TOPIC 3

SALARY

Name: Akash Mandal

Roll. No. : 16MA20007

Problem Statement -

Database contains salary information of different employees in different organisations. It is required to test whether Overtime Pay, Other Pay and benefits altogether increases with Basic Pay for the year 2014.

**Professor's instructions-**

Take sum of Basic Pay, Overtime Pay and Other Pay and benefits ( Which is the attribute TotalPayBenefits ) and regress it with Base Pay, to find the relationship.

Processing data

We filter data from 2014 only, and remove cases with values missing for our concerned attributes BasePay, and TotalPayBenefits.

> SALARY\_14 <- SALARY[SALARY$Year == 2014, ]

> SALARY\_14\_NEW <- SALARY\_14[!is.na(SALARY\_14$BasePay) & !is.na(SALARY\_14$TotalPayBenefits), ]

> basepay <- SALARY\_14\_NEW$BasePay

> totalpay <- SALARY\_14\_NEW$TotalPayBenefits

Linear regression

We use linear regression to find relationship between BasePay, and TotalPayBenefits. Here BasePay is our predictor variable, and TotalPayBenefits is our response.

Correlation r = 0.9677

Applying linear model, we have

TotalPayBenefits = 3412 + 1.455 \* BasePay

> cor(basepay, totalpay)

[1] 0.9677359

> model <- lm(totalpay ~ basepay)

> summary(model)

Call:

lm(formula = totalpay ~ basepay)

Residuals:

Min 1Q Median 3Q Max

-74286 -8149 -3526 4203 319410

Coefficients:

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

(Intercept) 3.412e+03 1.549e+02 22.03 <2e-16 \*\*\*

basepay 1.455e+00 1.940e-03 749.85 <2e-16 \*\*\*

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

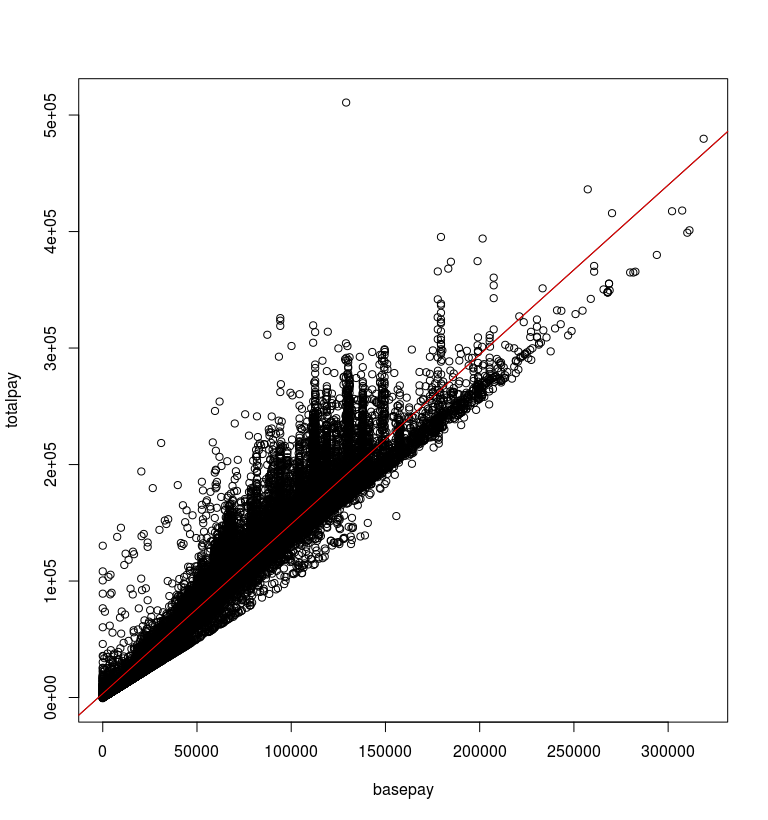
Residual standard error: 16690 on 38117 degrees of freedom

Multiple R-squared: 0.9365,Adjusted R-squared: 0.9365

F-statistic: 5.623e+05 on 1 and 38117 DF, p-value: < 2.2e-16

Conclusion

We obtain an r value 0.9677, which shows h hogh positive correlation between the attributes. R2 value of 0.9365, which idicates 93.65% of variation in TotalPayBenefits can be attributed to variation in BasePay.

A p value of 2.2e-16 << 0.05 tells us this result is highly significant.