Session 6 - Least Squares and Statistical Inference for Linear Regression Models

Brian Leroux

Wednesday, February 20, 2019

Outline

1. Least Squares and Interpretation of Regression Coefficients
2. Statistical Inference for Linear Regression Models
3. Checking Assumptions for Linear Regression Models

1. Least Squares and Interpretation of Regression Coefficients

- Least-squares
- Interpretation of regression coefficients with multiple predictors

Least-Squares

The usual method of fitting a regression model to data is called "least-squares" (LS). The LS estimates of the regression coefficients are the values that minimize the sum of squares of residuals.

For a simple linear regression model with just one predictor, $Y = \alpha + \beta X + \epsilon$, the LS method finds estimates for the intercept ($\hat{\alpha}$) and slope ($\hat{\beta}$) thatminimize the sum of squares of the residuals.

The residuals are the differences between observations and the predicted (`fitted`) values from the model, i.e., $e_i = Y_i - (\hat{\alpha} + \hat{\beta} X_i)$.

One way to think of it is that the estimates of slope and intercept determine a line that best fits the data according to a certain criterion.

There are other criteria besides the sum of squares of residuals (e.g., sum of absolute value of residuals), which lead to alternative methods of fitting regression models. However, the LS method is by far the most popular and has been used for over 200 years. One of the reasons for the popularity of LS is that it is a special case of the *maximum-likelihood* (ML) method of estimation (we will see ML next week).

Formulas for Least-Squares estimates for simple linear regression

A linear regression model with a single predictor variable is called "simple linear regression."

In this case, there are simple explicit formulas for the LS estimators of the parameters α and β:

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

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

With multiple predictor variables, the formulas involve matrix algebra.

Derivation of LS estimates for simple linear regression

The LS estimates minimize the sum of squares of residuals:

$$\hat{\beta} = \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2$$

To minimize set derivatives to 0:

For α:

$$0 = -\sum_{i=1}^{n} 2(Y_i - \hat{\alpha} - \hat{\beta} X_i)$$

For β:

$$0 = -\sum_{i=1}^{n} 2X_i(Y_i - \hat{\alpha} - \hat{\beta} X_i)$$

Derivation (continued)

Solve first equation for $\hat{\alpha}$:

$$\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\beta}X_i)$$

Second equation becomes

$$\sum_{i=1}^{n} X_i Y_i = n\hat{\alpha} - \hat{\beta}\sum_{i=1}^{n} X_i^2$$

Now solve for $\hat{\beta}$ and plug in formula for $\hat{\alpha}$ to get

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

(a few steps of algebra not shown)

For more on derivations of formulas for least-squares, see, for example, *Applied Linear Regreesion* by Weisberg, *Applied Regression Analysis* by Draper and Smith, *Design and Analysis of Experiments* by Montgomery.

LS estimates for models with multiple predictors

A general linear regression model can be written as

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots \beta_p X_{p-1,i} + \epsilon_i$$

where

Y_i is the response variable for the ith observation

$X_1, X_2, \ldots, X_{p-1}$ are $p-1$ predictor variables for the ith observation

β_0 is the intercept

β_i is the regression coefficient for the predictor variable X_i $(i = 1, \ldots, p-1)$

ϵ_i is the error, a random variable with mean 0 and variance σ^2

LS estimates for the general linear regression model

The model is written in vector notation as $\underline{Y} = \underline{\beta}'\mathbf{X} + \underline{\epsilon}$, where

$\underline{Y}$ is the $n \times 1$ vector of responses $\underline{Y} = (Y_1, \ldots, Y_n)$

$\underline{\beta}$ is the vector of coefficients $\beta = (\beta_0, \beta_1, \ldots, \beta_{p-1})$

$\underline{\epsilon}$ is the vector of errors $\underline{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$

$\mathbf{X}$ is the $n \times p$ matrix of predictors with jth column $(X_{j1}, \ldots, X_{jn})$ (for $j = 0$, $X_{0i} = 1$ represents the intercept)

The vector of LS estimates is

$$\underline{\hat{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\underline{\mathbf{Y}}$$

Derivation: a matrix version of the derivation shown above for simple linear regression.

Interpretation of regression coefficients with multiple predictors

We have seen with the Sales data (HW 6) that regression coefficient estimates can be counter-intuitive when there are multiple predictor variables in the model.

The key point is the following:

> *The interpretation of a regression coefficient in a model with multiple predictors is different than the interpretation of a regression coefficient in a simple linear regression model (with just one predictor).*

The interpretation of each coefficient in a model has to take account of the presence of the *other predictor variables in the model.*

The FEV data also illustrates this phenomenon.

Example - FEV data

```
f=read.csv("fev.csv")
summary(lm(fev ~ smoke, data=f))$coef
```

```
             Estimate Std. Error   t value      Pr(>|t|)
(Intercept) 2.5661426 0.03466043 74.036674 1.487335e-319
smoke       0.7107189 0.10994262  6.464453  1.992846e-10
```

```
summary(lm(fev ~ smoke + age, data=f))$coef
```

```
              Estimate  Std. Error   t value      Pr(>|t|)
(Intercept)  0.3673730 0.081435716  4.511203  7.647680e-06
smoke       -0.2089949 0.080745337 -2.588321  9.859773e-03
age          0.2306046 0.008184372 28.176209 8.279537e-115
```

First model: The coefficient of 'smoke' is > 0 and statistically significant

Second model: The coefficient of 'smoke' is < 0 and not statistically significant

Interpretation

In the first model the coefficient of `smoke` is the average difference in fev comparing smokers with non-smokers.

In the second model the coefficient of 'smoke' is interpreted as the average difference in fev comparing a smoker with a non-smoker **of the same age**.

In effect we have "controlled" age (or adjusted for age). We say that age is "confounding" the association between fev and smoking. By adjusting for age, we remove its confounding effect.

Justification for interpretation of coefficients

The FEV model is

$$E(Y|X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

where Y is FEV, X_1 is the indicator of being a smoker and X_2 is age. (Note: subscript i is left out for simplicity)

The mean Y for a smoker and non-smoker of a given age A are:

$$E(Y|X_1 = 1, X_2 = A) = \beta_0 + \beta_1 1 + \beta_2 A$$

and

$$E(Y|X_1 = 0, X_2 = A) = \beta_0 + \beta_1 0 + \beta_2 A$$

and the difference is β_1, the coefficient of the smoking indicator **in the model that includes age**.

Example - ToothGrowth data

```
summary(lm(len ~ dose, data=ToothGrowth))$coef
```

```
            Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 7.422500  1.2600826  5.890487 2.064211e-07
dose        9.763571  0.9525329 10.250114 1.232698e-14
```

```
summary(lm(len ~ dose + supp, data=ToothGrowth))$coef
```

```
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept)  9.272500  1.2823649  7.230781 1.312335e-09
dose         9.763571  0.8768343 11.135025 6.313519e-16
suppVC      -3.700000  1.0936045 -3.383307 1.300662e-03
```

In this case, the coefficient of `dose` does **not** change when we add `supp` to the model. (But note that the SE of the `dose` coefficient changes.)

Why is this example different?

It depends on whether predictor variables are correlated

The coefficient of `smoke` changes when `age` is added to the model because `smoke` and `age` are correlated in the sample: smokers tend to be older than non-smokers.

But the variables `dose` and `supp` are uncorrelated, i.e., the distribution of `dose` is the same for both levels of `supp`. There can be no confounding effect if variables are uncorrelated.

Why Randomization is Important

Confounding is why we use randomization in experiments. By randomizing treatments we make them uncorrelated with other variables and avoid biased estimates of treatment effects due to confounding.

To assess the effect of a treatment on FEV we would want to randomly assign it to the children so that it would not be correlated with age (or gender or any other variable).

Note: randomization produces balanced treatment groups on average. However, if the sample size is not large we might have some correlation between treatment and certain patient characteristics *just by chance*. So sometimes we perform adjusted analyses using regression as a sensitivity analysis.

Why the LS Estimates Depend on Correlation Between Predictors

It is because of the X'X matrix in the LS formula. For the FEV data:

```
Y=f$fev
n=nrow(f)
X=cbind(rep(1,n),f$smoke,f$age)
# X'X inverse
solve(t(X) %*% X)
```

```
              [,1]          [,2]          [,3]
[1,]  0.020770005  0.0062798400 -0.0020002729
[2,]  0.006279840  0.0204193377 -0.0008366855
[3,] -0.002000273 -0.0008366855  0.0002097865
```

```
# beta-hat
solve(t(X) %*% X) %*% t(X) %*% Y
```

```
           [,1]
[1,]  0.3673730
[2,] -0.2089949
[3,]  0.2306046
```

Example - the Tooth Growth data

Note the 0 values in the off-diagonal elements of $(X'X)^{-1}$, corresponding to dose and supp.

```
Y=ToothGrowth$len
n=nrow(ToothGrowth)
X=cbind(rep(1,n),ToothGrowth$dose,ToothGrowth$supp=="OJ")
# X'X inverse
solve(t(X) %*% X)
```

```
            [,1]        [,2]        [,3]
[1,]  0.09166667 -0.05000000 -0.03333333
[2,] -0.05000000  0.04285714  0.00000000
[3,] -0.03333333  0.00000000  0.06666667
```

```
# beta-hat
solve(t(X) %*% X) %*% t(X) %*% Y
```

```
         [,1]
[1,] 5.572500
[2,] 9.763571
[3,] 3.700000
```

Summary

The interpretation of regression coefficients is in general quite difficult. This is particularly true when there are multiple predictor variables in the model that are correlated with each other. (As we will see correlation between predictors also impacts the standard error estimates of the coefficient estimates.)

The term "multi-collinearity" refers to predictors that are correlated with. In extreme cases (when predictors are very strongly correlated) it can produce very misleading results.

If results can change so much when we add new predictors to a model **how are we to decide which predictors should be included in our model**. This is a very difficult question and leads to the subject of Model Selection which we will cover later. The strategies used depend on the context, eg, randomized experiment, analysis of associations in observational data, or prediction.

2. Statistical Inference for Regression Models

- Inference for linear regression coefficients
- Composite hypotheses and F-tests
- Regression vs ANOVA revisited

Statistical Inference for Linear Regression

Summary:

t-test for individual coefficients

F-test for simultaneous tests of sets of coefficients ("composite hypothesis")

Estimating the sampling variance of $\underline{\hat{\beta}}$

The LS estimate of the vector β is

$$\underline{\hat{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\underline{\mathbf{Y}}$$

By the properties of covariance matrices*

$$\text{cov}(\underline{\hat{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{cov}(\underline{\mathbf{Y}})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

The variances of the individual coefficient estimates are on the diagonal of this covariance matrix.

To estimate the covariance matrix we substitute an estimate for the error variance σ^2.

- Note: for any matrix A and random vector $\underline{Y}$, $\text{cov}(AY) = A\text{cov}(\underline{Y})A'$.

Estimating the error variance

The error variance is the variance of the errors ϵ_i, so we estimate it using the sum of squares of residuals:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-p}$$

The residuals have sample mean 0 (by definition), so $\hat{\sigma}^2$ is almost the sample variance of the residuals, except that it has $n-p$ in the denominator rather than $n-1$. The subtraction of p in the denominator reflects the fact that p parameters (the regression coefficients) have been estimated. For example, with 1 predictor, the value becomes $n-2$. Although it is not exactly equal to the sample variance of the residuals we sometimes call $\hat{\sigma}^2$ the *residual variance* and $\hat{\sigma}$ the *residual SD*.

Estimated standard errors for simple linear regression

For the simple linear regression model with one predictor, the covariance matrix simplifies and we get the following formulas for calculating estimated standard errors of the LS estimates:

$$\mathrm{SE}(\hat{\beta}) = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

$$\mathrm{SE}(\hat{\alpha}) = \frac{\hat{\sigma}\sqrt{\sum X_i^2}}{\sqrt{n\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

Use of SE($\hat{\beta}$) in experimental design

$$\text{SE}(\hat{\beta}) = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

This formula is important for design of experiments because it helps us to determine how large a sample size we need. The forumula tells us that the SE of our slope estimate depends on 2 things:

1. The error SD $\hat{\sigma}$ is in the numerator. Hence, if we can reduce the SD of our errors (not easy to do!) we can make our slope estimate more precise.
2. The sums of squares of the X_i's from their mean is in the denominator. This actually has two components: (i) how spread out the X_i's are, and (ii) the sample size n. Therefore, we will make our slope estimate more precise by (i) having X values more spread out, and (ii) having more of them. (We will explore this in an Exercise today.)

When there are multiple predictors, the SEs depend also on the sums of cross-products of different X variables (the off-diagonal terms of $\mathbf{X}'\mathbf{X}$). This is why SEs can change when a variable is added to a model even if the coefficient estimate does not change (see Tooth-Growth example above).

Example - FEV ~ age

```
f=read.csv("fev.csv")
n=nrow(f)
Y=f$fev
X=cbind(rep(1,n),f$age)
p=ncol(X)
beta=solve(t(X) %*% X) %*% t(X) %*% Y
muhat= X %*% beta
sigmasq=sum((Y - muhat)^2)/(n-p)
covbeta=sigmasq * solve(t(X) %*% X)
sebeta=sqrt(diag(covbeta))
print(covbeta)
```

```
               [,1]          [,2]
[1,]  0.0060676923 -5.613832e-04
[2,] -0.0005613832  5.652727e-05
```

```
data.frame(beta,sebeta)
```

```
       beta      sebeta
1 0.4316481 0.077895393
2 0.2220410 0.007518462
```

Compare to summary of 'lm' function

```
summary(lm(fev ~ age, data=f))
```

```
Call:
lm(formula = fev ~ age, data = f)

Residuals:
     Min       1Q   Median       3Q      Max
-1.57539 -0.34567 -0.04989  0.32124  2.12786

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.431648   0.077895   5.541 4.36e-08 ***
age         0.222041   0.007518  29.533  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5675 on 652 degrees of freedom
Multiple R-squared:  0.5722,    Adjusted R-squared:  0.5716
F-statistic: 872.2 on 1 and 652 DF,  p-value: < 2.2e-16
```

Example - FEV ~ age + smoke

```
X=cbind(rep(1,n),f$age,f$smoke)
p=ncol(X)
beta=solve(t(X) %*% X) %*% t(X) %*% Y
muhat= X %*% beta
sigmasq=sum((Y - muhat)^2)/(n-p)
covbeta=sigmasq * solve(t(X) %*% X)
sebeta=sqrt(diag(covbeta))
print(covbeta)
```

```
               [,1]          [,2]          [,3]
[1,]  0.0066317759 -6.386788e-04  0.0020051267
[2,] -0.0006386788  6.698394e-05 -0.0002671502
[3,]  0.0020051267 -2.671502e-04  0.0065198094
```

```
data.frame(beta,sebeta)
```

```
        beta      sebeta
1  0.3673730 0.081435716
2  0.2306046 0.008184372
3 -0.2089949 0.080745337
```

compare to summary...

```
summary(lm(fev ~ age + smoke, data=f))
```

```
Call:
lm(formula = fev ~ age + smoke, data = f)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6653 -0.3564 -0.0508  0.3494  2.0894

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.367373   0.081436   4.511 7.65e-06 ***
age          0.230605   0.008184  28.176  < 2e-16 ***
smoke       -0.208995   0.080745  -2.588  0.00986 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5651 on 651 degrees of freedom
Multiple R-squared:  0.5766,    Adjusted R-squared:  0.5753
F-statistic: 443.3 on 2 and 651 DF,  p-value: < 2.2e-16
```

Hypothesis testing for regression coefficients

For each coefficient in a regression model we can test the null hypothesis that the coefficient is equal to 0 using a t-statistic defined as the parameter estimate divided by its standard error, i.e., $T = \hat{\beta}/\text{SE}(\hat{\beta})$. This is referred to a t_{n-p-1} distribution to assess statistical significance.

Example: in the model E(FEV)$= \alpha + \beta$ Age, test $H_0 : \beta = 0$.

```
z=beta[2]/sebeta[2]
p=2*(1-pt(abs(z),df=n-2))
data.frame(z,p)
```

```
          z p
1 28.17621 0
```

We reject the null hypothesis and conclude that there is evidence for an association between age and FEV.

Hypothesis testing for the intercept

We can also test the null hypothesis that the intercept is equal to 0, i.e., $H_0 : \alpha = 0$.

```
z=beta[1]/sebeta[1]
p=2*(1-pt(abs(z),df=n-2))
data.frame(z,p)
```

```
          z             p
1 4.511203 7.645715e-06
```

This hypothesis is not usually of scientific interest. In this case the intercept is interpreted as the mean FEV at age 0.

Composite hypotheses

A "composite" hypothesis is one that involves two or more parameters in a regression model. These have many uses.

Example 1. Testing the effect of a categorical treatment variable with 3 or more levels ('A', 'B', and 'C').

Model: $\mathrm{E}(Y) = \beta_0 + \beta_1 I\{\text{Group A}\} + \beta_2 I\{\text{Group B}\}$

Null hypothesis: $H_0 : \beta_1 = \beta_2 = 0$.

This null hypothesis says that there are no differences between the 3 groups. Note that there are indicator variables for only 2 of the 3 levels of the group variable.

Example 2. Testing the overall effect of dose in a factorial experiment (e.g., Tooth-Growth).

Model: $\mathrm{E}(Y) = \beta_0 + \beta_1 I\{\text{Group VC}\} + \beta_2 X + \beta_3 I\{\text{Group VC}\} X$

Null hypothesis: $H_0 : \beta_2 = \beta_3 = 0$.

The meaning of this null hypothesis is that dose has no effect at all on the response, tooth length, neither on the level of response or an interactive effect with 'supp'.

More examples of composite hypotheses

Example 3. Testing for no effect of any factor (e.g., NPK experiment).

Model: $\mathrm{E}(Y) = \beta_0 + \beta_1 N + \beta_2 P + \beta_3 K$

Null hypothesis: $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$.

This null hypothesis implies that none of the factors has an effect.

Example 4. Testing for no interactions (e.g., NPK experiment).

Model: $\mathrm{E}(Y) = \beta_0 + \beta_1 N + \beta_2 P + \beta_3 K + \beta_4 NP + \beta_5 NK + \beta_6 PK + \beta_7 NPK$

Null hypothesis: $H_0 : \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$.

This null hypothesis says that there are no 2-way or 3-way interactions. However, the null hypothesis still allows for main effects of the 3 factors.

The F-test

To test a null hypothesis we compare the sums of squares of residuals for the *full* model, which includes the coefficients being tested, with the *reduced* model, which has those coefficients set to 0.

If we define SSE_0 and SSE_1 as the sums of squares of residuals for the reduced and full models, respectively, the F-statistic is defined as:

$$F = \frac{(SSE_0 - SSE_1)/(p_1 - p_0)}{SSE_1/(n - p_1)}$$

The F-statistic is referred to the $F_{p_1 - p_0, n - p_1}$ distribution for calculation of the p-value.

Estimation of the Error Variance from the Full Model

The denominator is an estimate of the error variance using the **full**. Note that the full model more fully captures the effects of the predictors and so may give a better estimate of pure error variance, compared with the reduced model.

Recall that the error variance is estimated using the MSE which is the residual sum of squares divided by its df, which for the full model is:

$$\hat{\sigma}^2 = \text{MSE}_1 = \text{SSE}_1/(n - p_1)$$

Example

Testing the overall effect of dose in the Tooth-Growth experiment

Model: $\mathsf{E}(Y) = \beta_0 + \beta_1 I\{\text{Group VC}\} + \beta_2 X + \beta_3 I\{\text{Group VC}\}X$

Null hypothesis: $H_0 : \beta_2 = \beta_3 = 0$.

Reduced Model: $\mathsf{E}(Y) = \beta_0 + \beta_1 I\{\text{Group VC}\}$

The ANOVA tables for the full and reduced models

```
anova(lm(len ~ supp*dose, data=ToothGrowth))
```

```
Analysis of Variance Table

Response: len
          Df  Sum Sq Mean Sq  F value    Pr(>F)
supp       1  205.35  205.35  12.3170 0.0008936 ***
dose       1 2224.30 2224.30 133.4151 < 2.2e-16 ***
supp:dose  1   88.92   88.92   5.3335 0.0246314 *
Residuals 56  933.63   16.67
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(lm(len ~ supp, data=ToothGrowth))
```

```
Analysis of Variance Table

Response: len
          Df Sum Sq Mean Sq F value  Pr(>F)
supp       1  205.4  205.35  3.6683 0.06039 .
Residuals 58 3246.9   55.98
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$
\begin{aligned}
F &= \frac{(SSE(\text{Reduced}) - SSE(\text{Full}))/(df_R - df_F)}{SSE(\text{Full})/df_F} \\
&= \frac{(3246.9 - 933.63)/(58 - 56)}{933.63/56} \\
&= 69.37
\end{aligned}
$$

The p-value is obtained by the tail probability for the value 69.37 in the F-distribution with 2 numerator df and 56 denominator df.

```
1-pf(69.37,2,56)
```

```
[1] 6.661338e-16
```

We would reject the null hypothesis of no dose effect ($p < 0.001$).

Performing the F-test with the 'anova' function

We can calculate the F-statistic using the 'anova' function to compare the two models.

```
full=lm(len ~ supp*dose, data=ToothGrowth)
reduced=lm(len ~ supp, data=ToothGrowth)
anova(reduced,full)
```

```
Analysis of Variance Table

Model 1: len ~ supp
Model 2: len ~ supp * dose
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     58 3246.9
2     56  933.6  2    2313.2 69.374 6.982e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note: we are not doing ANOVA (the method) here, it is linear regression, but we are using a function called "anova" to do the calculations.

Regression versus ANOVA revisited

For testing the effect of a categorical treatment variable with 3 or more levels.

Full Model: $\mathsf{E}(Y) = \beta_0 + \beta_1 I\{\text{Group A}\} + \beta_2 I\{\text{Group B}\}$

Null hypothesis: $H_0 : \beta_1 = \beta_2 = 0$.

Reduced Model: $\mathsf{E}(Y) = \beta_0$.

```
full=lm(len ~ factor(dose), data=ToothGrowth)
reduced=lm(len ~ 1, data=ToothGrowth)
anova(reduced,full)
```

```
Analysis of Variance Table

Model 1: len ~ 1
Model 2: len ~ factor(dose)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     59 3452.2
2     57 1025.8  2    2426.4 67.416 9.533e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We get the same result using ANOVA (i.e., the method ANOVA)

```
summary(aov(len ~ factor(dose), data=ToothGrowth))
```

```
             Df Sum Sq Mean Sq F value   Pr(>F)
factor(dose)  2   2426    1213   67.42 9.53e-16 ***
Residuals    57   1026      18
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The explanation is the ANOVA decomposition of total sum of squares into sum of squares between and within groups: $SS_Total = SS_B + SS_W$. But SS_W is the SSE for the full model and $SS_B = SS_Total - SS_B = SSE_0 - SSE_1$, the difference in SSE between reduced and full models, so

$$F = \frac{(SSE_0 - SSE_1)/2}{SSE_1/n - 3} = \frac{MS_B}{MSE_1}$$

Summary: 1-way ANOVA gives the same result as regression with indicator ("dummy") variables.

Reparametrization of regression models

We have seen how comparison of 3 or more group means with ANOVA can be done using regression with indicator ("dummy") variables to represent the groups.

However, the **interpretation** of the regression model will depend on how we define the dummy variables, i.e, by which *parametrization* we use.

Consider the Iris data. We choose one group to be the reference group (here it is "virginica") and define indicator variables for the other two groups:

$$I_{\text{set}} = I\{\text{Species} = \text{"setosa"}\}$$

$$I_{\text{ver}} = I\{\text{Species} = \text{"versicolor"}\}$$

The model is

$$Y = \beta_0 + \beta_1 I_{\text{set}} + \beta_2 I_{\text{ver}} + \epsilon$$

ANOVA via multiple regression

The F-statistic (49.16) is the same for regression as for ANOVA.

```
iris$setosa=as.numeric(iris$Species=="setosa")
iris$versicolor=as.numeric(iris$Species=="versicolor")
summary(lm(Sepal.Width ~ setosa + versicolor, data=iris))
```

```
Call:
lm(formula = Sepal.Width ~ setosa + versicolor, data = iris)

Residuals:
   Min     1Q Median     3Q    Max
-1.128 -0.228  0.026  0.226  0.972

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.97400    0.04804  61.908  < 2e-16 ***
setosa       0.45400    0.06794   6.683 4.54e-10 ***
versicolor  -0.20400    0.06794  -3.003  0.00315 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3397 on 147 degrees of freedom
Multiple R-squared:  0.4008,    Adjusted R-squared:  0.3926
F-statistic: 49.16 on 2 and 147 DF,  p-value: < 2.2e-16
```

Using 'factor' for categorical variables

```
summary(lm(Sepal.Width ~ factor(Species), data=iris))
```

```
Call:
lm(formula = Sepal.Width ~ factor(Species), data = iris)

Residuals:
   Min     1Q Median     3Q    Max
-1.128 -0.228  0.026  0.226  0.972

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 3.42800    0.04804  71.359  < 2e-16 ***
factor(Species)versicolor -0.65800    0.06794  -9.685  < 2e-16 ***
factor(Species)virginica  -0.45400    0.06794  -6.683 4.54e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3397 on 147 degrees of freedom
Multiple R-squared:  0.4008,    Adjusted R-squared:  0.3926
F-statistic: 49.16 on 2 and 147 DF,  p-value: < 2.2e-16
```

How does 'factor' decide which dummy variables to use?

Note that the F-statistics are the same for both analyses. However, the models are parametrized differently: the 'factor' function chooses the "first" value as the reference category. First is defined as the lowest value if the values are numerical, or as the first alphabetically if the values are character strings.

Sometimes we may want to see the explicit comparisons with a given reference group (e.g., virginca), which we can do by defining the dummy variables ourselves. In the first analysis we see that mean sepal width is different for setosa than virginica (difference between means 0.454, SE=0.068, $p < 0.001$) and that there is also a difference between versicolor and virginica (difference=-0.204, SE=0.068, $p = 0.003$).

Using 'factor' we get different comparisons:

versicolor vs setosa: difference=-0.658, SE=0.068, $p < 0.001$

virginica vs setosa: difference=-0.454, SE=0.068, $p < 0.001$.

The overall F-test revisited

The overall F-test is a special case of an F-test for a composite hypothesis - it involves the comparison of the full model with the reduced model that contains only an intercept (called the "null" model).

```
full=lm(len ~ supp*dose, data=ToothGrowth)
null=lm(len ~ 1, data=ToothGrowth)
anova(null,full)
```

```
Analysis of Variance Table

Model 1: len ~ 1
Model 2: len ~ supp * dose
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     59 3452.2
2     56  933.6  3    2518.6 50.355 6.521e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(lm(len ~ supp*dose, data=ToothGrowth))$fstatistic
```

```
   value    numdf    dendf
50.35522  3.00000 56.00000
```

3. Checking assumptions for ANOVA and linear regression

- Independence
- Linearity
- Constant variance
- Normality or large sample size
- What to do when assumptions do not hold

The Independence Assumption

Independence of the observations is an assumption of linear regression and ANOVA and is typically verified through our knowledge of the study design and data collection mechanisms.

For example, the independence assumption for the vitamin C experiment would be valid if a "completely-randomized" experimental design was used. In this example, it means that the pigs were randomly assigned to the different doses of vitamin C completely randomly.

The NPK experiment used a randomized block design, in which treatments were assigned to plots of land that were pre-arranged into blocks of homogeneous plots. This invalidates the analysis presented earlie. A correct analysis would include block effects.

Example of Non-Independence: Nested Designs

In a factorial design the levels of two factors are **crossed** (i.e., all combinations of the levels of the factors are used in the experiment).

Sometimes factors are nested rather than crossed. For example, in a cluster-randomized experiment on smoking prevention, schools are randomly assigned to receive two versions of a smoking prevention program. The outcome is daily smoking of the students at the end of 12th grade. We think of this as randomly assigning treatments to a `cluster` of students (all students in a given school). The analysis for this type of design is more complicated than the analyses considered so far and will be considered later.

The Linearity Assumption for Linear Regression

In one respect, the assumption of linearity is necessary for linear regression.

However, we should never treat the assumption of linearity too strictly because it rarely if ever is exactly true. (Recall that "All models are wrong".)

Like other assumptions such as constant variance and normality, it is a matter of degree. Linear regression is widely used in the presence of some non-linearity in the relationship. In these settings it is important to interpret the results accordingly. For example, the regression coefficient β is thought of as an `average` slope, averaged over the range of the independent variable.

Constant Variance and Normality Assumptions

These two assumptions have similar effects on the validity of our inference for linear regression as they do on t-tests, ANOVA and linear regression.

In particular, non-constant variance can have a large effect on the performance of our significance test for our regression coefficient. The effect can be to make the test either overly conservative (type I error probability too small) or anti-conservative (type I error probability too large). Non-constant variance (like non-independence) is a problem **no matter how large the sample size**.

In contrast with the assumptions of independence and equal variances, normality of the error distribution is only necessary if the sample sizes are not sufficiently large.

Residuals

Methods for checking both the constant-variance and normality assumptions rely on the residuals. For linear regression the form of the residual is

$$e_i = Y_i - (\hat{\alpha} + \hat{\beta} X_i).$$

The residuals can be thought of as estimates of the errors in the model. We use the residuals to check constant variance and normality in very similar ways for t-tests, ANOVA and regression.

Residuals for the OJ data

Below are a few of the residuals from the OJ group to illustrate the calculations.

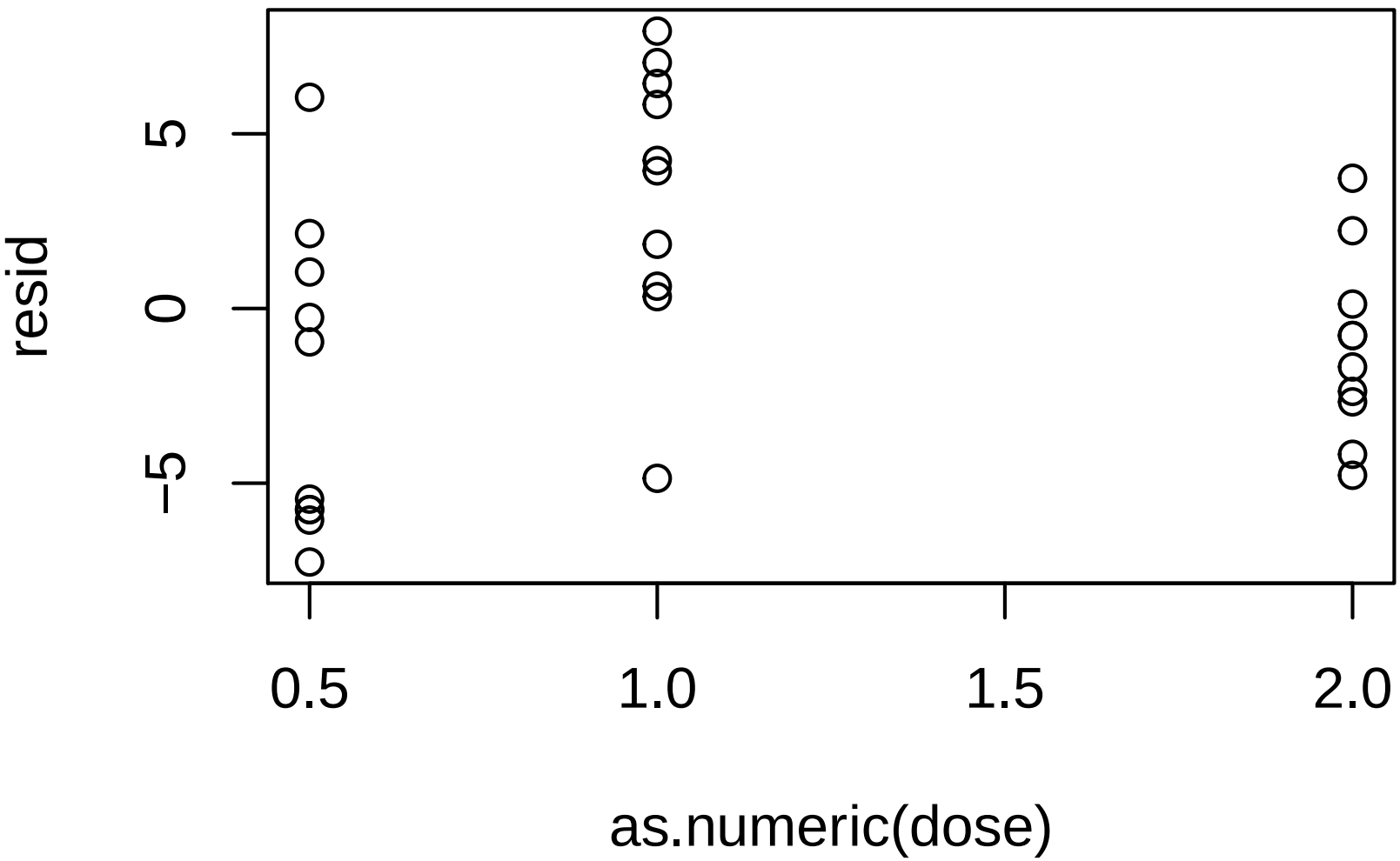
```
oj=subset(ToothGrowth,supp=="OJ")
oj$predicted = 11.55+7.881*oj$dose
oj$residual = oj$len - oj$predicted
oj[c(1:3,11:13,21:23),]
```

```
    len supp dose predicted residual
31 15.2   OJ  0.5   15.4905  -0.2905
32 21.5   OJ  0.5   15.4905   6.0095
33 17.6   OJ  0.5   15.4905   2.1095
41 19.7   OJ  1.0   19.4310   0.2690
42 23.3   OJ  1.0   19.4310   3.8690
43 23.6   OJ  1.0   19.4310   4.1690
51 25.5   OJ  2.0   27.3120  -1.8120
52 26.4   OJ  2.0   27.3120  -0.9120
53 22.4   OJ  2.0   27.3120  -4.9120
```

Using Residuals to Check Constant Variance

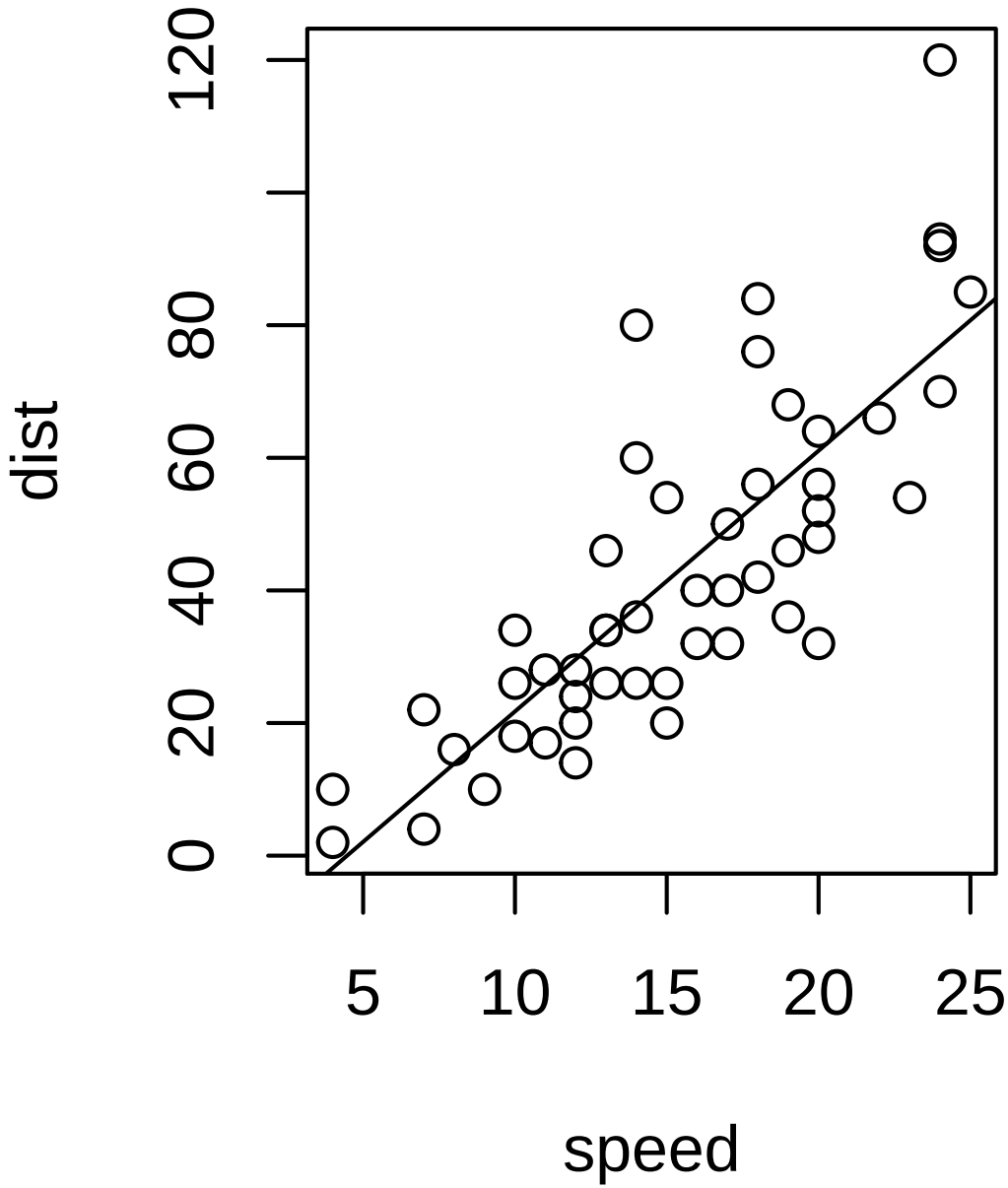
For the vitamin C experiment we can plot the residuals against dose (OJ group only). The plot suggests no clear evidence against constant variance.

```
fit=lm(len ~ dose, data=oj)
oj$resid = fit$residuals
plot(resid ~ as.numeric(dose), data=oj)
```



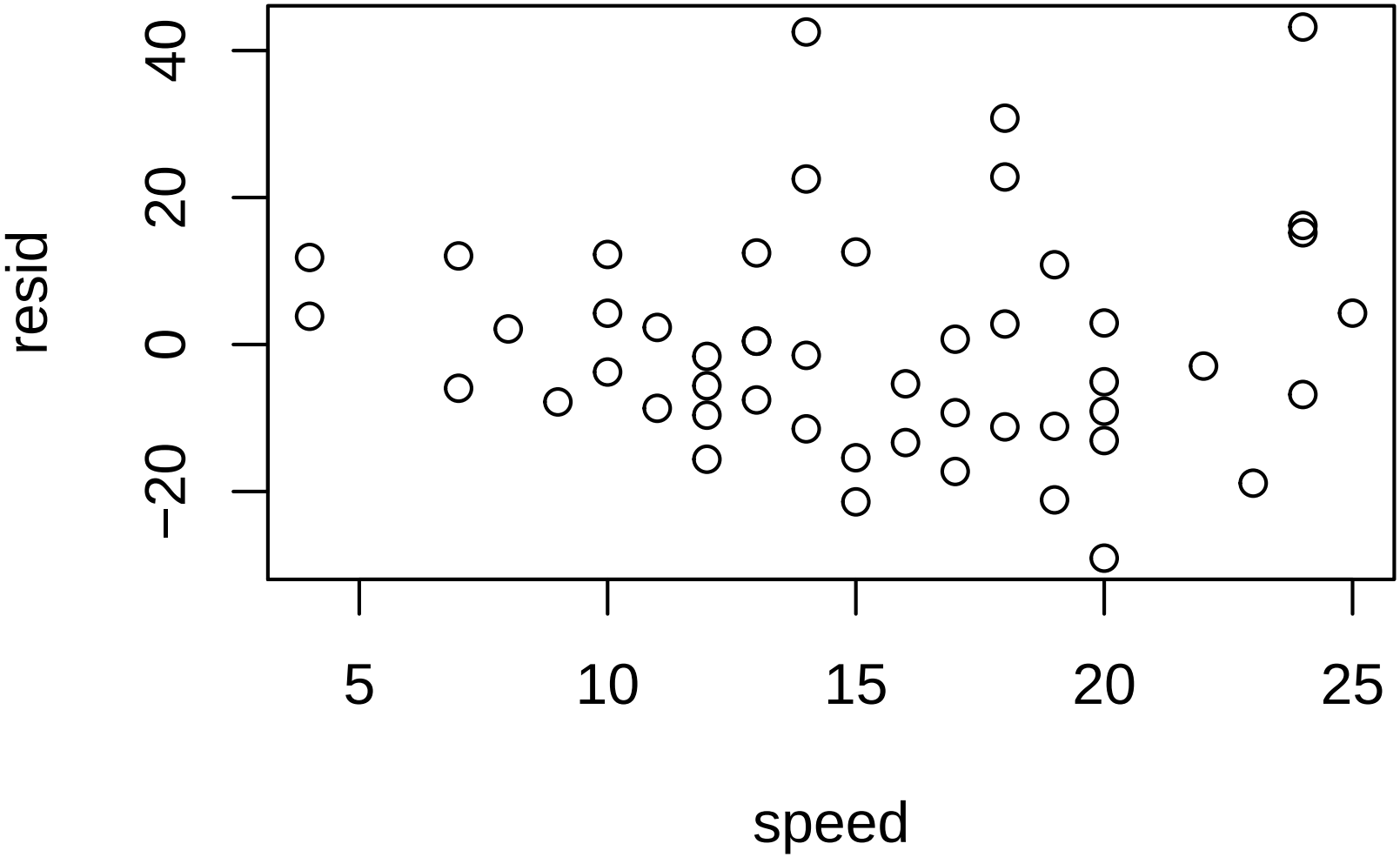
The ‘cars’ data

The cars data consists of data on speed (mph) and distance required to stop (ft) for 50 cars.



The residual plot for the 'cars' data

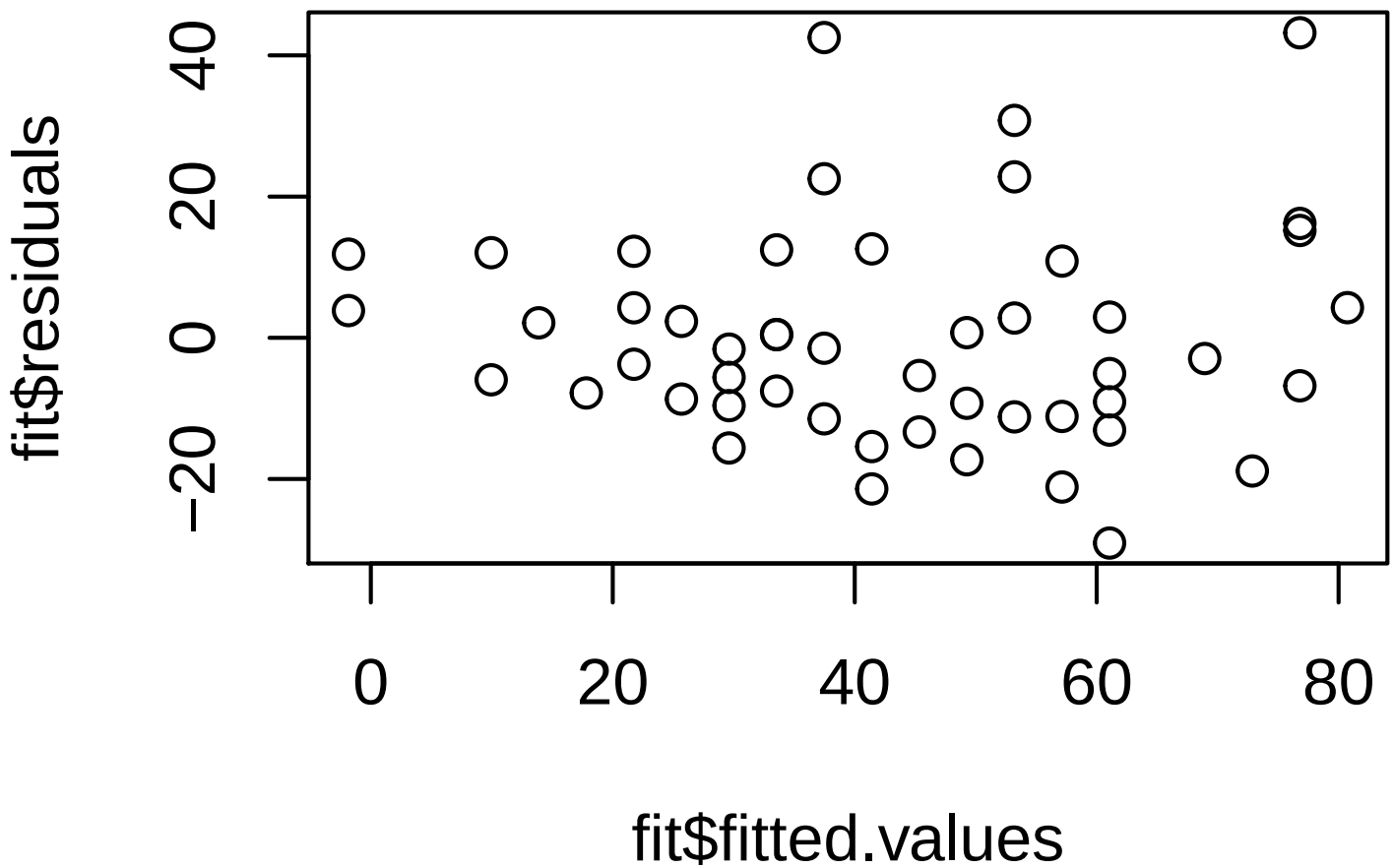
The plot of residuals against speed shows some evidence of non-constant variance - the residuals are more spread out for the larger speeds compared to the lower speeds.



Residuals versus fitted values plot

The cars data suggests that there is a relationship between the mean response and the error variance. One way to visualize this relationship is to plot the residuals (which represent variance) against the fitted values from the regression model (which represent the mean response). This is particularly useful for experiments with 2 or more factors.

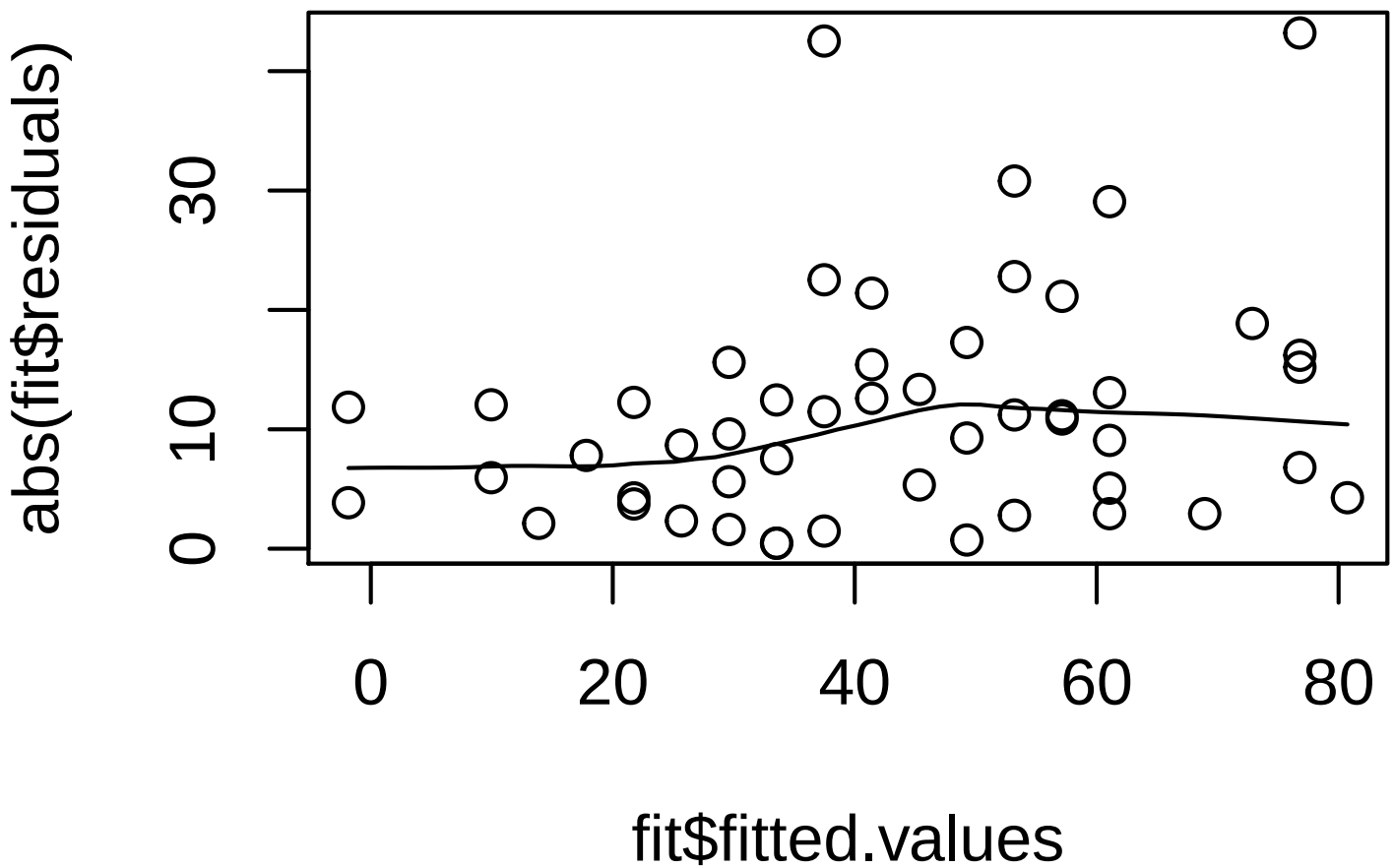
```
fit=lm(dist ~ speed, data=cars)
plot(fit$fitted.values, fit$residuals)
```



Smoothing the residual plot

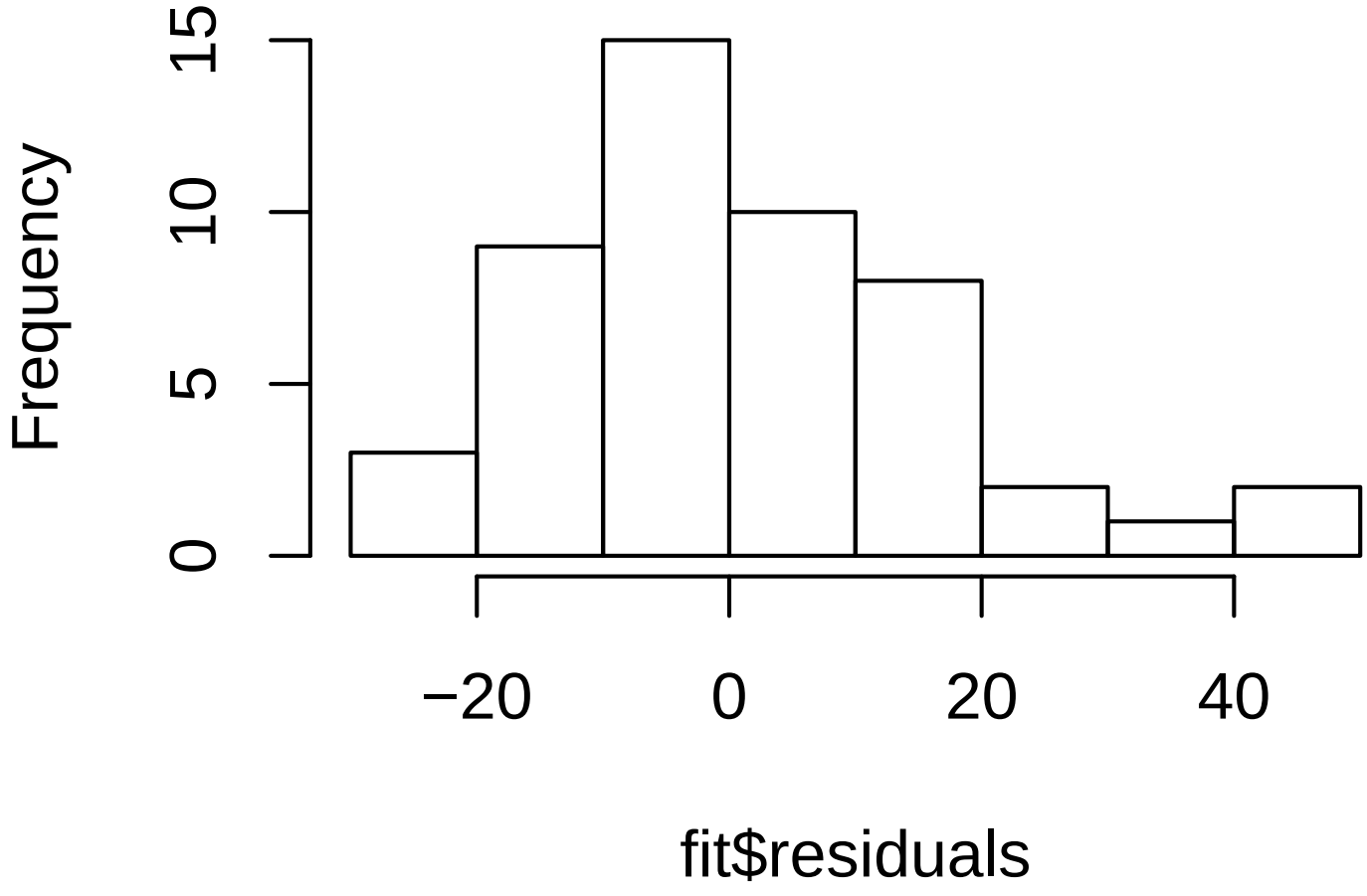
To illustrate the potential relationship between mean and variance we plot the absolute value of the residuals against the fitted values and apply a smoothing procedure to assess the trend. In this example, there is a suggestion of a positive relationship but it may be due to 2 large outliers (with absolute residuals above 40).

```
scatter.smooth(fit$fitted.values, abs(fit$residuals))
```



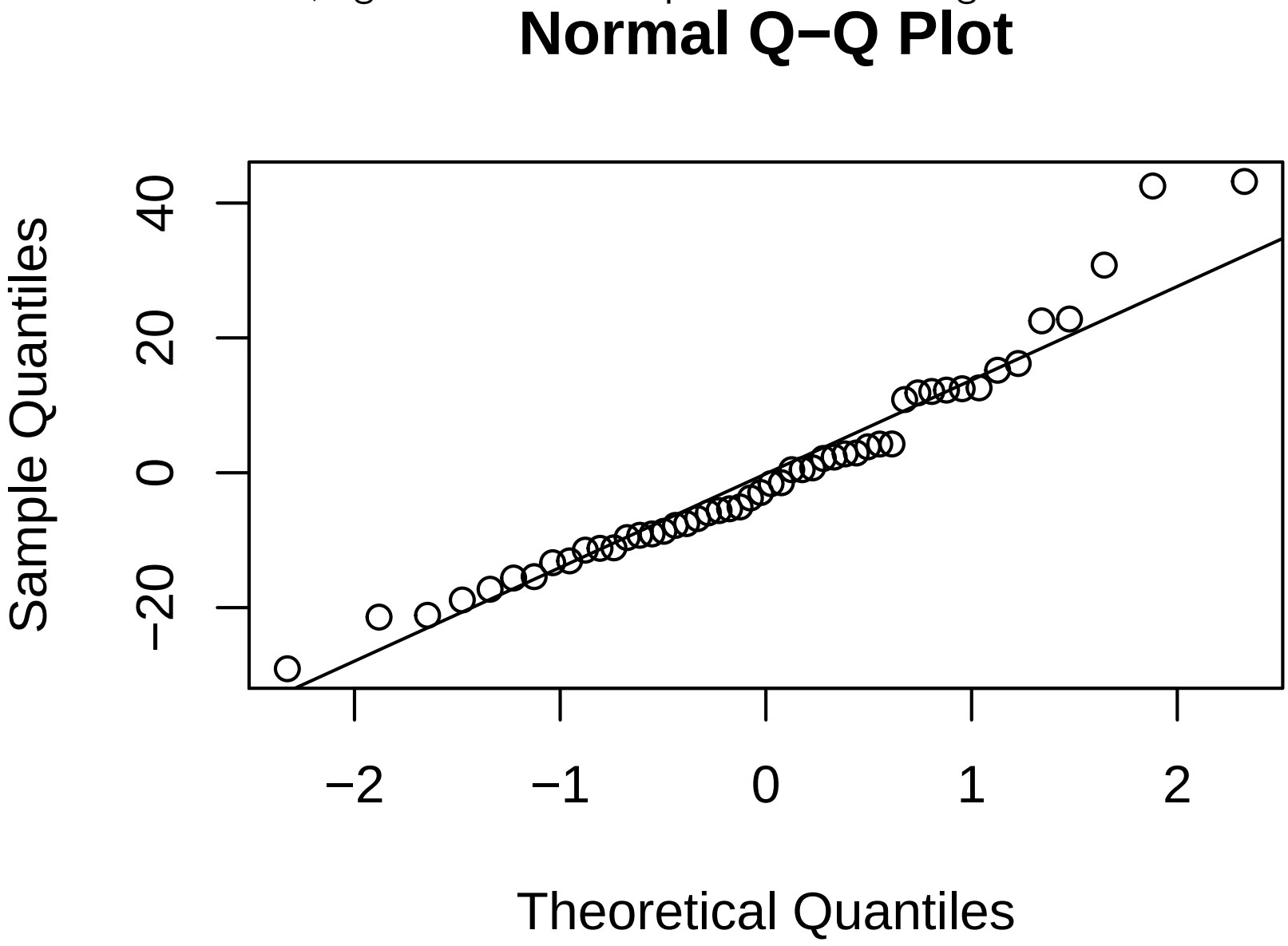
Using Residuals to Check Normality

We use the residuals to check normality by applying histograms and q-q plots to the residuals. The histogram of the residuals from the cars data looks fairly close to normal (with possible exception of a few large outliers).



The Q-Q Plot of the Residuals

The q-q plot of the residuals also suggests a reasonable approximation to a normal distribution, again with the exception of a few large outliers.



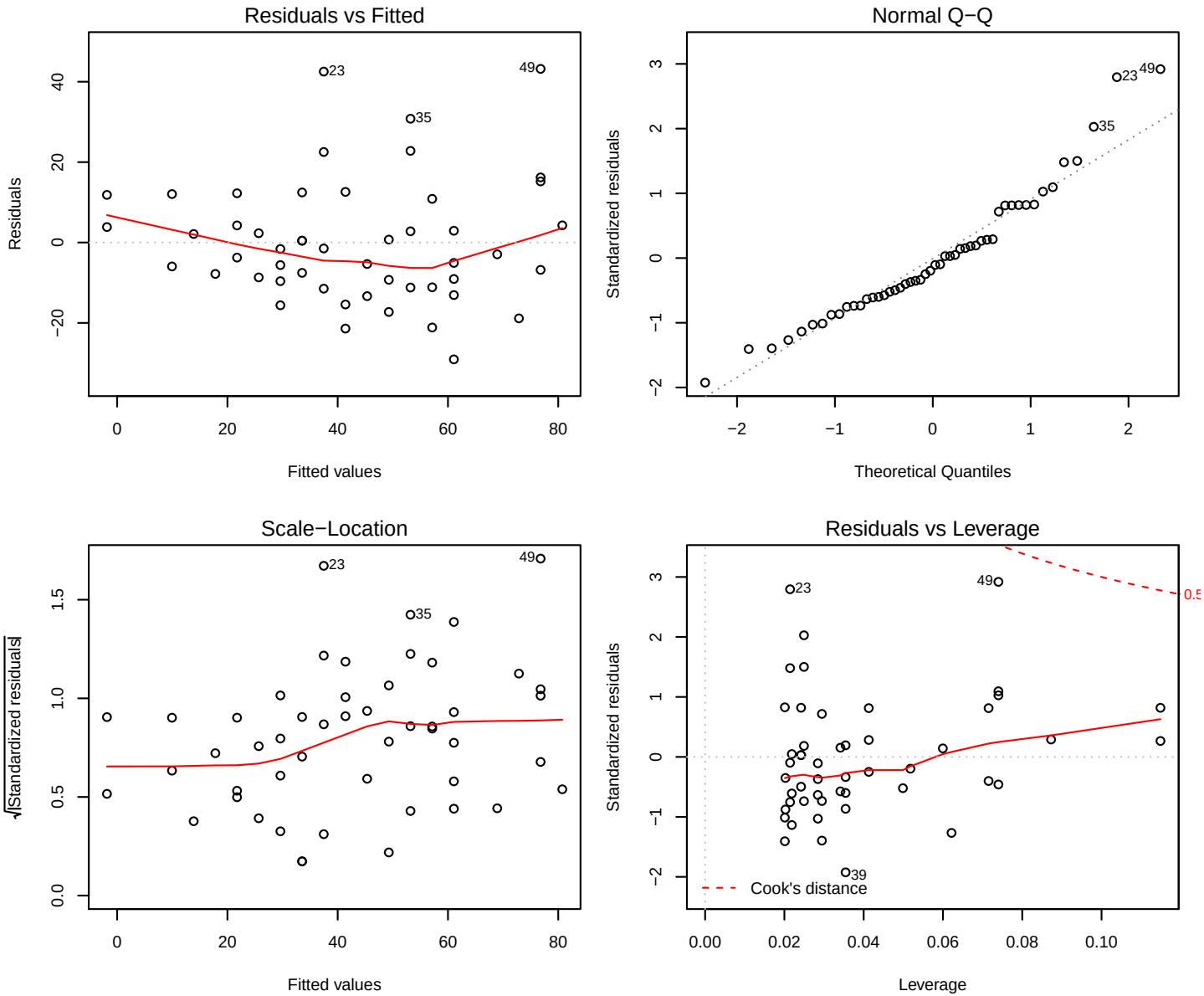
Using the 'plot' function to obtain residual diagnostic plots

The 'plot' function applied to a linear regression model fit gives us four residual plots:

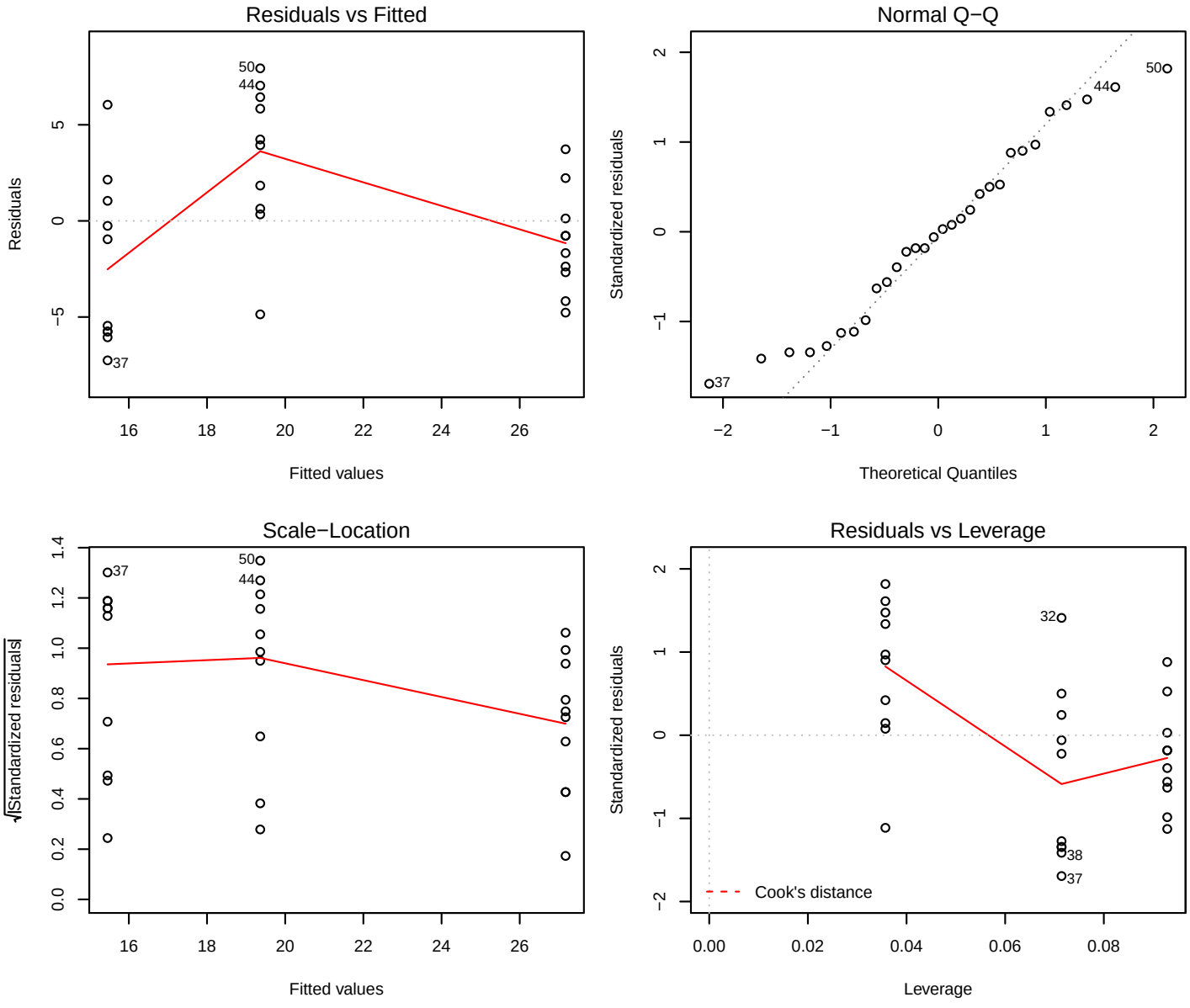
1. residuals vs fitted values plot - use to assess linearity and constant variance
2. q-q plot of residuals - use to assess normality
3. a scale-location plot - use to assess the relationship between mean and variance
4. a "leverage" plot - plots residuals versus leverage (also called influence - we will see this later)

Residual plots for the cars data

```
par(mfrow=c(2,2),mar=c(5,4,2,1))
plot(lm(dist ~ speed,data=cars))
```

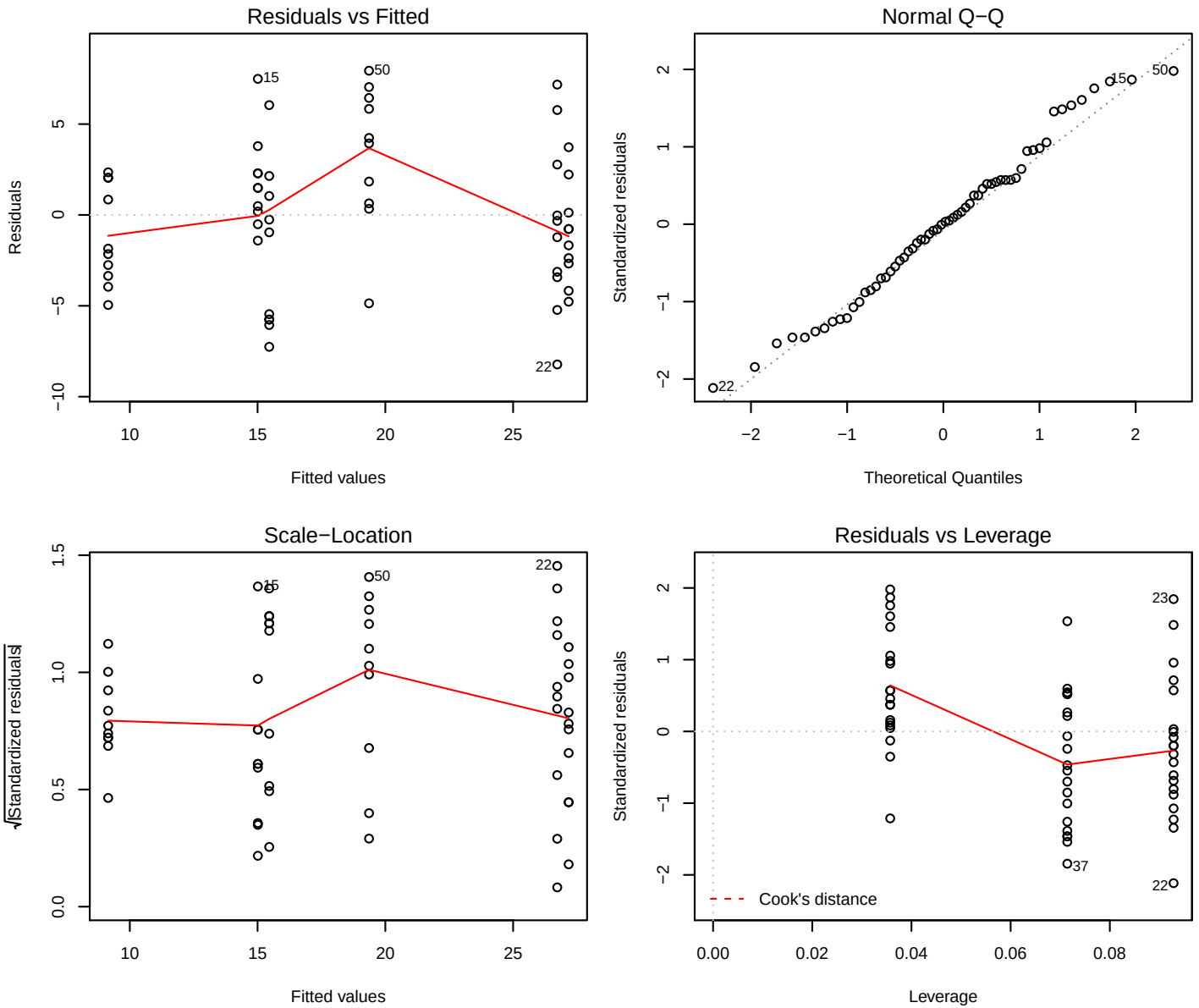


Residual plots for the OJ Data: len ~ dose



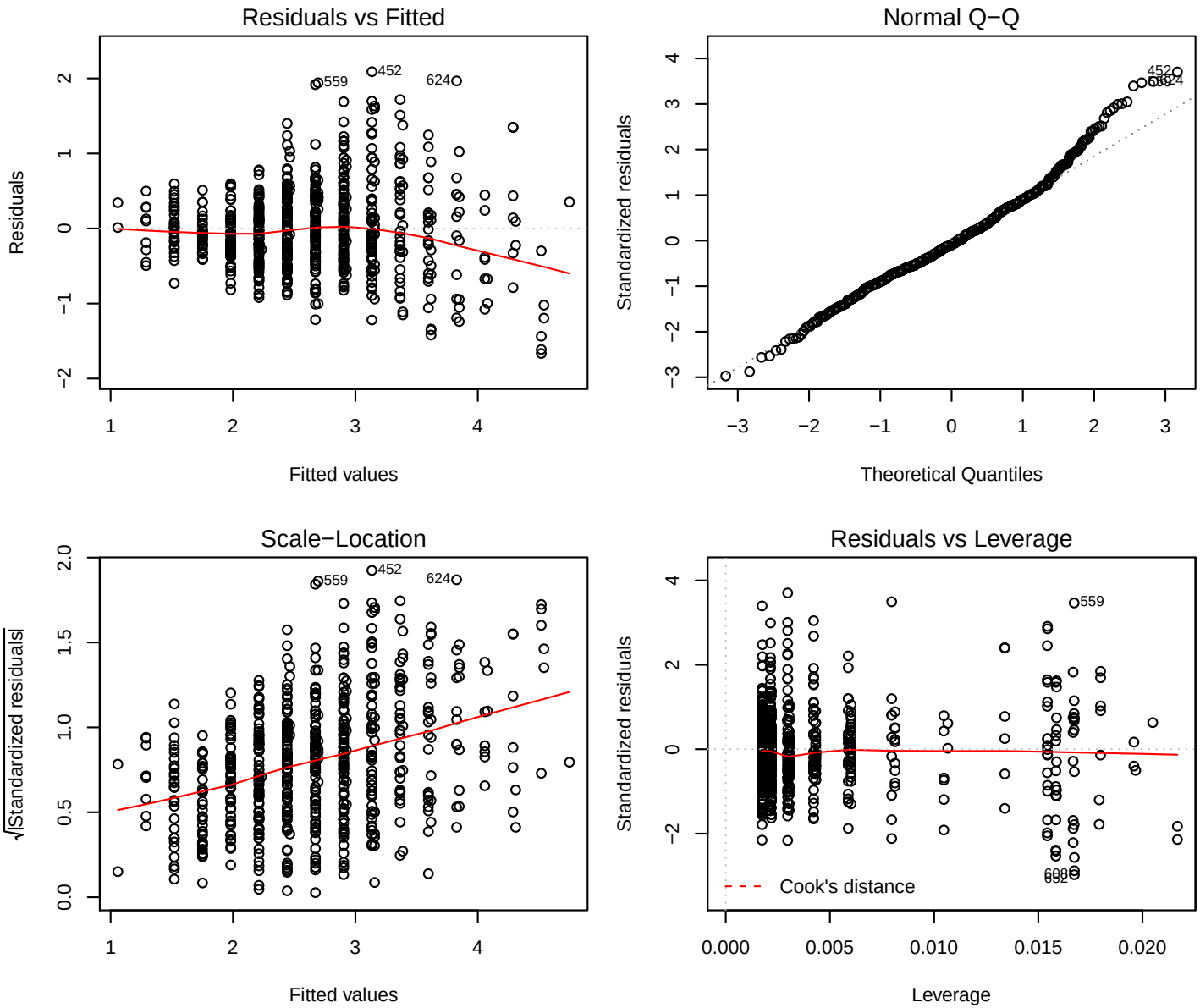
Residual plots for the interaction model: len ~ dose*supp

```
par(mfrow=c(2,2),mar=c(5,4,2,1))
plot(lm(len ~ factor(supp)*dose, data=ToothGrowth))
```



Residual plots for FEV ~ age + smoke

```
par(mfrow=c(2,2),mar=c(5,4,2,1))
plot(lm(fev ~ age+smoke, data=f))
```



What to do when assumptions do not hold?

First. . . .take a deep breath. . . .and remember that. . . .

> *All models are wrong - some are useful. . . . George E.P. Box*

> *Since all models are wrong the scientist cannot obtain a "correct" one by excessive elaboration. . . . . Since all models are wrong the scientist must be alert to what is importantly wrong.* [1]

So we should ask: **is my model wrong enough to matter?**

Let's see what we can do if the answer is yes, that our assumptions are far from met.

[1] Science and statistics, by GEP Box, J. Amer. Stat. Assoc., 76(356):791-799, 1976.

What to do about non-normality

This one is "easy" - just make sure you have a large enough sample so that the CLT takes effect and the non-normality does not matter! (Easier said than done.)

If the sample size is not large enough, some options are:

- use an alternative type of model, e.g., Generalized Linear Model - see next week
- use a transformation (see example below)
- use specialized methods, e.g., permutation test

What to do about non-constant variance?

This one is harder - large sample size does not make up for non-constant variance.

GLMs and tranformations may help here. Typically, we need to choose an appropriate GLM or transformation that *simultaneously* deals with non-normality and non-constant variance.

A more general solution is **robust standard errors**.

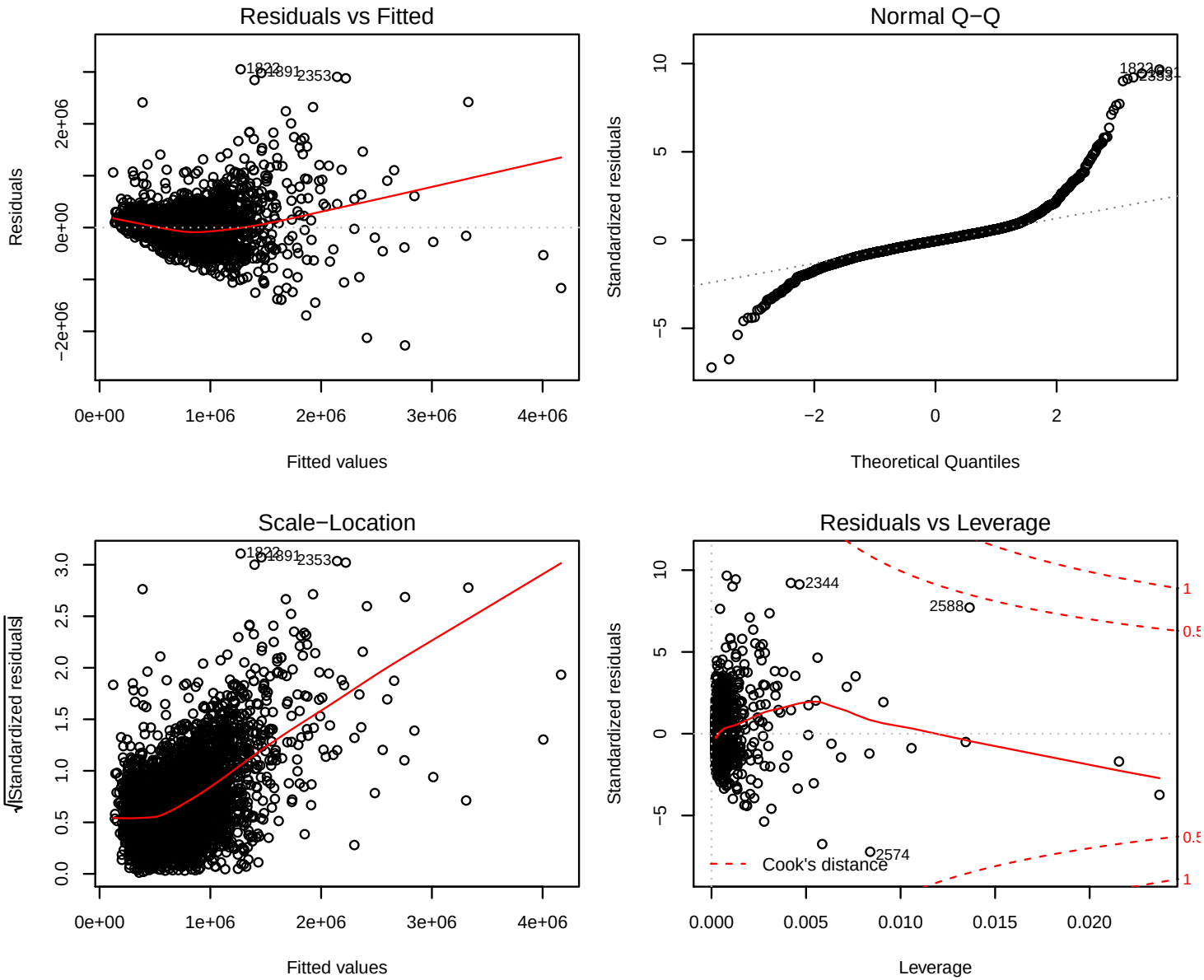
What to do about non-linearity

Similar strategies to the ones used for non-constant variance may be useful, e.g., GLM or transformation.

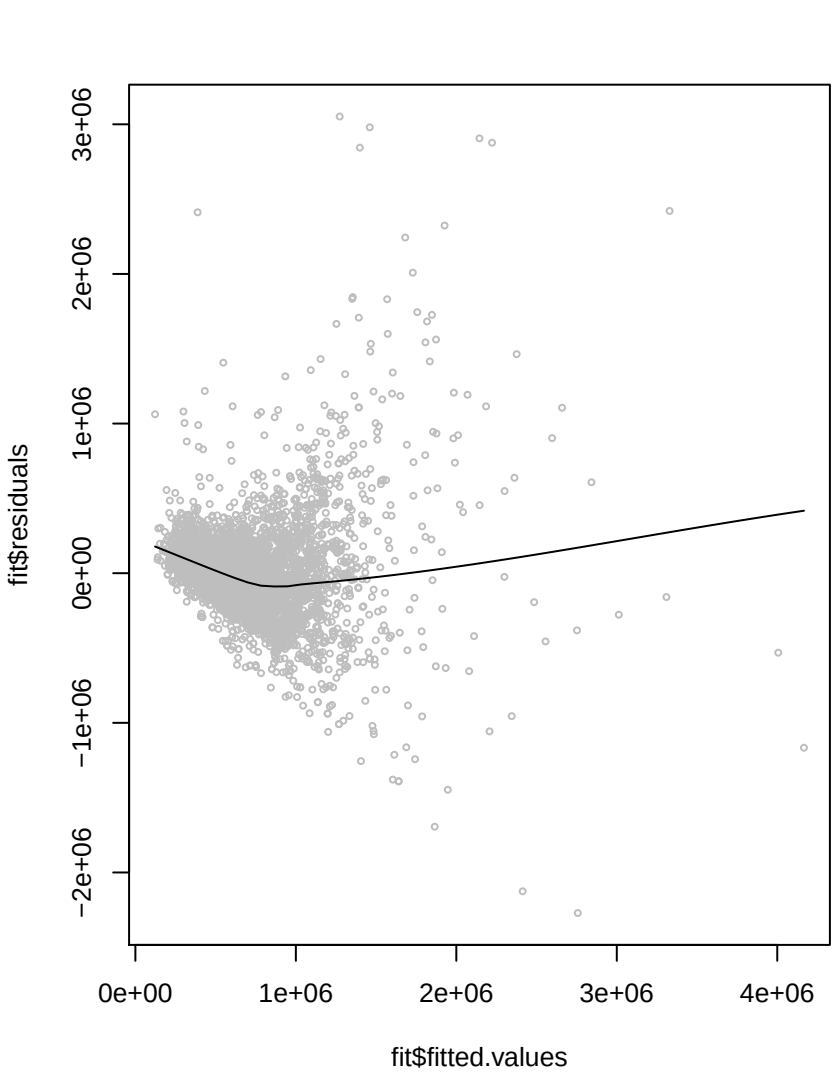
Another approach is to describe the association using a scatterplot smoother. There are many types of smoothers, including splines, which fit piece-wise polynomials to the data.

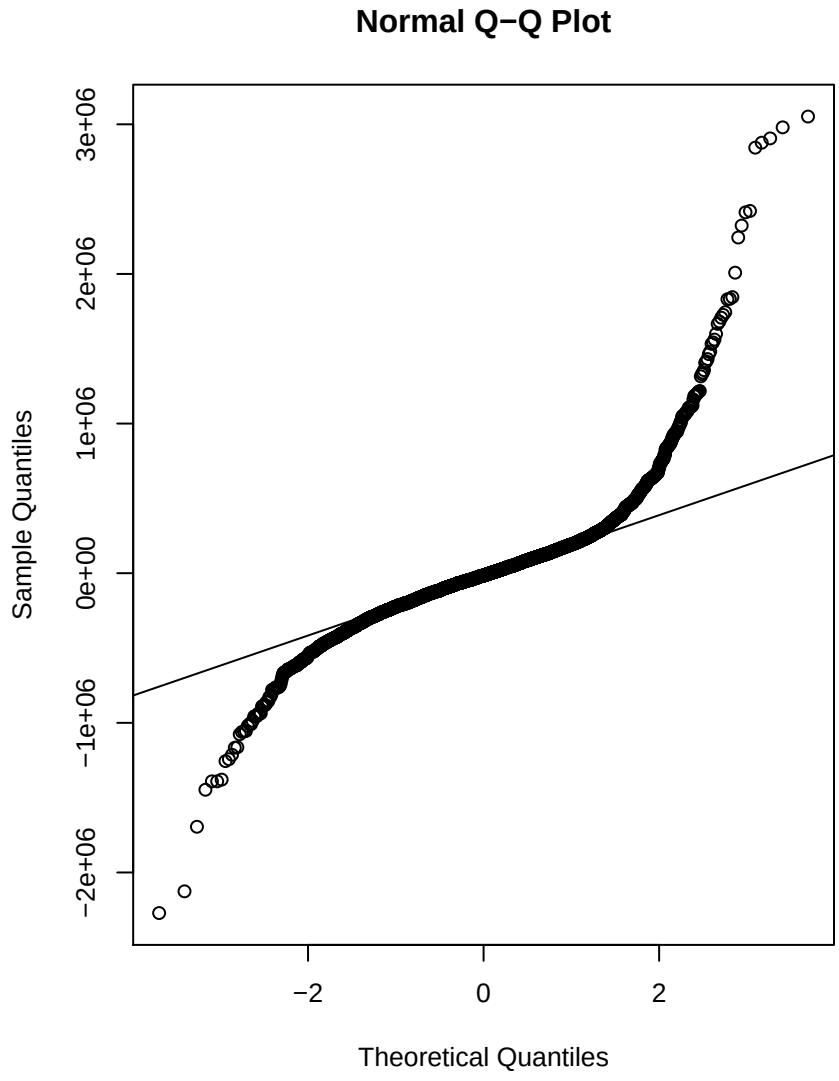
Example - the house price data

The house-price data is a good illustration of the difficulties in dealing with the assumptions of linear regression, because it exhibits non-normality, non-constant variance and non-linearity.



Residuals vs fitted values and Q-Q plot





Non-linearity, non-constant variance and non-normality

The plot of residuals vs fitted values shows clear evidence of non-constant variance, with greater variance for larger fitted values.

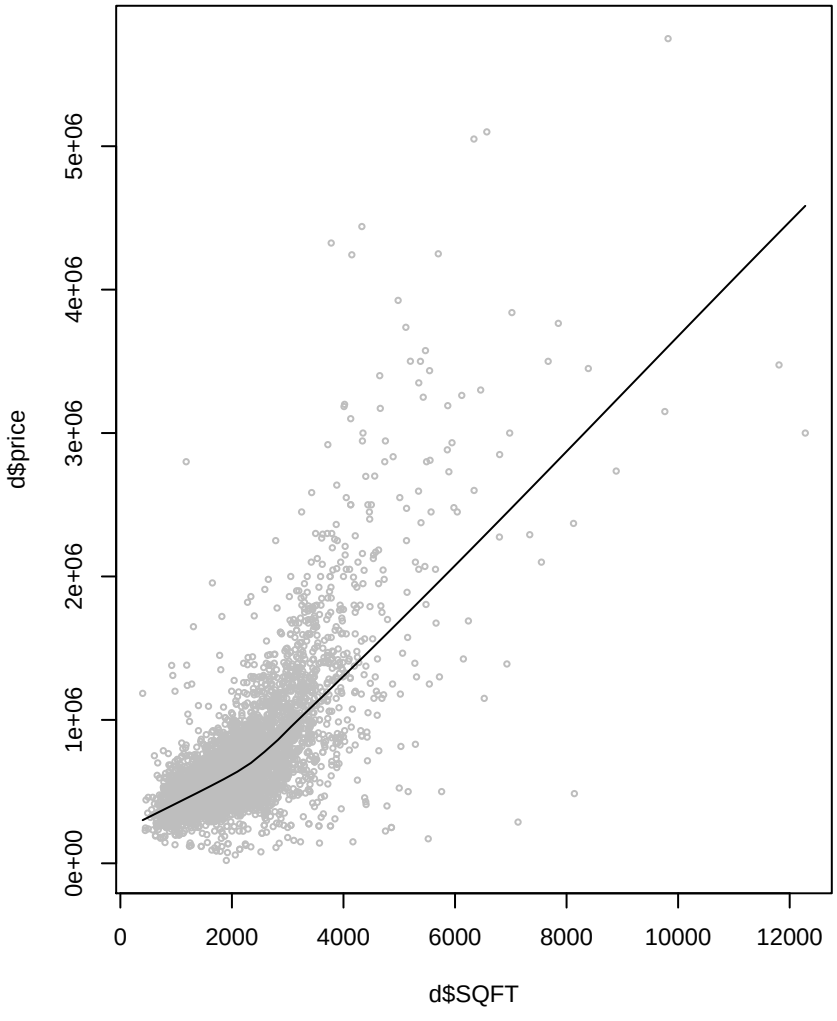
The non-linearity of the smooth curve also indicates a departure from linearity in the association between price and square feet.

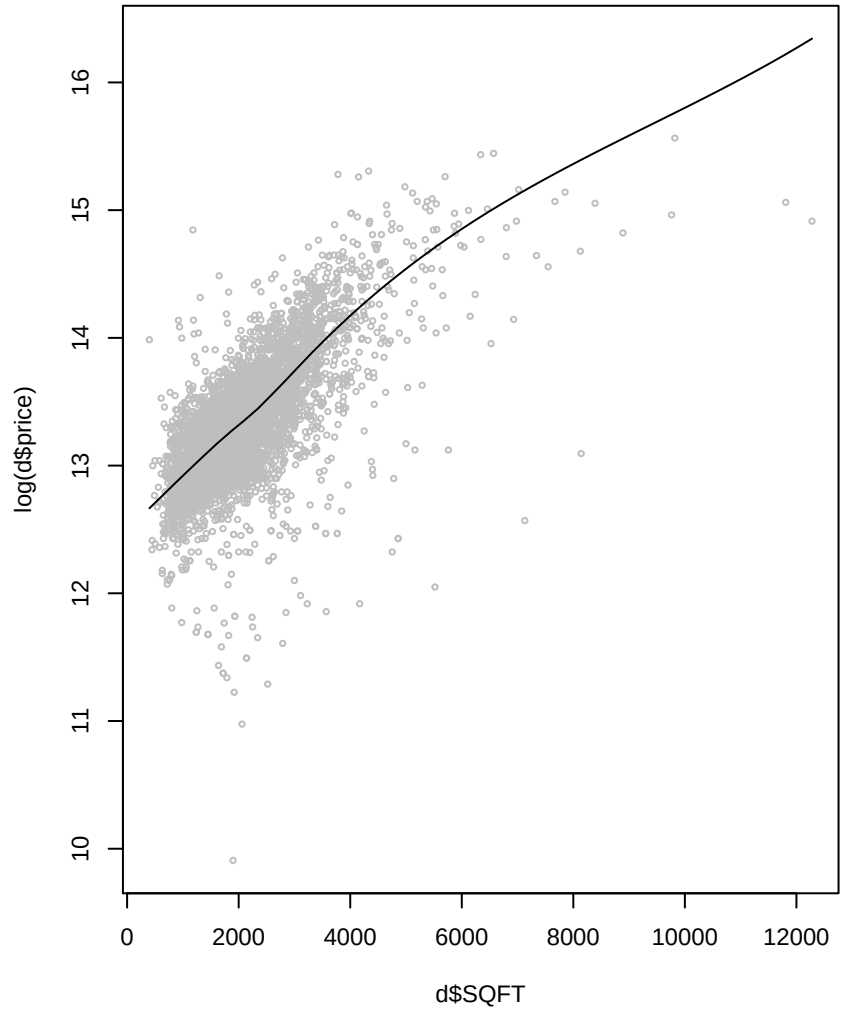
The distribution of the residuals is far from normal.

The extreme departures from assumptions suggest that a different modeling approach might be more appropriate. Two options in this situation are to use a transformation or to use a different type of model, such as a Generalized Linear Model (GLM). A GLM can account for non-linear associations as well as non-constant variance. For now we will consider the transformation approach.

Examining the non-linearity

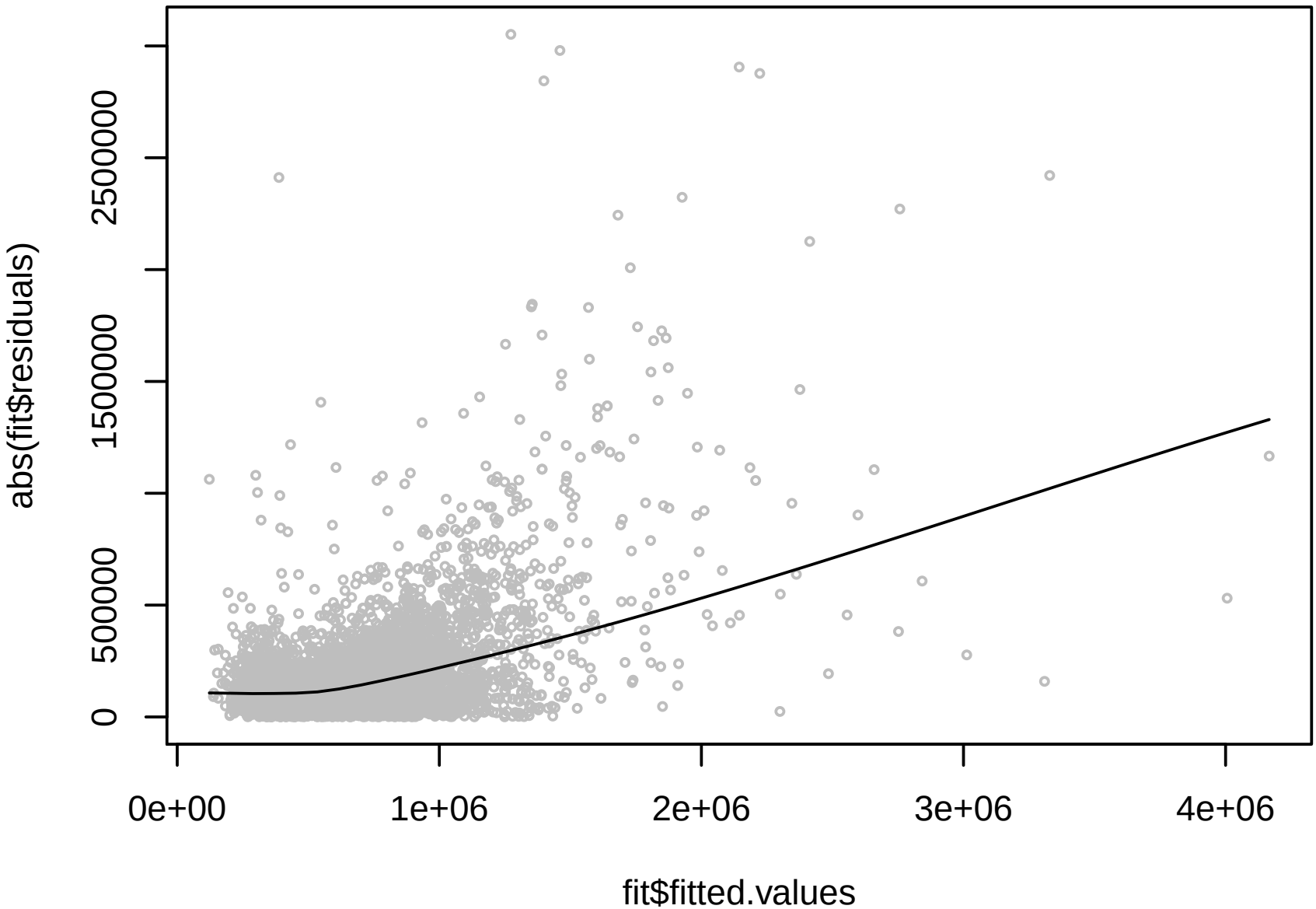
The plot of price vs sqft appears to be exponential so let's try plotting log(price) vs sqft, which is close to linear (except for the extreme values). This is a hint that we might try analyzing price on the logarithmic scale.





Relationship between mean and variance

There is a strong association between the spread of the residuals and the fitted values. This is interpreted as a mean-variance relationship, i.e., there is a positive relationship between the *mean* sales price and the *variance* of sales price.



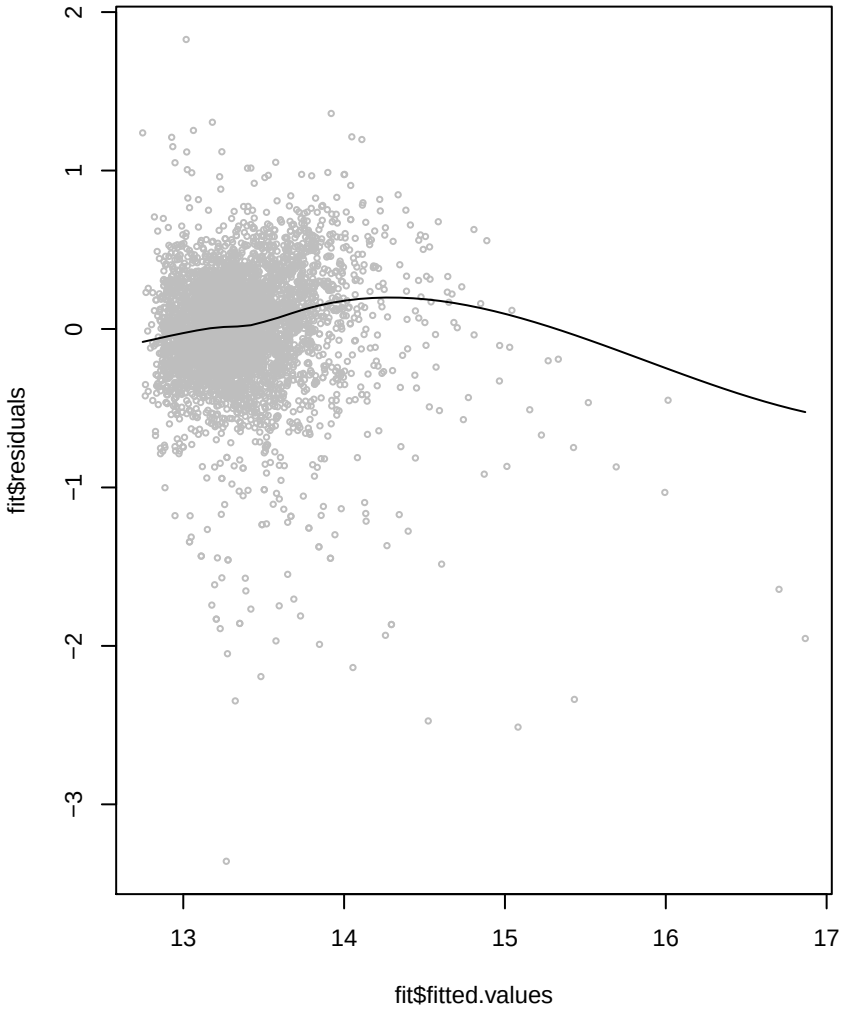
Fitting the model to log-transformed price

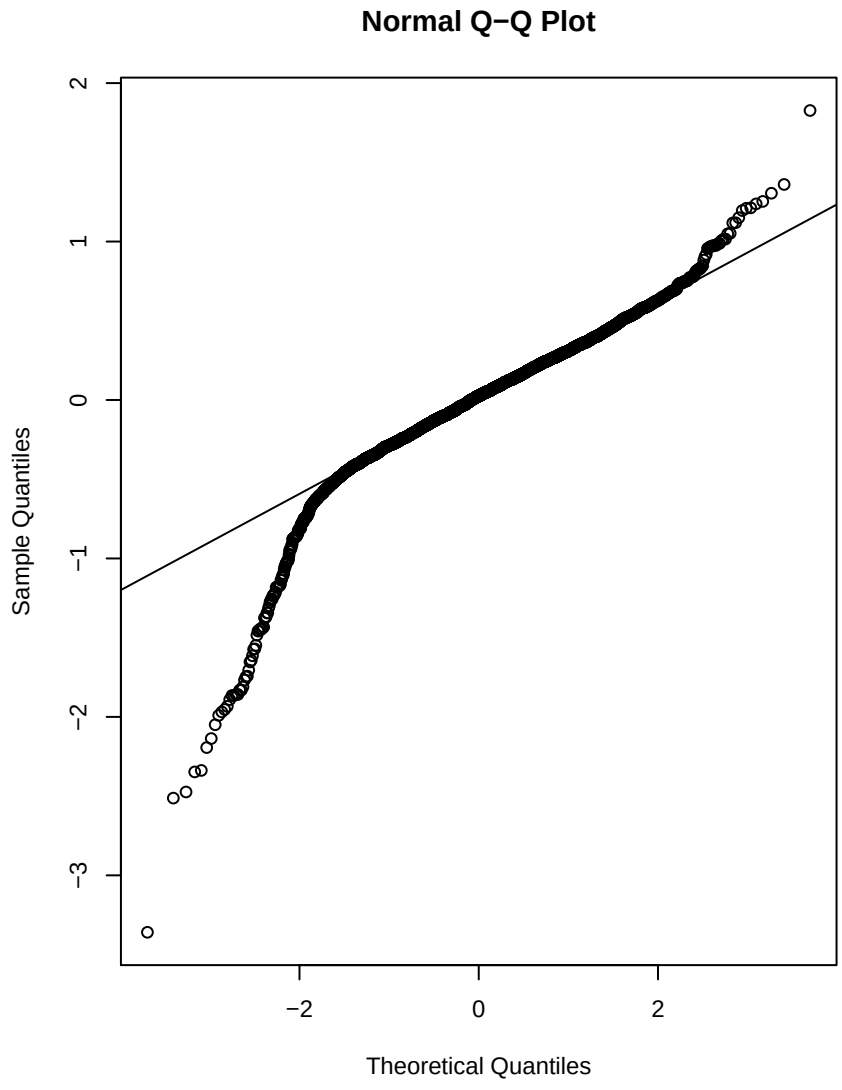
```
d$logprice=log(d$price)
summary(lm(logprice ~ SQFT,data=d))$coef
```

```
                Estimate   Std. Error   t value Pr(>|t|)
(Intercept) 1.260875e+01 1.328558e-02 949.05505        0
SQFT        3.467958e-04 5.560685e-06  62.36567        0
```

One disadvantage of the transformation approach is that the interpretation of the regression coefficients is not clear. We have to remember to interpret in terms of "log-dollars" rather than dollars. This can be a serious disadvantage in some contexts (e.g., when we want to attach a price per additional bedroom) but perhaps not in other (e.g., when we are just interested in prediction).

Checking assumptions for the log-transformed response





Assumptions still not met!