Simple Linear Regression and Correlation.

Corresponds to
Chapter 10
Tamhane and Dunlop

Slides prepared by Elizabeth Newton (MIT)
with some slides by
Jacqueline Telford (Johns Hopkins University)

Simple linear regression analysis estimates the relationship between two variables.

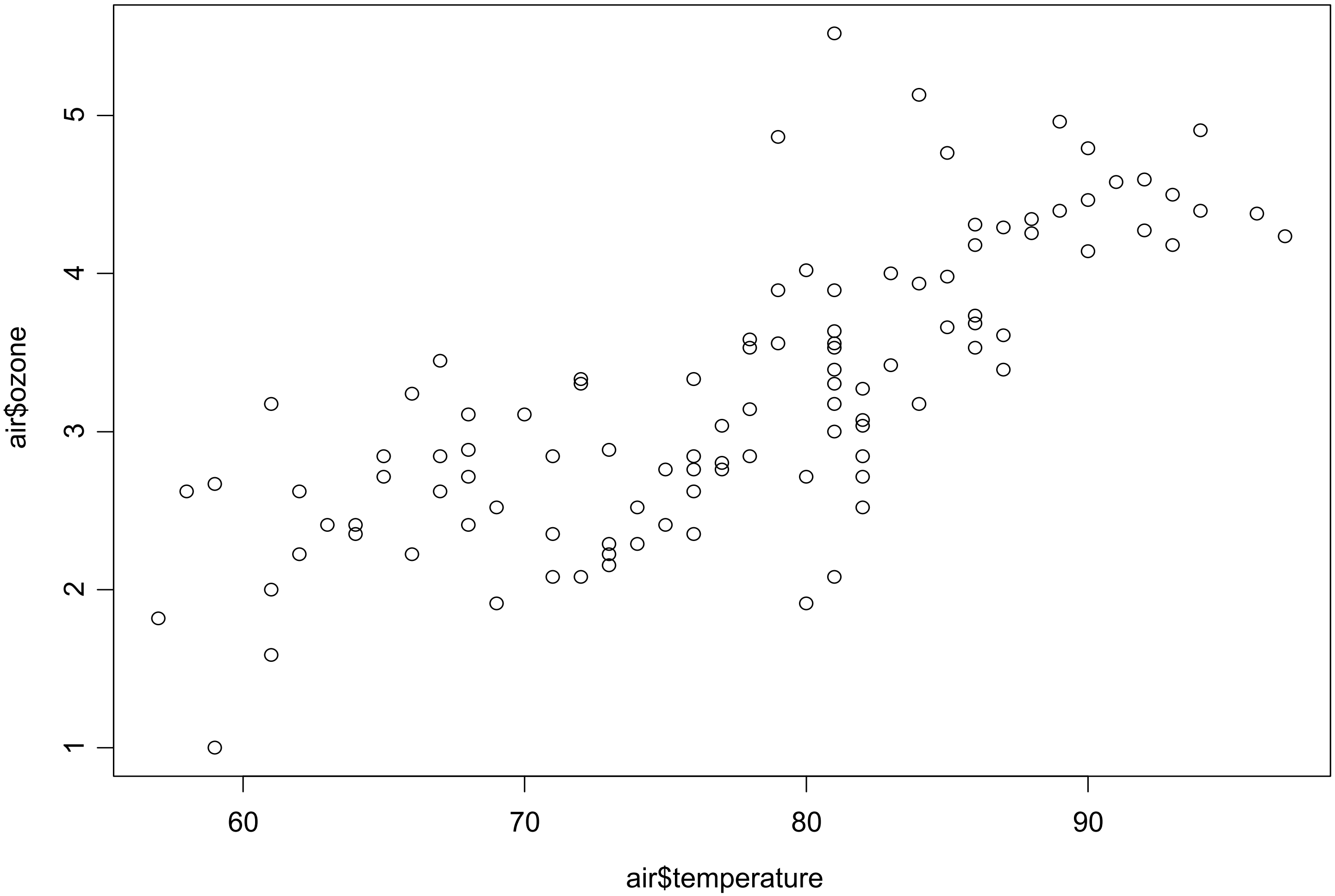
One of the variables is regarded as a **response** or **outcome variable (y).**

The other variable is regarded as **predictor** or **explanatory variable (x).**

Sometimes it is not clear which of two variable should be the response (e.g. height and weight). In this case, correlation analysis may be used.

Simple linear regression estimates relationships of the form $y = a + bx$.

Scatter plot of ozone concentration by temperature



This graph was created using S-PLUS(R) Software. S-PLUS(R) is a registered trademark of Insightful Corporation.

A Probabilistic Model for Simple Linear Regression

Let $x_1, x_2, ..., x_n$ be specific settings of the predictor variable.

Let $y_1, y_2, ..., y_n$ be the corresponding values of the response variable.

Assume that y_i is the observed value of a random variable (r.v.) Y_i, which depends X on according to the following model:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (i = 1, 2, \ldots, n)$$

Here ε_i is the random error with $E(\varepsilon_i)=0$ and $Var(\varepsilon_i)=\sigma^2$.

Thus, $E(Y_i) = \mu_i = \beta_0 + \beta_1 x_i$ (true regression line).

The x_i's usually are assumed to be fixed (not random variables).

A Probabilistic Model for Simple Linear Regression

See Figure 10.1, p. 348 and also see page 348 for the four assumptions of a simple linear regression model.

Least Square Line Mathematics (invented by Gauss)

Find the line, i.e., values of β_0 and β_1 that minimizes the sum of the squared deviations:

$$Q = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

How?

Solve for values of β_0 and β_1 for which

$$\frac{\partial Q}{\partial \beta_0} = 0 \text{ and } \frac{\partial Q}{\partial \beta_1} = 0$$

Finding Regression Coefficients

$$\frac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n}[y_i - (\beta_0 + \beta_1 x_i)]$$

$$\frac{\partial Q}{\partial \beta_1} = -2\sum_{i=1}^{n} x_i [y_i - (\beta_0 + \beta_1 x_i)]$$

Normal Equations

$$n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

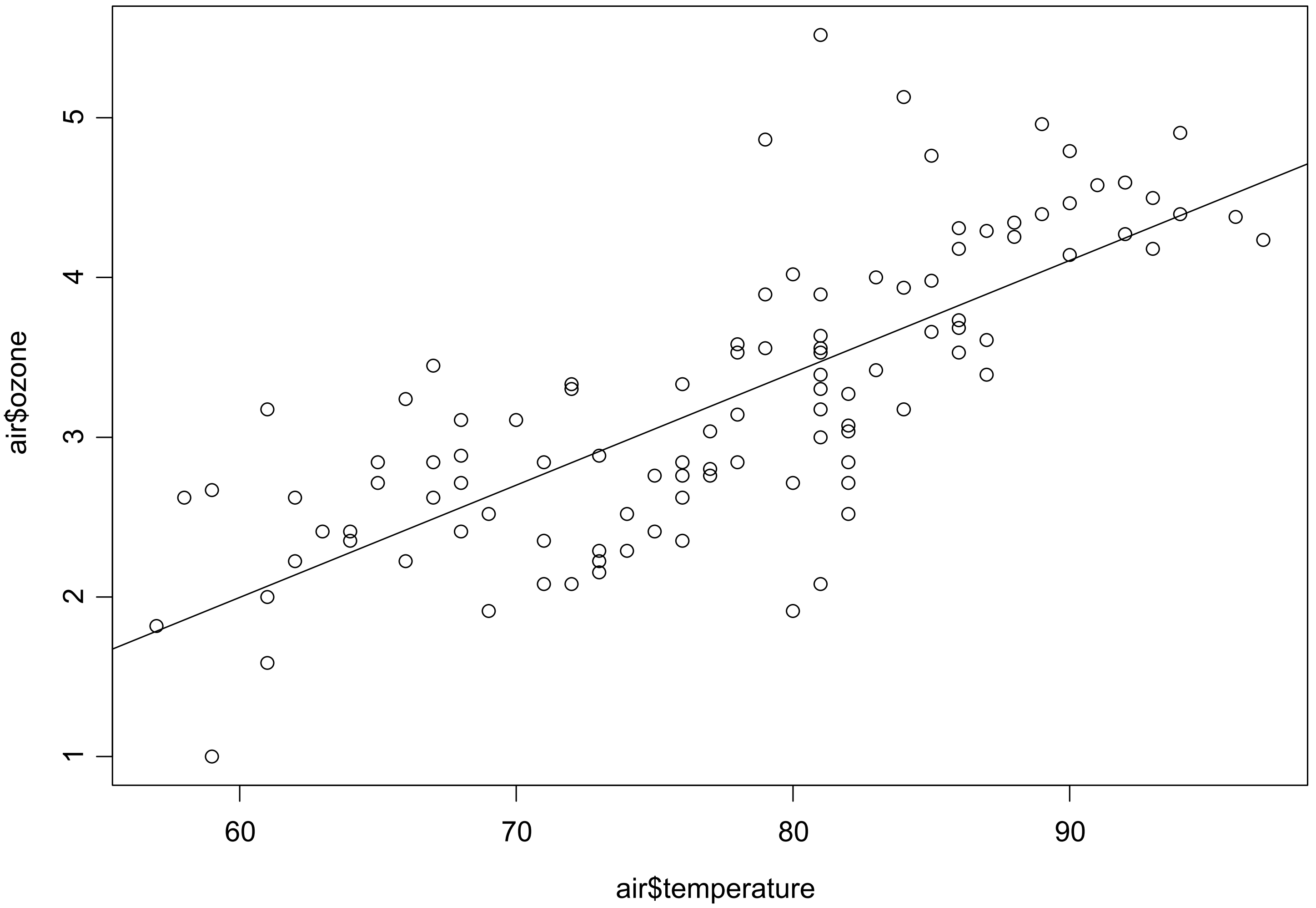
Solution to Normal Equations

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Note that least squares line goes through $(\bar{x}, \bar{y})$.

Fitted regression line



This graph was created using S-PLUS(R) Software. S-PLUS(R) is a registered trademark of Insightful Corporation.

Fitted values of y_i : $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, $i = 1, 2, \ldots, n$

Residuals : $e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$, $i = 1, 2, \ldots, n$

temperature	ozone	fitted	resid
67	3.45	2.49	0.96
72	3.30	2.84	0.46
74	2.29	2.98	-0.69
62	2.62	2.14	0.48
65	2.84	2.35	0.50

Matrix Approach to Simple Linear Regression (what your regression package is really doing)

The model: $y = X\beta + \varepsilon$

y is n by 1
X is n by 2
β is 2 by 1
ε is n by 1

$Y = X\beta + \varepsilon$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & x_4 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \varepsilon_3 \\ \varepsilon_4 \end{bmatrix}$$

Solution of linear equations

In linear algebra:

Find x which solves Ax=b.

In regression analysis:

Find β which solves Xβ=y

Why can't we do this?

Least Squares

$Q=(y-X\beta)'(y-X\beta)$

$= y'y - \beta'X'y - y'X\beta + \beta'X'X\beta$

$= y'y - 2\,\beta'X'y + \beta'X'X\beta$

$\partial Q/\,\partial\beta = -2X'y + 2X'X\beta$

$\partial Q/\,\partial\beta = 0 \rightarrow X'y = X'Xb$, where $b = \hat{\beta}$

Least Squares continued

For simple linear regression:

$$X'X = \begin{bmatrix} n & \sum \mathrm{x}_\mathrm{i} \\ \sum x_i & \sum \mathrm{x}_\mathrm{i}^2 \end{bmatrix}$$

$$\mathrm{X'y} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

Least Squares continued

$X'Xb = X'y$

$$\begin{bmatrix} n & \sum \mathrm{x_i} \\ \sum x_i & \sum \mathrm{x_i^2} \end{bmatrix} \mathrm{b} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

The Normal Equations as before

Least Squares continued

$X'Xb = X'y$

$b = (X'X)^{-1}X'y$ (if X has linearly independent columns)

Solution by QR decomposition

$X=QR$, Q orthonormal, R upper triangular and invertible

$b=(X'X)^{-1}X'y = (R'Q'QR)^{-1}R'Q'y$

$=(R'R)^{-1}R'Q'y = R^{-1}Q'y$

The Hat Matrix

$b=(X'X)^{-1} X'y$

$\hat{y}=Xb = X(X'X)^{-1}X'y =Hy$

H (n by n) is the Hat matrix

Takes y to $\hat{y}$

H is symmetric and idempotent HH=H

Diagonal elements of the hat matrix are useful in detecting influential observations.

Expected value of b

$$E(b) = E((X'X)^{-1}X'y]$$
$$= E[(X'X)^{-1}X'(X\beta+\varepsilon)]$$
$$= E[(X'X)^{-1}X'X\,\beta + (X'X)^{-1}X'\varepsilon]$$
$$= \beta$$

Hence b is an unbiased estimator of β.

Covariance of b

The covariance matrix of y is $\sigma^2 I$

$b=(X'X)^{-1}X'y = Ay$ (where A is k by n)

$Cov(b) = A\, Var(y)\, A' = A\, \sigma^2 I\, A = \sigma^2 AA'$

$= \sigma^2\, (X'X)^{-1}X'X(X'X)^{-1}$

$= \sigma^2\, (X'X)^{-1}$

Covariance of b

For simple linear regression, $\sigma^2(X'X)^{-1}=$

$$\sigma^2\begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}^{-1} = \frac{\sigma^2}{n\sum x_i^2 - (\sum x_i)^2}\begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix}$$

$$SD(b_0) = \sigma\sqrt{\frac{\sum x_i^2}{nS_{xx}}}; \quad \mathrm{SD}(\mathrm{b}_1) = \sigma\sqrt{\frac{1}{S_{xx}}}$$

Estimation of σ^2

$$S^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2}$$

Note: The denominator is n - 2 since two parameters are being estimated (β_0 and β_1).

$E[S^2]=\sigma^2$ (See proof in Seber, Linear Regression Analysis)

Statistical Inference for β_0 and β_1

$$SE(\hat{\beta}_0) = s\sqrt{\frac{\sum x_i^2}{nS_{xx}}} \quad \text{and} \quad SE(\hat{\beta}_1) = \frac{s}{\sqrt{S_{xx}}}$$

$$\frac{\hat{\beta}_0 - \beta_0}{SE(\hat{\beta}_0)} \sim t_{n-2} \quad \text{and} \quad \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim t_{n-2}$$

For ozone example:

Coefficients:

	Value	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.2260	0.4614	-4.8243	0.0000
temperature	0.0704	0.0059	11.9511	0.0000

Sums of Squares

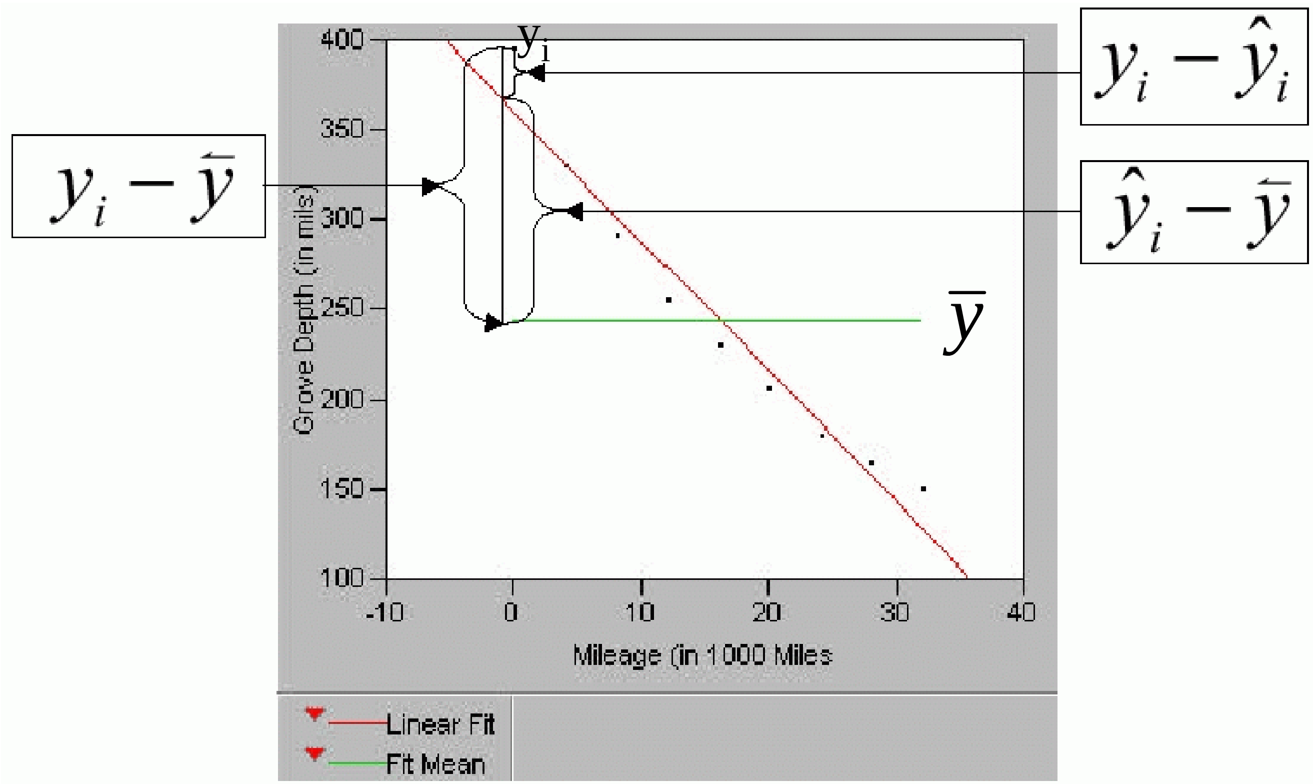
Sum of Squares Total (SST) : $\sum_{i=1}^{n}(y_i - \bar{y})^2$

Sum of Squares for Error (SSE) : $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

Sum of Squares for Regression (SSR) : $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$

Geometry of the Sums of Squares

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$



SST = SSR + SSE, see derivation on p. 354

Coefficient of Determination (R-squared)

$$r^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} =$$

proportion of the variance in y that is accounted for by the regression on x

= square of correlation between y and $\hat{y}$

For ozone example:
Multiple R-Squared: 0.5672

Analysis of Variance (ANOVA)

$$H_0 : \beta_1 = 0 \ vs. \ H_0 : \beta_1 \neq 0$$

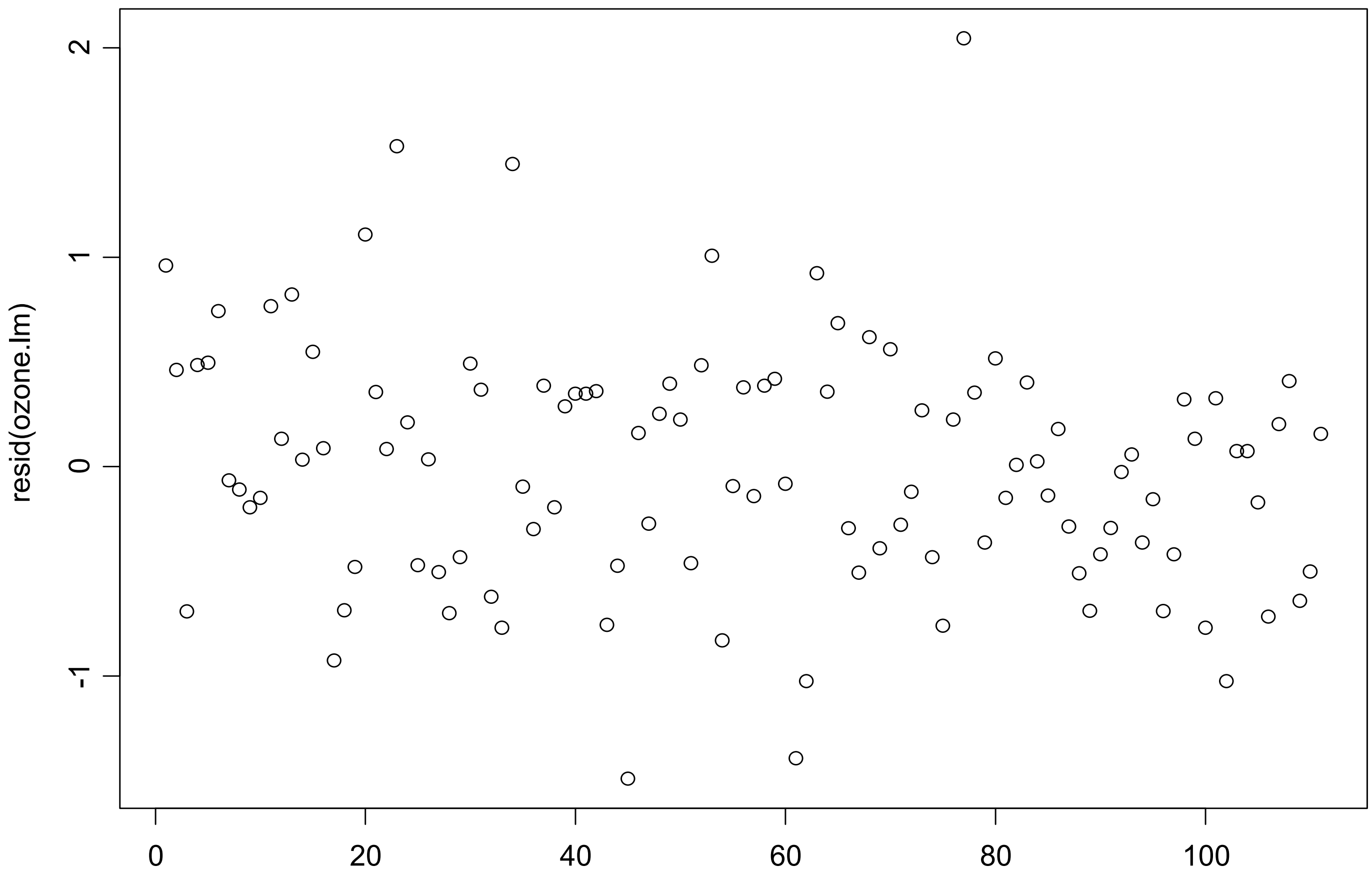
$$F = \frac{\text{SSR}/1}{\text{SSE}/(\text{n} - 2)} = \frac{\text{MSR}}{\text{MSE}} = t^2$$

For ozone example:

summary.aov(tmp)

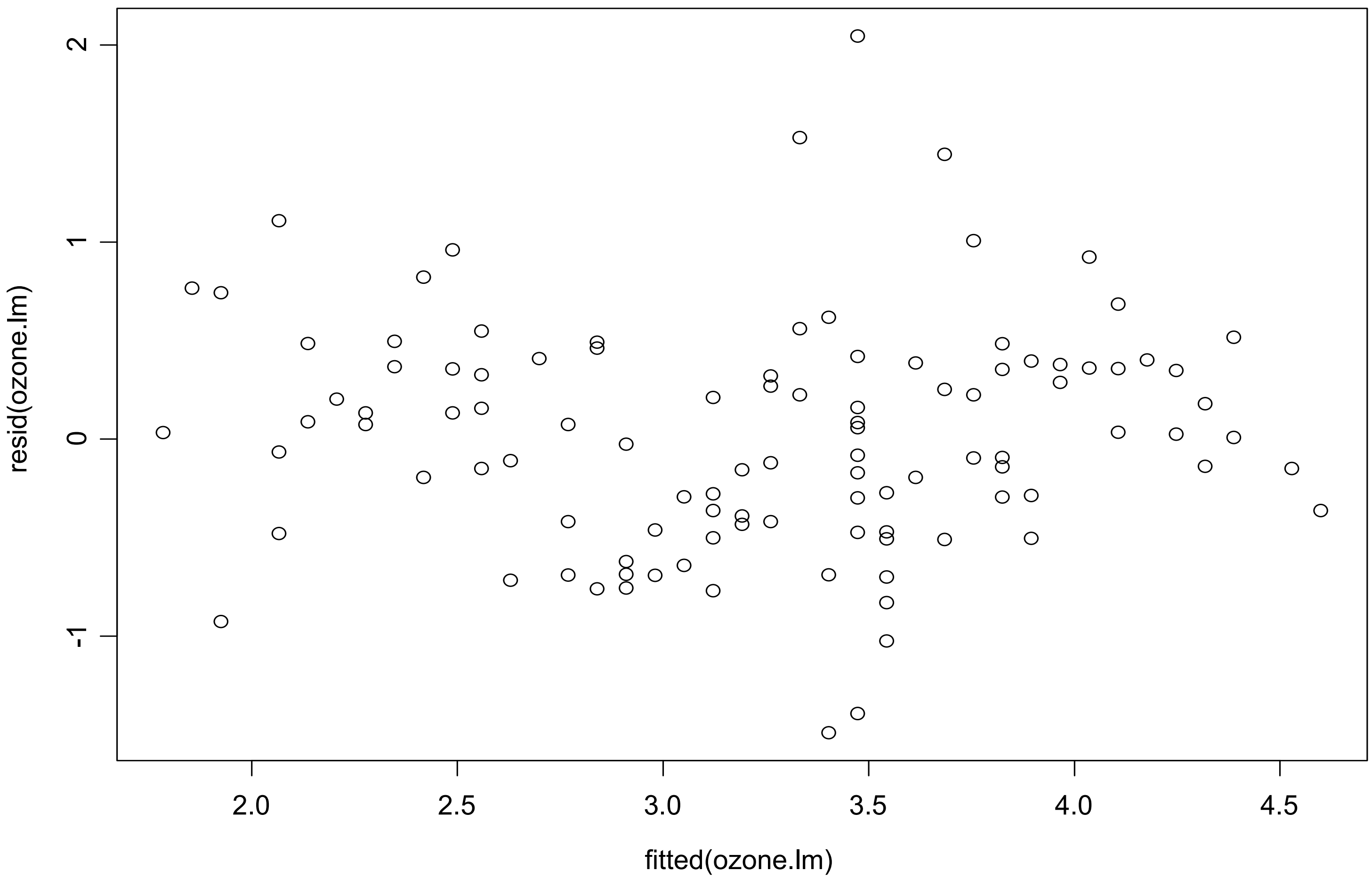
	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
temperature	1	49.46178	49.46178	142.8282	0
Residuals	109	37.74698	0.34630		

Regression Diagnostics
Residual vs. observation number

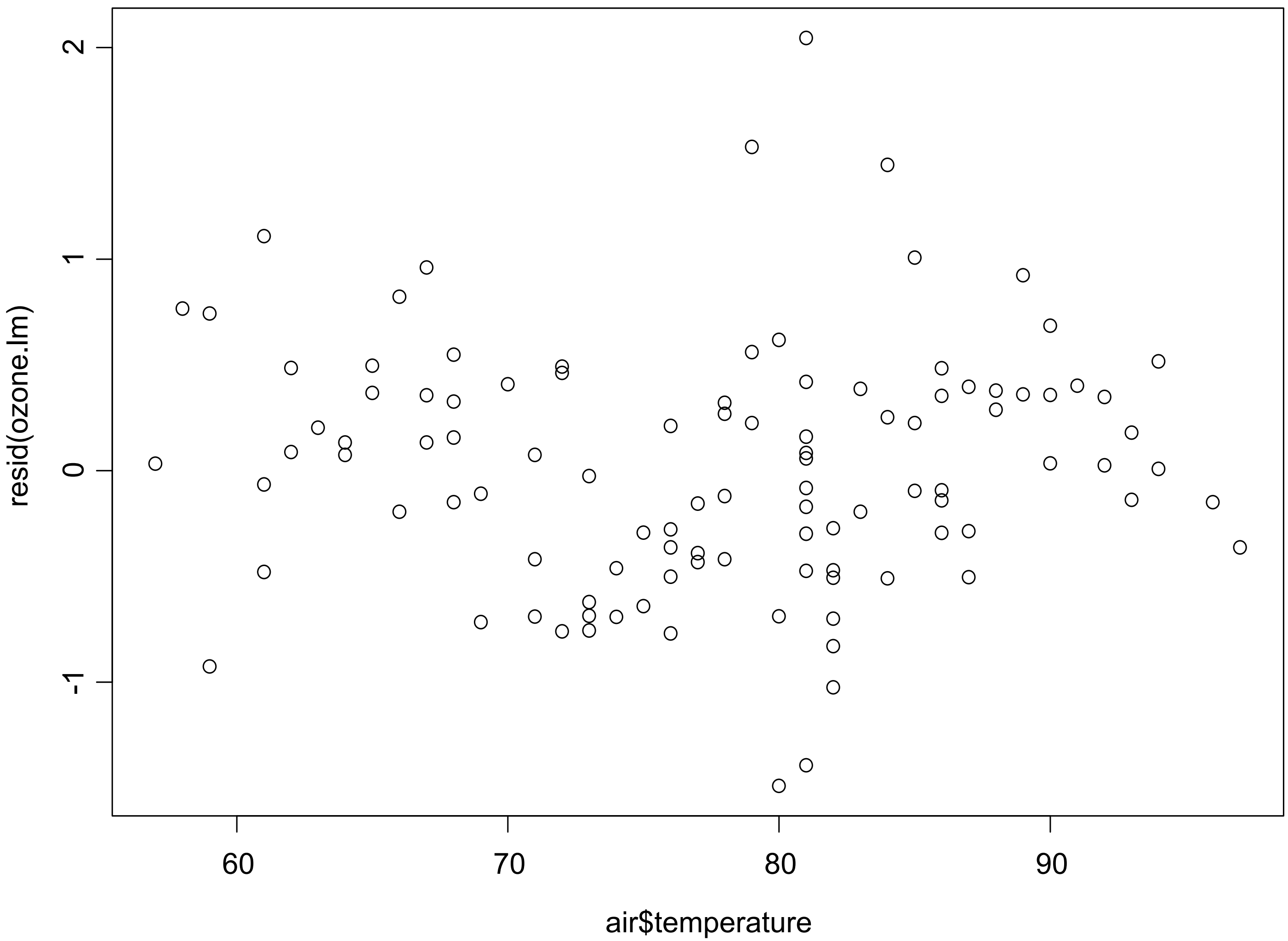


This graph was created using S-PLUS(R) Software. S-PLUS(R) is a registered trademark of Insightful Corporation.

Regression Diagnostics residual vs. fitted value

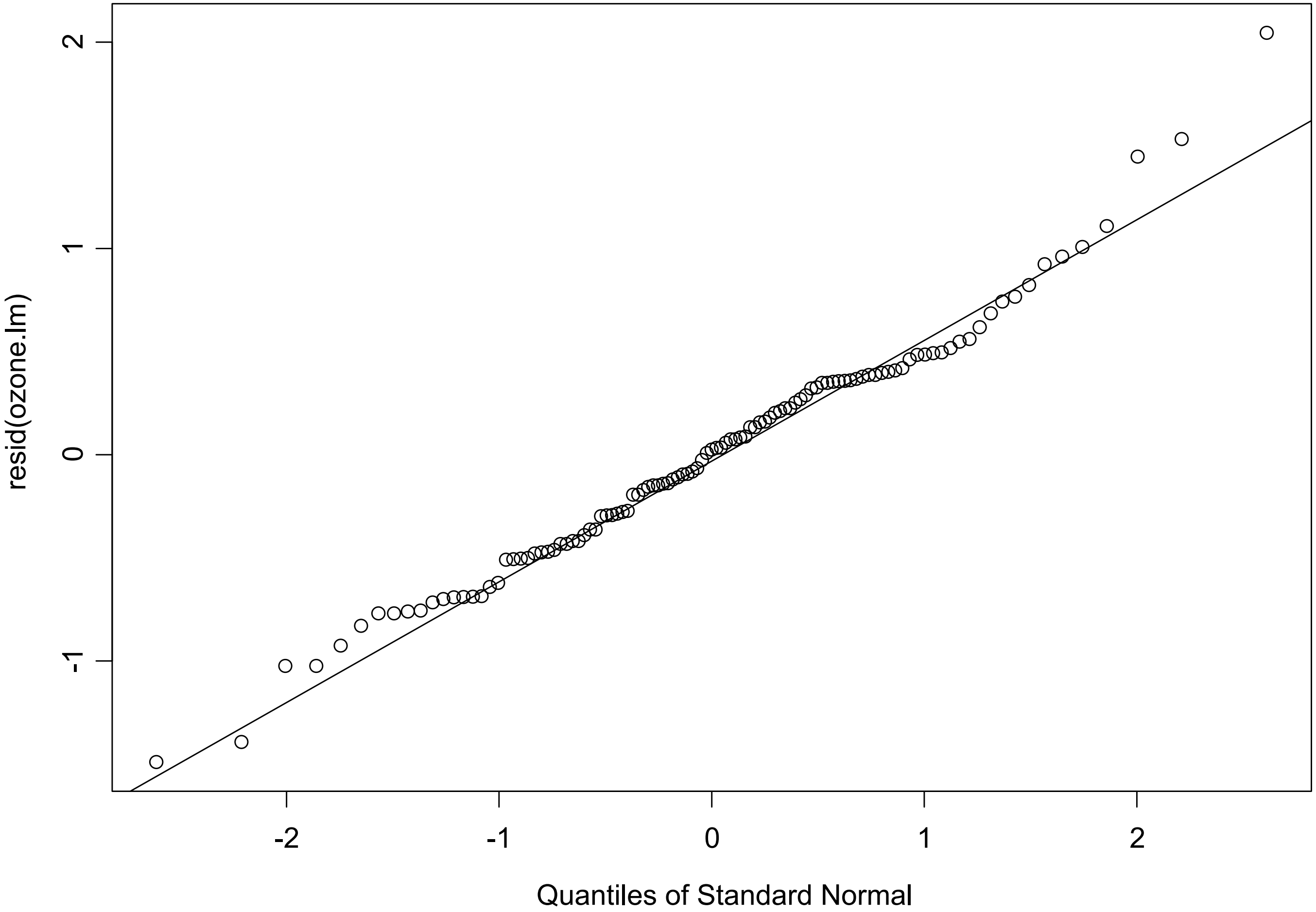


Regession Diagnostics
residual vs. x



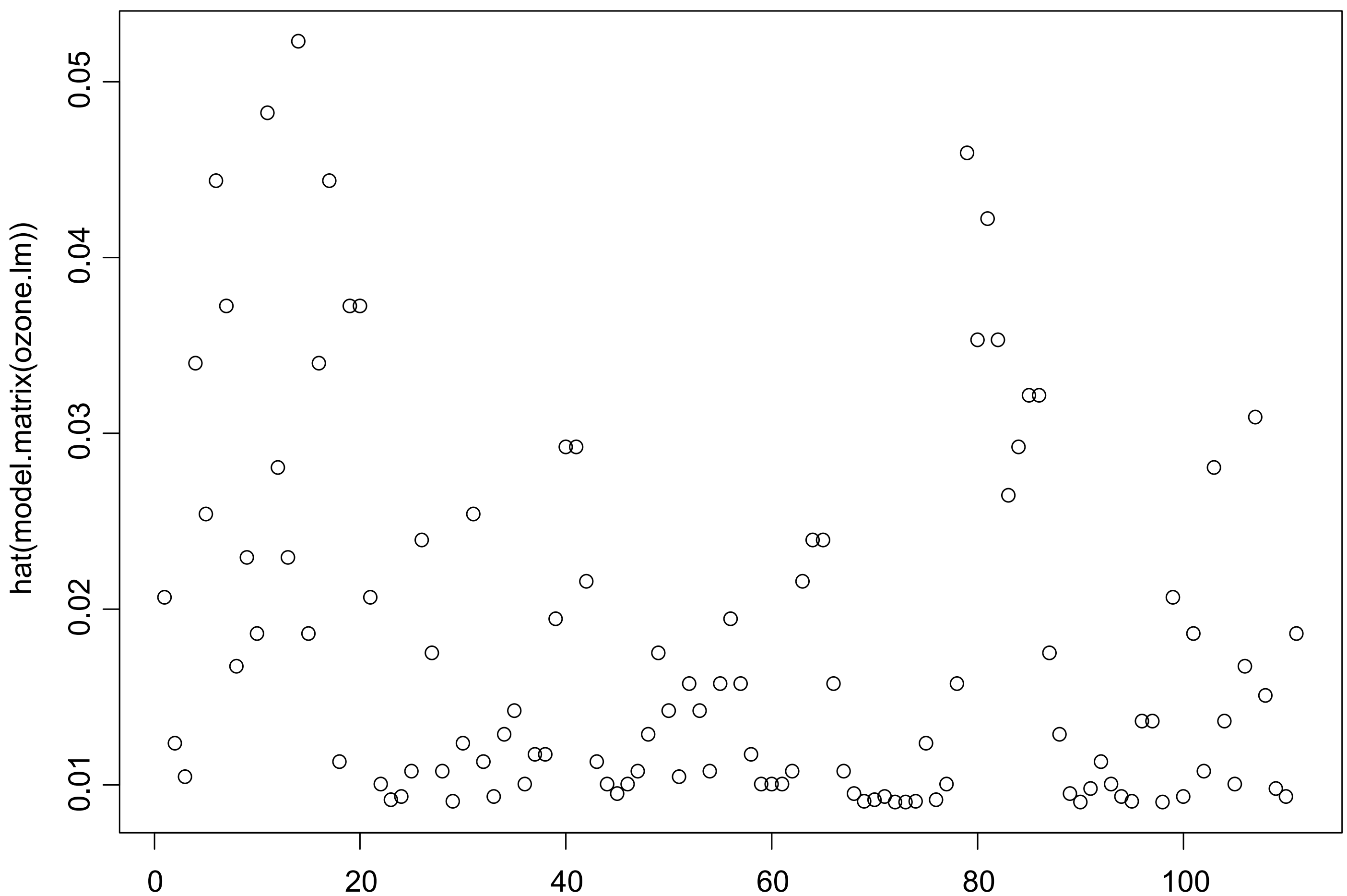
This graph was created using S-PLUS(R) Software. S-PLUS(R) is a registered trademark of Insightful Corporation.

Regression Diagnostics
qq plot of residuals



This graph was created using S-PLUS(R) Software. S-PLUS(R) is a registered trademark of Insightful Corporation.

Hat Matrix Diagonals



This graph was created using S-PLUS(R) Software. S-PLUS(R) is a registered trademark of Insightful Corporation.

Some useful S-Plus commands

my.lm <- lm(y~x, data=mydata, na.action=na.omit)
  includes intercept term by default

summary(my.lm)
  gives coefficients, correlation of coefficients, R-square, F-statistic, residual standard error

summary.aov(my.lm)
  gives ANOVA table

resid(my.lm)
  gives residuals

fitted(my.lm)
  gives fitted values

model.matrix(my.lm)
  gives model matrix