BOSTON UNIVERSITY

AD699

Data Mining for Business Analytics

Spring 2021

Professor Greg Page

Assignment 2

Kunfei Chen

U15575304

Task 1

The data file “nba_contracts.csv” is downloaded from the class blackboard site and imported to R-studio for analysis.

Task 2

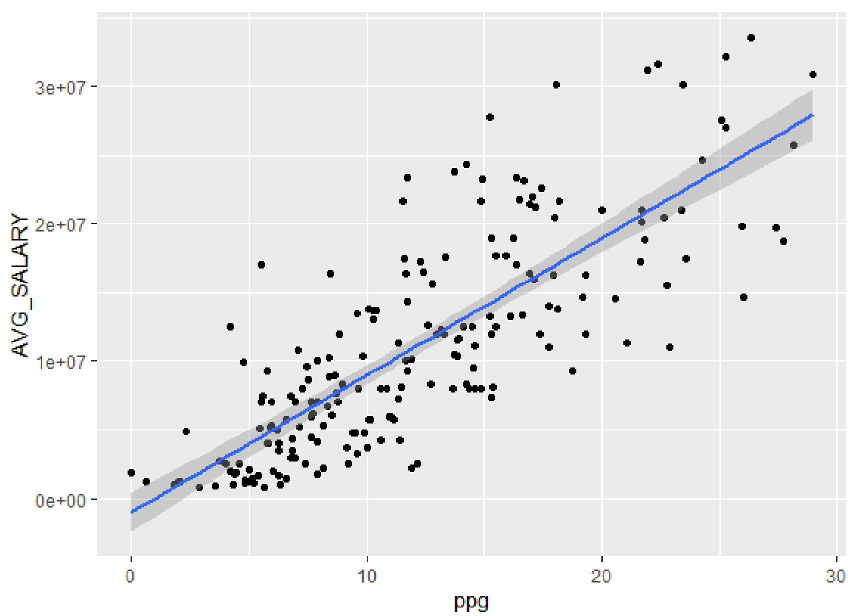
Figure 1

	BLK	PF	X...	ppg
4	17	160	-90	12.173913
3	126	206	-104	17.166667
1	142	255	-100	7.075000

As shown in Figure 1, the new ppg variable is created by dividing “PTS” by “GP”, which means dividing total number of points scored by the number of games played.

Task 3

Figure 2



As shown in Figure 2, the scatterplot with bet-fit line is created to display relationship between points per game and average salary. The slop of this line is positive, which make intuitive sense to us because usually the player who can get more points per game, which means the player makes more contribution to the team, deserves higher salary.

Task 4

Figure 3

```
        Pearson's product-moment correlation

data:  data$AVG_SALARY and data$ppg
t = 18.777, df = 197, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7448472 0.8458187
sample estimates:
      cor
0.8009573
```

As shown in Figure 3, the correlation coefficient between “AVG_SALARY” and “ppg” is 0.8, which means the correlation between these two variables is significant (the scaler of correlation coefficient is in a range from 0 to 1).

Task 5

Figure 4

```
# 5
set.seed(30)

length = count(data)$n

shuffle <- sample_n(data, length)

index = round(length*0.6)
train <-slice(shuffle, 1:index)
valid <- slice(shuffle, index+1:length)
```

As shown in Figure 4, after the seed value “30” is set, the data set is split into training set with 60% size and validation set with 40% by “slice ()” function.

Task 6

Figure 5

```
Call:
lm(formula = AVG_SALARY ~ ppg, data = train)

Residuals:
      Min        1Q    Median        3Q       Max
-10694174  -3409118   -792950   2894186  13555152

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -554470    1017509  -0.545    0.587
ppg           971665      73915  13.146   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4805000 on 117 degrees of freedom
Multiple R-squared:  0.5963,	Adjusted R-squared:  0.5928 
F-statistic: 172.8 on 1 and 117 DF,  p-value: < 2.2e-16
```

Figure 5 illustrates summary information of the simple linear regression model between input variable of "AVG_SALARY" and output variable of "ppg";

Task 7

Figure 6

```
> # 7
> data_2 <- data
> data_2$predict_salary <- predict(simple_linear_model, data_2)
> data_2$residuals <- data$AVG_SALARY-data_2$predict_salary
> # 7.a
> filter(data_2, data_2$residuals == max(data_2$residuals)) %>% select(NAME,ppg,predict_salary,AVG_SALARY,residuals)
         NAME      ppg predict_salary AVG_SALARY residuals
1 Al Horford 15.23171       14245651   27800804  13555152
> # 7.b
> filter(data_2, data_2$residuals == min(data_2$residuals)) %>% select(NAME,ppg,predict_salary,AVG_SALARY,residuals)
             NAME      ppg predict_salary AVG_SALARY residuals
1 Stephen Curry 22.89744       21694174    1.1e+07 -10694174
```

As shown in Figure 6, the player AL Horford generated the highest residual value at $13555152, his actual salary is $27800804 and the predicted value is $14245651, the residual values is calculated by subtracting the predicted value from the true salary value. Similarly, the player Stephen Curry generated the lowest residual value at -$10694174, with actual salary at $11000000 and predicted value at $21694174. It might be unfair to say that AL Horford is overpaid or Stephen Curry is overpaid. This is because, firstly, this single linear regression model is too simple to predict a very accurate or reliable result, only one variable is not enough to explain a player's salary. Secondly, there are so many other factors that a club would consider when they make salary plan for a player, such as assists ability and blocks ability.

Task 8

Figure 7

```
> # 8
> 971665 * 30 - 554470
[1] 28595480
```

According to Figure 5, the regression equation of this single liner regression model should be "predicted salary = 971665 * ppg - 554470". For example, if the input value is 30 points per game, then its predicted salary should be $28595480(See Figure 7).

Task 9

The accuracy performance of this model on training set and validation set is shown in Figure 8, both value of RMSE and MAE on validation set are lower than that in training set.

Figure 8

```
> train_acc <- accuracy(train$AVG_SALARY, predict(simple_linear_model,train))
> valid_acc <- accuracy(valid$AVG_SALARY, predict(simple_linear_model,valid))
> train_acc
                   ME    RMSE     MAE        MPE     MAPE
Test set 7.340757e-09 4764427 3755591 -0.3903789 39.17594
> valid_acc
               ME    RMSE     MAE      MPE     MAPE
Test set 303097.5 4656131 3650631 -39.4557 96.87756
>
```

Task 10

As shown in Figure 9, the standard deviation of average salary in the dataset is $7877951, wherease the model's RMSE is $4721190. Generally, the standard deviation is used to measure the dispersion of the input, while the root mean square error is used to measure the deviation between the predicted value and the true value. The two items' research objects and purposes are different, but their calculation process are similar.

Figure 9

```
> # 10
> # std <- sd(data$AVG_SALARY)
> std <- sqrt(sum((data_2$AVG_SALARY - mean(data_2$AVG_SALARY))^2)/length)
> std
[1] 7877951
> model_acc <- accuracy(data_2$AVG_SALARY, data_2$predict_salary)
> model_acc
               ME    RMSE     MAE       MPE     MAPE
Test set 121848.2 4721190 3713396 -16.09503 62.37257
```

Multiple Linear Regression

Task 1

Figure 10

```
> # PART 2
> # 1
> # identical name  no discipline
> table(data$NAME)

     Al-Farouq Aminu          Al Horford          Alec Burks     Andre Iguodala      Anthony Davis         Aron Baynes
                   1                   2                   1                  1                  1                   2
       Austin Rivers       Avery Bradley        Ben McLemore    Bismack Biyombo      Blake Griffin     Boban Marjanovic
                   3                   3                   1                  2                  1                   1
     Bojan Bogdanovic         Brook Lopez         Bryn Forbes    Carmelo Anthony         Chris Paul      Christian Wood
                   1                   2                   1                  2                  1                   1
```

The “NAME” variable will not be used as an input variable in this multiple linear regression model. This is because the “count()” function shows that there are so many categorical value here and each category contains a few number of value(See Figure 10), thus the dimension of the converted dummies or one-hot variable will be very large which significantly increase the calculation pressure of the model. In addition, if the input variable of “NAME” is a new category (a new name), then the model cannot generate a result.

Task 2

As shown in Figure 11, the exhaustive subsets method is used to build model. The 24th model is the one with the highest r-squared value among the options.

Figure 11

```
1 subsets of each size up to 24
Selection Algorithm: exhaustive
```

	CONTRACT_START	CONTRACT_END	AGE	GP	W	L	MIN	PTS	FGM	FGA	FG.	X3PM	X3PA	X3P.	FTM	FTA	FT.	OREB	DREB	REB	AST	TOV	STL	BLK	PF	X...	ppg
1 (1)	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	"*"
2 (1)	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	"*"	" "	" "	" "	" "	" "	" "	" "	"*"
3 (1)	" "	" "	" "	" "	" "	"*"	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	"*"	" "	" "	" "	" "	" "	" "	" "	"*"
4 (1)	" "	"*"	" "	" "	" "	"*"	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "	"*"	" "	" "	" "	" "	" "	" "	"*"
5 (1)	" "	"*"	" "	" "	" "	"*"	" "	" "	" "	" "	" "	" "	" "	" "	" "	"*"	" "	" "	" "	"*"	" "	" "	" "	" "	" "	" "	"*"
6 (1)	"*"	"*"	" "	" "	"*"	" "	" "	" "	" "	"*"	" "	" "	" "	" "	" "	" "	" "	" "	"*"	" "	" "	" "	" "	" "	" "	" "	"*"
7 (1)	"*"	"*"	" "	" "	"*"	" "	" "	" "	" "	"*"	" "	" "	" "	" "	" "	"*"	" "	" "	"*"	" "	" "	" "	" "	" "	" "	" "	"*"
8 (1)	"*"	"*"	" "	" "	"*"	" "	" "	" "	" "	"*"	" "	" "	" "	"*"	" "	" "	" "	" "	"*"	" "	"*"	" "	" "	" "	" "	" "	"*"
9 (1)	"*"	"*"	" "	" "	"*"	" "	" "	" "	" "	"*"	" "	" "	" "	"*"	" "	" "	" "	" "	"*"	" "	"*"	" "	" "	" "	"*"	" "	"*"
10 (1)	"*"	"*"	"*"	" "	"*"	" "	"*"	" "	" "	" "	" "	" "	" "	" "	" "	"*"	" "	" "	" "	"*"	"*"	"*"	" "	" "	" "	" "	"*"
11 (1)	"*"	"*"	"*"	" "	"*"	" "	"*"	" "	" "	" "	" "	" "	" "	" "	" "	"*"	" "	" "	" "	"*"	"*"	"*"	" "	"*"	" "	" "	"*"
12 (1)	"*"	"*"	"*"	" "	"*"	" "	"*"	" "	" "	" "	"*"	" "	" "	" "	" "	"*"	" "	" "	" "	"*"	"*"	"*"	" "	"*"	" "	" "	"*"
13 (1)	"*"	"*"	"*"	" "	"*"	" "	"*"	" "	" "	" "	"*"	" "	" "	"*"	" "	"*"	" "	" "	" "	"*"	"*"	"*"	" "	"*"	" "	" "	"*"
14 (1)	"*"	"*"	"*"	" "	"*"	" "	"*"	" "	" "	" "	"*"	" "	" "	"*"	" "	"*"	" "	" "	" "	"*"	"*"	"*"	" "	"*"	"*"	" "	"*"
15 (1)	"*"	"*"	"*"	" "	"*"	" "	"*"	" "	" "	"*"	"*"	" "	" "	"*"	" "	"*"	" "	" "	" "	"*"	"*"	"*"	" "	"*"	"*"	" "	"*"
16 (1)	"*"	"*"	"*"	"*"	"*"	" "	"*"	" "	" "	"*"	"*"	" "	" "	"*"	" "	"*"	" "	" "	" "	"*"	"*"	"*"	" "	"*"	"*"	" "	"*"
17 (1)	"*"	"*"	"*"	"*"	"*"	" "	"*"	" "	" "	"*"	"*"	" "	" "	"*"	" "	"*"	"*"	" "	" "	"*"	"*"	"*"	" "	"*"	"*"	" "	"*"
18 (1)	"*"	"*"	"*"	"*"	" "	"*"	"*"	" "	" "	"*"	"*"	" "	" "	"*"	" "	"*"	"*"	" "	" "	"*"	"*"	"*"	"*"	"*"	"*"	" "	"*"
19 (1)	"*"	"*"	"*"	"*"	"*"	" "	"*"	" "	" "	"*"	"*"	" "	" "	"*"	" "	"*"	"*"	" "	" "	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"
20 (1)	"*"	"*"	"*"	"*"	"*"	" "	"*"	"*"	" "	"*"	"*"	" "	" "	"*"	" "	"*"	"*"	" "	" "	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"
21 (1)	"*"	"*"	"*"	"*"	"*"	" "	"*"	" "	"*"	"*"	"*"	" "	"*"	"*"	" "	"*"	"*"	" "	" "	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"
22 (1)	"*"	"*"	"*"	"*"	"*"	" "	"*"	" "	"*"	"*"	"*"	"*"	"*"	"*"	" "	"*"	"*"	" "	" "	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"
23 (1)	"*"	"*"	"*"	"*"	"*"	" "	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"	" "	"*"	"*"	" "	" "	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"
24 (1)	"*"	"*"	"*"	"*"	"*"	" "	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"	" "	"*"	"*"	"*"	"*"	" "	"*"	"*"	"*"	"*"	"*"	"*"	"*"

```
> which.max(rgsbst_sum$rsq) #24
[1] 24
```

Task 3

As shown in Figure 12, except the variable “AGE”, “X3P.”, “FT.”, “BLK”, all other variables’ VIF values are greater than 5.

Figure 12

```
> vif <- vif(multiple_linear_model)
> vif <- sort(vif, decreasing=FALSE)
> barplot(vif, horiz=TRUE)
> vif
        AGE        X3P.         FT.         BLK         STL          PF        X...  CONTRACT_END         FG. CONTRACT_START
   1.901834    2.339904    2.709395    3.164501    5.283571    5.505619    5.756909      5.765360    6.223125       6.936130
       DREB         AST        OREB           W         TOV         ppg          GP           MIN         FTA           X3PM
   7.091497    8.158917    8.317521   10.684661   14.490982   15.645222   18.564358     32.402034   54.840585      93.104219
       X3PA         FGA         FGM         PTS
 114.544810  258.930724  533.505027  735.437820
```

Task 4

Figure 13 calculates the VIF value of "PTS" variable step by step. The final result is 735.4378. Although "PTS" variable has the highest VIF value, it might not be very useful in this multiple regression model since its high VIF value means there are strong multicollinearity relationship between "PTS" and some of other variables, which create problems for interpretability of particular coefficients.

Figure 13

```
> # 4 calculate vif value for PTS, since its vif is the highest
> model_PTS <- lm(PTS ~ CONTRACT_START + CONTRACT_END + AGE + GP +
+                  W+MIN + FGM + FGA + FG. + X3PM + X3PA + X3P. + FTA + FT. + OREB + DREB + AST+TOV+STL+BLK+PF+
X...+ppg,train)
> SST_PTS <- sum((train$PTS - mean(train$PTS))^2)
> SSR_PTS <- sum((model_PTS$fitted.values - mean(train$PTS))^2)
> R_squared_PTS <- SSR_PTS / SST_PTS
> R_squared_PTS  # 0.996
[1] 0.9986403
> vif_PTS <- 1/(1-R_squared_PTS)
> vif_PTS # 256.194
[1] 735.4378
```

Task 5

- **a**

As shown in Figure 14, there are many correlations higher than 0.7,

`Figure 14

```
> # 5
> cor_df <- select(train, CONTRACT_START , CONTRACT_END ,AGE,GP ,
+                  W,MIN,PTS , FGM , FGA , FG. , X3PM , X3PA , X3P. , FTA , FT. , OREB , DREB , AST,TOV,STL,BL
K,PF,X...,ppg)
> cor(cor_df)
```

	CONTRACT_START	CONTRACT_END	AGE	GP	W	MIN	PTS
CONTRACT_START	1.00000000	0.8521248486	0.3319469418	-0.07330439	-0.1734140940	-0.213335960	-0.214211902
CONTRACT_END	0.85212485	1.0000000000	0.2757646957	0.03923707	-0.0008506974	0.008999581	0.007836711
AGE	0.33194694	0.2757646957	1.0000000000	0.20084976	0.2255508300	0.236948063	0.097891317
GP	-0.07330439	0.0392370736	0.2008497559	1.00000000	0.7665771453	0.826780337	0.623369136
W	-0.17341409	-0.0008506974	0.2255508300	0.76657715	1.0000000000	0.690016365	0.564056538
MIN	-0.21333596	0.0089995807	0.2369480628	0.82678034	0.6900163654	1.000000000	0.882316512
PTS	-0.21421190	0.0078367111	0.0978913175	0.62336914	0.5640565384	0.882316512	1.000000000
FGM	-0.22582610	-0.0056181706	0.0803488747	0.65351014	0.5590383650	0.892991904	0.979705778
FGA	-0.20923303	0.0177639585	0.0940325786	0.64755726	0.5421760528	0.898581809	0.979463405
FG.	-0.10231711	-0.0807757218	-0.0006035641	0.23400251	0.2676235510	0.179262542	0.189615171
X3PM	0.06400892	0.1903449694	0.2997344381	0.36319966	0.4059170505	0.518906818	0.557189509
X3PA	0.07296079	0.2075856189	0.2958743760	0.38703715	0.4163644752	0.542474438	0.573618579
X3P.	0.10831855	0.1587318506	0.2306174979	0.16112459	0.1547991968	0.280618133	0.250240340
FTA	-0.25044429	-0.0685521909	0.0193386588	0.45125403	0.4496032237	0.701449938	0.846242633
FT.	0.02764650	0.1099502710	0.1574669873	0.29653433	0.2577915130	0.299769781	0.366219319
OREB	-0.21050117	-0.0920770781	-0.0272973500	0.44416320	0.3521640928	0.440708351	0.322233111
DREB	-0.18958909	-0.0108038161	0.1327713742	0.60453373	0.5487705525	0.739813443	0.668402351
AST	-0.08688523	0.0479296498	0.2747124614	0.38431224	0.3778815739	0.621783116	0.617911876
TOV	-0.18862467	-0.0194872974	0.1525303606	0.54917157	0.4811790152	0.809149964	0.870584410
STL	-0.26093216	-0.0802486597	0.2718239266	0.59062607	0.5349101701	0.793090431	0.712948982
BLK	-0.21804507	-0.1608132124	0.0279360238	0.36947045	0.4146122652	0.410253368	0.364595020
PF	-0.17489512	-0.0449181260	0.1699734607	0.80064047	0.6315278543	0.812513057	0.636522135
X...	-0.18472891	-0.0133621826	0.1470555770	0.18627882	0.6823612090	0.311116368	0.339487988
ppg	-0.24077832	-0.0353201052	0.0813388293	0.27435989	0.3166960861	0.666686222	0.883547050

	FGM	FGA	FG.	X3PM	X3PA	X3P.	FTA
CONTRACT_START	-0.225826104	-0.20923303	-0.1023171066	0.06400892	0.07296079	0.108318551	-0.25044429
CONTRACT_END	-0.005618171	0.01776396	-0.0807757218	0.19034497	0.20758562	0.158731851	-0.06855219
AGE	0.080348875	0.09403258	-0.0006035641	0.29973444	0.29587438	0.230617498	0.01933866
GP	0.653510136	0.64755726	0.2340025125	0.36319966	0.38703715	0.161124585	0.45125403
W	0.559038365	0.54217605	0.2676235510	0.40591705	0.41636448	0.154799197	0.44960322

- **b**

As shown in Figure 15, for each variable, the number of other highly correlated variables are counted. Generally, for each highly correlated variable pair, the one with higher count value will be remove, so that maximize the number of remained variables. After remove 12 variables, the remained variables include “CONTRACT_START”, “AGE”, “W”. etc.

Figure 15

```
> # >0.7 :
> # CONTRACT_START & CONTRACT_END                    1  y
> # CONTRACT_END & CONTRACT_START                    1  y
> # GP & W,MIN,PF                                    3  y
> # W & GP                                           1  y
> # MIN & GP | PTS,FGM,FGA,FTA,DREB,TOV,STL,PF       9  y
> # PTS & MIN| FGM,FGA,FTA,TOV,STL,ppg               7  y
> # FGM & MIN,PTS | FGA,FTA,DREB,TOV,ppg             7  y
> # FGA & MIN,PTS,FGM | FTA,TOV,STL,TOV,STL,ppg      9  y
> # FG. & OREB                                       1  y
> # X3PM & X3PA                                      1  y
> # X3PA & X3PM |                                    1  y
> # FTA & PTS,FGM,FGA | TOV,ppg                      5  y
> # OREB & FG. | DREB                                2  y
> # DREB & MIN,FGM,OREB | BLK,PF                     5  y
> # AST & TOV,STL                                    2  y
> # TOV & MIN,PTS,FGM,FGA,FTA,AST | STL,ppg          8  y
> # STL & MIN,FGA,AST,TOV |                          4  y
> # BLK & DREB  |                                    1  y
> # PF & GP,MIN,DREB |                               3  y
> # ppg & PTS,FGM,FGA,FTA,TOV |                      5  y
>
> # remove  CONTRACT_END, MIN, FGA, TOV, PTS,  FTA, X3PA, ppg, DREB, OREB,STL,GP  # 12
>
> # remain 12
> cor_df_2 <- select(train, CONTRACT_START, AGE, W, FGM, FG., X3PM, X3P., FT., AST, BLK, PF, X...)
> cor(cor_df_2)
               CONTRACT_START          AGE          W         FGM            FG.       X3PM        X3P.
CONTRACT_START     1.00000000 0.3319469418 -0.1734141 -0.22582610 -0.1023171066 0.06400892 0.108318551
AGE                0.33194694 1.0000000000  0.2255508  0.08034887 -0.0006035641 0.29973444 0.230617498
W                 -0.17341409 0.2255508300  1.0000000  0.55903837  0.2676235510 0.40591705 0.154799197
```

Task 6

The summary of new model is shown in Figure 16

Figure 16

```
> # 6
> new_model <- lm(AVG_SALARY ~ CONTRACT_START+ AGE+W+ FGM+ FG.+ X3PM+ X3P.+ FT.+ AST+ BLK+ PF+ X..., train)
> summary(new_model)  # R-squared: 0.6566

Call:
lm(formula = AVG_SALARY ~ CONTRACT_START + AGE + W + FGM + FG. +
    X3PM + X3P. + FT. + AST + BLK + PF + X..., data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-10279074 -3347732  -609672  2943040 12155909

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -511583737  506525710  -1.010 0.314802
CONTRACT_START     253332     252339   1.004 0.317696
AGE                353912     194148   1.823 0.071138 .
W                 -104935      60092  -1.746 0.083669 .
FGM                 30108       5061   5.949 3.52e-08 ***
FG.                 -6895     112484  -0.061 0.951238
X3PM                 3182      12282   0.259 0.796082
X3P.              -132420      46993  -2.818 0.005768 **
FT.                 16528      42976   0.385 0.701312
AST                  4839       4005   1.208 0.229592
BLK                 29673      16196   1.832 0.069748 .
PF                 -15302      12377  -1.236 0.219085
X...                10951       3156   3.470 0.000753 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4804000 on 106 degrees of freedom
Multiple R-squared:  0.6344,	Adjusted R-squared:  0.593
F-statistic: 15.33 on 12 and 106 DF,  p-value: < 2.2e-16
```

Task 7

As shown in Figure 17, all of the VIF values in the new model are less than 5.

Figure 17

```
> # 7
> vif(new_model)
CONTRACT_START            AGE              W            FGM            FG.           X3PM           X3P.
      1.325351       1.472939       3.975937       4.055294       2.393421       3.073716       1.741559
           FT.            AST            BLK             PF           X...
      1.496167       2.254605       2.458623       3.192831       2.694182
```

Task 8

The calculated SST of the new model is 6.69103e+15 (See Figure 18).

Figure 18

```
> SST_new_model <- sum((train$AVG_SALARY - mean(train$AVG_SALARY) )^2)
> SST_new_model
[1] 6.69103e+15
```

Task 9

The calculated SSR of the new model is 4.244933e+15.

Figure 19

```
> # 9
> SSR_new_model <- sum((new_model$fitted.values - mean(train$AVG_SALARY))^2)
> SSR_new_model
[1] 4.244933e+15
```

Task 10

As shown in Figure 20, the calculated result of new model's SSR/SST, which is R-Squared, equals 0.6344, which shows the same result as the summary information of the new model.

Figure 20

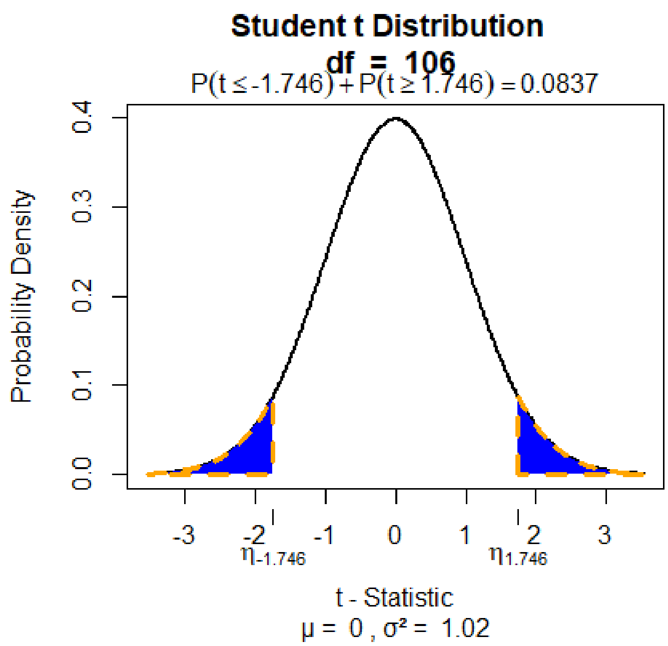
```
> summary(new_model)  # R-squared : 0.6344

Call:
lm(formula = AVG_SALARY ~ CONTRACT_START + AGE + W + FGM + FG. +
    X3PM + X3P. + FT. + AST + BLK + PF + X..., data = train)

Residuals:
      Min        1Q    Median        3Q       Max
-10279074  -3347732   -609672   2943040  12155909

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -511583737  506525710  -1.010 0.314802
CONTRACT_START     253332     252339   1.004 0.317696
AGE                353912     194148   1.823 0.071138 .
W                 -104935      60092  -1.746 0.083669 .
FGM                 30108       5061   5.949 3.52e-08 ***
FG.                 -6895     112484  -0.061 0.951238
X3PM                 3182      12282   0.259 0.796082
X3P.              -132420      46993  -2.818 0.005768 **
FT.                 16528      42976   0.385 0.701312
AST                  4839       4005   1.208 0.229592
BLK                 29673      16196   1.832 0.069748 .
PF                 -15302      12377  -1.236 0.219085
X...                10951       3156   3.470 0.000753 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4804000 on 106 degrees of freedom
Multiple R-squared:  0.6344,    Adjusted R-squared:  0.593
F-statistic: 15.33 on 12 and 106 DF,  p-value: < 2.2e-16
```

Task 11

The t value of “W” variable is -1.746, its t-distribution plot is shown in Figure 21, its freedom degree equals “n-p-1”, which is (119 - 12 -1) = 106. The shaded percent equals 0.08367, which means there is about 92% percent of confidence interval to think the “W” variable is correlated with “AVG_SALARY” variable.

Figure 21

```
> # 11
> library(visualize)
> # df means degrees of freedom
> train_length <- count(train)$n
> freedom_degree <- train_length - 12 - 1
> visualize.t(stat = c(-1.746,1.746), df=freedom_degree, section= "tails")
```



Task 12

The F-value of new model is 15.33 (See Figure 22).

Figure 22

```
# 12
SSE_new_model <- SST_new_model - SSR_new_model
F_value <- (SSR_new_model/12) / (SSE_new_model/(train_length-12-1))
F_value # 15.33
visualize.f(stat=15.32928, section = "upper")
```

Task 13

A fictional basketball player called IVAN is created, with several assumed input values. The predicted salary is $12898792 (See Figure 23).

Figure 23

```
> # 13
> # CONTRACT_START:2019 AGE: 24  Name:Ivan  w:40, FGM: 250  FG. 50  x3pm:30  x3p.:30
> #  FT. : 80  AST:200  BLK:20   pf: 50  X...: 300
> # get salary: 12898792
> predict(new_model, data.frame(CONTRACT_START=2019,AGE=24,W=40,FGM=250,FG.=50,X3PM=30,X3P.=30,FT.=80,AST=200,BLK=20,PF=50,X...=300))
       1
12898792
```

Task 14

To compare the accuracy between SLR model and MLR model, the RMSE and MAE are decreased on both test set and validation set, which means the performance of MLR model is better than SLR model. In addition, the MAPE of SLR reaches 96.88, which demonstrates that the performance of SLR on validation set is not good.

Figure 24

```
> # train
> accuracy(train$AVG_SALARY, predict(simple_linear_model, train))
                   ME    RMSE     MAE        MPE     MAPE
Test set 7.340757e-09 4764427 3755591 -0.3903789 39.17594
> accuracy(train$AVG_SALARY, predict(new_model, train))
                    ME    RMSE     MAE       MPE     MAPE
Test set -1.080722e-07 4533810 3697392 -3.448333 41.61474
> # valid
> accuracy(valid$AVG_SALARY, predict(simple_linear_model, valid))
               ME    RMSE     MAE      MPE     MAPE
Test set 303097.5 4656131 3650631 -39.4557 96.87756
> accuracy(valid$AVG_SALARY, predict(new_model, valid))
               ME    RMSE     MAE      MPE     MAPE
Test set 637880.6 4440697 3439789 8.148562 40.13201
```