STATISTICAL LEARNING

CHAPTER 3: LINEAR REGRESSION

INSTRUCTOR: SEOKHO LEE

HANKUK UNIVERSITY OF FOREIGN STUDIES

2015 SPRING

Linear Regression

- **Linear regression** is a very simple approach for supervised learning
  - Linear regression is a useful tool for predicting a quantitative response
  - It is still important although we will see many fancy statistical learning approaches
- Some questions on linear regression for Advertising data
  - Is there a relationship between advertising budget and sales?
  - How strong is the relationship between advertising budget and sales?
  - Which media contribute to sales?
  - How accurately can we estimate the effect of each medium on sales?
  - How accurately can we predict future sales?
  - Is the relationship linear?
  - Is there synergy among the advertising media?

Simple Linear Regression

- **Simple linear regression** is a very straightforward approach for predicting a quantitative response Y on the basis of a single predictor variable X with a linear relationship assumption

$$Y \approx \beta_0 + \beta_1 X \tag{3.1}$$

  - We say that we are **regressing** Y **on** X
  - Unknown constants β_0 and β_1 represent the **intercept** and the **slope** terms in the linear model. They are called model **coefficients** or **parameters**
  - $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimates of β_0 and β_1 from the training data
  - We can predict future response on the basis of a particular value of x by computing

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{3.2}$$

Estimating the Coefficients

- Training data: $(x_1, y_1),\ (x_2, y_2),\ \ldots\ ,\ (x_n, y_n)$
- Our goal is to obtain coefficient estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ such that the linear model (3.1) fits the available data well, that is $y_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i$ for $i = 1, 2, \ldots, n$.
- Measure of **closeness** : minimizing the **least squares** criterion
  - $e_i = y_i - \hat{y}_i$: **residuals**
  - **Residual sum of squares** (RSS):

$$\begin{aligned} \text{RSS} &= e_1^2 + e_2^2 + \ldots + e_n^2 \\ &= (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2 \end{aligned} \tag{3.3}$$

- Least square estimates

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \end{aligned} \tag{3.4}$$

with $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$

Estimating the Coefficients

- Advertising data example
  - Y-sales, X-TV
  - model: sales $\approx \beta_0 + \beta_1 \times$ TV
  - $\hat{\beta}_0 = 7.03$ and $\hat{\beta}_1 = 0.0475$
  - According to this approximation, an additional \$1,000 spent on TV advertising is associated with selling approximately 47.5 additional units of the product (See Figure 3.1 and 3.2)

Estimating the Coefficients

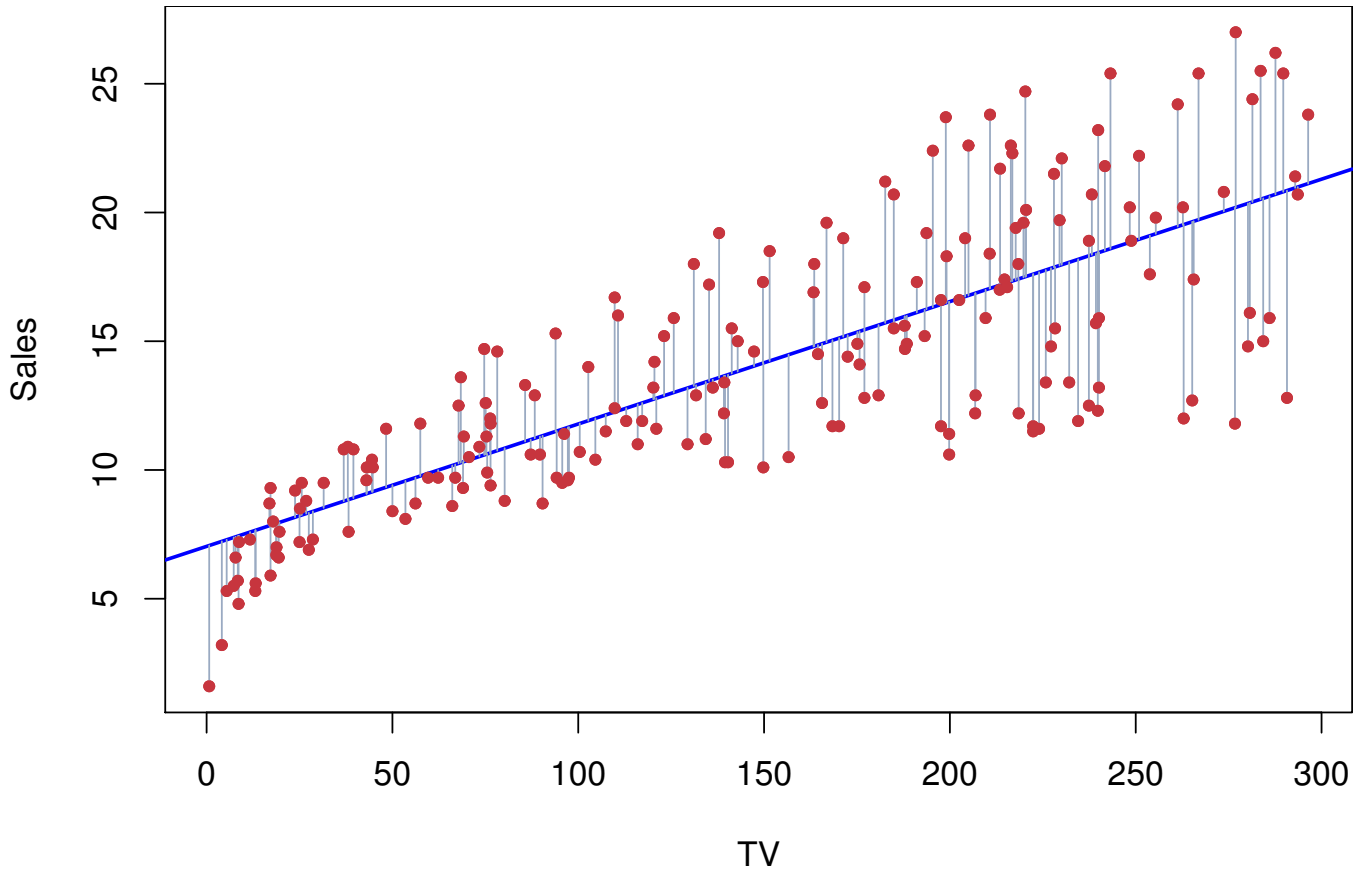

Figure 3.1: For the Advertising data, the least squares fit for the regression of sales onto TV is shown. The fit is found by minimizing the sum of squared errors. Each grey line segment represents an error, and the fit makes a compromise by averaging their squares. In this case a linear fit captures the essence of the relationship although it is somewhat deficient in the left of the plot.

Estimating the Coefficients

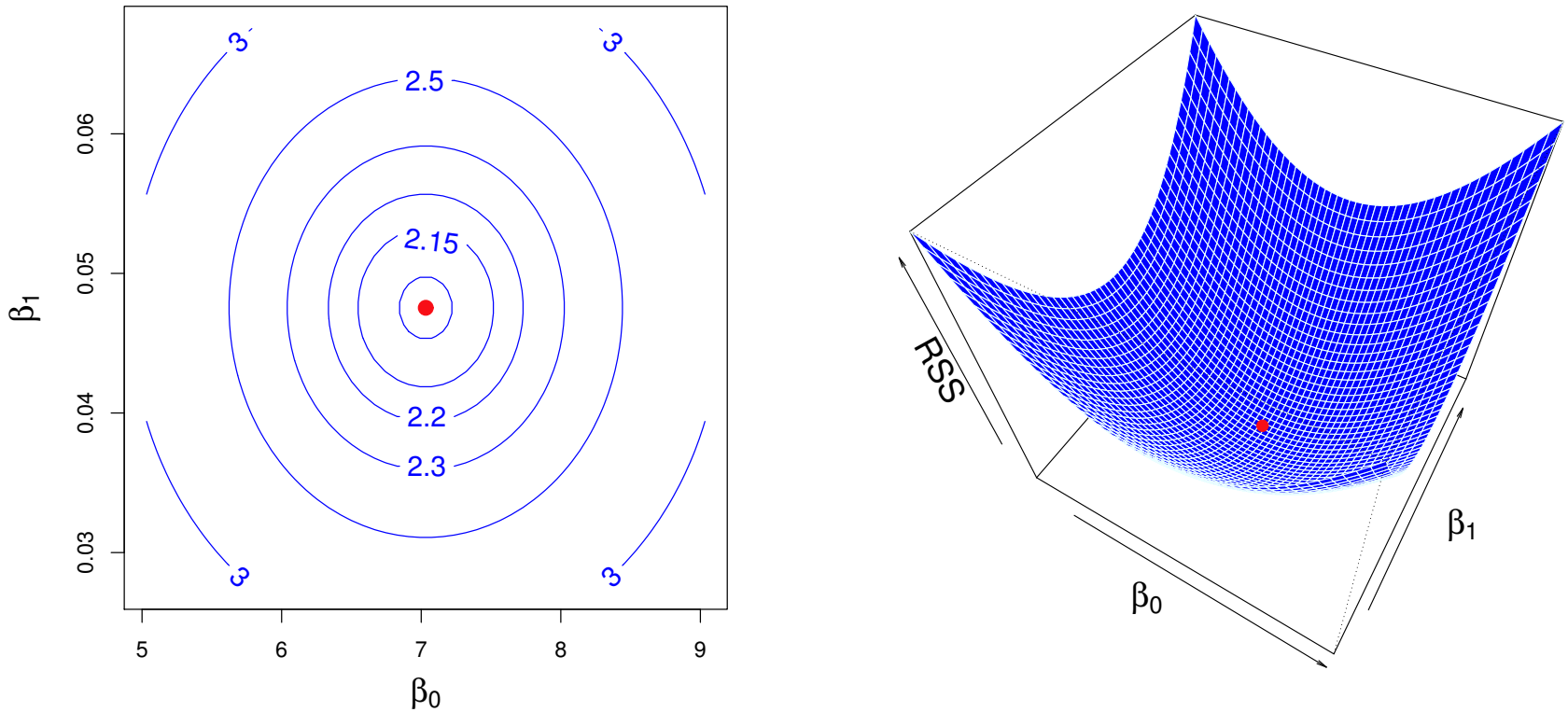

Figure 3.2: Contour and three-dimensional plots of the RSS on the Advertising data, using sales as the response and TV as the predictor. The red dots correspond to the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, given by (3.4).

Assessing the Accuracy of the Coefficient Estimates

- **Population regression line**

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{3.5}$$

  - Population regression line is the best linear approximation to the true relationship between X and Y, say $Y = f(X) + \epsilon$

- We need to assess bias and standard error of estimate
  - **Standard error** of estimates:

$$\text{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right], \quad \text{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3.6}$$

  - **Residual standard error**: estimate of σ^2

$$\text{RSE} = \sqrt{\text{RSS}/(n-2)}$$

Assessing the Accuracy of the Coefficient Estimates

- 95% confidence interval for β_1:

$$\hat{\beta}_1 \pm 2 \cdot \text{SE}(\hat{\beta}_1) \tag{3.7}$$

  - This means that there is approximately a 95% chance that the interval

$$\left[\hat{\beta}_1 - 2 \cdot \text{SE}(\hat{\beta}_1), \hat{\beta}_1 + 2 \cdot \text{SE}(\hat{\beta}_1)\right] \tag{3.8}$$

- 95% confidence interval for β_0:

$$\hat{\beta}_0 \pm 2 \cdot \text{SE}(\hat{\beta}_0) \tag{3.9}$$

- In Advertising data,
  - 95% confidence intervals for β_0 and β_1 are [6.130, 7.935] and [0.042, 0.053]
  - In the absence of any advertising, sales will, on average, fall somewhere between 6,130 and 7,940 unites
  - For each \$1,000 increase in television advertising, there will be an average increase in sales of between 42 and 53 units

Assessing the Accuracy of the Coefficient Estimates

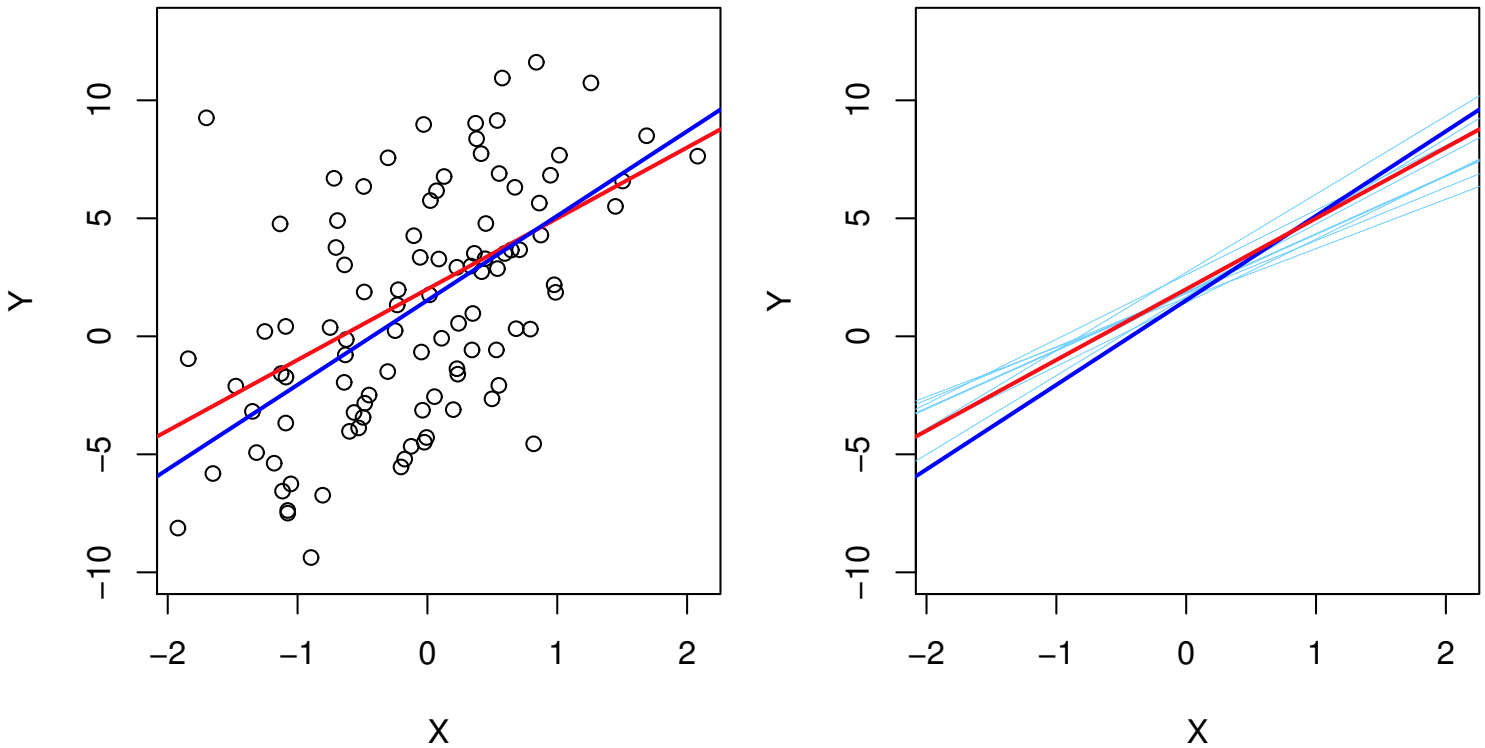

Figure 3.3: A simulated data set. Left: The red line represents the true relationship, $f(X) = 2 + 3X$, which is known as the population regression line. The blue line is the least squares line; it is the least squares estimate for $f(X)$ based on the observed data, shown in black. Right: The population 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.

Assessing the Accuracy of the Coefficient Estimates

- Hypothesis test
  - **Null hypothesis**

$$H_0 \ : \ \text{There is no relationship between } X \text{ and } Y \tag{3.10}$$

  - **Alternative hypothesis**

$$H_a \ : \ \text{There is some relationship between } X \text{ and } Y \tag{3.11}$$

  - Mathematically, this is corresponds to testing

$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_a : \beta_1 \neq 0$$

  - If $\beta_1 = 0$ then the model (3.5) reduces to $Y = \beta_0 + \epsilon$, and X is not associated with Y
  - ***t*-statistic**

$$t = \frac{\hat{\beta}_1 - 0}{\text{SE}(\hat{\beta}_1)} \ \sim \ t(n-2) \tag{3.12}$$

Assessing the Accuracy of the Coefficient Estimates

	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).

Assessing the Accuracy of the Model

- **Residual standard error** (RSE)

$$\text{RSE} = \sqrt{\frac{1}{n-2}\text{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.13}$$

- R^2 **Statistic**

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} \tag{3.14}$$

where $\text{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the **total sum of squares**, and $\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the **residual sum of squares**

  - $R^2 = r^2$ where r is the sample correlation coefficient between X and Y

Multiple Linear Regression

- Multiple linear regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon \tag{3.15}$$

- Advertising data:

$$\text{sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times \text{newspaper} + \epsilon \tag{3.16}$$

Estimating the Regression Coefficients

- Prediction formula:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p \tag{3.17}$$

- We choose β_0, β_1, ..., β_p to minimize the sum of squared residuals

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 && (3.18) \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2 && (3.19) \end{aligned}$$

  - Figure 3.4 illustrates an example of the least squares fit to a toy data set with $p = 2$ predictors

Estimating the Regression Coefficients

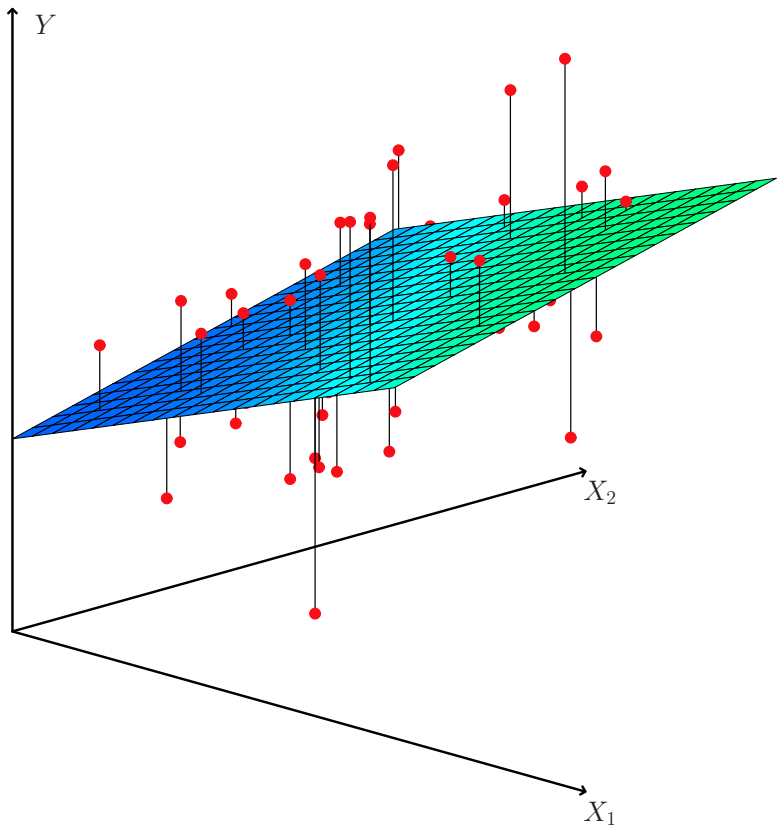

Figure 3.4: In a three-dimensional setting, with two predictors and one response, the least squares regression line becomes a plane. The plane is chosen to minimize the sum of the squared vertical distances between each observation (shown in red) and the plane.

Estimating the Regression Coefficients

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	<0.0001

Table 3.2: 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).

Estimating the Regression Coefficients

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

- Advertising data
  - Coefficients for newspaper from simple linear regression and multiple linear regression are different

Estimating the Regression Coefficients

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

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

Some Important Questions

- A few important questions from multiple linear regression:
  1. Is at least one of the predictors X_1, X_2, ..., X_p useful in predicting the response?
  2. Do all the predictors help to explain Y, or is only a subset of the predictors useful?
  3. How well does the model fit the data?
  4. Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

One: Is There a Relationship Between the Response and Predictors?

- Null and alternative hypotheses

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

versus

$$H_a \ : \ \text{at least one } \beta_j \text{ is non-zero}$$

- ***F*-statistic**

$$F = \frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(n-p-1)} \ \sim \ F(p, n-p-1) \qquad (3.20)$$

One: Is There a Relationship Between the Response and Predictors?

- Test that a particular subset of q of the coefficients are zero
  - Null hypothesis

$$H_0 \ : \ \beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0$$

  - F-**statistic**

$$F = \frac{(\text{RSS}_0 - \text{RSS})/q}{\text{RSS}/(n-p-1)} \ \sim \ F(q, n-p-1) \qquad (3.21)$$

    where RSS_0 is the residual sum of squares for the model under H_0

Two: Deciding on Important Variables

- Related to **variable selection**, which will be discussed in Chapter 6
- All possible models
  - Fit 2^p models and find the best
  - Measures of model fit:
    Mallow's C_p, **Akaike Information criterion** (AIC), **Bayesian Information criterion** (BIC), **adjusted R^2**, etc.
  - Too many models we need to fit: even for moderate p, trying out every possible subset of the predictors is infeasible.
    For instance, $p = 30$, then we must consider $2^{30} = 1,073,741,824$!
- Stepwise approach
  - **Forward selection**
  - **Backward selection**
  - **Mixed selection** (or **Stepwise selection**)

Three: Model Fit

- Two of the most common numerical measures of model fit are the R^2 and RSE
  - RSE for multiple linear regression

$$\text{RSE} = \sqrt{\frac{1}{n-p-1}\text{RSS}} \tag{3.22}$$

- In addition to looking at the RSE and R^2, it is useful to plot the data

Three: Model Fit

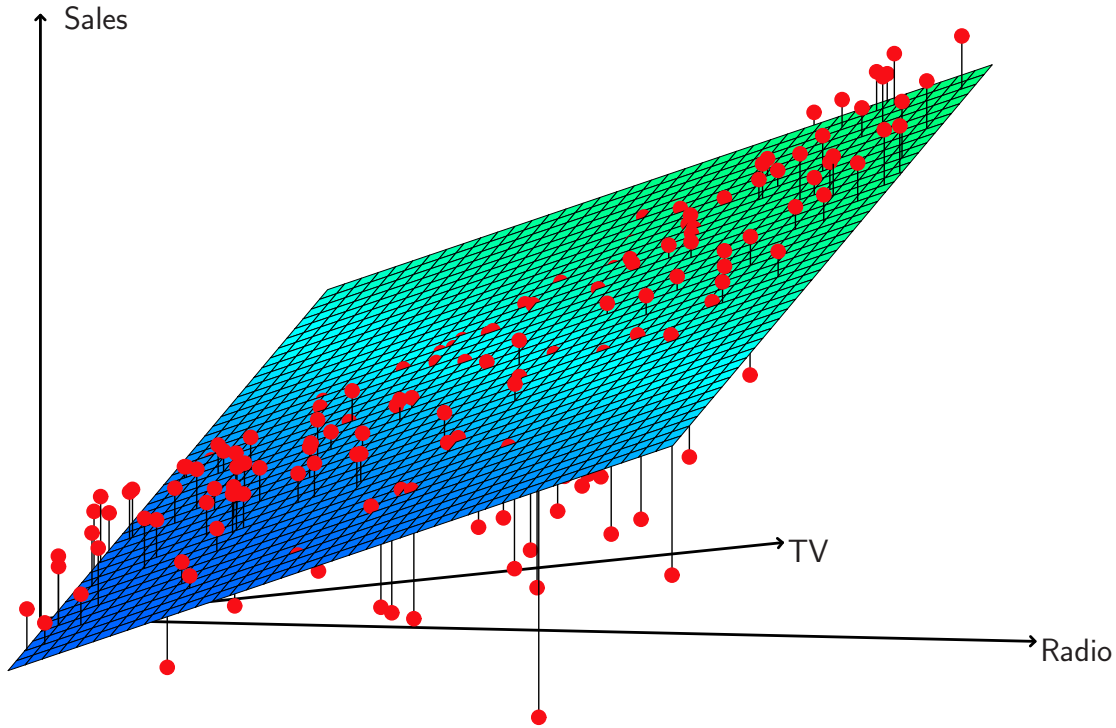

Figure 3.5: For the Advertising data, a linear regression fit to sales using TV and radio as predictors. From the pattern of the residuals, we can see that there is a pronounced non-linear relationship in the data.

Four: Predictions

- For prediction, it is straightforward to apply (3.17)
- Three sorts of uncertainty associated with this prediction
  1. **Reducible error**:
     It comes from the accuracy of estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ for $\beta_0, \beta_1, \ldots, \beta_p$
  2. **Model bias**:
     Assuming a linear model for $f(X)$ is almost always an approximation of reality, so there is an additional source of potentially reducible error
  3. **Irreducible error**:
     Even if we know the true values for $\beta_0, \beta_1, \ldots, \beta_p$, the response value cannot be predicted perfectly because of the random error ϵ in the model (3.15)

Qualitative Predictors

- In practice, often some predictors are **qualitative** (also known as **factor** variables)
- Credit data set
  - Response: balance (average credit card debt for a number of individuals)
  - Quantitative predictors: age, cards (number of credit cards), education (years of education), income (in thousands of dollars), limit (credit limit), and rating (credit rating)
  - Qualitative predictors: gender, student (student status), status (marital status), and ethnicity (Caucasian, African American or Asian)

Qualitative Predictors

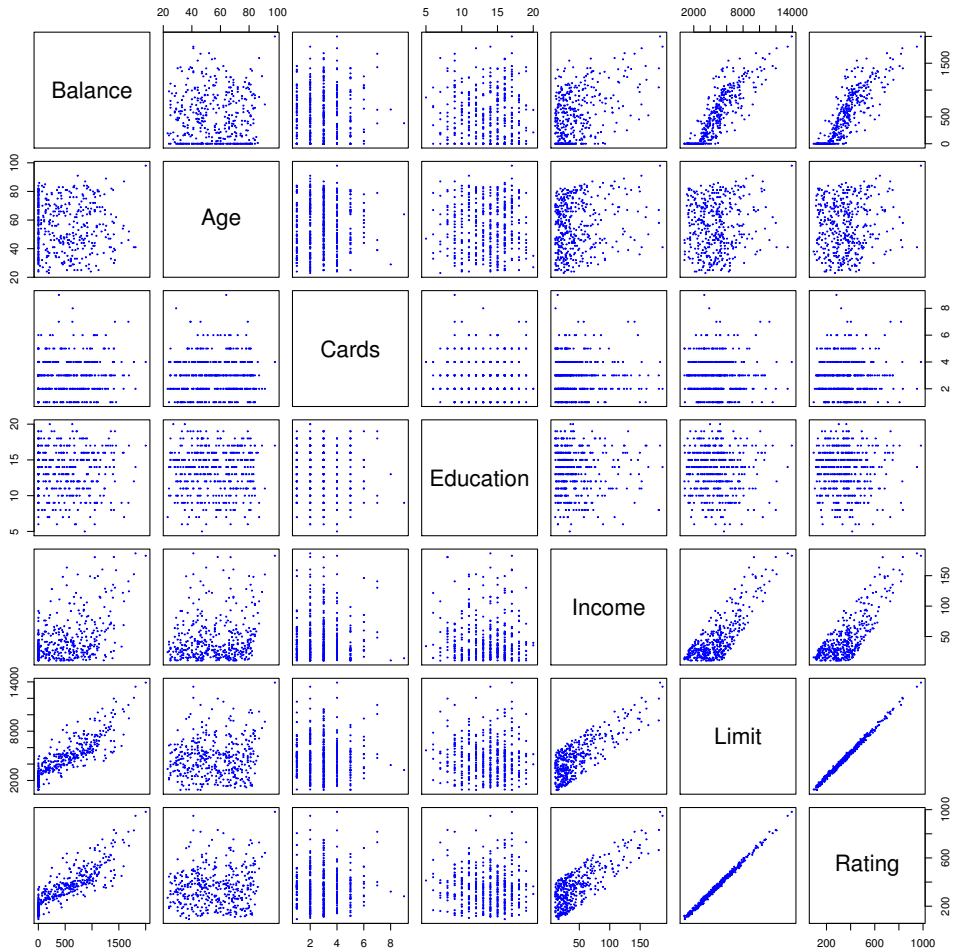

Figure 3.6: The Credit data set contains information about balance, age, cards, education, income, limit, and rating for a number of potential customers.

Predictors with Only Two Levels

- Suppose we wish to investigate difference in credit card balance between males and females, ignoring the other variables for the moment
- **Dummy variable** for gender

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male} \end{cases} \tag{3.23}$$

- Linear model using (3.23)

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male} \end{cases} \tag{3.24}$$

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

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

Table 3.5: Least squares coefficient estimates associated with the regression of balance onto gender in the Credit data set. The linear model is given in (3.24). That is, gender is encoded as a dummy variable, as in (3.23)

- We may use a difference coding system, which has no effect on the regression fit
  - $x_i = 1$ for male and 0 for female
  - $x_i = 1$ for female and -1 for male
  - Try to imagine what results will be given under the above coding fashions

Qualitative Predictors with More than Two Levels

- Consider ethnicity consisting of three levels. We need to create two dummy variables
  - The first variable

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is Asian} \\ 0 & \text{if } i\text{th person is not Asian} \end{cases} \tag{3.25}$$

  - The second variable

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is Caucasian} \\ 0 & \text{if } i\text{th person is not Caucasian} \end{cases} \tag{3.26}$$

  - Linear model will be

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is African American} \end{cases} \tag{3.27}$$

  - β_0: average credit card balance for African American
  - β_1: difference in the average balance between the Asian and African American categories
  - β_2: difference in the average balance between the Caucasian and African American categories
  - African American category, with no dummy variable, is known as the **baseline**

Qualitative Predictors with More than Two Levels

	Coefficient	Std. error	t-statistic	p-value
Intercept	531.00	46.32	11.464	<0.0001
ethnicity[Asian]	-18.69	65.02	-0.287	0.7740
ethnicity[Caucasian]	-12.50	56.68	-0.221	0.8260

Table 3.6: Least square coefficient estimates associated with the regression of balance onto ethnicity in the Credit data set. The linear model is given in (3.27). That is, ethnicity is encoded via two dummy variables (3.25) and (3.26).

Extension of the Linear Model

- The standard linear regression model (3.15) provides interpretable results and works quite well on many real-world problems
- It makes several highly restrictive assumptions that are often violated in practice
- Two of the most important assumptions:
  - **Additivity**: 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
  - **Linearity**: 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
- We examine a number of sophisticated methods that relax these two assumptions

Removing the Additive Assumption

- Advertising data (See Figure 3.5):

  "The linear model (3.16) 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. However, this simple model may be incorrect. Suppose that spending money on radio advertising actually increasing 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 of TV may increase sales more than allocating the entire amount to either TV or to radio."

- This phenomena is known as **synergy** effect in marketing, and **interaction** effect in statistics

- Model with interaction term

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon \tag{3.28}$$

- Another form

$$\begin{aligned} Y &= \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon \\ &= \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \epsilon \end{aligned} \tag{3.29}$$

  - Since $\tilde{\beta}_2 = \beta_2 + \beta_3 X_2$ changes with X_2, the effect of X_1 on Y is no longer constant

Removing the Additive Assumption

	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 3.7: For the Advertising data, least squares coefficient estimates associated with the regression of sales onto TV and radio, with an interaction term, as in (3.30).

- Advertising data fit with interaction

$$\begin{aligned} \text{sales} &= \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times (\text{TV} \times \text{radio}) + \epsilon \\ &= \beta_0 + (\beta_1 + \beta_3 \times \text{radio}) \times \text{TV} + \beta_2 \times \text{radio} + \epsilon \end{aligned} \quad (3.30)$$

- **Hierarchical principal**

  The hierarchical principal 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

Removing the Additive Assumption

- Credit data without interaction

$$\begin{aligned} \text{balance}_i &\approx \beta_0 + \beta_1 \times \text{income}_i + \begin{cases} \beta_2 & \text{if } i\text{th person is a student} \\ 0 & \text{if } i\text{th person is not a student} \end{cases} \\ &= \beta_1 \times \text{income}_i + \begin{cases} \beta_0 + \beta_2 & \text{if } i\text{th person is a student} \\ \beta_0 & \text{if } i\text{th person is not a student} \end{cases} \end{aligned} \tag{3.31}$$

- Credit data with interaction

$$\begin{aligned} \text{balance}_i &\approx \beta_0 + \beta_1 \times \text{income}_i + \begin{cases} \beta_2 + \beta_3 \times \text{income}_i & \text{if student} \\ 0 & \text{if not student} \end{cases} \\ &= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times \text{income}_i & \text{if student} \\ \beta_0 + \beta_1 \times \text{income}_i & \text{if not student} \end{cases} \end{aligned} \tag{3.32}$$

Removing the Additive Assumption

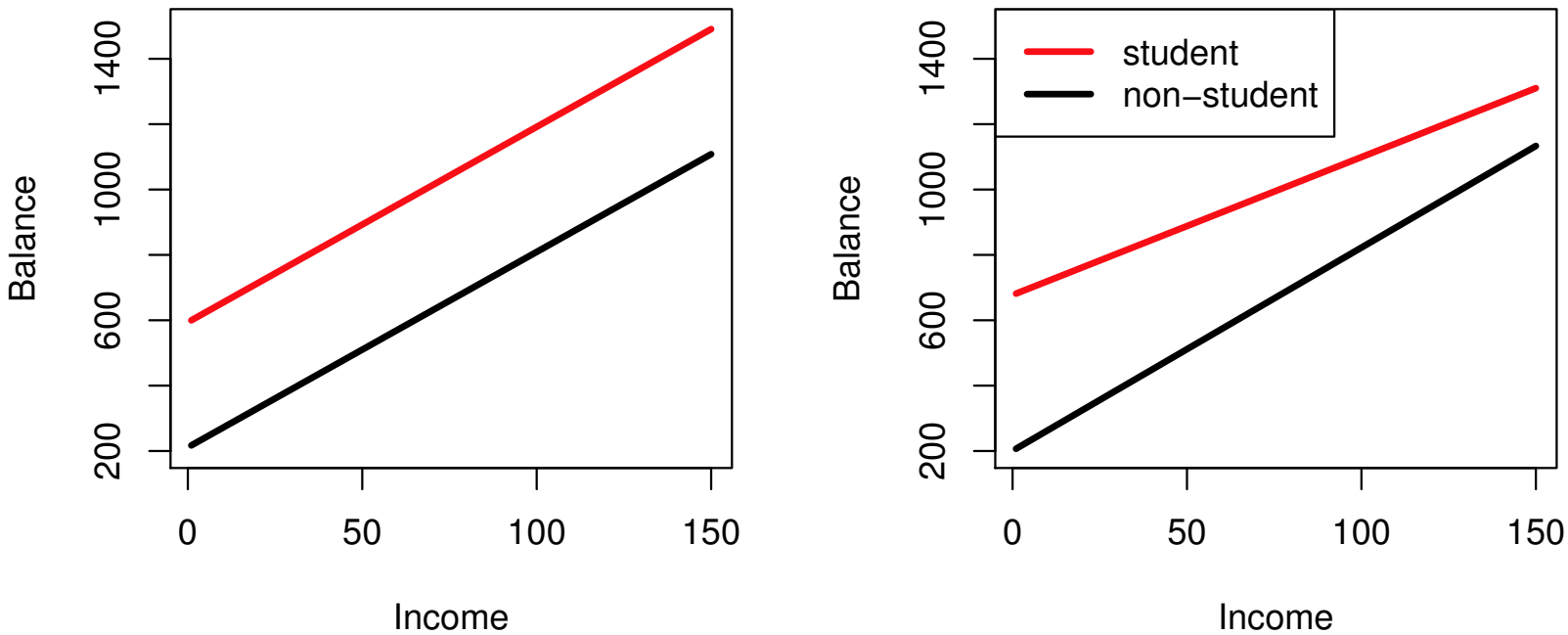

Figure 3.7: For the Credit data, the least squares lines are shown for prediction of balance from income for students and non-students. Left: The model (3.31) was fit. There is no interaction between income and student. Right: The model (3.32) was fit. There is an interaction term between income and student.

Non-linear Relationships

- A simple classical way to directly extend the linear model to accommodate non-linear relationships is to use **polynomial regression**
- Auto data set (See Figure 3.8)
  - Response: mpg (gas mileage in miles per gallon)
  - Predictor: horsepower
  - Linear model

$$\text{mpg} = \beta_0 + \beta_1 \times \text{horsepower} + \epsilon$$

  - Quadratic model

$$\text{mpg} = \beta_0 + \beta_1 \times \text{horsepower} + \beta_2 \times \text{horsepower}^2 + \epsilon \qquad (3.33)$$

	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

Table 3.8: For the Auto data set, least squares coefficient estimates associated with the regression of mpg onto horsepower and horsepower^2.

Non-linear Relationships

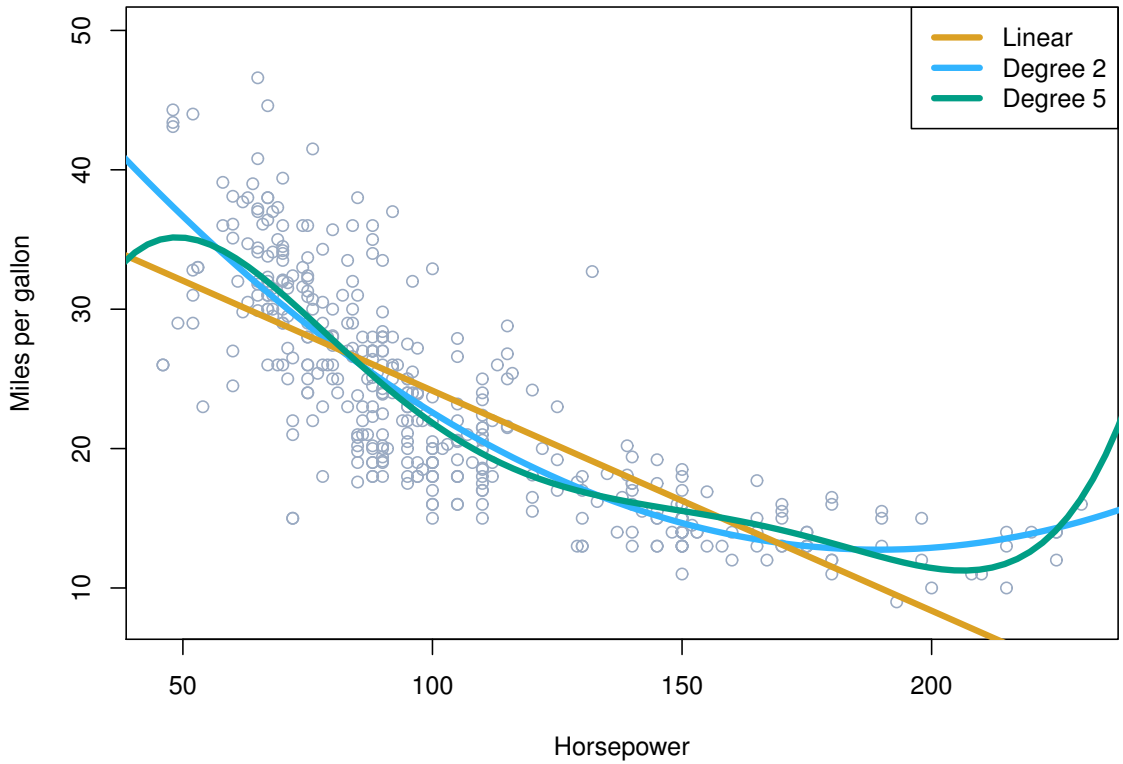

Figure 3.8: The Auto data set. For a number of cars, mpg and horsepower are shown. The linear regression fit is shown in orange. The linear regression fit for a model that includes $horsepower^2$ is shown as a blue curve. The linear regression fit of a model that includes all polynomials of horsepower up to fifth-degree is shown in green.

Potential Problems

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. Collinearity

1. Non-linearity of the Data

- **Residual pots** are a useful graphical tool for identifying non-linearity

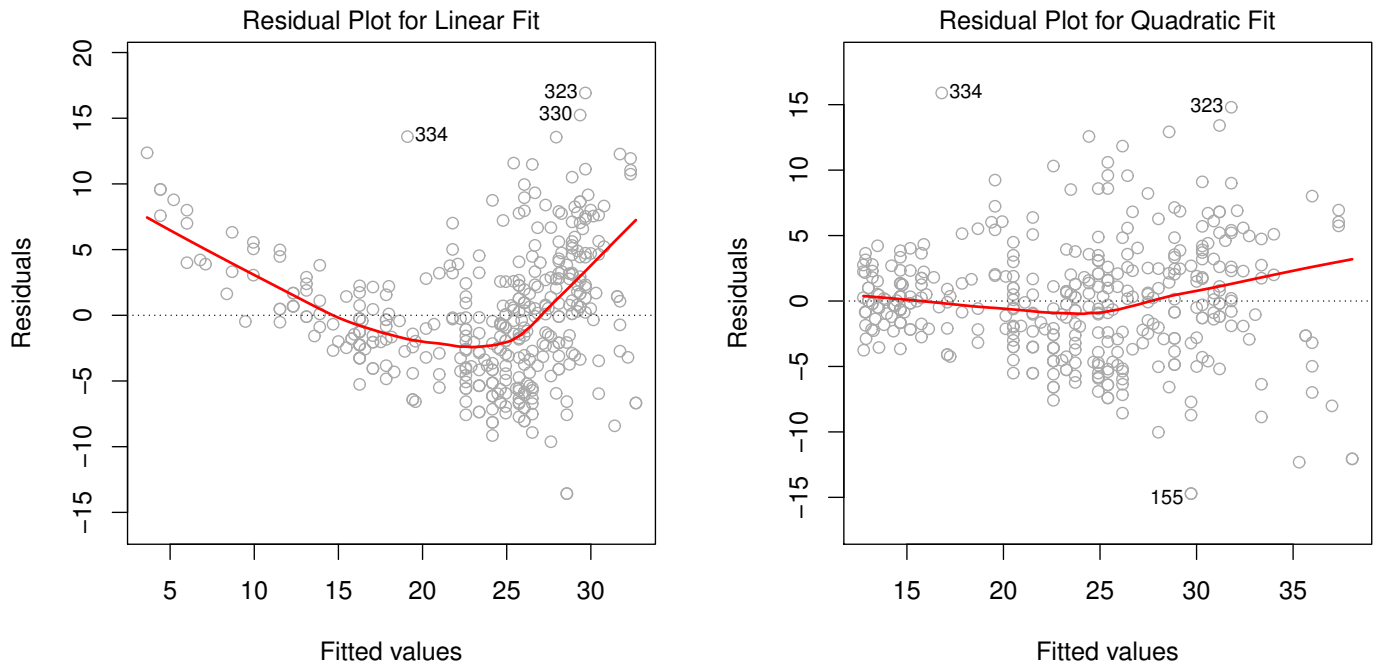

Figure 3.9: Plots of residuals versus predicted (or fitted) values for the Auto data set. In each plot, the red line is a smooth fit to the residuals, intended to make it easier to identify a trend. Left: A linear regression of mpg on horsepower. A strong pattern in the residuals indicates non-linearity in the data. Right: A linear regression of mpg on horsepower and horsepower^2. There is little pattern in the residuals.

2. Correlation of Error Terms

- Most of **time series** data are subject to this subject

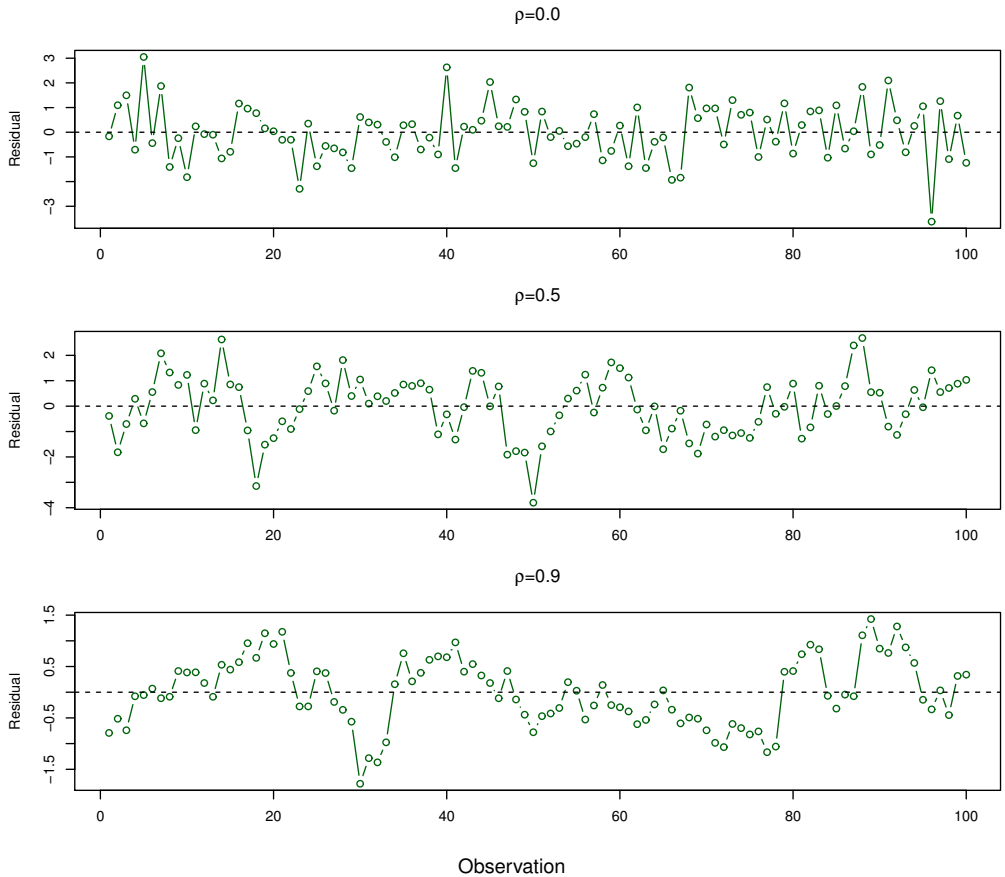

Figure 3.10: Plots of residuals from simulated time series data sets generated with differing levels of correlation ρ between error terms for adjacent time points.

3. Non-constant Variance of Error Terms

- **Heteroscedasticity** - transformation of response or **weighted least squares** are simple remedy

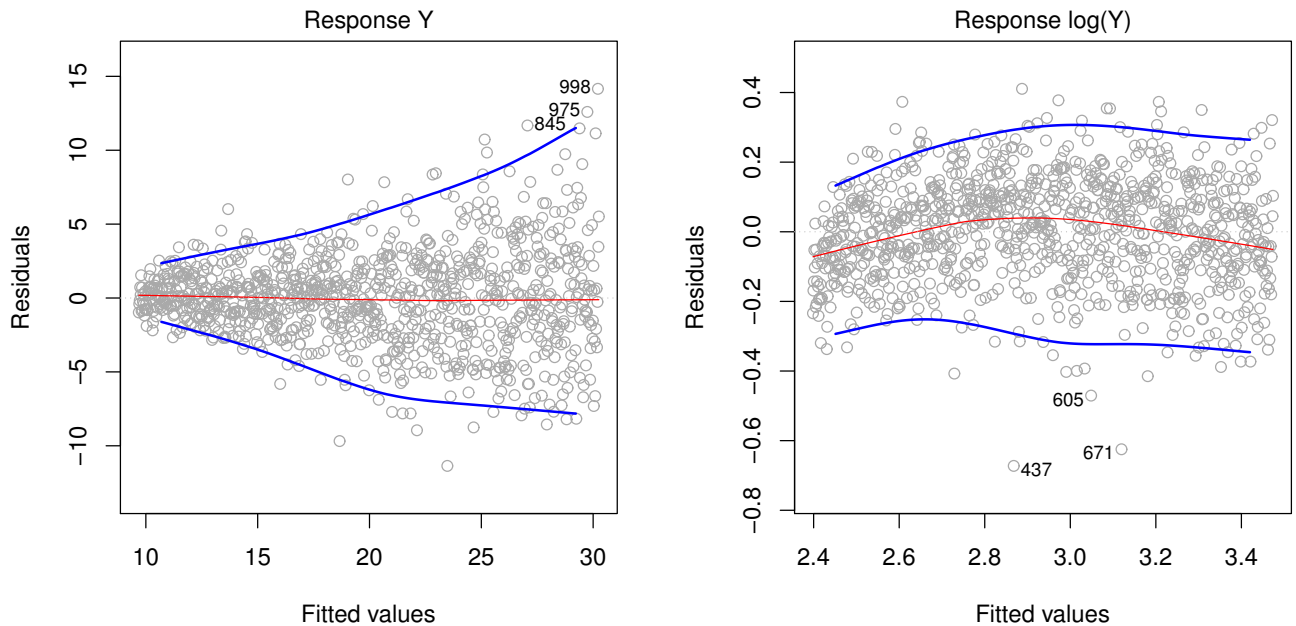

Figure 3.11: Residual plots. In each plot, the red line is a smooth fit to the residuals, intended to make it easier to identify a trend. The blue lines track the outer quantiles of the residuals, and emphasize patterns. Left: The funnel shape indicates heteroscedasticity. Right: The predictor has been log-transformed, and there is now no evidence of heteroscedasticity.

4. Outliers

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

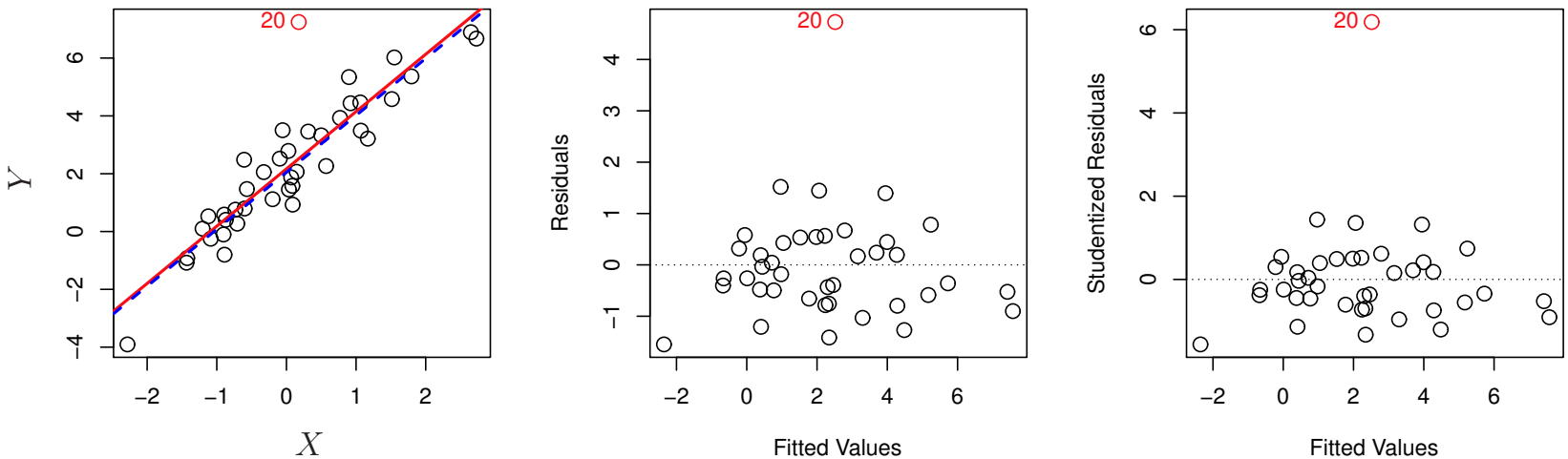

Figure 3.12: Left: The least squares regression line is shown in red, and the regression line after removing the outlier is shown in blue. Center: The residual plot clearly identifies the outlier. Right: The outlier has a studentized residual of 6; typically we expect values between -3 and 3.

5. High Leverage Points

- Observations with **high leverage** have an unusual vale for x_i
  - **Leverage statistic** is defined as

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2} \tag{3.34}$$

  - High leverage is claimed if h_i greatly exceeds $(p+1)/n$

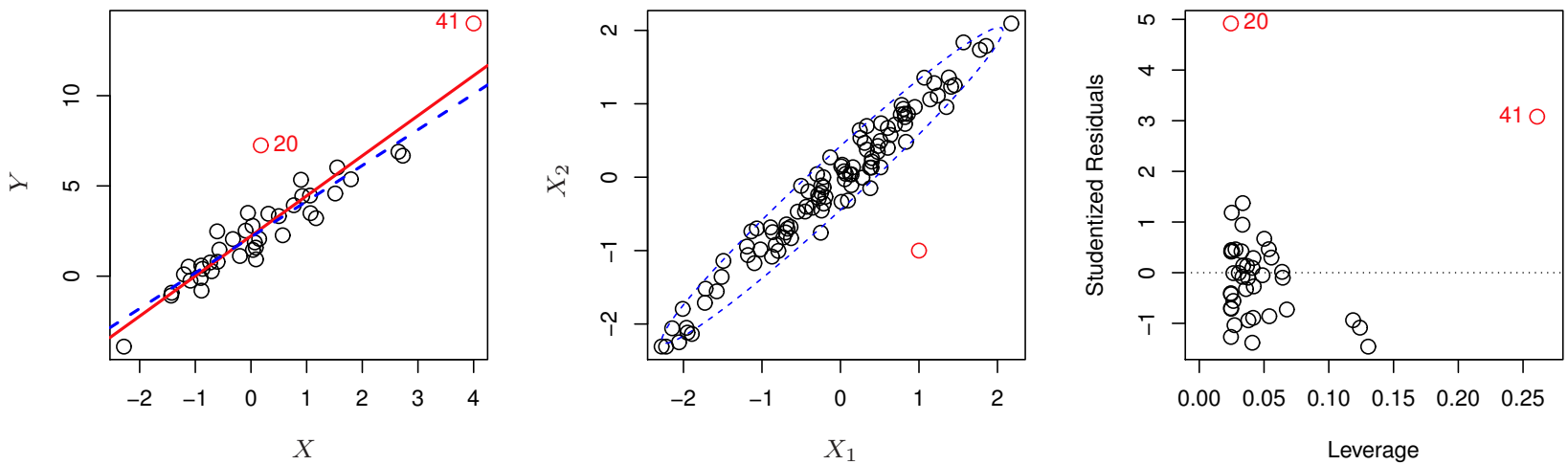

Figure 3.13: Left: Observation 41 is a high leverage point, while 20 is not. The red line is the fit to all the data, and the blue line is the fit with observation 41 removed. Center: The red observation is not unusual in terms of its X_1 value or its X_2 value, but still falls outside the bulk of the data, and hence has high leverage. Right: Observation 41 has a high leverage and a high residual.

6. Collinearity

- **Collinearity** refers to the situation in which two or more predictor variables are closely related to one another
  - Estimation is unstable under high collinearity, and impossible under perfect collinearity
  - The **power** of hypothesis test-the probability of correctly detecting a non-zero coefficient-is reduced by collinearity (See Table 3.11)
  - **Multicollinearity** refers to the situation where collinearity exits between three or more variables even if no pair of variables has a particularly high correlation

- Detecting collinearity
  - Look at the correlation matrix of the predictors
  - **Variance inflation factor** (VIF)

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

where $R^2_{X_j|X_{j-1}}$ is the R^2 from a regression of X_j onto all the other predictors. A VIF value that exceeds 5 or 10 indicates a problematic amount of collinearity

6. Collinearity

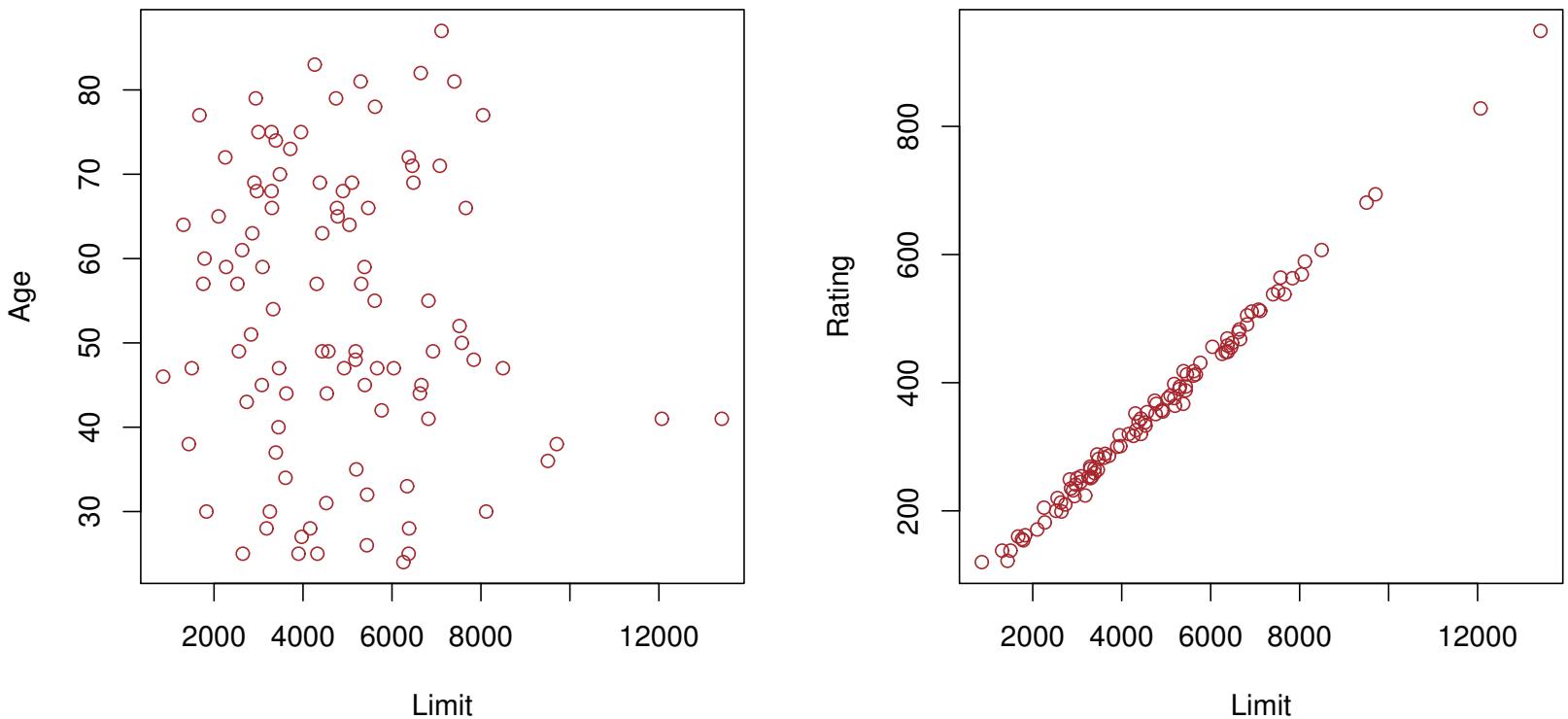

Figure 3.14: Scatterplots of the observations from the Credit data set. Left: A plot of age versus limit. These two variables are not collinear. Right: A plot of rating versus limit. There is high collinearity.

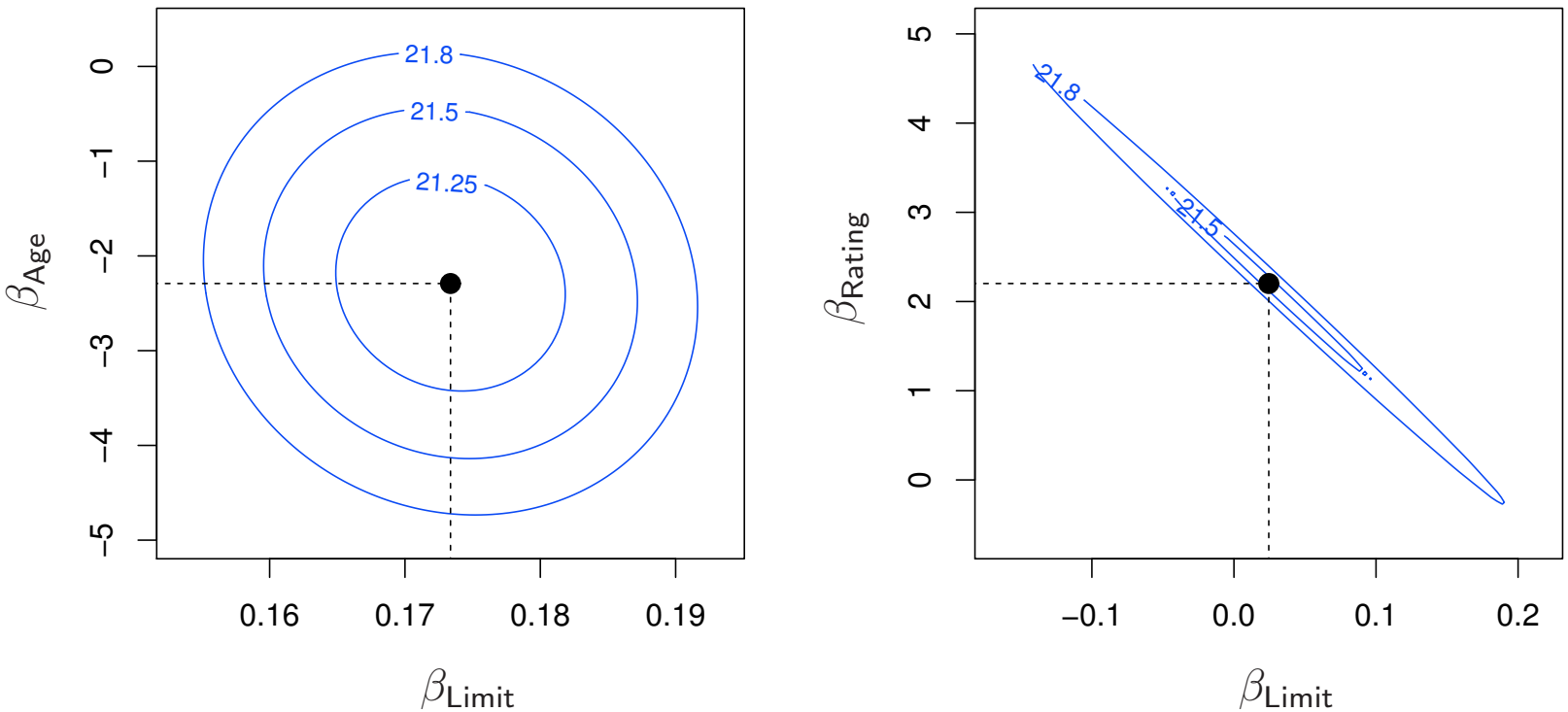

Figure 3.15: Contour plots for the RSS values as a function of the parameters β for various regression involving the Credit data set. In each plot, the black dots represent the coefficient values corresponding to the minimum RSS. Left: A contour plot of RSS for the regression of balance onto age and limit. The minimum value is well defined. Right: A contour plot of RSS for the regression of balance onto rating and limit. Because of the collinearity, there are many pairs $(\beta_{\text{Limit}}, \beta_{\text{Rating}})$ with a similar value for RSS.

		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 3.9: 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.

The Marketing Plan

Find the answer, form the textbook, for the seven questions about the Advertising data at the beginning of this chapter.

- Is there a relationship between advertising budget and sales?
- How strong is the relationship between advertising budget and sales?
- Which media contribute to sales?
- How accurately can we estimate the effect of each medium on sales?
- How accurately can we predict future sales?
- Is the relationship linear?
- Is there synergy among the advertising media?

Comparison of Linear Regression with K-Nearest Neighbors

- **K-nearest neighbors regression** (KNN regression) is one of the simplest and best-known non-parametric method for regression.
  - Prediction for x_0, we estimate $f(x_0)$ using the average of all the training responses in neighbors $\mathcal{N}_0$. In other words,

$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in \mathcal{N}_0} y_i$$

- The optimal value for K will depend on the **bias-variance trade-off**
  - A small value for K provides the most flexible fit, which will have low bias but high variance
  - Larger values of K provide a smoother and less variable fit. However, the smoothing may cause bias by masking some of the structure in $f(X)$

Comparison of Linear Regression with K-Nearest Neighbors

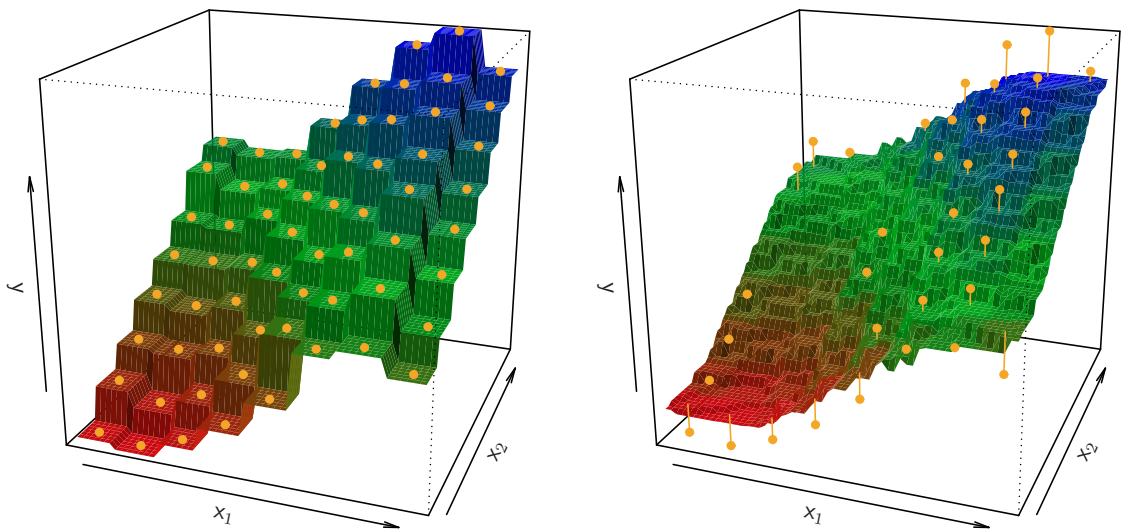

Figure 3.16: Plots of $\hat{f}(X)$ using KNN regression on a two-dimensional data set with 64 observations (orange dots). Left: $K = 1$ results in a rough step function fit. Right: $K = 9$ produces a much smoother fit.

Comparison of Linear Regression with K-Nearest Neighbors

- In what setting will a parametric approach (such as least squares linear regression) outperform a non-parametric approach (such as KNN regression)?
- The parametric approach will outperform the non-parametric approach if the parametric form that has been selected is close to the true form of f
  - See Figures 3.17 and 3.18 for the case that the true relationship is linear
  - See Figure 3.19 for the case that the true relationship is non-linear
- The parametric methods will tend to outperform non-parametric approaches when there is a small number of observations per predictor
  - **Curse of dimensionality**: With increasing p, a given observation has no *nearby neighbors*
  - See Figure 3.20

Comparison of Linear Regression with K-Nearest Neighbors

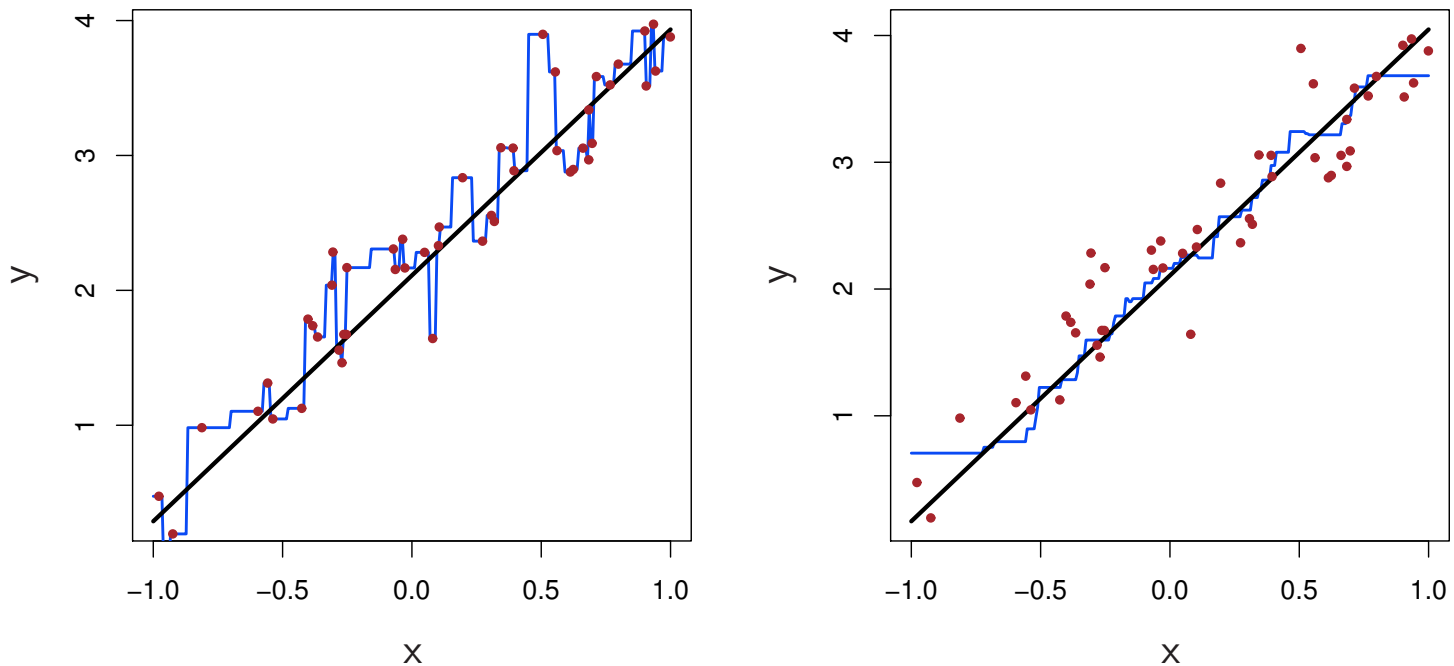

Figure 3.17: Plot of $\hat{f}(X)$ using KNN regression on a one-dimensional data set with 100 observations. The true relationship is given by the black solid line. Left: The blue curve corresponds to $K = 1$ and interpolates (i.e. passes directly though) the training data. Right: The blue curve corresponds to $K = 9$, and represents a smoother fit.

Comparison of Linear Regression with K-Nearest Neighbors

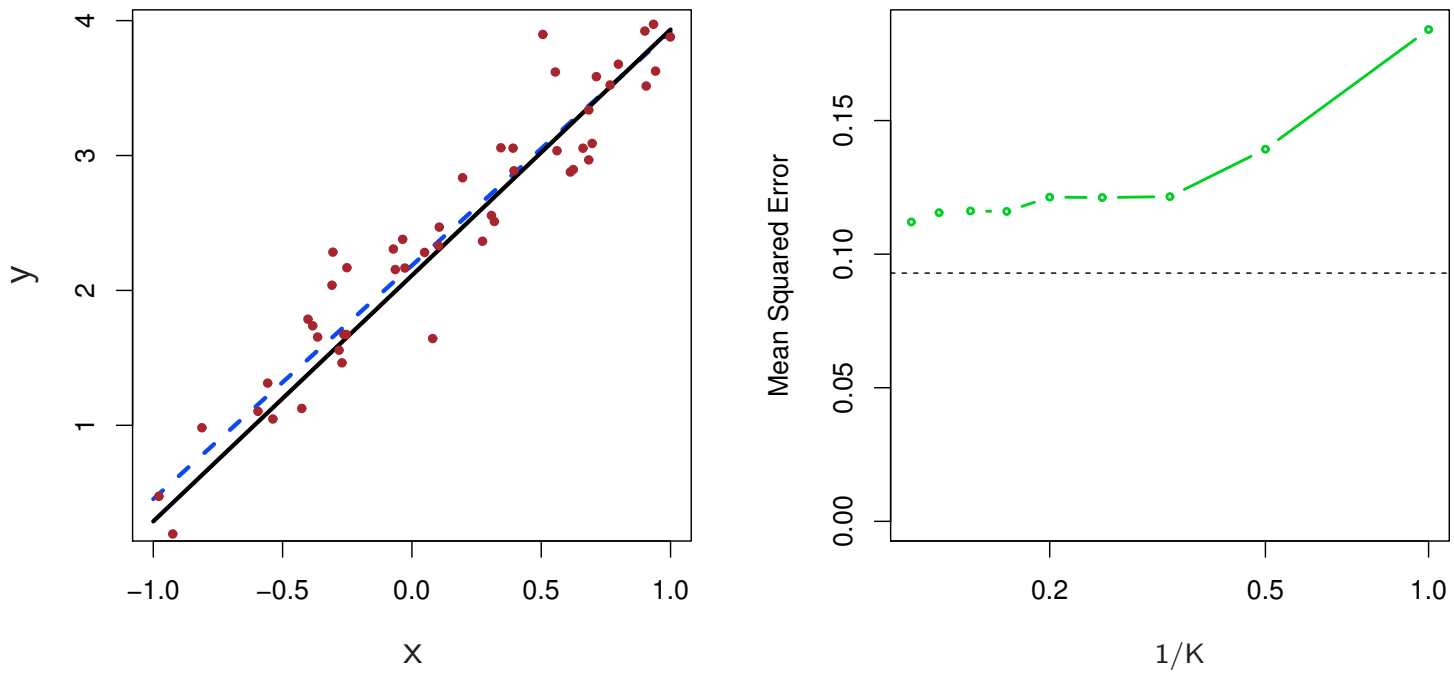

Figure 3.18: The same data set shown in Figure 3.17 is investigated further. Left: The blue dashed line is the least squares fit to the data. Since $f(X)$ is in fact linear (dispelled as the black line), the least squares regression line provides a very good estimate of $f(X)$. Right: The dashed horizontal line represents the least squares test set MSE, while the green solid line corresponds to the MSE for KNN as a function of $1/K$ (on the log scale). Linear regression achieves a lower test MSE than does KNN regression, since $f(X)$ is in fact linear. For KNN regression, the best result occur with a very large value of K, corresponding to a small value of $1/K$.

Comparison of Linear Regression with K-Nearest Neighbors

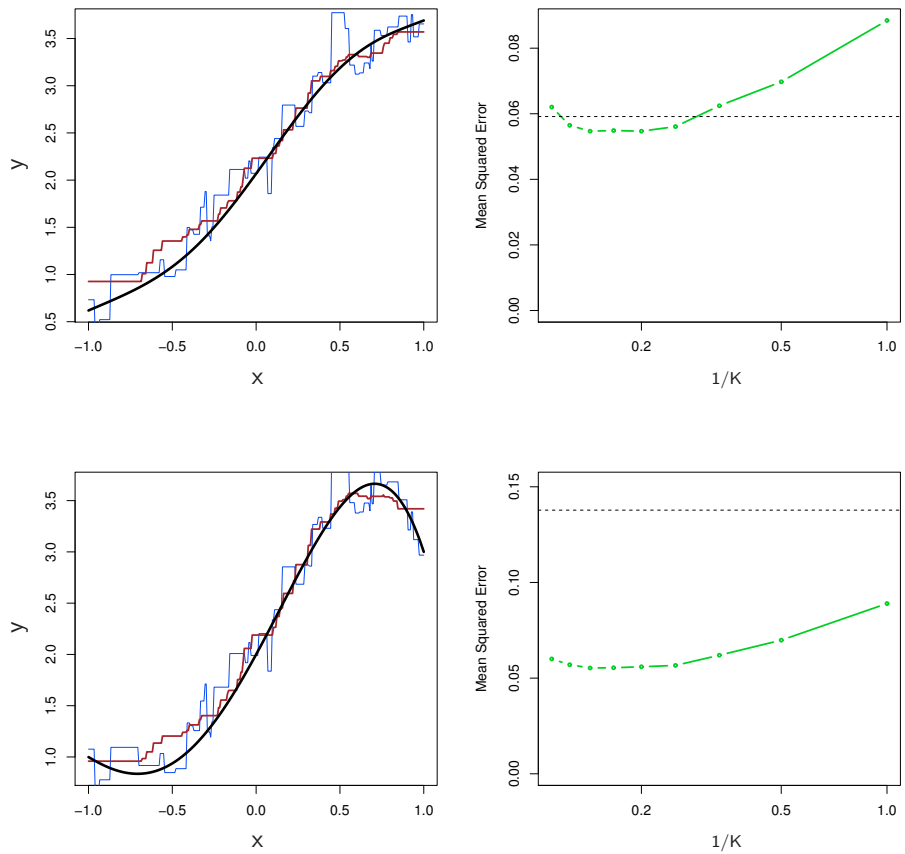

Figure 3.19: Top Left: In a setting with a slightly non-linear relationship between X and Y (solid black line), the KNN fits with $K = 1$ (blue) and $K = 9$ (red) are displayed. Top Right: For the slightly non-linear data, the test set MSE for least squares regression (horizontal black) and KNN with various values of $1/K$ (green) are displayed. Bottom Left and Bottom Right: As in the top panel, but with a strong non-linear relationship between X and Y.

Comparison of Linear Regression with K-Nearest Neighbors

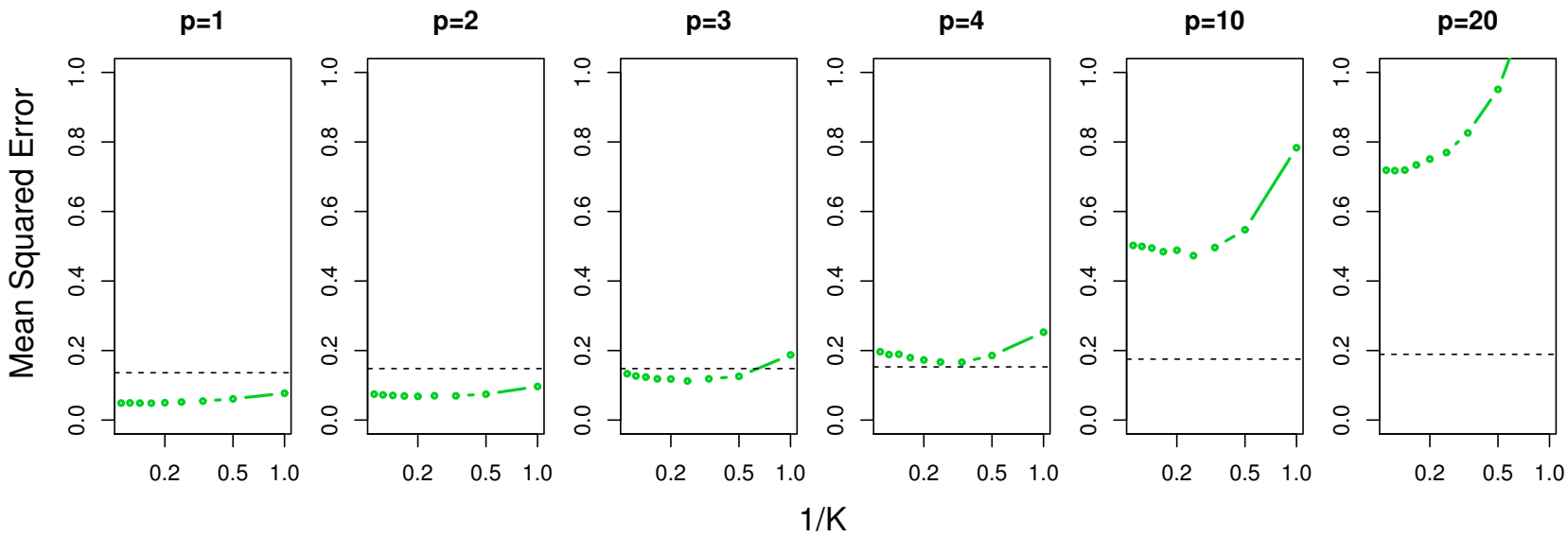

Figure 3.20: Test MSE for linear regression (black dashed lines) and KNN (green curves) as the number of variables p increases. The true function is non-linear in the first variable, as in the lower panel in Figure 3.19, and does not depend on the additional variables. The performance of linear regression deteriorates slowly in the presence of these additional noise variables, whereas KNN's performance degrades much more quickly as p increases.

Lab: Linear Regression