# ST 5203: Experimental Design

(Semester 1, AY 2017/2018)

**Text book:** *Experiments: Planning, Analysis, and Optimization (2nd. edition)*

by Jeff Wu and Mike Hamada

# Topic 1: Introduction to Design & Analysis of Experiments; Regression Review

- Historical perspectives.
- Observational studies vs. Experimental studies.
- Basic terminologies.
- Basic principles of experimentation.
- Regression review.

# Brief history of statistical design and analysis of experiments

- Founder: R. A. Fisher in the 1920s and 1930s (agricultural experimentation)
- Frank Yates: novel block designs and factorial designs
- George Box: statistical method for process optimization
- Raj Chandra Bose: mathematical theory of construction of experimental designs
- Jack Kiefer: theory of optimal designs
- A. Bradford Hill: randomized assignments of patients in clinical trials
- Genichi Taguchi: engineering applications, robust design

# Brief history of statistical design and analysis of experiments

- (1930s) Fisher's pioneering work on design of experiments and analysis of variance (ANOVA) and application to agricultural research.
- (After World War II) Box's work motivated in chemical industries and applicable to other processing industries, regression modeling and response surface methodology.
- (Recent) Taguchi's work on robust parameter design to reduce system variation.

# Observational study

- Researcher is a passive observer who records variables of interest
- Variables of interest
    - Independent/explanatory variables (or factor)
    - Dependent/response variables
- Variables are not controlled and hence may **confound** the outcome
- May establish *association* but not *causation*, between the factors of interest (**treatment factors**) and the response variable

# Observational study (cont.)

- Possible confounding caused by
  - Noise factors: all other factors that are not controlled
  - Lurking variables: not recognized to be important
- Epidemiological studies: an important class of observational studies
  - Treatment factors: the suspected **risk factors** of a disease
  - Objective: find out whether they are associated with the disease
  - Two types
    1. Prospective studies: subjects followed forward; their disease outcome recorded
    2. Retrospective studies (case-control studies): subjects followed backward; their exposure to suspected risk factors recorded

# Experimental study

- Researcher actively manipulates the factors and evaluates their effects on the response variables.
- May (or may not) establish causation.
- Goal: knowledge and discovery about the phenomenon under study; better understanding of the phenomenon; discover which factors affect the outcomes and how to improve performance by tuning key design factors
  - (a) Screen the treatment factors
  - (b) Determine the factor space
  - (c) Select the best combination of the treatment factor settings
  - (d) Fit a model
  - (e) Determine or expand the scope of applicability

# Basic terminology and concepts

- **Factor:** variable whose influence upon a response variable is being studied in the experiment.
- **Factor level (treatment):** selected numerical values or settings for a factor.
- **Controllable factors:** we can set the levels
- **Noise factors and blocking factors**
- **Qualitative factor and quantitative factor**
- **Experimental unit:** object to which a treatment is applied.
- **Trial** (or **run:**) application of a treatment to an experimental unit.
- **Randomization:** random assignment of treatments to experimental units (or run orders).

# Basic terminology and concepts (cont.)

- **Replicate:** an independent run carried out under the same conditions.
- **Replication error**
- **Repeat measurement:** another measurement of the same response; not an independent replicate.
- **Measurement error**
- **Treatment group:** all experimental units receiving the same treatment.
- **Standard (or control) treatment:** benchmark; active control and placebo (or passive control).
- **Control group**

# Example 1

In order to investigate how the temperature will affect the growth of a certain type of plant, three temperatures, namely $10^\circ C$, $20^\circ C$ and $30^\circ C$, are examined in a laboratory.

- Factor: temperature.
- Factor levels: 3 levels of temperatures, i.e. $10^\circ C$, $20^\circ C$ and $30^\circ C$.
- Trial: growth procedure of each plant.
- Treatment: each temperature examined, the same as “factor level” in this specific example.
- Experimental unit: each plant.
- Randomization: randomly assign available plants to each of the temperature levels.

# Checklist for Planning an Experiment

1. State the objective of the experiment.
2. Choose the response variable (based on the objective).
3. Choose factors and levels.
4. Plan of the experiment; Determine the set of treatments and order to run; Perform the experiment.
5. Analyze data (carefully choose models and technologies).
6. Draw conclusion.

# Example 2

A car company would like to test whether a new designed car model (A) is safer than an old one (B).They decide to perform the following experiment:

- Randomly sample a number of new and old designed cars..
- Collide cars to walls with the same speed.
- Quantify and collect the damage levels.
- Due to budget, they can afford 6 old cars, 6 new cars and 6 walls.

# Example 2 (cont.)

- **Objective**: compare safety of two types of cars.
- **Response**: damage levels of each car.
- **Factors and levels**: "car model" (with levels "A" and "B"); "wall" (with levels $1, 2, \ldots, 6$).
- *Noise factors, blocking factors?*

# Example 2 (cont.)

- **Several designs**: Randomize "model A" cars and "model B" cars within the "model", then collide cars to walls by the following order.
  1. Convenient runs:

     1. A,B; 2. A,B; 3. A,B; 4. A,B; 5. A,B; 6. A,B.

     (A always followed by B. Is this good?)
  2. Randomizing the order of A and B in each group and result in new sequence:

     1. A,B; 2. B,A; 3. A,B; 4. B,A; 5. A,B; 6. A,B.

     (Better? But still there are four with "A,B" and two with "B,A". Is this good?)
  3. Randomly choose three walls with order "A,B" and the rest with "B,A". (Try to produce one of such plan on your own. Is this good?)
- **Data analysis and conclusion.**

# Basic Principles: Randomization, Blocking and Replication

**Randomization:** Use of a chance mechanism (e.g.: random number generators) to assign treatments to units or to run order. It has the following advantages.

- Protect against latent variables or “lurking” variables.
- Reduce influence of subjective bias in treatment assignments.
- Ensure validity of statistical inference.

# Blocking

**Blocking** provides local control of the environment to reduce experimental error. A **block** refer to a collection of homogeneous units. (Examples: hours, lots, street blocks, couples, locations, etc.)

- Carefully assign blocks; Randomize treatments within the same blocks to eliminate block-block variation and reduce variability of treatments effects estimates.
- **Block what you can; Randomize what you cannot.**
- Revisit car collision experiment to demonstrate the idea of blocking and randomization.

# Revisit Example 2

**Several designs**: Randomize “model A” cars and “model B” cars within the “model”, then collide cars to walls by the following order.

1. Convenient runs:

   1. A,B; 2. A,B; 3. A,B; 4. A,B; 5. A,B; 6. A,B.

   (A always followed by B. Is this good?)

2. Randomizing the order of A and B in each group and result in new sequence:

   1. A,B; 2. B,A; 3. A,B; 4. B,A; 5. A,B; 6. A,B.

   (Better? But still there are four with “A,B” and two with “B,A”. Is this good?)

3. Randomly choose three walls with order “A,B” and the rest with “B,A”. (Try to produce one of such plan on your own. Is this good?)

# Replication

- **Replication:** Apply each treatment to different units.
- Different from repetition (example: measurements of 3 units versus 3 repeated measurements of 1 unit).
- Enable the estimation of experimental standard error (Use sample standard error).
- Increase the accuracy of the estimates (decrease the variance of the estimator). For example, $Y_1, Y_2, \ldots, Y_n$, to estimate $\mu = E(Y_i)$, we use $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$:

$$\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\mathrm{Var}(Y_1).$$

# How to minimize biases and variability?

- Strategy for dealing with noise factors at the **design stage**
    - **Blocking**: if a noise factor is controllable or its value for each experimental unit is known.
    - **Randomization**: other noise factors; randomly assign experimental units to the different treatments.
- Covariates: noise factors not observed before assignment of experimental units to the treatments; or their levels cannot be fixed.
- **Replication**: can reduce the effect of random errors.
- **Repeated measurements**: can reduce the effect of measurement errors.

# How to minimize biases and variability?

- Main component of the noise:
  - Systematic biases/errors: caused by systematic differences between the experimental units in different treatment groups; **confound** or **bias** the treatment comparisons.
  - Replication or random errors: caused by the inherent variability in the responses.
  - Measurement errors: caused by imprecise measuring.
- Experimental errors: replication and measurement errors together.

# Example 3: Heat treatment of steel

A metallurgical engineer designing a heat treatment method wants to study the effects of furnace temperature (high or low) and quench oil bath temperature (high or low) on the surface hardness of the steel, which is the response variable. 20 steel samples and 5 furnaces are available for experimentation.

- What are the treatment factors? What are the factor levels?
  furnace temperature and quench oil bath temparature.
  $2 \times 2 = 4$ factor levels: (high,high), (high,low), (low, high), (low,low).
- What are the experimental units?
  20 steel samples.
- What are some blocking factors?
  5 furnaces. The steel samples are likely to be more similar if they are tested in the same furnace.
- What are some noise factors?
  For example, deviations from constant furnace and quench oil bath temperatures, variations between steel samples, etc.

# Example 3: Heat treatment of steel (cont.)

Two types of designs:

- **Completely randomized (CR) design:** all experimental units are assigned at random.

  In Example 3, a CR design randomly assigns 20 steel samples to one of the four treatments. It is not even required to assign an equal number of sample (5) to each level, although usually randomization is done under this restriction.

- **Randomized block (RB) design:** randomization is done subject to a blocking restriction.

  In Example 3, a RB design uses the furnace as a blocking variable. So we can form blocks of 4 samples from each of the 5 furnaces (batches) and randomly assign them to 4 treatments. Then we have 1 replicate per treatment from each batch.

# CR and RB designs for Example 3

A=(low,low), B=(high,low), C=(low, high), D=(high,high).

Completely Randomized Design

| Batch 1 | Batch 2 | Batch 3 | Batch 4 | Batch 5 |
|---|---|---|---|---|
| A | C | D | D | B |
| B | A | A | C | D |
| C | D | B | C | A |
| B | C | D | B | A |

Randomized Block Design

| Batch 1 | Batch 2 | Batch 3 | Batch 4 | Batch 5 |
|---|---|---|---|---|
| A | D | C | D | B |
| C | A | B | C | C |
| B | C | D | B | A |
| D | B | A | A | D |

# Regression Review

# Simple Linear Regression

- The **simple linear regression model** assumes a relationship between the response values $y_1, y_2, \ldots, y_n$ and the corresponding predictor values $x_1, x_2, \ldots, x_n$:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ \epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2),\ i = 1, 2, \ldots, n.$$

- $\epsilon_i$: the **random error**.
- The line $y = \beta_0 + \beta_1 x$: **true (or population) regression line**.
- The parameters $\beta_0$, $\beta_1$ and $\sigma^2$ need to be estimated.

# Data for Regression Analysis and Scatter Plot

- Let $x_1, x_2, \ldots, x_n$ denote $n$ different settings of the independent variable $x$ and $y_1, y_2, \ldots, y_n$ denote the corresponding response values.
- The available data: $n$ pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.
- A **scatter plot** uses data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ and plots each $(x_i, y_i)$ as a point on a two-dimensional coordinate system.

# Estimation in the Simple Linear Regression Model

- $\beta_0$ and $\beta_1$ are estimated by minimizing

$$\sum_{i=1}^{n} \left\{ y_i - (\beta_0 + \beta_1 x_i) \right\}^2$$

- Estimated coefficients:

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}, \qquad \mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \qquad \mathrm{Var}(\hat{\beta}_0) = \sigma^2 \left\{ \frac{1}{n} + \frac{(\bar{x})^2}{\sum (x_i - \bar{x})^2} \right\},$$

with $\bar{x} = \sum x_i / n, \bar{y} = \sum y_i / n.$

- The estimate of $\sigma^2$ is

$$\widehat{\sigma^2} = s^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n - 2}$$

- The estimator of $\sigma$ is $s$.

# Explanatory Power of the Model

- The total variation in $y$ can be measured by "Corrected Total Sum of Squares (CTSS)", $CTSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$.
- This can be decomposed into two parts (Analysis of Variance (ANOVA)):

  $$CTSS = RegrSS + RSS$$

  where

  $$RegrSS = \text{Regression Sum of Squares} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2,$$

  $$RSS = \text{Residual Sum of Squares} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

  $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ is the predicted value of $y_i$ at $x_i$.
- $R^2 = \frac{RegrSS}{CTSS} = 1 - \frac{RSS}{CTSS}$ measures the proportion of variation in $y$ explained by the fitted model.

# Hypothesis Testing

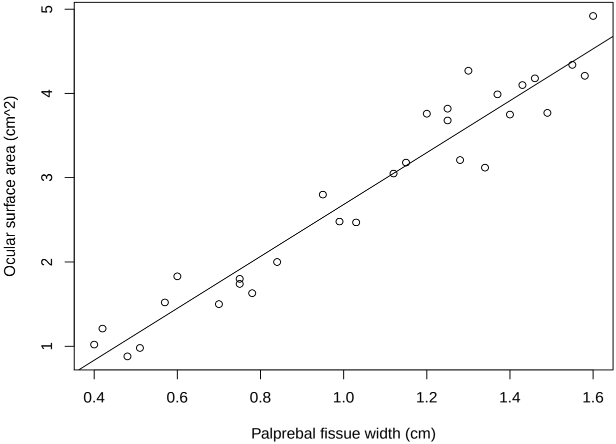
- To test $H_0 : \beta_j = 0$ versus $H_a : \beta_j \neq 0$, $j = 0$ or 1, use the test statistic
$$t_j = \frac{\hat{\beta}_j}{s.e.(\hat{\beta}_j)} \sim t_{n-2}.$$
- $s.e.(\hat{\beta}_j)$ is the standard error of $\hat{\beta}_j$.
- For the 2-sided alternatives above, p-value $=$ $2 \times \text{Prob}(T > |t_{obs}|)$, where $T \sim t_{n-2}$, the $t$ distribution with degree of freedom $n-2$ and $t_{obs}$ is the computed statistic from the formula above with the observed sample.

# Example 12.01 from “Devore7”

This example gives the measurements of $y$, the ocular surface area (OSA), and $x$, the palprebral fissure width, for 30 subjects. The dataset is called “xmp12.01” from the R-package `Devore7`.

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| $x_i$ | 0.40 | 0.42 | 0.48 | 0.51 | 0.57 | 0.60 | 0.70 | 0.75 |
| $y_i$ | 1.02 | 1.21 | 0.88 | 0.98 | 1.52 | 1.83 | 1.50 | 1.80 |
| $i$ | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| $x_i$ | 0.75 | 0.78 | 0.84 | 0.95 | 0.99 | 1.03 | 1.12 | 1.15 |
| $y_i$ | 1.74 | 1.63 | 2.00 | 2.80 | 2.48 | 2.47 | 3.05 | 3.18 |
| $i$ | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 |
| $x_i$ | 1.20 | 1.25 | 1.25 | 1.28 | 1.30 | 1.34 | 1.37 | 1.40 |
| $y_i$ | 3.76 | 3.68 | 3.82 | 3.21 | 4.27 | 3.12 | 3.99 | 3.75 |
| $i$ | 25 | 26 | 27 | 28 | 29 | 30 | | |
| $x_i$ | 1.43 | 1.46 | 1.49 | 1.55 | 1.58 | 1.60 | | |
| $y_i$ | 4.10 | 4.18 | 3.77 | 4.34 | 4.21 | 4.92 | | |

# Scatter Plot and Fitted line



# Regression Analysis

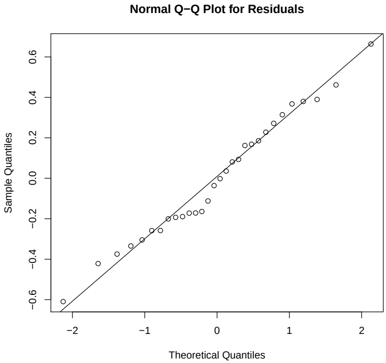
Outputs from R:

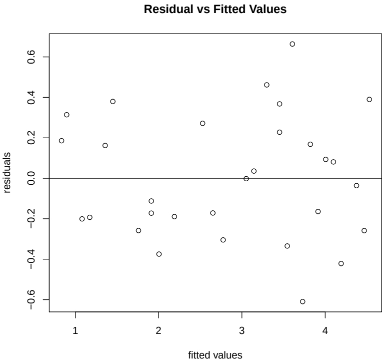
```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.3977     0.1680  -2.367   0.0251 *
x             3.0800     0.1506  20.453   <2e-16 ***
---

Residual standard error: 0.308 on 28 degrees of freedom
Multiple R-squared: 0.9373,     Adjusted R-squared: 0.935
F-statistic: 418.3 on 1 and 28 DF,  p-value: < 2.2e-16
```

# Predicted Values and Residuals

- $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ is the predicted value of $y_i$ at $x_i$.
- $r_i = y_i - \hat{y}_i$ is the residual.
- Use residual plots to judge the “goodness” of fitted model.
  - Normal probability plot (Q-Q plot) (will be discussed later).
  - Residuals versus predicted values.
  - Other residual plots.

# Residual Plots


Normal Q−Q Plot for Residuals


Residual vs Fitted Values

# Multiple Linear Regression

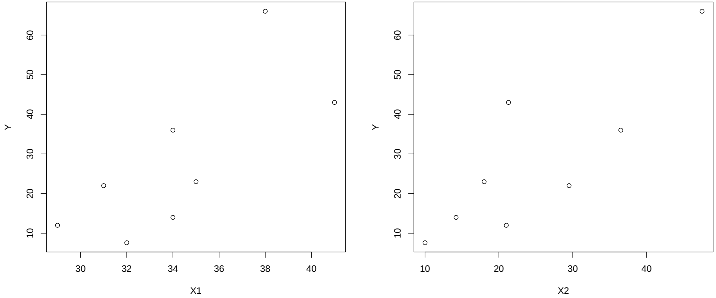
A toy example:

- Eight runs were made at various conditions of saturation ($X_1$) and transisomers ($X_2$). The response, SCI is listed below as $Y$ for the corresponding levels of $X_1$ and $X_2$.

| $Y$ | $X_1$ | $X_2$ |
|---|---|---|
| 66.0 | 38 | 47.5 |
| 43.0 | 41 | 21.3 |
| 36.0 | 34 | 36.5 |
| 23.0 | 35 | 18.0 |
| 22.0 | 31 | 29.5 |
| 14.0 | 34 | 14.2 |
| 12.0 | 29 | 21.0 |
| 7.6 | 32 | 10.0 |

# Multiple Linear Regression

- There are 8 responses.
- Each response has more than one predictor ($X_1$ and $X_2$ in this example).
- So, the data available are $(y_1, x_{1,1}, x_{1,2}), \ldots, (y_8, x_{8,1}, x_{8,2})$.
- Scatter plots can be drawn by $Y$ vector versus $X_1$ vector and $X_2$ vector respectively.



# Model and Multiple Regression Fitting

- Multiple regression model: for each $i = 1, 2, \ldots, n$,

$$y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_k x_{i,k} + \epsilon_i, \quad \epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2).$$

- $\beta_0, \beta_1, \ldots, \beta_k$ are estimated by minimizing

$$\sum_{i=1}^{n} \{y_i - (\beta_0 + \beta_1 x_{i,1} + \ldots + \beta_k x_{i,k})\}^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

- Least Square Estimates:

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

- Variance-covariance matrix of $\hat{\boldsymbol{\beta}}$: $\Sigma_{\hat{\boldsymbol{\beta}}} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

# ANOVA

- Similar to simple linear regression: $CTSS = RegrSS + RSS$, where

$$\begin{aligned} CTSS &= \sum (y_i - \bar{y})^2 = \mathbf{y}'\mathbf{y} - n\bar{y}^2, \\ RegrSS &= \sum (\hat{y}_i - \bar{y})^2 = \hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta} - n\bar{y}^2, \\ RSS &= \sum (y_i - \hat{y}_i)^2 = (\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta}). \end{aligned}$$

ANOVA Table

| Source | D.F. | Sum of Squares | Mean Squares |
|---|---|---|---|
| regression | $k$ | $RegrSS$ | $RegrSS/k$ |
| residual | $n-k-1$ | $RSS$ | $RSS/(n-k-1)$ |
| total | $n-1$ | $CTSS$ | |

# Explanatory Power of the Model

- Multiple correlation coefficient $R^2 = \frac{RegrSS}{CTSS} = 1 - \frac{RSS}{CTSS}$ measures the proportion of variation in $y$ explained by the fitted model.

- Adjusted $R^2$:
  $R_a^2 = 1 - \frac{RSS/\{n-(k+1)\}}{CTSS/(n-1)} = 1 - \frac{n-1}{n-k-1}\frac{RSS}{CTSS}$.

- $R^2$ is always increasing when additional predictors are added to the model, however, $R_a^2$ may decrease if the included variable is not an informative predictor. In general, $R_a^2$ is a better measure for model fit.

# Hypothesis Testing

- $\sigma^2$ is estimated by $\hat{\sigma}^2 = MSE = RSS/(n-k-1)$.

- To test $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$ versus $H_a$: not so. use the test statistic

$$F = \frac{RegrSS/k}{RSS/(n-k-1)},$$

which follows a $F_{k,n-k-1}$ distribution when $H_0$ is true.

- The p-value of the F-test above is Prob($F > f_{obs}$), where $F$ follows F-distribution with degrees of freedom $k$ and $n-k-1$, $f_{obs}$ is the computed statistic from the formula above with the observed sample.

# Hypothesis Testing (cont.)

- To test $H_0 : \beta_j = 0$ versus $H_a : \beta_j \neq 0$, use the test statistic: for any $j = 0, 1, \ldots, k$,

$$T_j = \frac{\hat{\beta}_j}{s.e.(\hat{\beta}_j)} \sim t_{n-k-1}.$$

- For the 2-sided alternatives above, p-value $=$ 2Prob$(T > |t_{obs}|)$, where $T \sim t_{n-k-1}$, the $t$ distribution with degree of freedom $n-k-1$ and $t_{obs}$ is the computed statistic from the formula above with the observed sample.

# Analysis of the Toy Example

Outputs from R:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -94.55203    9.96343  -9.490 0.000220 ***
X1            2.80155    0.30098   9.308 0.000241 ***
X2            1.07268    0.09323  11.505  8.7e-05 ***
---
Residual standard error: 2.938 on 5 degrees of freedom
Multiple R-squared: 0.9838,     Adjusted R-squared: 0.9773
F-statistic: 151.7 on 2 and 5 DF,  p-value: 3.347e-05

Analysis of Variance Table
          Df  Sum Sq Mean Sq F value    Pr(>F)
X1         1 1476.36 1476.36  171.03 4.664e-05 ***
X2         1 1142.62 1142.62  132.37 8.696e-05 ***
Residuals  5   43.16    8.63
```

# Notes on Board