**ANALYSIS**

**Problem:** Data for a random sample of profit measurements on 25 individuals. However, observation 14 has a missing entry, so we are excluding the entire observation. Input the remaining 24 observations. Our object is to characterize the dependent variable PROFIT (P) (in $100). Along with the value of Profit for 24 individuals, we have been provided with the value of two independent variables: Material A (gallons), Material B (lbs), and the days that the data was collected over 25 consecutive days and is listed in Day. Besides characterizing the relationship between these four variables, we wish to predict PROFIT for an individual with Material A=35 gallons and Material B=95 lbs on day 26, and give a 90% confidence interval for these predictions.

Data: the data has been entered into the computer and printed out (p.1a). 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 222.58, standard deviation of 54.34, ranges from a minimum value of 129.60 to a maximum of 324.40. The shape of the distribution appears unimodal. (p.1c)

Day: an independent variable, has an average of 12.96, standard deviation of 7.52, ranges from a minimum value of 1 to a maximum of 25. The shape of the distribution appears uniform. (p.1c)

Material A: an independent variable, has an average of 43.78 gallons, standard deviation of 9.71 gallons, ranges from a minimum value of 30.20 gallons to a maximum of 62 gallons. The shape of the distribution appears unimodal, skewed to the right. (p.1c)

Material B: an independent variable, has an average of 71.13 lbs, standard deviation of 2.94 lbs, ranges from a minimum value of 66.70 lbs to a maximum of 76.50 lbs. The shape of the distribution appears unimodal, skewed to the right. (p.1c)

**PROFIT (in $100) vs independent variables**: examining the corr. matrix (p.1d), we see the following significant results: the corr. between:

PROFIT vs DAY: r= 0.57, about 32.49% of the variability in the y scores around y-bar is explained by simple regression between PROFIT and DAY. The std. dev. of the y scores about a simple linear regression using DAY is approximately 0.82 times the std. dev. of the y scores about y-bar. As DAY increases, y-hat increases in this fitted simple regression model.

PROFIT vs Material B: r= 0.36, about 12.96% of the variability in the y scores around y-bar is explained by simple regression between PROFIT and Material B. The std. dev. of the y scores about a simple linear regression using Material B is approximately 0.93 times the std. dev. of the y scores about y-bar. As Material B increases, y-hat increases in this fitted simple regression model.

PROFIT vs Material A: r= 0.016, about 0.0256% of the variability in the y scores around y-bar is explained by simple regression between PROFIT and Material A. The std. dev. of the y scores about a simple linear regression using Material A is approximately 0.99 times the std. dev. of the y scores about y-bar. As Material A increases, y-hat increases in this fitted simple regression model.

**Correlations between pairs of independent variables**: Examining the correlation matrix (p1d), we see the following significant results: for DAY vs Material B, r= 0.71, about 50.41% of the variability in the y scores around y-bar is explained by simple regression between DAY and Material B. The std. dev. of the y scores about a simple linear regression using Material B is approximately 0.70 times the std. dev. of the y scores about y-bar. As Material B increases, y-hat increases in this fitted simple regression model.

For Material B vs Material A, r= 0.31, about 9.61% of the variability in the y scores around y-bar is explained by simple regression between Material B and Material A. The std. dev. of the y scores about a simple linear regression using Material A is approximately 0.95 times the std. dev. of the y scores about y-bar. As Material A increases, y-hat increases in this fitted simple regression model.

For DAY vs Material A, r= 0.16, about 2.56% of the variability in the y scores around y-bar is explained by simple regression between DAY and Material A. The std. dev. of the y scores about a simple linear regression using Material A is approximately 0.99 times the std. dev. of the y scores about y-bar. As Material A increases, y-hat increases in this fitted simple regression model.

**Scatterplots of PROFIT vs independent variables:** Describe the relationship:

PROFIT vs Material A: There doesn’t appears to be a linear relationship between these two variables.

PROFIT vs Material B: There appears to be a positive linear relationship between these two variables.

PROFIT vs DAY: There appears to be a positive linear relationship between these two variables.

**Scatterplots between pairs of independent variables:**

Material A vs DAY: Describe the relationship: There appears to be a slightly positive linear relationship between these two variables.

Material B vs DAY: Describe the relationship: There appears to be a positive linear relationship with a jump between these two variables.

Material A vs Material B: Describe the relationship: There appears to be a slightly positive linear relationship between these two variables.

There is significant relationship found between DAY and Material B and PROFIT after examining correlation matrix and scatterplots.

**Fit of zero order model w/o DAY**: E(P)=β0 (p.4c). Examining the EXCEL results for this model, we find sP,A= $ 52.78 vs SP=54.34, R2= 0.047. For testing H0: β0=μ=0, (since A holds), we find the p for the tobs= 0<0.05=α. Thus, the zero order model is a significant improvement.

**Fit of first order model w/o DAY**: E(P)=β0+β1\*A+β2\*B (p.4a). Examining the EXCEL results for this model, we find sP,A= $ 54.26 vs SP=54.34, R2= 0.14. For testing H0: β1=β2=0, (since ~~A~~B holds), we find the p for the Global F= 0.21> 0.05=α. Thus, the first order model is not a significant improvement over the model E(P)= β0.

**Fit of second order model w/o DAY**: (p.4b) E(P)=β0+β1\*A+β2\*B+β3\*A2+β4\*B2+β5\*A\*B.

We see the standard deviation of the residuals about the second order model is $ 50.03, which we note is not significantly lower than the standard deviation of the y scores about the first order model. R2= 0.34. For testing H0: β1=β2=β3=β4=β5=0, (since ~~AB~~C holds), we find the partial F is obtained by [(64767-45056)/(22-18)]/ 2503= 1.97. We reject this value if it is too large. Note since the cutoff point with α=0.05 on an F5,18 is approximately 2.77. (Reject H0 if Fobs>2.77, accept H0 otherwise.) We accept H0, and thus P>0.05. We conclude the second order model is not significantly better than the zero order model.

**Residual plot for the Highest order model fitted**: Examining the residual plot for the second model (p.4b), we note: the mean of the residuals seems to be zero regardless of the value if the predicted. 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 zero order model ( the current best model w/o DAY):** We now examine the zero order model, to see if we can drop any of variables. E(P)=β0, where β0 is equal to the mean of PROFIT, β0=μ=222.575.

We only look at the zero order terms, i.e. terms not included in higher order terms in the model, and individually test; p.4c (A holds since we are testing a single β=0.)

H0: β0=μ=0 vs H0: not H0; (A holds) [p= 0 <0.05, reject H0]

If we find one or more p values above α=0.05, we drop the one with the highest p-value, and refit without this term. In this case, we drop do not drop any term.

Examining the residual plot for zeroth order model, the mean of the residuals is 0. However, we can’t check the variability of the residuals for A2 since there is only one value.

**Histogram of residuals for the FINAL model w/o DAY:** Examining the histogram of the residuals (p.4c-3), we see that they appear to be more or less symmetric ad unimodal, and we can’t reject that they are normally distributed.

**Independence of the residuals for the FINAL model w/o DAY:** Given that the data has been ordered by the order it was collected, we examine the plot of residual t vs residual t-1 (p.4c-4). There doesn’t appear to be a line through this data which has a very positive or negative slope, so we conclude the residuals are independent. Indeed, if we regress residual t on residual t-1, we get a correlation of r=-0 with a p=0.73>0.05=α and conclude this is not significantly different from zero (p.4b-5). We could also calculate the Durbin-Watson statistics =133438.771/67926.165=1.96, which ranges from 0-4 with values near 2 indicating no correlation, 0 pos. corr., and 4 neg. corr. Therefore, there is enough evidence to prove the independence of the residuals for the final model, at most a very slightly positive correlation.

The **prediction model w/o DAY** is given by PROFIT= 222.575.

The predicted profit for an individual with Material A=35 gallons and Material B=95 lbs on day 26 is 222.575 ($100), obtained by substituting this value in the above prediction equation. A 90% confidence interval for this prediction is given by the above prediction plus or minus tdferror, α/2\*Sβ0, 222.575+/-1.717\*11.08=(203.56,241.59)(in $100).

**Conclusion:** We recommend using the zero order model given above to describe the relationships if do not include DAY variable.

However, since we found a significant relationship between DAY and Material B and PROFIT after examining correlation matrix and scatterplots. We should also consider to include DAY as an independent variable into our model.

**Fit of first order model w DAY**: E(P)=β0+β1\*A+β2\*B+β3\*D (p.2a). Examining the EXCEL results for this model, we find sP,A= $47.71, R2= 0.33. For testing H0: β1=β2=β3=0, (since ~~A~~B holds), we find the p for the Global F= 0.042 < 0.05=α. Thus, the first order model is a significant improvement over the model E(P)= β0.

**Fit of second order model w DAY**: (p.2b) E(P)=β0+β1\*A+β2\*B+β3\*D+β4\*A2+β5\*B2+β6\*A\*B+β7\*A\*D+β8\*B\*D.

We see the standard deviation of the residuals about the second order model is $52.08, which we note is not significantly lower than the standard deviation of the y scores about the first order model. R2= 0.66. For testing H0: β4=β5=β6=β7=β8=0, (since ~~AB~~C holds), we find the partial F is obtained by [(45529-37966)/(20-14)]/ 2712= 0.4648 . We reject this value if it is too large. Note since the cutoff point with α=0.05 on an F9,14 is approximately 2.65. (Reject H0 if Fobs>2.65, accept H0 otherwise.) We accept H0, and thus P>0.05. We conclude the second order model is not significantly better than the first order model.

**Residual plot for the Highest order model fitted w DAY**: Examining the residual plot for the second model (p.2b), we note: the mean of the residuals seems to be zero regardless of the value if the predicted. 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 first order model w DAY:** We now examine the first order model, to see if we can drop any of 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.2a (A holds since we are testing a single β=0.)

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

H0: β2=0 vs H0: not H0; (A holds) [p= 0.79>0.05, accept H0]

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

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

**Fit of reduced first order of form w DAY:** E(P)=β0+β1\*B+β2\*D (p.3a). Examining the EXCEL results for this model, we find sP,A= $ 46.67, R2= 0.33.

We only look at the higher order terms and individually test

H0: β2=0 vs H0: not H0; (A holds) [p= 0.02<0.05, reject H0]

H0: β1=0 vs H0: not H0; (A holds) [p= 0.71>0.05, accept H0]

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

**Fit of reduced first order of form w DAY:** E(P)=β0+β1\*D (p.3b). Examining the EXCEL results for this model, we find sP,A= $ 45.75, R2= 0.32. We only look at the higher order terms and individually test

H0: β1=0 vs H0: not H0; (A holds) [p= 0.004<0.05, reject H0]

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

**Residual plot for the FINAL model w DAY:** Examining the residual plot for this model (3b-2), 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 model w DAY:** Examining the histogram of the residuals (p.3b-3), we see that they appear to be more or less symmetric ad unimodal, and we can’t reject that they are normally distributed.

**Independence of the residuals for the FINAL model w DAY:** Given that the data has been ordered by the order it was collected, we examine the plot of residual t vs residual t-1 (p.3b-4). a) It only appears to be a line through this data with a slight negative slope. b) Indeed, if we regress residual t on residual t-1 to fit E(et)= β0+β1et-1, we get a correlation of r=-0.43 with a p=0.044<0.05= α. Therefore, we reject Ho: β1=0, and conclude this is significantly different from zero, and β1 <0 (p.3b-5). c) calculate the Durbin-Watson statistics =130957.706/46039.91=2.84, which ranges from 0-4 with values near 2 indicating no correlation, 0 pos. corr., and 4 neg. corr. Therefore, there is not enough evidence to prove the independence of the residuals for the final model. There appears a possible negative correlation of the residuals. Given a,b,c we conclude that it is very possible that the A4 violated and there exists negative corr. residuals.

The **prediction model w DAY** is given by PROFIT= 169.39+ 4.1038\* DAY.

The predicted profit for an individual with Material A=35 gallons and Material B=95 lbs on day 26 is 275.99 ($100), obtained by substituting this value in the above prediction equation. A 90% confidence interval for this prediction is given by the above prediction plus or minus tdferror, α/2\*SP,A\*sqrt(1+(1/n)+0.1309)), 275.99 +/-1.717\*45.75\*1.0829=(190.93,361.05)(in $100).

**Conclusion:** We recommend using the first order model given above to describe the relationships if include DAY variable.