# Developing a Statistical Model

## Step 1: Introduction

Go to the following URL: https://www.stat2games.sites.grinnell.edu/games/greenhouse.html
(Note: *This site may take a few minutes to load.*)

Input a **Player ID**: _____
This will be on the internet, so you probably *do not want to use a name that will identify you.*

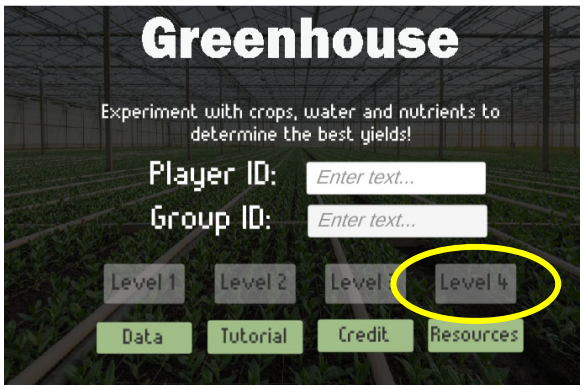
Record your **Group ID**: _____
Your instructor will give you a **Group ID,** which will be identical for every person in the class.

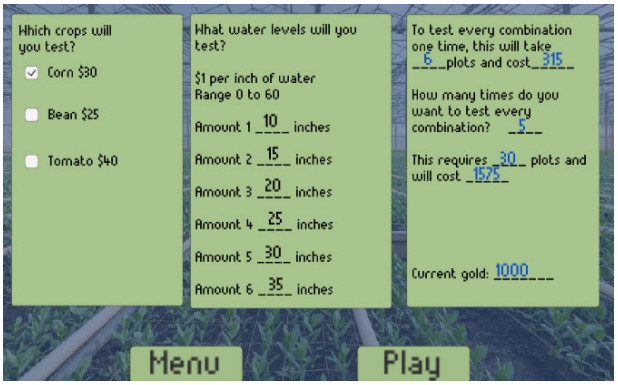
To start, click on the **Tutorial** button and watch the tutorial.

In this lab, you will be using an online **Greenhouse** game to collect your own data. This lab was developed so that the yields (amount of crops produced) follow models that reflect actual crop growth in the United States. Each test plot represents 1/10$^{th}$ of an acre of land and yields are the number of bushels produced on that plot.

As you play the game and complete this lab, **your first goal is to create a model that estimates the amount of yield based upon the amount of water**. Select **Level 4** and enter the levels as you see below. **Level 4** will assign each of your selected treatments (crop type and water amount) to a randomly selected plot.

NOTE: At this point, do not worry about winning or profits; just carefully follow the instructions to collect data.




1. Move your cursor over the plots and examine the amount of water, fertilizer and yield.
2. Click the Harvest button on the right, followed by sell and then you can see how much profit you made.
3. After you have harvested and sold your crops. Enter your data below:

| Income | | Expenses | | Profit | |
|---|---|---|---|---|---|

4. For this study, identify the following:

    a. Population:

    b. Sample:

5. If we wish to analyze this data (by making a scatterplot for example), which of our variables is the explanatory (independent) variable and which is the response (dependent) variable?

- Explanatory (independent) variable is: _____
- Response (dependent) variable is: _____

6. Notice that we collected several measurements (several plots) for each amount of water. (i.e. There were five plots watered with 20 inches of water.) Why do you think it could be useful to have repeated measurements?

## Step 2: Developing and Interpreting a Linear Model

Go to https://shiny.grinnell.edu/Greenhouse_Models/ and use the following settings to calculate your model.

- Group ID: **Group ID = sample1** [Do NOT use your class Group ID in this section]
- Remove Player ID: (should be blank)
- X Variable: **WaterAdded**
- Y Variable: **Yield**
- Select Crop: **Corn**
- Facet by: **None**
- Statistical Model: **Linear**
- Ignore other items, such as the Remove Interaction Terms, Nitrate Levels, and X-axis Limits.

7. Notice that the yields vary, even when the same amount of water is used. Explain why this should be expected.

8. Do you see a pattern in your scatter plot? If so, describe it.

9. Use the online App identify the slope of the linear regression model. _____ (Round the XVariable estimate to 2 decimal places)

10. What is the interpretation of the slope in the context of this experiment? That is, what does this imply about corn yields?

11. Use the online App identify the y-intercept? _____ (Round the Intercept Estimate to 2 decimal places)

12. What is the interpretation of the y-intercept in the context of this experiment? Why is this not a very practical value in the context of this question?

13. What is the equation of the least squares regression line?

14. Find the **r-squared (the coefficient of determination)** value: _____ (round to 3 decimal places) and explain what this value tells us.

15. Calculate the **correlation coefficient (r)**: _____ (round to 3 decimal places) and explain what this value tells us.

16. Use **your model** to predict the corn yield for each of the following cases, then indicate whether or not you expect the estimate to be accurate.
    a. Water = 0, Yield = _____ Would you expect this to be accurate? Yes/No

    b. Water = 28, Yield = _____ Would you expect this to be accurate? Yes/No

    c. Water = 40, Yield = _____ Would you expect this to be accurate? Yes/No

## Step 3: Developing a Better Model?

**Instead of using yields, we will now create a model based upon profits.**
Go to https://shiny.grinnell.edu/Greenhouse_Models/

- Group ID: Use the **Group ID** for your class.
- Remove Player ID: (should initially be blank)
- X Variable: **WaterAdded**
- Y Variable: **Profit**
- Select Crop: **Corn**
- Facet by: **None**
- Statistical Model: **Linear**
- Ignore other items, such as the Remove Interaction Terms, Nitrate Levels, and X-axis Limits.

**Restricting our data**

15. If there are any PlayerIDs that did not appear to properly follow the instructions in Step 1, remove them from the dataset. If you removed a PlayerID, list the names you removed.

16. Are there any outliers or unusual observations that you believe should be removed from the model? If so, list the Player IDs corresponding to these points.

17. If you made any modifications to the data, provide your new regression equation. Provide the new regression line and R-squared value when PlayerIDs from Question 15 and 16 are removed.

18. Choose Statistical Model: **Quadratic**. Provide your new regression equation and the R-squared value. Notice the quadratic model has the following form: $Y = Bo + B_1*X + B_2*X^2$

19. Which model has the best R-Squared value?
    a. the linear model in Question 17 b. the quadratic model Question 18

20. Explain why you might expect Question 18 to have a better R-squared value.

21. Use your models to predict the corn yield for each of the following cases.

| Water | Profit from linear model (Question 17) | Profit from quadratic model (Question 18) |
|---|---|---|
| 0 | | |
| 28 | | |
| 40 | | |

22. Explain why you might expect Question 18 to be able to better predict the actual profit when WaterAdded = 40.

23. Go back to the Greenhouse game and plant a few plots with Corn and 40 inches of water. Is a linear or quadratic model better at predicting Profit when the WaterAdded = 40?

24. Using the quadratic model, roughly estimate the amount of water needed for maximum profits for Corn. Briefly describe a strategy for guaranteeing the most profit when planting corn in the Greenhouse game.

## Step 4: Making a Recommendation

Assume that your neighbor is interested in knowing the best conditions for growing tomatoes. In particular, what amount of water would optimize yield for tomato plants? Create a one-page report that includes the following:

- At least one graph
- At least one regression model
- A measure and interpretation of the accuracy of your model
- A recommendation for your neighbor
- If your neighbor follows your recommendation, give an approximate estimate of the predicted yield for each plot. Assume your neighbor's growing conditions are essentially the same as the greenhouse.
- Explain to your neighbor why they should not expect exactly the same yield from every plot. Give an approximate range for the expected yield if they follow your recommendation.