

Consider the multiple regression model:

$$E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

$Y$  = amount of money in pocket  
 $X_1$  = # of coins in pocket  
 $X_2$  = # of pennies, nickels, dimes in pocket

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)
```

```
> num.coins <- c(9,10,3,5,10,37,28,9,11,4,6,17,15)
```

```
> num.lowcoins <- c(4,8,0,4,9,34,9,3,2,2,5,12,11)
```

```
> summary( lm(amount ~ num.coins) )
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.73222	0.66848	1.095	0.2968
num.coins	0.10812	0.04237	2.552	0.0269 *

```
> summary( lm(amount ~ num.lowcoins) )
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.73038	0.67992	2.545	0.0272 *
num.lowcoins	0.04617	0.05909	0.781	0.4512

```
> coin.lm <- lm(amount ~ num.coins + num.lowcoins)
```

```
> summary(coin.lm)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.30724	0.46569	0.660	0.524321
num.coins	0.29648	0.05778	5.132	0.000443 ***
num.lowcoins	-0.24629	0.06561	-3.754	0.003762 **

Residual standard error: 0.9781 on 10 degrees of freedom

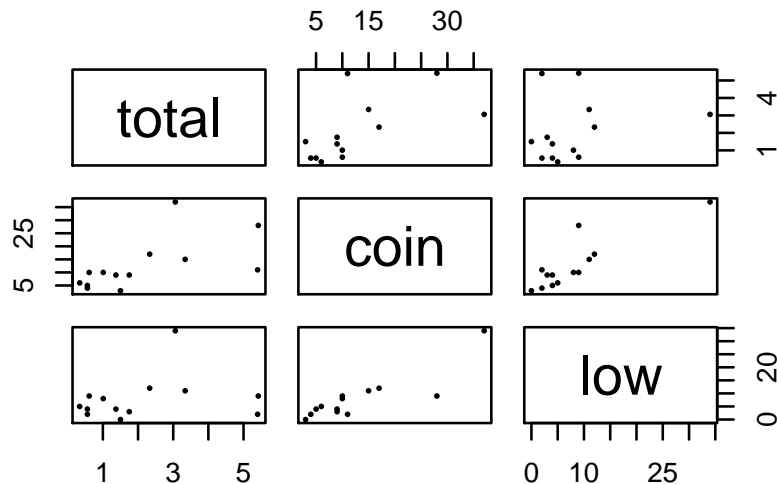
Multiple R-squared: 0.7392, Adjusted R-squared: 0.6871

F-statistic: 14.17 on 2 and 10 DF, p-value: 0.001206

```
> anova(coin.lm)
Analysis of Variance Table
```

Response: amount

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
num.coins	1	13.6423	13.6423	14.260	0.003624 **
num.lowcoins	1	13.4794	13.4794	14.089	0.003762 **
Residuals	10	9.5670	0.9567		



### 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.