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	37	17	20	74
5th	21	24	19	64
Total	58	41	39	138

	R	P	S
1st	31.1	22.0	20.9
5th	26.9	19.0	18.1

$$\Rightarrow \chi^2 = \frac{(37-31.1)^2}{31.1} + \frac{(17-22)^2}{22} + \frac{(20-20.9)^2}{20.9} + \frac{(21-26.9)^2}{26.9} + \frac{(24-19)^2}{19} + \frac{(19-18.1)^2}{18.1} = 4.95$$

(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(4.95, df = 2)

(f) Since $0.0848 < 0.1$, we reject H_0 in favor of H_a, that there is an association between these variables.

2. (a) The variable with the highest (in magnitude) correlation coefficient when compared with DietaryChol (the response variable) is Fat, with $r = 0.7098$. So, a linear model with 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:

- a few extra large residuals on the positive side (right-skewness?)
- a bit of deviation from normality (normal quantile plot has some arc in it)

These noted, the F-score for the model is 108.2, 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: $\widehat{\text{DietaryChol}} = 8.41 + 2.2(\text{Fat}) + 0.033(\text{Calories}) + 0.108(\text{Age})$.

So, at (65, 2000, 47), $\widehat{\text{DietaryChol}} = 8.41 + (2.2)(65) + (0.033)(2000) + (0.108)(47) = 222.48$ mg.

(d) The model in (c) explains about 50–51% of variability in response values, as reflected in the coefficient of determination, R^2.

(e) A good reason for trying a linear model with Calories omitted (still keeping Fat and Age as explanatory variables) is the high correlation, $r = 0.872$, between Calories and Fat. It seems changes in Fat go a long way toward explaining both changes in Calories and changes in DietaryChol.

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 (51, 68, and 222). And the ratio

$$\frac{s_{max}}{s_{min}} = \frac{13.55}{8.75} < 2.$$

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	1166.64	538.32	4.621
338	42622.36	126.22	
340	43829		

(d) 1 - pf(4.621, 2, 338) should produce this P-value, which is statistically significant at the 5% level, since $0.0105 < 0.05$. We conclude there is at least one pair of means that is different

(e) Option (iv) 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).