1. (a) The figure below shows histograms and Q-Q plots for both the \( SO_2 \) measurements (top row) and \( y = \log(SO_2) \) (bottom row). The \( SO_2 \) measurements do not appear to be normally distributed given the shape of the histogram. It does, however, seem reasonable to assume that \( y \) is normally distributed. These figures can be constructed using the following R code:

```r
par(mfrow=c(2,2))
hist(air.data[,1],main="Histogram of SO_2",xlab="SO_2",prob=T)
norm(air.data[,1],main="Q-Q Plot -- SO_2")
norm(y,main="Q-Q Plot -- log(SO_2)")
y <- log(air.data[,1])
hist(y,main="Histogram of log(SO_2)",xlab="y",prob=T)
norm(y,main="Q-Q Plot -- log(SO_2)")
norm(y)
```

(b) The code below can be used to construct a \textit{pairs plot} of \( y \) and the explanatory variables.
X <- as.matrix(air.data[, -1])
pairs(cbind(y, X))

(c) NOTE: There is not one correct answer to this question. Here is one reasonable model
based on the pairs plot in part 1(b).

The pairs plot shows that there is strong positive association between manufacturing and
population size (correlation=0.95). As a result, in order to avoid problems with multi-
collinearity, population size is not used in the regression model. In addition, the remaining
explanatory variables are centered (i.e., the column means are subtracted from the observ-
ations). The model

$$E[\log(y_i)] = \beta_0 + \beta_1 c\text{Neg.Temp} + \beta_2 c\text{Manuf} + \beta_3 c\text{Wind} + \beta_4 c\text{Precip} + \beta_5 c\text{Days}$$

is fit to the data. The following table provides a summary of the fitted model:

| Estimate | Std. Error | t value | Pr(>|t|)   |
|----------|------------|---------|------------|
| (Intercept) | 3.1530036 | 0.0712867 | 44.230 < 2e-16 *** |
| X.cNeg.Temp | 0.0689735 | 0.0184040 | 3.748 0.000643 *** |
| X.cManuf | 0.0005550 | 0.0001331 | 4.170 0.000191 *** |
| X.cWind | -0.1797441 | 0.0562091 | -3.198 0.002936 ** |
| X.cPrecip | 0.0194181 | 0.0112308 | 1.729 0.092620 . |
| X.cDays | 0.0003444 | 0.0050524 | 0.068 0.946037 . |
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 .’ 0.1 ‘ ’ 1

Residual standard error: 0.4565 on 35 degrees of freedom
Multiple R-Squared: 0.6304,    Adjusted R-squared: 0.5776
F-statistic: 11.94 on 5 and 35 DF,  p-value: 8.74e-07

The positive linear relationships between both $c_{Neg.Temp}$ and $c_{Manuf}$ and $y$ are highly significant. $c_{Wind}$ is negatively related to $y$ at the 0.01 level, and $c_{Precip}$ is positively related to $y$ at the 0.05 level. Given the $R^2$ value, the explanatory variables are explaining around 63 percent of the variation in $y$.

The following R code was used to fit the linear regression model:

```r
X.c <- X[, -3] - matrix(rep(apply(X[, -3], 2, mean), N), N, dim(X[, -3])[2], byrow=T)
lm.fit <- lm(y ~ X.c)
summary(lm.fit)
```

(d) The figure below shows a scatterplot of the standardized residuals versus the fitted values. The dashed lines denote 95 percent confidence intervals for the residuals. Since the points are random randomly scattered on either side of the solid line at zero and are mostly contained between the two dashed lines, it appears that the assumptions of the linear regression model hold. The following code was used to construct this residual plot.

```r
plot(lm.fit$fitted.values, lm.fit$residuals, ylim=c(-1.2, 1.2),
     xlab="Fitted Values", ylab="Residuals")
abline(h=0)
abline(h=2*summary(lm.fit)$sigma, lty=2)
abline(h=-2*summary(lm.fit)$sigma, lty=2)
```
2. Since \( g(x) = c_1 x^{1-b} + c_2, \) \( g'(x) = c_1 (1-b) x^{-b}. \) Therefore,

\[
\sigma_X^2 \approx (g'(\mu_X)\sigma(\mu_X))^2 \\
= (a(\mu_X)^b \cdot c_1 (1-b)(\mu_X)^{-b})^2 \\
= (ac_1 (1-b))^2 \\
= \text{constant},
\]

since \( a, b, \) and \( c_1 \) are constants.

3. Since \( Y_i \overset{i.i.d}{\sim} N(\mu, \sigma^2) \) for \( i = 1, \ldots, n, \) the maximum likelihood estimates for \( \mu \) and \( \sigma \) are

\[
\hat{\mu} = \bar{y} \quad \text{and} \quad \hat{\sigma} = \left[ \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\mu})^2 \right]^{1/2}.
\]

Plugging in \( y_i = \log(x_i) \) shows that the MLEs of \( \hat{\mu} \) and \( \hat{\sigma} \), are

\[
\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \log(x_i) \quad \text{and} \quad \hat{\sigma} = \left[ \frac{1}{n} \sum_{i=1}^{n} (\log(x_i) - \hat{\mu})^2 \right]^{1/2}.
\]

4. (a) \( \hat{\mu} = 2.10 \) and \( \hat{\sigma} = 0.46 \)

(b) \( \hat{\mu} = 1.96 \) and \( \hat{\sigma} = 0.48 \)

(c) The method in (a) overestimates \( \mu \) and underestimates \( \sigma \). The method in part(b) overestimates \( \mu \) and underestimates \( \sigma \), but is not as bad as the estimates from part (a).