STAT 435 HW2

Chongyi Xu

April 12, 2018

1. Suppose we have a quantitative response Y, and a single feature $X \leq \mathbb{R}$. Let RSS_1 denote the residual sum of squares that results from fitting the modedl

$$Y = \beta_0 + \beta_1 X + \epsilon$$

using least squares. Let RSS_{12} denote the residual sum of squares that results from fitting the model

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

using least squares.

(a) Prove that $RSS_{12} \leq RSS_1$.

Denote RSS to be the residual sum of squares. The method of least squares fitting is to minimize RSS.

(b) Prove that R^2 of the model containing just the feature X is no greater than the R^2 of the model containing both X and X^2.

Since from part(a), we have concluded that $RSS_{12} \leq RSS_1$. And R^2 is definded to be $R^2 = 1 - \frac{RSS}{TSS}$

$$R_1^2 = 1 - \frac{RSS_1}{TSS}$$
$$R_{12}^2 = 1 - \frac{RSS_{12}}{TSS}$$

TSS are the same to two R^2 since they are from the same response Y.

$$RSS_{12} \leq RSS_1 \Rightarrow 1 - \frac{RSS_1}{TSS} \leq 1 - \frac{RSS_{12}}{TSS} \Rightarrow R_1^2 \leq R_{12}^2$$

2. Describe the null hypotheses to which the p-value in Table 3.4 of the text book correspond. Explain what conclusion you can draw based on these p-values. Your explanation should be phrased in term of `sales`, `TV`, `radio`, and `newspaper`.

- `TV`
- H_0: The sales is not related with TV advertising.
- H_1: The sales is related with TV advertising.

From the p-value we found in Table 3.4 for TV ($p - value < 0.0001$), it indicates strong evidence against the null hypothesese, the null hypothese is rejected, in the other word, the sales is related with TV advertising.

- `radio`
- H_0: The sales is not related with radio advertising.
- H_1: The sales is related with radio advertising.

From the p-value we found in Table 3.4 for radio ($p - value < 0.0001$), it indicates strong evidence against the null hypothesese, the null hypothese is rejected, in the other word, the sales is related with radio advertising.

- `newspaper`
- H_0: The sales is not related with newspaper advertising.
- H_1: The sales is related with newspaper advertising.

From the p-value we found in Table 3.4 for radio ($p - value = 0.8599$), it indicates that there is no sufficient evidence that supports rejecting the null hypotheses with a level of $\alpha = 0.01$, the rejection fails. In the other word, we can not reject the hypotheses that there is no relationship between newpaper advertising and sales.

3. Consider a linear model with just one feature.

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Suppose we have n observations from this model, $(x_1, y_1), \cdots, (x_n, y_n)$. The least squares estimators is given in (3.4) of the textbook. Furthermore, we saw in class that if we construct a n x 2 matrix $\tilde{\mathbf{X}}$. If we let $\mathbf{y}$ denote the vector with elements $y_1, \cdots, y_n$, then the least squares estimator takes the form

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^T \mathbf{y}$$

Prove that the equation agrees with equation (3.4) of the textbook.

$$\begin{aligned}
(\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^T \mathbf{y} &= \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}^{-1} \begin{pmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \\
&= \frac{1}{n \sum x_i^2 - (\sum x_i)^2} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix} \begin{pmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \\
&= \frac{1}{n \sum x_i^2 - (\sum x_i)^2} \sum_{j=1}^{n} \begin{pmatrix} y_j \sum x_i^2 - x_j y_j \sum x_i \\ -y_j \sum x_i + n x_j y_j \end{pmatrix}
\end{aligned}$$

Consider the bottom of $\frac{1}{n \sum x_i^2 - (\sum x_i)^2}$, use that $\sum x_i = n\bar{x}$

$$\begin{aligned}
n \sum x_i^2 - (\sum x_i)^2 &= n \sum x_i^2 - n^2 \bar{x}^2 \\
&= n^2 (\frac{1}{n} \sum x_i^2 - \bar{x}^2) \\
&= n^2 \frac{1}{n} (\sum x_i^2 - 2n\bar{x}\bar{x} + n\bar{x}^2) \\
&= n^2 \frac{1}{n} (\sum x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\
&= n \sum (x_i - \bar{x})^2
\end{aligned}$$

Now consider the second row of summation part(β_1). We know that $n\bar{x}\bar{y} = \bar{x} \sum_i y_i = \bar{y} \sum_i x_i$

$$\begin{aligned}
\sum_{j=1}^{n}(-y_j\sum x_i + nx_jy_j) &= \sum_{j=1}^{n}(nx_jy_j - n\bar{x}y_j) \\
&= n(\sum_j x_jy_j - \bar{x}\sum_j y_j) \\
&= n(\sum_j x_jy_j - \bar{x}\sum_j y_j + n\bar{x}\bar{y} - \bar{y}\sum_j x_j) \\
&= n\sum_j(x_jy_j - \bar{y}x_j - \bar{x}y_j + \bar{x}\bar{y}) \\
&= n\sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y})
\end{aligned}$$

Therefore,

$$\beta_1 = \frac{n\sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y})}{n\sum(x_i - \bar{x})^2} = \frac{\sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y})}{\sum(x_i - \bar{x})^2} = \hat{\beta}_1$$

Now consider the first row of summation part(β_0). We want to check if $\beta_0 = \bar{y} - \beta_1\bar{x}$

$$\begin{aligned}
\frac{\sum_{j=1}^{n}(y_j\sum x_i^2 - x_jy_j\sum x_i)}{n\sum(x_i - \bar{x})^2} &= \frac{\sum_j(y_j\sum x_i^2)}{n\sum(x_i - \bar{x})^2} - \frac{\sum_j(x_jy_j\sum x_i)}{n\sum(x_i - \bar{x})^2} \\
&= \frac{n\bar{y}\bar{x^2}}{\sum(x_i - \bar{x})^2} - \frac{\bar{x}\sum_j x_jy_j}{\sum(x_i - \bar{x})^2} \\
&= \bar{y} - \beta_1\bar{x} \\
\beta_0 &= \bar{y} - \beta_1\bar{x}(QED)
\end{aligned}$$

4. This question involves the use of multiple linear regression on the Auto data set.

(a) Use the `lm()` function to perform a multiple linear regression with `mpg` as the response and all other variables except `name` as the predictors.

```r
library(ISLR)
```

```
## Warning: package 'ISLR' was built under R version 3.4.4
```

```r
dat <- Auto
dat$origin.f <- factor(dat$origin,labels=c('American','European','Japanese'))

model.fit <- lm(data=dat, mpg~. - name - origin)

summary(model.fit)
```

```
##
## Call:
## lm(formula = mpg ~ . - name - origin, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0095 -2.0785 -0.0982  1.9856 13.3608
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)      -1.795e+01  4.677e+00  -3.839 0.000145 ***
## cylinders        -4.897e-01  3.212e-01  -1.524 0.128215
## displacement      2.398e-02  7.653e-03   3.133 0.001863 **
## horsepower       -1.818e-02  1.371e-02  -1.326 0.185488
## weight           -6.710e-03  6.551e-04 -10.243  < 2e-16 ***
## acceleration      7.910e-02  9.822e-02   0.805 0.421101
## year              7.770e-01  5.178e-02  15.005  < 2e-16 ***
## origin.fEuropean  2.630e+00  5.664e-01   4.643 4.72e-06 ***
## origin.fJapanese  2.853e+00  5.527e-01   5.162 3.93e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.307 on 383 degrees of freedom
## Multiple R-squared:  0.8242, Adjusted R-squared:  0.8205
## F-statistic: 224.5 on 8 and 383 DF,  p-value: < 2.2e-16
```

From the result above, we are able to give the following conclusions:

- With a significant level of $\alpha = 0.001$, we can not reject the hypotheses that `cylinders`, `displacement`, `horsepower`, `acceleration` have no relationship with `mpg`. However, if we are using a siginificant level of $\alpha = 0.01$, `displacement` might considered to be a factor of `mpg`.
- For every ≈ 75 years, a vehicle will be able to drive 100 more miles per gallon.
- I also make the `origin` data to be categorical for appropriate use.

```r
is.factor(dat$origin.f)
```

```
## [1] TRUE
```

(b) Try out some models to predict mpg using functions of variable horsepower.

```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.4
```

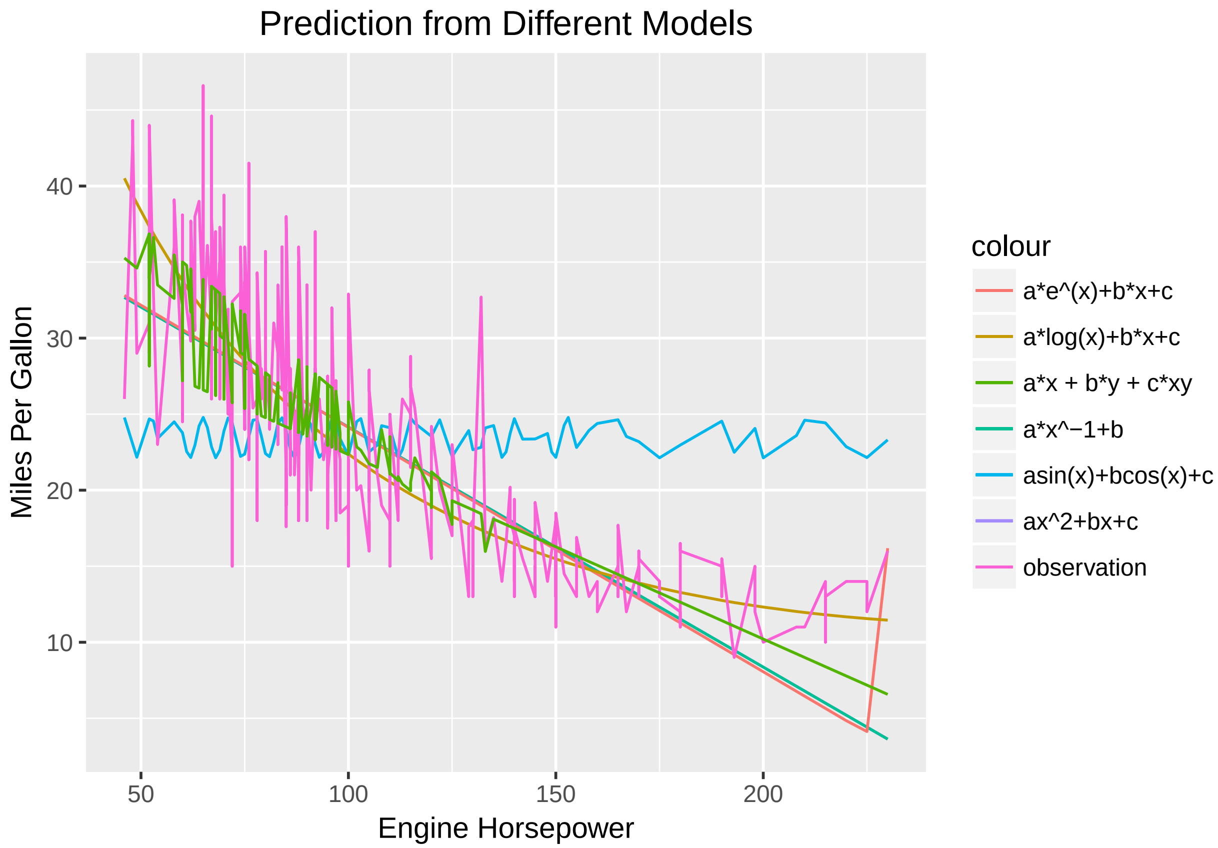
```r
# mpg = a * horsepower + b
fit1 <- lm(data=dat, mpg~horsepower^2 + horsepower)
fit1.predict <- predict(fit1, newdata=dat)

# mpg = a * horsepower^-1 + b
fit2 <- lm(data=dat, mpg~1 / horsepower + horsepower)
fit2.predict <- predict(fit2, newdata=dat)

# mpg = a * e^(horsepower) + b * horsepower + c
fit3 <- lm(data=dat, mpg~exp(horsepower) + horsepower)
fit3.predict <- predict(fit3, newdata=dat)

# mpg = alog(horsepower) + b * horsepower + c
fit4 <- lm(data=dat, mpg~log(horsepower) + horsepower)
fit4.predict <- predict(fit4, newdata=dat)

# mpg = asin(horsepower) + bcos(horsepower) + c
fit5 <- lm(data=dat, mpg~sin(horsepower) + cos(horsepower))
fit5.predict <- predict(fit5, newdata=dat)


p <- ggplot() + geom_line(aes(dat$horsepower, fit1.predict, color= 'ax^2+bx+c')) +
```

```
  geom_line(aes(dat$horsepower, fit2.predict, color='a*x^-1+b')) +
  geom_line(aes(dat$horsepower, fit3.predict, color='a*e^(x)+b*x+c')) +
  geom_line(aes(dat$horsepower, fit4.predict, color='a*log(x)+b*x+c')) +
  geom_line(aes(dat$horsepower, fit5.predict, color='asin(x)+bcos(x)+c')) +
  geom_line(aes(dat$horsepower, dat$mpg, color='observation')) +
  xlab('Engine Horsepower') + ylab('Miles Per Gallon') +
  ggtitle('Prediction from Different Models') +
  theme(plot.title=element_text(hjust=0.5))
p
```

From the plot, we can see that in fact none of the models has an acceptable result. In general, the $alog(x) + bx + c$ function has a relatively better result.

(c) Now fit a model to predict `mpg` using `horsepower`, `origin` and an interaction between them.

```
fit.model <- lm(data=dat, mpg ~ horsepower + origin.f + origin.f * horsepower)
fit.predict <- predict(fit.model, newdata=dat)
p + geom_line(aes(dat$horsepower, fit.predict, color='a*x + b*y + c*xy'))
```



We can see that the new model fits the data much better than single feature models in the preivous part.

```
summary(fit.model)
```

```
##
## Call:
## lm(formula = mpg ~ horsepower + origin.f + origin.f * horsepower,
##     data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.7415  -2.9547  -0.6389   2.3978  14.2495
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 34.476496   0.890665  38.709  < 2e-16 ***
## horsepower                  -0.121320   0.007095 -17.099  < 2e-16 ***
## origin.fEuropean            10.997230   2.396209   4.589 6.02e-06 ***
## origin.fJapanese            14.339718   2.464293   5.819 1.24e-08 ***
## horsepower:origin.fEuropean -0.100515   0.027723  -3.626 0.000327 ***
## horsepower:origin.fJapanese -0.108723   0.028980  -3.752 0.000203 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.422 on 386 degrees of freedom
## Multiple R-squared:  0.6831, Adjusted R-squared:  0.679
## F-statistic: 166.4 on 5 and 386 DF,  p-value: < 2.2e-16
```

The model tells that consider at significant level $\alpha = 0.001$, all of the factors we have used to build the model are statistically significant, including the interaction between origins and horsepower. The model says that for American vehicles, loosing every ≈ 0.121 unit of horsepower, the vehicle will be also run 1 more mile per gallon. For those vehicles that are from Europe and Japan, loosing every $\approx 0.101, 0.109$ unit of horsepower will make the vehicle run 1 more mile per gallon.

5. Consider fitting a model to predict credit card `balance` using `income` and `student`, where `student` is a quantative variable that takes on one of three values

(a) Encode the student variable using two dummy variables, one of which equals 1 if `student=graduate` (and 0 otherwise), and one of which equals 1 is `student=undergraduate` (and 0 otherwise). Write out an expression for a linear model to predict `balance` using `income` and `student`, using this coding of dummy variables.

```r
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
```

```r
dat <- Credit

students <- which(dat$Student == "Yes")
dat$NotStudent <- 1
dat$NotStudent[students] <- 0
dat$NotStudent <- factor(dat$NotStudent)

# Graduate: years of education is greater than 16
dat$Graduate <- 0
graduates <- students[dat$Education >= 16]
under <- students[!(students %in% graduates)]

dat$Graduate[graduates] <- 1
dat$Graduate <- factor(dat$Graduate)

dat$Undergraduate <- 0
dat$Undergraduate[under] <- 1
dat$Undergraduate <- factor(dat$Undergraduate)


dat_a <- dat%>% select(Balance, Income, Graduate, Undergraduate)

model.fit_a <- lm(data=dat_a, Balance ~ .)

summary(model.fit_a)
```

```
##
## Call:
## lm(formula = Balance ~ ., data = dat_a)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -762.21 -331.47  -44.35  323.58  818.34
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    211.2992    32.5139   6.499 2.44e-10 ***
## Income           5.9809     0.5578  10.722  < 2e-16 ***
## Graduate1      366.5230   125.7985   2.914  0.00378 **
## Undergraduate1 388.0637    74.5947   5.202 3.17e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 392.3 on 396 degrees of freedom
## Multiple R-squared:  0.2775, Adjusted R-squared:  0.272
## F-statistic:  50.7 on 3 and 396 DF,  p-value: < 2.2e-16
```

From the result above, we can see that, with a significant level $\alpha \geq 0.001$, we could say that whether the student is graduate student or not affects the credit balance. In general, all factors we are considering have sufficient evidence to support that they are related to the credit card balance. With \$598 higher income, the credit balance will increase by \$100.

(b) Now encode the student variable using two dummy variables, one of which equals 1 if `student=not student` (and 0 otherwise), and one of which equals 1 is `student=graduate` (and 0 otherwise). Write out an expression for a linear model to predict `balance` using `income` and `student`, using this coding of dummy variables.

```r
dat_b <- dat%>% select(Balance, Income, NotStudent, Graduate)

model.fit_b <- lm(data=dat_b, Balance ~ .)

summary(model.fit_b)
```

```
##
## Call:
## lm(formula = Balance ~ ., data = dat_b)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -762.21 -331.47  -44.35  323.58  818.34
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  599.3629    76.8483   7.799 5.57e-14 ***
## Income         5.9809     0.5578  10.722  < 2e-16 ***
## NotStudent1 -388.0637    74.5947  -5.202 3.17e-07 ***
## Graduate1    -21.5407   143.3608  -0.150    0.881
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 392.3 on 396 degrees of freedom
## Multiple R-squared:  0.2775, Adjusted R-squared:  0.272
## F-statistic:  50.7 on 3 and 396 DF,  p-value: < 2.2e-16
```

With a significant level of 0.1, we are not able to reject the hypothese that whether is graduate student or not does not affect the balance of credit card. In general, both income and the fact that is student or not play roles in credit card balance analysis. With \$598 higher income, the credit balance will increase by \$100.

(c) Using the coding in (a), write out an expression for a linear model to predict `balance` using `income`, `student` and interaction between `income` and `student`.

```
dat_c <- dat%>% select(Balance, Income, Graduate, Undergraduate)

model.fit_c <- lm(data=dat_c, Balance ~ . + Income*Graduate + Income*Undergraduate)

summary(model.fit_c)
```

```
##
## Call:
## lm(formula = Balance ~ . + Income * Graduate + Income * Undergraduate,
##     data = dat_c)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -773.39 -325.70  -41.13  321.65  814.04
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            200.6232    33.7788   5.939 6.29e-09 ***
## Income                   6.2182     0.5935  10.477  < 2e-16 ***
## Graduate1              420.6187   206.1444   2.040    0.042 *
## Undergraduate1         496.3933   119.5927   4.151 4.06e-05 ***
## Income:Graduate1        -1.3420     4.1408  -0.324    0.746
## Income:Undergraduate1  -2.1922     1.8889  -1.161    0.247
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 392.6 on 394 degrees of freedom
## Multiple R-squared:  0.2801, Adjusted R-squared:  0.271
## F-statistic: 30.66 on 5 and 394 DF,  p-value: < 2.2e-16
```

From the result above, we can see that the incomes of graduate(undergraduate) student or not are not statistically significant at a significant level $\alpha = 0.1$. But the income itself and the fact if the customer is a undergraduate student or not are related to the credit card balance. For not graduate nor undergraduate customer, with \$622 higher income, the credit balance will increase by \$100

(d) Using the coding in (b), write out an expression for a linear model to predict `balance` using `income`, `student` and interaction between `income` and `student`.

```
dat_d <- dat%>% select(Balance, Income, Graduate, NotStudent)

model.fit_d <- lm(data=dat_d, Balance ~ . + Income*Graduate + Income*NotStudent)

summary(model.fit_d)
```

```
##
## Call:
## lm(formula = Balance ~ . + Income * Graduate + Income * NotStudent,
##     data = dat_d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -773.39 -325.70  -41.13  321.65  814.04
```

```
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         697.0165   114.7232   6.076 2.91e-09 ***
## Income                4.0260     1.7932   2.245   0.0253 *
## Graduate1           -75.7746   233.4865  -0.325   0.7457
## NotStudent1        -496.3933   119.5927  -4.151 4.06e-05 ***
## Income:Graduate1      0.8502     4.4732   0.190   0.8494
## Income:NotStudent1    2.1922     1.8889   1.161   0.2465
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 392.6 on 394 degrees of freedom
## Multiple R-squared:  0.2801, Adjusted R-squared:  0.271
## F-statistic: 30.66 on 5 and 394 DF,  p-value: < 2.2e-16
```

With a significant level of 0.1, we are not able to reject the hypothese that whether is graduate student or not does not affect the balance of credit card. The fact that is student or not will affect the credit card balance. With a significant level $\alpha \geq 0.05$, we will say the income is related to the credit card balance. In general, the income for student but not graduate student, with \$402 higher income, the credit balance will increase by \$100.

(e) Using simulated data to show that the fitted values from the models in (a) - (d) do not depend on the coding of the dummy variables.

```r
dat<- dat%>% select(Balance, Income, Graduate, NotStudent, Undergraduate)
model.predict_a <- predict(model.fit_a, newdata=dat)
model.predict_b <- predict(model.fit_b, newdata=dat)
model.predict_c <- predict(model.fit_c, newdata=dat)
model.predict_d <- predict(model.fit_d, newdata=dat)
```

We want to see if the predictions are identical. Set the tolerence to be 10^{-10}

```r
tol = 1e-10
any(abs(model.predict_a - model.predict_b) > tol)
```

```
## [1] FALSE
```

So we know that the prediction from the model in part(a) is identical with the model in part(b). Similarly, we have

```r
any(abs(model.predict_c - model.predict_d) > tol)
```

```
## [1] FALSE
```

So the prediction from the model in part(c) is also identical with the model in part(d).