1. Consider a sample on n subjects divided into two groups indicated by $X_i = 0$ for $i = 1, \ldots, n/2$ and $X_i = 1$ for $i = n/2 + 1, \ldots, n$, and suppose each subject has measurement $Y_i \in \mathbb{R}$.

   (a) Consider the regression model $Y = \beta_0 + X\beta_1 + \varepsilon$, for $\varepsilon \sim N(0, \sigma^2 I_n)$. Compute and simplify expressions for $\hat{\beta}_0$ and $\hat{\beta}_1$.

   (b) Compute and simplify expressions for $\hat{\text{se}}(\hat{\beta}_0)$ and $\hat{\text{se}}(\hat{\beta}_1)$.

   (c) Compute an expression for the Wald test of $H_0 : \beta_1 = 0$. Interpret the null hypothesis of this test.

   (d) Consider the reparamterized form of the model $Y = \alpha_0 I(X = 0) + \alpha_1 I(X = 1) + \varepsilon$. State the test in terms of these parameters that is equivalent to $H_0 : \beta_1 = 0$. (Note that you do **not** need to calculate the expressions for $\hat{\alpha}_0$, $\hat{\alpha}_1$.)

2. Use the *PlantGrowth* data in `R` for this problem. (Load with `data("PlantGrowth")`.)

   (a) Create side-by-side boxplots of the distribution of weights in each group. What groups do you anticipate having significantly different mean weights?

   (b) Fit a linear regression model to the Plant Growth data with *weight* as the response and *group* as the predictor.

   (c) Calculate a 95% confidence interval for the mean weight in the control group.

   (d) Calculate a 95% confidence interval for the difference in effect between treatments 1 and 2.

   (e) Use `plot()` on your fitted model to examine the diagnostic plots. Comment on any potential violations of the model assumptions.

3. Use the U.S. government data on the nationwide, county-level environmental quality index available at `https://edg.epa.gov/data/Public/ORD/NHEERL/EQI/Eqidata_all_domains_2014March11.csv`. We are interested in quantifying the effects of unemployment, education levels, and lead concentration on the violent crime rate.

   You can load the data into `R` using

   ```
   dat <-
   read_csv("https://edg.epa.gov/data/Public/ORD/NHEERL/EQI/Eqidata_all_domains_2014March11.csv")}
   ```

   (a) Create histograms of `violent_rate_log`, `pct_hs_more`, `pct_unemp`, and `mean_pb_ln`. Do you notice anything problematic?

   (b) Create scatterplots of `violent_rate_log` against the three other variables.

   (c) Regress `violent_rate_log` on the other three variables and print the summary of the fit.

   (d) Compute 95% confidence intervals for the covariates that are significant at the $\alpha = 0.05$ level.

   (e) Interpret the significant coefficients in real-world terms.

   (f) Use the `plot` function to produce diagnostics plots of the fitted model. Comment on any potential issues.

   (g) Do you think the linear regression model is appropriate for these data? Are any of the regression model assumptions likely violated? Do you think the coefficient confidence intervals you computed are valid?

   (h) Can you think of a more appropriate model for these data?