Department of Statistics

COURSE STATS 330

Model Answers for Assignment 2, 2004

The viscometer is a scientific instrument that measures the viscosity of a fluid by measuring the time taken for an inner cylinder in the mechanism to perform a fixed number of revolutions in response to an actuating weight. The viscometer is calibrated by measuring the time taken with varying weights while the mechanism is suspended in fluids of accurately known viscosity. The data overleaf come from such a calibration. The variables are

Viscosity: Viscosity of fluid

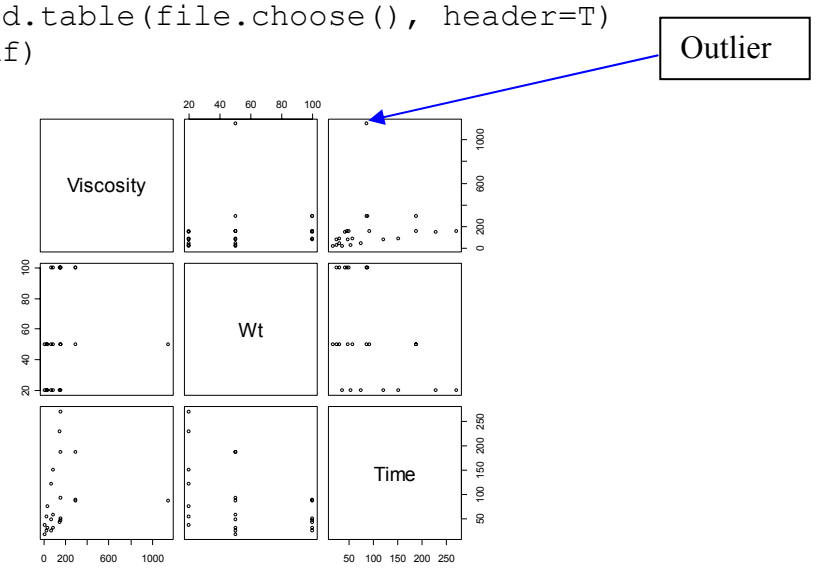
Wt: Actuating weight

Time: Time taken

The data (in the form of a tab-delimited text file) are available on the course web page under the title viscosity.txt.

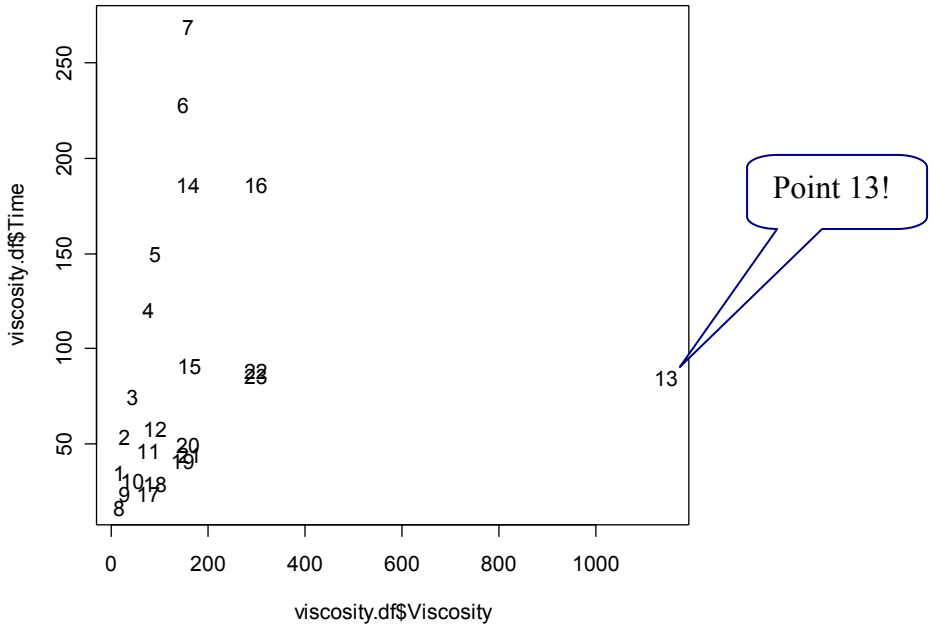
Reading in data, outlier checks:

```
viscosity.df<-read.table(file.choose(), header=T)
pairs(viscosity.df)
```



Looks like there is one outlier. We can identify it by plotting Viscosity versus Time, labeling the points by their row in the data frame:

```
plot(viscosity.df$Viscosity,viscosity.df$Time, type="n")
text (viscosity.df$Viscosity,viscosity.df$Time, 1:23)
# (there are 23 data points)
```



Looks like point 13 is the culprit:

```
> viscosity.df[13,]
   Viscosity Wt Time
13    1146.6 50 85.6
```

Checking the printed data sheet, we see that the viscosity should be 146.6. We correct the mistake:

```
viscosity.df[13,1]<-146.6
> viscosity.df[13,]
Viscosity Wt Time
   13  146.6 50 85.6
```

The rest of the data are OK. Now we can answer the questions.

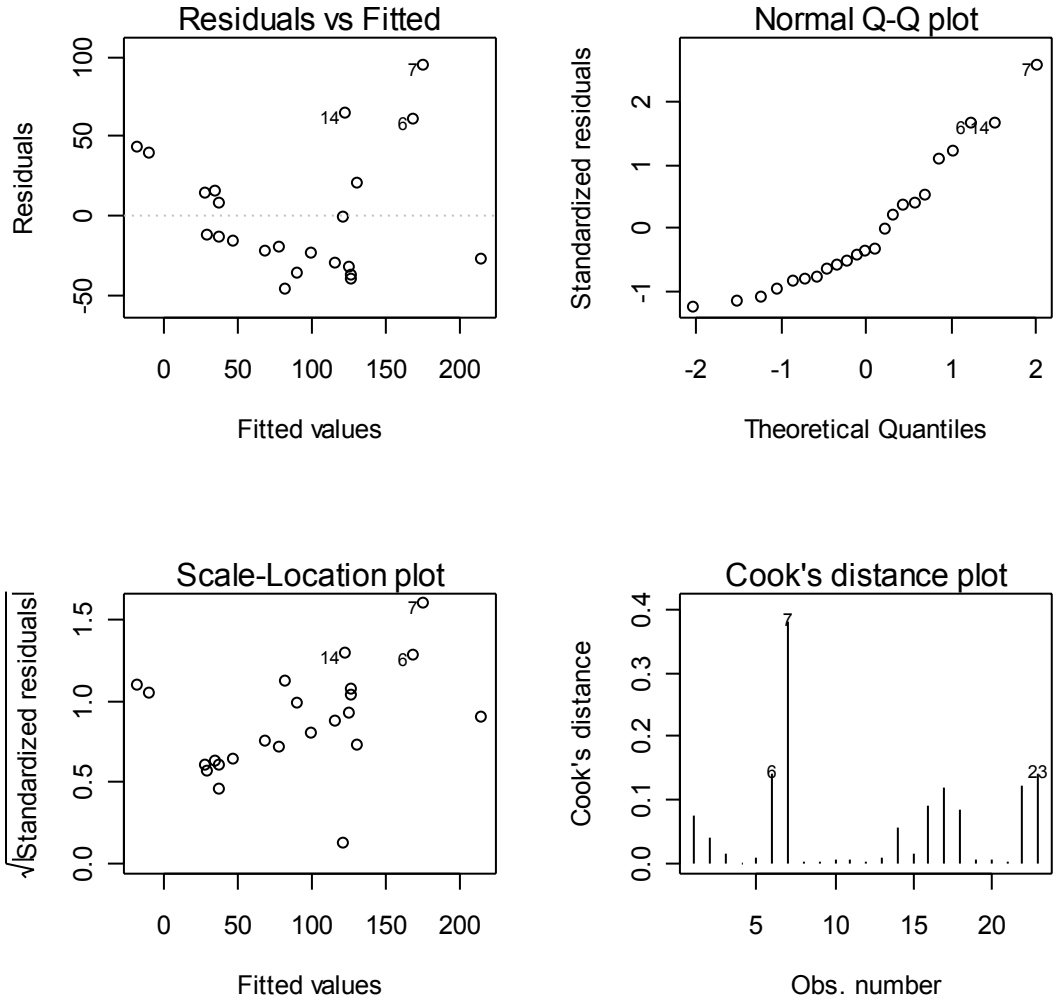
1. Fit a regression model to the data, using time as the response. Then, having fitted the model, examine the fit for

- *Non-planar regression*
- *Non-constant variance*
- *Outliers and high-leverage points*
- *Lack of normality*

Make a list of the defects in the fit that you have found. Show any plots used, together with the code used to produce them

We fit the model and draw the plots using the code

```
viscosity.lm<-lm(Time~Viscosity + Wt, data=viscosity.df)
layout(2,2)
plot(viscosity.lm)
```



From these plots, we can draw the following conclusions:

- From the plot of residuals versus fitted values, there seems to be a definite non-linear effect in these data.
- From the normal plot, there seems to be a degree of non-normality.
- From the scale-location plot, some hint of variance increasing with mean.
- Lets check for outliers and influential points:

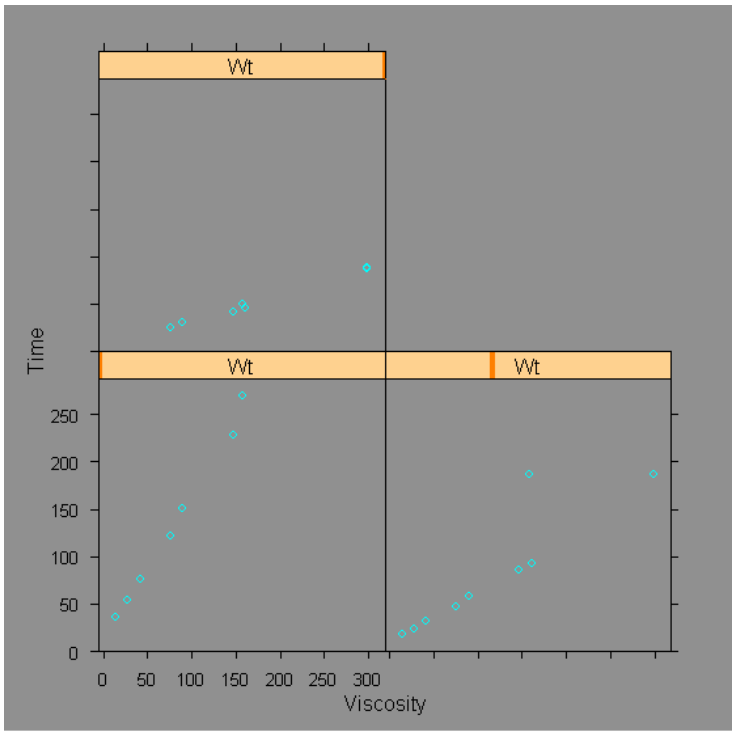
```
> qf(0.5,2,20)
[1] 0.7177346
> qf(0.9,2,20)
```

```
[1] 2.589254
```

so the spikes in the Cooks D plot are not indicative of outliers or influential points. Moreover, none show up on the other plots.

Thus, the main problem seems to be a non-planar regression. Note that there are only three distinct values of the variable Wt, so to get more insight into the nature of the regression surface, let's try a coplot, conditioning on Wt :

```
xyplot(Time~Viscosity|Wt, data=viscosity.df)
```



Now the situation is clearer: there is a strong relationship between Viscosity and Time, but the slope depends on Wt.

2. Find a suitable transformation that will cure (or at least partially cure) the defects you listed in 1. Document the reasoning that led you to your transformation.

The slopes are approximately (estimating from the coplot) in the ratio 5:2:1, which is inversely proportional to Wt. The intercepts are approximately zero. Thus the relationship is roughly of the form

$$\text{Time} = \text{Viscosity}/\text{Wt}.$$

We can make this a linear relationship by taking logs, and getting

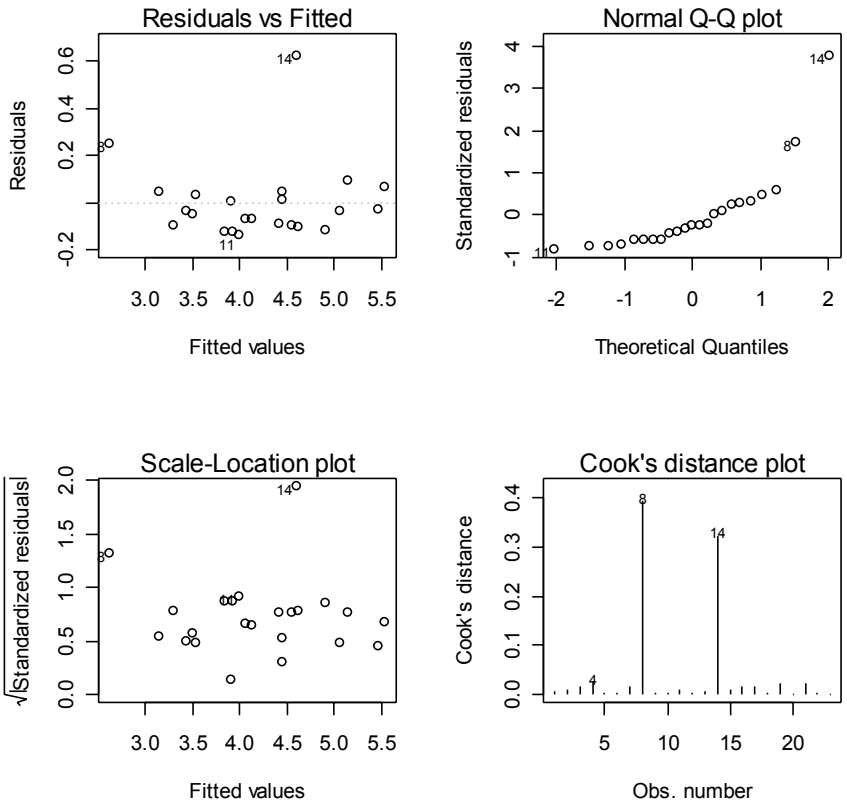
$$\log(\text{Time}) = \log(\text{Viscosity}) - \log(\text{Wt}),$$

which suggests transforming all the variables by taking logs and refitting:

```
logs.lm<-lm(log(Time)~log(Viscosity) + log(Wt),
data=viscosity.df)
```

This gives a fit with an R^2 of 95.5%, compared with an R^2 of 70.5% for the original fit. Lets check the new fit for problems:

```
plot(logs.lm)
```



Points 8 and 13 are clearly outliers: deleting them gives a fit with an R^2 of 99.0%. This is a very good fit. The model is

```
> logs2.lm<-lm(log(Time)~log(Viscosity) + log(Wt),
data=viscosity.df, subset=(1:23)[-c(8,14)])
# (1:23)[-c(8,14)] is the vector of numbers 1,2,…, 23
#  with 8 and 14 deleted
> summary(logs2.lm)

Call:
lm(formula = log(Time) ~ log(Viscosity) + log(Wt),
     data = viscosity.df, subset = (1:23)[-c(8, 14)])
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.09103 -0.06120 -0.01207  0.04871  0.11351

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.23514    0.10532   40.21   <2e-16 ***
log(Viscosity)  0.85555    0.02218   38.57   <2e-16 ***
log(Wt)        -1.02031    0.02710  -37.65   <2e-16 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 `
' 1

Residual standard error: 0.07164 on 18 degrees of freedom
Multiple R-Squared: 0.9909,     Adjusted R-squared: 0.9899
F-statistic:   979 on 2 and 18 DF, p-value: < 2.2e-16
```

We take this as our final model.

3. Use the model you have developed in Questions 1 and 2 to predict the time corresponding to a viscosity of 150 and a weight of 60.

```
> predict.stuff<-predict(logs2.lm,
     data.frame(Viscosity=150, Wt=60),
     interval="prediction")
> predict.stuff
          fit      lwr      upr
[1,] 4.344477 4.189243 4.499712
```

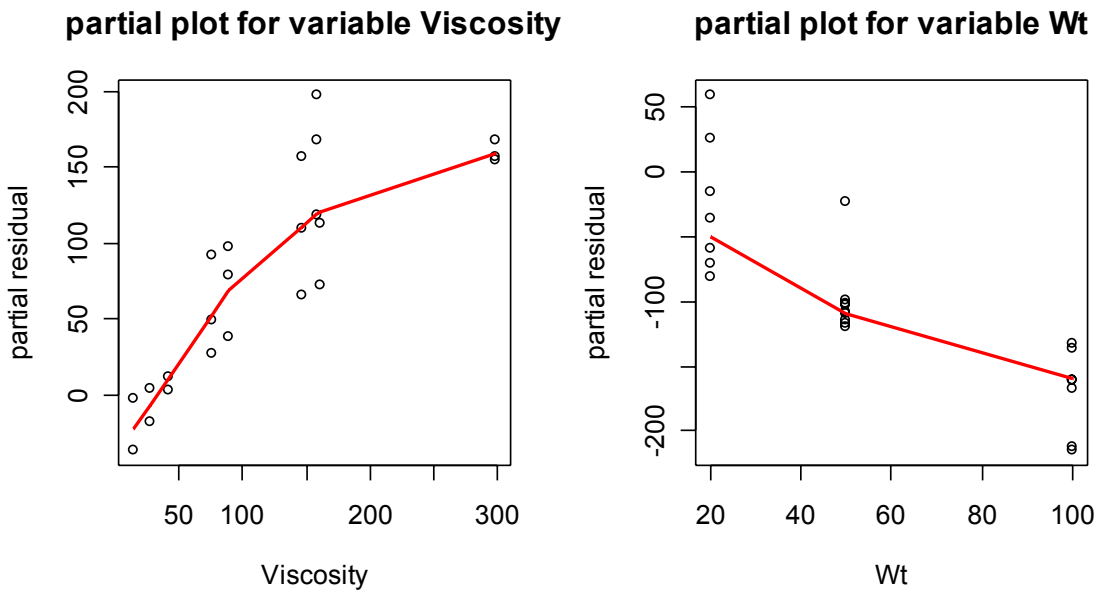
This is an interval for log(Time). To get an interval for Time, we must "exp" the interval:

```
> exp(predict.stuff)
          fit     lwr      upr
[1,] 77.05175 65.9728 89.99122
```

Thus, our prediction interval is (65.97, 89.99).

Other comments: The usual methods of finding transformations are either fitting additive models or partial residual plots. The former don't work in this example, because there are not enough distinct values in the data. The latter produce plot s that roughly suggest a log transformation for Viscosity and a –log transformation for Wt. To make the partial residual plots, "source" the 330 functions and use the function partial.res.plot. This produces

```
layout(1,2)
partial.res.plot(viscosity.lm)
```



If we make these transformations, and refit, we get

```
logs3.lm<-lm(Time~log(Viscosity)  + log(Wt),
data=viscosity.df)
```

This model has an R2 of about 80%. The plot of residuals versus fitted values shows a "curved funnel effect", so we might consider a transformation on Y. A Box-Cox plot has a minimum at p=-1/3, or pretty close to a log. This suggests logging all the variables might be a good idea.

Mark Scheme

Preliminary work
Reading in data, noting outlier, correcting outlier: 10 marks

Question 1
Fitting model, drawing diagnostic plots: 5 marks
Correctly interpreting plots: 5 marks

Question 2
Finding a transformation that improves the R^2: 15 marks. This can be done in many different ways. I have suggested 2. You will have to use your judgment.

Question 3
Making a prediction using the fitted model: 5 marks

Total marks: 40