Stat 849: Analysis of variance

Sündüz Keleş

Department of Statistics
Department of Biostatistics and Medical Informatics
University of Wisconsin, Madison

Analysis of variance

Consists of special cases of multiple linear regression model.
Regressors (predictors/covariates) are qualitative.
Example: Dataset comes from a study of blood coagulation times. 24 animals were randomly assigned to four different diets and the samples were taken in a random order.

A	B	C	D
4	6	6	8

Table: Number of animals in each diet group.

`diet` is referred to as a *factor* or *treatment*, and the four kinds of diet types are called *levels* of the factor.
This entire experiment is referred to as a *single factor experiment* or *one-way layout*.

I levels of the factor; n_i observations ith sample, $i = 1, \cdots, I$.
Each sample is from a particular underlying population which is normally distributed. We assume a common variance σ^2 across different populations.
Let Y_{ij} be the j-th observation ($j = 1, \cdots, n_i$) on the i-th normal population $\mathcal{N}(\mu_i, \sigma^2)$, $i = 1, \cdots, I$.

		Sample mean
Population 1:	$Y_{11}, Y_{12}, \cdots, Y_{1n_1}$	$\bar{Y}_{1.}$
Population 2:	$Y_{21}, Y_{32}, \cdots, Y_{2n_2}$	$\bar{Y}_{2.}$
$\cdots$	$\cdots$	$\cdots$
Population I:	$Y_{I1}, Y_{I2}, \cdots, Y_{In_I}$	$\bar{Y}_{I.}$

Consider the following single model

$$Y_{ij} = \mu_i + \epsilon_i, \quad i = 1, \cdots, I; \quad j = 1, \cdots, n_j,$$

where ϵ_{ij} are i.i.d. as $\mathcal{N}(0, \sigma^2)$.
Now, consider rewriting this in the form of $Y = X\beta + \epsilon$, where $\epsilon \sim \mathcal{N}_n(0, \sigma^2 I_n)$, Y represents all the observations and $\beta = (\mu_1, \mu_2, \cdots, \mu_I)$ represent the unknown parameters.

$$X = \begin{pmatrix} \mathbf{1}_{n_1} & 0 & 0 & \cdots & 0 \\ 0 & \mathbf{1}_{n_2} & 0 & \cdots & 0 \\ 0 & 0 & \mathbf{1}_{n_3} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & \mathbf{1}_{n_I} \end{pmatrix},$$

where $\mathbf{1}_{n_i}$ is a vector of n_i 1s.
We now have a regular linear regression model *without an intercept* and with a special design matrix. X has orthogonal columns and is of rank I. We can do estimation and inference in this model using the least squares theory.

Source	Sum of squares (SS)	Df
Between populations	$\mathrm{SSR} = \sum_i \sum_j (\bar{Y}_{i.} - \bar{Y}_{..})^2$	$I - 1$
Within populations	$\mathrm{SSE} = \sum_i \sum_j (Y_{ij} - \bar{Y}_{i.})^2$	$N - I$
Total	$\mathrm{SST} = \sum_i \sum_j (Y_{ij} - \bar{Y}_{..})^2$	$N - 1$

Thus, we have the following test statistic for testing $H_0 : \mu_1 = \mu_2 = \cdots = \mu_I$,

$$F = \frac{SSR/(I-1)}{SSE/(N-I)} \sim F_{(I-1),(N-I)}.$$

Source	Sum of squares (SS)	Df
Between populations	$SSR = \sum_i \sum_j (\bar{Y}_{i.} - \bar{Y}_{..})^2$	$I - 1$
Within populations	$SSE = \sum_i \sum_j (Y_{ij} - \bar{Y}_{i.})^2$	$N - I$
Total	$SST = \sum_i \sum_j (Y_{ij} - \bar{Y}_{..})^2$	$N - 1$

Thus, we have the following test statistic for testing
$H_0 : \mu_1 = \mu_2 = \cdots = \mu_I$,

$$F = \frac{SSR/(I-1)}{SSE/(N-I)} \sim F_{(I-1),(N-I)}.$$

Here, SSE is also the residual sum of squares from the full (larger) model and SSR equals the difference between the residual sum of squares of the null model and the full model.
$RSS_{H_0} = \sum_i \sum_j (Y_{ij} - \hat{\mu})^2 = \sum_i \sum_j (Y_{ij} - \bar{Y}_{..})^2$.
$RSS_{H_a} = \sum_i \sum_j (Y_{ij} - \hat{\mu}_i)^2 = \sum_i \sum_j (Y_{ij} - \bar{Y}_{i.})^2$.
$RSS_{H_0} - RSS_{H_a} = \sum_i n_i (\bar{Y}_{i.} - \bar{Y}_{..})^2$.

Alternative parameterizations

Let's try to include an intercept:

$$Y_{ij} = \mu + \alpha_i + \epsilon_i, \quad i = 1, \cdots, I$$

where $\alpha_1, \cdots, \alpha_I$ are I contrasts defined as $\alpha_i = \mu_i - \mu$, where $\mu = \sum_{i=1}^{I} \mu_i / I$.

Alternative parameterizations

Let's try to include an intercept:

$$Y_{ij} = \mu + \alpha_i + \epsilon_i, \quad i = 1, \cdots, I$$

where $\alpha_1, \cdots, \alpha_I$ are I contrasts defined as $\alpha_i = \mu_i - \mu$, where $\mu = \sum_{i=1}^{I} \mu_i / I$.
However, now α_i are mathematically dependent, i.e.,
$\alpha_\cdot = \sum_i \alpha_i = 0$ (identifiability constraint).
Hence, we have $Y = X_2\gamma + \epsilon$, where $\gamma = (\mu, \alpha_1, \cdots, \alpha_I)'$,
$X_2 = \texttt{cbind}(1, X)$ is an $n \times (I+1)$ matrix of rank I.
Set $\alpha_I = -\sum_{i=1}^{I-1} \alpha_i$ to turn X_2 into a full rank matrix.

So, we can have several parameterizations (driven by the hypothesis we are after).

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad i = 1, \cdots, I, \quad j = 1, \cdots, n_i.$$

Set $\mu = 0$ and use I dummy variables.
Set $\alpha_1 = 0$, then μ represents μ_1 (reference treatment) and $\alpha_i = \mu_i - \mu_1$ (treatment contrasts).

Example: Diet A will be the reference group.

```
plot(coag ~ diet, data = coagulation)
> lm1 = lm(coag ~ diet, data = coagulation)
> summary(lm1)

Call: lm(formula = coag ~ diet, data = coagulation)

Residuals:
       Min          1Q      Median          3Q         Max
-5.000e+00 -1.250e+00  1.488e-16  1.250e+00  5.000e+00

Coefficients:
             Estimate Std. Error   t value Pr(>|t|)
 (Intercept)  6.100e+01  1.183e+00    51.554  < 2e-16 ***
 dietB        5.000e+00  1.528e+00     3.273 0.003803 **
 dietC        7.000e+00  1.528e+00     4.583 0.000181 ***
 dietD       -1.071e-14  1.449e+00 -7.39e-15 1.000000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

 Residual standard error: 2.366 on 20 degrees of freedom
 Multiple R-Squared: 0.6706,     Adjusted R-squared: 0.6212
 F-statistic: 13.57 on 3 and 20 DF,  p-value: 4.658e-05
```

```
> model.matrix(lm1)
   (Intercept) dietB dietC dietD
1            1     0     0     0
2            1     0     0     0
3            1     0     0     0
4            1     0     0     0
5            1     1     0     0
6            1     1     0     0
7            1     1     0     0
8            1     1     0     0
9            1     1     0     0
10           1     1     0     0
11           1     0     1     0
12           1     0     1     0
13           1     0     1     0
14           1     0     1     0
15           1     0     1     0
16           1     0     1     0
17           1     0     0     1
18           1     0     0     1
19           1     0     0     1
20           1     0     0     1
21           1     0     0     1
22           1     0     0     1
23           1     0     0     1
24           1     0     0     1
#Coding of treatment contrasts with 4 factor levels.
> contr.treatment(4)
  2 3 4
1 0 0 0
2 1 0 0
3 0 1 0
4 0 0 1
```

```
> lm2 = lm(coag ~ diet - 1, data = coagulation)
> summary(lm2)

Call: lm(formula = coag ~ diet - 1, data = coagulation)

Residuals:
       Min          1Q      Median          3Q         Max
-5.000e+00 -1.250e+00  1.743e-16  1.250e+00  5.000e+00

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
 dietA  61.0000     1.1832   51.55   <2e-16 ***
 dietB  66.0000     0.9661   68.32   <2e-16 ***
 dietC  68.0000     0.9661   70.39   <2e-16 ***
 dietD  61.0000     0.8367   72.91   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

 Residual standard error: 2.366 on 20 degrees of freedom
 Multiple R-Squared: 0.9989,     Adjusted R-squared: 0.9986
 F-statistic: 4399 on 4 and 20 DF,  p-value: < 2.2e-16
```

Levene's test for homogeneity of the variances

Compute absolute value of the residuals

$$Z_{ij} = |Y_{ij} - \bar{Y}_{i.}|, \quad j = 1, 2, \cdots, n_i.$$

Then, fit one-way anova model using Z as response.

```
> summary(lm(abs(lm1$resid) ~ diet))
Call: lm(formula = abs(lm1$resid) ~ diet)
Residuals:
       Min          1Q      Median          3Q         Max
-2.000e+00 -1.000e-00 -1.361e-16  6.250e-01  3.000e+00

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
 (Intercept)   1.5000     0.7159   2.095   0.0491
   dietB       0.5000     0.9242   0.541   0.5945
   dietC      -0.5000     0.9242  -0.541   0.5945
   dietD       0.5000     0.8768   0.570   0.5748
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1.432 on 20 degrees of freedom
Multiple R-Squared: 0.09559,    Adjusted R-squared: -0.04007
F-statistic: 0.7046 on 3 and 20 DF,  p-value: 0.5604
```

If we cannot reject the overall null hypothesis (no difference between diets), then there is no evidence against the assumption of homogeneous variance.

Two-way ANOVA

Consider an experiment in which two factors A and B are allowed to vary, e.g., type of drug and dosage.
Let factor A have I levels and factor B have J levels.
Y_{ijk}: k-th observation ($k = 1, \cdots, n_{ij}$) and $\sum_i \sum_j n_{ij} = n$ observations. We have IJ independent samples, each from a different population. We have

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk}, \quad i = 1, \cdots, I, \quad j = 1, \cdots, J, \quad k = 1, \cdots, n_{ij},$$

where the $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$.
Similar to one-way ANOVA, we can rewrite this as

$$Y = X\beta + \epsilon,$$

where X is $n \times IJ$ matrix of rank IJ and $\beta = (\mu_{11}, \mu_{12}, \cdots, \mu_{IJ})'$.
The least squares estimates of μ_{ij} are $\hat{\mu}_{ij} = \bar{Y}_{ij.}$.

Testing $H_0 : \mu_{ij} = \mu, \forall i, j$ versus H_a : at least one μ_{ij} is different.
Corresponding test statistic is

$$T = \frac{\sum_i \sum_j n_{ij}(\bar{Y}_{ij.} - \bar{Y}_{...})^2/(IJ-1)}{\sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{ij.})^2/(n-IJ)} \sim F_{(IJ-1),(n-IJ)} \text{under } H_0.$$

Modeling interactions

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}, \quad (***)$$

where $(\alpha\beta)_{ij}$ is part of the mean response not attributable to the additive effect of α_i and β_j.

- Balanced layout: $n_{ij} = n^* \ \forall \ i, j$.

- If $n_{ij} = n^* = 1$, we have as many observations as the number of parameters in (***) [number of parameters equal $1 + (I-1) + (J-1) + (I-1)(J-1) = IJ$]. Can estimate the parameters but cannot make any further inference, e.g., testing. If we set $(\alpha\beta)_{ij} = 0$, then can make inference for α_i and β_j.

If we have $n_{ij} > 1$,then we can test for the interactions using the following hypothesis

$$\begin{aligned} H_o &: Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}, \\ H_a &: Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}. \end{aligned}$$

Example. 48 rats were allocated to 3 poisons (I, II, III) and 4 treatments (A, B, C, D). Outcome is the survival time in 10 hours with $n_{ij} = 4$.

```
> par(mfcol = c(2, 2))
> plot(time ~ poison + treat, data = rats)
> interaction.plot(treat, poison, time)
> interaction.plot(poison, treat, time)
```

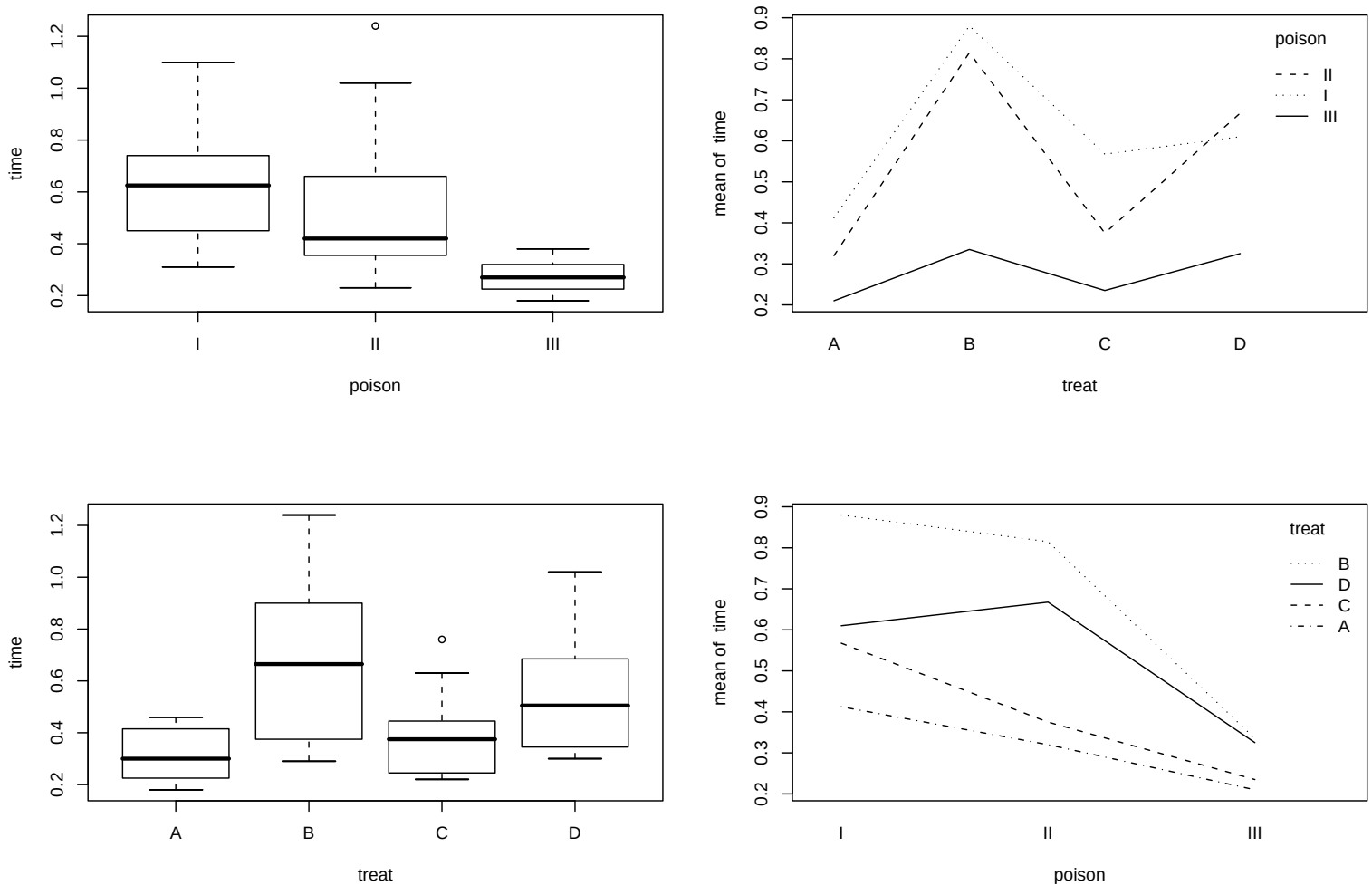

Figure: Do the interaction plots look parallel?

```
> lm0 = lm(time ~ poison + treat)
> lm1 = lm(time ~ poison * treat)
> anova(lm0, lm1)
Analysis of Variance Table

Model 1: time ~ poison + treat
Model 2: time ~ poison * treat
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     42 1.05086
2     36 0.80073  6   0.25014 1.8743 0.1123
```

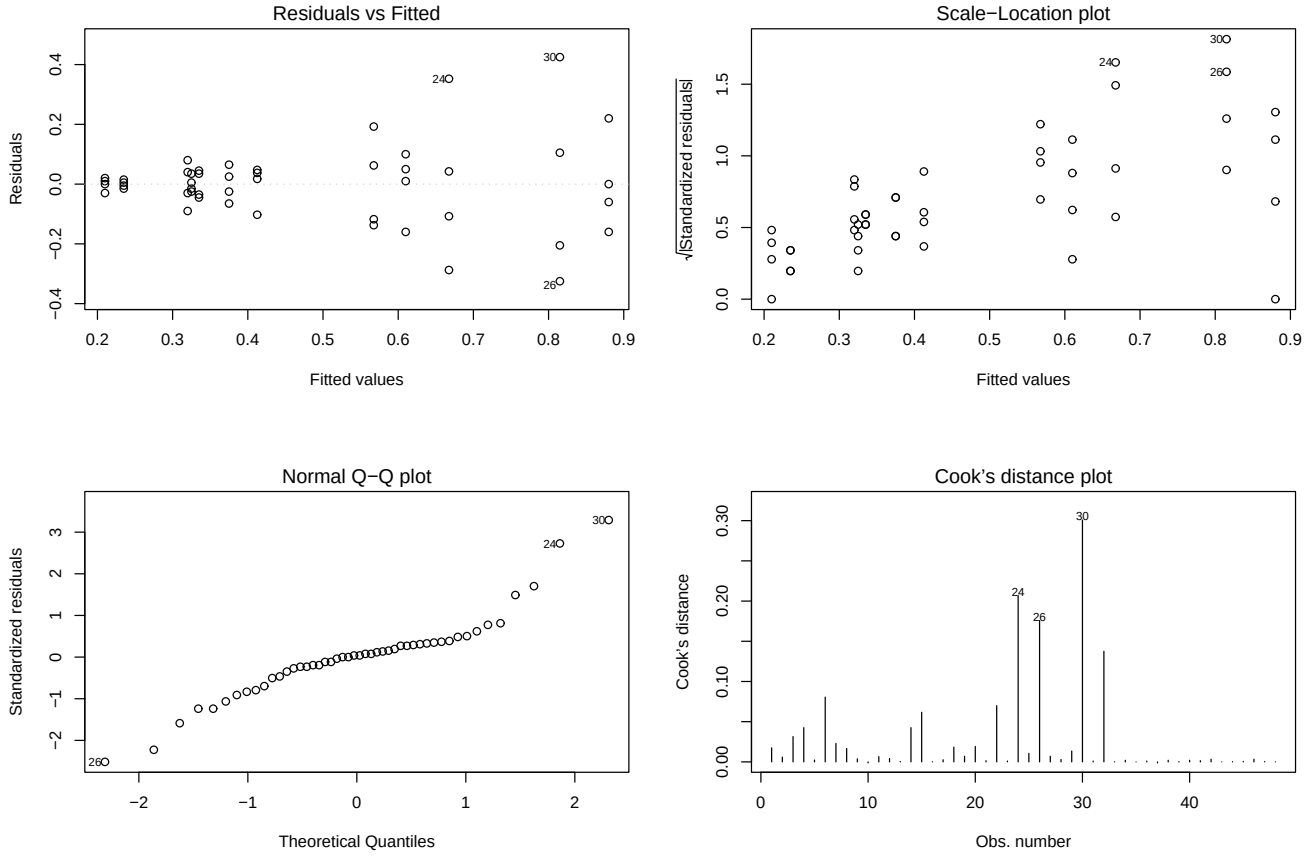

Figure: Are the linear model assumptions satisfied?

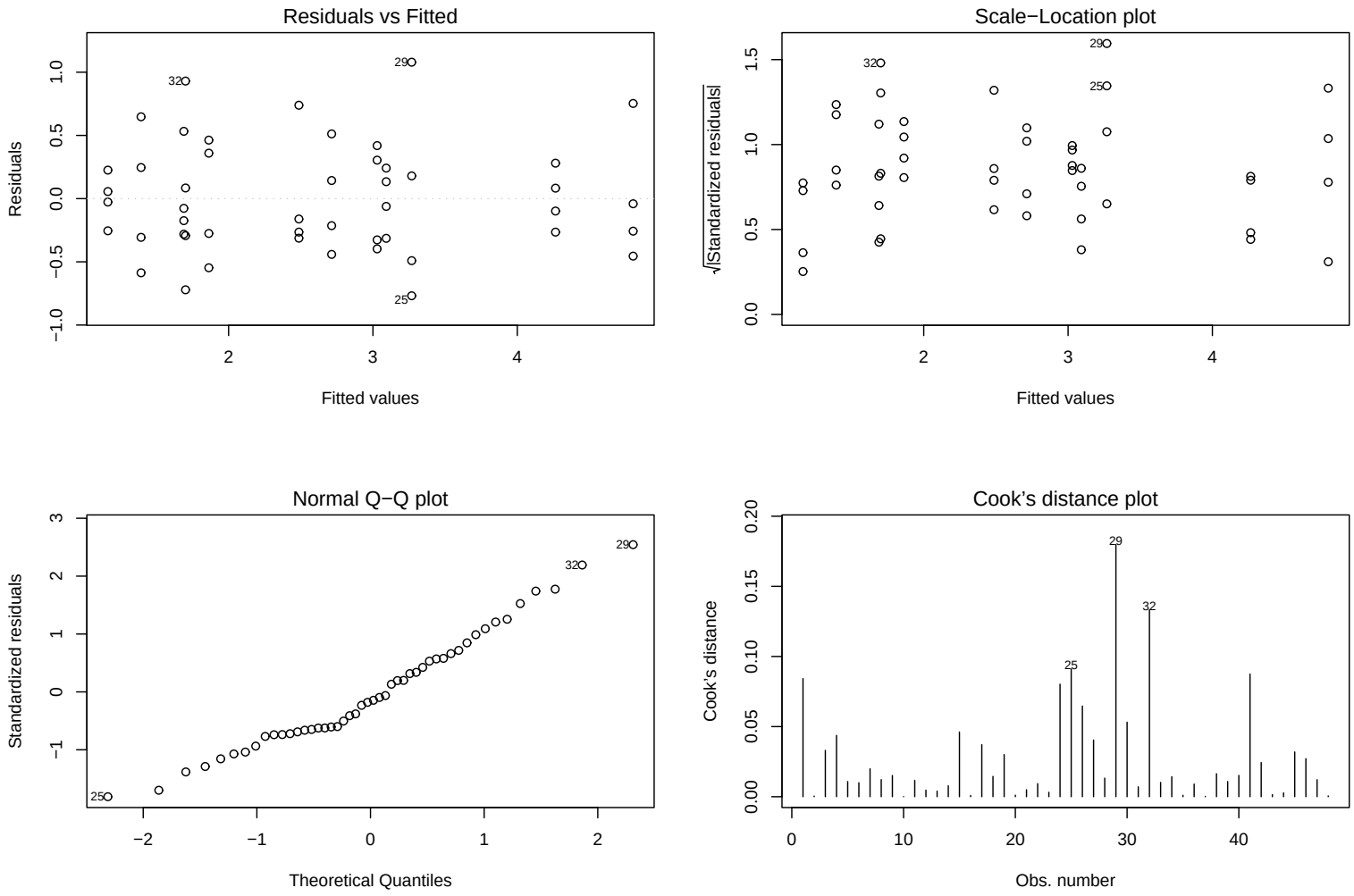

Figure: Diagnostic plots after reciprocal transformation.

```
> lm2 = lm(1/time ~ poison * treat)
> lm3 = lm(1/time ~ poison + treat)
> anova(lm3, lm2)
Analysis of Variance Table

Model 1: 1/time ~ poison + treat
Model 2: 1/time ~ poison * treat
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     42 10.2139
2     36  8.6431  6    1.5708 1.0904 0.3867
```

Final model

```
> summary(lm3)
Call: lm(formula = 1/time ~ poison + treat)
Residuals:
     Min       1Q   Median       3Q      Max
-0.82757 -0.37619  0.02116  0.27568  1.18153

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
 (Intercept)   2.6977     0.1744  15.473  < 2e-16 ***
 poisonII      0.4686     0.1744   2.688  0.01026 *
 poisonIII     1.9964     0.1744  11.451 1.69e-14 ***
 treatB       -1.6574     0.2013  -8.233 2.66e-10 ***
 treatC       -0.5721     0.2013  -2.842  0.00689 **
 treatD       -1.3583     0.2013  -6.747 3.35e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
 Residual standard error: 0.4931 on 42 degrees of freedom
 Multiple R-Squared: 0.8441,     Adjusted R-squared: 0.8255
 F-statistic: 45.47 on 5 and 42 DF,  p-value: 6.974e-16
```

In class exercise: Handout on how to specify contrasts in R.