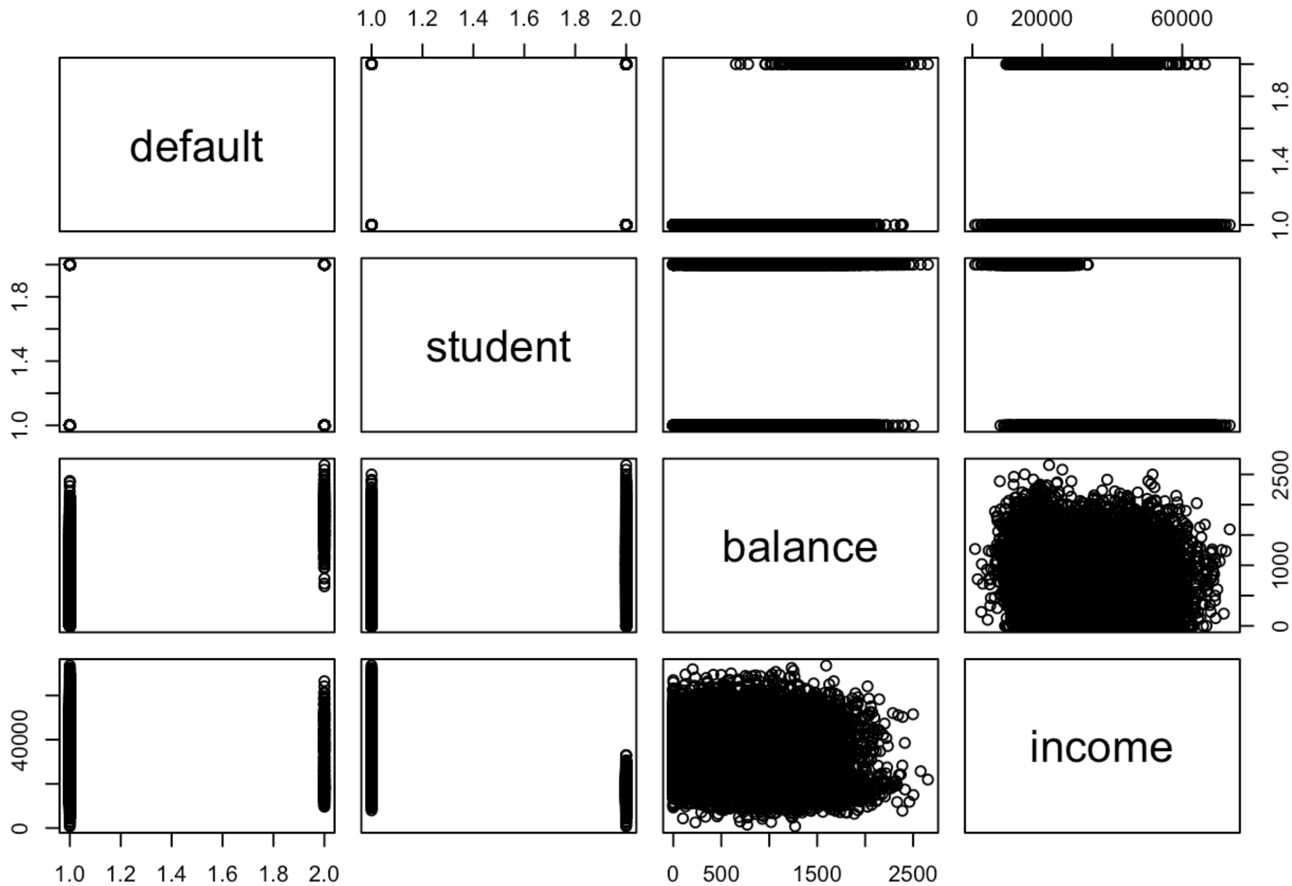
Problem 1

Chapter 5, Exercise 5 (Sec. 5.4, p. 198).

Part A




```
fit = glm(default ~ income + balance, family = 'binomial')
coef(fit)
```

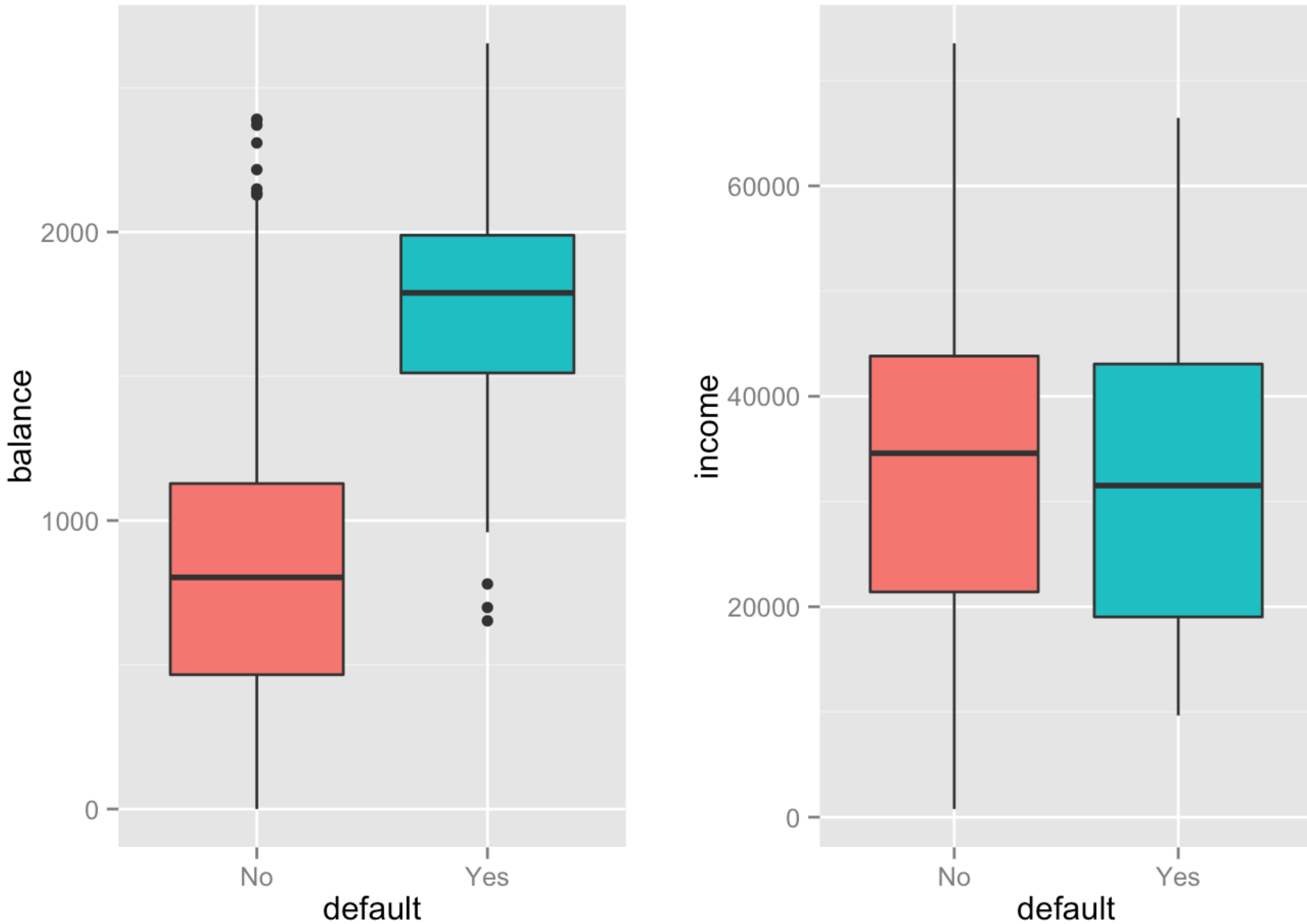
```
##   (Intercept)        income       balance
## -1.154047e+01  2.080898e-05  5.647103e-03
```

```
tmp = table(Default$default)
percent_defaults = (tmp[[2]]/tmp[[1]])*100
cat(percent_defaults, "percent of people default")
```

```
## 3.444709 percent of people default
```

```
# The following code is inspired by: rpubs.com/ryankelly/21379
x = qplot(x = balance, y = income, color = default, geom = 'point') + scale_shap
e(solid = FALSE)
y = qplot(x = default, y = balance, fill = default, geom = 'boxplot') + guides(fil
l = FALSE)
z = qplot(x = default, y = income, fill = default, geom = 'boxplot') + guides(fill
= FALSE)
x
```


```
grid.arrange(y, z, nrow=1)
```



Part B

```
set.seed(1)

total = nrow(Default)
num   = floor(0.9 * total)

sampled       = Default[sample(total), ]
Default.train = sampled[1:num, ]
Default.test  = sampled[(num + 1):total, ]
fit = multinom(default ~ income + balance, data = Default.train, family = 'binomia
l')
```

```
## # weights:  4 (3 variable)
## initial  value 6238.324625
## iter  10 value 716.944726
## final  value 716.865607
## converged
```

```
print(summary(fit))
```

```
## Call:
## multinom(formula = default ~ income + balance, data = Default.train,
##     family = "binomial")
##
## Coefficients:
##                    Values    Std. Err.
## (Intercept) -1.152515e+01 4.576857e-08
## income       2.027518e-05 4.415690e-06
## balance      5.668073e-03 9.492261e-05
##
## Residual Deviance: 1433.731
## AIC: 1439.731
```

```
pred = predict(fit, Default.test)
print(confusionMatrix(pred, Default.test$default)$table)
```

```
##           Reference
## Prediction  No Yes
##        No  965  19
##        Yes   6  10
```

Part C

```
##
## ====== Cross validation run # 2 =========================
## # weights:  4 (3 variable)
## initial  value 6238.324625
## iter  10 value 685.892190
## final  value 685.885544
## converged
##           Reference
## Prediction  No Yes
##        No  956  29
##        Yes   2  13
## MSE = 0.0475
## ====== Cross validation run # 3 =========================
## # weights:  4 (3 variable)
## initial  value 6238.324625
## iter  10 value 702.235743
## final  value 702.233405
## converged
##           Reference
## Prediction  No Yes
##        No  966  22
##        Yes   4   8
## MSE = 0.0443
## ====== Cross validation run # 4 =========================
## # weights:  4 (3 variable)
## initial  value 6238.324625
## iter  10 value 691.160350
## final  value 691.157602
## converged
##           Reference
## Prediction  No Yes
##        No  960  29
##        Yes   4   7
## MSE = 0.0433
```

Each of the 4 runs gave in similar results with just a little bit of variation:

- The 0-1 loss for each run was `25/1000`, `31/1000`, `26/1000`, and `33/1000` respectively.
- Of these errors, the respective ratios of false positives to false negatives were `6:19`, `2:29`, `4:22`, and `4:29`.
- We missed `10/(10 + 19) = .35`, `13/(29 + 13) = .31`, `8/(8 + 22) = .27`, and `7/(7 + 29) = .19` of defaults.
- We missed `6/(6 + 965) = .0062`, `2/(2 + 956) = .0021`, `4/(4 + 966) = .0041`, and `4/(960 + 4) = .0041` of non-defaults.

Our model does a pretty good job at categorizing non-defaults correctly (missing < 1%), but it fails miserably on actual defaults (missing upwards of 20% and as bad as 35%).

Part D

```
##
## ====== Cross validation run # 1 =========================
## # weights:  5 (4 variable)
## initial  value 6238.324625
## iter  10 value 712.052175
## final  value 711.984415
## converged
## MSE = 0.0465
## ====== Cross validation run # 2 =========================
## # weights:  5 (4 variable)
## initial  value 6238.324625
## iter  10 value 683.842024
## final  value 683.604493
## converged
## MSE = 0.0467
## ====== Cross validation run # 3 =========================
## # weights:  5 (4 variable)
## initial  value 6238.324625
## iter  10 value 697.797874
## final  value 697.507099
## converged
## MSE = 0.0443
## ====== Cross validation run # 4 =========================
## # weights:  5 (4 variable)
## initial  value 6238.324625
## iter  10 value 687.959598
## final  value 687.757482
## converged
## MSE = 0.0433
```

Including the student variable didn’t affect our MSE significantly. Since it has no notable effect on our results, it’s best to just remove the student variable from our analysis entirely.