Family Name:_____

Given Name:_____

Student Number:_____

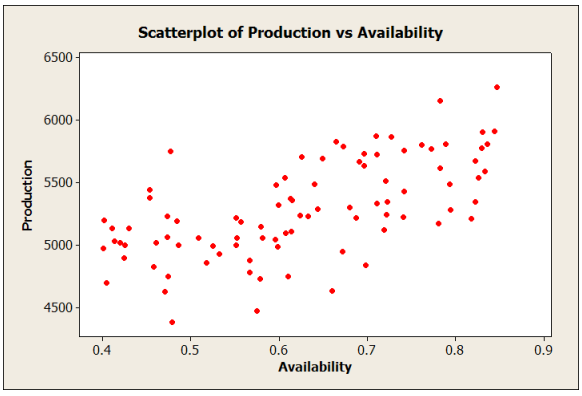
MIE237S Term Test
Examination Type B; Calculator Type 2 Permitted
March 19, 2013, 11:10 A.M.
50 minutes; 20 Marks Available

You should have two booklets. This booklet has 5 pages and consists of the questions and space for you to answer each question. You should have ample space to answer. You may use the backs of pages for extra rough work space. Please do not detach any pages from this booklet.

The second booklet consists of 2 pages printed on both sides. The first side is the aid sheet and the other sides consist of tables of probabilities for the standard normal and t distributions. This booklet is yours to keep.

1.**(16 marks total)** One way to measure the performance of a fleet of haul trucks in an open pit mine is by "availability". The availability of a fleet is (roughly speaking) the proportion of the amount of time the trucks are not being repaired and are actually able to operate.

A reliability engineer at a mine is going to analyze the relationship between haul truck availability and overall mine production in tonnes of ore produced. Here is a plot of $n = 90$ days' worth of data:



The usual simple linear regression model is fit to the data. In case you need it, $S_{xx} = 1.545$. Here is the Minitab output with many entries missing:

```
The regression equation is
Production = 4119 + 1866 Availability
```

```
Predictor       Coef  SE Coef      T      P
Constant      4119.2    154.7  26.63  0.000
Availability  1866.2    241.9   7.72  0.000


S = 300.665   R-Sq = 40.4%   R-Sq(adj) = 39.7%


Analysis of Variance

Source          DF        SS       MS      F      P
Regression       1   5381961  5381961  59.54  0.000
Residual Error  88   7955163    90400
Total           89  13337124


Unusual Observations

Obs  Availability  Production     Fit  SE Fit  Residual  St Resid
  4         0.477      5748.0  5008.7    48.0     739.3      2.49R
  7         0.660      4633.6  5351.2    32.8    -717.6     -2.40R
 75         0.575      4469.8  5191.6    34.0    -721.8     -2.42R
 89         0.478      4380.6  5011.5    47.8    -630.9     -2.13R

R denotes an observation with a large standardized residual.


Predicted Values for New Observations

New
Obs     Fit  SE Fit       95% CI            95% PI
  1  5287.5    31.7  (5224.5, 5350.5)  (4686.7, 5888.3)


Values of Predictors for New Observations

New
Obs  Availability
  1         0.626

Predicted Values for New Observations

New
Obs     Fit  SE Fit       95% CI            95% PI
  1  5518.9    43.7  (5432.1, 5605.6)  (4915.1, 6122.7)


Values of Predictors for New Observations

New
```

```
Obs  Availability
  1         0.750
```

(a) **(4 marks)** Provide the missing entries for the Analysis of Variance part of the Minitab output.

See above. One strategy would be to enter the degrees of freedom, then use the square of S to get the MSE, then get SSE, then use the R^2 to get SST and SSR etc. Note that your answers are likely to differ from the above in the less significant digits due to rounding error.

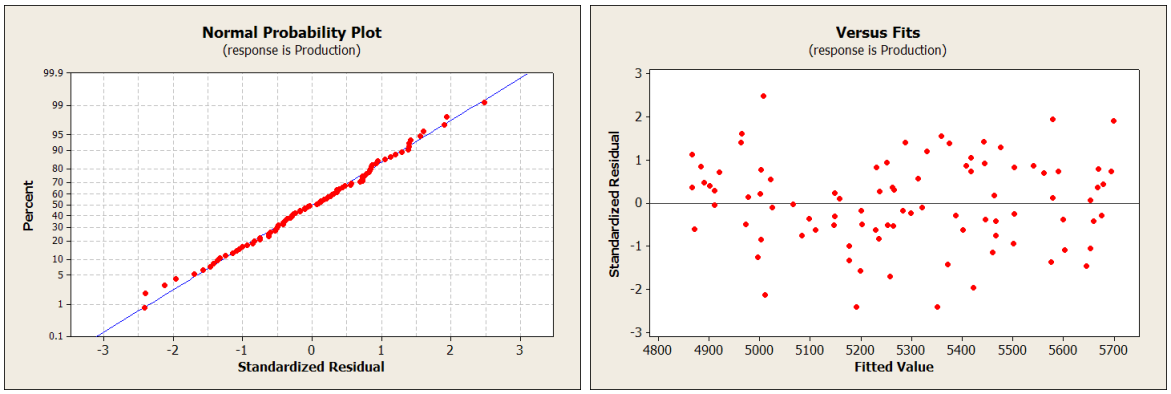
(b) **(3 marks)** Provide a 99% confidence interval for the slope of the regression line.

You need to figure out the estimated slope, probably via $T^2 = F$ etc. Or there might be other ways. Once you have that the answer is

$$1866.2 \pm t_{88,0.005} \cdot 241.9$$

again with possible rounding error will carry over. And people will choose either $t_{80,0.005} = 2.639$ or $t_{100,0.005} = 2.626$ or even 2.576 from the normal table, any of which is fine.

(c) **(3 marks)** Here is a normal probability plot of the standardized residuals and a plot of the standardized residuals versus the fitted values. Comment on whether the usual assumptions of the simple regression model have been satisfied or not.



The plots look great to me. Some students might imagine a pattern in the residuals vs. fitted values plot, but it really isn't anything there. There just aren't many observations around $x = 0.525$ or so. Don't remove marks for imagining a pattern here.

(d) **(2 marks)** Comment on the possible existence of any outliers or influential points in this dataset.

Minitab identifies a few candidates. Give 0.5 for pointing this out. But the plots reveal none of them to be really outlying. It's OK if they say the one with the largest residual is an outlier.

(e) **(2 marks)** The sample average availability over the 90 days was $\overline{x} = 0.626$. Produce a 95% confidence interval for the mean response at 0.626.

See Minitab output above. Again there will be rounding error carryover.

(f) **(2 marks)** The sample average availability over the 90 days was $\overline{x} = 0.626$. The mining company is going to spend thirty million dollars to upgrade the maintenance facility to try to tincrease the average availability to 0.75. Provide a range of plausible values for daily production of ore in tonnes at this new average availability of 0.75, with 95% confidence.

Answer is above. If the students found a problem with the model assumptions, however, they should have declined to answer this question. If they did anyway, remove 1 mark.

2.**(4 marks total)** Some data $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ is analyzed using the usual simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ with slope estimator $\hat{\beta}_1 = S_{xy}/S_{xx}$.

Suppose a new variable w_i is introduced that is just a linear transformation of the y_i variable. In other words, $w_i = c + dy_i$ for each i (assume $d \neq 0$.) Consider the new simple linear regression model $w_i = \beta_0^{(w)} + \beta_1^{(w)} x_i + \varepsilon_i$.

(a) **(2 marks)** Show that the new slope estimator $\hat{\beta}_1^{(w)}$ is equal to d multiplied by the old slope estimator $\hat{\beta}_1$.

$$\begin{aligned} S_{xw} &= \sum (x_i - \overline{x})(w_i - \overline{w}) \\ &= \sum (x_i - \overline{x})(c + dy_i - (c + d\overline{y}) \\ &= \sum (x_i - \overline{x})(dy_i - d\overline{y}) \\ &= d \sum (x_i - \overline{x})(y_i - \overline{y}) \\ &= dS_{xy} \end{aligned}$$

And the result follows.

Give 0.5 marks for the incorrect (in fact circular) argument involving $\hat{w}_i = c + d\hat{y}_i$ and similar.

(b) **(2 marks)** Show why the p-value obtained for the hypothesis test $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$ will be exactly the same as the p-value for the hypothesis test $H_0 : \beta_1^{(w)} = 0$ versus $H_1 : \beta_1^{(w)} \neq 0$.

The mean square error for $\hat{\beta}_1^{(w)}$ has numerator:

$$
\begin{aligned}
&= \sum (w_i - \hat{w}_i)^2 \\
&= \sum (w_i - (\hat{\beta}_0^{(w)} + \hat{\beta}_1^{(w)} x_i))^2 \\
&= \sum (w_i - (\overline{w} - \hat{\beta}_1^{(w)} \overline{x} + \hat{\beta}_1^{(w)} x_i))^2 \\
&= \sum (c + d y_i - (c + d\overline{y} - d\hat{\beta}_1 \overline{x} + d\hat{\beta}_1 x_i))^2 \\
&= d \sum (y_i - (\overline{y} - \hat{\beta}_1 \overline{x} + \hat{\beta}_1 x_i))^2 \\
&= d \sum (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 \\
&= d \sum (y_i - \hat{y}_i)^2
\end{aligned}
$$

which is d times the mean square error for $\hat{\beta}_1$. So the test statistics for both tests will be identical (the ds cancel) and they have the same distribution, so they will result in the same p-value.

Give full marks for a written explanation that captures the correct reasoning.