1. (a) H_0: Choice of rock, paper or scissors is independent of 1st or 5th grade age

H_a: The two variables have an association

(b) 1st and 5th graders at this school.

(c) The two way table has table of expected counts

	R	P	S	Total
1st	30	9	15	54
5th	24	17	23	64
Total	54	26	38	118

	R	P	S
1st	24.71	11.90	17.39
5th	29.29	14.1	20.61

$$\Rightarrow \chi^2 = \frac{(30-24.7)^2}{24.7} + \frac{(9-11.9)^2}{11.9} + \frac{(15-17.4)^2}{17.4} + \frac{(24-29.3)^2}{29.3} + \frac{(17-14.1)^2}{14.1} + \frac{(23-20.6)^2}{20.6} = 3.99$$

(d) Each expected count is ≥ 5, so it is appropriate to use a theoretical χ^2 distribution as our null distribution: the one with

$$df = (3-1)\cdot(2-1) = 2.$$

(e) 1 - pchisq(3.99, df = 2)

(f) Since $0.1358 > 0.1$, we fail to reject H_0. That is, there is insufficient to reject that these variables are independent.

2. (a) The variable with the highest (in magnitude) correlation coefficient when compared with Calories (the response variable) is Dietary Fat, with $r = 0.872$. So, a linear model with Dietary Fat as the lone explanatory variable would have the largest coefficient of determination R^2.

(b) There are a few aspects about the residual plots that draw our attention:

- one extra large positive residual
- one (probably the same) major outlier in the normal quantile plot

These noted, the F-score for the model is 336.1, with P-value 2.2×10^{-16}. We can reject H_0: the model is not useful in favor of H_a: it is useful.

(c) The model: $\widehat{\text{Calories}} = 512.9 + 16.26(\text{Dietary Fat}) + 0.42(\text{Cholesterol}) - 1.42(\text{Age})$.

So, at (55, 200, 37), $\widehat{\text{Calories}} = 512.9 + (16.26)(55) + (0.42)(200) - (1.42)(37) = 1438.66$ cal.

(d) The model in (c) explains about 76% of variability in response values, as reflected in the coefficient of determination, R^2.

(e) A good reason for trying a linear model with Cholesterol omitted (still keeping Age and DietaryFat as explanatory variables) is the high correlation, $r = 0.710$, between Cholesterol and DietaryFat. It seems changes in DietaryFat go a long way toward explaining both changes in Calories and changes in Cholesterol.

3. (a) It seems reasonable that individuals from the 3 samples should behave independently. The sample means should have approximately normal distributions, owing to the reasonably large sample sizes (37, 61, and 285). And the ratio

$$\frac{s_{max}}{s_{min}} = \frac{12.56}{10.38} < 2.$$

So, a theoretical F-distribution is reasonable to use.

(b) If μ_1, μ_2, μ_3 represent population mean SCI for the 3 groups

1: management, 2: skilled workers, 3: unskilled workers,

then $H_0: \mu_1 = \mu_2 = \mu_3$ (these means are all the same)

$H_a: \mu_i \neq \mu_j$ for at least one pairing.

(c)

DF	SS	MS	F
2	1411	705.5	5.867
380	45695	120.25	
382	47106		

(d) 1 - pf(5.867, 2, 380) should produce this P-value, which is statistically significant at the 5% level, since $0.0031 < 0.05$. We conclude there is at least one pair of means that is different.

(e) Option (ii) is best.

(f) We see evidence to conclude $\mu_2 \neq \mu_3$ (skilled vs. unskilled) only, as this pairing alone has P-value < 0.05 (and, correspondingly, 0 is not inside the family-rate 95% CI).