# Homework 3, Math455: Due Mon, 02/26/2018

**Your Name: ... (replace this)**

March 5, 2018

**Instructions**: The homework assignment editing this LaTeX document. Download the LaTeX source from the class web page and study it to learn more about LaTeX. Replace the text with appropriate information. Run "pdflatex" on this document.

You will submit this assignment in two parts:

1. Print out the PDF file and bring it to class, and

2. Send an e-mail to:

   gang@math.binghamton.edu

   *before class* on the due date with two attachments:

   - The LaTeX source file, and
   - The generated PDF document.

Please complete the following:

1. Finish R exercises 1-5 on page 12 of the textbook. (exercises from chapter 1). Choose 1 out these 5 exercises to submit as your homework.

   Solution: put your solution here.

   After some exploration, I explored how gender played a role into the gambling data. I split them into two groups.

```
males = teengamb[teengamb$sex == 0,]
females = teengamb[teengamb$sex == 1,]
```

   I then checked their mean and variance to see how they varied.

```
mmean = weighted.mean(males$gamble)
fmean = weighted.mean(females$gamble)
mvar = var(males$gamble)
fvar = var(females$gamble)
> mmean
[1] 29.775
> fmean
[1] 3.865789
> mvar
[1] 1393.095
> fvar
[1] 26.53001
```

   I found that males had a higher variance and higher mean expenditure gambling. I plotted the data sets based on gender to show this effect. The graph is below.

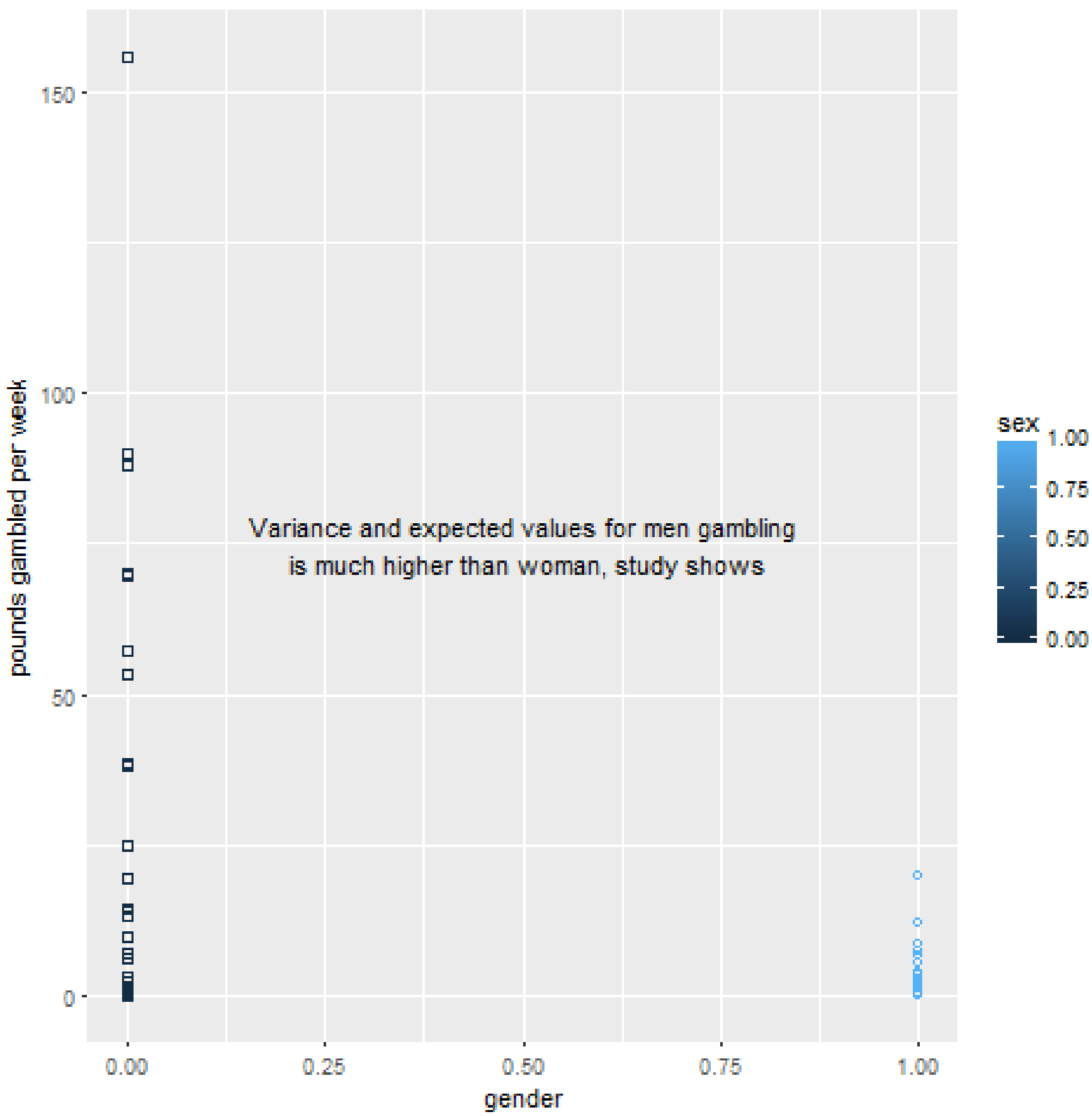


Figure 1: plot of gambling amounts split by gender

2. Finish R exercises 1,2,4,6 on page 30-31 of the textbook. (exercises from chapter 2). Submit your answers for ALL questions.

   1

   (a)
```
> lTG<-lm(gamble ~ sex + status + income + verbal, teengamb)
> summary(lTG)


Call:
```

```
lm(formula = gamble ~ sex + status + income + verbal, 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
```

$R^2 = .53$, so 53 percent of the model is explained by the predictors

(b)
```
> rTG = residuals(lTG)
> mTG = max(rTG)
> which(rTG==mTG)
[1] 94.25222
24
```

Thus, case 24 has the highest residual with a residual of 94.252

(c)
```
> weighted.mean(rTG)
[1] -3.07083e-17
```

mean   -3.07e-17, median = -1.451

(d)
```
> var(rTG,fTG)
```

```
[1] -5.309559e-14
```

Cov(Residuals,Fitted Values) = −5.3096e-14

(e) 
```
> var(rTG,teengamb$income)
[1] -5.576533e-15
```

Cov(Residuals,income) = −5.577e-15

(f) 
```
> print(mmean-fmean)
[1] 25.90921
```

MaleMean - FemaleMean = 25.90921 pounds

2

(a) 
```
> usModel <- lm(wage~educ+exper,uswages)
> usLModel <- lm(log(wage)~educ+exper,uswages)
> summary(usModel)

Call:
lm(formula = wage ~ educ + exper, data = uswages)

Residuals:
Min      1Q  Median      3Q     Max
-1018.2  -237.9   -50.9   149.9  7228.6

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -242.7994    50.6816  -4.791 1.78e-06 ***
educ          51.1753     3.3419  15.313  < 2e-16 ***
exper          9.7748     0.7506  13.023  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 427.9 on 1997 degrees of freedom
```

```
Multiple R-squared:  0.1351,Adjusted R-squared:  0.1343
F-statistic:   156 on 2 and 1997 DF,  p-value: < 2.2e-16

> summary(usLModel)

Call:
lm(formula = log(wage) ~ educ + exper, data = uswages)

Residuals:
Min      1Q  Median      3Q     Max
-2.7533 -0.3495  0.1068  0.4381  3.5699

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.650319   0.078354   59.35   <2e-16 ***
educ        0.090506   0.005167   17.52   <2e-16 ***
exper       0.018079   0.001160   15.58   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.6615 on 1997 degrees of freedom
Multiple R-squared:  0.1749,Adjusted R-squared:  0.174
F-statistic: 211.6 on 2 and 1997 DF,  p-value: < 2.2e-16
```

The t value is quite high for the education coefficient and thus we can trust it. Since the coefficient is at 51.17 approximately we can say that for every year of education they make, they get a boost, on average, of 51.17 dollars per week.

(b) If we take the log of weekly wages, we have a much reduced weekly wage for each case. Thus, the coefficient, now measured at approximately 0.091 is the additional logged pay per week added per year of education

4

3.

```
> prostateModel = lm(lpsa~lcavol,prostate)
> tempS=summary(prostateModel)
> rVals = double()
> rSE = double()
> for( nam in vars)
+ {
+   prostateModel = update(prostateModel, as.formula(paste('~ . +', nam)))
+   tempS = (summary(prostateModel))
+   print(tempS$r.squared)
+   rVals=append(rVals, c(tempS$r.squared))
+   rSE=append(rSE,tempS$sigma)
+
+ }
[1] 0.5394319
[1] 0.5859345
[1] 0.5892177
[1] 0.597575
[1] 0.6441024
[1] 0.645113
[1] 0.650644
[1] 0.6547541
[1] 0.6547541
Warning messages:
1: In model.matrix.default(mt, mf, contrasts) :
the response appeared on the right-hand side and was dropped
2: In model.matrix.default(mt, mf, contrasts) :
problem with term 9 in model.matrix: no columns are assigned
> #need to remove the last elements
> rVals = head(rVals,-1)
> rSE = head(rSE,-1)
> png("C:/Users/alexander/Documents/GitHub/regressions/rVals.png")
```

```
> qplot(1:8,rVals)
> dev.off()
> png("C:/Users/alexander/Documents/GitHub/regressions/rSE.png")
> qplot(1:8,rSE)
> dev.off()
null device
1
```

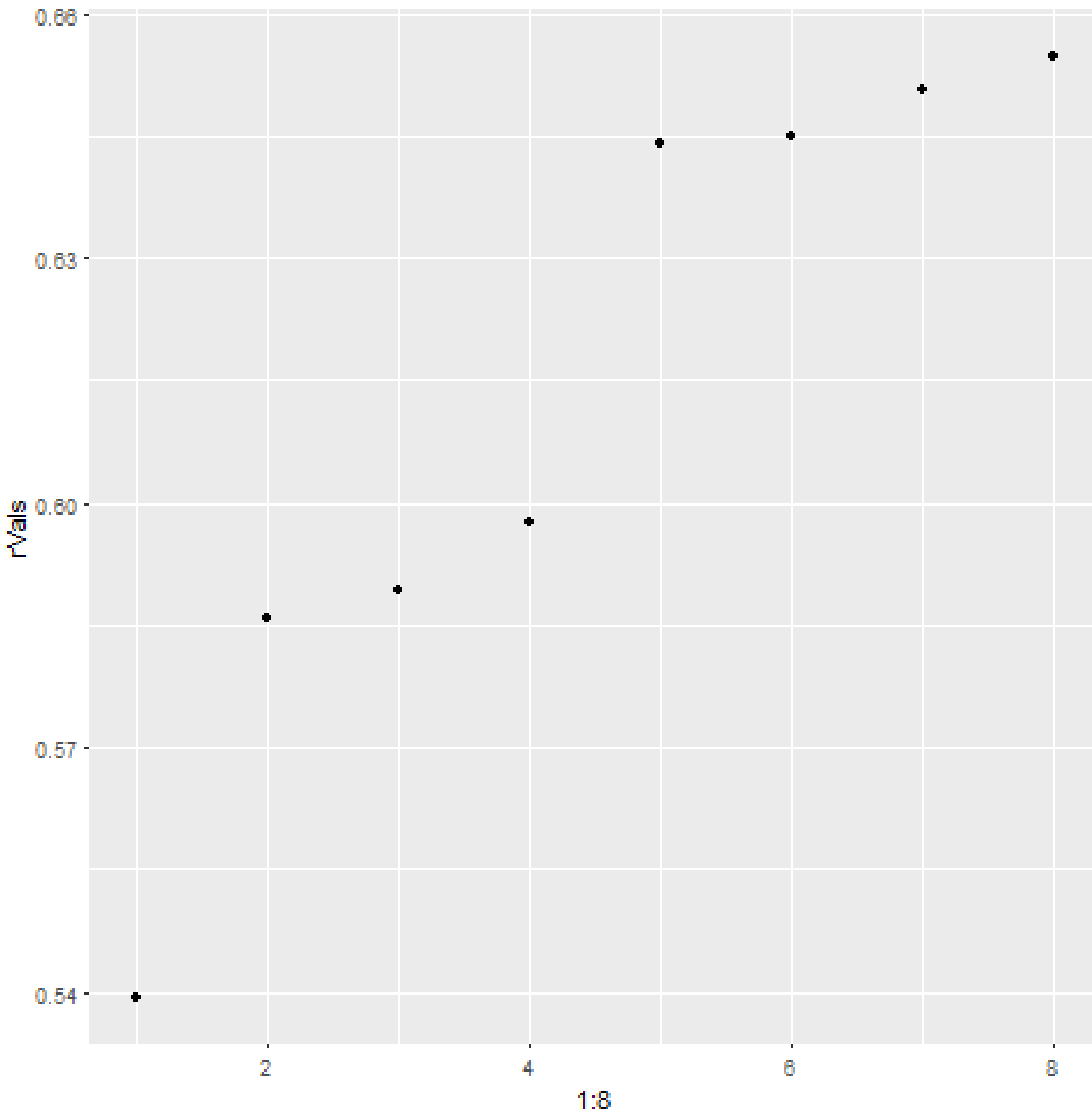


Figure 2: R squared value over each element being added

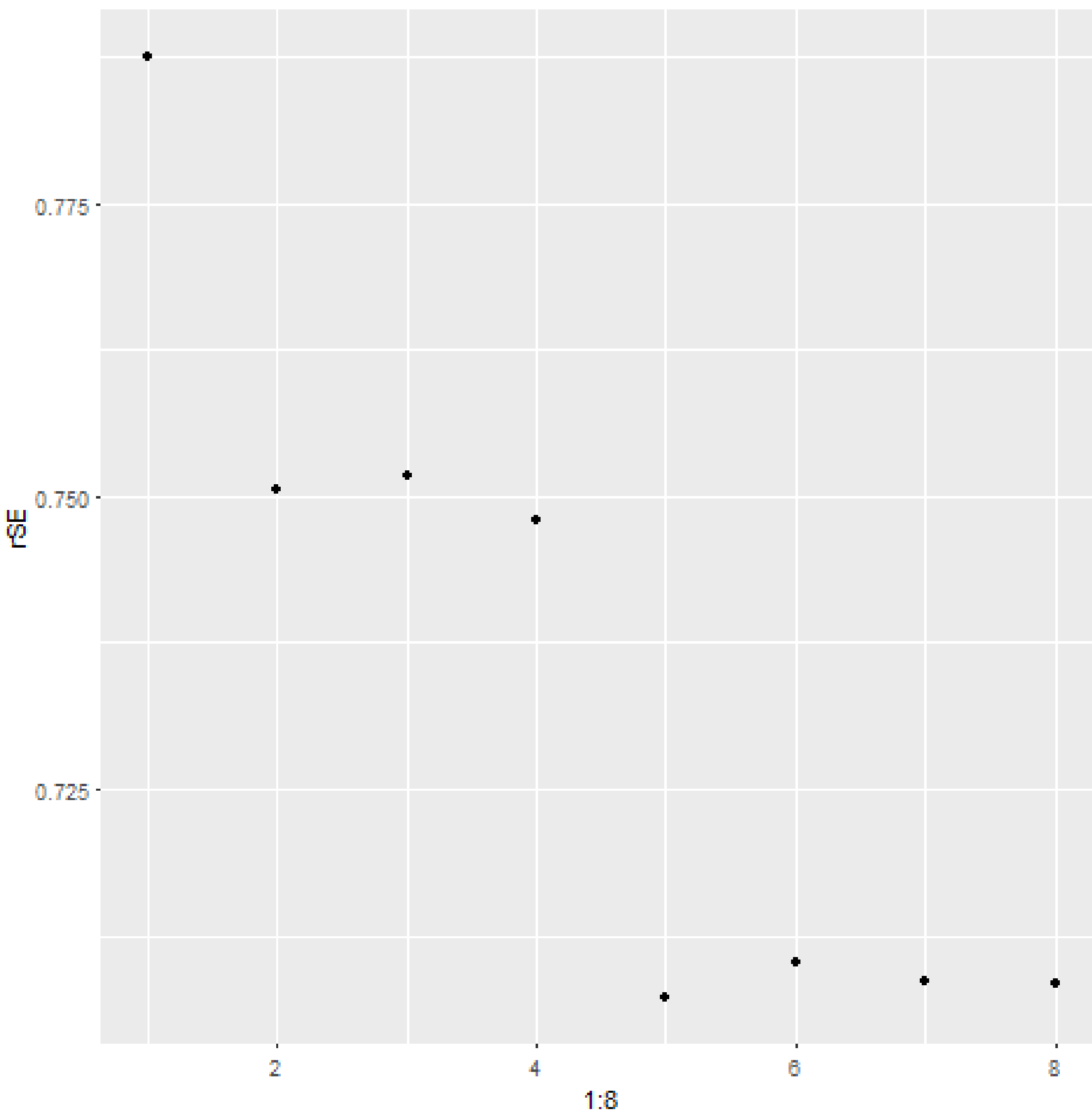


Figure 3: residual squared error

4. 6

(a)
```
> cheeseModel = lm(taste~Acetic + H2S + Lactic, cheddar)
> summary(cheeseModel)

Call:
lm(formula = taste ~ Acetic + H2S + Lactic, data = cheddar)
```

```
Residuals:
Min      1Q  Median      3Q     Max
-17.390  -6.612  -1.009   4.908  25.449

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -28.8768    19.7354  -1.463  0.15540
Acetic        0.3277     4.4598   0.073  0.94198
H2S           3.9118     1.2484   3.133  0.00425 **
Lactic       19.6705     8.6291   2.280  0.03108 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 10.13 on 26 degrees of freedom
Multiple R-squared:  0.6518,Adjusted R-squared:  0.6116
F-statistic: 16.22 on 3 and 26 DF,  p-value: 3.81e-06
```

intercept -29, acetic 0.3277, H2S 3.9, Lactic 19.7

(b)
```
> cheeseFit = fitted(cheeseModel)
> actual = cheddar$taste
> corrCheese = cor(actual,cheeseFit)
> print(corrCheese^2)
[1] 0.6517747
```

we get a value of .6518 which is the multiple R Squared found in the summary

(c)
```
> cheeseModelNoInt = lm(taste~Acetic + H2S + Lactic+0, cheddar)
> summary(cheeseModelNoInt)

Call:
lm(formula = taste ~ Acetic + H2S + Lactic + 0, data = cheddar)

Residuals:
```

```
Min       1Q   Median       3Q      Max
-15.4521  -6.5262  -0.6388   4.6811  28.4744

Coefficients:
Estimate Std. Error t value Pr(>|t|)
Acetic   -5.454      2.111  -2.583  0.01553 *
H2S       4.576      1.187   3.854  0.00065 ***
Lactic   19.127      8.801   2.173  0.03871 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 10.34 on 27 degrees of freedom
Multiple R-squared:  0.8877,Adjusted R-squared:  0.8752
F-statistic: 71.15 on 3 and 27 DF,  p-value: 6.099e-13

> CF = fitted(cheeseModelNoInt)
> print(cor(actual,CF)^2)
[1] 0.6244075
```

R squared is now .8877, much higher than the previous version, using corr squared we get .6244 which makes more sense

(d)
```
> qrc = qr(cheeseModel)
> qrCC = t(qr.Q(qrc)) %*% actual
> backsolve(qr.R(qrc),qrCC)
[,1]
[1,] -28.8767696
[2,]   0.3277413
[3,]   3.9118411
[4,]  19.6705434
```