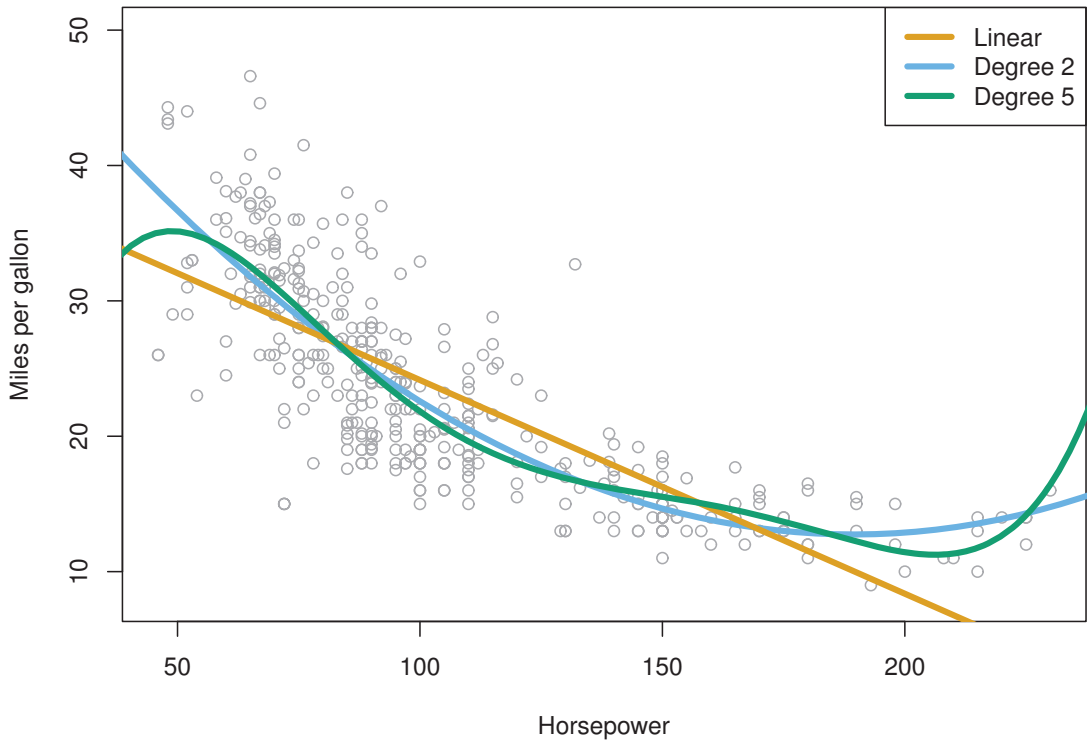
Non-linear Relationships

- In some cases, the true relationship between the response and the predictors may be non-linear.
- We can directly extend the linear model to accommodate non-linear relationships, using polynomial regression.
- Consider the Figure in which the mpg (gas mileage in miles per gallon) versus horsepower is shown for a number of cars in the Auto data set.
- The points in the Figure seem to have a quadratic shape, suggesting that a model of the form

$$\text{mpg} = \beta_0 + \beta_1 \times \text{horsepower} + \beta_2 \times (\text{horsepower})^2 + \epsilon \qquad (*)$$

may provide a better fit.

(*) is still a linear model! That is, the model is simply a multiple linear regression model with $X_1 = \text{horsepower}$ and $X_2 = (\text{horsepower})^2$. So we can use standard linear regression software to estimate β_0, β_1, and β_2 in order to produce a non-linear fit.



The Auto data set. For a number of cars, mpg and horsepower are shown. The linear regression fit is shown in orange. The linear regression fit for a model that includes horsepower^2 is shown as a blue curve. The linear regression fit for a model that includes all polynomials of horsepower up to fifth-degree is shown in green.

```
> auto <- read.table("Auto.data", header = T, na.strings = "?",
   stringsAsFactors = T)
> names(auto)
[1] "mpg"          "cylinders"    "displacement" "horsepower"
[5] "weight"       "acceleration" "year"         "origin"
[9] "name"
> dim(auto)
[1] 397   9
> auto <- na.omit(auto) # removes rows with missing observations
> dim(auto)
[1] 392   9
> lm_fit_auto_1 = lm(mpg ~ horsepower , data=auto)
```

```
> summary(lm_fit_auto_1)

Call:
lm(formula = mpg ~ horsepower, data = auto)

Residuals:
     Min       1Q   Median       3Q      Max
-13.5710  -3.2592  -0.3435   2.7630  16.9240

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.935861   0.717499   55.66   <2e-16 ***
horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.906 on 390 degrees of freedom
Multiple R-squared:  0.6059,Adjusted R-squared:  0.6049
F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16

> lm_fit_auto_2 = lm(mpg ~ horsepower + I(horsepower^2), data=auto)
> summary(lm_fit_auto_2)

Call:
lm(formula = mpg ~ horsepower + I(horsepower^2), data = auto)

Residuals:
     Min       1Q   Median       3Q      Max
-14.7135  -2.5943  -0.0859   2.2868  15.8961

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     56.9000997  1.8004268   31.60   <2e-16 ***
horsepower      -0.4661896  0.0311246  -14.98   <2e-16 ***
I(horsepower^2)  0.0012305  0.0001221   10.08   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.374 on 389 degrees of freedom
Multiple R-squared:  0.6876,Adjusted R-squared:  0.686
F-statistic:   428 on 2 and 389 DF,  p-value: < 2.2e-16
```

```
> anova(lm_fit_auto_1,lm_fit_auto_2)
Analysis of Variance Table

Model 1: mpg ~ horsepower
Model 2: mpg ~ horsepower + I(horsepower^2)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    390 9385.9
2    389 7442.0  1    1943.9 101.61 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The anova() function performs a hypothesis test comparing the two models. The null hypothesis is that the two models fit the data equally well, and the alternative hypothesis is that the full model is superior. Here the F-statistic is 101.61 and the associated p-value is virtually zero. This provides very clear evidence that the model containing the predictors horsepower and horsepower^2 is far superior to the model that only contains the predictor horsepower.

- In order to create a higher order fit, we can use the poly() function to create the polynomial with a given order within lm().

```
> lm_fit_auto_5 = lm(mpg ~ poly(horsepower, 5, raw = TRUE),  data=auto)
> summary(lm_fit_auto_5)

Call:
lm(formula = mpg ~ poly(horsepower, 5, raw = TRUE), data = auto)

Residuals:
     Min       1Q   Median       3Q      Max
-15.4326  -2.5285  -0.2925   2.1750  15.9730

Coefficients:
                                    Estimate Std. Error t value
(Intercept)                       -3.223e+01  2.857e+01  -1.128
poly(horsepower, 5, raw = TRUE)1  3.700e+00  1.303e+00   2.840
poly(horsepower, 5, raw = TRUE)2 -7.142e-02  2.253e-02  -3.170
poly(horsepower, 5, raw = TRUE)3  5.931e-04  1.850e-04   3.206
poly(horsepower, 5, raw = TRUE)4 -2.281e-06  7.243e-07  -3.150
```

```
poly(horsepower, 5, raw = TRUE)5  3.330e-09  1.085e-09   3.068
                                 Pr(>|t|)
(Intercept)                       0.26003
poly(horsepower, 5, raw = TRUE)1  0.00475 **
poly(horsepower, 5, raw = TRUE)2  0.00164 **
poly(horsepower, 5, raw = TRUE)3  0.00146 **
poly(horsepower, 5, raw = TRUE)4  0.00176 **
poly(horsepower, 5, raw = TRUE)5  0.00231 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.326 on 386 degrees of freedom
Multiple R-squared:  0.6967,Adjusted R-squared:  0.6928
F-statistic: 177.4 on 5 and 386 DF,  p-value: < 2.2e-16
```

Note: We are in no way restricted to using polynomial transformations of the predictors. Here we try a log transformation.

```
> summary(lm(mpg ~  log(horsepower) , data = auto))

Call:
lm(formula = mpg ~ log(horsepower), data = auto)

Residuals:
     Min       1Q   Median       3Q      Max
-14.2299  -2.7818  -0.2322   2.6661  15.4695

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     108.6997     3.0496   35.64   <2e-16 ***
log(horsepower) -18.5822     0.6629  -28.03   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.501 on 390 degrees of freedom
Multiple R-squared:  0.6683,Adjusted R-squared:  0.6675
F-statistic: 785.9 on 1 and 390 DF,  p-value: < 2.2e-16
```

Potential Problems

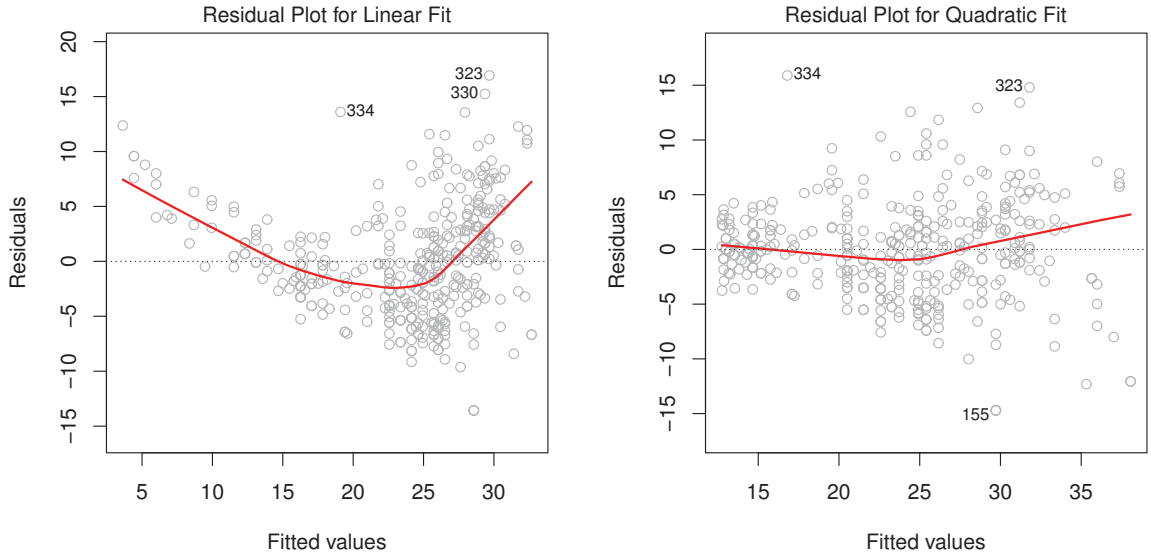
When we fit a linear regression model to a particular data set, many problems may occur. Most common among these are the following:

- Non-linearity of the response-predictor relationships.
- Correlation of error terms.
- Non-constant variance of error terms.
- Outliers.
- High-leverage points.
- Collinearity.

Non-linearity of the Data

- If the true relationship between the predictors and the response is far from linear, then virtually all of the conclusions that we draw from the fit are suspect.
- In addition, the prediction accuracy of the model can be significantly reduced.
- Residual plots are a useful graphical tool for identifying non-linearity. Given a simple linear regression model, we can plot the residuals, $e_i = y_i - \hat{y}_i$, versus the predictor x_i.
- In the case of a multiple regression model, since there are multiple predictors, we instead plot the residuals versus the predicted (or fitted) values $\hat{y}_i$.
- Ideally, the residual plot will show no discernible pattern.

The presence of a pattern may indicate a problem with some aspect of the linear model.



Plots of residuals versus predicted (or fitted) values for the Auto data set. Left: A linear regression of mpg on horsepower. A strong pattern in the residuals indicates non-linearity in the data. Right: A linear regression of mpg on horsepower and horsepower^2. There is little pattern in the residuals, suggesting that the quadratic term improves the fit to the data.

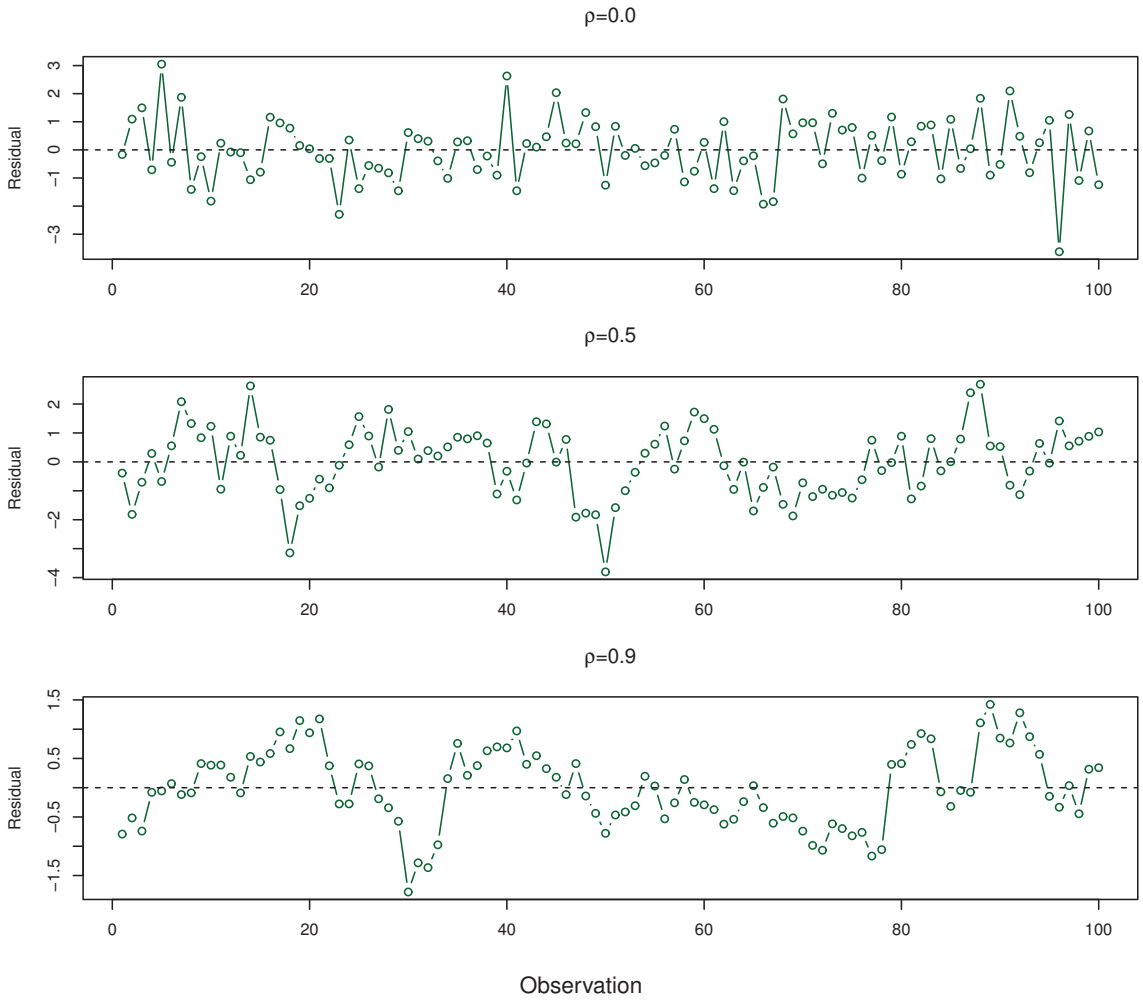
- If the residual plot indicates that there are non-linear associations in the data, then a simple approach is to use non-linear transformations of the predictors, such as $\log X$, $\sqrt{X}$, and X^2, in the regression model.

Correlation of Error Terms

- An important assumption of the linear regression model is that the error terms, $\epsilon_1, \epsilon_2, ..., \epsilon_n$, are uncorrelated.
- If there is correlation among the error terms, then the estimated standard errors will tend to underestimate the true standard errors. As a result, confidence and prediction intervals will be narrower than they should be.

If the error terms are correlated, we may have an unwarranted sense of confidence in our model.

- Such correlations frequently occur in the context of time series data, which consists of observations for which measurements are obtained at discrete points in time.
- In many cases, observations that are obtained at adjacent time points will have positively correlated errors.
- In order to determine if this is the case for a given data set, we can plot the residuals from our model as a function of time. If the errors are uncorrelated, then there should be no discernible pattern.
- On the other hand, if the error terms are positively correlated, then we may see tracking in the residuals—that is, adjacent residuals may have similar values.



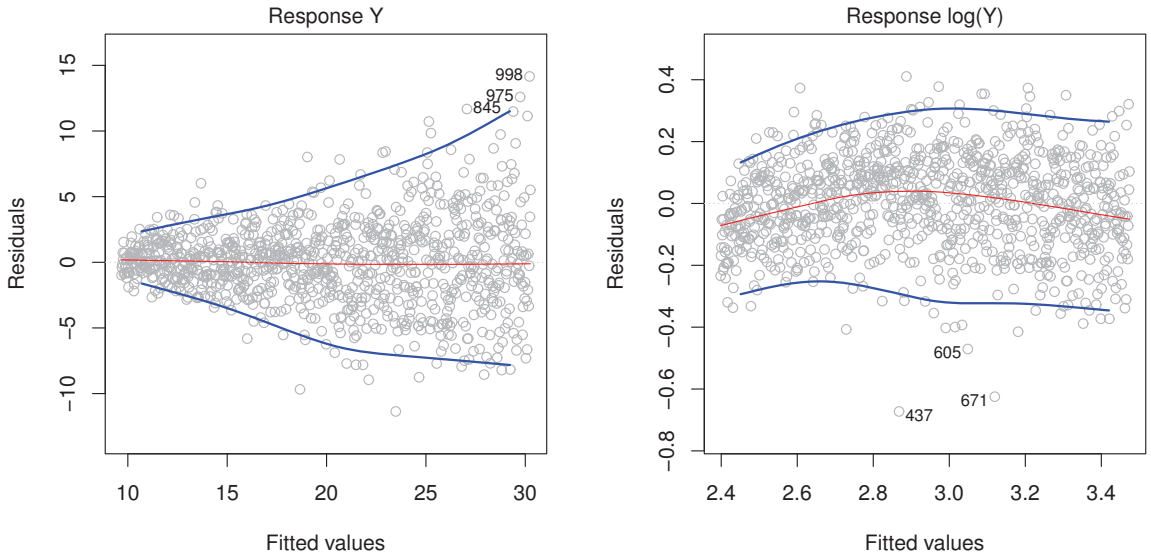
Plots of residuals from simulated time series data sets generated with differing levels of correlation ρ between error terms for adjacent time points.

Non-constant Variance of Error Terms

- Another important assumption of the linear regression model is that the error terms have a constant variance, $\text{Var}(\epsilon_i) = \sigma^2$. The standard errors, confidence intervals, and hypothesis tests associated with the linear model rely upon this assumption.

- Unfortunately, it is often the case that the variances of the error terms are non-constant. For instance, the variances of the error terms may increase with the value of the response.
- One can identify non-constant variances in the errors, or heteroscedasticity, from the presence of a funnel shape in the residual plot.
- When faced with this problem, one possible solution is to transform the response Y using a concave function such as $\log Y$ or $\sqrt{Y}$.

such a transformation results in a greater amount of shrinkage of the larger responses, leading to a reduction in heteroscedasticity.



Residual plots. In each plot, the red line is a smooth fit to the residuals, intended to make it easier to identify a trend. The blue lines track the outer quantiles of the residuals, and emphasize patterns. Left: The funnel shape indicates heteroscedasticity. Right: The response has been log transformed, and there is now no evidence of heteroscedasticity.