

**Question 10)** The plots to the left shows diagnostics from the scatter plot in the previous question.

**10a)** Which **TWO** plots provide information on the normality of the data? (circle 2 answers)

1. Plot a
2. Plot b
3. Plot c
4. Plot e

**10b)** Which diagnostic shown in the previous question provides information about whether there are any **outliers** and/or **influental points** in the data set?

1. Plot a
2. Plot b
3. Plot c
4. Plot e

**10c)** Is there an **outerlier/influential point** in this data set, and if so what is it? (1 point)

1. There is not an outlier/influential point
2. There is an influential point and it is \_\_\_\_\_\_\_\_\_

**Question 11:** When doing regression, the process involves calculating the residuals of the regression model and doing what mathematically to them to determine the line that best fits the data?

1. Doing nothing to them. The line of best fit is determined using the raw residuals
2. Getting rid of the negative sign (-) by taking the absolute value (abs) of the residual
3. Getting rid of the negative sign (-) by squaring the residual (eg, residual^2)
4. Cubing the residuals, ie I(residual^3)

**Question 14:** **TRUE / FALSE:** If you have a very low p-value (highly significant difference), you must also have a very high R2 value. Therefore, R2 is highly correlated with p-values. (1 point)

**Question 15:** What are the key assumptions of regression? (4 points)

1) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

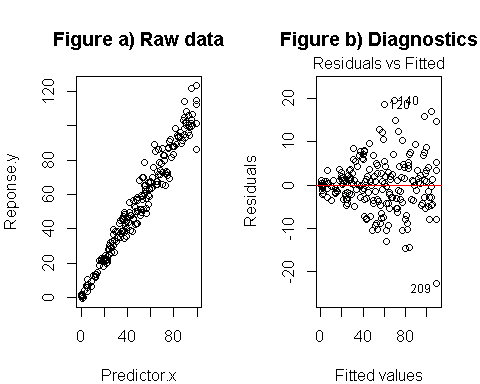
2) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

3) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

4) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Question 16:** Which of your 4 key assumptions is the most important to pay attention to (1 point; refer to to the list you made for question 17)

1)Assumption 1 2)Assumption 2 3) Assumption 3 4) Assumption 4

**Question 17:** The left-hand graph shows a plot raw data, and the right hand graph shows a **diagnostic plot**.

**17a)** What assumption of regression analysis does this diagnostics plot tell us about? (refer to the list you made above)

1)Assumption 1

2)Assumption 2

3) Assumption 3

4) Assumption 4

**17b)** Do these data in the previous plot appear to violate this assumption?

1. Yes, assumption violated.
2. No

**17c)** Assuming these data do indeed violate this assumption, what could be done to try to fix the problem?

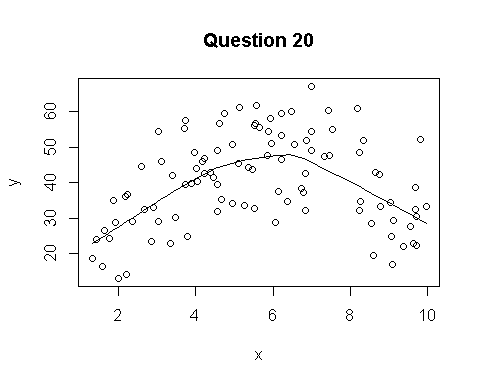
1. Nothing
2. Remove outliers
3. Log transformation of predictor (x)
4. Log transformation of response (y)
5. Collect more data

**Question 18:** Which of the following things does the **log transformation** **NOT** do

1. Improve normality of the residuals
2. Make the variance more constant
3. Reduce impact of non-random sampling
4. Reduce the impact of outliers

**Question 19)** Which **two** statements are true about outliers ( circle 2)

1. All outliers are bad data points that should be removed
2. Regression models assume that there are no outliers
3. All outliers occur due to errors during data collection or data entry.
4. Diagnostics techniques such as plot cannot identify outliers
5. Outliers can occur due to real but extreme observations



**Question 20)** For the following questions consider this scatter plot of variable y plotted against variable x.

**20b)** What is the technical name for a such as this that is not straight, as in the previous plot? (1/2 point)

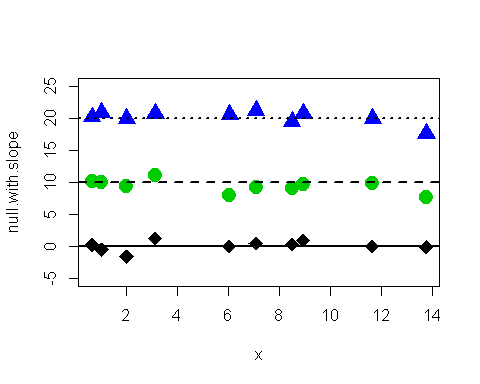
1. Logistic
2. Curvey
3. Residual
4. Asymptote
5. Non-linear

**Question 21)** What **R code** would you add to a regression to model a curvey line? Assume the predictor is called "x" (1 point)

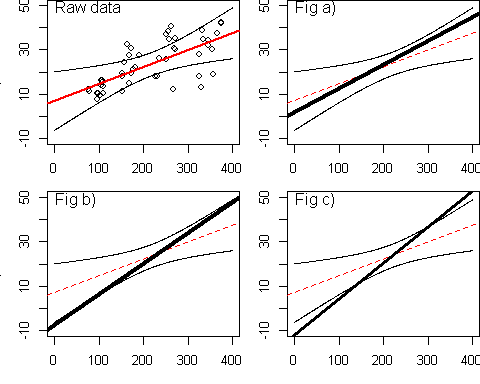
1. I
2. I()
3. I(x)
4. I(x^2)
5. I(x^3)
6. I(1/x)
7. none of the above

**Question 23)** Which of these **TWO** statements are true about **logistic regression** (circle 2)

1. Confidence intervals cannot be greater than 1
2. Confidence intervals cannot be less than 0
3. Confidence intervals cannot be calculated
4. Confidence intervals should be calculated from percentages

**Question 24)** This plot to the left shows three sets of data (diamonds, circles, triangles) and regression lines running through them. **What statement is true about these lines?**

1. They all have the same intercept
2. They all have the same slope
3. They are all all significantly different from 0
4. They all represent an alternative hypothesis.
5. They are all random

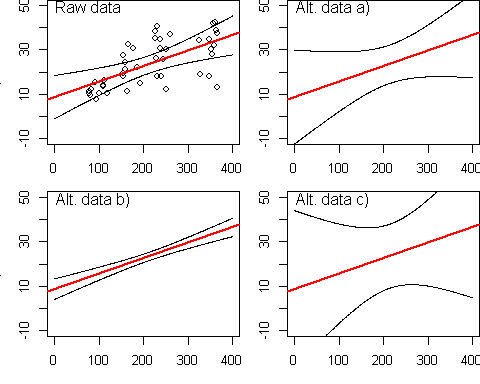


**Question 25)** The figure to the left,, in the top-left panel, shows scatter plot of raw data with a regression line and a "confidence band" (aka "confidence ellipse") around the line. This line represents uncertainty about the true values of the parameters that define the line. In panels a, b, and c are shown 3 other possible regression lines as thick lines plotted with the original regression line (now thin and dotted) and confidence band.

**Of the 3 alternative regression lines, which ones are possible alternatives that are consistent with the data?** (1 point)

1. Fig a (upper right)
2. Fig b (lower left)
3. Fig c (lower right)
4. All are consistent with the data
5. None are consistent with the data

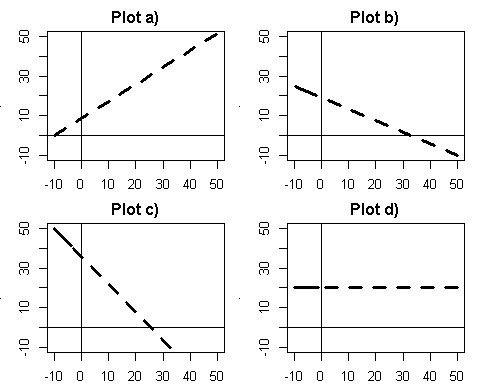
**-----------------------------------------------------------------------------------------------------------------------------------------------**

**Question 26)** In the plot to the left, the raw data used previously is plotted with it 95% Confidence Band. Panels a, b and c represent similar regression lines fit to alternative data sets with **different sample sizes.**

**Why do the confidence bands change size (get wider or narrower)? (1 point)**

1. When sample sizes are small there are more **outliers** and confidence bands are big. When sample sizes are large confidence bands ar narrow.
2. When sample sizes are small parameters are *not* estimated with precision and confidence bands are big. When sample sizes are large parameters can be estimated precisely and confidence bands are narrow.
3. Change in the error bands is not related to sample size and is just due to random error.
4. None of the above
5. All of the above

**-----------------------------------------------------------------------------------------------------------------------------------------------**

**Question 27** The plot to the left shows four hypothetical regression lines with different intercepts and slopes

**27a)** Which plot(s) have positive (+) slopes? (circle all that apply; 1 point)

1. Plot a)
2. Plot b)
3. Plot c)
4. Plot d)
5. None

**27b)** Which plot(s) on the previous page have negative (-) slopes)? (circle all that apply; 1 point)

1. Plot a)
2. Plot b)
3. Plot c)
4. Plot d)
5. None

**27c)** Which plot(s) have slopes of zero (0)? (circle all that apply; 1 point)

1. Plot a)
2. Plot b)
3. Plot c)
4. Plot d)
5. None

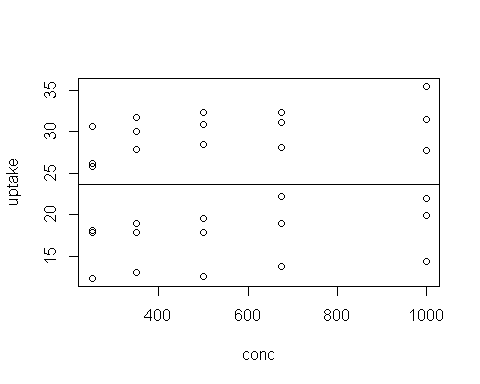
**27d)** Which plot(s) have positive intercepts? (circle all that apply; 1 point)

1. Plot a)
2. Plot b)
3. Plot c)
4. Plot d)
5. None

**27e)** Which plot(s) have negative intercepts? (circle all that apply; 1 point)

1. Plot a)
2. Plot b)
3. Plot c)
4. Plot d)
5. None

**-----------------------------------------------------------------------------------------------------------------------------------------------**

 **Question 28)** In a lab experiment researchers were interested in the effect of carbon dioxide concentration (CO2) in the air the rate at which plants can use CO2 for photosynthesis. Their response variable was CO2 uptake rate ("uptake") and their predictor variable was CO2 concentation ("conc").

**28a)** Write the R code to represents a "null" hypothesis (Ho); that is, a model that assumes CO2" "update" **does not** change with "conc" (1 point)

**28a)** Write the R code to represent an "**alternative**" hypothesis (Ha); that is, a model that assumes CO2"uptake" **does** change with CO2 "conc" (1 point)

**28a)** Assume that the null model is called "m.null" and the alternative model is called "m.alt". Write the one line of R code used to test whether the null hypothesis should be rejected.

**-----------------------------------------------------------------------------------------------------------------------------------------------**

ANOVA TABLE:

## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 29 1470.7   
## 2 28 1436.8 1 33.933 0.6613 0.423

## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 21.420631 3.037488 7.052 1.14e-07 \*\*\*  
## conc 0.004017 0.004940 0.813 0.423   
## Multiple R-squared: 0.02307, Adjusted R-squared: 0.02307   
## F-statistic: 0.6613 on 1 and 28 DF, p-value: 0.423

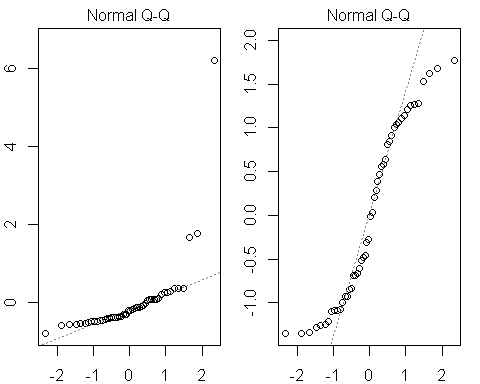
**Question 29** Above is partial R output from an analysis of the data in **question 28**.

**Question 29a)** What hypothesis does this data support?

1. Ho
2. Ha

**Question 29b)** Write a sentence (as best you can, not knowing all of the biology and the study situation) using this output to describe the results of this study. Include all relevant info we want to include in the results section of a paper.

**-----------------------------------------------------------------------------------------------------------------------------------------------**



**Question 30)** Which of these Normal Q-Q plots indicates the **worst** violation of an assumption of regression modeling? That is, which plot indicates the biggest problem with the model?

1. The left-hand plot
2. The right-hand plot
3. Neither plot
4. Both plots are about the same.

**-----------------------------------------------------------------------------------------------------------------------------------------------**

**Question 31)** You are reading an old paper from the Journal of Aquatic Ecology and the author's state "The regression analysis we conducted was highly significant (p < 0.0001), supporting our hypothesis that pH impacted the abundance of insects in southwestern PA streams."

**31a)** What can you conclude about the **slope** of the regression model they are referring to from this p-value?

1. The slope is very likely to be positive (+)
2. The slope is very likely to be negative (-)
3. The slope is very likely to be zero (0)
4. The slope is very likely to be different than zero (0)

**31b)** What can you conclude about the R2 value of the regression model from this p-value?

1. The R2 value is very high (near 0.8).
2. The R2 value is very low (near 0.1).
3. The R2 value is highly significant.
4. Nothing can be conclude about about R2