In your writeups, we expect clear explanations of models chosen, hypotheses tested, and findings analogous to what you would produce for a consulting project.

1. **(Concepts)** Answer the following questions, referring to concrete examples for support.

   (a) The coefficient of determination $R^2$ for a simple linear model describes the proportion of variance in the data that the model explains. When $R^2$ is close to one, must the simple linear model be an adequate explanation of the data? Why or why not?

   (b) When we perform hypothesis tests that the intercept and slope of a simple linear model are nonzero, we assume that these coefficients, once standardized, have a $t$-distribution. What properties of the data are required to ensure that this assumption is reasonable? Describe at least two distinct cases.

   (c) When we calculate a prediction interval, we assume that the standardized difference between the fitted value and the response has a $t$-distribution. What properties of the data are required to ensure that this assumption is reasonable?

   (d) Suppose that we add a new term to a simple linear model. What can we say about the residual sum of squares in the larger model? Does the larger model necessarily provide a more adequate explanation for the data? Why or why not?

   (e) Suppose that we have paired data $(x_i, y_i)$. Consider the model

   $$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i.$$  

   Why does this model remain linear? What can this model explain that the simple linear model does not? When does it make sense to add a term of the form $x^2$?

   (f) Suppose that we have paired data $(x_i, y_i)$ with $x_i$ distinct for $i = 1, 2, \ldots, n$. One can show that the model

   $$\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_n x_i^{n-1}$$

   for the mean response can explain all $n$ observations perfectly. Why is this model likely to be a poor choice?

   (g) Consider a linear model

   $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p.$$  

   Describe a concrete example of a response variable and at least three terms that might be relevant to modeling the mean response. You might look to your own research for inspiration. Answer the following questions with reference to your example. What does the coefficient $\beta_0$ represent? What does it mean for the coefficient $\beta_1$ to be positive? Negative? If we increase the value of the variable $X_p$ by one while holding the other variables fixed, what does the model tell us about the mean response?
(h) We developed a formula for the LS estimate of the coefficient vector in a linear model:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

where $Y$ is the vector of responses and $X$ is the data matrix. Is this formula valid when the number of variables (there are $p$ variables plus the constant variable for the intercept) exceeds the number of observations $n$? Why or why not? Are there circumstances where this formula does not make sense?

2. (We few, we happy few). The satisfaction with life survey (SWLS) is known to be a robust measure of well-being (Pavot and Diener, Psychological Assessment, 1993). We will attempt to predict national SWLS scores using social and economic data. For 36 countries, the file swls-data.txt contains information on life satisfaction, material well-being, health, political stability, family life, and climate. Respectively, the specific variables are satisfaction with life (swls) in comparison with the worst country (Burundi = 100), GDP in dollars per capita (gdp), life expectancy in years (le), political stability (pol) on a scale from worst (-2.5) to best (+2.5), annual divorce rate per 1,000 population (fam), and latitude in degrees (lat).

(a) Make a scatterplot matrix to compare each variable against the others. Are there any variables that look especially valuable for predicting swls? Any that seem irrelevant?

(b) Which predictor has the strongest correlation with the response variable? Fit a simple linear model using this predictor. Make a scatterplot and add the regression line. What proportion of the variance does this predictor explain?

(c) Which predictor has the strongest correlation with the residuals from the model in (b)? Make an added-variable plot to show how well this predictor explains the residuals.

(d) Fit a linear model for swls using the two predictors you have identified. Is each of the three terms in this model statistically significant? Compute a 95% confidence interval for each coefficient.

(e) In nontechnical language, what does the intercept in the model from (d) represent? What about the other two coefficients? What do the signs of the coefficients mean? Their magnitudes? Please be concrete.

(f) What proportion of the total variance does the model in (d) explain? Perform an overall analysis of variance. Is the improvement over the constant model

$$Y = \beta_0 + \epsilon$$

statistically significant?

(g) Use analysis of variance to compare the models from (b) and (d). Is the improvement in the larger model statistically significant?

(h) Fit a model for swls using all available predictors. Is each term statistically significant? Is this result surprising in view of the scatterplot matrix in (a)? Why or why not?

(i) What does an overall analysis of variance tell us about this model in comparison with the constant model? How does it compare with the model in (d)?

Consider the hypothetical Gondwanaland whose state of affairs is summarized by

$$\text{gdp} = \$12,100 \text{ per capita}, \quad \text{le} = 47 \text{ yrs}, \quad \text{pol} = +2.0,$$
$$\text{fam} = 1.1 \text{ per 1,000}, \quad \text{pop lat} = 12^\circ.$$
(j) According to the model in (d), what is the expected value of swls in Gondwanaland? Compute a 95% prediction interval.

(k) According to the model in (h), what is the expected value of swls in Gondwanaland? Compute a 95% prediction interval. How does this compare with (j)? Why?

(l) Using the model from (d) and then from (h), compute 95% confidence intervals for the mean swls in countries that share the same statistics as Gondwanaland. Compare.