UNIVERSITY OF TEXAS AT SAN ANTONIO

# Linear Regression for Air Pollution Data

Liang Jing

April 2008

# 1 Goal

The increasing health problems caused by traffic-related air pollution have caught more and more attention nowadays. As a result, analysis and prediction of the air quality are widely studied. Several methodologies, both deterministic and statistical, have been proposed. In this project we use the linear model to detect the relationship between the concentration of an air pollutant at a specific site and traffic volume as well as other meteorological variables. The procedure of model building and validating is demonstrated along with a variety of coefficient tests.

# 2 Introduction of Data

The data are a sub-sample of 500 observations from a data set collected by the Norwegian Public Roads Administration. The response variable consist of hourly values of the logarithm of the concentration of NO2 (particles), measured at Alnabru in Oslo, Norway, between October 2001 and August 2003. The predictor variables are the logarithm of the number of cars per hour, temperature 2 meter above ground (degree C), wind speed (meters/second), temperature difference between 25 and 2 meters above ground (degree C), wind direction (degrees between 0 and 360), hour of day and day number from October 1. 2001.

**Size** : $n = 500$

**Response Variable** $Y$ : concentration of $NO_2$

**Predictors** : $k = 7$

- $x_1$: number of cars per hour
- $x_2$: temperature 2 meter above ground (degree C)
- $x_3$: wind speed (meters/second)
- $x_4$: temperature difference between 25 and 2 meters (degree C)
- $x_5$: wind direction (degrees between 0 and 360)
- $x_6$: hour of day
- $x_7$: day number

Table 3.1: Normality test for response variable

| Tests for Normality | | |
|---|---|---|
| Test | Statistic | p-value |
| Shapiro-Wilk | W 0.995731 | Pr < W 0.6039 |

# 3 Methodology & Results

## 3.1 Normality Test for Response Variable

$H_0$ : y follows normal distribution

$H_1$ : y doesn't follow normal distribution

Shapiro-Wilk test, proposed by Samuel Shapiro and Martin Wilk 1965, was conducted to test the null hypothesis. The test statistic is

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{3.1}$$

where $x_{(i)}$ is the order statistic, $\bar{x}$ is the sample mean and the constant $a_i$ is given by

$$(a_1, ..., a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{1/2}} \tag{3.2}$$

where $m = (m_1, ..., m_n)^T$ are the expected values of the order statistics of i.i.d random variables sampled from standard normal distribution and $V$ is the covariance matrix of those order statistics. The test result is summarized in table 3.1.

Thus, with p-value is 0.6039 we fail to reject Null hypothesis which means response variable follows normal distribution. More evidences are shown in the histogram plot and normal percentile plot listed below.

## 3.2 Correlation Analysis for Predictors

The correlation coefficient between two random variables $X$ and $Y$ is defined as:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_x)(Y - \mu_Y)]}{\sigma_X \sigma_Y} \tag{3.3}$$

Figure 3.1: Histogram for response variable

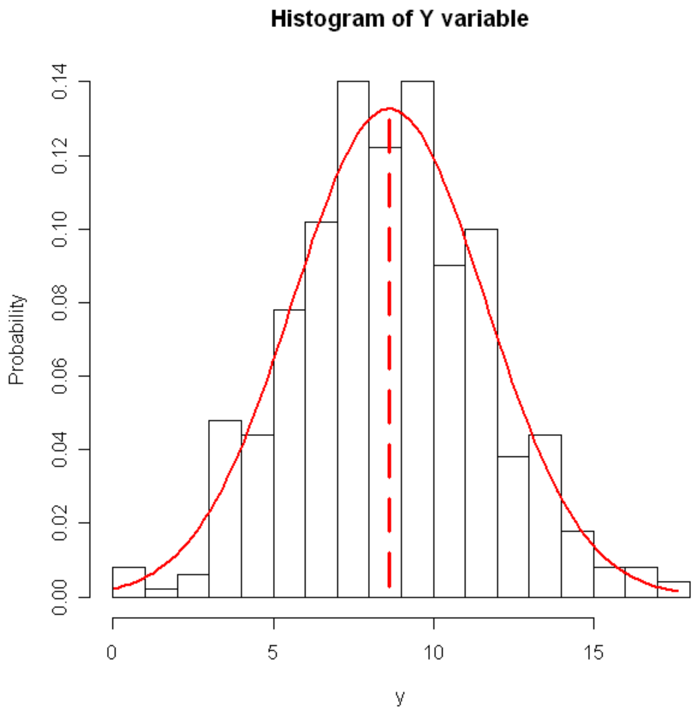

Figure 3.2: QQ-plot for response variable

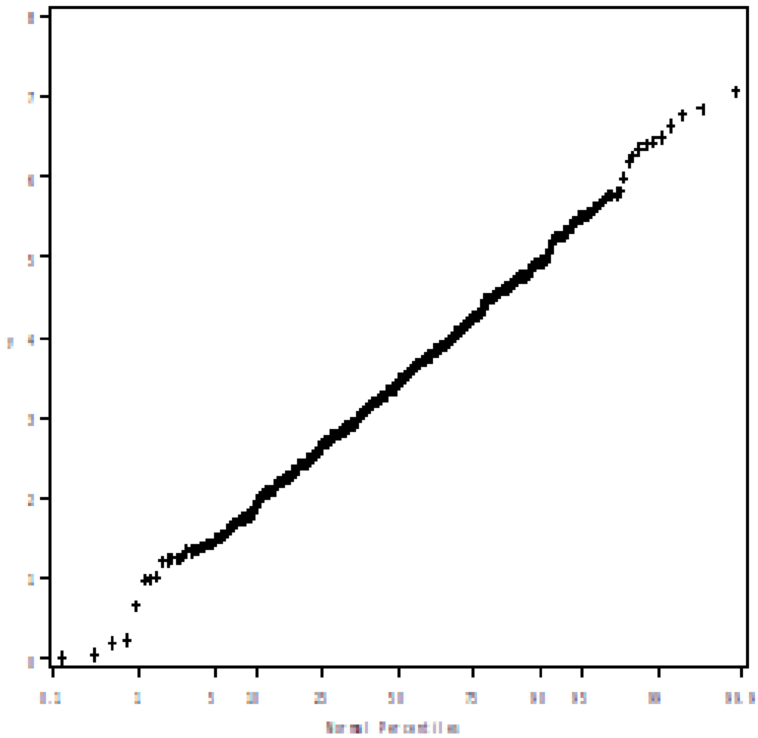

Table 3.2: Correlation test for predictors

| | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ |
|---|---|---|---|---|---|---|---|
| $X_1$ | 1.00000 | 0.07848 | 0.05951 | 0.00434 | -0.04960 | -0.04477 | 0.01040 |
| $X_2$ | | 1.00000 | -0.10547 | -0.04630 | -0.01578 | -0.02564 | 0.02230 |
| $X_3$ | | | 1.00000 | -0.04205 | 0.00242 | 0.00492 | -0.10630 |
| $X_4$ | | | | 1.00000 | -0.01460 | -0.01250 | -0.01640 |
| $X_5$ | | | | | 1.00000 | 0.05492 | -0.06042 |
| $X_6$ | | | | | | 1.00000 | 0.02780 |
| $X_7$ | | | | | | | 1.00000 |

The correlation indicates the degree of linear dependence between these two variables: it is 1 in the case of an increasing linear relationship; -1 in the case of a decreasing linear relationship; and the values in between for all other cases. The closer the coefficient is to either -1 or 1, the stronger the correlation between the variables.

In table 3.2, we can see that all the values are small enough for us to say none of the predictor pair is remarkably correlated. So we will keep all of them in the initial model.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 \tag{3.4}$$

## 3.3 Prescreen of X Variables

Test of significance for all the X variables is performed for the initial model. The result shown in figure 3.3 suggests that $X_5$ and $X_7$ are not significant due to large p-values.

After removing the most insignificant predictor $X_7$, the test is conducted again and the result is shown in figure 3.4 which suggests removal of $X_6$.

## 3.4 Test of $\beta_{00}$

To test whether any of the predictors has impact on response variable, the following hypotheses are tested.

$H_0 \ : \beta_{00} = 0$

Figure 3.3: QQ-plot for response variable

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| x1 | 1 | 353.4795783 | 353.4795783 | 1049.35 | <.0001 |
| x2 | 1 | 31.4967990 | 31.4967990 | 93.50 | <.0001 |
| x3 | 1 | 81.7733151 | 81.7733151 | 242.76 | <.0001 |
| x4 | 1 | 42.5806095 | 42.5806095 | 126.41 | <.0001 |
| x5 | 1 | 1.1035829 | 1.1035829 | 3.28 | 0.0709 |
| x6 | 1 | 64.9333130 | 64.9333130 | 192.76 | <.0001 |
| x7 | 1 | 0.0407115 | 0.0407115 | 0.12 | 0.7283 |

Figure 3.4: QQ-plot for response variable

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| x1 | 1 | 353.4441610 | 353.4441610 | 1051.12 | <.0001 |
| x2 | 1 | 31.5189400 | 31.5189400 | 93.74 | <.0001 |
| x3 | 1 | 82.3191957 | 82.3191957 | 244.81 | <.0001 |
| x4 | 1 | 42.6558151 | 42.6558151 | 126.86 | <.0001 |
| x5 | 1 | 1.1341596 | 1.1341596 | 3.37 | 0.0669 |
| x6 | 1 | 65.1078688 | 65.1078688 | 193.63 | <.0001 |

$H_1 \ : \beta_{00} \neq 0$

where $\beta_{00} = 0 = (\beta_1, ...\beta_k)$ is the coefficient parameters of predictors.

When $H_0$ is true the F test statistic will follow a central F-distribution. Thus, $H_0$ will be rejected if $F > F_{\alpha,n-k-1}$ where $F_{\alpha,n-k-1}$ is the upper $\alpha$ percentage point of the (central) F distribution. Alternatively, p-value, the tail area of the central F distribution beyond the calculated F test statistic, can be calculated. A p-value whose value is less than $\alpha$ is equivalent to $F > F_{\alpha,n-k-1}$.

By using SAS IML procedure, the result is

$$p = 0.0007662 < 0.05 \tag{3.5}$$

which suggests rejection of $H_0 : \beta_{00} = 0$.

## 3.5 Bonferroni Test

Typical inferences are performed using the 95% confidence level or 5% significance level. In either case, the comparison-wise error rate (CER) is 5%. The statement "$H_0$" refers to a "null hypothesis" concerning a parameter or parameters of interest, which we shall always assume to be a strict equality. Suppose that we have defined a family of inferences (tests or intervals) containing k elements. The Family-wise Error Rate (FWE) is the probability of at least one erroneous inference. This is defined for simultaneous confidence intervals as

FWE = P (at least one interval is incorrect)
= 1- P(all intervals are correct)

To simplify the presentation of multiple tests, the p-values are often displayed as adjusted p-values. By definition, the adjusted p-values for any hypothesis is equal the smallest FWE at which the hypothesis would be rejected. Therefore, adjusted p-values are readily interpretable as evidence against the corresponding null hypotheses, when all tests are considered as a family. To make a decision on any hypothesis $H_{0j}$, we can simply compare its corresponding adjusted p-values with the desired FWE level. The Bonferroni procedure rejects any $H_{0j}$ whose corresponding p-value, $p_j$ , is less than or equal to $\alpha/k$. This is equivalent to rejecting any $H_{0j}$ for

Table 3.3: Multi-Bonferroni test for predictors

| | $p_1$ | $p_2$ | $p_3$ | $p_4$ | $p_6$ |
|---|---|---|---|---|---|
| Initial model | 1 | $5.2527 \times 10^{-8}$ | $1.623 \times 10^{-44}$ | 1 | $1.701 \times 10^{-36}$ |
| $X_1$ removed | | 0.0001765 | $2.8 \times 10^{-13}$ | 1 | $1.672 \times 10^{-16}$ |
| $X_4$ removed | | 0.0000652 | $2.055 \times 10^{-13}$ | | $8.646 \times 10^{-16}$ |

which $kp_j$ is less than or equal to $\alpha$. Thus $kp_j$ is the Bonferroni adjusted p-value for $H_{0j}$. We require any p-value to be less than 1, and therefore define Bonferroni adjusted p-value for hypothesis $H_0$ more specifically as,

$$p_j = \begin{cases} k \cdot prob(t_i, k) \\ 1, & \text{if } p_j \geq 1 \end{cases} \tag{3.6}$$

$H_0$ : $X_j$ is insignificant in the model if $p_j > \alpha$

$H_1$ : $X_j$ is not insignificant in the model if $p_j \leq \alpha$

The rational for this method is the well known Bonferroni inequality.

When it comes to our project, $X_1, X_2, X_3, X_4, X_6$ are initially in the model. We use repeated Bonferroni test to remove any insignificant variable. The result of multi-Bonferroni test is shown in table 3.3.

In the table, we can see for the model with $k = 5$, $p_1$ and $p_4$ are the largest ones. Either of them can be removed at first. We remove $X_1$ for convenience, then we run the Bonferroni test again with the number of variable is reduced to $k = 4$. ItâĂŹs not difficult to find that the $p_4$ are still the largest value. So we remove $X_4$ in the second step. After removing $X_4$, the Bonferroni test for the rest of the three variables are far less than .05, which means that $X_2$,$X_3$,$X_6$ are all significant in this stage.

To validate this assumption, we can do the t-test of the model consisting of $X_2$,$X_3$,$X_6$ only. The output of *SAS GLM* in figure 3.5 shows that they are truly significant within the model, since their p-values are all less than 0.05 which means the null hypothesis that the corresponding predictors are insignificant in the model should be rejected.

Furthermore, the estimated parameters from SAS output in figure 3.6 suggests the following model,

$$Y = 5.132332387 - 0.030364128X_2 - 0.205725423X_3 - 0.058798271X_6 \tag{3.7}$$

Figure 3.5: Significance test for $X_2$,$X_3$,$X_6$

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| x2 | 1 | 19.36174588 | 19.36174588 | 17.02 | <.0001 |
| x3 | 1 | 65.93453225 | 65.93453225 | 57.95 | <.0001 |
| x6 | 1 | 79.78175934 | 79.78175934 | 70.11 | <.0001 |

Figure 3.6: Estimated parameters for $X_2$,$X_3$,$X_6$

| Parameter | Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|
| Intercept | 5.132332387 | 0.15890613 | 32.30 | <.0001 |
| x2 | -0.030364128 | 0.00736098 | -4.13 | <.0001 |
| x3 | -0.205725423 | 0.02702581 | -7.61 | <.0001 |
| x6 | -0.058798271 | 0.00702199 | -8.37 | <.0001 |

## 3.6 ANOVA Table for the Final Model

*SAS IML* procedure is used to calculate the ANOVA table shown in table 3.4. The counterpart of SAS GLM out put is shown in figure 3.7. Compare these two outputs, we can see that they are exactly the same which means the procedure based on the SAS IML is correct. (The code is included in Appendix.)

Table 3.4: ANOVA table

| | Degree of freedom | Sum Square | Mean Square | F statistic |
|---|---|---|---|---|
| Regression 3 | 157.1688400 | 52.3896133 | 46.04 | |
| Residual Error | 496 | 564.3859382 | 1.1378749 | |
| Total | 499 | 721.5547783 | | |

Figure 3.7: ANOVA from SAS GLM

The GLM Procedure

Dependent Variable: y y

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 3 | 157.1688400 | 52.3896133 | 46.04 | <.0001 |
| Error | 496 | 564.3859382 | 1.1378749 | | |
| Corrected Total | 499 | 721.5547783 | | | |

## 3.7 Confidence Interval of $\beta_j$

For the final model, because $\hat{\beta}_j \sim N(\beta_j, var(\beta)_{jj})$, $\frac{\hat{\beta}_j - \beta_j}{var(\beta)_{jj}}$ has a t-distribution with $n-k-1$ degree of freedom, where $var(\beta)_{jj}$ is the $i^{th}$ diagonal entry of the covariance matrix of $\beta$. So the 95% confidence interval for $\beta_j$ is

$$[\hat{\beta}_j - t_{\alpha/2,n-k-1} var(\beta)_{jj}, \hat{\beta}_j + t_{\alpha/2,n-k-1} var(\beta)_{jj}]. \tag{3.8}$$

The result from SAS output for $\beta_2, \beta_3, \beta_6$ is: [-0.044827, -0.015902], [-0.258825, -0.152626], and [-0.072595 , -0.045002] respectively.

## 3.8 Conclusion

Response variable approximately follows normal distribution. Though many predictors are significant, under Bonferroni test response variable is most closely related to three predictors only. The final model with estimated parameter is

$$\begin{aligned} \text{Concentration of NO2} = 5.132332387 &- 0.030364128 \times \text{temp above ground} \\ &- 0.205725423 \times \text{wind speed} \\ &- 0.058798271 \times \text{hour of day} \end{aligned} \tag{3.9}$$

This model suggests that the concentration of $NO_2$ is negative proportional to the temperature above ground and wind speed which means: when wind speed is increasing, the density is decreasing; and the higher the temperature is, the bigger the volume of $NO_2$ inflates and the lower the density is as a result.

# 4 APPENDIX

SAS code

```
/* Normality Test for Y variable */
proc univariate data=no2 normal;
var  y; probplot y; run;
/* Correlation of X variables */
proc corr data=no2;
var x1-x7; run;
/* GLM Analysis */
proc glm data=no2;
model y=x1-x6 ; run;
/* IML Analysis */
proc iml;
use no2;
read  all var {x1 x2 x3 x4 x6}  into x;
read  all var {y}  into y;
/*print x y; run;*/
n=nrow(x);            /* Number of observations;*/
k=ncol(x);            /* Number of parameters including the intercept; */
j=j(n,1,1);
x10=j||x;
          /* Display the design matrix  */
cov_x=inv(x10`*x10);
xpy=x10`*y;           /* The vector (X'X)^-1*Y  ;*/
beta=cov_x*xpy;
PRINT beta;
          /* The estimated regression parameters;*/
/*Table 1 ANOVA for fitting regression*/
/* The fitted values, the residuals, SSE, and MSE ;*/
ssr=beta`*x10`*y;      /*SSR= sum of square of residual;*/
dfreq=k+1;             /*Degree of freedom of SSR;*/
```

```
print ssr dfreq;
msr=ssr/dfreq;         /*Mean square of residual;*/
sse=y`*y-beta`*x10`*y;  /* SSE = Sum of squares of residuals;*/
dferr=n-k-1;           /* Degrees of freedom of SSE;*/
mse=sse/dferr;        /* MSE = SSE/dferror;*/
print msr sse dferr mse;
sst=y`*y;             /* SST= sum square of total;*/
dftot=n;               /*Degrees of freedom of total;*/
fstat=msr/mse;         /* F-statistics; */
print sst dftot fstat;
/*Table 2 ANOVA*/
J=j(n,n,1);
beta00=beta[2:k+1];
xbar_t=j`*x/n;
x_b=(I(n)-J/n)*x;

SSRm=beta00`*x_b`*y;
MSRm=SSRm/k;
print SSRm MSRm;
fstatRm=MSRm/MSE;
print SSE MSE fstatRm;
SSTm=y`*(I(n)-J/n)*y;
print SSTm;
/*Table 3 ANOVA showing in the term mean*/
SSM=y`*J*y/n;
MSM=SSM/1;
print SSM MSM;
fstatM=MSM/MSE;
print SSE MSE fstatM fstatRm;
SST=y`*y;
print SST;
/*Bonferroni Test of beta_j*/
var_beta=MSE*cov_x;
print var_beta;
b_1=beta[2,1]/sqrt(var_beta[2,2]);
b_2=beta[3,1]/sqrt(var_beta[3,3]);
b_3=beta[4,1]/sqrt(var_beta[4,4]);
```

```
b_4=beta[5,1]/sqrt(var_beta[5,5]);
b_5=beta[6,1]/sqrt(var_beta[6,6]);
print b_1 b_2 b_3 b_3 b_4 b_5;
p1=5*probt(b_1,n-k-1);
p2=5*probt(b_2,n-k-1);
p3=5*probt(b_3,n-k-1);
p4=5*probt(b_4,n-k-1);
p5=5*probt(b_5,n-k-1);
print p1 p2 p3 p4 p5;
```