14-32 (p.470)
Remember that "deaths" in this problem means deaths for white males aged 15-64.
(a)
Average number of deaths per state (inspection states): $\frac{5144}{15} \approx 343$
Average number of deaths per state (non-inspection states): $\frac{13380}{33} \approx 405$
As a percentage: $\frac{\frac{13380}{33}-\frac{5144}{15}}{\frac{13380}{33}} \approx 15.4\%$
So non-inspection states have a 15.4% higher number of deaths per state.
This isn't such a useful statistic because we haven't taken into account any of the various factors that are different between inspection and non-inspection states that might explain some of this difference (such as population, population density, differences in vehicle usage/mileage).

(b) Average number of deaths per million (inspection states): $\frac{5144}{16.7} \approx 308$
Average number of deaths per million (non-inspection states): $\frac{13380}{30.0} \approx 446$
As a percentage: $\frac{\frac{13380}{30.0}-\frac{5144}{16.7}}{\frac{13380}{30.0}} \approx 30.9\%$
This is an observational study. To make it a randomized controlled experiment, we'd have to randomly assign white males to inspection and non-inspection states. This isn't feasible, so we'd do better to try to discover other factors that might influence the number of deaths in inspection vs. non-inpsection states.

(c) I would say that very young and very old drivers are more likely to be in accidents or cause accidents of others, so if the inspection states and non-inspection states were different in age distributions, this could make a difference. Areas with higher population densities will have different traffic patterns and vehicle usage patterns (how many miles driven, how prevalent is the use of public transportation, etc.). I would initially guess that areas with higher population density may have more traffic and therefore more accidents, even if the average person drives less. (This is just a guess...)

(d) $H_0 : \beta_{insp} = 0 \qquad H_A : \beta_{insp} \neq 0$
$T^{45} = \frac{63.4-0}{30.6} \approx 2.07$
Use T-table with 40 degress of freedom (closest to 45). This gives us the critical value 2.02 (for α=0.05). This means our test statistic falls in the rejection region, and we reject the null hypothesis. The effect of inspection is significant (with α=0.05). Inspection makes a difference of 63.4 deaths per million white males (with all other factors being equal), according to our model in part (d).

(e) $H_0 : \beta_{insp} = 0 \qquad H_A : \beta_{insp} \neq 0$
$T = \frac{8.1-0}{39.8} \approx 0.20$
No matter how many degrees of freedom (and since they removed some of the data points, we're not sure exactly how many degrees of freedom there are), we cannot conclude that inspection makes a difference (fail to reject H_0). All else being equal, having inspection reduces deaths by 8.1 deaths per million white

males.

(f)

i.)Unlike my proposition in part (c), it seems from this regression that increases in population density correspond to decreases in mortality. Maybe people in higher population density areas drive shorter distances and/or take public transportation more (or any of a variety of other reasons one might think of).

ii.) All else being equal, introducing inspection reduces mortality reate by 8.1 deaths per million white males. If you double the population density (with all else) equal, you get the following equations:

Before doubling the population density x, you have:

$MR60 = 483 - 62.5 \log x + c$ (where c equals contributions from the other regressors)

Before doubling the population density x, you have:

$MR60 = 483 - 62.5 \log 2x + c = 483 - 62.5(\log 2 + \log x) + c$

So, doubling the population density lowers the mortality rate by $62.5 \log 2 \approx 43.3$ deaths per million white males.

(g)

part d	63.4 per million	takes population density (log) into account
part e	8.1 per million	takes into accout mortality rates from 1950

(h) Factors that we might like to control for (obviously, this list is **n**ot exclusive: how many cars there are, how many licensed drivers there are, total/average miles driven per capita, etc.

15-22 (p. 508)

(a) R^2 = SS explained by all regressors/total SS

(b) Multicollinearity occurs when the regressors are highly correlated with each other.

(c) True.

(d) True.

(e) True.

15-24 (p. 510)

(a)

i.) (0.34)(2) = 0.68 $\rightarrow$ 0.68 SDs more police officers than group with less income disparity

ii.) (0.34)(2) + (0.27)(-1) = 0.41 $\rightarrow$ 0.41 SDs more police officers

iii.) (0.63)(2) = 1.26 $\rightarrow$ 1.26 SDs more police officers. If the one is drawn at random, the other variable values will probably change, so we use the simple regression fit of P' on D'.

(b)

I would disagree with Sue. Once you take into account all the factors that we're given (disparity in income, mean family income annually, etc.), it looks like an increased number of riots is correlated with a small decrease in policemen.

(c)

$$\begin{aligned} X' &= \frac{X - \bar{X}}{s_X} \\ 0.15I' &= 0.15\frac{I - 10000}{1200} \\ 0.15I' &= \frac{1}{8000}I - 1.25 \end{aligned}$$

where -1.25 goes into the intercept term.

(d)Starting with the 2nd answer: large, number of small stores S, correlation coefficient

The 1st answer is dependent on how you read the question. I read it as asking why it can sometimes be helpful to standardize the variables. If you don't standardize the variables, the units that the explanatory variable is measured in will affect the regression coefficient. However, I think most people would read the question to ask "When you standardize the variable, is the regression coefficient closely related to the units in which the variable is measured"? The answer to this question is no, leading to the choice "independent of".

15-26 (p. 512)

(a) In the sense that this coefficient will change as the units change, this isn't necessarily true.

(b) False. Would be 5.4 minutes.

(c)

i.)

$$\begin{aligned} \bar{x} &\pm z_{.025}\frac{s}{\sqrt{n}} \\ 32 &\pm 1.96\frac{15}{\sqrt{200}} \\ (29.9 &\;,\; 34.1) \end{aligned}$$

ii.)

$$b \pm z_{.025}SE_b$$

$$
\begin{array}{ccc}
-0.24 & \pm & 1.96(0.05) \\
(-0.338 & , & -0.142)
\end{array}
$$

(d) False. Patients who are late are less likely to wait longer (according to this model) → coefficient of PALATE is negative

(e) R^2 would be larger, but $\bar{R}^2$ might not be (because $\bar{R}^2$ has a "penalty" for introduction for new varaibles → will only increase if new variable explains a sufficient amount of the variation)