# ENGR 599 Homework 4

## Adrian Henle

```
using DataFrames    , PlutoUI
```

# Exercise 5.2

Show that $\mathbf{X}^T\mathbf{X} = \begin{bmatrix} n & \sum \mathbf{X}_i \\ \sum \mathbf{X}_i & \sum \mathbf{X}_i^2 \end{bmatrix}$ and $\mathbf{X}^T\mathbf{y} = \begin{bmatrix} \sum y_i \\ \sum Xiy_i \end{bmatrix}$

The matrix $\mathbf{X}$ is $\mathbf{X} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ X_1 & X_2 & \dots & X_n \end{bmatrix}$

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ X_1 & X_2 & \dots & X_n \end{bmatrix} \times \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}$$

$\mathbf{X}^T\mathbf{X}_{1,1}$ is the sum of the element-wise product of two $n$-dimensional unit vectors, so it equals $n$.

The off-diagonal elements are the sum of the element-wise product of an $n$-dimensional unit vector and the vector of values $X$, which is $X$ itself.

$\mathbf{X}^T\mathbf{X}_{2,2}$ is the element-wise product of the vector of values $X$ with itself.

$$\mathbf{X}^T\mathbf{y} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ X_1 & X_2 & \dots & X_n \end{bmatrix} \times \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

The first term of $\mathbf{X}^T\mathbf{y}$ is the sum of $\mathbf{y}$, and the second term is the sum of the element-wise product of $\mathbf{X}$ and $\mathbf{y}$.

# Exercise 5.4

Distinguish the models that are linear from those that are not.

*(a)* $y = b_o \sin X + b_1 \cos(b_2 X)$

nonlinear

*(b)* $pV = constant$

linear

*(c)* $\frac{\bar{V}}{RT} = 1 + B'P + C'P^2 + D'P^3 + ...$

linear

*(d)* $2^k$ factorial design model

linear

*(e)* $[A] = [A]_0 e^{-kt}$

nonlinear

# Exercise 5.6

The fit of a statistical model to the results of a factorial design can also be done by the least-squares method, solving $\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.

With $y = \begin{bmatrix} 59 \\ 90 \\ 54 \\ 68 \end{bmatrix}$ and $\mathbf{X} = \begin{bmatrix} 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$, we can solve for $\mathbf{b}$:

```
[67.75, 11.25, -6.75, -4.25]
```

```
begin
    y = [59; 90; 54; 68]
    X = [1 -1 -1 1; 1 1 -1 -1; 1 -1 1 -1; 1 1 1 1]
    b = inv(X' * X) * X' * y
end
```

# Exercise 5.8

Start from **Eq. 5.9** and show that $b_1 = \frac{\sum(X_i - \bar{X})y_i}{S_{xx}}$

**Eq. 5.9**:

$$b_1 = \frac{\sum(X_i - \bar{X})(y_i - \bar{y})}{\sum(X_i - \bar{X})^2}$$

$$= \frac{\sum(X_i - \bar{X})y_i - \bar{y}\sum(X_i - \bar{X})}{\sum(X_i - \bar{X})^2} = \frac{\sum(X_i - \bar{X})y_i}{\sum(X_i - \bar{X})^2}$$

$$S_{xx} = \sum(X_i - \bar{X})^2$$

$$b_1 = \frac{\sum(X_i - \bar{X})y_i}{S_{xx}} \therefore$$

# Exercise 5.10

Complete this table and verify if the linear model is satisfactory:

| Source of variation | Sum of squares | Degree of freedom | Mean square |
|---|---|---|---|
| Regression | $2.95146 \times 10^{-1}$ | ? | ? |
| Residual | ? | ? | ? |
| Lack of fit | ? | ? | ? |
| Pure error | $1.09355 \times 10^{-4}$ | ? | ? |
| Total | $2.95425 \times 10^{-1}$ | 11 | |
| % explained variation: ? | | | |
| Maximum % explainable variation: ? | | | |

| | Source of variation | Sum of squares | Degree of freedom | Mean square |
|---|---|---|---|---|
| 1 | "Regression" | 0.295145 | 1 | 0.295146 |
| 2 | "Residual" | 0.000279 | 10 | 2.8e-5 |
| 3 | "Lack of fit" | 0.00017 | 4 | 4.3e-5 |
| 4 | "Pure error" | 0.000109355 | 6 | 1.8e-5 |

```
DataFrame(
    "Source of variation" => ["Regression", "Residual", "Lack of fit", "Pure error"],
    "Sum of squares" => [0.295145, 2.79e-4, 1.7e-4, 1.09355e-4],
    "Degree of freedom" => [1, 10, 4, 6],
    "Mean square" => [0.295146, 2.8e-5, 4.3e-5, 1.8e-5]
)
```

explained_variation = 99.906

```
explained_variation = 99.906
```

max_explainable_variance = 99.963

```
max_explainable_variance = 99.963
```

There is no lack of fit. It is a good model.