# Special Modelling Methods for Non-Ordinary Responses

Martin Bezener, PhD
President & Chief Technology Officer
martin@statease.com

March 2022

## Making the most of this learning opportunity

Hide the Control Panel
Mute your line

To prevent audio disruptions, all attendees will be muted.

Questions can be posted in the **Question** area. If they are not addressed during the webinar, I will reply via email afterwards.

Questions may also be sent to stathelp@statease.com. Please provide your company name and, if you are using Design-Expert, the serial number (found under Help, About).

**Note:** The slides and a recording of this webinar will be posted on the Webinars page of the Stat-Ease website within a few days.

## Agenda

- Introduction to Stat-Ease 360
- Logistic Regression (Binary Data)
- Poisson Regression (Count Data)
- Gaussian Process Models (Simulation)
- Final Comments & Conclusion



## Agenda

- **Introduction to Stat-Ease 360**
- Logistic Regression (Binary Data)
- Poisson Regression (Count Data)
- Gaussian Process Models (Simulation)
- Final Comments & Conclusion

## About Stat-Ease

- Stat-Ease was founded in 1985 by Pat Whitcomb.
- Headquartered in Minneapolis, Minnesota USA.
- Owned by Constellation Software, Inc. in Toronto, Canada.
- Developers and publishers of the following leading software packages for design and analysis of experiments:
  - Design-Expert® (DX)
  - Stat-Ease®360 (SE360)
- We also provide DOE training as well.



## Our Software – Going Forward

- Over the years, Stat-Ease has received steady requests for more advanced DOE tools.
- We've also received demand for tools that are adjacent to "old school" DOE, such as design and analysis of computer experiments, Gage R&R (measurement systems analysis), more computing capabilities, statistical process control, and many more.
- In the interest of keeping Design-Expert as simple and accessible as possible, we have historically tabled these types of requests.
- The demand was simply too great!

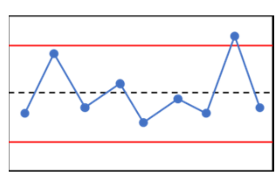
STAT-EASE 360

- Stat-Ease 360 (SE360) was released **October 4, 2021**.
- It contains a **full copy** of Design-Expert software.
- The first release of SE360 **also** contains some of our most highly-requested advanced DOE features:
  - Space-filling designs
  - **Gaussian process models for zero-error data**
  - Python scripting
  - Advanced classification tools for logistic regression
  - **Coming soon:** Report generator, Excel import/export, and more!

Special Models 7

STAT-EASE 360

- Going forward, Design-Expert will continue to be targeted towards engineers, formulators, and others working in R&D. Remember, SE360 contains all of DX.
- Going forward, easy-to-use and bread-and-butter techniques will go into both DX and SE360.
- Highly-advanced, non-DOE, and highly technical & computational features will only go into SE360. Some ideas we have for the future:
  - Measurement Systems Analysis
  - Multivariate Analysis
  - Advanced Scripting/Computing
  - Reliability Analysis



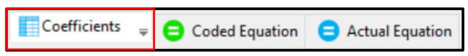
Special Models 8

## Preview of Today’s Webinar

- In many industrial experiments, **ordinary least squares (OLS)** regression is an adequate modelling tool for the collected experimental data. The standard assumption is that:

$$y_i \sim Normal(x_i^T \boldsymbol{\beta}, \sigma^2)$$

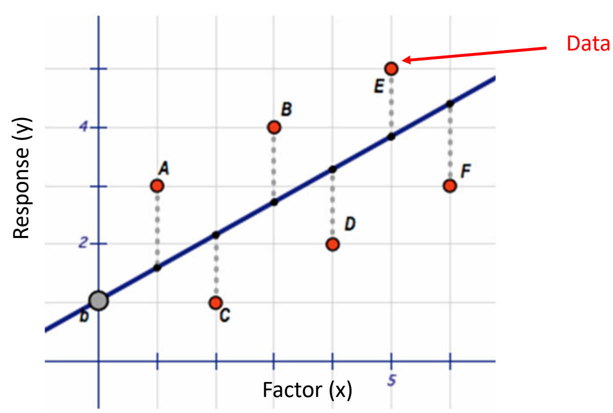
- This means that at each combination of factor settings, we expect the response to be **bell-shaped** with constant variability.
- This model is usual written as: $\boldsymbol{y} = x^T \boldsymbol{\beta}$
- Ordinary least squares finds the “best” $\boldsymbol{\beta}$ for the given experimental design $x$ and measured responses $\boldsymbol{y}$. These estimates of $\boldsymbol{\beta}$ are found on the ANOVA in the **Coefficients** tab.






## Preview of Today’s Webinar

Here is an illustration of how OLS works. The red points are the data.



## Preview of Today's Webinar

- In some situations, however, OLS should not be used.
- Remember that OLS requires
  - Continuous data
  - Bell-shaped error with constant variance
- We'll cover the following **alternative modelling methods** today:
  - Logistic Regression – 0/1 data (pass/fail)
  - Poisson Regression – count data
  - Gaussian Process Models – simulation data (no error)
- These are only 3 common scenarios where OLS doesn't work – there are many more!



## Agenda

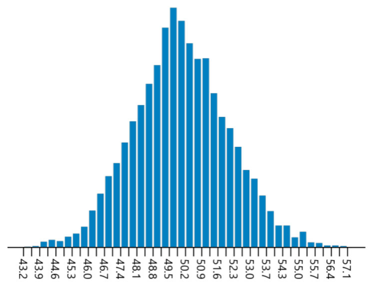
- Introduction to Stat-Ease 360
- **Logistic Regression (Binary Data)**
- Poisson Regression (Count Data)
- Gaussian Process Models (Simulation)
- Final Comments & Conclusion



## Logistic Regression

- As mentioned on the previous slides, in many cases the response(s) of interest in an industrial experiment are **continuous** and can be analyzed using OLS.
- A continuous response is one that can (theoretically) take any value in some feasible range:

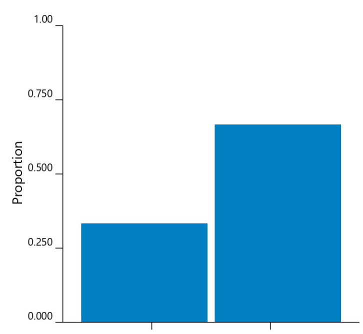



Response Distribution at Time=20, Temp=300



## Logistic Regression

- However, in other cases, at least one response many be a **binary response**. This is a response that can take one of two outcomes:




Response Distribution at Pressure=200

## Logistic Regression

- Some common examples of binary responses:
  - **Quality Engineering:** {Pass, Fail}
  - **Pharma:** {Survived, Didn't Survive}
  - **Medical:** {Disease Present, Disease Absent}
  - **Marketing:** {Clicked on Ad, Didn't Click on Ad}
  - **Food:** {Dough Rises, Dough Doesn't Rise}
  - …
- Experiments may include both continuous and binary responses.
- A binary response is often considered a **categorical** response.

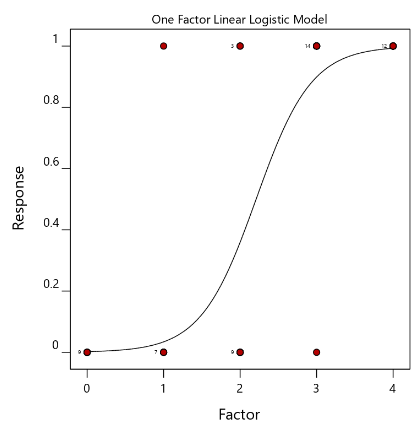


## Logistic Regression

- Here is a one factor example:

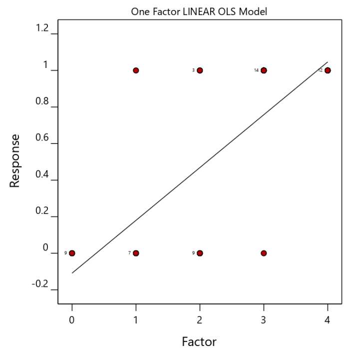
| Factor | Reps | # Success |
|---|---|---|
| 0 | 9 | 0 |
| 1 | 8 | 1 |
| 2 | 12 | 3 |
| 3 | 15 | 14 |
| 4 | 12 | 12 |

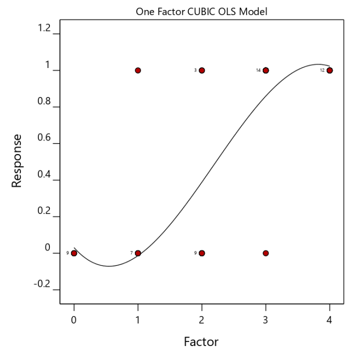
A linear logistic model can be fit to this data (right).



## Logistic Regression

- What’s the problem with using OLS?






## Logistic Regression

- Logistic Regression overcomes the issues that OLS suffers from when analyzing **binary data**.
- The logistic model looks like:

$$p_i = \frac{1}{1 + e^{-x_i^T \beta}}$$

- $p_i$ is the **probability** of observing a ‘1’ under factor settings $x_i$.
- Some **key** points:
  - Logistic regression models a 0/1 response. The predicted values must be between 0 and 1, and the represent a probability.
  - It’s a very flexible model, and widely-used in industry.
  - Available in both Design-Expert and Stat-Ease® 360.

## Logistic Regression

**Here is an example:**

- An aircraft fastener manufacturer wanted to study the ability of a new **fastener** to withstand different loads.
- Two **factors** were varied: **Load** from 2000 psi to 4000 psi, and **Thickness** of the fastener from 1 to 5 (no units given).
- 8 fasteners were tested at 9 different load and factor combinations.
- The goal was to see how **Load** and **Thickness** affected the probability of a fastener failing.
- Each of the fasteners was an independent run – these were not batches of 8 used to calculate a proportion!
- **Demo Time**



## Logistic Regression

**Replicates vs Repeats**

- Logistic regression is used when each **independent** run of the experiment generates a 0 or 1 response.
- If you are measuring a **proportion** success/failure (like in a batch) under a single run, OLS can be used, possibly with the arcsine square root transform.

**How many runs?**

- Sizing a design for binary data is trickier than for continuous data. There are more variables that affect the precision and power of the design under a logistic model. We've found that DOEs for binary data require 3-5x as many runs as for continuous data.
- **Free webinar:** www.youtube.com/watch?v=C_8geR7H1-0

## Agenda

- Introduction to Stat-Ease 360
- Logistic Regression (Binary Data)
- **Poisson Regression (Count Data)**
- Gaussian Process Models (Simulation)
- Final Comments & Conclusion



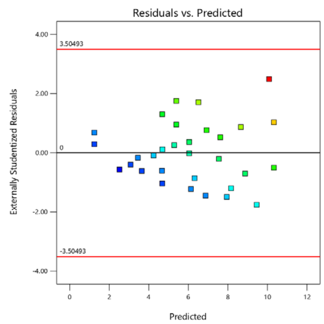
## Poisson Regression

- A **count response** is often measured in industrial experiments. Some examples:
  - The number of defects on a metallic surface
  - The number of hits on a webpage
  - Colony forming units for an antiseptic formulation
  - And many more….
- Notice that counts **must** take the values 0, 1, 2, 3, 4….
- In theory there is **no** upper limit… but practically there likely will be one.

## Poisson Regression

- Count data usually has the property that the variation of the data **increases** as expected counts go up. Count data tends not to be bell-shaped for processes where low counts are expected.
- OLS assumes bell-shaped data and constant variance throughout the design space. Therefore, it's not always the best option for modelling count data.

Typical count data modelled using OLS

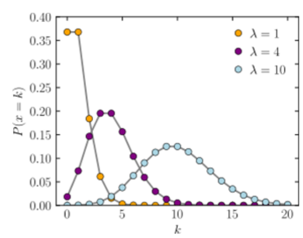





## Poisson Regression

- Poisson regression assumes that each run has the following distribution:

$$y_i \sim \text{Poisson}(x_i^T \boldsymbol{\beta})$$




- The model takes the form:

$$\boldsymbol{y} = \log(\boldsymbol{x}^T \boldsymbol{\beta})$$

- Maximum likelihood is used to find the values of $\boldsymbol{\beta}$. These coefficients are reported on the ANOVA tab in Design-Expert and Stat-Ease 360.

## Poisson Regression

**Here is an example:**

- A mixture experiment was performed to formulate an antiseptic:

| Component | Name | Units | Type | Minimum | Maximum |
|---|---|---|---|---|---|
| A | N-Propanol | wt % | Mixture | 66 | 73 |
| B | Polawax | wt % | Mixture | 4 | 8 |
| C | Anti Lipid | wt % | Mixture | 3 | 7 |
| D | Zinc omadine | wt % | Mixture | 1 | 3 |
| E | Thickening agent | wt % | Mixture | 2 | 2 |
| F | Antioxidant | wt % | Mixture | 4 | 4 |
| G | Buffer | wt % | Mixture | 1 | 1 |
| H | Distilled water | wt % | Mixture | 10 | 10 |
| | | | | Total = | 100.00 |

- **Demo Time: Antiseptic**

Special Models 25

## Poisson Regression

- Poisson regression is used when the response being analyzed is a count.
- OLS may work okay in certain situations for count data, but it's safer to use Poisson regression instead.
- Here is a webinar on the topic of count data and Poisson regression: www.youtube.com/watch?v=FiazrjhIWkk

Special Models 26

## Agenda

- Introduction to Stat-Ease 360
- Logistic Regression (Binary Data)
- Poisson Regression (Count Data)
- **Gaussian Process Models (Simulation)**
- Final Comments & Conclusion

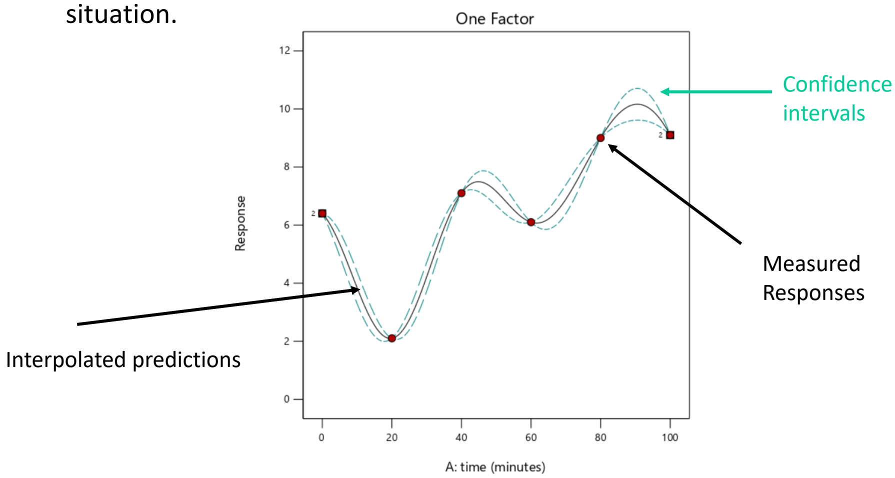


## Gaussian Process Models

- This last case is somewhat specialized, so we won't spend too much time on it.
- **Simulation experiments** are heavily-used in certain industries (medical devices, aerospace, etc.).
- In a simulation experiment, a computer program simulates a response under specific factor settings. These simulations might be very time-intensive.
- Repeating a simulation under the same factor settings will produce an identical response – there is **no error**.
- OLS is not a great tool to analyze this type of data.

## Gaussian Process Models

- Gaussian Process Models (GPMs) are one tool we can use in this situation.






## Gaussian Process Models

**Here is an example:**

- A simulation was run to determine the expected throughput of a process based on a specified warm-up time:

| Run | Time | Throughput |
|---|---|---|
| 1 | 0 | 6.4 |
| 2 | 20 | 2.1 |
| 3 | 40 | 7.1 |
| 4 | 60 | 6.1 |
| 5 | 80 | 9 |
| 6 | 100 | 9.1 |

- The goal was to determine the time where the throughput would reach a **minimum** of 8.5.
- **Demo Time: One Factor GPM**

## Agenda

- Introduction to Stat-Ease 360
- Logistic Regression (Binary Data)
- Poisson Regression (Count Data)
- Gaussian Process Models (Simulation)
- **Final Comments & Conclusion**



## Conclusion

Today we looked at three alternatives to ordinary least squares regression:

- **Logistic regression** for binary data
- **Poisson regression** for count data
- **Gaussian process models** for simulation and low-error data

These are just a handful of the alternatives to OLS that are out there. Here are a few more that you may run into:

- Quantile regression
- Penalized regression (e.g. LASSO)
- Weighted least squares regression
- And many more!



## Conclusion

Here are some resources for further information:

- Stat-Ease Academy
- Stat-Ease Live Public Workshops
- Official Stat-Ease YouTube Channel
    - Logistic Regression Webinar
    - Poisson Regression Webinar
    - GPM Webinar
- Case Studies and White Papers



Make the most from every experiment!℠

**Thanks for listening!**

martin@statease.com