Linear Regression - Student Survey

Demond Love

Using statistical correlation and R programming, I will analyze the results of a survey recently given to college students. Research to investigage: "Is there a significant relationship between the amount of time spent reading and the time spent watching television?"

```
library(ggplot2)
library(readr)
library(pastecs)
library(ggm)
setwd('/Users/Love/Documents/Projects')
file = read_csv('./student-survey.csv')
```

Calculating the covariance of the Survey variables.

```
cov(file)
```

```
##              TimeReading       TimeTV  Happiness      Gender
## TimeReading   3.05454545 -20.36363636 -10.350091 -0.08181818
## TimeTV      -20.36363636 174.09090909 114.377273  0.04545455
## Happiness   -10.35009091 114.37727273 185.451422  1.11663636
## Gender       -0.08181818   0.04545455   1.116636  0.27272727
```

Covariance is a measure of the 'average' relationship between two variables. It is the average cross-product deviation. If we assume that two variables are related, then we can use covariance to measure the validity of our expectation; since when one variable deviates from its mean, then we would expect the other variable to deviate from its mean in a similar way.

Examining the Survey data variables.

TimeReading is measuring the amount of time a student is spending reading, with a unit of measurement of hours. TimeTV is measiring the amount of time a student is spending watching TV, with a unit of measurement of minutes. Happiness is measuring the amount of happiness each student feels, on a scale of 1 to 100. Gender is a binary variable, with 1 representing one gender and 0 representing the other.

There are a few issues with the units of measurement being used. First, that TimeReading are all whole numbers, so these are presumably being rounded to the nearest integer. Whereas TimeTV are not represented in multiples of 60. Lastly, it is unclear which gender is assigned to which binary variable. Therefore, we will not be able to do any meaningful analysis comparing males to females on these dimensions.

The obvious transformation would be to convert TimeReading to minutes, and rerunning the covariance calculations.

```
transfile = data.frame(file$TimeReading*60, file$TimeTV, file$Happiness, file$Gender)
names(transfile) = c("TimeReading", "TimeTV", "Happiness", "Gender")

transfile
```

```
##    TimeReading TimeTV Happiness Gender
## 1           60     90     86.20      1
## 2          120     95     88.70      0
## 3          120     85     70.17      0
## 4          120     80     61.31      1
## 5          180     75     89.52      1
## 6          240     70     60.50      1
## 7          240     75     81.46      0
```

```
## 8           300     60     75.92      1
## 9           300     65     69.37      0
## 10          360     50     45.67      0
## 11          360     70     77.56      1
```

```r
cov(transfile)
```

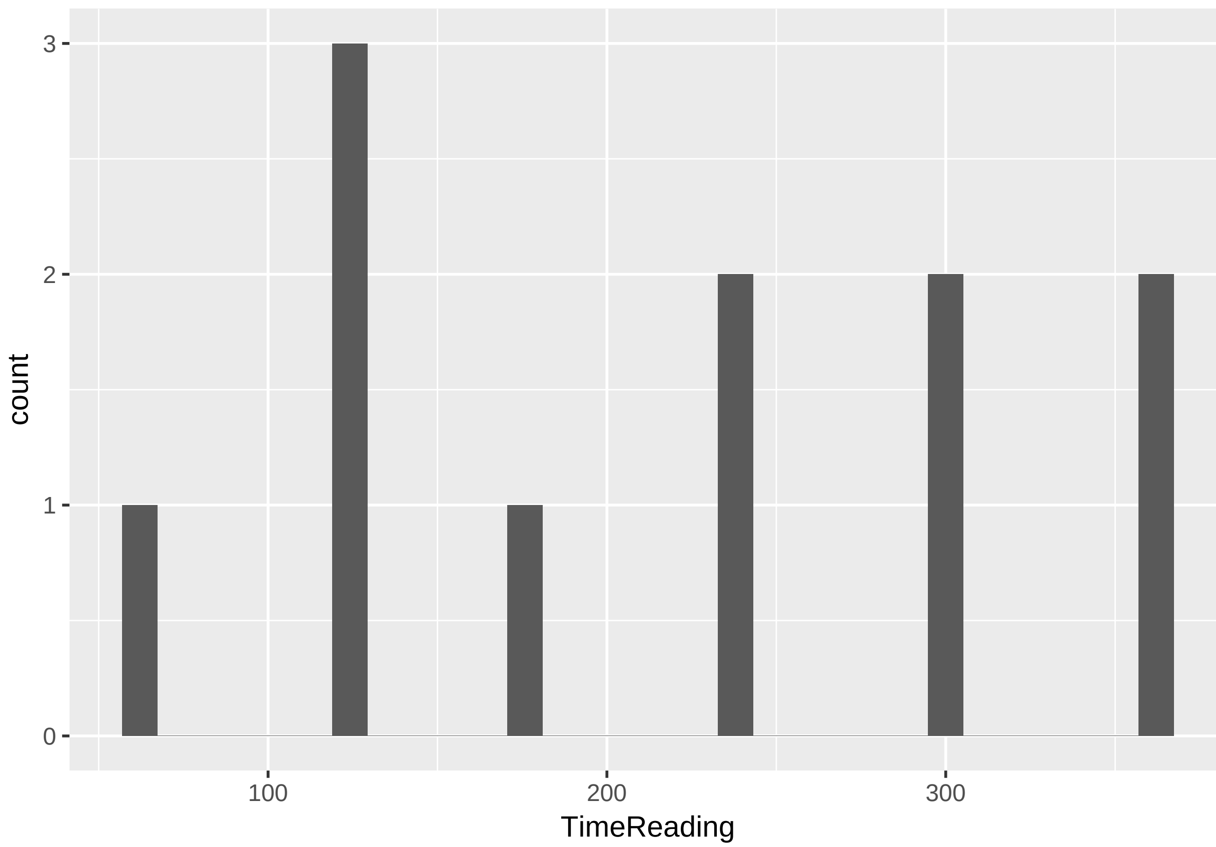
```
##               TimeReading        TimeTV   Happiness      Gender
## TimeReading 10996.363636 -1.221818e+03 -621.005455 -4.90909091
## TimeTV      -1221.818182  1.740909e+02  114.377273  0.04545455
## Happiness    -621.005455  1.143773e+02  185.451422  1.11663636
## Gender         -4.909091  4.545455e-02    1.116636  0.27272727
```

Here, you can see that the covariance values have changed. This does expose a problem with covariance. It depends upon the scales of measurement used, which is why covariance is not a standardized measure. Therefore, this second measurement is a more accuate measure of the relationship of TimeTv and TimeReading.

Investigating the correlation of the variables.

```r
ggplot(transfile, aes(TimeReading)) + geom_histogram()
```

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



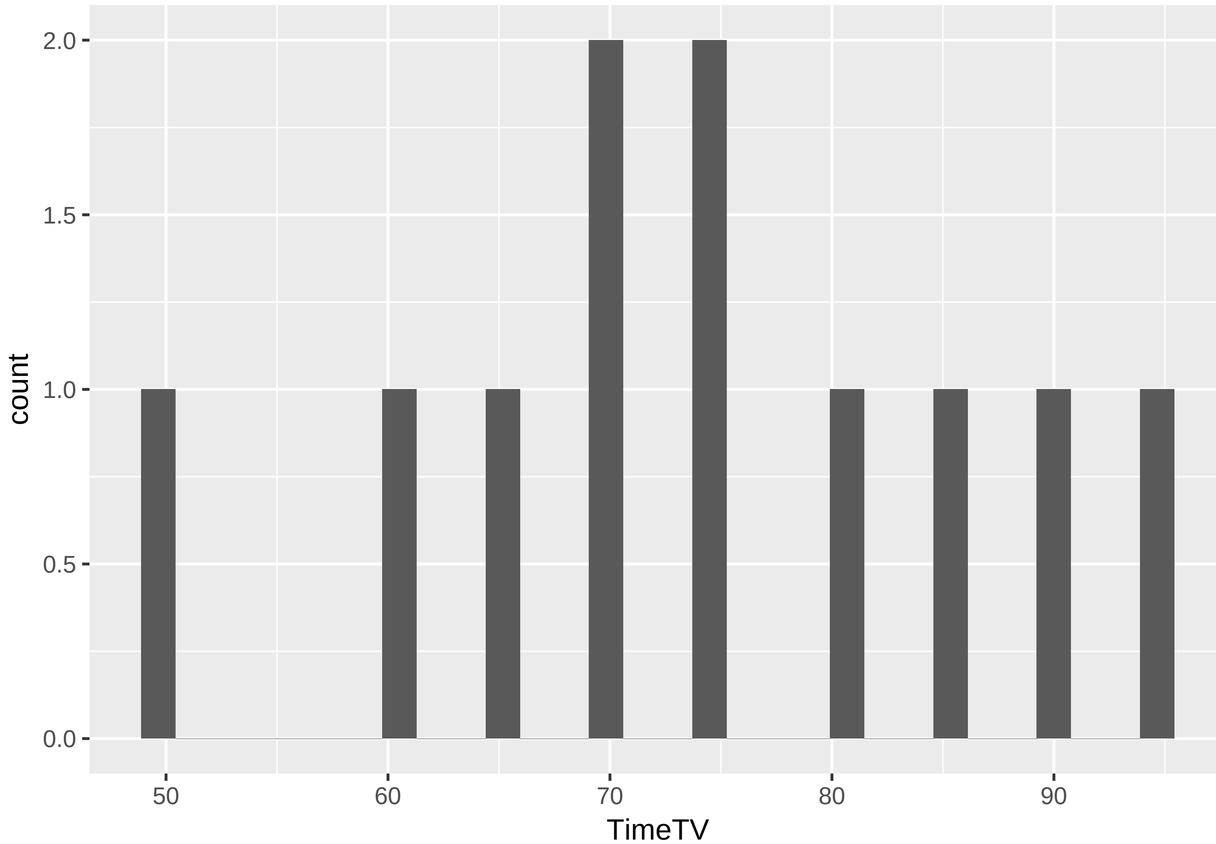

```r
round(stat.desc(transfile$TimeReading, basic = FALSE, norm = TRUE), digits = 3)
```

```
##       median         mean      SE.mean CI.mean.0.95          var
##      240.000      218.182       31.618       70.448    10996.364
##      std.dev     coef.var     skewness     skew.2SE     kurtosis
##      104.864        0.481       -0.003       -0.002       -1.642
##     kurt.2SE   normtest.W   normtest.p
```

```
##       -0.642        0.921        0.326
```

```r
ggplot(transfile, aes(TimeTV)) + geom_histogram()
```

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




```r
round(stat.desc(transfile$TimeTV, basic = FALSE, norm = TRUE), digits = 3)
```

```
##       median         mean      SE.mean CI.mean.0.95          var
##       75.000       74.091        3.978        8.864      174.091
##      std.dev     coef.var     skewness     skew.2SE     kurtosis
##       13.194        0.178       -0.118       -0.090       -1.038
##     kurt.2SE   normtest.W   normtest.p
##       -0.406        0.987        0.992
```

I have chosen to use the Pearson correlation coefficient because the underlying data meets the assumptions of Pearson's correlation coefficient. Namely that the variables TimeReading and TimeTV are both normally distributed interval variables. If skew.2SE or kurt.2SE are greater than 1 (ignoring the plus or minus sign) then you have significant skew/kurtosis. However, in neither of the above distributions is this significance level met. Therefore, our z-scores do not dictate that our skewness or kurtosis are statistically significant.

All variables

```r
cor(transfile)
```

```
##             TimeReading       TimeTV  Happiness       Gender
## TimeReading  1.00000000 -0.883067681 -0.4348663 -0.089642146
## TimeTV      -0.88306768  1.000000000  0.6365560  0.006596673
## Happiness   -0.43486633  0.636555986  1.0000000  0.157011838
## Gender      -0.08964215  0.006596673  0.1570118  1.000000000
```

A single correlation between two a pair of the variables

```
cor(transfile$TimeReading, transfile$TimeTV)
```

```
## [1] -0.8830677
```

Repeat with a confidence interval at 99%

```
cor.test(transfile$TimeReading, transfile$TimeTV, method = 'pearson', conf.level = 0.99)
```

```
##
##  Pearson's product-moment correlation
##
## data:  transfile$TimeReading and transfile$TimeTV
## t = -5.6457, df = 9, p-value = 0.0003153
## alternative hypothesis: true correlation is not equal to 0
## 99 percent confidence interval:
##  -0.9801052 -0.4453124
## sample estimates:
##        cor
## -0.8830677
```

Analyzing the calculations.

```
cor(transfile)
```

```
##             TimeReading       TimeTV  Happiness       Gender
## TimeReading  1.00000000 -0.883067681 -0.4348663 -0.089642146
## TimeTV      -0.88306768  1.000000000  0.6365560  0.006596673
## Happiness   -0.43486633  0.636555986  1.0000000  0.157011838
## Gender      -0.08964215  0.006596673  0.1570118  1.000000000
```

The correlation coefficients in the above matrix above can range between -1 and 1, with higher positive numbers meaning a closer relationship between the two variables, lower negative numbers meaning an inverse relationship and numbers near zero meaning no relationship.

There are really 3 figures that are most interesting in the dataset. The relationships of TimeReading and TimeTv, the relationship of TimeReading and Happiness, then lastly the relationship of TimeTV and Happiness.

TimeReading and TimeTv have a high negative value of -.88, which implies a strong inverse relationship. Meaning that the more someone reads, then the less television that they read on average. And vice versa.

Interestingly, TimeReading and Happiness have a moderate negative value of -.43, which implies a moderate inverse relationship. Meaning that the more someone reads, then the less happy that they are on average. And vice versa.

Aso, surprisingly, TimeTV and Happiness have a moderate positive value of .64, which implies a moderate positive relationship. Meaning that the more someone watches TV, then the happier that they are on average. And vice versa.

e. Calculate the correlation coefficient and the coefficient of determination, describe what you conclude about the results.

```
cor(transfile)
```

```
##             TimeReading       TimeTV  Happiness       Gender
## TimeReading  1.00000000 -0.883067681 -0.4348663 -0.089642146
## TimeTV      -0.88306768  1.000000000  0.6365560  0.006596673
## Happiness   -0.43486633  0.636555986  1.0000000  0.157011838
## Gender      -0.08964215  0.006596673  0.1570118  1.000000000
```

```r
cor(transfile)^2*100
```

```
##             TimeReading       TimeTV  Happiness       Gender
## TimeReading 100.0000000  77.98085292  18.910873   0.80357143
## TimeTV       77.9808529 100.00000000  40.520352   0.00435161
## Happiness    18.9108726  40.52035234 100.000000   2.46527174
## Gender        0.8035714   0.00435161   2.465272 100.00000000
```

As opposed to the correlation coefficient, described earlier, the coefficient of determination is a measure of the amount of variablility in one variable that is shared by others.

From this, you can see that 80% of the variance in TimeReading is accounted for by TimeTV. 19% of the variance in Happiness is accounted for by TimeReading. 41% of the variance in Happiness is accounted for by TimeTV.

f. Based on your analysis can you say that watching more TV caused students to read less? Explain.

No, correlation doesn't imply causation.

g. Picking three variables and performing a partial correlation.

I have chosen to perform a partial correlation on TimeReading and Happiness, while controlling for TimeTV.

```r
pc = pcor(c("Happiness", "TimeReading", "TimeTV"), var(transfile))
pc
```

```
## [1] 0.3516355
```

```r
pcor.test(pc, 1, 11)
```

```
## $tval
## [1] 1.062425
##
## $df
## [1] 8
##
## $pvalue
## [1] 0.319059
```

Whereas previously the correlation coefficient between Happiness and TimeReading was -.43; when controlling for the time spent watching TV, the correlation coefficient becomes 0.35. Although its relationship has improved, this isn't statisitcally significant (its p-value of .32 is quite a bit higher than the .05 confidence interval).