Assignment_1

Mrunali Vikas Patil
2023-09-29

#PROBLEM SET 1

```
BankingData = read.table("Banking.txt", header = T)
print(BankingData)
```

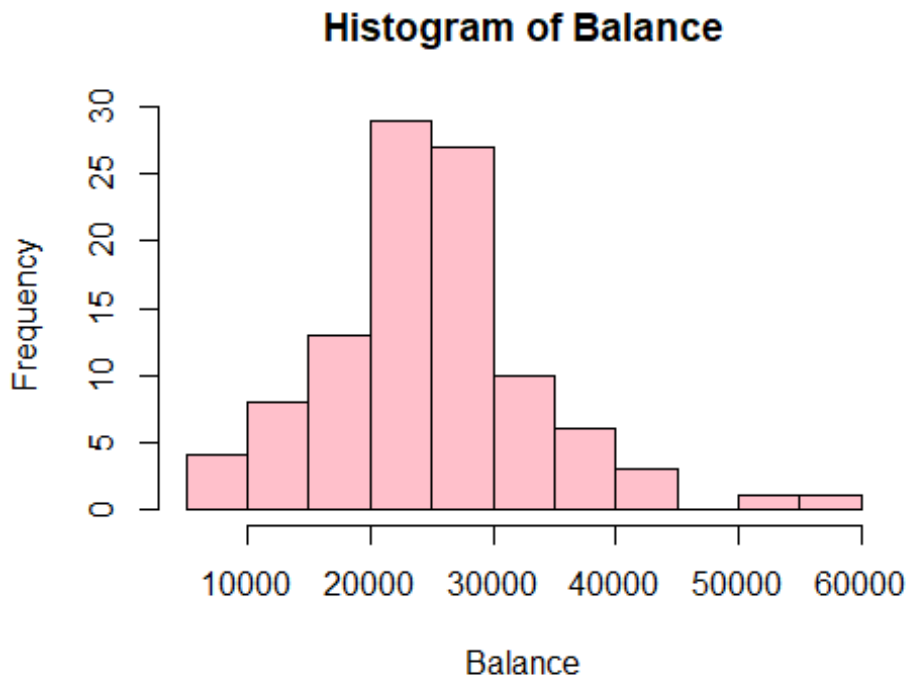
```
##      Age Education Income Balance
## 1   35.9      14.8  91033   38517
## 2   37.7      13.8  86748   40618
## 3   36.8      13.8  72245   35206
## 4   35.3      13.2  70639   33434
## 5   35.3      13.2  64879   28162
## 6   34.8      13.7  75591   36708
## 7   39.3      14.4  80615   38766
## 8   36.6      13.9  76507   34811
## 9   35.7      16.1 107935   41032
## 10  40.5      15.1  82557   41742
## 11  37.9      14.2  58294   29950
## 12  43.1      15.8  88041   51107
## 13  37.7      12.9  64597   34936
## 14  36.0      13.1  64894   32387
## 15  40.4      16.1  61091   32150
## 16  33.8      13.6  76771   37996
## 17  36.4      13.5  55609   24672
## 18  37.7      12.8  74091   37603
## 19  36.2      12.9  53713   26785
## 20  39.1      12.7  60262   32576
## 21  39.4      16.1 111548   56569
## 22  36.1      12.8  48600   26144
## 23  35.3      12.7  51419   24558
## 24  37.5      12.8  51182   23584
## 25  34.4      12.8  60753   26773
## 26  33.7      13.8  64601   27877
## 27  40.4      13.2  62164   28507
## 28  38.9      12.7  46607   27096
## 29  34.3      12.7  61446   28018
## 30  38.7      12.8  62024   31283
## 31  33.4      12.6  54986   24671
## 32  35.0      12.7  48182   25280
## 33  38.1      12.7  47388   24890
## 34  34.9      12.5  55273   26114
## 35  36.1      12.9  53892   27570
## 36  32.7      12.6  47923   20826
```

```
## 37  37.1      12.5  46176   23858
## 38  23.5      13.6  33088   20834
## 39  38.0      13.6  53890   26542
## 40  33.6      12.7  57390   27396
## 41  41.7      13.0  48439   31054
## 42  36.6      14.1  56803   29198
## 43  34.9      12.4  52392   24650
## 44  36.7      12.8  48631   23610
## 45  38.4      12.5  52500   29706
## 46  34.8      12.5  42401   21572
## 47  33.6      12.7  64792   32677
## 48  37.0      14.1  59842   29347
## 49  34.4      12.7  65625   29127
## 50  37.2      12.5  54044   27753
## 51  35.7      12.6  39707   21345
## 52  37.8      12.9  45286   28174
## 53  35.6      12.8  37784   19125
## 54  35.7      12.4  52284   29763
## 55  34.3      12.4  42944   22275
## 56  39.8      13.4  46036   27005
## 57  36.2      12.3  50357   24076
## 58  35.1      12.3  45521   23293
## 59  35.6      16.1  30418   16854
## 60  40.7      12.7  52500   28867
## 61  33.5      12.5  41795   21556
## 62  37.5      12.5  66667   31758
## 63  37.6      12.9  38596   17939
## 64  39.1      12.6  44286   22579
## 65  33.1      12.2  37287   19343
## 66  36.4      12.9  38184   21534
## 67  37.3      12.5  47119   22357
## 68  38.7      13.6  44520   25276
## 69  36.9      12.7  52838   23077
## 70  32.7      12.3  34688   20082
## 71  36.1      12.4  31770   15912
## 72  39.5      12.8  32994   21145
## 73  36.5      12.3  33891   18340
## 74  32.9      12.4  37813   19196
## 75  29.9      12.3  46528   21798
## 76  32.1      12.3  30319   13677
## 77  36.1      13.3  36492   20572
## 78  35.9      12.4  51818   26242
## 79  32.7      12.2  35625   17077
## 80  37.2      12.6  36789   20020
## 81  38.8      12.3  42750   25385
## 82  37.5      13.0  30412   20463
## 83  36.4      12.5  37083   21670
## 84  42.4      12.6  31563   15961
## 85  19.5      16.1  15395    5956
## 86  30.5      12.8  21433   11380
```

```
## 87  33.2       12.3  31250   18959
## 88  36.7       12.5  31344   16100
## 89  32.4       12.6  29733   14620
## 90  36.5       12.4  41607   22340
## 91  33.9       12.1  32813   26405
## 92  29.6       12.1  29375   13693
## 93  37.5       11.1  34896   20586
## 94  34.0       12.6  20578   14095
## 95  28.7       12.1  32574   14393
## 96  36.1       12.2  30589   16352
## 97  30.6       12.3  26565   17410
## 98  22.8       12.3  16590   10436
## 99  30.3       12.2   9354    9904
## 100 22.0       12.0  14115    9071
## 101 30.8       11.9  17992   10679
## 102 35.1       11.0   7741    6207
```

Question A)

```
library(psych)
hist(BankingData$Balance,main= "Histogram of Balance",xlab = "Balance",col =
"pink",border = "black")
```




```
summary(BankingData$Balance)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5956   20036   24661   24888   29180   56569
```

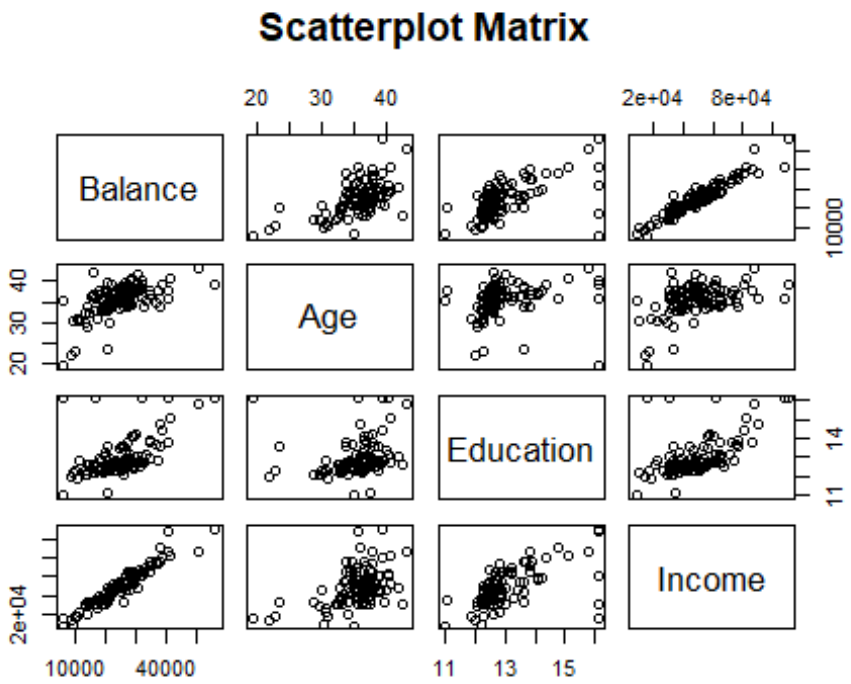
```
describe(BankingData$Balance)

##    vars   n     mean      sd  median  trimmed     mad  min   max range
skew
## X1    1 102 24887.88 8697.81 24660.5 24546.44 6834.04 5956 56569 50613
0.59
##    kurtosis     se
## X1     1.27 861.21
```

#The histogram of average account balance appears to be slightly right-skewed #Mean of the data is 24888 which is slightly higher than the median of the data which is 24660 as these values are almost similar we can say that the data is slightly normally distributed

Question B)

```
pairs(~Balance+Age+Education+Income,data = BankingData, main="Scatterplot
Matrix")
```



#There is a significantly stronger positive linear relationship observed in the scatterplot of income and balance. #There is not a clear linear relationship observed in the Education vs. Balance scatterplot. #A rather weak linear relationship between age and balance is visible in the scatterplot. #There appear to be a few outliers in the scatterplots between education and balance.

Question C)

```
cor(BankingData)
```

```
##                 Age Education    Income   Balance
## Age       1.0000000 0.1734071 0.4771474 0.5654668
## Education 0.1734071 1.0000000 0.5753940 0.5548807
## Income    0.4771474 0.5753940 1.0000000 0.9516845
## Balance   0.5654668 0.5548807 0.9516845 1.0000000
```

#The correlation between Income and Balance is positive and moderate, indicating Strong Corelation.

#The correlation between Education and Balance as well as with age and balance is low, suggesting a weak association.

Question D)

#Balance is dependent variable while the Age, Education and Income are Independent Variable.

Question E)

```
Rfit <- lm(Balance~Age+Education+Income,data = BankingData)
summary(Rfit)
```

```
##
## Call:
## lm(formula = Balance ~ Age + Education + Income, data = BankingData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7722.0 -1547.4   -56.1  1167.9  8480.2
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -9.540e+03  4.423e+03  -2.157   0.0335 *
## Age          3.325e+02  7.234e+01   4.597 1.28e-05 ***
## Education    2.887e+02  3.005e+02   0.960   0.3392
## Income       3.871e-01  1.748e-02  22.137  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2458 on 98 degrees of freedom
## Multiple R-squared:  0.9225, Adjusted R-squared:  0.9201
## F-statistic: 388.8 on 3 and 98 DF,  p-value: < 2.2e-16
```

#Income has the most significant effect on balance as the pvalue for this coefficient in the summary output is <2e-16 *

Question F)

#As per the summary pvalue for the education is greatest, we remove this predictor and refit the new regression model.

```
Rfit2 <- lm(Balance~Age+Income,data = BankingData)
summary(Rfit2)
```

```
##
## Call:
## lm(formula = Balance ~ Age + Income, data = BankingData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7385.1 -1577.9  -119.2  1200.6  8362.9
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.912e+03  2.301e+03  -2.570   0.0117 *
## Age          3.227e+02  7.159e+01   4.508  1.8e-05 ***
## Income       3.966e-01  1.437e-02  27.600  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2457 on 99 degrees of freedom
## Multiple R-squared:  0.9218, Adjusted R-squared:  0.9202
## F-statistic: 583.2 on 2 and 99 DF,  p-value: < 2.2e-16
```

#The fitted regression model without Education is: #Balance = B0 + B1*AGE + B2*INCOME
#Balance = -5.912e+03 + 3.227e+02*AGE + 3.966e-01*INCOME

Question G)

B0 = -5.912e+03 (Y Intercept)

B1 = 3.227e (Represents the Age)

B2 = 3.966e-01 (Represents the Income)

Question H)

R-squared

```
summary(Rfit2)$r.squared
```

```
## [1] 0.9217638
```

#R Squared for the model is 0.9218 which is higher. A higher R2 indicates that more of the variance in the dependent variable has been determined by the model.

Question I)

```
predicted_balance <- predict(Rfit2, newdata = data.frame(Age = 34.8,
Education = 12.5, Income = 42401))
predicted_balance
```

```
##        1
## 22135.22
```

#The model prediction error is the difference between the predicted value and the observed value

```
predicted_balance <-22135.22
observed_balance <-21572

Error <-predicted_balance - observed_balance
print(Error)

## [1] 563.22

Error <-563.22
```

Question J)

Global F-test for model adequacy

```
anova(Rfit2)

## Analysis of Variance Table
##
## Response: Balance
##           Df     Sum Sq    Mean Sq F value    Pr(>F)
## Age        1 2443180856 2443180856  404.61 < 2.2e-16 ***
## Income     1 4599872720 4599872720  761.78 < 2.2e-16 ***
## Residuals 99  597790568    6038289
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

H0 = Independent variables are not significant
Ha = Independent variables are significant
As observed, P value for both the variables is less than 0.05 hence both the variables are important.

p-value: < 2.2e-16 for the F test. Hence, we can conclude that model is significant.

#PROBLEM SET 2

Question 1 :

Numsex is 1 if sex is Male, Numsex is 0 if sex is Female. # College_Business = 1 if College = Business. College_Business = 0 if College not = Business

Question 2 :

#Refer below mentioned equation which shows general regression equation.

y = B0 + numsex*B1 + College_Business*B2

Question 3 :

That means that the category business and engineering have the same effect on salary, so we can use one variable for both categories. # y =B0 + numsex*B1 + college_Business_Engineering*B2 + college_Nursing*B2 + college_liberal_Arts*B4

#PROBLEM SET 3

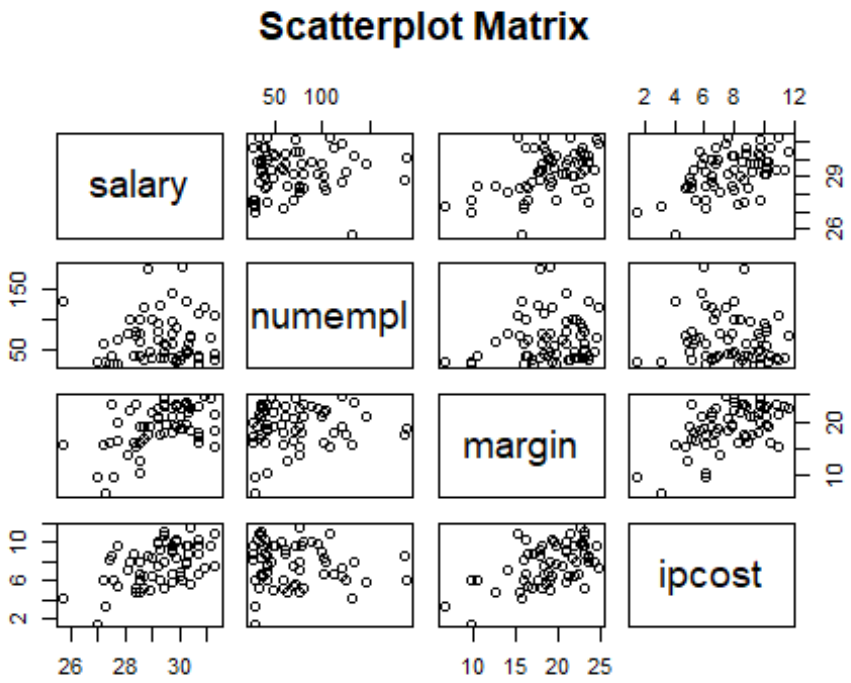
```
SalaryData = read.table("salary_IS.txt", header = T)
print(SalaryData)
```

```
##    salary numempl margin ipcost
## 1    29.5      58   19.4  10.14
## 2    29.3      37   17.7   9.18
## 3    29.2      69   20.5   7.59
## 4    28.9      48   19.1   4.96
## 5    27.5      42   23.4   8.61
## 6    29.4      37   23.1  10.72
## 7    30.4      71   18.5   5.65
## 8    27.7      69   16.4   5.46
## 9    30.9     121   24.6   7.37
## 10   29.7      99   20.9   9.05
## 11   30.3      62   23.0   8.81
## 12   31.3     107   15.3  10.94
## 13   30.0      42   18.8   6.84
## 14   30.0      35   21.0   6.45
## 15   28.5      42   10.5   6.06
## 16   29.9      31   19.3  10.20
## 17   29.7      78   18.0   9.60
## 18   30.2     132   23.5   7.88
## 19   29.7      37   22.4   6.71
## 20   29.9      89   22.8  10.04
## 21   29.0     101   21.7   8.39
```

```
## 22   29.4      60   18.0   5.24
## 23   30.3      48   21.9   9.60
## 24   30.4      75   22.6  11.63
## 25   31.1      71   24.5   9.65
## 26   29.4      47   24.2   7.94
## 27   30.7      39   22.7   9.67
## 28   30.2      50   23.1   9.66
## 29   30.7      40   16.1  10.31
## 30   28.5     102   16.2   6.67
## 31   29.0      38   21.9   6.45
## 32   29.2      80   20.9  10.07
## 33   28.1      77   14.0   7.06
## 34   27.7      28   19.8   9.70
## 35   27.3      30    6.7   3.16
## 36   31.3      34   21.4  10.91
## 37   27.4      28   16.0   8.19
## 38   28.7     121   19.3   6.42
## 39   29.7     146   20.9   5.74
## 40   29.3     124   17.6   6.13
## 41   28.3      40   16.3   8.86
## 42   25.7     130   15.6   4.11
## 43   27.2      60   15.9   6.13
## 44   29.2      94   22.6   9.95
## 45   30.2      43   19.6   7.83
## 46   30.7     111   18.2   6.70
## 47   29.4      37   23.0  11.25
## 48   28.4      76   15.5   4.77
## 49   30.1     188   18.9   5.94
## 50   28.5      64   12.6   4.81
## 51   28.8     185   17.7   8.66
## 52   28.4      81   23.1   5.14
## 53   29.7      62   20.9   9.26
## 54   27.0      30    9.8   1.44
## 55   28.2     103   22.1   7.98
## 56   27.6      29    9.7   6.09
## 57   30.7      28   17.1   8.71
## 58   28.7      34   16.8   5.11
## 59   29.9      35   23.4   8.42
## 60   31.3      43   18.3   7.52
## 61   28.5      77   19.0   7.85
```

Question 1 :

```
pairs(~salary+numempl+margin+ipcost,data = SalaryData, main="Scatterplot
Matrix")
```




```
cor(SalaryData)
```

```
##              salary     numempl    margin      ipcost
## salary  1.00000000  0.04267267 0.4988443  0.52975765
## numempl 0.04267267  1.00000000 0.1257754 -0.09667573
## margin  0.49884432  0.12577542 1.0000000  0.55409931
## ipcost  0.52975765 -0.09667573 0.5540993  1.00000000
```

#Salary is not linearly related to the three predictors

#Margin and Ipcost are more strongly related in comparison with others.

Question 2 :

```
Fit1 <- lm(salary~numempl+margin+ipcost,data = SalaryData)
summary(Fit1)
```

```
##
## Call:
## lm(formula = salary ~ numempl + margin + ipcost, data = SalaryData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.60018 -0.68845  0.04395  0.57401  2.16937
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.900686   0.681101  38.028  < 2e-16 ***
```

```
## numempl      0.001356   0.003471   0.391  0.69747
## margin       0.087484   0.040529   2.159  0.03511 *
## ipcost       0.208865   0.073137   2.856  0.00598 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9997 on 57 degrees of freedom
## Multiple R-squared:  0.3432, Adjusted R-squared:  0.3087
## F-statistic: 9.929 on 3 and 57 DF,  p-value: 2.307e-05
```

Question 3 :

```
summary(Fit1)

##
## Call:
## lm(formula = salary ~ numempl + margin + ipcost, data = SalaryData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.60018 -0.68845  0.04395  0.57401  2.16937
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.900686   0.681101  38.028  < 2e-16 ***
## numempl      0.001356   0.003471   0.391  0.69747
## margin       0.087484   0.040529   2.159  0.03511 *
## ipcost       0.208865   0.073137   2.856  0.00598 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9997 on 57 degrees of freedom
## Multiple R-squared:  0.3432, Adjusted R-squared:  0.3087
## F-statistic: 9.929 on 3 and 57 DF,  p-value: 2.307e-05
```

#The variable ipcost has most significant effect on salary as it has less p value for the t-test.

Question 4 :

#We can see from the model's prior report that the numempl influence is less important as it has greater p-value. Particular variable is therefore not included in the refit model.

```
Fit2 <- lm(salary~ipcost+margin,data = SalaryData)
summary(Fit2)

##
## Call:
## lm(formula = salary ~ ipcost + margin, data = SalaryData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -2.5260 -0.5797  0.0490  0.6611  2.1359
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.97304    0.65063  39.920  < 2e-16 ***
## ipcost       0.20311    0.07111   2.856  0.00594 **
## margin       0.09091    0.03928   2.315  0.02420 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9923 on 58 degrees of freedom
## Multiple R-squared:  0.3415, Adjusted R-squared:  0.3188
## F-statistic: 15.04 on 2 and 58 DF,  p-value: 5.479e-06
```

#The slope coefficient for "ipcost" is 0.20311, and the slope coefficient for "margin" is 0.09091.

Question 5 :

```
r_squared <- summary(Fit2)$r.squared
print(r_squared)

## [1] 0.3414687
```

#Adjusted R squared for the first model with all 3 independent variables is 0.3087, while for the second fit model is 0.3188. As R Squared value for the second model is greater than the first model we can say that second model performed better.

Question 6 :

```
Anova <- anova(Fit2)
print(Anova)

## Analysis of Variance Table
##
## Response: salary
##           Df Sum Sq Mean Sq F value    Pr(>F)
## ipcost     1 24.340 24.3396 24.7176 6.225e-06 ***
## margin     1  5.275  5.2753  5.3572    0.0242 *
## Residuals 58 57.113  0.9847
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

#H0 : The vairables does not have relation with salary #Ha : The variables have relation with salary

#As per summary of Anova test we can see that pvalue for both attributes is less than significance level hence both variables are important.

Question 7 :

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(QuantPsyc)
```

```
## Loading required package: boot
```

```
##
## Attaching package: 'boot'
```

```
## The following object is masked from 'package:psych':
##
##     logit
```

```
## Loading required package: purrr
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
##
## Attaching package: 'QuantPsyc'
```

```
## The following object is masked from 'package:base':
##
##     norm
```

```
lm.beta(Fit2)
```

```
##    ipcost    margin
## 0.3655958 0.2962680
```

#The variable “ipcost” has a major effect on salary, as determined by the second refit model’s standardized coefficients.

#PROBLEM SET 4

H0 is mean heart rate = 71 beats/minute

H1 is mean heart rate > 71 beats/minute

significance level (α) = 0.05

sample mean (X) is 73.5 beats/minute

sample standard deviation (s) = 6 beats/minute

sample size (n) = 90

population mean (μ) = 71

$Z = X-\mu/(s/\sqrt{n})$

$Z = 73.5-71/(6/\sqrt{90})$

$Z = 3.95$

The Z value at $\alpha = 0.05$ is 1.645.

Since the calculated Z-value (3.95) is greater than the Z value at $\alpha = 0.05$ (1.6622), we reject the null hypothesis (H0).

The t-test results provide sufficient evidence to conclude that, at the 0.05 level of significance, the true mean heart rate during laughter exceeds 71 beats per minute.