# BUAN 6356.501 - Business Analytics with R (Spring 2019)
# Problem Set 2

## Question 1 <– vote1

1. Every 1% rise in A's campaign expenditure will result in $\beta_1/100$ units increase in percent votes for A
2. Given is a scenario that 1% increase in A's expenditure is offset by 1% increase in B's expenditure. The null hypothesis for this above scenario can be stated as $H_0: \beta_2 = -\beta_1 \; vs \; H_1: \beta_2 \neq -\beta_1$
3. The regression is $voteA = \beta_0 + \beta_1 \ln(expendA) + \beta_2 \ln(expendB) + \beta_3 prtystrA + u$

```
Call:
lm(formula = voteA ~ log(expendA) + log(expendB) + prtystrA,
    data = vote1)

Residuals:
     Min       1Q   Median       3Q      Max
-20.3990  -5.4184  -0.8737   4.9563  26.0575

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  45.08788    3.92680  11.482   <2e-16 ***
log(expendA)  6.08136    0.38211  15.915   <2e-16 ***
log(expendB) -6.61563    0.37889 -17.461   <2e-16 ***
prtystrA      0.15201    0.06203   2.451   0.0153 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.713 on 169 degrees of freedom
Multiple R-squared:  0.7925,    Adjusted R-squared:  0.7888
F-statistic: 215.2 on 3 and 169 DF,  p-value: < 2.2e-16
```

   The model predicts that if A spends 10% more, then A's vote share goes up by 0.608 percent points and if B spends 10% more on the campaign then A's vote share goes down by 0.662 percent points. We can use the estimated values from the regression and can test the null hypothesis stated above.
4. Our hypothesis can be written as $H_0: \theta_1 = 0 \; vs \; H_1: \theta_1 \neq 0 \; where \; \theta_1 = \beta_1 + \beta_2$. We can derive a new model, $voteA = \beta_0 + \theta_1 \ln(expendA) + \beta_2[\ln(expendB) - \ln(expendA)] + \beta_3 prtystrA + u$

```
Call:
lm(formula = voteA ~ log(expendA) + I(log(expendB) - log(expendA)) +
    prtystrA, data = vote1)

Residuals:
     Min       1Q   Median       3Q      Max
-20.3990  -5.4184  -0.8737   4.9563  26.0575

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 45.08788    3.92680  11.482   <2e-16 ***
log(expendA)                -0.53427    0.53311  -1.002   0.3177
I(log(expendB) - log(expendA)) -6.61563    0.37889 -17.461   <2e-16 ***
prtystrA                     0.15201    0.06203   2.451   0.0153 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.713 on 169 degrees of freedom
Multiple R-squared:  0.7925,    Adjusted R-squared:  0.7888
F-statistic: 215.2 on 3 and 169 DF,  p-value: < 2.2e-16
```

5. For $\theta_1$, we get t-statistic as -1.002 and p-value as 0.3177 (> 0.05). We cannot reject null hypothesis and can conclude that 1% increase in A's expenditure is offset by 1% increase in B's expenditure.

## Question 2 <– lawsch85

1. Given model is $\ln(salary) = \beta_0 + \beta_1 LSAT + \beta_2 GPA + \beta_3 \ln(libvol) + \beta_4 \ln(cost) + \beta_5 rank + u$
   Null hypothesis that rank has no effect on starting salary can be stated as $H_0: \beta_5 = 0 \; vs \; H_1: \beta_5 \neq 0$

```
Call:
lm(formula = log(salary) ~ LSAT + GPA + log(libvol) + log(cost) +
    rank, data = lawsch85)

Residuals:
      Min        1Q    Median        3Q       Max
-0.301356 -0.084982 -0.004359  0.077935  0.288614

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.3432262  0.5325192  15.667  < 2e-16 ***
LSAT         0.0046965  0.0040105   1.171  0.24372
GPA          0.2475239  0.0900371   2.749  0.00683 **
log(libvol)  0.0949932  0.0332544   2.857  0.00499 **
log(cost)    0.0375539  0.0321061   1.170  0.24427
rank        -0.0033246  0.0003485  -9.541  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1124 on 130 degrees of freedom
  (20 observations deleted due to missingness)
Multiple R-squared:  0.8417,    Adjusted R-squared:  0.8356
F-statistic: 138.2 on 5 and 130 DF,  p-value: < 2.2e-16
```

For $\beta_5$, this model gives t-statistic of -9.541 and p-value close to 0. We can reject the null hypothesis and can conclude that the rank does affect the median starting salary with 0.01 level of significance.

***2.*** LSAT p-value is 0.24 and insignificant at 5% level & GPA p-value is 0.007 and significant at 5% level
LSAT and GPA joint F-statistic is 8.05 and p-value is 0.005. They are jointly significant at 5% level

$$F = \frac{(R_{ur}^2 - R_r^2)/J}{(1 - R_{ur}^2)/(N - K)} = \frac{(0.8417 - 0.8221)/2}{(1 - 0.8417)/130} = 8.048$$

***3.*** clsize and faculty joint F-statistic is 0.94, p-value is 0.392. They are jointly insignificant at 5% level

$$F = \frac{(R_{ur}^2 - R_r^2)/J}{(1 - R_{ur}^2)/(N - K)} = \frac{(0.844 - 0.8417)/2}{(1 - 0.844)/128} = 0.9436$$

***4.*** Many factors can influence rank of law school like school facilities or faculty and student standards

## Question 3 <– hprice1

***1.*** $\ln(price) = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + u \; and \; \theta_1 = 150\beta_1 + \beta_2 \Rightarrow \theta_1 = 0.0858$

***2.*** The equation is rewritten as $\ln(price) = \beta_0 + \beta_1(sqrft - 150bdrms) + \theta_1 bdrms + u$

***3.*** The standard error of $\theta_1$ is 0.0268 and its 95% level confidence interval is 0.0326 and 0.1390

## Question 4 <– wage2

***1.*** $\ln(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u \; and \; H_0: \beta_2 = \beta_3 \; vs \; H_1: \beta_2 \neq \beta_3$

***2.*** Our hypothesis can be written as $H_0: \theta_1 = 0 \; vs \; H_1: \theta_1 \neq 0 \; where \; \theta_1 = \beta_2 - \beta_3$
We get new model, $\ln(wage) = \beta_0 + \beta_1 educ + \theta_1 exper + \beta_3(exper + tenure) + u$

***3.*** For $\theta_1$, we get t-statistic as 0.412 and p-value as 0.681 (> 0.05). We cannot reject null hypothesis

```
Call:
lm(formula = log(wage) ~ educ + exper + I(exper + tenure), data = wage2)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8282 -0.2401  0.0203  0.2569  1.3400

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     5.496696   0.110528  49.731  < 2e-16 ***
educ            0.074864   0.006512  11.495  < 2e-16 ***
exper           0.001954   0.004743   0.412    0.681
I(exper + tenure) 0.013375   0.002587   5.170 2.87e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3877 on 931 degrees of freedom
Multiple R-squared:  0.1551,    Adjusted R-squared:  0.1524
F-statistic: 56.97 on 3 and 931 DF,  p-value: < 2.2e-16
```

## Question 5 <- 401ksubs

1. There are 2017 single-person households in the given dataset
2. $nettfa = \beta_0 + \beta_1 inc + \beta_2 age + u$

```
Call:
lm(formula = nettfa ~ inc + age, data = t401ksubs[fsize == 1])

Residuals:
    Min      1Q  Median      3Q     Max
-179.95  -14.16   -3.42    6.03 1113.94

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -43.03981    4.08039 -10.548   <2e-16 ***
inc           0.79932    0.05973  13.382   <2e-16 ***
age           0.84266    0.09202   9.158   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 44.68 on 2014 degrees of freedom
Multiple R-squared:  0.1193,    Adjusted R-squared:  0.1185
F-statistic: 136.5 on 2 and 2014 DF,  p-value: < 2.2e-16
```

   Net wealth increases by $800 for $1000 rise in income and increases by $843 for 1 year rise in age
3. The intercept could be the predicted net wealth of newly born person with age = 0 and income = 0
4. The p-value for the test $H_0: \beta_2 = 1 \; vs \; H_1: \beta_2 < 1$ is 0.0437 (<0.5 but >0.01). It is significant at 5% level but not significant at 1% level. We can reject null hypothesis for 5% but we can't reject for 1%
5. Its almost the same. Estimate for income coefficient is 0.8 for first model and 0.82 for second model

## Question 6 <- kielmc

1. $\ln(price) = \beta_0 + \beta_1 \ln(dist) + u$. 10% rise in distance results in 3.65% increase in the value of price

```
Call:
lm(formula = log(price) ~ log(dist), data = kielmc[year == 1981])

Residuals:
     Min       1Q   Median       3Q      Max
-0.87318 -0.22657 -0.01985  0.25687  0.95045

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.04716    0.64624  12.452  < 2e-16 ***
log(dist)    0.36488    0.06576   5.548 1.39e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3543 on 140 degrees of freedom
Multiple R-squared:  0.1803,    Adjusted R-squared:  0.1744
F-statistic: 30.79 on 1 and 140 DF,  p-value: 1.395e-07
```

2. The estimate decreased from 0.365 to 0.055. Adding all other variables can reduce the significance
3. No change is observed because we already have the log(distance) variable included in the model
4. There is absolutely no impact from both the logarithmic squares and hence they are insignificant

## Question 7 <- wage1

1. $\ln(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + u$

```
Call:
lm(formula = log(wage) ~ educ + exper + I(exper^2), data = wage1)

Residuals:
     Min       1Q   Median       3Q      Max
-1.96516 -0.30073 -0.04635  0.29939  1.30407

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.1263226  0.1061081   1.191    0.234
educ         0.0906207  0.0074804  12.114  < 2e-16 ***
exper        0.0409731  0.0052051   7.872 2.04e-14 ***
I(exper^2)  -0.0007121  0.0001160  -6.141 1.63e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4466 on 522 degrees of freedom
Multiple R-squared:  0.3003,    Adjusted R-squared:  0.2963
F-statistic: 74.67 on 3 and 522 DF,  p-value: < 2.2e-16
```

2. Yes, estimate for experience squared is statistically significant at 1% level as it has p-value close to 0
3. Approximate return to 5th year of experience is 0.0339 and return to 20th year of experience is 0.0125
4. When experience crosses 28.7705 years then ln(wage) starts to decline. There are 121 such workers

## Question 8 <- wage2

1. Return to another year education holding experience fixed is the derivative of the given equation

$$\partial \ln(wage)/\partial educ = \beta_1 + \beta_3 exper$$

2. Null hypothesis that return to education won't depend on experience can be $H_0: \beta_3 = 0 \; vs \; H_1: \beta_3 \neq 0$
3. For $\beta_3$, the t-statistic value is 2.095 and p-value is 0.0365. We can reject null hypothesis at 5% level
4. New model using $\beta_1 = \theta_1 - 10\beta_3$ is $\ln(wage) = \beta_0 + \theta_1 educ + \beta_2 exper + \beta_3 educ(exper - 10)$. The standard error of $\theta_1$ is 0.0066 and its 95% level confidence interval is 0.0631 and 0.0891

## Question 9 <- gpa2

1. $sat = \beta_0 + \beta_1 hsize + \beta_2 hsize^2 + u$

```
Call:
lm(formula = sat ~ hsize + I(hsize^2), data = gpa2)

Residuals:
    Min      1Q  Median      3Q     Max
-562.38  -93.07   -3.71   90.62  507.72

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  997.981      6.203 160.875  < 2e-16 ***
hsize         19.814      3.991   4.965 7.14e-07 ***
I(hsize^2)    -2.131      0.549  -3.881 0.000106 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 138.9 on 4134 degrees of freedom
Multiple R-squared:  0.00765,   Adjusted R-squared:  0.007169
F-statistic: 15.93 on 2 and 4134 DF,  p-value: 1.279e-07
```

The quadratic term is statistically significant at 1% level of significance as the p-value is 0.0001 (<0.01)

2. Optimal high school size is when derivative is 0 and by evaluating derivative, the optimal value is 4.65
3. Its not representative of all high school seniors as the analysis is only for students who took SAT exam
4. Optimal high school size using ln(sat) is 4.696. It is almost same as the above optimal high school size

## Question 10 <- hprice1

1. Regression model is $\ln(price) = \beta_0 + \beta_1 \ln(lotsize) + \beta_2 \ln(sqrft) + \beta_3 bdrms + u$

```
Call:
lm(formula = log(price) ~ log(lotsize) + log(sqrft) + bdrms,
    data = hprice1)

Residuals:
     Min       1Q   Median       3Q      Max
-0.68422 -0.09178 -0.01584  0.11213  0.66899

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.29704    0.65128  -1.992   0.0497 *
log(lotsize)  0.16797    0.03828   4.388 3.31e-05 ***
log(sqrft)    0.70023    0.09287   7.540 5.01e-11 ***
bdrms         0.03696    0.02753   1.342   0.1831
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1846 on 84 degrees of freedom
Multiple R-squared:  0.643,     Adjusted R-squared:  0.6302
F-statistic: 50.42 on 3 and 84 DF,  p-value: < 2.2e-16
```

2. The predicted value of price when lotsize = 20000, sqrft = 2500 and bdrms = 4 is $400,574
3. $R^2$ for log model is 0.643 and $R^2$ for normal one is 0.672. Both are same, so we can prefer 2nd model