1. Linerboard and regression tests

My fitted model was:

$$\log(product) = -1.762 + .126\log(rawmat) + .543\log(energy) + .286\log(deprec) + .34\log(labor)$$

$$product = \exp(-1.762)(rawmat)^{0.126}(energy)^{0.543}(deprec)^{0.286}(labor)^{0.34}$$

Variable	DF	Estimate	Std Error	t Value	Pr > |t|
Intercept	1	-1.76206	0.79507	-2.22	0.0384
lograwmat	1	0.12634	0.04316	2.93	0.0083
logenergy	1	0.54267	0.07886	6.88	<.0001
logdeprec	1	0.28626	0.06516	4.39	0.0003
loglabor	1	0.33994	0.14459	2.35	0.0291

(a) This is the F-test of the entire regression model:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	4.83013	1.20753	647.29	<.0001
Error	20	0.03731	0.00187		
Corrected Total	24	4.86744			

The test statistic is F=647, with p < 0.0001. There is very strong evidence that at least one regression slope differs from zero.

* Again, remember the overall test does not tell you that all slopes are non-zero, just at least one.

(b) Test $H_0 : \beta_1 + \beta_2 + \beta_3 + \beta_4 = 1$

You can test this hypothesis three ways. Each gives you the same p-value and conclusion, but the test statistics may not be the same. P < 0.017. There is evidence that the sum of the coefficients is not 1. In econometric terms, there is not constant return to scale.

1) Estimate $\beta_1 + \beta_2 + \beta_3 + \beta_4$ and its standard error, then construct a t-test:

T = (estimate of sum - 1)/se = (1.295-1)/0.114 = 2.59.

You can get the estimate and s.e. from proc glm; … estimate rawmat 1 energy 1 deprec 1 labor 1;

Parameter	Estimate	Error	t Value	Pr > |t|
sum1	1.29521599	0.11377588	11.38	<.0001

* The t-test provided by SAS tests the wrong hypothesis (sum = 0).

2) Use a test statement in proc reg to construct the F statistic for a linear constraint.

Source	DF	Mean Sq	F Value	Pr > F
Numerator	1	0.01256	6.73	0.0173
Denominator	20	0.00187		

3) Construct a reduced model corresponding to the null hypothesis.

Full model: Model in (a)

Reduced model: $\log(Y) = \beta_0 + \beta_1 \log X_1 + \beta_2 \log X_2 + \beta_3 \log X_3 + (1 - \beta_1 - \beta_2 - \beta_3)\log X_4 + \varepsilon$

Or: $\log(Y) - \log(X_4) = \beta_0 + \beta_1[\log X_1 - \log X_4] + \beta_2[\log X_2 - \log X_4] + \beta_3[\log X_3 - \log X_4] + \varepsilon$

(c) You need the s.e. of the sum, which you can get from the estimate (above) command or by computing Var b = MSE $(X'X)^{-1}$ then Var Cb = C Var b C' where C = [0 1 1 1 1], then taking the square root. This leads to: $1.295 \pm t_{20,0.975}(0.114) = (1.058, 1.532)$

Parameter	Estimate	Error	t Value	Pr > |t|
sum1	1.29521599	0.11377588	11.38	<.0001

(d) $s.e. = \sqrt{Var(\beta_1 + \beta_2 + \beta_3 + \beta_4)} = \sqrt{\sum_{i=1}^{4} s.e._{\beta_i}^2 + 2 \sum_{1 \le i < j \le 4} \mathrm{cov}(\beta_i, \beta_j)}$

If you just add up the sum of the variances, you are ignoring the covariances.

2) added variable plots. My SAS code for all parts:

```
data avp;
  infile 'avp.txt' firstobs=2;
  input x1 x2 x3 y;
  x1q = x1*x1;
  x2q = x2*x2;
  x3q = x3*x3;
  x12 = x1*x2;
  x13 = x1*x3;
  x23 = x2*x3;
run;
proc reg;
  model y = x1 x2 x3;
  output out=resids r=resid p=yhat;
  run;
proc plot;
  plot resid*(x1 x2 x3);
  title '2a';
  run;
proc reg data=avp;
  model y = x1 x2 x3 x1q x2q x3q x12 x13 x23;
lof: test x1q=0, x2q=0, x3q=0, x12=0, x13=0, x23=0;
  title '2b';
  run;
proc loess data= avp;
  model y = x1 x2 x3 /smooth=0.6 dfmethod=exact;
  title 'data for 2c';
  run;
proc reg data=avp;
  model y = x1 x2 x3 /partial;
  title '2d';
run;
```

a) No sign of lack of fit: all plots are a flat band of points.

b) Two ways to construct the lack of fit test: by using model comparison to compare the linear-only model to the full quadratic model, or by testing that all 6 quadratic and crossproduct coefficients are 0 in the full quadratic model. Both give the same F statistic: F = 4.70 with p < 0.0003. Strong evidence of lack of fit.

c) This can only be done by model comparison.

Model	error SS	error d.f.	MS	Fstat	Pvalue
Linear	43,206	96			
Loess	37,885	88.2	429.2343		
change	5,321	7.7	687.6285	1.60	0.14

The F statistic for lack-of-fit is 687.6285/429.2343=1.60. This compared to a F 7.7, 88.2 distribution if you have a computer, or a F 8, 90 or F 7, 60 distribution. Using 7,60 is rounding both d.f. down to the nearest tabulated values. This is conservative. The details of which tabled value to use are not important; anything reasonable is acceptable. The p-value is 0.14. There is no evidence of lack of fit.

Note: If you used R for these computations, you got a different answer.

```
avp.lo <- loess(y~x1+x2+x3, span = 0.6, data=avp)
avp.lo1 <- loess(y~x1+x2+x3, data=avp, span=100, degree=1)
anova(avp.lo1, avp.lo)
```

She got the following output:

```
# Model 1: loess(formula = y ~ x1 + x2 + x3, data = avp, span = 100,
degree
= 1)
# Model 2: loess(formula = y ~ x1 + x2 + x3, data = avp, span = 0.6)

# Analysis of Variance:   denominator df 75.56

#       ENP   RSS F-value Pr(>F)
# [1,]    4 43201
# [2,]   19 32715       1 0.5293
```

This is because of different default values (SAS default degree=1, R default = 2). If you used R, your answer was probably marked as wrong, even though it was correct according to R. It you got this answer, bring your HW to Chuanlong to be regraded.

d) The answer is obvious if you calculated the plots correctly: lack of fit in X1. The dependence of Y on X1 is oscillatory. The dependence on X2 is reasonably linear; the dependence on X3 is reasonably linear.

3. Website development. My SAS code:

```
data website;
  infile 'c:/philip/stat500/data/website.txt' firstobs=2;
  input id deliver backlog team experience process year quarter;
  i = _n_;

proc reg;
  model deliver = backlog experience process year / vif influence;
  output out=resids r=resid p=yhat rstudent = rs
                    cookd = cookd dffits = dffit h = h;
  run;

proc gplot;
  plot rs*yhat;
  plot (cookd dffit h)*i;
  run;

/* proc standard standardizes each variable to the specified mean and var */
/*   you could also multiply each coefficient from the original regression by */
/*   sd(X)/sd(Y).  */
proc standard data=website mean=0 std=1 out= webstand;
  var deliver backlog experience process year;
run;

proc reg;
  model deliver = backlog experience process year;
  run;
```

a) The best answer: no problems with regression outliers.

- You should be looking at residuals or externally studentized residuals. Plotting either against the predicted values or the id number indicates some large values, but no values far from the general pattern.
- The error ms has 73 – 5 = 68 d.f. The advantage of the externally studentized residuals is that they have a t distribution distribution, if the model is correct. Here, there are 4 observations with externally studentized residuals larger than 1.995 (or < -1.995), the 0.975 quantile of a t distribution with 68 d.f. These may be worth quickly investigating, especially if one is very large (or small), but you expect 73*0.05 = 3.65 points to fall outside the 0.025 and 0.975 t quantiles, even if there are no outliers.
- If you were really concerned about falsely identifying an outlier, you would use a multiple comparisons adjustment. A Bonferroni adjustment is appropriate here (multiple tests). The 1-0.025/73 quantile is 3.56. None of the residuals are more extreme than that.
- Since this whole approach is exploratory, I don't get too concerned about the details of p-values and whether or not to adjust for multiple comparisons. The argument against always using a mc adjustment is that you lose power.

b) Yes, id 65 has a large DFFITS value (larger than 1), indicating that id 65 has a large influence on its own fitted value. The 0.2 quantile of the F 5,68 distribution is 0.466. Values of Cook's D less than this are no concern. There are no points with a Cook's D larger than 0.466.

- I told you to consider this a small-medium data set. If it were large, you would use a smaller critical value of DFFITS to identify influential points.

- If you told me about points with high leverage (h_{ii}) , that was fine but not necessary. Leverage is potential influence. The question asked about points with large influence.

c) Yes, id 65 has a large influence on the slope for backlog because it has a DFBETA > 1. It is the only point with a DFBETA < -1 or > 1.

d) No, the VIF values for each variable (2.98, 3.29, 2.48, and 3.05) are all considerably less than 10.

e). There is no sign of an unusual regression residual. The variability does seem to increase slightly with the mean, but the s.d.'s in the two groups are quite similar. I would probably assume equal variances. If I did transform, I would make very sure that the transformed values had no problems. The log transform is probably too strong here.

- The residual plot has two clusters of observations. These correspond to the two PROCESS groups. We'll look more carefully at this next week.

f) The standardized regression slopes are 0.098 for backlog, -0.10 for experience, 0.64 for process, and 0.099 for year. Relative to a 1 s.d. change in X, process had the biggest change in productivity.

g) The predicted change in delivery for a 12 month change in backlog is 12*0.0877 = 1.05 , for a 9 month increase in experience is 9*(-0.128) = -1.15, for the change in process = 9.38 and for a change of 1 year = 1.39. Using this 'relevant change in X' scale, process has the largest effect on productivity