1. Forty-six mountains in the Adirondacks of upstate New York are known as the High Peaks with elevations near or above 4000 feet. Below is some R output from a linear regression model of Y = *Time* (expected trip time to hike the peak, in hours) on *X* = *Ascent* (in feet).

```
> summary(time.lm)
Call:
lm(formula = Time ~ Ascent, data = HighPeaks)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.2100541  1.8661683   2.256  0.02909 *
Ascent      0.0020805  0.0005909   3.521  0.00101 **
```

What is the fitted regression model?

2. Forty-six mountains in the Adirondacks of upstate New York are known as the High Peaks with elevations near or above 4000 feet. Below is some R output from a linear regression model of *Y* = *Time* (expected trip time to hike the peak, in hours) on *X* = *Ascent* (in feet).

```
> summary(time.lm)
Call:
lm(formula = Time ~ Ascent, data = HighPeaks)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.2100541  1.8661683   2.256  0.02909 *
Ascent      0.0020805  0.0005909   3.521  0.00101 **
```

Interpret the slope of the model in context.

3. Forty-six mountains in the Adirondacks of upstate New York are known as the High Peaks with elevations near or above 4000 feet. We regress *Y* = *Time* (expected trip time to hike the peak, in hours) on *X* = *Ascent* (in feet), and the fitted model is

Using this model, predict the hiking time for a mountain with an ascent of 3000 feet and explain how much faith you have in that prediction.

A) 6.3 hours
B) Somewhere between 7 and 12 hours, but we can't be more specific
C) 10.5 hours
D) 3004.2 hours

4. Forty-six mountains in the Adirondacks of upstate New York are known as the High Peaks with elevations near or above 4000 feet. We regress $Y = Time$ (expected trip time to hike the peak, in hours) on $X = Ascent$ (in feet), and the fitted model is

   Using this model, predict the hiking time for a mountain with an ascent of 300 feet and explain how much faith you have in that prediction.
   A) 0.63 hours
   B) Somewhere between 0 and 18 hours, but we can't be more specific
   C) 4.8 hours
   D) 304.2 hours

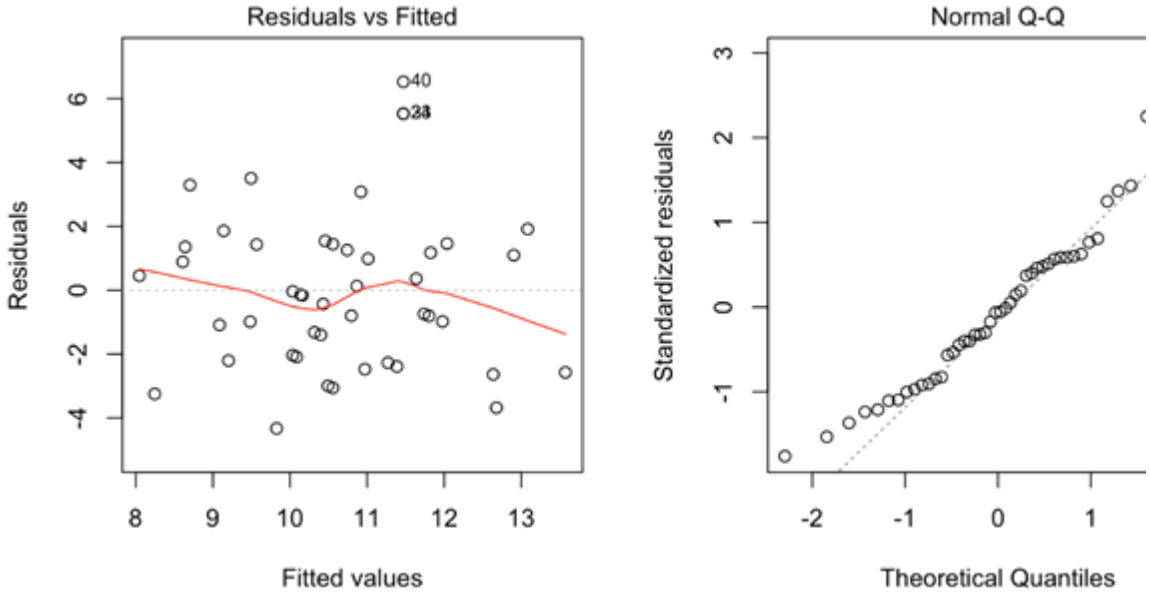
5. Forty-six mountains in the Adirondacks of upstate New York are known as the High Peaks with elevations near or above 4000 feet. Below is some R output from a linear regression model of $Y = Time$ (expected trip time to hike the peak, in hours) on $X = Ascent$ (in feet).

```
> summary(time.lm)
Call:
lm(formula = Time ~ Ascent, data = HighPeaks)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.2100541  1.8661683   2.256  0.02909 *
Ascent      0.0020805  0.0005909   3.521  0.00101 **

Residual standard error: 2.496 on 44 degrees of freedom
Multiple R-squared:  0.2198,     Adjusted R-squared:  0.2021
F-statistic:  12.4 on 1 and 44 DF,  p-value: 0.001014
```

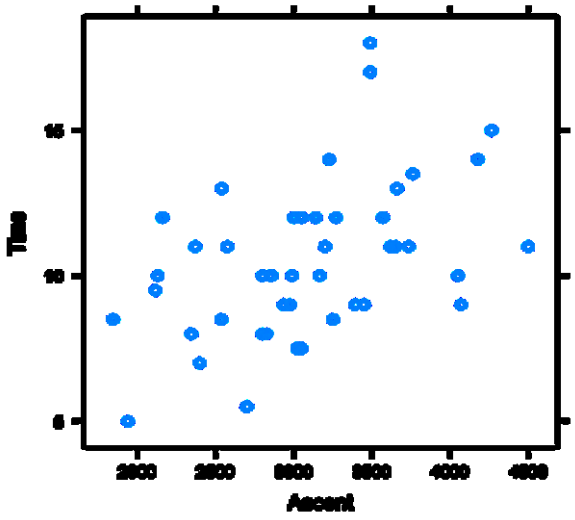
   Report the standard error of regression.
   A) 1.86617
   B) 0.00059
   C) 2.496
   D) Values between –4.327 and 6.529

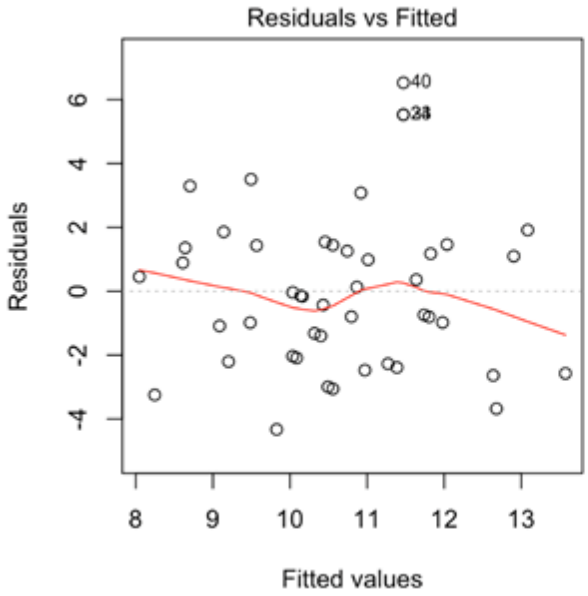
6. Forty-six mountains in the Adirondacks of upstate New York are known as the High Peaks with elevations near or above 4000 feet. A linear regression model of $Y = Time$ (expected trip time to hike the peak, in hours) on $X = Ascent$ (in feet) results in a standard error of regression of 2.496. Interpret this value.

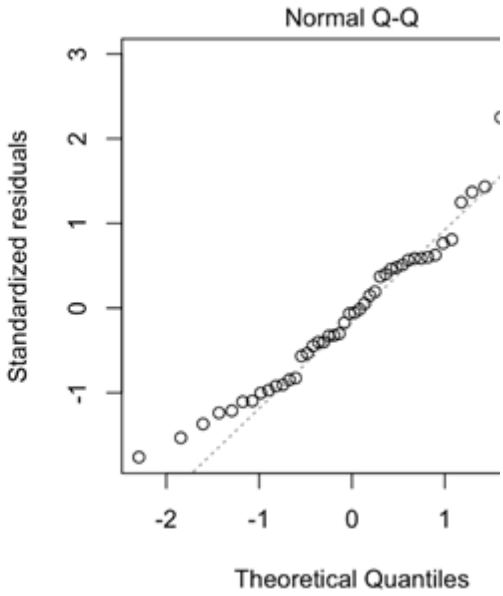
7. Forty-six mountains in the Adirondacks of upstate New York are known as the High Peaks with elevations near or above 4000 feet. A linear regression model of $Y = Time$ (expected trip time to hike the peak, in hours) on $X = Ascent$ (in feet) results in the residual plots below. Are there any outliers in this situation? If so, identify these points and explain what leads you to believe they are outliers.



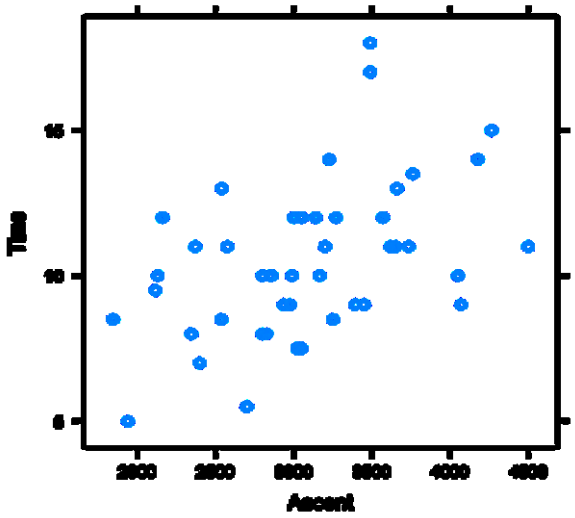
8. Forty-six mountains in the Adirondacks of upstate New York are known as the High Peaks with elevations near or above 4000 feet. A linear regression model of $Y = Time$ (expected trip time to hike the peak, in hours) on $X = Ascent$ (in feet) results in the scatterplot and residual plots below. Are there any influential points in this situation? If so, identify these points and explain what leads you to believe they are influential.

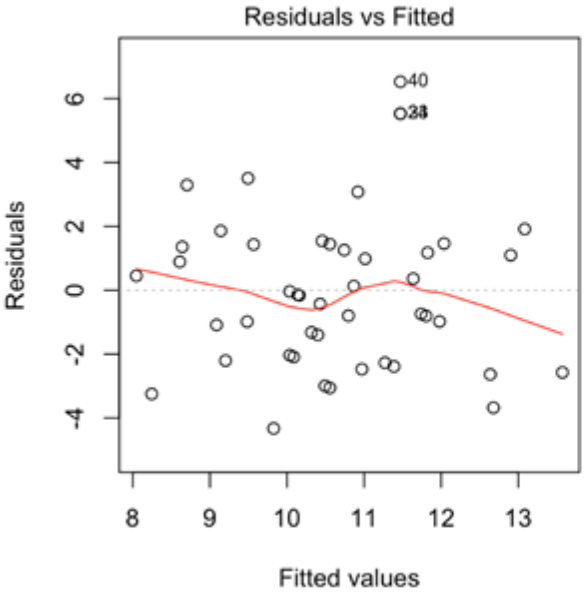


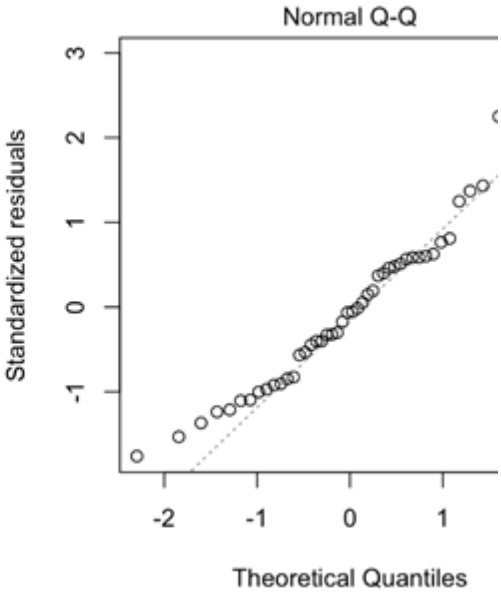




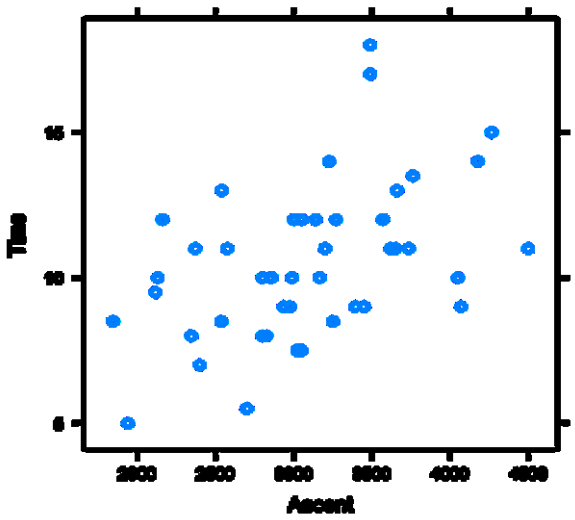
9. Forty-six mountains in the Adirondacks of upstate New York are known as the High Peaks with elevations near or above 4000 feet. A linear regression model of $Y = Time$ (expected trip time to hike the peak, in hours) on $X = Ascent$ (in feet) results in the scatterplot and residual plots below. Are the conditions for inference met in this case? Make sure you discuss each condition.

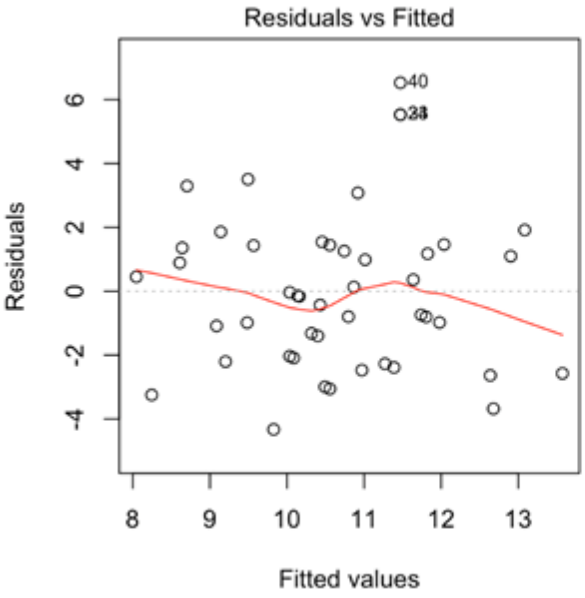


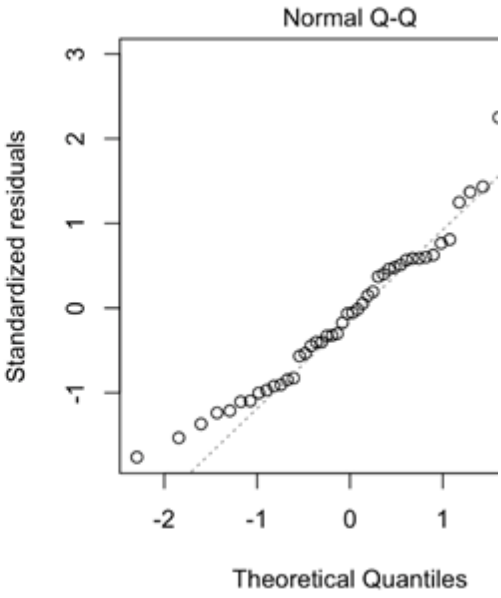




10. Forty-six mountains in the Adirondacks of upstate New York are known as the High Peaks with elevations near or above 4000 feet. A linear regression model of $Y = Time$ (expected trip time to hike the peak, in hours) on $X = Ascent$ (in feet) results in the scatterplot and residual plots below. Do you think a transformation for the explanatory or response variable is needed here? Explain your answer. If you think a transformation is needed, suggest a specific transformation and explain why you recommend it specifically.

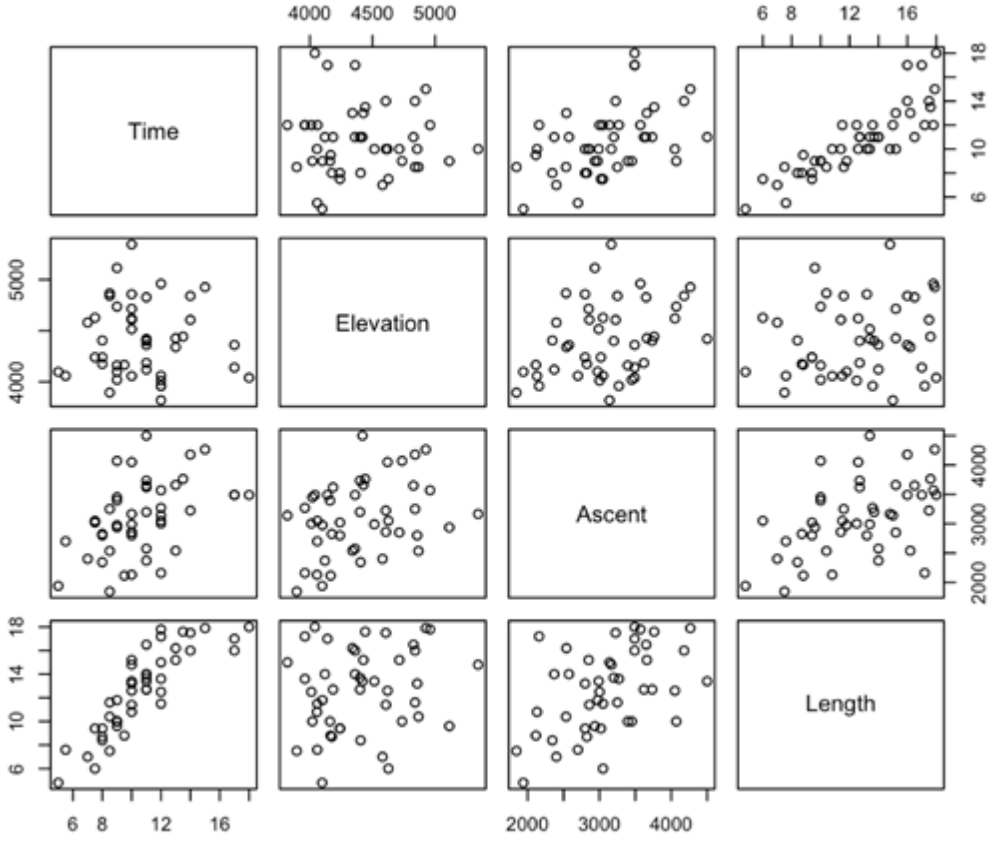






11. Forty-six mountains in the Adirondacks of upstate New York are known as the High Peaks with elevations near or above 4000 feet. The variables include *Elevation* (in feet) of each peak, *Difficulty* rating (on a 2 to 7 scale with 7 being the most difficult), *Ascent* (in feet), *Length* of a round-trip (in miles), and expected hike *Time* (in hours). Below is a scatterplot matrix. This is a matrix that contains a scatterplot for each pair of numerical variables in the data set. For instance, the graph in the second column of the first row shows a scatterplot with *Elevation* on the *x* axis and *Time* on the *y* axis.

Based on these plots, which variable is the best single predictor of *Time*?



A) *Elevation*
B) *Ascent*
C) *Length*

12. Cholesterol levels are measured on a sample of 21 volunteers. HDL (high-density lipoprotein, or "good" cholesterol) is regressed on total cholesterol (*Chol*). Below is some R output from a regression model. Write down the least-squares regression line for these data.

```
> summary(chol.lm)
Call:
lm(formula = HDL ~ Chol, data = HDL)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 24.32224    8.35551   2.911  0.00896 **
Chol         0.11599    0.03436   3.376  0.00317 **
```

13. Cholesterol levels are measured on a sample of 21 volunteers. HDL (high-density lipoprotein, or "good" cholesterol) is regressed on total cholesterol (*Chol*). Below is some R output from a regression model. Interpret the slope of the model in context.

```
> summary(chol.lm)
Call:
lm(formula = HDL ~ Chol, data = HDL)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 24.32224    8.35551   2.911  0.00896 **
Chol         0.11599    0.03436   3.376  0.00317 **
```

14. Cholesterol levels are measured on a sample of 21 volunteers. HDL (high-density lipoprotein, or "good" cholesterol) is regressed on total cholesterol (*Chol*). The fitted model is

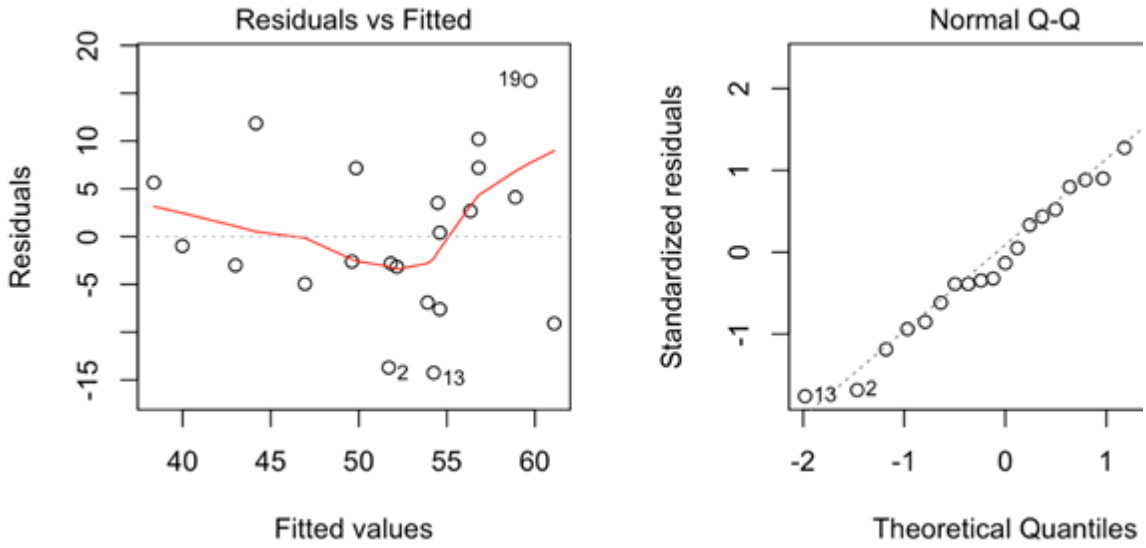
    Does it make sense to interpret the intercept of this model?

15. Cholesterol levels are measured on a sample of 21 volunteers. HDL (high-density lipoprotein, or "good" cholesterol) is regressed on total cholesterol (*Chol*). The fitted model is

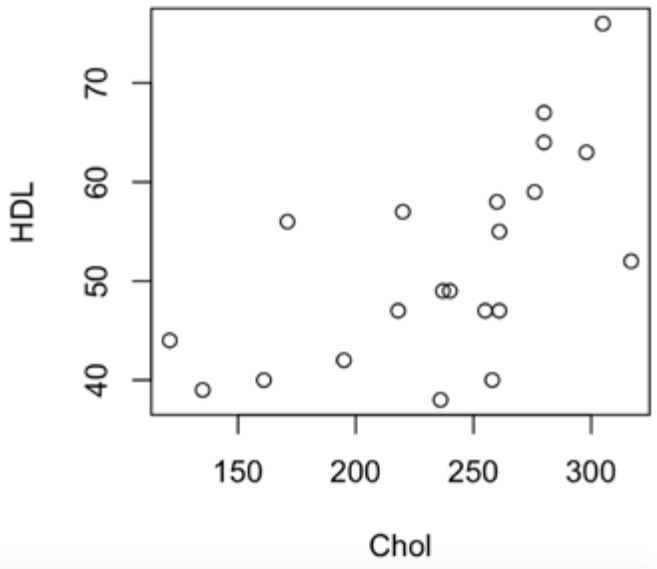
    A new patient shows a total cholesterol level of 280 mg/dl. Using this model, what would you predict as the HDL value for this patient?

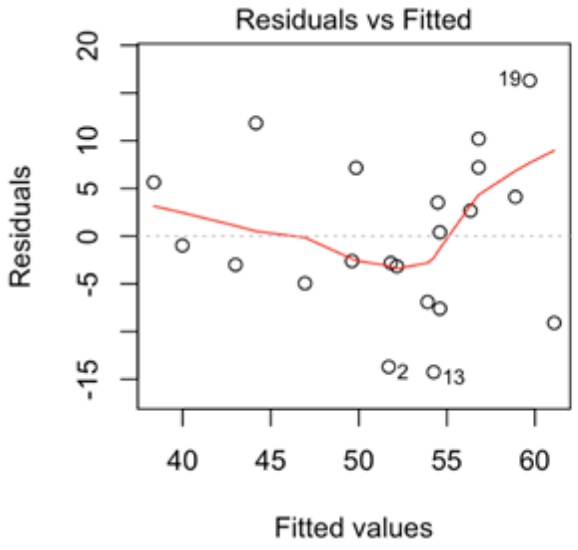
    A) 32.5 mg/dl
    B) 56.8 mg/dl
    C) 57.2 mg/dl
    D) 304.3 mg/dl

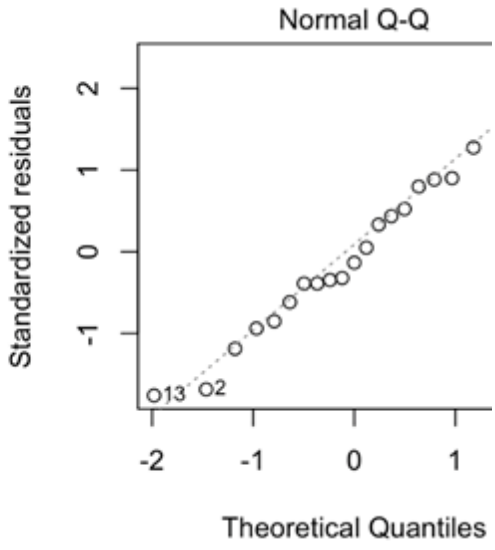
16. Cholesterol levels are measured on a sample of 21 volunteers. HDL (high-density lipoprotein, or "good" cholesterol) is regressed on total cholesterol (*Chol*), which results in the residual plots below. Are there any outliers in this situation? If so, identify these points and explain what leads you to believe they are outliers.



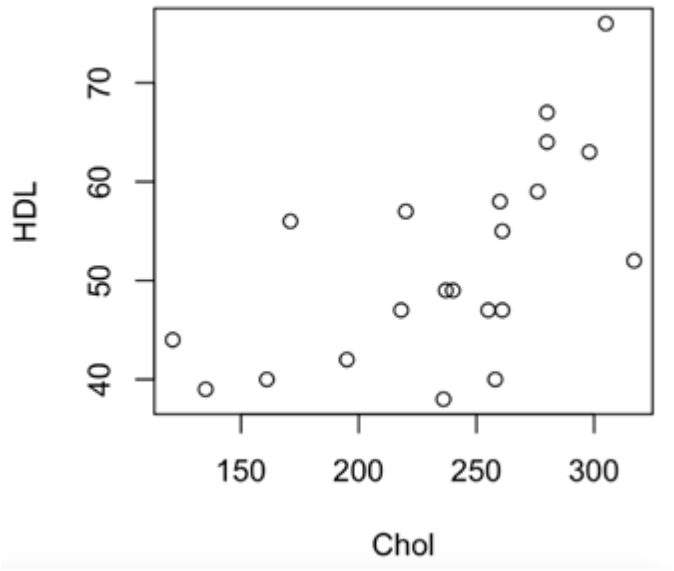
17. Cholesterol levels are measured on a sample of 21 volunteers. HDL (high-density lipoprotein, or "good" cholesterol) is regressed on total cholesterol (*Chol*), which results in the scatterplot and residual plots below. Are there any influential points in this situation? If so, identify these points and explain what leads you to believe they are influential.

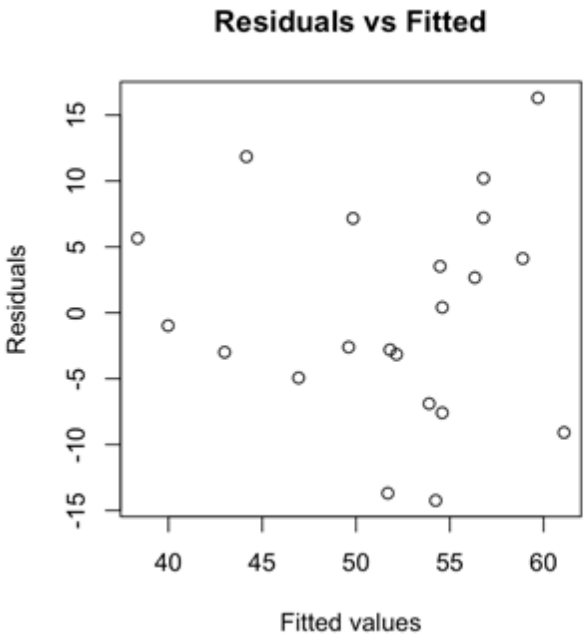


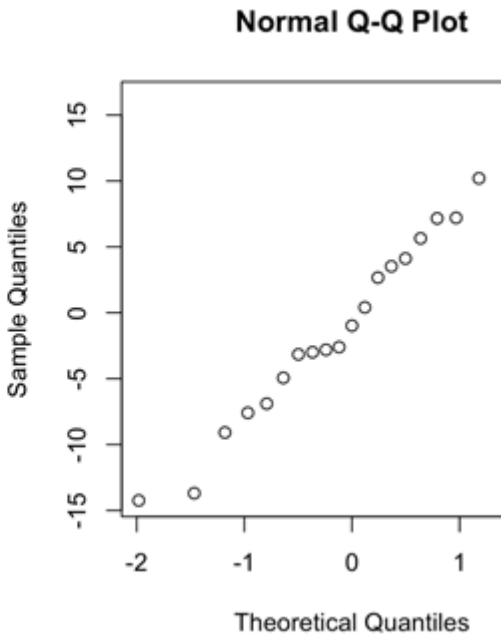




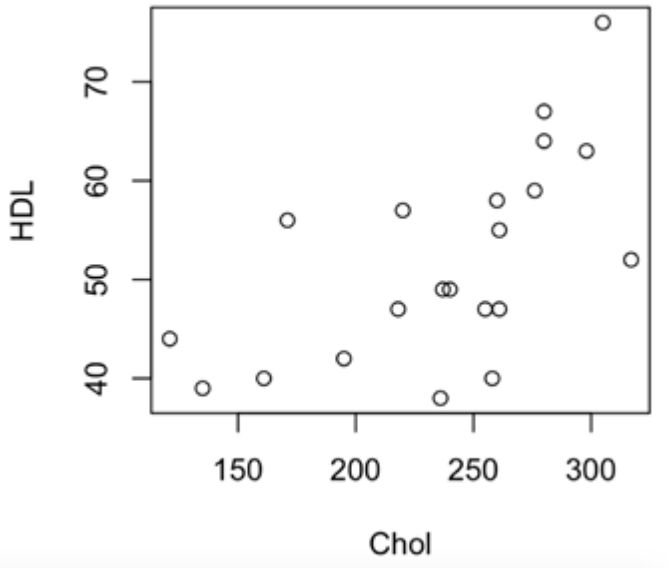
18. Cholesterol levels are measured on a sample of 21 volunteers. HDL (high-density lipoprotein, or "good" cholesterol) is regressed on total cholesterol (*Chol*), which results in the scatterplot and residual plots below. Are the conditions for inference met in this case? Make sure you discuss each condition.

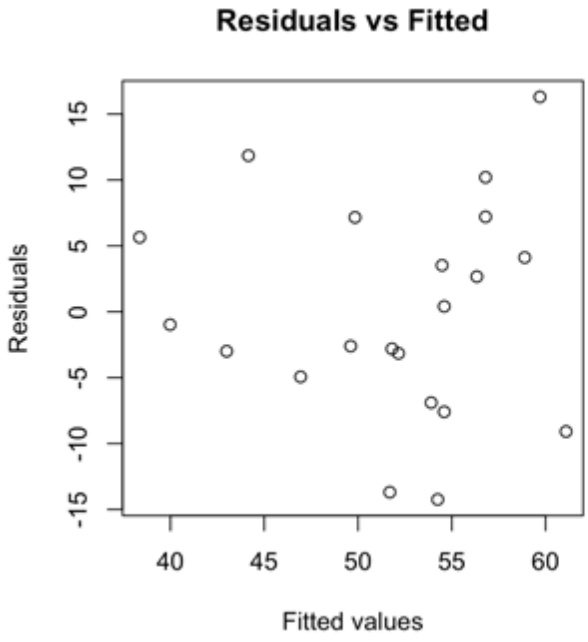


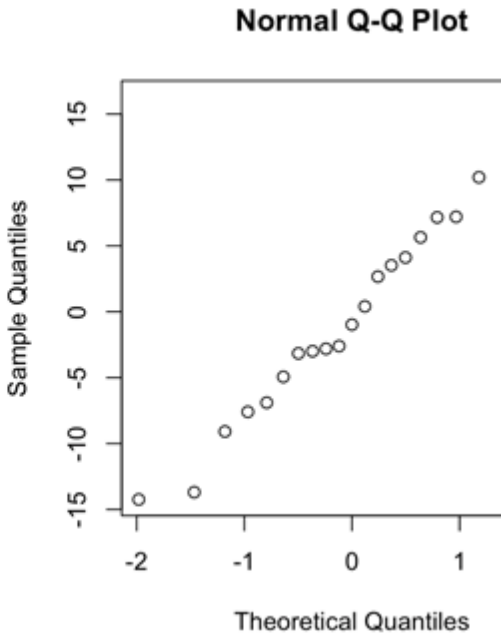




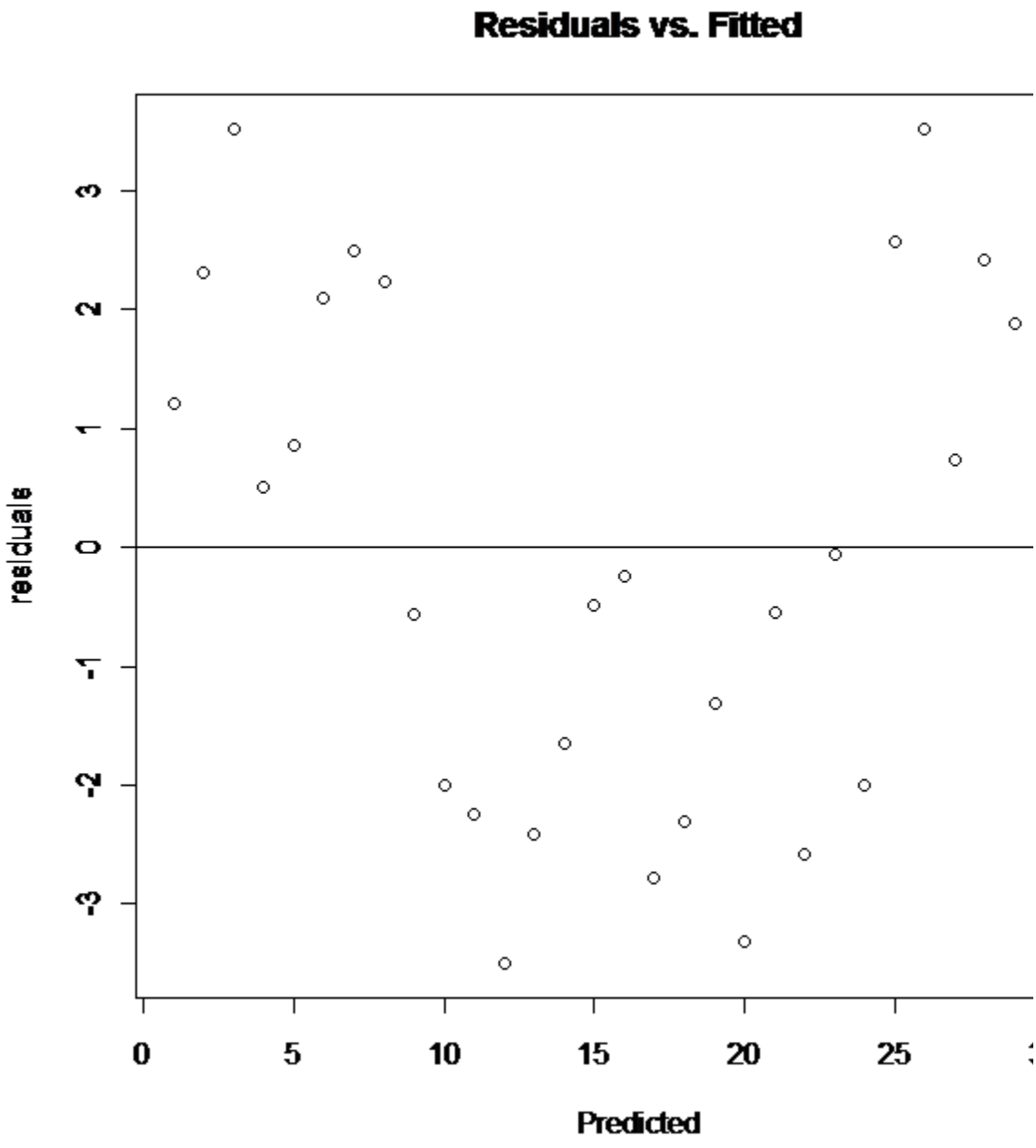
19. Cholesterol levels are measured on a sample of 21 volunteers. HDL (high-density lipoprotein, or "good" cholesterol) is regressed on total cholesterol (*Chol*), which results in the scatterplot and residual plots below. Do you think a transformation for the explanatory or response variable is needed here? Explain your answer. If you think a transformation is needed, suggest a specific transformation and explain why you recommend it specifically.





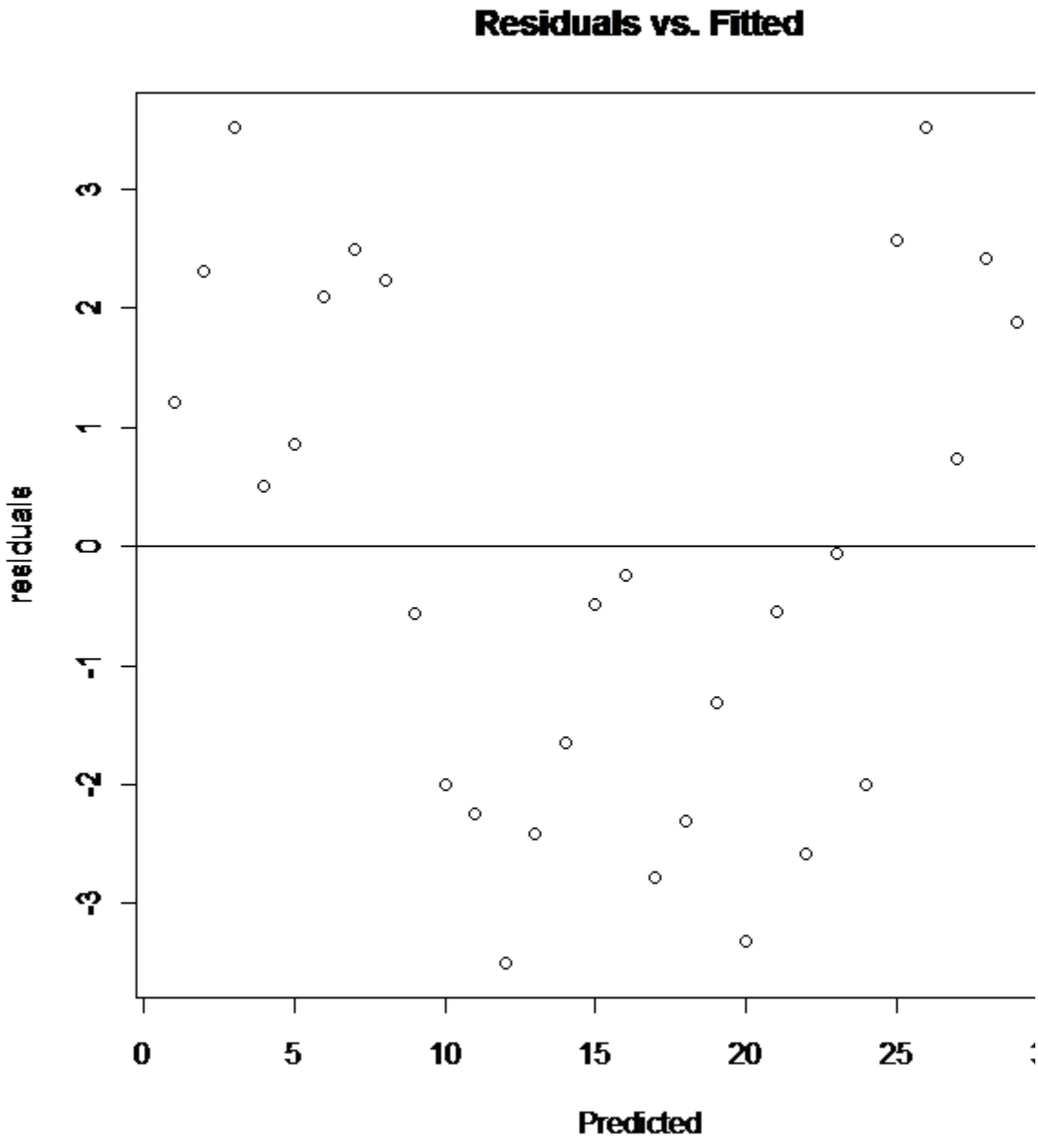


20. A residuals vs. fitted value plot for a regression model is shown below. Based only on the information in this plot, do you feel the condition of linearity is *reasonable*, *problematic*, or you *can't judge* (from the plot shown)?



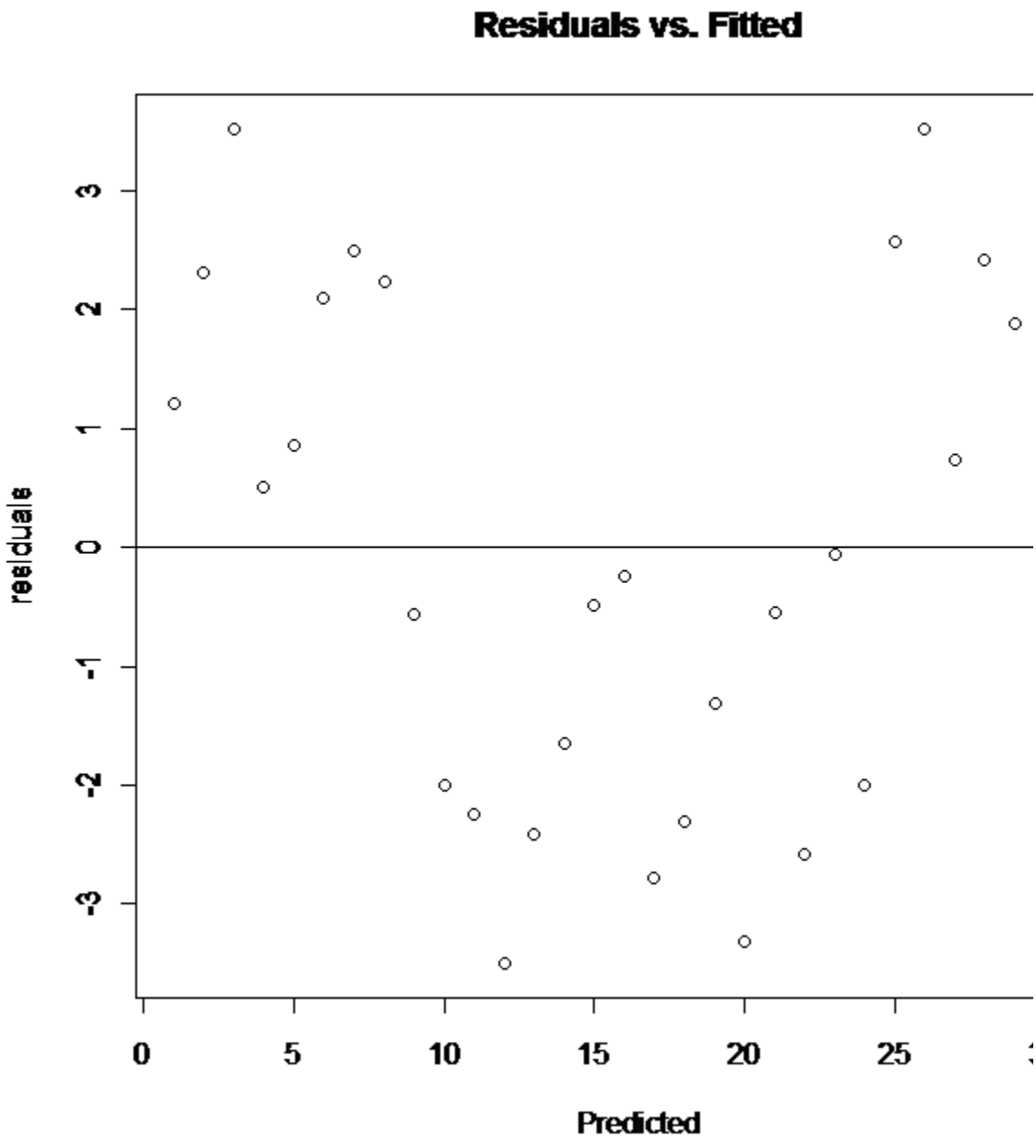
A) Reasonable
B) Problematic
C) Can't judge

21. A residuals vs. fitted value plot for a regression model is shown below. Based only on the information in this plot, do you feel the condition of equal variance is *reasonable*, *problematic*, or you *can't judge* (from the plot shown)?



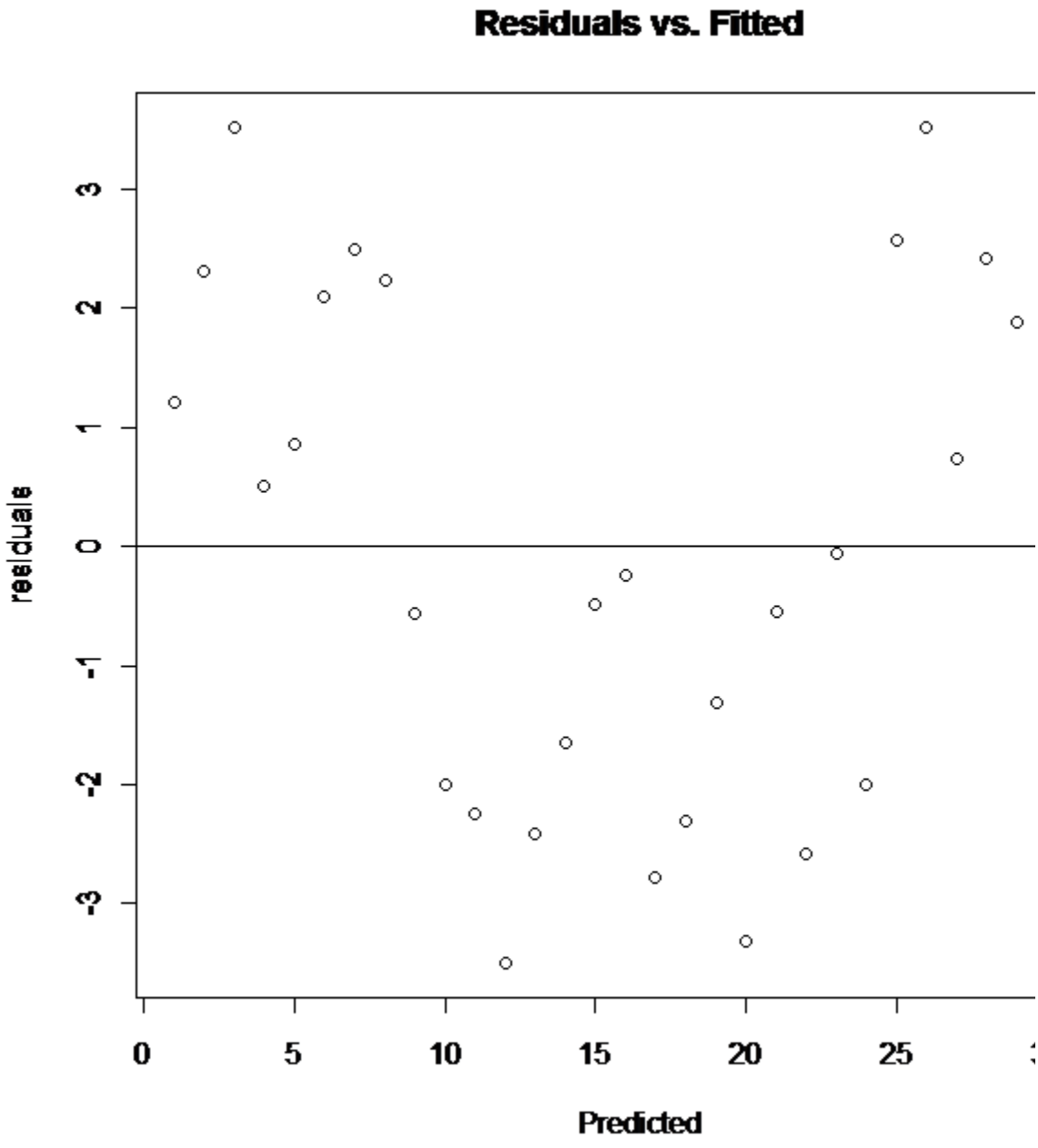
A) Reasonable
B) Problematic
C) Can't judge

22. A residuals vs. fitted value plot for a regression model is shown below. Based only on the information in this plot, do you feel the condition of normality is *reasonable*, *problematic*, or you *can't judge* (from the plot shown)?



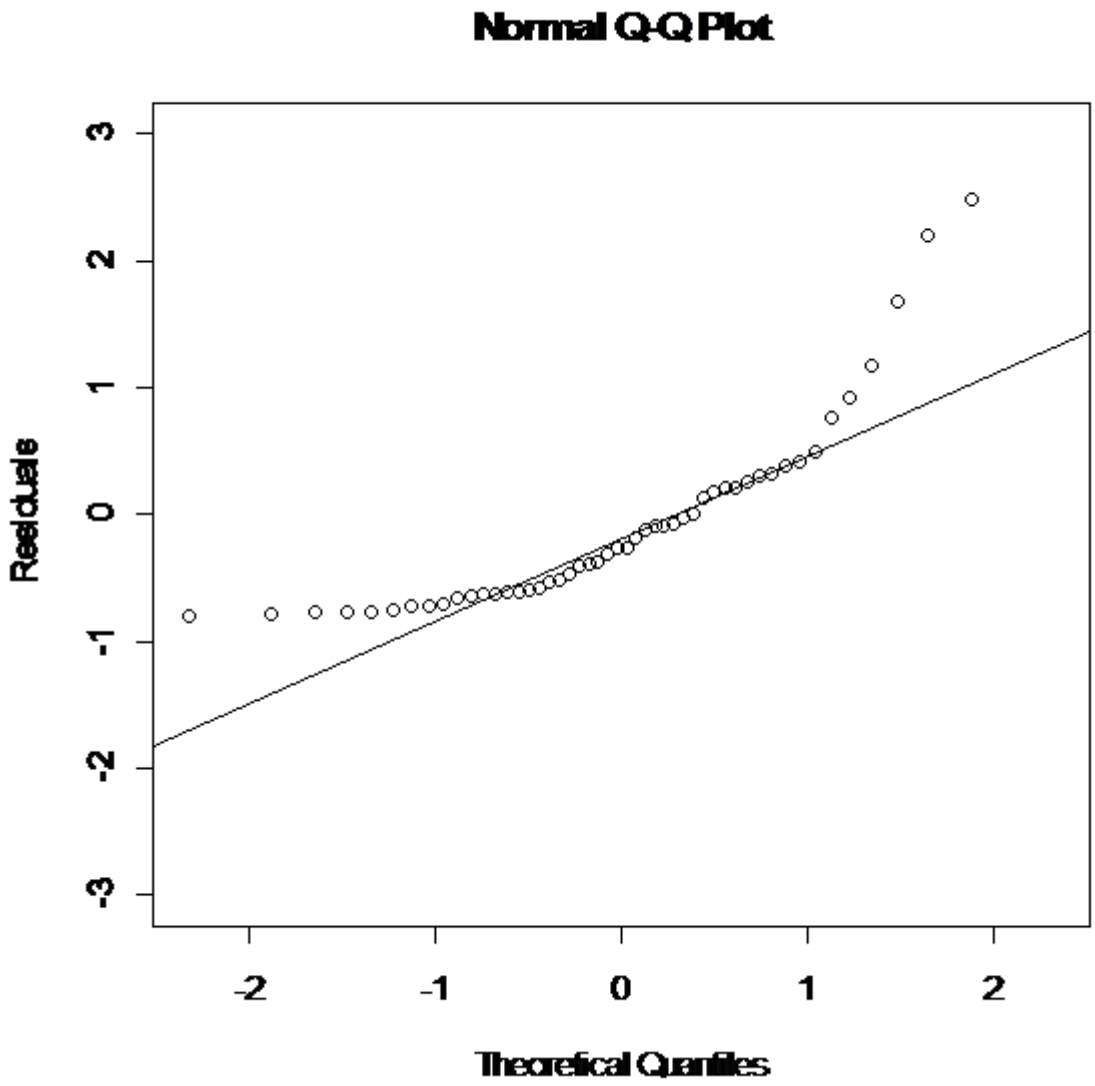
A) Reasonable
B) Problematic
C) Can't judge

23. A residuals vs. fitted value plot for a regression model is shown below. Based only on the information in this plot, do you feel the condition of independence is *reasonable*, *problematic*, or you *can't judge* (from the plot shown)?



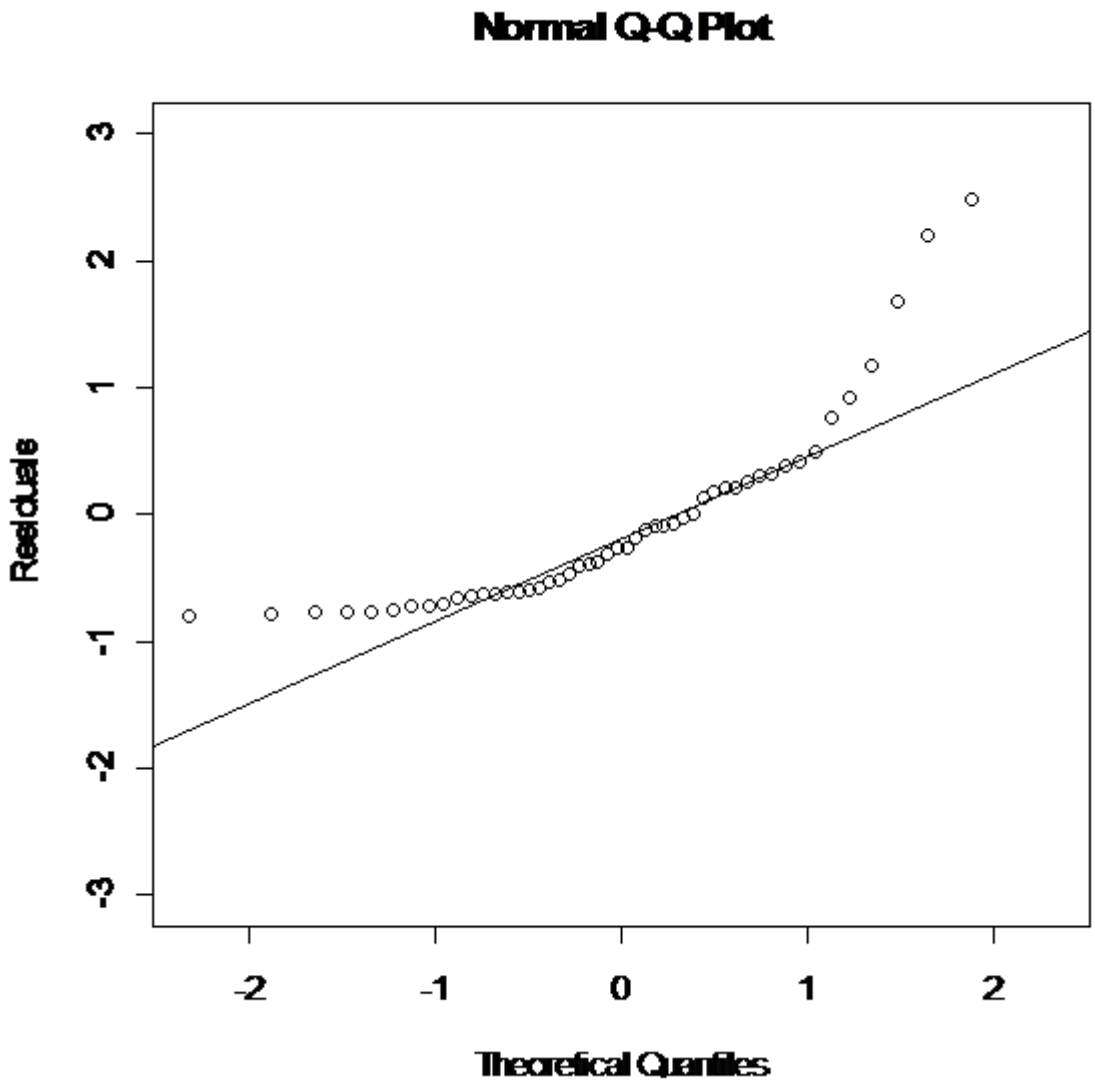
A) Reasonable
B) Problematic
C) Can't judge

24. A normal quantile plot for a regression model is shown below. Based only on the information in this plot, do you feel the condition of linearity is *reasonable*, *problematic*, or you *can't judge* (from the plot shown)?



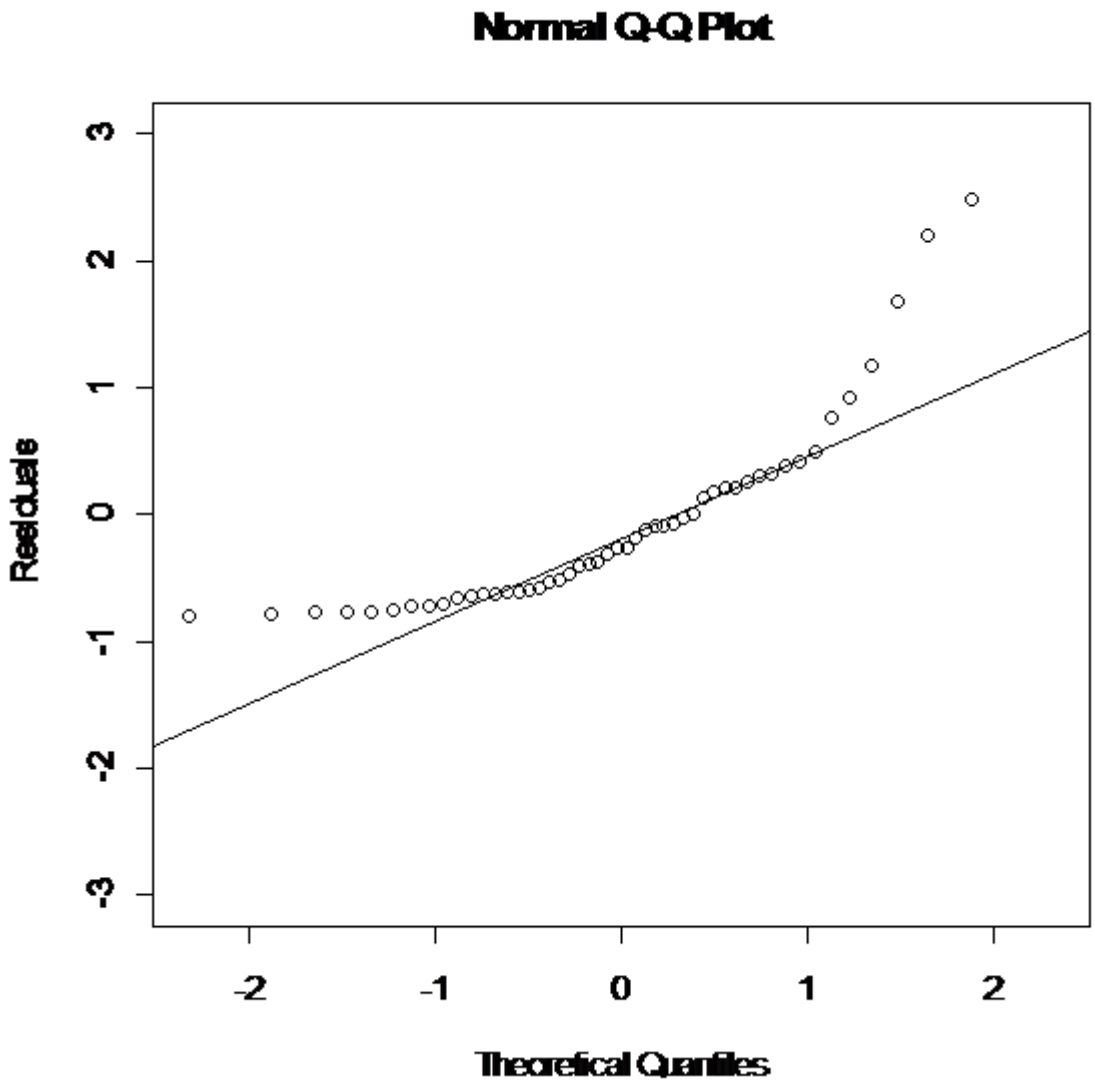
A) Reasonable
B) Problematic
C) Can't judge

25. A normal quantile plot for a regression model is shown below. Based only on the information in this plot, do you feel the condition of equal variance is *reasonable*, *problematic*, or you *can't judge* (from the plot shown)?



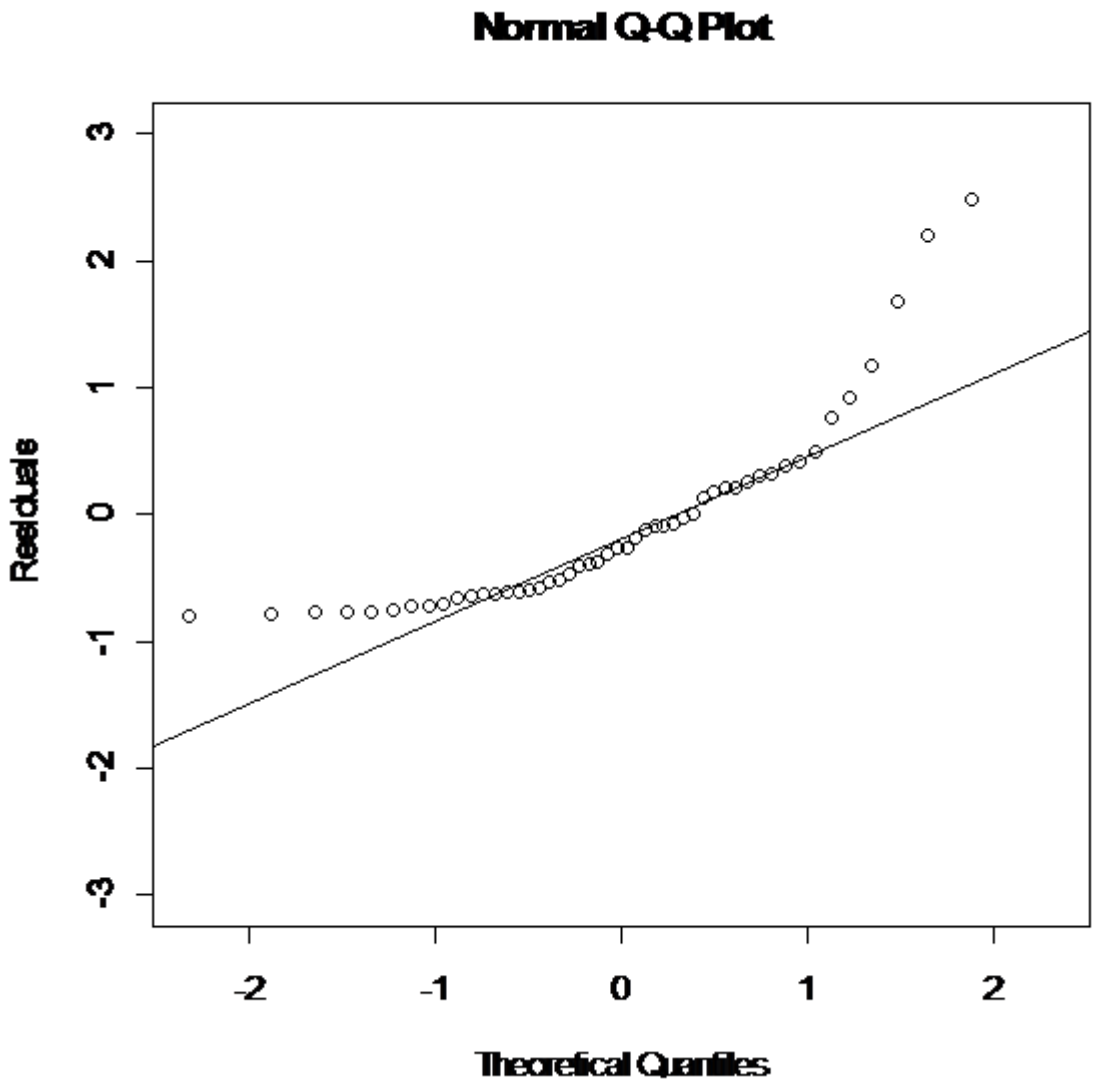
A) Reasonable
B) Problematic
C) Can't judge

26. A normal quantile plot for a regression model is shown below. Based only on the information in this plot, do you feel the condition of normality is *reasonable*, *problematic*, or you *can't judge* (from the plot shown)?



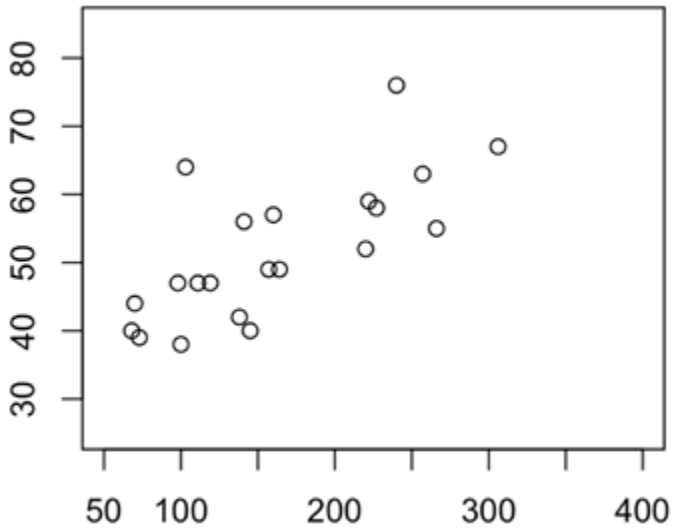
A) Reasonable
B) Problematic
C) Can't judge

27. A normal quantile plot for a regression model is shown below. Based only on the information in this plot, do you feel the condition of independence is *reasonable*, *problematic*, or you *can't judge* (from the plot shown)?



A) Reasonable
B) Problematic
C) Can't judge

28. Below is a scatterplot. You will be adding two points to the graph. Specifically, you will add (1) a point that is an outlier but not influential, which you will draw as an "O," and (2) a point that is influential but not an outlier, which you will draw as an "I."



29. Why might we prefer to use *standardized* residuals (rather than residuals) when looking for unusual points in a regression? Explain.

Answer Key

1.
2.
3. C
4. C
5. C
6.
7.
8.
9.
10.
11. C
12.
13.
14.
15. B
16.
17.
18.
19.
20. B
21. A
22. C
23. C
24. C
25. C
26. B
27. C
28.
29.