Statistical Learning for Data Science

Lecture 10

唐晓颖
电子与电气工程系
南方科技大学

March 27, 2023

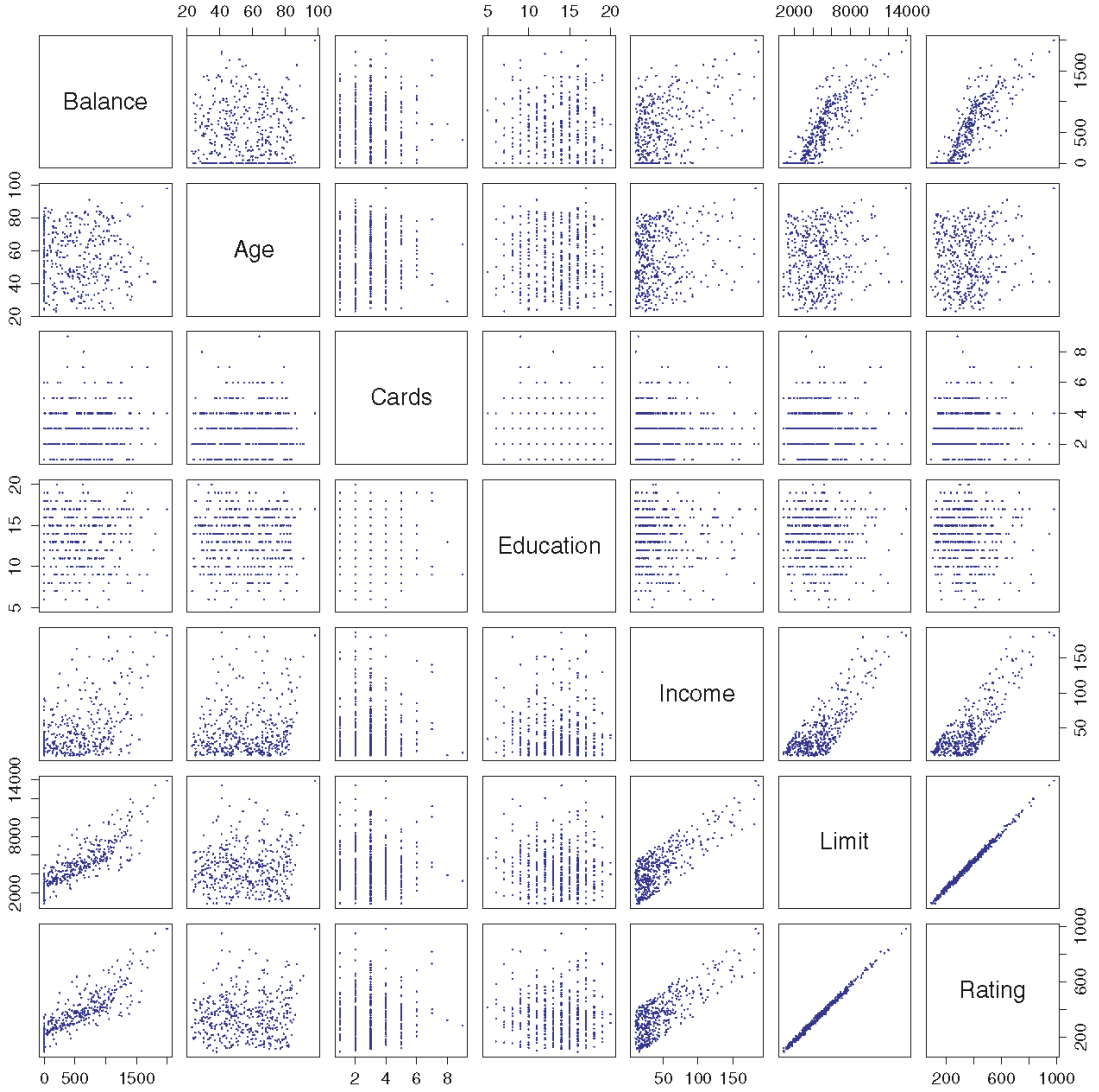
Other considerations in the regression model

- Qualitative predictors
  - Some predictors are not quantitative but are qualitative, taking a discrete set of values.
  - These are also called *categorical predictors* or *factor variables*.
  - Example: geographical position of a house in a house price prediction model
  - In regression problems, qualitative predictors usually need to be converted into numerical variables to be used by the model.

Other considerations in the regression model

- Qualitative predictors

In addition to the 7 quantitative variables, we also have four qualitative variables: gender, student (student status), marital status (married or single), and ethnicity (Caucasian, African American, or Asian).



Other considerations in the regression model

- Qualitative predictors

  Predictors with only two levels

  Investigate differences in credit card balance between males and females, ignoring the other variables

  - Create an indicator or dummy variable that takes on two possible numerical values.

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male} \end{cases}$$

Other considerations in the regression model

- Qualitative predictors

  Predictors with only two levels

  Investigate differences in credit card balance between males and females, ignoring the other variables

  - Use this dummy variable as a predictor in the regression equation

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \varepsilon_i & \text{if } i\text{th person is male} \end{cases}$$

Other considerations in the regression model

- Qualitative predictors

  Predictors with only two levels

  Investigate differences in credit card balance between males and females, ignoring the other variables

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \varepsilon_i & \text{if } i\text{th person is male} \end{cases}$$

  - β_0 -- the average credit card balance among males
  - $\beta_0 + \beta_1$ -- the average credit card balance among females
  - β_1 -- the average difference in credit card balance between females and males

Other considerations in the regression model

- Qualitative predictors

Predictors with only two levels

	Coefficient	Std. error	t-statistic	p-value
Intercept	509.80	33.13	15.389	<0.0001
Gender [Female]	19.73	46.05	0.429	0.6690

The p-value for the dummy variable is very high. This indicates that there is no statistical evidence of a difference in average credit card balance between the genders.

Other considerations in the regression model

- Qualitative predictors

  Predictors with only two levels

  - The decision to code females as 1 and males as 0 is arbitrary, and has no effect on the regression fit.

  - Instead of a 0/1 coding scheme, we could create a dummy variable

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ -1 & \text{if } i\text{th person is male} \end{cases}$$

Other considerations in the regression model

- Qualitative predictors

  Predictors with only two levels

  - Instead of a 0/1 coding scheme, we could create a dummy variable

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{if } i\text{th person is female} \\ \beta_0 - \beta_1 + \varepsilon_i & \text{if } i\text{th person is male} \end{cases}$$

    - β_0 -- the overall average credit card balance (ignoring the gender effects)

    - β_1 -- the amount that females are above the average and males are below the average

Other considerations in the regression model

- Qualitative predictors

  Predictors with only two levels

  Note: the final predictions for the credit balances of males and females will be identical regardless of the coding scheme used. The only difference is in the way that the coefficients are interpreted.

Other considerations in the regression model

- Qualitative predictors

  Predictors with more than two levels

  Investigate differences in credit card balance among different ethnicity groups.

  - Create additional dummy variables

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is Asian} \\ 0 & \text{if } i\text{th person is not Asian} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is Caucasian} \\ 0 & \text{if } i\text{th person is not Caucasian} \end{cases}$$

Other considerations in the regression model

- Qualitative predictors

  Predictors with more than two levels

  Investigate differences in credit card balance among different ethnicity groups.

  - Use these variables in the regression equation

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \varepsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \varepsilon_i & \text{if } i\text{th person is African American} \end{cases}$$

β_0 -- the average credit card balance for African Americans

β_1 -- the difference in the average balance between Asian and African American categories

β_2 -- the difference in the average balance between Caucasian and African American categories

Other considerations in the regression model

- Qualitative predictors

  Predictors with more than two levels

  Note: there will always be one fewer dummy variables than the number of levels.

Other considerations in the regression model

- Qualitative predictors

  Predictors with more than two levels

  Investigate differences in credit card balance among different ethnicity groups.

	Coefficient	Std. error	t-statistic	p-value
Intercept	531.00	46.32	11.464	<0.0001
Ethnicity [Asian]	-18.96	65.02	-0.287	0.7740
Ethnicity [Caucasian]	-12.50	56.68	-0.221	0.826

Other considerations in the regression model

- Qualitative predictors

  Predictors with more than two levels

	Coefficient	Std. error	t-statistic	p-value
Intercept	531.00	46.32	11.464	<0.0001
Ethnicity [Asian]	-18.96	65.02	-0.287	0.7740
Ethnicity [Caucasian]	-12.50	56.68	-0.221	0.826

Investigate differences in credit card balance among different ethnicity groups.

- The estimated balance for African American is $531.00.
- The Asian category will have $18.69 less debt than the African American category.
- The Caucasian category will have $12.50 less debt than the African American category.
- The p-values associated with the coefficient estimates for the two dummy variables are very large, suggesting no statistical evidence of a real difference in credit card balance between the ethnicities.

Other considerations in the regression model

- Qualitative predictors

  Predictors with more than two levels

  Investigate differences in credit card balance among different ethnicity groups.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \varepsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \varepsilon_i & \text{if } i\text{th person is African American} \end{cases}$$

  Alternatively, we can use an F-test to perform hypothesis testing $H_0 : \beta_1 = \beta_2 = 0$

  - This hypothesis testing does not depend on the coding
  - This F-test has a p-value of 0.96, indicating that we cannot reject the null hypothesis that there is no relationship between balance and ethnicity

Other considerations in the regression model

- Qualitative predictors

Note:

- This dummy variable approach also works when there are both quantitative and qualitative predictors.

- There are many different ways of coding qualitative variables besides the dummy variable approach taken here. They all lead to equivalent model fits, but the coefficients are different and have different interpretations.

Other considerations in the regression model

- Extensions of the linear model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

  - Why do we need extension?
    - In linear regression, we assume that the relationship between the predictor variables and the response variable is linear.
    - However, in many real-life situations, a linear model may not adequately capture the underlying relationship between the variables.
  - The extension is to improve the predictive performance of the linear model.

Other considerations in the regression model

- Extensions of the linear model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

  - Extensions of linear model
    - Examples: Polynomial regression, generalized linear models, kernel regression, …
    - Help better fit the data to improve the predictive ability of the model
    - Consider about overfitting and underfitting
  - Removing additive assumptions of linear model
    - Allowing for non-additive effects of predictor variables on the response variable, such as interactions between predictors or nonlinear relationships.
    - Does not necessarily mean that the effect of each predictor variable is dependent on the values of the other predictor variables.

Other considerations in the regression model

- Extensions of the linear model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \varepsilon$$

Two highly restrictive assumptions:

- Additive: the effect of changes in the predictor on the response is independent of the values of the other predictors (there may be interactions).

- Linear: the change in the response Y due to a one-unit change in X_j is constant, regardless of the value of X_j (there may be nonlinearity).

Other considerations in the regression model

- Extensions of the linear model

Removing the additive assumption

- In our previous analysis of the advertising data, we assumed that the effect on sales of increasing one advertising medium is independent of the amount spent on the other media.

- For example, the linear model

$$\text{sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times \text{newspaper} + \varepsilon$$

states that the average effect on sales of a one-unit increase in TV is always β_1 regardless of the amount spent on radio.

Other considerations in the regression model

- Extensions of the linear model

Removing the additive assumption

- But suppose that spending money on radio advertising actually increase the effectiveness of TV advertising, so that the slope term for TV should increase as radio increases.

- In this situation, given a fixed budget of $100,000, spending half on radio and half on TV may increase sales more than allocating the entire amount to either TV or to radio.

- In marketing, this is known as a ***synergy*** effect, and in statistics it is referred to as an ***interaction*** effect.

Other considerations in the regression model

- Extensions of the linear model

  Removing the additive assumption

  The standard linear regression model with two variables:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

  Note: if we increase X_1 by one unit, then Y will increase by an average of β_1 units. The presence of X_2 does not alter this statement.

Other considerations in the regression model

- Extensions of the linear model

  Removing the additive assumption

  Extending the model to allow for interaction effects:

  The standard linear regression model with two variables:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \underbrace{\beta_3 X_1 X_2} + \varepsilon$$

Interaction term

Other considerations in the regression model

- Extensions of the linear model

  Removing the additive assumption

  Extending the model to allow for interaction effects:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$
$$= \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \varepsilon$$

  The effect of X_1 on Y is no longer constant: adjusting X_2 will change the impact of X_1 on Y .

Other considerations in the regression model

- Extensions of the linear model

  Removing the additive assumption

  Example: To predict the number of units produced on the basis of the number of production lines and the total number of workers.

  - The effect of increasing the number of production lines will depend on the number of workers, since if no workers are available to operate the lines, then increasing the number of lines will not increase production.

  - It would be appropriate to include an interaction term between lines and workers in a linear model to predict units.

Other considerations in the regression model

- Extensions of the linear model

Removing the additive assumption

Example: To predict the number of units produced on the basis of the number of production lines and the total number of workers.

Suppose that when we fit the model, we obtain

$$\begin{aligned}\text{units} &\approx 1.2 + 3.4 \times \text{lines} + 0.22 \times \text{workers} + 1.4 \times (\text{lines} \times \text{workers}) \\ &= 1.2 + (3.4 + 1.4 \times \text{workers}) \times \text{lines} + 0.22 \times \text{workers}\end{aligned}$$

Conclusion: adding an additional line will increase the number of units produced by $3.4 + 1.4 \times \text{workers}$. Hence, the more workers we have, the stronger will be the effect of lines.

Other considerations in the regression model

- Extensions of the linear model

Removing the additive assumption

Example: Advertising data (sales versus TV & radio)

$$\begin{aligned}\text{sales} &\approx \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times (\text{TV} \times \text{radio}) + \varepsilon \\ &= \beta_0 + (\beta_1 + \beta_3 \times \text{radio}) \times \text{TV} + \beta_2 \times \text{radio} + \varepsilon\end{aligned}$$

β_3 : the increase in the effectiveness of TV advertising for a one unit increase in radio advertising (or vice-versa).

Other considerations in the regression model

- Extensions of the linear model

Removing the additive assumption

Example: Advertising data (sales versus TV & radio)

	Coefficient	Std. Error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	< 0.0001

Conclusion: the p-value for the interaction term, $\text{TV} \times \text{radio}$, is extremely low, indicating that there is strong evidence for $H_a : \beta_3 \neq 0$. In other words, it is clear that the true relationship is not additive.

Other considerations in the regression model

- Extensions of the linear model

Removing the additive assumption

Example: Advertising data (sales versus TV & radio)

R^2 *Statistic (the fraction of variance explained)*

With interaction term (sales versus TV and radio)

$R^2 = 0.968$

Without interaction term (sales versus TV and radio)

$R^2 = 0.897$

Other considerations in the regression model

- Extensions of the linear model

Removing the additive assumption

Example: Advertising data (sales versus TV & radio)

R^2 *Statistic (0.968 versus 0.897)*

Conclusion:

- The model that includes the interaction term is superior to the model that contains only main effects.
- (96.8-89.7)/(100-89.7) = 69% of the variability in sales that remains after fitting the additive model has been explained by the interaction term.

Other considerations in the regression model

- Extensions of the linear model

Removing the additive assumption

Example: Advertising data (sales versus TV & radio)

$$\text{sales} = 6.7502 + 0.0191 \times \text{TV} + 0.0289 \times \text{radio} + 0.0011 \times (\text{TV} \times \text{radio}) + \varepsilon$$

Conclusion:

- An increase in TV advertising of \$1,000 is associated with increased sales of $(\beta_1 + \beta_3 \times \text{radio}) \times 1000 = 19 + 1.1 \times \text{radio}$ units.
- An increase in radio advertising of \$1,000 is associated with increased sales of $(\beta_2 + \beta_3 \times \text{TV}) \times 1000 = 29 + 1.1 \times \text{TV}$ units.

Other considerations in the regression model

- Extensions of the linear model

Removing the additive assumption

Example: Advertising data (sales versus TV & radio)

	Coefficient	Std. Error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	< 0.0001

Conclusion: both the main effects (TV, radio) and the interaction effect (TV×radio) are statistically significant, so it is obvious that all three variables should be included in the model.

Other considerations in the regression model

- Extensions of the linear model

Removing the additive assumption

Question: What if an interaction term has a very small p-value, but the associated main effects (in this case, TV and radio) do not?

Answer: The hierarchical principle states that if we include an interaction in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant

Other considerations in the regression model

- Extensions of the linear model

Removing the additive assumption

If we include an interaction in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant. **(Why?)**

Rationale: If $X_1 \times X_2$ is related to the response, then whether or not the coefficients of X_1 or X_2 are exactly zeros is of little interest. Also $X_1 \times X_2$ is typically correlated with X_1 and X_2, and so leaving them out tends to alter the meaning of the interaction.

Other considerations in the regression model

- Extensions of the linear model

  Removing the additive assumption

  Interactions of qualitative variables, or a combination of quantitative and qualitative variables.

  Example: Credit data (balance versus income (quantitative) & student (qualitative))

  No interaction term:

$$\text{balance}_i \approx \beta_0 + \beta_1 \times \text{income}_i + \begin{cases} \beta_2, & \text{if } i\text{th person is a student} \\ 0, & \text{if } i\text{th person is not a student} \end{cases}$$

$$= \beta_1 \times \text{income}_i + \begin{cases} \beta_0 + \beta_2, & \text{if } i\text{th person is a student} \\ \beta_0, & \text{if } i\text{th person is not a student} \end{cases}$$

Other considerations in the regression model

- Extensions of the linear model

Removing the additive assumption

Example: Credit data (balance versus income (quantitative) & student (qualitative))

No interaction term:

$$\text{balance}_i = \beta_1 \times \text{income}_i + \begin{cases} \beta_0 + \beta_2, & \text{if } i\text{th person is a student} \\ \beta_0, & \text{if } i\text{th person is not a student} \end{cases}$$

- Fitting two lines, one for students and one for non-students
- The two lines have different intercepts, $\beta_0 + \beta_2$ versus β_0 , but the same slope, β_1

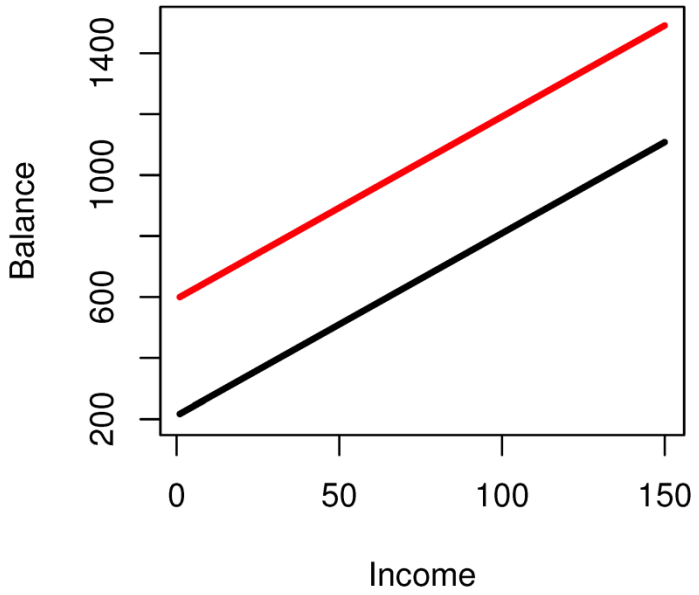
Other considerations in the regression model

- Extensions of the linear model

Removing the additive assumption

Example: Credit data (balance versus income (quantitative) & student (qualitative))

No interaction term:



Conclusion: the[illegible] effect on balanc[illegible]n income[illegible]ether or not the individual is a student.

NOT REASONABLE!

Other considerations in the regression model

- Extensions of the linear model

Removing the additive assumption

Interactions of qualitative variables, or a combination of quantitative and qualitative variables.

Example: Credit data (balance versus income (quantitative) & student (qualitative))

With interaction term:

- A change in income may have a very different effect on the credit card balance of a student versus a non-student.
- Adding an interaction variable, created by multiplying income with the dummy variable for student.

Other considerations in the regression model

- Extensions of the linear model

Removing the additive assumption

Example: Credit data (balance versus income (quantitative) & student (qualitative))

With interaction term:

$$\text{balance}_i \approx \beta_0 + \beta_1 \times \text{income}_i + \begin{cases} \beta_2 + \beta_3 \times \text{income}_i, & \text{if student} \\ 0, & \text{if not student} \end{cases}$$

$$= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times \text{income}_i, & \text{if student} \\ \beta_0 + \beta_1 \times \text{income}_i, & \text{if not student} \end{cases}$$

- Fitting two lines, one for students and one for non-students
- The two lines have different intercepts, $\beta_0 + \beta_2$ versus β_0, as well as different slopes $\beta_1 + \beta_3$ versus β_3

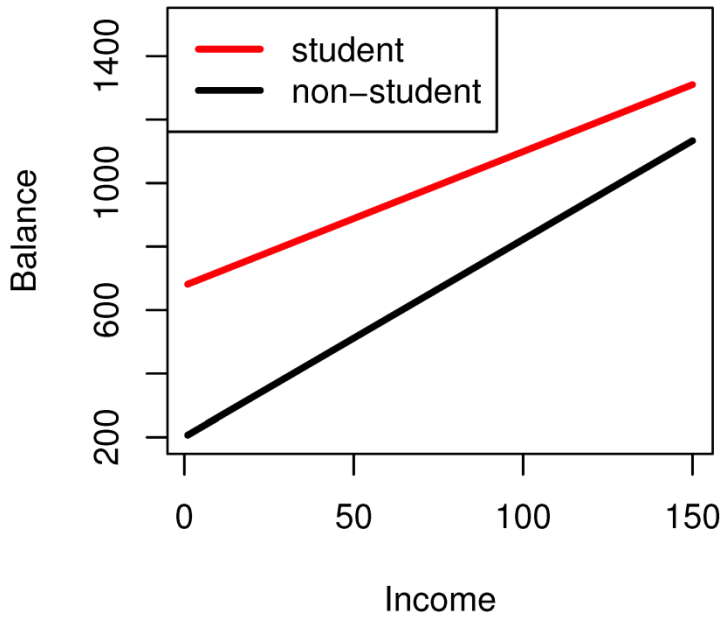
Other considerations in the regression model

- Extensions of the linear model

Removing the additive assumption

Example: Credit data (balance versus income (quantitative) & student (qualitative))

With interaction term:



Conclusion:

- Changes in income affect the credit card balances of students and non-students differently.
- The slope for students is lower than the slope for non-students, suggesting that increases in income are associated with smaller increase in credit card balance among students as compared to non-students

Other considerations in the regression model

- Extensions of the linear model

  Non-linear relationships

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

The above equation assumes a linear relationship between the response and predictors. What if the true relationship between the response and the predictors may be non-linear?

Other considerations in the regression model

- Extensions of the linear model

  Non-linear relationships

  A very simple way to directly extend the linear model to accommodate non-linear relationships, using ***polynomial regression***.

  Later, we will present more complex approaches for performing non-linear fits in more general settings.

Other considerations in the regression model

- Extensions of the linear model

Non-linear relationships

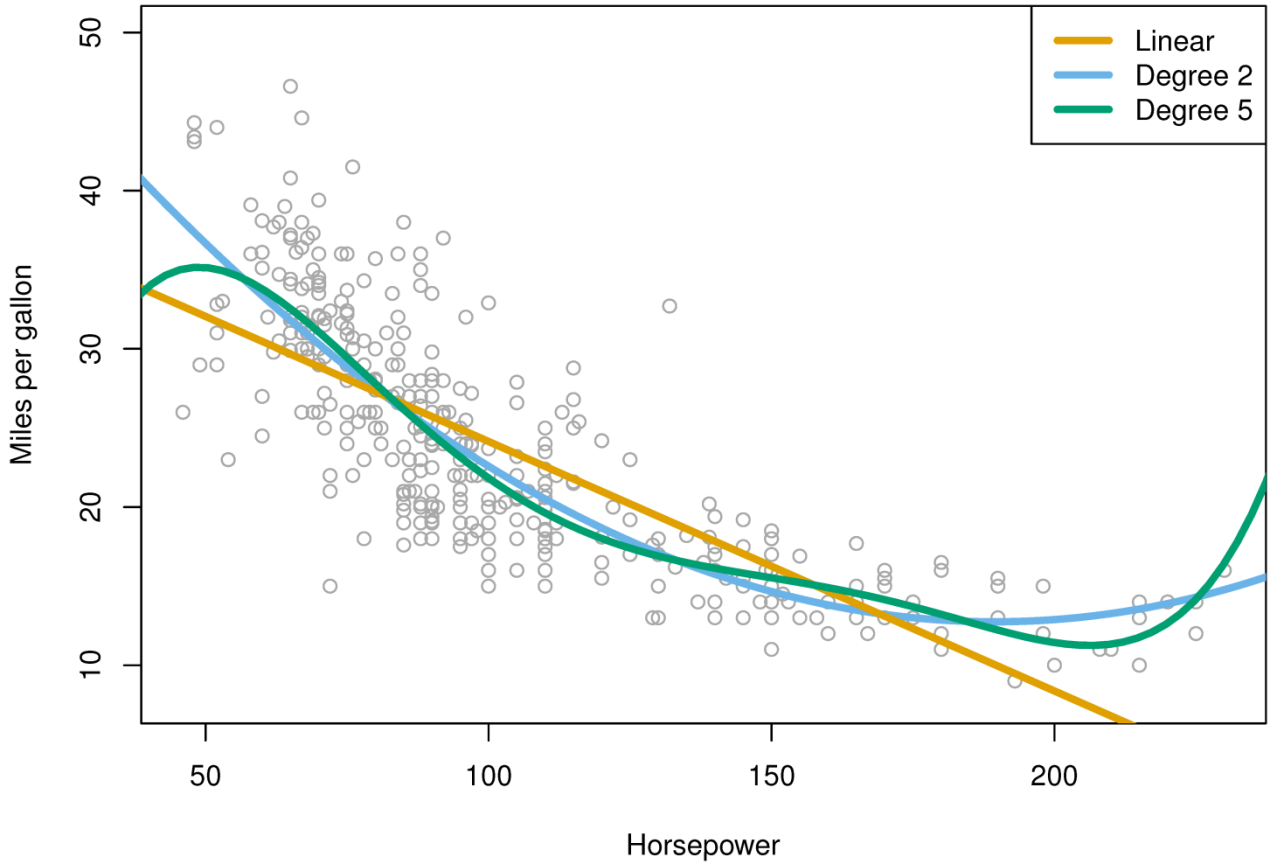
A very simple way to directly extend the linear model to accommodate non-linear relationships, using ***polynomial regression***.

Polynomials	Form	Degree	Examples
Linear Polynomial	$p(x): ax+b, a \neq 0$	Polynomial with Degree 1	$x + 8$
Quadratic Polynomial	$p(x): ax^2+b+c, a \neq 0$	Polynomial with Degree 2	$3x^2-4x+7$
Cubic Polynomial	$p(x): ax^3+bx^2+cx, a \neq 0$	Polynomial with Degree 3	$2x^3+3x^2+4x+6$

Other considerations in the regression model

- Extensions of the linear model

Polynomial regression



The data suggests a curved relationship between mpg and horsepower.

Other considerations in the regression model

- Extensions of the linear model

Polynomial regression (quadratic)

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

Questions:

1. Is the relationship between horsepower and mpg linear? **NO**
2. Is the model linear? **YES**

Other considerations in the regression model

- Extensions of the linear model

Polynomial regression (quadratic)

$$\text{mpg} = \beta_0 + \beta_1 \times \underbrace{\text{horsepower}}_{X_1} + \beta_2 \times \underbrace{\text{horsepower}^2}_{X_2} + \varepsilon$$

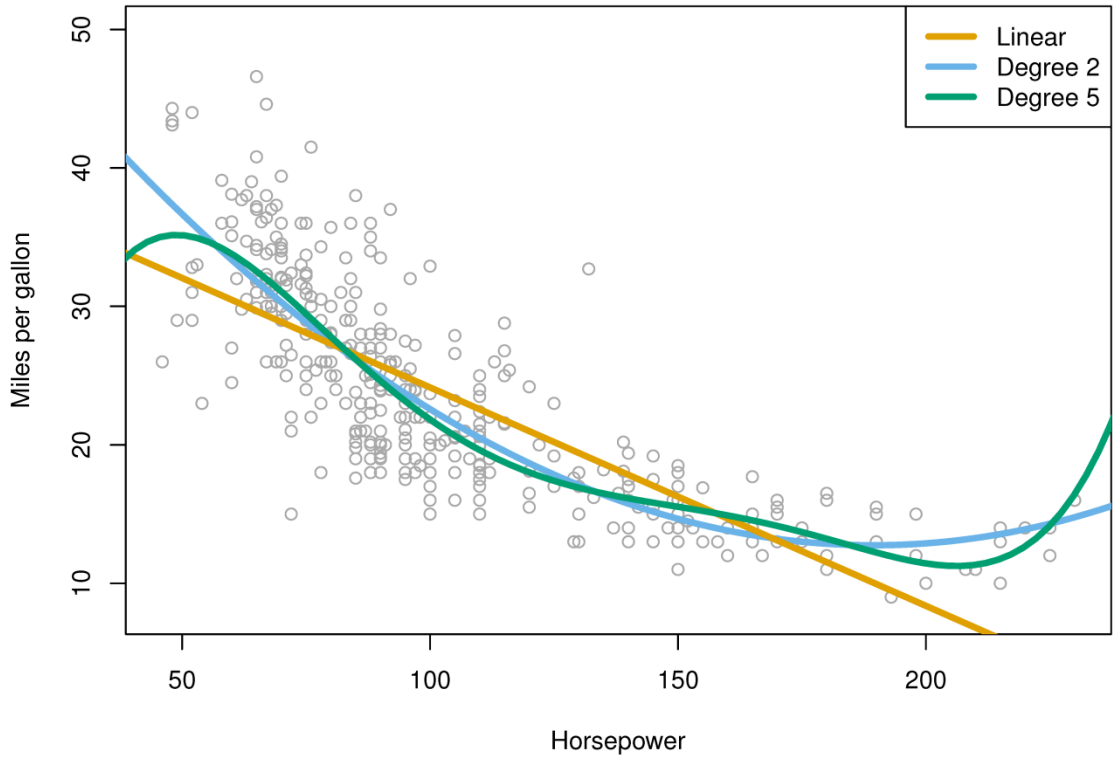
We can still use standard linear regression software to estimate the model coefficients in order to produce a non-linear fit.

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

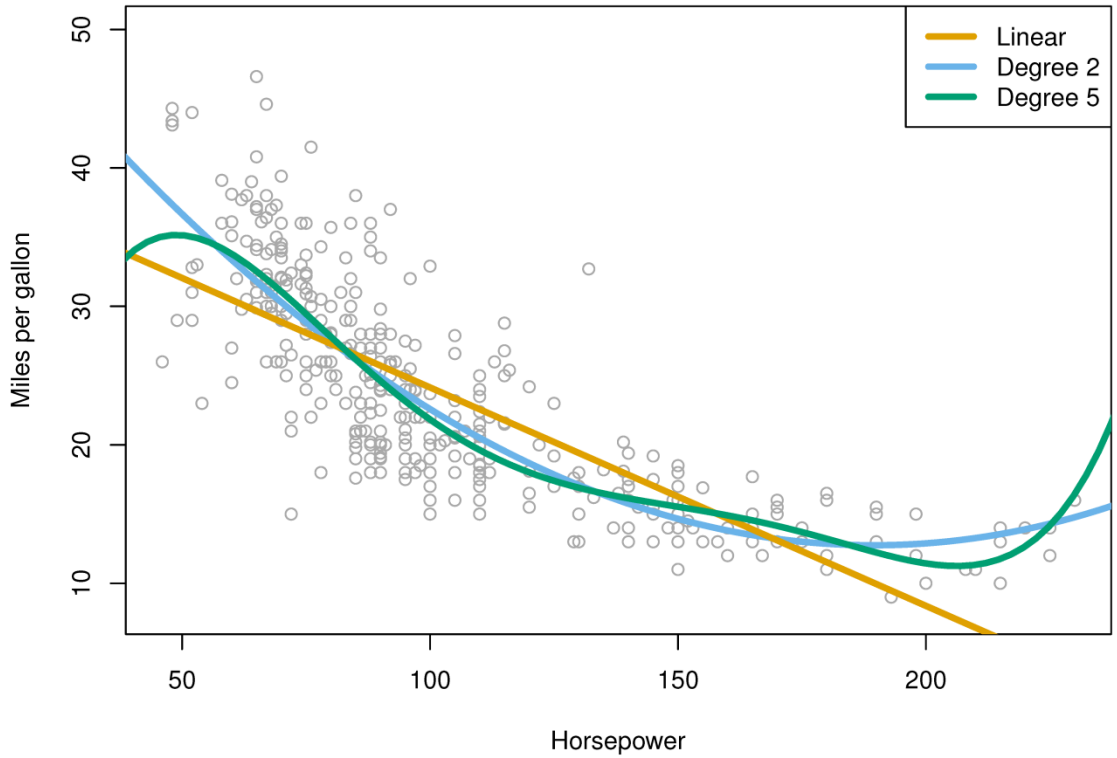
Conclusion: the p-value for the quadratic term is highly significant.

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

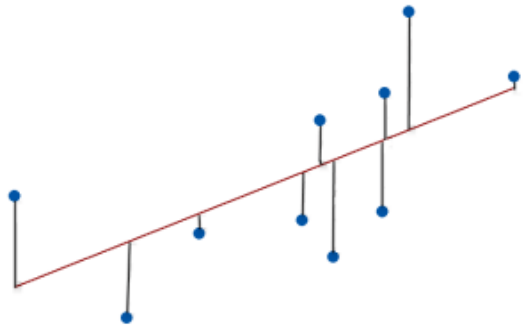
  - If the true relationship is far from linear, then virtually all of the conclusions that we draw from the fit are suspect.

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

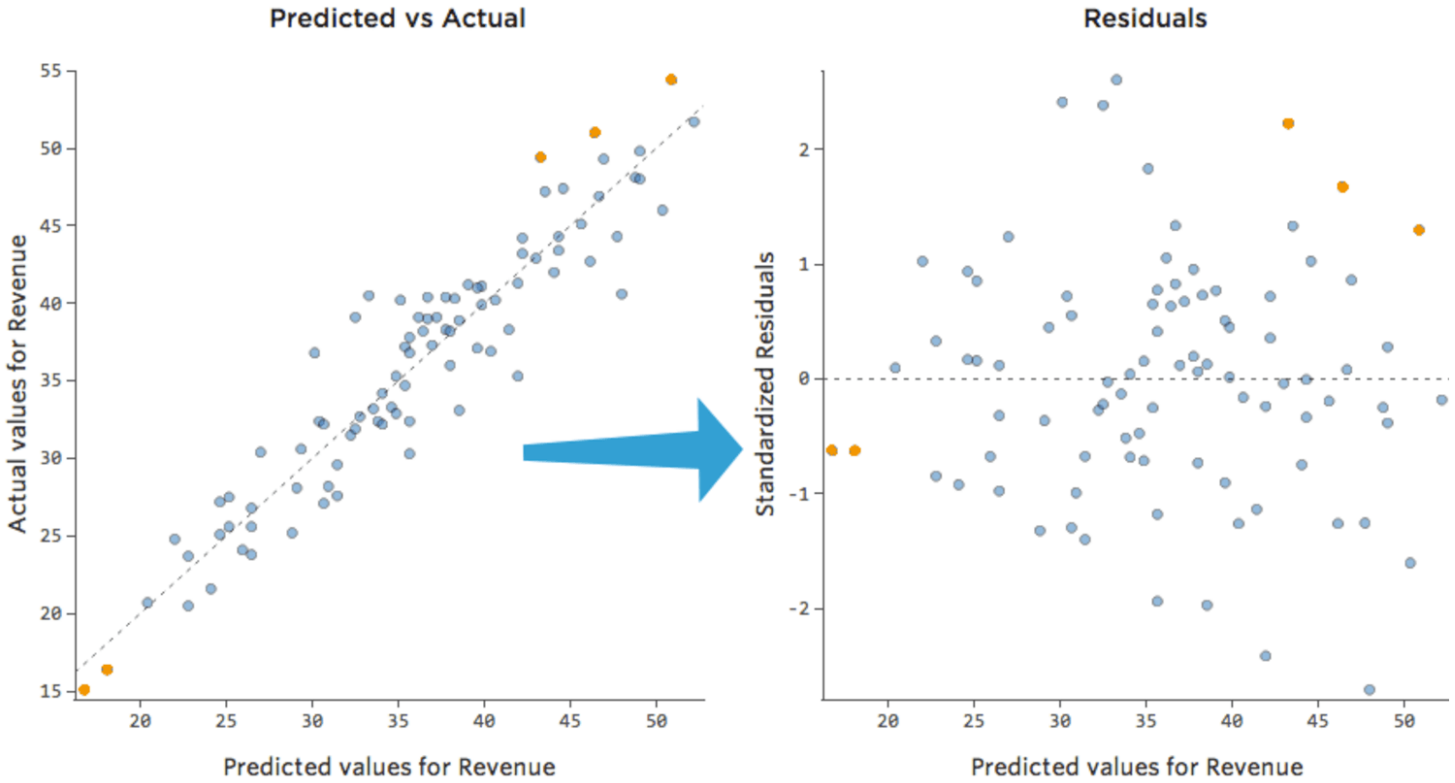
  - Residual plots (residuals vs. predictor OR residuals vs. predicted values) are a useful graphical tool for identifying non-linearity. **Ideally, the residual plot will show no discernible pattern.** The presence of a pattern may indicate a problem with some aspect of the linear model.

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.

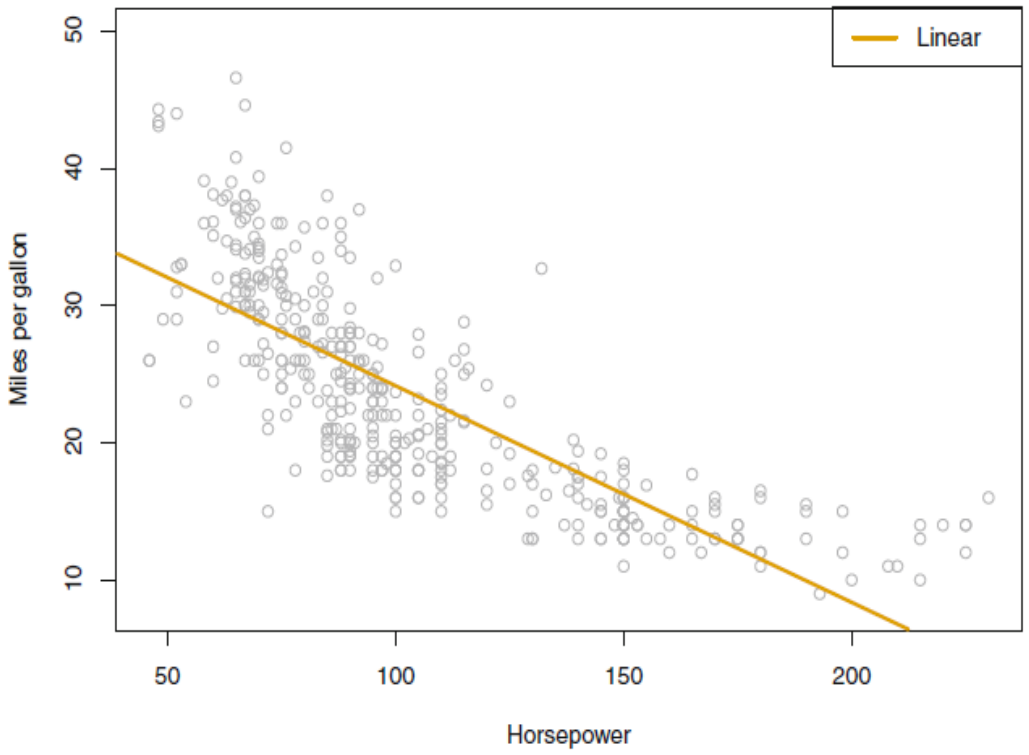
Other considerations in the regression model

- Potential problems of fitting a linear regression model

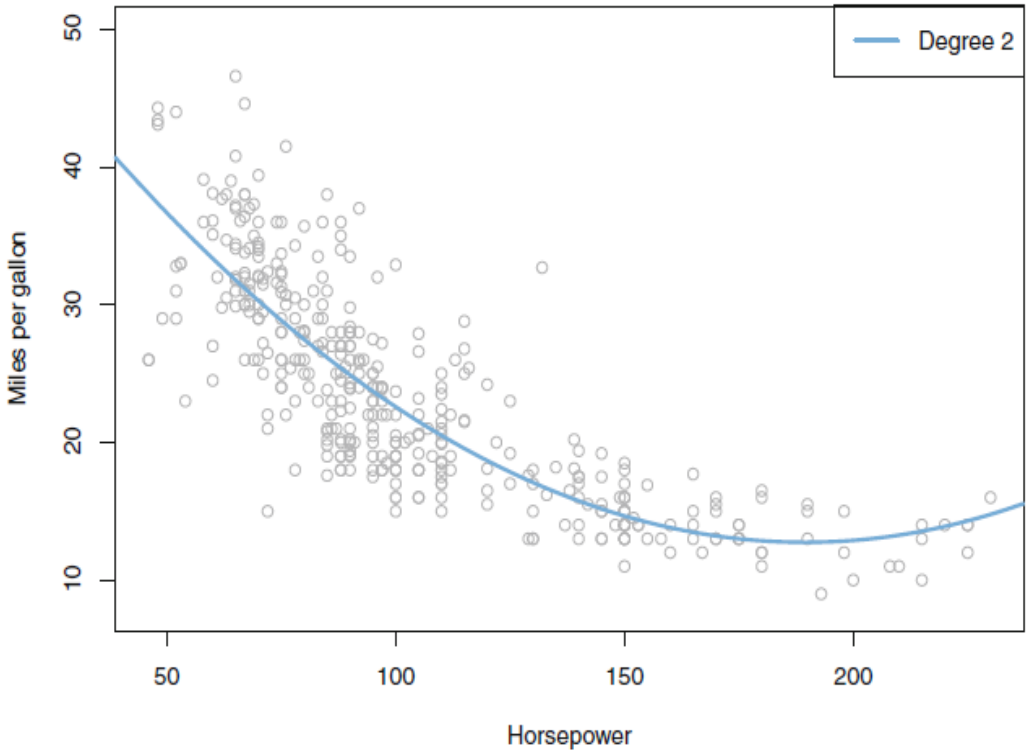
  Non-linearity of the data

  mpg vs. horsepower

Linear Fit



Quadratic Fit

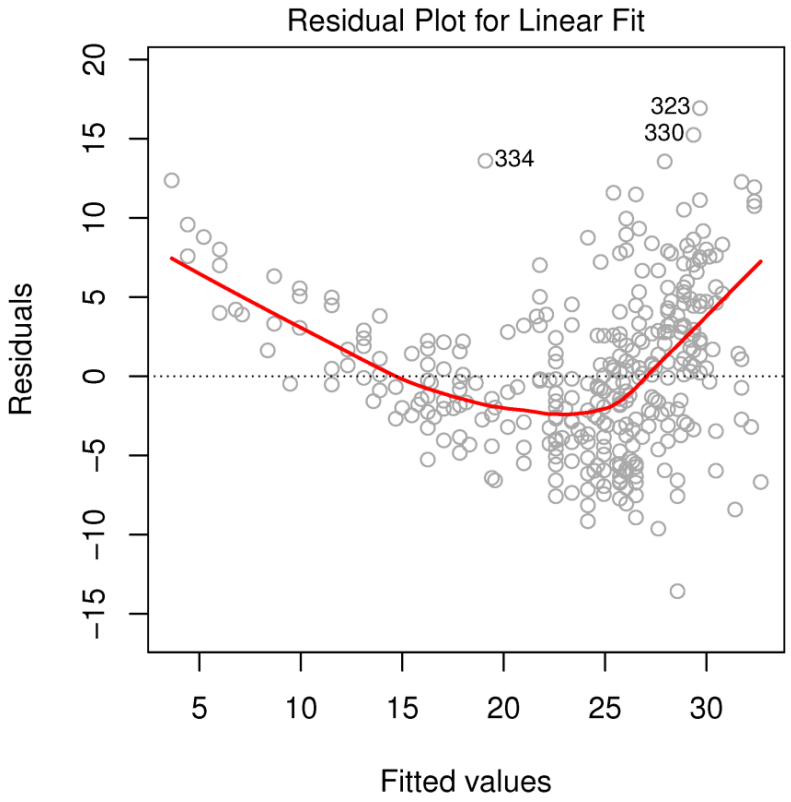


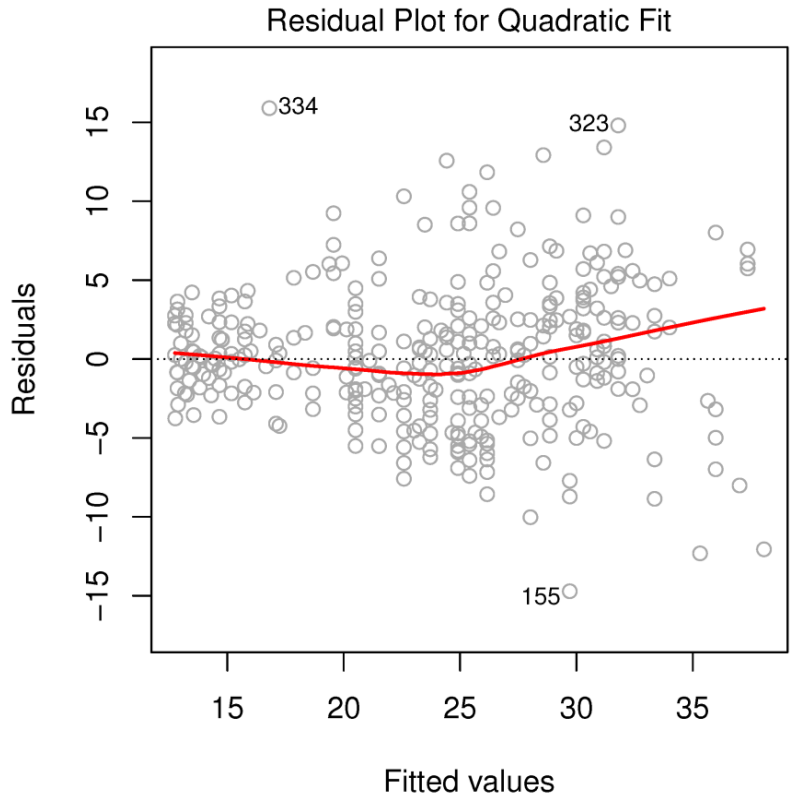
Other considerations in the regression model

- 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.

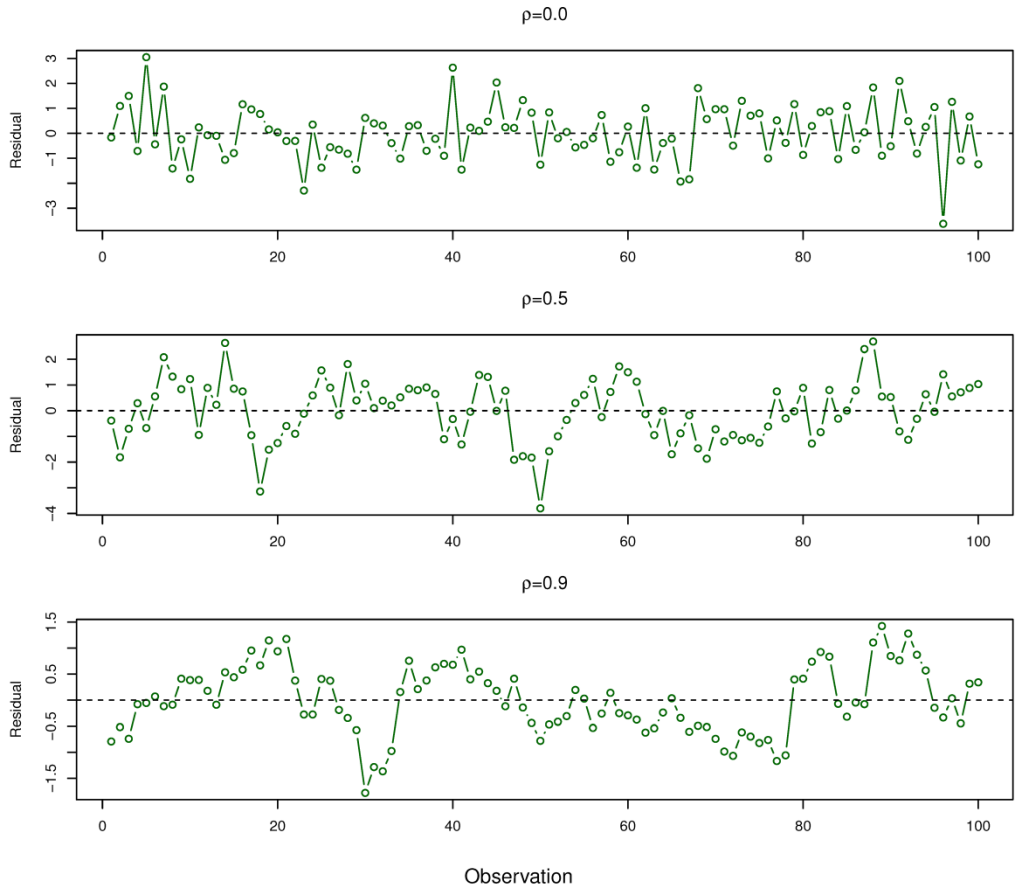
  - Observations that are obtained at adjacent time points will have positively correlated errors.

  - 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

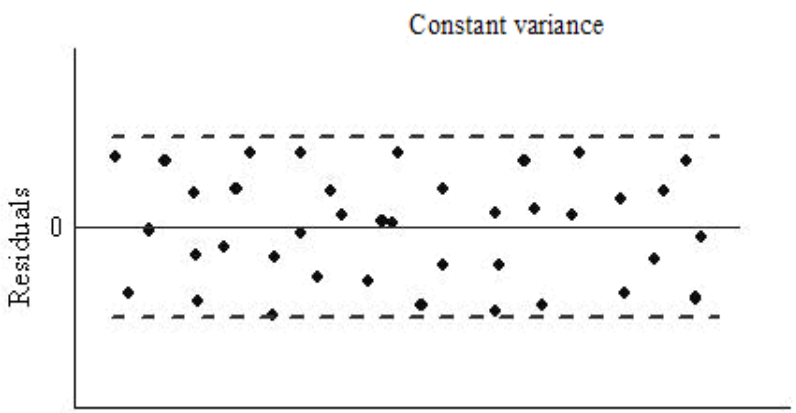
Another important assumption of the linear regression model is that the error terms have a constant variance, $\text{var}(\varepsilon_i) = \sigma^2$. The SEs, confidence intervals, and hypothesis testing associated with the linear model rely upon this assumption.

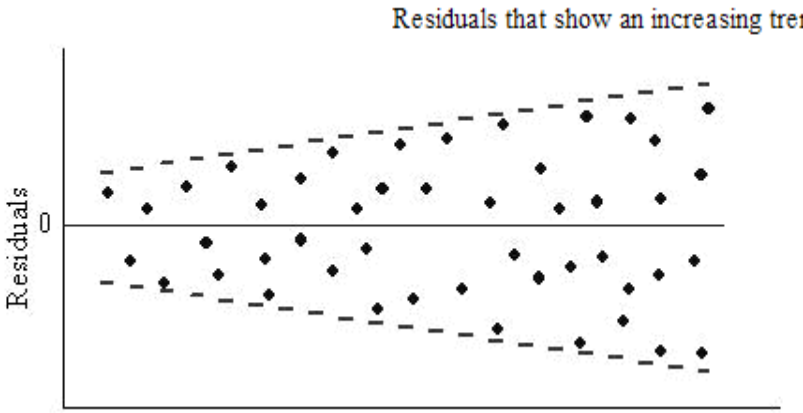
- It is often the case that the variances of the error terms are non-constant.
- One can identify non-constant variances in the errors, or *heteroscedasticity*, from the presence of a *funnel shape*（漏斗形状）in the residual plot.

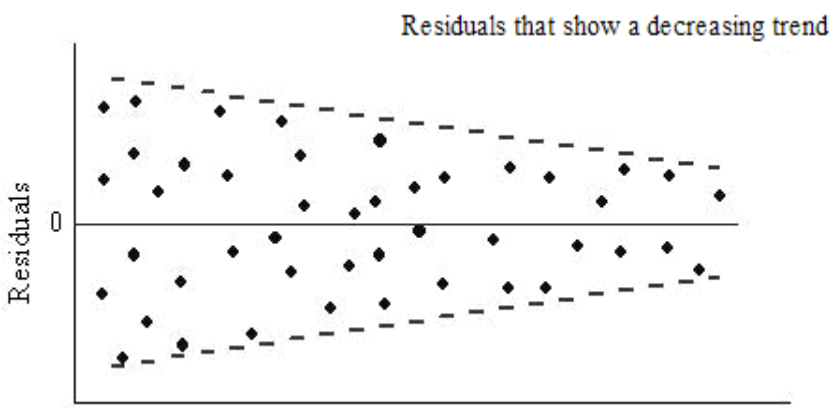
Other considerations in the regression model

- Potential problems of fitting a linear regression model

  Non-constant variance of error terms



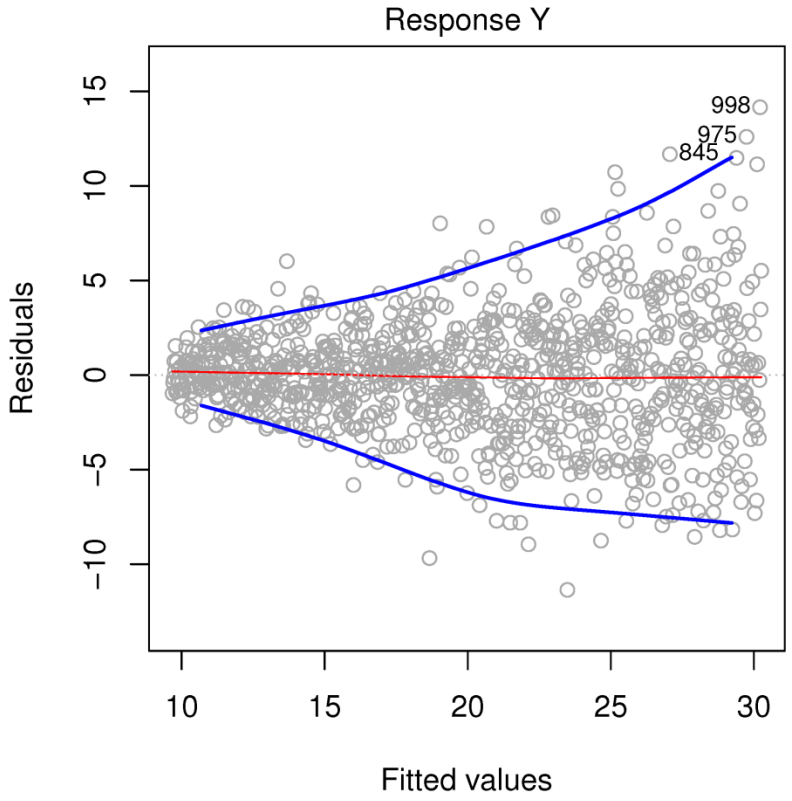




Other considerations in the regression model

- Potential problems of fitting a linear regression model

  Non-constant variance of error terms

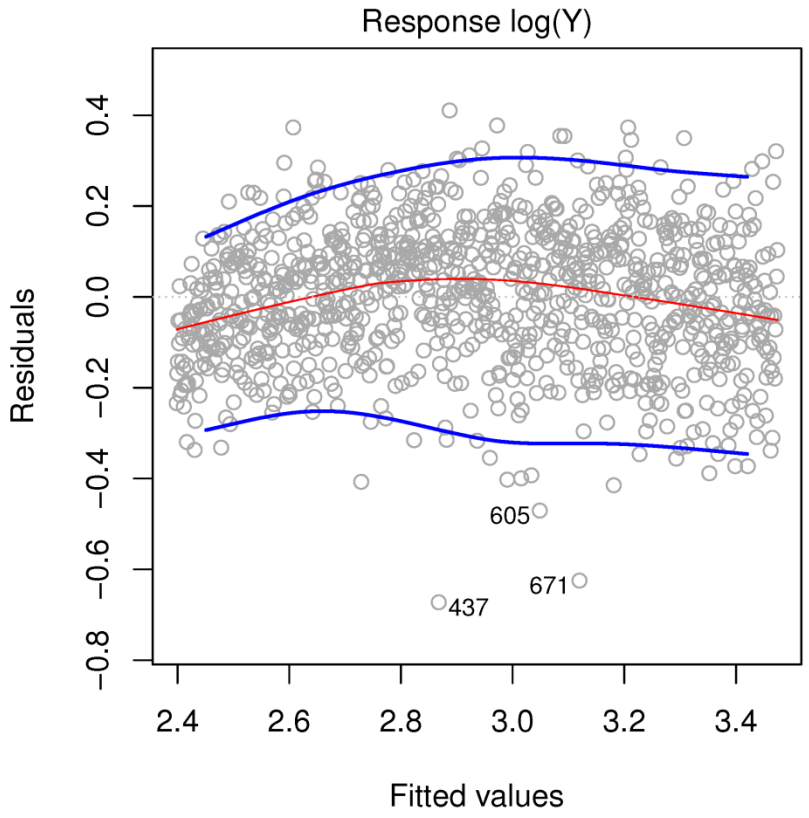


Other considerations in the regression model

- 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

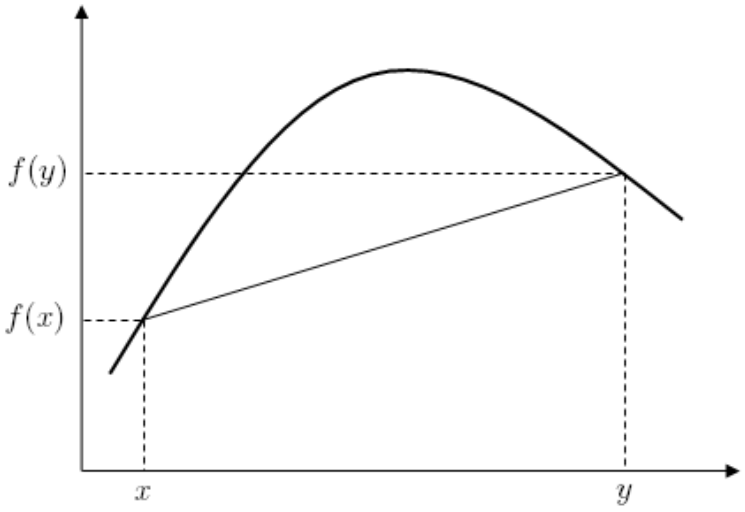
- Potential problems of fitting a linear regression model

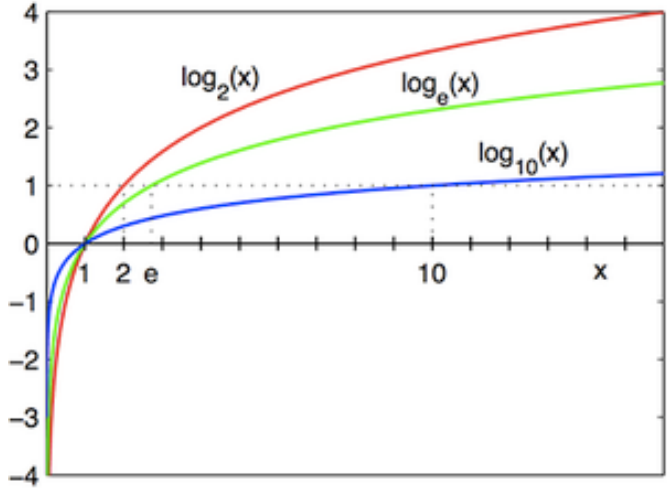
  Non-constant variance of error terms

  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)

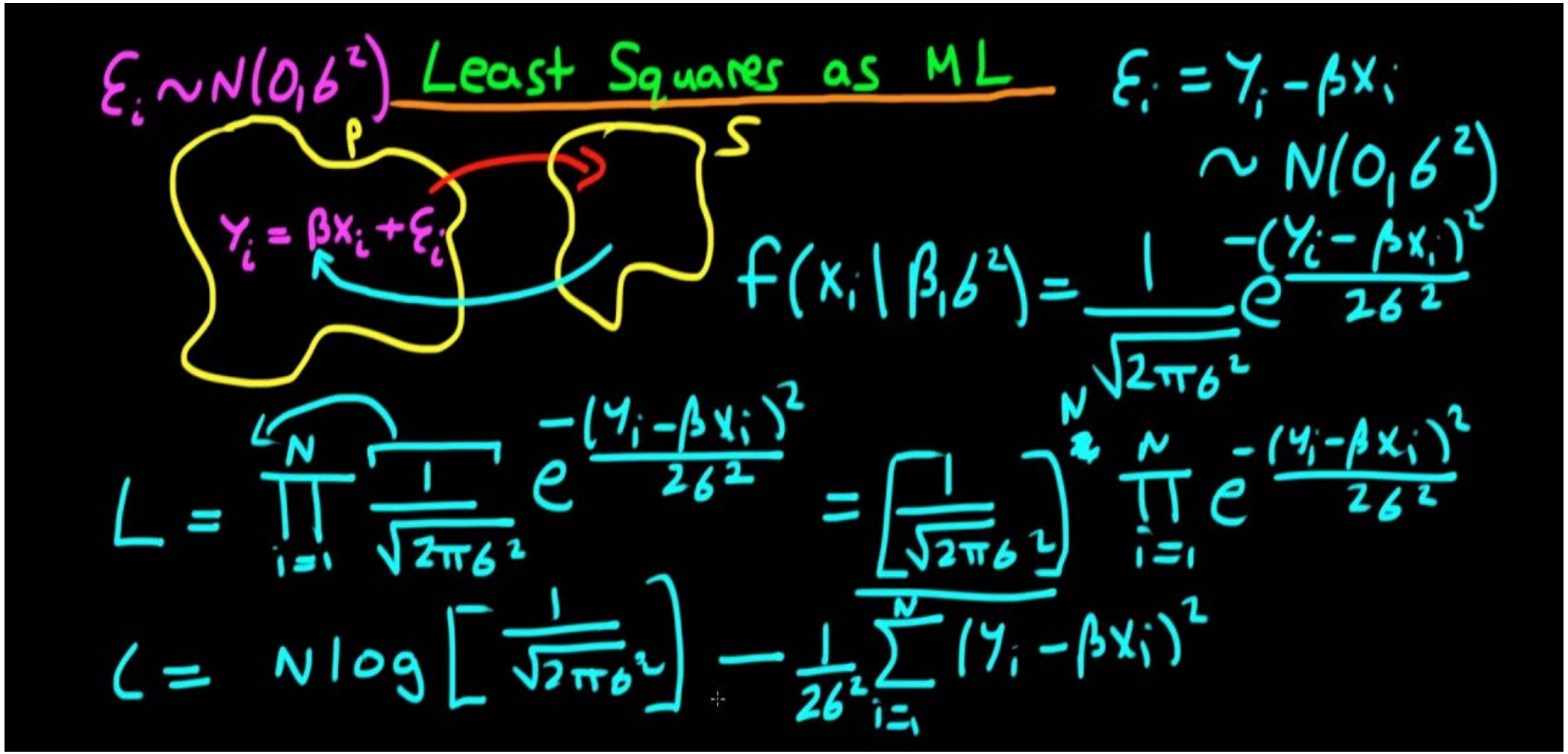
$$\varepsilon \sim \begin{pmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma^2 \end{pmatrix}$$

(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)

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

(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

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

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