In-Class Exercise 01/23/2020

Team 7

1/27/2020

A regression with all variables to predict balance

```
fullBalanceModel <- lm(
  balance ~ age + job + marital + education + default + housing + loan + contact + day + month + duratio
  data=bank
)
```

Using general intuition to subset variables to predict balance

Using some general intuition about variables that may be related to predicting balance, we attempt to subset the full model to fewer predictors by eliminating variables which have no effect on predicting a customer's account balance:

```
intuitiveBalanceModel <- lm(
  balance ~ age + default + marital,
  data=bank
)
```

With the null hypothesis being intuitiveBalanceModel

```
anova(intuitiveBalanceModel, fullBalanceModel)
```

yields an F value of 6.1 and a p value of 2.2e-16, thus all of the intuitive variables we picked well predict balance.

Updating our intuitive model

Using some more intuitive thinking, we update the intuitiveBalanceModel to attempt to even better predict balance:

```
updatedIntuitiveBalanceModel <- lm(
  balance ~ age + default + marital + job + education,
  data=bank
)
```

With the null hypothesis being updatedIntuitiveBalanceModel

```
anova(updatedIntuitiveBalanceModel, fullBalanceModel)
```

yields an F value of 7.8, which tells us that this model better predicts balance than the first intuitive model we created. This ANOVA test also yields a p value of 2.2e-16, which denotes that the additional variables of job and education that were added to this model both help better predict balance.

Statistically relevant variables

Now, we can attempt to see what variables could be meaningful in predicting balance by looking at the p value column of the summary of the fullBalanceModel, and choose all categories with a p value < 0.01:

```
statisticallyImportantModel <- lm(
  balance ~ age + default + marital + loan,
  data=bank
)
```

With the null hypothesis being staisticallyImportantModel

```
anova(statisticallyImportantModel, fullBalanceModel)
```

yields an F value of 5.7 and a p value of 2.2e-16, thus we find that this model is once again, a good predictor, but not as good of a predictor.

Discussion

The results demonstrate that the age of the customer, whether the customer has any credit in default, the marital status of the customer, what job the customer holds, and the education level of the customer, all hold meaningful correlations when attempting to predict the balance a customer has in an account with the bank. Which further leads to a model which can also be useful in predicting any of the aforementioned variables, as correlations work in both directions.