MIE237S Term Test; February 11, 2014

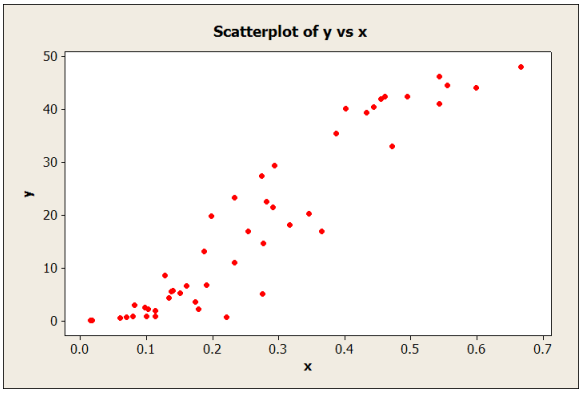
General Marking Notes: Try to remove marks only once for any error unless the result is also completely unreasonable and they don't notice the obvious problem.

There is a large total number of marks available. The allocation of marks to questions is in proportion to what I want them to be, so a straightforward question could be worth 4 marks, say, which means one error could result in 2/4 or something like that.

1.**(20 marks total)** A paper was published in 2002 in the journal *Metallurgical and Materials Transactions B* that described some relationships among many characteristics of chromium oxides in a steel-making process.

The details of the process are not important, except to note that all of the measured variables are positive.

In this question we will focus on the two variables called "activity coefficient" (y) and "reciprocal of amount" (x). Here is a plot of the data, which consists of $n = 48$ records:



We will analyze the data using the usual simple linear regression model as specified on the aid sheet.

The sample average of the x values is $\overline{x} = 0.266$ and $S_{xx} = 1.35$.

Here is the Minitab output with some of the entries replaced with * symbols:

```
The regression equation is
y = ****** + **** x

Predictor     Coef  SE Coef      T      P
Constant    -6.600    1.538  -4.29  0.000
x           92.478   ******  *****  *****
```

```
S = 5.64328   R-Sq = 88.5%

Analysis of Variance

Source          DF     SS     MS       F      P
Regression      **  *****  *****  ******  *****
Residual Error  **  *****  *****
Total           **  12775

Unusual Observations

Obs      x       y     Fit  SE Fit  Residual  St Resid
  7  0.221   0.700  13.860   0.843   -13.160     -2.36R
 29  0.276   5.100  18.947   0.816   -13.847     -2.48R
 48  0.667  48.000  55.052   2.129    -7.052     -1.35 X

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large leverage.
```

(a) **(4 marks)** Provide the missing entries for the Analysis of Variance part of the Minitab output, but do not attempt to compute the p-value.

Use either $R^2 \cdot \text{SST}$ to complete the SS decomposition, or use S^2 as MSE.

Source	DF	SS	MS	F
Regression	1	11310	11310	355.13
Residual Error	46	1465	32	
Total	47	12775		

Marking notes: There are three tasks here: the DF, the SS, and then putting it all together. Give 1 for the DF, 2 for the SS, and 1 for finishing the table, removing marks only once for each error as long as the result is reasonable

(b) **(4 marks)** Hypothesis test $H_0 : \beta_1 = 0$ versus $\beta_1 \neq 0$ using a p-value....

Observed value of the test statistic is $T = \sqrt{F} = 18.84$. When compared to a t distribution with 46 degrees of freedom on the table, the p-value must be less than $2 \cdot 0.0005 = 0.001$. There is very strong evidence against the null hypothesis.

It is also possible to compute T using the paramater estimate along with the MSE and the given S_{xx}.

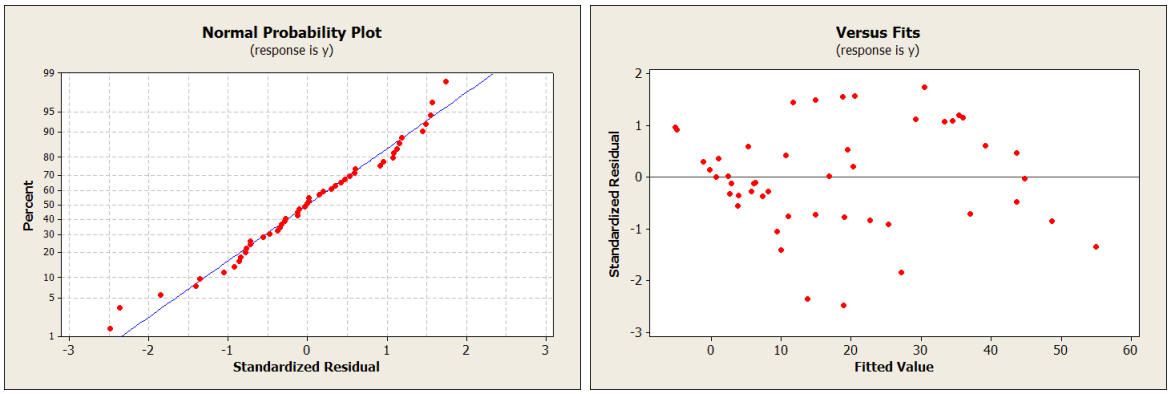
Marking notes: 2 marks for T, 1 for using the correct distribution, and 1 for reporting the p-value. I'm not concerned with how they word the conclusion as long as they have used reasonable judgment. If they messed up (a), the answer here could be wrong, but only remove marks if the conclusion is badly affected, such as they conclude that the slope is not different from 0 when it obviously is from the plot.

(c) **(4 marks)** Comment on the possible existence of unusual observations in the data.

Minitab suggests two possible outliers and one possible influential observation. The outliers are not obviously far from the rest of the data. The possibly influential observation doesn't seem to have a large effect on the parameter estimates, as it is consistent with the model suggested by the rest of the data.

Marking notes: If they note the points Minitab suggests, they can have 1 point. The other three are for pointing out there is nothing really unusual about the points.

(d) **(4 marks)** Here are a normal probability plot of the standardized residuals and a plot of the standardized residuals versus the fitted values:



Comment on the extent to which the usual lineal regression model assumptions have been satisfied.

The normal plot is close to a straight line, so the normality assumption is satisfied. There are clear problems suggested by the other plot. The equal variance assumption is not satisfied. There is very little variation near fitted values close to 0, corresponding to x values close to 0. Not only that, but the regression line goes below zero at $x = 0.071$, which clearly isn't good since both x and y are supposed to be positive, so the linear model is not a good fit.

Two points for the normality assuption and two for the equal variance. It isn't necessary to point out that the line goes below zero.

(e) **(4 marks)** Predict the value of the response y at an x value of 0.05. If possible, produce a 95% prediction interval for a new response at this x value of 0.05. If not, state why not.

It is not possible due to the model assumption violation, and the situation is made even more obviously problematic given that the predicted value is less than 0, which is not even an allowed value for y.

Max 2 points if for some reason they thought the model assumptions were OK, because the resulting interval would be mostly under 0 and therefore obviously unreasonable.

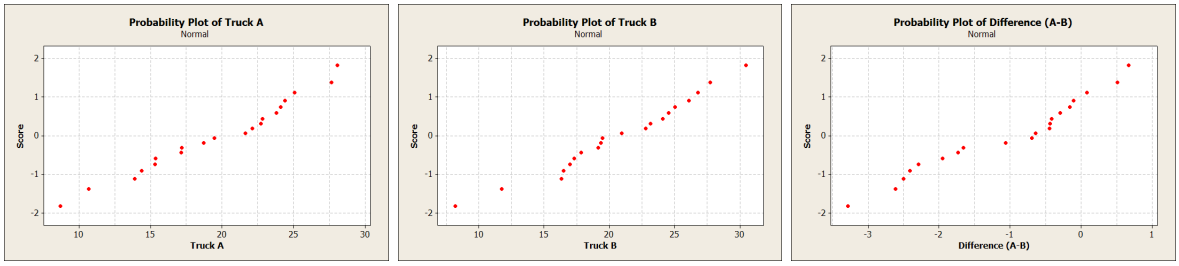
2. **(15 marks)** A mining company wants to evaluate the productivity of two brands (Brand A and Brand B) of haul truck in an open pit mining environment. They own one truck of each brand (Truck A and Truck B). Each truck performs the same task: moves raw material from the mine site to the processing area. Productivity is measured by the amount of raw material the truck is able to move in a given time period.

The company plans to operate both trucks in the same mine for the same 20 days. Each day they will measure the total amount of raw material moved by each truck (in hundreds of thousands of kilograms).

Here is an excerpt of the raw data they collected, just to show you how it was organized. Most rows of the raw data were omitted to save space. The last two rows contain the sample averages and sample standard deviations of each column.

Day	Truck A	Truck B	Difference (A-B)
1	15.335	16.989	-1.655
2	28.082	30.493	-2.412
3	27.653	27.755	-0.102
⋮	⋮	⋮	⋮
19	19.462	19.381	0.082
20	8.716	8.199	0.517
Average	19.67	20.74	-1.07
SD	5.45	5.52	1.15

Here are normal quantile plots of the data from the three columns:



Use the following page to perform the hypothesis test that answers the question "*Is there a difference in the average productivity between Truck A and Truck B?*". Include the following:

- perform the hypothesis test using a p-value in your conclusion;
- comment on whether or not the model assumptions you used in your calculations have been satisfied, and if they have not been satisfied, whether any violations cast doubt on the validity of your conclusions.

The data are collected as pairs, since the two trucks are each being operated in the same mine on the same days. So the analysis should be done on the single sample of differences.

So we are testing $H_0 : \mu_d = \mu_A - \mu_B = 0$ versus $H_1 : \mu_d \neq 0$. The test statistic is:

$$T = \frac{\overline{Y}_d - \mu_d}{S_d/\sqrt{n}} \sim t_{19}$$

The observed value of T under the null hypothesis is $-1.07/(1.15/\sqrt{20}) = 4.16$. The p-value is less than $2 \cdot 0.0005 = 0.001$. There is strong evidence against the null hypothesis.

The normal quantile plot of the differences is nearly a straight line, so the assumption of normality is satisfied.

(Note: the incorrect analysis using two independent samples would proceed as follows: $s_p = 5.4851$, observed value of test statistic is $T_{\text{obs}} = -0.62$, compared with the t_{38} distribution the p-value is betten 0.5 and 0.6. No evidence against the null hypothesis. Both groups look normal and the observed SD ratio is much less than 3 (close to 1) so the equal variance assumption is OK.)

Marking notes: 5 points for correctly identifying the paired design (which didn't have to be stated directly but can be inferred from the analysis. 6 points for the calculations. 4 points for verifying the model assumption. Max 2/4 for the model assumptions if they wrongly mention some sort of equal variance assumption in the paired case.

3.**(5 marks total)** Linear regression models are used to model bivariate data, i.e. data collected as $\{(y_1, x_1), (y_2, x_2), \ldots (y_n, x_n)\}$. A substantial part of this course so far has been spent on the details of such an analysis.

It turns out that not all bivariate data can be analyzed in the way we've learned, and not in the sense that the linear model is a poorly fitting model. I mean that you literally cannot do one or more of the computations at all.

In this question you will invent two datasets with $n = 4$ records each, that in some specified way cannot be analyzed as described in class.

(a) **(3 marks)** Invent a dataset with $n = 4$ records in such a way that it is not possible to compute the slope estimator $\hat{\beta}_1$, with a brief explanation of how the computation fails.

You can't compute $\hat{\beta}_1 = S_{xy}/S_{xx}$ if and only if $S_{xx} = 0$, which only happens when the x values are all the same. So any dataset with identical x values will do.

(b) **(2 marks)** Invent a dataset with $n = 4$ records in such a way that it *is* possible to compute the slope estimator $\hat{\beta}_1$ but it is *not* possible to perform the hypothesis test $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$, with a brief explanation of how the computation fails.

We can't do the hypothesis test if the denominator of the test statistic $\sqrt{MSE/S_{xx}}$ equals 0. We need $S_{xx} \neq 0$ in order for $\hat{\beta}_1$ to exist, so we'll need $MSE = 0$ instead. This happens when the data lies on a perfect straight line. The easiest example might be to let the x values be anything at all as long as they aren't the same, and to make the y values all identical, making the data fall on a horizontal line. But any line would do, so a dataset like $\{(1,1),(2,2),(3,3),(4,4)\}$ would also be fine.

Marking notes for this entire question: Depending on how the class performs on this question, consider generous partial credit for nice incorrect attempts that seem to come close to the right idea. But if a large % of the class is answering this question well, be less generous with partial credit.