1. (23 points) Consider the four pairs plots below. Each is generated as follows: \( X_1 \sim N(0,1) \). Then \( X_2 \) is sampled in one of four ways, and \( Y \) is generated from \( X_1 \) and \( X_2 \) in one of four ways. The functions for \( X_2 \) and \( Y \) are given in the table below. Indicate which set of plots corresponds to each row of the table. Then indicate in which cases:

- \( X_1 \) and \( X_2 \) are related
- \( X_1 \) and \( X_2 \) are correlated
- There is an interaction between \( X_1 \) and \( X_2 \)
- There is a nonlinear relationship between \( Y \) and the \( X \)'s

Then answer the following question.

\[
X_2 = \begin{cases} 
1(X_{1i} + \tau_i > 0) & X_i - X_1 \ast X_2 \\
\tau_i \sim N(0,1) & \\
\end{cases}
\]

(e) For which of the above cases would fitting a standard linear model (including interactions but not transformations) be inappropriate? In this case, would fitting a standard linear model of the form \( \log(Y) \sim X_1 + X_2 \) be appropriate? Why or why not?
2. The information below relates $y$, a second measurement on wood volume, to $x_1$, a first measurement on wood volume, $x_2$, the number of trees, $x_3$, the average age of trees, and $x_4$, the average volume per tree. Note that $x_4 = x_1/x_2$. Some of the information has not been reported, so that you can figure it out on your own.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>$\hat{\beta}_k$</th>
<th>SE($\hat{\beta}_k$)</th>
<th>$t$</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>23.45</td>
<td>14.90</td>
<td>0.122</td>
<td></td>
</tr>
<tr>
<td>$x_1$</td>
<td>0.93209</td>
<td>0.08602</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>$x_2$</td>
<td>0.4721</td>
<td>1.5554</td>
<td>0.126</td>
<td></td>
</tr>
<tr>
<td>$x_3$</td>
<td>-0.4982</td>
<td>0.1520</td>
<td>0.002</td>
<td></td>
</tr>
<tr>
<td>$x_4$</td>
<td>3.486</td>
<td>2.274</td>
<td>0.132</td>
<td></td>
</tr>
</tbody>
</table>

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>$F$</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>4</td>
<td>887994</td>
<td></td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>Error</td>
<td>50</td>
<td>902773</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>54</td>
<td>902773</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Sequential SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>1</td>
<td>883880</td>
</tr>
<tr>
<td>$x_2$</td>
<td>1</td>
<td>183</td>
</tr>
<tr>
<td>$x_3$</td>
<td>1</td>
<td>3237</td>
</tr>
<tr>
<td>$x_4$</td>
<td>1</td>
<td>694</td>
</tr>
</tbody>
</table>

(a) (3 points) How many observations are in the data?

(b) (3 points) What is $R^2$ for this model?

(c) (3 points) What is the mean squared error?

(d) (3 points) Give a 95% confidence interval for $\beta_2$.

(e) (3 points) Test the null hypothesis $\beta_3=0$ with $\alpha = .05$ (use a two-sided test).

(f) (3 points) Give the $F$ statistic for testing the null hypothesis $\beta_3=0$.

(g) (3 points) Set up the test of the model with only variables $x_1$ and $x_2$ against the model with all of variables $x_1$, $x_2$, $x_3$, and $x_4$. Be sure to calculate the test statistic and describe which quantity you would read from a table or from R, including any necessary degrees of freedom.

(h) (3 points) Consider testing the model with only variables $x_1$ and $x_2$ against the model with variables $x_1$, $x_2$, and $x_3$. Should this test be the same as the one in the previous part? Why or why not?

(i) (3 points) For estimating the point on the regression surface at $(x_1, x_2, x_3, x_4) = (100,25,50,4)$, the standard error of the estimate for the point on the surface is 2.62. Give the estimated point on the surface, a 95% confidence interval for the point on the surface, and a 95% prediction interval for a new point with these $x$ values.

(j) (3 points) Test the null hypothesis $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ with $\alpha = 0.5$. 

3. Fred recruited 150 subjects from a local shopping mall, all age 18 to 70, asked them their sex, age, and years of education, and had them take a test of reaction time. The following variables were therefore available for each subject:

- **sex**: 0 = male, 1 = female
- **age**: The subject’s age, in years
- **education**: Years of education: eg, 16 = bachelor’s degree but no higher
- **time**: Average reaction time in the test, in seconds (lower is better)

Pairwise scatterplots of age, education, and time are shown in the figure below, with sex shown by the symbol used, and values for education randomly jittered to avoid overlap.

Fred fit a linear regression model for time with the other three variables as covariates, using the standard least squares method (Output 1). Fred thought that more education might indicate greater intelligence, and that greater intelligence and better (i.e., lower) reaction times might go together. Fred fit another regression model with only education as a covariate, (Output 2). Fred also fit a regression model with only age as a covariate (Output 3).

Answer the following questions about this study:

(a) (5 points) Is this an experiment or an observational study? What does your answer say about the ability of this study to answer questions about the causal factors influencing reaction time? Discuss this with reference to the specific variables measured in this study.

(b) (5 points) Do you see any unusual data points? What would be an appropriate action to take with regard to any such data points? Does it appear that the regression model fits have been strongly influenced by any such unusual point or points?

(c) (5 points) What conclusions do you think Fred should draw from these data? Discuss what can and cannot be concluded from the plots and model fits in output 1, 2, and 3, as well as your answers to (a) and (b). Discuss for each of the covariates (sex, age, and education) whether or not there is good reason to believe that that covariate is associated with time. In particular, discuss what the data show regarding Fred’s idea that greater years of education might be associated with lower reaction time.
Output 1

Call:
lm(formula = time ~ sex + age + education)

Residuals:
     Min      1Q  Median       3Q      Max
-0.11421 -0.04709 -0.01406  0.03942  0.36180

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)    
(Intercept) 0.2243380  0.0306130   7.328 1.47e-11 ***
  sex         0.0170760  0.0113831   1.500 0.13574      
  age         0.0010038  0.0003542   2.834 0.00524 **
education   0.0032729  0.0020417   1.603 0.11109      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06787 on 146 degrees of freedom
Multiple R-Squared: 0.08698,  Adjusted R-squared: 0.06822
F-statistic: 4.636 on 3 and 146 DF,  p-value: 0.003973

Output 2

Call:
lm(formula = time ~ education)

Residuals:
     Min      1Q  Median       3Q      Max
-0.11398 -0.05058 -0.01411  0.03568  0.37186

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)    
(Intercept) 0.265212  0.027756   9.555  <2e-16 ***
  education  0.003692  0.002009   1.838     0.0681 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06976 on 148 degrees of freedom
Multiple R-Squared: 0.02231,  Adjusted R-squared: 0.0157
F-statistic: 3.377 on 1 and 148 DF,  p-value: 0.06814
Output 3

Call:
ln(formula = time ~ age)

Residuals:
    Min     1Q  Median     3Q    Max
-0.12504 -0.04773 -0.01424  0.04140  0.37449

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.274167   0.0142250  19.270 < 2e-16 ***
age        0.0010964  0.0003498   3.134  0.00208 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06832 on 148 degrees of freedom
Multiple R-Squared: 0.06224,  Adjusted R-squared: 0.0559
F-statistic: 9.823 on 1 and 148 DF,  p-value: 0.002079
4. The general form of an over-dispersed exponential family distribution is given by:

\[ f_Z(z|\theta, \tau) = h(z, \tau) \exp \left( \frac{b(\theta)T(z) - A(\theta)}{d(\tau)} \right), \]

where \( z \) represents the data, \( \theta \) is the parameter vector of interest, and \( \tau \) is the dispersion parameter. We will consider this form as a generalization of the Poisson regression framework, in which:

\[ \lambda_{Y|x} = E(Y|x) = e^{\beta^T x} \]

\[ Y|x \sim Poisson(\lambda_{Y|x}). \]

Note that in the general exponential family expression (1), \( z \) represents all of the data, while in the regression expression (2), the data are partitioned into a vector of outcome variables \( y \) and a matrix of predictors, \( x \), which are considered fixed. Since \( z \) only represents the random variables, you may consider \( z = y \), and treat \( x \) as fixed constants.

(a) (5 points) Write the distribution for the Poisson regression model in (2) in the form in (1). Give expressions for \( b(\theta), T(y), A(\theta), d(\tau), \) and \( h(y, \tau) \) in terms of \( y, x \), and \( \beta \). You may do this for a single observation (rather than for a data set of many observations).

(b) (3 points) The dispersion parameter allows a distribution to have more or less variance, for fixed mean. Knowing what you know about the Poisson distribution and Poisson regression, why might this be helpful?

(c) (6 points) Consider the following examples where researchers would like to use Poisson regression as in (2). For each of the examples, consider whether you feel Poisson regression is appropriate. Choose the two examples which are most worrisome for using Poisson regression. For each of these, describe your concerns.

(i) An environmental scientist counts bird nests in 1-acre plots in many areas of Massachusetts, and models the count of bird nests as a function of features of the local environment.

(ii) A meteorologist counts the hours per day of rain (that is, the number of one-hour blocks in which there is more than 0 rain, so each day the count is an integer between 0 and 24) in Amherst and each of 6 surrounding towns for each day in 2017. She models the count of hours of rain as a function of daily high and low temperature, as well as features of the towns.

(iii) A physiologist uses 3 tests of balance to evaluate 30 research subjects. She models the number of tests ‘passed’ as a function of individual characteristics.
5. Let $X_1, \ldots, X_n$ be independent and identically distributed samples according to the Pareto distribution with parameters $a$ and $b$. Suppose that $b$ is known to be 1. We would like to compare the mean squared error (MSE) of two estimators of $a$, i.e., the maximum likelihood estimator (MLE) and the moment estimator. Note for a given sample $x_1, \ldots, x_n$, the maximum likelihood estimator $\hat{a}_1 = n/(\log x_1 + \ldots + \log x_n)$ and the moment estimator is $\hat{a}_2 = 1/(1 - 1/\bar{x})$ with $\bar{x} = \sum_{i=1}^n x_i/n$. Also note the Pareto distribution is related to the Exponential distribution, that is, if $U$ follows an Exponential distribution with mean $1/a$, then $X = b \exp(U)$ follows a Pareto distribution with parameters $a$ and $b$.

(a) (10 points) Design a simulation study to compare the MSE of the MLE and the moment estimator of $a$ when $b = 1$. To answer this part, please provide a conceptual outline of the steps needed for the simulation study.

(b) Write R code to implement the simulation. (5 points for workable code, 3 points for elegance and efficiency of the code.)

hint: You may wish to use the function `rexp(n, rate=[rate])`. There is no R function to sample from a Pareto distribution directly.