1. I have a cauldron of coins (pennies, nickels and dimes) in my office. I had $n = 26$ students each draw a small handful of coins and record $X_1$, the number of coins, $X_2$, the number of pennies, and $Y$, the total value in cents (data in the file coins.txt). They also recorded the weight in grams of their handfuls.

   a) Discuss the interpretation of simple and multiple regression coefficients for regressions of $Y$ on $X_1$ and $X_2$. Explain why we would be justified in fitting a no-intercept model for some of these regressions. Also consider regressions of the weights on the numbers of different coins.

   b) For a less contrived example, consider the Centennial Conference men’s basketball data in CCBB.txt. Each row represents one season for one of the ten teams. Let $Y$ be the winning percentage (proportion of conference games won in the season), $X_1$ the field goal percentage (proportion of successful shots, excluding free throws) and $X_2$ the average number of offensive rebounds (recovering the ball after missing a shot) per game. Show that this yields another instance of “Simpson’s paradox”, as well as a much more unusual quality: the $R^2$ value for the multiple regression is larger than the sum of the $R^2$ values from the two simple regressions. Odd things happen with turnover-margin as well.

   c) Show how to get the multiple regression coefficients from simple regression output. Fit a simple regression of $Y$ on $X_1$ and save the residuals. Then fit a regression of $X_2$ on $X_1$ and save the residuals. Finally, fit a simple regression of the first set of residuals on the second set of residuals. The resulting slope is the slope of $X_2$ in the multiple regression. An analogous procedure gives the coefficient of $X_1$ in the multiple regression. You can find the correct standard errors by adjusting to the correct degrees of freedom for the multiple regression.

   d) Find a good model for predicting a team’s winning percentage from statistics other than points scored (once you include points scored, no other statistics will seem important at all). There is a stepwise regression procedure that automates the choice of variables. Discuss what it means for a predictor to be important in the context of a multiple regression (and what it means to “find a good model”). Explain the difference between $R^2$ and adjusted $R^2$.

   e) Find a confidence interval for the mean winning percentage for teams with a particular set of covariates (i.e., specific values of the variables you included in your model). Also find a prediction interval for a particular team with these statistics.

2. The ANOVA table and $F$ tests for regression.

   a) Show that a 1—sample t analysis is equivalent to linear regression on a constant, and that a pooled 2-sample t analysis is equivalent to linear regression on a variable that takes values 0 and 1 (an indicator variable). Use the data for height and gender as an example.

   b) Describe the ANOVA table in the context of regression on indicator variables. Use CCBB data and compare winning percentage by team. Describe the whole-model $F$-test and find an equivalent test based on $R^2$. 

c) Analysis of Covariance (ANCOVA) refers to a multiple regression with a numeric response $Y$, and with indicator variables as well as numeric variables as $X$’s, along with possible interactions. As an example, consider winning percentage regressed on team and field goal percentage allowed (FG% Def), with and without an interaction.

d) Explain the “Extra Sum of Squares $F$ test” for multiple coefficients, such as categorical predictors with more than two categories (or interactions with such categorical predictors).

3. Swarthmore Biology professor Sara Hiebert Burch came to me with this problem. Theory predicts that, within a certain range of temperature values, the metabolic rate $Y_i$ of a hamster will decrease linearly with temperature until a lower critical temperature $\theta$ is reached. At that point, the decrease remains linear, but with a less negative slope. For each hamster, measurements of metabolic rate $Y_i$ are made at $n = 7$ temperature values $x_i$ ($i = 1, \ldots, n$). Consider the following model:

$$
Y_i = \mu_i + \epsilon_i; \quad \epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2); \quad \mu_i = \begin{cases} 
\beta_0 + \beta_1 x_i, & x_i < \theta, \\
\beta_0 + \beta_1 x_i + \delta(x_i - \theta), & x_i \geq \theta.
\end{cases}
$$

a) Write out the likelihood function for the five unknown parameters.

b) For a given value of $\theta$, find expressions for the maximum likelihood estimates of $\beta_0$, $\beta_1$, $\delta$ and $\sigma^2$. You may express the estimate of $\sigma$ in terms of the other estimates, and all will depend on $\theta$.

c) The table below gives temperatures (degrees C) and metabolic rate measurements for one of the hamsters.

<table>
<thead>
<tr>
<th>Temp</th>
<th>5.3</th>
<th>11.8</th>
<th>19.3</th>
<th>23.8</th>
<th>25.6</th>
<th>27.6</th>
<th>31</th>
</tr>
</thead>
<tbody>
<tr>
<td>MR</td>
<td>6.7</td>
<td>4.4</td>
<td>3.3</td>
<td>2.12</td>
<td>2.02</td>
<td>1.82</td>
<td>1.74</td>
</tr>
</tbody>
</table>

Graph the metabolic rate values against temperature and draw in the best fitting lines (based on the MLE) for $\theta = 13.59$ and for $\theta = 23.24$.

d) Make a graph of the maximized likelihood function against a grid of $\theta$ values between $t_1 = 12$ and $t_2 = 27$ (imagine you have theoretical justification for setting those boundaries). If you scale the likelihood values to have a maximum value of 1.0, this is a graph of the GLR statistic for testing different values of $\theta$. Which of the values from part c is more consistent with the data according to this criteria?

**Problems to turn in:**

1. Fit a regression of height on shoe length, male and an interaction:
   ```
   \texttt{out = lm(height ~ shoe + male + shoe*male)}
   ```
   The interaction variable is literally the product of male and shoe length (0 for female, shoe length for males). Create a graph of the data with the two fitted lines, along with a residual plot. State hypotheses and carry out a test to determine whether an interaction is needed in the model. Explain the conclusion in the context of the example (i.e., don’t just say, “reject $H_0$” or “fail to reject $H_0$”).

2. Find the sampling distribution of $\hat{\beta}$, the MLE for the weighted regression model ($W$ is a known symmetric $n \times n$ positive definite matrix and $X$ is $n \times p$):
   $$
   Y = X\beta + \epsilon, \quad \epsilon \sim N_n(0, \sigma^2W^{-1}).
   $$