LINEAR REGRESSION

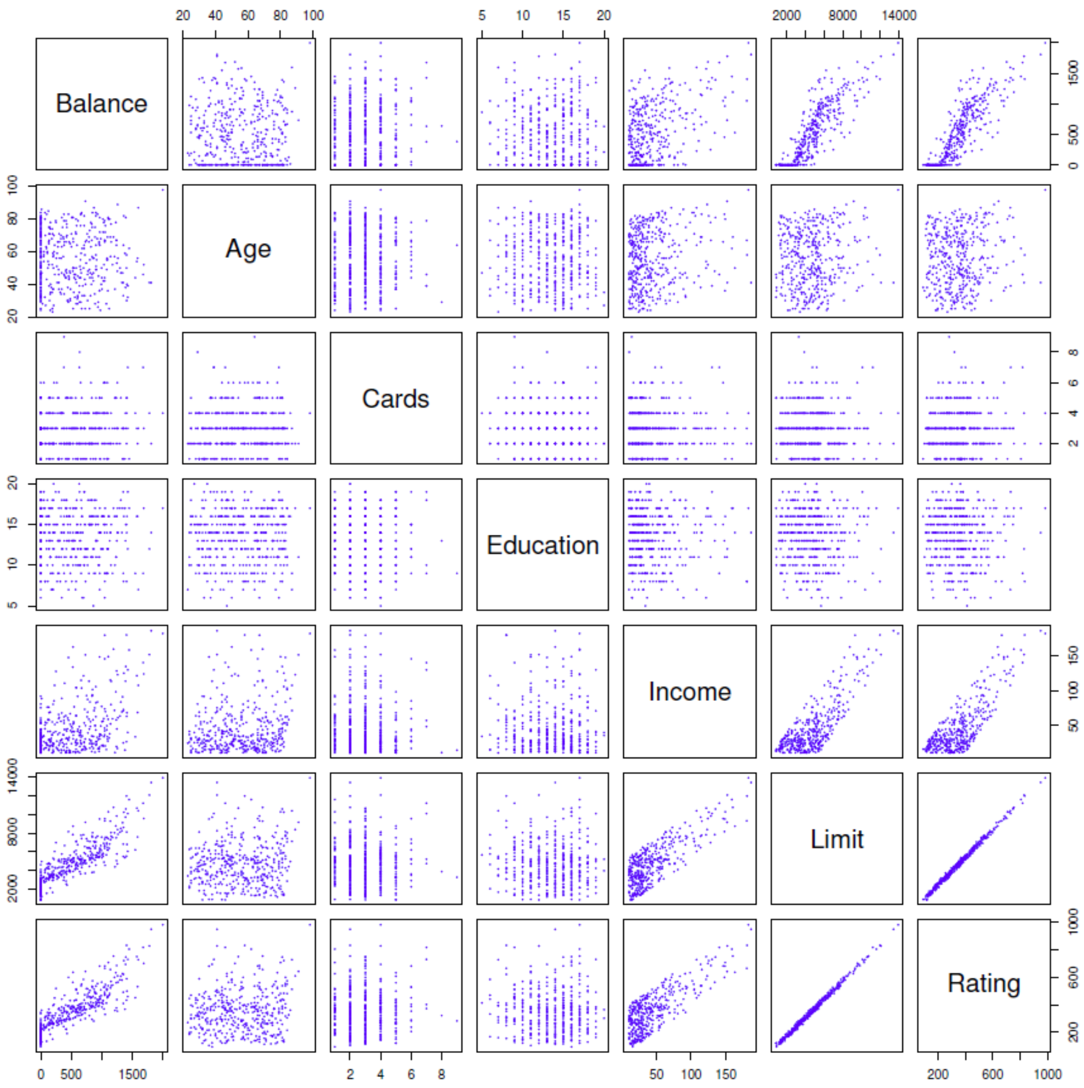
Part III

Outline

- Qualitative predictors
- Extensions of the linear model
- Regression diagnostics

Credit card data

- Information about credit card debt for 10,000 customers (credit)
- Understand the association between a customer's balance and a number of variables, such as age, cards, education, income, limit, rating, own, student, status, and region



Some of the figures and tables in this presentation are taken from "*An Introduction to Statistical Learning, with Applications in R*" (Springer) with permission from the authors: G. James, D. Witten, T. Hastie, and R. Tibshirani

- Response: balance
- Quantitative predictors: age, cards, education, income, limit, and rating
- Qualitative predictors: own, student, status, and region

Qualitative predictors

Dummy variables

- Predictors with two levels (e.g., own)

$$x_{i1} = \begin{cases} 1, & \text{if } i\text{th person owns a house} \\ 0, & \text{if } i\text{th person dose not} \end{cases}$$

- The linear regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$$
$$= \begin{cases} \beta_0 + \beta_1 + \epsilon_i, & \text{if } i\text{th person owns a house} \\ \beta_0 + \epsilon_i, & \text{if } i\text{th person does not} \end{cases}$$

	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).*

- Predictors with more than two levels (e.g., region)

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

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

- The linear model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

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

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

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).*

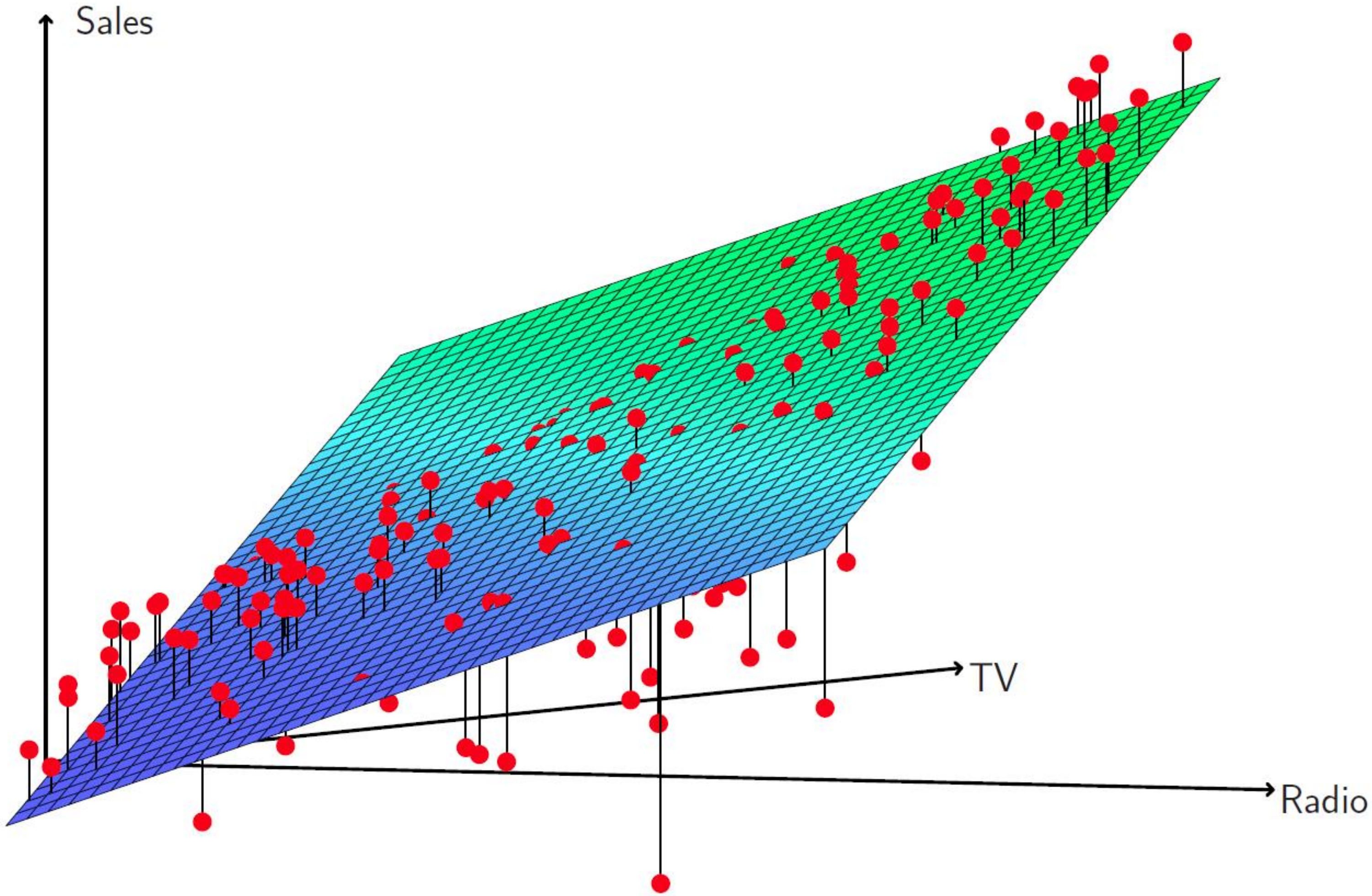
Extensions of the linear model

- The linear model with two predictors

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

Two important assumptions

- Additivity—the effect of changes in a predictor on the response is independent of the values of the other predictors
- Linearity—the change in the response due to a one-unit change in a predictor is constant



For the Advertising data, a linear regression fit to sales using TV and radio as predictors

Removing the additive assumption

- Inclusion of an interaction term

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \underbrace{\beta_3 X_1 X_2}_{\substack{\text{interaction} \\ \text{or} \\ \text{synergy}}} + \epsilon$$

$$\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 = (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2$$

$$\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 = \beta_1 X_1 + (\beta_2 + \beta_3 X_1) X_2$$

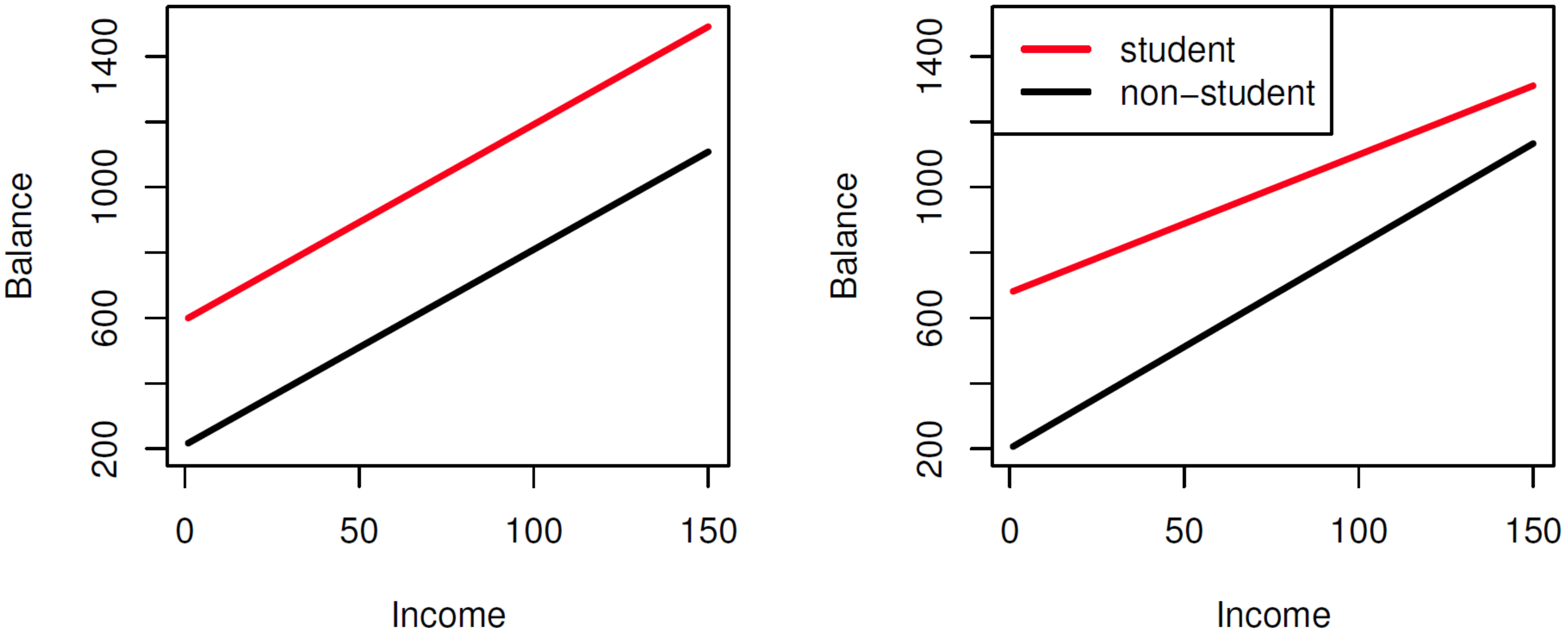
	Coefficient	Std. error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	< 0.0001

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).*

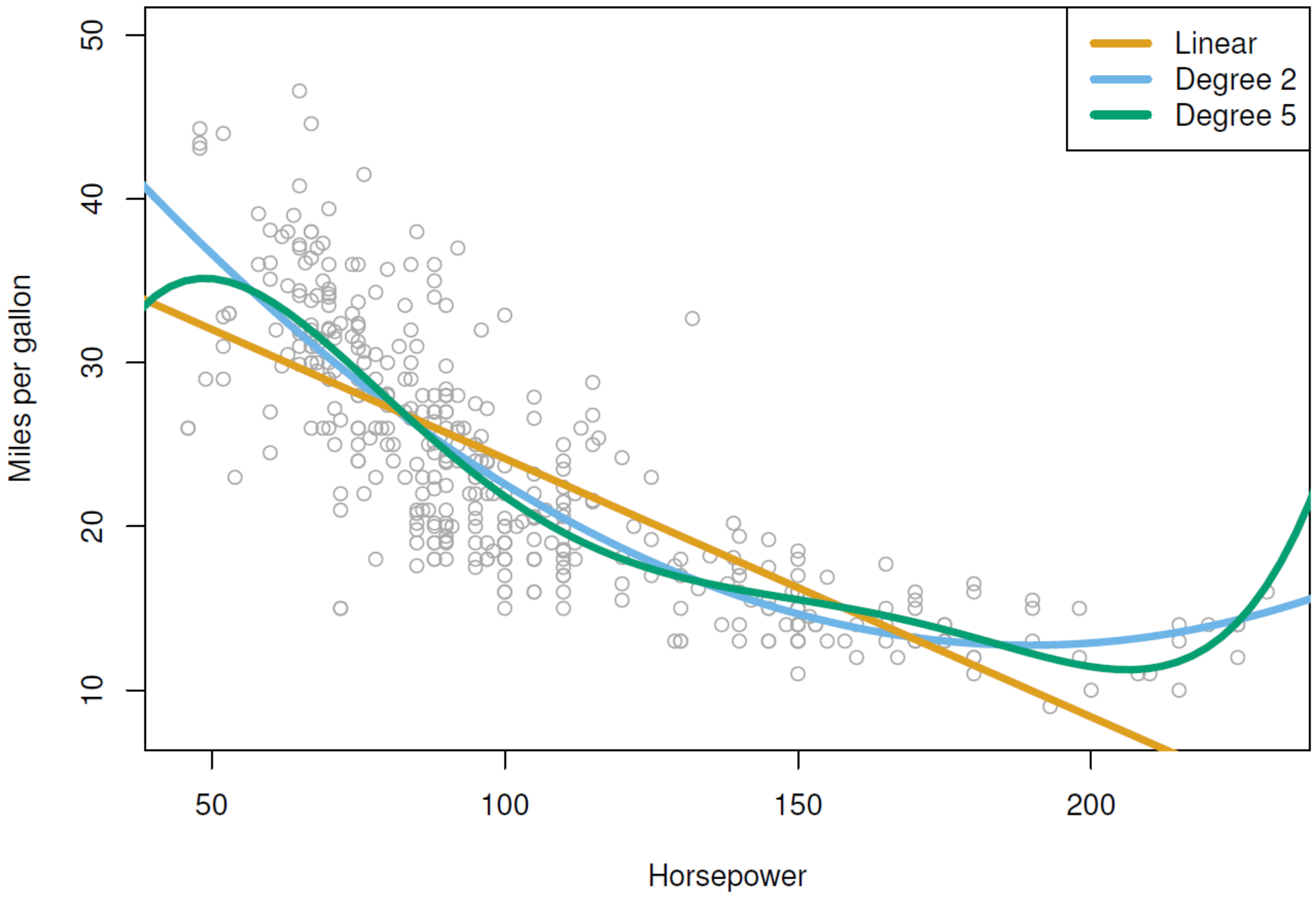
Credit card data

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

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


Balance
1400
1000
600
200
0
50
100
150
Income
student
non–student

Non-linear relationships



	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

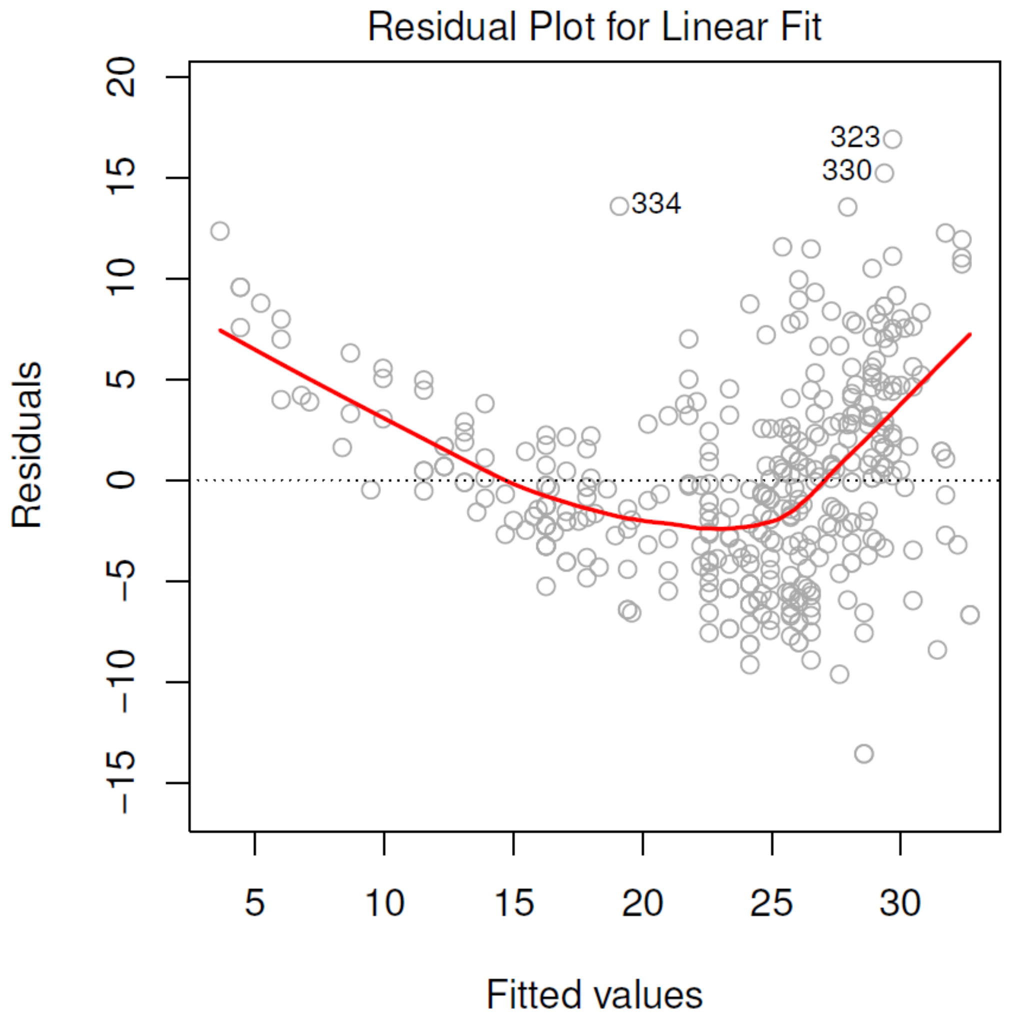
TABLE 3.10. *For the* `Auto` *data set, least squares coefficient estimates associated with the regression of* `mpg` *onto* `horsepower` *and* `horsepower`^{2}.

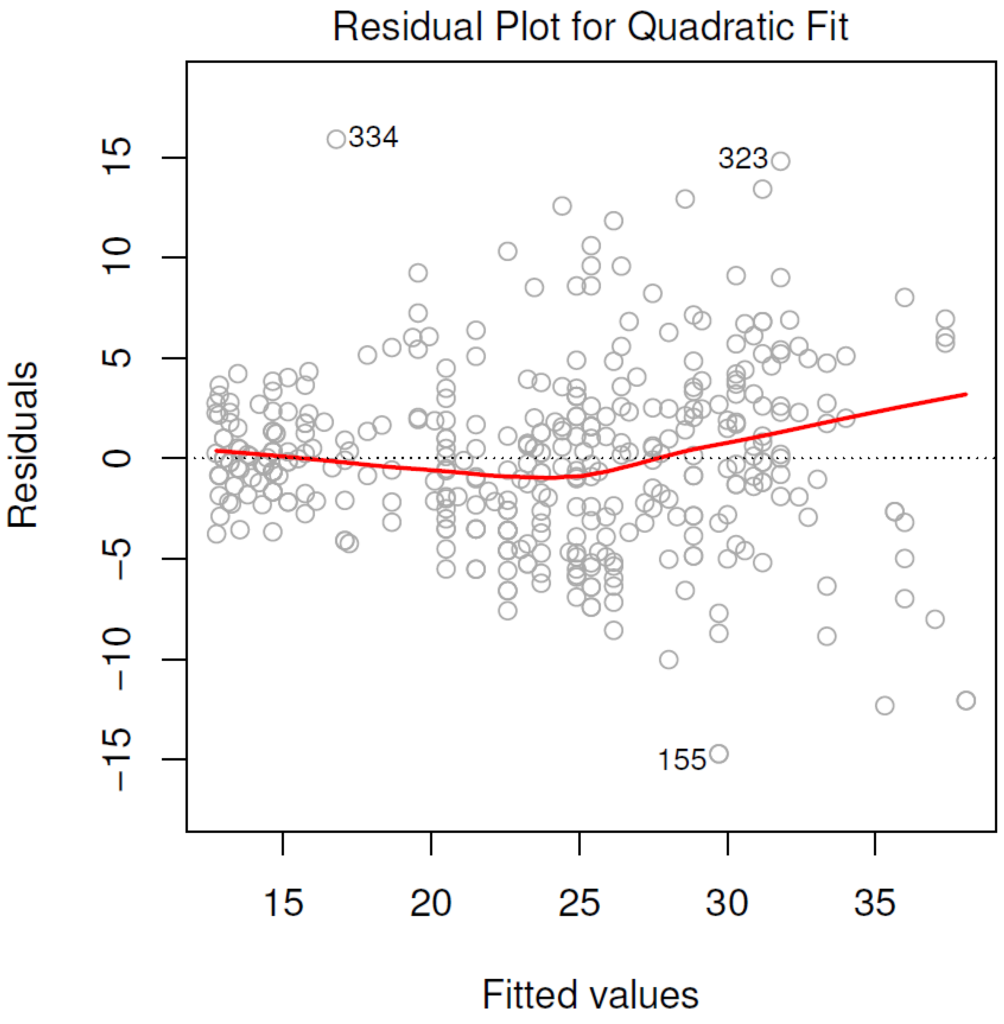
Regression diagnostics

- *Non-linearity of the response-predictor relationships*
- *Correlation of error terms*
- *Non-constant variance of error terms*
- *Outliers*
- *High-leverage points*
- *Collinearity*

Non-linearity of the data

- The linear model assumes that there is a straight-line relationship between the predictors and the response
- Residual plots

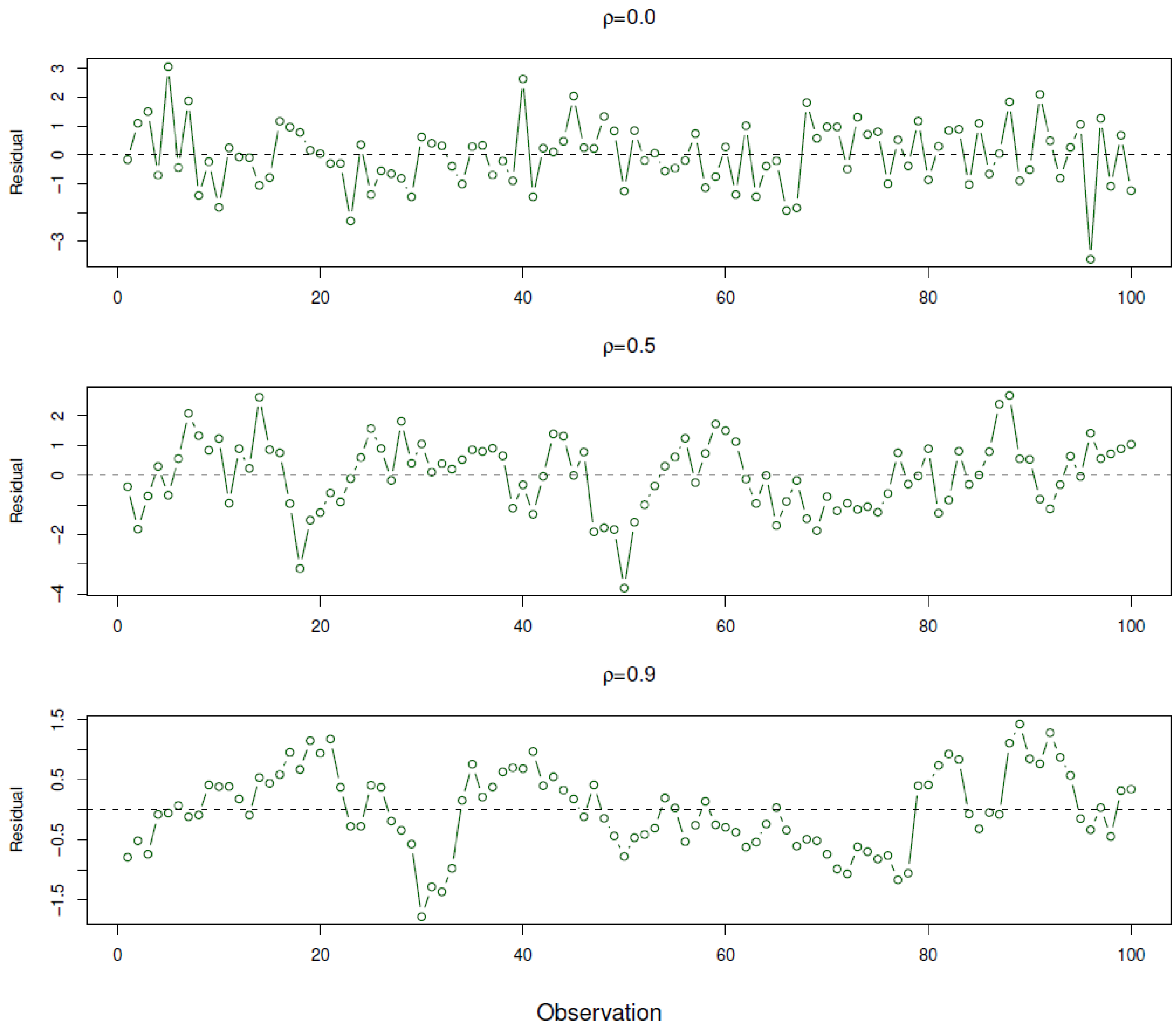




Plots of residuals versus predicted values for the Auto data set

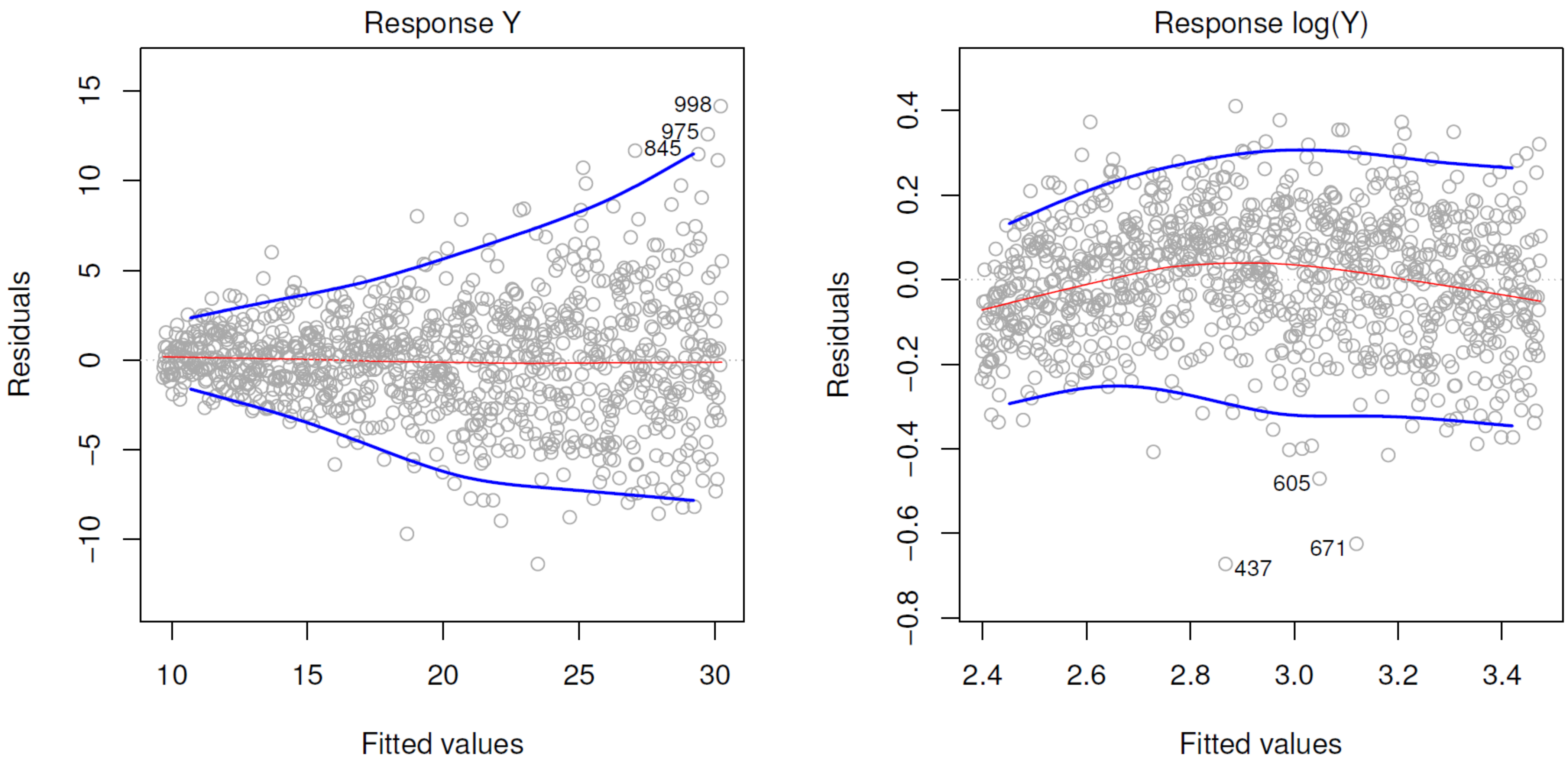
Correlation of error terms

- An important assumption of the linear model is that the error terms are uncorrelated
- Residual plots


ρ=0.0
Residual
ρ=0.5
Residual
ρ=0.9
Residual
Observation

Non-constant variance of error terms

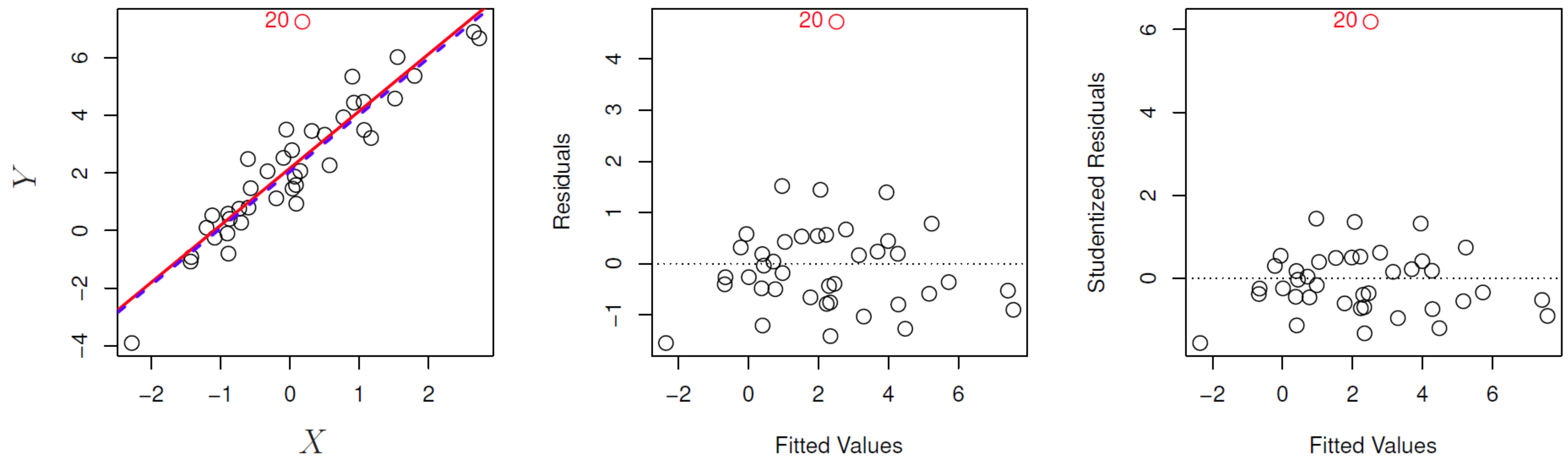
- Another important assumption of the linear model is that the error terms have a constant variance
- Residual plots



Response Y
Residuals
Fitted values
998
975
845
Response log(Y)
Residuals
Fitted values
605
671
437

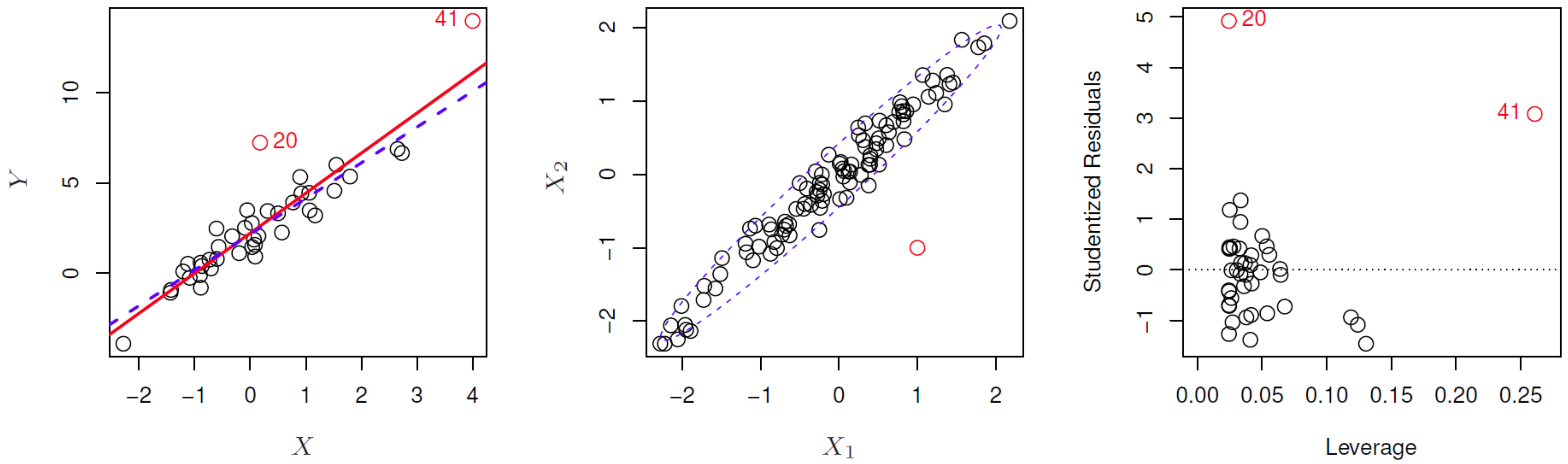
Outliers

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


20
Y
X
Residuals
Fitted Values
Studentized Residuals
Fitted Values

High leverage points

- Observations whose predictor values are unusual
- Have a sizable impact on the estimated regression fit



41
20
Y
X
X₂
X₁
20
41
Studentized Residuals
Leverage

Leverage statistics

- For a simple linear regression, the *leverage statistic* for the ith observation is

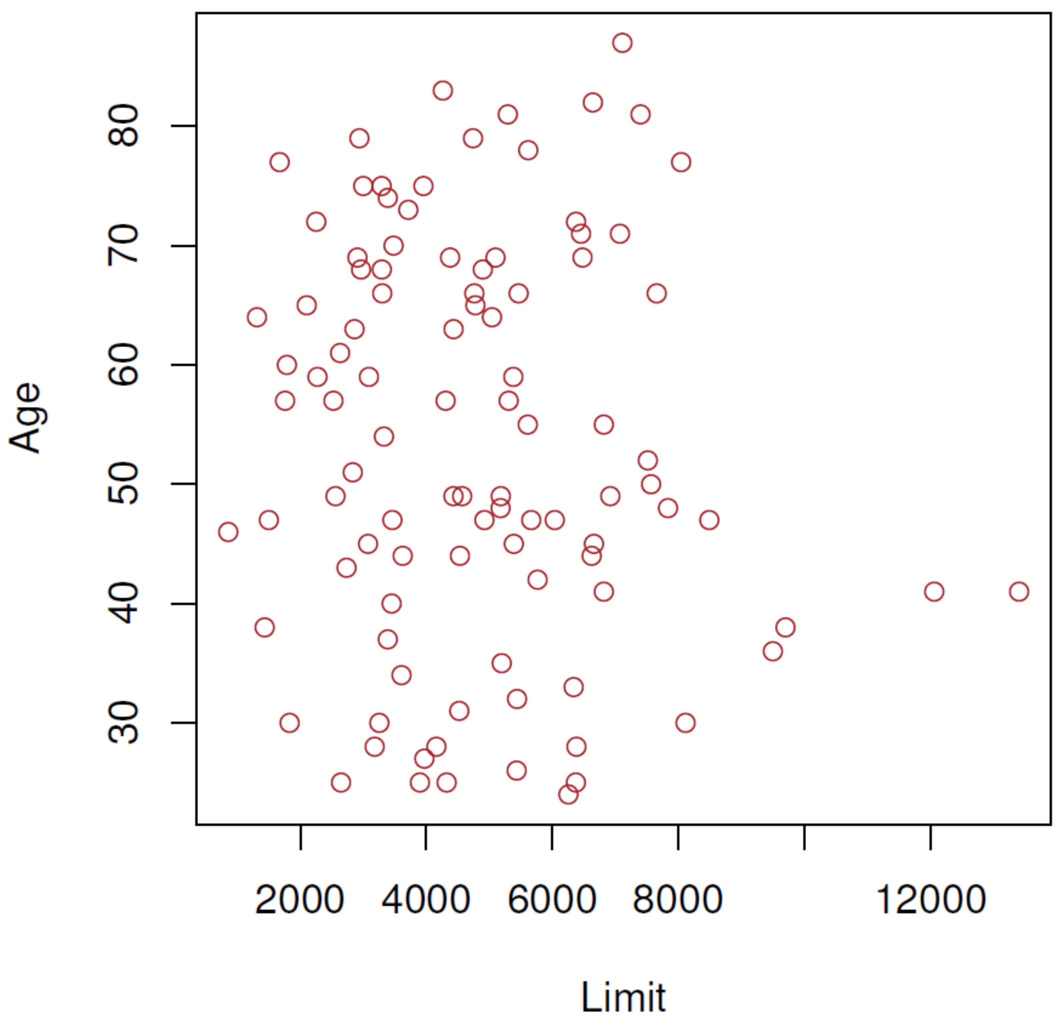
$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

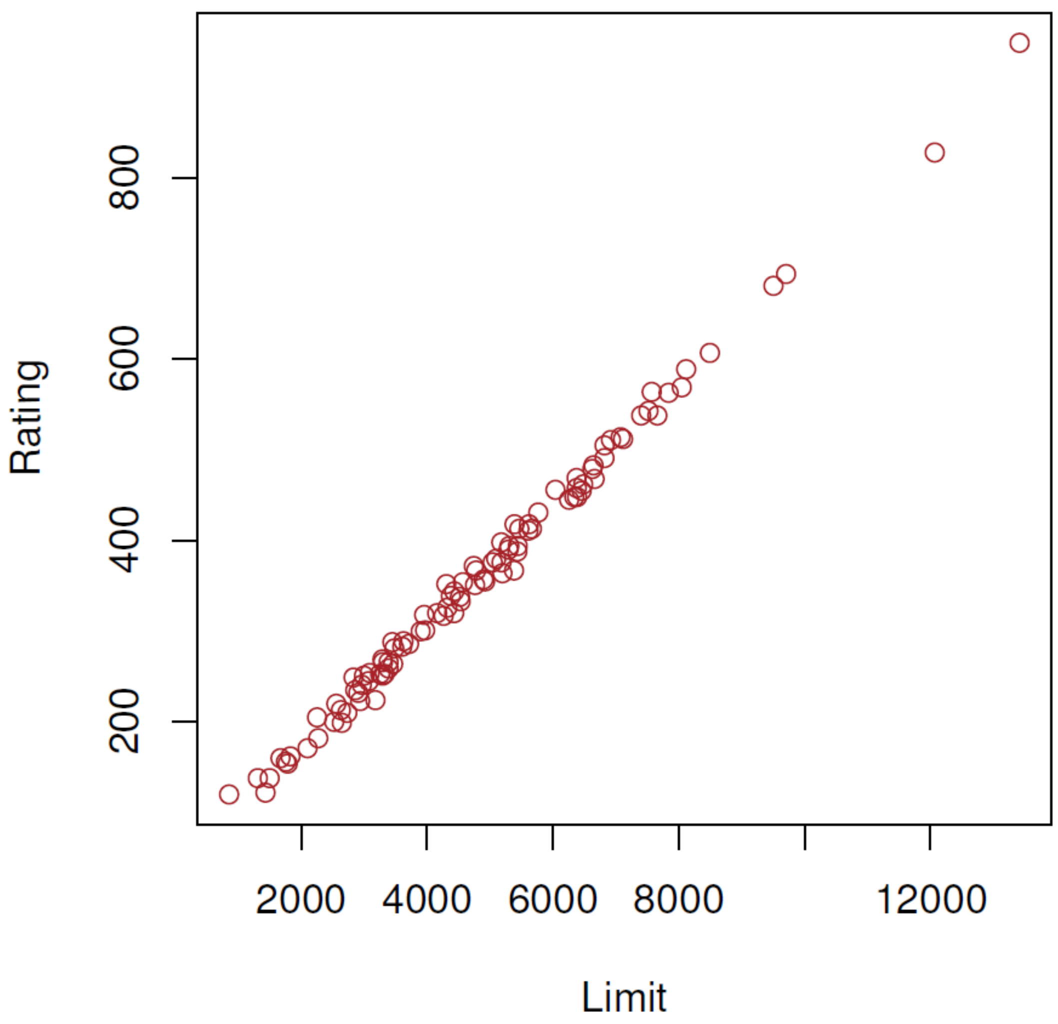
- The ith *studentized* residual is computed by dividing e_i by its estimated standard error

$$\frac{e_i}{\hat{\sigma}\sqrt{1 - h_i}}$$

Collinearity

- The situation in which two or more predictor variables are closely related to one another


Age
80
70
60
50
40
30
2000
4000
6000
8000
12000
Limit


Rating
800
600
400
200
2000
4000
6000
8000
12000
Limit

		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.*

- How could we detect collinearity?
- Multi-collinearity

Variance inflation factor (VIF)

- The ratio of the variance of $\hat{\beta}_j$ when fitting the full model divided by the variance of $\hat{\beta}_j$ if fit on its own

- Let R_j^2 be the R^2 statistic from the linear regression of X_j onto all of the other predictors

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2}$$

- In the Credit data, age, rating, and limit have VIF values of 1.01, 160.67, and 160.59