1. **Split Dataset** Let $\mathbf{y}$ and $\mathbf{X}$ contain n observations and come from the true model:

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}).$$

Consider subdividing the data into two pieces :

$$\mathbf{y}_1 = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n_1} \end{pmatrix}, \mathbf{X}_1 = \begin{pmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_{n_1}' \end{pmatrix}$$

$$\mathbf{y}_2 = \begin{pmatrix} y_{n_1+1} \\ y_{n_1+2} \\ \vdots \\ y_{n_1+n_2} \end{pmatrix}, \mathbf{X}_2 = \begin{pmatrix} \mathbf{x}_{n_1+1}' \\ \mathbf{x}_{n_1+2}' \\ \vdots \\ \mathbf{x}_{n_1+n_2}' \end{pmatrix}$$

where $n_1 + n_2 = n$.

Let $\hat{\boldsymbol{\beta}}_1 = [\hat{\beta}_{11}, \hat{\beta}_{12}, \ldots, \hat{\beta}_{1p}]'$ be the estimated regression parameters obtained by fitting the linear model

$$\mathbf{y}_1 \sim N(\mathbf{X}_1\boldsymbol{\beta}, \sigma^2\mathbf{I})$$

and $\hat{\boldsymbol{\beta}}_2[\hat{\beta}_{21}, \hat{\beta}_{22}, \ldots, \hat{\beta}_{2p}]'$ be the estimate obtained by fitting the linear model

$$\mathbf{y}_2 \sim N(\mathbf{X}_2\boldsymbol{\beta}, \sigma^2\mathbf{I}).$$

(a) What is the distribution of $\theta_k = \hat{\beta}_{1k} - \hat{\beta}_{2k}$?

(b) Construct a hypothesis test to test:

$$H_0: \ \theta_k = 0 \quad vs. \quad \theta_k \neq 0.$$

Clearly state your test statistic, the distribution of the test statistic, and the conditions needed to reject the null hypothesis at the $\alpha = .05$ level.

2. **Estimating the Mean with a Smoother**. Let $\mathbf{y} = [y_1, y_2, \ldots, y_n]'$ be n observations from the true model:

$$y_t = \beta_0 + \beta_1 \cdot t + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma^2)$$

with $\epsilon_i \perp\!\!\!\perp \epsilon_j$ whenever $i \neq j$.

Consider estimating $E(y_u)$ by smoothing (averaging over the data at nearby locations). Let your estimate of the mean of y at the observed location u be:

$$\widehat{E(y_u)}_{sm} = \frac{1}{4}y_{u-1} + \frac{1}{2}y_u + \frac{1}{4}y_{u+1}.$$

You may assume that u is NOT equal to 1 or n. Construct a 95% confidence interval for $E(y_u)$ using this estimate. Clearly describe any estimates you use and justify your confidence interval.

3. **Anscombe data sets**. Consider the "anscombe" data in R:

```
data(anscombe)
str(anscombe)
?anscombe
```

This data set contains four pairs of predictor (x_1, x_2, x_3, x_4) and response (y_1, y_2, y_3, y_4) variables.

(a) Fit a simple linear regression model to the first pair of variables: $\{\mathbf{x_1}, \mathbf{y_1}\}$:

$$\mathbf{y_1} \sim N(\mu_1 \mathbf{1} + \beta_1 \mathbf{x_1}, \sigma_1^2 \mathbf{I}).$$

Repeat for the other three pairs of predictor and response variables. Report the estimates for regression parameters for all four pairs. You do not need to do any residual analysis. Just fit the models and estimate the regression parameters.

(b) Analyze each of the four pairs of x and y variables within a linear modeling framework. You should consider transformations, nonlinear relationships, potential outliers, influential points, and residual analysis. Provide any plots you think are appropriate, but do not include plots that you do not discuss.

(c) Use your analysis from (b) to predict y at $x = 13$ using the fitted models. Provide an expected value of y and a 95% prediction interval for y for each of the four fitted models when $x = 13$. Comment on whether you think you can trust each of the prediction intervals you have constructed.