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+rec
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_expend
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+rec
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_expendi
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
```

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.

```r
clothing_model = lm(clothing_expenditure~income+sex+food_expenditure+rec
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
```

```r
clothing_model_reordered = lm(clothing_expenditure~miscellaneous_expend
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
```

```r
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
```

```r
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
```

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

layout: false

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 dertermination $R_{Y1|2}$ 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_group
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

```r
clothing_model = lm(clothing_expenditure~income+sex+food_expenditure+rec
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
```

```r
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
```

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

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):

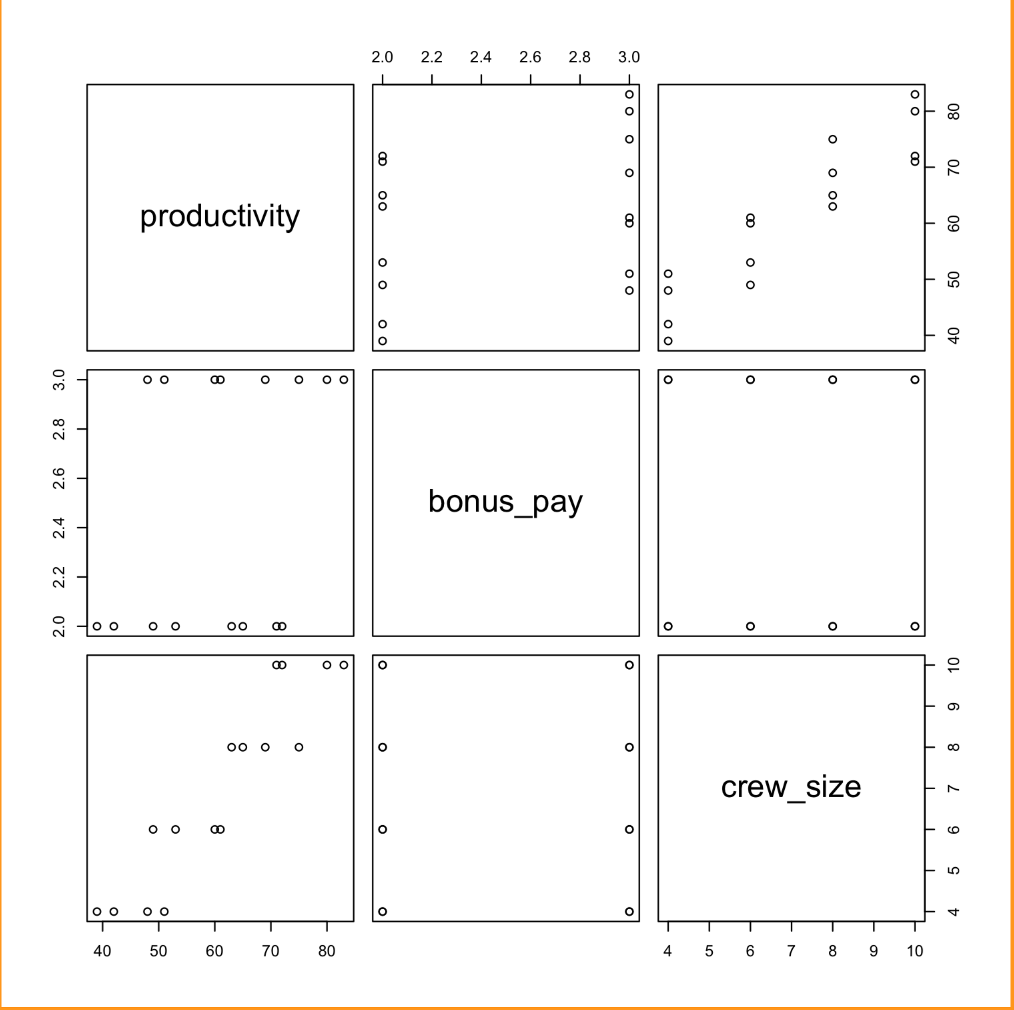
Show 20 entries Search:

	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	0.353	0.924
bonus_pay	0.353	1	0
crew_size	0.924	0	1

Showing 1 to 3 of 3 entries   Previous 1 Next

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

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.137      3.084    0.37     0.72
## crew_size       5.338      0.231   23.15  5.9e-12
## bonus_pay       9.125      1.031    8.85  7.3e-07
##
## Residual standard error: 2.06 on 13 degrees of freedom
## Multiple R-squared:  0.979,    Adjusted R-squared:  0.976
## F-statistic:  307 on 2 and 13 DF,  p-value: 1.14e-11
```

```r
anova(mod_full)
```

```
## Analysis of Variance Table
##
## Response: productivity
##           Df Sum Sq Mean Sq F value  Pr(>F)
## crew_size  1   2279    2279   536.1 5.9e-12
## bonus_pay  1    333     333    78.3 7.3e-07
## Residuals 13     55       4
```

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

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   23.950      4.327    5.54  7.3e-05
## crew_size      5.338      0.589    9.06  3.1e-07
##
## Residual standard error: 5.27 on 14 degrees of freedom
## Multiple R-squared:  0.854,    Adjusted R-squared:  0.844
## F-statistic: 82.2 on 1 and 14 DF,  p-value: 3.11e-07
```

```r
anova(mod_x1)
```

```
## Analysis of Variance Table
##
## Response: productivity
##           Df Sum Sq Mean Sq F value  Pr(>F)
## crew_size  1   2279    2279    82.2 3.1e-07
## Residuals 14    388      28
```

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

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    38.50      16.46    2.34    0.035
## bonus_pay       9.12       6.46    1.41    0.179
##
## Residual standard error: 12.9 on 14 degrees of freedom
## Multiple R-squared:  0.125,    Adjusted R-squared:  0.0624
## F-statistic:    2 on 1 and 14 DF,  p-value: 0.179
```

```r
anova(mod_x2)
```

```
## Analysis of Variance Table
##
## Response: productivity
##           Df Sum Sq Mean Sq F value Pr(>F)
## bonus_pay  1    333     333       2   0.18
## Residuals 14   2334     167
```

Example of the Problem with Perfect Multicollinearity

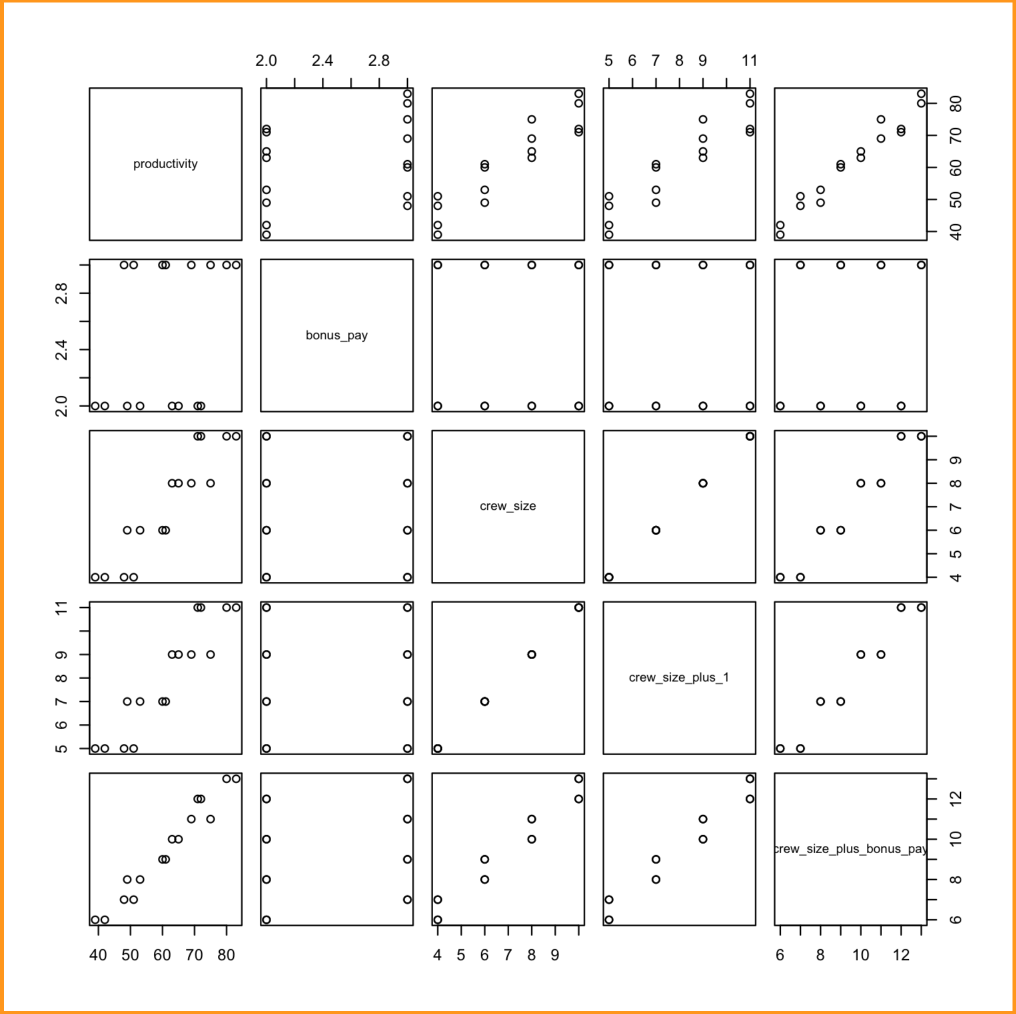
Show 20 entries Search:

	productivity	bonus_pay	crew_size	crew_size_plus_1	crew_size_
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)
```



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

Show 20 entries

Search:

	productivity	bonus_pay	crew_size	crew_siz
productivity	1	0.353	0.924	
bonus_pay	0.353	1	0	
crew_size	0.924	0	1	
crew_size_plus_1	0.924	0	1	
crew_size_plus_bonus_pay	0.979	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)$coefficien
```

```
##      (Intercept)        crew_size crew_size_plus_1
##           23.950            5.338               NA
```

```
lm(productivity~crew_size + bonus_pay + crew_size_plus_bonus_pay, data=c
```

```
##              (Intercept)                crew_size                bonus_pay
##                    1.137                    5.338                    9.125
## crew_size_plus_bonus_pay
##                       NA
```

Example of the Problem with Multicollinearity

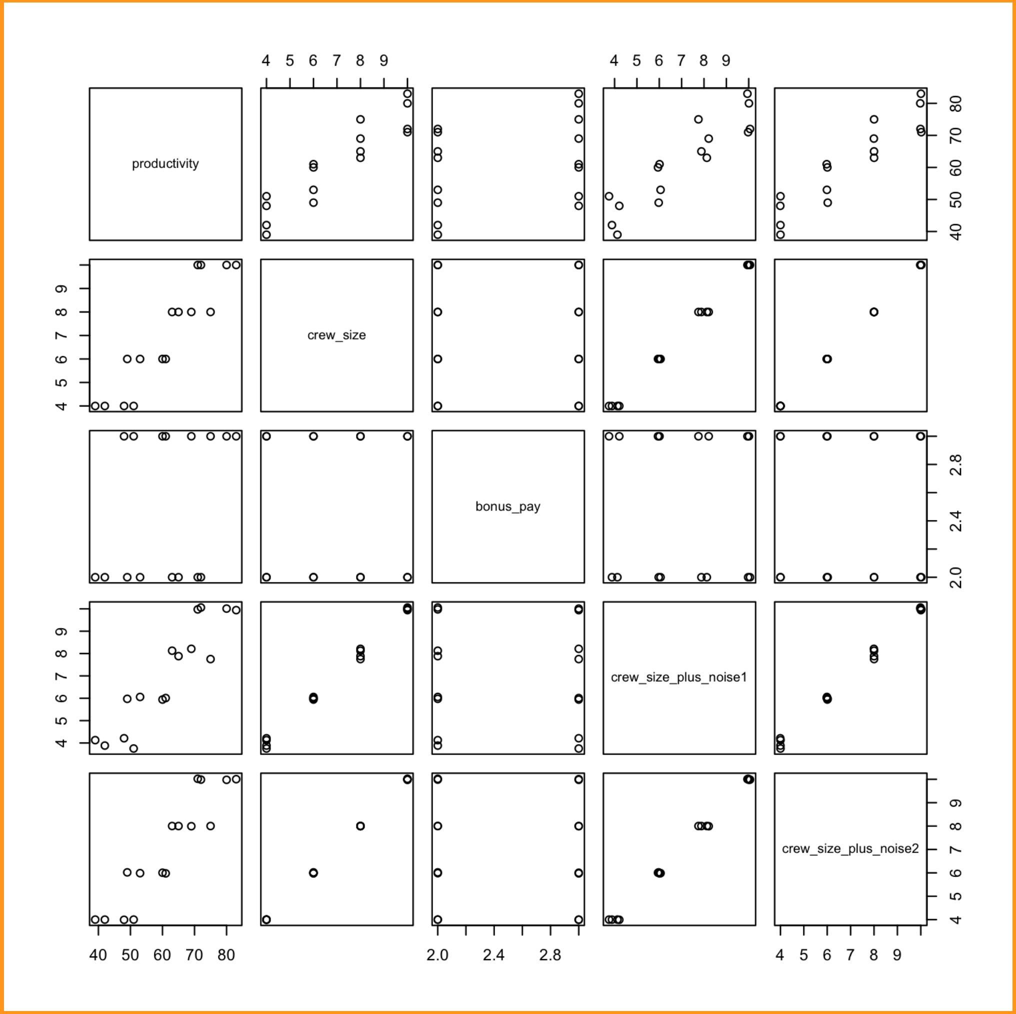
Show 20 entries Search:

	productivity	crew_size	bonus_pay	crew_size_plus_noise1	crew_
1	42	4	2	3.883084297181991	
2	39	4	2	4.126828678773032	
3	48	4	3	4.210909800590375	
4	51	4	3	3.752254820613921	
5	49	6	2	5.974150176281689	
6	53	6	2	6.058061814800459	

Showing 1 to 16 of 16 entries

Previous 1 Next

```
plot(crew_data)
```



```
cor(crew_data)       %>% round(7) %>% datatable(options=list(scrollY=350)
```

Show 20 entries Search:

	productivity	crew_size	bonus_pay	crew_size_plu
productivity	1	0.9243485	0.3533587	
crew_size	0.9243485	1	0	
bonus_pay	0.3533587	0	1	
crew_size_plus_noise1	0.9146658	0.9982196	-0.0070659	
crew_size_plus_noise2	0.9236645	0.9999857	-0.0017596	

Showing 1 to 5 of 5 entries

Previous 1 Next

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

```
## (Intercept)   crew_size
##      23.950       5.338
```

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

```
##            (Intercept)             crew_size crew_size_plus_noise1
##                  23.85                 18.36                -13.02
```

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

```
##            (Intercept)             crew_size crew_size_plus_noise2
##                  24.02                140.59               -135.29
```

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

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   23.950      4.327    5.54  7.3e-05
## crew_size      5.338      0.589    9.06  3.1e-07
##
## Residual standard error: 5.27 on 14 degrees of freedom
## Multiple R-squared:  0.854,    Adjusted R-squared:  0.844
## F-statistic: 82.2 on 1 and 14 DF,  p-value: 3.11e-07
```

```
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)                23.85       4.20    5.68  7.6e-05
## crew_size                  18.36       9.58    1.92    0.078
## crew_size_plus_noise1     -13.02       9.56   -1.36    0.197
##
## Residual standard error: 5.11 on 13 degrees of freedom
## Multiple R-squared:  0.873,    Adjusted R-squared:  0.853
## F-statistic: 44.5 on 2 and 13 DF,  p-value: 1.53e-06
```

```
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)                24.02       4.24    5.66  7.8e-05
## crew_size                 140.59     107.84    1.30     0.21
## crew_size_plus_noise2    -135.29     107.87   -1.25     0.23
##
## 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: 1.73e-06
```

Consider the effects of multicollinearity on Extra Sums of Squares

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

```
## Analysis of Variance Table
##
## Response: productivity
##                       Df Sum Sq Mean Sq F value Pr(>F)
## crew_size              1   2279    2279   87.17  4e-07
## crew_size_plus_noise1  1     48      48    1.85    0.2
## Residuals             13    340      26
```

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

```
## Analysis of Variance Table
##
## Response: productivity
##                       Df Sum Sq Mean Sq F value  Pr(>F)
## crew_size_plus_noise1  1   2232    2232   85.35 4.5e-07
## crew_size              1     96      96    3.67   0.078
## Residuals             13    340      26
```

Consider the effects of multicollinearity on Simultaneous Tests of β_k:

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

```
## Anova Table (Type II tests)
##
## Response: productivity
##                       Sum Sq Df F value Pr(>F)
## crew_size                 96  1    3.67  0.078
## crew_size_plus_noise1     48  1    1.85  0.197
## Residuals                340 13
```

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

```
## 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 2667
## 2     13  340  2      2328 44.5 1.5e-06
```

Consider the effects of multicollinearity on Fitted Values and Predictions:

```
predict(lm(productivity~crew_size
```

```
##       fit   lwr   upr
## 1  45.30 33.06 57.54
## 2  45.30 33.06 57.54
## 3  45.30 33.06 57.54
## 4  45.30 33.06 57.54
## 5  55.98 44.26 67.69
## 6  55.98 44.26 67.69
## 7  55.98 44.26 67.69
## 8  55.98 44.26 67.69
## 9  66.65 54.94 78.36
## 10 66.65 54.94 78.36
## 11 66.65 54.94 78.36
## 12 66.65 54.94 78.36
## 13 77.32 65.08 89.57
## 14 77.32 65.08 89.57
## 15 77.32 65.08 89.57
## 16 77.32 65.08 89.57
```

```
predict(lm(productivity~crew_size
```

```
##       fit   lwr   upr
## 1  46.74 34.55 58.93
## 2  43.57 31.28 55.85
## 3  42.47 29.69 55.26
## 4  48.44 35.47 61.42
## 5  56.24 44.78 67.70
## 6  55.15 43.62 66.68
## 7  55.77 44.31 67.23
## 8  56.66 45.15 68.16
## 9  68.11 56.42 79.79
## 10 64.93 53.16 76.71
## 11 63.84 51.55 76.13
## 12 69.81 57.31 82.31
## 13 77.60 65.62 89.59
## 14 76.51 64.47 88.56
## 15 77.13 65.15 89.11
## 16 78.02 66.00 90.05
```

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.

```r
## 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_group
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

```r
clothing_model = lm(clothing_expenditure~income+sex+food_expenditure+rec
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
```

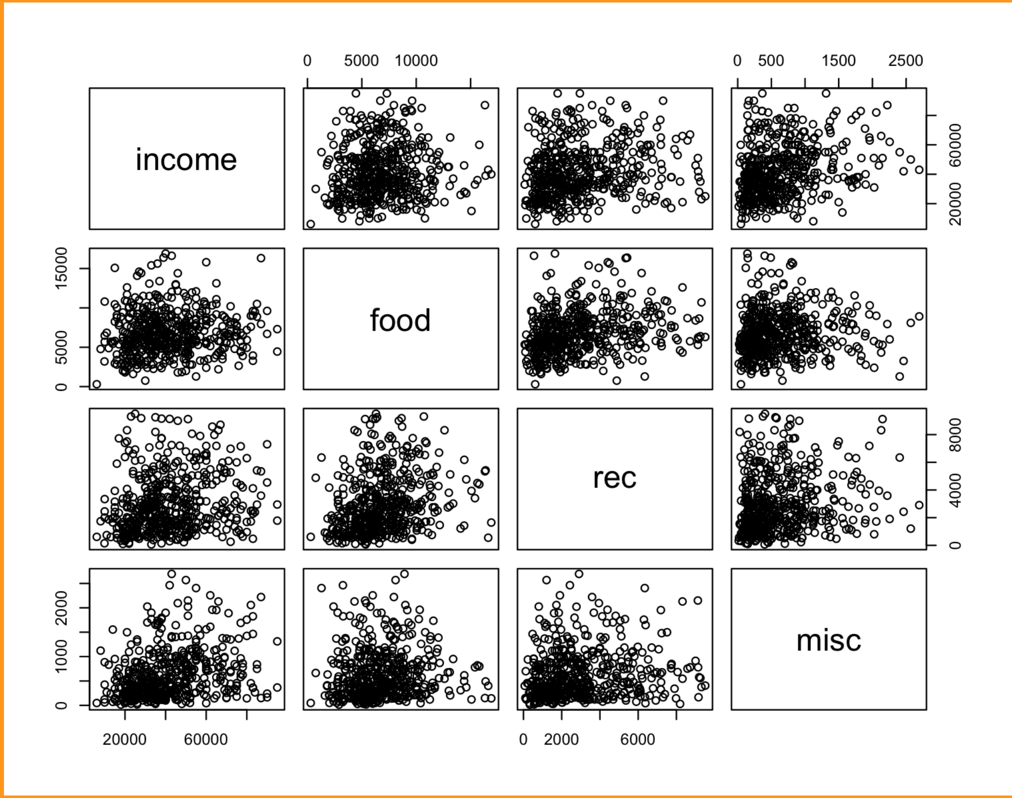
```r
anova(clothing_model)
```

```
## 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_expenditu

plot(cor.data)
```



```r
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
```

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