Statistical Methods
統計方法

NOVEMBER 07, 2023

I-CHEN LEE

Simple Linear Regression

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

The model defines the *population regression line*, which population regression line is the best linear approximation to the true relationship between X and Y.

By the LSE, the estimators are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

where $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$ are the sample means.

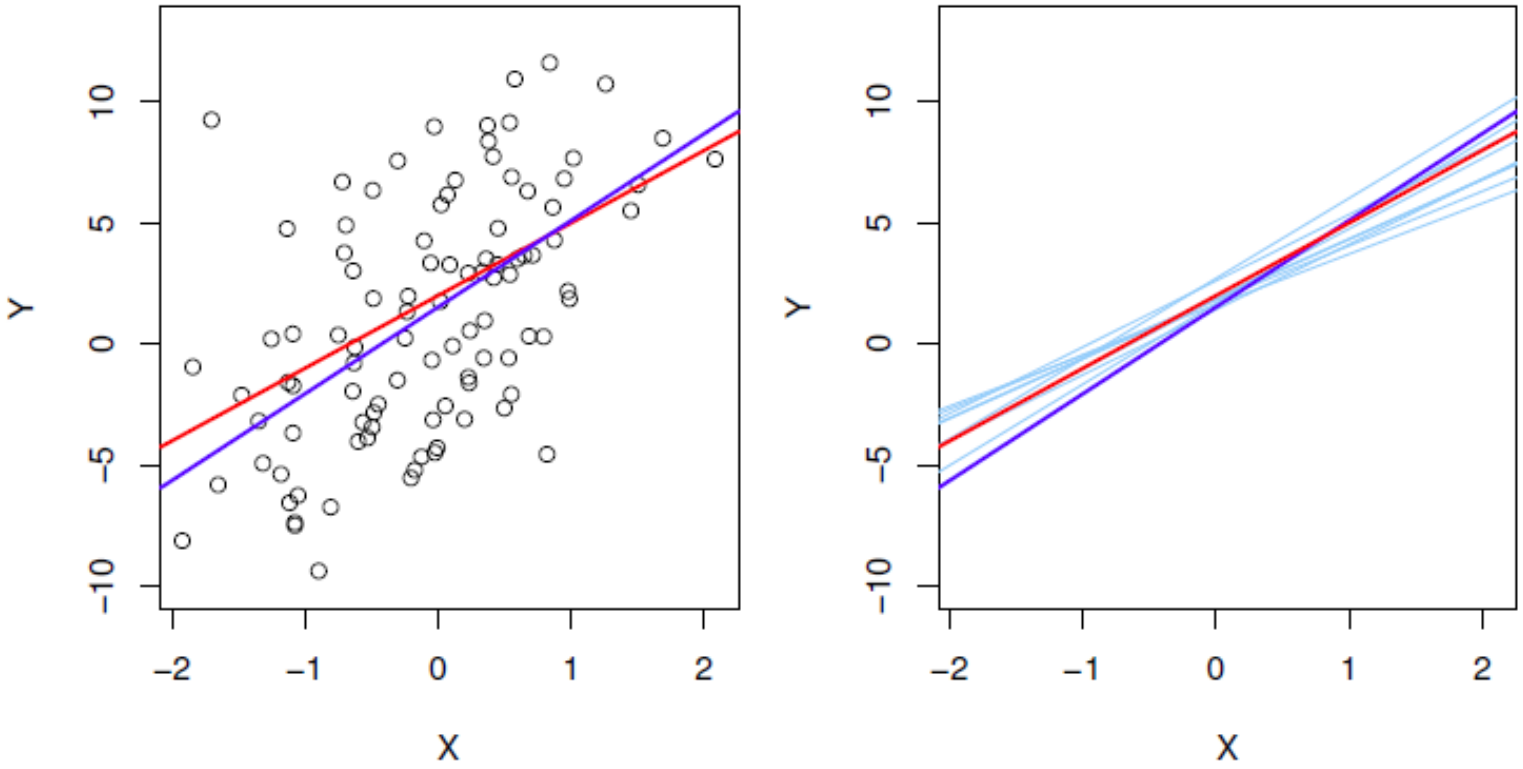


FIGURE 3.3. *A simulated data set.* Left: *The red line represents the true relationship,* $f(X) = 2 + 3X$, *which is known as the population regression line. The blue line is the least squares line; it is the least squares estimate for* $f(X)$ *based on the observed data, shown in black.* Right: *The population regression line is again shown in red, and the least squares line in dark blue. In light blue, ten least squares lines are shown, each computed on the basis of a separate random set of observations. Each least squares line is different, but on average, the least squares lines are quite close to the population regression line.*

Relative quantities

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right], \quad \mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$\hat{\sigma}^2 = \sqrt{\mathrm{RSS}/(n-2)}. \qquad \mathrm{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

For linear regression, the 95 % confidence interval for β_1 approximately a 95 % chance that the interval

$$\left[\hat{\beta}_1 - 2 \cdot \mathrm{SE}(\hat{\beta}_1),\ \hat{\beta}_1 + 2 \cdot \mathrm{SE}(\hat{\beta}_1) \right]$$

Hypothesis testing

$$H_0 : \text{There is no relationship between } X \text{ and } Y$$

versus the *alternative hypothesis*

$$H_a : \text{There is some relationship between } X \text{ and } Y.$$

Mathematically, this corresponds to testing

$$H_0 : \beta_1 = 0$$

versus

$$H_a : \beta_1 \neq 0$$

Q: Can we test for $\beta_0 = 0$?

Example 1: Simple linear regression

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

TABLE 3.1. *For the* Advertising *data, coefficients of the least squares model for the regression of number of units sold on TV advertising budget. An increase of* \$1,000 *in the TV advertising budget is associated with an increase in sales by around 50 units. (Recall that the* sales *variable is in thousands of units, and the* TV *variable is in thousands of dollars.)*

Table 3.1 provides details of the least squares model for the regression of number of units sold on TV advertising budget for the Advertising data. From the p-value, we can conclude that $\beta_0 \neq 0$ and $\beta_1 \neq 0$.

Example2. Multiple regression

Simple regression of `sales` on `radio`

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	9.312	0.563	16.54	< 0.0001
`radio`	0.203	0.020	9.92	< 0.0001

Simple regression of `sales` on `newspaper`

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	12.351	0.621	19.88	< 0.0001
`newspaper`	0.055	0.017	3.30	0.00115

TABLE 3.3. *More simple linear regression models for the* `Advertising` *data. Coefficients of the simple linear regression model for number of units sold on* Top: *radio advertising budget and* Bottom: *newspaper advertising budget. A* \$1,000 *increase in spending on radio advertising is associated with an average increase in sales by around 203 units, while the same increase in spending on newspaper advertising is associated with an average increase in sales by around 55 units. (Note that the* `sales` *variable is in thousands of units, and the* `radio` *and* `newspaper` *variables are in thousands of dollars.)*

Do not use several simple linear regressions

- Dependent variable: Sales
- Independent variables: TV, Radio, Newspaper

- Multiple regression:

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper} + \epsilon.$$

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	−0.001	0.0059	−0.18	0.8599

TABLE 3.4. *For the* Advertising *data, least squares coefficient estimates of the multiple linear regression of number of units sold on TV, radio, and newspaper advertising budgets.*

Which model is better?

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \epsilon.$$

Quantity	Value
Residual standard error	3.26
R^2	0.612
F-statistic	312.1

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper} + \epsilon.$$

Quantity	Value
Residual standard error	1.69
R^2	0.897
F-statistic	570

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$

$H_a :$ at least one β_j is non-zero.

Predictions

The coefficient estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ are estimates for $\beta_0, \beta_1, \ldots, \beta_p$. That is, the *least squares plane*

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$$

is only an estimate for the *true population regression plane*

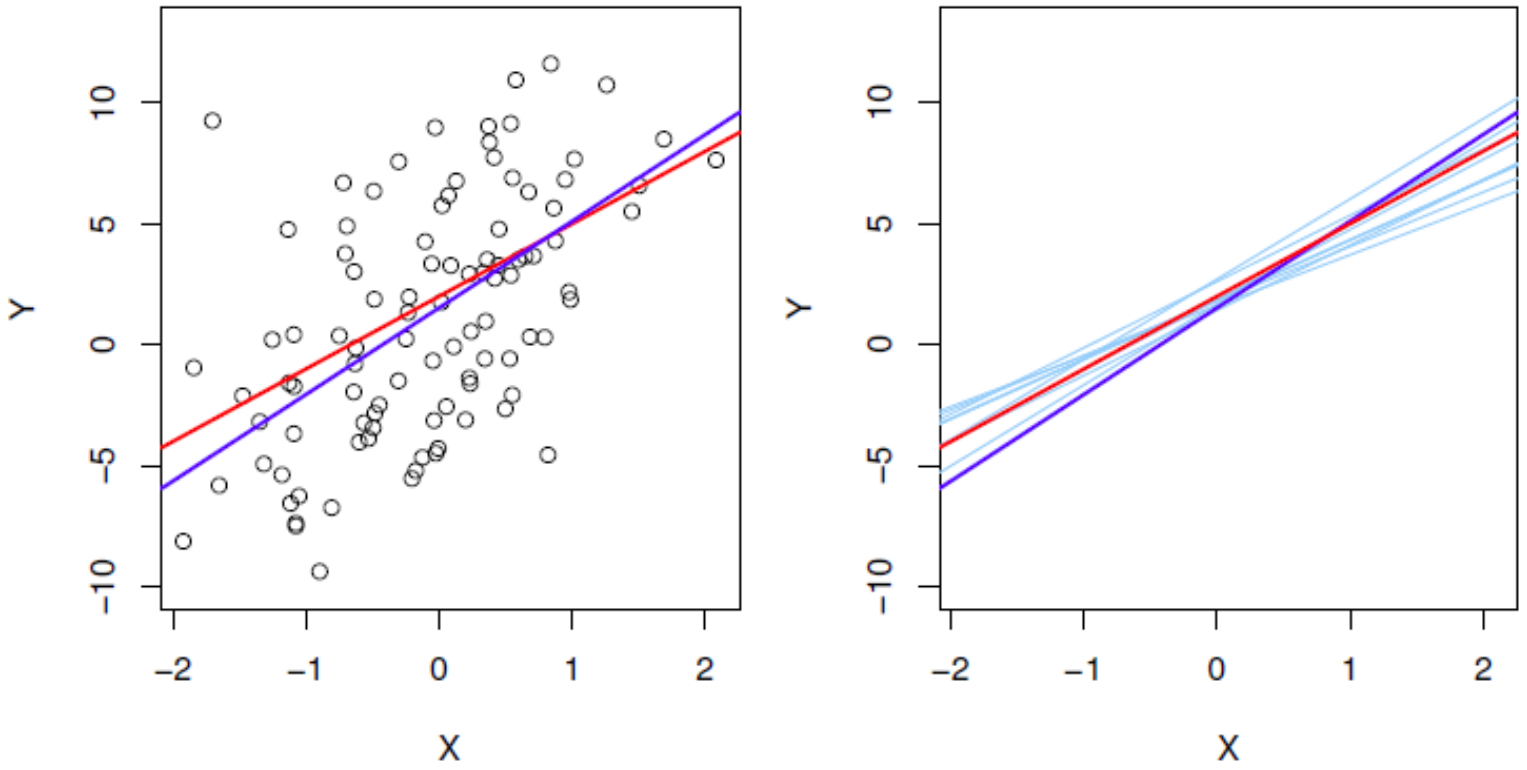
$$f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

The inaccuracy in the coefficient estimates is related to the *reducible error* from Chapter 2. We can compute a *confidence interval* in order to determine how close $\hat{Y}$ will be to $f(X)$.

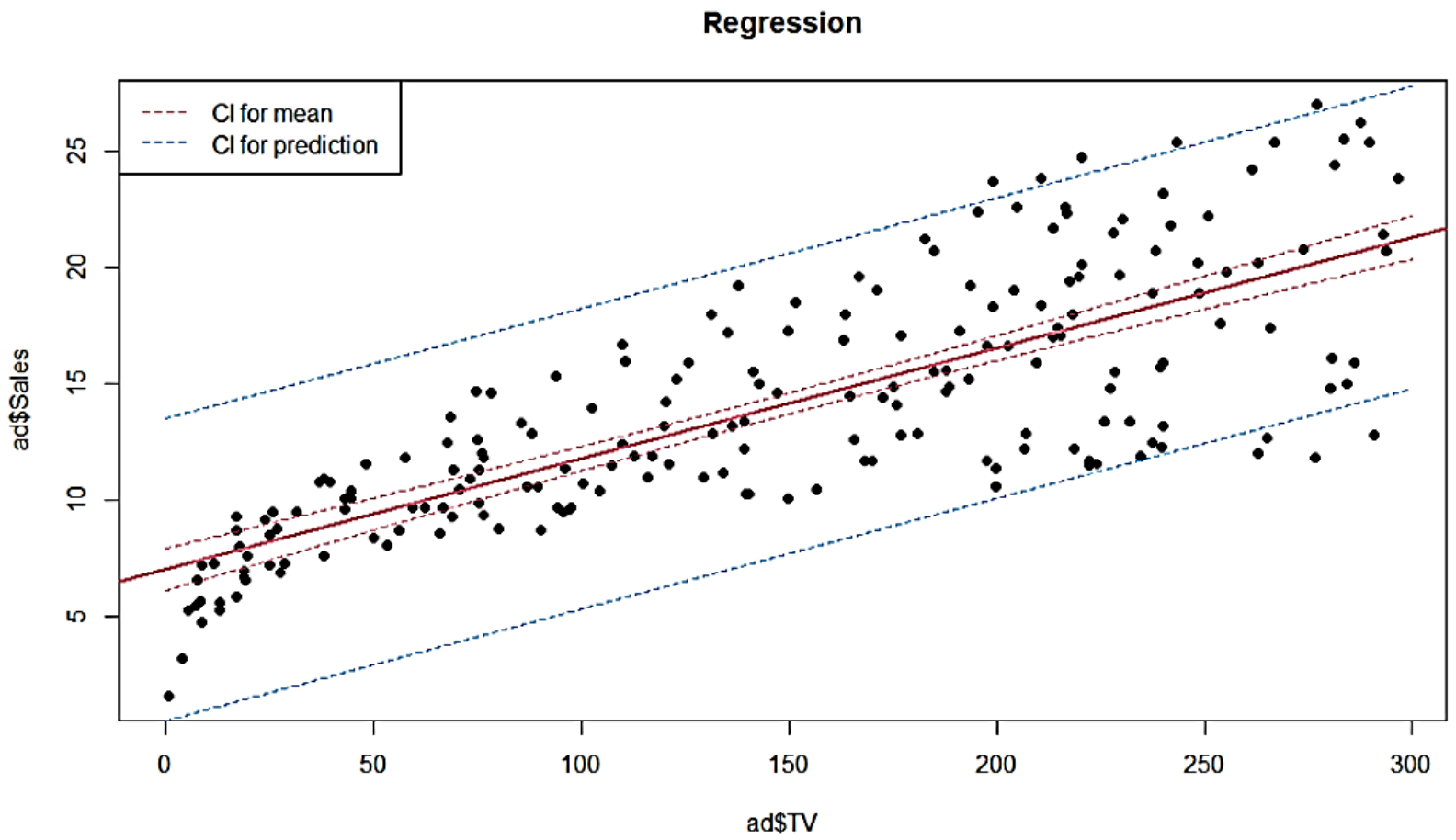
Differences???

How much will Y vary from $\hat{Y}$?

We use *prediction intervals* to answer this question.


Y
10
5
0
−5
−10
−2
−1
0
1
2
X
Y
10
5
0
−5
−10
−2
−1
0
1
2
X

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \epsilon.$$



3.3.1 Qualitative Predictors

For example, the `Credit` data set displayed in Figure 3.6 records variables for a number of credit card holders. The response is `balance` (average credit card debt for each individual) and there are several quantitative predictors: `age`, `cards` (number of credit cards), `education` (years of education), `income` (in thousands of dollars), `limit` (credit limit), and `rating` (credit rating). Each panel of Figure 3.6 is a scatterplot for a pair of variables whose identities are given by the corresponding row and column labels. For example, the scatterplot directly to the right of the word "Balance" depicts `balance` versus `age`, while the plot directly to the right of "Age" corresponds to `age` versus `cards`. In addition to these quantitative variables, we also have four qualitative variables: `own` (house ownership), `student` (student status), `status` (marital status), and `region` (East, West or South).

Credit dataset

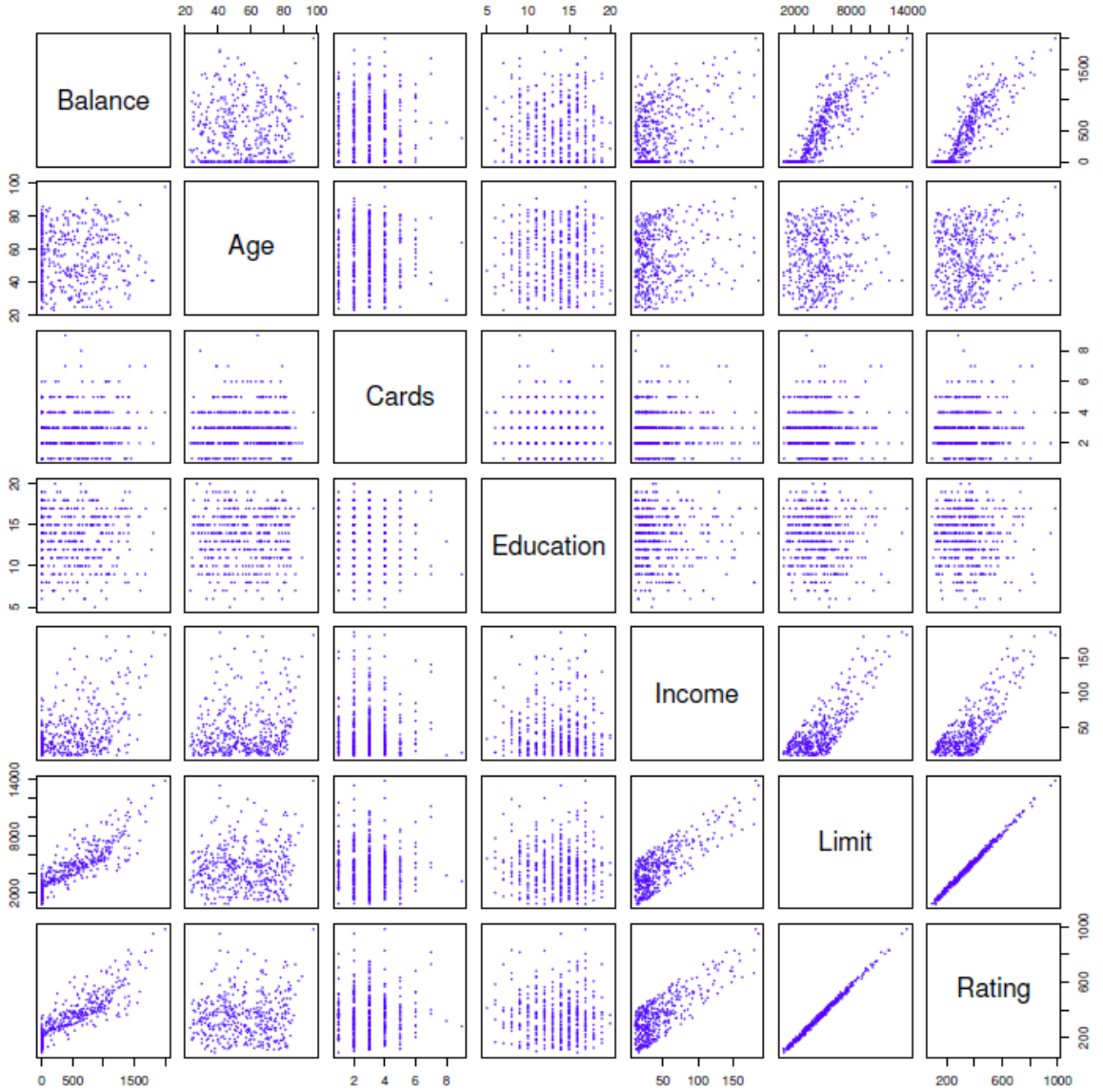
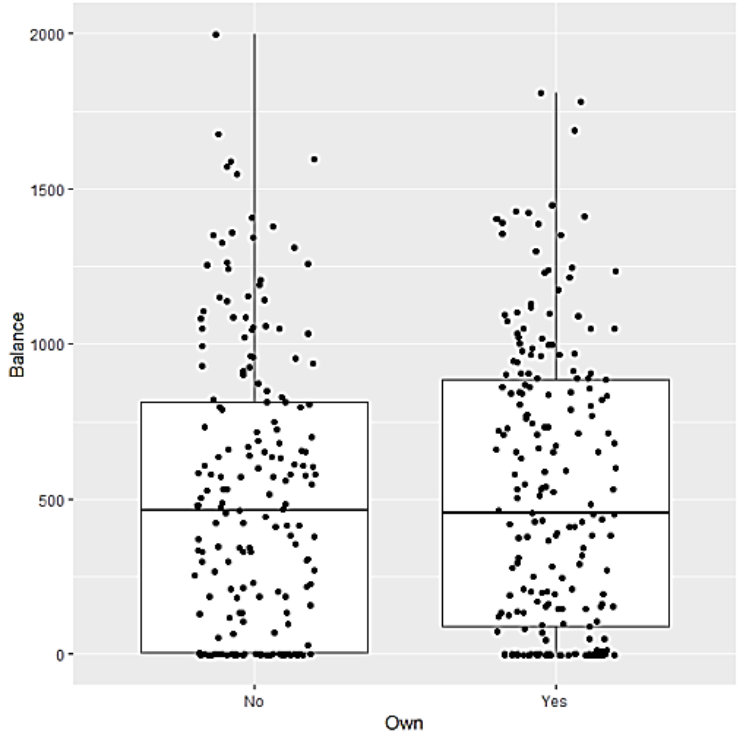


FIGURE 3.6. *The* `Credit` *data set contains information about* `balance`, `age`, `cards`, `education`, `income`, `limit`, *and* `rating` *for a number of potential customers.*

Predictors with Only Two Levels

Use the variable “Own” to develop a regression model.



Coding first. (Dummy variable)

$$x_i = \begin{cases} 1 & \text{if } i\text{th person owns a house} \\ 0 & \text{if } i\text{th person does not own a house} \end{cases}$$

Predictors with Only Two Levels

Regression model would be

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person owns a house} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person does not.} \end{cases}$$

- β_0 is the average balance among those who do not own a house.
- $\beta_0 + \beta_1$ is the average balance among those who own a house.
- β_1 is the average difference in balance between owners and non-owners.

Results

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

TABLE 3.7. *Least squares coefficient estimates associated with the regression of* balance *onto* own *in the* Credit *data set. The linear model is given in (3.27). That is, ownership is encoded as a dummy variable, as in (3.26).*

This indicates that there is no statistical evidence of a difference in average credit card balance based on house ownership.

Q: Supported by the figure???

Other coding: sum-coding

a dummy variable

$$x_i = \begin{cases} 1 & \text{if } i\text{th person owns a house} \\ -1 & \text{if } i\text{th person does not own a house} \end{cases}$$

and use this variable in the regression equation. This results in the model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person owns a house} \\ \beta_0 - \beta_1 + \epsilon_i & \text{if } i\text{th person does not own a house.} \end{cases}$$

One-hot coding

$$x_{own,i} = \begin{cases} 1, & \text{if the i-th person owns a house} \\ 0, & \text{if the i-th person does not own a house..} \end{cases}$$

$$x_{not-own,i} = \begin{cases} 0, & \text{if the i-th person owns a house} \\ 1, & \text{if the i-th person does not own a house..} \end{cases}$$

The regression model is

$$y_i = \alpha_0 + \alpha_1 x_{own,i} + \alpha_2 x_{not-own,i} + \varepsilon_i,$$

where

$$y_i = \begin{cases} \alpha_0 + \alpha_1 + \varepsilon_i, & \text{if the i-th person owns a house} \\ \alpha_0 + \alpha_2 + \varepsilon_i, & \text{if the i-th person does not own a house.} \end{cases}$$

- $\alpha_0 + \alpha_1$ is the average balance among those who own a house.
- $\alpha_0 + \alpha_2$ is the average balance among those who do not own a house.
- Model assumption: $\varepsilon_i \sim N(0, \sigma^2)$.

Need 3 parameters $(\alpha_0, \alpha_1, \alpha_2)$?

Let $\mu_1 = \alpha_0 + \alpha_1$ and $\mu_2 = \alpha_0 + \alpha_2$. It implies the regression model is

$$y_i = \mu_1 x_{own,i} + \mu_2 x_{not-own,i} + \varepsilon_i.$$

Short summary:
The coding depends on the interpretation of the intercept.

Qualitative Predictors with More than Two Levels

When a qualitative predictor has more than two levels, a single dummy variable cannot represent all possible values. Then, we can create additional dummy variables.

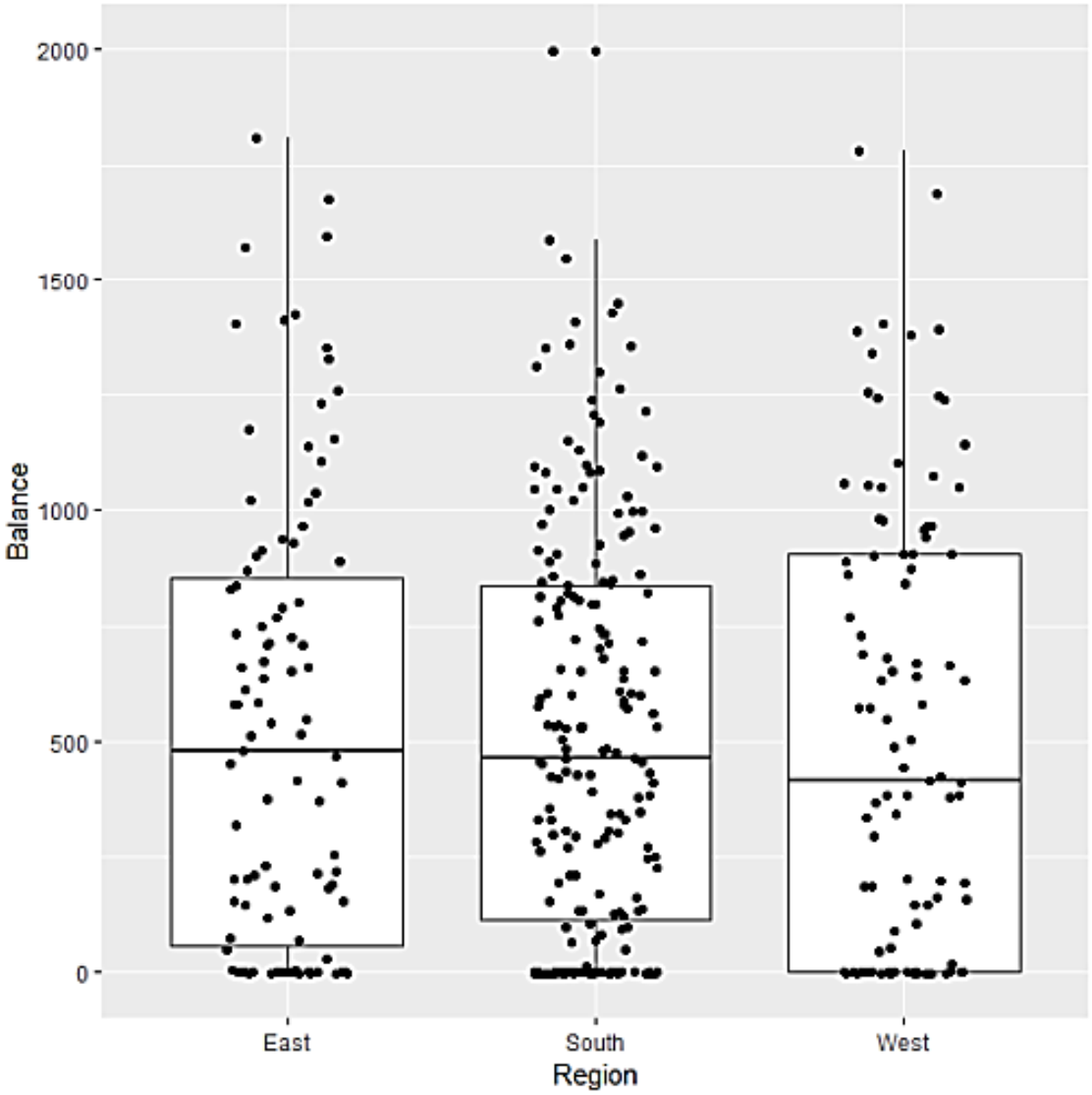
Region (3 levels for East, South, or West):

$$x_{i1} = \begin{cases} 1, & \text{if the i-th person is from the South} \\ 0, & \text{if the i-th person is not from the South.} \end{cases}$$

$$x_{i2} = \begin{cases} 1, & \text{if the i-th person is from the West} \\ 0, & \text{if the i-th person is not from the West.} \end{cases}$$

In a regression model, $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$

$$y_i = \begin{cases} \beta_0 + \varepsilon_i, & \text{if the i-th person is from East} \\ \beta_0 + \beta_1 + \varepsilon_i, & \text{if the i-th person is from South} \\ \beta_0 + \beta_2 + \varepsilon_i, & \text{if the i-th person is from West.} \end{cases}$$


2000
1500
1000
500
0
Balance
East
South
West
Region

Results

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	531.00	46.32	11.464	< 0.0001
`region[South]`	−12.50	56.68	−0.221	0.8260
`region[West]`	−18.69	65.02	−0.287	0.7740

TABLE 3.8. *Least squares coefficient estimates associated with the regression of* `balance` *onto* `region` *in the* `Credit` *data set. The linear model is given in (3.30). That is, region is encoded via two dummy variables (3.28) and (3.29).*

Q: Supported by the figure???

Interaction terms: Quantitative × Quantitative

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

Alternatively,

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

$$\text{where } \tilde{\beta}_1 = \beta_1 + \beta_3 X_2.$$

The association between X_1 and Y is no longer constant: a change in the value of X_2 will change the association between X_1 and Y .

Example

- $R^2 = 0.968$ with interaction term
- $R^2 = 0.897$ without interaction term

	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

TABLE 3.9. *For the* Advertising *data, least squares coefficient estimates associated with the regression of* sales *onto* TV *and* radio*, with an interaction term, as in (3.33).*

We now return to the Advertising example. A linear model that uses radio, TV, and an interaction between the two to predict sales takes the form

$$\begin{aligned} \texttt{sales} &= \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times (\texttt{radio} \times \texttt{TV}) + \epsilon \\ &= \beta_0 + (\beta_1 + \beta_3 \times \texttt{radio}) \times \texttt{TV} + \beta_2 \times \texttt{radio} + \epsilon. \end{aligned} \quad (3.33)$$

We can interpret β_3 as the increase in the effectiveness of TV advertising associated with a one-unit increase in radio advertising (or vice-versa).

Interpretation

$$\begin{aligned} \texttt{sales} &= \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times (\texttt{radio} \times \texttt{TV}) + \epsilon \\ &= \beta_0 + (\beta_1 + \beta_3 \times \texttt{radio}) \times \texttt{TV} + \beta_2 \times \texttt{radio} + \epsilon. \end{aligned} \quad (3.33)$$

- An increase in TV advertising of \$1,000 is associated with increased sales of $(\hat{\beta}_1 + \hat{\beta}_3 \times \text{radio}) \times 1{,}000 = 19 + 1.1 \times \text{radio}$ units.
- An increase in radio advertising of \$1,000 will be associated with an increase in sales of $(\hat{\beta}_2 + \hat{\beta}_3 \times \text{TV}) \times 1{,}000 = 29 + 1.1 \times \text{TV}$ units.

Interaction terms: Quantitative + Qualitative

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

$Y: Balance;$
$X_1: income; X_2: Student$ (using 0/1 coding)

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

Interaction terms: Quantitative × Qualitative

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

$Y: Balance;$
$X_1: income; X_2: Student$ (using 0/1 coding)

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

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

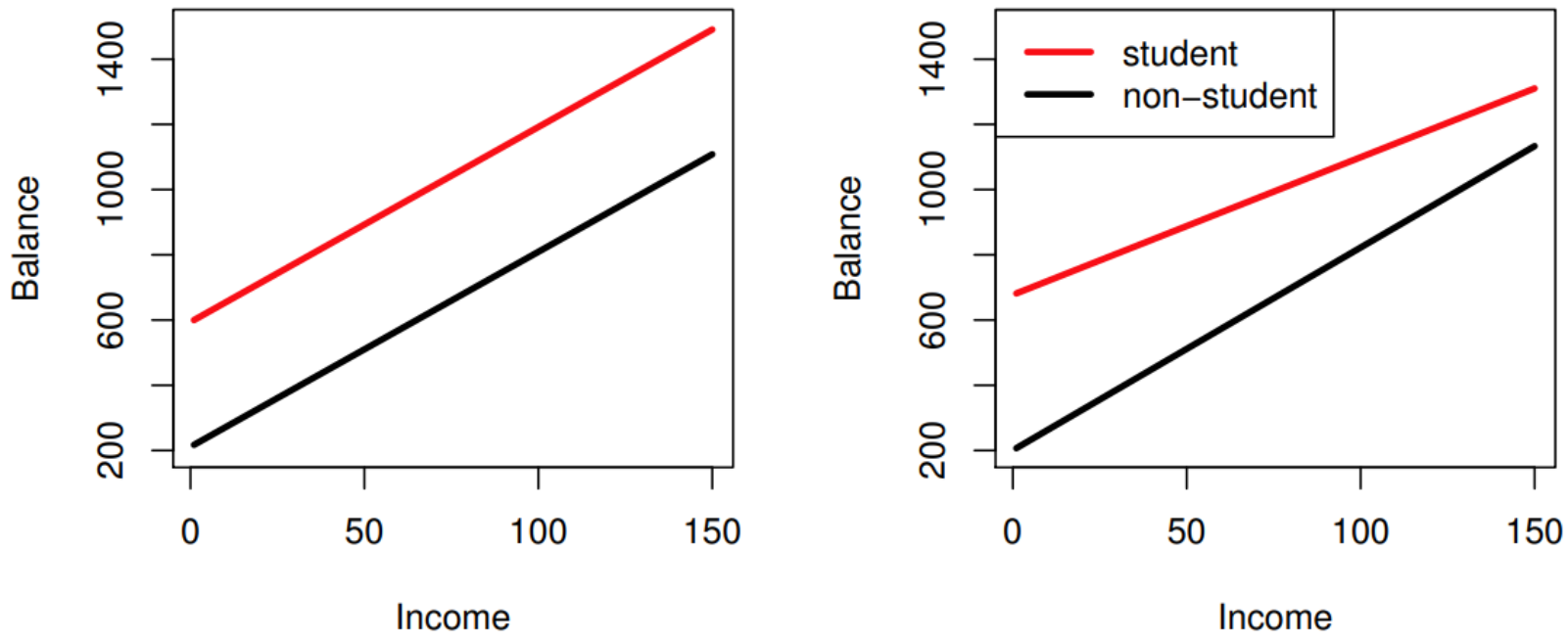
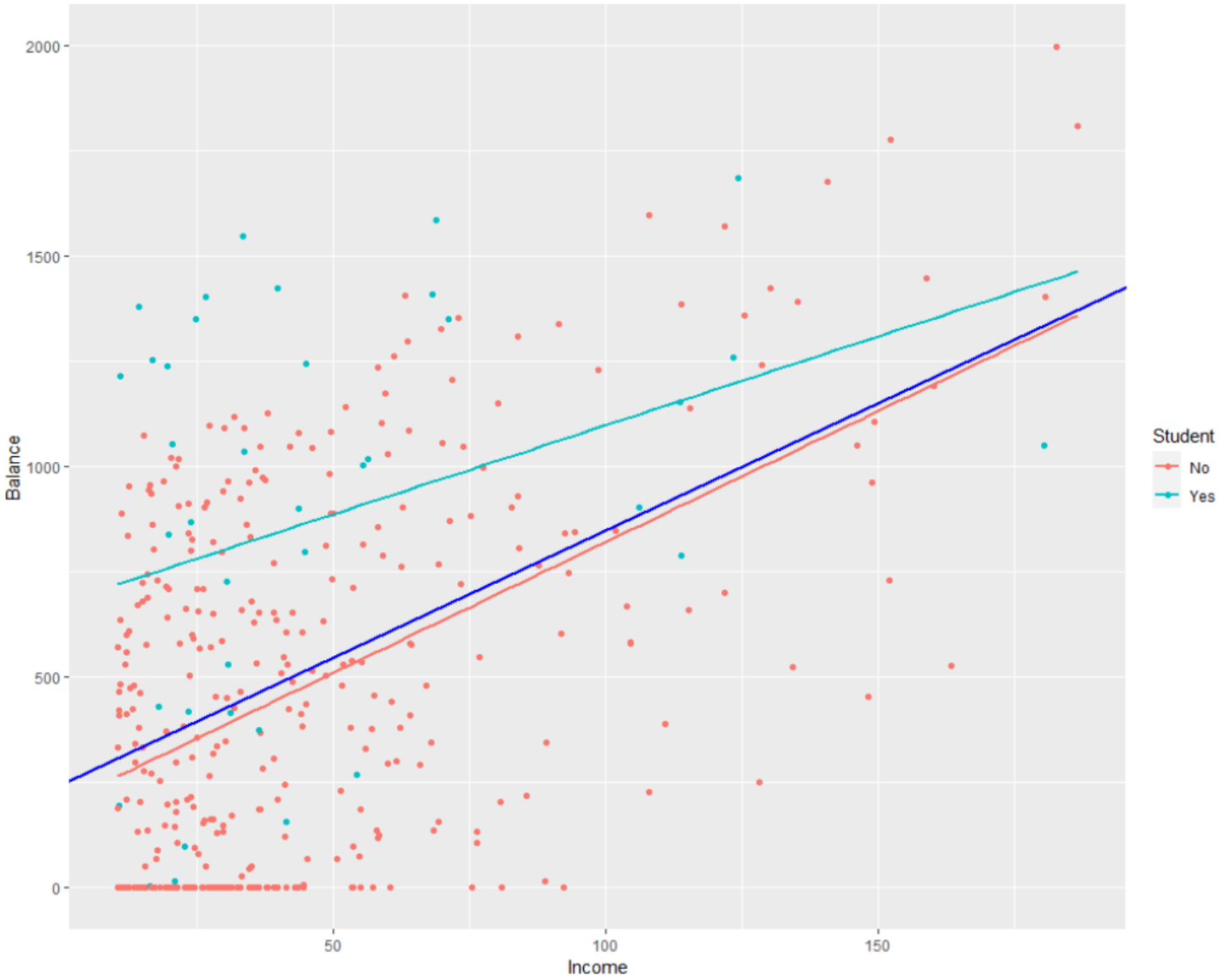


FIGURE 3.7. *For the* `Credit` *data, the least squares lines are shown for prediction of* `balance` *from* `income` *for students and non-students.* Left: *The model (3.34) was fit. There is no interaction between* `income` *and* `student`. Right: *The model (3.35) was fit. There is an interaction term between* `income` *and* `student`.


2000
1500
1000
500
0
Balance
50
100
150
Income
Student
No
Yes


2000
1500
1000
500
0
Balance
50
100
150
Income
Student
No
Yes


2000
1500
1000
500
0
Balance
50
100
150
Income
Student
No
Yes


2000
1500
1000
500
0
Balance
50
100
150
Income
Student
No
Yes

Which model is better?

Model	Int.Y	S.Y	Int.N	S.N	$\hat{\sigma}$	n.parameters
M1	246.51	6.05	246.51	6.05	407.86	3
M2-Y	667.30	4.22	NA	NA	468.27	3
M2-N	NA	NA	200.62	6.22	382.59	3
M3	593.81	5.98	211.14	5.98	391.79	4
M4	250.33	9.70	250.33	5.53	401.30	4
M5	677.30	4.22	200.62	6.22	391.62	5

Q1: What are differences between (M2-Y, M2-N) and M5?

Q2: Is M5 better?

Q3: Can we use technique of hypothesis testing between M3 and M5?

Short summary for interaction terms

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

- If the interaction between X_1 and X_2 seems important, then we should include both X_1 and X_2 in the model even if their coefficient estimates have large *p*-values.

3.3.6 Collinearity

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

- The two predictors limit and age appear to have no obvious relationship.
- The predictors limit and rating are very highly correlated with each other, and we say that they are collinear.

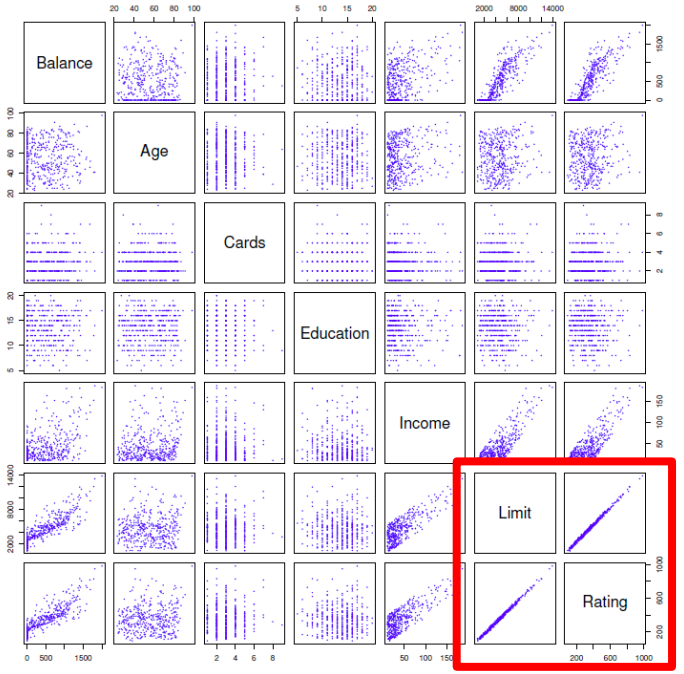


FIGURE 3.6. *The* Credit *data set contains information about* balance, age, cards, education, income, limit, *and* rating *for a number of potential customers.*

3.3.6 Collinearity

Since limit and rating tend to increase or decrease together, it can be difficult to determine how each one separately is associated with balance.

		Coefficient	Std. error	t-statistic	p-value
Model 1	Intercept	−173.411	43.828	−3.957	< 0.0001
	age	−2.292	0.672	−3.407	0.0007
	limit	0.173	0.005	34.496	< 0.0001
Model 2	Intercept	−377.537	45.254	−8.343	< 0.0001
	rating	2.202	0.952	2.312	0.0213
	limit	0.025	0.064	0.384	0.7012

TABLE 3.11. *The results for two multiple regression models involving the* Credit *data set are shown. Model 1 is a regression of* balance *on* age *and* limit*, and Model 2 a regression of* balance *on* rating *and* limit*. The standard error of* $\hat{\beta}_{\text{limit}}$ *increases 12-fold in the second regression, due to collinearity.*

Q: Supported by the figure???

3.3.6 Collinearity

$$Y = \beta_0 + \beta_1 rating + \beta_2 limit + \varepsilon$$

$$= \beta_0 + \beta_1 rating + \beta_2(\gamma_0 + \gamma_1 rating) + \varepsilon \quad (\because \text{collinearity})$$

$$= (\beta_0 + \beta_2\gamma_0) + (\beta_1 + \beta_2\gamma_1) rating + \varepsilon$$

		Coefficient	Std. error	t-statistic	p-value
Model 1	Intercept	−173.411	43.828	−3.957	< 0.0001
	age	−2.292	0.672	−3.407	0.0007
	limit	0.173	0.005	34.496	< 0.0001
Model 2	Intercept	−377.537	45.254	−8.343	< 0.0001
	rating	2.202	0.952	2.312	0.0213
	limit	0.025	0.064	0.384	0.7012

TABLE 3.11. *The results for two multiple regression models involving the* `Credit` *data set are shown. Model 1 is a regression of* `balance` *on* `age` *and* `limit`*, and Model 2 a regression of* `balance` *on* `rating` *and* `limit`*. The standard error of* $\hat{\beta}_{\text{limit}}$ *increases 12-fold in the second regression, due to collinearity.*

Q: Supported by the figure???

3.3.6 Collinearity

- The importance of the limit variable has been masked due to the presence of collinearity.
- To avoid such a situation, it is desirable to identify and address potential collinearity problems while fitting the model.
- A better way to assess multi-collinearity is to compute the *variance inflation factor (VIF).*

Detecting multicollinearity

Use variance inflation factors (VIFs) to examine the possible multicollinearity.

$$\text{VIF}(\hat{\beta}_k) = \frac{1}{1 - R_k^2},$$

where R_k^2 is the R^2 from the regression model

$$x_k = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_{k-1} x_{k-1} + \alpha_{k+1} x_{k+1} + \cdots + \epsilon.$$

General rule of thumb:

- VIF > 4: further investigation.
- VIF > 10: serious multicollinearity requiring correction.
- In the Credit data, a regression of balance on age, rating, and limit indicates that the predictors have VIF values of 1.01, 160.67, and 160.59. Collinearity in the data!

Reading for midterm exam

3.4 The Marketing Plan

3.6 Lab: Linear Regression