2.9 Potential problems

Correlation of the error terms

An important assumption of the linear model is that $\varepsilon_1, \ldots, \varepsilon_N$ are **uncorrelated**.

If there is correlation among the error terms, then the estimated standard errors of the coefficients will tend to underestimate the true standard errors.

When does it happen? A classic situation is for **time series** (Observations at adjacent time points are positively correlated).

We investigate the correlation of the errors by **plotting the residuals** w.r.t. time:

● If errors are uncorrelated, there should be no discernible pattern;

● If the error terms are positively correlated, we may see a trend, in particular for adjacent residuals.

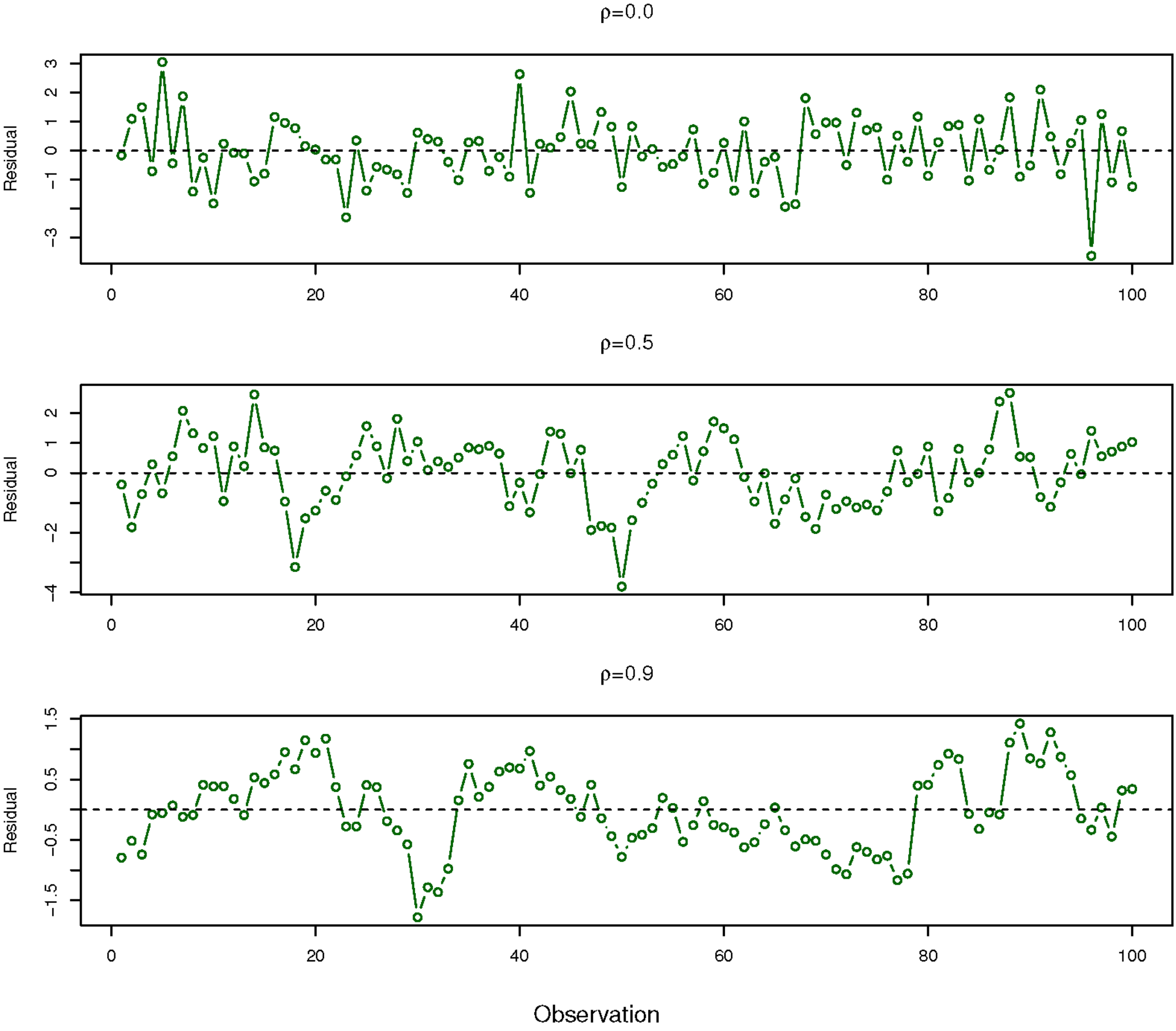


Figure 2.9.1: Residuals from a linear regression fit to data generated with error correlation $\rho = 0, 0.5$ and 0.9.

Non-constant variance

Another important assumption of the linear model is that the *error terms have constant variance*, $\mathbb{V}\mathrm{ar}(\varepsilon_i) = \sigma^2$ for every i.

The case of non-constant variance is called **heteroscedasticity**.

A possible solution is to use a **concave** transformation of the observations, like $\log \mathrm{Y}$ or $\sqrt{\mathrm{Y}}$, which results in a greater amount of shrinkage of the larger responses.

Sometimes we have an idea of the variance of each response: for example, each observation could be an average of n_i observations, there the average can have variance $\sigma_i^2 = \frac{\sigma^2}{n_i}$. A solution to this situation is to fit **weighted least squares**, with weights proportional to $w_i = n_i$.

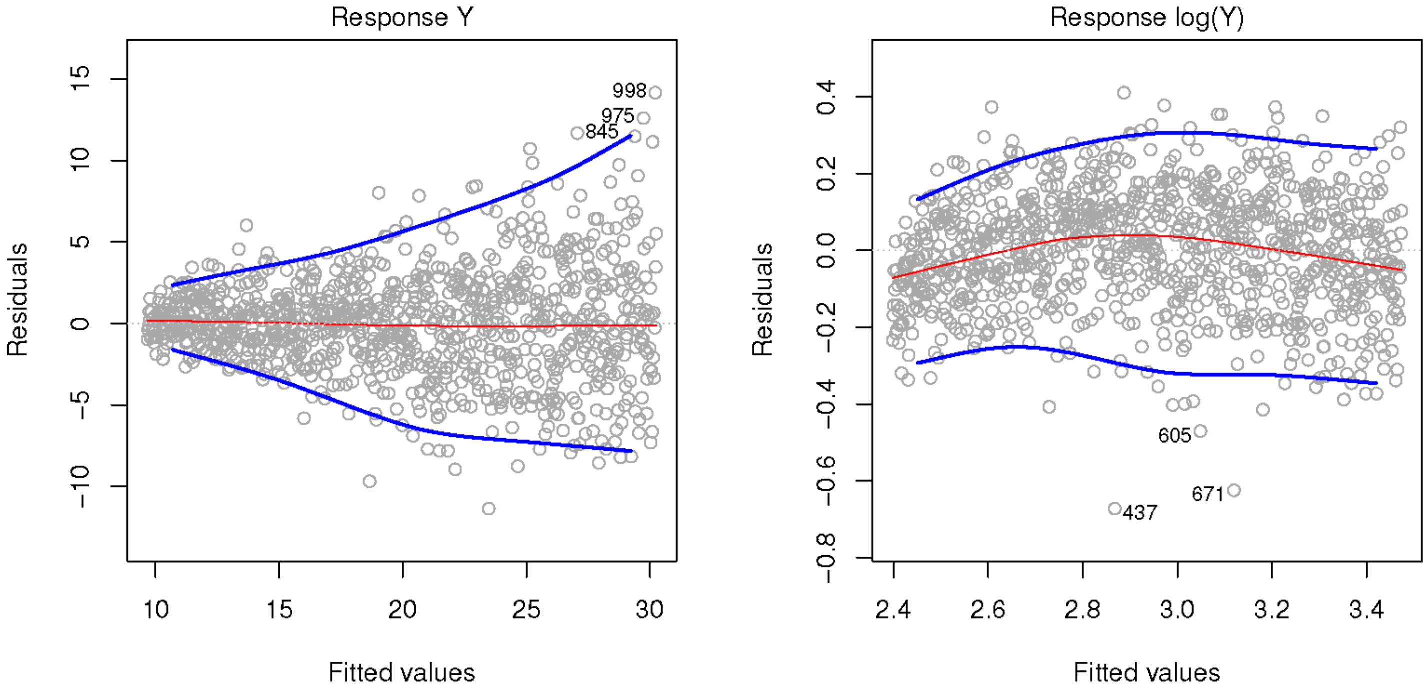


Figure 2.9.2: Residuals for non-transformed (left) and log-transformed data (right). Red line is a smooth fit to the residuals, blue lines give the outer quantile.

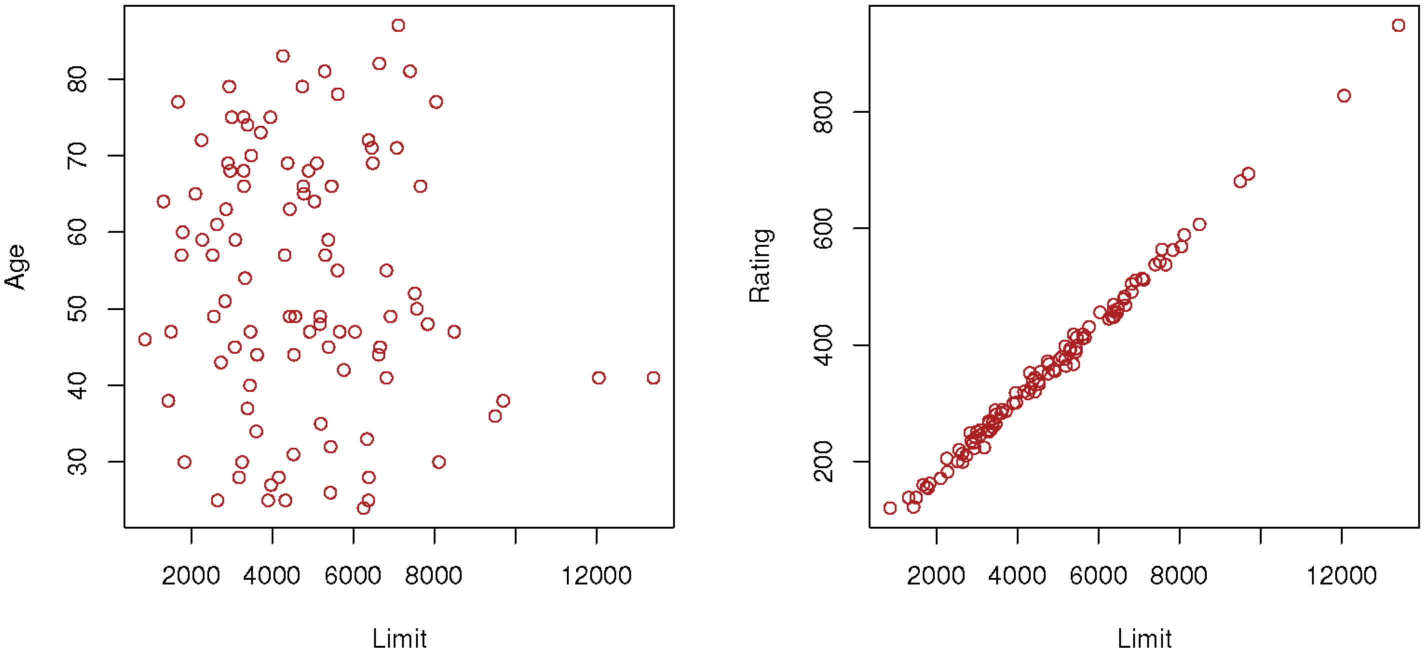
Collinearity

Collinearity refers to the situation in which two or more predictors are closely related to each other.

The presence of collinearity makes it difficult to separate out the individual effects of collinear variables on the response. Moreover, *a small change in the data can cause the coefficient values to be estimated very differently*. This means there is a great uncertainty in the estimates in presence of collinearity.

```
library(ISLR)
data(Credit)
attach(Credit)

par(mfrow=c(1,2))
plot(Limit, Age, xlab="Limit", ylab="Age", col="red")
plot(Limit, Rating, xlab="Limit", ylab="Rating", col="red")
```



A simple way to detect collinearity is to look at the **correlation matrix** of the predictors.

```
library(ISLR)
data(Credit)
attach(Credit)

cor(cbind(Limit, Age, Rating))
```

However, it is possible that collinearity exists among three or more variables even when no pair of variables has high correlation. This situation is called **multicollinearity**.

A way to inspect multicollinearity is the **variance inflation factor**, i.e. the ratio of the variance of $\hat{\beta}_j$ when fitting the full model divided by the variance of $\hat{\beta}_j$ if fit on its own.

$$\mathrm{VIF}(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|x_{-j}}} \geq 1 \tag{2.9.1}$$

where $R^2_{x_j|x_{-j}}$ is the R^2 statistic from regression of X_j on all the other predictors.

- If $\mathrm{VIF}(\hat{\beta}_j) = 1$ there is no collinearity
- If $\mathrm{VIF}(\hat{\beta}_j) > 5$ there is a problem

Solutions:

- drop one of the problematic variables
- combine the collinear variables into a single predictor (e.g. taking the average of each pair of predictors)

In the **`Credit`** dataset, if we were to regress **`balance`** on **`limit`**, **`rating`** and **`age`**, we could then compute the VIF for each of the three predictors as follows:

```
library(ISLR)
data(Credit)
attach(Credit)

lm.limit <- lm(Limit ~ Rating + Age)
1 / (1- summary(lm.limit)$r.squared) # VIF Limit

lm.rating <- lm(Rating ~ Limit + Age)
1 / (1- summary(lm.rating)$r.squared) # VIF Rating

lm.age <- lm(Age ~ Rating + Limit)
1 / (1- summary(lm.age)$r.squared) # VIF Age
```

Alternatively one could install the **`car`** package in **`R`** and directly use the **`vif()`** function.

```
library(ISLR)
library(car)

data(Credit)
attach(Credit)

lm.full <- lm(Balance ~ Limit + Rating + Age)
vif(lm.full)
```