1. Consider the following linear regression model with two predictor variables ($\mathbf{x}_1$ and $\mathbf{x}_2$)

$$\mathbf{y} \sim N(\beta_0\mathbf{1} + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2, \sigma^2\mathbf{I}).$$

If there are n observations in $\mathbf{y}$ (and similarly in the predictor variables), define a $n \times n$ symmetric matrix

$$\mathbf{C} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'$$

and a related regression model

$$\mathbf{y} \sim N(\alpha_0\mathbf{1} + \alpha_1\mathbf{w}_1 + \alpha_2\mathbf{w}_2, \tau^2\mathbf{I})$$

where

$$\mathbf{w}_1 = \mathbf{C}\mathbf{x}_1 \quad , \quad \mathbf{w}_2 = \mathbf{C}\mathbf{x}_2.$$

(a) Explain in words how $\mathbf{w}_1$ is related to $\mathbf{x}_1$.

(b) Express $\alpha_0, \alpha_1, \alpha_2$ and τ^2 in terms of $\beta_0, \beta_1, \beta_2$ and σ^2. How does transforming the predictor variables in this way change the interpretation of the regression parameters?

(c) Now consider scaling each predictor variable. Let $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$ be the scaled predictor variables defined by:

$$x_{i1}^* = \frac{x_{i1}}{c_1} \quad , \quad x_{i2}^* = \frac{x_{i2}}{c_2}$$

where c_1 and c_2 are nonzero real numbers. The corresponding regression model is

$$\mathbf{y} \sim N(\delta_0\mathbf{1} + \delta_1\mathbf{x}_1^* + \delta_2\mathbf{x}_2^*, \kappa^2\mathbf{I})$$

Express $\delta_0, \delta_1, \delta_2$ and κ^2 in terms of $\beta_0, \beta_1, \beta_2$ and σ^2. How does transforming the predictor variables in this way change the interpretation of the regression parameters?

2. Consider a linear regression model with response variable $\mathbf{y}$ and two predictor variables: a continuous covariate ($\mathbf{x}$ = age) and a categorical covariate ($\mathbf{z}$ = gender). Consider the following six models. In each model, the error terms ϵ_i are assumed to be independent, identically-distributed normal random variables.

Model A

$$y_i = \begin{cases} \theta_0 + \epsilon_i & \text{, for men } (z_i = M) \\ \theta_1 + \epsilon_i & \text{, for women } (z_i = W) \end{cases}.$$

Model B

$$y_i = \theta_0 + \theta_1 x_{i1} + \epsilon_i$$

Model C

$$y_i = \begin{cases} \theta_0 + \theta_1 x_i + \epsilon_i & \text{, for men } (z_i = M) \\ \theta_2 + \theta_1 x_i + \epsilon_i & \text{, for women } (z_i = W) \end{cases}.$$

Model D

$$y_i = \begin{cases} \theta_0 + \theta_1 x_i + \epsilon_i & \text{, for men } (z_i = M) \\ \theta_0 + \phi_0 + \theta_1 x_i + \epsilon_i & \text{, for women } (z_i = W) \end{cases}.$$

Model E

$$y_i = \begin{cases} \theta_0 + \theta_1 x_i + \epsilon_i & \text{, for men } (z_i = M) \\ \theta_0 + \theta_2 x_i + \epsilon_i & \text{, for women } (z_i = W) \end{cases}.$$

Model F

$$y_i = \begin{cases} \theta_0 + \theta_1 x_i + \epsilon_i & \text{, for men } (z_i = M) \\ \theta_2 + \theta_3 x_i + \epsilon_i & \text{, for women } (z_i = W) \end{cases}.$$

(a) Rewrite each model as a linear regression model without cases by introducing a dummy variable $z_i = 0$ for men and $z_1 = 1$ for women. In each case, provide a careful interpretation of the parameters in the model and give a description of the assumed form of the relationship between the predictor and response variables.

(b) Which of the six models are equivalent?

3. **Particle Displacement.** The mean-squared displacement (MSD) of a particle in a fluid with position $x(t)$ at time t is defined as $MSD(t) = E(x^2(t))$. In many fluids, the MSD scales according to the power law

$$MSD(t) = \gamma \cdot t^{\alpha}.$$

A reseacher interested in estimating α for a particular fluid has conducted a number of trials in which she measured a particle's final displacement $x(t)$ for a fixed time t. This data is given in the "fluid.csv" file. Use the data to estimate α, the exponent in the power law relationship between time and MSD. Carefully explain the model you fit, and consider the assumptions of the model.

4. **Olympic 100m Gold Medal Times.** Read in the data in the "Olympics.csv" file from Angel.

```
oly=read.csv("Olympics.csv")
str(oly)
```

The "oly" data frame contains three columns. "year" is the year of the Olympic games, "goldtime" is the winning gold-medal time in the 100m race for that Olympics, and "gender" indicates whether the winning time is for the mens race ("M") or the womens race ("W").

(a) Fit the following regression model to the data:

$$\text{goldtime}_i = \text{year}_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

Report the fitted model, and comment on whether the model seems to be appropriate. Interpret the effect of year on gold medal time in the 100m race.

(b) Now fit a regression model with "goldtime" as response and both "year" and "gender" as predictor variables. Clearly write the model you are fitting. You should consider transformations and interactions and find a model that you think is appropriate for the data. Report only your final model, with estimated parameters. Comment on whether the model seems to be appropriate. Interpret the effects of the predictor variables on the gold medal time in the 100m race.

(c) The Olympic games were not held in 1944. Use your model from part (b) to predict the expected gold medal times for the mens and womens 100m races, if those races had been held.