MIE237 Term Test 2 Solutions

2016-03-22

Examination Type B; Calculator Type 2 Permitted

50 Minutes; 40 Marks Available

Family Name:_____

Given Name:_____

Student Number:_____

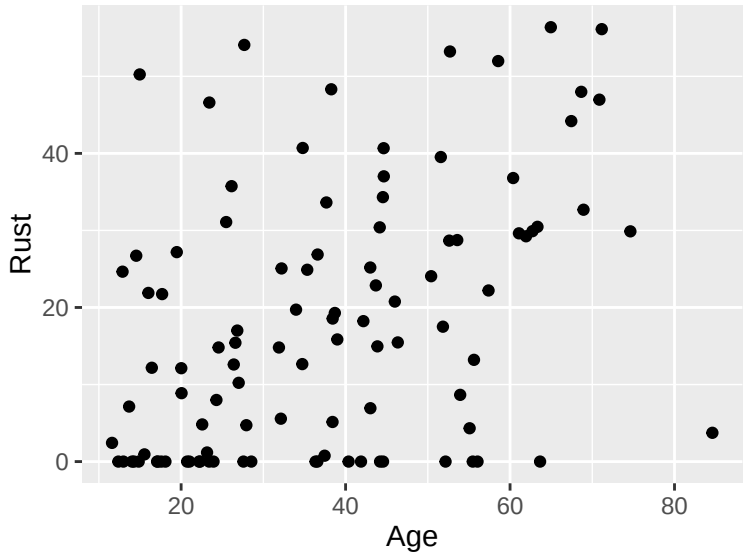
This test contains 5 pages. Pages 6–8 are tables. Page 9 is a formula sheet. You can detach the formula sheet if you like, but please don't detach the tables. (Detaching too many pages causes the test to fall apart.) You may use the backs of pages for rough work.

An electricity distribution company (a company that delivers electricity to homes and businesses) has accumulated a dataset related to 102 failed small transformers and wants to analyse some aspects of the data.

The dataset is similar in structure to the one from the first test, with one new variable added, similar to the first assignment. But the numbers are all different now.

The dataset has 5 variables: `ID, Manufacturer, Size, Age, Rust`. The variable `ID` contains the serial number of the transformer. The variable `Manufacturer` contains the manufacturer name, one of: `Camtran, Nema, Moloney`. The variable `Size` contains a description of the transformer's power rating. The variable `Age` contains the age in years of the transformer at the time of its failure. The variable `Rust` contains a number from 0 to 100 indicating the amount of corrosion that the transformer has.

1. **(25 marks total)** Here is a plot of the data along with the `R` output for the linear regression model fit with `Rust` as the output variable and `Age` as the input variable. Some of the entries have been removed and replaced with `XXX` (at least as many `X` as required).




```
Coefficients:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  2.80224    3.58909   0.781     0.437
Age          XXXXXXX    0.08673   4.617 XXXXXXXXX
---

Residual standard error: 15.35 on XXXX degrees of freedom
Multiple R-squared:  XXXXXX,    Adjusted R-squared:  XXXXXX
F-statistic: XXXXX on XX and XXXX DF,  p-value: XXXXXXXX
```

a) **(5 marks)** Perform the hypothesis test $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$ **using an F distribution and estimating the p-value using the 0.05 and 0.01 probability F tables provided**.

```
```
##
Call:
lm(formula = Rust ~ Age, data = tx)
##
Residuals:
Min   1Q Median   3Q   Max
-32.947 -10.697 -1.299   9.718 41.450
##
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.80224   3.58909   0.781   0.437
Age   0.40045   0.08673   4.617 1.16e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```
```

```
## Residual standard error: 15.35 on 100 degrees of freedom
## Multiple R-squared:  0.1757, Adjusted R-squared:  0.1675
## F-statistic: 21.32 on 1 and 100 DF,  p-value: 1.159e-05
```

b) **(3 marks)** Now, give the *best possible* estimate for the p-value in a) using all available information in this midterm package (tables and formula sheet).

The p-value is the same for the t approach as with the F approach. The best possible estimate is to use the 100 df line from the t table: the p-value is less than 0.001.

c) **(3 marks)** Produce a 95% confidence interval for β_1.

$0.4004478 \pm 1.9839715 \cdot 0.0867311$

$[0.2283759, 0.5725198]$

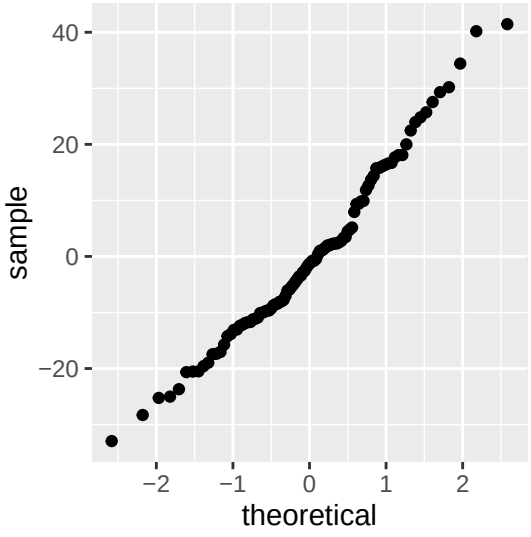
d) **(4 marks)** Denote the observed values of the 'Age' variable by $\{x_1, \ldots, x_n\}$. It turns out that $\overline{x} = 37.5$ and $S_{xx} = 31304.1$. Produce a 95% confidence interval for the mean value of 'Rust' at an 'Age' of 50.

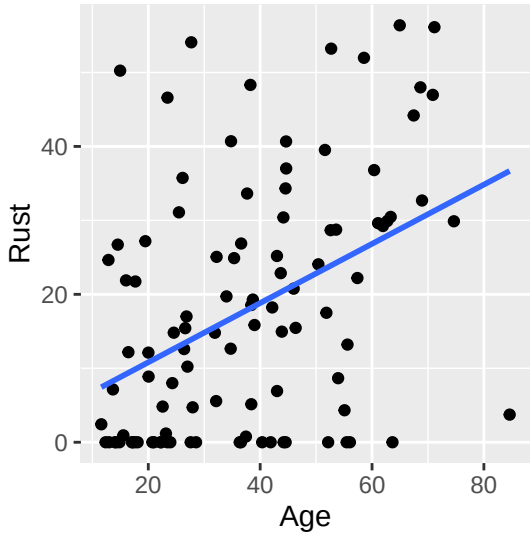
```
##        fit      lwr      upr
## 1 22.82463 19.12055 26.52872
```

e) **(3 marks)** Produce a 95% prediction interval for the value of 'Rust' for a transformer with an 'Age' of 15.

```
##       fit       lwr      upr
## 1 8.80896 -22.02836 39.64628
```

f) **(4 marks)** Here is a normal quantile plot of the residuals, along with a plot of the raw data with the fitted regression line included.





In light of these plots, which of the calculations done in a) to e) above do you think is probably the *least* accurate, and why?

The normal quantile plot is straight, so there is no problem with normality. The 'Rust' variable looks like it has a minimum value of 0. There is a bunching of observations at the bottom of the plot, especially on the left. There would appear to be much less variation possible below the regression line to the left.

All this suggests a big problem with the regression model (unequal variance problem) affecting all calculations. But the prediction interval will probably be the least accurate, since it depends the most strongly on everything being correct. It was also clear there was a problem with the interval itself, going so far below 0 where there are are no actual observations.

g) **(3 marks)** Compute the sample correlation coefficient between the 'Age' and 'Rust' variables.

0.4191882

2. **(10 marks total)** Let's call the `Rust` variable y and the `Age` variable x_1. Two new variables x_2 and x_3 are added to the data. If the transformer was manufactured by Camtran, both x_2 and x_3 are set to 0. If the transformer was manufactured by Nema, x_2 is set to 1 and x_3 is set to 0. If the transformer was manufactured by Moloney, x_2 is set to 0 and x_3 is set to 1.

The multiple regression model $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}$ is considered.

a) **(7 marks)**When the model is fit to the data, the value of MSE turns out to be 152.73. Perform the hypothesis test that answers the question "is there any linear relationship between the output variable and the input variables?". (You will need information from 1.a). If you are completely stuck you just can go ahead and use a total sum of squares of 30000 for a maximum of 4 out of 7 marks.)

	DF	SS	MS	F	p-value
Regression	3	13599.77	4533.26	29.68	0.00
Error	98	14967.94	152.73		
Total	101	28567.71	4685.99		

b) **(3 marks)**The p-value for the hypothesis test $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$ is 0.000000148. Give a brief practical conclusion you can make from this hypothesis test.

There is a significant relationship between 'Rust' and 'Age', over and above the information provided by knowing the value of 'Manufacturer' (as encoded in x_2 and x_3.)

	DF	SS	MS	F	p-value
Regression	3	15032.06	5010.69	32.81	0.00
Error	98	14967.94	152.73		
Total	101	30000.00	4685.99		