**ANALYSIS**

Problem: Data for a random sample of (bonus) income measurements on 18 individuals. Our object is to characterize the dependent variable INCOME (I) (in $100). Along with the value of Income for 18 individuals, we have been provided with the value of two independent variables: age of the individual AGE (yrs), and the order that the data was collected over one week and is listed in ORDER. Besides characterizing the relationship between these three variables, we wish to predict INCOME for an individual with AGE=35 or AGE=50, and give a 95% 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.

INCOME: (in $100), the dependent variable, has an average of 19736, standard deviation of 9.94, ranges from a minimum value of 17670 to a maximum of 21130. The shape of the distribution appears unimodal, skewed to the left. (p.1c)

ORDER: an independent variable, has an average of 9.5, standard deviation of 5.34, ranges from a minimum value of 1 to a maximum of 18. The shape of the distribution appears uniform. (p.1c)

AGE: an independent variable, has an average of 40 years, standard deviation of 13.28 years, ranges from a minimum value of 20 years to a maximum of 60 years. The shape of the distribution appears uniform. (p.1c)

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

INCOME vs ORDER: r=-0.14, about 2% of the variability in the y scores around y-bar is explained by simple regression between INCOME and ORDER. The std. dev. of the y scores about a simple linear regression using ORDER is approximately 0.99 times the std. dev. of the y scores about y-bar. As ORDER increases, y-hat decreases in this fitted simple regression model.

INCOME vs AGE: r=0.48, about 23% of the variability in the y scores around y-bar is explained by simple regression between INCOME and AGE. The std. dev. of the y scores about a simple linear regression using AGE is approximately 0.88 times the std. dev. of the y scores about y-bar. As AGE 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 ORDER vs AGE, r=-0.20, about 4% of the variability in the y scores around y-bar is explained by simple regression between ORDER and AGE. The std. dev. of the y scores about a simple linear regression using AGE is approximately 0.98 times the std. dev. of the y scores about y-bar. As AGE increases, y-hat decreases in this fitted simple regression model.

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

INCOME vs AGE: There appears to be a positive linear relationship between these two variables.

INCOME vs ORDER: There appears to be a slightly negative linear relationship between these two variables.

**Scatterplots between pairs of independent variables:**

AGE vs ORDER: Describe the relationship: There appears to be a slightly negative linear relationship between these two variables.

No significant relationship found between Order and any of the other variables after examining correlation matrix and scatterplots.

**Fit of first order model**: E(I)=β0+β1\*A (p.2a). Examining the EXCEL results for this model, we find sI,A= $8.96, R2= 0.23. For testing H0: β1=0, (since ~~A~~B holds), we find the p for the Global F= 0.13>0.05=α. Thus, the first order model is a significant improvement over the model E(I)= β0.

**Fit of second order model**: (p.2b) E(I)=β0+β1\*A+β2\*A2.

We see the standard deviation of the residuals about the second order model is $9.22, which we note is not significantly lower than the standard deviation of the y scores about the first order model. R2= 0.24. For testing H0: β2=0, (since ~~AB~~C holds), we find the p for the partial F is obtained by [(1284-1276)/(16-15)]/85= 0.094 . We reject this value if it is too large. Note since the cutoff point with α=0.05 on an F2,15 is approximately 3.68. (Reject H0 if Fobs>3.68, 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.

**Fit of third order model**: (p.2c) E(I)=β0+β1\*A+β2\*A2+β3\*A3.

We see the standard deviation of the residuals about the third order model is $6, which we note is slightly significantly lower than the standard deviation of the y scores about the first order model. R2= 0.70. For testing H0: β3=0, (since ~~AB~~C holds), we find the p for the partial F is obtained by [(1284-504)/(16-14)]/36= 10.83 . We reject this value if it is too large. Note since the cutoff point with α=0.05 on an F3,14 is approximately 3.34. (Reject H0 if Fobs>3.34, accept H0 otherwise.) We reject H0, and thus P<0.05. We conclude the third order model is significantly better than the first order model.

**Fit of fourth order model**: (p.2d) E(I)=β0+β1\*A+β2\*A2+β3\*A3+β4\*A4.

We see the standard deviation of the residuals about the third order model is $5.93, which we note is not significantly lower than the standard deviation of the y scores about the third order model. R2= 0.73. For testing H0: β4=0, (since ~~AB~~C holds), we find the p for the partial F is obtained by [(504-457)/(14-13)]/ 35= 1.34 . We reject this value if it is too large. Note since the cutoff point with α=0.05 on an F4,13 is approximately 3.18. (Reject H0 if Fobs>3.18, accept H0 otherwise.) We accept H0, and thus P>0.05. We conclude the fourth order model is not significantly better than the third order model.

**Residual plot for the Highest order model fitted**: Examining the residual plot for the fourth model (p.2d), 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 third order model:** We now examine the third 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.2c (A holds since we are testing a single β=0.)

H0: β3=0 vs H0: not H0; (A holds) [p= 0.0003<0.05, reject H0]

H0: β2=0 vs H0: not H0; (A holds) [p= 0.0004<0.05, reject H0]

H0: β1=0 vs H0: not H0; (A holds) [p= 0.0004<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 do not drop any term.

**Residual plot for the FINAL model:** Examining the residual plot for this model (2c-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:** Examining the histogram of the residuals (p.2c-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:** 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.2c-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=-0.042 with a p=0.88 and conclude this is not significantly different from zero (p.2c-5). We could also calculate the Durbin-Watson statistics =872.45/457.42=1.91, which ranges from 0-4 with values near 2 indicating no correlation, 0 pos. corr., and 4 neg. corr.

The **prediction model** is given by INCOME= -41.25+ 19.50\*AGE -0.5043\* AGE2+ 0.004163\* AGE3.

The predicted income for individuals who are 35 years old is 203.07 ($100), obtained by substituting this value in the above prediction equation. A 95% confidence interval for this prediction is given by the above prediction plus or minus tdferror, α/2\*SI,A\*sqrt(1+(1/n)), 203.07+/-2.145\*6.00\*1.027=(189.85,216.29)(in $100).

The predicted income for individuals who are 50 years old is 193.38 ($100), obtained by substituting this value in the above prediction equation. A 95% confidence interval for this prediction is given by the above prediction plus or minus tdferror, α/2\*SI,A\*sqrt(1+(1/n)), 193.38 +/-2.145\*6.00\*1.027=(180.16,206.60)(in $100).

**Conclusion:** We recommend using the third order model given above to describe the relationships.