sex ($p - value \approx 0.01$) and income ($p - value \approx 0$) have p-values less than 0.05. Thus, sex and income are statistically significant at the 0.05 level for this model

h) Prediction of the amount that a male with average status, income and verbal score that would gamble along with a 95 percent prediction interval.

```
> x=data.frame(sex=0,status=mean(status),
+              income=mean(income),verbal=mean(verbal))
>
> predict.lm(gamble.full.lm,x,level=0.95,interval="prediction")
       fit       lwr      upr
1 28.24252 -18.51536 75.00039
```

- Predicted = 28.243, Lower bound = -18.516, Upper bound =75.00

Prediction for a male with maximal values of status, income and verbal score would gamble along with a 95 percent prediction interval.

```
> y=data.frame(sex=0,status=max(status),
+              income=max(income),verbal=max(verbal))
>
> predict.lm(gamble.full.lm,y,level=0.95,interval="prediction")
       fit      lwr    upr
1 71.30794 17.06588 125.55
```

- Predicted = 71.308, Lower bound = 17.066, Upper bound =125.55
- The prediction interval is wider for maximum values and narrower for the mean values because most of the data points are around the mean, which means the model will have a higher confidence predicting around the mean.

i) F-test comparison of full model and model with just income as predictor

```
> gamble.income.lm=lm(gamble~income, data=teengamb)
> summary(gamble.income.lm)

Call:
lm(formula = gamble ~ income, data = teengamb)

Residuals:
    Min      1Q  Median      3Q     Max
-46.020 -11.874  -3.757  11.934 107.120

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -6.325      6.030  -1.049      0.3
income         5.520      1.036   5.330 3.05e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24.95 on 45 degrees of freedom
Multiple R-squared: 0.387,      Adjusted R-squared: 0.3734
F-statistic: 28.41 on 1 and 45 DF,  p-value: 3.045e-06

> anova(gamble.income.lm, gamble.full.lm)
Analysis of Variance Table

Model 1: gamble ~ income
Model 2: gamble ~ sex + status + income + verbal
  Res.Df   RSS Df Sum of Sq      F  Pr(>F)
1     45 28009
2     42 21624  3    6384.8 4.1338 0.01177 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Since the Pr(>F) = 0.01177 is less than 0.05, reject the null hypothesis at a 0.05 level
- The small P-value tells us that, assuming Model 1 is correct, the probability of randomly obtaining the data that fits model 2 much better is really small.
- Thus, reject Model 1 (reduced) in favor of the significantly better Model 2 (full)
- (Note: if p-value were not small, then there would not be evidence supporting the full/more complex model 2, so accept the reduced/simpler model (model 1).

PROBLEM 6

a) Fit a model with lpsa as the response and lcavol as the predictor. Report the residual standard error.

```
> model1=lm(lpsa~lcavol, data=prostate)
> summary(model1)

Call:
lm(formula = lpsa ~ lcavol, data = prostate)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.6762 -0.4165  0.0986  0.5071  1.8967 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.50730    0.12194   12.36   <2e-16 ***
lcavol       0.71932    0.06819   10.55   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 0.7875 on 95 degrees of freedom
Multiple R-squared: 0.5394,     Adjusted R-squared: 0.5346 
F-statistic: 111.3 on 1 and 95 DF,  p-value: < 2.2e-16
```

- Residual Standard Error = 0.7875

b) Report R^2

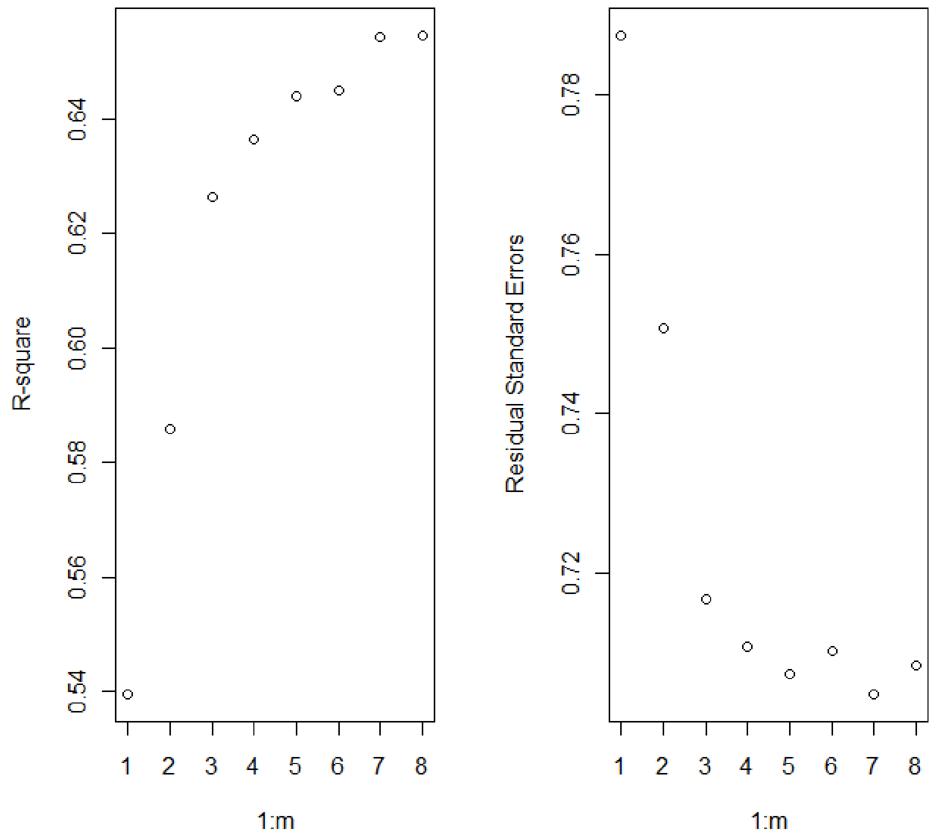
- R^2 =0.5394

c) Add variables one at a time to the model and record the residual standard error and R^2 for each model. Graph and comment trends

- Recorded Residual standard error (mysd) and R^2 (myR2) for each model. See code for details on how this was obtained.

```
> varlist
[1] "lcavol"  "lweight" "svi"     "lbph"    "age"     "lcp"     "pgg45"  
[8] "gleason"
> mysd
[1] 0.7874994 0.7506469 0.7168094 0.7108232 0.7073054 0.7102135 0.7047533
[8] 0.7084155
> myR2
[1] 0.5394319 0.5859345 0.6264403 0.6366035 0.6441024 0.6451130 0.6544317
[8] 0.6547541
```

- Graphs (see code for details)



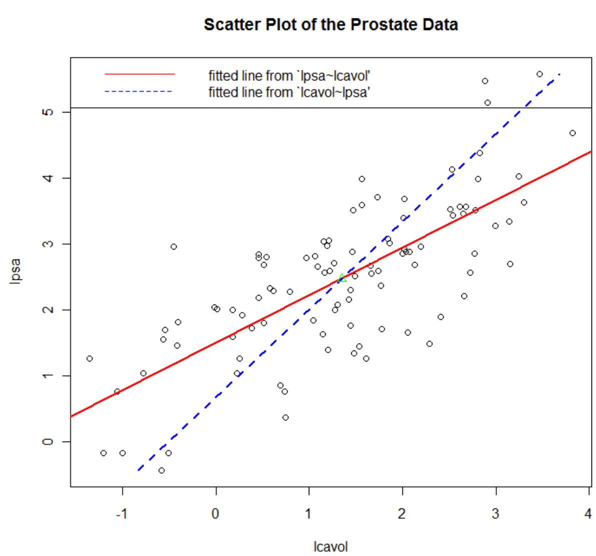
- Comments:
  - The Residual Standard Errors seem to decrease as the number of variables (m) in the model increases. In general, this does not have to be true.
  - The R^2 seems to increase as the number of variables (m) in the models increases. This is expected since R^2 is " the proportion of variability in a data set that is accounted for by the statistical model". Therefore, as more predictors are added to the model, R^2 always increases.

d) Fit the regressions of lpsa on lcavol, and lcavol on lpsa. Display both regression on the scatter plot of lpsa (y-coordinate) against lcavol (x-coordinate)

- Regression fits

```
> summary(prostate.lm1)

Call:
lm(formula = lpsa ~ lcavol, data = prostate)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.6762 -0.4165  0.0986  0.5071  1.8967 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.50730    0.12194   12.36   <2e-16 ***
lcavol       0.71932    0.06819   10.55   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7875 on 95 degrees of freedom
Multiple R-squared: 0.5394,	Adjusted R-squared: 0.5346 
F-statistic: 111.3 on 1 and 95 DF,  p-value: < 2.2e-16 

> summary(prostate.lm2)

Call:
lm(formula = lcavol ~ lpsa, data = prostate)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.15948 -0.59383  0.05034  0.50826  1.67751 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.50858    0.19419  -2.619   0.0103 *  
lpsa         0.74992    0.07109  10.548   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8041 on 95 degrees of freedom
Multiple R-squared: 0.5394,	Adjusted R-squared: 0.5346 
F-statistic: 111.3 on 1 and 95 DF,  p-value: < 2.2e-16 
```

- Graph



e) Intersection point of the two lines

- The lines intersect at the means of the variables (x,y)=(mean(lcavol),mean(lpsa))
- This is because the line of the simple regression model contains the point (x,y)=(mean(predictor),mean(response)). See question 1 c) for the proof.

- Calculation:

```
> mean(lcavol)
[1] 1.350010
> predict(prostate.lm1,data.frame(lcavol=mean(lcavol)))
       1
2.478387
> mean(lpsa)
[1] 2.478387
> predict(prostate.lm2,data.frame(lpsa=mean(lpsa)))
       1
1.350010
```

PROBLEM 7

a) Test the hypothesis that $\beta_{salary} = 0$

```
> sat.lmodel1=lm(total~expend+ratio+salary, data=sat)
> summary(sat.lmodel1)

Call:
lm(formula = total ~ expend + ratio + salary, data = sat)

Residuals:
     Min       1Q   Median       3Q      Max 
-140.911  -46.740   -7.535   47.966  123.329 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1069.234    110.925   9.639 1.29e-12 ***
expend        16.469     22.050   0.747   0.4589    
ratio          6.330      6.542   0.968   0.3383    
salary        -8.823      4.697  -1.878   0.0667 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 68.65 on 46 degrees of freedom
Multiple R-squared: 0.2096,     Adjusted R-squared: 0.1581 
F-statistic: 4.066 on 3 and 46 DF,  p-value: 0.01209 
```

- Looking at the table from the summary of the coefficients, we see p-value=0.0667 for the salary coefficient.
- Since p-value= 0.0667 > α =0.05 , no evidence to reject the null $\beta_{salary} = 0$ at α =0.05

b) Test the hypothesis that $\beta_{salary} = \beta_{ratio} = \beta_{expend} = 0$

- Looking at the table from the summary of the coefficients, we see p-value=0.01209 for for the overall model.
- Since p-value= 0.01209 < α =0.05 , reject the null $\beta_{salary} = \beta_{ratio} = \beta_{expend} = 0$ at α =0.05

c) Test the hypothesis that $\beta_{takers} = 0$

```
> sat.lmodel2=lm(total~expend+ratio+salary+takers, data=sat)
> summary(sat.lmodel2)

Call:
lm(formula = total ~ expend + ratio + salary + takers, data = sat)

Residuals:
    Min      1Q  Median      3Q     Max 
-90.531 -20.855  -1.746  15.979  66.571 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1045.9715    52.8698  19.784  < 2e-16 ***
expend         4.4626    10.5465   0.423    0.674    
ratio         -3.6242     3.2154  -1.127    0.266    
salary         1.6379     2.3872   0.686    0.496    
takers        -2.9045     0.2313 -12.559 2.61e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 32.7 on 45 degrees of freedom
Multiple R-squared: 0.8246,     Adjusted R-squared: 0.809 
F-statistic: 52.88 on 4 and 45 DF,  p-value: < 2.2e-16 
```

- Looking at the table from the summary of the coefficients, we see p-value$\approx$ 0 for the takers coefficient.
- Since p-value$\approx$ 0 < α =0.05 , reject the null $\beta_{takers} = 0$. Coefficient is statistically significant in the model with salary, ratio, expend and takers as predictors.

d) Model comparison ((c) vs original model) using F-test

```
> anova(sat.lmodel1, sat.lmodel2)
Analysis of Variance Table

Model 1: total ~ expend + ratio + salary
Model 2: total ~ expend + ratio + salary + takers
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     46 216812
2     45  48124  1    168688 157.74 2.607e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Looking at the ANOVA, we see p-value$\approx$ 0 (Fcalc=157.74 > Fcrit)
- The small P-value tells us that, assuming Model 1 is correct, the probability of randomly obtaining the data that fits model 2 much better is really small.
- Thus, reject Model 1 (original with 3 predictors) in favor of the significantly better Model 2 (with the addition of takers as predictor)

e) Show numerically $F = t^2$ (t-statistic in c, F-statistic in d)

t-statistic: $t_{takers} = \frac{\hat{\beta}_{takers}}{se(\hat{\beta}_{takers})}$

F-statistic: $Fstat = (RSS_1 - RSS_2)/(RSS_2/45)$

```
> se.takers=0.2313
> coeff.takers=sat.lmodel2$coef[5]
> t=coeff.takers/se.takers
> tsquared=t^2
>
> rss.sat.lmodel1=sum(sat.lmodel1$res^2)
> rss.sat.lmodel2=sum(sat.lmodel2$res^2)
> Fstat=(rss.sat.lmodel1-rss.sat.lmodel2)/(rss.sat.lmodel2/45)
>
> tsquared
  takers
157.6833
> Fstat
[1] 157.7379
```

- Thus, $F = t^2$= 157.7