**Name: ERIC AGYEMANG**

HOMEWORK 2

3.(a). Correct option is (iii).For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.

The least squares regression line is

Y = 50 + 20GPA + 0.07IQ + 37Gender + 0.001(GPA\*IQ) – 10(GPA\*Gender)

= (35 – 10GPA)Gender

(b).

= 85 + 20(4) + 0.07(110) + 0.01(4\*110)-10(4\*1)

= 85 + 80 + 7.7 + 4.4 – 40

= 137.1

(c). False – The size of the coefficient for the interaction term doesn’t necessarily imply little evidence of an interaction effect. The P value will help us determine significance of the term in the model, and the size of the coefficients of the GPA and IQ main effects will give us a relative scale of which we will see the actual effects of the interaction.

10.(a)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 13.043469 0.651012 20.036 < 2e-16 \*\*\*

Price -0.054459 0.005242 -10.389 < 2e-16 \*\*\*

UrbanYes -0.021916 0.271650 -0.081 0.936

USYes 1.200573 0.259042 4.635 4.86e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.472 on 396 degrees of freedom

Multiple R-squared: 0.2393, Adjusted R-squared: 0.2335

F-statistic: 41.52 on 3 and 396 DF, p-value: < 2.2e-16

(b)

* The intercept is the baseline in sales, in this particular case this includes non-urban and non-US people
* Price variable may be interpreted by saying the average effect of price increase of a dollar is a decrease in unit sales all other predictors remained fixed
* UrbanYes: the extra effect that being in an urban area has on sales vs the baseline of being in rural area. This value is not predicted to be a statistically significant
* USYes: the added effect that being in the US has on sales. This is predicted to be significant, and represents an extra unit of sales

(c) Sales=13.0434689+(−0.0544588) ×Price+(−0.0219162) ×Urban+(1.2005727) ×US+ ε

*with*Urban=1*if the store is in an urban location and*0*if not, and*US=1*if the store is in the US and*0*if not.*

(d) We can reject the null hypothesis for the “price” and “US” variables

(e)

Call:

lm (formula = Sales ~ Price + US, data = Carseats)

Residuals:

Min 1Q Median 3Q Max

-6.9269 -1.6286 -0.0574 1.5766 7.0515

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 13.03079 0.63098 20.652 < 2e-16 \*\*\*

Price -0.05448 0.00523 -10.416 < 2e-16 \*\*\*

USYes 1.19964 0.25846 4.641 4.71e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.469 on 397 degrees of freedom

Multiple R-squared: 0.2393, Adjusted R-squared: 0.2354

F-statistic: 62.43 on 2 and 397 DF, p-value: < 2.2e-16

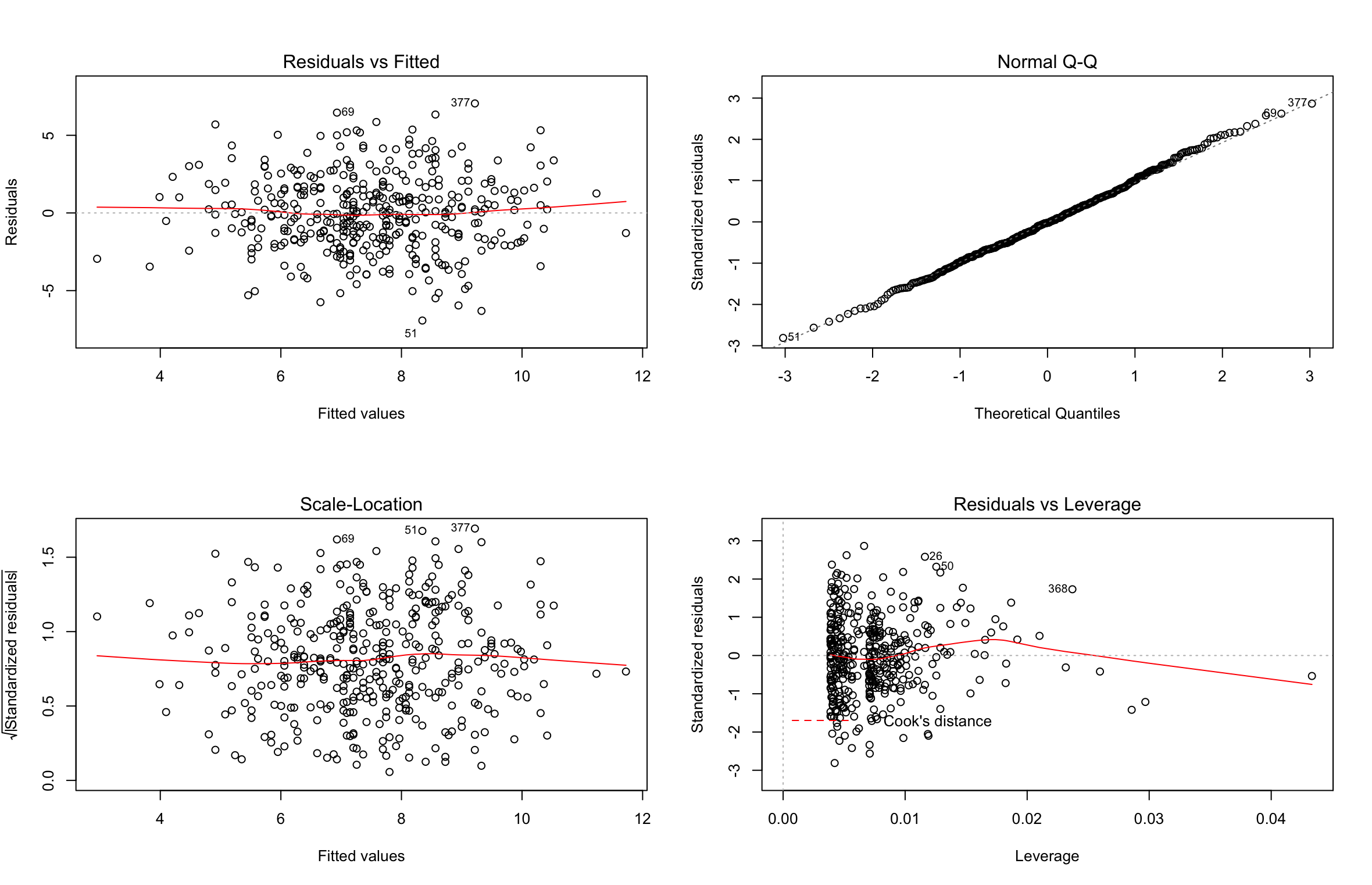
(f) The second model in part e fits better. The adjusted is better, and the p value is better. With fewer degrees of freedom it has a very similar unadjusted even, which is less prone to overfitting and likely more robust.

(g) 2.5 % 97.5 %

(Intercept) 11.79032020 14.27126531

Price -0.06475984 0.04419543

USYes 0.69151957 1.70776632

(h) 

**INTERPRETATION,** the points do not seem to have a particularly high leverage, there are few outliers just above a residual of 3.Also, the plot standardized residuals versus leverage indicate the presence of some outliers

14.(a)

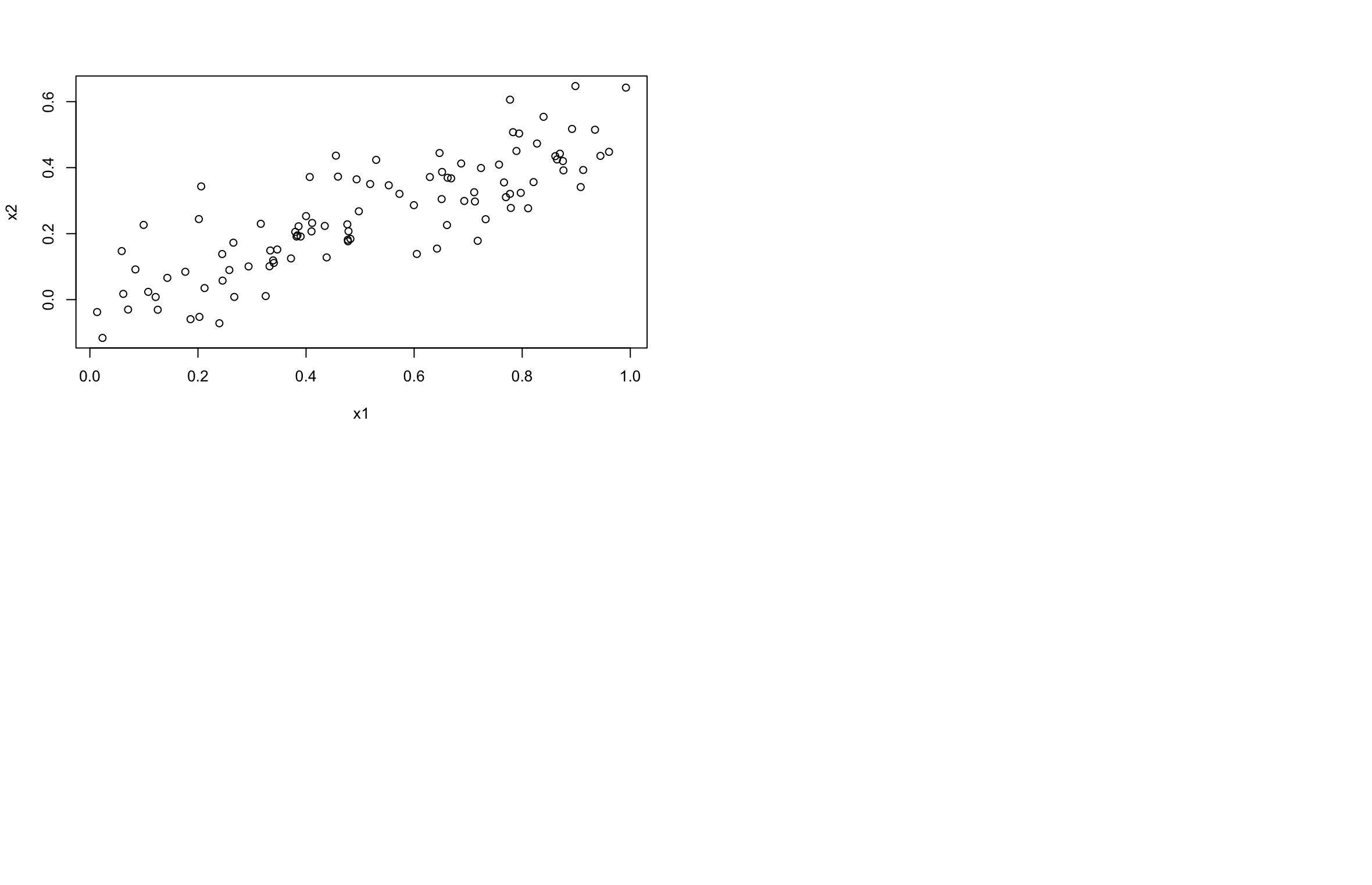
The form of the linear model is

Y=2+2X1+0.3X2+ε

with ε a N (0,1) random variable. The regression coefficients are respectively 2, 2 and 0.3.

(b). i. [1] 0.8351212

ii.



(c) Residuals:

Min 1Q Median 3Q Max

-2.8311 -0.7273 -0.0537 0.6338 2.3359

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.1305 0.2319 9.188 7.61e-15 \*\*\*

x1 1.4396 0.7212 1.996 0.0487 \*

x2 1.0097 1.1337 0.891 0.3754

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.056 on 97 degrees of freedom

Multiple R-squared: 0.2088, Adjusted R-squared: 0.1925

F-statistic: 12.8 on 2 and 97 DF, p-value: 1.164e-05

(d) Residuals:

Min 1Q Median 3Q Max

-2.89495 -0.66874 -0.07785 0.59221 2.45560

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.1124 0.2307 9.155 8.27e-15 \*\*\*

x1 1.9759 0.3963 4.986 2.66e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.055 on 98 degrees of freedom

Multiple R-squared: 0.2024, Adjusted R-squared: 0.1942

F-statistic: 24.86 on 1 and 98 DF, p-value: 2.661e-06

(e) Residuals:

Min 1Q Median 3Q Max

-2.62687 -0.75156 -0.03598 0.72383 2.44890

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.3899 0.1949 12.26 < 2e-16 \*\*\*

x2 2.8996 0.6330 4.58 1.37e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.072 on 98 degrees of freedom

Multiple R-squared: 0.1763, Adjusted R-squared: 0.1679

F-statistic: 20.98 on 1 and 98 DF, p-value: 1.366e-05

(f) They are not contradictory. There is high interaction between X1 and X2, so it is hard for the linear model to tease those interactions apart. When we feed in either X1 or X2 by itself, the interaction is clear between those and Y so a fit is possible.

(g) i.

Residuals:

Min 1Q Median 3Q Max

-2.8311 -0.7273 -0.0537 0.6338 2.3359

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.1305 0.2319 9.188 7.61e-15 \*\*\*

x1 1.4396 0.7212 1.996 0.0487 \*

x2 1.0097 1.1337 0.891 0.3754

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.056 on 97 degrees of freedom

Multiple R-squared: 0.2088, Adjusted R-squared: 0.1925

F-statistic: 12.8 on 2 and 97 DF, p-value: 1.164e-05

ii.

Residuals:

Min 1Q Median 3Q Max

-2.89495 -0.66874 -0.07785 0.59221 2.45560

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.1124 0.2307 9.155 8.27e-15 \*\*\*

x1 1.9759 0.3963 4.986 2.66e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.055 on 98 degrees of freedom

Multiple R-squared: 0.2024, Adjusted R-squared: 0.1942

F-statistic: 24.86 on 1 and 98 DF, p-value: 2.661e-06

iii. Residuals:

Min 1Q Median 3Q Max

-2.62687 -0.75156 -0.03598 0.72383 2.44890

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.3899 0.1949 12.26 < 2e-16 \*\*\*

x2 2.8996 0.6330 4.58 1.37e-05 \*\*\*

---

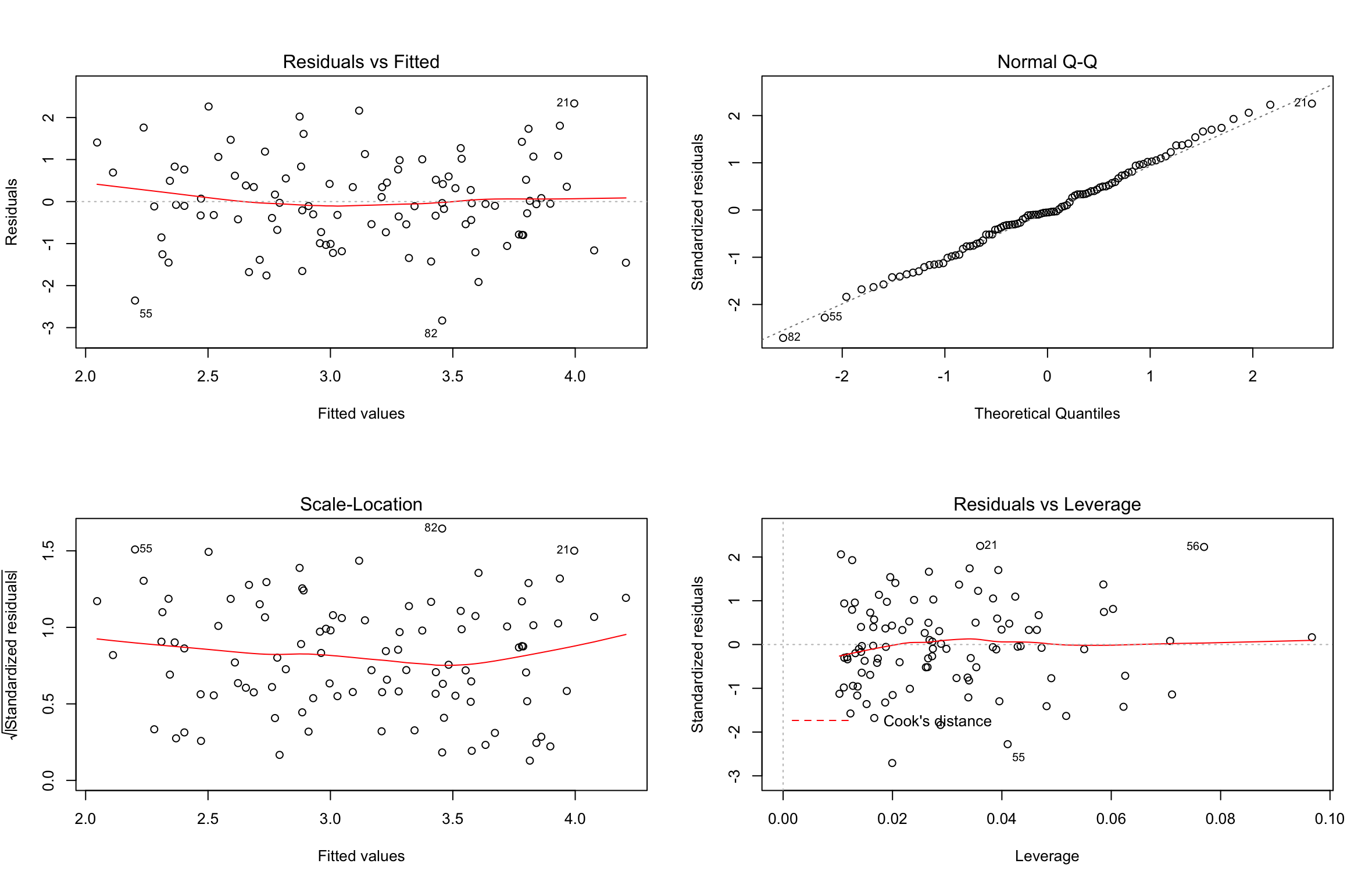
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.072 on 98 degrees of freedom

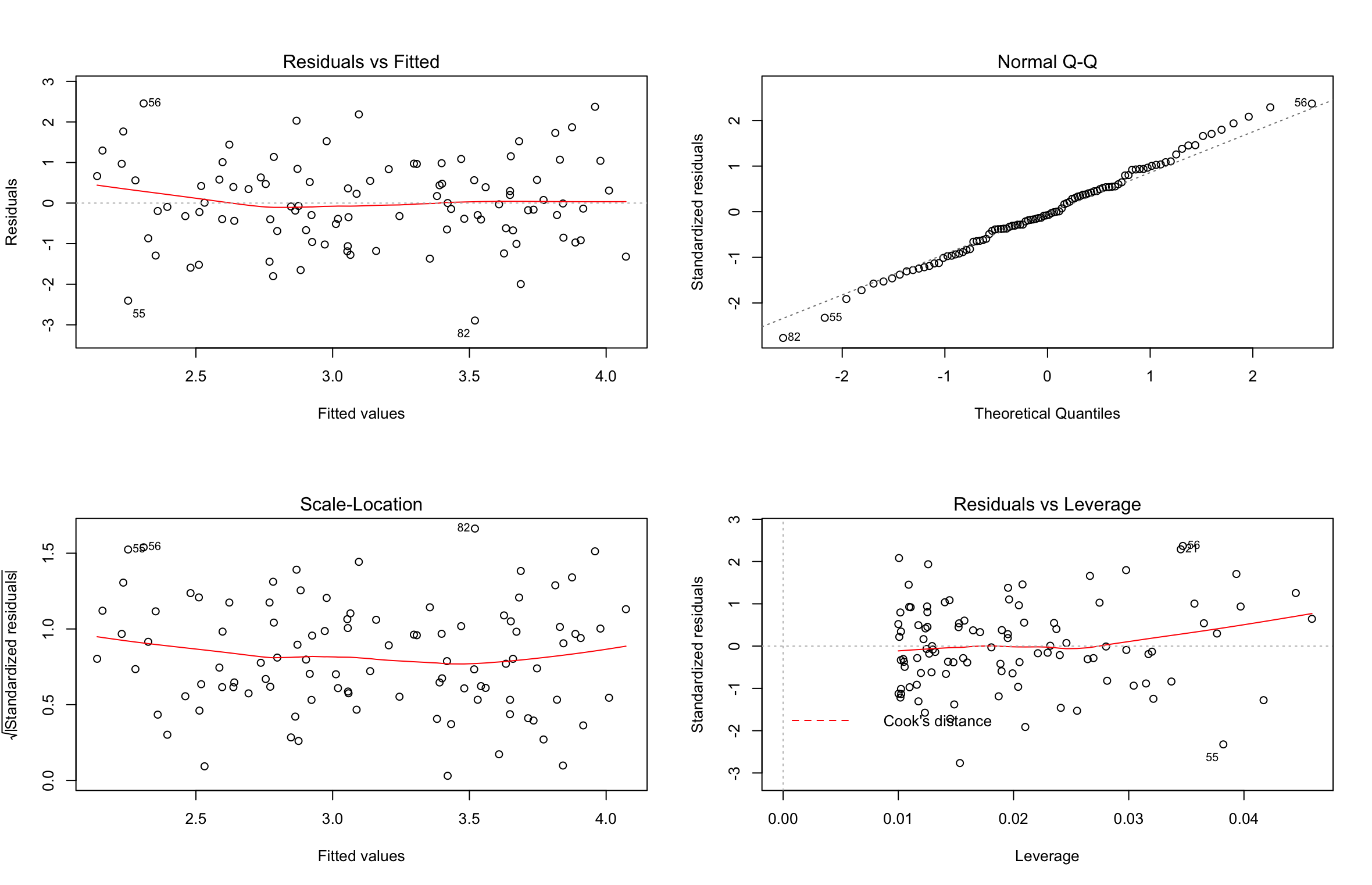
Multiple R-squared: 0.1763, Adjusted R-squared: 0.1679

F-statistic: 20.98 on 1 and 98 DF, p-value: 1.366e-05

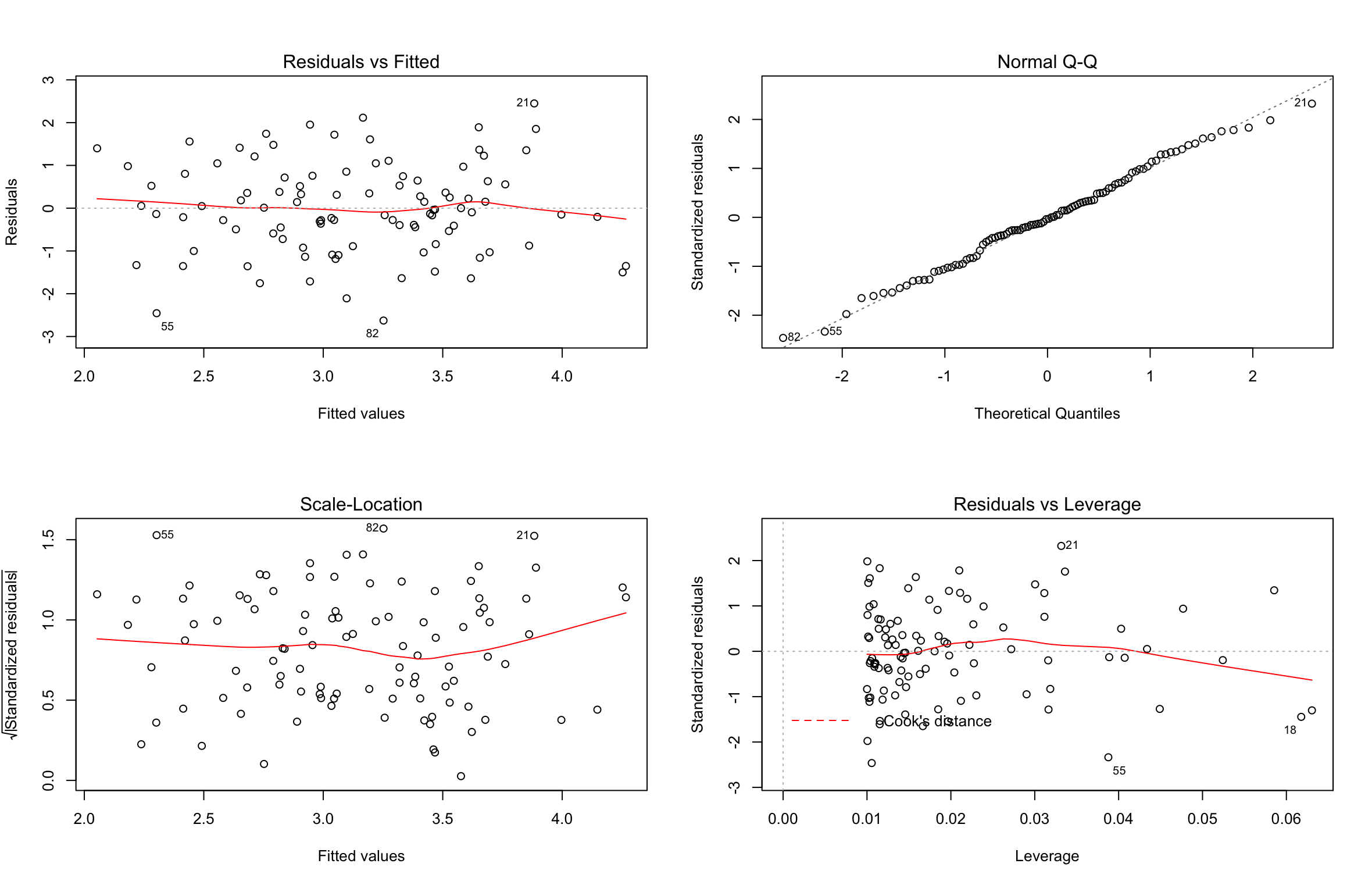
iv.



v.



vi.



**INTERPRETATION,** In the model with two predictors, the last point is a high-leverage point. In the model with “x1” as sole predictor, the last point is an outlier. In the model with “x2” as sole predictor, the last point is a high leverage point.

########start of question 10########

library(ISLR)

data("Carseats")

lm.fit.a <- lm(Sales ~ Price + Urban + US, data=Carseats)

summary(lm.fit.a)

View(Carseats)

attach(Carseats)

str(data.frame(Price, Urban, US))

fit4 <- lm(Sales ~ Price + US, data = Carseats)

summary(fit4)

confint(fit4)

par(mfrow = c(2, 2))

plot(fit4)

########end of question 10#######

########start of question 14######

set.seed(1)

x1 <- runif(100)

x2 <- 0.5 \* x1 + rnorm(100)/10

y <- 2 + 2 \* x1 + 0.3 \* x2 + rnorm(100)

cor(x1, x2)

plot(x1, x2)

fit13 <- lm(y ~ x1 + x2)

summary(fit13)

fit14 <- lm(y ~ x1)

summary(fit14)

fit15 <- lm(y ~ x2)

summary(fit15)

fit16 <- lm(y ~ x1 + x2)

fit17 <- lm(y ~ x1)

fit18 <- lm(y ~ x2)

summary(fit16)

summary(fit17)

summary(fit18)

plot(fit16)

plot(fit17)

plot(fit18)

######end of question 14#######