Please submit HW4
Reminder
\[ Y = \beta_0 + \beta_1 X + \epsilon \]

**Figure 11-1** Scatter diagram of oxygen purity versus hydrocarbon level from Table 11-1.

\[ Y = 75 + 15 \cdot X + \epsilon \]
$Y = \beta_0 + \beta_1 X + \epsilon$; $E(\epsilon | x) = 0 \forall x$

How does one find $\beta_0$ & $\beta_1$?

$\text{Cov}(Y, X) = \text{Cov}((\beta_0 + \beta_1 X + \epsilon), X) = \text{Cov}(\beta_0, X) + \beta_1 \text{Cov}(X, X) + \text{Cov}(\epsilon, X)$

$\text{Cov}(\beta_0, X) = 0$ since $\beta_0$ is constant

$\text{Cov}(X, X) = E(X^2) - E(X)^2 = \text{Var}(X)$

$\text{Cov}(\epsilon, X) = E(\epsilon \cdot X) - E(\epsilon) \cdot E(X) = E(\epsilon \cdot X) = \sum_{\text{all } x} x \cdot E(\epsilon | x) = 0$

Thus $\beta_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}$; $\beta_0 = E(Y) - \beta_1 E(X)$
Method of least squares

• The **method of least squares** is used to estimate the parameters, $\beta_0$ and $\beta_1$ by minimizing the sum of the squares of the vertical deviations in Figure 11-3.

*Figure 11-3* Deviations of the data from the estimated regression model.
Traditional notation

**Definition**

The **least squares estimates** of the intercept and slope in the simple linear regression model are

\[
\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (11-7)
\]

\[
\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n} = \frac{S_{xy}}{S_{xx}} \quad (11-8)
\]

where \( \bar{y} = (1/n) \sum_{i=1}^{n} y_i \) and \( \bar{x} = (1/n) \sum_{i=1}^{n} x_i \).
11-2: Simple Linear Regression

**Definition**

The **least squares estimates** of the intercept and slope in the simple linear regression model are

\[
\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{11-7}
\]

\[
\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \left( \sum_{i=1}^{n} y_i \right) \left( \sum_{i=1}^{n} x_i \right)}{n \left( \sum_{i=1}^{n} x_i^2 - \frac{\left( \sum_{i=1}^{n} x_i \right)^2}{n^2} \right)} = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \tag{11-8}
\]

where \( \bar{y} = (1/n) \sum_{i=1}^{n} y_i \) and \( \bar{x} = (1/n) \sum_{i=1}^{n} x_i \).
11-4: Hypothesis Tests in Simple Linear Regression

11-4.2 Analysis of Variance Approach to Test Significance of Regression

The analysis of variance identity is

\[
\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2
\]  

(11-24)

Symbolically,

\[
SS_T = SS_R + SS_E
\]

(11-25)
11-7: Adequacy of the Regression Model

11-7.2 Coefficient of Determination ($R^2$)

**VERY COMMONLY USED**

• The quantity

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

is called the **coefficient of determination** and is often used to judge the adequacy of a regression model.

• $0 \leq R^2 \leq 1$;

• We often refer (loosely) to $R^2$ as the amount of variability in the data explained or accounted for by the regression model.
11-7: Adequacy of the Regression Model

11-7.2 Coefficient of Determination ($R^2$)

- For the oxygen purity regression model,
  \[ R^2 = \frac{SS_R}{SS_T} = \frac{152.13}{173.38} = 0.877 \]
- Thus, the model accounts for 87.7% of the variability in the data.
11-2: Simple Linear Regression

Estimating $\sigma^2_{\epsilon}$

An unbiased estimator of $\sigma^2_{\epsilon}$ is

$$\hat{\sigma}^2 = \frac{SS_E}{n - 2} \tag{11-13}$$

where $SS_E$ can be easily computed using

$$SS_E = SS_T - \hat{\beta}_1 S_{xy} \tag{11-14}$$
11-3: Properties of the Least Squares Estimators

- **Slope Properties**

  \[ E(\hat{\beta}_1) = \beta_1 \]

  \[ V(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{S_{xx}} = \frac{\hat{\sigma}_\varepsilon^2}{n \hat{\sigma}_\varepsilon^2} \]

- **Intercept Properties**

  \[ E(\hat{\beta}_0) = \beta_0 \quad \text{and} \quad V(\hat{\beta}_0) = \hat{\sigma}_\varepsilon^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right] = \hat{\sigma}_\varepsilon^2 \left[ 1 + \frac{\mu^2}{\hat{\sigma}_\varepsilon^2} \right]^{\frac{1}{2}} \frac{1}{n} \]
11-4.1 Use of $t$-tests for smaller $n$.

The number of degrees of freedom is $n-2$.

One can always fit a straight line through two points so one needs $n \geq 3$.
11-4: Hypothesis Tests in Simple Linear Regression

\[ H_0: \beta_1 = 0 \]
\[ H_1: \beta_1 \neq 0 \]

\[ T = \frac{\hat{\beta}_1}{\hat{\sigma}_x \sqrt{n}} \]

Choose \( \alpha \) (e.g., \( \alpha = 0.05 \))

Confidence in rejecting \( H_0 \)

is such

\[ t_{\alpha/2, n-2} = \pm \text{cdf}(t_{\alpha/2, n-2}) \]

\[ 1 - \frac{\alpha}{2} = \pm \text{cdf}(t_{\alpha/2, n-2}) \]

Reject \( H_0 \) if \( |Z| = \left| \frac{\hat{\beta}_1}{\hat{\sigma}_x \sqrt{n}} \right| > t_{\alpha/2, n-2} \)
Gene Expression “Wheel of Fortune”

• Each group gets a pair of genes that are known to be correlated.
• Each group also gets a random pair of genes selected by the “Wheel of Fortune”. They may or may not be correlated.
• Download (log-transformed) expression_table.mat
• Run command fitlm(x,y) on assigned and random pairs
• Record $\beta_0$, $\beta_1$, $R^2$, P-value of the slope $\beta_1$ and write them on the blackboard
• Validate Matlab result for $R^2$ using your own calculations
• Look up gene names (see gene_description in your workspace) and write down a brief description of biological functions of genes. Does their correlation make biological sense?
Correlated pairs
plausible biological connection based on short description

\[
g1=1994; \quad g2=188; \quad \text{groups 1, 6}
g1=2872; \quad g2=1269; \quad \text{groups 2, 7}
g1=1321; \quad g2=10; \quad \text{groups 3, 8}
g1=886; \quad g2=819; \quad \text{groups 4, 9}
g1=2138; \quad g2=1364; \quad \text{groups 5, 10}
\]

no obvious biological common function

\[
g1=1+\text{floor}(\text{rand.}*3000); \quad g2=1+\text{floor}(\text{rand.}*3000);
disp([g1, g2])
\]
Random pairs

>> g1=floor(3000.*rand)+1; g2=floor(3000.*rand)+1; disp([g1,g2]);

>> g1=floor(3000.*rand)+1; g2=floor(3000.*rand)+1; disp([g1,g2]);

>> g1=floor(3000.*rand)+1; g2=floor(3000.*rand)+1; disp([g1,g2]);

>> g1=floor(3000.*rand)+1; g2=floor(3000.*rand)+1; disp([g1,g2]);
load expression_table.mat

\( g1=2907; \ g2=288; \)

\( x=\text{exp}\_t(g1,:); \ \ y=\text{exp}\_t(g2,:); \)

figure; plot(x,y,'ko');

\( \text{lm} = \text{fitlm}(x,y) \)

\( y\_fit=\text{lm} . \text{Fitted}; \)

hold on; plot(x,lm.Fitted,'r-');

\( \text{SST}=\text{sum}((y-\text{mean}(y)).^2); \)

\( \text{SSR}=\text{sum}((y\_fit-\text{mean}(y)).^2); \)

\( \text{SSE}=\text{sum}((y-y\_fit).^2); \)

\( \text{R}2=\text{SSR}./\text{SST} \)

\( \text{disp}([\text{gene}\_\text{names}(g1), \ \text{gene}\_\text{names}(g2)]); \)

\( \text{disp}(\text{gene}\_\text{description}(g1)); \ \text{disp}(\text{gene}\_\text{description} (g2)); \)
HW5 has been posted.
It is due next Tuesday, May 4
Multiple Linear Regression
(Chapters 12-13 in Montgomery, Runger)
12-1: Multiple Linear Regression Model

12-1.1 Introduction

• Many applications of regression analysis involve situations in which there are more than one regressor variable.
• A regression model that contains more than one regressor variable is called a **multiple regression model**.
Multiple Linear Regression Model

\[ Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + \varepsilon \]

One can also use powers and products of other variables instead of \( x_3, \ldots, x_k \).

Example: the general two-variable quadratic regression has 6 constants:
\[ Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1)^2 + \beta_4 (x_2)^2 + \beta_5 (x_1 x_2) + \varepsilon \]
12-1: Multiple Linear Regression Model

12-1.3 Matrix Approach to Multiple Linear Regression

Suppose the model relating the regressors to the response is

\[ y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \quad i = 1, 2, \ldots, n \]

In matrix notation this model can be written as

\[ y = X\beta + \varepsilon \quad (12-6) \]
12-1: Multiple Linear Regression Model

12-1.3 Matrix Approach to Multiple Linear Regression

where

\[ y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \text{and} \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \]
12-1.3 Matrix Approach to Multiple Linear Regression

We wish to find the vector $\hat{\beta}$ that minimizes:

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon'\varepsilon = (y - X\beta)'(y - X\beta)$$

$$0 = \frac{\partial L}{2\partial \beta} = -X'(y - X\beta) = -X'y + (X'X)\beta$$

The resulting least squares estimate is

$$\hat{\beta} = (X'X)^{-1}X'y$$

(12-7)

Analog of $\frac{1}{\text{Var}(x)}$ and $\text{Cov}(x,y)$.
12-1: Multiple Linear Regression Models

Estimating error variance $\sigma^2_e$

An unbiased estimator of error variance $\sigma^2_e$ is

$$
\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^{n} e_i^2 = \frac{SSE}{n-p}
$$

(12-16)

Here $p=k+1$ for $k$-variable multiple linear regression
R² and Adjusted R²

The coefficient of multiple determination R²

\[
R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}
\]

The adjusted R² is

\[
R_{adj}^2 = 1 - \frac{SS_E/(n - p)}{SS_T/(n - 1)}
\] (12-23)

- The adjusted R² statistic penalizes adding terms to the MLR model.
- It can help guard against overfitting (including regressors that are not really useful)
How to know where to stop?

• Adding new variables $x_i$ to MLR watch the adjusted $R^2$

• Once the adjusted $R^2$ no longer increases = stop.
  Now you did the best you can.