Brent Cottrell, Curt Winegar, Deepti Chevvuri, Jacob McCracken, Jessi Richter

STAT 526

Final project

**Inventory data for manufacturing**

For our final project, we started with a list of 1000+ raw material inventory items based on the following criteria:

1. Items must have at least 12 months of transaction history (no new parts)

2. Items will be stored in multiple manufacturing facilities

3. Items will be purchased from multiple suppliers

Response variable: Dollars on hand in current inventory

The following data was provided:

1. Item number

2. Supplier

3. Inventory Level at snapshot in quantity

4. Minimum Order Quantity

5. Box quantity or Lot size

6. Frequency of shipment from supplier (1x per week, 5x per week, etc)

7. Buyer/Planner

8. Item lead time from Supplier

9. Supplier On-time Delivery rate

10. Item Safety Stock level

11. Item cost

12. Annual usage or average daily/weekly usage

Key Question: Could we build a multiple regression analysis that would include explanatory variables which are statistically significant and relevant to inventory dollars on hand?

Our goal is to answer the following questions:

1. Based on a multiple regression analysis, which factor(s) are most important to inventory dollars on hand?

2. How much impact does a change (up or down) in the item lead time from the supplier have on inventory dollars on hand?

3. Example, for a 1 day increase/decrease in lead time, there is a xx% increase/decrease in inventory dollars on hand, assuming all other variables are constant.

4. How much impact does a change (up or down) in the item minimum order quantity from the supplier have on inventory dollars on hand?

5. How much impact does a change (up or down) in the item box or lot size quantity from the supplier have on inventory dollars on hand?

Finally, we want to build a prediction model that answers the following question (if applicable):

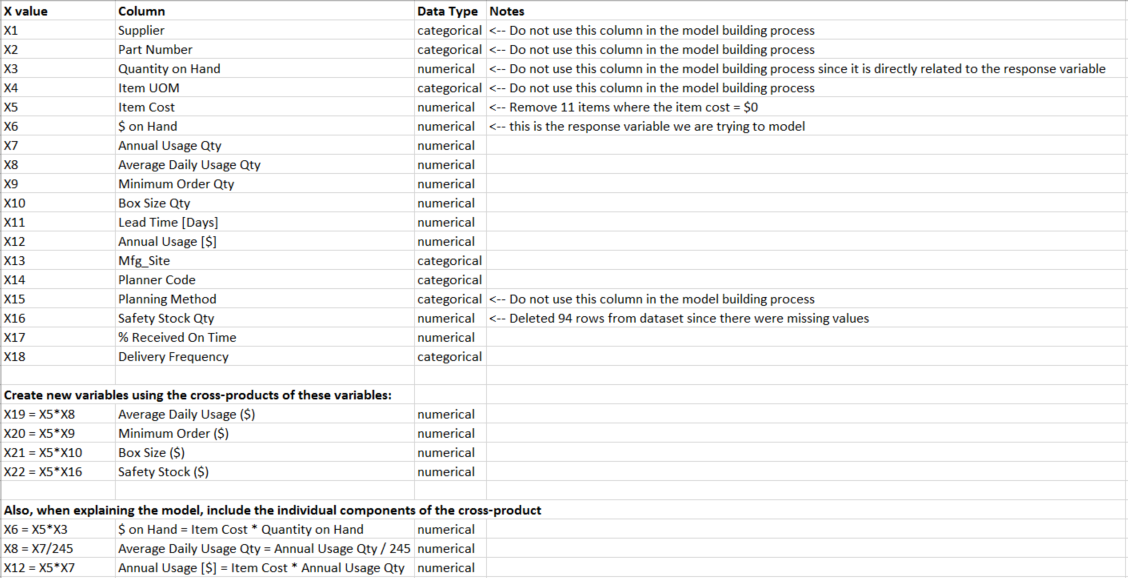
If a new item is going to be added to the production plan, what would be the expected inventory dollars on hand assuming the following attributes:

|  |  |
| --- | --- |
| 1. Lead time is 10 days | 6. Safety stock level is equal to 3 days' usage |
| 2. Item cost is $12 | 7. Supplier delivery rate is 98% on time |
| 3. Minimum Order Quantity is 1000 | 8. Supplier ships on 1 day each week |
| 4. Annual usage is expected to be 250,000 | 9. Site is equal to SITE1 |
| 5. Box quantity is 500 | 10. Buyer/Planner is BUYER11 |

**Getting Started:**

The initial process involved reviewing the raw dataset [InvDataSet.csv] and determining whether the data had missing values that would cause us concern. We found 94 rows of data that were missing values in the Safety Stock Qty column as well as 11 additional rows in the Item Cost column where the cost = $0. Since the data was incorrect or incomplete, we excluded these data from our dataset. In addition, the business wanted a model to predict only those items with a 30 day or less lead time, so we excluded those 42 rows of data as well. After reviewing the dataset and “cleansing” it of incomplete and incorrect data we created [InvDataSet2.csv].

Next, we assigned a label for each column to ease the building of the model in R Studio and identified which columns we would not use during the model building process based on input from the business.



**Analysis Phase 1:**

The first analysis was to review the data using only the numerical variables and excluding the categorical variables. We also chose to use the Annual usage data (X7, X12) and not the Daily usage data (X8, X19) since the daily data is calculated directly from the annual data.

Full Model: X6 ~ X5 + X7 + X9 + X10 + X11 + X12 + X16 + X17 + X20 + X21 + X22

Backward Elimination Model: X6 ~ X7 + X12 + X16 + X17 + X20 + X21 + X22

Forward Selection Model: X6 ~ X22 + X12 + X21 + X20 + X17

Mixed Selection Model: X6 ~ X22 + X12 + X21 + X20 + X17 [*Same as Forward model*]

All combinations: X6 ~ X7 + X16 + X22 + X12 + X21 + X20 + X17 [*Same as Backward model*]

Table 1: Table of Summaries for the 5 models fit

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Full Model | Forward Selection | Backward Elimination | Mixed | All Combinations |
| Error df | 851 | 857 | 855 | 857 | 855 |
|  | 0.7887 | 0.7888 | 0.7891 | 0.7888 | 0.7891 |
| RMSE | 1040 | 1040 | 1039 | 1040 | 1039 |
| Largest VIF | 7.49  (X16) | 2.892  (X22) | 7.29  (X16) | 2.892  (X22) | 7.29  (X16) |
| *p*-value for F-test | <2.2e-16 | <2.2e-16 | <2.2e-16 | <2.2e-16 | <2.2e-16 |
| Model Useful (Yes/No) | Yes | Yes | Yes | Yes | Yes |
| Largest p-value (t-test) | .8629  (X5) | 0.0613  (X17) | .1586  (X16) | 0.0613  (X17) | .1586  (X16) |
| AIC | 14453 | 14447 | 14448 | 14447 | 14448 |

**Results:**

This analysis resulted in essentially 2 different models, one model with 7 variables, and the other model with 5 explanatory variables.

**Analysis Phase 2:**

The second analysis was to review the data using only the numerical variables and excluding the categorical variables. This time we chose to use the Daily usage data (X8, X19) and not the Annual usage data (X7, X12) as tested in Phase 1.

Full Model: X6 ~ X5 + X8 + X9 + X10 + X11 + X16 + X17 + X19 + X20 + X21 + X22

Backward Elimination Model: X6 ~ X17 + X19 + X20 + X21 + X22

Forward Selection Model: X6 ~ X22 + X19 + X21 + X20 + X17

Mixed Selection Model: X6 ~ X22 + X19 + X21 + X20 + X17 [*Same as Forward model*]

All combinations: X6 ~ X17 + X19 + X20 + X21 + X22 [*Same as Backward model*]

Table 2: Table of Summaries for the 5 models fit

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Full Model | Forward Selection | Backward Elimination | Mixed | All Combinations |
| Error df | 851 | 857 | 857 | 857 | 857 |
|  | 0.7885 | 0.7888 | 0.7888 | 0.7888 | 0.7888 |
| RMSE | 1041 | 1040 | 1040 | 1040 | 1040 |
| Largest VIF | 7.45  (X16) | 2.89  (X22) | 2.89  (X22) | 2.89  (X22) | 2.89  (X22) |
| *p*-value for F-test | <2.2e-16 | <2.2e-16 | <2.2e-16 | <2.2e-16 | <2.2e-16 |
| Model Useful (Yes/No) | Yes | Yes | Yes | Yes | Yes |
| Largest p-value (t-test) | 0.9644  (X5) | 0.0613  (X17) | 0.0613  (X17) | 0.0613  (X17) | 0.0613  (X17) |
| AIC | 14454 | 14447 | 14447 | 14447 | 14447 |

**Results:**

This analysis resulted in 5 models created, however all the models use the same explanatory variables, just in a different order based on each selection criteria.

**Analysis Phase 3:**

The third analysis was to review the data now including both the numerical variables and the categorical variables. We chose to include both the Daily usage data (X8, X19) and the Annual usage data (X7, X12) in this model selection process.

Full Model: X6 ~ X5 + X7 + X8 + X9 + X10 + X11 + X12 + X13 + X16 + X17 + X18 + X19 + X20 + X21 + X22

Backward Elimination Model: X6 ~ X12 + X13 + X20 + X21 + X22

Forward Selection Model: X6 ~ X22 + X19 + X21 + X13 + X20

Mixed Selection Model: X6 ~ X22 + X19 + X21 + X13 + X20 [*Same as Forward model*]

All combinations: X6 ~ X12 + X13 + +X19 + X20 + X21 + X22

Table 3: Table of Summaries for the 5 models fit

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Full Model | Forward Selection | Backward Elimination | Mixed | All Combinations |
| Error df | 836 | 850 | 850 | 850 | 849 |
|  | 0.7911 | 0.7914 | 0.7913 | 0.7914 | 0.7911 |
| RMSE | 1034 | 1034 | 1034 | 1034 | 1034 |
| Largest VIF | 1.52e+10  (X7 & X8) | 1.73  (X22) | 3.017  (X22) | 1.73  (X22) | 2096  (X19) |
| *p*-value for F-test | <2.2e-16 | <2.2e-16 | <2.2e-16 | <2.2e-16 | <2.2e-16 |
| Model Useful (Yes/No) | No | Yes | Yes | Yes | No |
| Largest p-value (t-test) | 0.998  (X5) | 0.6399  (X13) | 0.676  (X13) | 0.6399  (X13) | 0.9113  (X5) |
| AIC | 14454 | 14443 | 14444 | 14443 | 14445 |

**Results:**

Phase 3 analysis resulted in 4 unique models created and all have a higher value than previous analysis in Phase 1 and Phase 2. After reviewing the models created, using the X13 explanatory variable did not provide a good fit for the business purpose of analyzing raw material inventory dollars on hand.

**Analysis Phase 4:**

The fourth and final analysis was to review the business case again and we decided to simplify the model to include only the explanatory variables that are externally set by the suppliers (Lead time, minimum order, box size, item cost) and exclude the variables that are internally managed by the business (safety stock, planner code, manufacturing site).

Full Model: X6 ~ X5 + X7 + X8 + X9 + X10 + X11 + X12 + X17 + X18 + X19 + X20 + X21

Backward Elimination Model: X6 ~ X11 + X17 + X19 + X20 + X21

Forward Selection Model: X6 ~ X19 + X21 + X20 + X17 + X11

Mixed Selection Model: X6 ~ X19 + X21 + X20 + X17 + X11 [*Same as Forward model*]

All combinations: X6 ~ X19 + X21 + X20 + X17 + X11 [*Same as Forward model*]

Table 4: Table of Summaries for the 5 models fit

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Full Model | Forward Selection | Backward Elimination | Mixed | All Combinations |
| Error df | 846 | 857 | 857 | 857 | 857 |
|  | 0.7403 | 0.7405 | 0.7405 | 0.7405 | 0.7405 |
| RMSE | 1153 | 1153 | 1153 | 1153 | 1153 |
| Largest VIF | 1.499e+10  (X7 & X8) | 1.526  (X21) | 1.526  (X21) | 1.526  (X21) | 1.526  (X21) |
| *p*-value for F-test | <2.2e-16 | <2.2e-16 | <2.2e-16 | <2.2e-16 | <2.2e-16 |
| Model Useful (Yes/No) | No | Yes | Yes | Yes | Yes |
| Largest p-value (t-test) | 0.8928  (X18) | 0.090  (X11) | 0.090  (X11) | 0.090  (X11) | 0.090  (X11) |
| AIC | 14636 | 14625 | 14625 | 14625 | 14625 |

**Results:**

This analysis resulted in 1 model created, each selection criteria selected the same 5 explanatory variables, just in a slightly different order.

**Final Model Output:**

X6 = -260.37632 + 6.47288(X19) + 0.89401(X21) + 0.07447(X20) + 3.06564(X17) + 13.52305(X11)

|  |  |
| --- | --- |
| Variable | Description |
| X6 | $ on Hand |
| X11 | Lead Time [Days] |
| X17 | % Received On Time |
| X19 | Average Daily Usage ($) |
| X20 | Minimum Order ($) |
| X21 | Box Size ($) |

**Final Model - Testing:**

X6 = -260.37632 + 6.47288(X19) + 0.89401(X21) + 0.07447(X20) + 3.06564(X17) + 13.52305(X11)

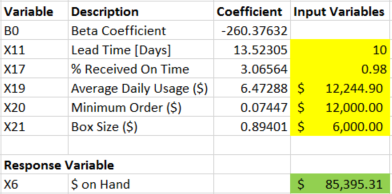
|  |  |
| --- | --- |
| Variable | Description |
| X6 | $ on Hand |
| X11 | Lead Time [Days] |
| X17 | % Received On Time |
| X19 | Average Daily Usage ($) |
| X20 | Minimum Order ($) |
| X21 | Box Size ($) |

Using this model we built an excel spreadsheet for the business that answers the following question:

If a new item is going to be added to the production plan, what would be the expected inventory dollars on hand assuming the following attributes:

|  |  |
| --- | --- |
| 1. Lead time is 10 days | 6. Safety stock level is equal to 3 days' usage |
| 2. Item cost is $12 | 7. Supplier delivery rate is 98% on time |
| 3. Minimum Order Quantity is 1000 | 8. Supplier ships on 1 day each week |
| 4. Annual usage is expected to be 250,000 | 9. Site is equal to SITE1 |
| 5. Box quantity is 500 | 10. Buyer/Planner is BUYER11 |

**Output:**



Based on our model, we would expect that an item with these explanatory variables would have approximately $85,395 on hand after 12 months of usage. Since our model was built based only on items that have been used for at least 12 months, we must specify to the user as such in our explanation of the model.

**This is the Homework 6 data that we pulled together, there are some key take-aways from this assignment that could be valuable in providing our final project output.**

1. (a) i. Variables which appear to have a strong linear relationship to MidPrice: Horsepower, FuelTank, Weight

ii. Variables which appear to have a positive linear relationship with MidPrice: EngineSize, Horsepower, Fueltank, Length, Wheelbase, Width, Weight

iii. The strongest linear relationship between two explanatory variables is Weight with FuelTank, where r = .8905.

1. (b) Table 1: Table of Summaries for the 5 models fit

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Full Model | Forward Selection | Backward Elimination | Mixed | All Combinations |
| Error df | 82 | 85 | 84 | 85 | 86 |
|  | .7089 | .7064 | .7096 | .7064 | .7044 |
| RMSE | 4.378 | 4.397 | 4.373 | 4.397 | 4.412 |
| Largest VIF | 14.153  (Weight) | 5.262  (Width) | 6.888  (Width) | 5.262  (Width) | 6.040  (Weight) |
| *p*-value for F-test | <2.2e-16 | <2.2e-16 | <2.2e-16 | <2.2e-16 | <2.2e-16 |
| Model Useful (Yes/No) | Yes | Yes | Yes | Yes | Yes |
| Largest p-value (t-test) | .366  (FuelTank) | .077  (FuelTank) | .123  (Engine Size) | .077  (FuelTank) | 0.023  (Length) |
| AIC | 277.28 | 275.329 | 275.223 | 275.329 | 274.99 |

i. Yes, we would consider eliminating one or more of the explanatory variables because we have identified at least one explanatory variable with a VIF > 10 which suggests that multicollinearity may be a problem. Additionally, one could argue that we should always consider eliminating explanatory variables if we’re serious about model selection. This is one of the reasons why we may be interested in AIC and SBC/BIC analysis. In this example, we would consider removing REV from the model first since it has a high p-value and the lowest correlation with the response variable.

ii. We can check for multicollinearity by seeing if the direction of correlation is opposite of the slope (beta coefficient) in the least regression model for any of the explanatory variables. In other words if we have an explanatory variable x1 which is observed to have a positive association with some response variable y, we should expect that this variable’s beta coefficient is also positive. In this case, we see two variables that fit this definition. Width has a positive correlation but a negative slope (-1.231). REV has a negative correlation but a positive slope in the least squares regression model. The correlations here are not strong so these are not of great concern. Another check is to look at variables with strong correlation and their slope coefficient significance. In this case, an individual t-test for a beta coefficient may suggest that it is not significantly associated with y, even though theoretically that variable should be highly correlated with y. In the context of this example we have several explanatory variables with low t-values despite moderate correlations to MidPrice. The most notable (3) include FuelTank (t value = .908, p = .366), Wheelbase (t value = .98, p = .33), and Weight (t value = .938, p = .351).

1. (c) Backward Elimination Model:

MidPrice ~ EngineSize + Horsepower + REV + Length + Width + Weight

See values in Table 1

1. (d) Forward Selection Model:

MidPrice ~ Horsepower + Wheelbase + Width + Length + FuelTank

See values in Table 1

1. (e) Tool 1 – VIF: The highest VIF of the variables used the forward model is 5.23 which is considerably less than 10 so the results of this tool do not create concern for multicollinearity.

Tool 2 - We check for variables were the direction of correlation is opposite the slope in least square regression model. We see that Width has a negative slope but positive correlation. Because the correlation is not strong, it is not of great concern.

Tool 3 – We check for variables with strong correlation but non-significant slope coefficients. The only variable with a strong correlation and appears in our model is Horsepower. It’s coefficient is statistically significant so it does not cause concern for multicollinearity.

Based on all three tests, multicollinearity does not appear to be present in the final model using the Forward Selection process.

1. (f) Mixed Selection Model:

MidPrice ~ Horsepower + Wheelbase + Width + Length + FuelTank

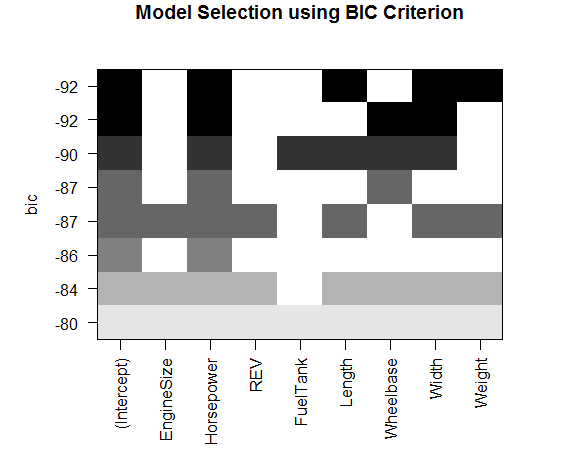
See values in Table 1

1. (g) All Combinations Model:

MidPrice ~ Horsepower + Length + Width + Weight

See values in Table 1

1. (h) The variables are shown on the x-axis and BIC criterion is shown on y-axis. The color shading indicates which variables were included in each model and gets darker as the BIC gets smaller. We are interested in the model which produces the lowest BIC value so we look at the top of the graph (-92) for this model selection. Horsepower, Width, & Length were included in all of the “best” models



1. (i) Backward

1. (j) All Combinations

1. (k) Backward

1. (l) Forward or Mixed (this was the same model for both selection criterion)

1. (m) When creating a model an analyst would ideally have previous knowledge of typical relationships between the variables of interest. If there is a theoretical explanation of the form of the relationships between variables it should be used to our advantage when building a model. At the very least, the knowledge of important variables should be used when deciding which variables to include in a model.

Keeping your intended audience in mind is important. If you have identified multicollinearity and need to choose between two variables to keep in the model, you may want to consider which variable your audience would find more familiar. Presenting data that is not easily understood by your audience makes it less valuable.

Another reason to choose one variable over another would be the ongoing cost and effort associated with obtaining and measuring the values associated with the variables.