# WRITING A FINAL REPORT

Professor Jennifer Rose
Wesleyan University

# TITLE

- **Should be concise**
- **Clearly summarizes the research question**
- **Avoids redundant or non-informative text and the use of causal language**

  - Bad title: ***A Study of Weather and Caterpillars***
    - Does not clearly summarize the research question
    - Non-informative

  - Better title: ***The Association between Weather Patterns and Caterpillar Reproduction***

# TITLE

- Bad title: ***Major Depression Causes Nicotine Dependence after Controlling for Smoking Exposure***
  - Causal language

- Better title: ***Major Depression Is Associated with Nicotine Dependence Independent of Smoking Exposure***

# INTRODUCTION TO THE RESEARCH QUESTION

- **Clear statement of your research question**
  - Include a description of your response variable and predictors
- **Motivation/rationale for testing the research question**
- **Implications of the research question**
  - For you
  - For your industry or field
  - For society in general

# INTRODUCTION TO THE RESEARCH QUESTION

The purpose of this study was to identify the best predictors of drug manufacturing lead time from multiple production related factors such as inventory status, equipment failure, amount of operator experience, and operator fatigue.

As a process engineer, it is my responsibility to minimize manufacturing lead times. Having a better understanding of factors that are most likely to increase or decrease lead times will allow me to identify which factors to focus on in order to decrease manufacturing lead time.

Shorter manufacturing lead times can result in increased drug production and decreased manufacturing costs, which could translate into lower costs and fewer drug shortages for consumers.

# METHODS

- **Sample**
  - Population from which sample was drawn
    - describe your selection criteria
  - Sample size
  - Enough descriptive information for readers to understand the population
- **Measures**
  - Definitions for the variables that were analyzed
  - How variables were managed
- **Analysis**
  - Summary of statistical methods used and their purpose
  - How data was split into training/test sets and/or type of cross-validation method

# METHODS - SAMPLE

Methods

Sample

The sample included N=435 injection drug production batches manufactured at the Chicago plant from Jan 1, 2015 to December 31, 2015. All batches were high yield batches, meaning that each batch produced between 500,000 and 1 million 0.5 mg drug units.

# METHODS - MEASURES

## Measures

The manufacturing lead time response variable was measured for each drug batch by calculating the number of hours between release of the batch manufacturing order and completion of product packaging.

Predictors included 1) an average of the number of units of each drug ingredient on the bill of materials that was in stock at the time of release of the batch manufacturing order, 2) any equipment failure during production (yes/no) based on Engineering reports, and 3) the number of production steps that were required to complete the manufacturing process.

Employee records were used to determine 4) whether or not trainees were involved during the production process, with trainees defined as production operators who had been working less than 6 months in their current job at the time of manufacturing. Operator fatigue was assessed by 5) the average number of hours of sleep the night before batch production that each production operator reported, and 6) the average of the number of shift hours production operators had already worked prior to beginning batch production.

# METHODS - ANALYSES

## Analyses

The distributions for the predictors and the manufacturing lead time response variable were evaluated by examining frequency tables for categorical variables and calculating the mean, standard deviation and minimum and maximum values for quantitative variables.

Scatter plots and box plots were also examined, and Pearson correlation and Analysis of Variance (ANOVA) were used to test bivariate associations between individual predictors and the manufacturing lead time response variable.

Lasso regression with the least angle regression selection algorithm was used to identify the subset of variables that best predicted manufacturing lead time. The lasso regression model was estimated on a training data set consisting of a random sample of 60% of the batches (N=411). A test data set included the other 40% of the batches (N=273). All predictor variables were standardized to have a mean=0 and standard deviation=1 prior to conducting the lasso regression analysis. Cross validation was performed using k-fold cross validation specifying 10 cross validation folds. The change in the cross validation mean squared error rate at each step was used to identify the best subset of predictor variables. Predictive accuracy was assessed by determining the mean squared error rate of the training data prediction algorithm when applied to observations in the test data set.

# RESULTS

- **Summarize and interpret results of each statistical analysis**
  - Start with simplest analysis (Data Management and Visualization)
  - Bivariate analyses (Data Analysis Tools)
  - Multivariable analyses (Regression Modeling in Practice, Machine Learning for Data Analysis)

  - **Results of descriptive and bivariate analyses can be brief**
    - Provide more detail for results of multivariable analyses

  - **Tables can be useful for summarizing a lot of results**

  - **Figures should be *relevant* to the research question and *enhance* understanding of the results**
    - Refer to figures in text

# RESULTS - DESCRIPTIVES

Results

## Descriptive Statistics

Table 1 shows descriptive statistics for the quantitative data analytic variables. The average manufacturing lead time was 21.45 hours (sd=3.86), with a minimum lead time of 9.40 hours and a maximum of 33.37 hours. In addition, equipment failure occurred in manufacturing for nearly half of the batches (48%; N=329) and trainees were involved in the production of 51% (N=351) of the batches.

**Table 1. Descriptive Statistics for Data Analytic Variables**

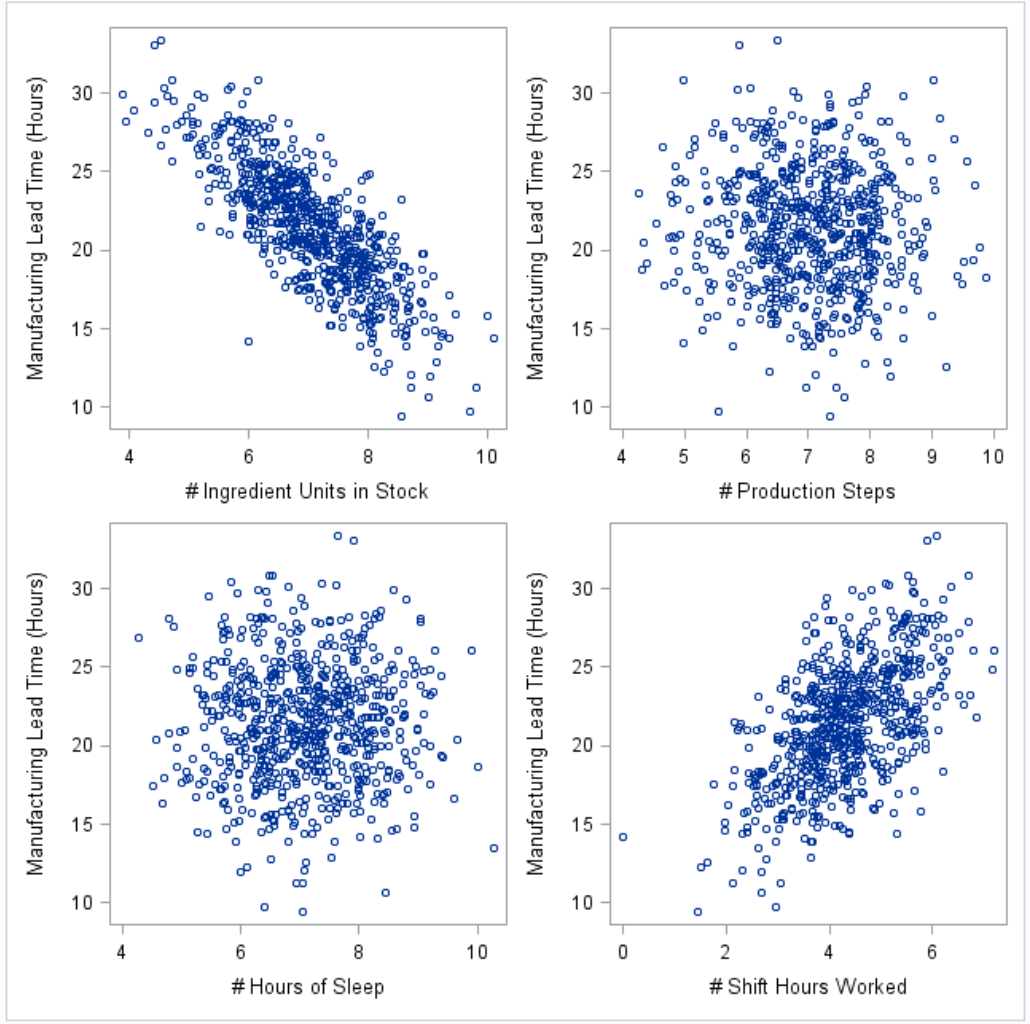
| Analysis Variables | N | Mean | Std Dev | Minimum | Maximum |
|---|---|---|---|---|---|
| # Ingredient Units in Stock | 684 | 7.02 | 1.02 | 3.89 | 10.10 |
| # Production Steps | 684 | 7.02 | 0.99 | 4.25 | 9.86 |
| # Hours of Sleep | 684 | 7.08 | 1.03 | 4.26 | 10.28 |
| # Shift Hours Worked | 684 | 4.32 | 1.01 | 0.00 | 7.19 |
| Manufacturing Lead Time (Hours) | 684 | 21.45 | 3.86 | 9.40 | 33.37 |

# RESULTS - BIVARIATE

Bivariate Analyses

Scatter plots for the association between the manufacturing lead time response variable and quantitative predictors (Figure 1) revealed that manufacturing lead times were shorter when there was a greater number of ingredient units in stock (Pearson r=-0.79, p<.0001), but increased when production workers had worked more hours on their shift before beginning production (Pearson r=0.57, p<.0001). Manufacturing lead time was not significantly associated with the number of steps involved in the production of a batch (Pearson r=-0.05, p=.176) and the number of hours of sleep that production workers reported getting the night before batch production began (Pearson r=0.01, p=.710).

Figure 1. Association Between Quantitative Predictors and Manufacturing Lead Time



# RESULTS – BIVARIATE (CONT.)

Analysis of variance indicated that average manufacturing lead times did not differ significantly as a function of equipment failure (F(1,682)=1.25, p=.264; R2<0.01) or trainee involvement in production (F(1,682)=0.04, p=.833; R2<0.01) (see Figures 2 -3).

Figure 2. Association between Equipment Failure and Manufacturing Lead Time

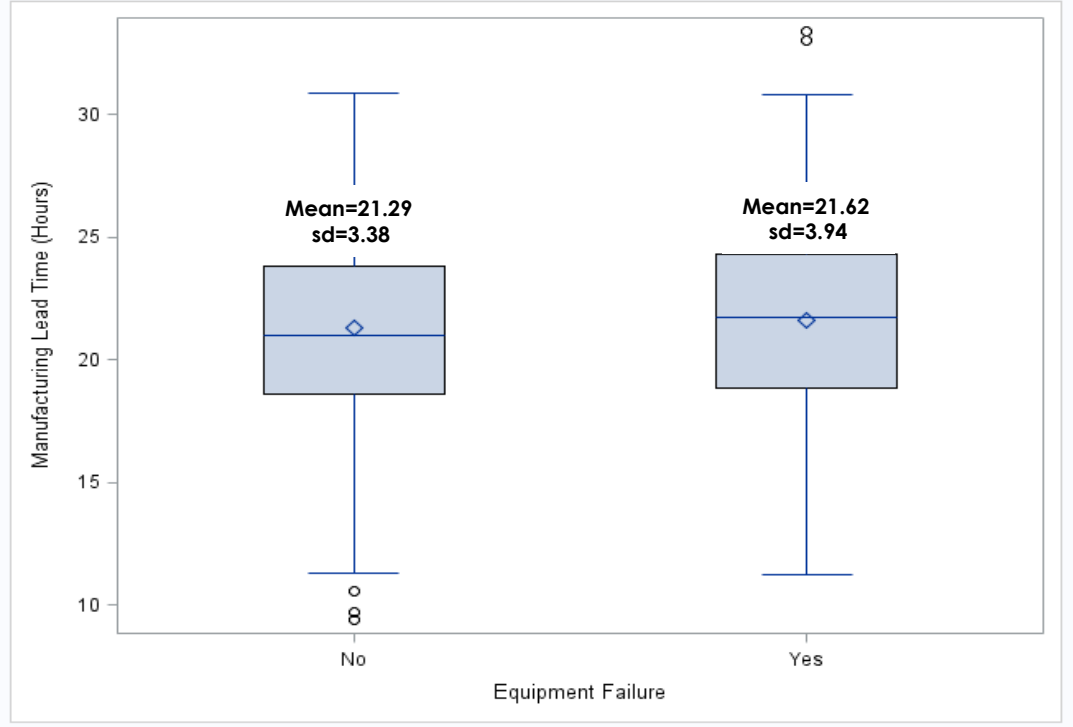
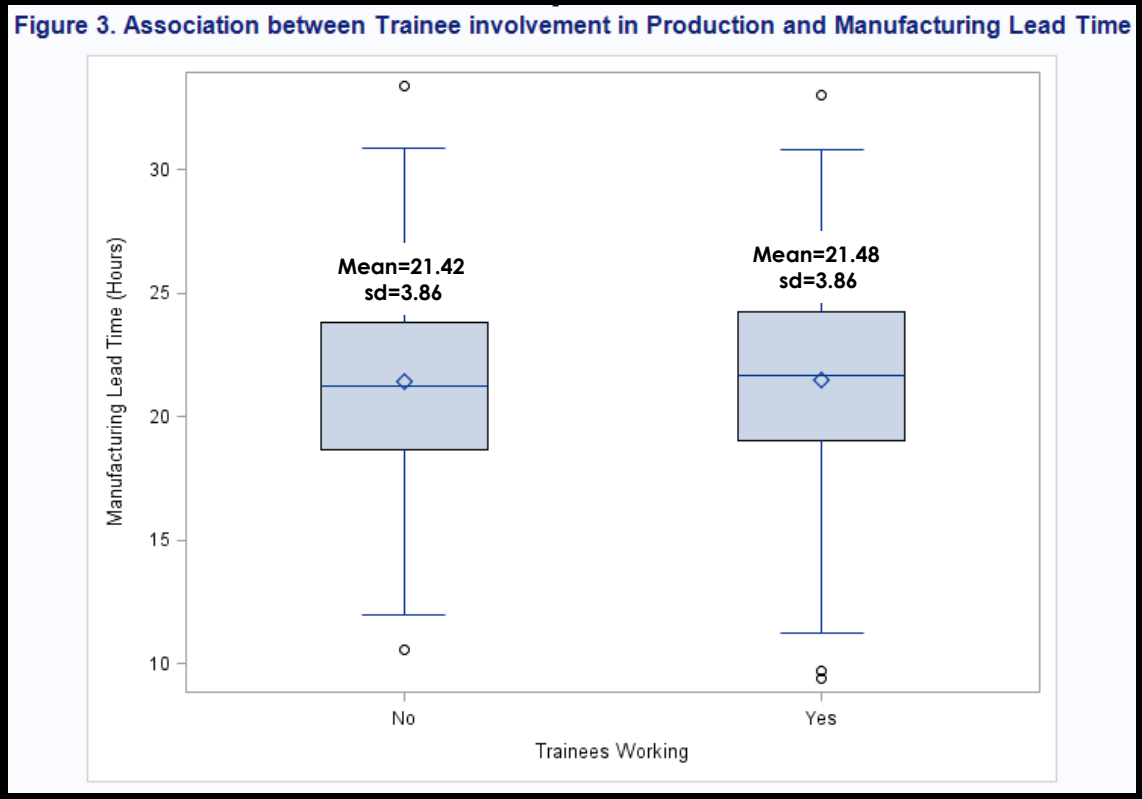



Figure 3. Association between Trainee involvement in Production and Manufacturing Lead Time



# RESULTS – MULTIVARIABLE

Figure 4 shows that 5 of the 6 variables were retained in the model selected by the lasso regression analysis. Only the number of production steps predictor was excluded. The number of ingredient units in stock and the number of shift hours employees worked before beginning production were most strongly associated with manufacturing lead time, followed by equipment failure, trainee involvement in production, and the number of hours of sleep that production workers reported getting the night before their shift began (Table 2). Manufacturing lead times were shorter for batches that had a greater number of ingredients in stock and when production operators reported sleeping for more hours the night prior to batch production. Working more shift hours prior to manufacturing, equipment failure, and having trainees involved in batch production was associated with increased lead times. Together, these 5 predictors accounted for 93.3% of the variance in manufacturing lead time. The mean squared error (MSE) for the test data (MSE=1.03) differed very little from the MSE for the training data (MSE=1.00), which suggests that predictive accuracy did not decline when the lasso regression algorithm developed on the training data set was applied to predict lead manufacturing times in the test data set (Figure 4).

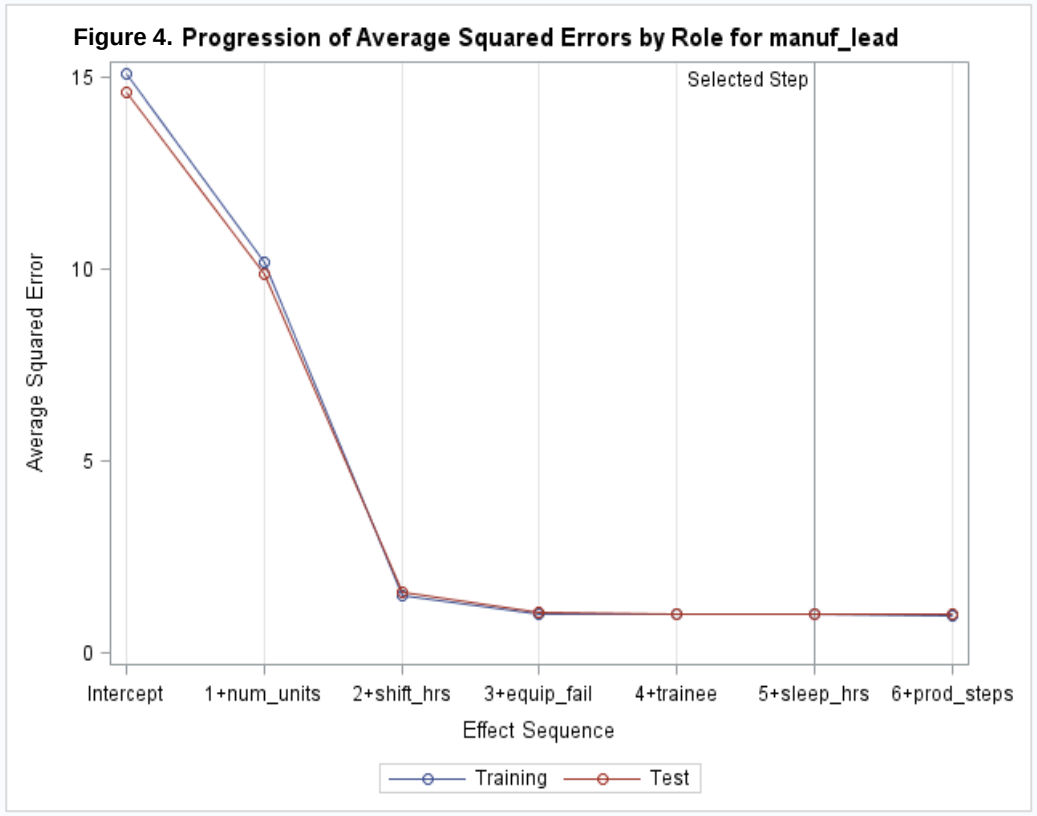



**Table 2. Lasso Least Angle Regression Selection Summary**

| Step | Effect Entered | Number of Effects in Model | ASE | Test ASE | CV PRESS |
|---|---|---|---|---|---|
| 0 | Intercept | 1 | 15.0916 | 14.5909 | 6227.5271 |
| 1 | # Ingredient Units in Stock | 2 | 10.1987 | 9.8558 | 2247.8603 |
| 2 | # Shift Hours Worked | 3 | 1.5187 | 1.5933 | 501.8472 |
| 3 | Equipment Failure | 4 | 1.0180 | 1.0535 | 421.2354 |
| 4 | Trainees Involved in Production | 5 | 1.0014 | 1.0327 | 419.1475 |
| 5 | # Hours of Sleep | 6 | 1.0002 | 1.0310 | 419.1431* |
| 6 | # Production Steps | 7 | 0.9871 | 1.0217 | 421.5203 |

Note: * Optimal Value Of Criterion. ASE=Average Squared Error. CV PRESS=Cross validation selection criterion

# CONCLUSIONS/LIMITATIONS

- **Provide a brief overview of key findings**
  - Should be concise, comprehensive, and easy to understand
- **Discuss implications**
  - Suggestions for how the results might be used
- **Limitations**
  - Summary of the problems identified during the data analytic process
  - Limitations resulting from sampling, data collection, variables, statistical analysis, etc.
  - Cautions (e.g., generalization beyond the study population, unmeasured confounding variables)
- **Future directions**

# CONCLUSIONS/LIMITATIONS

This project used lasso regression analysis to identify a subset of production related variables that best predicted manufacturing lead times in N=435 injection drug production batches manufactured at the Chicago plant from Jan 1, 2015 to December 31, 2015. Manufacturing lead times for these batches ranged from 9.4 hours to 33.4 hours, indicating that there was considerable variability in the amount of time it took to complete production of each batch.

The lasso regression analysis indicated that 5 of the 6 production related predictor variable were selected in the final model. These 5 predictors accounted for 93% of the observed variability in manufacturing lead times. Only the number of production steps involved in the manufacturing of the batch was excluded. The strongest predictors of manufacturing lead times were number of ingredient units in stock and the number of shift hours employees worked before beginning production. Manufacturing lead times were shorter for batches that had a greater number of ingredients in stock and when production operators reported sleeping for more hours the night prior to batch production. Equipment also failure emerged as a predictor of longer manufacturing lead times.

There was minimal increase in the MSE when the training set lasso regression algorithm was used to predict manufacturing lead times in the test data set. This suggests that the predictive accuracy of the algorithm may be stable in future samples of injection drug batches.

# CONCLUSIONS/LIMITATIONS (CONT.)

The results of this project indicate that ensuring an ample supply of product ingredients in stock prior to batch production and reducing operator fatigue by timing production to occur closer to the beginning of production operators' shifts are a priority for achieving consistently shorter manufacturing lead times. These results should be considered carefully in the proposal to adopt a Just In Time (JIT) approach in which product ingredient inventory is kept at a minimum. Although a JIT approach to inventory is estimated to reduce plant operating costs, it may actually increase costs due to increased batch manufacturing lead times if inventory shortages occur. Therefore, if a JIT approach is to be implemented, it must be done so in a way that minimizes the possibility of inventory shortages.

Although having trainees involved in batch production and the number of hours of sleep the night before batch production reported by production operators were both retained in the final lasso regression model, Figure 4 showed that the MSE did not appear to decrease much when these factors were included in the subset of predictors. In addition, neither of these predictors was significantly associated with manufacturing lead time in the bivariate analyses. This suggests that these two production related factors may not have a considerable impact on manufacturing lead times, and re-analysis of these factors should be conducted on future batches to determine whether this conclusion is supported.

This project successfully developed a predictive algorithm for manufacturing lead times that appears to have little bias and variance in a different sample. In addition, it provides more information on which production related factors are most likely to have a significant impact on manufacturing lead times. However, there are some limitations that should be taken into account when considering changes in the manufacturing process based on the results of this project. First, we analyzed only data from a single year, but changes in the manufacturing process are ongoing. So, it is important to test this algorithm in batches that are manufactured in multiple years to determine whether the algorithm remains relatively unbiased and stable despite these ongoing changes. Second, the analysis was conducted on production of injection drugs only. Different factors may emerge as important predictors of manufacturing lead times for drugs with different routes of administration. Therefore, we cannot assume that the predictive algorithm developed in this study will be valid or useful for predicting manufacturing lead times for non-injection drugs. Finally, there is a large number of production related factors in the manufacturing process that could impact manufacturing lead times, but the current project examined only a few of these factors. It is possible that the factors identified as important predictors of manufacturing lead time among the set of predictors analyzed in this project are confounded by other factors not considered in this analysis, As a result. these same factors may not emerge as important factors when other factors are taken into consideration. Therefore, future efforts to develop a solid predictive algorithm for manufacturing lead times should expand the algorithm by adding more production related predictors to the statistical model, and testing the applicability of the algorithm for non-injection drugs.

# WHAT TO AVOID

1. **Using actual data set variable names:**
   a. in place of a description in the measures section
   b. when interpreting results and conclusions
   c. In graphs and figures if possible

2. **Writing about variables and/or analyses that:**
   a. you did not describe in the Methods section
   b. don't relate to the research question

3. **Not describing the direction of an association**

4. **Including figures that:**
   a. are overly complicated
   b. are not relevant to answering the research question
   c. do not enhance understanding of the results

5. **Drawing conclusions that aren't supported by the research**

6. **Using causal language**