BUAN 6356.501 - Business Analytics with R (Spring 2019)
Problem Set 3

Question 1 <– mlb1

1. Null hypothesis is $H_0: \beta_{13} = 0 \; vs \; H_1: \beta_{13} \neq 0$ with p-value of t-statistic as 0.05432 (> 0.05). Hence, we cannot reject the null hypothesis and can conclude that β_{13} is insignificant at 5% level of significance. However, for a 10% level of significance β_{13} becomes significant. When controlling for all other factors, average salary difference for outfielders and catchers can be derived as $\left(e^{\beta_{13}} - 1\right) = 0.2886 \approx 29\%$

```
Call:
lm(formula = log(salary) ~ years + gamesyr + bavg + hrunsyr +
    rbisyr + runsyr + fldperc + allstar + frstbase + scndbase +
    thrdbase + shrtstop + catcher, data = mlb1)

Residuals:
     Min       1Q   Median       3Q      Max
-2.42088 -0.42665 -0.03092  0.47925  2.74975

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.1295536  2.3044545   4.830 2.07e-06 ***
years        0.0584178  0.0122732   4.760 2.87e-06 ***
gamesyr      0.0097670  0.0033776   2.892  0.00408 **
bavg         0.0004814  0.0011411   0.422  0.67340
hrunsyr      0.0191459  0.0159638   1.199  0.23124
rbisyr       0.0017875  0.0074755   0.239  0.81116
runsyr       0.0118707  0.0045264   2.623  0.00912 **
fldperc      0.0002833  0.0023078   0.123  0.90239
allstar      0.0063351  0.0028828   2.198  0.02866 *
frstbase    -0.1328008  0.1309243  -1.014  0.31115
scndbase    -0.1611010  0.1414296  -1.139  0.25547
thrdbase     0.0145271  0.1430352   0.102  0.91916
shrtstop    -0.0605672  0.1302031  -0.465  0.64210
catcher      0.2535592  0.1313128   1.931  0.05432 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7092 on 339 degrees of freedom
Multiple R-squared:  0.6535,    Adjusted R-squared:  0.6403
F-statistic: 49.19 on 13 and 339 DF,  p-value: < 2.2e-16
```

2. The null hypothesis is $H_0: \beta_9 = 0, \beta_{10} = 0, \beta_{11} = 0, \beta_{12} = 0, \beta_{13} = 0 \; vs \; H_1: at\ least\ one\ is\ not\ zero$ with p-value of F-statistic as 0.1168 (> 0.10). Hence, we cannot reject the null hypothesis and can conclude all estimates of $\beta_9, \beta_{10}, \beta_{11}, \beta_{12}, \beta_{13}$ are insignificant at both 5% and 10% level of significance.

```
Model 1: log(salary) ~ years + gamesyr + bavg + hrunsyr + rbisyr + runsyr +
    fldperc + allstar
Model 2: log(salary) ~ years + gamesyr + bavg + hrunsyr + rbisyr + runsyr +
    fldperc + allstar + frstbase + scndbase + thrdbase + shrtstop +
    catcher
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    344 174.99
2    339 170.52  5    4.4703 1.7774 0.1168
```

3. Above results are inconsistent for 10% level of significance but consistent for 5% level of significance. This inconsistency could be arising because we are calculating the joint significance of β_{13} which has moderate p-value along with the coefficients that are individually insignificant with very high p-values.

Question 2 <– gpa2

1. We can expect β_3 to be negative as $hsperc$ is lower for better students and β_4 to be positive as sat is higher for better students. We cannot say anything about the coefficients of $hsize, female, athlete$.
2. $colgpa = \beta_0 + \beta_1 hsize + \beta_2 hsize^2 + \beta_3 hsperc + \beta_4 sat + \beta_5 female + \beta_6 athlete + u$

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

Residuals:
     Min       1Q   Median       3Q      Max
-2.69216 -0.34954  0.03416  0.38806  1.90159

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.241e+00  7.949e-02  15.616  < 2e-16 ***
hsize       -5.685e-02  1.635e-02  -3.477 0.000512 ***
I(hsize^2)   4.675e-03  2.249e-03   2.079 0.037722 *
hsperc      -1.321e-02  5.728e-04 -23.068  < 2e-16 ***
sat          1.646e-03  6.682e-05  24.640  < 2e-16 ***
female       1.549e-01  1.800e-02   8.602  < 2e-16 ***
athlete      1.693e-01  4.235e-02   3.998  6.5e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5544 on 4130 degrees of freedom
Multiple R-squared:  0.2925,	Adjusted R-squared:  0.2915
F-statistic: 284.6 on 6 and 4130 DF,  p-value: < 2.2e-16
```

   Being an athlete improves the GPA by 0.1693 points and it is statistically significant even at 0.1% level.
3. If sat is dropped, coefficient of $athlete$ drops to 0.005 and becomes insignificant with 0.90318 p-value.

```
Call:
lm(formula = colgpa ~ hsize + I(hsize^2) + hsperc + female +
    athlete, data = gpa2)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5164 -0.3819  0.0205  0.4204  1.8809

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.0476980  0.0329148  92.594  < 2e-16 ***
hsize       -0.0534038  0.0175092  -3.050  0.00230 **
I(hsize^2)   0.0053228  0.0024086   2.210  0.02716 *
hsperc      -0.0171365  0.0005892 -29.086  < 2e-16 ***
female       0.0581231  0.0188162   3.089  0.00202 **
athlete      0.0054487  0.0447871   0.122  0.90318
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5937 on 4131 degrees of freedom
Multiple R-squared:  0.1885,	Adjusted R-squared:  0.1875
F-statistic: 191.9 on 5 and 4131 DF,  p-value: < 2.2e-16
```

   Since we are not accounting for sat scores, being an athlete does not show a significant effect on GPA. When sat scores are taken, only then can we observe that athletes have better GPA than non-athletes.

4. By adding an interaction variable $female * athlete$ to initial model, we get $\frac{\partial calgpa}{\partial athlete} = \beta_6 + \beta_7 female$

$\widehat{colgpa} = \beta_0 + \beta_1 hsize + \beta_2 hsize^2 + \beta_3 hsperc + \beta_4 sat + \beta_5 female + \beta_6 athlete + \beta_7 female * athlete$

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

Residuals:
     Min       1Q   Median       3Q      Max
-2.69202 -0.34944  0.03446  0.38799  1.90139

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.242e+00  7.955e-02  15.608  < 2e-16 ***
hsize          -5.680e-02  1.637e-02  -3.470 0.000525 ***
I(hsize^2)      4.670e-03  2.251e-03   2.075 0.038060 *
hsperc         -1.321e-02  5.730e-04 -23.056  < 2e-16 ***
sat             1.646e-03  6.687e-05  24.618  < 2e-16 ***
female          1.546e-01  1.831e-02   8.443  < 2e-16 ***
athlete         1.674e-01  4.849e-02   3.453 0.000560 ***
female:athlete  7.692e-03  9.617e-02   0.080 0.936257
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5545 on 4129 degrees of freedom
Multiple R-squared:  0.2925,    Adjusted R-squared:  0.2913
F-statistic: 243.9 on 7 and 4129 DF,  p-value: < 2.2e-16
```

The null hypothesis that the women athletes and women non-athletes have no difference in $colgpa$ is $H_0: \beta_6 + \beta_7 = \beta_7 \Rightarrow H_0: \beta_6 = 0 \; vs \; H_1: \beta_6 \neq 0$ with p-value of t-statistic as 0.00056 (< 0.001). The coefficient estimate of $athlete$ is significant even at 0.1% and we can reject null hypothesis. The effect of $athlete$ on $colgpa$ does not differ by gender as the coefficient of interaction variable is insignificant.

5. By adding an interaction variable $female * sat$ to initial model, we get $\frac{\partial calgpa}{\partial sat} = \beta_4 + \beta_7 female$

$\widehat{colgpa} = \beta_0 + \beta_1 hsize + \beta_2 hsize^2 + \beta_3 hsperc + \beta_4 sat + \beta_5 female + \beta_6 athlete + \beta_7 female * sat$

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

Residuals:
     Min       1Q   Median       3Q      Max
-2.69877 -0.35033  0.03414  0.38919  1.89876

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.264e+00  9.750e-02  12.962  < 2e-16 ***
hsize       -5.691e-02  1.635e-02  -3.480 0.000506 ***
I(hsize^2)   4.686e-03  2.250e-03   2.083 0.037307 *
hsperc      -1.323e-02  5.737e-04 -23.053  < 2e-16 ***
sat          1.625e-03  8.516e-05  19.089  < 2e-16 ***
female       1.023e-01  1.338e-01   0.765 0.444547
athlete      1.678e-01  4.253e-02   3.944 8.14e-05 ***
sat:female   5.121e-05  1.291e-04   0.397 0.691730
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5545 on 4129 degrees of freedom
Multiple R-squared:  0.2925,    Adjusted R-squared:  0.2913
F-statistic: 243.9 on 7 and 4129 DF,  p-value: < 2.2e-16
```

Effect of sat on $colgpa$ does not differ by gender as coefficient of interaction variable is insignificant.

Question 3 <– loanapp

1. If there is discrimination against minorities, β_1 will be positive raising approval probability for whites.
2. Coefficient estimate for $white$ is 0.2 with high t-statistic of 10.11 and can be concluded as significant. A white person has 20% more approval probability and it is high discrimination against the minorities.

```
Call:
lm(formula = approve ~ white, data = loanapp)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.90839  0.09161  0.09161  0.09161  0.29221 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.70779    0.01824   38.81   <2e-16 ***
white        0.20060    0.01984   10.11   <2e-16 ***
```

3. Coefficient estimate of $white$ reduces to 0.1288 and is significant, acting as evidence of discrimination.

```
Call:
lm(formula = approve ~ white + hrat + obrat + loanprc + unem + 
    male + married + dep + sch + cosign + chist + pubrec + mortlat1 + 
    mortlat2 + vr, data = loanapp)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.06482  0.00781  0.06387  0.13673  0.71105 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.936731   0.052735  17.763  < 2e-16 ***
white        0.128820   0.019732   6.529 8.44e-11 ***
hrat         0.001833   0.001263   1.451   0.1469    
obrat       -0.005432   0.001102  -4.930 8.92e-07 ***
loanprc     -0.147300   0.037516  -3.926 8.92e-05 ***
unem        -0.007299   0.003198  -2.282   0.0226 *  
male        -0.004144   0.018864  -0.220   0.8261    
married      0.045824   0.016308   2.810   0.0050 ** 
dep         -0.006827   0.006701  -1.019   0.3084    
sch          0.001753   0.016650   0.105   0.9162    
cosign       0.009772   0.041139   0.238   0.8123    
chist        0.133027   0.019263   6.906 6.72e-12 ***
pubrec      -0.241927   0.028227  -8.571  < 2e-16 ***
mortlat1    -0.057251   0.050012  -1.145   0.2525    
mortlat2    -0.113723   0.066984  -1.698   0.0897 .  
vr          -0.031441   0.014031  -2.241   0.0252 *  
```

4. Interaction term has coefficient estimate of 0.008 with a low p-value and is significant at 0.1% level.

```
Call:
lm(formula = approve ~ white + hrat + obrat + loanprc + unem + 
    male + married + dep + sch + cosign + chist + pubrec + mortlat1 + 
    mortlat2 + vr + white:obrat, data = loanapp)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.05523  0.01253  0.06320  0.12692  0.83284 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.180648   0.086808  13.601  < 2e-16 ***
white       -0.145975   0.080263  -1.819 0.069109 .  
hrat         0.001790   0.001260   1.421 0.155521    
obrat       -0.012226   0.002216  -5.518 3.88e-08 ***
loanprc     -0.152536   0.037436  -4.075 4.79e-05 ***
unem        -0.007528   0.003189  -2.360 0.018352 *  
male        -0.006015   0.018817  -0.320 0.749241    
married      0.045536   0.016260   2.800 0.005154 ** 
dep         -0.007630   0.006686  -1.141 0.253905    
sch          0.001777   0.016601   0.107 0.914787    
cosign       0.017709   0.041081   0.431 0.666458    
chist        0.129855   0.019227   6.754 1.90e-11 ***
pubrec      -0.240325   0.028149  -8.538  < 2e-16 ***
mortlat1    -0.062782   0.049891  -1.258 0.208400    
mortlat2    -0.126845   0.066891  -1.896 0.058071 .  
vr          -0.030540   0.013993  -2.183 0.029188 *  
white:obrat  0.008088   0.002290   3.531 0.000423 ***
```

5. The confidence interval for the linear combination $\frac{\partial approve}{\partial white} = \beta_1 + 32\beta_{16}$ is $(0.07325, 0.15243)$

Question 4 <– hprice1

1. Compared to OLS, Robust errors increased by 1013% for $lotsize$, 207% for $sqrft$, 28% for $bdrms$.

```
> model41 <- lm(price~lotsize+sqrft+bdrms,data=hprice1)
> summary(model41)

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

Residuals:
     Min       1Q   Median       3Q      Max
-120.026  -38.530   -6.555   32.323  209.376

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.177e+01  2.948e+01  -0.739  0.46221
lotsize      2.068e-03  6.421e-04   3.220  0.00182 **
sqrft        1.228e-01  1.324e-02   9.275 1.66e-14 ***
bdrms        1.385e+01  9.010e+00   1.537  0.12795
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 59.83 on 84 degrees of freedom
Multiple R-squared:  0.6724,    Adjusted R-squared:  0.6607
F-statistic: 57.46 on 3 and 84 DF,  p-value: < 2.2e-16

> sqrt(diag(vcov(model41)))
 (Intercept)      lotsize        sqrft        bdrms
2.947504e+01 6.421258e-04 1.323741e-02 9.010145e+00
> sqrt(diag(vcovHC(model41)))
 (Intercept)      lotsize        sqrft        bdrms
41.032694404  0.007148464  0.040732542 11.561790104
```

2. Compared to OLS, Robust errors increased 39% for $\ln(lotsize)$, 30% for $\ln(sqrft)$, 29% for $bdrms$.

```
> model42 <- lm(log(price)~log(lotsize)+log(sqrft)+bdrms,data=hprice1)
> summary(model42)

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

> sqrt(diag(vcov(model42)))
 (Intercept) log(lotsize)   log(sqrft)        bdrms
  0.65128361   0.03828115   0.09286525   0.02753131
> sqrt(diag(vcovHC(model42)))
 (Intercept) log(lotsize)   log(sqrft)        bdrms
  0.85045733   0.05327497   0.12139232   0.03557555
```

3. Using log transformation reduced the effect of heteroskedasticity and reduced the marginal change between heteroskedasticity corrected robust standard errors and the normal OLS standard errors.

Question 5 <– gpa1

1. The OLS regression of the model $colGPA = \beta_0 + \beta_1 hsGPA + \beta_2 ACT + \beta_3 skipped + \beta_4 PC + u$

```
Call:
lm(formula = colGPA ~ hsGPA + ACT + skipped + PC, data = gpa1)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.84006 -0.20392 -0.03352  0.25346  0.74558 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.35651    0.32750   4.142 6.01e-05 ***
hsGPA        0.41295    0.09243   4.468 1.65e-05 ***
ACT          0.01334    0.01044   1.278  0.20353    
skipped     -0.07103    0.02625  -2.706  0.00768 ** 
PC           0.12444    0.05731   2.171  0.03165 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3251 on 136 degrees of freedom
Multiple R-squared:  0.2593,	Adjusted R-squared:  0.2375 
F-statistic:  11.9 on 4 and 136 DF,  p-value: 2.553e-08
```

2. $\widehat{u_i}^2 = \delta_0 + \delta_1 \widehat{colGPA} + \delta_2 \left(\widehat{colGPA}\right)^2 + e$

```
Call:
lm(formula = model51$resid^2 ~ model51$fitted + I(model51$fitted^2))

Residuals:
     Min       1Q   Median       3Q      Max 
-0.13286 -0.07802 -0.04020  0.04954  0.60632 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         -0.321837   2.005841  -0.160    0.873
model51$fitted       0.129599   1.316763   0.098    0.922
I(model51$fitted^2)  0.002946   0.215660   0.014    0.989

Residual standard error: 0.1237 on 138 degrees of freedom
Multiple R-squared:  0.04934,	Adjusted R-squared:  0.03557 
F-statistic: 3.581 on 2 and 138 DF,  p-value: 0.03045
```

3. All the above fitted values from part 2 are positive with 0.02738 as their minimum value.
The WLS regression of the model $colGPA = \beta_0 + \beta_1 hsGPA + \beta_2 ACT + \beta_3 skipped + \beta_4 PC + u$

```
Call:
lm(formula = colGPA ~ hsGPA + ACT + skipped + PC, data = gpa1, 
    weights = 1/fitted(model52))

Weighted Residuals:
    Min      1Q  Median      3Q     Max 
-2.6994 -0.6892 -0.1191  0.7963  2.5098 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.401564   0.298430   4.696 6.39e-06 ***
hsGPA        0.402506   0.083362   4.828 3.65e-06 ***
ACT          0.013162   0.009827   1.339 0.182698    
skipped     -0.076365   0.022173  -3.444 0.000762 ***
PC           0.126005   0.056339   2.237 0.026945 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.013 on 136 degrees of freedom
Multiple R-squared:  0.3062,	Adjusted R-squared:  0.2858 
F-statistic: 15.01 on 4 and 136 DF,  p-value: 3.488e-10
```

There is very minor difference between OLS and WLS coefficient estimates for $skipped$ and PC. Both the OLS and WLS estimates are significant at 5% level for PC and are significant at 1% level for $skipped$.

4. Heteroskedasticity robust WLS errors are slightly more when compared to normal WLS errors.