Statistical Learning for Data Science

Lecture 11

唐晓颖

电子与电气工程系

南方科技大学

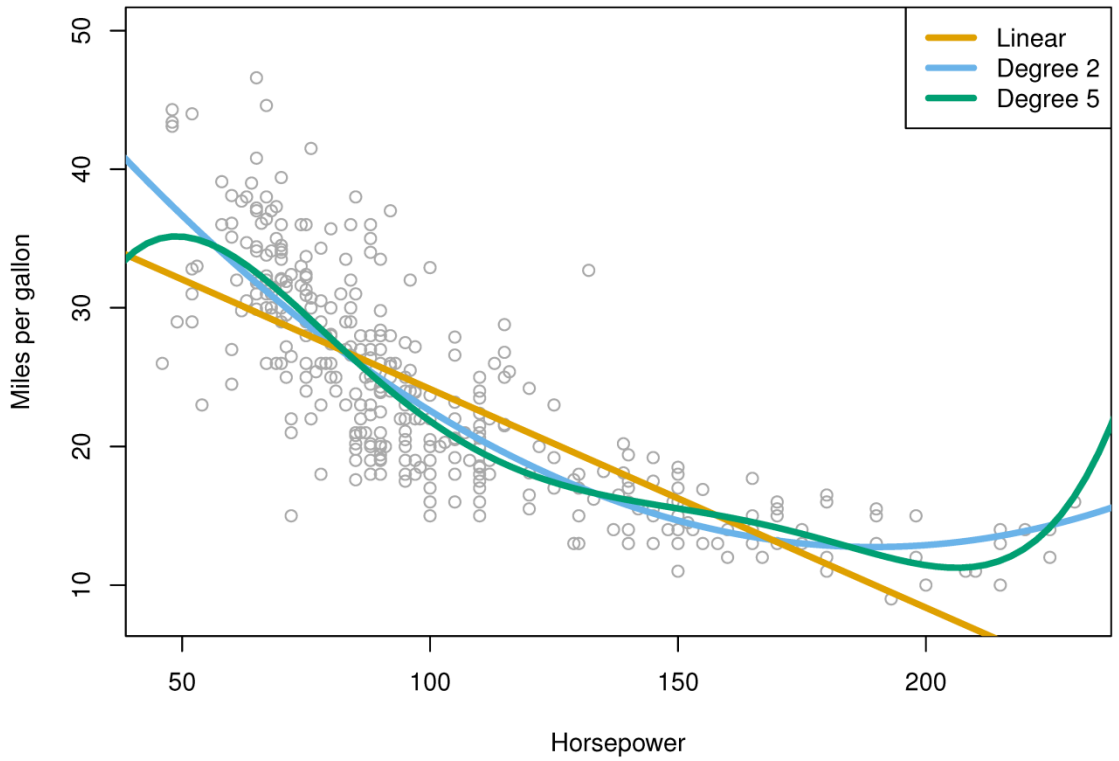
March 29, 2023

Other considerations in the regression model

- Extensions of the linear model

Polynomial regression (quadratic)

$$\text{mpg} = \beta_0 + \beta_1 \times \text{horsepower} + \beta_2 \times \text{horsepower}^2 + \varepsilon$$



Linear

$R^2 = 0.606$

Degree 2

$R^2 = 0.688$

Other considerations in the regression model

- Extensions of the linear model

Polynomial regression (quadratic)

$$\text{mpg} = \beta_0 + \beta_1 \times \text{horsepower} + \beta_2 \times \text{horsepower}^2 + \varepsilon$$

	Coefficient	Std. Error	t-statistic	p-value
`Intercept`	56.9001	1.8004	31.6	< 0.0001
`horsepower`	-0.4662	0.0311	-15.0	< 0.0001
`horsepower`^{2}	0.0012	0.0001	10.1	< 0.0001

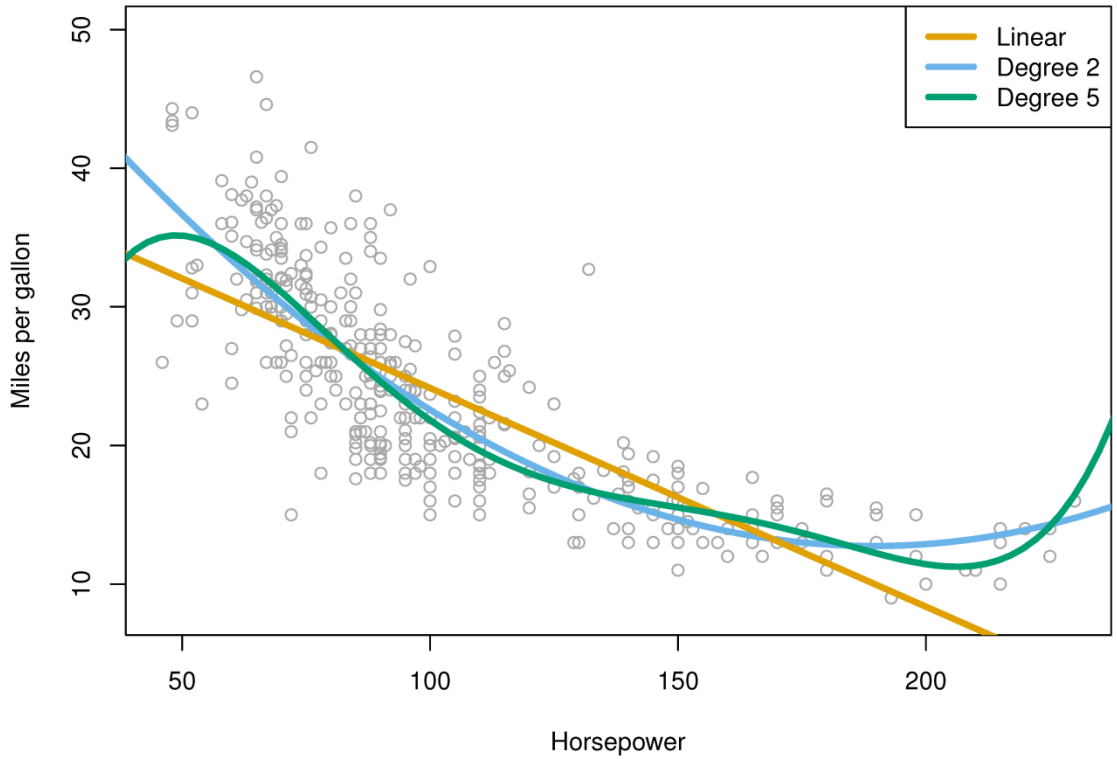
Conclusion: the p-value for the quadratic term is highly significant.

Other considerations in the regression model

- Extensions of the linear model

Polynomial regression (quadratic)

Why not even higher degree (e.g. 5)?



- It is unclear (not obvious) that including the additional terms really has led to a better fit to the data.
- It increased the fitting difficulty.

Other considerations in the regression model

- Potential problems of fitting a linear regression model
  - Non-linearity of the response-predictor relationships.
  - Correlation of error terms.
  - Non-constant variance of error terms.
  - Outliers.
  - High-leverage points.
  - Co-linearity.

Other considerations in the regression model

- Potential problems of fitting a linear regression model

  Non-linearity of the data

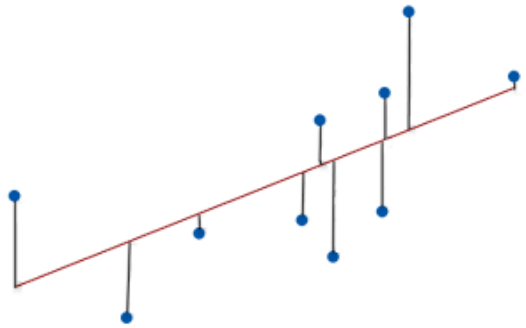
  - If the true relationship is far from linear, then virtually all of the conclusions that we draw from the fit are suspect.

  - In addition, the prediction accuracy of the model can be significantly reduced.

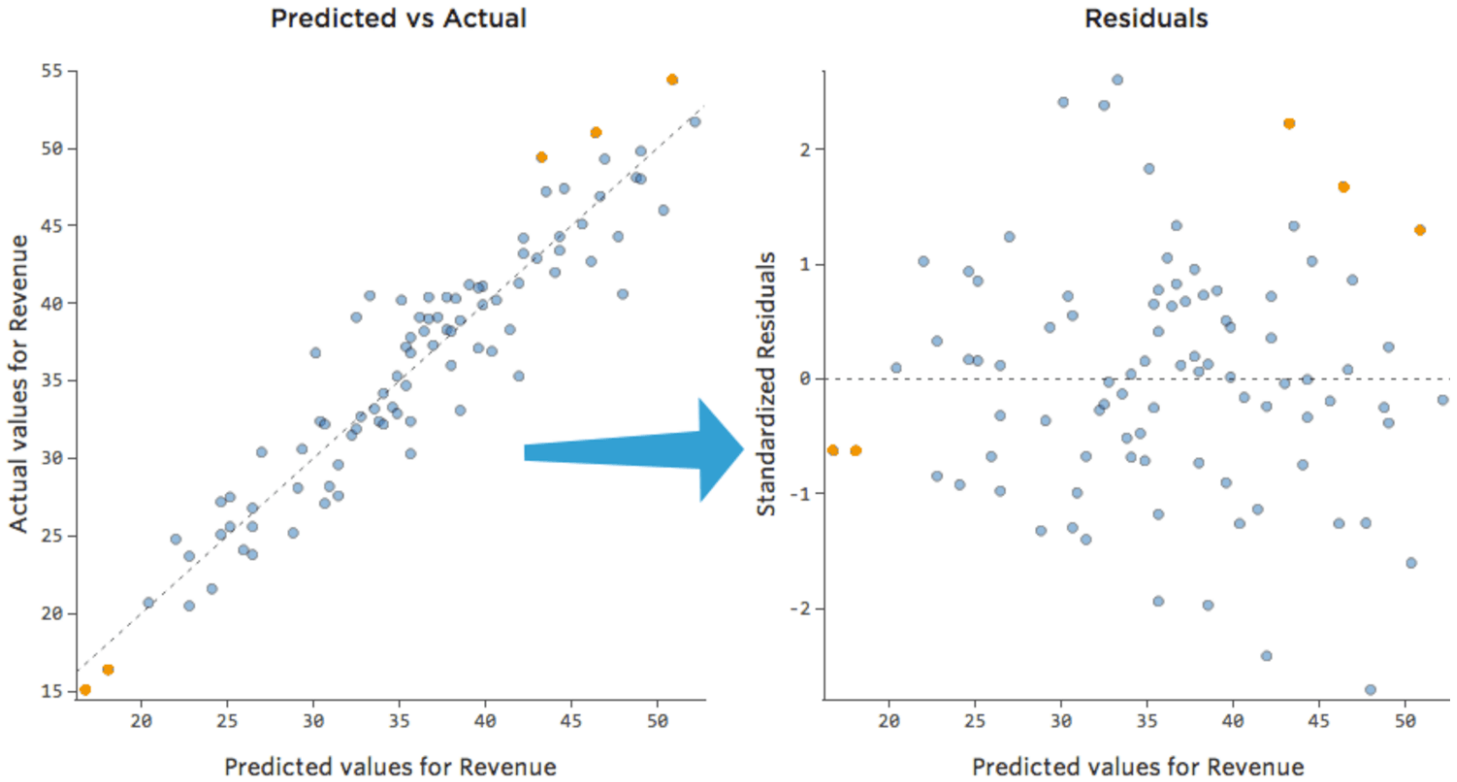
  - Residual plots (residuals vs. predictor OR residuals vs. predicted values) are a useful graphical tool for identifying non-linearity. **Ideally, the residual plot will show no discernible pattern.** The presence of a pattern may indicate a problem with some aspect of the linear model.

Other considerations in the regression model

- Residual plots



Residuals = Observed value - Fitted value



Note that we've colored in a few dots in orange so you can get the sense of how this transformation works.

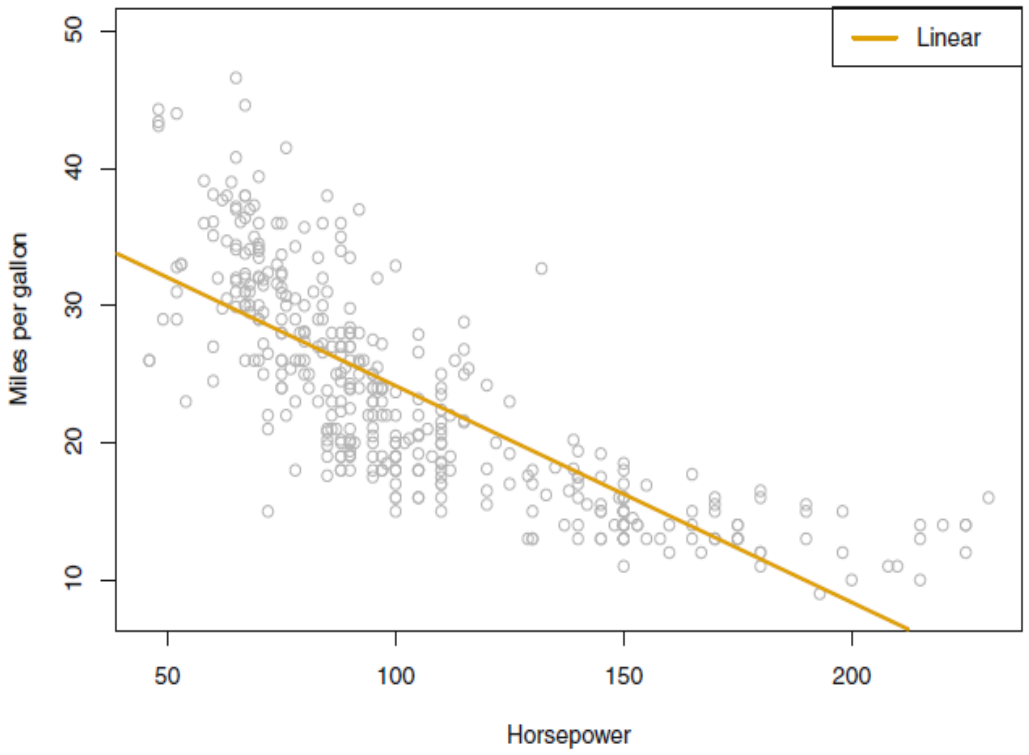
Other considerations in the regression model

- Potential problems of fitting a linear regression model

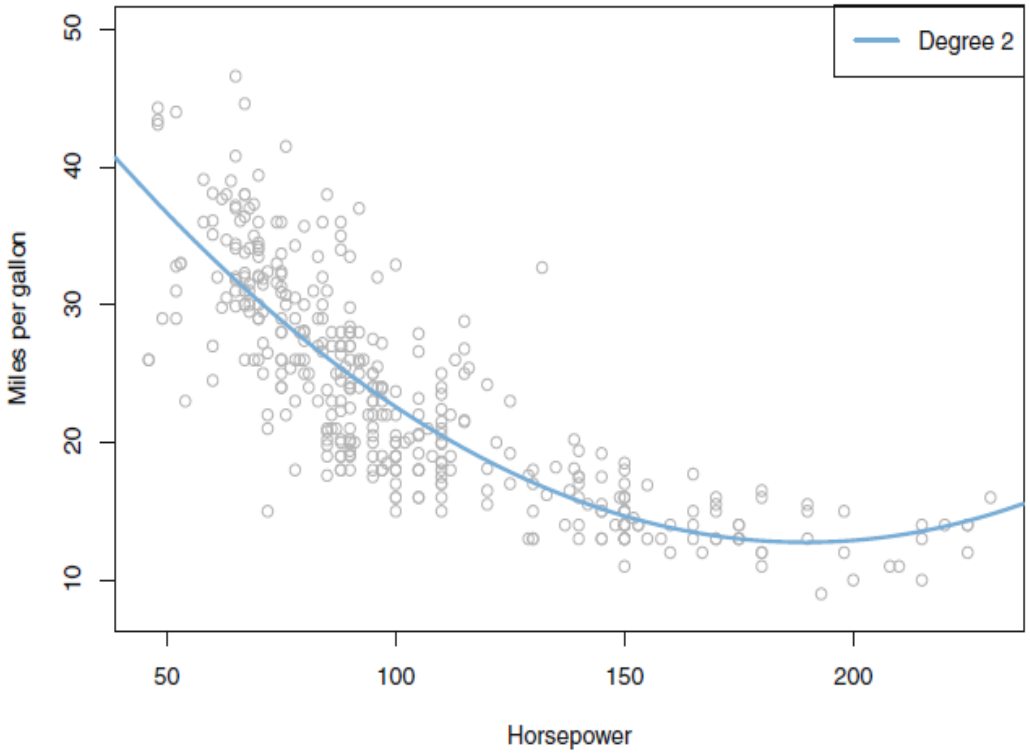
  Non-linearity of the data

  mpg vs. horsepower

Linear Fit



Quadratic Fit

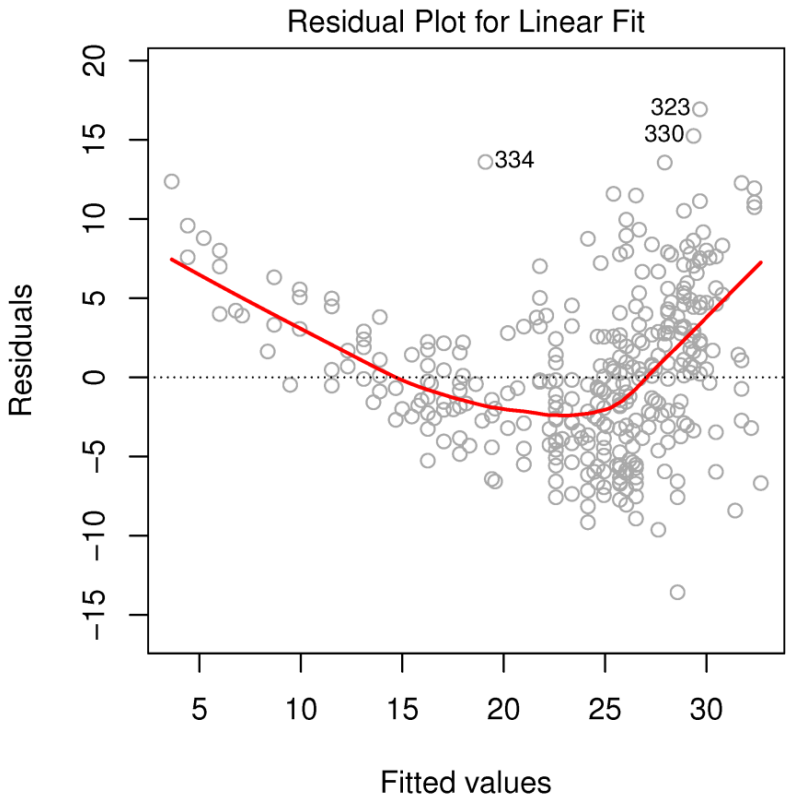


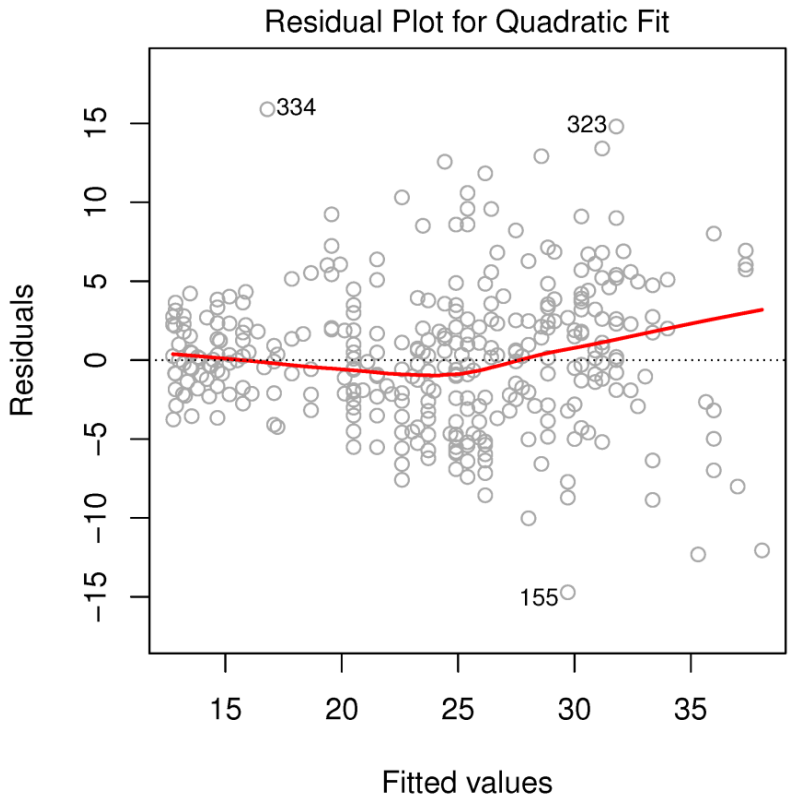
Other considerations in the regression model

- Potential problems of fitting a linear regression model

  Non-linearity of the data

  mpg vs. horsepower





Other considerations in the regression model

- Potential problems of fitting a linear regression model

  Non-linearity of the data

  Suggestion: 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.

Other considerations in the regression model

- Potential problems of fitting a linear regression model

  Correlation of error terms

  An important assumption of the linear regression model is that the error terms, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$, are uncorrelated.

  Question: What is the difference between **independent** and **uncorrelated** random variables? Describe your understanding without any math formula.

  Answer: If two random variables are uncorrelated, they have no linear dependence, but they might have a dependence that is nonlinear. If two random variables are independent they have no dependence at all. So being independent means being uncorrelated as well. But being uncorrelated does not necessarily guarantee being independent.

Other considerations in the regression model

- Potential problems of fitting a linear regression model

  Correlation of error terms

  An important assumption of the linear regression model is that the error terms, $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$, are uncorrelated.

  - The SEs that are computed for the estimated regression coefficients or the fitted values are based on the assumption of uncorrelated error terms.
  - If in fact there is correlation among the error terms, then the estimated SEs will tend to underestimate the true standard errors.
  - Confidence and prediction intervals will be narrow than they should be.
  - P-values associated with the model will be lower than they should be.

Other considerations in the regression model

- Potential problems of fitting a linear regression model

  Correlation of error terms

  An important assumption of the linear regression model is that the error terms, $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$, are uncorrelated.

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

Other considerations in the regression model

- Potential problems of fitting a linear regression model

  Correlation of error terms

  Example: suppose we accidently doubled our data, leading to observations and error terms identical in pairs. If we ignored this, our SE calculations would be as if we had a sample of size 2n, when in fact we have only n samples. Our estimated parameters would be the same for the 2n samples as for the n samples, but the confidence intervals would be narrower by a factor of $\sqrt{2}$!

Other considerations in the regression model

- Potential problems of fitting a linear regression model

  Correlation of error terms

  Why might correlations among the error terms occur?

  - In the context of ***time series data***, it consists of observations for which measurements are obtained at discrete points in time.

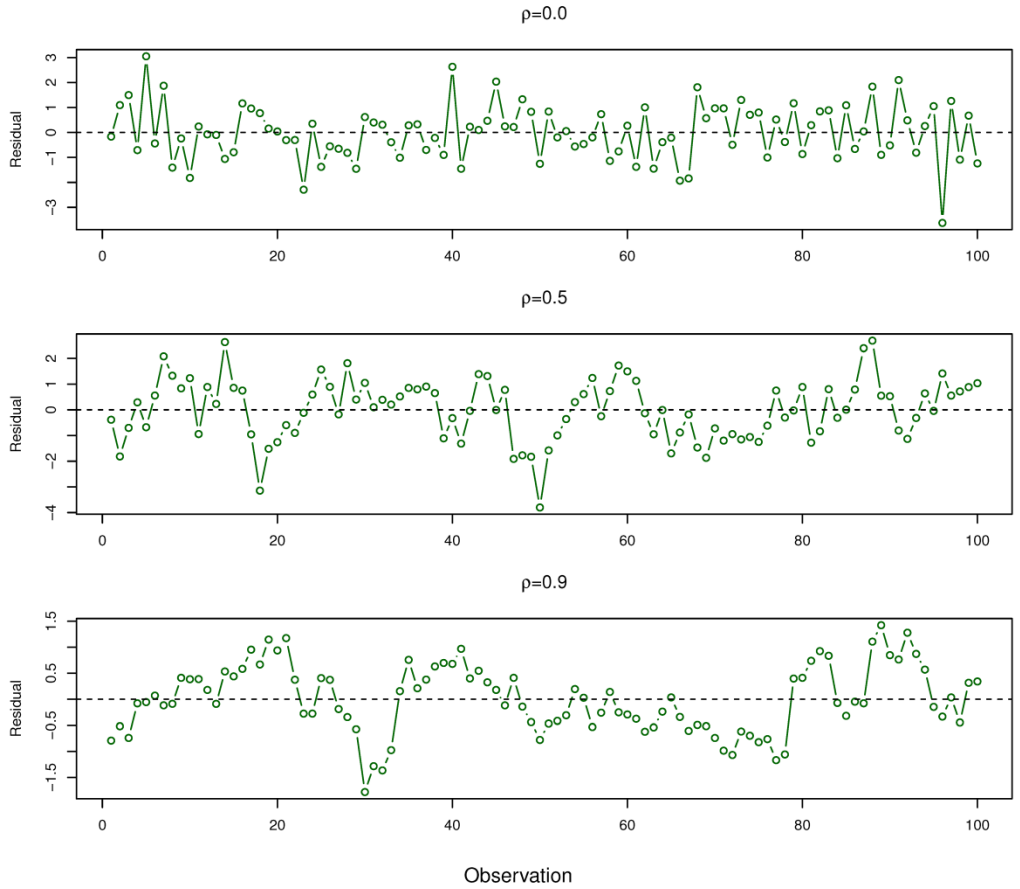
  - Observations that are obtained at adjacent time points will have positively correlated errors.

  - To determine whether this is the case, we can plot the residuals from our model as a function of time. If the error terms are positively correlated, then we may see ***tracking*** in the pattern (adjacent residuals may have similar values).

Other considerations in the regression model

- Potential problems of fitting a linear regression model

Correlation of error terms



No evidence of a time-related trend in the residuals.

Evidence of tracking, but the pattern is less clear.

A clear pattern in the residuals – adjacent residuals tend to take on similar values.

Other considerations in the regression model

- Potential problems of fitting a linear regression model

  Correlation of error terms

  Why might correlations among the error terms occur?

  - Outside of time series data.

  Example: consider a study in which individuals' heights are predicted from their weights. The assumption of uncorrelated errors should be violated if some of the individuals in the study are members of the same family, or eat the same diet, or have been exposed to the same environmental factors.

Other considerations in the regression model

- Potential problems of fitting a linear regression model

  Correlation of error terms

  Suggestion: the assumption of uncorrelated errors is extremely important for linear regression as well as for other statistical methods, and good **experimental design** is crucial in order to mitigate the risk of such correlations.

Other considerations in the regression model

- Potential problems of fitting a linear regression model

  Non-constant variance of error terms

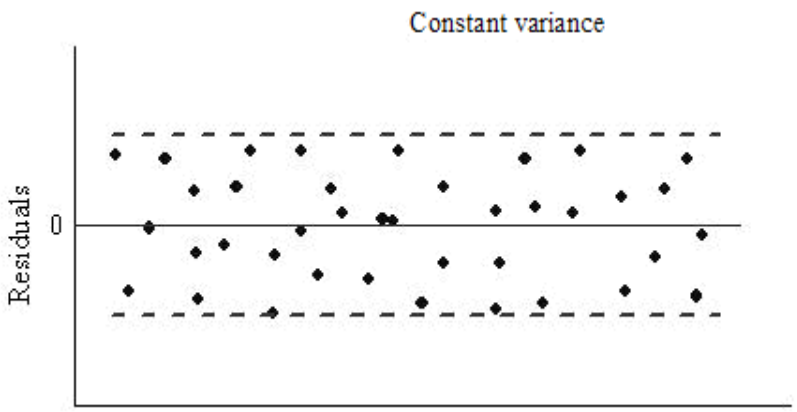
  Another important assumption of the linear regression model is that the error terms have a constant variance, $\text{var}(\varepsilon_i) = \sigma^2$. The SEs, confidence intervals, and hypothesis testing associated with the linear model rely upon this assumption.

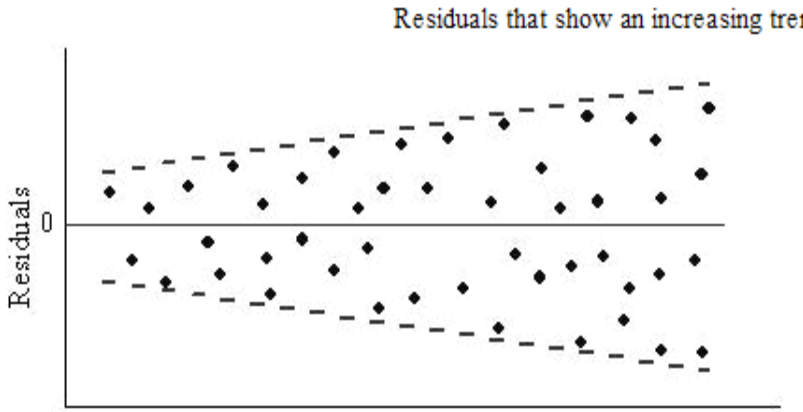
  - It is often the case that the variances of the error terms are non-constant.
  - One can identify non-constant variances in the errors, or *heteroscedasticity*, from the presence of a *funnel shape*（漏斗形状）in the residual plot.

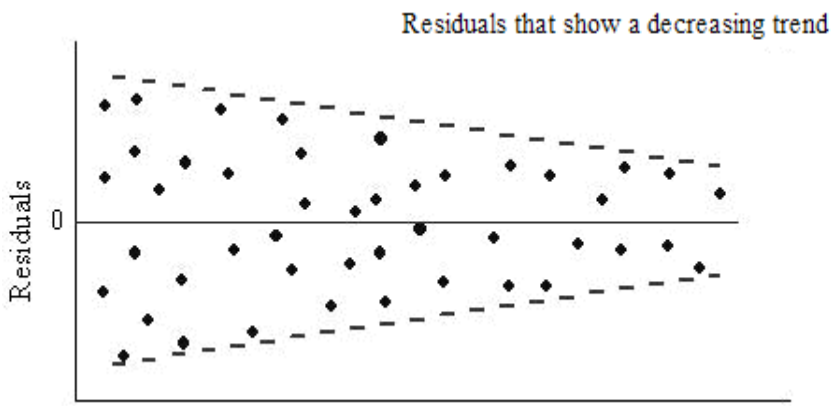
Other considerations in the regression model

- Potential problems of fitting a linear regression model

  Non-constant variance of error terms



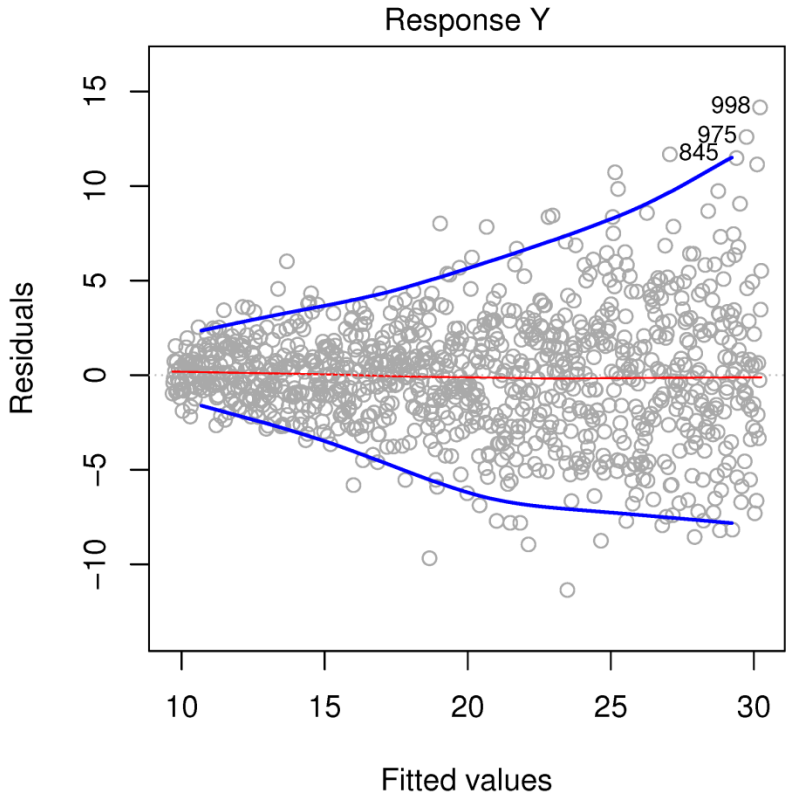




Other considerations in the regression model

- Potential problems of fitting a linear regression model

  Non-constant variance of error terms

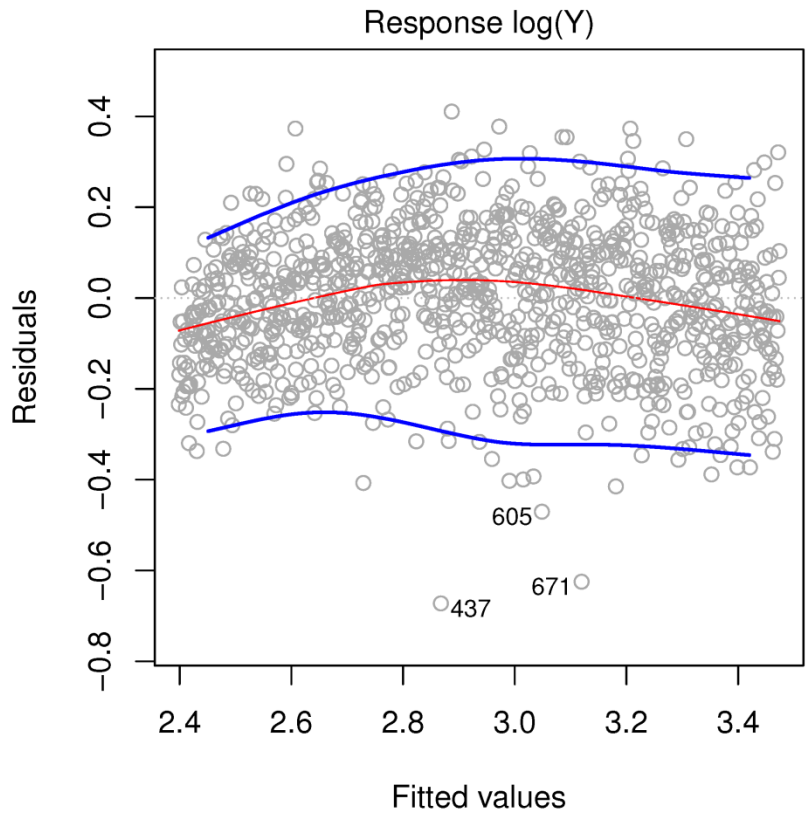


Other considerations in the regression model

- Potential problems of fitting a linear regression model

  Non-constant variance of error terms

Potential solution: transform the response Y using a **concave function** such as $\log Y$ or $\sqrt{Y}$



Other considerations in the regression model

- Potential problems of fitting a linear regression model

  Non-constant variance of error terms

  Potential solution: transform the response Y using a **concave function** such as $\log Y$ or $\sqrt{Y}$.

  Questions:

  1. What's the definition of a concave function?

     $$f((1-a)x+ay) \geq (1-a)f(x)+af(y), \forall a \in [0,1]$$

  2. Quickly show that $\log Y$ is a concave function.

Other considerations in the regression model

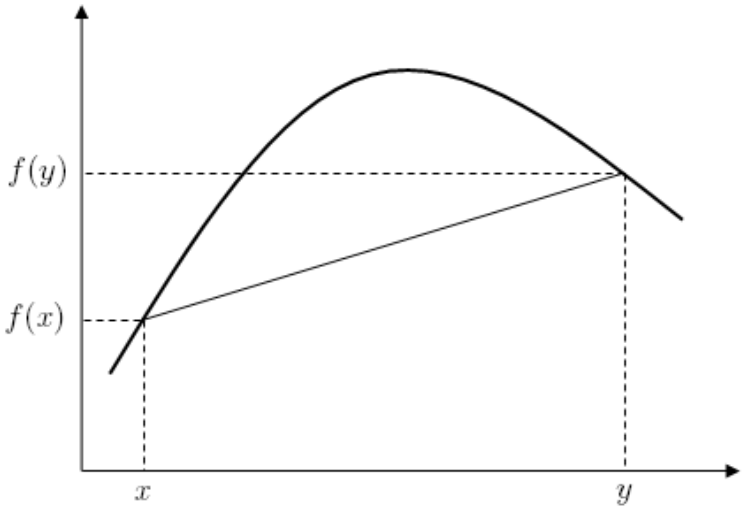
- Potential problems of fitting a linear regression model

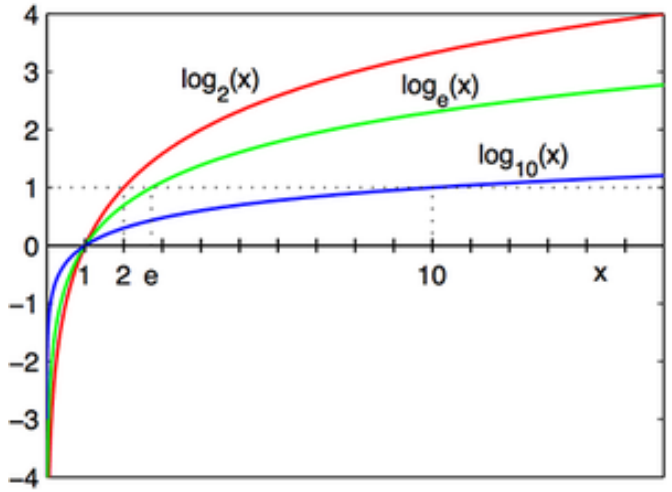
  Non-constant variance of error terms

  Potential solution: transform the response Y using a **concave function** such as $\log Y$ or $\sqrt{Y}$.

  An important property of concave functions:

  For a function f: R→R, the definition of a concave function merely states that for every z between x and y, the point (z, f(z)) on the graph of f is above the straight line joining the points (x, f(x)) and (y, f(y)).





Other considerations in the regression model

- Potential problems of fitting a linear regression model

  Non-constant variance of error terms

  Sometimes we have a good idea of the variance of each response.

  For example: The i-th response could be an average of n_i raw observations. If each of these raw observations is uncorrelated with variance σ^2, then their average has variance $\sigma_i^2 = \sigma^2 / n_i$. In this case a simple remedy is to fit our model by ***weighted least squares***, with weights proportional to the inverse variances – i.e. $w_i = n_i$ in this case. Most linear regression software allows for observation weights.

Other considerations in the regression model

- Potential problems of fitting a linear regression model

  Non-constant variance of error terms

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

  - Ordinary least squares (constant variance)

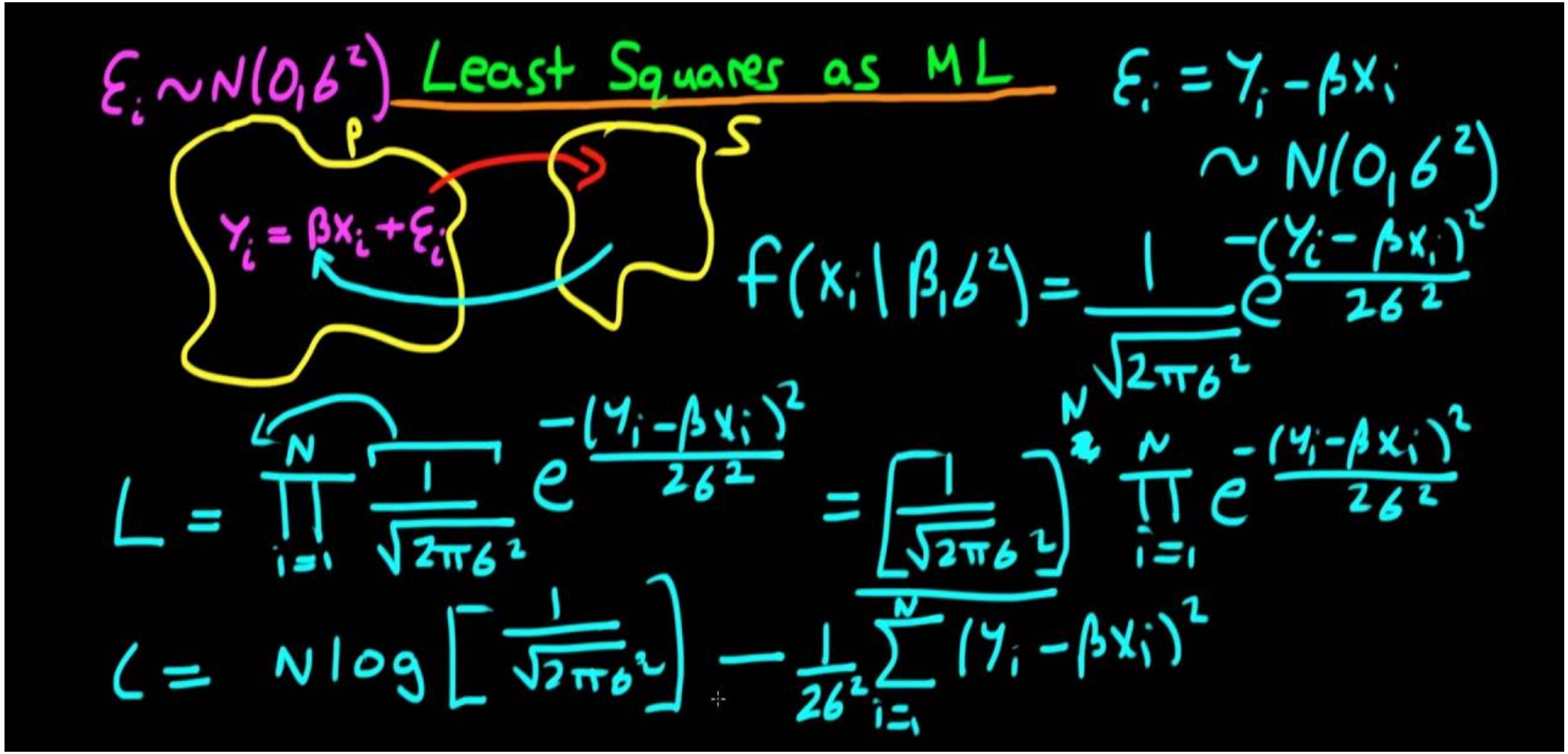
$$\varepsilon \sim \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix}$$

(multivariate) normally distributed with mean vector **0** and constant variance-covariance matrix

$$\hat{\beta}_{\text{OLS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Other considerations in the regression model

- Potential problems of fitting a linear regression model



Other considerations in the regression model

- Potential problems of fitting a linear regression model

  Non-constant variance of error terms

  $$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

  - Weighted least squares (non-constant variance)

$$\varepsilon \sim \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_n^2 \end{pmatrix}$$

(multivariate) normally distributed with mean vector **0** and nonconstant variance-covariance matrix

Other considerations in the regression model

$$\frac{\partial \mathbf{a}^T\mathbf{y}}{\partial \mathbf{y}} = \mathbf{a}$$

$$\frac{\partial \mathbf{a}^T\mathbf{y}\mathbf{b}}{\partial \mathbf{y}} = \mathbf{a}\mathbf{b}^T$$

$$\frac{\partial \mathbf{a}^T\mathbf{y}^T\mathbf{b}}{\partial \mathbf{y}} = \mathbf{b}\mathbf{a}^T$$

- Potential problems of fitting a linear regression model

  Non-constant variance of error terms

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

  - Weighted least squares (non-constant variance)

$$\mathbf{W} = \begin{pmatrix} 1/\sigma_1^2 & 0 & \dots & 0 \\ 0 & 1/\sigma_2^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1/\sigma_n^2 \end{pmatrix}$$

$$\hat{\beta}_{\text{WLS}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}$$

$$\varepsilon \sim N(\mathbf{0}, \mathbf{W}^{-1})$$

$$p(\mathbf{y}, \beta) = \frac{1}{(2\pi)^{\frac{N}{2}}\det^{\frac{1}{2}}(\mathbf{W}^{-1})}\exp\left[-\frac{1}{2}(\mathbf{y}-\mathbf{X}\beta)^T\mathbf{W}(\mathbf{y}-\mathbf{X}\beta)\right]$$

$$\begin{aligned}&(\mathbf{y}-\mathbf{X}\beta)^T\mathbf{W}(\mathbf{y}-\mathbf{X}\beta)\\ &= \mathbf{y}^T\mathbf{W}\mathbf{y} - \beta^T\mathbf{X}^T\mathbf{W}\mathbf{y} - \mathbf{y}^T\mathbf{W}\mathbf{X}\beta + \beta^T\mathbf{X}^T\mathbf{W}\mathbf{X}\beta\end{aligned}$$

$$\begin{aligned}&\frac{\partial(\mathbf{y}-\mathbf{X}\beta)^T\mathbf{W}(\mathbf{y}-\mathbf{X}\beta)}{\partial\beta}\\ &= -\mathbf{X}^T\mathbf{W}\mathbf{y} - \mathbf{X}^T\mathbf{W}^T\mathbf{y} + \mathbf{X}^T\mathbf{W}\mathbf{X}\beta + \mathbf{X}^T\mathbf{W}^T\mathbf{X}\beta\end{aligned}$$

Other considerations in the regression model

- Potential problems of fitting a linear regression model

  Non-constant variance of error terms

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

  - Weighted least squares (non-constant variance)

$$\mathbf{W} = \begin{pmatrix} 1/\sigma_1^2 & 0 & \dots & 0 \\ 0 & 1/\sigma_2^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1/\sigma_n^2 \end{pmatrix}$$

$$\hat{\beta}_{\text{WLS}} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}$$

$$\mathbf{W} = \mathbf{W}^T$$

$$\frac{\partial(\mathbf{y}-\mathbf{X}\beta)^T\mathbf{W}(\mathbf{y}-\mathbf{X}\beta)}{\partial\beta} = -2\mathbf{X}^T\mathbf{W}\mathbf{y} + 2\mathbf{X}^T\mathbf{W}\mathbf{X}\beta$$

$$\frac{\partial(\mathbf{y}-\mathbf{X}\beta)^T\mathbf{W}(\mathbf{y}-\mathbf{X}\beta)}{\partial\beta} = 0$$

$$\hat{\beta}_{MLE} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}$$

Other considerations in the regression model

- Potential problems of fitting a linear regression model

  Non-constant variance of error terms

  One situation that we can easily identify the weight matrix:

  If the i-th response is an average of n_i equally variable observations, then

$$\text{var}(y_i) = \sigma^2 / n_i \quad \text{and} \quad w_i = n_i$$

$$\mathbf{W} = \begin{pmatrix} n_1 & 0 & \ldots & 0 \\ 0 & n_2 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & n_n \end{pmatrix}$$

Other considerations in the regression model

- Potential problems of fitting a linear regression model

  Outliers

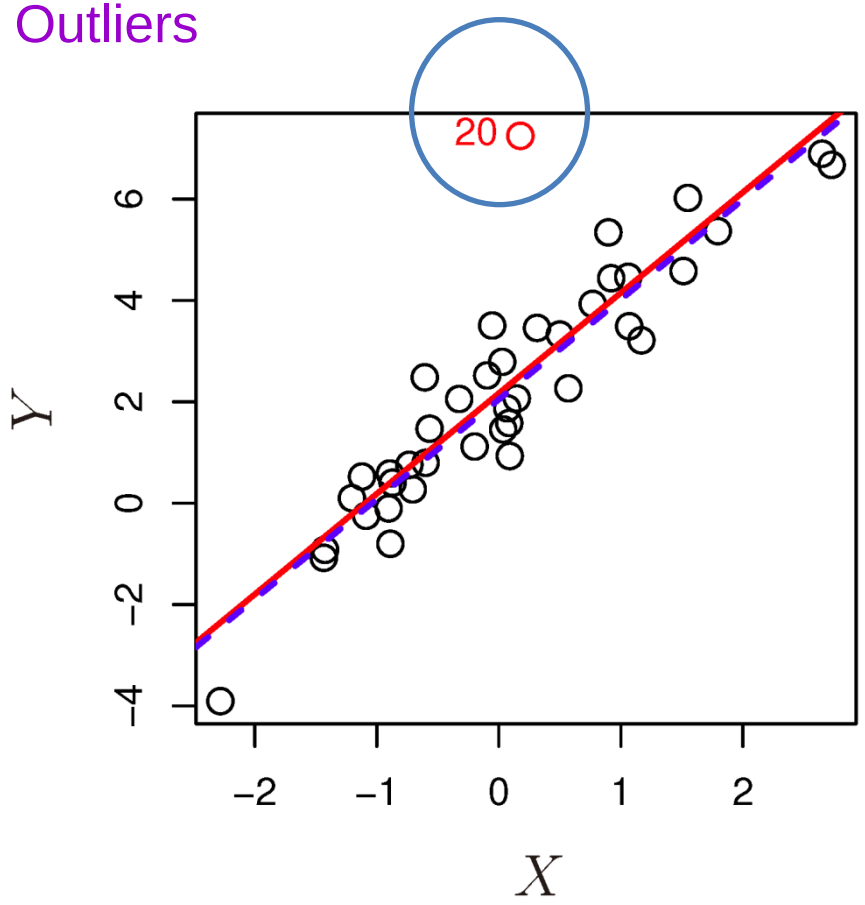
  An outlier is a point for which y_i is far from the value predicted by the model.

  Note: outliers can arise for a variety of reasons, such as incorrect recording of an observation during data collection.

Other considerations in the regression model

- Potential problems of fitting a linear regression model

Outliers



--red solid line: the least squares regression fit of all data

--blue dash line: the least squares regression fit after removal of the outlier

No difference or very mild difference?

Other considerations in the regression model

- Potential problems of fitting a linear regression model

  Outliers

  Even if an outlier does not have much effect on the least squares fit, it can cause other problems.

  Without the outlier

  $RSE = 0.77,\ R^2 = 0.892$

  With the outlier

  $RSE = 1.09,\ R^2 = 0.805$

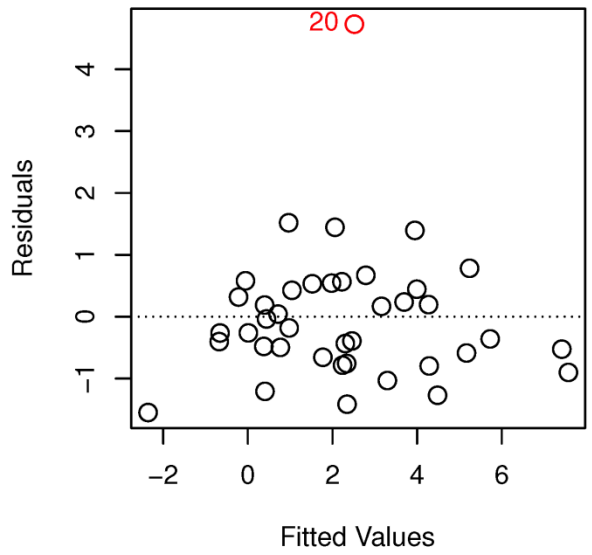
Other considerations in the regression model

- Potential problems of fitting a linear regression model

  Outliers

  How to identify?

  Using residual plots



Potential issue: in practice, it can be difficult to decide how large a residual needs to be before we consider the point to be an outlier

Other considerations in the regression model

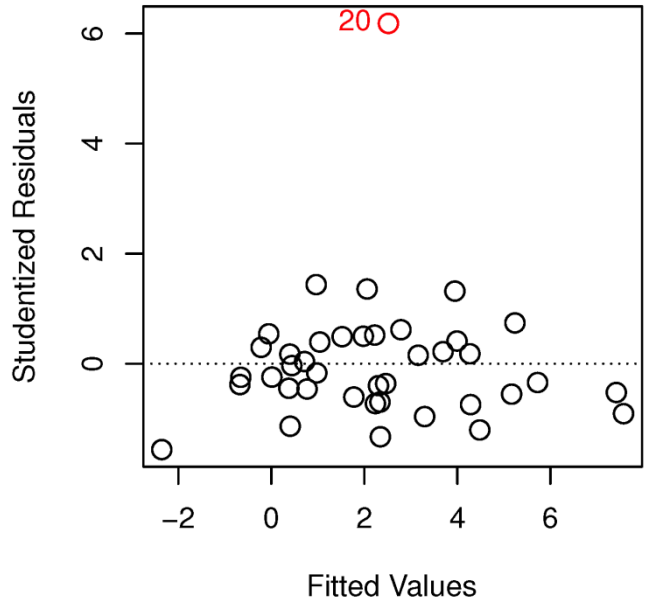
- Potential problems of fitting a linear regression model

  Outliers

  How to identify?

  Using residual plots (plot the **studentized residuals**)

the division of a residual by an estimate of its standard deviation



Observations whose studentized residuals are greater than 3 in absolute value are possible outliers.

https://newonlinecourses.science.psu.edu/stat462/node/247/

Other considerations in the regression model

- Potential problems of fitting a linear regression model

  Outliers

  Suggestion: if we believe that an outlier has occurred due to an error in data collection or recording, then one solution is to simply remove the observation. However, care should be taken, since an outlier may instead indicate a deficiency with the model, such as a missing predictor.