Some exercises in inference for MLR

Let's read a datafile containing nutrition information on a bunch of breakfast cereals. We wan to construct an MLR model to predict calories from the other variables. After that we will caryy out some inferences.

In [4]:

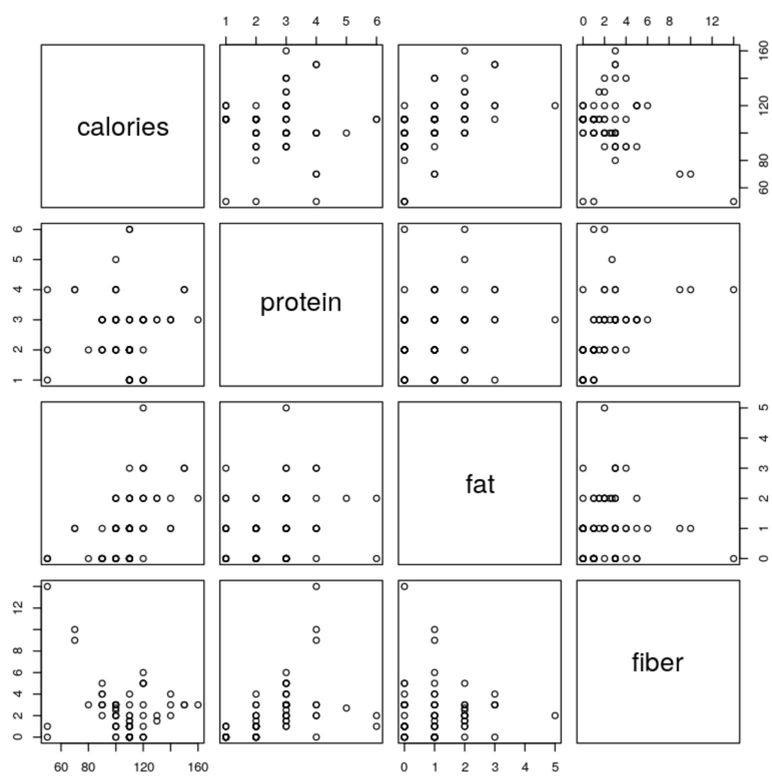
```
# Read data file
ex1dat = read.csv(file="https://cs.earlham.edu/~pardhan/sage_and_r/
breakfast_cereals.csv", header=TRUE, sep=",")
head(ex1dat)

# Make scatterplot and correlation matrix:
plot(ex1dat)
cor(ex1dat)
```

A data.frame: 6 × 4

	calories	protein	fat	fiber
	<int>	<int>	<int>	<dbl>
1	70	4	1	10.0
2	120	3	5	2.0
3	70	4	1	9.0
4	50	4	0	14.0
5	110	2	2	1.0
6	110	2	2	1.5

A matrix: 4 × 4 of type dbl

	calories	protein	fat	fiber
calories	1.00000000	0.01906607	0.49860981	-0.29341275
protein	0.01906607	1.00000000	0.20843099	0.50033004
fat	0.49860981	0.20843099	1.00000000	0.01671924
fiber	-0.29341275	0.50033004	0.01671924	1.00000000



In [6]:
```r
# Fit MLR model for calories vs all other variables
lmresults = lm(calories ~ protein+fat+fiber, data=ex1dat)
summary (lmresults)
```

```
Call:
lm(formula = calories ~ protein + fat + fiber, data = ex1dat)

Residuals:
    Min      1Q  Median      3Q     Max 
-50.955  -6.907  -0.350   8.017  45.511 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  99.3222     4.7886  20.742  < 2e-16 ***
protein       1.6328     2.0002   0.816  0.41698    
fat           9.3948     1.8841   4.986 4.02e-06 ***
fiber        -2.8402     0.8987  -3.160  0.00229 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.08 on 73 degrees of freedom
Multiple R-squared:  0.3457,	Adjusted R-squared:  0.3188 
F-statistic: 12.85 on 3 and 73 DF,  p-value: 7.847e-07
```

Inference questions

- Is the model as a whole a significant predictor of the response?
- Is fat a significant predictor of calories?
- Is protein a significant predictor of calories?

Hypothesis test for the model as a whole
Null hypothesis: None of the slopes is a significant predictor
Alt hypothesis: At least one slope is a significant predictor

In symbols:
$H_0 : \beta_p = \beta_{fat} = \beta_{fib} = 0$
H_A : At leat one of the $\beta' s \neq 0$

Look at the F-statistic with the indicated df, and the P-value.
Reject H_0 if the P-value is below significance level.

Conclusion Since the P-value is less than α, we reject H_0 and conclude that the model as a whole a significant predictor of calories.

Is fat a significant predictor of calories?
$H_0 : \beta_{fat} = 0$ when the model includes all other variables.
$H_A : \beta_{fat} \neq 0$ when the model includes all variables.

Find the relevant t-statistic and P-value:
$t = \frac{b_{fat} - 0}{SE_{fat}} = \frac{9.3948 - 0}{1.884} = 4.99$
With $df = n - k - 1 = 77 - 3 - 1 = 73$, the P-value=4.02×10^{-6}

Conclusion Since the P-value is less than α, we reject H_0 and conclude that fat is a significant predictor of calories.

Is protein a significant predictor of calories?
$H_0 : \beta_p = 0$ when the model includes all other variables.
$H_A : \beta_p \neq 0$ when the model includes all variables.

Find the relevant t-statistic and P-value:
From the regression output: $t = 0.816$ and the P-value=0.417

Conclusion Since the P-value is high we retain H_0 and conclude that protein is not a significant predictor of calories.

Confidence intervals

We can compute a confidence interval for the slope of each predictor.
Suppose we want a 90% confidence interval for the slope of fat.

CI = estimate $\pm\ t^*_{df} \cdot SE$

From the regression output: estimate = 9.3948, $SE = 1.8841$, df=73, $t^*_{73} = 1.666$ (for 90% confidence).

CI $= 9.3948 \pm 1.666 \times 1.8841 = (6.26, 12.53)$

Conclusion: We are 90% confident that, all else being equal, the model predicts that for each additional gram of fat, the increase in calories is between 6.26 and 12.53.

Exercise/Lab project for turn in

The file "movies.cvs" contains data on some movies, and the
variables that might be related to the amount of money those movies generated. We want to construct an MLR model to predict the amount of money the movie made from the 3 predictor variables *Budget*, *Stars*, and *Run_Time*. The units of *USGross* and

Budget are million dollars, and *Run_Time* is in minutes. Carry out each of the following tasks:

1. Make a matrix of scatterplots and correlations for these variables. Comment on what these plots and correlations suggest about the relationship between *USGross* and the 3 predictor variables.
2. Construct an MLR model, and write the model in the form of an equation.
3. Interpret each slope in context.
4. Interpret the adjusted R^2 in context.
5. Is the model as a whole a significant predictor of the response? Carry out a hypothesis test and state your conclusion.
6. Carry out a hypothesis test to determine whether the *Budget* is a significant predictor.
7. Compute and interpret a confidence interval for the slope of the *Budget* predictor.

In [12]:
```
movdat = read.csv(file="https://cs.earlham.edu/~pardhan/sage_and_r/movies.csv", header=TRUE, sep=",")
head(movdat)
```

A data.frame: 6 × 7

	Movie	USGross	Budget	Stars	Rating	Genre	Run_Time
	<fct>	<fct>	<fct>	<dbl>	<fct>	<fct>	<int>
1	White Noise	56.09436	30	2	PG-13	Horror	101
2	Coach Carter	67.264877	45	3	PG-13	Drama	136
3	Elektra	24.409722	65	2	PG-13	Action	100
4	Racing Stripes	49.772522	30	3	PG	Comedy	110
5	Assault on Precinct 13	20.040895	30	3	R	Action	109
6	Are We There Yet?	82.674398	20	2	PG	Comedy	94

In [ ]: