Assignment 3 — Due Oct 18, 2018

NAIQING CAI

ncai5@wisc.edu

1. Consider the model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ where $\varepsilon_i \sim$ ind $N(0, \sigma_i^2)$ for observations $i = 1,\ldots,n$.

Suppose $X_i > 0$ and $\sigma_i^2 = \sigma^2 X_i$.

$Y = (Y_1, \ldots, Y_n)'$: $n \times 1$ vector of response variables

X: $n \times 2$ design matrix with 1's in the first column and $(X_1, \ldots, X_n)'$ in the second column

$\beta = (\beta_0, \beta_1)'$: 2×1 vector of regression coefficients

(a) Derive the distribution (i.e., the type of distribution, mean vector, and the variance-covariance matrix) of the weighted least squares estimates of β.

$$\tilde{Q}(\beta) = \sum_{i=1}^{n} \frac{\{Y_i - (\beta_0 + \beta_1 X_i)\}^2}{\sigma_i^2} \qquad w_i = \frac{1}{\sigma_i^2}, \sigma_i^2 = cX_i^2$$

$$\Rightarrow \tilde{Q}(\beta) = \sum_{i=1}^{n} \frac{\{Y_i - (\beta_0 + \beta_1 X_i)\}^2}{X_i^2}$$

$$\Rightarrow \tilde{\beta} = (X^T \Sigma^{-1} X)^{-1}(X^T \Sigma^{-1} Y) \quad where \Sigma = diag(1/w_i)$$

Thus, we have:

$$\begin{cases} E(\tilde{\beta}) = \beta \\ \text{var}(\tilde{\beta}) = \sigma^2 (X^T \Sigma^{-1} X)^{-1} \end{cases} \Rightarrow \tilde{\beta} \sim N(\beta, \sigma^2 (X^T \Sigma^{-1} X)^{-1})$$

(b) Despite the weighted variances, derive the ordinary least squares estimates of β by minimizing the (unweighted) error sum of squares all in matrix terms.

$$Q(\beta) = \sum_{i=1}^{n} \{Y_i - (\beta_0 + \beta_1 X_i)\}^2$$

$$\min Q(\beta) = \min \sum_{i=1}^{n} \{Y_i - (\beta_0 + \beta_1 X_i)\}^2$$

$$\Rightarrow \begin{cases} \dfrac{\partial Q(\beta)}{\partial \beta_0} = \dfrac{\sum_{i=1}^{n} \{Y_i - (\beta_0 + \beta_1 X_i)\}^2}{\partial \beta_0} = 0 \\ \dfrac{\partial Q(\beta)}{\partial \beta_1} = \dfrac{\sum_{i=1}^{n} \{Y_i - (\beta_0 + \beta_1 X_i)\}^2}{\partial \beta_1} = 0 \end{cases} \Rightarrow \begin{cases} \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \\ \hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \end{cases}$$

$$\Rightarrow \hat{\beta} = (X^T X)^{-1} X^T Y$$

(c) Find the distribution of the (ordinary) least squares estimates

$$E(\hat{\beta}) = E((X^T X)^{-1} X^T Y) = (X^T X)^{-1} X^T E(Y)$$
$$= (X^T X)^{-1} X^T E(X\beta + \varepsilon) = (X^T X)^{-1} X^T [X\beta + E(\varepsilon)]$$
$$= (X^T X)^{-1} X^T X\beta = \beta$$

$$\text{var}(\hat{\beta}) = E[(\hat{\beta} - E(\hat{\beta}))(\hat{\beta} - E(\hat{\beta}))^T] = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T]$$
$$= E\{[(X^T X)^{-1} X^T \varepsilon][(X^T X)^{-1} X^T \varepsilon]^T\} = (X^T X)^{-1} X^T E[\varepsilon\varepsilon^T] X^T (X^T X)^{-1}$$
$$= \sigma^2 (X^T X)^{-1}$$

$$\Rightarrow \hat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1})$$

(d) Draw a connection

$$when \Sigma^{-1} = I_n, \text{var}(\tilde{\beta}) \text{ can be reduced to } \text{var}(\hat{\beta})$$

That is the weighted least squares estimation be reduced to ordinary least squares estimation

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

$$\tilde{\beta} = (X^T \Sigma^{-1} X)^{-1} (X^T \Sigma^{-1} Y)$$

$$\hat{\beta} = \tilde{\beta} + cY \Rightarrow \text{var}\,\hat{\beta} = \text{var}(\tilde{\beta} + cY) = \text{var}(\tilde{\beta}) + \text{var}(cY)$$

$$\Rightarrow \text{var}\,\hat{\beta} \geq \text{var}\,\tilde{\beta}$$

$$\Rightarrow \tilde{\beta} : BLUE$$

2. A chemist studied the concentration (Y) of a solution over time (X). Fifteen (15) identical solutions were prepared. The 15 solutions were randomly divided into five sets of three, and the five sets were measured, respectively, after 1, 3, 5, 7, and 9 hours.

(a) Identify ingredients of the experiment/study including population vs. sample, study/experimental units, and whether cause-and-effect relationships can be established.

Population vs. Sample:

population: total concentration (Y) of a solution & time (X)

sample: fifteen concentration of identical solutions (y) & time (x)

Study/experimental Units:

5 sets of concentration of identical solutions & their time (each set of 3 size)

Cause-and-Effect Relationships:

We can estimate the relationships between cause-and-effect.

(b) (Hand calculation with calculator) Perform a simple linear regression analysis with concentration (Y) as the response variable and time (X) as the explanatory variable. Obtain the regression coefficients estimates along with their standard errors, an unbiased estimate of error variance, and the coefficient of determination R^2.

Simple Linear Regression Model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \varepsilon_i \sim i.i.d. N(0, \sigma^2)$$

Regression Coefficients Estimates:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \Rightarrow \hat{\beta}_1 = -0.324$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \Rightarrow \hat{\beta}_0 = 2.575$$

Standard Errors:

$$\hat{se}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}} \Rightarrow \hat{se}(\hat{\beta}_1) = 0.0433$$

$$\hat{se}(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)} \Rightarrow \hat{se}(\hat{\beta}_0) = 0.2487$$

Unbiased Estimate of Error Variance:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-2} \Rightarrow \hat{\sigma}^2 = 0.225$$

Coefficient of Determination R^2:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \Rightarrow R^2 = 0.8116$$

(c) (Hand calculation with calculator) Perform a hypothesis test to determine whether there is evidence that the mean concentrations are different for the different hours since the solutions are prepared. Indicate the assumptions underlying the test.

Assumptions (one sample):

1) Independence within the sample

2) Normality for the sample

3) Outliers: can be sensitive

Hypothesis:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

T-test (two-sided, 95 CI):

$$t^* = \frac{\hat{\beta}_1}{\hat{se}(\hat{\beta}_1)} = -7.483$$

P-value:

$$p = 2 * P(T_{n-2} > |t^*|) = 4.61\text{e-}06$$

Conclusions:

Since p-value<0.05, we should reject the null hypothesis. There is evidence that the mean concentrations are different for the different hours since the solutions are prepared.

(d) Indicate the model underlying the regression analysis in (b) and assess the model assumptions by using suitable graphical techniques. How reasonable are the assumptions? What remedial measures are desirable, if any?

Simple Linear Regression Model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \varepsilon_i \sim i.i.d. N(0, \sigma^2)$$

Y_i : concentration of a solution (response variables)

X_i : time (explanatory variables)

β_0 : intercept

β_1 : slope

ε_i : random errors

Model Assumptions:

- A straight line relationship between the response variable Y and the explanatory variable X

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad E(\varepsilon_i)=0$$

- Equal Variance

$$\text{var}(\varepsilon_i) = \sigma^2$$

- Independence

$$\text{cov}(\varepsilon_i, \varepsilon_{i'})=0, i \neq i'$$

- Normal Distribution

$$\varepsilon_i \sim i.i.d.N(0,\sigma^2)$$

Graphical Techniques to Assess Model Assumptions:

1) Linear Assumption: Y v.s X

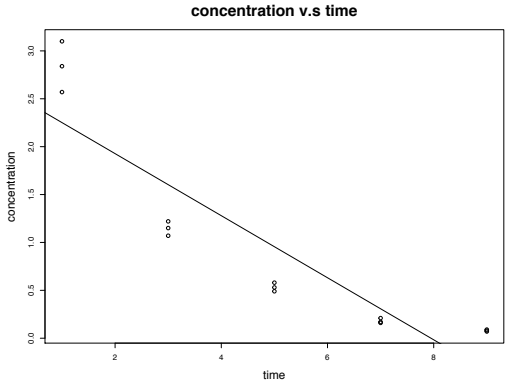


figure 2.1 Linear Plot

Based on the plot of y against x, we can conclude the nonlinearity of the regression function. Thus, the assumption linear regression is not reasonable.

2) Equal Variance Assumption: residual against fitted values

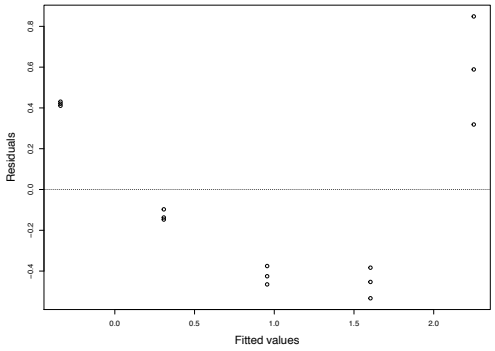


figure 2.2 Residual against Fitted Values

Based on the plot of residual against fitted values, we can conclude that the residuals do not have equal variance. Thus, the assumption equal variance is not reasonable.

3) Normal Distribution Assumption: QQ plot of residuals

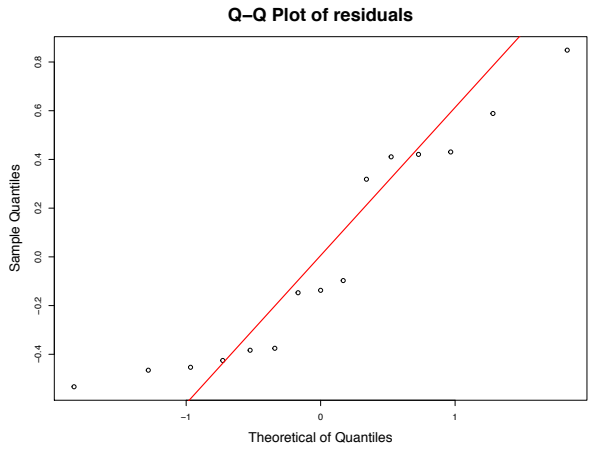


figure 2.3 QQ Plot of Residuals

Based on the QQ plot of residual above, the residuals are not completely from normal distribution, so the assumption normal distribution is not reasonable.

Remedial Measures:

1) Transformation:

Consider a log transformation for the data and make the model to be linear regression model.

2) Weighted Least Squares:

Consider extend the model to a weighted linear regression model so that the assumption that equal error variance could be not true.

(e) What kind of variable transformations would you recommend to the chemist? Give reasoning.

I recommend the log transformation of the data Y because it can not only change the nonlinear regression model into a linear one, but also control unequal variance.

(f) Repeat (b) but now with the concentration after the transformation in (e).

Simple Linear Regression Model:

$$\log(Y_i) = \beta_0 + \beta_1 X_i + \varepsilon_i, \varepsilon_i \sim i.i.d. N(0, \sigma^2)$$

Regression Coefficients Estimates:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \Rightarrow \hat{\beta}_1 = -0.44993$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \Rightarrow \hat{\beta}_0 = 1.50792$$

Standard Errors:

$$\hat{se}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}} \Rightarrow \hat{se}(\hat{\beta}_1) = 0.01049$$

$$\hat{se}(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)} \Rightarrow \hat{se}(\hat{\beta}_0) = 0.06028$$

Unbiased Estimate of Error Variance:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-2} = 0.013225$$

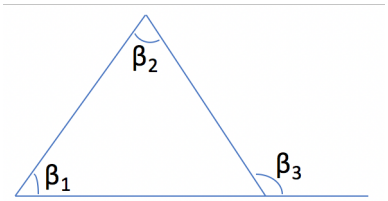
Coefficient of Determination R^2:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \Rightarrow R^2 = 0.993$$

(g) Compare the results in (b) and (f).

The standard error and unbiased estimate of error variance are much smaller after the transformation, and R^2 are much bigger, which means the regression model is more approximate to the true model.

3. Suppose we wish to measure the three angles β1, β2, and β3 as depicted in the diagram below.



Elementary geometry shows that β1 + β2 = β3. Suppose, as a check on the accuracy of the results, we decide to measure all three angles, with measurement error. Let bj be the actual measurement for βj, j = 1, 2, 3. Due to measurement error, b1 + b2 might not be equal to b3. Assume that the measurement errors are independent and follow normal distribution with mean 0 and variance σ2. Formulate this as a multiple linear regression model, and derive the least squares estimates β^1, β^2, and β^3, with their standard deviations.

We can construct a multiple linear regression model as follows:

$$b_1 = \beta_1 + \varepsilon_1 \qquad b_1 = \beta_1 + \varepsilon_1$$
$$b_2 = \beta_2 + \varepsilon_2 \qquad b_2 = \beta_2 + \varepsilon_2$$
$$b_3 = \beta_3 + \varepsilon_3 \qquad b_3 = \beta_1 + \beta_2 + \varepsilon_3$$

$\varepsilon_1, \varepsilon_2, \varepsilon_3$ *assumed independent with mean* 0 *and variance* σ^2

$$\beta = (\beta_1, \beta_2)'$$

$$X = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix} \qquad X'Y = \begin{pmatrix} b_1 + b_3 \\ b_2 + b_3 \end{pmatrix}$$

$$\begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix}$$

$$(X'X)^{-1} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}^{-1} = \begin{pmatrix} \frac{2}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{2}{3} \end{pmatrix}$$

$$\hat{\beta}_1 = \frac{2b_1 - b_2 + b_3}{3}$$
$$\hat{\beta}_2 = \frac{-b_1 + 2b_2 + b_3}{3}$$
$$\hat{\beta}_3 = \frac{b_1 + b_2 + 2b_3}{3}$$

$$\textit{standard deviation} \sqrt{\frac{2\sigma^2}{3}}$$

4. Consider the multiple regression model

Y =β0 +β1X1 +···βp−1Xp−1 +ε

(a) Find the distribution

$$\sum_{j=0}^{p-1} c_j \hat{\beta}_j = c'\hat{\beta}, where\ c = (c_0,...,c_{p-1})', \beta = (\beta_0,...,\beta_{p-1})'$$

$$E(c'\hat{\beta}) = c'\beta$$

$$\text{var}(c'\hat{\beta}) = c'\,\text{var}(c'\beta)c = \sigma^2 c'(X'X)^{-1}c$$

$$\Rightarrow c'\hat{\beta} \sim MVN(c'\beta, \sigma^2 c'(X'X)^{-1}c)$$

(b) Consider the hypothesis. Explain how to test theses hypotheses at significance level α. Construct a suitable test statistic, find its distribution, and specify the rejection region.

Hypotheses:

$$H_0 : \sum_{j=0}^{p-1} c_j \beta_j = h$$

$$H_1 : \sum_{j=0}^{p-1} c_j \beta_j \neq h$$

Test Statistic and its Distribution:

$$T^* = \frac{c'\hat{\beta} - c'\beta}{\sqrt{\hat{\text{var}}(c'\hat{\beta})}} \sim T_{n-p}, where\ c = (c_0,...,c_{p-1})', \beta = (\beta_0,...,\beta_{p-1})'$$

$$T^* = \frac{c'\hat{\beta} - h}{\sqrt{\hat{\text{var}}(c'\hat{\beta})}} \sim T_{n-p}, where\ c = (c_0,...,c_{p-1})', \beta = (\beta_0,...,\beta_{p-1})'$$

Rejection Region:

$$(-\infty, c'\hat{\beta} - t_{n-p,\alpha/2}\sqrt{\hat{\text{var}}(c'\hat{\beta})}) \cup (c'\hat{\beta} + t_{n-p,\alpha/2}\sqrt{\hat{\text{var}}(c'\hat{\beta})}, +\infty)$$

(c) Find the distribution of Yn+1 − Yn+1(hat)

$$Y_{n+1} = z\beta + \varepsilon, \varepsilon \sim i.i.d.N(0,\sigma^2)$$

$$\hat{Y}_{n+1} = z\hat{\beta}$$

$$E(Y_{n+1} - \hat{Y}_{n+1}) = E(Y_{n+1}) - E(\hat{Y}_{n+1}) = z\beta - z\beta = 0$$

$$\text{var}(Y_{n+1} - \hat{Y}_{n+1}) = \text{var}(z\beta + \varepsilon - z\hat{\beta}) = \text{var}(z\beta - z\hat{\beta}) + \text{var}(\varepsilon)$$

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

$$\Rightarrow Y_{n+1} - \hat{Y}_{n+1} \sim N(0, \sigma^2(1 + z'(X'X)^{-1}z))$$

(d) Show that the MSE (mean square error) of the prediction Yn+1(hat) = zβ(hat) is strictly greater than σ2.

$$MSE\ of\ \hat{Y}_{n+1} = E[(\hat{Y}_{n+1} - Y_{n+1})^2] = E[(\hat{Y}_{n+1} - E(\hat{Y}_{n+1}) + E(\hat{Y}_{n+1}) - Y_{n+1})^2]$$
$$= \text{var}(\hat{Y}_{n+1}) + bias^2(\hat{Y}_{n+1}) = \text{var}(\hat{Y}_{n+1}) + (E(\hat{Y}_{n+1}) - Y_{n+1})^2$$
$$= \text{var}(\hat{Y}_{n+1}) + (E(\hat{Y}_{n+1}) - z\beta - \varepsilon)^2 > \sigma^2$$

(e) Given the vector z of predictor variables for the future observation Yn+1, find an interval I

$$P(Y_{n+1} \in \text{I}) = 1 - \alpha$$

$$\frac{\hat{Y}_{n+1} - Y_{n+1}}{\hat{\sigma}_{pred}} \sim T_{n-p}$$

$$\text{I} = [\hat{Y}_{n+1} - t_{n-p,\alpha/2}\hat{\sigma}_{pred}, \hat{Y}_{n+1} + t_{n-p,\alpha/2}\hat{\sigma}_{pred}] \quad where\ \hat{\sigma}_{pred} = \hat{\sigma}\sqrt{1 + z'(X'X)^{-1}z}$$

(f) Find an expression for the residual vector rp in terms of X and xp

$$Y = \gamma_0 + \gamma_1 X_1 + \cdots \gamma_{p-1} X_{p-1} + \gamma_p X_p + \varepsilon$$

$$\tilde{X} = [X, x_p] \qquad \hat{x}_p = a_0 x_0 + \cdots + a_{p-1} x_{p-1}$$

$$x_p = \begin{pmatrix} 1 & \cdots & x_{1,p-1} \\ \vdots & \ddots & \vdots \\ 1 & \cdots & x_{n,p-1} \end{pmatrix} \begin{pmatrix} a_0 \\ \vdots \\ a_{p-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_1^r \\ \vdots \\ \varepsilon_n^r \end{pmatrix} \Rightarrow x_p = XA + \varepsilon^r$$

$$\hat{A} = (X'X)^{-1}X'x_p$$

$$r_p = x_p - \hat{x}_p = x_p - X\hat{A} = x_p - X(X'X)^{-1}X'x_p$$

(g) Expression the least squares estimates

$$\hat{\gamma} = (\tilde{X}'\tilde{X})\tilde{X}'Y, \tilde{X} = [X, x_p]$$

$$\hat{\gamma} = \left(\begin{bmatrix} X' \\ x_p' \end{bmatrix} \begin{bmatrix} X & x_p \end{bmatrix} \right)^{-1} \begin{bmatrix} X'Y \\ x_p'Y \end{bmatrix} = \begin{pmatrix} X'X & X'x_p \\ x_p'X & x_p'x_p \end{pmatrix}^{-1} \begin{bmatrix} X'Y \\ x_p'Y \end{bmatrix} = \begin{pmatrix} \hat{\beta} - \dfrac{\hat{A}x_p'N_xY}{x_p'N_xx_p} \\ \dfrac{x_p'N_xY}{x_p'N_xx_p} \end{pmatrix}$$

$$where\ \hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_{p-1})'\ and\ \hat{A} = (\hat{a}_0, \hat{a}_1, ..., \hat{a}_{p-1})\ and\ N_x = I - X(X'X)^{-1}X'$$

$$N_x x_p = r_p \Rightarrow x_p'N_x x_p = r_p'r_p$$

$$\hat{\gamma} = \begin{pmatrix} \hat{\gamma}_0 \\ \vdots \\ \hat{\gamma}_p \end{pmatrix} = \begin{pmatrix} \hat{\beta} - \dfrac{\hat{A}r_p'Y}{r_p'r_p} \\ \dfrac{r_p'Y}{r_p'r_p} \end{pmatrix}$$

Appendix:

```
#2 linear
x=c(9,9,9,7,7,7,5,5,5,3,3,3,1,1,1)
y=c(0.07,0.09,0.08,0.16,0.17,0.21,0.49,0.58,0.53,1.22,1.15,1.07,2.84,2.57,3.10)
plot(x,y,xlab="time",ylab="concentration",main = "concentration v.s time",cex.lab=1.5,
cex.main=2)
lsfit=lm(y ~ x)
abline(lsfit)
```

```
#2 equal variance
x=c(9,9,9,7,7,7,5,5,5,3,3,3,1,1,1)
y=c(0.07,0.09,0.08,0.16,0.17,0.21,0.49,0.58,0.53,1.22,1.15,1.07,2.84,2.57,3.10)
lsfit=lm(y ~ x)
lsfit
confint(lsfit)
e=lsfit$residuals
fitted.values=lsfit$coefficients[1]+lsfit$coefficients[2]*x
plot(fitted.values,e,xlab = "Fitted values",ylab = "Residuals",cex.lab=1.6)
abline(h=0,lty="dotted")
```

```
#2 qq-plot
x=c(9,9,9,7,7,7,5,5,5,3,3,3,1,1,1)
y=c(0.07,0.09,0.08,0.16,0.17,0.21,0.49,0.58,0.53,1.22,1.15,1.07,2.84,2.57,3.10)
lsfit=lm(y ~ x)
lsfit
confint(lsfit)
e=lsfit$residuals
qqnorm(e,main = "Q-Q Plot of residuals",xlab = "Theoretical of Quantiles",ylab = "Sample
Quantiles",cex.lab=1.5,cex.main=2)
qqline(e, col=2, lwd=2)
```