3.3 Other Considerations in the Regression Model

Zongyi Liu

2023-05-14

3.3 Other Considerations in the Regression Model

3.3.1 Qualitative Predictors

In our discussion so far, we have assumed that all variables in our linear regression model are quantitative, but in practice, there are predictors that are qualitative.

Predictors with Only Two Levels

Dummy variable: it takes on two possible numerical values to indicate two categories.

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

If we plug this into the model:

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

In practice, we code females as 1 and males as -1, instead of 1 and 0.

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

and use this variable in regression equation, we get

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

Qualitative Predictors with More than Two Levels

If a qualitative predictor has more than two levels, then a single dummy variable might not enough represent all values, thus we would utilize more than one dummy variables. The first could be

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

and the second could be

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

Include both of them in the regression model:

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

Here, we can interpret estimates as follows:

- β_1: the average credit card balance for African Americans
- β_1: the difference in the average balance between the Asian and African American categories
- β_2: the difference in the average balance between the Caucasian and African American categories
  Baseline: the level with no dummy variables: African American

3.3.2 Extensions of the Linear Model

There are two assumptions of linear model, **additive** and **linear**.

- The **additive assumption** means that the effect of changes in a predictor X_j on the response Y is independent of the values of the other predictors.
- The **linear assumption** states that the change in the response Y due to a one-unit change in X_j is constant, regardless of the value of X_j.

Removing the Additive Assumption

The reason to remove this assumption is because of the existence of **interaction term**.

Consider the standard linear regression with two variables:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon.$$

If we increase X_1 by one unit, Y would increase by β_1 unit, and the same mechanism applies for X_2 too. But the change in both of them doesn't influence each other, which is not really correct. Thus we will introduce the interaction term:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon.$$

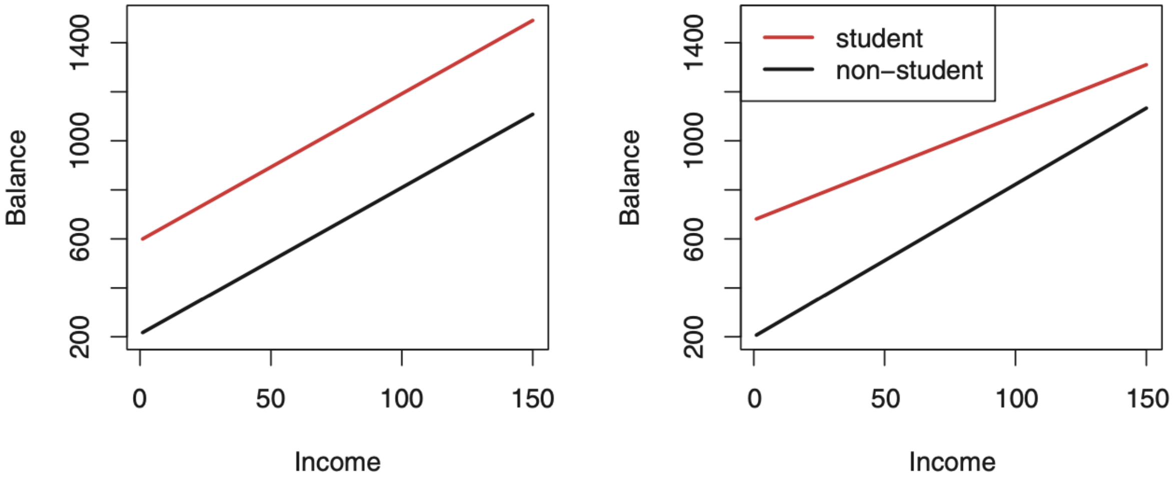
Which can be written as

$$
\begin{aligned}
Y &= \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon \\
&= \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \epsilon
\end{aligned}
$$

From test, we can know that the model containing the interaction term is much better than the model contains only **main effects**.

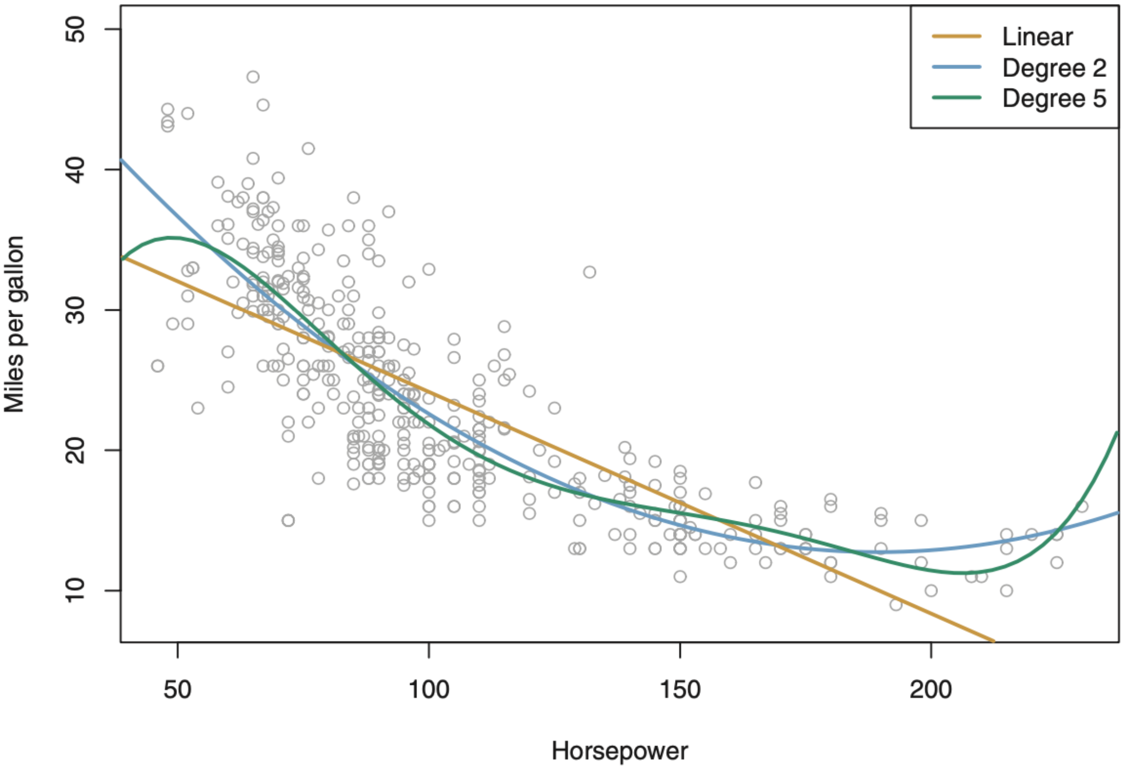
Hierarchical Principle: if we include an interaction in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant.

In the plot below, the left-hand panel shows the model which doesn't have interaction term, and the right-hand panel has the interaction term included.



Non-linear Relationship

As we discussed before, we assume there exists a linear relationship in the model, which is not quite correct because the reality is complicated.



We may add degrees on the model, which is known as **polynomial regression**. We then need to decide which model is the best when fitting the data.

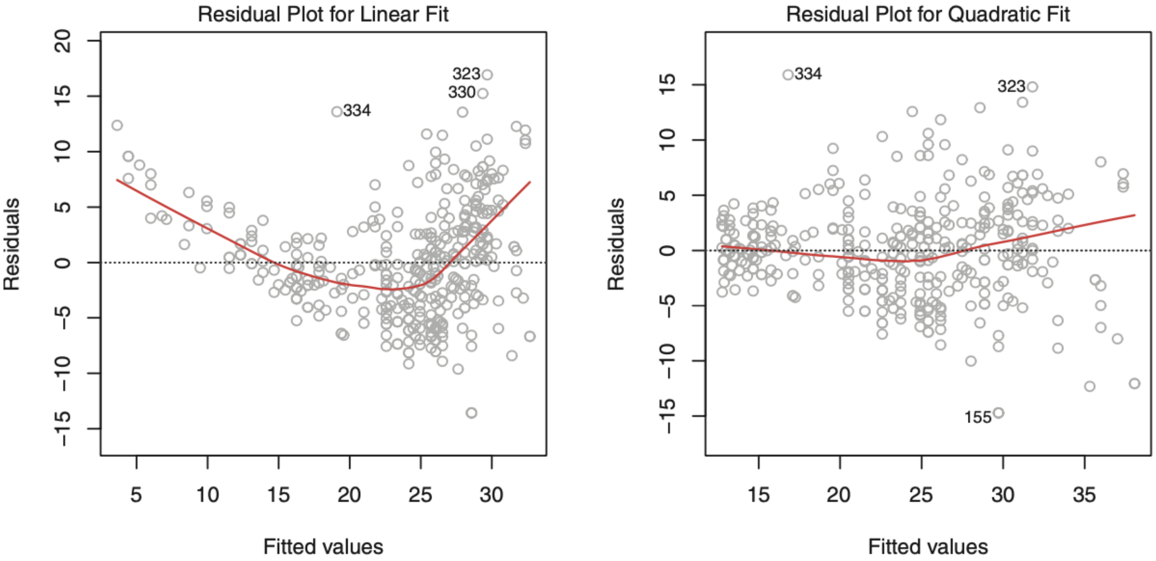
3.3.3 Potential Problems

There are many problems we might face when fitting the model:

1. Non-linearity of the response-predictor relationships
2. Correlation of error terms
3. Non-constant variance of error terms
4. Outliers
5. High-leverage points
6. Col-linearity

Non-linearity

Non-linearity means that there might be a non linear relationship between the predictors and the response. We will use the residual plot to identify it.



In each plot, the red line is a smooth fit to the residuals, intended to make it easier to identify a trend. 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.

To solve this problem, we tend to use transformation, such as log, sqrt, or power.

Correlation of Error Terms

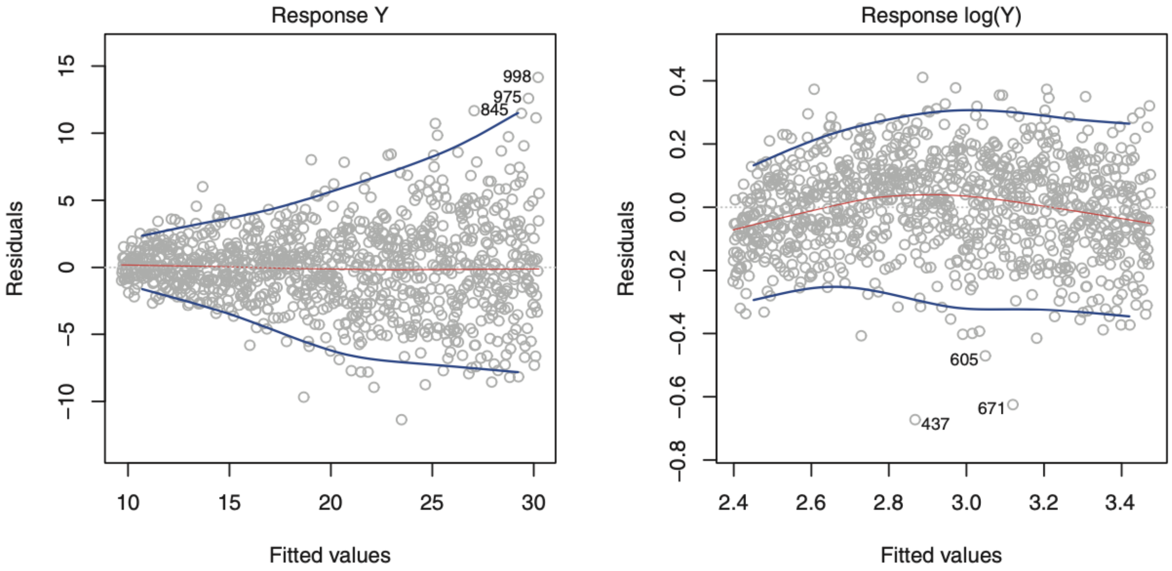
An important assumption of the linear model is that its error terms are not correlated. A counter-example is the time series data, in which error terms might be correlated.

Non-constant Variance of Error Terms

Another assumption is that the error term has a constant variance, $Var(\epsilon_i) = \sigma^2$. All other things, like standard errors, confidence interval, and hypothesis testing, are all related to this.

We will summarize this trend as **heteroscedasticity**.

In residual plot, we can see it as

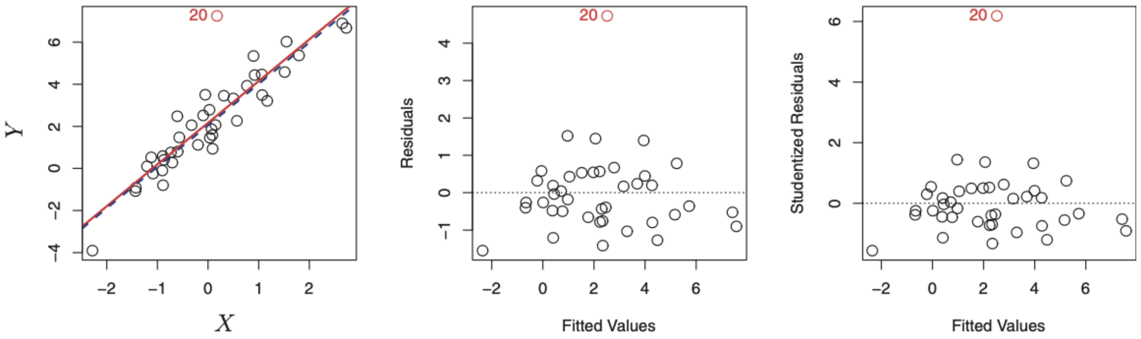


- Left: Strong heteroscedasticity
- Right: constant variance in the error term

Sometimes we have a idea of the variance of each response. For example, the ith 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 $\sigma_i^2 = \frac{\sigma^2}{n_i}$. In this case a simple remedy is to fit our model by **weighted least squares**, with weights proportional to the inverse variances.

Outliers

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



In the plot above, the point 20 is illustrated as the outlier. The central plot is the residual plot, and the right plot is the **studentzied residual** plot, computed by dividing each residual e_i by its estimated standard error. Observations beyond -3 or 3 are considered outliers.

High Leverage Points

Outliers are unusual values for y, whereas **high leverage points** are unusual points for x. In the plot, the point 41 is the high leverage point whereas point 20 is not.

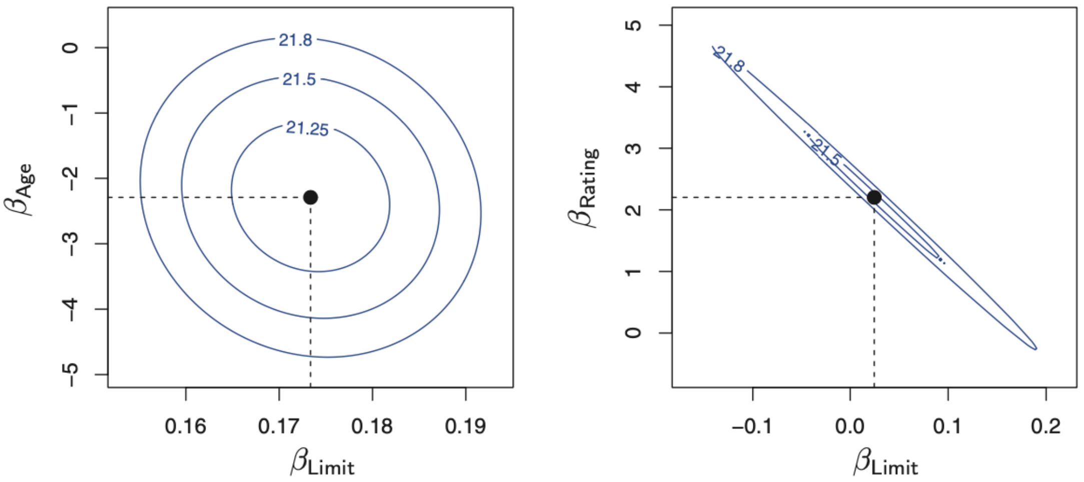
To identify an observation’s leverage, we need to compute the leverage statistic:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2}.$$

From the equation we can see that h_i increases with the distance of x_i from $\bar{x}$. The leverage statistic is always between $1/n$ and 1.

Collinearity

The collinearity refers to the situation in which two or more predictor variables are closely related to one another.



In the left-hand panel of, the two predictors appear to have no obvious relationship. In contrast, in the right-hand panel, the predictors limit and rating are very highly correlated with each other, and we say that they are **collinear**.

In plots, the black dots represent the coefficient value corresponding to the **minimum RSS**. Because of the collinearity, in the right panel, there are many pairs with similar value for RSS.

There are two ways to inspect the collinearity. First we can have a look at the correlation matrix, and the second way is to compute the **variance inflation factor (VIF)**, which can be used to detect the **multicollinearity**:

$$\text{VIF}(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}},$$

$R^2_{X_j|X_{-j}}$ is the R^2 from a regression of X_j onto all of the other predictors, if it is closet to 1, then collinearity is present, and so the VIF will be large.