Family Name:____________________

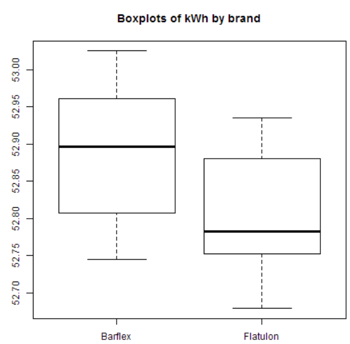
Given Name:____________________

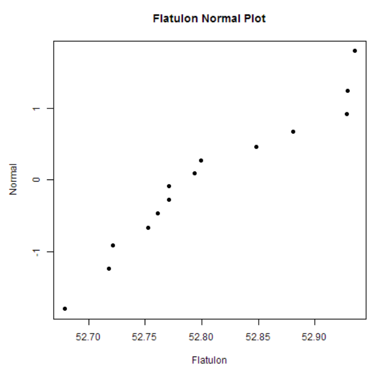
Student Number:____________________

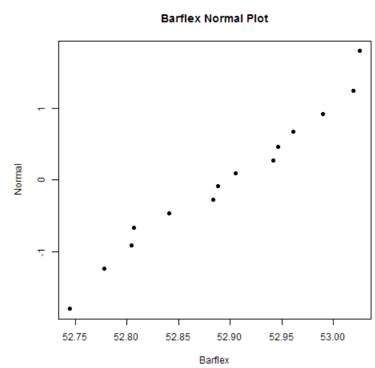
MIE237S Term Test 2012 Solutions

1.**(10 marks total)** Your dog biscuit factory is going to purchase a dehumidifier for its dog biscuit warehouse. Two suppliers, Flatulon (1) and Barflex (2), offer to provide units for you to test for three months. First you decide to measure how much electricity is consumed by each unit. You run the units on alternating days for 28 days, resulting in fourteen daily kWh measurements for each unit.

Here are some plots and some numerical summaries (consisting of observed sample averages and observed sample standard deviations) you can use, if necessary:







Brand	$\overline{y}$	s
Flatulon	52.806	0.08424
Barflex	52.895	0.09016

(a) **(3 marks)** Compute a 99% confidence interval for the difference $\mu_1 - \mu_2$ of the mean kWh usage for the two units.

Two sample t interval.

$$\begin{aligned} s_p &= \sqrt{\frac{(14-1)(0.08424)^2 + (14-1)(0.09016)^2}{14+14-2}} \\ &= 0.0873 \\ t_{26,0.005} &= 2.779 \end{aligned}$$

The interval:

$$(52.806 - 52.895) \pm 0.0873\sqrt{\frac{1}{14} + \frac{1}{14}}$$

or $(-0.1806, 0.0026)$

(b) **(3 marks)** Comment on the validity of the confidence interval in (a) with respect to any relevant model assumptions you would need to make.

The normal quantile plots for the two samples are both roughly straight, so the normality assumption is satisfied. The sample standard deviations are well within the 3:1 ratio, so the equal variance assumption is satisfied. Therefore the confidence interval in (a) is likely accurate.

(c) **(4 marks)** Test $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$ by using an F distribution, producing a p-value as part of your conclusion. (Note: I have not provided tables of F probabilities, but you should be able to obtain accurate bounds on your p-value nevertheless.)

There are a few ways to compute the observed test statstic under $\boldsymbol{H_0}$. The quickest would be to use the computations from (a) and using the fact that $\boldsymbol{T^2 = F}$ as in:

$$\boldsymbol{F = \frac{(52.806 - 52.895)^2}{0.0873^2\left(\frac{1}{14} + \frac{1}{14}\right)} = 7.29}$$

The other way would be to use the formula for $\boldsymbol{F}$ directly, which will end up being virtually the same calculation anyway (and, of course, the same result.)

Since there is only a table of $\boldsymbol{t}$ probabilities, one will have to use the fact that $\boldsymbol{P(F_{1,26} > 7.29) = 2P(t_{26} > 2.70)}$, which will be between 0.01 and 0.02. (The actual value is 0.012.)

The p-value is somewhat small, so there is some evidence against the null hypothesis.

2.**(5 marks total)** The next test of the dehumidifiers will be to measure their effectiveness in removing moisture from the air. The units are run at opposite ends of the same warehouse on 10 consecutive days, as follows. At noon on each day, the starting humidity in the warehouse is measured. Both units are run for one hour, and the humidity is measured again near each unit. The reduction in humidity is recorded as that unit's *score* for the day.

You are interested in whether the amount of moisture reduction is the same on average for each unit or if it is different.

Here is the data along with a row of observed sample averages and a row of observed sample standard deviations, for each column:

Day	Starting	Barflex	Barflex Score	Flatulon	Flatulon Score	Score Difference
1	70.00	65.54	4.46	67.68	2.32	2.14
2	74.00	68.17	5.83	70.75	3.25	2.57
3	77.33	71.76	5.57	72.70	4.64	0.93
4	80.00	75.01	4.99	76.07	3.93	1.06
5	82.00	75.79	6.21	78.13	3.87	2.34
6	83.33	77.77	5.57	79.37	3.97	1.60
7	84.00	78.60	5.40	80.42	3.58	1.82
8	84.00	80.68	3.32	80.89	3.11	0.21
9	83.33	80.30	3.03	81.61	1.72	1.31
10	82.00	78.15	3.85	79.15	2.85	1.00
Average	80.00	75.18	4.82	76.68	3.32	1.50
SD	4.78	5.14	1.10	4.76	0.86	0.73

Perform the appropriate hypothesis test to determine if there is a difference on average between the units. Use a p-value in your conclusion, and state how you would check any relevant model assumptions to verify the validity of the p-value you obtained.

There are 10 independent observations—the daily differences between the humidifiers that come in "pairs". So it is a one-sample t test with observed test statistic under H_0:

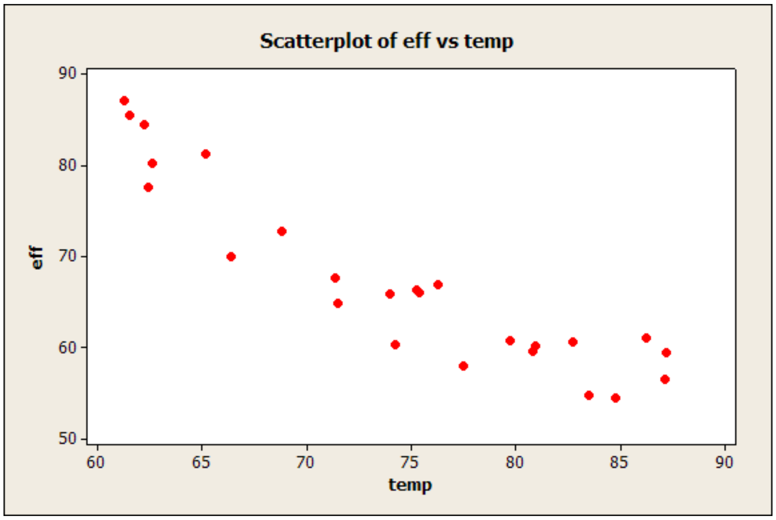
$$\boldsymbol{T = \frac{1.50}{0.73/\sqrt{10}} = 6.50}$$

Comparing 6.50 with a t_9 distribution gives a p-value of less than 0.001. (In fact, it is 0.00011.) This is a very small p-value giving strong evidence of a difference between the means.

There is one assumption—that the 10 differences come from a normal distribution. To check this assumption I would make a normal quantile plot of the 10 daily score differences. Note: making normal plots of the two sets of 10 scores might not work (as the scores might be non-normal while the differences themselves are.)

3.**(15 marks total)** The efficiency of a dehumidifier can depend on the air temperature. The efficiency of the Barfulon humidifier is going to be assessed. For 25 days the average air temperature (x_i) in degrees Celsius in the dog biscuit warehouse is measured, along with an efficiency score out of 100 (y_i) for the Barfulon unit.

Here is a scatterplot of the data:



The "usual" simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon$ with $\varepsilon_i \sim N(0, \sigma^2)$ is fit to the data. The "total sum of squares" is 2410.1 and R^2 is 0.848 (which on Minitab output would be written as 84.8%). Also, $S_{xx} = 1821.24$.

(a) **(4 marks)** Fill in the blanks in the following table. (I left out the p-value column that usually appears in such tables.)

```
Analysis of Variance

Source          DF      SS      MS      F

Regression       1  2043.8  2043.8  128.3

Residual Error  23   366.3    15.9

Total           24  2410.1
```

(b) **(3 marks)** Perform the hypothesis test for $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$, with your conclusion being either to accept or reject the null hypothesis at the $\alpha = 0.01$ level.

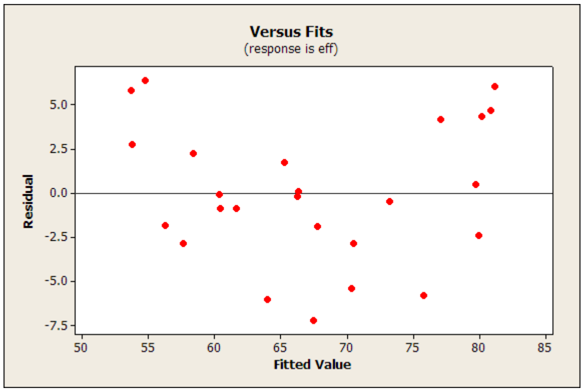
The observed test statistic is $F = 128.3$, which would be compared with an $F_{1,23}$ distibution. To find the 0.01 critical value one can use the fact that $P(F_{1,23} > x^2) = 0.01$ and $2P(t_{23} > x) = 0.01$ have the same solution, which is $x = 2.807$. So the critical value is 7.87. Since 128.3 is way bigger, clearly the conclusion is to reject the null hypothesis.

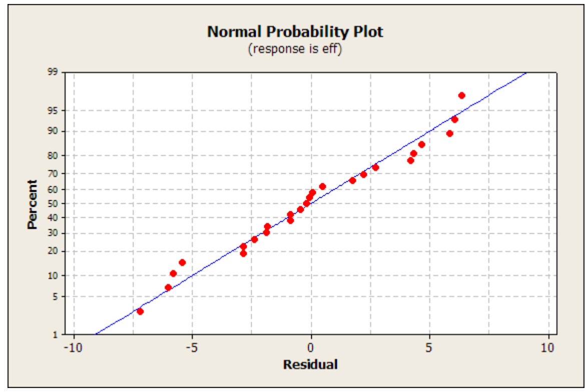
(c) **(3 marks)** Compute a 95% confidence interval for the slope of regression line.

First, figure out $\hat{\beta}_1$. From the plot of the data, the slope is clearly negative. So value of the test statsistic for the hypothesis test in (a) using the t distribution approach is $\hat{\beta}_1/SE(\hat{\beta}_1) = -\sqrt{(128.3)} = -11.33$. Since $SE(\hat{\beta}_1) = \sqrt{MSE/S_{xx}} = \sqrt{15.9/1821.24} = 0.0934$, that gives $\hat{\beta}_1 = -1.06$.

So the confidence interval is: $-1.06 \pm 2.069 \cdot 0.0934$, or $(0.87, 1.25)$

(d) **(5 marks)** Use the following plots to comment on the validity of using the "usual" simple linear regression model on this dataset, paying particular attention to the computations done in parts (b) and (c) of this question on the previous page.





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

On the plot of residuals versus fitted values, the amount of scatter is even, so the equal variance assumption is also satisfied.

However, there is a clear nonlinear pattern on the plot of residuals versus fitted values, so the linear model is not appropriate for this data. (This is subtly evident from the raw data plot as well.)

Both computations in (a) and (b) are made highly questionable by the fact the linear model is not a good fit to the data.

4.**(5 marks total)** Occasionally in class I mentioned the idea of *consistency*, even if I didn't always use that term. Roughly speaking, an estimator for a parameter is consistent if it gets closer and closer to the true parameter value with larger and larger sample sizes.

If the expected value of the estimator is equal to the parameter being estimated, checking for consistency is then a simple matter of seeing if the variance of the estimator goes to 0 as the sample size goes to infinity.

For example, consider a sample $Y_1, \ldots, Y_n$ from a $N(\mu, \sigma^2)$ distribution and consider estimating μ with $\overline{Y}$ as usual. The expected value of $\overline{Y}$ is μ. Also, the variance of $\overline{Y}$ is σ^2/n, which clearly decreases to 0 as the sample size tends to infinity. so $\overline{Y}$ is a *consistent* estimator of μ.

Now, consider the simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. The slope estimator $\hat{\beta}_1$ has expected value β_1 and variance σ^2/S_{xx} where $S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2$.

(Hint to answering the questions below: recall that the inputs (x values) are arbitrary fixed numbers, and one input x can have multiple outputs (y values) measured there. My suggestion is to use examples of datasets that only have two distinct inputs, each with one or more outputs measured at each input.)

(a) **(3 marks)** Describe an example of how a regression dataset's sample size can increase so that the variance of $\hat{\beta}_1$ tends to 0, therefore making the slope estimator *consistent*.

Suppose there were only two inputs, say, $x = 0$ and $x = 1$, but that there are an equal number, say, m outputs (y values) observed at both inputs. The sample size is $n = 2m$. In this case $S_{xx} = \sum_{i=1}^{n} \frac{1}{4}$, which diverges as n grows big, so σ^2/S_{xx} goes to 0.

(b) **(2 marks)** Describe an example of how a regression dataset's sample size can increase so that the variance of does not tend to zero, therefore making the slope estimator *not consistent*.

Again, suppose there are two inputs $x = 0$ and $x = 1$, but this time there is only one output observed at $x = 1$ and $n - 1$ observed at $x = 0$. In this case $S_{xx} = \frac{n-1}{n^2} + \left(1 + \frac{1}{n}\right)^2$ (this takes a little work to determine) which converges to 1 as n gets large. So the variance of $\hat{\beta}_1$ does not converge to 0.

I noticed a student came up with an easier example, in which there are three x values only at, say, -1, 0, and 1. There is one output only observed at -1 and 1, and all the other $n - 2$ are observed at 0. In this case S_{xx} is simply 2, no matter how big n gets.