**EXECUTIVE SUMMARY**

**Problem:** Data for 25 days’ profit measurements have been obtained. However, since day 14 had a missing entry, we excluded this observation from the analysis, and did the analysis only on remaining 24 observations. These 24 observations consist of four variables: PROFIT (dependent variable), and three (independent) variables MATERIAL A, MATERIAL B, and DAY (where PROFIT, MATERIAL A, MATERIAL B, and DAY are referred to as P, A, B, and O respectively). We are trying to predict the profit (p) (in $100) on a certain day given material A (in gallons), material B (in lbs). We have been asked to give a predicted profit (in $100) on day 26 when material A = 35 gallons and material B = 95 lbs. Also, we are asked to give a 90% confidence interval for this prediction.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Mean** | **Std. Dev.** | **Minimum** | **Maximum** | **Shape of Distribution** |
| Profit (in $100) | 240.86 | 57.54 | 142.4 | 348.7 | Unimodal and less symmetric |
| Material A (in gallons) | 78.625 | 16.19 | 56 | 109 | Bimodal and non-symmetric |
| Material B (in lbs) | 62 | 2.519 | 58.2 | 66.6 | Unimodal and less symmetric |
| Day\* | 12.958 | 7.515 | 1 | 25 | Uniform |

**Recommended Model without the Order Variable:**

(Profit (P) in units of $100; Material A (A) in gallons; Material B (B) in lbs.) = 57.538 (in $100), R2=0 (because the recommended model is 0th order);

Using the 0th order model, on day 26 when Material A is 35 gallons and Material B is 95 lbs, we would **predict a profit** of 240.86 (in $100) with 90% confidence interval for this prediction given by the interval (140.21, 341.51), (in $100). (caution: Predictions outside of the domain on which the best model was developed can be extremely wrong because the model developed may not be at all correct.)

Histogram of the residuals for the recommended model appears to be uniform and less symmetric, and therefore, we reject that they are normally distributed. (Assumption 3 violated)

Not sure if the residuals are correlated positively or negatively. (Not enough evidence to conclude if Assumption 4 is violated)

While examining the scatter plot of Profit Vs Day scatterplot, we found that as Day increases, the predicted profit increases. Also, examining the correlation matrix, we found that, the correlation between Profit and order variable Day is higher than the correlation between profit and any other independent variable. Therefore, we ran a second set of models to find the best model with all the independent variables including the order variable.

**Recommended Model with the Order Variable:**

Profit (P) in units of $100, Material A (A) in gallons; Material B (B) in lbs.) (in $100), R2=0.32191

Using this first order model (without the variables A and B), on day 26 when Material A is 35 gallons and material B is 95 lbs, we would **predict a profit** of 297.52 (in $100) with 90% confidence interval for this prediction given by the interval (207.45, 387.59), (in $100). (caution: Predictions outside of the domain on which the best model was developed can be extremely wrong because the model developed may not be at all correct.)

Histogram of the residuals for the recommended model appears to be unimodal and less symmetric, and therefore, we reject that they are normally distributed. (Assumption 3 violated)

Negatively correlated residuals. (Possible Assumption 4 violation)

**ANALYSIS**

**Problem:** Data for 25 days’ profit measurements have been obtained. However, since day 14 had a missing entry, we excluded this observation from the analysis, and did the analysis only on remaining 24 observations. These 24 observations consist of four variables: PROFIT (dependent variable), and three (independent) variables MATERIAL A, MATERIAL B, and DAY (where PROFIT, MATERIAL A, MATERIAL B, and DAY are referred to as P, A, B, and O respectively). We are trying to predict the profit (p) (in $100) on a certain day given material A (in gallons), material B (in lbs). We have been asked to give a predicted profit (in $100) on day 26 when material A = 35 gallons and material B = 95 lbs. Also, we are asked to give a 90% confidence interval for this prediction.

**Data:** The data has been entered into the computer and printed out (p. 2a). The data has been checked for accuracy and has been verified to be the same as the data provided to us.

**Profit:** (in $100), The dependent variable has an average of 240.86, standard deviation of 57.54, ranges from a minimum value of 142.4 to a maximum of 348.7. The shape of the distribution appears to be unimodal and less symmetric. (p. 2b,3a)

**Material A:** (in gallons), An independent variable has an average of 78.625, standard deviation of 16.19, ranges from a minimum value of 56 to a maximum value of 109. The shape of the distribution appears to bimodal and non-symmetric. (p. 2b,3a)

**Material B:** (in lbs), An independent variable has an average of 62, standard deviation of 2.519, ranges from a minimum value of 58.2 to a maximum value of 66.6. The shape of the distribution appears to be unimodal and less symmetric. (p. 2b,3a)

**Day:** An independent variable has an average of 12.958, standard deviation of 7.515, ranges from a minimum value of 1 to a maximum value of 25. The shape of the distribution appears to be uniform. (p. 2b,3a)

**Correlation Results:** Calculated the correlations of all pairs of variables including the order variables. Significant results are described below:

**Income (I) Vs Independent Variables:** Examining the correlation matrix (p. 2c), we notice the following significant result: the correlation between:

Profit Vs Day: r = 0.5674, about 32% of the variability in the y scores around is explained by simple regression between Profit and Day. The standard deviation of the y scores about a simple regression using Day is approximately 0.82 times the standard deviation of the y scores about . As Day increases, predicted profit () increases in this fitted simple regression model.

**Correlation Between Pairs of Independent Variables:** Examining the correlation matrix (p. 2c), we notice the following significant result: the correlation between:

Material B Vs Day: r = 0.7143, about 51% of the variability in the y scores around is explained by simple regression between Material B and Day. The standard deviation of the y scores about a simple regression using Day is approximately 0.70 times the standard deviation of the y scores about . As Day increases, predicted amount of Material B increases in this fitted simple regression model.

**Scatterplot Results:** Calculated the scatterplots (p. 3b) of all pairs of variables including the order variables. Significant results are described below:

**Scatterplots of Income Vs Independent Variable:** Describe the relationship:

Profit Vs Day: As Day increases, predicted Profit increases.

**Scatterplots of Independent Variables:** Describe the relationship:

Material B Vs Day: As Day increases from 1 to 13, the predicted Material B remains constant below the average Material B; as Day increases from 15 to 25, the predicted Material B remains constant above the average Material B.

**Fit of First Order Model (without the order variable):** E(P) = β0 + β1A + β2B (p. 5a). Examining the EXCEL results for this model, we that find =55.888 (in $100) vs =57.538 (in $100). R2=0.13856. For testing H0: β1= β2=0 (since ~~A~~, B holds), we find the p for Global F = 0.21 > 0.05 = α. Therefore, we accept H0. We conclude that the first order model is not significantly better than the model E(I) = β0. Thus, the current best model is still E(P) = β0.

**Fit of Second Order Model (without the order variable):** E(P) = β0 + β1A + β2B + β3A2 + β4B2 + β5AB (p. 6a). Examining the EXCEL results for this model, we find that standard deviation of the residual about the second order model is 52.976 (in $100), which we note is not significantly lower than =57.538. R2=0.33656. For testing H0: β1= β2= β3= β4= β5=0 (since ~~A~~, B holds), we find the p for Global F = 0.16 > 0.05 = α. Therefore, we accept H0. We conclude that the second order model is not significantly better than the model E(I) = β0. Thus, the current best model is still E(P) = β0.

**Residual Plot for the Second Order Model Fitted (without the order variable):** Examining the residual plot for the second order model (p. 6c), we note that the mean of the residuals seems to be zero regardless of the value of the predicted. Assumption 1 holds. If assumption 1 was violated, we would consider fitting a higher order model than has been fit so far and compare it to the last acceptable model.

**Since the current best model is the 0th order mode. Since, , for this particular data set, 0th order model is .**

**Residual Plot for the FINAL (0th Order) Model (without the order variable):** Examining the residual plot for this model (p. 4c), we note:

1. The mean of the residuals regardless of the value of the predicted seems to be 0.
2. The variance of the residuals regardless of the value of the predicted seems to be constant.

**Histogram of Residuals for the FINAL (0th Order) Model (without the order variable):** Examining the histogram of the residuals for this model (p. 4b), we see that they appear to be uniform and less symmetric. Thus, we can reject that they are normally distributed.

**Independence of the Residuals for the FINAL (0th Order) Model (without the order variable):** Given that the data has been ordered by the order it was collected, we calculate the Durbin-Watson statistic (p. 4e) and examine the plot of residual(t) vs residual(t-1) (p. 4d). Below are the interpretations:

1. Based on the plot residual(t) vs residual(t-1), it appears that a line passing thru this data possibly have a slightly negative slope. Which implies that there is a slightly negative correlation.
2. Calculating the Durbin-Watson statistic, we get the value of d=1.933, which indicates that there is a possibility of slightly positive correlation between residual(t) and residual(t-1).
3. Since we are fitting the regressing model (p. 4f), for testing H0: β1=0 (since A holds), we find p for t = 0.97 > 0.05 = α. Which means we accept H0 (that is β1=0). r = -0.00789. r2=6x10-5. Sqrt(1-r2) = 0.99994.

After reviewing the best case for a violation of A4 in **a**, **b** and **c** above, there is not sufficient evidence to conclude that A4 is violated.

The **predicted model (without order variable)** is given by

A 90% confidence interval for the predicted income is given by:

1. Predicted profit on day 26, when Material A is 35 gallons (A\*=35) and Material B is 95 lbs (B\*=95): **(in $100**)

(caution: Predictions outside of the domain on which the best model was developed can be extremely wrong because the model developed may not be at all correct.)

1. A 90% confidence interval for this prediction is given by:

tdferror = n – (# of β’s in E(y))

= 24-1 = 23

[predicted profit evaluated at A=35, B=95]

240.86 = 240.86 **140.21, 341.51) (in $100)**

(caution: Predictions outside of the domain on which the best model was developed can be extremely wrong because the model developed may not be at all correct.)

**Conclusion:** We recommend using the 0th order model given above to describe the relationship (without the order variable).

**[Since we have observed a relationship between the Profit and the order variable Day, described above as Day increases, predicted profit increases, and observed the correlation between these two variables which indicate that there is a positive correlation between Profit and Day, we are going to repeat the model building part of the Analysis to find the best model with all independent variables including the Order variable Day.]**

**Fit of First Order Model (with the order variable):** E(P) = β0 + β1A + β2B + β3D (p. 7a). Examining the EXCEL results for this model, we that find =50.522 (in $100) vs =57.538 (in $100). R2=0.3296. For testing H0: β1= β2= β3=0 (since ~~A~~, B holds), we find the p for Global F = 0.042 < 0.05 = α. Therefore, we reject H0. We conclude that the first order model is significantly better than the model E(P) = β0. Thus, the current best model moves to E(P) = β0 + β1A + β2B + β3D.

**Fit of Second Order Model (with the order variable):** E(P) = β0 + β1A + β2B + β3D + β4A2 + β5B2 + β6D2 + β7AB + β8AD + β9BD (p. 8a). Examining the EXCEL results for this model, we find that standard deviation of the residual about the second order model is 55.137 (in $100), which we note is not significantly lower than the standard deviation of the y scores about the first order model. R2=0.4410. For testing H0: β4= β5= β6= β7= β8= β9=0 (since ~~A~~, ~~B~~, C holds), we find that partial F is obtain by [(SSEred – SSEcomplete)/(dferror,red-dferror,complete)]/MSEcompletea = [(51045-42561)/(20-14)]/3040 = [1414]/[3040] = 0.465. We reject this value if it is too large. Note since the cutoff point with α=0.05 on an F6,14 is approximately 2.85 (Reject H0 if Fobs > 2.85, accept H0 otherwise), we accept H0, and thus p > 0.05. We conclude that the second order model is not significantly better than the first order model. Therefore, the current model is still the first order model.

**Residual Plot for the Second Order Model Fitted (with the order variable):** Examining the residual plot for the second order model (p. 8c), we note that the mean of the residuals seems to be zero regardless of the value of the predicted. Assumption 1 holds. If assumption 1 was violated, we would consider fitting a higher order model than has been fit so far and compare it to the last acceptable model.

**Return to the 1st Order Model (with the order variable):** we now examine the 1st order model, to see if we can drop any of the variables. We only look at the higher order terms, i.e. terms not included in higher order terms in the model, and individually test; p. 7a (A holds since we are testing single β=0).

H0: β1=0 vs H1: Not H0; (A holds) [p=0.76 > 0.05, accept H0];

H0: β2=0 vs H1: Not H0; (A holds) [p=0.80 > 0.05, accept H0];

H0: β3=0 vs H1: Not H0; (A holds) [p=0.027 < 0.05, reject H0];

If we find one or more p value above α=0.05, we drop the one with the highest p-value and refit the model without this term. In this case we drop the B term and refit.

**Fit for the Reduced 1st Order of Form (with order variable):** E(P) = β0 + β1A + β3D (p. 9a)

The standard deviation of the y scores about this model is 49.389. R2 = 0.3273

We only look at the higher order terms and individually test

H0: β1=0 vs H1: Not H0; (A holds) [p=0.69 > 0.05, accept H0];

H0: β3=0 vs H1: Not H0; (A holds) [p=0.0044 < 0.05, reject H0];

If we find one or more p value above α=0.05, we drop the one with the highest p-value and refit the model without this term. In this case we drop the A term and refit.

**Fit for the Reduced 1st Order of Form (with order variable):** E(P) = β0 + β3D (p. 10a)

The standard deviation of the y scores about this model is 48.445. R2 = 0.3219

We only look at the higher order terms and individually test

H0: β3=0 vs H1: Not H0; (A holds) [p=0.0038 < 0.05, reject H0];

If we find one or more p value above α=0.05, we drop the one with the highest p-value and refit the model without this term. In this case, we have no more terms to drop.

**Residual Plot for the FINAL (1st Order) Reduced Model (without the order variable):** Examining the residual plot for this model (p. 10c), we note:

1. The mean of the residuals regardless of the value of the predicted seems to be 0.
2. The variance of the residuals regardless of the value of the predicted seems to be constant.

**Histogram of Residuals for the FINAL (1st Order) Reduced Model (with the order variable):** Examining the histogram of the residuals for this model (p. 10b), we see that they appear to unimodal and less symmetric. Thus, we can reject that they are normally distributed.

**Independence of the Residuals for the FINAL (1st Order) Reduced Model (with the order variable):** Given that the data has been ordered by the order it was collected, we calculate the Durbin-Watson statistic (p. 10d) and examine the plot of residual(t) vs residual(t-1) (p. 10e). Below are the interpretations:

1. Based on the plot residual(t) vs residual(t-1), it appears that a line passing thru this data possibly have a slightly negative slope. Which implies that there a slightly negative correlation.
2. Calculating the Durbin-Watson statistic, we get the value of d=2.843, which indicates that there is a negative correlation between residual(t) and residual(t-1).
3. Since we are fitting the regressing model (p. 10f), for testing H0: β1=0 (since A holds), we find p for t = 0.044 < 0.05 = α. Which means we reject H0 (that is, β1<0). r = -0.424. r2=0.1798. Sqrt(1-r2) = 0.9057.

After reviewing the best case for a violation of A4 in **a**, **b** and **c** above, there is sufficient evidence to conclude that A4 is possibly violated, negatively correlated residuals.

The **predicted model (with order variable)** is given by

A 90% confidence interval for the predicted income is given by:

[predicted profit evaluated at A=A\*, B=B\*, D=D\*]

1. Predicted profit on day 26 (D\*=26) when Material A is 35 gallons (A\*=35) and Material B is 95 lbs (B\*=95):

**(in $100**)

(caution: A=35 [56,109], B=95 [58.2,66.6], D=26 [1,25])

1. A 90% confidence interval for this prediction is given by:
2. [predicted profit evaluated at A=A\*, B=B\*, D=D\*]

tdferror = n – (# of β’s in E(y))

= 24-2 = 22

297.52

297.52 **207.45, 387.59) (in $100)**

(caution: Predictions outside of the domain on which the best model was developed can be extremely wrong because the model developed may not be at all correct.)

**Conclusion:** We recommend using the 1st order model (without variable A and B) given on above to describe the relationship (with the order variable).