Fall 2019 Applied Statistics Comprehensive Examination Solutions

1. (23 points) A recent study in Yellowstone National Park aimed to compare the proportion of black and brown bear cubs surviving to at least one year in age. Of the 181 black bear cubs studied, 150 survived to a year, while out of the 42 brown bear cubs studied, 32 survived.

   (a) (10 points) Create a 95% confidence interval for the difference in the proportion of each species of bear cub surviving to one year.

   Solution:
   (a) 95% CI: $(\frac{150}{181} - \frac{32}{42}) \pm 1.96\sqrt{\frac{\frac{150}{181}\left(1-\frac{150}{181}\right)}{181} + \frac{\frac{32}{42}\left(1-\frac{32}{42}\right)}{42}} = (-0.073, 0.207)$

   We are 95% confident that the difference between the proportions of black bear cubs and brown bear cubs surviving to at least a year (black−brown) is between -0.073 and 0.207.

   (b) (3 points) For inference for two proportions, we must assume that the sample sizes are "large enough." Briefly explain the rationale behind one common check of this assumption, $n_1\widehat{p}_1 \geq 5$, $n_1(1-\widehat{p}_1) \geq 5$, $n_2\widehat{p}_2 \geq 5$, and $n_2(1-\widehat{p}_2) \geq 5$.

   Solution:
   We need large enough samples so that the sample proportions are approximately normally distributed (implying that the difference in sample proportions is also approximately normal) so that the z-stat will be approximately N(0,1) under H_0. The sample size checks for this assumption of "large enough samples" depend not only on the number of observations but also on the estimated proportions. This is due to the fact that when the true proportions for each population are close to 0 or 1, larger sample sizes n_1 and n_2 are needed so that the distributions of the sample proportions are not skewed right or left. Simulations have established that requiring $n_1\widehat{p}_1 \geq 5$, $n_1(1-\widehat{p}_1) \geq 5$, $n_2\widehat{p}_2 \geq 5$, and $n_2(1-\widehat{p}_2) \geq 5$ will tend to lead to correct coverage probabilities and Type I error rates for the confidence interval and hypothesis test procedures, respectively.

   (c) (10 points) Suppose that the main goal of the study was actually to test at the 0.05 level whether the one-year survival rate for black bear cubs is different than the one-year survival rate for brown bear cubs. Use the confidence interval to draw a conclusion for the test of interest. Make sure to do each of the following:

      i. write the hypotheses of interest symbolically, defining symbols as is necessary.

ii. briefly explain what the conclusion is and how the confidence interval helped you reach that conclusion.

iii. interpret the conclusion in terms of the problem.

Solution:

We want to test $H_0 : \pi_{black} = \pi_{brown}$ vs. $H_0 : \pi_{black} \neq \pi_{brown}$ where π_{black} is the true proportion of black bear cubs surviving to a year and π_{brown} is the true proportion of brown bear cubs surviving to a year (answer to part i).

Since 0 (the hypothesized difference in proportions under H_0) is included in the confidence interval, we would fail to reject H_0 (answer to part ii).

That is, we do not have enough evidence at the 0.05 level to conclude that there is a difference in the proportion of black and brown bear cubs surviving to a year (answer to part iii).

2. (30 points) Four local amateur golfers participated in a long drive contest where each golfer hit four drives. The drive lengths (in yards) and the sample means and sample standard deviations are given in the table below. You are interested in comparing the mean drive lengths for the four golfers.

Golfer	Drive Lengths (Yards)				$\bar{x}$	s
A	248	125	193	229	199	54
B	271	251	247	225	249	19
C	163	149	177	189	170	17
D	253	252	284	301	273	24

(a) (10 points) Using level 0.05, test for evidence that the mean drive length differs from one golfer to another.

Solution:

Let μ_A, μ_B, μ_C, and μ_D be the mean drive lengths of golfers A, B, C, and D, respectively. Then, we want to test

$H_0 : \mu_A = \mu_B = \mu_C = \mu_D$ vs.
$H_a : \mu_A$, μ_B, μ_C, and μ_D are not all the same.

$$\begin{aligned} SSE &= (4-1)(s_A^2 + s_B^2 + s_C^2 + s_D^2) \\ &= 3(54^2 + 19^2 + 17^2 + 24^2) \\ &= 3(4142) = 12,426 \end{aligned}$$

$$\begin{aligned} SST &= 4\sum_i^4 (\overline{x}_{i\cdot} - \overline{x}_{\cdot\cdot})^2 \\ &= 4\left\{(199-222.8)^2 + (249-222.8)^2 + (170-222.8)^2 + (273-222.8)^2\right\} \\ &\approx 26,243 \end{aligned}$$

Note: Using raw data, you should get $SST \approx 26,196$ and $SSE \approx 12,524$. These numbers are slightly different than the SST and SSE based on the mean and standard deviation summary statistics due to rounding (in the drive lengths and/or the summary statistics).

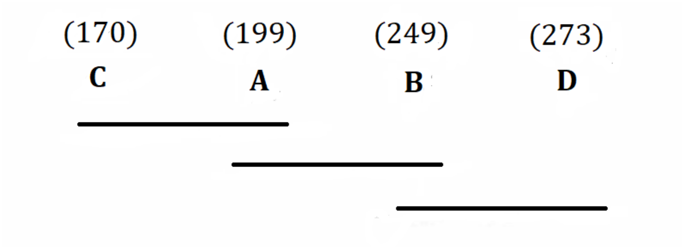
F-Stat $= \frac{26{,}243/(4-1)}{12{,}426/(16-4)} \approx 8.4$.

Then, since $F_{3,12,0.05} = 3.49 < 8.4$, we have enough evidence at the 0.05 level to conclude that the mean drive lengths differ across the golfers.

(b) (10 points) Apply Tukey's method at level 0.05 to compare the mean drive lengths for the four golfers. Present your conclusions either in words or with an appropriate diagram.

Solution:

From the table, $q_{4,12,0.05} = 4.20$, and thus the least significant difference is $\frac{4.20}{\sqrt{2}}\sqrt{\frac{12{,}426}{12}\left(\frac{1}{4}+\frac{1}{4}\right)} \approx 67.6$. This leads to the following underline diagram



(c) (10 points) List all the assumptions required for the inference procedures that you applied in parts (a) and (b). For each assumption that you list, provide one way to check the assumption. Please be specific. It is not necessary that you actually do the assumption-checking.

Solution:

i. Independent errors. We could check this by plotting the residuals against the order in which the drives were hit. A pattern in the plot might indicate a lack of independence.

ii. Normal errors. We could make a histogram or normal probability plot of the residuals either overall or by treatment.

iii. Equal variances. We could compare the sample variances. If the ratio between the biggest and smallest exceeds 3 or so, then the population variances may be sufficiently different to lead to an inflated α level.

3. (20 points) In the 1980-1981 and 1981-1982 NBA basketball seasons, Larry Bird shot 338 pairs of free throws from the foul line (from *Wardrop 1995, The American Statistician*). The observed distribution of the number of free throws (out of two) that he made each time at the foul line is given by the following table:

Number of Free Throws Made	0	1	2
Number of Occasions	5	82	251

(a) (15 points) If all of the free throws Larry Bird took over his career were independent with the same probability of success, we would expect the number of free throws made out of two to follow a binomial distribution with success probability 0.89 (Larry Bird made 89% of his career free throws). Conduct a 0.05 level goodness-of-fit test to decide whether the binomial distribution with $\pi = 0.89$ is reasonable for these two seasons.

Solution:
First, note that if the Y =number of free throws out of two is binomial with $\pi = 0.89$ then,

$$\begin{aligned}\pi_{00} = P(Y=0) &= 0.11^2 = 0.0121\\ \pi_{10} = P(Y=1) &= 2(0.89)(0.11) = 0.1958\\ \pi_{20} = P(Y=2) &= 0.89^2 = 0.7921\end{aligned}$$

Then, we want to test

$H_0 : \pi_0 = 0.0121, \pi_1 = 0.1958, \pi_2 = 0.7921$ vs.
$H_a : \pi_i \neq \pi_{io}$ for some $i \in \{0, 1, 2\}$,

where π_i is the true probability of making i out of the 2 free throws

The expected values for each of the three cells are $338(0.0121) = 4.090$, $338(0.1958) = 66.180$, and $338(0.7921) = 267.730$.

Then, χ^2-stat$= \frac{(5-4.090)^2}{4.090} + \frac{(82-66.180)^2}{66.180} + \frac{(251-267.730)^2}{267.730} = 5.030$

Rejection/Critical Region: χ^2-stat$> \chi^2_{2,0.05} = 5.991$

Since the test statistic is not in the rejection region, we do not have enough evidence at the 0.05 level to conclude that the distribution of the number of free throws made by Larry Bird out of 2 is different than a binomial with success probability 0.89.

(b) (5 points) What assumptions are required for the test in part a? Comment on whether you have any concerns about the validity of these assumptions.

Solution:

We need to assume that our sample sizes are large enough. I would have some concern about this assumption since the expected count for the 0 case is less than 5. The rule of thumb is to have all expected counts above 1 and no more than 20% below 5. Since the expected value 4.090 is only just below five, the validity of the test may still be reasonable (note: simulations show this to be the case). Alternatively, one may combine the 0 and 1 cells before conducting the goodness-of-fit test.

We also need to assume that we have collected 338 pairs of free throws that are independent from one another. This assumption may or may not be reasonable (injuries, hot hand, etc.) but is fine if we are willing to make the assumption as stated in the problem that each and every individual free throw is independent. Of course, if we are willing to assume this, then it may make more sense to do a one sample proportion test to see if the overall proportion of free throws made in the 1980-1981 and 1981-1982 seasons differs from 0.89.

4. (25 points) A no-intercept simple linear regression model (or a regression model through the origin), that is $y_i = \beta x_i + \epsilon_i$, $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, $i = 1, 2, ..., n$, is often appropriate in analyzing data from chemical and other manufacturing processes.

   a. (10 points) Show that the least-squares estimator for the slope β is $\hat{\beta} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$.

   Solution:

$$\begin{aligned} S(\beta) &= \sum_{i=1}^n (y_i - \beta x_i)^2 \\ \frac{dS(\beta)}{d\beta} &= 2\sum_{i=1}^n (y_i - \beta x_i)(-x_i) = -2\left(\sum_{i=1}^n x_i y_i - \beta \sum_{i=1}^n x_i^2\right) \overset{\text{set}}{=} 0 \\ \hat{\beta} &= \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}. \end{aligned}$$

   b. (7 points) Derive the mean and variance of $\hat{\beta}$.

   Solution:

$$E(\hat{\beta}) = E\left(\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}\right) = \frac{\sum_{i=1}^n x_i E(y_i)}{\sum_{i=1}^n x_i^2} = \frac{\sum_{i=1}^n x_i (x_i \beta)}{\sum_{i=1}^n x_i^2} = \beta$$

$$V(\hat{\beta}) = V\left(\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}\right) = \frac{\sum_{i=1}^n x_i^2 V(y_i)}{\left(\sum_{i=1}^n x_i^2\right)^2} = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{\left(\sum_{i=1}^n x_i^2\right)^2} = \frac{\sigma^2}{\sum_{i=1}^n x_i^2}$$

   c. (3 points) Since two points determine a straight line, a student proposes another estimator for β that uses only the first observation in the dataset, i.e., (x_1, y_1), and the origin, i.e., $(0, 0)$, as $\tilde{\beta} = \frac{y_1 - 0}{x_1 - 0} = \frac{y_1}{x_1}$. Is $\tilde{\beta}$ an unbiased estimator for β? Show your work.

   Solution: $\tilde{\beta}$ is an unbiased estimator for β.

$$E(\tilde{\beta}) = E\left(\frac{y_1}{x_1}\right) = \frac{\beta x_1}{x_1} = \beta.$$

   d. (5 points) Which estimator is better, $\hat{\beta}$ or $\tilde{\beta}$? Explain.

   Solution: I conclude that $\hat{\beta}$ is better because both estimators are unbiased for estimating β, but $\hat{\beta}$ is more efficient with a smaller variance.

$$V(\tilde{\beta}) = V\left(\frac{y_1}{x_1}\right) = \frac{\sigma^2}{x_1^2} \geq \frac{\sigma^2}{\sum_{i=1}^n x_i^2},$$

where the equality only holds when $x_2, x_3, ..., x_n$ are all 0 which is not a realistic scenario. Comments such as "$\hat{\beta}$ is more efficient because it uses all data while $\tilde{\beta}$ only uses the first observation" would be worth partial credit.

5. (37 points) From a random sample of all domestic flights departing from Newark Liberty International Airport (EWR) and John F. Kennedy International Airport (JFK) in the year 2013, the following summary statistics were observed for the departure delay in (minutes):

	Mean	Standard Deviation	n
EWR	18.5	41.7	178
JFK	20.3	44.6	153

(a) (10 points) Conduct an appropriate statistical test to determine if there is evidence that the variances in departure delay time from each airport are unequal, at $\alpha = 0.05$. Note: since the provided distribution table does not actually contain the proper degrees of freedom based on the data provided, please use the closest that you can.

Solution:
$H_0 : \sigma^2_{EWR} = \sigma^2_{JFK}$ vs. $H_a : \sigma^2_{EWR} \neq \sigma^2_{JFK}$,
where σ^2_{EWR} is the variance in flight delays at EWR airport and σ^2_{JFK} is the variance in flight delays at JFK airport.

F-stat $= \frac{44.6^2}{41.7^2} = 1.144$

While the rejection region is two-sided, since I chose to put the larger sample variance in the numerator, we would reject H_0 if F-stat$> F_{152,177,0.025} \approx F_{120,120,0.025} = 1.433$. Alternatively, if one were to choose to put the smaller sample variance in the numerator, we would then reject if F-stat$< F_{152,177,0.975} = 1/F_{177,152,0.025} \approx 1/F_{120,120,0.025} = 1/1.433 = 0.698$

Since our F-stat is not in the rejection region, we do not have enough evidence at the 0.05 level to conclude that the variance in departure delays is different at the EWR and JFK airports.

(b) (8 points) Since the desired degrees of freedom for the test in part (a) are not in the table, the Type I and Type II Error rates will be different than the ones resulting from the decisions based on the proper degrees of freedom. Answer each of the following:

i. Using the provided distribution table, the Type I Error rate will be:

lower than the same as higher than

that of the decisions made with the proper degrees of freedom (choose one, and explain).

ii. Using the provided distribution table, the Type II Error rate will be:

lower than   the same as   higher than

that of the decisions made with the proper degrees of freedom (choose one, and explain).

Solution:

The Type I Error rate is lower than if we had used the correct degrees of freedom. This is because the F upper percentiles decrease as df increases in both numerator and denominator. Thus, when H_0 is true, we would not reject H_0 in certain cases when the F-stat is actually above the true rejection region cut-off if we are using the incorrect df.

The Type II Error rate would be higher than if we had used the correct degrees of freedom. Again, since the F upper percentiles decrease as df increases, using artificially low df makes the cut-off to reject H_0 too high. Thus, when H_a is true, we would not reject H_0 in certain cases when the F-stat is above the true rejection region cut-off. This means that the chances of making a Type II Error are inflated (and power is decreased).

(c) (10 points) Based on your answer to part (a), conduct an appropriate statistical test to determine if there is evidence that the mean departure delays from each airport are different, at $\alpha = 0.05$. If a required degrees of freedom is not exactly available, please use the closest that you can.

Solution:

Using the equal variance assumption, we want to test $H_0 : \mu_{EWR} = \mu_{JFK}$ vs. $H_a : \mu_{EWR} \neq \mu_{JFK}$, where μ_{EWR} (μ_{JFK}) is the mean flight delay at EWR (JFK) airport.

$s_p = \sqrt{\frac{177(41.7^2)+152(44.6)^2}{329}} = 43.064$ so that t-stat$= \frac{20.3-18.5}{43.196\sqrt{\frac{1}{178}+\frac{1}{153}}} = 0.379$

RR: |t-stat|$> t_{329,0.025} \approx t_{120,0.025} = 1.98$

Since the t-stat is not in the rejection region, we do not have enough evidence at the 0.05 level to conclude that there is a difference in the mean departure delays at EWR and JRK airports.

(d) (6 points) Based on the information given, do either of the tests performed in parts (a) or (c) require an assumption that departure delays follow a normal distribution? Why or why not?

Solution:

The test for unequal variances requires normally distributed departure times. This is because the F-statistic is derived under this assumption (as sample sizes get larger, the F-stat will not necessarily become close to a true F-distribution if the departure times are non-normal). The test for unequal means does not require a normality assumption to work well, just a large enough sample such that the distributions of the sample mean will be normally distributed (so that the t-stat t_{df} under H_0).

Note: since departure delays are non-negative and the sample standard deviations given in the table are well greater than the means, the normality assumption is not reasonable for the test of unequal variances. It is also possible that the data are so skewed that the two-sample t-test may have inflated Type I Error rates as well, but this is less likely given the large sample sizes.

(e) (3 points) List any other required assumptions for each of the tests in parts (a) and (c), and comment briefly on whether they appear to be reasonable here.

Solution:

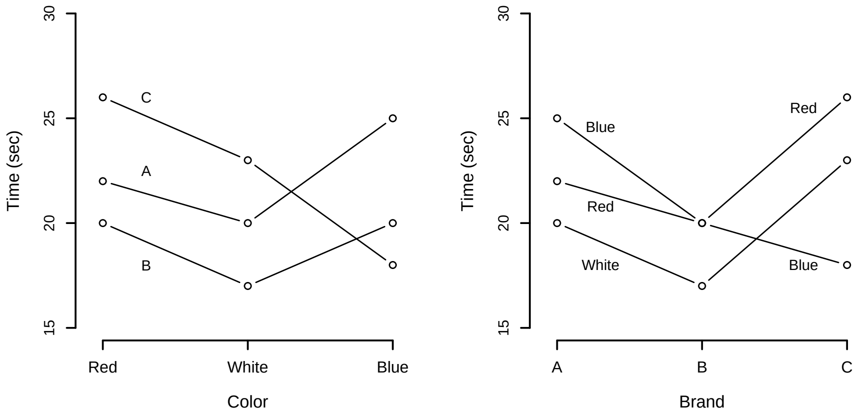
For both tests in part (a) and (c), we need the samples to be independently collected and randomly sampled from the populations of flight delays. We are told that the samples are random which should imply in this context independence between the samples. For the test in (c), we additionally assumed that the variances are equal. The test for unequal variances probably doesn't help much here as the samples of flight delays are very skewed. However, with the sample sizes being relatively similar, there is likely no issue with using the equal variance test.

6. (30 points) A student ran an experiment to study how the average time required for her to blow up a balloon varied with the color of the balloon (Red, White, or Blue) and the brand (A, B, or C). She blew up three balloons for each combination of a color and a brand, recording the time required (in seconds). The treatment sample means are given in the table below.

	Brand		
Color	A	B	C
Red	22	20	26
White	20	17	23
Blue	25	20	18

(a) (10 points) Create appropriate interaction plots, and comment on both the presence/absence of interaction and the type of interaction.

Solution:



Since there is crossing in the plots, it appears like there may be disorderly interaction.

(b) (10 points) Write down a complete set of orthogonal interaction contrasts.

Solution: Answers may vary. The following is one potential strategy. Start with an interaction contrast only involving the upper left 2 by 2 portion of the table. The coefficients for an interaction contrast can be easily found by multiplying the coefficients on the marginal simple effects of A vs. B and Red vs. White (in diagram below, coefficients to the left and above the lines represent the marginal effects while the coefficients inside the lines represent the coefficients for the interaction contrast).

		A	B	C
		1	-1	0
Red	1	1	-1	0
White	-1	-1	1	0
Blue	0	0	0	0

The associated interaction contrast is $\mu_{Red,A} - \mu_{Red,B} - \mu_{White,A} + \mu_{White,B}$.

Next, keep one of the marginal simple effects the same while changing the other to a new marginal comparison of means that on its own will be orthogonal to the first, such as the average of A and B vs. C. Then cross-multiply the coefficients to get a new interaction contrast:

		A	B	C
		1	1	-2
Red	1	1	1	-2
White	-1	-1	-1	2
Blue	0	0	0	0

The associated interaction contrast is $\mu_{Red,A} - \mu_{Red,B} - 2\mu_{Red,C} - \mu_{White,A} - \mu_{White,B} + \mu_{White,C}$.

Similarly, going back to the original A vs. B but now considering the average of Red and White vs. Blue:

		A	B	C
		1	-1	0
Red	1	1	-1	0
White	1	1	-1	0
Blue	-2	-2	2	0

The associated interaction contrast is $\mu_{Red,A} - \mu_{Red,B} + \mu_{White,A} - \mu_{White,B} - 2\mu_{Blue,A} + 2\mu_{Blue,B}$.

Finally, use both of the more complicated marginal comparisons and cross-multiply to get the coefficients for the final interaction contrast:

		A	B	C
		1	1	-2
Red	1	1	1	-2
White	1	1	1	-2
Blue	-2	-2	-2	4

The associated interaction contrast is $\mu_{Red,A} + \mu_{Red,B} - 2\mu_{Red,C} + \mu_{White,A} + \mu_{White,B} - 2\mu_{White,C} - 2\mu_{Blue,A} - 2\mu_{Blue,B} + 4\mu_{Blue,C}$.

Putting everything together, this gives a set of orthogonal contrasts on

$$(\mu_{Red,A}, \mu_{Red,B}, \mu_{Red,C}, \mu_{White,A}, \mu_{White,B}, \mu_{White,C}, \mu_{Blue,A}, \mu_{Blue,B}, \mu_{Blue,C})$$

with coefficient matrix

$$\begin{bmatrix} 1 & -1 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & -2 & -1 & -1 & 2 & 0 & 0 & 0 \\ 1 & -1 & 0 & 1 & -1 & 0 & -2 & 2 & 0 \\ 1 & 1 & -2 & 1 & 1 & -2 & -2 & -2 & 4 \end{bmatrix}$$

(c) (10 points) Suppose that the SSE for an effects model with interaction is 72. Using level 0.05, test for an interaction involving colors White and Blue and brands B and C.

Solution:
$H_0 : \mu_{White,B} - \mu_{White,C} - \mu_{Blue,B} + \mu_{Blue,C} = 0$ vs.
$H_a : \mu_{White,B} - \mu_{White,C} - \mu_{Blue,B} + \mu_{Blue,C} \neq 0$

MSE $= \frac{72}{9(3-1)} = 4$ (18 df for error)

$t = \frac{17-23-20+18}{\sqrt{4(\frac{1^2}{3}+\frac{(-1)^2}{3}+\frac{(-1)^2}{3}+\frac{1^2}{3}}} = \frac{-8}{2.309} \approx -3.5$

Then, since $|-3.5| > t_{18,0.025} = 2.10$, we reject H_0 and conclude at the 0.05 level that there is interaction in the effect on time between the colors white and blue and the brands B and C.

7. (35 points) Researchers from the Netherlands were interested in whether smoking by the mother in pregnancy is related to higher childhood blood pressure in their offspring. In a recent study, they recruited 200 mothers and their newborns (one newborn per mother). They considered the following regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon,\ \epsilon \overset{iid}{\sim} N(0, \sigma^2),$$

to relate the infant's systolic blood pressure in mm Hg (y) to the infant's age in weeks (x_1) and weight in kg (x_2), and the mother's smoking status in pregnancy (no smoking, passive exposure to smoking, or smoking). Mothers who did not smoke in pregnancy but were exposed to smoke by others were considered as having "passive exposure to smoking". Two indicator variables were defined to account for mother's smoking status:

$$x_3 = \begin{cases} 1, & \text{passive exposure to smoking,} \\ 0, & \text{otherwise;} \end{cases} \qquad x_4 = \begin{cases} 1, & \text{smoking,} \\ 0, & \text{otherwise.} \end{cases}$$

The R output of the regression analysis for this full model is presented in Table 1. Table 2 presents the ANOVA tables for the full model, the reduced model 1 (only including the independent variables x_1 and x_2), and the reduced model 2 (only including the independent variables x_3 and x_4). Figure 1 shows the residual plots for the full model.

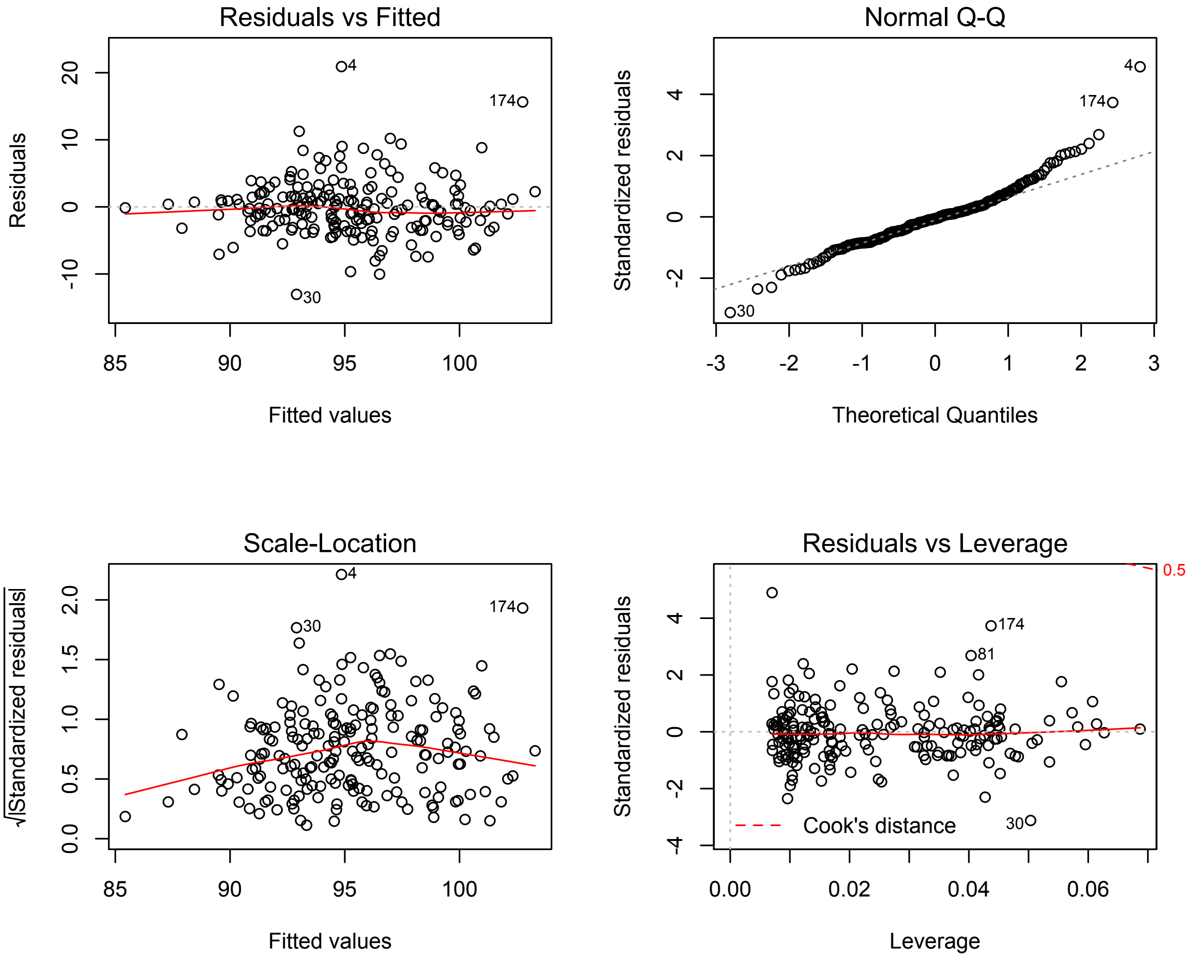
Table 1: R output of regression analysis for the full model.

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4, data = sdata)

Residuals:
   Min       1Q     Median     3Q      Max
-13.0408  -2.6090  -0.2574   1.6563  20.9125

Coefficients:
               Estimate   Std.Error  t value    Pr(>|t|)
(Intercept)     72.0164     2.5112    28.678     < 2e-16 ***
   x1            1.0006     0.2393     4.181    4.38e-05 ***
   x2            4.0882     0.5007     8.166    3.92e-14 ***
   x3            0.2803     0.9379     0.299       0.765
   x4            5.6025     0.8522     6.574    4.36e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 4.286 on 195 degrees of freedom
F-statistic: 30.08 on 4 and 195 DF,  p-value: < 2.2e-16
```

Table 2: ANOVA tables for the full model (including x_1, x_2, x_3, and x_4), the reduced model 1 (including x_1 and x_2), and the reduced model 2 (including x_3 and x_4).

Models		Degrees of Freedom	Sum of Squares	Mean Squares
Full Model	Regression	4	2210.1	552.53
	Error	195	3581.4	18.37
Reduced Model 1	Regression	2	1403.5	701.75
	Error	197	4388.0	22.27
Reduced Model 2	Regression	2	699.3	349.65
	Error	197	5092.2	25.85

Figure 1: Residual plots for the full model.



a. (5 points) Interpret the estimated coefficient associated with the variable age (x_1) in the full model.

Solution: Each one week increase in the infant's age is associated with an estimated mean systolic blood pressure increase of 1.006 mm Hg, given that all other variables are held constant.

b. (5 points) Calculate and interpret the coefficient of determination, R^2, for the full model.

Solution:

$$R^2 = \frac{SSR}{SST} = \frac{2210.1}{2210.1 + 3581.4} = 0.382$$

This means that 38.2% of the variability in the infant's systolic blood pressure can be explained by a linear relationship with the infant's age, weight, and the mother's smoking status.

c. (10 points) Use appropriate sums of squares or mean squares in Table 2 to test whether the infant's systolic blood pressure is associated with the mother's smoking status given that the age and weight of the infant are in the model at the significance level of 0.05.

Solution:

$H_0 : \beta_3 = \beta_4 = 0$

H_a: at least one of the regression coefficients is not 0.

$$F = \frac{(\text{SSR}_{\text{full}} - \text{SSR}_{\text{reduced}})/df}{\text{MSE}_{\text{full}}} = \frac{(2210.1 - 1403.5)/2}{18.37} = 21.95 \overset{H_0}{\sim} F_{2,195}$$

Since $21.95 > F_{0.05,2,195} \approx 3.07$, reject H_0. There is sufficient evidence at the 0.05 level to claim that the infant's systolic blood pressure is associated with the mother's smoking status after adjusting for the infant's age and weight.

d. (9 points) List all model assumptions. Based on the residual plots for the full model in Figure 1, comment on whether the model assumptions appear to be satisfied. Provide a way to address each assumption violation if there are any.

Solution:

Assumption	**Violation**	**Solution**
Linearity	No obvious violation as the residuals seem to be scattered randomly about 0 for all predicted values of the response	
Constant variance	This assumption may not quite be met, though since most of the data is between the predicted values of 90 and 100, we would expect to see observations further from 0 in this range. That said, there appear to exist outliers, such as observations 4, 30, and 174, and one could argue that the variances at the center of the fitted values tend to be slightly greater than the two ends.	weighted least squares (transformation on y may work in some cases of changing variance)
Normality	There is a clear violation of normality. The tails of the distribution tend to be heavier than those of normal distributions.	Robust regression or (transformation on y may work in some cases)
Independence	Whether or not independence is reasonable cannot be diagnosed from Figure 1. No clear violation is found from the study description.	

Additionally, note that while influential observations are not strictly part of the model assumptions and thus discussion of influential points are not required as part of a complete answer here, it is still good practice to consider their effect. Based on the "Residual vs Leverage" plot, we see that observation 30 may be identified as an influential observation. However, according to Cook's distance, none of the observations are influential.

e. (6 points) Later on, one researcher suspected that the association between the infant's systolic blood pressure and the age might be different across different smoking statuses. Propose a model that allows the slope of age to differ among the three smoking statuses and write down the null and alternative hypotheses that would be used to test this researcher's idea.

Solution: The proposed model is

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_1 x_3 + \beta_6 x_1 x_4 + \epsilon,\ \epsilon \overset{iid}{\sim} N(0, \sigma^2)$$

The tested hypotheses are $H_0 : \beta_5 = \beta_6 = 0$ v.s. $H_a : \beta_5$ or $\beta_6 \neq 0$.