Chapter 7: Multiple Regression II

STAT 3240

Michael McIsaac

UPEI

Learning Objectives for Sections 7.1-7.3

After Sections 7.1-7.3, you should be able to

- Understand the concept of the extra sums of squares principle
- Conduct and interpret tests concerning regression coefficients using ESS principle

7.1 Extra Sums of Squares

An extra sum of squares measures the marginal reduction in the error sum of squares when one or several predictor variables are added to the regression model, given that other predictor variables are already in the model.

Equivalently, one can view an extra sum of squares as measuring the marginal increase in the regression sum of squares when one or several predictor variables are added to the regression model.

An extra sum of squares involves the difference between

- the regression sum of squares for the regression model containing both the original X variable(s) and the new X variable(s) and
- the regression sums of squares for the regression model containing the X variable(s) already in the model

E.g., if X_1 is the "extra" variable:

$$SSR(X_1|X_2) = SSR(X_1, X_2) - SSR(X_2)$$

If X_2 is the "extra" variable:

$$SSR(X_2|X_1) = SSR(X_1, X_2) - SSR(X_1)$$

Decomposition of *SSR* into Extra Sums of Squares

Notice that we can decompose $SSR(X_1, X_2)$ as

$$SSR(X_1, X_2) = SSR(X_1|X_2) + SSR(X_2)$$

or

$$SSR(X_1, X_2) = SSR(X_2|X_1) + SSR(X_1)$$

These get at different questions:

- How much variability in Y is explained by X_2 alone? how much *additional* variability is explained by adding in X_1?

vs

- How much variability in Y is explained by X_1 alone? how much *additional* variability is explained by adding in X_2?

Note that the `R` function `anova` provides *Sequential* or *Extra sums of squares* which reports how much variation is explained by the variable after accounting for everything that has *previously* been added to the model

- (e.g., $SSR(X_1), SSR(X_2|X_1), SSR(X_3|X_1, X_2)$, etc).

However, very similar looking functions (e.g., `Anova` or even the t-tests reported in the `summary` of `lm`) will commonly report *Adjusted* or *Type II sums of squares* that show how much variation is explained by the variable after accounting for everything else that *will be* added to the model

- (e.g., $SSR(X_1|X_2, X_3)$, $SSR(X_2|X_1, X_3)$, $SSR(X_3|X_1, X_2)$).

Notice how the *Sequential sums of squares* differ when the order in which variables are added changes:

```
clothing_model = lm(clothing_expenditure~income+sex+food_expenditure+
anova(clothing_model)
```

```
## Analysis of Variance Table
##
## Response: clothing_expenditure
##                            Df   Sum Sq  Mean Sq F value  Pr(>F)
## income                      1 2.75e+07 27477746   12.71 0.00046
## sex                         1 1.15e+07 11511976    5.33 0.02205
## food_expenditure            1 5.94e+07 59433610   27.50 4.1e-07
## recreation_expenditure      1 4.05e+07 40522542   18.75 2.4e-05
## miscellaneous_expenditure   1 8.71e+03     8711    0.00 0.94944
## Residuals                 194 4.19e+08  2161097
```

Notice how the *Sequential sums of squares* differ when the order in which variables are added changes:

```
clothing_model_reordered = lm(clothing_expenditure~miscellaneous_exp
anova(clothing_model_reordered)
```

```
## Analysis of Variance Table
##
## Response: clothing_expenditure
##                            Df   Sum Sq  Mean Sq F value  Pr(>F)
## miscellaneous_expenditure   1 4.98e+06  4984244    2.31  0.1305
## income                      1 2.32e+07 23171360   10.72  0.0013
## sex                         1 1.15e+07 11459552    5.30  0.0224
## food_expenditure            1 5.92e+07 59168738   27.38 4.3e-07
## recreation_expenditure      1 4.02e+07 40170692   18.59 2.6e-05
## Residuals                 194 4.19e+08  2161097
```

Notice how the *Adjusted sums of squares* don't differ when the order in which variables are added changes:

```
clothing_model = lm(clothing_expenditure~income+sex+food_expenditure+
Anova(clothing_model)
```

```
## Anova Table (Type II tests)
##
## Response: clothing_expenditure
##                             Sum Sq  Df F value  Pr(>F)
## income                    1.80e+07   1    8.34  0.0043
## sex                       1.85e+07   1    8.58  0.0038
## food_expenditure          3.80e+07   1   17.57 4.2e-05
## recreation_expenditure    4.02e+07   1   18.59 2.6e-05
## miscellaneous_expenditure 8.71e+03   1    0.00  0.9494
## Residuals                 4.19e+08 194
```

Notice how the *Adjusted sums of squares* don't differ when the order in which variables are added changes:

```
clothing_model_reordered = lm(clothing_expenditure~miscellaneous_exp
Anova(clothing_model_reordered)
```

```
## Anova Table (Type II tests)
##
## Response: clothing_expenditure
##                              Sum Sq  Df F value  Pr(>F)
## miscellaneous_expenditure 8.71e+03   1    0.00  0.9494
## income                    1.80e+07   1    8.34  0.0043
## sex                       1.85e+07   1    8.58  0.0038
## food_expenditure          3.80e+07   1   17.57 4.2e-05
## recreation_expenditure    4.02e+07   1   18.59 2.6e-05
## Residuals                 4.19e+08 194
```

Notice again how the *Adjusted sums of squares* don't differ:

```
msummary(clothing_model)
```

```
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  3.90e+02   3.68e+02    1.06   0.2898
## income                       1.85e-02   6.39e-03    2.89   0.0043
## sexmale                     -6.31e+02   2.15e+02   -2.93   0.0038
## food_expenditure             1.61e-01   3.85e-02    4.19  4.2e-05
## recreation_expenditure       2.38e-01   5.52e-02    4.31  2.6e-05
## miscellaneous_expenditure -1.47e-02   2.32e-01   -0.06   0.9494
##
## Residual standard error: 1470 on 194 degrees of freedom
## Multiple R-squared:  0.249,    Adjusted R-squared:  0.23
## F-statistic: 12.9 on 5 and 194 DF,  p-value: 8.32e-11
```

```
msummary(clothing_model_reordered)
```

```
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  3.90e+02   3.68e+02    1.06   0.2898
## miscellaneous_expenditure -1.47e-02   2.32e-01   -0.06   0.9494
## income                       1.85e-02   6.39e-03    2.89   0.0043
## sexmale                     -6.31e+02   2.15e+02   -2.93   0.0038
## food_expenditure             1.61e-01   3.85e-02    4.19  4.2e-05
## recreation_expenditure       2.38e-01   5.52e-02    4.31  2.6e-05
##
## Residual standard error: 1470 on 194 degrees of freedom
## Multiple R-squared:  0.249,    Adjusted R-squared:  0.23
## F-statistic: 12.9 on 5 and 194 DF,  p-value: 8.32e-11
```

Test Whether All $\beta_k = 0$

This is the *overall F test* of whether or not there is a regression relation between the response variable Y and the set of X variables:

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$

H_a : not all $\beta_k (k = 1, \ldots, p-1)$ equal 0

and the test statistic is:

$$F^* = \frac{SSR(X_1, \ldots, X_{p-1})}{p-1} \div \frac{SSE(X_1, \ldots, X_{p-1})}{n-p} = \frac{MSR}{MSE}$$

If H_0 holds, $F^* \sim F(p-1, n-p)$. Large values of F^* lead to conclusion H_a.

Test Whether a Single $\beta_k = 0$

This is a *partial F test* of whether a particular regression coefficient β_k equals 0:

$H_0 : \beta_k = 0$

$H_a : \beta_k \neq 0$

and the test statistic is:

$$\begin{aligned} F^* &= \frac{SSR(X_k|X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_{p-1})}{1} \div \frac{SSE(X_1, \ldots, X_{p-1})}{n-p} \\ &= \frac{MSR(X_k|X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_{p-1})}{MSE} \end{aligned}$$

If H_0 holds, $F^* \sim F(1, n-p)$. Large values of F^* lead to conclusion H_a.

An equivalent test statistic is

$$t^* = \frac{b_k}{s\{b_k\}}$$

Test Whether Some $\beta_k = 0$

This is another *partial F test*:

$H_0 : \beta_q = \beta_{q+1} = \cdots = \beta_{p-1} = 0$

H_a : not all of these β_k equal 0

and the test statistic is:

$$\begin{aligned} F^* &= \frac{SSR(X_q, \ldots, X_{p-1} | X_1, \ldots, X_{q-1})}{p-q} \div \frac{SSE(X_1, \ldots, X_{p-1})}{n-p} \\ &= \frac{MSR(X_q, \ldots, X_{p-1} | X_1, \ldots, X_{q-1})}{MSE} \end{aligned}$$

If H_0 holds, $F^* \sim F(p-q, n-p)$. Large values of F^* lead to conclusion H_a.

Notice that the previous two tests were just special cases of this one with $q = 1$ and $p - q = 1$.

Other Tests

These extra sums of squares tests - where we are testing whether one or several β_k is equal to 0 - are special cases of the general linear test appraoch.

However, we can answer an even broader range of questions using the general linear test approach.

Consider testing whether $\beta_1 = \beta_2$ in the full model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$$

This is equivalent to testing the adequacy of the reduced model

$$Y_i = \beta_0 + \beta_1 (X_{i1} + X_{i2}) + \beta_3 X_{i3} + \varepsilon_i$$

which we can accomplish using the general F^* test statistic (2.70) with 1 and $n - 4$ degrees of freedom.

Similarly, we might want to test whether $\beta_1 = 3$ and $\beta_3 = 5$ in the full model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$$

which could be tested by testing the adequacy of the reduced model

$$Y_i - 3X_{i1} - 5X_{i3} = \beta_0 + \beta_2 X_{i2} + \varepsilon_i$$

using the general linear test statistic F^* with 2 and $n - 4$ degrees of freedom.

```
clothing_model = lm(clothing_expenditure~income+sex+food_expenditure+
msummary(clothing_model)
```

```
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  2.64e+02   2.32e+02    1.14  0.25474
## income                       1.46e-02   3.97e-03    3.69  0.00025
## sexmale                     -4.25e+02   1.35e+02   -3.15  0.00174
## food_expenditure             1.51e-01   2.45e-02    6.18  1.4e-09
## recreation_expenditure       2.29e-01   3.36e-02    6.81  2.8e-11
## miscellaneous_expenditure    1.60e-01   1.39e-01    1.16  0.24846
##
## Residual standard error: 1480 on 494 degrees of freedom
## Multiple R-squared:  0.237,    Adjusted R-squared:  0.229
## F-statistic: 30.7 on 5 and 494 DF,  p-value: <2e-16
```

```
clothing_model_reordered = lm(clothing_expenditure~miscellaneous_exp
msummary(clothing_model_reordered)
```

```
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  2.64e+02   2.32e+02    1.14  0.25474
## miscellaneous_expenditure    1.60e-01   1.39e-01    1.16  0.24846
## income                       1.46e-02   3.97e-03    3.69  0.00025
## sexmale                     -4.25e+02   1.35e+02   -3.15  0.00174
## food_expenditure             1.51e-01   2.45e-02    6.18  1.4e-09
## recreation_expenditure       2.29e-01   3.36e-02    6.81  2.8e-11
##
## Residual standard error: 1480 on 494 degrees of freedom
## Multiple R-squared:  0.237,    Adjusted R-squared:  0.229
## F-statistic: 30.7 on 5 and 494 DF,  p-value: <2e-16
```

```
anova(clothing_model)
```

```
## Analysis of Variance Table
##
## Response: clothing_expenditure
##                             Df   Sum Sq  Mean Sq F value  Pr(>F)
## income                       1 7.13e+07 7.13e+07   32.50 2.1e-08
## sex                          1 1.25e+07 1.25e+07    5.71   0.017
## food_expenditure             1 1.44e+08 1.44e+08   65.66 4.2e-15
## recreation_expenditure       1 1.06e+08 1.06e+08   48.42 1.1e-11
## miscellaneous_expenditure    1 2.93e+06 2.93e+06    1.34   0.248
## Residuals                  494 1.08e+09 2.19e+06
```

```
anova(clothing_model_reordered)
```

```
## Analysis of Variance Table
##
## Response: clothing_expenditure
##                             Df   Sum Sq  Mean Sq F value  Pr(>F)
## miscellaneous_expenditure    1 2.89e+07 2.89e+07   13.20 0.00031
## income                       1 5.16e+07 5.16e+07   23.54 1.6e-06
## sex                          1 1.31e+07 1.31e+07    5.98 0.01479
## food_expenditure             1 1.41e+08 1.41e+08   64.49 7.2e-15
## recreation_expenditure       1 1.02e+08 1.02e+08   46.41 2.8e-11
## Residuals                  494 1.08e+09 2.19e+06
```

- Based on the extra sums of squares principle, is β_{misc} (the slope parameter associated with miscellaneous_expenditure) equal to zero? Justify your answer and explain what your finding means in the context of the problem.

- The general linear test statistic to test if Bminc = 0 is F* = 1.335069394 (1st ANOVA table), which has a p-value 0.2484639164 > 0.1, so we fail to prove Bmisc different than 0 (fail to reject the null hypothesis). So, adding Xmisc to the model, when the other 4 X variables are already in the model, does not give a good marginal increase for SSR to be valuable to the model. So, the reduced model that excludes BmiscXmisc is appropriate, and we can drop BmiscXmisc.

- With the F value being 13.19 for miscellaneous expenditure, and the p value being close to zero it tells us that the chance of Bmisc being equal to zero is very unlikely. We conclude that miscellaneous expenditure is an important predictor variable for our model (it cannot be dropped) which is designed to predict response behavior of clothing expenditure.

- Based on the extra sums of squares principle, is β_{misc} (the slope parameter associated with miscellaneous_expenditure) equal to zero? Justify your answer and explain what your finding means in the context of the problem.

- Yes, based on the difference in the sum of squares and also the p values between the full model and the reduced model, we see that there is no relation associated the miscellaneous expenditure and therefore, we conclude the slope equals zero.

- When miscellaneous expenditure is added first there is way more error associated with it compared to when it is added last. The large p value when it is added last compared to first also tell us that adding miscellaneous expenditure to the model doesn't really make it any better, because of this I would say that we can't say for sure that Beta for miscellaneous expenditure isn't equal to zero. Miscellaneous expenditure isn't a good predictor of clothing expenditure.

```
with(spending_subset, data.frame(income, food=food_expenditure, sex=
```

```
##         income    food     sex    rec    misc
## income 1.00000 0.08276 0.16925 0.1960 0.29163
## food   0.08276 1.00000 0.04664 0.2528 0.05981
## sex    0.16925 0.04664 1.00000 0.1050 0.07485
## rec    0.19598 0.25284 0.10501 1.0000 0.15591
## misc   0.29163 0.05981 0.07485 0.1559 1.00000
```

- Are men and women different in terms of the amount they spend on clothing? Justify your answer.
- Different in that the expected average clothing expenditure for males with zero income is less than that for females of similar income. The rate of change for both however is equal.
- Yes. As we move from considering women to considering men, we see a decrease of $425. This means that, on average, a male will spend $425 less on clothing than a female.
- We need to do a F test, using one reduced model(drop the predictor variable sex) and a full model(using all predictor variable), SSE of reduced model - SSE of full model and divide the value by # of degree freedom and divided the value by the mean square error of the full model, if the F value is large then we conclude the men and women are different in terms of the amount they spend on clothing.
- The sex category seems to not be equal to zero so I would say that it is a predictor that adds to the error reduction, therefore men and women must spend differently.

- Are men and women different in terms of the amount they spend on clothing? Justify your answer.
- Men and women are different in terms of the amount they spend on clothing. When income is zero, the average amount women spend on clothing is 263.9887 dollars, and for men is 0.
- Yes, they are different. We know this because the slope of the male regression line is negative, meaning the females in our dataset spend more on clothes than males.

- Are all β_k equal to 0 here?
- No, for the F-statistics is 30.72419 on 5 and 494 DF, with p-value almost 0, which implies rejection of null.
- No, the f-statistic is too large. Also the very small p-value (2.2204e-16) indicates that we should reject the null hypothesis and conclude that at least some of the slope parameters are not equal to zero.

Recap: Sections 7.1-7.3

After Sections 7.1-7.3, you should be able to

- Understand the concept of the extra sums of squares principle
- Conduct and interpret tests concerning regression coefficients using ESS principle

Learning Objectives for Sections 7.4, 7.6

After Sections 7.4 and 7.6, you should be able to

- Compute and interpret coefficients of partial determination
- Understand multicollinearity and its effects

7.4: Coefficients of Partial Determination

Recall that the *coefficient of multiple determination*, R^2, measures the proportionate reduction in the variation of Y achieved by the introduction of the entire set of X variables considered in the model.

A *coefficient of partial determination*, in contrast, measures the marginal contribution of one X variable when all others are already included in the model.

For example, the coefficient of partial determination between Y and X_2, given that X_1 is in the model is

$$R^2_{Y2|1} = \frac{SSR(X_2|X_1)}{SSE(X_1)}$$

That is, the coefficient of partial determination is the percent of variation that cannot be explained in the reduced model, but can be explained by the predictors specified in a fuller model.

Coefficients of partial determination can take on values between 0 and 1

The coefficient of partial derermination $R_{Y1|2}$ measures the relation between Y and X_1 when both of these variables have been adjusted for their linear relationships to X_2.

I.e., a coefficient of partial determination can be interpreted as a coefficient of simple determination of these residuals

Consider a multiple regression model with two X variables. Suppose we regress Y on X_2 and obtain the residuals:

$$e(Y|X_2) = Y_i - \hat{Y}_i(X_2)$$

where $\hat{Y}_i(X_2)$ denotes the fitted values of Y when X_2 is in the model.

Suppose we further regress X_1 on X_2 and obtain the residuals:

$$e(X_1|X_2) = X_{i1} - \hat{X}_{i1}(X_2)$$

where $\hat{X}_{i1}(X_2)$ denotes the fitted values of X_1 in the regression of X_1 on X_2.

The coefficient of simple determination R^2 between these two sets of residuals equals the coefficient of partial determination $R_{Y1|2}$

- The plot of the residuals $e(Y|X_2)$ against $e(X_1|X_2)$ provides a graphical representation of the strength of the relationship between Y and X_1, adjusted for X_2. Such plots of residuals, called added variable plots or partial regression plots, are discussed in Section 10.1.

Coefficients of Partial Correlation

The square root of a coefficient of partial determination is called a coefficient of partial correlation.

It is given the same sign as that of the corresponding regression coefficient in the fitted regression function.

Coefficients of partial correlation are frequently used in practice, although they do not have as clear a meaning as coefficients of partial determination.

One use of partial correlation coefficients is in computer routines for finding the best predictor variable to be selected next for inclusion in the regression model. We discuss this use in Chapter 9.

```
## Use all data from now on.
spending_subset=spending_subset_all[1:500,]
spending_subset %>% datatable()
```

Show 20 entries Search:

	province	type_of_dwelling	income	marital_status	age_gr
1	NL	single_detached	68000	never_married	30-34
2	NL	single_detached	48000	never_married	25-29
3	NL	single_detached	30000	married	35-39
4	NL	row_house	30000	never_married	30-34
5	NL	single_detached	35000	married	25-29
6	NL	single_detached	26000	married	25-29

Showing 1 to 20 of 500 entries

Previous 1 2 3 4 5 ... 25 Next

```
clothing_model = lm(clothing_expenditure~income+sex+food_expenditure+
msummary(clothing_model)
```

```
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  2.64e+02   2.32e+02    1.14  0.25474
## income                       1.46e-02   3.97e-03    3.69  0.00025
## sexmale                     -4.25e+02   1.35e+02   -3.15  0.00174
## food_expenditure             1.51e-01   2.45e-02    6.18  1.4e-09
## recreation_expenditure       2.29e-01   3.36e-02    6.81  2.8e-11
## miscellaneous_expenditure    1.60e-01   1.39e-01    1.16  0.24846
##
## Residual standard error: 1480 on 494 degrees of freedom
## Multiple R-squared:  0.237,    Adjusted R-squared:  0.229
## F-statistic: 30.7 on 5 and 494 DF,  p-value: <2e-16
```

```
anova(clothing_model)
```

```
## Analysis of Variance Table
##
## Response: clothing_expenditure
##                            Df   Sum Sq  Mean Sq F value  Pr(>F)
## income                      1 7.13e+07 7.13e+07   32.50 2.1e-08
## sex                         1 1.25e+07 1.25e+07    5.71   0.017
## food_expenditure            1 1.44e+08 1.44e+08   65.66 4.2e-15
## recreation_expenditure      1 1.06e+08 1.06e+08   48.42 1.1e-11
## miscellaneous_expenditure   1 2.93e+06 2.93e+06    1.34   0.248
## Residuals                 494 1.08e+09 2.19e+06
```

$$
\begin{aligned}
R^2_{Y2|1} &= \frac{SSR(X_2|X_1)}{SSE(X_1)} \\
&= \frac{2927550.874}{1083247161.692 + 2927550.874} \\
&\approx 0.002695
\end{aligned}
$$

- Compute and interpret the coefficient of partial determination for miscellaneous expenses.

- (Using the 1st ANOVA table) The coefficient of partial determination for miscellaneous expenses is equal to SSR (Xmisc | the other 4 Xs are in the model) divided by SSE(The 4 first Xs in the model). The answer is 0.0027026 (rounded). So, the variation in Y (that remained after including income, sex, food expenditures, and recreation expenditures in the model) is reduced by 0.2706% when Xmisc is introduced into the model.

- The coefficient of partial determination for miscellaneous expenses is 0.00270256962. This indicates that the marginal contribution of miscellaneous expenses after all other variables are included in the model is insignificant.

7.6: Multicollinearity and Its Effects

When the predictor variables are correlated among themselves, *intercorrelation* or *multi-collinearity* among them is said to exist.

- multi-collinearity generally refers to situations where the correlation among the predictor variables is very high.

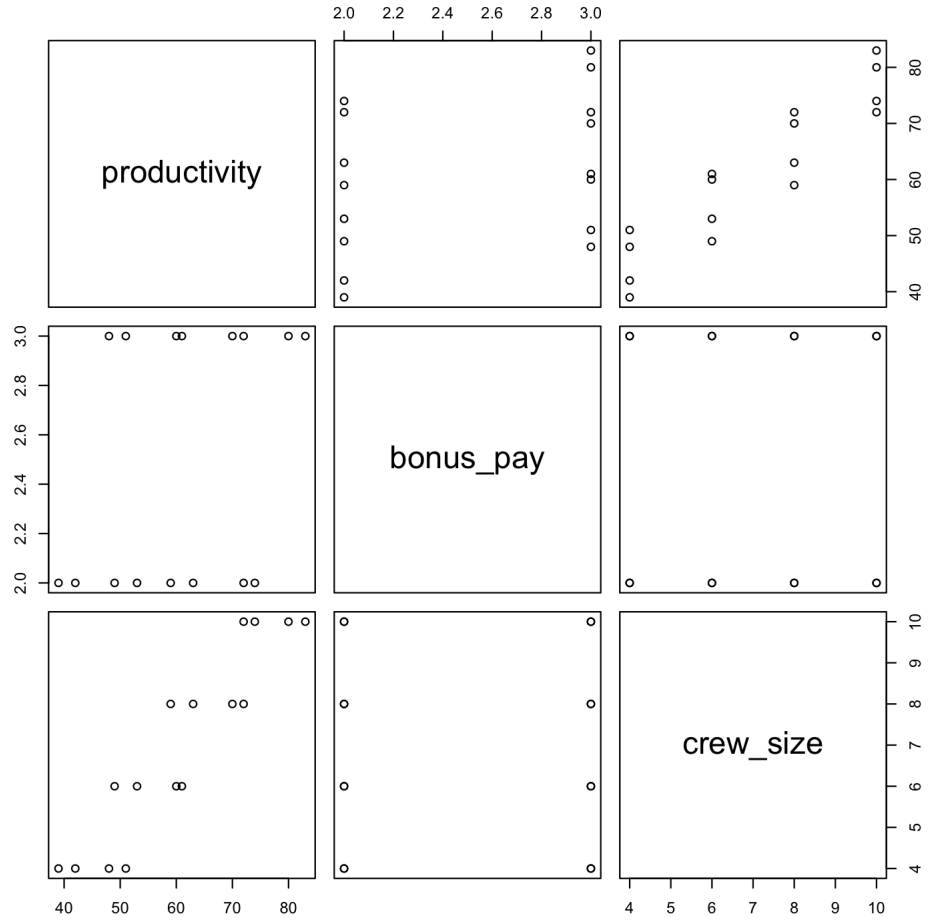
Example with uncorrelated predictor variables (Table 7.6):

Show 20 entries Search:

	productivity	bonus_pay	crew_size
1	42	2	4
2	39	2	4
3	48	3	4
4	51	3	4
5	49	2	6
6	53	2	6

Showing 1 to 16 of 16 entries Previous 1 Next

```
plot(crew_data)
```



```
cor(crew_data)      %>% round(3) %>% datatable()
```

Show 20 entries Search:

	productivity	bonus_pay	crew_size
productivity	1	0.358	0.926
bonus_pay	0.358	1	0
crew_size	0.926	0	1

Showing 1 to 3 of 3 entries Previous 1 Next

```
mod_full = lm(productivity~crew_size + bonus_pay, data=crew_data)
msummary(mod_full)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.425      2.487    0.17     0.87
## crew_size      5.350      0.186   28.78  3.7e-13
## bonus_pay      9.250      0.831   11.12  5.2e-08
##
## Residual standard error: 1.66 on 13 degrees of freedom
## Multiple R-squared:  0.987,    Adjusted R-squared:  0.984
## F-statistic:  476 on 2 and 13 DF,  p-value: 6.95e-13
```

```
anova(mod_full)
```

```
## Analysis of Variance Table
##
## Response: productivity
##           Df Sum Sq Mean Sq F value  Pr(>F)
## crew_size  1   2290    2290     828 3.7e-13
## bonus_pay  1    342     342     124 5.2e-08
## Residuals 13     36       3
```

```
mod_x1 = lm(productivity~crew_size, data=crew_data)
msummary(mod_x1)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   23.550      4.270    5.51  7.6e-05
## crew_size      5.350      0.581    9.21  2.6e-07
##
## Residual standard error: 5.2 on 14 degrees of freedom
## Multiple R-squared:  0.858,    Adjusted R-squared:  0.848
## F-statistic: 84.8 on 1 and 14 DF,  p-value: 2.57e-07
```

```
anova(mod_x1)
```

```
## Analysis of Variance Table
##
## Response: productivity
##           Df Sum Sq Mean Sq F value  Pr(>F)
## crew_size  1   2290    2290    84.8 2.6e-07
## Residuals 14    378      27
```

```
mod_x2 = lm(productivity~bonus_pay, data=crew_data)
msummary(mod_x2)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    37.88      16.43    2.31    0.037
## bonus_pay       9.25       6.44    1.44    0.173
##
## Residual standard error: 12.9 on 14 degrees of freedom
## Multiple R-squared:  0.128,    Adjusted R-squared:  0.066
## F-statistic: 2.06 on 1 and 14 DF,  p-value: 0.173
```

```
anova(mod_x2)
```

```
## Analysis of Variance Table
##
## Response: productivity
##           Df Sum Sq Mean Sq F value Pr(>F)
## bonus_pay  1    342     342    2.06   0.17
## Residuals 14   2326     166
```

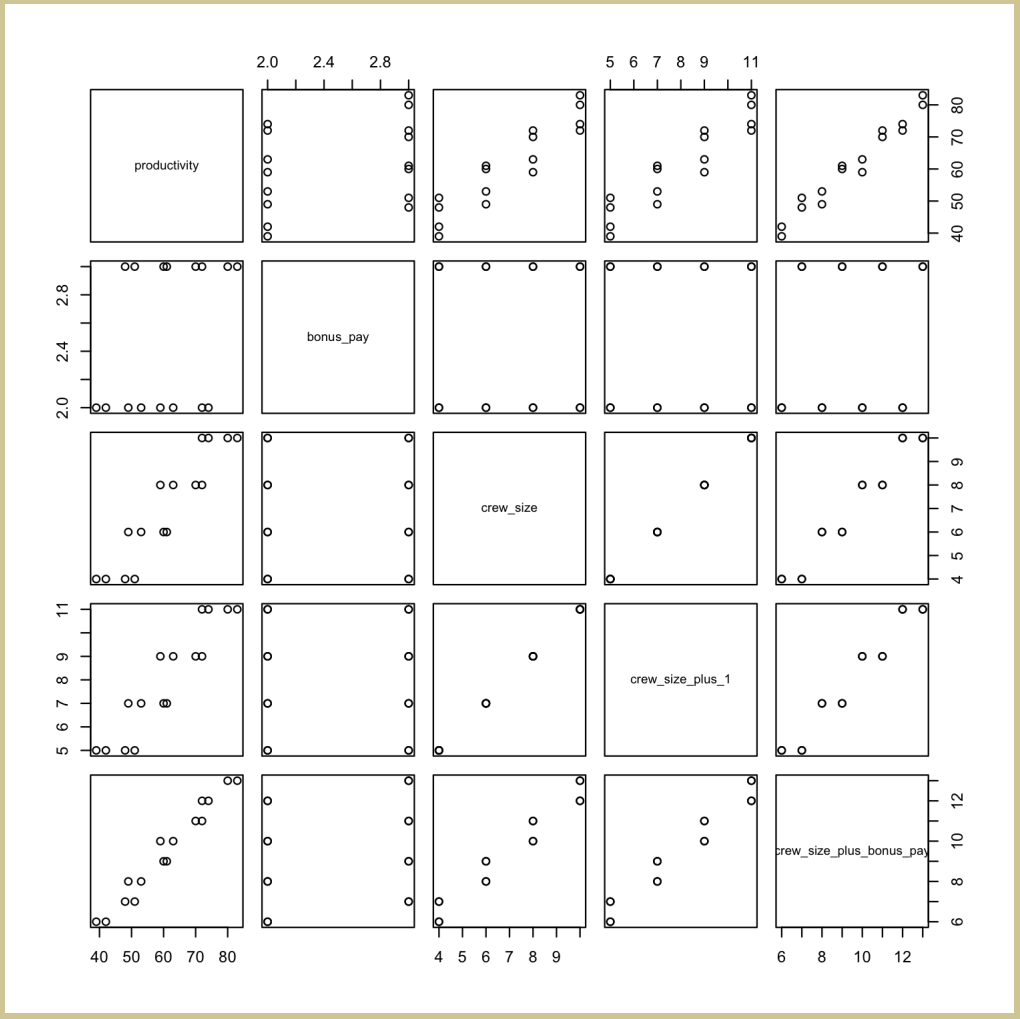
Example of the Problem with Perfect Multicollinearity

Show 20 entries Search:

	productivity	bonus_pay	crew_size	crew_size_plus_1	crew_s
1	42	2	4	5	
2	39	2	4	5	
3	48	3	4	5	
4	51	3	4	5	
5	49	2	6	7	
6	53	2	6	7	

Showing 1 to 16 of 16 entries Previous 1 Next

```
plot(crew_data)
```



```
cor(crew_data)      %>% round(3) %>% datatable()
```

Show 20 entries Search:

	productivity	bonus_pay	crew_size	cr
productivity	1	0.358	0.926	
bonus_pay	0.358	1	0	
crew_size	0.926	0	1	
crew_size_plus_1	0.926	0	1	
crew_size_plus_bonus_pay	0.982	0.218	0.976	

Showing 1 to 5 of 5 entries Previous 1 Next

```
lm(productivity~crew_size + crew_size_plus_1, data=crew_data)$coeffi
```

```
##      (Intercept)        crew_size crew_size_plus_1
##            23.55             5.35               NA
```

```
lm(productivity~crew_size + bonus_pay + crew_size_plus_bonus_pay, da
```

```
##              (Intercept)                crew_size                bonus_pay
##                    0.425                    5.350                    9.250
## crew_size_plus_bonus_pay
##                       NA
```

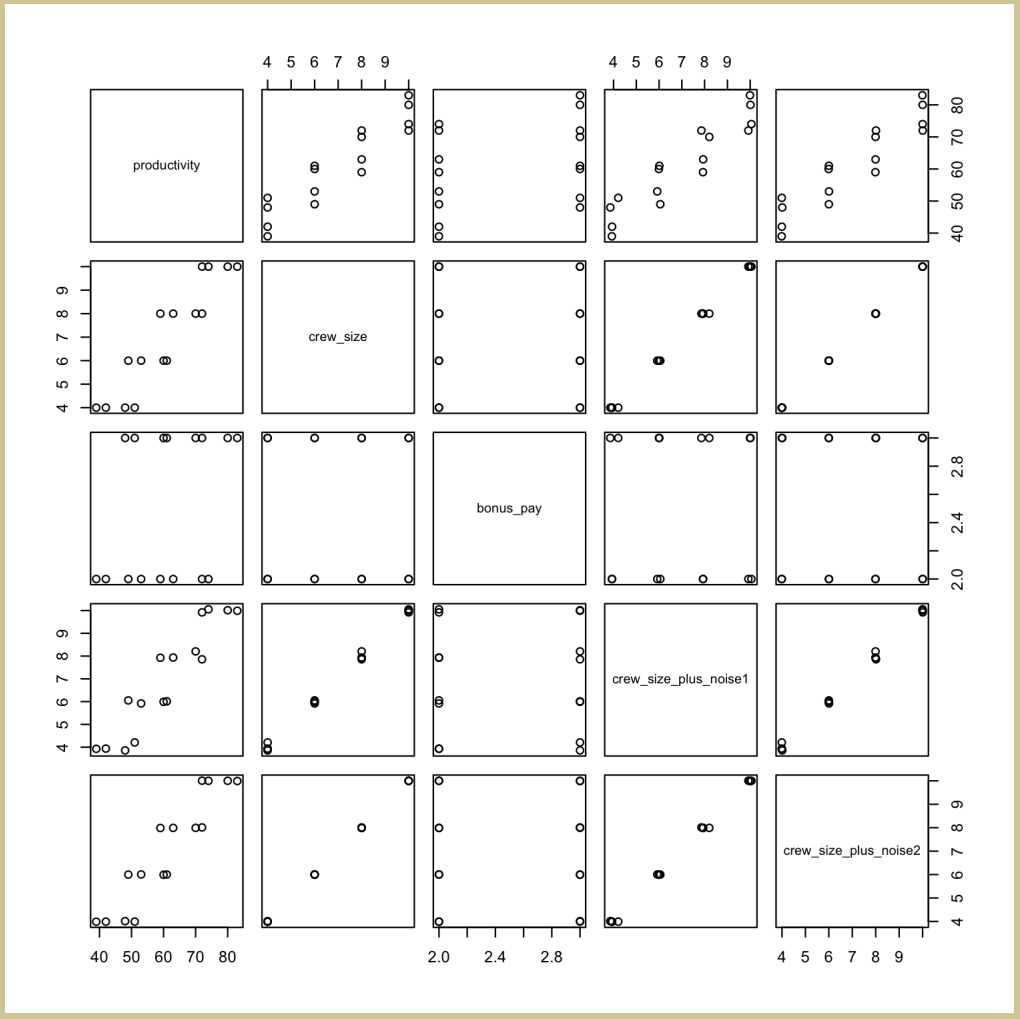
Example of the Problem with Multicollinearity

Show 20 entries Search:

	productivity	crew_size	bonus_pay	crew_size_plus_noise1	c
1	42	4	2	3.93662930363056	
2	39	4	2	3.92641348950914	
3	48	4	3	3.85884083266299	
4	51	4	3	4.2063792956606	
5	49	6	2	6.05311494981267	
6	53	6	2	5.92008599612028	

Showing 1 to 16 of 16 entries Previous 1 Next

```
plot(crew_data)
```



```
cor(crew_data)      %>% round(7) %>% datatable(options=list(scrollY=
```

Show 20 entries Search:

	productivity	crew_size	bonus_pay	crew_s
productivity	1	0.9264156	0.3581614	
crew_size	0.9264156	1	0	
bonus_pay	0.3581614	0	1	
crew_size_plus_noise1	0.9303856	0.9990012	0.0129906	
crew_size_plus_noise2	0.9266909	0.9999943	0.0006637	

Showing 1 to 5 of 5 entries Previous 1 Next

```
lm(productivity~crew_size, data=crew_data)$coefficients
```

```
## (Intercept)   crew_size
##       23.55        5.35
```

```
lm(productivity~crew_size + crew_size_plus_noise1, data=crew_data)$c
```

```
##           (Intercept)             crew_size crew_size_plus_noise1
##                23.838                -8.795                14.127
```

```
lm(productivity~crew_size + crew_size_plus_noise2, data=crew_data)$c
```

```
##           (Intercept)             crew_size crew_size_plus_noise2
##                 24.16               -135.83                141.13
```

Consider the effects of multicollinearity on $s\{b_k\}$:

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   23.550      4.270    5.51  7.6e-05
## crew_size      5.350      0.581    9.21  2.6e-07
##
## Residual standard error: 5.2 on 14 degrees of freedom
## Multiple R-squared:  0.858,    Adjusted R-squared:  0.848
## F-statistic: 84.8 on 1 and 14 DF,  p-value: 2.57e-07
```

```
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)              23.84       4.25    5.61  8.5e-05
## crew_size                -8.79      12.91   -0.68     0.51
## crew_size_plus_noise1    14.13      12.88    1.10     0.29
##
## Residual standard error: 5.16 on 13 degrees of freedom
## Multiple R-squared:  0.87,    Adjusted R-squared:  0.85
## F-statistic: 43.6 on 2 and 13 DF,  p-value: 1.72e-06
```

```
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)              24.16       4.39    5.51   0.0001
## crew_size              -135.83     173.64   -0.78   0.4481
## crew_size_plus_noise2   141.13     173.58    0.81   0.4308
##
## Residual standard error: 5.26 on 13 degrees of freedom
## Multiple R-squared:  0.865,    Adjusted R-squared:  0.844
## F-statistic: 41.7 on 2 and 13 DF,  p-value: 2.21e-06
```

Consider the effects of multicollinearity on Extra Sums of Squares

```
anova(lm(productivity~crew_size + crew_size_plus_noise1, data=crew_d
```

```
## Analysis of Variance Table
##
## Response: productivity
##                      Df Sum Sq Mean Sq F value  Pr(>F)
## crew_size             1   2290    2290    86.0 4.3e-07
## crew_size_plus_noise1  1     32      32     1.2    0.29
## Residuals             13    346      27
```

```
anova(lm(productivity~crew_size_plus_noise1 + crew_size, data=crew_d
```

```
## Analysis of Variance Table
##
## Response: productivity
##                      Df Sum Sq Mean Sq F value  Pr(>F)
## crew_size_plus_noise1  1   2309    2309   86.73 4.1e-07
## crew_size             1     12      12    0.46    0.51
## Residuals             13    346      27
```

Consider the effects of multicollinearity on Simultaneous Tests of β_k:

```
Anova(lm(productivity~crew_size + crew_size_plus_noise1, data=crew_d
```

```
## Anova Table (Type II tests)
##
## Response: productivity
##                       Sum Sq Df F value Pr(>F)
## crew_size                 12  1    0.46   0.51
## crew_size_plus_noise1     32  1    1.20   0.29
## Residuals                346 13
```

```
anova(lm(productivity~1, data=crew_data), lm(productivity~crew_size +
```

```
## Analysis of Variance Table
##
## Model 1: productivity ~ 1
## Model 2: productivity ~ crew_size + crew_size_plus_noise1
##   Res.Df  RSS Df Sum of Sq    F  Pr(>F)
## 1     15 2668
## 2     13  346  2      2322 43.6 1.7e-06
```

Consider the effects of multicollinearity on Fitted Values and Predictions:

```
predict(lm(productivity~crew_siz
```

```
##      fit   lwr   upr
## 1  44.95 32.87 57.03
## 2  44.95 32.87 57.03
## 3  44.95 32.87 57.03
## 4  44.95 32.87 57.03
## 5  55.65 44.09 67.21
## 6  55.65 44.09 67.21
## 7  55.65 44.09 67.21
## 8  55.65 44.09 67.21
## 9  66.35 54.79 77.91
## 10 66.35 54.79 77.91
## 11 66.35 54.79 77.91
## 12 66.35 54.79 77.91
## 13 77.05 64.97 89.13
## 14 77.05 64.97 89.13
## 15 77.05 64.97 89.13
## 16 77.05 64.97 89.13
```

```
predict(lm(productivity~crew_siz
```

```
##      fit   lwr   upr
## 1  44.27 32.12 56.43
## 2  44.13 31.94 56.32
## 3  43.17 30.59 55.76
## 4  48.08 34.51 61.66
## 5  56.58 44.88 68.29
## 6  54.71 43.00 66.41
## 7  56.02 44.43 67.60
## 8  55.71 44.15 67.27
## 9  65.60 53.95 77.26
## 10 65.46 53.77 77.15
## 11 64.51 52.39 76.62
## 12 69.42 56.37 82.46
## 13 77.92 65.71 90.12
## 14 76.04 63.79 88.28
## 15 77.35 65.25 89.44
## 16 77.04 64.96 89.12
```

Need for More Powerful Diagnostics for Multicollinearity

As we have seen, multicollinearity among the predictor variables can have important consequences for interpreting and using a fitted regression model.

The diagnostic tool considered here for identifying multicollinearity - namely, the pairwise coefficients of simple correlation between the predictor variables - is frequently helpful.

Often, however, serious multicollinearity exists without being disclosed by the pairwise correlation coefficients.

In Chapter 10, we present a more powerful tool for identifying the existence of serious multicollinearity. Some remedial measures for lessening the effects of multicollinearity will be considered in Chapter 11.

```
## Use all data from now on.
spending_subset=spending_subset_all[1:500,]
spending_subset %>% datatable()
```

Show 20 entries Search:

	province	type_of_dwelling	income	marital_status	age_gr
1	NL	single_detached	68000	never_married	30-34
2	NL	single_detached	48000	never_married	25-29
3	NL	single_detached	30000	married	35-39
4	NL	row_house	30000	never_married	30-34
5	NL	single_detached	35000	married	25-29
6	NL	single_detached	26000	married	25-29

Showing 1 to 20 of 500 entries

Previous 1 2 3 4 5 ... 25 Next

```
clothing_model = lm(clothing_expenditure~income+sex+food_expenditure+
msummary(clothing_model)
```

```
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  2.64e+02   2.32e+02    1.14  0.25474
## income                       1.46e-02   3.97e-03    3.69  0.00025
## sexmale                     -4.25e+02   1.35e+02   -3.15  0.00174
## food_expenditure             1.51e-01   2.45e-02    6.18  1.4e-09
## recreation_expenditure       2.29e-01   3.36e-02    6.81  2.8e-11
## miscellaneous_expenditure    1.60e-01   1.39e-01    1.16  0.24846
##
## Residual standard error: 1480 on 494 degrees of freedom
## Multiple R-squared:  0.237,    Adjusted R-squared:  0.229
## F-statistic: 30.7 on 5 and 494 DF,  p-value: <2e-16
```

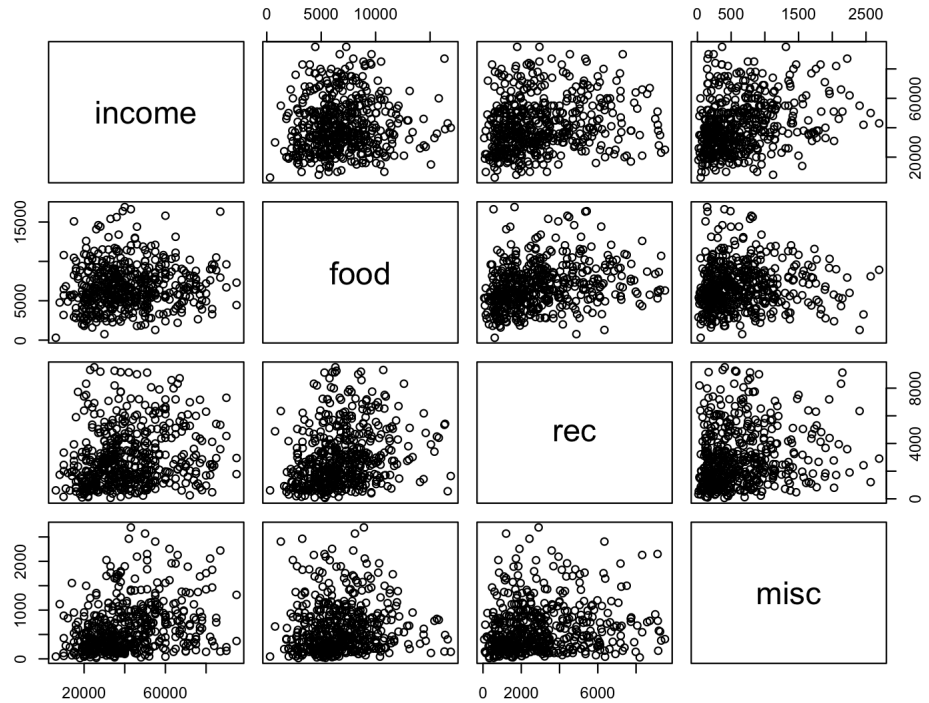
```
anova(clothing_model)
```

```
## Analysis of Variance Table
##
## Response: clothing_expenditure
##                             Df   Sum Sq  Mean Sq F value  Pr(>F)
## income                       1 7.13e+07 7.13e+07   32.50 2.1e-08
## sex                          1 1.25e+07 1.25e+07    5.71   0.017
## food_expenditure             1 1.44e+08 1.44e+08   65.66 4.2e-15
## recreation_expenditure       1 1.06e+08 1.06e+08   48.42 1.1e-11
## miscellaneous_expenditure    1 2.93e+06 2.93e+06    1.34   0.248
## Residuals                  494 1.08e+09 2.19e+06
```

We can examine the correlation among our continuous predictor variables by producing a scatterplot and correlation matrix:

```
cor.data <- with(spending_subset, data.frame(income, food=food_expen

plot(cor.data)
```



```
cor(cor.data)
```

```
##          income    food    rec    misc
## income 1.00000 0.08276 0.1960 0.29163
## food   0.08276 1.00000 0.2528 0.05981
## rec    0.19598 0.25284 1.0000 0.15591
## misc   0.29163 0.05981 0.1559 1.00000
```

- Comment on whether there appears to be multicollinearity in this model and what effect that would have in this setting if it were present.

- Multicollinearity does exist because the predictor variables are correlated among themselves. If they were uncorrelated, the model would give them the same effects regardless of the other predictor variables. We see that, in the above tables, the values for miscellaneous expenditure vary depending on its ordering.

- According to the correlation matrix, we see that there is correlation between the X variables, but the correlation in very low (< 0.3). So, there is very little multicollinearity in this model. Since the multicollinearity is low, we are still able to make inferences on clothing expenditure, and we can get more precise estimates for Bi than if the multicollinearity was high. Also, the slope of an Xi will change as more predictor variables are introduced into the model, and it will change according the the X variables already present in the model. As a result, the s{bi} will change depending on the variables in the model and the variables introduced into the model. Also, multicollinearity will affect the extra sums if squares, where SSR(Xi) will be different than SSR(Xi given other variables are in the model). Lastly, the fitted values will change with every X introduced into the model, to decrease the variability between Y and the fitted values.

- Comment on whether there appears to be multicollinearity in this model and what effect that would have in this setting if it were present.
- No I do not think there is multicollinearity present in this model. If it was present then there would be correlation between supposedly INDEPENDENT variables in the data set.
- Multicollinearity would cause regression coefficients and extra sum of squares to change drastically with how variables are placed in a sequence.
- there certainly appears to be multicollinearity, although it is not very strong between Y and SOME of the independent variables, it is definetly still present.
- There is no multicollinearity for we didn't allow any interaction between the factors.

Recap: Sections 7.4, 7.6

After Sections 7.4 and 7.6, you should be able to

- Compute and interpret coefficients of partial determination
- Understand multicollinearity and its effects