Who got how much salary?

For this case, a subset of people who have a salary(have a job) is taken.

#subset of people who have a job(have a salary)

#excludes entries where salary is 0,998 or 999

got\_a\_job <- mba[mba$salary!=0 & mba$salary!=998 & mba$salary!=999, ]

got\_a\_job

# A tibble: 103 x 13

age sex gmat\_tot gmat\_qpc gmat\_vpc gmat\_tpc s\_avg f\_avg quarter work\_yrs frstlang salary

<int> <int> <int> <int> <int> <int> <dbl> <dbl> <int> <int> <int> <int>

1 22 2 660 90 92 94 3.5 3.75 1 1 1 85000

2 27 2 700 94 98 98 3.3 3.25 1 2 1 85000

3 25 2 680 87 96 96 3.5 2.67 1 2 1 86000

4 25 2 650 82 91 93 3.4 3.25 1 3 1 88000

5 27 1 710 96 96 98 3.3 3.50 1 2 1 92000

6 28 2 620 52 98 87 3.4 3.75 1 5 1 93000

7 24 1 670 84 96 95 3.3 3.25 1 0 1 95000

8 25 2 560 52 81 72 3.3 3.50 1 1 1 95000

9 25 2 530 50 62 61 3.6 3.67 1 3 1 95000

10 25 1 650 79 93 93 3.3 3.50 1 1 1 96000

# ... with 93 more rows, and 1 more variables: satis <int>

summary(got\_a\_job)

age sex gmat\_tot gmat\_qpc gmat\_vpc gmat\_tpc

Min. :22.00 Min. :1.000 Min. :500 Min. :39.00 Min. :30.00 Min. :51.00

1st Qu.:25.00 1st Qu.:1.000 1st Qu.:580 1st Qu.:72.00 1st Qu.:71.00 1st Qu.:78.00

Median :26.00 Median :1.000 Median :620 Median :82.00 Median :81.00 Median :87.00

Mean :26.78 Mean :1.301 Mean :616 Mean :79.73 Mean :78.56 Mean :84.52

3rd Qu.:28.00 3rd Qu.:2.000 3rd Qu.:655 3rd Qu.:89.00 3rd Qu.:92.00 3rd Qu.:93.50

Max. :40.00 Max. :2.000 Max. :720 Max. :99.00 Max. :99.00 Max. :99.00

s\_avg f\_avg quarter work\_yrs frstlang

Min. :2.200 Min. :0.000 Min. :1.000 Min. : 0.00 Min. :1.000

1st Qu.:2.850 1st Qu.:2.915 1st Qu.:1.000 1st Qu.: 2.00 1st Qu.:1.000

Median :3.100 Median :3.250 Median :2.000 Median : 3.00 Median :1.000

Mean :3.092 Mean :3.091 Mean :2.262 Mean : 3.68 Mean :1.068

3rd Qu.:3.400 3rd Qu.:3.415 3rd Qu.:3.000 3rd Qu.: 4.00 3rd Qu.:1.000

Max. :4.000 Max. :4.000 Max. :4.000 Max. :16.00 Max. :2.000

salary satis

Min. : 64000 Min. :3.000

1st Qu.: 95000 1st Qu.:5.000

Median :100000 Median :6.000

Mean :103031 Mean :5.883

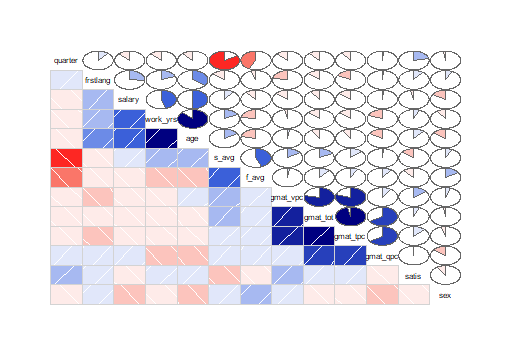
3rd Qu.:106000 3rd Qu.:6.000

Max. :220000 Max. :7.000

library(corrgram)

> corrgram(got\_a\_job, order=TRUE, lower.panel=panel.shade,

+ upper.panel=panel.pie, text.panel=panel.txt)



Hypothesis 1:

Males get a higher starting salary compared to females.

aggregate(got\_a\_job$salary, by=list(got\_a\_job$sex) , mean)

Group.1 x

1 1 104970.97

2 2 98524.39

As seen, it is the males who have a higher average salary as compared to the females.

Let us run a t test to confirm the same.

T TEST

H0: Females and males have the same salary. Difference between the mean salary of females and mean salary of males is 0

H1: Males have a higher salary compared to the females. Difference between the mean salary of females and mean salary of males is not 0.

t.test(got\_a\_job$salary, got\_a\_job$sex)

Welch Two Sample t-test

data: got\_a\_job$salary and got\_a\_job$sex

t = 58.517, df = 102, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

99537.17 106521.71

sample estimates:

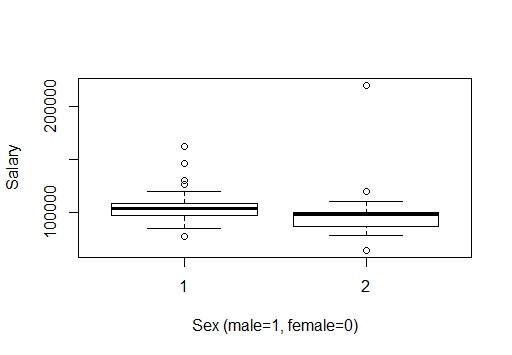
mean of x mean of y

1.030307e+05 1.300971e+00

**P value is very less(<0.001). Therefore we reject the null hypothesis and conclude that the males have a higher salary compared to the females.**

boxplot(salary ~ sex, data=got\_a\_job,

+ xlab="Sex (male=1, female=0)", ylab="Salary")



Let us run a chi square test on the salary and sex.

Null hypothesis states that the salary and sex are independent of each other.

Alternate hypothesis states that there is a dependency between salary and sex

mytable <- xtabs(~salary+sex, data=got\_a\_job)

> chisq.test(mytable)

Pearson's Chi-squared test

data: mytable

X-squared = 52.681, df = 41, p-value = 0.1045

P value is found to be more than 0.05. Therefore we are unable to reject the null hypothesis. We conclude that the null hypothesis and the alternate hypothesis are plausible.

Hypothesis 2:

People who have English as their first language earn a better salary than other people.

aggregate(got\_a\_job$salary, by=list(got\_a\_job$frstlang) , mean)

Group.1 x

1 1 101748.6

2 2 120614.3

As seen, it is the people who have English as first language who have a higher average salary as compared to the others.

Let us run a t test to confirm the same.

T TEST

H0: People who have English as first language as well as people who have other languages as first language have the same salary. Difference between the mean salary of both of them is 0

H1: People who have English as first language have a higher salary compared to the people who do not have English as their first language. Difference between the mean salary of both of them is not 0.

t.test(got\_a\_job$salary, got\_a\_job$frstlang)

Welch Two Sample t-test

data: got\_a\_job$salary and got\_a\_job$frstlang

t = 58.517, df = 102, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

99537.4 106521.9

sample estimates:

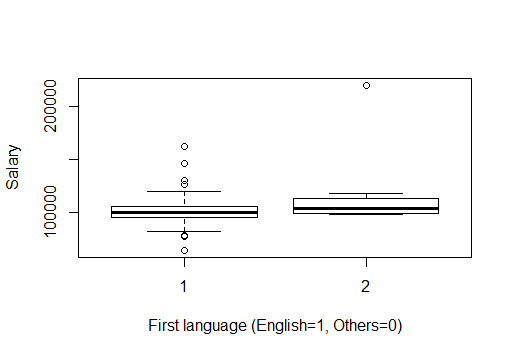
mean of x mean of y

1.030307e+05 1.067961e+00

**P value is very less(<0.001). Therefore we reject the null hypothesis and conclude that the people who have English as their first language have a higher salary compared to the others.**

boxplot(salary ~ frstlang, data=got\_a\_job,

+ xlab="First language (English=1, Others=0)", ylab="Salary")



Let us run a chi square test on the salary and first language.

Null hypothesis states that the salary and first language are independent of each other.

Alternate hypothesis states that there is a dependency between salary and first language

mytable <- xtabs(~salary+frstlang, data=got\_a\_job)

> chisq.test(mytable)

Pearson's Chi-squared test

data: mytable

X-squared = 69.847, df = 41, p-value = 0.003296

P value is found to be less than 0.05. Therefore we reject the null hypothesis.

**SUMMARY OF THE ABOVE ANALYSIS USING CONTINGENCY TABLES**

xtabs(~sex+frstlang, got\_a\_job)

frstlang

sex 1 2

1 68 4

2 28 3

From the above analysis, we can see that there are more number of males having English as first language.

That can also explain the higher salary of males as compared to females.

**Regression model**

1. **The salary depends on the age and the work experience.**

cor(got\_a\_job$salary, got\_a\_job$age)

[1] 0.4996428

cor(got\_a\_job$salary, got\_a\_job$work\_yrs)

[1] 0.4546663

fit <- lm(salary~work\_yrs+age,data=got\_a\_job)

> summary(fit)

Call:

lm(formula = salary ~ work\_yrs + age, data = got\_a\_job)

Residuals:

Min 1Q Median 3Q Max

-31675 -8099 -2108 4411 80650

Coefficients:

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

(Intercept) 36967.5 23323.8 1.585 0.1161

work\_yrs 388.8 1084.0 0.359 0.7206

age 2413.8 997.4 2.420 0.0173 \*

---

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

Residual standard error: 15620 on 100 degrees of freedom

Multiple R-squared: 0.2506, Adjusted R-squared: 0.2356

F-statistic: 16.72 on 2 and 100 DF, p-value: 5.438e-07

**Regression model formulated is**

**Salary= [388.8(work experience)]+ [2413.8(age)] + 36967.5**

got\_a\_job$salary

[1] 85000 85000 86000 88000 92000 93000 95000 95000 95000 96000 96000 100000 100000

[14] 100000 105000 105000 105000 105000 105000 105000 106000 106000 107500 108000 110000 112000

[27] 115000 115000 118000 120000 120000 120000 120000 146000 162000 82000 92000 93000 95000

[40] 95000 96000 96500 98000 98000 98000 99000 100000 100000 101000 103000 104000 105000

[53] 105000 105000 107000 112000 115000 115000 130000 145800 78256 88500 90000 90000 93000

[66] 95000 97000 97000 98000 98000 98000 98000 98000 98000 100000 100000 101000 101100

[79] 102500 105000 106000 107300 108000 112000 64000 77000 85000 85000 86000 90000 92000

[92] 95000 96000 98000 100000 100000 100400 101600 104000 105000 115000 126710 220000

fitted(fit)

1 2 3 4 5 6 7 8 9

90459.01 102916.64 98089.12 98477.96 102916.64 106496.91 94897.69 97700.29 98477.96

10 11 12 13 14 15 16 17 18

97700.29 101280.55 93261.60 95675.36 104471.98 98089.12 98477.96 100891.72 100502.88

19 20 21 22 23 24 25 26 27

111324.43 114127.02 112490.93 111713.26 103305.48 98477.96 106885.74 137325.45 103694.31

28 29 30 31 32 33 34 35 36

102527.81 120509.88 103305.48 106496.91 111324.43 112490.93 139350.37 97700.29 100114.05

37 38 39 40 41 42 43 44 45

98089.12 114904.69 98089.12 98477.96 95675.36 95675.36 93261.60 95675.36 100891.72

46 47 48 49 50 51 52 53 54

106496.91 95675.36 113349.35 98089.12 103305.48 106108.07 100891.72 103305.48 125256.65

55 56 57 58 59 60 61 62 63

100502.88 109299.50 95286.53 103694.31 101280.55 95675.36 92872.77 103305.48 98089.12

64 65 66 67 68 69 70 71 72

98089.12 98866.79 108910.67 103694.31 105719.24 95675.36 98089.12 98089.12 102916.64

73 74 75 76 77 78 79 80 81

106108.07 109688.33 96064.20 98089.12 95675.36 110077.17 111713.26 95675.36 109299.50

82 83 84 85 86 87 88 89 90

114985.44 106108.07 95675.36 95675.36 93261.60 98477.96 100502.88 100891.72 104083.15

91 92 93 94 95 96 97 98 99

98089.12 98089.12 100502.88 95675.36 95675.36 100502.88 108133.00 100891.72 113349.35

100 101 102 103

93261.60 98477.96 100891.72 139350.37

residuals(fit)

1 2 3 4 5 6 7

-5459.00732 -17916.64155 -12089.12173 -10477.95640 -10916.64155 -13496.90548 102.30753

8 9 10 11 12 13 14

-2700.28706 -3477.95640 -1700.28706 -5280.55098 6738.39810 4324.63818 -4471.98024

15 16 17 18 19 20 21

6910.87827 6522.04360 4108.28369 4497.11836 -6324.42530 -9127.01989 -6490.92932

22 23 24 25 26 27 28

-5713.25997 4194.52378 9522.04360 3114.25985 -25325.44590 11305.68911 12472.19312

29 30 31 32 33 34 35

-2509.87840 16694.52378 13503.09452 8675.57470 7509.07068 6649.62886 64299.71294

36 37 38 39 40 41 42

-18114.04697 -6089.12173 -21904.68923 -3089.12173 -3477.95640 324.63818 824.63818

43 44 45 46 47 48 49

4738.39810 2324.63818 -2891.71631 -7496.90548 4324.63818 -13349.35054 2910.87827

50 51 52 53 54 55 56

-305.47622 -2108.07081 4108.28369 1694.52378 -20256.64634 6497.11836 2700.49994

57 58 59 60 61 62 63

19713.47286 11305.68911 28719.44902 50124.63818 -14616.76723 -14805.47622 -8089.12173

64 65 66 67 68 69 70

-8089.12173 -5866.79107 -13910.66539 -6694.31089 -8719.23613 2324.63818 -89.12173

71 72 73 74 75 76 77

-89.12173 -4916.64155 -8108.07081 -11688.33473 3935.80351 1910.87827 5324.63818

78 79 80 81 82 83 84

-8977.16941 -9213.25997 9324.63818 -3299.50006 -7685.44111 1891.92919 16324.63818

85 86 87 88 89 90 91

-31675.36182 -16261.60190 -13477.95640 -15502.88164 -14891.71631 -14083.14557 -6089.12173

92 93 94 95 96 97 98

-3089.12173 -4502.88164 2324.63818 4324.63818 -502.88164 -7732.99605 708.28369

99 100 101 102 103

-9349.35054 11738.39810 16522.04360 25818.28369 80649.62886

confint(fit)

2.5 % 97.5 %

(Intercept) -9306.2381 83241.147

work\_yrs -1761.8204 2539.490

age 434.8646 4392.655

Some more regression models are proposed and the best model is found out.

In the first model, we saw the dependency of salary on both age and work experience.

In the subsequent models, we will see the dependency of salary individually on age and work experience.

**MODEL 2**

**Dependency of salary on age only**

fit <- lm(salary~age,data=got\_a\_job)

> summary(fit)

Call:

lm(formula = salary ~ age, data = got\_a\_job)

Residuals:

Min 1Q Median 3Q Max

-31454 -8533 -2182 4546 80886

Coefficients:

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

(Intercept) 29962.6 12697.8 2.360 0.0202 \*

age 2728.8 470.7 5.797 7.75e-08 \*\*\*

---

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

Residual standard error: 15550 on 101 degrees of freedom

Multiple R-squared: 0.2496, Adjusted R-squared: 0.2422

F-statistic: 33.6 on 1 and 101 DF, p-value: 7.748e-08

got\_a\_job$salary

[1] 85000 85000 86000 88000 92000 93000 95000 95000 95000 96000 96000 100000 100000

[14] 100000 105000 105000 105000 105000 105000 105000 106000 106000 107500 108000 110000 112000

[27] 115000 115000 118000 120000 120000 120000 120000 146000 162000 82000 92000 93000 95000

[40] 95000 96000 96500 98000 98000 98000 99000 100000 100000 101000 103000 104000 105000

[53] 105000 105000 107000 112000 115000 115000 130000 145800 78256 88500 90000 90000 93000

[66] 95000 97000 97000 98000 98000 98000 98000 98000 98000 100000 100000 101000 101100

[79] 102500 105000 106000 107300 108000 112000 64000 77000 85000 85000 86000 90000 92000

[92] 95000 96000 98000 100000 100000 100400 101600 104000 105000 115000 126710 220000

fitted(fit)

1 2 3 4 5 6 7 8 9

89996.11 103640.08 98182.49 98182.49 103640.08 106368.88 95453.70 98182.49 98182.49

10 11 12 13 14 15 16 17 18

98182.49 100911.29 92724.90 95453.70 103640.08 98182.49 98182.49 100911.29 100911.29

19 20 21 22 23 24 25 26 27

111826.46 114555.26 111826.46 111826.46 103640.08 98182.49 106368.88 136385.62 103640.08

28 29 30 31 32 33 34 35 36

103640.08 120012.85 103640.08 106368.88 111826.46 111826.46 139114.41 98182.49 100911.29

37 38 39 40 41 42 43 44 45

98182.49 114555.26 98182.49 98182.49 95453.70 95453.70 92724.90 95453.70 100911.29

46 47 48 49 50 51 52 53 54

106368.88 95453.70 114555.26 98182.49 103640.08 106368.88 100911.29 103640.08 122741.64

55 56 57 58 59 60 61 62 63

100911.29 109097.67 95453.70 103640.08 100911.29 95453.70 92724.90 103640.08 98182.49

64 65 66 67 68 69 70 71 72

98182.49 98182.49 109097.67 103640.08 106368.88 95453.70 98182.49 98182.49 103640.08

73 74 75 76 77 78 79 80 81

106368.88 109097.67 95453.70 98182.49 95453.70 109097.67 111826.46 95453.70 109097.67

82 83 84 85 86 87 88 89 90

117284.05 106368.88 95453.70 95453.70 92724.90 98182.49 100911.29 100911.29 103640.08

91 92 93 94 95 96 97 98 99

98182.49 98182.49 100911.29 95453.70 95453.70 100911.29 109097.67 100911.29 114555.26

100 101 102 103

92724.90 98182.49 100911.29 139114.41

residuals(fit)

1 2 3 4 5 6 7 8

-4996.1066 -18640.0804 -12182.4909 -10182.4909 -11640.0804 -13368.8751 -453.6961 -3182.4909

9 10 11 12 13 14 15 16

-3182.4909 -2182.4909 -4911.2856 7275.0986 4546.3039 -3640.0804 6817.5091 6817.5091

17 18 19 20 21 22 23 24

4088.7144 4088.7144 -6826.4647 -9555.2594 -5826.4647 -5826.4647 3859.9196 9817.5091

25 26 27 28 29 30 31 32

3631.1249 -24385.6175 11359.9196 11359.9196 -2012.8489 16359.9196 13631.1249 8173.5353

33 34 35 36 37 38 39 40

8173.5353 6885.5878 63817.5091 -18911.2856 -6182.4909 -21555.2594 -3182.4909 -3182.4909

41 42 43 44 45 46 47 48

546.3039 1046.3039 5275.0986 2546.3039 -2911.2856 -7368.8751 4546.3039 -14555.2594

49 50 51 52 53 54 55 56

2817.5091 -640.0804 -2368.8751 4088.7144 1359.9196 -17741.6437 6088.7144 2902.3301

57 58 59 60 61 62 63 64

19546.3039 11359.9196 29088.7144 50346.3039 -14468.9014 -15140.0804 -8182.4909 -8182.4909

65 66 67 68 69 70 71 72

-5182.4909 -14097.6699 -6640.0804 -9368.8751 2546.3039 -182.4909 -182.4909 -5640.0804

73 74 75 76 77 78 79 80

-8368.8751 -11097.6699 4546.3039 1817.5091 5546.3039 -7997.6699 -9326.4647 9546.3039

81 82 83 84 85 86 87 88

-3097.6699 -9984.0542 1631.1249 16546.3039 -31453.6961 -15724.9014 -13182.4909 -15911.2856

89 90 91 92 93 94 95 96

-14911.2856 -13640.0804 -6182.4909 -3182.4909 -4911.2856 2546.3039 4546.3039 -911.2856

97 98 99 100 101 102 103

-8697.6699 688.7144 -10555.2594 12275.0986 16817.5091 25798.7144 80885.5878

confint(fit)

2.5 % 97.5 %

(Intercept) 4773.600 55151.644

age 1794.966 3662.624

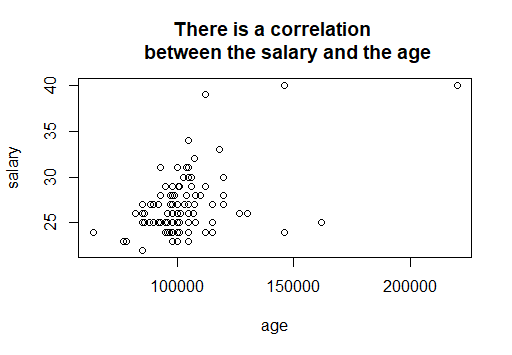
plot(got\_a\_job$salary,got\_a\_job$age,main="There is a correlation

+ between the salary and the age",xlab="age"

+ ,ylab="salary")

>

> abline(fit)



**MODEL 3**

**Dependency of salary on work experience only**

fit <- lm(salary~work\_yrs,data=got\_a\_job)

> summary(fit)

Call:

lm(formula = salary ~ work\_yrs, data = got\_a\_job)

Residuals:

Min 1Q Median 3Q Max

-34498 -7745 -498 3803 86419

Coefficients:

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

(Intercept) 93101 2496 37.30 < 2e-16 \*\*\*

work\_yrs 2699 526 5.13 1.4e-06 \*\*\*

---

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

Residual standard error: 15990 on 101 degrees of freedom

Multiple R-squared: 0.2067, Adjusted R-squared: 0.1989

F-statistic: 26.32 on 1 and 101 DF, p-value: 1.403e-06

got\_a\_job$salary

[1] 85000 85000 86000 88000 92000 93000 95000 95000 95000 96000 96000 100000 100000

[14] 100000 105000 105000 105000 105000 105000 105000 106000 106000 107500 108000 110000 112000

[27] 115000 115000 118000 120000 120000 120000 120000 146000 162000 82000 92000 93000 95000

[40] 95000 96000 96500 98000 98000 98000 99000 100000 100000 101000 103000 104000 105000

[53] 105000 105000 107000 112000 115000 115000 130000 145800 78256 88500 90000 90000 93000

[66] 95000 97000 97000 98000 98000 98000 98000 98000 98000 100000 100000 101000 101100

[79] 102500 105000 106000 107300 108000 112000 64000 77000 85000 85000 86000 90000 92000

[92] 95000 96000 98000 100000 100000 100400 101600 104000 105000 115000 126710 220000

fitted(fit)

1 2 3 4 5 6 7 8 9

95799.31 98498.00 98498.00 101196.68 98498.00 106594.05 93100.63 95799.31 101196.68

10 11 12 13 14 15 16 17 18

95799.31 103895.36 98498.00 98498.00 109292.73 98498.00 101196.68 101196.68 98498.00

19 20 21 22 23 24 25 26 27

106594.05 109292.73 114690.10 109292.73 101196.68 101196.68 109292.73 136279.57 103895.36

28 29 30 31 32 33 34 35 36

95799.31 120087.47 101196.68 106594.05 106594.05 114690.10 133580.88 95799.31 95799.31

37 38 39 40 41 42 43 44 45

98498.00 114690.10 98498.00 101196.68 98498.00 98498.00 98498.00 98498.00 101196.68

46 47 48 49 50 51 52 53 54

106594.05 98498.00 103895.36 98498.00 101196.68 103895.36 101196.68 101196.68 136279.57

55 56 57 58 59 60 61 62 63

98498.00 109292.73 95799.31 103895.36 103895.36 98498.00 95799.31 101196.68 98498.00

64 65 66 67 68 69 70 71 72

98498.00 103895.36 106594.05 103895.36 101196.68 98498.00 98498.00 98498.00 98498.00

73 74 75 76 77 78 79 80 81

103895.36 111991.41 101196.68 98498.00 98498.00 114690.10 109292.73 98498.00 109292.73

82 83 84 85 86 87 88 89 90

98498.00 103895.36 98498.00 98498.00 98498.00 101196.68 98498.00 101196.68 106594.05

91 92 93 94 95 96 97 98 99

98498.00 98498.00 98498.00 98498.00 98498.00 98498.00 101196.68 101196.68 103895.36

100 101 102 103

98498.00 101196.68 101196.68 133580.88

residuals(fit)

1 2 3 4 5 6 7 8

-10799.3142 -13497.9977 -12497.9977 -13196.6811 -6497.9977 -13594.0481 1899.3692 -799.3142

9 10 11 12 13 14 15 16

-6196.6811 200.6858 -7895.3646 1502.0023 1502.0023 -9292.7315 6502.0023 3803.3189

17 18 19 20 21 22 23 24

3803.3189 6502.0023 -1594.0481 -4292.7315 -8690.0984 -3292.7315 6303.3189 6803.3189

25 26 27 28 29 30 31 32

707.2685 -24279.5661 11104.6354 19200.6858 -2087.4653 18803.3189 13405.9519 13405.9519

33 34 35 36 37 38 39 40

5309.9016 12419.1174 66200.6858 -13799.3142 -6497.9977 -21690.0984 -3497.9977 -6196.6811

41 42 43 44 45 46 47 48

-2497.9977 -1997.9977 -497.9977 -497.9977 -3196.6811 -7594.0481 1502.0023 -3895.3646

49 50 51 52 53 54 55 56

2502.0023 1803.3189 104.6354 3803.3189 3803.3189 -31279.5661 8502.0023 2707.2685

57 58 59 60 61 62 63 64

19200.6858 11104.6354 26104.6354 47302.0023 -17543.3142 -12696.6811 -8497.9977 -8497.9977

65 66 67 68 69 70 71 72

-10895.3646 -11594.0481 -6895.3646 -4196.6811 -497.9977 -497.9977 -497.9977 -497.9977

73 74 75 76 77 78 79 80

-5895.3646 -13991.4150 -1196.6811 1502.0023 2502.0023 -13590.0984 -6792.7315 6502.0023

81 82 83 84 85 86 87 88

-3292.7315 8802.0023 4104.6354 13502.0023 -34497.9977 -21497.9977 -16196.6811 -13497.9977

89 90 91 92 93 94 95 96

-15196.6811 -16594.0481 -6497.9977 -3497.9977 -2497.9977 -497.9977 1502.0023 1502.0023

97 98 99 100 101 102 103

-796.6811 403.3189 104.6354 6502.0023 13803.3189 25513.3189 86419.1174

confint(fit)

2.5 % 97.5 %

(Intercept) 88149.26 98052.006

work\_yrs 1655.18 3742.187

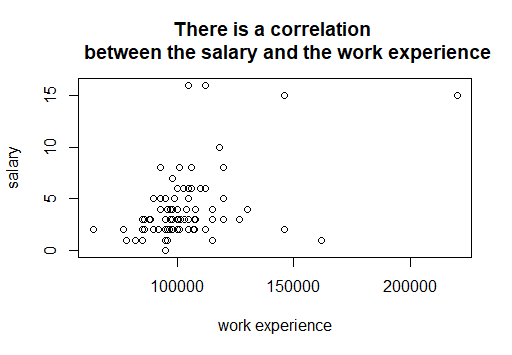
plot(got\_a\_job$salary,got\_a\_job$work\_yrs,main="There is a correlation

+ between the salary and the work experience",xlab="work experience"

+ ,ylab="salary")

>

> abline(fit)



**Which is the best model?**

The higher the R-squared, the better the model fits your data.

Model 1: R squared was 0.2356

Model 2: R squared was 0.2422

Model 3: R squared was 0.1989

**So model 2 was the best model that fits the data.**

For the next set of problems, we have to analyse the subsets of people having a job and those who don’t have a job.

Therefore we eliminate entries where people’s salaries are 998 and 999.

clean <- mba[mba$salary!=998 & mba$salary!=999, ]

To simplify the dataset, we add another variable to the data set called job which is a categorical variable and it says 1 if the person has a job(has a salary) and 0 if does not have a salary(does not have a job)

job <- ifelse(clean$salary==0,0,1)

clean <- cbind(clean,job)

**CONTINGENCY TABLES**

**Does gender(sex) play a role in getting/not getting a job?**

mytable <- xtabs(~job+sex, data=clean)

> addmargins(mytable)

sex

job 1 2 Sum

0 67 23 90

1 72 31 103

Sum 139 54 193

prop.table(mytable,1)

sex

job 1 2

0 0.7444444 0.2555556

1 0.6990291 0.3009709

Almost 70% of the jobs have gone to males and only 30% of the jobs have gone to females.

prop.table(mytable,2)

sex

job 1 2

0 0.4820144 0.4259259

1 0.5179856 0.5740741

51% of males have a job while 57% of females have a job

Null hypothesis is that the job and the sex are independent

Alternate hypothesis is that the job and the sex are not independent

> chisq.test(mytable)

Pearson's Chi-squared test with Yates' continuity correction

data: mytable

X-squared = 0.29208, df = 1, p-value = 0.5889

We get a relatively high p value therefore we fail to reject the null hypothesis.

**Does English as first language play a role in getting/not getting a job?**

mytable <- xtabs(~job+frstlang, data=clean)

> addmargins(mytable)

frstlang

job 1 2 Sum

0 82 8 90

1 96 7 103

Sum 178 15 193

> prop.table(mytable,1)

frstlang

job 1 2

0 0.91111111 0.08888889

1 0.93203883 0.06796117

93% of people who have got jobs have English as their first language

> prop.table(mytable,2)

frstlang

job 1 2

0 0.4606742 0.5333333

1 0.5393258 0.4666667

Almost 54% of people who have English as their first language have got jobs

Let us run a chi square test where

H0: Job and first language of the person are independent of each other

H1: Job and first language of the person are not independent of each other

chisq.test(mytable)

Pearson's Chi-squared test with Yates' continuity correction

data: mytable

X-squared = 0.074127, df = 1, p-value = 0.7854

The p value is relatively high so we fail to reject the null hypothesis.

**Logistic Regression**

We split the data into two chunks: training and testing set. The training set will be used to fit our model which we will be testing over the testing set.

According to the example on the titanic data set, almost 90% of the data set is used for training and the remaining is for testing.

Here we have 193 rows so 90% of 193 is almost 174.

train <- clean[1:174,]

test <- clean[175:193,]

model <- glm(job ~age+work\_yrs,family=binomial(link='logit'),data=train)

summary(model)

Call:

glm(formula = job ~ age + work\_yrs, family = binomial(link = "logit"),

data = train)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.4759 -1.1674 -0.5629 1.1316 1.7677

Coefficients:

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

(Intercept) 5.26635 1.98814 2.649 0.00808 \*\*

age -0.21512 0.08353 -2.575 0.01002 \*

work\_yrs 0.14511 0.08997 1.613 0.10677

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 241.01 on 173 degrees of freedom

Residual deviance: 231.40 on 171 degrees of freedom

AIC: 237.4

Number of Fisher Scoring iterations: 4

anova(model, test="Chisq")

Analysis of Deviance Table

Model: binomial, link: logit

Response: job

Terms added sequentially (first to last)

Df Deviance Resid. Df Resid. Dev Pr(>Chi)

NULL 173 241.01

age 1 6.9171 172 234.09 0.008537 \*\*

work\_yrs 1 2.6950 171 231.40 0.100661

---

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

fitted.results <- predict(model,newdata=subset(test,select=c(1,10)),type='response')

> fitted.results <- ifelse(fitted.results > 0.5,1,0)

>

> misClasificError <- mean(fitted.results != test$job)

> print(paste('Accuracy',1-misClasificError))

[1] "Accuracy 0.684210526315789"