Question 1 **[45 marks]**

(a) Plot *cases* against *week*. What does the plot suggest? [5 marks]

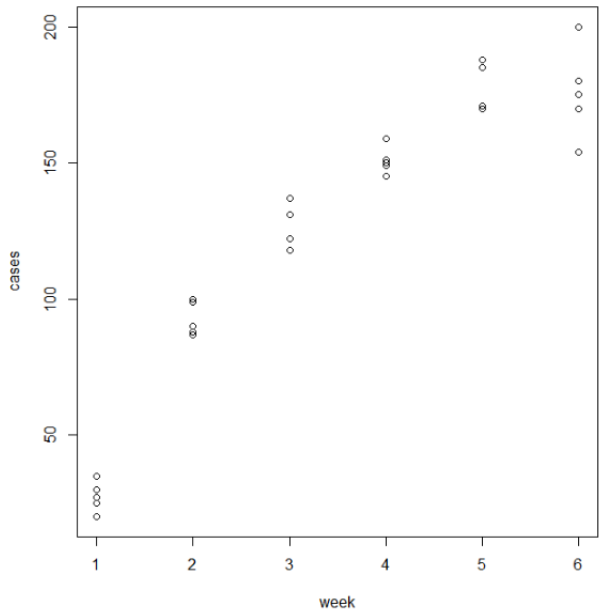

Figure 1: Number of COVID-19 cases recorded over 6 weeks

Plot of `cases` against `week`, Figure 1, suggests that the association between `cases` and `week` would obviously not be well described by a simple linear regression model, instead a quadratic model or a higher polynomial model may be approriate. There are only 6 distinct values in the predictor, so the highest degree of the polynomial model should not be excess 5.

(b) Determine the order of the polynomial model required to fit the data. With reference to relevant outputs, justify each step in the process. [15 marks]

The quadratic model was the starting point: $\text{E(cases)} = \beta_0 + \beta_1 Week + \beta_2 Week^2$. Outputs for this model are given in Table 1.

Table 1: Summary table for the quadratic model

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -37.080      7.904   -4.69  7.0e-05
week          74.209      5.171   14.35  3.7e-14
I(week^2)     -6.443      0.723   -8.91  1.6e-09

Residual standard error: 9.88 on 27 degrees of freedom
Multiple R-squared:  0.969,Adjusted R-squared:  0.967
F-statistic:  419 on 2 and 27 DF,  p-value: <2e-16
```

From Table 1, the quadratic model is useful to predict the number of cases (F-statistic = 419 on 2 and 27 df, p-value $< 2e-16$) and it explains about 96.7% of the variability in the response variable. The second-order term ($week^2$) is proven to be a significant predictor (p-value =1.6e-09 which is almost 0). No need to check for the ANOVA table.

Next, the cubic model was checked with the summary of the coefficients given in Table 2.

Table 2: Summary table for the cubic model

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -49.867     15.976   -3.12   0.0044
week          90.293     18.203    4.96  3.7e-05
I(week^2)    -11.771      5.825   -2.02   0.0537
I(week^3)      0.507      0.550    0.92   0.3651

Residual standard error: 9.91 on 26 degrees of freedom
Multiple R-squared:  0.97,Adjusted R-squared:  0.966
F-statistic:  278 on 3 and 26 DF,  p-value: <2e-16
```

From Table 2, it can be seen that the cubic model does not improve the adjusted R^2, and the cubic term ($week^3$) is also not significant (p-value = 0.365). So we check the ANOVA table (Table 3).

Table 3: ANOVA table for the cubic model

```
Analysis of Variance Table

Response: cases
           Df Sum Sq Mean Sq F value   Pr(>F)
week        1  74140   74140  755.25   < 2e-16
I(week^2)   1   7749    7749   78.93   2.3e-09
I(week^3)   1     83      83    0.85   0.37
Residuals 26   2552      98
```

The p-value for $week^3$ in Table 3 is not significant (p-value = 0.37). This suggests that after having $week$ and $week^2$ in the model, the cubic term does not add significant information.

Therefore, the term $week^3$ should be dropped, and we consider the quadratic model as the final model.

From Table 1, the fitted equation for the quadratic model is:

$$E(cases) = -37.08 + 74.21week - 6.44week^2$$

Optional: you can go further and test the quartic model with $week^4$.

Table 4: Summary table for the quartic model

```
Coefficients:
             Estimate Std. Error  t value  Pr(>|t|)
(Intercept) -127.867     35.471   -3.60    0.0014
week         224.627     58.078    3.87    0.0007
I(week^2)    -84.044     30.399   -2.76    0.0105
I(week^3)     15.674      6.300    2.49    0.0199
I(week^4)     -1.083      0.449   -2.42    0.0234

Residual standard error: 9.1 on 25 degrees of freedom
Multiple R-squared:  0.976,Adjusted R-squared:  0.972
F-statistic:  249 on 4 and 25 DF,  p-value: <2e-16
```

Outputs for the quartic model are given in Table 4. This model slightly improves the adjusted R^2 (97.2%) compared to the quadratic model (96.7%). All terms $week$, $week^2$, $week^3$ and $week^4$ are significant (p-values of 0.0007, 0.0105, 0.0199 and 0.0234 respectively). So you can also use this as the final model.

The fitted model is:

$$E(cases) = -127.87 + 224.63week - 84.04week^2 + 15.67week^3 - 1.08week^4$$

(c) For the final model in (b), check the model assumptions, and also identify any potential outliers or influential points. [15 marks]

The residuals plots for the quadratic model are given in Figure 2.
There is no obvious pattern in the residuals vs fitted plot, the residuals are randomly scattered around 0 so the assumption of equal variances appears valid. The scale location plot does show a slight positive trend, suggesting a slight increase in variance.

The QQ-plot is an approximate straight line so that the errors appear to have a normal distribution. This is also confirmed from the Shapiro-Wilk test (Table 5).

Observation numbers 28 and 29 could be the potential outliers, since their standardized residuals more than 2 and less than -2, respectively. These 2 observations also have the largest Cook's distance.

Table 5: Shapiro-Wilk normality test for the quadratic model

```
Shapiro-Wilk normality test

data:  VM2$residuals
W = 1, p-value = 0.5
```

Observation 28 has a Cook's distance of ≈ 0.5, which corresponds to 31^{th} percentile on the $F_{3,27}$ distribution - this observation is not an influential. Thus, we can conclude that there are no influential points for this model. This can be confirmed using the Residual

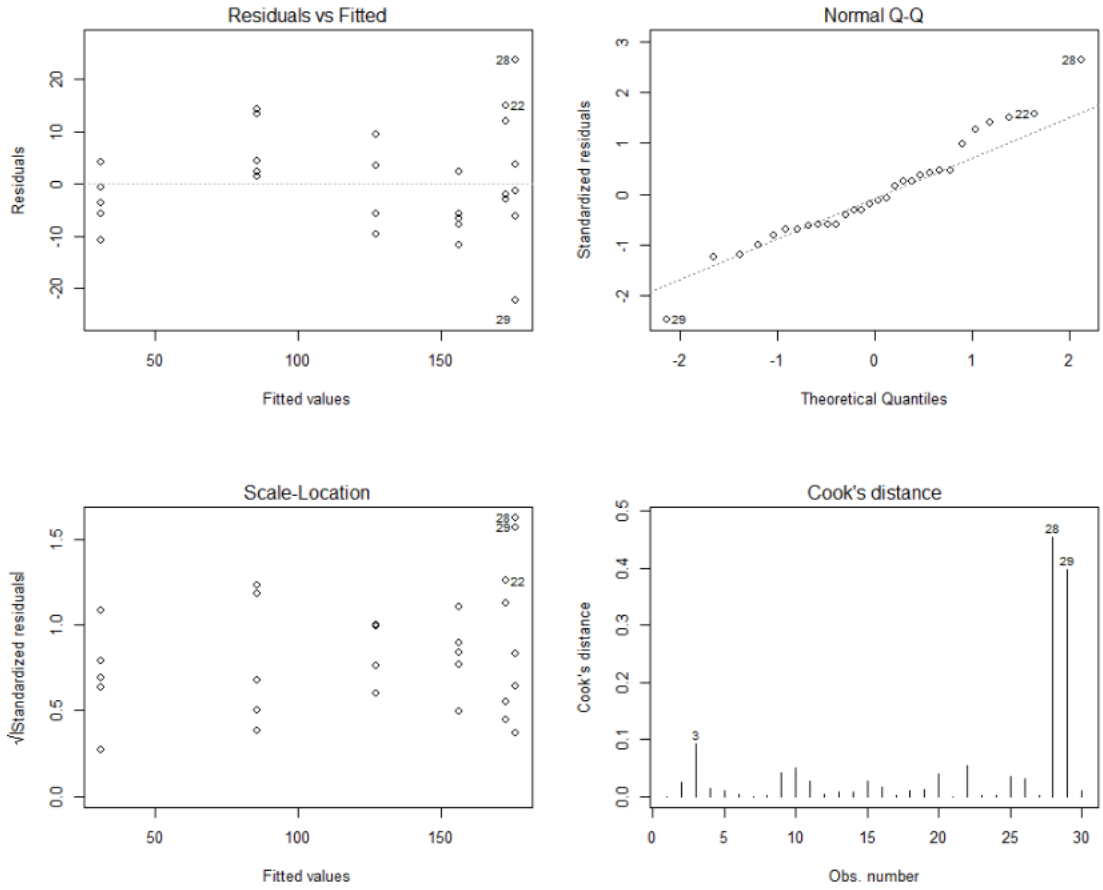

Figure 2: Diagnostic plots for the quadratic model

vs Leverage plot (Figure 3). There are no observations outside of the 0.5 contour i.e. at the upper right corner and lower right corner, so there are no influential points.

Overall, all model assumptions are reasonable, thus the quadratic model appears appropriate.

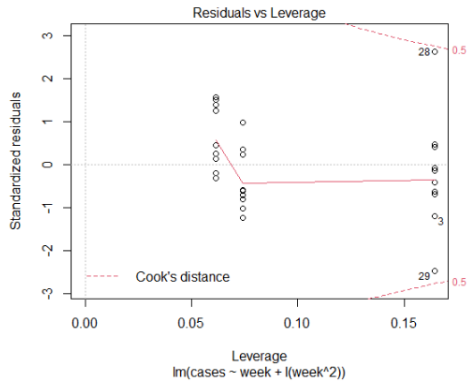

Figure 3: Residual vs Leverage plot for the quadratic model

(d) Plot the fitted values on a scatter plot of the data. Include a plot of the 95% confidence bands on your graph [10 marks]

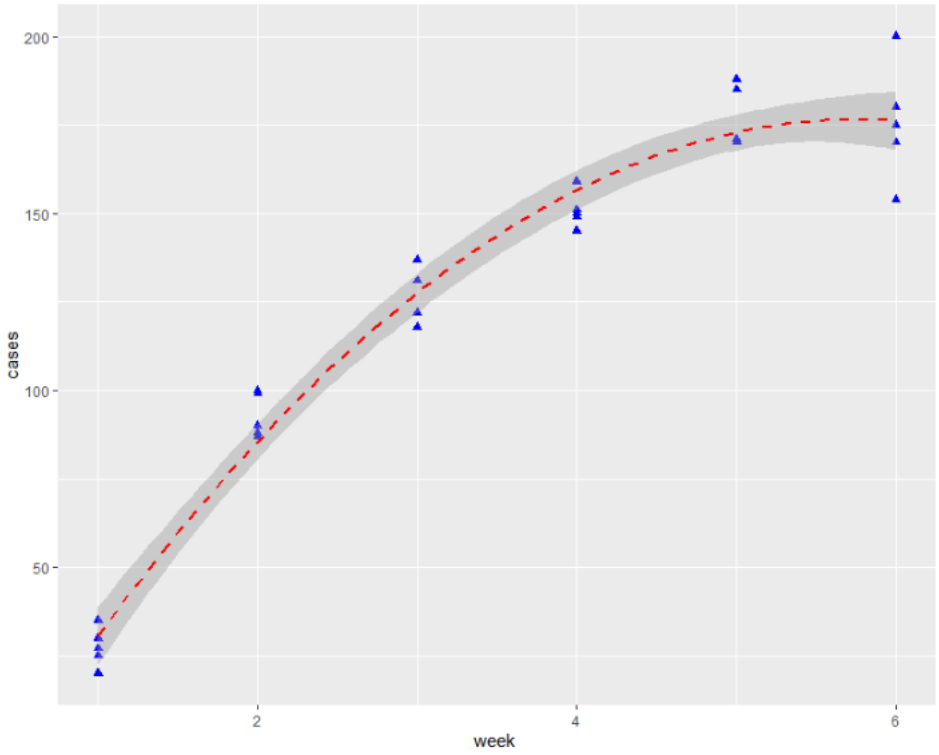

Figure 4: Quadratic Polynomial and 95% confidence bands

Figure 4 is a ggplot of the quadratic model and 95% confidence bands. This model gives a reasonable representation of the association between *cases* and *week* although a few points lie outside the confidence bands. Notice that the confidence bands get slightly wider towards week 6.

Question 2 **[50 marks]**

(a) Produce and interpret a pairs plot. What does the plot suggest as an appropriate model? Explain your response. [10 marks]

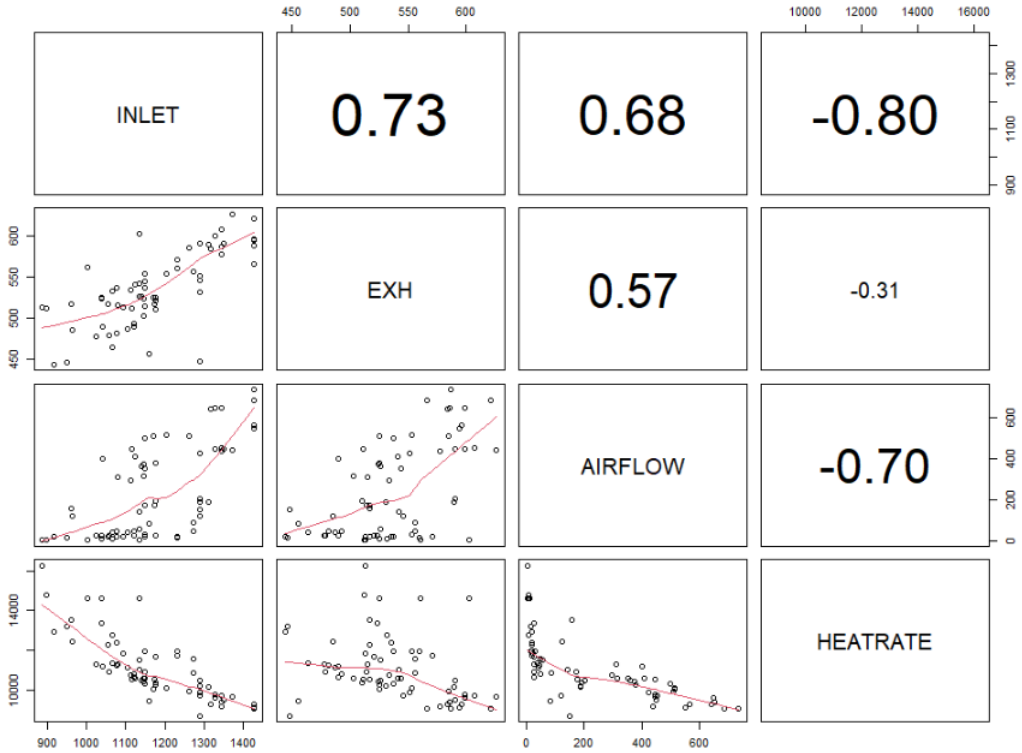

Figure 5: Pairs plot of the gas turbine data

From the pairs plot (Figure 5), it can be seen that the response variable `HEATRATE` has a strong negative correlation with `INLET` and `AIRFLOW`, $r = -0.80$ and $r = -0.70$ respectively. Also, there appears to be a weak negative correlation between `HEATRATE` and `EXH`, $r = -0.31$.

It is also evident from the plot that the explanatory variables are highly correlated. The predictor `INLET` has a strong positive correlation with both `EXH` ($r = 0.73$) and `AIRFLOW` ($r = 0.68$). In addition, `AIRFLOW` also has a moderate positive association with `EXH` ($r = 0.57$). This suggests that multicollinearity may exist when we use all 3 predictors. Therefore, using a subset of these 3 measurements maybe sufficient to model the variability in `HEATRATE`. Correlation matrix is given in Table 6.

Table 6: Correlation matrix for the gas turbine data, with corresponding p-values

	INLET	EXH	AIRFLOW	HEATRATE
INLET	1.000	2.85e-12	2.31e-10	4.44e-16
EXH	0.728	1.00e+00	5.82e-07	9.59e-03
AIRFLOW	0.681	5.67e-01	1.00e+00	3.28e-11
HEATRATE	-0.801	-3.14e-01	-7.03e-01	1.00e+00

(b) Fit a *main effects* model. Produce relevant outputs that will allow you to check the four indicators of multicollinearity. Summarise your findings. [15 marks]

The summary table for the main effects model with all 3 predictors is given in Table 7. Consider the four indicators of multicollinearity for the main effects model.

Table 7: Summary table for the main effects model.

```
Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept) 13170.666   1049.618   12.55    < 2e-16
INLET         -11.627      0.875  -13.28    < 2e-16
EXH            22.736      2.423    9.38    1.4e-13
AIRFLOW        -2.655      0.441   -6.02    9.9e-08

Residual standard error: 589 on 63 degrees of freedom
Multiple R-squared:  0.87,Adjusted R-squared:  0.863
F-statistic:  140 on 3 and 63 DF,  p-value: <2e-16
```

1. **Significant correlations between pairs of independent variables:**
   Yes, from Table 6, there are 3 pairs that are significantly correlated:
   - INLET & EXH ($r = 0.73$, p-value = 2.85e - 12)
   - INLET & AIRFLOW ($r = 0.68$, p-value = 2.31e-10)
   - EXH & AIRFLOW ($r = 0.57$, p-value = 5.82e-0.7)

   This suggests multicollinearity may be an issue.

2. **Non-significant t-tests for all (or nearly all) of the coefficients β_i when the F-test for model adequacy is significant:**
   From Table 7, each of the predictors is statistically significant, given that the other two predictors have been fitted, with p-value $< 2e-16$, 1.4e-13 and 9.9e-08, respectively.
   Thus, these results do not suggest multicollinearity being an issue because for indicating multicollinearity, we would expect to see all or nearly all parameters to be non-significant.

3. **Opposite signs from what is expected for regression coefficients**
   According to the pairs plot in Figure 5, there is a negative correlation between HEATRATE and EXH ($r = -0.31$). We would expect the coefficient for EXH to be negative, but looking at the summary table in Table 7, this coefficient is positive ($\beta_{EXH} = 22.7$) which is opposite of what is expected. So this is an indicator of multicollinearity.

4. **β's with VIF > 10**
   Table 8 shows the variation inflation factor (VIF) for the three predictors. The VIF for INLET is 2.75, for EXH is 2.17 and for AIRFLOW is 1.90. None of these VIF values is greater than 10, so we don't see any indication of multicollinearity based on VIF.

Table 8: Summary table for the main effects model.

```
> vif(mod1)
INLET   EXH    AIRFLOW
2.75    2.17    1.90
```

Conclusion: We checked the four indications of multicollinearity above, only indicators 1 and 3 appear to be an issue here. However, since all VIF values are very small(less than 5), it is likely that multicollinearity is not a big issue in this model.

(c) Run *forward stepwise* model selection with the "upper" model containing all possible interactions. Include all relevant outputs and the summary table for the final model. [15 marks]

Table 9: Outputs from stepwise regression

```
Start:  AIC=989
HEATRATE ~ 1

Df Sum of Sq      RSS AIC
+ INLET    1  1.08e+08 6.03e+07 923
+ AIRFLOW  1  8.30e+07 8.49e+07 946
+ EXH      1  1.66e+07 1.51e+08 984
<none>                1.68e+08 989

Step:  AIC=923
HEATRATE ~ INLET

Df Sum of Sq      RSS AIC
+ EXH      1  25831694 34463802 887
+ AIRFLOW  1   7818472 52477024 915
<none>                60295496 923

......
Step:  AIC=839
HEATRATE ~ INLET + EXH + AIRFLOW + INLET:AIRFLOW

Df Sum of Sq      RSS AIC
+ EXH:AIRFLOW  1   4806495 11021838 817
+ INLET:EXH    1   1213545 14614788 836
<none>                     15828333 839

Step:  AIC=817
HEATRATE ~ INLET + EXH + AIRFLOW + INLET:AIRFLOW + EXH:AIRFLOW

Df Sum of Sq     RSS AIC
<none>                  1.1e+07 817
+ INLET:EXH  1     20129 1.1e+07 819
```

Table 10: Summary table for the final model

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.39e+04   1.04e+03   13.35  < 2e-16
INLET         -1.51e+01   7.77e-01  -19.47  < 2e-16
EXH            2.88e+01   2.30e+00   12.52  < 2e-16
AIRFLOW       -6.89e-01   3.63e+00   -0.19     0.85
INLET:AIRFLOW  2.28e-02   3.00e-03    7.59  2.2e-10
EXH:AIRFLOW   -5.43e-02   1.05e-02   -5.16  2.9e-06

Residual standard error: 425 on 61 degrees of freedom
Multiple R-squared:  0.934,Adjusted R-squared:  0.929
F-statistic:  174 on 5 and 61 DF,  p-value: <2e-16
```

Using stepwise variable selection we find that the preferred model (AIC = 817) is:

$$E(HEATRATE) = 13900 - 15.1INLET + 28.8EXH - 0.689AIRFLOW$$
$$+0.023\ INLET : AIRFLOW - 0.054\ EXH : AIRFLOW$$

Selected steps in the variable screening process are given in Table 9. The AIC for the null model is 989 whereas the AIC for the final model is 817. Both of the interaction terms INLET:AIFLOW, and EXH:AIRFLOW are useful. The interaction INLET:EXH does not add significant additional information about HEATRATE.

The summary table for the final model is given in Table 10. From Table 10, all terms are significant, except AIRFLOW. Since AIRFLOW is in the interaction terms, we have to keep it in the model. The adjusted R^2 for the final model is 0.93 which shows a big improvement compared to the main effects model, which had adjusted R^2 of 0.86 (Table 7).

(d) Write a concise (one to two paragraphs), informative conclusion based on your analysis and results. [10 marks]

The exploratory analysis suggested that all three independent variables may be good predictors for modeling HEATRATE. After fitting the main effects model, it was found that all 3 predictors are significant. The main effects model explains about 86% of the varibility in the HEATRATE.

We observed some indications of multicollinearity among the independent variables, however, the VIFs for all parameters were less than 5 which suggest multicollinearity is not likely to be a big issue.

Stepwise regression with forward selection incorporating all predictors and their interactions produced:

$$E(HEATRATE) = 13900 - 15.1INLET + 28.8EXH - 0.689AIRFLOW$$
$$+0.023\ INLET : AIRFLOW - 0.054\ EXH : AIRFLOW$$

We can see there are two significant interaction terms (INLET*AIRFLOW and EXH*AIRFLOW) in the model. This suggests that the effects of inlet temperature (INLET) and exhaust gas temperature (EXH) on heat rate depend on the value of air mass flow rate (AIRFLOW). This model explains about 93% of the variability observed in the response. This model explains about 93% of the variability observed in the response. This model is useful (F-statistic = 174 on 5 and 61 df, p-value ¡2e-16), subject to checking model assumptions.

Appendix: R code

Question 1

```
## Q1 covid-19 ###############################################
## Fit polynomial models: quadratic, cubic & quartic ######
rm(list=ls())   # remove all variables in the working space
options(digits=3, show.signif.stars=F)
Virus <- read.table("Virus.txt",header=T)
# scatterplot
plot(cases ~ week, data=Virus)
VM1<-lm(cases ~ week, data=Virus)
summary(VM1)
VM2<-lm(cases ~ week + I(week^2), data=Virus)
summary(VM2)
VM3<-lm(cases ~ week + I(week^2)+I(week^3), data=Virus)
summary(VM3)
anova(VM3)
VM4<-lm(cases ~ week + I(week^2)+I(week^3) + I(week^4), data=Virus)
summary(VM4)
anova(VM4)
#Diagnostic Plots for quadratic model
par(mfrow=c(2,2))
plot(VM2,which=1:4,add.smooth = F)
dev.off()
# Shapiro test
shapiro.test(VM2$residuals)
# Cook's distance: observation 28
pf(0.5,3,27)

# Residuals vs Leverage plot
dev.off()
plot(VM2,which=5)
##
library(ggplot2)
ggplot( data=Virus,aes(x=week, y=cases)) +
geom_point(pch=17, color="blue", size=2) +
geom_smooth( method = "lm", formula = y ~ poly(x,2), color="red", linetype=2) +
labs(title="Quadratic polynomial with 95% confidence bands",
```

```
x="week ", y="cases")
```

Question 2:

```
rm(list=ls())
options(digits=3, show.signif.stars=F)
gas <- read.table("GASTURBINE.txt",header=T)
source("Rfunctions.r")
pairs(gas[,2:5],lower.panel = panel.smooth, upper.panel = panel.cor)
# correlation matrix
cor.prob(gas[,2:5])

## main effects model
mod1 <- lm(HEATRATE ~ INLET + EXH + AIRFLOW,data=gas)
summary(mod1)
# obtain VIF
install.packages("car")
library(car)
vif(mod1)

# stepwise regression
formU <- formula(~INLET*EXH*AIRFLOW)
formL <- formula (~ 1)
start.mod <- lm(HEATRATE~1, data=gas)
step.model <- step(start.mod, direction = "forward",scope = list(lower = formL,
                              upper = formU))
summary(step.model)
```