Swetha Adike

Venu Goud Raparti

Problems from Chapter 10

MSIS 545

11.

In this problem, regression output data for the data containing log of number of butterfly species observed on log of size of the reserve and number of days of observation from 16 reserves is given

a. Two-sided p value for the test on size of reserve and effect on number of species, after account for days of observation is 0.2443 and one-sided p value is 0.2443/2 = 1.222. Since this is greater than level of significance, there is no evidence that the median number of species is related to the reserve size. And so, it does not account to relationship between species number and reserve sizes. When researchers spend more days searching for butterflies in large reserves and not small reserves existing no relation on median number of species.
b. Two-sided p value for the test if coefficient of lsize is 1 is calculated estimate and standard error as (0.0809-1)/0.1131 = -8.1264
c. 95% confidence interval for coefficient of lsize is calculated by 0.0809+-(2.160*0.1131) = 0.0809+-0.2443 = -0.1634 and 0.3252
d. R square for the test is 11.41% which means that 0.1141 proportion of variation in log number of species remains unexplained by log size and days of observation which is 100-11.41 = 88.59%.

14.

This problem is regarding the study of joint toxicity of copper and zinc. The study was randomly conducted on 25 beakers containing minnow larvae and gave one of 25 treatment combinations. The treatment combinations consisted of combination of adding 5 levels of zinc and 5 levels of copper in beaker. The sample is then analyzed for a protein after the minnow larvae were homogenized.

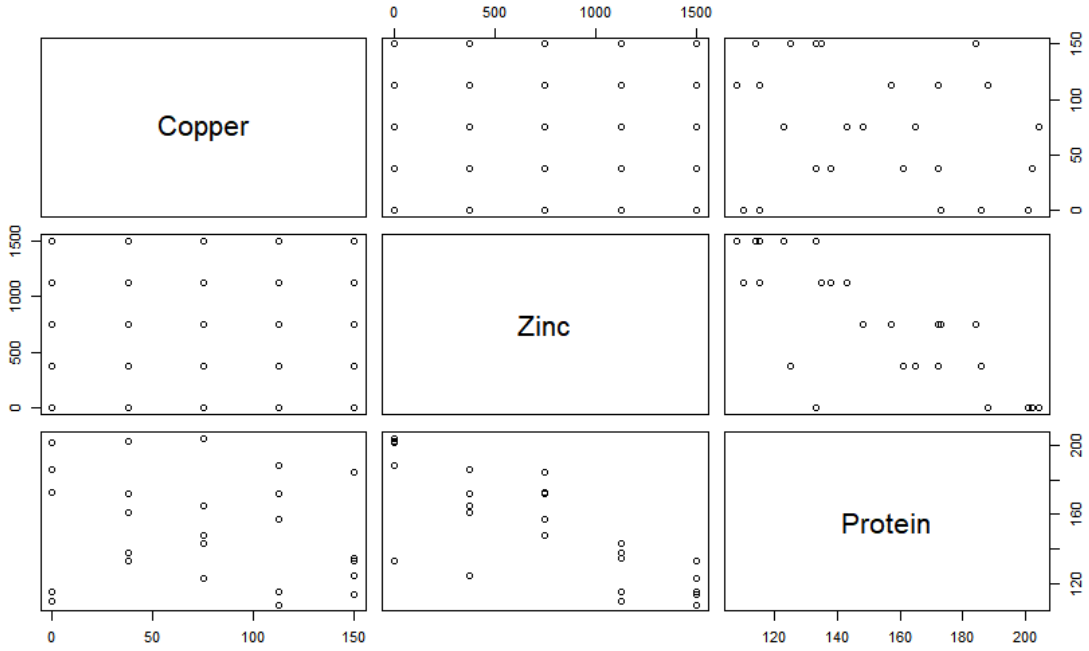
Here analysis is made on the data samples for fitting a full second-order model for regression of protein on copper and zinc. For this, lets head to the data summary statistics

```
> summary(ex1014)
     Copper           Zinc         Protein
 Min.   :  0.0   Min.   :   0   Min.   :108.0
 1st Qu.: 38.0   1st Qu.: 375   1st Qu.:125.0
 Median : 75.0   Median : 750   Median :148.0
 Mean   : 75.2   Mean   : 750   Mean   :152.2
```

```
 3rd Qu.:113.0   3rd Qu.:1125   3rd Qu.:173.0
 Max.   :150.0   Max.   :1500   Max.   :204.0
```

From the above summary, the model clearly suggests transformation since there is wide spread in the data, especially, in zinc. Let's make a scatterplot to find the relationships in the fields

```
> plot(ex1014)
```



From the above scatter plot, copper with protein and zinc with protein graphs are widely spread. Let's fit a second order regression equation for protein on copper and zinc. Below is the code for it

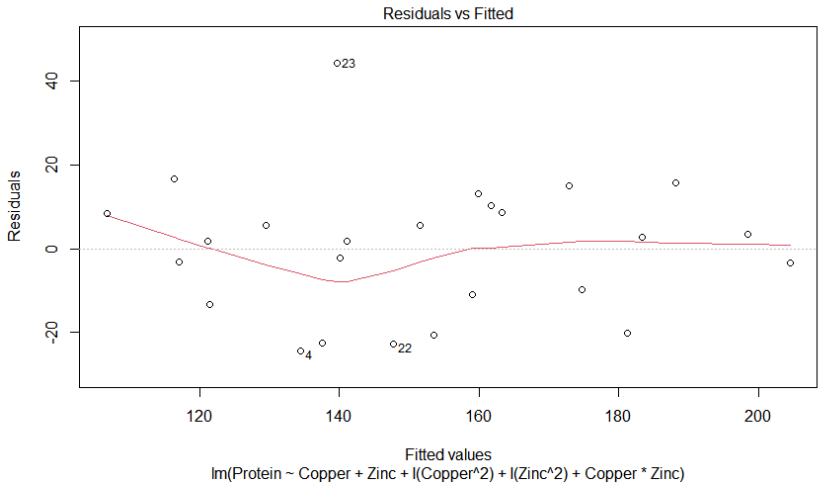
```
> pcz.lm <- lm(Protein ~ Copper + Zinc + I(Copper^2) + I(Zinc^2)
+ Copper*Zinc, data = ex1014)

> summary(pcz.lm)

Call:

lm(formula = Protein ~ Copper + Zinc + I(Copper^2) + I(Zinc^2) +

    Copper * Zinc, data = ex1014)

Residuals:

    Min      1Q  Median      3Q     Max

-24.397 -11.001   1.903   8.688  44.338
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.045e+02  1.236e+01  16.548 9.66e-13 ***
Copper      -9.418e-02  2.567e-01  -0.367   0.7178
Zinc        -5.362e-02  2.567e-02  -2.089   0.0504 .
I(Copper^2) -1.636e-03  1.522e-03  -1.075   0.2959
I(Zinc^2)   -7.721e-06  1.523e-05  -0.507   0.6179
Copper:Zinc  2.727e-04  1.274e-04   2.140   0.0455 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 17.91 on 19 degrees of freedom
Multiple R-squared:  0.7389,  Adjusted R-squared:  0.6702
F-statistic: 10.75 on 5 and 19 DF,  p-value: 5.095e-05
```

From the above results, it shows that the coefficients copper, copper^2 and zinc^2 are non-significant while the remaining coefficients are moderately significant. However, the p value for the model is <0.00001 with 73.89% as R square.

Below is the plot of the residuals Vs fitted values

```
> plot(pcz.lm,1)
```



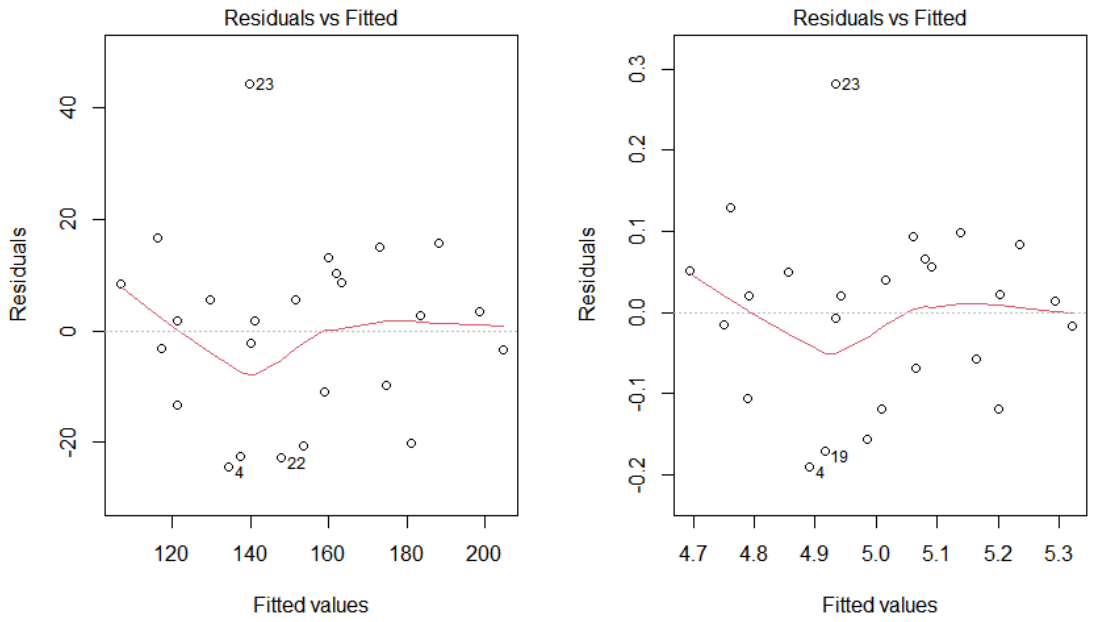
Let's repeat the model by applying log transformations to protein and see if there is any difference in the two models. The below is the code and result

```
> lnpcz.lm <- lm(log(Protein) ~ Copper + Zinc + I(Copper^2) +
I(Zinc^2) + Copper*Zinc, data = ex1014)

> summary(lnpcz.lm)

Call:

lm(formula = log(Protein) ~ Copper + Zinc + I(Copper^2) + I(Zinc^2)
+

    Copper * Zinc, data = ex1014)

Residuals:

     Min       1Q   Median       3Q      Max

-0.18983 -0.06898  0.02019  0.05710  0.28185

Coefficients:

              Estimate Std. Error t value Pr(>|t|)

(Intercept)  5.321e+00  8.306e-02  64.060   <2e-16 ***

Copper      -2.163e-04  1.726e-03  -0.125   0.9016

Zinc        -2.770e-04  1.725e-04  -1.605   0.1249

I(Copper^2) -1.237e-05  1.023e-05  -1.209   0.2416

I(Zinc^2)   -9.382e-08  1.023e-07  -0.917   0.3708

Copper:Zinc  1.633e-06  8.562e-07   1.907   0.0717 .

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1204 on 19 degrees of freedom

Multiple R-squared:  0.7309,  Adjusted R-squared:  0.6601

F-statistic: 10.32 on 5 and 19 DF,  p-value: 6.69e-05
```

Comparing the two models, R square has not changed much while p value is not significant for all the coefficients. However, the p value for the model is <0.00001.

Let’s compare the residual plots for both the above models and see if there is really any difference.

```
> par(mfrow = c(1,2))

> plot(pcz.lm,1)

> plot(lnpcz.lm, 1)

> par(mfrow = c(1,1))
```



The above graph shows that there is not much difference in both the models which is quite suggested by no change in R square value. Hence log transformations on the protein value is not much effected to the results.

15.

This is the list of data on Kentucky Derby horse race winners from 1896 to 2011 on 1.25 miles length of race. The winning time here is inversely related to the speed. Let's head to summary statistics in the data involving year, speed and track

```
> dim(ex0920)
[1] 116   8
> summary(ex0920[c(1,6,7)])
      Year          Speed          Track
 Min.   :1896   Min.   :33.28   Dusty : 1
 1st Qu.:1925   1st Qu.:35.84   Fast  :85
 Median :1954   Median :36.47   Good  : 9
 Mean   :1954   Mean   :36.22   Heavy : 6
 3rd Qu.:1982   3rd Qu.:36.89   Muddy : 6
 Max.   :2011   Max.   :37.69   Sloppy: 4
                                Slow  : 5
```

Out of the 8 fields year, winner, (race) starters, Net to winner, time, speed, track and conditions with 116 records, the above are the summary statistics for year, speed and track.

Clearly, the data contained the race details from 1896 to 2011 with a mean speed 36.22 units where we have 7 categories of track. We can see that the dusty track is used only one time out of 116 records while fast track is mostly used one.

Let's test if there is any effect of track categories on winning speed after accounting for year. For this, let's fit a regression model for year and track categories on speed and below is the code for it

```
> syt.lm <- lm(Speed ~ Year + factor(Track), data = ex0920)
> summary(syt.lm)
Call:
lm(formula = Speed ~ Year + factor(Track), data = ex0920)
Residuals:
     Min       1Q   Median       3Q      Max
-1.51083 -0.25661  0.04018  0.25578  0.95203
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         6.094652   2.606132   2.339  0.02120 *
Year                0.015367   0.001354  11.346  < 2e-16 ***
factor(Track)Fast   0.324727   0.455649   0.713  0.47759
factor(Track)Good   0.020822   0.473941   0.044  0.96504
factor(Track)Heavy -1.325379   0.480915  -2.756  0.00687 **
factor(Track)Muddy -0.765976   0.484498  -1.581  0.11681
factor(Track)Sloppy -0.372564   0.506774  -0.735  0.46383
factor(Track)Slow  -0.371396   0.489466  -0.759  0.44964
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.4449 on 108 degrees of freedom
Multiple R-squared:  0.7701,  Adjusted R-squared:  0.7552
F-statistic: 51.68 on 7 and 108 DF,  p-value: < 2.2e-16
```

The p value for the model is <0.00001 while with R square as 77% while many coefficients under tracks are non-significant

Similarly, let's fit a regression model for year on speed

```
> sy.lm <- lm(Speed ~ Year, data = ex0920)
> summary(sy.lm)
Call:
lm(formula = Speed ~ Year, data = ex0920)
Residuals:
     Min       1Q   Median       3Q      Max
-2.03396 -0.35691  0.05172  0.43039  1.18001
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.51752    3.28191  -0.767    0.445
Year         0.01983    0.00168  11.804   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.6058 on 114 degrees of freedom
Multiple R-squared:   0.55,   Adjusted R-squared:  0.5461
F-statistic: 139.3 on 1 and 114 DF,  p-value: < 2.2e-16
```

It can be observed that the R square is 55% and less when compared to previous model. Let's make the comparison clearer by applying anova on the above models

```
> anova(syt.lm, sy.lm)
Analysis of Variance Table
Model 1: Speed ~ Year + factor(Track)
Model 2: Speed ~ Year
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    108 21.373
2    114 41.837 -6   -20.464 17.234 6.785e-14 ***
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The anova model suggests that the difference is significant with p value < 0.0001 and sum of squares is 20.464 with F statistic 17.234.

29.

This is problem is dealt with data of 25,437 males between age 18 and 70 who worked full-time in the year 1987 with years of education, years of experience and whether they are black, whether they worked in standard metropolitan statistical area and code for the region in US where they worked like north-east, Midwest, south and west.

The data is collected in a survey and the present study is to analyze the data to see to what extent black males were paid less than nonblack males in same region and with same level of education and experience.

Let’s head to the summary of the data

```
> summary(ex1029)
```

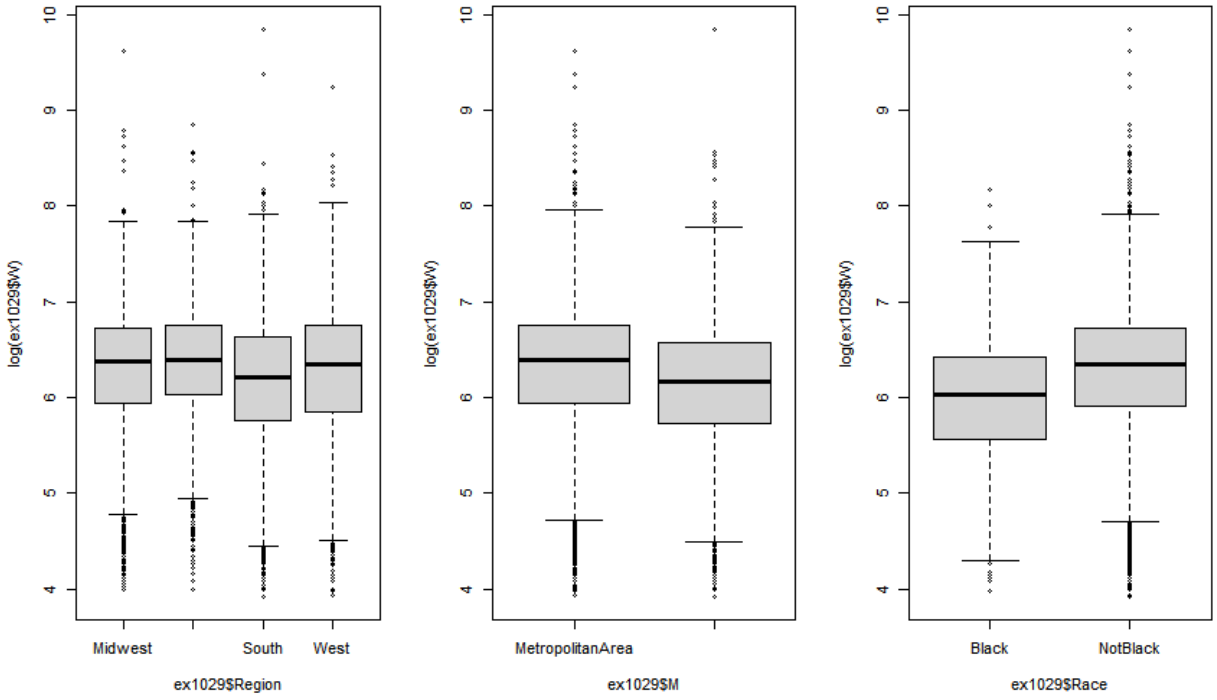
```
Region               MetropolitanStatus     Exper             Educ
Race
 Midwest  :6185   MetropolitanArea   :18880    Min.   : 0.00   Min.
: 0.00   Black    : 1978
 Northeast:5885   NotMetropolitanArea: 6557    1st Qu.: 9.00   1st
Qu.:12.00   NotBlack:23459
 South    :7933                                Median :16.00
Median :12.00
 West     :5434                               Mean   :18.74   Mean
:13.06
                                               3rd      Qu.:27.00
3rd Qu.:16.00
                                               Max.        :63.00
Max.   :18.00
WeeklyEarnings
 Min.   :   50.39
 1st Qu.:  356.13
 Median :  569.80
 Mean   :  642.59
 3rd Qu.:  830.96
```

```
 Max.    :18777.20
```

It can be observed that the ratio of black to nonblack is very high approximately 1:20.

From the above statistics, it can be seen that wage should be log transformed but other than that, let's see if there is any other transformations required or anything more to be analyzed in other fields by plotting log(WeeklyEarnings) across the other remaining fields (Region, metropolitan area status, education, experience and Race)

```
> par(mfrow = c(1,3))
> boxplot(log(ex1029$W) ~ ex1029$Region)
> boxplot(log(ex1029$W) ~ ex1029$M)
> boxplot(log(ex1029$W) ~ ex1029$Race)
> par(mfrow = c(1,1))
```

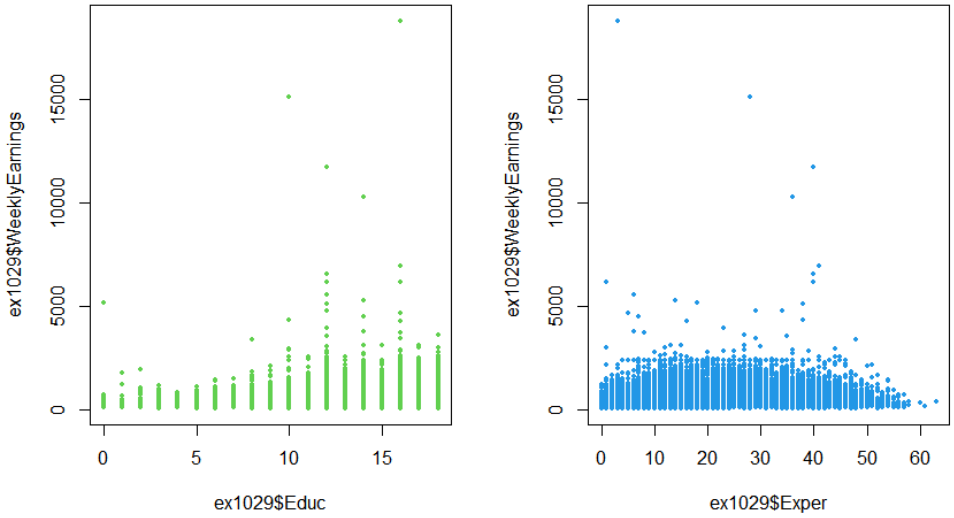


The race boxplot (3rd boxplot) shows that relation between black males and wages, which is of interest

Below is the plot made for education and experience fields across log(WeeklyEarnings)

```
> par(mfrow = c(1,2))
> plot(ex1029$Educ, ex1029$WeeklyEarnings,  col = 3, pch = 19, cex
= 0.5)
```

```
> plot(ex1029$Exper, ex1029$WeeklyEarnings,  col = 4, pch = 19, cex = 0.5)

> par(mfrow = c(1,1))
```



The above boxplots and plots are showing different relation with log of weekly earnings, each one explaining distinct relation with wage. For this reason, in regression fit, all should be included in the analysis.

Let's see if there is any specific relation between the race and region

```
> rr.lm <- lm(log(ex1029$W) ~ factor(ex1029$Region) + factor(ex1029$Race))

> summary(rr.lm)

Call:

lm(formula = log(ex1029$W) ~ factor(ex1029$Region) + factor(ex1029$Race))

Residuals:

    Min      1Q  Median      3Q     Max

-2.3994 -0.3836  0.0499  0.4174  3.6121

Coefficients:

                                Estimate  Std.  Error  t  value Pr(>|t|)
```

```
(Intercept)                      6.03644    0.01594 378.774  < 2e-
16 ***
factor(ex1029$Region)Northeast  0.07702    0.01121   6.869 6.59e-
12 ***
factor(ex1029$Region)South     -0.08767    0.01053  -8.323  < 2e-
16 ***
factor(ex1029$Region)West      -0.02232    0.01145  -1.949   0.0513
.
factor(ex1029$Race)NotBlack     0.27955    0.01464  19.097  < 2e-
16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.6157 on 25432 degrees of freedom
Multiple R-squared:  0.02675, Adjusted R-squared:  0.0266
F-statistic: 174.7 on 4 and 25432 DF,  p-value: < 2.2e-16
```

There is a significant relation between race and region with p value <0.00001 but the R square is not too high there is 38.43% chance that the relation is not significant.

We have made few combinations to fit a regression model for all the factors with log of weekly earnings.

```
> black <- ex1029$Race == "Black"
```

```
> wage.lm <- lm(log(ex1029$W) ~ black + ex1029$Exper + ex1029$Educ
+ factor(ex1029$Region))
```

```
> summary(wage.lm)
```

```
Call:
lm(formula = log(ex1029$W) ~ black + ex1029$Exper + ex1029$Educ +
    factor(ex1029$Region))
Residuals:
    Min      1Q  Median      3Q     Max
-2.6653 -0.3013  0.0432  0.3488  3.5868
Coefficients:
                                 Estimate  Std.  Error  t  value
Pr(>|t|)
```

```
(Intercept)                          4.6723340  0.0193688 241.230  <
2e-16 ***
blackTRUE                           -0.2156000  0.0126999 -16.977  <
2e-16 ***
ex1029$Exper                         0.0177516  0.0002822  62.914  <
2e-16 ***
ex1029$Educ                          0.0990952  0.0012020  82.443  <
2e-16 ***
factor(ex1029$Region)Northeast       0.0619622  0.0097088   6.382
1.78e-10 ***
factor(ex1029$Region)South          -0.0575524  0.0091288  -6.304 2.94e-
10 ***
factor(ex1029$Region)West            0.0034676  0.0099176   0.350
0.727
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.533 on 25430 degrees of freedom
Multiple R-squared:  0.2707,  Adjusted R-squared:  0.2705
F-statistic:  1573 on 6 and 25430 DF,  p-value: < 2.2e-16
```

> wage2.lm <- lm(log(ex1029$W) ~ black + ex1029$Exper + I(ex1029$Exper^2) +

+ ex1029$Educ + factor(ex1029$Region))

> summary(wage2.lm)

```
Call:
lm(formula = log(ex1029$W) ~ black + ex1029$Exper + I(ex1029$Exper^2) +
    ex1029$Educ + factor(ex1029$Region))

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-2.6977 -0.2929  0.0383  0.3317  3.7817


Coefficients:
                                  Estimate  Std.  Error  t  value
Pr(>|t|)
(Intercept)                        4.510e+00  1.913e-02 235.739  <
2e-16 ***
blackTRUE                         -2.170e-01  1.228e-02 -17.666  <
2e-16 ***
ex1029$Exper                       5.455e-02  9.198e-04  59.311  <
2e-16 ***
I(ex1029$Exper^2)                 -8.280e-04  1.976e-05 -41.898  <
2e-16 ***
ex1029$Educ                        9.074e-02  1.180e-03  76.930  <
2e-16 ***
factor(ex1029$Region)Northeast    6.870e-02   9.392e-03     7.315
2.66e-13 ***
factor(ex1029$Region)South    -5.920e-02  8.829e-03  -6.705 2.06e-
11 ***
factor(ex1029$Region)West         -5.001e-03   9.594e-03   -0.521
0.602
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 0.5155 on 25429 degrees of freedom
Multiple R-squared:  0.3178,  Adjusted R-squared:  0.3176
F-statistic:  1692 on 7 and 25429 DF,  p-value: < 2.2e-16


> wage3.lm <- lm(log(ex1029$W) ~ black + ex1029$Exper + ex1029$Educ
+ ex1029$Educ*ex1029$Exper
+                + factor(ex1029$Region))
> summary(wage3.lm)
```

```
Call:
lm(formula = log(ex1029$W) ~ black + ex1029$Exper + ex1029$Educ +
    ex1029$Educ * ex1029$Exper + factor(ex1029$Region))

Residuals:
    Min      1Q  Median      3Q     Max
-2.6710 -0.3022  0.0441  0.3490  3.5789

Coefficients:
                                   Estimate  Std.  Error  t  value
Pr(>|t|)
(Intercept)                         4.633e+00  3.174e-02 145.963  <
2e-16 ***
blackTRUE                          -2.159e-01  1.270e-02 -16.996  <
2e-16 ***
ex1029$Exper                        1.940e-02  1.084e-03  17.895  <
2e-16 ***
ex1029$Educ                         1.022e-01  2.295e-03  44.516  <
2e-16 ***
factor(ex1029$Region)Northeast    6.168e-02   9.710e-03     6.352
2.16e-10 ***
factor(ex1029$Region)South    -5.783e-02  9.130e-03  -6.334 2.43e-
10 ***
factor(ex1029$Region)West          3.541e-03   9.917e-03    0.357
0.721
ex1029$Exper:ex1029$Educ          -1.335e-04   8.485e-05   -1.574
0.116
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.533 on 25429 degrees of freedom
```

Multiple R-squared: 0.2707, Adjusted R-squared: 0.2705

F-statistic: 1349 on 7 and 25429 DF, p-value: < 2.2e-16

```
> wage4.lm <- lm(log(ex1029$W) ~ black + I(ex1029$Exper^3) +
+                   I(ex1029$Educ^3) + factor(ex1029$Region))
> summary(wage4.lm)
```

```
Call:
lm(formula = log(ex1029$W) ~ black + I(ex1029$Exper^3) +
I(ex1029$Educ^3) +
    factor(ex1029$Region))

Residuals:
    Min      1Q  Median      3Q     Max
-2.7960 -0.3326  0.0474  0.3776  3.5579

Coefficients:
                                  Estimate  Std.  Error  t  value
Pr(>|t|)
(Intercept)                        5.802e+00  1.016e-02 571.228  <
2e-16 ***
blackTRUE                         -2.153e-01  1.338e-02 -16.099  <
2e-16 ***
I(ex1029$Exper^3)                  4.144e-06  1.392e-07  29.774  <
2e-16 ***
I(ex1029$Educ^3)                   1.732e-04  2.442e-06  70.939  <
2e-16 ***
factor(ex1029$Region)Northeast    5.821e-02   1.022e-02     5.693
1.26e-08 ***
factor(ex1029$Region)South    -7.487e-02  9.603e-03  -7.796 6.63e-
15 ***
```

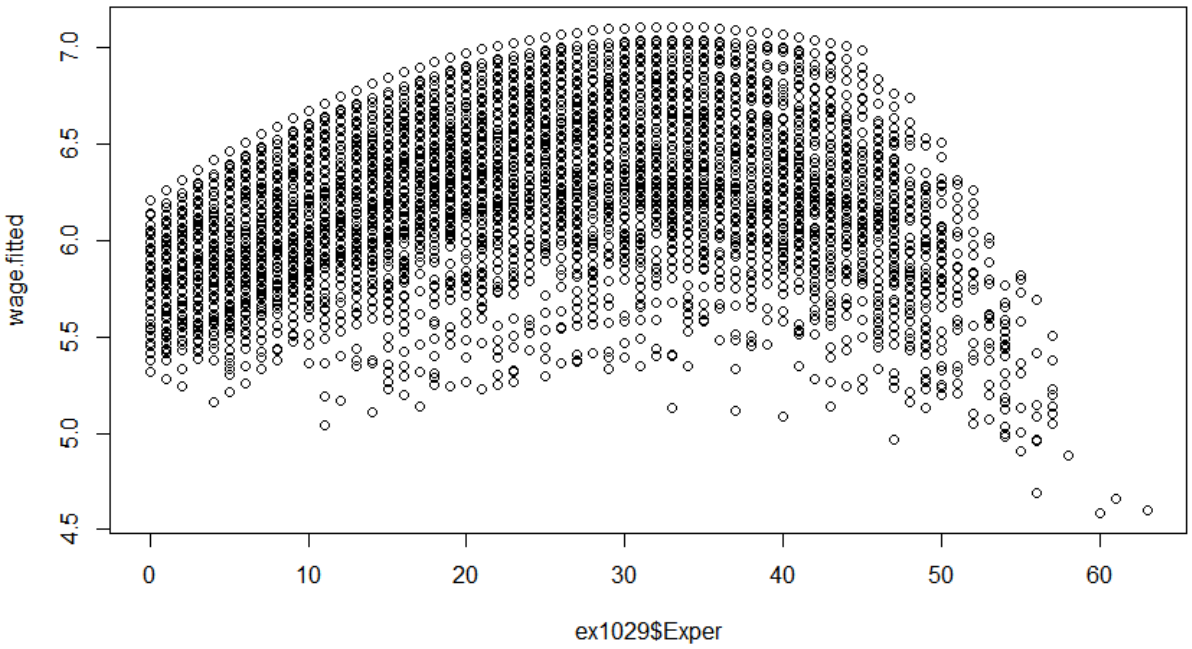
```
factor(ex1029$Region)West          -1.939e-02   1.044e-02   -1.858
0.0632 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 0.5611 on 25430 degrees of freedom
Multiple R-squared:  0.1916,  Adjusted R-squared:  0.1914
F-statistic:  1005 on 6 and 25430 DF,  p-value: < 2.2e-16
```

We have considered metro first in these models but R square has quite increased when we removed metro factor.

From the above results, the last model where experience ^3 is considered gives high R square out of the other models 56.11% but the fitted values curve for this model has got reversed and completely got complicated. We have repeated the same process for all the other models and then only wage.lm plot is like a curve without any deviations. So, let’s go with the second model of wage2.lm

```
> exp(wage2.lm$coefficients)
Intercept)              blackTRUE                ex1029$Exper
90.9070323              0.8049348                1.0560694
I(ex1029$Exper^2)     ex1029$Educ factor(ex1029$Region)Northeast
0.9991723             1.0949854                     1.0711103
factor(ex1029$Region)South      factor(ex1029$Region)West
  0.9425197                      0.9950113
> confint(wage2.lm)
                                        2.5 %        97.5 %
(Intercept)                        4.4723402469  4.5473344749
blackTRUE                         -0.2410698860 -0.1929180055
ex1029$Exper                       0.0527510885  0.0563568048
I(ex1029$Exper^2)                 -0.0008667871 -0.0007893119
ex1029$Educ                        0.0884291015  0.0930529974
factor(ex1029$Region)Northeast     0.0502876091  0.0871040178
factor(ex1029$Region)South        -0.0765046176 -0.0418923072
```

```
factor(ex1029$Region)West       -0.0238066502  0.0138042093
> wage.fitted <- wage2.lm$fitted.values
> plot(ex1029$Exper, wage.fitted)
```



The above graph shows the that the years of experience give high wages till person reaches approximately 45 years and then decreases. This can be explained that the person cannot be productive after they get older.

The average weekly earning of black male with all the factors keeping zero is 90.97. it is with 95% confidence that the average weekly earnings for this case falls between the confidence interval as stated above.

For blacks with zero years of education and experience who work in metropolitan area has a median weekly wage is 20.51% smaller than white men.

The whole model is explained with 31.78% R square which is not so great but an average estimation. Finally, the data gives no evidence that race effect on wages in different regions.