BUAN 6356.501 - Business Analytics with R (Spring 2019)
Problem Set 1

Question 1 <– wage1

1. Average years of education is 12.56 with lowest and highest years being 0 and 18 years respectively
2. Average hourly wage is 5.909 dollars which seems high enough as the wage density is right-skewed with the mean = 5.909 lying between the median (second quartile) = 4.7 and the third quartile = 6.9
3. Consumer Price Index for the year 1976 is 57% and Consumer Price Index for the year 2010 is 214%
4. Using above CPI values, the hourly wage in 2010 could be $22 which is considered as a decent wage
5. The total number of women is 252 and the total number of men is 274 in the sample data provided

Question 2 <– meap01

1. Smallest and largest values of math4 are 0 and 100 respectively. The range makes sense because it is the percentage of students in given schools that passed 4th grade math test with a satisfactory score
2. A total of 38 schools have perfect pass rate on the math test which is only 2.08% in the whole sample
3. Out of the schools provided in the sample, 17 schools have exactly 50% of pass rate in the math test
4. Avg pass rates for math and reading are 72% and 60% respectively indicating that reading is harder
5. There is a high correlation of 0.84 between math4 and read4. But this does not necessarily imply that they are dependent on each other. They might be being affected by other variables like school quality
6. Average of exppp is $5195 and its standard deviation is $1092. It does not seem to be a high variation
7. School A spending exceeds School B spending by 9.09% and logarithmic percentage difference is 8.7%

Question 3 <– 401k

1. Average participation rate from the workers is 87.36% and average match rate from the firms is $0.73
2. Equation is $\widehat{prate} = 83.0755 + 5.8611 * mrate \; where \; N = 1534, df = 1532 \; and \; R^2 = 0.0747$
3. The intercept means 83% workers will participate even if the firms don't contribute anything and the mrate coefficient means that $1 increase in match rate will add 5.86 to the participation rate metric
4. The predicted prate with $3.5 mrate is 103.59% which we cannot consider as a reasonable prediction because the prate is participation rate in percent values and a prate > 100 is not a practical prediction
5. Only 7.47% of variation in participation rate is explained by match rate in this linear regression model

Question 4 <– ceosal2

1. In the sample, average annual salary of CEO is $856,900 and average tenure as the CEO is 7.955 years
2. A total of 5 CEOs are in their first year as CEO. Among all CEOs, the longest tenure as CEO is 37 years
3. Equation is $ln(\widehat{salary}) = 6.5055 + 0.0097 * ceoten \; where \; N = 177, df = 175 \; and \; R^2 = 0.0132$
   Given one more year as CEO, the model predicts there would be a 0.97% increase in the annual salary

Question 5 <- wage2

1. Average salary is $958 and average IQ score is 101.3 and sample standard deviation of IQ is 15.05264
2. The equation is $\widehat{wage} = 116.9916 + 8.3031 * IQ \; where \; N = 935, df = 933 \; and \; R^2 = 0.09554$
   Increase in wage is $125 for 15 points IQ increase and the IQ explains only 9.55% variation in wage
3. The equation is $ln(\widehat{wage}) = 5.887 + 0.0088 * IQ \; where \; N = 935, df = 933 \; and \; R^2 = 0.09909$
   Increase in wage is 14% for 15 points IQ increase and the IQ explains only 9.91% variation in wage

Question 6 <- meap93

1. $math10 = 13.36 + 0.002456 * expend \; where \; N = 408 \; and \; df = 406 \; and \; R^2 = 0.03296$
   Every extra dollar spent has same but minimal effect on pass rate giving just 0.0025 point increment
2. $math10 = \beta_0 + \beta_1 ln(expend) + u \Rightarrow \frac{\partial(math10)}{\partial(expend)} = \frac{\beta_1}{expend}$ (differential of math10 over the expend)
   Change in math10 for 100% rise in expend is $\frac{\partial(math10)}{\partial(expend)} * (100\% \; expend) = \frac{\beta_1}{expend} * expend = \beta_1$
   Change in math10 for only 10% rise in expend is $\frac{\partial(math10)}{\partial(expend)} * (10\% \; expend) = \frac{\beta_1}{expend} * \frac{expend}{10} = \frac{\beta_1}{10}$
3. $math10 = -69.341 + 11.164 * ln(expend) \; where \; N = 408 \; and \; df = 406 \; and \; R^2 = 0.02966$
4. For 100% increase in expend we get 11.164 percentage point increase in the pass rate. We don't see high change for 10% increase in expend as we get only 1.1164 percentage point increase in pass rate
5. Maximum math pass rate is 66.7 and it won't touch 100 unless there is a 300% increase in the expend

Question 7 <- hprice1

1. $\widehat{price} = -19.315 + 0.1284 * sqrft + 15.1982 * bdrms \; where \; N = 88, df = 85 \; and \; R^2 = 0.632$
2. Addition of 1 bedroom increases the total price by $15,200 if we keep the square footage as constant
3. Its estimated that the addition of 1 bedroom and 140 square feet increases the total price by $33,180
4. 63% of variation in the selling price is explained by the square footage and the number of bedrooms
5. The predicted selling price for the first house with 2438 square footage and 4 bedrooms is $354,605
6. Residual for first house is -54,605. It is underpaid than predicted indicating influence of other factors

Question 8 <- ceosal2

1. $ln(\widehat{salary}) = 4.62 + 0.16 * ln(sales) + 0.11 * ln(mktval) \; where \; N = 177, df = 174, R^2 = 0.3$
2. $ln(\widehat{salary}) = 4.69 + 0.16 * ln(sales) + 0.1 * ln(mktval) + 0.00 * profits \; (N = 177, df = 173, R^2 = 0.3)$
   We cannot include profits in a logarithmic form because there are negative values in the profits field
   The firm performance variables explain only 30% of variations in CEO salaries and 70% is unexplained
3. $ln(\widehat{salary}) = 4.56 + 0.16ln(sales) + 0.1ln(mktval) + 0.00profits + 0.01ceoten \; (df = 172, R^2 = 0.32)$
   When tenure of the CEO increase by 1 year then it is estimated that salary of the CEO increases by 1%
4. The sample correlation coefficient between logarithm of market value, ln(mktval) and profits is 0.777
   Both are highly correlated and it is tough to estimate their independent effect on the salary logarithm

Question 9 <- attend

1. Minimum of atndrte is 6.25 and maximum of atndrte is 100 and the average of atndrte is 81.71
   Minimum of priGPA is 0.857 and maximum of priGPA is 3.93 and the average of priGPA = 2.587
   Minimum of ACT is 13 and maximum of ACT is 32 and the average of ACT is 22.51
2. $\widehat{atndrte} = 75.7 + 17.261 * priGPA - 1.717 * ACT \; where \; N = 680, df = 677 \; and \; R^2 = 0.2906$
   The intercept means predicted attendance rate for student with 0 prior GPA and 0 ACT score is 75.7%
   Although the intercept is useful in the linear regression model, its interpretation isn't very meaningful
3. Estimated slope coefficients indicate that 1 extra GPA point will increase attendance rate by 17% and 1 extra ACT score will result in 1.7% decreased attendance rate. The GPA slope makes sense, but ACT slope is little surprising. We can interpret that students with higher ACT score might skip more classes
4. The predicted attendance rate of a student with prior GPA of 3.65 and 20 ACT score is 104% which is impractical as it exceeds 100% attendance. One student with the exact values attends 87.5% classes
5. Predicted attendance rates of A and B are 93.16% and 67.32% respectively. Their difference is 25.84%

Question 10 <- htv

1. Range of educ variable is 14 (max = 20 & min = 6). Out of all men, 41.626% have 12th as their highest grade. None of the men or their parents have higher levels of education (Avg are 13.04, 12.18, 12.45)
2. $\widehat{educ} = 6.9644 + 0.3042 * motheduc + 0.1903 * fatheduc \; where \; N = 1230, df = 1227 \; and \; R^2 = 0.2493$
   Only 25% of variation in the men's education is explained by mother education and father education
   We can interpret that 3 grades rise in mother education will lead to 1 grade increase in son education
3. $\widehat{educ} = 8.45 + 0.19 * motheduc + 0.11 * fatheduc + 0.5 * abil \; (N = 1230, df = 1226 \; and \; R^2 = 0.4275)$
   Yes, the variable ability helps to explain variations in educ. Its addition increased R^2 from 0.25 to 0.43
4. $educ = \beta_0 + \beta_1 motheduc + \beta_2 fatheduc + \beta_3 abil + \beta_4 abil^2 + u$
   $\frac{\partial(educ)}{\partial(abil)} = \beta_3 + 2\beta_4 abil = 0 \Rightarrow abil = -\frac{\beta_3}{2\beta_4} \; (educ \; has \; a \; minimum \; at \; a \; negative \; abil \; value)$
   $\frac{\partial^2(educ)}{\partial(abil)^2} = 2\beta_4 > 0 \left(we \; indeed \; have \; a \; minimum \; for \; educ \; at \; abil = -\frac{\beta_3}{2\beta_4}\right)$
5. Only 1.22% men of the whole sample have an ability less than the above calculated value at minima
6. Below is the relationship plot between education level and measure of cognitive ability of the sample