A1Q8

Undergraduate Student

```
overdue <- read.csv("overdue.txt",sep = "\t")
type <- c(rep(0, 48), rep(1,48))
overdue$TYPE <- type
summary(lm(LATE~BILL+TYPE, data = overdue))
```

```
##
## Call:
## lm(formula = LATE ~ BILL + TYPE, data = overdue)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.7637 -11.4760   0.4037  12.4812  29.0765
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.28599    3.91286   8.507 2.93e-13 ***
## BILL        -0.01264    0.01901  -0.665    0.508
## TYPE        37.39583    2.94375  12.703  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 14.42 on 93 degrees of freedom
## Multiple R-squared:  0.635,  Adjusted R-squared:  0.6272
## F-statistic: 80.91 on 2 and 93 DF,  p-value: < 2.2e-16
```

We found that BILL predictor has a p-value greater than 0.05, this indicates that it is not significant. We should preceed to reduced models.

Reduced model:

```
summary(lm(LATE~BILL, data = overdue))
```

```
##
## Call:
## lm(formula = LATE ~ BILL, data = overdue)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.846 -17.212  -0.793  19.007  47.774
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 51.98390    5.96405   8.716 9.84e-14 ***
## BILL        -0.01264    0.03128  -0.404    0.687
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 23.72 on 94 degrees of freedom
## Multiple R-squared:  0.001734,   Adjusted R-squared:  -0.008885
## F-statistic: 0.1633 on 1 and 94 DF,  p-value: 0.687
```

```
summary(lm(LATE~TYPE, data = overdue))
```

```
##
## Call:
## lm(formula = LATE ~ TYPE, data = overdue)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -29.4792 -11.6302   0.5208  12.0677  30.5208
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   31.083      2.075   14.98   <2e-16 ***
## TYPE          37.396      2.935   12.74   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 14.38 on 94 degrees of freedom
## Multiple R-squared:  0.6333, Adjusted R-squared:  0.6294
## F-statistic: 162.3 on 1 and 94 DF,  p-value: < 2.2e-16
```

We see that reduced models do not improve the model results. We have to check we can add some interaction terms.

Full interaction model:

```
summary(lm(LATE~BILL*TYPE, data = overdue))
```

```
##
## Call:
## lm(formula = LATE ~ BILL * TYPE, data = overdue)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.1211  -2.2163   0.0974   1.9556   8.6995
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.209624   1.198504   1.844   0.0685 .
## BILL         0.165683   0.006285  26.362   <2e-16 ***
## TYPE        99.548561   1.694940  58.733   <2e-16 ***
## BILL:TYPE   -0.356644   0.008888 -40.125   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
```

```
## Residual standard error: 3.371 on 92 degrees of freedom
## Multiple R-squared:  0.9803, Adjusted R-squared:  0.9796
## F-statistic:  1524 on 3 and 92 DF,  p-value: < 2.2e-16
```

First we tried full interaction model. p-values of all predictors including interaction term are significant. However, the p-value of the intercept indicates that this is not a appropriate model even it has a really high R-squared value.

We have to try one predictors and one interaction term.

One predictor with one interaction:

```
summary(lm(LATE~BILL+BILL:TYPE, data = overdue))
```

```
##
## Call:
## lm(formula = LATE ~ BILL + BILL:TYPE, data = overdue)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.427 -15.964   0.964  15.136  44.161
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 51.98390    5.22973   9.940 2.74e-16 ***
## BILL        -0.07286    0.02960  -2.461   0.0157 *
## BILL:TYPE    0.12043    0.02227   5.408 4.91e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.8 on 93 degrees of freedom
## Multiple R-squared:  0.2406, Adjusted R-squared:  0.2243
## F-statistic: 14.73 on 2 and 93 DF,  p-value: 2.768e-06
```

```
summary(lm(LATE~TYPE+BILL:TYPE, data = overdue))
```

```
##
## Call:
## lm(formula = LATE ~ TYPE + BILL:TYPE, data = overdue)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.0833  -3.8647  -0.2568   4.7023  20.9167
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 31.08333    1.41538   21.96   <2e-16 ***
## TYPE        70.67485    3.76268   18.78   <2e-16 ***
## TYPE:BILL   -0.19096    0.01828  -10.45   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.806 on 93 degrees of freedom
## Multiple R-squared:  0.8313, Adjusted R-squared:  0.8276
## F-statistic: 229.1 on 2 and 93 DF,  p-value: < 2.2e-16
```

For these two models, we can see that the second model is better than the first one because it has good R-squared value (0.8313), and all predictors including intercept have really small p-value. This means that the model can not be further optimized.

Therefore, we will use LATE~TYPE+TYPE:BILL as our regression model