6 Dealing with Model Assumption Violations

If the regression diagnostics have resulted in the removal of outliers and influential observations, but the residual and partial residual plots still show that model assumptions are violated, it is necessary to make further adjustments either to the model (including or excluding predictors), or transforming the response and/or predictors, and/or weighting the measurements, and if this all does not help switching to a different model or estimation method entirely.
If the inclusion or exclusion of predictors do not resolve the concerns about the violation of the model assumptions further approaches can be used.
Depending on the type of violation different remedies can help.

6.1 Transformations

Transformations can help when

1. the homoscedasticity assumption, or
2. the linearity assumption, or
3. normality

is violated.

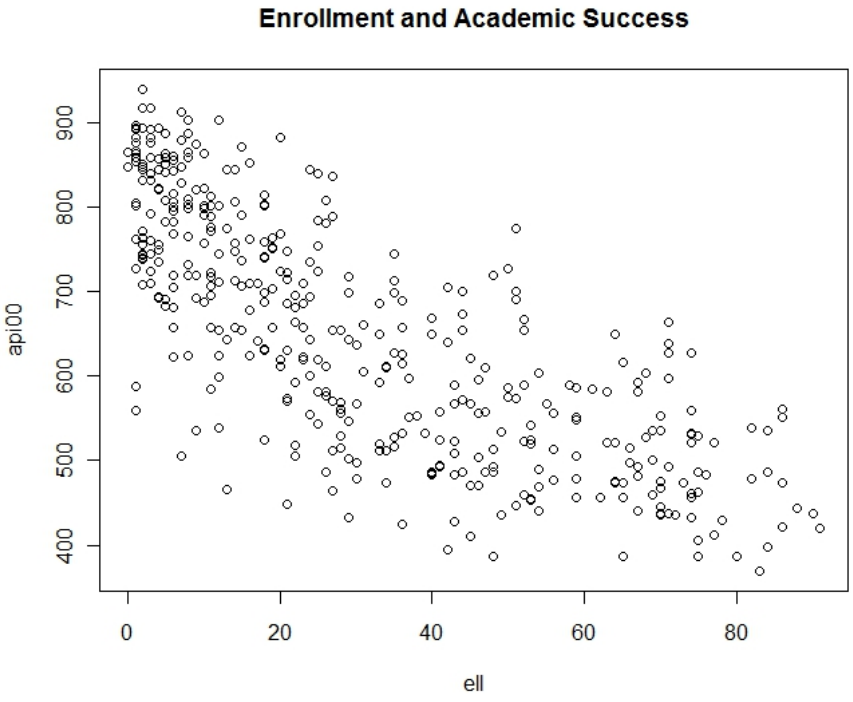
6.1.1 Heteroscedasticity

If the assumption of constant variance is violated, the least squares estimators are still unbiased, but the Gauss-Markov theorem does not hold anymore, and standardized scores do not have the assumed distribution, and therefore test results and confidence intervals are unreliable. Usually the standard errors of the regression coefficients are too large.
A transformation of the response variable can help to resolve such a problem. Depending on the type of violation, different transformations are helpful:
Useful transformations:

Relationship between the error variance and the mean response	Transformation of Y
$\sigma^2 \propto E(Y)$	square root
$\sigma^2 \propto E(Y)^2$	log
$\sigma^2 \propto E(Y)^3$	reciprocal square root $(1/\sqrt(y))$
$\sigma^2 \propto E(Y)^4$	reciprocal
$\sigma^2 \propto E(Y)(1-E(Y))$	if $0 \leq Y \leq 1$, arcsin, $(sin^{-1}(\sqrt(y))$

However, after applying the transformation the interpretation of the regression coefficients is not straight forward anymore, and the inverse transformation is not necessarily resulting in unbiased estimates on the original scale.
Confidence intervals and prediction intervals though can be transformed back to the original scale.

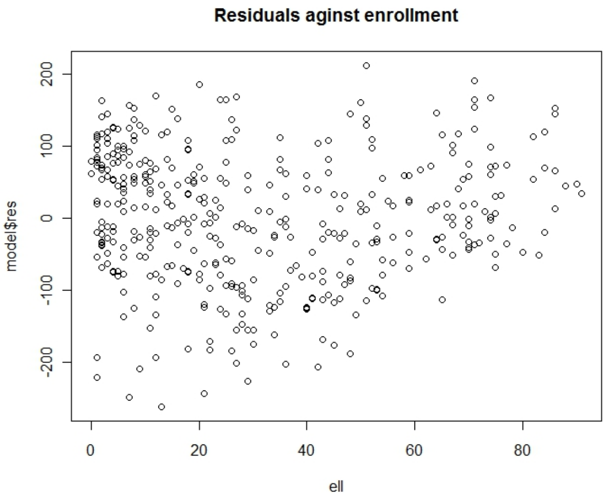
Example 6.1.
The data set consists on 400 randomly sampled elementary schools from the California Department of Education's API in the 2000. This data file contains a measure of school academic performance (api00) as well as other attributes of the elementary schools, such as, class size, enrollment, poverty, etc.
Does enrollment have an effect on academic performance?

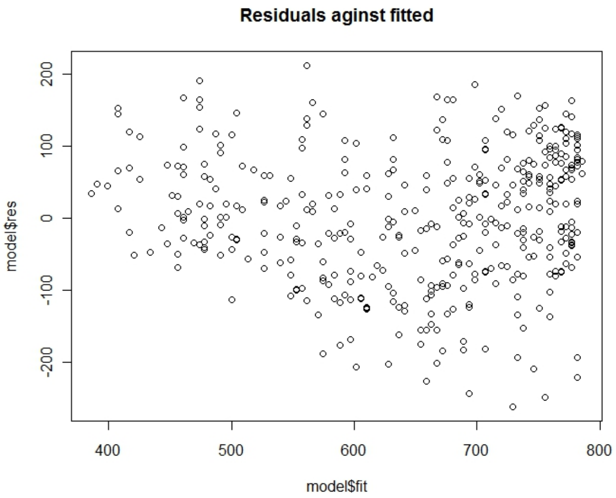


The scatterplot indicates the higher the enrollment the lower the academic success.
Fitting the model

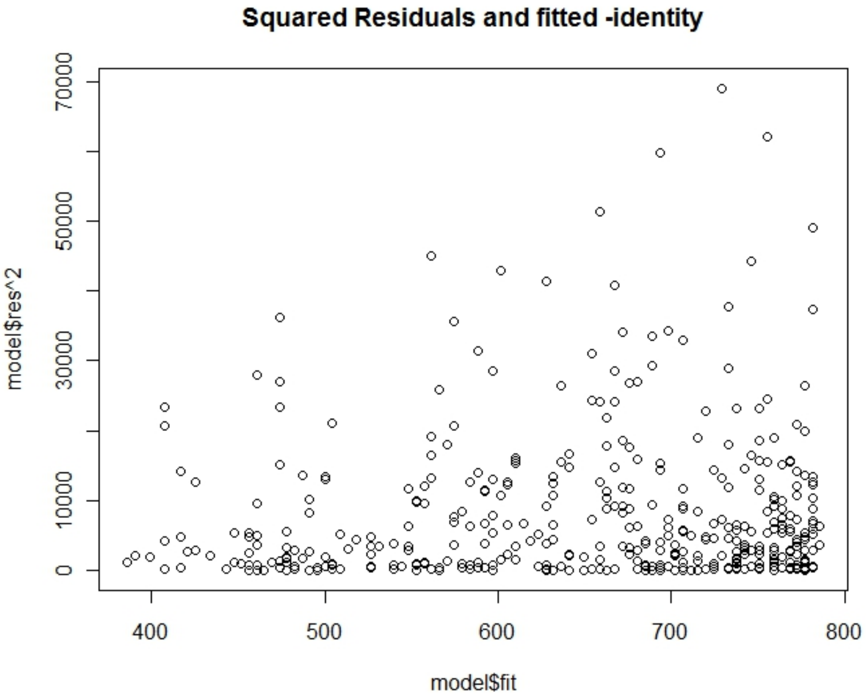
$$api00 = \beta_0 + \beta_1 enrollment + \varepsilon$$

results in the following residual plots





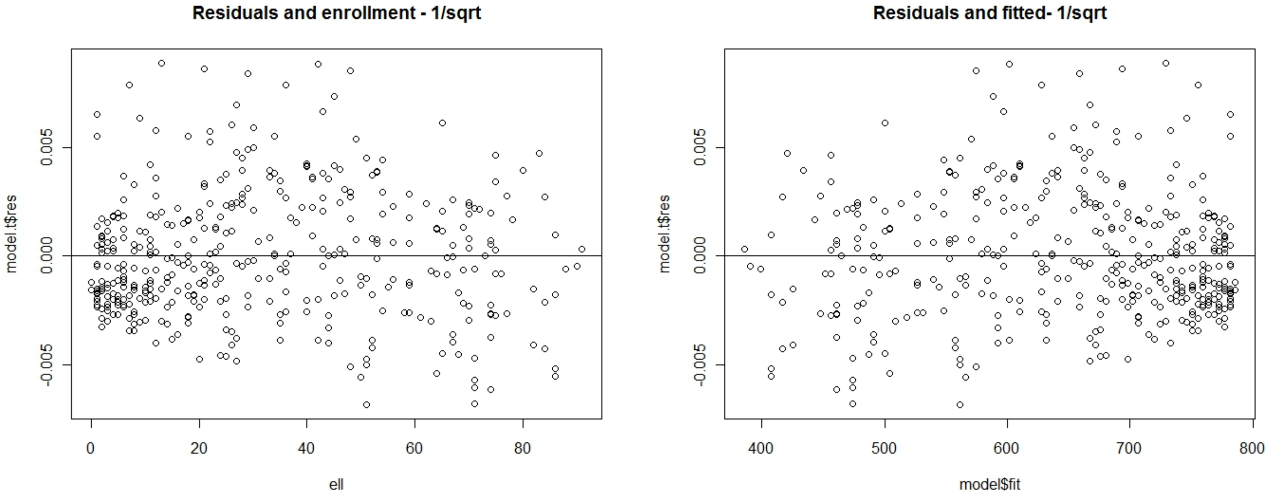
Both indicate a violation of the assumption of homoscedasticity. To investigate the nature of the relationship of the violation plot the squared residuals against the fitted values.



Trying the different transformations suggested in the table above

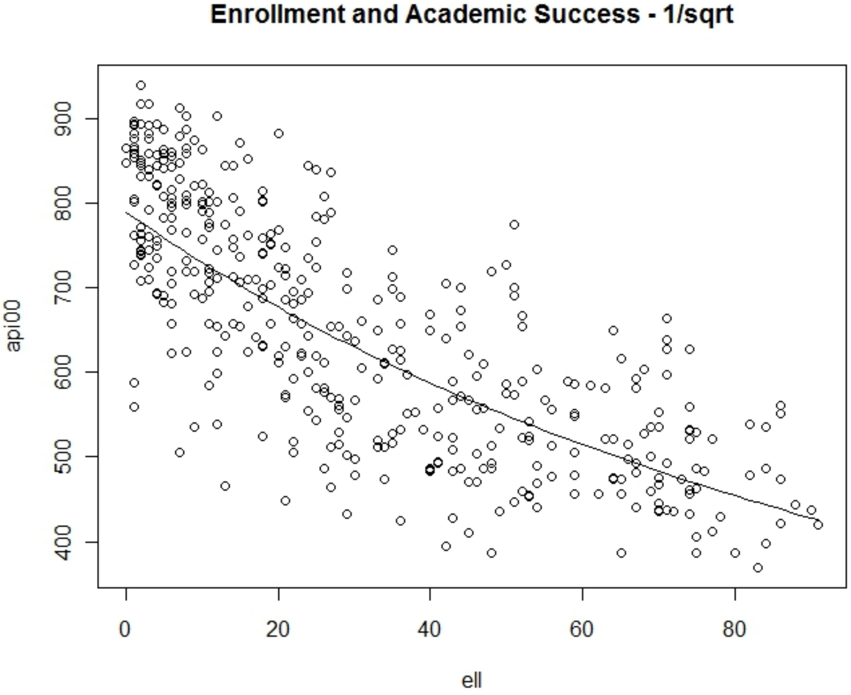
$$1/\sqrt{api00} = \beta_0 + \beta_1 enrollment + \varepsilon$$

results in the following residual plots



the best of the four.

The fitted curve based on the transformed model is



The transformation of the model ensures that the model assumptions are met, and improves the fit of the model to data and hopefully therefore to the population. Tests and confidence intervals become reliable statements which hold the associated error probabilities.
But by transforming the response variable we loose the straight forward interpretation of the estimate of $\hat{\beta}$.

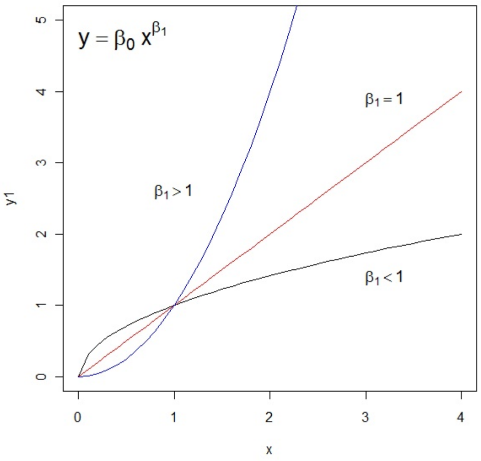
6.1.2 Non-linear Relationships

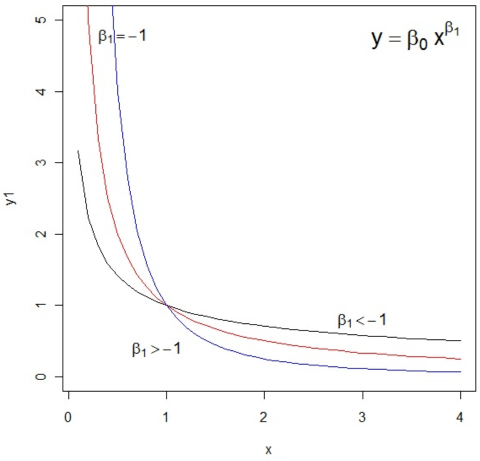
If the scatterplot for response and predictor does indicate a non linear relationship transformations of the response and/or predictor can result in a linear model for fitting the non-linear relationship.
The easiest example is to permit for a polynomial relationship, Where the model becomes:

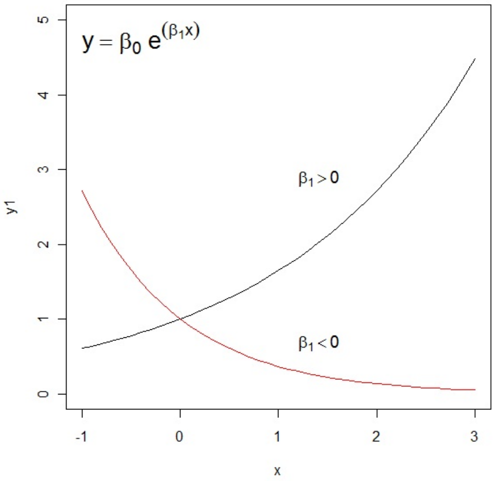
$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

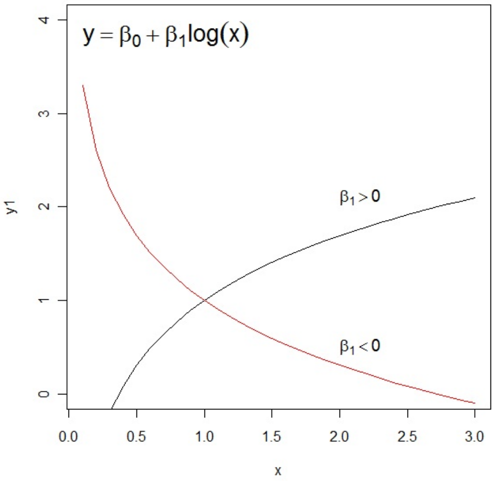
Some other non-linear relationships we can deal with by transforming the response and/or predictor variable(s) are given in the graphs and table below.
In general a non-linear regression model should be considered.
Some non-linear relationships which can be dealt with by transformations are:

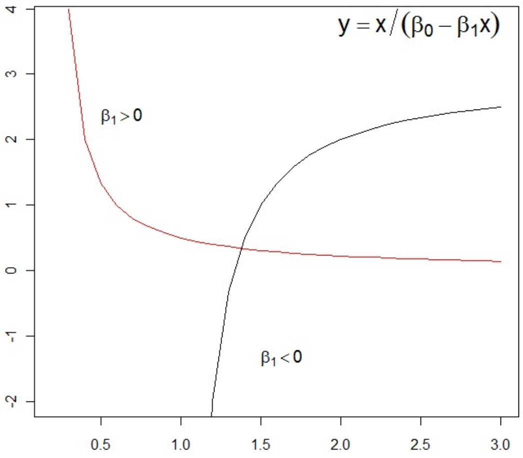
Function	Transformations of x and/or y	Resulting model
$y = \beta_0 x^{\beta_1}$	$y' = log(y), x' = log(x)$	$y' = log(\beta_0) + \beta_1 x'$
$y = \beta_0 e^{\beta_1 x}$	$y' = ln(y)$	$y' = ln(\beta_0) + \beta_1 x$
$y = \beta_0 + \beta_1 log(x)$	$x' = log(x)$	$y = \beta_0 + \beta_1 x'$
$y = \dfrac{x}{\beta_0 x - \beta_1}$	$y' = \dfrac{1}{y}, x' = \dfrac{1}{x}$	$y' = \beta_0 - \beta_1 x'$











The transformation have a large influence on the error structure. The assumption is that the error

in the transformed model is normal with mean 0 and constant variance σ^2.
E.g. looking closer at the second transformation:
The MLRM is

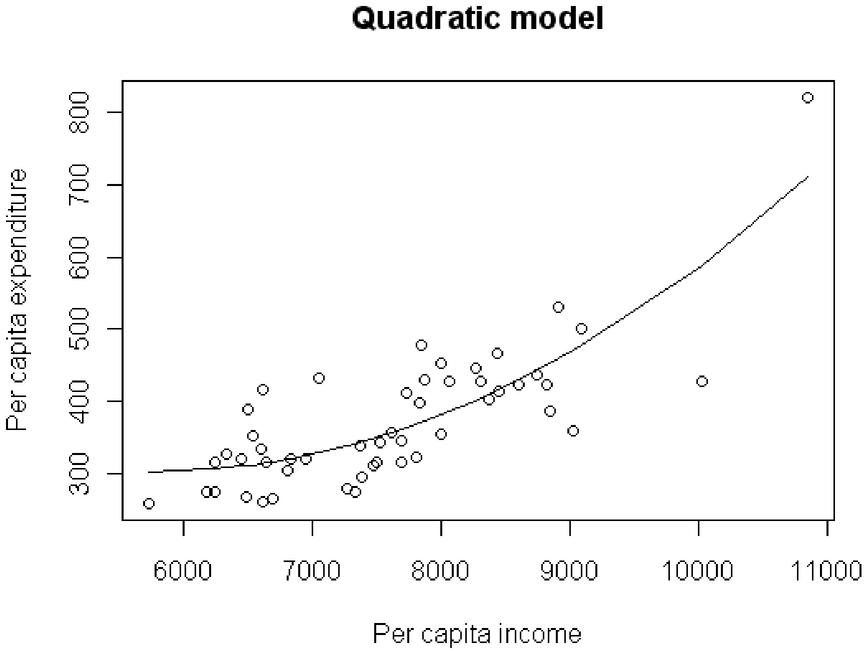
$$log(Y) = log(\beta_0) + \beta_1 x + \varepsilon$$

which is equivalent to (take e to the power of sides):

$$Y = \beta_0 e^{\beta_1 x} e^{\varepsilon}$$

Since we assume that ε is normal, this implies that the multiplicative error in the model for Y is log normal (it's logarithm is normal).
To check model assumption the transformed model has to be checked using residuals and influence measures.

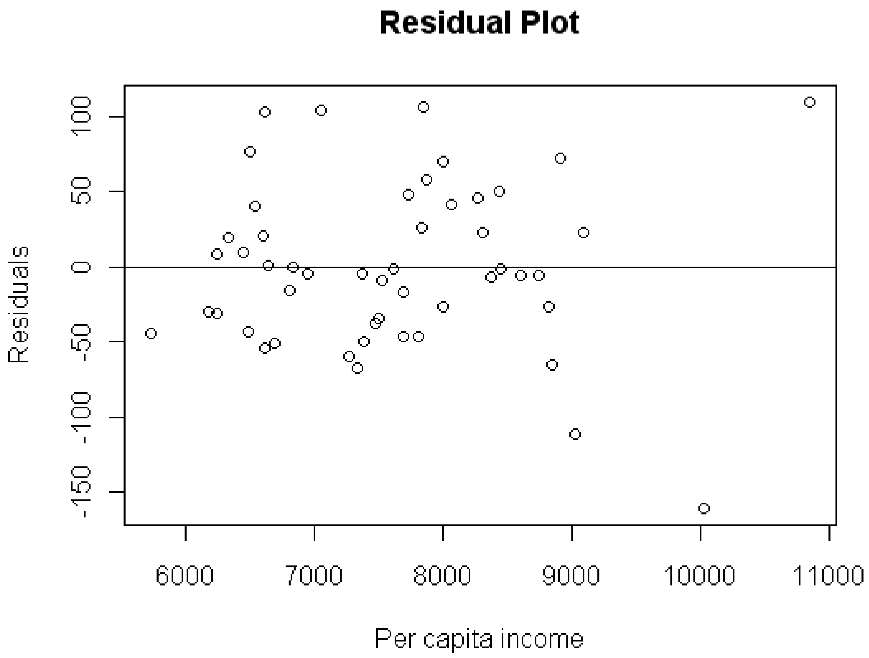
Example 6.2.
The sample analyzed consists of 50 observations of per capita expenditure on public schools and per capita income for each state and the District of Columbia in 1979. Wisconsin was removed because no information was available on the per capita expenditure for public schools.



The scatterplot suggest a quadratic relationship. Use the data to fit the model

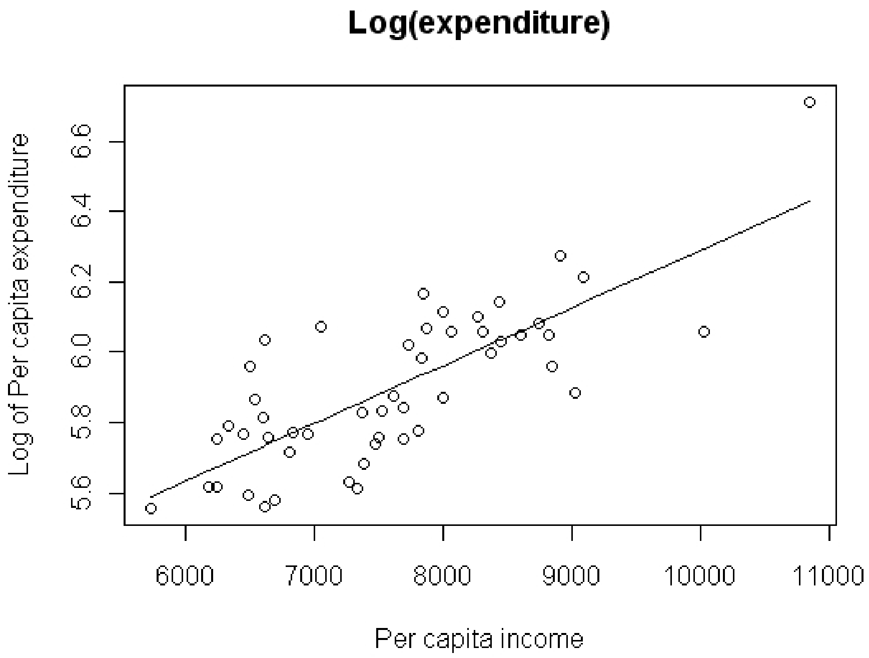
$$expend = \beta_0 + \beta_1(income) + \beta_2(income)^2 + \varepsilon$$

Resulting in:



The residual plot shows that for states with larger per capita income the variation in the per capita expenditure is larger than for states with lower per capita income. This is in violation of the MLRM assumption.

Using the log transformation for the per capita expenditure results in the following scatterplot:

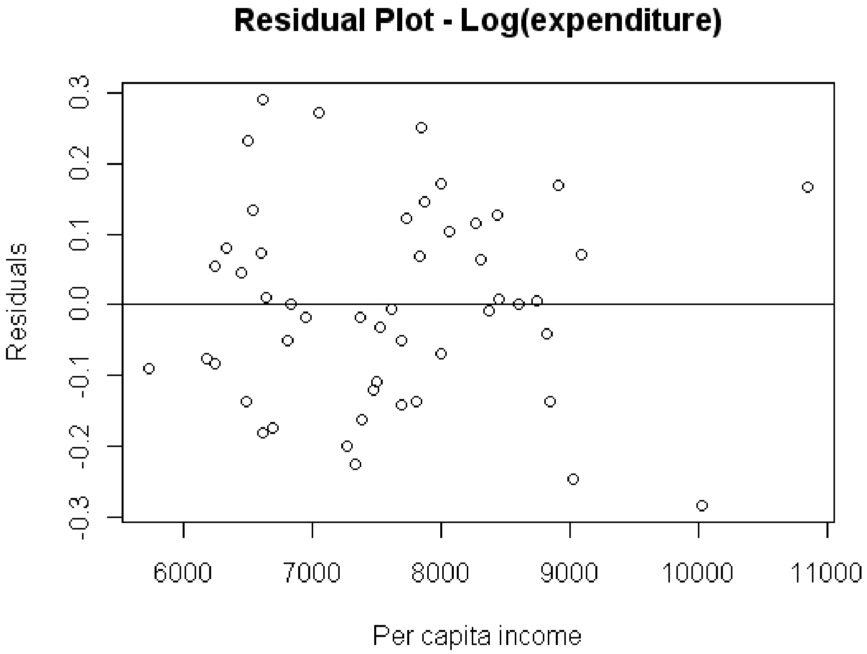


Showing a linear relationship.

Fitting

$$log(expend) = \beta_0 + \beta_1(income) + \varepsilon$$

results in the following residual plot:



This residual plot for the transformed model shows a big improvement over the residual plot for the initial model. Tests and confidence intervals based on this model are more reliable than for the first model.

6.1.3 Choosing a Transformation

6.1.4 Box-Cox: transforming the response

Another violation which can be coped with by transforming the data is non-normality of the error. Box and Cox(1964) developed a method for choosing the "best" transformation from the set of power transformations to correct for this violation.
The set of power transformations can be parameterized with the following definition

Definition 6.1.
Let $\lambda \in \mathbb{R}$, then

$$y^{(\lambda)} = \begin{cases} \dfrac{y^\lambda - 1}{\lambda\, \tilde{y}^{\lambda-1}}, & \lambda \neq 0 \\ \tilde{y}\, ln(y), & \lambda = 0 \end{cases}$$

where

$$\tilde{y} = e^{\dfrac{1}{n}\sum_{i=1}^{n} ln(y_i)}$$

is the geometric mean of the observations.

The geometric mean is a scaling factor, which enters the equation when finding the *maximum likelihood estimators* for λ and $\vec{\beta}$ simultaneously. It makes the residual sum of squares for different λ comparable.

Choosing λ
No formula exists which will produce the value of λ, which will produce the smallest residual sum of squares for a transformation $y^{(\lambda)}$, $SS_{Res}(\lambda)$. Therefore usually $SS_{Res}(\lambda)$ is found for a number of different values for λ, which are then plotted against λ. From the graph then the λ with the smallest $SS_{Res}(\lambda)$ can be read. This procedure can be done iteratively, by first determining a rough estimate and then investigating the neighbourhood of the rough estimate using a finer grid.
Once λ has been decided on, sometimes after rounding so that the result is an easier to interpret model (use $\lambda = 0.5$ instead of $\lambda = 0.473$). The transformation y^{λ} is used in the analysis, omitting the scaling factor, which was only necessary to make the sum of squares comparable.

A confidence interval for λ
Based on the theory about maximum likelihood a $(1-\alpha)\times 100\%$ confidence interval for λ is given by all λ with

$$L(\hat{\lambda}) - L(\lambda) \leq \frac{1}{2}\chi^2_{\alpha,1}$$

where

$$L(\lambda) = -\frac{1}{2}nln[SS_{Res}(\lambda)]$$

is the log likelihood function.
This is equivalent to all λ with

$$L(\lambda) \geq L(\hat{\lambda}) - \frac{1}{2}\chi^2_{\alpha,1}$$

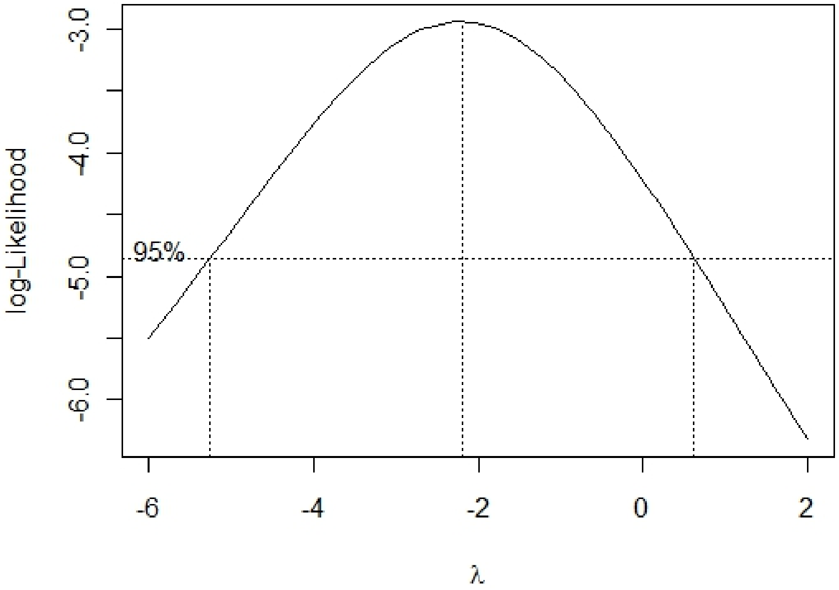
or all λ with

$$SS_{Res}(\lambda) \leq SS_{Res}(\hat{\lambda})\, e^{\chi^2_{\alpha,1}/n}$$

The bounds can be found graphically by including a horizontal line with the plot of λ against $SS_{Res}(\lambda)$.

Continue Example. Go back to the blood pressure example and find the confidence interval for λ.

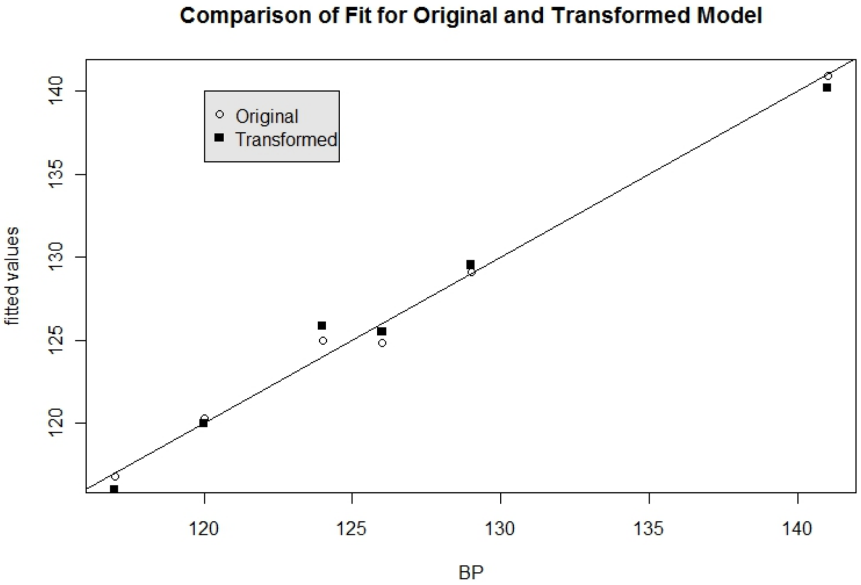
λ	$L(\lambda)$
-6.07	-5.49
-5.5	-5.07
-5.0	-4.62
-4.5	-4.18
-4.0	-3.76
-3.5	-3.40
-3.0	-3.12
-2.5	-2.97
-2.0	-2.96
-1.5	-3.10
-1.0	-3.38
-0.5	-3.76
0.0	-4.28
0.5	-4.74
1.0	-5.26
1.5	-5.80
2.0	-6.32



According to the graph the 95% confidence interval for λ falls between -5.3 and 0.5, with $\hat{\lambda} \approx -2.1$. In particular 1 is not the confidence interval indicating that some transformation should be done to make the residuals more normal, and the residual sum of squares smaller, and therefore result in a better fit of the model. A good choice for λ seems to be -2.
The model would be

$$\frac{1}{Y^2} = \beta_0 + \beta_1(weight) + \beta_2(age) + \varepsilon$$

Comparing the fit of the two models give an adjusted R^2 for the original model of 0.9729, and for the transformed model of 0.9858. The already well fitting model could be improved by the transformation.
Plotting the observed versus the fitted blood pressure for the two models illustrates the same fact, the model based on the transformation fits better than the original model.



6.1.5 Transforming a predictor

In some instances the residual plots or the initial scatter plots indicate a non linear relationship between a regressor and the response, which can be linearized by transforming the regressor.
Consider the following model (one regressor to illustrate, but also applies to MLRMs)

$$Y = \beta_0 + \beta_1 t(x) + \varepsilon,\ \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

Where t is a function $t : \mathbb{R} \mapsto \mathbb{R}$.
Box and Tidwell (1962) proposed an analytic approach to choosing the transformation t.
Let

$$t_\alpha(x) = \begin{cases} x^\alpha, & \alpha \neq 0 \\ ln(x), & \alpha = 0 \end{cases}$$

then for the model given above with $t = t_\alpha$

$$E(Y) = \beta_0 + \beta_1 t_\alpha(x) = f(\beta_0, \beta_1, t_\alpha)$$

Their iterative approach is based on the Taylor series for f about an initial guess $\alpha_0 = 1$, see MPV(pg.186).
The process:

1. Start with an initial guess $\alpha_0 = 1$, and let $i = 0$.

2. Until you are convinced to have found a proper value for α, repeat

   Find $\hat{\beta}_{1,i}$ from fitting the model $Y = \beta_{0,i} + \beta_{1,i} x^{\alpha_i} + \varepsilon$

   Let $w = x^{\alpha_i} ln(x^{\alpha_i})$

   Find $\hat{\gamma}$ from fitting the model $Y = \beta^*_{0,i} + \beta^*_{1,i} x^{\alpha_i} + \gamma w + \varepsilon^*$

The new guess for α is

$$\alpha_{i+1} = \frac{\hat{\gamma}}{\hat{\beta}_{1,i}} + 1$$

Let $i <- i + 1$.

Continue Example. Use R to find the Box Tidwell transformation for WEIGHT in the blood pressure model.

```
library(car)
attach(BP.data)
boxTidwell(BP~WEIGHT,~AGE,verbose=TRUE)
```

Generating the following output:

```
 iter = 1     powers = 2.782409
 iter = 2     powers = 2.789993
 iter = 3     powers = 2.790411
 Score Statistic   p-value MLE of lambda
       0.8651332 0.3869657      2.790411

iterations =  3
```

The output give $\alpha \approx 2.79$, so we could use transformations WEIGHT^2 or WEIGHT^3. Both do not improve the model fit, which was expected since the Score Statistic does not indicate a significant improvement.

Continue Example. Analyze again the per capita expenses on education and the mean per capita income in the different American states.

```
> boxTidwell(expend~income, verbose=TRUE)
 iter = 1     powers = 5.053827
 iter = 2     powers = 5.143801
 iter = 3     powers = 5.162594
 iter = 4     powers = 5.166507
 Score Statistic   p-value MLE of lambda
        2.939926 0.0032829      5.166507

iterations =  4
```

The Box Tidwell procedure suggest a polynomial of degree 5.

```
summary(lm(expend~poly(income,5)))
```

Shows improved model fit for the model of expenditure being a polynomial of degree 5 in income.

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)       373.260      6.819  54.739  < 2e-16 ***
poly(income, 5)1  507.011     48.217  10.515 1.39e-13 ***
poly(income, 5)2  173.291     48.217   3.594 0.000818 ***
poly(income, 5)3  100.842     48.217   2.091 0.042294 *  
poly(income, 5)4  108.461     48.217   2.249 0.029537 *  
poly(income, 5)5  163.584     48.217   3.393 0.001474 ** 
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 48.22 on 44 degrees of freedom
Multiple R-squared: 0.7665,Adjusted R-squared:  0.74 
F-statistic: 28.89 on 5 and 44 DF,  p-value: 7.308e-13
```

6.2 Generalized and Weighted Least-Squares

In order to deal with a violation of the homoscedasticity assumption is to generalize model, by permitting, a less restrictive form for the covariance matrix of the error.
The General Multiple Linear Regression Model (GMLRM) is given by

$$\vec{Y} = X\vec{\beta} + \varepsilon, \quad \varepsilon \sim \mathcal{N}(\vec{0}, \sigma^2 V)$$

For this model the (ordinary) least squares estimator $\hat{\beta}$ is no longer appropriate. It is still unbiased, but not any longer providing the estimate with smallest variance.
The matrix V

1. V is non singular and positive definite. (A matrix A is called positive definite, iff $\vec{x}'A\vec{x} > 0$, for all $\vec{x} \in \mathbb{R}^n \setminus \{\vec{0}\}$). V has to be positive definite since $\sigma^2\vec{x}'A\vec{x} = Var(\vec{x}'\vec{\varepsilon}) > 0$ if $x \in \mathbb{R}^n \setminus \{\vec{0}\}$. Positive matrices are not singular.

2. $V = KK$, for some symmetric matrix $K \in \mathbb{R}^{n \times n}$. (This matrix exists because V is positive definite, K is called the square root of V.)

3. It is usually assumed that V is known, but σ^2 is not. Therefore V gives the structure of variances and covariances in the sample.

In order to be able to apply the finding on least squares estimator from earlier. The matrix K is used to transform the model in such a way that it fits a MLRM, for which we know the least squares estimator etc..

Theorem 6.1.
Let

$$\vec{Z} = K^{-1}\vec{Y}, \quad B = K^{-1}X, \quad \vec{g} = K^{-1}\vec{\varepsilon}$$

Then the by K^{-1} transformed model is

$$\vec{Z} = B\vec{\beta} + \vec{g}$$

is a MLRM with $\vec{g} \sim \mathcal{N}(\vec{0}, \sigma^2 I_n)$

Proof:
Since $\vec{g} = K^{-1}\vec{\varepsilon}$ it is a linear combination of normally distributed random variables and therefore normal, with

$$E(\vec{g}) = E(K^{-1}\vec{\varepsilon}) = K^{-1}E(\vec{\varepsilon}) = K^{-1}\vec{0} = \vec{0}$$

and

$$Cov(\vec{g}) = Cov(K^{-1}\vec{\varepsilon}) = K^{-1}Cov(\vec{\varepsilon})K^{-1} = K^{-1}\sigma^2 V K^{-1} = \sigma^2 K^{-1}KKK^{-1} = \sigma^2 I_n$$

q.e.d.

The least squares estimate for $\vec{\beta}$ derived from this model must then be the least squares estimate of the GMLRM.
The normal equation for the transformed model is

$$\begin{aligned} B'B\hat{\beta} &= B'\vec{Z} \\ \Leftrightarrow \quad X'K^{-1}K^{-1}X\hat{\beta} &= X'K^{-1}K^{-1}\vec{Y} \\ \Leftrightarrow \quad X'V^{-1}X\hat{\beta} &= X'V^{-1}\vec{Y} \end{aligned}$$

with solution

$$\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}\vec{Y}$$

$\hat{\beta}$ is called the generalized least squares estimator of $\vec{\beta}$.

Lemma 1.
The generalized least squares estimator of $\vec{\beta}$, $\hat{\beta}$, is unbiased and has covariance matrix $Cov(\hat{\beta}) = \sigma^2(X'V^{-1}X)^{-1}$.

The residual sum of squares for the GMLRM is

$$SS_{Res} = \vec{Z}'\vec{Z} - \hat{\beta}'B\vec{Z} = \vec{Y}'V^{-1}\vec{Y} - \vec{Y}'V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}\vec{Y}$$

The regression sum of squares

$$SS_R = \hat{\beta}'B\vec{Z} - \frac{1}{n}\vec{Z}'J_n\vec{Z} = \vec{Y}'V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}\vec{Y} - \frac{1}{n}\vec{Y}'K^{-1}J_nK^{-1}\vec{Y}$$

The total sum of squares

$$SS_T = \vec{Z}'\vec{Z} - \frac{1}{n}\vec{Z}'J_n\vec{Z} = \vec{Y}'V^{-1}\vec{Y} - \frac{1}{n}\vec{Y}'K^{-1}J_nK^{-1}\vec{Y}$$

In general the error covariance matrix is not known, but choosing some structure to it permits the fit of more general models, for example in time series.
Here we will focus on Weighted Least Squares Regression, which shall help in the case when the homoscedasticity assumption is violated in the ordinary least squares model.

6.2.1 Weighted Least Squares

If we can assume that the errors are uncorrelated, but that the variance of the errors are NOT the same, then the covariance matrix of $\vec{\varepsilon}$ can be written as

$$Cov(\vec{\varepsilon}) = \sigma^2 V = \sigma^2 \begin{pmatrix} 1/w_1 & 0 & \dots & 0 \\ 0 & 1/w_2 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \dots & 0 & 1/w_n \end{pmatrix}$$

Using this shape for the covariance matrix of $\vec{\varepsilon}$ in the general least squares model, is called weighted least squares regression model or weighted multiple regression model(WMLRM).
Let

$$W = V^{-1} = \begin{pmatrix} w_1 & 0 & \dots & 0 \\ 0 & w_2 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \dots & 0 & w_n \end{pmatrix}$$

then the least squares estimator for $\vec{\beta}$ in the WMLRM is given by

$$\hat{\beta} = (X'WX)^{-}1X'W\vec{Y}$$

To obtain this estimate from R and get the proper output add an option **weight** to lm

```
lm(BP~AGE+ WEIGHT, data=BP.data, weights=wghts)
```

with wghts being the variable including the weights for each measurement.

What happens when the ordinary least squares estimator is used, when $V \neq I_n$?
In this cases $\hat{\beta}_o = (X'X)^{-1}X'\vec{Y}$ is still unbiased, but does not have smallest variance anymore. (Gauss Markov fails for $\hat{\beta}_o$). Therefore weighted least squares is preferable over ordinary least squares in this situation.

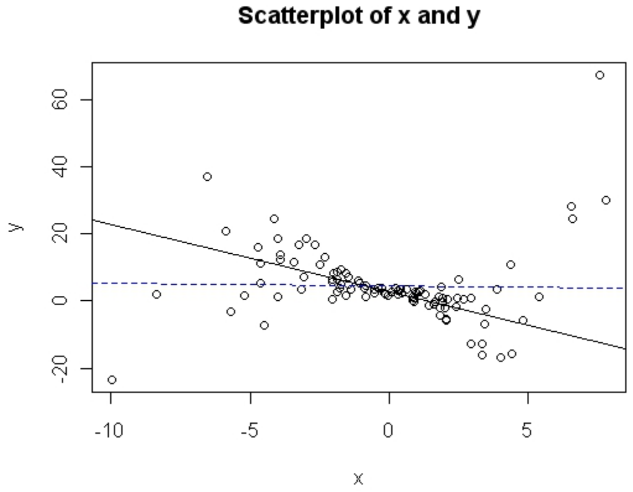
The biggest **problem** with weighted least squares consists of the assumption that the weights are assumed to be known. The most common application of weighted least squares when measurement y_i represents the average of n_i measurements for $x_{1i}, \dots, x_{ki}$. In this case the sample sizes n_i are used as the weights.

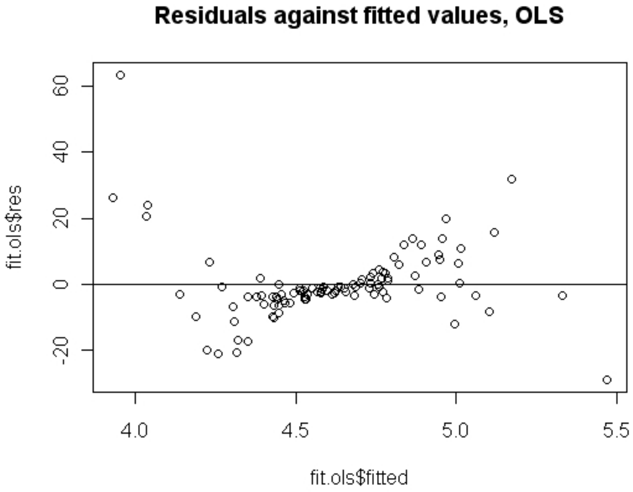
In other instances a residual analysis leads to the observation that $Var(\varepsilon)$ is proportional to one of the predictor variables, x_j, so that $w_i = 1/x_{ij}$ is a good choice.
In many practical situations the weights have to be guessed, and an analysis might show how to improve on the guess until a proper fit can be achieved. See MPV example 5.5 on page 192.

Example 6.3.
(Inspired by http://www.stat.cmu.edu/ cshalizi/350/lectures/18/lecture-18.pdf)

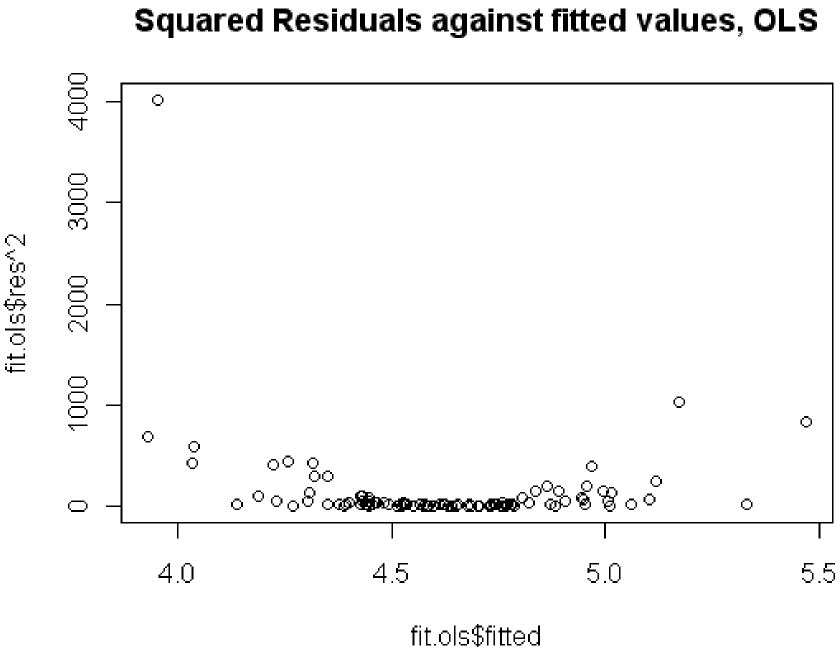
Simulate data with heteroscedastic error, the further x falls away from 0, the larger the variance.

```
x = rnorm(100,0,3)
y = 3-2*x + rnorm(100,0,sapply(x,function(x){1+0.5*x^2}))
plot(x,y)
abline(a=3,b=-2,lty=1)
fit.ols = lm(y~x)
abline(fit.ols$coefficients,lty=2,col="blue")
```


Scatterplot of x and y

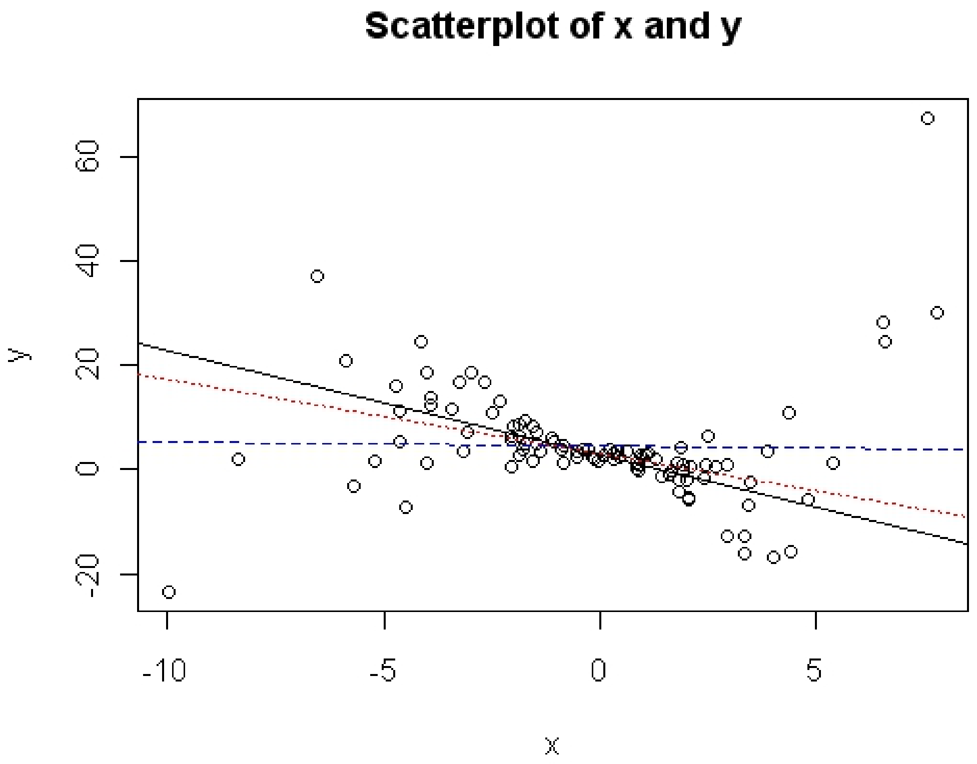

Residuals against fitted values, OLS

The solid line gives the true line, the dashed is the fitted line. And the residual plot clearly shows the violation of homoscedasticity.
The squared residuals are indicative of the variance in the error


Squared Residuals against fitted values, OLS

Use

```
wght<-1/(1+0.5*x^2)
fit.wls = lm(y~x,weight=wght)
```



The solid line is the correct fit $y = 3 - 2x$, the blue dashed line is the fitted line using OLS $y = 4.60 - 0.09x$, and the red dotted line is the fitted line using WLS $y = 3.07 - 1.43x$.
What do the confidence intervals say?

```
> confint(fit.ols) # Confidence intervals for OLS
                2.5 %    97.5 %
(Intercept)  2.358142 6.8555990
x           -0.775504 0.6022992
> confint(fit.wls)  # Confidence intervals for WLS
               2.5 %     97.5 %
(Intercept)  2.15419  3.9945709
x           -1.96547 -0.9086006
```