Math 158, Spring 2018
Jo Hardin
Multiple Regression III – R code
Multicollinearity

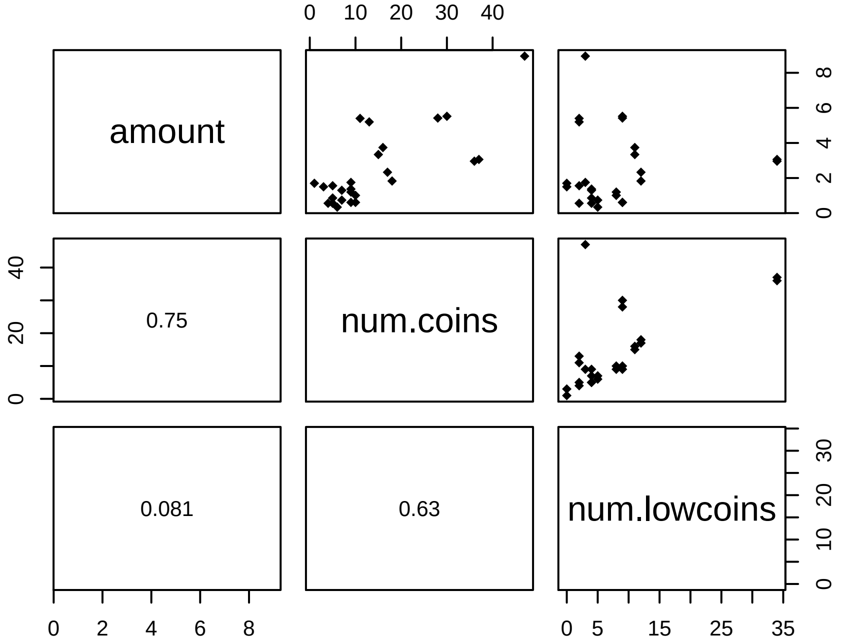
Consider the multiple regression model:

$$\begin{aligned} E[Y] &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 \\ Y &= \text{amount of money in pocket} \\ X_1 &= \#\text{ of coins in pocket} \\ X_2 &= \#\text{ of pennies, nickels, dimes in pocket} \end{aligned}$$

Using a completely non-random sample, I got the following data:

```
amount <- c(1.37, 1.01, 1.5, 0.56, 0.61, 3.06, 5.42, 1.75, 5.4, 0.56, 0.34, 2.33,
            3.34, 1.3, 1.2, 1.7, 0.86, 0.61, 2.96, 5.52, 8.95, 5.2, 1.56, 0.74, 1.83, 3.74)
num.coins <- c(9,10,3,5,10,37,28,9,11,4,6,17,15,7,9,1,5,9,36,30,47,13,5,7,18,16)
num.lowcoins <- c(4,8,0,4,9,34,9,3,2,2,5,12,11,4,8,0,4,9,34,9,3,2,2,5,12,11)

pairs(cbind(amount, num.coins, num.lowcoins), lower.panel=panel.cor, pch=18)
```



```
summary( lm(amount ~ num.coins) )$coef

##               Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) 0.5455854 0.44310536 1.231277 2.301490e-01
## num.coins   0.1341547 0.02422195 5.538560 1.070416e-05
```

```
summary( lm(amount ~ num.lowcoins) )$coef

##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  2.27979939 0.58478523 3.8985242 0.0006806597
## num.lowcoins 0.02012241 0.05082503 0.3959153 0.6956652745

summary( lm(amount ~ num.coins + num.lowcoins))$coef

##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)   0.7981116 0.30066739  2.654467 1.416556e-02
## num.coins     0.2062375 0.02085594  9.888671 9.438697e-10
## num.lowcoins -0.1602916 0.02908845 -5.510489 1.326852e-05

anova(lm(amount ~ num.coins + num.lowcoins))

## Analysis of Variance Table
##
## Response: amount
##              Df Sum Sq Mean Sq F value    Pr(>F)
## num.coins     1 63.363  63.363  68.209 2.473e-08 ***
## num.lowcoins  1 28.208  28.208  30.366 1.327e-05 ***
## Residuals    23 21.366   0.929
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Effects of Multicollinearity

In reality, there is always some degree of correlation between the explanatory variables (pg 283). for regression models, it is important to understand the entire context of the model, particularly for correlated variables.

1. Regardless of the degree of multicollinearity, our ability to obtain a good fit and make predictions (mean or individual) is not inhibited.

2. If the variables are highly correlated, many different linear combinations of them will produce equally good fits. That is, different samples from the same population may produce wildly different estimated coefficients. For this reason, the variability associated with the coefficients can be quite high. Additionally, the explanatory variables can be statistically not significant even though a definite relationship exists between the response and the set of predictors.

3. We can no longer interpret the coefficient to mean "the change in response when this variable increases by one unit and the others are held constant" because it may be impossible to hold the other variables constant. The regression coefficients do not reflect any inherent effect of the particular predictor variable on the response but rather a marginal or partial effect given whatever other correlated predictor variables are included in the model.