

# 228371 Statistical Modelling for Engineerings and Technologists

## Assignment 1

Name: Minghua Yang

SID: 02067315

**Q 1:**

>gas<-read.csv(file.choose(),header=T)

>attach(gas)

> t.test(gas\$HeatRate, mu=9750, alternative='greater')

p-value = 0.0003562 .P value is much smaller than 0.05, and the NULL hypothesis is rejected. I

can say the heat rate for turbine exceeds 9751KJ/kwh.

conclusion about your set of  
gas turbines --- sample is not  
representative and conclusions  
should not be applied to gas  
turbines in general

**Q 2:**

> t.test(InletTemp,ExhTemp,paired=TRUE)

p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

618.6944 703.7431

P value is much smaller than 0.05, and then NULL hypothesis is rejected. The default value for NULL hypothesis is zero. So i can say the mean reduction between InletTemp and ExhTemp exist, which value is between 618.6944 and 703.7431.



### Q3:

```
> tr = Power[Engine == 'Traditional']
```

```
> ad = Power[Engine == 'Advanced']
```

alternative = "greater"

```
> t.test(ad,tr,paired=F,var.equal=FALSE)
```

As the data shows, p-value = 0.0001507. alternative hypothesis: true difference in means is not equal to 0. So i can say the advanced engine have more power than the other.



-1

The assumptions are the different size of sample group, and the variance is unequal.

This depends on the engines in our population being a random (or at least representative) sample of all engines of their respective types. Because the sample sizes are relatively small, we also need the distribution of the measurements within each group to be approximately normal.

### Q4:

```
> t.test(InletTemp[Engine == 'Aeroderiv'],ExhTemp[Engine == 'Aeroderiv'],paired=TRUE)
```

```
>t.test(InletTemp[Engine == 'Advanced'],ExhTemp[Engine == 'Advanced'],paired=TRUE)
```

```
> t.test(InletTemp[Engine == 'Traditional'],ExhTemp[Engine == 'Traditional'],paired=TRUE)
```

As summary, three different types of turbines have three different estimated mean temperature reduction. The values are 587, 775.3636 and 608.6429.

-2

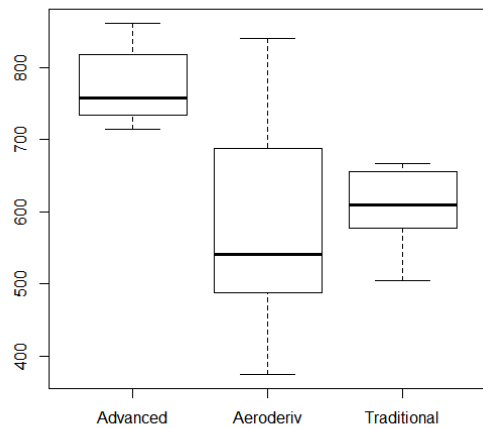
```
> diff <- (InletTemp-ExhTemp)
```

```
> boxplot(diff~Engine)
```

```
Mod <- lm(InletTemp-ExhTemp ~  
Engine)
```

```
anova(Mod)
```

Is Engine significant?



The graph shows the mean of the heat difference between inlet and outlet is different significantly.



### Q5

> air\_p<-lm(Airflow~Power, data=gas)

> summary(air\_p)

Formal hypothesis test  
for correlation required.  
>cor.test(Airflow,Power)

As conclusion, P value is 2e-16, so i would reject NULL hypothesis, and say this linear model is suitable for the relationship between Airflow and Power.

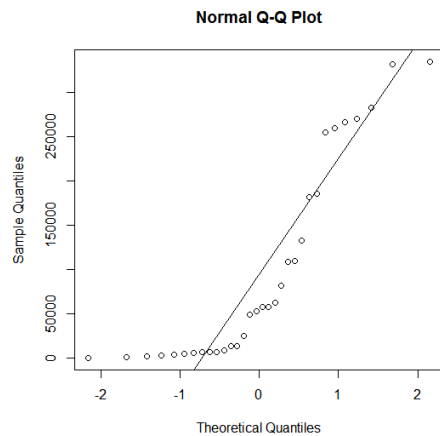


### Q6

> qqnorm(Power)

> qqline(Power)

-1



```
> t.test(Power)
```

```
> shapiro.test(gas$Power)
```

p-value = 5.304e-05



From the graphs, the trend of data points do not follow qqline, and the p-value is much smaller than 0.05, which means NULL hypothesis has been rejected. Null hypothesis stands for Power is normally distributed, so i can say this is not a normal distribution of the power of the turbines.



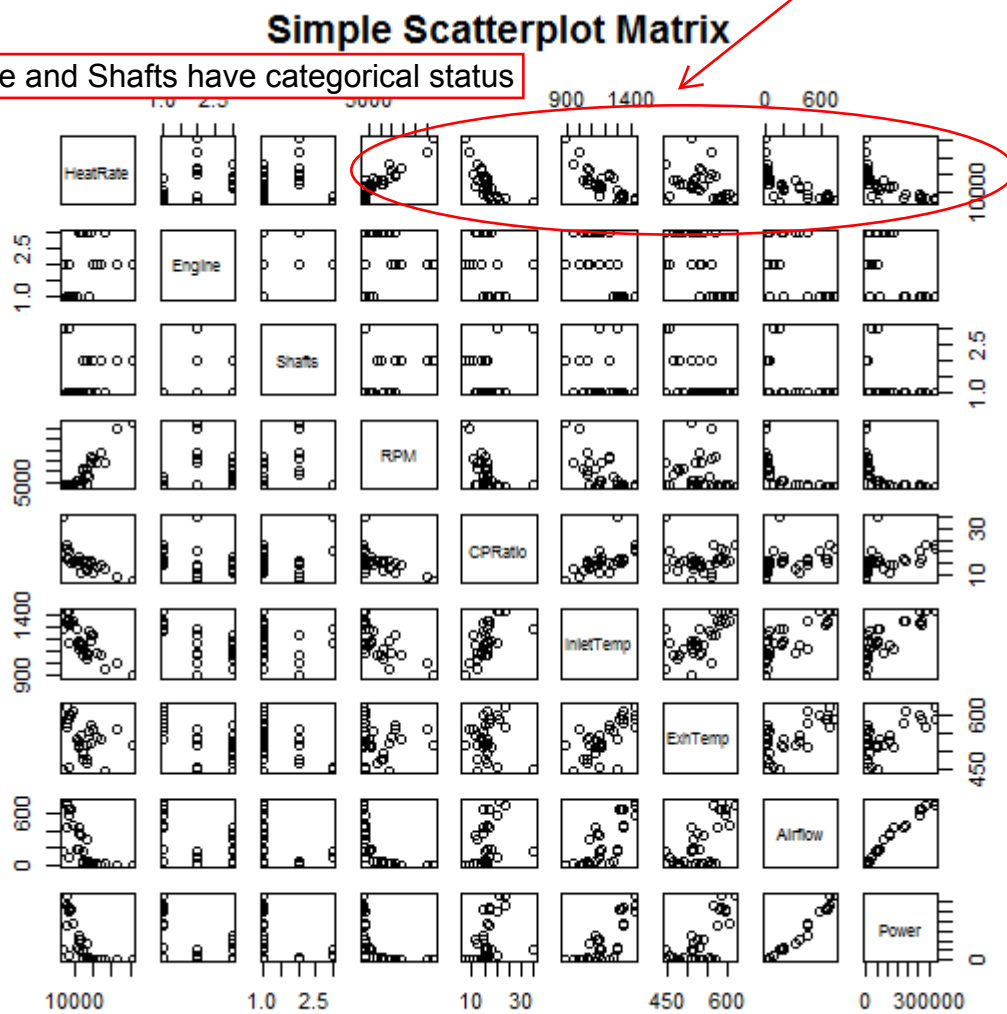
### **Q7:**

```
> pairs(HeatRate~Engine+Shafts+RPM+CPRatio+InletTemp+ExhTemp+Airflow+Power,data=gas,
main="Simple Scatterplot Matrix")
```

-2

Engine and Shafts have categorical status

Only these plots are required



From the matrix graphs, i find two graphs which hold data points around a line. They are HeatRate vs RPM and HeatRate vs InletTemp. So i can say RPM and InletTemp are the numeric variables are useful as linear regression predictor.

**Q 8:**

```
> he_sp<-lm(HeatRate~RPM)
```

Equation :  $\text{HeatRate} = 0.2019 \cdot \text{RPM} + 9108.6573$

```
> summary(he_sp)
```

P-value is  $3.393e-14$  and Multiple R-squared: 0.8568 which value is close to one, so it mean these two data sets fit on a linear model.

-2

Q9.

```
> he_sp1<-lm(HeatRate~ExhTemp)
```

```
> summary(he_sp1)
```

Multiple R-squared: 0.07478, which means holds smaller correlation coefficient than the one on Q8. So ExhTemp is worse predictor of heatrate.

Why did you choose this variable? You mentioned in Q7 that the only other suitable variable was InletTemp

Q10.

```
> he_sp2<-lm(HeatRate~InletTemp*RPM)
```

```
> summary(he_sp2)
```

That's better

Obviously, p-value for interaction model of both predictors is 0.485491. So this model is less suitable model than the additive model. As conclusion, the interaction mode is not required.



-1

Fit the additive model aswell for completeness

Q11.

```
> lm1<-lm(HeatRate~Shafts+RPM+CPRatio+InletTemp+ExhTemp+Airflow+Power)
```

Coefficients:

(Intercept)	Shafts	RPM	CPRatio	InletTemp	ExhTemp
1.378e+04	-2.078e+02	1.501e-01	-3.399e+01	-3.684e+00	1.839e+00
Airflow	Power				
1.609e+00	-4.025e-03				

So the model is;

-2

Comment on significance of the terms included.  
Discuss the value of this model vs the added complexity

HeatRate=

$$(-2.078e+02)*Shafts+(1.501e-01)*RPM+(-3.399e+01)*CPRatio+(-3.684e+00)*InletTemp+(1.839e+00)*ExhTemp+(1.609)*Airflow+*(-4.025e-03)Power+1.378e+04$$

### Q12

-1 According to seven different beta values, i figure our that, when predictor variables of Shaft, CPRatio, InletTemp, Power increase, the response variable decrease with each particular ratio. When predictor variables of RPM, ExhTemp, Airflow increase, the repose variable increase with each particular ratio as well.

??

By how much?  
Give an example  
for atleast one  
beta value

### Q13

> he\_sp<-lm(HeatRate~InletTemp+RPM)

Residual standard error: 472.9 on 29 degrees of freedom

> lm1<-lm(HeatRate~Shafts+RPM+CPRatio+InletTemp+ExhTemp+Airflow+Power)

Residual standard error: 408.1 on 24 degrees of freedom

Based on least squared theory, Seven predictor model has the smaller Residual standard error, which mean this model has better predictive than the two predictor model.

### Q14

Give unit 95.44%

-5 For seven predictors model,  $R^2 = 0.9544$ . And for two predictors model,  $R^2 = 0.07478$ . So seven predictors has the better fit in this model.

Interpret this R-sq. 95.44% of the variation in the heat rate of the turbines is explained by the fitted model that includes all numerical variables

### Q15

The overall p-value:  $1.539e-14$ , and it is much smaller than 0.01. so Null hypothesis is rejected. Which means this model is good to predict HeatRate.



### Q16


```
> lm1<-lm(HeatRate~Shafts+RPM+CPRatio+InletTemp+ExhTemp+Airflow+Power)
```

```
> hatvalues(lm1)
```

```
> which((hatvalues(lm1)/mean(hatvalues(lm1))>2))
```


```
1 4 24
```

```
1 4 24
```

There are three of leverages more than twice the mean leverage, and we need to concern about it. However there is no enough evidence to prove point 1,4, 24 are significantly influential, but it could be potentially influential. 

### Q17

```
> lm1<-lm(HeatRate~Shafts+RPM+CPRatio+InletTemp+ExhTemp+Airflow+Power)
```

The p-value for ExhTemp is 0.7253, so i take this predictor away from equation. 

```
> lm1<-lm(HeatRate~Shafts+RPM+CPRatio+InletTemp+Airflow+Power)
```

-1

The p-value for Power is 0.47024, so i take this predictor away from equation to

Should take away  
Airflow first since  
p = 0.47104

```
> lm1<-lm(HeatRate~Shafts+RPM+CPRatio+InletTemp+Airflow)
```

The p-value for Airflow is 0.94522 , so i take this predictor away from equation too.

```
> lm1<-lm(HeatRate~Shafts+RPM+CPRatio+InletTemp)
```

The p-value for Shafts is 0.193756 , so i take this predictor away from equation too.

```
> lm2<-lm(HeatRate~CPRatio+RPM+InletTemp), then there is no big p-value shows.
```

So my reduced model is HeaRate vs CPRatio, RPM ,InletTemp.

### Q18

```
> anova(lm1)
```

Analysis of Variance Table

Response: HeatRate

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Shafts	1	2877374	2877374	17.2740	0.0003548 ***
RPM	1	73314150	73314150	440.1336	< 2.2e-16 ***
CPRatio	1	4375057	4375057	26.2652	3.03e-05 ***
InletTemp	1	2939898	2939898	17.6494	0.0003166 ***
ExhTemp	1	18693	18693	0.1122	0.7405365
Airflow	1	4340	4340	0.0261	0.8731211
Power	1	85211	85211	0.5116	0.4813675
Residuals	24	3997740	166573		

`>anova(FullModel,ReducedModel)`

Reduced model is better.

There are big p-values existing in Power, Airflow ,ExhTemp, so seven predictor model is not good as reduced model.

You need a formal test for this statement

### Q19

`> par(mfrow=c(2,2))`

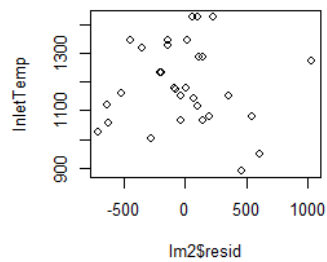
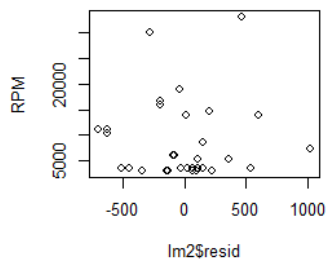
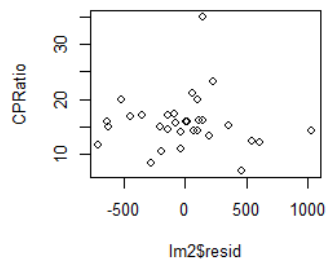
`> plot(CPRatio~lm2$resid)`

`> plot(RPM~lm2$resid)`

`> plot(InletTemp~lm2$resid)`

`> par(mfrow=c(2,2))`  
`> plot(ReducdModel)`

Will give you a 4 in One residual plot.



Resids vs fits?  
Normal QQ plot  
Resids vs leverage  
plot?  
etc

**-4**

There are random data points showing in the graph, and there is no particular trend in these graphs data pattern, so it is no need to concern about.

Comment on normality and equal variance assumptions

## Q20

```
> which(cooks.distance(lm2)>1)
```

```
named integer(0)
```

```
> library(car)
```

```
> vif(lm2)
```

```
Shafts      RPM InletTemp
1.119740  1.831977  1.793526
```

All the cooks distance values are smaller than 1, and all the variance inflation factor are smaller than 10, so there is nothing to worry about.

## Q21

```
> poly3<-lm(HeatRate~poly(ExhTemp,3,raw=TRUE))
```

```
> poly2<-lm(HeatRate~poly(ExhTemp,2,raw=TRUE))
```

```
> poly1<-lm(HeatRate~poly(ExhTemp,1,raw=TRUE))
```

```
> anova(poly1,poly2)
```

P-value = 0.02783

```
> anova(poly1,poly3)
```

P-value = 0.0923

```
> anova(poly2,poly3)
```

P-value = 0.907

Small p-value 0.02782 proves this polynomial regression model with degree of two is the best model in three. Because the other two p-value is significantly large.

## Q22

```
> poly1<-lm(HeatRate~poly(CPRatio,RPM,degree=2,raw=TRUE))
```

```
> summary(RPM)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
3000	3600	5400	8710	11920	33000

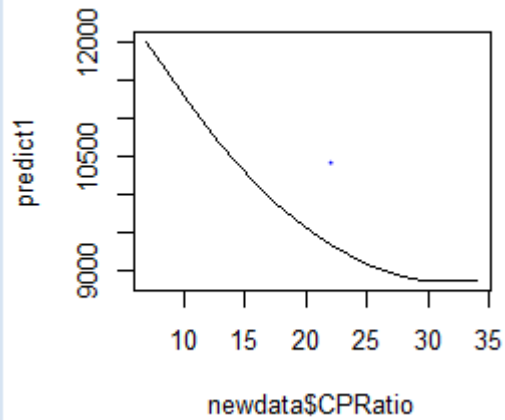
```
> summary(CPRatio)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
6.90	13.85	15.50	15.93	17.00	35.00

```
> newdata<-data.frame(RPM=rep(4500,19),CPRatio=seq(7,35,1.5))
```

```
> predict1<-predict.lm(poly1,newdata)
```

```
> plot(newdata$CPRatio,predict1,type='l')
```



The graph shows the new predicted data when RPM equals to 4900 rpm, and CPRatio is between 7 and 35.

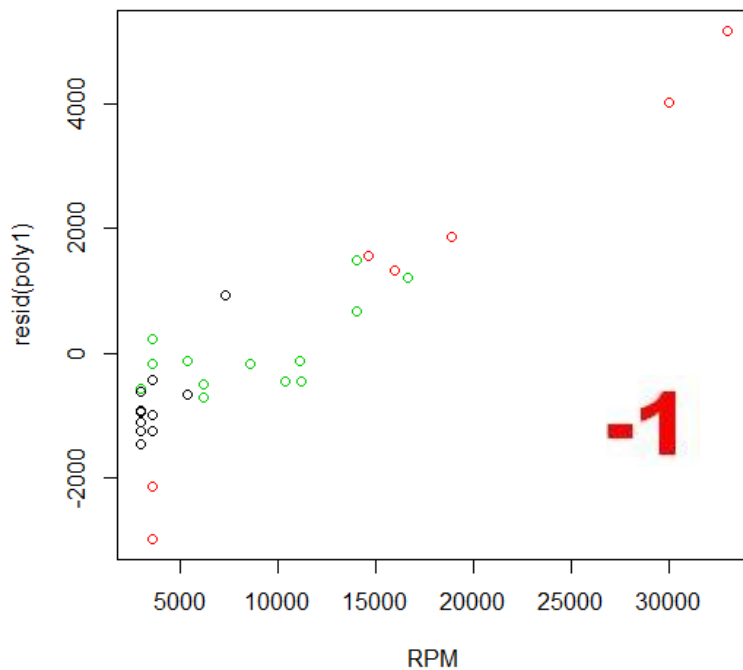
### Q23

```
> plot(RPM,resid(poly1),col=Engine)
```



-1

This code is correct but it is not producing the correct plot. Make sure that 'poly1' is your second order model



Comment - is engine important?

#### Q24

```
> poly1<-lm(HeatRate~poly(CPRatio,RPM,degree=2,raw=TRUE)+Engine)
```



```
> summary(poly1)
```

-2

Obviously, the overall p-value is  $5.108e-13$ , and it does improve the ability of this model to predict HeatRate.

The model shows some predictive capability but does not improve the ability of this model to predict HeatRate look at anova(poly1,second order model)

#### Q25

```
> poly1<-lm(HeatRate~poly(CPRatio,RPM,degree=2,raw=TRUE)*Engine)
```



```
> anova(poly1)
```

-1

anova to compare this model against the 2nd order model

In the additive model, the response surface shift up and down on vertical axis. Bu in multiplication model, the response surface changes in 3-D dimensions. So as conclusion, there is other different response surface existing.

-2

Interaction is not significant so there is insufficiency evidence that there is a different response surface for each engine type