1. Since $\sigma_1 = \sigma_2$ and the two samples are independent, we apply the pooled t test. The test statistic is

$$\begin{aligned} |T| &= \frac{|5.5 - 4.3|}{\sqrt{\frac{(20-1)(1.5^2) + (30-1)(1.6^2)}{20+30-2}\left(\frac{1}{20} + \frac{1}{30}\right)}} \\ &= 2.66267473. \end{aligned}$$

$t_{0.05/2,20+30-2} = t_{0.025,48} = 2.011 < 2.66267473$, so we can conclude that the mean driving times via Route 1 and Route 2 are different ($\mu_1 \neq \mu_2$) at level 0.05.

2. The p-value is $2(P(t(48) > 2.66267473)$. From the table "Quantiles for t distributions", we have

$$t_{0.01,48} < 2.66267473 < t_{0.005,48},$$

so

$$0.005 < P(t(48) > 2.66267473) < 0.01$$

and

$$\underbrace{2 \times 0.005}_{=0.01} < \underbrace{2P(t(48) > 2.66267473)}_{p\text{-value}} < \underbrace{2 \times 0.01}_{=0.02}.$$

Since the p-value is less than 0.02 and greater than 0.01, we can conclude that the mean driving times via Route 1 and Route 2 are different ($\mu_1 \neq \mu_2$) at level a when $a \geq 0.02$, but we cannot conclude $\mu_1 \neq \mu_2$ at level a when $a \leq 0.01$. The answers: (a) Yes; (b) Yes; (c) Yes; (d) No; (e) No.

3. Let

$$T_1 = \frac{\bar{X} - \bar{Y} - 1}{\sqrt{\hat{\sigma}^2(1/n_1 + 1/n_2)}},$$

where

$$\hat{\sigma} = \sqrt{\frac{(n_1 - 1)S_X^2 + (n_2 - 1)S_Y^2}{n_1 + n_2 - 2}}.$$

The proposed test rejects $H_0 : \mu_1 \leq \mu_2 + 1$ at level a if $T_1 > t_{a,n_1+n_2-2}$.

Below we will verify that the proposed test is of size a. Since the size of a test is the largest Type I error probability under H_0, to find the size of the test, we need to find the Type I error probability of the proposed test, which is

$$P(T_1 > t_{a,n_1+n_2-2})$$

when $\mu_1 \leq \mu_2 + 1$. Note that when $\mu_1 \leq \mu_2 + 1$,

$$\begin{aligned} &P(T_1 > t_{a,n_1+n_2-2}) \\ &= P\left(\frac{\bar{X} - \bar{Y} - 1}{\sqrt{\hat{\sigma}^2(1/n_1 + 1/n_2)}} > t_{a,n_1+n_2-2}\right) \\ &= P\left(\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\hat{\sigma}^2(1/n_1 + 1/n_2)}} + \frac{\mu_1 - \mu_2 - 1}{\sqrt{\hat{\sigma}^2(1/n_1 + 1/n_2)}} > t_{a,n_1+n_2-2}\right) \\ &\leq P\left(\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\hat{\sigma}^2(1/n_1 + 1/n_2)}} > t_{a,n_1+n_2-2}\right) \end{aligned} \tag{1}$$

since

$$\frac{\mu_1 - \mu_2 - 1}{\sqrt{\hat{\sigma}^2(1/n_1 + 1/n_2)}} \leq 0$$

when $\mu_1 \leq \mu_2 + 1$. Since

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\hat{\sigma}^2(1/n_1 + 1/n_2)}} \sim t(n_1 + n_2 - 2),$$

(1) implies that

$$P(T_1 > t_{a,n_1+n_2-2}) \leq P(t(n_1 + n_2 - 2) > t_{a,n_1+n_2-2}) = a \tag{2}$$

when $\mu_1 \leq \mu_2 + 1$. Moreover, when $\mu_1 = \mu_2 + 1$,

$$\begin{aligned} & P(T_1 > t_{a,n_1+n_2-2}) \\ &= P\left(\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\hat{\sigma}^2(1/n_1 + 1/n_2)}} > t_{a,n_1+n_2-2} \right) \\ &= P(t(n_1 + n_2 - 2) > t_{a,n_1+n_2-2}) = a. \end{aligned} \tag{3}$$

From (2) and (3), the largest Type I error probability of the proposed test under $H_0 : \mu_1 \leq \mu_2 + 1$ is a, so the proposed test is of size a.

4. The test rejects H_0 at level α if and only if

$$\begin{aligned} & \text{observed } T < -t_{\alpha,n_1+n_2-2} \\ \Leftrightarrow\ & -\text{observed } T > t_{\alpha,n_1+n_2-2} \\ \Leftrightarrow\ & P(t(n_1 + n_2 - 2) > -\text{observed } T) < \underbrace{P(t(n_1 + n_2 - 2) > t_{\alpha,n_1+n_2-2})}_{=\alpha} \\ \Leftrightarrow\ & P(t(n_1 + n_2 - 2) > -\text{observed } T) < \alpha, \end{aligned}$$

so the p-value of the test is

$$P(t(n_1 + n_2 - 2) > -\text{observed } T),$$

which is the same as

$$P(t(n_1 + n_2 - 2) < \text{ observed } T).$$

5. By Fact 2 given in Problem 5, we have

$$\frac{(n_1 - 1)S_X^2}{\sigma^2} = \sum_{i=1}^{n_1-1} W_i^2$$

and

$$\frac{(n_2 - 1)S_Y^2}{\sigma^2} = \sum_{i=1}^{n_2-1} V_i^2,$$

where $W_1, \ldots, W_{n_1-1}$ are IID $N(0, 1)$ random variables, and $V_1, \ldots, V_{n_2-1}$ are IID $N(0, 1)$ random variables. Thus by Fact 1 given in Problem 5, we have

$$\frac{(n_1 - 1)S_X^2}{\sigma^2} \sim \chi^2(n_1 - 1)$$

and

$$\frac{(n_2 - 1)S_Y^2}{\sigma^2} \sim \chi^2(n_2 - 1).$$

Since $(n_1 - 1)S_X^2/\sigma^2$ and $(n_2 - 1)S_Y^2/\sigma^2$ are independent, the distribution of

$$\frac{(n_1 - 1)S_X^2 + (n_2 - 1)S_Y^2}{\sigma^2}$$

is the same as the distribution of the sum of two independent χ^2 random variables of degrees of freedom $(n_1 - 1)$ and $(n_2 - 1)$ respectively. Let $\mathcal{D}$ denote this distribution and we have

$$\frac{(n_1 - 1)S_X^2 + (n_2 - 1)S_Y^2}{\sigma^2} \sim \mathcal{D}. \tag{4}$$

Let $Z_1, \ldots, Z_{n_1-1}$, $U_1, \ldots, U_{n_2-1}$ be IID $N(0, 1)$ random variables, then $\sum_{i=1}^{n_1-1} Z_i^2$ and $\sum_{i=1}^{n_2-1} U_i^2$ are independent, $\sum_{i=1}^{n_1-1} Z_i^2 \sim \chi^2(n_1 - 1)$, and $\sum_{i=1}^{n_2-1} U_i^2 \sim \chi^2(n_2 - 1)$, so

$$\left(\sum_{i=1}^{n_1-1} Z_i^2\right) + \left(\sum_{i=1}^{n_2-1} U_i^2\right) \sim \mathcal{D}. \tag{5}$$

From (4) and (5), the distribution of

$$\frac{(n_1 - 1)S_X^2 + (n_2 - 1)S_Y^2}{\sigma^2}$$

and the distribution of

$$\left(\sum_{i=1}^{n_1-1} Z_i^2\right) + \left(\sum_{i=1}^{n_2-1} U_i^2\right)$$

are both $\mathcal{D}$. By Fact 1,

$$\left(\sum_{i=1}^{n_1-1} Z_i^2\right) + \left(\sum_{i=1}^{n_2-1} U_i^2\right) \sim \chi^2(n_1 + n_2 - 2),$$

so the distribution $\mathcal{D}$ is $\chi^2(n_1 + n_2 - 2)$, and (4) gives

$$\frac{(n_1 - 1)S_X^2 + (n_2 - 1)S_Y^2}{\sigma^2} \sim \chi^2(n_1 + n_2 - 2).$$

6. Since the two samples of weights can be dependent, we apply the paired t test. The sample of weight increment amounts is

$$(77 - 80, 54 - 55, 53 - 63, 48 - 48, 51 - 50) = (-3, -1, -10, 0, 1).$$

The sample mean and sample standard deviation are

$$\frac{-3 - 1 - 10 + 0 + 1}{5} = -2.6$$

and

$$\sqrt{\frac{(-3)^2 + (-1)^2 + (-10)^2 + (0)^2 + (1)^2 - 5 \times (2.6)^2}{5 - 1}} = \sqrt{19.3} = 4.393177$$

respectively. The test statistic is

$$|T| = \frac{\sqrt{5} \times |-2.6|}{\sqrt{19.3}} = 1.323365.$$

Since $t_{0.05/2,5-1} = t_{0.025,4} = 2.776 > 1.323365$, we cannot conclude that participating in the program has a significant effect on weight change at the 0.05 level.

7. The testing problem is

$$H_0 : \sigma_1 = \sigma_2 \text{ v.s. } H_1 : \sigma_1 \neq \sigma_2.$$

Let S_i be the sample standard deviation of the driving times for Rout i for $i = 1$, 2 and let $F = S_2^2/S_1^2$. We will use the test that rejects H_0 at level 0.1 if $F > f_{0.1/2,31-1,21-1} = f_{0.05,30,20}$ or $1/F > f_{0.1/2,21-1,31-1} = f_{0.05,20,30}$. From the table "0.95 quantiles for F distributions", $f_{0.05,20,30} = 1.93$ and $f_{0.05,30,20} = 2.04$, so

$$\text{observed } F = \frac{1.6^2}{1.5^2} = 1.137778 < 2.04 = f_{0.05,30,20}$$

and

$$\text{observed } \frac{1}{F} = \frac{1.5^2}{1.6^2} < 1 < 1.93 = f_{0.05,20,30}.$$

Therefore, we cannot conclude $\sigma_1 \neq \sigma_2$ at level 0.1 based on the test.

8. The F test rejects H_0 at level a when

$$\frac{S_Y^2}{S_X^2} > f_{a/2,n_2-1,n_1-1} \text{ or } \frac{S_X^2}{S_Y^2} > f_{a/2,n_1-1,n_2-1}. \tag{6}$$

For a constant C, and positive integers m and n, note that

$$\begin{aligned} C > f_{a/2,m,n} \quad &\Leftrightarrow \quad P(F(m,n) > C) < P(F(m,n) > f_{a/2,m,n}) = a/2 \\ &\Leftrightarrow \quad a > 2P(F(m,n) > C), \end{aligned}$$

so (6) holds if and only if

$$a > 2P\left(F(n_2-1,n_1-1) > \text{ observed } \frac{S_Y^2}{S_X^2}\right) \text{ or } a > 2P\left(F(n_1-1,n_2-1) > \text{ observed } \frac{S_X^2}{S_Y^2}\right),$$

which is equivalent to

$$a > \min\left(2P\left(F(n_2-1,n_1-1) > \text{ observed } \frac{S_Y^2}{S_X^2}\right), 2P\left(F(n_1-1,n_2-1) > \text{ observed } \frac{S_X^2}{S_Y^2}\right)\right).$$

Therefore, the p-value for the F test is

$$\min\left(2P\left(F(n_2-1,n_1-1) > \text{ observed } \frac{S_Y^2}{S_X^2}\right), 2P\left(F(n_1-1,n_2-1) > \text{ observed } \frac{S_X^2}{S_Y^2}\right)\right),$$

9. (a) Consider two random variables X and Y that are independent, where $X \sim \chi^2(m)$ and $Y \sim \chi^2(n)$. Let $F = (X/m)/(Y/n)$, then $F \sim F(m,n)$ and $1/F \sim F(n,m)$. We will show that $f_{1-a,n,m} = 1/f_{a,m,n}$ by verifying

$$P(F(n,m) > 1/f_{a,m,n}) = 1 - a. \tag{7}$$

Note that

$$\begin{aligned} P(F(n,m) > 1/f_{a,m,n}) &= P(1/F > 1/f_{a,m,n}) \\ &= P(F < f_{a,m,n}) = P(F(m,n) < f_{a,m,n}) \\ &= 1 - P(F(m,n) \geq f_{a,m,n}) \\ &= 1 - P(F(m,n) > f_{a,m,n}) = 1 - a, \end{aligned}$$

so (7) holds and we have shown that $f_{1-a,n,m} = 1/f_{a,m,n}$, which implies $f_{a,m,n} = 1/f_{1-a,n,m}$.

(b) Since W is a positive random variable,

$$1/W > f_{\alpha/2,n,m} \Leftrightarrow W < 1/f_{\alpha/2,n,m},$$

where $1/f_{\alpha/2,n,m} = f_{1-\alpha/2,m,n}$ by the result in Part (a). Thus

$$1/W > f_{\alpha/2,n,m} \Leftrightarrow W < f_{1-\alpha/2,m,n}. \tag{8}$$

Since for $\alpha \in (0,1)$, $\alpha/2 \in (0, 0.5)$ and $\alpha/2 < 1 - \alpha/2$, which implies that

$$f_{\alpha/2,m,n} > f_{1-\alpha/2,m,n}.$$

Therefore, if $W > f_{\alpha/2,m,n}$, then we cannot have $W < f_{1-\alpha/2,m,n}$, so

$$\{W > f_{\alpha/2,m,n}\} \cap \{W < f_{1-\alpha/2,m,n}\} = \emptyset,$$

which implies that

$$\{W > f_{\alpha/2,m,n}\} \cap \{1/W > f_{\alpha/2,n,m}\} = \emptyset$$

by (8).
Remark. If the condition $W > 0$ is replaced by the condition $P(W > 0) = 1$, we can say that

$$\{W > 0\} \cap W > f_{\alpha/2,m,n}\} \cap \{1/W > f_{\alpha/2,n,m}\} = \emptyset,$$

which implies that

$$P(\{W > f_{\alpha/2,m,n}\} \cap \{1/W > f_{\alpha/2,n,m}\}) = 0$$

since $P(W > 0) = 1$.

10. Let $C = -t_{\alpha,n-1}$. Under H_0, we have $\mu \geq \mu_0$, so the Type I error probability

$$\begin{aligned}
&P\left(\frac{\sqrt{n}(\bar{X} - \mu_0)}{S} < C\right) \\
&= P\left(\frac{\sqrt{n}(\bar{X} - \mu)}{S} < C - \underbrace{\frac{\sqrt{n}(\mu - \mu_0)}{S}}_{\geq 0}\right) \\
&\leq P\left(\frac{\sqrt{n}(\bar{X} - \mu)}{S} < C\right) \\
&= P\left(t(n-1) < C\right).
\end{aligned} \tag{9}$$

In addition, when $\mu = \mu_0$, the Type I error probability

$$P\left(\frac{\sqrt{n}(\bar{X} - \mu_0)}{S} < C\right) = P\left(t(n-1) < C\right). \tag{10}$$

From (9) and (10), the largest Type I error probability under $H_0 : \mu \geq \mu_0$ is

$$\max_{\mu \geq \mu_0} P\left(\frac{\sqrt{n}(\bar{X} - \mu_0)}{S} < C\right) = P\left(t(n-1) < C\right),$$

so the size of the test is

$$P(t(n-1) < C) = P(t(n-1) < -t_{\alpha,n-1}) = P(t(n-1) > t_{\alpha,n-1}) = \alpha.$$

11. (a) The p-value for the F test for testing

$$H_0 : \sigma_1 = \sigma_2 \text{ v.s. } H_1 : \sigma_1 \neq \sigma_2$$

is

$$\min(1.655241, 0.3447595) = 0.3447595.$$

(b) The p-value from Part (a) is greater than 0.05, so we do not have strong evidence for $\sigma_1 \neq \sigma_2$. We will assume $\sigma_1 = \sigma_2$, as required in the problem. When $\sigma_1 = \sigma_2$, the pooled t test can be applied for finding evidence for $\mu_1 \neq \mu_2$. From the R output given in the problem, the p-value for the pooled t test is

$$0.001820591 < 0.002,$$

so we can conclude $\mu_1 \neq \mu_2$ at level 0.002.

12. (a) The observed test statistic for the apporximate z test is

$$\text{observed } Z = \frac{10/100 - 20/1000}{\sqrt{(1/100 + 1/1000)(30/1100)(1 - 30/1100)}} = 4.683104,$$

so the p-value is $P(N(0,1) > 4.683104)$. Note that

$$P(N(0,1) > 4.683104) < P(N(0,1) > 3.09)$$

and

$$\begin{aligned} P(N(0,1) > 3.09) &= 0.5 - 0.4990 \\ &= 0.001, \end{aligned} \tag{11}$$

so the p-value is less than $0.001 < 0.05$ and there is strong evidence for $p_A > P_B$.

(b) The p-value is

$$2P(N(0,1) > |4.683104|) < 2P(N(0,1) > 3.09) \overset{(11)}{=} 2 \cdot 0.001 < 0.05,$$

so there is strong evidence for $p_A \neq P_B$.

13. Let p_{pooled} and p_{paired} be the probabilites of rejecting $\mu_1 = \mu_2$ at level 0.05 when $(\mu_1, \mu_2, \sigma_1, \sigma_2) = (1, 1.2, 1, 1)$ based on the pooled t test and the paired t test, respectively. We are given data for the testing problem

$$H_0 : p_{\text{pooled}} \leq p_{\text{paired}} \text{ v.s. } H_1 : p_{\text{pooled}} > p_{\text{paired}}.$$

The observed Z statistic for the testing the difference of population proportions is

$$\frac{16942/10^5 - 16684/10^5}{\sqrt{(2/10^5) \cdot \hat{p}(1 - \hat{p})}},$$

where

$$\hat{p} = \frac{16942 + 16684}{2 \cdot 10^5}.$$

The observed Z statistics is 1.542604. From the table "Quantiles for t distributions", $z_{0.1} = 1.282 < 1.542604$, so we can conclude $p_{\text{pooled}} > p_{\text{paired}}$ at level 0.1. That is, at level 0.1, we can conclude that the two-sample pooled t test has larger power than the paired t test when $(\mu_1, \mu_2, \sigma_1, \sigma_2) = (1, 1.2, 1, 1)$.

14. Let p_{pooled} and p_{paired} be the probabilites of rejecting $\mu_1 = \mu_2$ at level 0.05 when $(\mu_1, \mu_2, \sigma_1, \sigma_2) = (0, 0, 1, 1)$ based on the pooled t test and the paired t test, respectively. We are given data for the testing problem

$$H_0 : p_{\text{pooled}} \leq p_{\text{paired}} \text{ v.s. } H_1 : p_{\text{pooled}} > p_{\text{paired}}.$$

The observed Z statistic for the testing the difference of population proportions is

$$\frac{5078/10^5 - 5039/10^5}{\sqrt{(2/10^5) \cdot \hat{p}(1-\hat{p})}},$$

where

$$\hat{p} = \frac{5078 + 5039}{2 \cdot 10^5}.$$

The observed Z statistics is $0.3979337 < z_{0.1} = 1.282$, so we cannot conclude $p_{\text{pooled}} > p_{\text{paired}}$ at level 0.1. That is, at level 0.1, we cannot conclude that the two-sample pooled t test has higher Type I error probability than the paired t test when $(\mu_1, \mu_2, \sigma_1, \sigma_2) = (0, 0, 1, 1)$.

15. (a) To prove the result in Part (a), we can apply the fact that for m values W_1, ..., W_m,

$$\sum_{j=1}^{m} W_j^2 = m(\bar{W})^2 + \sum_{j=1}^{m} (W_j - \bar{W})^2, \tag{12}$$

where $\bar{W} = \sum_{j=1}^{m} W_j/m$. For $i \in \{1, \ldots, k\}$, apply (12) with $m = n_i$ and

$$(W_1, \ldots, W_m) = (X_{i,1} - \bar{X}_G, \ldots, X_{i,n_i} - \bar{X}_G),$$

then we have

$$\bar{W} = \frac{1}{n_i} \sum_{j=1}^{n_i} (X_{i,j} - \bar{X}_G) = \bar{X}_i - \bar{X}_G,$$

and for $j \in \{1, \ldots, n_i\}$,

$$W_j - \bar{W} = X_{i,j} - \bar{X}_G - (\bar{X}_i - \bar{X}_G) = X_{i,j} - \bar{X}_i,$$

so (12) becomes

$$\sum_{j=1}^{n_i} (X_{i,j} - \bar{X}_G)^2 = n_i(\bar{X}_i - \bar{X}_G)^2 + \sum_{j=1}^{n_i} (X_{i,j} - \bar{X}_i)^2.$$

The proof of (12) is given below and can be omitted.
Proof of (12). Note that

$$\begin{aligned}
\sum_{j=1}^{m} (W_j - \bar{W})^2 &= \sum_{j=1}^{m} (W_j^2 + (\bar{W})^2 - 2\bar{W}W_j) \\
&= \left(\sum_{j=1}^{m} W_j^2 \right) + \left(\sum_{j=1}^{m} (\bar{W})^2 \right) + \left(\sum_{j=1}^{m} (-2\bar{W}W_j) \right) \\
&= \left(\sum_{j=1}^{m} W_j^2 \right) + m(\bar{W})^2 + \left(-2\bar{W} \underbrace{\sum_{j=1}^{m} W_j}_{=m\bar{W}} \right) \\
&= \left(\sum_{j=1}^{m} W_j^2 \right) - m(\bar{W})^2,
\end{aligned}$$

so

$$\sum_{j=1}^{m} W_j^2 = m(\bar{W})^2 + \sum_{j=1}^{m}(W_j - \bar{W})^2.$$

The proof of (12) is complete.

(b) From Part (a), we have

$$\sum_{j=1}^{n_i}(X_{i,j} - \bar{X}_G)^2 = n_i(\bar{X}_i - \bar{X}_G)^2 + \sum_{j=1}^{n_i}(X_{i,j} - \bar{X}_i)^2. \tag{13}$$

for $i \in \{1, \ldots, k\}$. Take the sum of each side of (13) over $i \in \{1, \ldots, k\}$, then we have

$$\underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{i,j} - \bar{X}_G)^2}_{\mathrm{SS}_{\mathrm{total}}} = \underbrace{\sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X}_G)^2}_{\mathrm{SS}_{\mathrm{treat}}} + \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{i,j} - \bar{X}_i)^2}_{\mathrm{SSE}}.$$

16. Let $\bar{X}$ and S_X be the sample mean and sample standard deviation of $(X_1, \ldots, X_{n_1})$ respectively, and let $\bar{Y}$ and S_Y be the sample mean and sample standard deviation of $(Y_1, \ldots, Y_{n_2})$ respectively. Then the two sample t test (pooled t test) is based on the test statistic T, where

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{(1/n_1 + 1/n_2)((n_1 - 1)S_X^2 + (n_2 - 1)S_Y^2)/(n_1 + n_2 - 2)}}, \tag{14}$$

as given in Problem 4.

For the ANOVA test, note that the grand mean is $(n_1\bar{X} + n_2\bar{Y})/(n_1 + n_2)$, so

$$\begin{aligned}
\mathrm{SS}_{\mathrm{treat}} &= n_1\left(\bar{X} - \frac{n_1\bar{X} + n_2\bar{Y}}{n_1 + n_2}\right)^2 + n_2\left(\bar{Y} - \frac{n_1\bar{X} + n_2\bar{Y}}{n_1 + n_2}\right)^2 \\
&= n_1\left(\frac{n_2(\bar{X} - \bar{Y})}{n_1 + n_2}\right)^2 + n_2\left(\frac{n_1(\bar{Y} - \bar{X})}{n_1 + n_2}\right)^2 \\
&= n_1 n_2(\bar{X} - \bar{Y})^2/(n_1 + n_2) = \frac{(\bar{X} - \bar{Y})^2}{1/n_1 + 1/n_2}
\end{aligned}$$

with degree of freedom $2 - 1 = 1$. Also,

$$\mathrm{SSE} = (n_1 - 1)S_X^2 + (n_2 - 1)S_Y^2$$

with degrees of freedom $n_1 + n_2 - 2$. Thus the F statistic is

$$\begin{aligned}
F &= \frac{\mathrm{SS}_{\mathrm{treat}}/1}{\mathrm{SSE}/(n_1 + n_2 - 2)} \\
&= \frac{(\bar{X} - \bar{Y})^2/(1/n_1 + 1/n_2)}{((n_1 - 1)S_X^2 + (n_2 - 1)S_Y^2)/(n_1 + n_2 - 2)} \\
&= \frac{(\bar{X} - \bar{Y})^2}{(1/n_1 + 1/n_2)((n_1 - 1)S_X^2 + (n_2 - 1)S_Y^2)/(n_1 + n_2 - 2)}.
\end{aligned} \tag{15}$$

From (14) and (15), we have $T^2 = F$.

17. The grand mean is

$$\frac{5(350/5) + 10(695/10) + 4(288/4)}{5 + 10 + 4} = 1333/19,$$

so

$$\begin{aligned} \mathrm{SS}_{\mathrm{treat}} &= 5(350/5 - 1333/19)^2 + 10(695/10 - 1333/19)^2 + 4(288/4 - 1333/19)^2 \\ &= 6507.5/361 \end{aligned}$$

with degrees of freedom $3 - 1 = 2$.

$$\mathrm{SSE} = (5-1)(\sqrt{6/4})^2 + (10-1)(\sqrt{8.5/9})^2 + (4-1)(\sqrt{6/3})^2 = 20.5$$

with degrees of freedom $5 + 10 + 4 - 3 = 16$. The observed F test statistic is

$$\text{observed } \frac{\mathrm{SS}_{\mathrm{treat}}/2}{\mathrm{SSE}/16} = \frac{(6507.5/361)/2}{20.5/16} = \frac{104120}{14801} = 7.03466.$$

Since $f_{0.05,2,16} = 3.63 < 7.03466$, we can conclude that the means of the score distributions for the three classes are not all the same at the 0.05 significant level.

18. Let $Z_i = (X_i - \mu)/\sigma$ for $i = 1, \ldots, n$, then $Z_1, \ldots, Z_n$ are IID and for $i = 1, \ldots, n$, $Z_i \sim N(0, 1)$ and $X_i = \mu + \sigma Z_i$. Let $\bar{Z} = \sum_{i=1}^n Z_i/n$, then

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n}(\mu + \sigma Z_i) = \mu + \sigma\bar{Z}$$

and

$$X_i - \bar{X} = \mu + \sigma Z_i - (\mu + \sigma\bar{Z}) = \sigma \cdot (Z_i - \bar{Z}), \tag{16}$$

so

$$\begin{aligned} \frac{(n-1)S^2}{\sigma^2} &= \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} \\ &\overset{(16)}{=} \sum_{i=1}^{n}(Z_i - \bar{Z})^2 \\ &\overset{(12);W_i=Z_i;m=n}{=} \sum_{i=1}^{n} Z_i^2 - n(\bar{Z})^2 \\ &= \sum_{i=1}^{n} Z_i^2 - (\sqrt{n}\bar{Z})^2. \end{aligned}$$

Note that $\sqrt{n}\bar{Z} = \sum_{i=1}^n Z_i/\sqrt{n}$ is a linear combination of $Z_1, \ldots, Z_n$, where $Z_1, \ldots, Z_n$ are IID $N(0, 1)$ random variables, and

$$Var(\sqrt{n}\bar{Z}) = (\sqrt{n})^2 Var(\bar{Z}) = n \cdot \frac{Var(Z_1)}{n} = 1,$$

so by Fact 2 in the handout "Difference between special sums of squares of IID N(0, 1) random variables",

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} Z_i^2 - (\sqrt{n}\bar{Z})^2 \sim \chi^2(n-1).$$

19. (a) Yes. The p-value is $4.156 \times 10^{-5} < 0.001$.

(b) The degrees of freedom for SS_{treat} is 2, so the total number of classes is $2 + 1 = 3$.

(c) The degrees of freedom for SSE is 96 and the degrees of freedom for SS_{treat} is 2, so the total number of students is $96 + 2 + 1 = 99$.

20. Let $\bar{\mu}$ denote the average of all $\mu_{i,j}$'s, then

$$\mu = \bar{\mu} = \frac{5 + 7 + 3 + 4 + 5 + 3}{6} = 4.5.$$

The α_i's are computed using

$$\alpha_i = \frac{\mu_{i,1} + \mu_{i,2} + \mu_{i,3}}{3} - \mu,$$

so

$$\alpha_1 = \frac{5 + 7 + 3}{3} - \mu = 5 - 4.5 = 0.5,$$

and

$$\alpha_2 = \frac{4 + 5 + 3}{3} - \mu = 4 - 4.5 = -0.5.$$

The β_j's are computed using

$$\beta_j = \frac{\mu_{1,j} + \mu_{2,j}}{2} - \mu,$$

so

$$\beta_1 = \frac{5 + 4}{2} - \mu = 4.5 - 4.5 = 0,$$

$$\beta_2 = \frac{7 + 5}{2} - \mu = 6 - 4.5 = 1.5,$$

and

$$\beta_3 = \frac{3 + 3}{2} - \mu = 3 - 4.5 = -1.5.$$

The $\gamma_{i,j}$'s are computed using

$$\gamma_{i,j} = \mu_{i,j} - \mu - \alpha_i - \beta_j,$$

so

$$\gamma_{1,1} = 5 - (4.5 + 0.5 + 0) = 0,$$

$$\gamma_{1,2} = 7 - (4.5 + 0.5 + 1.5) = 0.5,$$

$$\gamma_{1,3} = 3 - (4.5 + 0.5 - 1.5) = -0.5,$$

$$\gamma_{2,1} = 4 - (4.5 - 0.5 + 0) = 0,$$

$$\gamma_{2,2} = 5 - (4.5 - 0.5 + 1.5) = -0.5,$$

and

$$\gamma_{2,3} = 3 - (4.5 - 0.5 - 1.5) = 0.5.$$

21. Let $\bar{a} = (a_1 + a_2)/2$ and $\bar{b} = \sum_{j=1}^{3} b_j/3$. Then

$$\begin{aligned} \mu &= \frac{\sum_{i=1}^{2} \sum_{j=1}^{3} \mu_{i,j}}{6} \\ &= \frac{\sum_{i=1}^{2} \sum_{j=1}^{3} (a_i + b_j)}{6} \\ &= \frac{3 \sum_{i=1}^{2} a_i + 2 \sum_{j=1}^{3} b_j}{6} = \bar{a} + \bar{b}, \end{aligned}$$

$$\begin{aligned}\alpha_i &= \frac{\mu_{i,1}+\mu_{i,2}+\mu_{i,3}}{3} - \mu \\ &= \frac{\sum_{j=1}^{3}(a_i+b_j)}{3} - (\bar{a}+\bar{b}) \\ &= a_i - \bar{a}\end{aligned}$$

for $i = 1, 2$,

$$\begin{aligned}\beta_j &= \frac{\mu_{1,j}+\mu_{2,j}}{2} - \mu \\ &= \frac{\sum_{i=1}^{2}(a_i+b_j)}{2} - (\bar{a}+\bar{b}) \\ &= b_j - \bar{b}\end{aligned}$$

for $j = 1, 2, 3$, and

$$\begin{aligned}\gamma_{i,j} &= \mu_{i,j} - \mu - \alpha_i - \beta_j \\ &= a_i + b_j - (\bar{a}+\bar{b}) - (a_i - \bar{a}) - (b_j - \bar{b}) = 0\end{aligned}$$

for $i = 1, 2$, $j = 1, 2, 3$.

22. (a) The sum of squares due to error (SSE) is

$$\begin{aligned}&(6-1)\times(0.3+0.3+0.3+0.4+0.5+0.5+0.3+0.4+0.4+0.4+0.3+0.3) \\ &= 22,\end{aligned}$$

(b) The grand mean (total mean) $\bar{X}_G$ is

$$\begin{aligned}&\frac{2.8+3.0+2.7+2.5+2.6+2.8+2.7+2.8+3.0+2.8+2.7+2.5}{12} \\ &= \frac{32.9}{12} \approx 2.7417,\end{aligned}$$

and the sample means for the four major groups are

$$\frac{2.8+3.0+2.7}{3} = \frac{8.5}{3} \approx 2.8333,$$

$$\frac{2.5+2.6+2.8}{3} = \frac{7.9}{3} \approx 2.6333,$$

$$\frac{2.7+2.8+3.0}{3} = \frac{8.5}{3} \approx 2.8333,$$

and

$$\frac{2.8+2.7+2.5}{3} = \frac{8}{3} \approx 2.6667.$$

The sum of squares due to major, denoted by $\mathrm{SS}_{\mathrm{major}}$, is

$$\begin{aligned}&3\times 6\times[(8.5/3-32.9/12)^2+((7.9/3-32.9/12)^2 \\ &+(8.5/3-32.9/12)^2+(8/3-32.9/12)^2] \\ =\ &0.615.\end{aligned}$$

(c) The sum of squares due to interaction (SSI) will be computed using

$$\mathrm{SSI} = \mathrm{SS}_{\mathrm{total}} - \mathrm{SSE} - \mathrm{SS}_{\mathrm{major}} - \mathrm{SS}_{\mathrm{year}},$$

where $\mathrm{SS}_{\mathrm{year}}$ denotes the sum of squares due to year. Since $\mathrm{SS}_{\mathrm{major}}$ has been computed in Part (b), we will first compute $\mathrm{SS}_{\mathrm{year}}$ and then compute $\mathrm{SS}_{\mathrm{total}} - \mathrm{SSE}$ to obatin SSI.
To compute $\mathrm{SS}_{\mathrm{year}}$, note that the sample means for the three year groups are

$$\frac{2.8 + 2.5 + 2.7 + 2.8}{4} = 2.7,$$

$$\frac{3.0 + 2.6 + 2.8 + 2.7}{4} = 2.775,$$

and

$$\frac{2.7 + 2.8 + 3.0 + 2.5}{4} = 2.75,$$

so $\mathrm{SS}_{\mathrm{year}}$ is

$$\begin{aligned} & 4 \times 6 \times [(2.7 - 32.9/12)^2 + (2.775 - 32.9/12)^2 \\ & +(2.75 - 32.9/12)^2] \\ = \; & 0.07. \end{aligned}$$

Next, we compute $\mathrm{SS}_{\mathrm{total}} - \mathrm{SSE}$. Let $\bar{X}_{i,j}$ be the sample mean of the data in the group of the i-th major and the j-th graduation year for $i = 1, \ldots, 4$, $j = 1, 2, 3$. Then

$$\begin{aligned} \mathrm{SS}_{\mathrm{total}} - \mathrm{SSE} &= \sum_{i=1}^{4}\sum_{j=1}^{3} 6(\bar{X}_{i,j} - \bar{X}_G)^2 \\ &= 6\sum_{i=1}^{4}\sum_{j=1}^{3} \bar{X}_{i,j}^2 - 72\bar{X}_G^2 \\ &= 6 \times [2.8^2 + 3.0^2 + 2.7^2 + 2.5^2 + 2.6^2 + 2.8^2 \\ &\quad +2.7^2 + 2.8^2 + 3.0^2 + 2.8^2 + 2.7^2 + 2.5^2] \\ &\quad -72 \times (32.9/12)^2 \\ &= 1.735, \end{aligned}$$

so the sum of squares due to interaction (SSI) is

$$\begin{aligned} & \mathrm{SS}_{\mathrm{total}} - \mathrm{SSE} - \mathrm{SS}_{\mathrm{major}} - \mathrm{SS}_{\mathrm{year}} \\ & = 1.735 - 0.615 - 0.07 = 1.05. \end{aligned}$$

(d) The degree of freedom for the sum of squares due to interaction is $(4 - 1)(3 - 1) = 6$.

(e)

$$\begin{aligned} F &= \frac{\mathrm{SS}_{\mathrm{major}}/(4-1)}{\mathrm{SSE}/(4 \times 3 \times 6 - 4 \times 3)} \\ &= \frac{0.615/3}{22/60} \approx 0.5591 < f_{0.05,3,60} = 2.76, \end{aligned}$$

so we cannot conclude that the initial mean salaries for the four majors are not all the same at level 0.05.

(f) The degrees of freedom for SSI and SSE are 6 and $4 \times 3 \times 6 - 4 \times 3 = 60$ respectively, so

$$\begin{aligned} F &= \frac{\text{SSI}/6}{\text{SSE}/60} \\ &= \frac{1.05/6}{22/60} \approx 0.4773 < f_{0.05,6,60} = 2.25. \end{aligned}$$

We cannot conclude that there is a major-year interaction effect on salary at the 0.05 level.

23. (a) Direct calculation gives

$$\begin{aligned} & \sum_{i=1}^{k}\sum_{j=1}^{b} \ell(\bar{X}_{i,j} - \bar{X}_G)(\bar{X}_{i\cdot} - \bar{X}_G) \\ = \;& \ell \sum_{i=1}^{k}\sum_{j=1}^{b} (\bar{X}_{i,j} - \bar{X}_{i\cdot} + \bar{X}_{i\cdot} - \bar{X}_G)(\bar{X}_{i\cdot} - \bar{X}_G) \\ = \;& \ell \sum_{i=1}^{k}\sum_{j=1}^{b} (\bar{X}_{i,j} - \bar{X}_{i\cdot})(\bar{X}_{i\cdot} - \bar{X}_G) + \underbrace{\ell \sum_{i=1}^{k}\sum_{j=1}^{b} (\bar{X}_{i\cdot} - \bar{X}_G)(\bar{X}_{i\cdot} - \bar{X}_G)}_{\text{SSA}} \\ = \;& \left(\ell \sum_{i=1}^{k} (\bar{X}_{i\cdot} - \bar{X}_G) \underbrace{\sum_{j=1}^{b} (\bar{X}_{i,j} - \bar{X}_{i\cdot})}_{=0} \right) + \text{SSA} \\ = \;& \text{SSA}, \end{aligned}$$

where the last equality follws from

$$\sum_{j=1}^{b} (\bar{X}_{i,j} - \bar{X}_{i\cdot}) = \sum_{j=1}^{b} \left(\frac{1}{\ell} \sum_{m=1}^{\ell} X_{i,j,m} \right) - b \left(\frac{1}{b\ell} \sum_{j=1}^{b}\sum_{m=1}^{\ell} X_{i,j,m} \right) = 0.$$

(b) Direct calculation gives

$$\begin{aligned} & \sum_{i=1}^{k}\sum_{j=1}^{b} \ell(\bar{X}_{i,j} - \bar{X}_G)(\bar{X}_{\cdot j} - \bar{X}_G) \\ = \;& \ell \sum_{i=1}^{k}\sum_{j=1}^{b} (\bar{X}_{i,j} - \bar{X}_{\cdot j} + \bar{X}_{\cdot j} - \bar{X}_G)(\bar{X}_{\cdot j} - \bar{X}_G) \\ = \;& \ell \sum_{i=1}^{k}\sum_{j=1}^{b} (\bar{X}_{i,j} - \bar{X}_{\cdot j})(\bar{X}_{\cdot j} - \bar{X}_G) + \underbrace{\ell \sum_{i=1}^{k}\sum_{j=1}^{b} (\bar{X}_{\cdot j} - \bar{X}_G)(\bar{X}_{\cdot j} - \bar{X}_G)}_{\text{SSB}} \\ = \;& \left(\ell \sum_{j=1}^{b} (\bar{X}_{\cdot j} - \bar{X}_G) \underbrace{\sum_{i=1}^{k} (\bar{X}_{i,j} - \bar{X}_{\cdot j})}_{=0} \right) + \text{SSB} \\ = \;& \text{SSB}, \end{aligned}$$

where the last equality follws from

$$\sum_{i=1}^{k}(\bar{X}_{i,j}-\bar{X}_{\cdot j})=\sum_{i=1}^{k}\left(\frac{1}{\ell}\sum_{m=1}^{\ell}X_{i,j,m}\right)-k\left(\frac{1}{k\ell}\sum_{i=1}^{k}\sum_{m=1}^{\ell}X_{i,j,m}\right)=0.$$

(c) By definition,

$$\begin{aligned}
\text{SSI} &= \sum_{i=1}^{k}\sum_{j=1}^{b}\sum_{m=1}^{\ell}(\bar{X}_{i,j}-\bar{X}_{i\cdot}-\bar{X}_{\cdot j}+\bar{X}_G)^2\\
&= \ell\sum_{i=1}^{k}\sum_{j=1}^{b}(\bar{X}_{i,j}-\bar{X}_{i\cdot}-\bar{X}_{\cdot j}+\bar{X}_G)^2\\
&= \ell\sum_{i=1}^{k}\sum_{j=1}^{b}(\bar{X}_{i,j}-\bar{X}_G-(\bar{X}_{i\cdot}-\bar{X}_G)-(\bar{X}_{\cdot j}-\bar{X}_G))^2\\
&= \ell\sum_{i=1}^{k}\sum_{j=1}^{b}(\bar{X}_{i,j}-\bar{X}_G)^2+\underbrace{\ell\sum_{i=1}^{k}\sum_{j=1}^{b}(\bar{X}_{i\cdot}-\bar{X}_G)^2}_{=\text{SSA}}+\underbrace{\ell\sum_{i=1}^{k}\sum_{j=1}^{b}(\bar{X}_{\cdot j}-\bar{X}_G)^2}_{=\text{SSB}}\\
&\quad\underbrace{-2\,\ell\sum_{i=1}^{k}\sum_{j=1}^{b}(\bar{X}_{i,j}-\bar{X}_G)(\bar{X}_{i\cdot}-\bar{X}_G)}_{=\text{SSA (by Part (a))}}\underbrace{-2\,\ell\sum_{i=1}^{k}\sum_{j=1}^{b}(\bar{X}_{i,j}-\bar{X}_G)(\bar{X}_{\cdot j}-\bar{X}_G)}_{=\text{SSB (by Part (b))}}\\
&\quad+2\ell\sum_{i=1}^{k}\sum_{j=1}^{b}(\bar{X}_{i\cdot}-\bar{X}_G)(\bar{X}_{\cdot j}-\bar{X}_G)\\
&= \ell\sum_{i=1}^{k}\sum_{j=1}^{b}(\bar{X}_{i,j}-\bar{X}_G)^2-\text{SSA}-\text{SSB}+2\ell\sum_{i=1}^{k}\sum_{j=1}^{b}(\bar{X}_{i\cdot}-\bar{X}_G)(\bar{X}_{\cdot j}-\bar{X}_G)\\
&= \ell\sum_{i=1}^{k}\sum_{j=1}^{b}(\bar{X}_{i,j}-\bar{X}_G)^2-\text{SSA}-\text{SSB}+2\ell\sum_{i=1}^{k}(\bar{X}_{i\cdot}-\bar{X}_G)\sum_{j=1}^{b}(\bar{X}_{\cdot j}-\bar{X}_G),
\end{aligned}$$

where

$$\sum_{j=1}^{b}(\bar{X}_{\cdot j}-\bar{X}_G)=b\left[\left(\frac{1}{b}\sum_{j=1}^{b}\bar{X}_{\cdot j}\right)-\bar{X}_G\right]=0,$$

so

$$\text{SSI}=\ell\sum_{i=1}^{k}\sum_{j=1}^{b}(\bar{X}_{i,j}-\bar{X}_G)^2-\text{SSA}-\text{SSB}. \tag{17}$$

(d) Apply (12) in the solution to Problem 15 with

$$\{W_1,\ldots,W_m\}=\{X_{i,j,m}: i\in\{1,\ldots,k\}, j\in\{1,\ldots,b\}, m\in\{1,\ldots,\ell\}\}$$

and

$$\{W_1,\ldots,W_m\}=\{\bar{X}_{i,j}: i\in\{1,\ldots,k\}, j\in\{1,\ldots,b\},\},$$

respectively, and for each $i\in\{1,\ldots,k\}$ and $j\in\{1,\ldots,b\}$, apply (12) with

$$\{W_1,\ldots,W_m\}=\{X_{i,j,1},\ldots,X_{i,j,\ell}\},$$

then we have

$$\mathrm{SS}_{\text{total}} = \sum_{i=1}^{k}\sum_{j=1}^{b}\sum_{m=1}^{\ell}(X_{i,j,m} - \bar{X}_G)^2 = \left(\sum_{i=1}^{k}\sum_{j=1}^{b}\sum_{m=1}^{\ell} X_{i,j,m}^2\right) - kb\ell(\bar{X}_G)^2, \quad (18)$$

$$\sum_{i=1}^{k}\sum_{j=1}^{b}(\bar{X}_{i,j} - \bar{X}_G)^2 = \left(\sum_{i=1}^{k}\sum_{j=1}^{b}(\bar{X}_{i,j})^2\right) - kb(\bar{X}_G)^2, \quad (19)$$

and for each $i \in \{1, \ldots, k\}$ and $j \in \{1, \ldots, b\}$,

$$\sum_{m=1}^{\ell}(X_{i,j,m} - \bar{X}_{i,j})^2 = \left(\sum_{m=1}^{\ell} X_{i,j,m}^2\right) - \ell(\bar{X}_{i,j})^2,$$

which gives

$$\mathrm{SSE} = \sum_{i=1}^{k}\sum_{j=1}^{b}\sum_{m=1}^{\ell}(X_{i,j,m} - \bar{X}_{i,j})^2 = \sum_{i=1}^{k}\sum_{j=1}^{b}\left(\sum_{m=1}^{\ell} X_{i,j,m}^2\right) - \ell\sum_{i=1}^{k}\sum_{j=1}^{b}(\bar{X}_{i,j})^2 \quad (20)$$

From (18) and (20), we have

$$\begin{aligned}\mathrm{SS}_{\text{total}} - \mathrm{SSE} &= \left(\sum_{i=1}^{k}\sum_{j=1}^{b}\sum_{m=1}^{\ell} X_{i,j,m}^2\right) - kb\ell(\bar{X}_G)^2 - \left[\sum_{i=1}^{k}\sum_{j=1}^{b}\left(\sum_{m=1}^{\ell} X_{i,j,m}^2\right) - \ell\sum_{i=1}^{k}\sum_{j=1}^{b}(\bar{X}_{i,j})^2\right] \\ &\overset{(19)}{=} \ell\sum_{i=1}^{k}\sum_{j=1}^{b}(\bar{X}_{i,j} - \bar{X}_G)^2.\end{aligned}$$

Replace $\ell\sum_{i=1}^{k}\sum_{j=1}^{b}(\bar{X}_{i,j} - \bar{X}_G)^2$ in (17) with $\mathrm{SS}_{\text{total}} - \mathrm{SSE}$, then (17) becomes

$$\mathrm{SSI} = \mathrm{SS}_{\text{total}} - \mathrm{SSE} - \mathrm{SSA} - \mathrm{SSB},$$

which gives

$$\mathrm{SS}_{\text{total}} = \mathrm{SSA} + \mathrm{SSB} + \mathrm{SSI} + \mathrm{SSE}.$$

24. (a) No. The p-value is $0.55981 > 0.01$.
    (b) No. The p-value is $0.01032 > 0.01$.
    (c) Yes. The p-value is $2.548 \times 10^{-7} < 0.01$.
    (d) The degrees of freedom for the sum of squares due to the drug factor is 4, so the drug factor has $4 + 1 = 5$ levels and there are 5 types of drugs.
    (e) The degrees of freedom for the sum of squares due to the exercise factor is 3, so the exercise factor has $3 + 1 = 4$ levels and there are 4 types of exercises.
    (f) The total number of participants is $4 + 3 + 12 + 80 + 1 = 100$.

25. We will use the following results in the solution to Problem 22 for this problem.

    (i) The SSE in the two-way ANOVA is 22 with 60 degrees of freedom.

    (ii) $\mathrm{SS}_{\text{total}} - \mathrm{SSE} = 1.735$ in the two-way ANOVA, so the total sum of squares in the two-way ANOVA is

$$\mathrm{SS}_{\text{total}} = 1.735 + \mathrm{SSE} \overset{\text{(i)}}{=} 1.735 + 22 = 23.735$$

with 71 degrees of freedom.

(iii) The sum of squares due to major in the two-way ANOVA is 0.615 with three degrees of freedom.

(a) The total sum of squares and the sum of squares due to major in one-way ANOVA are the same as those in two-way ANOVA, which are 23.735 and 0.615 with degrees of freedom 71 and 3 respectively (based on (ii) and (iii)). Thus the sum of squares due to error in the one-way ANOVA is

$$\text{SSE} = \text{SS}_{\text{total}} - \text{SS}_{\text{major}} = 23.735 - 0.615 = 23.12.$$

(b) The total sum of squares in the one-way ANOVA is 23.735, as explained in Part (a).

(c) The sum of squares due to major and the sum of squares due to error in one-way ANOVA are 0.615 and 23.12 respectively from Part (a), so the observed F statistic is

$$F = \frac{0.615/3}{23.12/(71-3)} = 0.6029412.$$

Since

$$2.68 = f_{0.05,3,120} < f_{0.05,3,68} < f_{0.05,3,60} = 2.76,$$

$0.6029412 < f_{0.05,3,68}$. Therefore, we cannot conclude that inital expected wages for students of different majors are not all the same at the 0.05 level based on the one-way ANOVA.

26. (a) The total sum of squares in the one-way ANOVA is the same as that in the two-way ANOVA in Problem 24), which is

$$43.784 + 127.562 + 33.027 + 247.244 = 451.617.$$

(b) The sum of squares due to the drug factor is the same as that in the two-way ANOVA in Problem 24), which is 43.784 with 4 degrees of freedom. Thus the SSE in the one-way ANOVA is

$$\text{SS}_{\text{total}} - 43.784 \overset{(a)}{=} 451.617 - 43.784 = 407.833.$$

(c) From Part (b), the sum of squares due to the drug factor in the one-way ANOVA is 43.784 with 4 degrees of freedom, and the SSE is 407.833 with $100 - 1 - 4 = 95$ degrees of freedom since the total number of participants is 100 from Part (f) of Problem 24. Therefore, the observed F test statistic in the one-way ANOVA is

$$F = \frac{43.784/4}{407.833/95} = 2.549745.$$

From the table of 0.99 quantiles for F distributions,

$$3.48 = f_{0.01,4,120} < f_{0.01,4,95} < f_{0.01,4,60} = 3.65,$$

so $2.549745 < f_{0.01,4,95}$ and we cannot conclude that not all drugs have the same effect on weight loss at level 0.01 based on the one-way ANOVA.

27. (a) The total sum of squares is

$$43.784 + 127.562 + 33.027 + 247.244 = 451.617.$$

(b) The SSE is

$$451.617 - 43.784 - 127.562(= 33.027 + 247.244) = 280.271.$$

(c) The degrees of freedom for the SSE is $12 + 80 = 92$. The observed F statistics is

$$\frac{43.784/4}{280.271/92} = 3.593065.$$

From the R output, $P(F(4, 92) > 3.59) = 0.009128695 < 0.01$, so

$$f_{0.01,4,92} < 3.59 < 3.593065$$

and we can conclude that not all drugs have the same effect on weight loss at level 0.01 based on the two-way ANOVA without interaction.

Note that the table "0.99 quantiles for F distributions" only gives two relevant quantiles $f_{0.01,4,60} = 3.65$ and $f_{0.01,4,120} = 3.48$, which implies

$$3.48 < f_{0.01,4,92} < 3.65. \tag{21}$$

However, we cannot tell whether $f_{0.01,4,92} < 3.593065$ from (21) and it is necessary to use the R output for $P(F(4, 92) > 3.59)$ to solve this problem.

28. (a) The grand mean (total mean) $\bar{X}_G$ is

$$\frac{2.8 + 3.0 + 2.7 + 2.5 + 2.6 + 2.8 + 2.7 + 2.8 + 3.0 + 2.8 + 2.7 + 2.5}{12}$$
$$= \frac{32.9}{12} \approx 2.7417,$$

and the sample means for the four major groups are

$$\frac{2.8 + 3.0 + 2.7}{3} = \frac{8.5}{3},$$
$$\frac{2.5 + 2.6 + 2.8}{3} = \frac{7.9}{3},$$
$$\frac{2.7 + 2.8 + 3.0}{3} = \frac{8.5}{3},$$

and

$$\frac{2.8 + 2.7 + 2.5}{3} = \frac{8}{3}.$$

The sum of squares due to major, denoted by $\mathrm{SS}_{\mathrm{major}}$, is

$$\begin{aligned} & 3 \times [(8.5/3 - 32.9/12)^2 + ((7.9/3 - 32.9/12)^2 \\ & +(8.5/3 - 32.9/12)^2 + (8/3 - 32.9/12)^2] \\ = \; & 0.1025. \end{aligned}$$

To compute the sum of squares due to graduation year, note that the sample means for the three year groups are

$$\frac{2.8 + 2.5 + 2.7 + 2.8}{4} = 2.7,$$

$$\frac{3.0 + 2.6 + 2.8 + 2.7}{4} = 2.775,$$

and

$$\frac{2.7 + 2.8 + 3.0 + 2.5}{4} = 2.75,$$

so the sum of squares due to graduation year, denoted by SS_{year}, is

$$\begin{aligned} & 4 \times [(2.7 - 32.9/12)^2 + (2.775 - 32.9/12)^2 \\ & +(2.75 - 32.9/12)^2] \\ = \quad & \frac{0.07}{6}. \end{aligned}$$

The total sum of squares, denoted by SS_{total}, is

$$2.8^2 + 3.0^2 + 2.7^2 + 2.5^2 + 2.6^2 + 2.8^2 + 2.7^2 + 2.8^2 + 3.0^2 + 2.8^2 + 2.7^2 + 2.5^2 - 12\left(\frac{32.9}{12}\right)^2$$

$$= 90.49 - \frac{1082.41}{12}$$

The SSE is

$$\begin{aligned} & \text{SS}_{\text{total}} - \text{SS}_{\text{major}} - \text{SS}_{\text{year}} \\ & = 90.49 - \frac{1082.41}{12} - 0.1025 - \frac{0.07}{6} = 0.175 \end{aligned}$$

with degrees of freedom $12 - 1 - (4 - 1) - (3 - 1) = 6$. The observed F statistic (for testing the main effect of major) is

$$\frac{\text{SS}_{\text{major}}/(4 - 1)}{\text{SSE}/6} = \frac{0.1025/3}{0.175/6} = 1.171429.$$

From the table "0.95 quantiles for F distributions", $f_{0.05,3,6} = 4.76 > 1.171429$, so we cannot conclude that inital expected wages for students of different majors are not allthe same at level 0.05.

(b) No. There is only one observation for each major-year combination, so the SSE in the two-way ANOVA with interaction is zero and the F statistic for testing the interaction effect cannot be computed.

29. (a) Since

$$X_1 + a - E(X_1 + a) = X_1 + a - (E(X_1) + a) = X_1 - E(X_1),$$

we have

$$\begin{aligned} Cov(X_1 + a, Y) &= E[(X_1 + a - E(X_1 + a))(Y - E(Y))] \\ &= E[(X_1 - E(X_1))(Y - E(Y))] \\ &= Cov(X_1, Y). \end{aligned}$$

(b) Let $\mu_1 = E(X_1)$, $\mu_2 = E(X_2)$ and $\mu_Y = E(Y)$, then

$$E(X_1 + X_2) = \mu_1 + \mu_2 \tag{22}$$

and

$$E(aX_1) = a\mu_1. \tag{23}$$

Direct calculation gives

$$\begin{aligned} Cov(Y, X_1 + X_2, Y) &= E[(Y - E(Y))(X_1 + X_2 - E(X_1 + X_2))] \\ &\overset{(22)}{=} E[(Y - \mu_Y)(X_1 + X_2 - (\mu_1 + \mu_2))] \\ &= E[(Y - \mu_Y)((X_1 - \mu_1) + (X_2 - \mu_2))] \\ &= E[(Y - \mu_Y)(X_1 - \mu_1)] + E[(Y - \mu_Y)(X_2 - \mu_2)] \\ &= E[(Y - E(Y)(X_1 - E(X_1))] + E[(Y - E(Y))(X_2 - E(X_2))] \\ &= Cov(Y, X_1) + Cov(Y, X_2) \end{aligned}$$

and

$$\begin{aligned} Cov(Y, aX_1) &= E[(Y - \mu_Y)(aX_1 - E(aX_1))] \\ &\overset{(23)}{=} E[(Y - \mu_Y)(aX_1 - a\mu_1)] \\ &= E[(Y - \mu_Y) \cdot a(X_1 - \mu_1)] \\ &= aE[(Y - \mu_Y)(X_1 - \mu_1)] \\ &= aE[(Y - E(Y))(X_1 - E(X_1))] \\ &= aCov(Y, X_1). \end{aligned}$$

30. Direct calculation gives

$$E(X) = (-1) \times 0.3 + 0 \times 0.4 + a \times 0.3 = 0.3(a - 1),$$

$$E(X^2) = (-1)^2 \times 0.3 + 0^2 \times 0.4 + a^2 \times 0.3 = 0.3(a^2 + 1),$$

$$E(Y) = a \times 0.3 + 0 \times 0.4 + (-a) \times 0.3 = 0,$$

$$E(Y^2) = a^2 \times 0.3 + 0^2 \times 0.4 + (-a)^2 \times 0.3 = 0.6a^2,$$

and

$$E(XY) = (-1) \times a \times 0.3 + 0 \times 0 \times 0.4 + a \times (-a) \times 0.3 = -0.3a(a + 1),$$

so

$$Var(X) = E(X^2) - (E(X))^2 = 0.3(a^2 + 1) - (0.3)^2(a - 1)^2 = 0.3(0.7a^2 + 0.6a + 0.7),$$

$$Var(Y) = E(Y^2) - (E(Y))^2 = 0.6a^2 - 0^2 = 0.6a^2,$$

and

$$Cov(X, Y) = E(XY) - E(X)E(Y) = -0.3a(a - 1) - E(X) \times 0 = -0.3a(a + 1).$$

Therefore,

$$\begin{aligned} Corr(X, Y) &= \frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} \\ &= \frac{-0.3a(a + 1)}{\sqrt{0.3(0.7a^2 + 0.6a + 0.7)} \cdot \sqrt{0.6a^2}} \\ &= -\frac{a}{|a|} \cdot \frac{a + 1}{\sqrt{1.4a^2 + 1.2a + 1.4}}. \end{aligned}$$

31. Let $S(b) = Var(Y - bX)$ for $b \in R$, then we have

$$\begin{aligned} S(b) &= Var(Y) + Var(-bX) + 2Cov(Y, -bX) \\ &= Var(Y) + b^2 Var(X) - 2bCov(Y, X) \\ &= Var(X)\left(b - \frac{Cov(Y,X)}{Var(X)}\right)^2 + Var(Y) - Var(X)\left(\frac{Cov(Y,X)}{Var(X)}\right)^2. \quad (24) \end{aligned}$$

Since $S(b) = Var(Y - bX) \geq 0$ for all $b \in R$, when $b = Cov(Y, X)/Var(X)$, we have

$$s(b) = S\left(\frac{Cov(Y,X)}{Var(X)}\right) \geq 0. \quad (25)$$

Using (24) to compute $S(b)$ with $b = Cov(Y, X)/Var(X)$, then (25) becomes

$$Var(Y) - Var(X)\left(\frac{Cov(Y,X)}{Var(X)}\right)^2 \geq 0.$$

Multiply both sides of the above inequality by $Var(X)$, and we have

$$Var(X)Var(Y) - [Var(X)]^2\left(\frac{Cov(Y,X)}{Var(X)}\right)^2 \geq 0,$$

which gives

$$Var(X)Var(Y) - [Cov(X, Y)]^2 \geq 0,$$

so $[Cov(X, Y)]^2 \leq Var(X)Var(Y)$.

32. (a)

$$\begin{aligned} Corr(X, Y) &= \frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} \\ &= \frac{0.5}{\sqrt{1}\sqrt{0.25}} = 1. \end{aligned}$$

(b)

$$\begin{aligned} Var(X - 2Y) &= Var(X) + 2Cov(X, -2Y) + Var(-2Y) \\ &= Var(X) + 2(-2)Cov(X, Y) + (-2)^2 Var(Y) \\ &= 1 + (-4)(0.5) + (4)(0.25) = 0. \end{aligned}$$

33. Since $Y = a + bX$, where $a = 2$ and $b = 3 > 0$, we have

$$\begin{aligned} Cov(X, Y) &= Cov(X, a + bX) \\ &\overset{\text{Problem 29(a)}}{=} Cov(X, bX) \\ &= bCov(X, X) = bVar(X) \end{aligned}$$

and

$$Var(Y) = Var(a + bX) = Var(bX) = b^2 Var(X).$$

From the definition of correlation and the above calculation, we have

$$\begin{aligned} Corr(X, Y) &= \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} \\ &= \frac{bVar(X)}{\sqrt{Var(X)b^2 Var(X)}} = \frac{b}{|b|} = 1 \end{aligned}$$

since $b > 0$.

34. In the following calculation, we will use the assumption that the kb random variables $\bar{X}_{i,j}$: $i = 1, \ldots, k$, $j = 1, \ldots b$ are independent with common variance σ^2/ℓ, which implies that

$$Cov(\bar{X}_{i,j}, \bar{X}_{i',j'}) = \begin{cases} \sigma^2/\ell & \text{if } (i,j) = (i',j'); \\ 0 & \text{otherwise.} \end{cases}$$

(a) For $i = 1, \ldots, k$, $j = 1, \ldots, b$,

$$\begin{aligned} Cov(\bar{X}_{i,j}, \bar{X}_{i\cdot}) &= Cov\left(\bar{X}_{i,j}, \frac{\bar{X}_{i,1} + \cdots + \bar{X}_{i,b}}{b}\right) \\ &= \frac{1}{b} Cov(\bar{X}_{i,j}, \bar{X}_{i,j}) = \frac{1}{b}\left(\frac{\sigma^2}{\ell}\right) = \frac{\sigma^2}{b\ell}. \end{aligned}$$

(b) For $i = 1, \ldots, k$, $j = 1, \ldots, b$,

$$\begin{aligned} Cov(\bar{X}_{i,j}, \bar{X}_G) &= Cov\left(\bar{X}_{i,j}, \frac{1}{kb}\sum_{i'=1}^{k}\sum_{j'=1}^{b} \bar{X}_{i',j'}\right) \\ &= \frac{1}{kb} Cov(\bar{X}_{i,j}, \bar{X}_{i,j}) = \frac{1}{kb}\left(\frac{\sigma^2}{\ell}\right) = \frac{\sigma^2}{kb\ell}. \end{aligned}$$

(c) For $j = 1, \ldots, b$,

$$\begin{aligned} Cov(\bar{X}_{\cdot j}, \bar{X}_G) &= Cov\left(\frac{1}{k}\sum_{i=1}^{k} \bar{X}_{i,j}, \bar{X}_G\right) \\ &= \frac{1}{k}\sum_{i=1}^{k} Cov(\bar{X}_{i,j}, \bar{X}_G) \\ &\overset{\text{Part (b)}}{=} \frac{1}{k}\sum_{i=1}^{k} \frac{\sigma^2}{kb\ell} = \frac{\sigma^2}{kb\ell}. \end{aligned}$$

(d) Since $\bar{X}_{i\cdot} = \sum_{j'=1}^{b} \bar{X}_{i,j'}/b$, $\bar{X}_{i,1}$, …, $\bar{X}_{i,b}$ are independent, and $Var(\bar{X}_{i,j'}) = \sigma^2/\ell$, we have

$$Var(\bar{X}_{i\cdot}) = \left(\frac{1}{b}\right)^2 \sum_{j'=1}^{b} Var(\bar{X}_{i,j'}) = \frac{1}{b^2}\left(\frac{\sigma^2}{\ell}\right) \cdot b = \frac{\sigma^2}{b\ell}. \tag{26}$$

From Equation (11) given in Part (d) of the problem, $Cov(\bar{X}_{i\cdot} - \bar{X}_G, \bar{X}_{\cdot j} - \bar{X}_G) = 0$, so

$$\begin{aligned} &Cov(\bar{X}_{i\cdot} - \bar{X}_G, \bar{X}_{i,j} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X}_G) \\ &= Cov(\bar{X}_{i\cdot} - \bar{X}_G, \bar{X}_{i,j} - \bar{X}_{i\cdot}) + \underbrace{Cov(\bar{X}_{i\cdot} - \bar{X}_G, -\bar{X}_{\cdot j} + \bar{X}_G)}_{=0} \\ &= Cov(\bar{X}_{i\cdot} - \bar{X}_G, \bar{X}_{i,j} - \bar{X}_{i\cdot}) \\ &= Cov(\bar{X}_{i\cdot}, \bar{X}_{i,j}) + Cov(\bar{X}_{i\cdot}, -\bar{X}_{i\cdot}) + Cov(-\bar{X}_G, \bar{X}_{i,j}) + Cov(-\bar{X}_G, -\bar{X}_{i\cdot}) \\ &= \frac{\sigma^2}{b\ell} - \frac{\sigma^2}{b\ell} - \frac{\sigma^2}{kb\ell} + \frac{\sigma^2}{kb\ell} = 0. \end{aligned}$$

Here the last second equality follows from Equations (6) and (8) given in the problem, the fact that $Cov(\bar{X}_{i\cdot}, \bar{X}_{i,j}) = \sigma^2/(k\ell)$ established in Part (a), and the result $Var(\bar{X}_{i\cdot}) = \sigma^2/(b\ell)$ in (26).

(e) In Equations (27) – (30) below, we state the results in Equations (9)–(12) given in Part (e) of this problem:

$$Cov(\bar{X}_{i,j}, \bar{X}_{\cdot j}) = \frac{\sigma^2}{k\ell}, \tag{27}$$

$$Cov(\bar{X}_{i\cdot}, \bar{X}_G) = \frac{\sigma^2}{kb\ell}, \tag{28}$$

$$Var(\bar{X}_G) = \frac{\sigma^2}{kb\ell}, \tag{29}$$

and

$$Cov(\bar{X}_{i\cdot}, \bar{X}_{\cdot j}) = \frac{\sigma^2}{kb\ell}. \tag{30}$$

Then

$$\begin{aligned} & Cov(\bar{X}_{i\cdot} - \bar{X}_G, \bar{X}_G) \\ = \quad & Cov(\bar{X}_{i\cdot}, \bar{X}_G) - \underbrace{Cov(\bar{X}_G, \bar{X}_G)}_{=Var(\bar{X}_G)} \\ \overset{(28),(29)}{=} \quad & \frac{\sigma^2}{kb\ell} - \frac{\sigma^2}{kb\ell} = 0, \end{aligned} \tag{31}$$

$$\begin{aligned} & Cov(\bar{X}_{i\cdot} - \bar{X}_G, \bar{X}_{\cdot j} - \bar{X}_G) \\ = \quad & Cov(\bar{X}_{i\cdot} - \bar{X}_G, \bar{X}_{\cdot j}) - \underbrace{Cov(\bar{X}_{i\cdot} - \bar{X}_G, \bar{X}_G)}_{\overset{(31)}{=}0} \\ = \quad & Cov(\bar{X}_{i\cdot}, \bar{X}_{\cdot j}) - Cov(\bar{X}_G, \bar{X}_{\cdot j}) \\ \overset{(30),(c)}{=} \quad & \frac{\sigma^2}{kb\ell} - \frac{\sigma^2}{kb\ell} = 0, \end{aligned} \tag{32}$$

and

$$\begin{aligned} & Cov(\bar{X}_{i\cdot} - \bar{X}_G, \bar{X}_{i,j} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X}_G) \\ = \quad & Cov(\bar{X}_{i\cdot} - \bar{X}_G, \bar{X}_{i,j} - \bar{X}_{i\cdot}) - \underbrace{Cov(\bar{X}_{i\cdot} - \bar{X}_G, \bar{X}_{\cdot j} - \bar{X}_G)}_{\overset{(32)}{=}0} \\ = \quad & Cov(\bar{X}_{i\cdot}, \bar{X}_{i,j}) - \underbrace{Cov(\bar{X}_{i\cdot}, \bar{X}_{i\cdot})}_{=Var(\bar{X}_{i\cdot})} - Cov(\bar{X}_G, \bar{X}_{i,j}) + Cov(\bar{X}_G, \bar{X}_{i\cdot}) \\ \overset{(a),(d),(b),(28)}{=} \quad & \frac{\sigma^2}{b\ell} - \frac{\sigma^2}{b\ell} - \frac{\sigma^2}{kb\ell} + \frac{\sigma^2}{kb\ell} = 0 \end{aligned}$$

for $i = 1, \ldots, k$ and $j = 1, \ldots, b$.

35. By assumption, we have

$$Cov(\bar{Z}_G, \bar{Z}_{i\cdot} - \bar{Z}_G) = 0 \tag{33}$$

for $i = 1, \ldots, k$ and

$$Cov(\bar{Z}_G, \bar{Z}_{\cdot j} - \bar{Z}_G) = 0 \tag{34}$$

for $j = 1, \ldots, b$. We will first establish the following results:

$$Cov\left(\sum_{i=1}^{k} a_i \bar{Z}_{i\cdot}, X\right) = Cov(\bar{Z}_G, X) \sum_{i=1}^{k} a_i \tag{35}$$

for a random variable X and

$$Cov\left(\sum_{j=1}^{b} c_j \bar{Z}_{\cdot j}, X\right) = Cov(\bar{Z}_G, X) \sum_{j=1}^{b} c_j \tag{36}$$

for a random variable X. To prove (35), note that

$$\begin{aligned}
& Cov\left(\sum_{i=1}^{k} a_i \bar{Z}_{i\cdot}, X\right) \\
&= \sum_{i=1}^{k} a_i Cov\left(\bar{Z}_{i\cdot}, X\right) \\
&= \sum_{i=1}^{k} a_i Cov\left(\bar{Z}_{i\cdot} - \bar{Z}_G + \bar{Z}_G, X\right) \\
&= \sum_{i=1}^{k} a_i \left\{ \underbrace{Cov\left(\bar{Z}_{i\cdot} - \bar{Z}_G, X\right)}_{\overset{(33)}{=}0} + Cov\left(\bar{Z}_G, X\right) \right\} \\
&= Cov(\bar{Z}_G, X) \sum_{i=1}^{k} a_i,
\end{aligned}$$

so (35) holds. Similarly,

$$\begin{aligned}
& Cov\left(\sum_{j=1}^{b} c_j \bar{Z}_{\cdot j}, X\right) \\
&= \sum_{j=1}^{b} c_j Cov\left(\bar{Z}_{\cdot j}, X\right) \\
&= \sum_{j=1}^{b} c_j Cov\left(\bar{Z}_{\cdot j} - \bar{Z}_G + \bar{Z}_G, X\right) \\
&= \sum_{j=1}^{b} c_j \left\{ \underbrace{Cov\left(\bar{Z}_{\cdot j} - \bar{Z}_G, X\right)}_{\overset{(34)}{=}0} + Cov\left(\bar{Z}_G, X\right) \right\} \\
&= Cov(\bar{Z}_G, X) \sum_{j=1}^{b} c_j,
\end{aligned}$$

so (36) holds.

Now we are ready to prove the results in Parts (a) and (b).

(a) Apply (35) with $X = \bar{Z}_G$, then

$$Cov\left(\sum_{i=1}^{k} a_i \bar{Z}_{i\cdot}, \bar{Z}_G\right) = Cov(\bar{Z}_G, \bar{Z}_G) \sum_{i=1}^{k} a_i = Var(\bar{Z}_G) \sum_{i=1}^{k} a_i,$$

so

$$Cov\left(\sum_{i=1}^{k} a_i \bar{Z}_{i\cdot}, \bar{Z}_G\right) = 0 \Rightarrow Var(\bar{Z}_G) \sum_{i=1}^{k} a_i = 0,$$

which implies that $\sum_{i=1}^{k} a_i = 0$ since $Var(\bar{Z}_G) > 0$ by assumption.

(b)

$$\begin{aligned}
Cov\left(\sum_{i=1}^{k} a_i \bar{Z}_{i\cdot}, \sum_{j=1}^{b} c_j \bar{Z}_{\cdot j}\right) &= Cov\left(\bar{Z}_G, \sum_{j=1}^{b} c_j \bar{Z}_{\cdot j}\right) \sum_{i=1}^{k} a_i \qquad \text{(apply (35) with } X = \textstyle\sum_{j=1}^{b} c_j \bar{Z}_{\cdot j} \\
&= \left(\sum_{i=1}^{k} a_i\right) Cov\left(\sum_{j=1}^{b} c_j \bar{Z}_{\cdot j}, \bar{Z}_G\right) \\
&= \left(\sum_{i=1}^{k} a_i\right) Cov\left(\bar{Z}_G, \bar{Z}_G\right) \sum_{j=1}^{b} c_j \qquad \text{(apply (36) with } X = \bar{Z}_G \\
&= Var(\bar{Z}_G)\left(\sum_{i=1}^{k} a_i\right)\left(\sum_{j=1}^{b} c_j\right).
\end{aligned}$$

36. (a) Apply the result $\sum_{i=1}^{n} w_i^2 = \sum_{i=1}^{n}(w_i - \bar{w})^2 + n\bar{w}^2$ with $w_i = Y_i - a - bX_i$, then we have $\bar{w} = \bar{Y} - a - b\bar{X}$ and

$$\begin{aligned}
&\sum_{i=1}^{n}(Y_i - a - bX_i)^2 \\
&= \sum_{i=1}^{n}(Y_i - a - bX_i - (\bar{Y} - a - b\bar{X}))^2 + n(\bar{Y} - a - b\bar{X})^2 \\
&= \sum_{i=1}^{n}(Y_i - \bar{Y} - b(X_i - \bar{X}))^2 + n(\bar{Y} - a - b\bar{X})^2 \\
&= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 - 2b\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X}) + b^2\sum_{i=1}^{n}(X_i - \bar{X})^2 + n(\bar{Y} - a - b\bar{X})^2 \\
&= (n-1)\left(S_Y^2 - 2rS_XS_Yb + S_X^2b^2\right) + n(\bar{Y} - a - b\bar{X})^2.
\end{aligned}$$

Here the last equality follows from the definitions of S_Y, S_X and r:

$$S_Y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2},$$

$$S_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2},$$

and

$$r = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})/(n-1)}{S_XS_Y}.$$

(b) Let

$$S(b) = S_X^2b^2 - 2rS_XS_Yb + S_Y^2,$$

then from the result in Part (a),

$$\sum_{i=1}^{n}(Y_i - a - bX_i)^2 = (n-1)S(b) + n(\bar{Y} - a - b\bar{X})^2. \tag{37}$$

Note that when $S_X \neq 0$, $S(b)$ is a quadratic function of b with a global minimum. Solving $S'(b) = 0$ gives

$$2bS_X^2 - 2rS_XS_Y = 0,$$

which gives $b = rS_Y/S_X = \hat{b}$, so $S(b)$ is minimized when $b = \hat{b}$. From (37), we have

$$\begin{aligned}
\sum_{i=1}^{n}(Y_i - a - bX_i)^2 &\geq (n-1)S(b) \\
&\geq (n-1)S(\hat{b}) \qquad (S(b) \text{ is minimized when } b = \hat{b}) \\
&= (n-1)S(\hat{b}) + \underbrace{n(\bar{Y} - \hat{a} - \hat{b}\bar{X})^2}_{=0 \text{ since } \hat{a}=\bar{Y}-\hat{b}\bar{X}} \\
&\overset{(37)}{=} \left.\sum_{i=1}^{n}(Y_i - a - bX_i)^2\right|_{(a,b)=(\hat{a},\hat{b})},
\end{aligned}$$

so $\sum_{i=1}^{n}(Y_i - a - bX_i)^2$ is minimized when $(a, b) = (\hat{a}, \hat{b})$.

(c) Since $Y_i = a_0 + b_0X_i$, we have

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n}(a_0 + b_0X_i) = a_0 + b_0\bar{X} \tag{38}$$

and

$$Y_i - \bar{Y} = a_0 + b_0X_i - (a_0 + b_0\bar{X}) = b_0(X_i - \bar{X}) \tag{39}$$

for $i = 1, \ldots, n$. Therefore, the sample stanard deviation

$$\begin{aligned}
S_Y &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \\
&\overset{(39)}{=} \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}[b_0(X_i - \bar{X})]^2} \\
&= |b_0|S_X,
\end{aligned} \tag{40}$$

and the sample correlation

$$\begin{aligned}
r &= \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})/(n-1)}{S_XS_Y} \\
&\overset{(39)}{=} \frac{\sum_{i=1}^{n}(X_i - \bar{X})[b_0(X_i - \bar{X})]/(n-1)}{S_XS_Y} \\
&= \frac{b_0S_X^2}{S_XS_Y} \\
&\overset{(40)}{=} \frac{b_0S_X^2}{S_X|b_0|S_X} = \frac{b_0}{|b_0|}.
\end{aligned} \tag{41}$$

Thus

$$\begin{aligned}
\hat{b} &= \frac{rS_Y}{S_X} \\
&\overset{(40),(41)}{=} \frac{b_0}{|b_0|} \cdot \frac{|b_0|S_X}{S_X} = b_0
\end{aligned} \tag{42}$$

and

$$\begin{aligned}
\hat{a} &= \bar{Y} - \hat{b}\bar{X} \\
&\overset{(38),(42)}{=} a_0 + b_0\bar{X} - b_0\bar{X} = a_0.
\end{aligned}$$

We have verified that $(\hat{a}, \hat{b}) = (a_0, b_0)$.

(d)

$$\begin{aligned}
&\text{RSS} \\
&= \sum_{i=1}^{n}(Y_i - \hat{a} - \hat{b}X_i)^2 \\
&= \sum_{i=1}^{n}(Y_i - (\bar{Y} - \hat{b}\bar{X}) - \hat{b}X_i)^2 \\
&= \sum_{i=1}^{n}(Y_i - \bar{Y} - \hat{b}(X_i - \bar{X}))^2 \\
&= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 + (\hat{b})^2\sum_{i=1}^{n}(X_i - \bar{X})^2 - 2\hat{b}\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X}). \quad (43)
\end{aligned}$$

Since

$$\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X}) = rS_XS_Y,$$

dividing both sides of (43) by $(n-1)$ gives

$$\begin{aligned}
\frac{\text{RSS}}{n-1} &= S_Y^2 + (\hat{b})^2S_X^2 - 2\hat{b}rS_XS_Y \\
&= S_Y^2 + \left(\frac{rS_Y}{S_X}\right)^2 S_X^2 - 2\left(\frac{rS_Y}{S_X}\right)rS_XS_Y \\
&= S_Y^2 + r^2S_Y^2 - 2r^2S_Y^2 \\
&= S_Y^2(1 - r^2),
\end{aligned}$$

which implies that

$$\text{RSS} = (n-1)S_Y^2(1 - r^2).$$

37. (a) To prove the result in Part (a), we will first show that

$$\hat{b} = b_0 + \frac{\sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}. \quad (44)$$

To prove (44), note that

$$\hat{b} = \frac{rS_Y}{S_X} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2},$$

so

$$\begin{aligned}
\hat{b} - b_0 &= \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} - b_0 \\
&= \frac{\sum_{i=1}^{n}(X_i - \bar{X})(a_0 + b_0X_i + \varepsilon_i - (a_0 + b_0\bar{X} + \bar{\varepsilon}))}{(n-1)S_X^2} - b_0 \\
&= \frac{\sum_{i=1}^{n}(X_i - \bar{X})(\varepsilon_i - \bar{\varepsilon})}{(n-1)S_X^2} + \underbrace{\frac{\sum_{i=1}^{n}(X_i - \bar{X})(b_0X_i - b_0\bar{X})}{(n-1)S_X^2} - b_0}_{=0} \\
&= \frac{\sum_{i=1}^{n}(X_i - \bar{X})(\varepsilon_i - \bar{\varepsilon})}{(n-1)S_X^2}
\end{aligned}$$

and (44) holds.

Now we will prove the result in Part (a). Direct calculation gives

$$\begin{aligned}
\text{RSS} &= \sum_{i=1}^{n}(Y_i - \hat{a} - \hat{b}X_i)^2 \\
&= \sum_{i=1}^{n}(Y_i - \hat{a} - \hat{b}X_i - (\bar{Y} - \hat{a} - \hat{b}\bar{X}))^2 \\
&= \sum_{i=1}^{n}(Y_i - \bar{Y} - \hat{b}(X_i - \bar{X}))^2 \\
&= \sum_{i=1}^{n}(a_0 + b_0X_i + \varepsilon_i - (a_0 + b_0\bar{X} + \bar{\varepsilon}) - \hat{b}(X_i - \bar{X}))^2 \\
&= \sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon} - (\hat{b} - b_0)(X_i - \bar{X}))^2 \\
&= \sum_{i=1}^{n}\left\{(\varepsilon_i - \bar{\varepsilon})^2 - 2(\varepsilon_i - \bar{\varepsilon})(\hat{b} - b_0)(X_i - \bar{X}) + [(\hat{b} - b_0)(X_i - \bar{X})]^2\right\} \\
&= \left(\sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})^2\right) - 2(\hat{b} - b_0)\left(\sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})(X_i - \bar{X})\right) + (\hat{b} - b_0)^2\sum_{i=1}^{n}(X_i - \bar{X})^2 \\
&\overset{(44)}{=} \left(\sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})^2\right) - 2(\hat{b} - b_0)^2\sum_{i=1}^{n}(X_i - \bar{X})^2 + (\hat{b} - b_0)^2\sum_{i=1}^{n}(X_i - \bar{X})^2 \\
&= \left(\sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})^2\right) - (\hat{b} - b_0)^2\underbrace{\sum_{i=1}^{n}(X_i - \bar{X})^2}_{=(n-1)S_X^2} \\
&\overset{(44)}{=} \left(\sum_{i=1}^{n}\varepsilon_i^2\right) - (\sqrt{n}\bar{\varepsilon})^2 - \left(\frac{\sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})(X_i - \bar{X})}{\sqrt{(n-1)S_X^2}}\right)^2 \\
&= \left(\sum_{i=1}^{n}\varepsilon_i^2\right) - (\sqrt{n}\bar{\varepsilon})^2 - \left(\frac{\sum_{i=1}^{n}(X_i - \bar{X})\varepsilon_i}{\sqrt{(n-1)S_X^2}}\right)^2,
\end{aligned}$$

where the last equality follows from the result that

$$\sum_{i=1}^{n}(X_i - \bar{X})\bar{\varepsilon} = \bar{\varepsilon}\underbrace{\sum_{i=1}^{n}(X_i - \bar{X})}_{=n\bar{X}-n\bar{X}=0} = 0. \tag{45}$$

(b) From (44) in the solution to Part (a), we have

$$\begin{aligned}
\hat{b} - b_0 &= \frac{\sum_{i=1}^{n}(X_i - \bar{X})(\varepsilon_i - \bar{\varepsilon})}{(n-1)S_X^2} \\
&\overset{(45)}{=} \frac{\sum_{i=1}^{n}\varepsilon_i(X_i - \bar{X})}{(n-1)S_X^2}.
\end{aligned}$$

It remains to show that for $x_0 \in R$,

$$\hat{a} + \hat{b}x_0 - (a_0 + b_0x_0) = \bar{\varepsilon} + (x_0 - \bar{X})(\hat{b} - b_0). \tag{46}$$

Direction calculation gives

$$\begin{aligned}
&\hat{a} + \hat{b}x_0 - (a_0 + b_0 x_0) \\
&= \bar{Y} - \hat{b}\bar{X} + \hat{b}x_0 - (a_0 + b_0 x_0) \\
&= a_0 + b_0\bar{X} + \bar{\varepsilon} - \hat{b}\bar{X} + \hat{b}x_0 - (a_0 + b_0 x_0) \\
&= b_0(\bar{X} - x_0) - \hat{b}(\bar{X} - x_0) + \bar{\varepsilon} \\
&= (x_0 - \bar{X})(\hat{b} - b_0) + \bar{\varepsilon},
\end{aligned}$$

so (46) holds.

(c) From Part (b), we have

$$\hat{b} - b_0 = \frac{\sum_{i=1}^n \varepsilon_i(X_i - \bar{X})}{(n-1)S_X^2}.$$

To obtain $Cov(\bar{\varepsilon}, \hat{b})$, we will first compute $Cov(\bar{\varepsilon}, \varepsilon_i)$:

$$\begin{aligned}
Cov(\bar{\varepsilon}, \varepsilon_i) &= \frac{1}{n} Cov(\varepsilon_1 + \cdots + \varepsilon_n, \varepsilon_i) \\
&= \frac{Cov(\varepsilon_i, \varepsilon_i)}{n} = \frac{\sigma_\varepsilon^2}{n}.
\end{aligned}$$

Thus

$$\begin{aligned}
Cov(\bar{\varepsilon}, \hat{b}) &= Cov(\bar{\varepsilon}, \hat{b} - b_0) \\
&= Cov\left(\bar{\varepsilon}, \frac{\sum_{i=1}^n \varepsilon_i(X_i - \bar{X})}{(n-1)S_X^2}\right) \\
&= \frac{1}{(n-1)S_X^2} \cdot \sum_{i=1}^n (X_i - \bar{X}) Cov\left(\bar{\varepsilon}, \varepsilon_i\right) \\
&= \frac{1}{(n-1)S_X^2} \cdot \sum_{i=1}^n (X_i - \bar{X}) \frac{\sigma_\varepsilon^2}{n} \\
&= \frac{1}{(n-1)S_X^2} \cdot \frac{\sigma_\varepsilon^2}{n} \underbrace{\sum_{i=1}^n (X_i - \bar{X})}_{=n\bar{X} - n\bar{X} = 0} = 0. \qquad (47)
\end{aligned}$$

To prove

$$Var(\hat{b}) = \frac{\sigma_\varepsilon^2}{(n-1)S_X^2}, \qquad (48)$$

note that from Part (b), we have

$$\hat{b} - b_0 = \frac{\sum_{i=1}^n \varepsilon_i(X_i - \bar{X})}{(n-1)S_X^2},$$

so

$$\begin{aligned}
Var(\hat{b}) &= Var\left(\frac{\sum_{i=1}^n \varepsilon_i(X_i - \bar{X})}{(n-1)S_X^2}\right) \\
&= \frac{1}{[(n-1)S_X^2]^2} \sum_{i=1}^n Var(\varepsilon_i(X_i - \bar{X})) \text{ (since } \varepsilon_1, \ldots, \varepsilon_n \text{ are independent)}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{[(n-1)S_X^2]^2} \sum_{i=1}^{n} (X_i - \bar{X})^2 \underbrace{Var(\varepsilon_i)}_{=\sigma_\varepsilon^2} \\
&= \frac{\sigma_\varepsilon^2}{[(n-1)S_X^2]^2} \underbrace{\sum_{i=1}^{n} (X_i - \bar{X})^2}_{=(n-1)S_X^2} \\
&= \frac{\sigma_\varepsilon^2}{(n-1)S_X^2}
\end{aligned}
$$

and (48) holds.

Next, we compute $Var(\hat{a} + \hat{b}x_0)$ for a given x_0:

$$
\begin{aligned}
Var(\hat{a} + \hat{b}x_0) &= Var\left(\hat{a} + \hat{b}x_0 - (a_0 + b_0 x_0)\right) \\
&\overset{(46)}{=} Var\left(\bar{\varepsilon} + (x_0 - \bar{X})(\hat{b} - b_0)\right) \\
&= Var(\bar{\varepsilon}) + Var((x_0 - \bar{X})(\hat{b} - b_0)) + 2Cov(\bar{\varepsilon}, (x_0 - \bar{X})(\hat{b} - b_0)) \\
&= Var(\bar{\varepsilon}) + (x_0 - \bar{X})^2 Var(\hat{b} - b_0) + 2(x_0 - \bar{X})Cov(\bar{\varepsilon}, \hat{b} - b_0) \\
&= \frac{\sigma_\varepsilon^2}{n} + (x_0 - \bar{X})^2 Var(\hat{b}) + 2(x_0 - \bar{X}) \underbrace{Cov(\bar{\varepsilon}, \hat{b})}_{\overset{(47)}{=}0} \\
&= \frac{\sigma_\varepsilon^2}{n} + (x_0 - \bar{X})^2 Var(\hat{b}) \\
&\overset{(48)}{=} \frac{\sigma_\varepsilon^2}{n} + \frac{(x_0 - \bar{X})^2 \sigma_\varepsilon^2}{(n-1)S_X^2} = \sigma_\varepsilon^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{X})^2}{(n-1)S_X^2} \right).
\end{aligned}
$$

38. (a) Since the distribution $(X, Y)^T$ is a bivariate normal distribution and

$$
\begin{pmatrix} Y - bX \\ X \end{pmatrix} = \begin{pmatrix} -b & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix},
$$

the distribution of $(Y - bX, X)^T$ is also a bivariate normal distribution. Thus $Y - bX$ and X are independent if and only if $Cov(Y - bX, X) = 0$. Solving $Cov(Y - bX, X) = 0$ for b gives $Cov(Y, X) - bCov(X, X) = 0$, so

$$
b = \frac{Cov(Y, X)}{Cov(X, X)} = \frac{6}{9} = \frac{2}{3},
$$

and $Y - (2/3)X$ is independent of X.

(b)

$$
E(U) = E(Y - (2/3)X) = E(Y) - \frac{2}{3}E(X) = 2 - \frac{2}{3} \cdot 1 = \frac{4}{3}.
$$

$$
\begin{aligned}
Var(U) &= Var(Y - (2/3)X) \\
&= Var(Y) + Var(-(2/3)X) + 2Cov(Y, -(2/3)X) \\
&= 16 + \left(-\frac{2}{3}\right)^2 Var(X) - \frac{4}{3}Cov(Y, X) \\
&= 16 + \frac{4}{9} \cdot 9 - \frac{4}{3} \cdot 6 = 12.
\end{aligned}
$$

39. (a) The sample correlation is

$$r = \frac{0.2895}{\sqrt{0.383 \cdot 5.273}} = 0.2037137.$$

(b) The test statistic is

$$T = \sqrt{(70-2)} \cdot \frac{r}{\sqrt{1-r^2}} = \frac{68 \cdot 0.2037137}{\sqrt{1-0.2037137^2}} = 1.715847.$$

Since $t_{0.05/2,70-2} = t_{0.025,68} = 1.995 > |1.715847|$, we cannot conclude that $b \neq 0$ at level 0.05.

40. (a) For brevity, will use "price" and "age" to denote the car price and car age respectively. The estimated regression equation is

$$\widehat{\text{price}} = \hat{a} + \hat{b} \cdot \text{age}.$$

where

$$\hat{b} = \frac{-0.6 \cdot 1.9}{2.2} = -0.518$$

and

$$\hat{a} = 6.9 - \hat{b} \cdot 8.9 = 6.9 - \left(\frac{-0.6 \cdot 1.9}{2.2}\right) \cdot 8.9 = 11.512,$$

so the estimated regression equation is

$$\widehat{\text{price}} = 11.512 - 0.518 \cdot \text{age}.$$

(b) The residual sum of squares (RSS) is

$$(30-1)S^2_{\text{price}} \cdot (1-(-0.6)^2) = 29 \cdot (1.9)^2 \cdot 0.64 = 67.0016,$$

where S_{price} is the sample standard deviation of the observed car prices.

41. Substract 88 and 98 from the Stock A prices and the Stock B prices respectively, then we have the modified samples: $(0.3, -0.4, -0.3, -0.1, 0.5)$ and $(0.4, -0.4, -0.3, -0.1, 0.4)$. For the two modified samples, the sample means are

$$\frac{0.3-0.4-0.3-0.1+0.5}{5} = 0$$

and

$$\frac{0.4-0.4-0.3-0.1+0.4}{5} = 0,$$

and the sample variances are

$$\frac{(0.3)^2+(-0.4)^2+(-0.3)^2+(-0.1)^2+(0.5)^2}{4} = \frac{0.6}{4} = 0.15$$

and

$$\frac{(0.4)^2+(-0.4)^2+(-0.3)^2+(-0.1)^2+(0.4)^2}{4} = \frac{0.58}{4} = 0.145.$$

The sample covariance between the two modified samples is

$$\frac{0.3 \cdot 0.4 + (-0.4) \cdot (-0.4) + (-0.3) \cdot (-0.3) + (-0.1) \cdot (-0.1) + 0.5 \cdot 0.4}{4} = \frac{0.58}{4} = 0.145,$$

so the sample correlation between the two modified samples is

$$\frac{0.145}{\sqrt{0.145 \cdot 0.15}} = 0.9831921.$$

The sample correlation between the orginial samples is the same as the sample correlation the two modified samples, which is 0.9831921.

42. Let $\bar{D}$ and S_D be the sample mean and sample standard deviation for the sample (D_1, D_2, D_3, D_4), then for a given distance D,

$$Var(\hat{a} + \hat{b}d_0) = \sigma_\varepsilon^2 \left(\frac{1}{4} + \frac{(d_0 - \bar{D})^2}{3S_D^2} \right).$$

The average variance of estimated expected signal losses is

$$\frac{1}{3} \sum_{d_0=6}^{8} Var(\hat{a} + \hat{b}d_0) \quad = \quad \frac{1}{3} \sum_{d_0=6}^{8} \sigma_\varepsilon^2 \left(\frac{1}{4} + \frac{(d_0 - \bar{D})^2}{3S_D^2} \right)$$

Apply the fact that

$$\sum_{i=1}^{n} w_i^2 = \sum_{i=1}^{n} (w_i - \bar{w})^2 + n(\bar{w})^2$$

with $n = 3$ and $(w_1, w_2, w_3) = (6 - \bar{D}, 7 - \bar{D}, 8 - \bar{D})$, then we have

$$\begin{aligned} \sum_{d_0=6}^{8} (d_0 - \bar{D})^2 &= \sum_{d_0=6}^{8} (d_0 - \bar{D} - (7 - \bar{D}))^2 + 3(7 - \bar{D})^2 \\ &= \sum_{d_0=6}^{8} (d_0 - 7)^2 + 3(7 - \bar{D})^2 = 2 + 3(7 - \bar{D})^2, \end{aligned}$$

so

$$\frac{1}{3} \sum_{d_0=6}^{8} Var(\hat{a} + \hat{b}d_0) = \sigma_\varepsilon^2 \left(\frac{1}{4} + \frac{2 + 3(7 - \bar{D})^2}{9S_D^2} \right).$$

Let

$$S(D_1, D_2, D_3, D_4) = \frac{2 + 3(7 - \bar{D})^2}{3S_D^2},$$

then the average variance of estimated expected signal losses is minimized when $S(D_1, D_2, D_3, D_4)$ is minimized. Direct calculation gives

$$(4 + 6 + 8 + 10)/4 = 7 = (5 + 6 + 8 + 9)/4,$$

$$S(4, 6, 8, 10) = \frac{2 + 3(7 - 7)^2}{(4 - 7)^2 + (6 - 7)^2 + (8 - 7)^2 + (10 - 7)^2} = \frac{2}{20},$$

$$S(5, 6, 8, 9) = \frac{2 + 3(7 - 7)^2}{(5 - 7)^2 + (6 - 7)^2 + (8 - 7)^2 + (9 - 7)^2} = \frac{2}{10},$$

It is clear that

$$S(4, 6, 8, 10) < S(5, 6, 8, 9),$$

so the (D_1, D_2, D_3, D_4) in (a) gives the smallest average variance of estimated expected signal losses.

43. (a) For brevity, will use "price" and "age" to denote the car price and car age respectively. The estimated regression equation is

$$\widehat{\text{price}} = \hat{a} + \hat{b} \cdot \text{age}.$$

where

$$\hat{b} = \frac{-0.6 \cdot 1.9}{2.2} = -0.518$$

and

$$\hat{a} = 6.9 - \hat{b} \cdot 8.9 = 6.9 - \left(\frac{-0.6 \cdot 1.9}{2.2}\right) \cdot 8.9 = 11.512,$$

so the estimated regression equation is

$$\widehat{\text{price}} = 11.512 - 0.518 \cdot \text{age}.$$

(b) The residual sum of squares (RSS) is

$$(30-1)S^2_{\text{price}} \cdot (1 - (-0.6)^2) = 29 \cdot (1.9)^2 \cdot 0.64 = 67.0016,$$

where S_{price} is the sample standard deviation of the observed car prices.

(c) A 95% confidence interval for the coefficent of car age is given by

$$\hat{b} \pm t_{0.05/2,30-2} \sqrt{\frac{\hat{\sigma}^2_\varepsilon}{(30-1) \cdot (2.2)^2}},$$

where $\hat{b} = -0.6 \cdot 1.9/2.2$ from Part (a), $t_{0.05/2,30-2} = t_{0.025,28} = 2.048$, and

$$\begin{aligned} \hat{\sigma}^2_\varepsilon &= \frac{\text{RSS}}{30-2} \\ &= \frac{(30-1)(1.9)^2(1-(-0.6)^2)}{30-2} \qquad \text{(using the RSS formula in Problem 38(b),} \end{aligned}$$

so

$$\begin{aligned} \sqrt{\frac{\hat{\sigma}^2_\varepsilon}{(30-1) \cdot (2.2)^2}} &= \sqrt{\frac{(30-1)(1.9)^2(1-(-0.6)^2)}{30-2} \cdot \frac{1}{(30-1) \cdot (2.2)^2}} \\ &= \frac{1.9}{2.2}\sqrt{\frac{1-(-0.6)^2}{28}}. \end{aligned}$$

The 95% confidence interval for the coefficient of car age is

$$\frac{-0.6 \cdot 1.9}{2.2} \pm 2.048 \cdot \frac{1.9}{2.2}\sqrt{\frac{1-(-0.6)^2}{28}},$$

which gives $[-0.786, -0.251]$. Using the open interval $(-0.786, -0.251)$ as a 95% confidence interval is also fine.

(d) A 95% confidence interval for the expected car price when the car age is 8 is given by

$$\hat{a} + \hat{b} \cdot 8 \pm t_{0.05/2,30-2} \cdot \hat{\sigma}_\varepsilon \sqrt{\frac{1}{30} + \frac{(8-\overline{\text{age}})^2}{(30-1)S^2_{\text{age}}}},$$

where

$$\hat{a} + \hat{b} \cdot 8 = 6.9 - \left(\frac{-0.6 \cdot 1.9}{2.2}\right) \cdot 8.9 + \frac{-0.6 \cdot 1.9}{2.2} \cdot 8 = \frac{16.202}{2.2},$$

$t_{0.05/2,30-2} = t_{0.025,28} = 2.048$,

$\hat{\sigma}_\varepsilon = \sqrt{\frac{(30-1)(1.9)^2(1-(-0.6)^2)}{30-2}}$ (from the solution to Problem 36(b)),

and

$$\begin{aligned}\hat{\sigma}_\varepsilon \sqrt{\frac{1}{30} + \frac{(8-\overline{\text{age}})^2}{(30-1)S^2_{\text{age}}}} &= \sqrt{\frac{(30-1)(1.9)^2(1-(-0.6)^2)}{30-2}} \sqrt{\frac{1}{30} + \frac{(8-8.9)^2}{29 \cdot (2.2)^2}} \\ &= \frac{1.9}{2.2} \cdot \sqrt{\frac{105.3824}{840}}.\end{aligned}$$

The upper bound for the 95% confidence interval is

$$\frac{16.206}{2.2} + 2.048 \cdot \frac{1.9}{2.2} \cdot \sqrt{\frac{105.3824}{870}} \approx 7.993,$$

and the lower bound for the 95% confidence interval is

$$\frac{16.206}{2.2} - 2.048 \cdot \frac{1.9}{2.2} \cdot \sqrt{\frac{105.3824}{870}} \approx 6.740,$$

so the 95% confidence interval is $[6.740, 7.993]$.

(e) A 95% prediction interval for car price when the car age is 8 is given by

$$\hat{a} + \hat{b} \cdot 8 \pm t_{0.05/2,30-2} \cdot \hat{\sigma}_\varepsilon \sqrt{1 + \frac{1}{30} + \frac{(8-\overline{\text{age}})^2}{(30-1)S^2_{\text{age}}}}$$

where

$$\hat{a} + \hat{b} \cdot 8 = 6.9 - \left(\frac{-0.6 \cdot 1.9}{2.2}\right) \cdot 8.9 + \frac{-0.6 \cdot 1.9}{2.2} \cdot 8 = \frac{16.202}{2.2},$$

$t_{0.05/2,30-2} = t_{0.025,28} = 2.048$ and

$$\begin{aligned}\hat{\sigma}_\varepsilon \sqrt{1 + \frac{1}{30} + \frac{(8-\overline{\text{age}})^2}{(30-1)S^2_{\text{age}}}} &= \sqrt{\frac{(30-1)(1.9)^2(1-(-0.6)^2)}{30-2}} \cdot \sqrt{1 + \frac{1}{30} + \frac{(8-8.9)^2}{29 \cdot (2.2)^2}} \\ &= \frac{1.9}{2.2} \cdot \sqrt{\frac{2800.2944}{840}}.\end{aligned}$$

The upper bound for the 95% prediction interval is

$$\frac{16.206}{2.2} + 2.048 \cdot \frac{1.9}{2.2} \cdot \sqrt{\frac{2800.2944}{840}} \approx 10.596,$$

and the lower bound for the 95% prediction interval is

$$\frac{16.206}{2.2} - 2.048 \cdot \frac{1.9}{2.2} \cdot \sqrt{\frac{105.3824}{840}} \approx 4.137,$$

so the 95% predition interval is $[4.137, 10.596]$.

(f) The observed T statistic is

$$\sqrt{30-2}\left(\frac{-0.6}{\sqrt{1-(-0.6)^2}}\right) = -3.968627$$

and $t_{0.05/2,30-2} = t_{0.025,28} = 2.048 < |-3.968627|$, so we can conclude there is a linear relation between car price and car age at level 0.05.

(g) $(-0.6)^2 = 0.36$, so 36% of variation in car sale price can be explained by car age.

44. (a) The p-value is $0.0121 > 0.01$, so we cannot conclude that there is a linear relation between height and weight at level 0.01.

(b) The estimated regression equation is

$$\widehat{\text{weight}} = -27.2405 + 0.4489 \times \text{height}.$$

(c) Since the multiple R-squared is 0.2439, 24.39% of the variation in weight can be explained by height.

(d)

$$\sqrt{\frac{\text{RSS}}{23}} = 4.314,$$

so the residual sum of squares (RSS) is $23 \cdot (4.314)^2 = 428.0437$.

(e) A 95% C.I. for the coefficient of height is $0.4489 \pm t_{0.025,23} \times 0.1648$, where $t_{0.025,23} = 2.069$. The lower bound of the 95% C.I. is

$$0.4489 - 2.069 \times 0.1648 \approx 0.108$$

and the upper bound of the 95% C.I. is

$$0.4489 + 2.069 \times 0.1648 \approx 0.790$$

The 95% C.I. for the coefficient of height is $[0.108, 0.790]$.

45. Note that when $k = 1$ and $(X_{1,1}, \ldots, X_{1,n}) = (X_1, \ldots, X_n)$,

$$\begin{aligned}
(\boldsymbol{X}^*)^T \boldsymbol{X}^* &= \begin{pmatrix} 1 & \cdots & 1 \\ X_{1,1} & \cdots & X_{n,1} \end{pmatrix} \begin{pmatrix} 1 & X_{1,1} \\ \vdots & \vdots \\ 1 & X_{n,1} \end{pmatrix} \\
&= \begin{pmatrix} 1 & \cdots & 1 \\ X_1 & \cdots & X_n \end{pmatrix} \begin{pmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix} \\
&= \begin{pmatrix} n & \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i & \sum_{i=1}^n X_i^2 \end{pmatrix},
\end{aligned}$$

so the covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$\sigma_\varepsilon^2((\boldsymbol{X}^*)^T \boldsymbol{X}^*)^{-1} = \sigma_\varepsilon^2 \begin{pmatrix} n & \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i & \sum_{i=1}^n X_i^2 \end{pmatrix}^{-1} = \frac{\sigma_\varepsilon^2}{n\sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2} \begin{pmatrix} \sum_{i=1}^n X_i^2 & -\sum_{i=1}^n X_i \\ -\sum_{i=1}^n X_i & n \end{pmatrix}.$$

Since $Var(\hat{\beta}_1)$ is the $(2,2)$-th element of the covariance matrix of $\hat{\boldsymbol{\beta}}$,

$$\begin{aligned} Var(\hat{\beta}_1) &= \frac{\sigma_\varepsilon^2}{n\sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2} \cdot n \\ &= \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n X_i^2 - n(\sum_{i=1}^n X_i/n)^2} \\ &= \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ &= \frac{\sigma_\varepsilon^2}{(n-1)S_X^2}, \end{aligned}$$

where $\bar{X} = \sum_{i=1}^n X_i/n$ and $(n-1)S_X^2 = \sum_{i=1}^n (X_i - \bar{X})^2$. Therefore, when $k = 1$ and $(X_{1,1}, \ldots, X_{1,n}) = (X_1, \ldots, X_n)$, the $Var(\hat{\beta}_1)$ computed using the $(2,2)$-th element of

$$\sigma_\varepsilon^2((\boldsymbol{X}^*)^T \boldsymbol{X}^*)^{-1}$$

is the same as the $Var(\hat{b})$ given in Problem 37(c).

46. (a) The estimated regression equation is

$$\widehat{\text{price}} = 160.90813 - 6.77199 \cdot \text{age} - 0.13038 \cdot \text{mileage}.$$

(b) Based on the value of multiple R-squared, 21.81% of the variation in price that can be explained by age and mileage.

(c) Yes. The p-value for the F test is $0.01056 < 0.05$, so we can conclude that the coefficients for age and mileage are not all zero at level 0.05.

(d) No. The p-value is $0.0278 > 0.01$, so we cannot conclude that the coefficient for age is not zero at level 0.01.

(e) An observed 95% confidence interval for the coefficient of mileage is

$$-0.13038 \pm t_{0.05/2,n-1-k} \cdot 0.06114,$$

where $n - 1 - k = 37$ since the residual standard error has 37 degrees of freedom, so $t_{0.05/2,n-1-k} = t_{0.025,37} = 2.026$. Plug in $t_{0.05/2,n-1-k} = 2.026$ in the formula for the 95% confidence interval, then the upper bound of the interval is

$$-0.13038 + 2.026 \cdot 0.06114 = -0.00651036$$

and the lower bound is

$$-0.13038 - 2.026 \cdot 0.06114 = -0.2542496$$

and we have the 95% confidence interval for the coefficient of mileage is $[-0.2542496, -0.00651036]$.

(f) Since $n - 1 - k = 37$ with $k = 2$, $n = 37 + 1 + 2 = 40$.

(g) $\sqrt{\text{RSS}/37} = 36.38$, so the residual sum of squares is $\text{RSS} = 37 \cdot (36.38)^2 = 48969.66$.

(h) The regression sum of squares SS_{reg} satisfies

$$\frac{\text{SS}_{\text{reg}}/k}{\text{RSS}/(n-1-k)} = F \text{ value},$$

where $k = 2$ and $n - 1 - k = 37$, so

$$\frac{\mathrm{SS}_{\mathrm{reg}}/2}{37 \cdot (36.38)^2/37} = 5.16$$

and the regression sum of squares $\mathrm{SS}_{\mathrm{reg}}$ is $5.16 \cdot 2 \cdot 37 \cdot (36.38)^2/37 = 13658.57$.

(i) The total sum of squares is

$$\mathrm{SS}_{\mathrm{total}} = \mathrm{RSS} + \mathrm{SS}_{\mathrm{reg}} = 37 \cdot (36.38)^2 + 5.16 \cdot 2 \cdot 37 \cdot (36.38)^2/37 = 62628.23.$$

Note. For the computation of $\mathrm{SS}_{\mathrm{total}}$ and $\mathrm{SS}_{\mathrm{reg}}$, one can also use

$$\frac{\mathrm{RSS}}{\mathrm{SS}_{\mathrm{total}}} = 1 - 0.2181 = 0.7819$$

and RSS = 48969.66 to find $\mathrm{SS}_{\mathrm{total}} = 48969.66/0.7819 = 62629.06$ and then compute

$$\mathrm{SS}_{\mathrm{reg}} = \mathrm{SS}_{\mathrm{total}} - \mathrm{RSS} = 62629.06 - 48969.66 = 13659.4.$$

47. In the first linear regression model where `V2`, `V3` and `V4` are used as explanatory variables to explain `V1` (`(V1~V2+V3+V4)`), the p-value for the overall F test is less than 2.2×10^{-16}, which gives strong evidence that the coefficients of `V2`, `V3` and `V4` are not all zero, which means `V1` can be explained by some nonzero linear function of `V2`, `V3` and `V4`. However, none of the coefficeints of `V2`, `V3` and `V4` is significant at level 0.05 (the p-values for t tests are 0.147, 0.665, 0.284, respectively), which suggests that there is a problem of colinearity and it is better to use remove some explanatory variable from the model. If we try to remove one explanatory variable from the regression model, we may consider one of the following three linear regression models:

```
(V1~V2+V3)
(V1~V2+V4)
(V1~V3+V4)
```

The multiple R-squared values for the above three models are 0.9626, 0.963 and 0.9622, respectively, so it is reasonable to choose the linear regression model with the largest multiple R-squared value, which is the model `(V1~V2+V4)` since `V1` can be slightly better explained by `V2` and `V4` than by `V1` and `V4` or by `V3` and `V4`.

48. We will apply chi-squared goodness of fit test to solve this problem. We first compute the probabilities that a $N(47, 40^2)$ variable falls in the 5 salary ranges.

$$\begin{aligned} P(N(47,40^2) < 25) &= P(N(0,1) < (25-47)/40) \\ &= P(N(0,1) < -0.55) \\ &= 0.5 - 0.2088 = 0.2912. \end{aligned}$$

$$\begin{aligned} P(25 < N(47,40^2) < 50) &= P((25-47)/40 < N(0,1) < (50-47)/40) \\ &= P(-0.55 < N(0,1) < 0.075) \\ &\approx P(-0.55 < N(0,1) < 0.08) \\ &= 0.2088 + 0.0319 = 0.2407. \end{aligned}$$

$$\begin{aligned} P(50 < N(47, 40^2) < 75) &= P((50-47)/40 < N(0,1) < (75-47)/40) \\ &= P(0.075 < N(0,1) < 0.7) \\ &\approx P(0.08 < N(0,1) < 0.7) \\ &= 0.2580 - 0.0319 = 0.2261. \end{aligned}$$

$$\begin{aligned} P(75 < N(47, 40^2) < 100) &= P((75-47)/40 < N(0,1) < (100-47)/40) \\ &= P(0.7 < N(0,1) < 1.325) \\ &\approx P(0.7 < N(0,1) < 1.33) \\ &= 0.4082 - 0.2580 = 0.1502. \end{aligned}$$

$$\begin{aligned} P(N(47, 40^2) > 100) &= P(N(0,1) > 1.325) \\ &\approx P(N(0,1) > 1.33) \\ &= 0.5 - 0.4082 = 0.0918. \end{aligned}$$

The test statistic for the chi-squared goodness of fit test is

$$\begin{aligned} &\frac{(90 - 1000 \cdot 0.2912)^2}{1000 \cdot 0.2912} + \frac{(450 - 1000 \cdot 0.2407)^2}{1000 \cdot 0.2407} \\ &+ \frac{(190 - 1000 \cdot 0.2261)^2}{1000 \cdot 0.2261} + \frac{(140 - 1000 \cdot 0.1502)^2}{1000 \cdot 0.1502} \\ &+ \frac{(130 - 1000 \cdot 0.0918)^2}{1000 \cdot 0.0918} \\ &= 343.3646. \end{aligned}$$

The degree of freedom of the test statistic is $5 - 1 = 4$. The 0.99 quantile of $\chi^2(4)$ is $13.277 < 343.3646$, so based on the chi-squared goodness of fit test, at level 0.01, we can conclude that the distribution of the annal salary (in 10,000 NTD) is not $N(47, 40^2)$.

49. We will apply the chi-square test of independence to solve this problem. The row totals are

$$1758 + 5619 + 1697 = 9074 \text{ and } 273 + 565 + 88 = 926,$$

and the column totals are

$$1758 + 273 = 2031, 5619 + 565 = 6184, \text{ and } 1697 + 88 = 1785.$$

The test statistic is

$$\begin{aligned} &\frac{(1758 - 10000 \cdot (2031/10000) \cdot (9074/10000))^2}{10000 \cdot (2031/10000) \cdot (9074/10000)} \\ &+ \frac{(5619 - 10000 \cdot (6184/10000) \cdot (9074/10000))^2}{10000 \cdot (6184/10000) \cdot (9074/10000)} \\ &+ \frac{(1697 - 10000 \cdot (1785/10000) \cdot (9074/10000))^2}{10000 \cdot (1785/10000) \cdot (9074/10000)} \\ &+ \frac{(273 - 10000 \cdot (2031/10000) \cdot (926/10000))^2}{10000 \cdot (2031/10000) \cdot (926/10000)} \\ &+ \frac{(565 - 10000 \cdot (6184/10000) \cdot (926/10000))^2}{10000 \cdot (6184/10000) \cdot (926/10000)} \\ &+ \frac{(88 - 10000 \cdot (1785/10000) \cdot (926/10000))^2}{10000 \cdot (1785/10000) \cdot (926/10000)} \\ &= 82.20876. \end{aligned}$$

The degree of freedom of the test statistic is $(3-1)(2-1) = 2$. The 0.99 quantile of $\chi^2(2)$ is $9.210 < 82.20876$, so at level 0.01, we can conclude that the job function and the the annal salary range are associated.

50. We will apply the chi-square goodness of fit test to solve this problem. The test statistic is

$$\begin{aligned}&\frac{(1758-10000\cdot 0.2)^2}{10000\cdot 0.2}+\frac{(5619-10000\cdot 0.5)^2}{10000\cdot 0.5}+\frac{(1697-10000\cdot 0.2)^2}{10000\cdot 0.2}\\&+\frac{(273-10000\cdot 0.03)^2}{10000\cdot 0.03}+\frac{(565-10000\cdot 0.06)^2}{10000\cdot 0.06}+\frac{(88-10000\cdot 0.01)^2}{10000\cdot 0.01}\\&=157.7304.\end{aligned}$$

The degree of freedom of the test statistic is $3\cdot 2-1=5$. The 0.95 quantile of $\chi^2(5)$ is $11.070 < 157.7304$, so at level 0.05, we can conclude that the actual proportions are not all the same as the proposed proportions.

51. (a) Sort the sample data and we have $1.5 < 1.9 < 2 < 2.1$. The rank of 2 is 3.

    (b) Sort the sample data and we have $1.5 < 1.9 < 2 = 2 = 2$. The rank of 2 is $(3+4+5)/3 = 4$.

52. We will first establish the following result:

$$P(\mathcal{D}\geq C_\alpha)\leq\alpha. \tag{49}$$

To see that (49) holds, for $\alpha\in(0,1)$, let

$$S_\alpha=\{k:\ k \text{ is an integer and } P(\mathcal{D}\geq k)\leq\alpha\}.$$

Then, by definition, C_α is the largest integer in S_α[1], so $C_\alpha\in S_\alpha$, which implies (49).

To show that

$$C\geq C_\alpha\Leftrightarrow P(\mathcal{D}\geq C)\leq\alpha, \tag{50}$$

we will show that

$$C\geq C_\alpha\Rightarrow P(\mathcal{D}\geq C)\leq\alpha \tag{51}$$

and

$$P(\mathcal{D}\geq C)\leq\alpha\Rightarrow C\geq C_\alpha. \tag{52}$$

Suppose that C is an integer. Then

$$C\geq C_\alpha\quad\Rightarrow\quad P(\mathcal{D}\geq C)\leq P(\mathcal{D}\geq C_\alpha)\overset{(49)}{\leq}\alpha,$$

so (51) holds. Moreover, suppose that C is an integer such that

$$P(\mathcal{D}\geq C)\leq\alpha,$$

then $C\in S_\alpha$. Since C_α is the largest integer in S_α, we have $C_\alpha\geq C$ and (52) holds.

Since both (51) and (52) hold true, (50) holds.

[1]Note that athe existence of C_α is guaranteed since the set S_α is nonempty and has an upper bound. To see this, note that the set S_α is nonempty since $-1\in S_\alpha$. Moreover, the set S_α has an upper bound $(m-1)$ since for an integer $k\geq m$, k cannot be in S_α as $P(\mathcal{D}\geq k)=1>\alpha$.

53. Apply the result in Problem 52 with $m = n$ and $\mathcal{D} = Bin(n, p_0)$, then we have

$$\begin{aligned} &\text{the proposed test rejects } H_0 \text{ at level } \alpha \\ \Leftrightarrow\ &\text{observed } X \geq C_\alpha \\ \Leftrightarrow\ &P(Bin(n, p_0) \geq \text{ observed } X) \leq \alpha, \end{aligned}$$

so $P(Bin(n, p_0) \geq$ observed $X)$ is the p-value of the proposed test.