2 Linear Regression

2.1 Other Considerations in the Regression Model

2.1.1 Qualitative Predictors

Predictors with Two levels:

Gender x_i as dummy variable:

$$x_i = \begin{cases} 1, & \text{the ith person is female} \\ 0, & \text{the ith person is male} \end{cases} \tag{2.1}$$

Model:

$$y_i = \beta_0 + \beta_1 \times x_i + \epsilon_i = \beta_0 + \begin{cases} \beta_1 + \epsilon_i, & \text{the ith person is female} \\ \epsilon_i, & \text{the ith person is male} \end{cases} \tag{2.2}$$

- β_0: average credit card balance among male
- $\beta_0 + \beta_1$: average credit card balance among female
- β_1: average difference in credit card balance between male and female

Predictors with more than Two Levels:

Ethnicity x_{i1}, x_{i2} as dummy variable:

$$x_{1i} = \begin{cases} 1, & \text{the ith person is Asian} \\ 0, & \text{the ith person is not Asian} \end{cases} \tag{2.3}$$

$$x_{2i} = \begin{cases} 1, & \text{the ith person is Caucasian} \\ 0, & \text{the ith person is not Caucasian} \end{cases} \tag{2.4}$$

Model:

$$y_i = \beta_0 + \beta_1 \times x_{1i} + \beta_2 \times x_{2i} + \epsilon_i \tag{2.5}$$

$$= \beta_0 + \begin{cases} \beta_1 + \epsilon_i, & \text{the ith person is Asian} \\ \beta_2 + \epsilon_i, & \text{the ith person is Caucasian} \\ \epsilon_i, & \text{the ith person is African American} \end{cases} \tag{2.6}$$

2.1.2 Extensions of the Linear Model

Interaction Effect:

Consider the linear regression model with one quantitative predictor and one qualitative predictor. We are using the credit data as our example, suppose we wish to predict balance using income(quantitative) and student(qualitative) variables. In the absence of an interaction term, the model takes the form:

$$\text{balance}_i \approx \beta_0 + \beta_1 \times \text{income}_i + \begin{cases} \beta_2, & \text{the ith person is a student} \\ 0, & \text{the ith person is not a student} \end{cases} \tag{2.7}$$

$$= \beta_1 \times \text{income}_i + \begin{cases} \beta_0 + \beta_2, & \text{the ith person is a student} \\ \beta_0, & \text{the ith person is not a student} \end{cases} \tag{2.8}$$

	Coefficient	Std. Error	t-statistic	p-value
Intercept	211.14	32.46	6.51	0
Income	5.98	0.56	10.75	0
Student	382.67	65.31	5.86	0
$R^2 = 27.75\%$				

Table 1: Coefficients table

Source	SS	df	MS	F
Regression	23400858	3-1	11700429	76.22
Residual Error	60939054	400-3	153499	
Total	84339912	400-1		

Table 2: ANOVA table

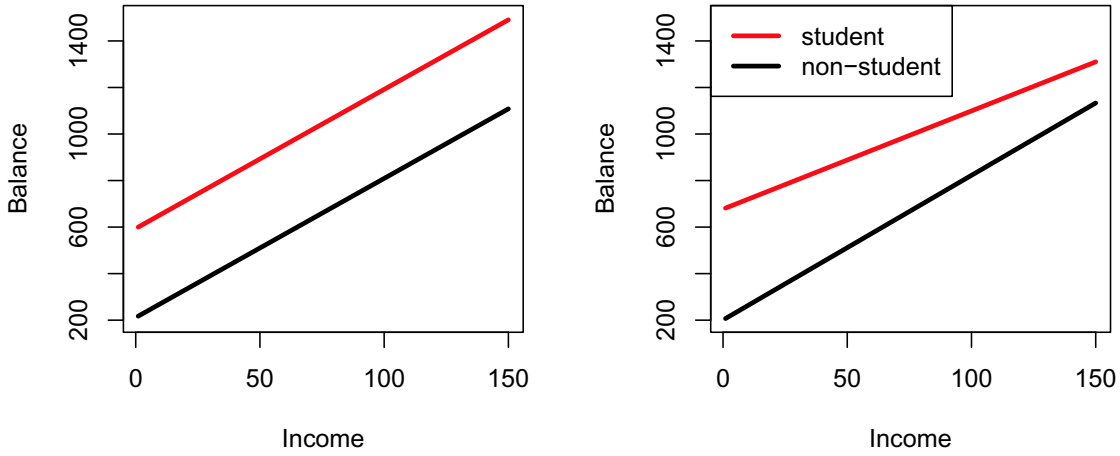

Figure 1: Left: The model (2.8) with no interaction was fit, Right: The model (2.11) with interaction was fit

In the left panel of Figure 1, we notice that Model (2.8) fit two parallel lines to the data, one for students and one for non-students. The lines for students and non-students have different interceptions, $\beta_0 + \beta_2$ versus β_0, but the same slope β_0. The fact that the lines are parallel means that according to Model (2.8), the average effect on balance of a one-unit increase in income does not depend on whether or not the person is a student. This assumption on the model is simply wrong, since in fact a change in income may have a very different effect on the credit card balance of a student versus a non-student.

This issue can be solved by adding an interaction variable into our model:

$$y_i \approx \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} \tag{2.9}$$

- x_{i1}: Income for the ith person (Quantitative variable)
- x_{i2}: whether the ith person is a student or not (Qualitative Dummy variable)
- $x_{i1}x_{i2}$: Interaction term between Income(x_{i1}) and Student(x_{i2})

The model with interaction term can be interpreted as below:

$$\text{balance}_i \approx \beta_0 + \beta_1 \times \text{income}_i + \begin{cases} \beta_2 + \beta_3 \times \text{income}_i, & \text{the ith person is a student} \\ 0, & \text{the ith person is not a student} \end{cases} \tag{2.10}$$

$$= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times \text{income}_i, & \text{the ith person is a student} \\ \beta_0 + \beta_1 \times \text{income}_i, & \text{the ith person is not a student} \end{cases} \tag{2.11}$$

In the right panel of Figure 1, we notice that when we added a interaction term in our model, we get two different regression lines for the students and the non-students. Those lines have different intercepts, $(\beta_0 + \beta_2)$ versus β_0 and different slopes, $(\beta_1 + \beta_3)$ versus β_1. This allows for the possibility that changes in income may affect the credit card balances of students and non-students differently.

2.1.3 Potential problems for the assumptions of linear regression

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

1. Non-linearity of the Data:

Ideally, the residual plot will show no obvious pattern. The presence of a pattern may indicate a problem with non-linearity.

How to detect Non-linearity:

- Plot the residuals$(y_i - \hat{y}_i)$ versus the fitted values $\hat{y}_i$ as a residual plot (Figure 2)

How to solve Non-linearity:

- If the residual plot indicates non-linear in the data, then adding a non-linear transformations of the predictors, such as $\log X, \sqrt{X}, X^2$, in the regression model.

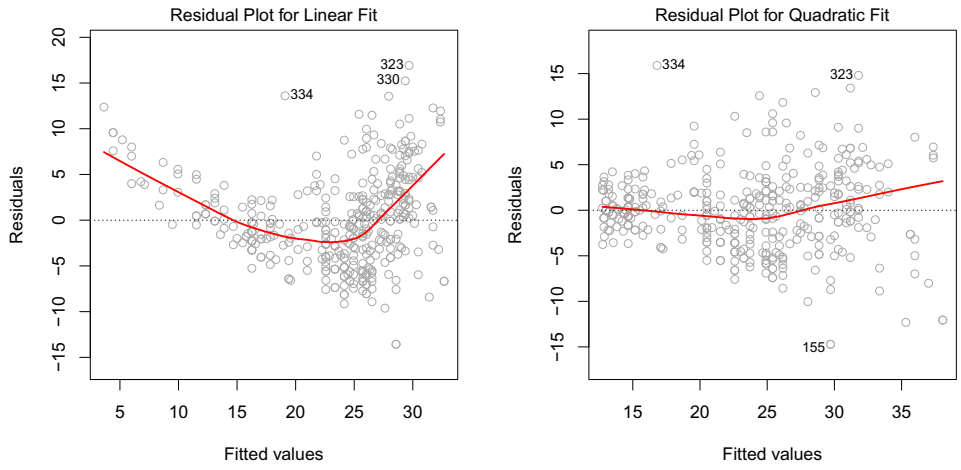

Figure 2: Residual Plot, Left: a strong pattern in the residual indicates non-linearity if we fit our data to a linear model. Right: There is little pattern in the residual when we fit a quadratic model to the data

2. Correlation of Error Terms:

In the assumption of the linear regression model, the error terms $\epsilon_1, \epsilon_1, \ldots, \epsilon_n$ are uncorrelated. If in fact there is correlation among the error term, then **the estimated standard**

errors will tend to underestimate the true standard error. Such correlations frequently occur in time series data. In many cases, adjacent observations will have positive correlated errors.

How to detect Correlation of Error Terms:

- Plot the residuals$(y_i - \hat{y}_i)$ versus the observation (Figure 3)
- Identify obvious pattern in Figure 3

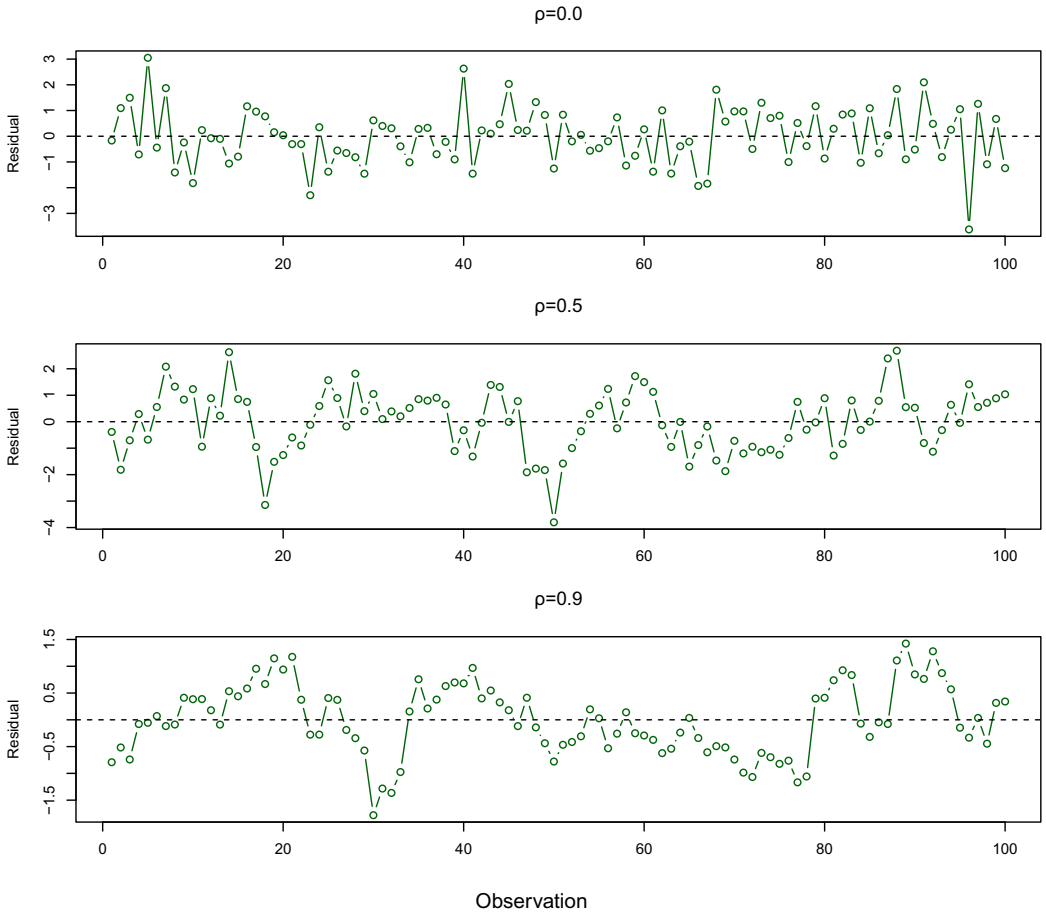

Figure 3: Plots of residuals versus simulated time series observations with different levels of ρ

How to solve Correlation of Error Terms:

- If the residual plot indicates an obvious patten, then there might be a problem experiment design.

3. Non-Constant Variance of Error Terms(Heteroscedasticity):

In the assumption of the linear regression model, the variance of error terms are constant($\text{Var}(\epsilon_i) = \sigma^2$). Unfortunately, it is often the case that the variances of the error terms are non-constant.

How to detect Non-Constant Variance of Error Terms:

- Plot residual plot to detect a funnel shape (Figure 4)

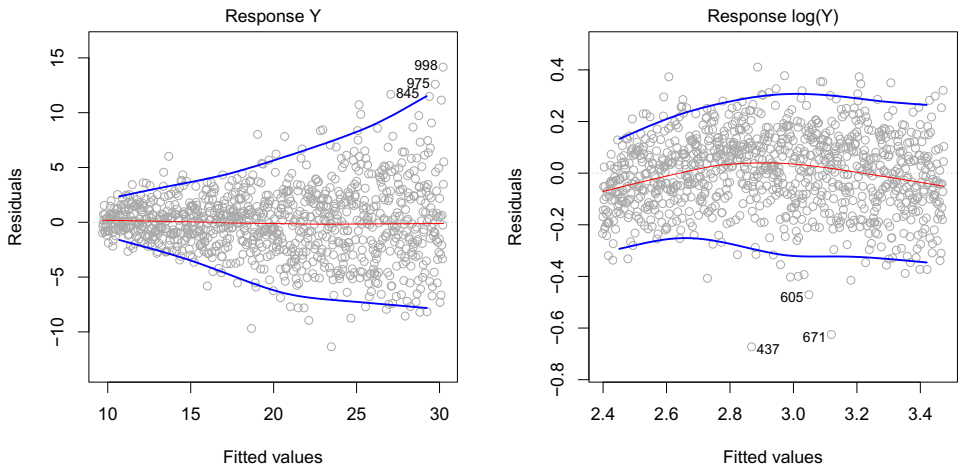

Figure 4

How to solve Non-Constant Variance of Error Terms:

- Transform the response Y using a concave function such as $\log Y, \sqrt{Y}$
- Or, fit our model by weighted least squares or generalized least squares

4. Outliers:

How to detect Outliers:

- Plot residual plot and Studentized residual plot to detect outliers (Figure 5)

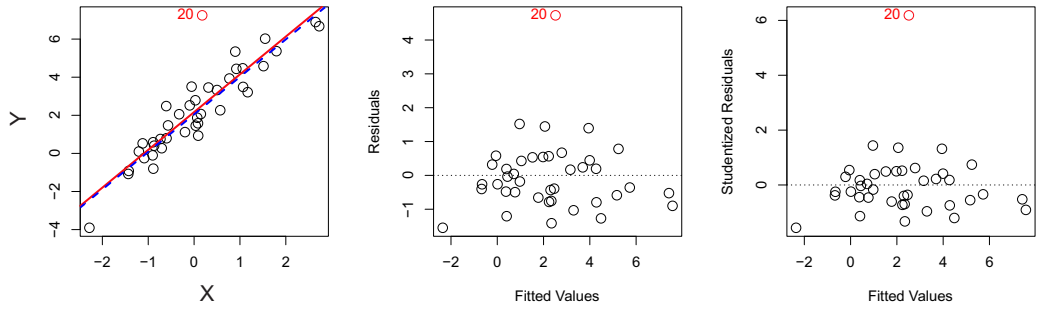

Figure 5

How to solve Outliers:

- Remove them

5. High Leverage Points:

Leverage points means the observations have an unusual value of x_i. High leverage observations tend to have a sizable impact on the estimated regression model.

How to detect High Leverage Points:

- In order to identify an observation is a high leverage point, we compute the leverage statistic. A large value of this statistic indicates an observation with high leverage. (Figure 6)

The leverage statistic for a simple linear regression where $y_i = \beta_0 + \beta_1 x_i$, ($p = 1$, not including intercept)

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n} (x_i - \bar{x})^2}, \text{where } \frac{1}{n} < h_i < 1 \tag{2.12}$$

For the average leverage $\bar{h} = \frac{p}{n}$. So if a given observation has a leverage statistic that greatly exceeds $\frac{p+1}{n}$, then we suspect the corresponding point has high leverage ($x_i > \frac{p+1}{n}$). The proof of average leverage for a simple linear regression :

$$\bar{h} = \sum_{i=1}^{n} \frac{h_i}{n} = \left(\sum_{i=1}^{n} \frac{1}{n} + \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n} (x_i - \bar{x})^2} \right) /n = 2/n \tag{2.13}$$

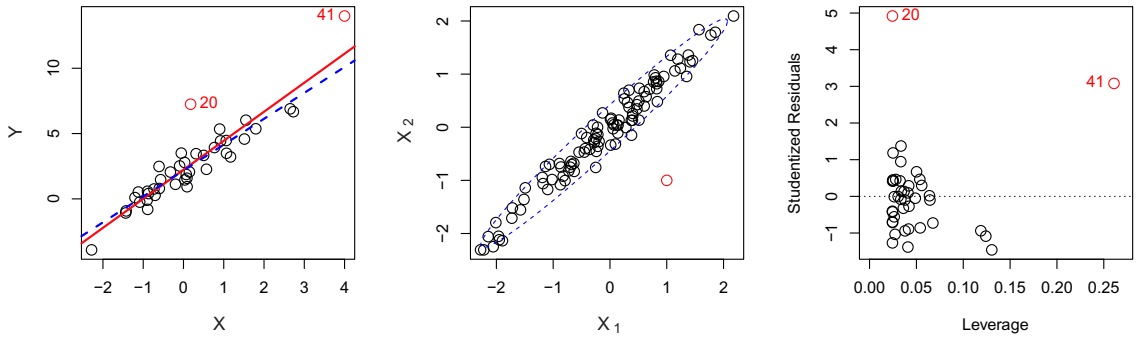

Figure 6

The leverage for multiple predictors:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

$$= \mathbf{H}\mathbf{y}$$

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

The leverage, h_{ii}, quantifies the influence that the observed response y_i has on its predicted value $\hat{y}_i$. That is, if h_{ii} is small, then the observed response y_i plays only a small role in the value of the predicted response $\hat{y}_i$. On the other hand, if h_{ii} is large, then the observed response y_i plays a large role in the value of the predicted response $\hat{y}_i$. It's for this reason that the h_{ii} are called the "leverage".

Here are some important properties of the leverage:

- The leverage h_{ii} is a measure of the distance between the x_i value for the ith data point and the mean of the x values for all n data points.
- The sum of the h_{ii} equals $p + 1$, the number of parameters (regression coefficients including the intercept).
- The leverage h_{ii} is a number between 0 and 1, inclusive.

Example

Obs.	X	Y
1	0.1	-0.0716
2	0.45401	4.1673
3	1.09765	6.5703
4	1.27936	13.815
5	2.20611	11.4501
6	2.50064	12.9554
7	3.0403	20.1575
8	3.23583	17.5633
9	4.45308	26.0317
10	4.1699	22.7573
11	5.28474	26.303
12	5.59238	30.6885
13	5.92091	33.9402
14	6.66066	30.9228
15	6.79953	34.11
16	7.97943	44.4536
17	8.41536	46.5022
18	8.71607	50.0568
19	8.70156	46.5475
20	9.16463	45.7762
21	4	40

R code:

```
x = matrix(c(0.1, 0.45401, 1.09765, 1.27936, 2.20611, 2.50064, 3.0403,
3.23583, 4.45308, 4.1699, 5.28474,5.59238 ,5.92091,6.66066,6.79953,
7.97943,8.41536,8.71607,8.70156,9.16463, 4), ncol = 1)
x = matrix(cbind(rep(1,21),x), ncol=2)
H = x %*% solve(t(x)%*%x) %*% t(x)
H = round(H,4)
```

6. Collinearity:

Collinearity refers to the situation in which two or more predictor variables are closely related to one another. The importance of the predictor variables may be masked if we did not consider Collinearity.

How to detect Collinearity:

- Compute correlation matrix for each predictor
- Or, compute the variance inflation factor(VIF)

How to solve Collinearity:

- Drop one of the problematic variables from the regression model, since the other variable provides enough information.
- Or, combine the collinear variables together to get a single predictor.

Reference

張翔 (2012), 《提綱契領學統計》, 四版, 鼎茂圖書

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1). Springer, Berlin: Springer series in statistics.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 6). New York: springer.

Wasserman, L. (2013). *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.