Fall 2023

BIF524/CSC463 Data Mining

Linear Regression

Eileen Marie Hanna, *PhD*

28/09/2023

Extensions of the linear model

- Two main restrictive assumptions are made on predictors in linear models:

  - the relationship between the predictors
  - the response is additive and linear

i.e., the effect of X_i on Y is independent of other predictors.

the change in Y due to a unit-change in X_i is constant, regardless of the value of X_i.

How can we relax these restrictions?

Removing the additive assumption

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

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

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

i.e., the coefficient of X_1 is dependent on X_2

A change in X_2 will change the effect of X_1 on Y.

Example – factory productivity

- The number of units produced based on the number of production lines and total number of workers.

- Obviously, increasing the number of lines also depends on the number of workers -> interaction term between lines and workers.

$$\texttt{units} \approx 1.2 + 3.4 \times \texttt{lines} + 0.22 \times \texttt{workers} + 1.4 \times (\texttt{lines} \times \texttt{workers})$$

$$\approx 1.2 + (3.4 + 1.4 \times \text{workers}) \times \text{lines} + 0.22 \times \text{workers}$$

How does the number of produced units change when we add a production line?

Example – factory productivity

- The number of units produced based on the number of production lines and total number of workers.

- Obviously, increasing the number of lines also depends on the number of workers -> interaction term between lines and workers.

$$\begin{aligned}\texttt{units} &\approx 1.2 + 3.4 \times \texttt{lines} + 0.22 \times \texttt{workers} + 1.4 \times (\texttt{lines} \times \texttt{workers}) \\ &= 1.2 + \underline{(3.4 + 1.4 \times \texttt{workers})} \times \texttt{lines} + 0.22 \times \texttt{workers}.\end{aligned}$$

How does the number of produced units change when we add a production line?

Example – factory productivity

- The number of units produced based on the number of production lines and total number of workers.
- Obviously, increasing the number of lines also depends on the number of workers -> interaction term between lines and workers.

$$
\begin{aligned}
\texttt{units} &\approx 1.2 + 3.4 \times \texttt{lines} + 0.22 \times \texttt{workers} + 1.4 \times (\texttt{lines} \times \texttt{workers}) \\
&= 1.2 + \underline{(3.4 + 1.4 \times \texttt{workers})} \times \texttt{lines} + 0.22 \times \texttt{workers}.
\end{aligned}
$$

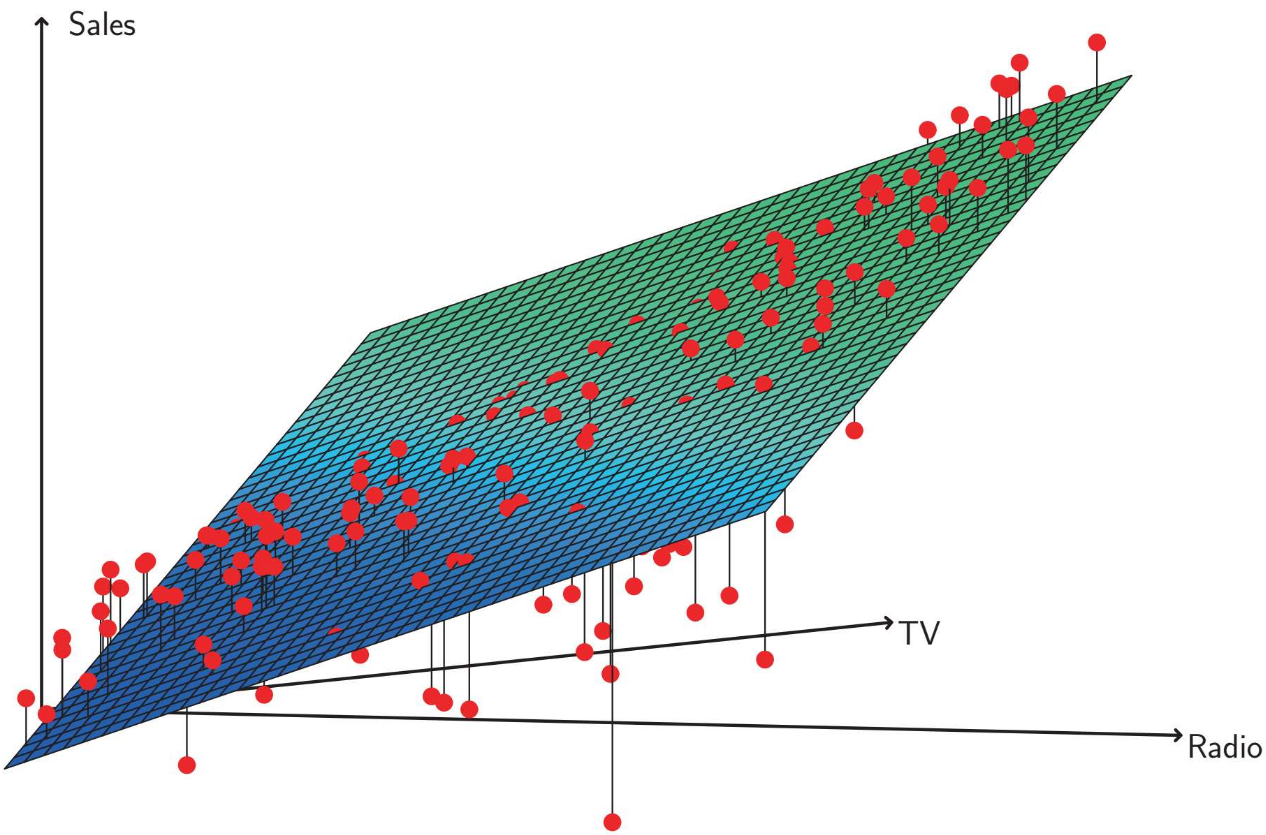
How does the number of produced units change when we add a production line?

Adding one line will increase sales by $(3.4 + 1.4 \times workers)$ units.

Removing the additive assumption

- Previous regression for the advertising data -> average effect of unit increase in TV budget on sales is β_1, regardless of the amount is spent on radio advertising.

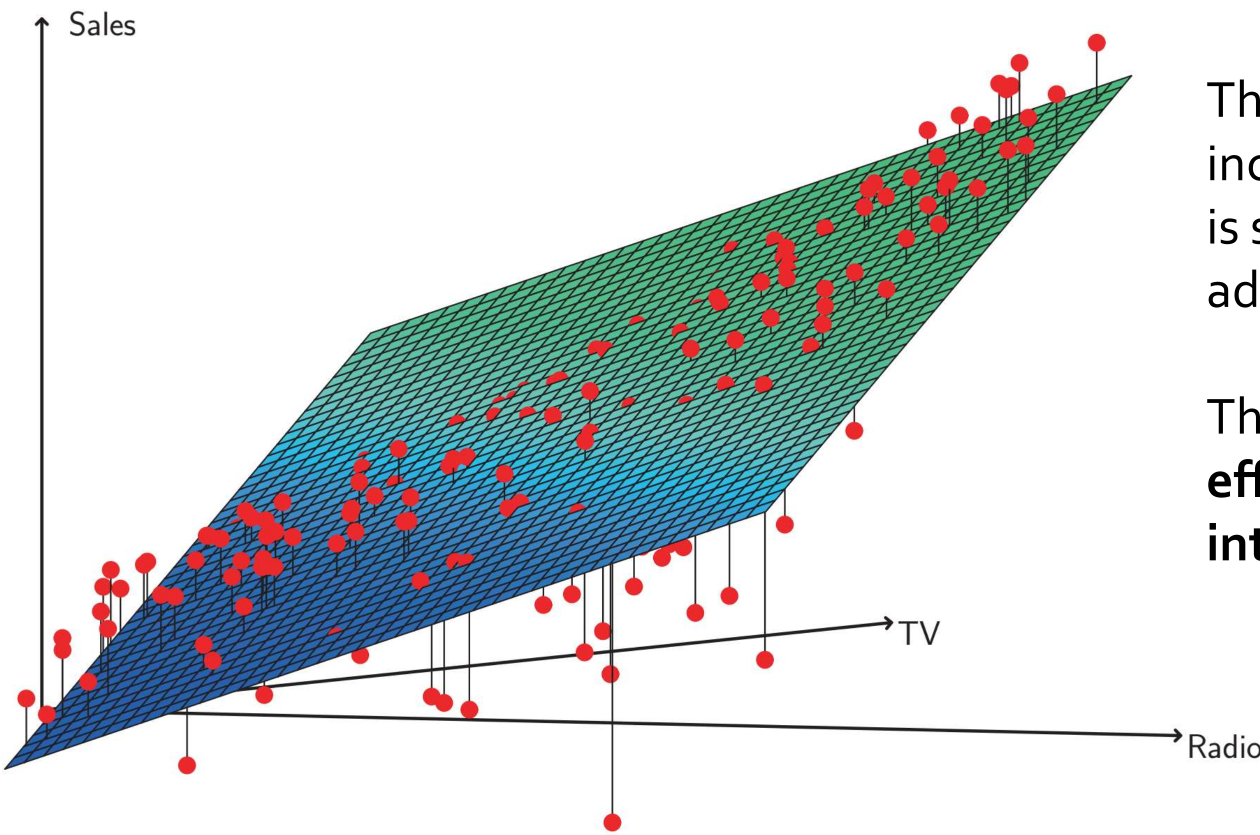
Does spending money on radio increase the effectiveness of TV advertising?



Removing the additive assumption

- Previous regression for the advertising data -> average effect of unit increase in TV budget on sales is β_1, regardless of the amount is spent on radio advertising.

What if spending money on radio increases the effectiveness of TV advertising?

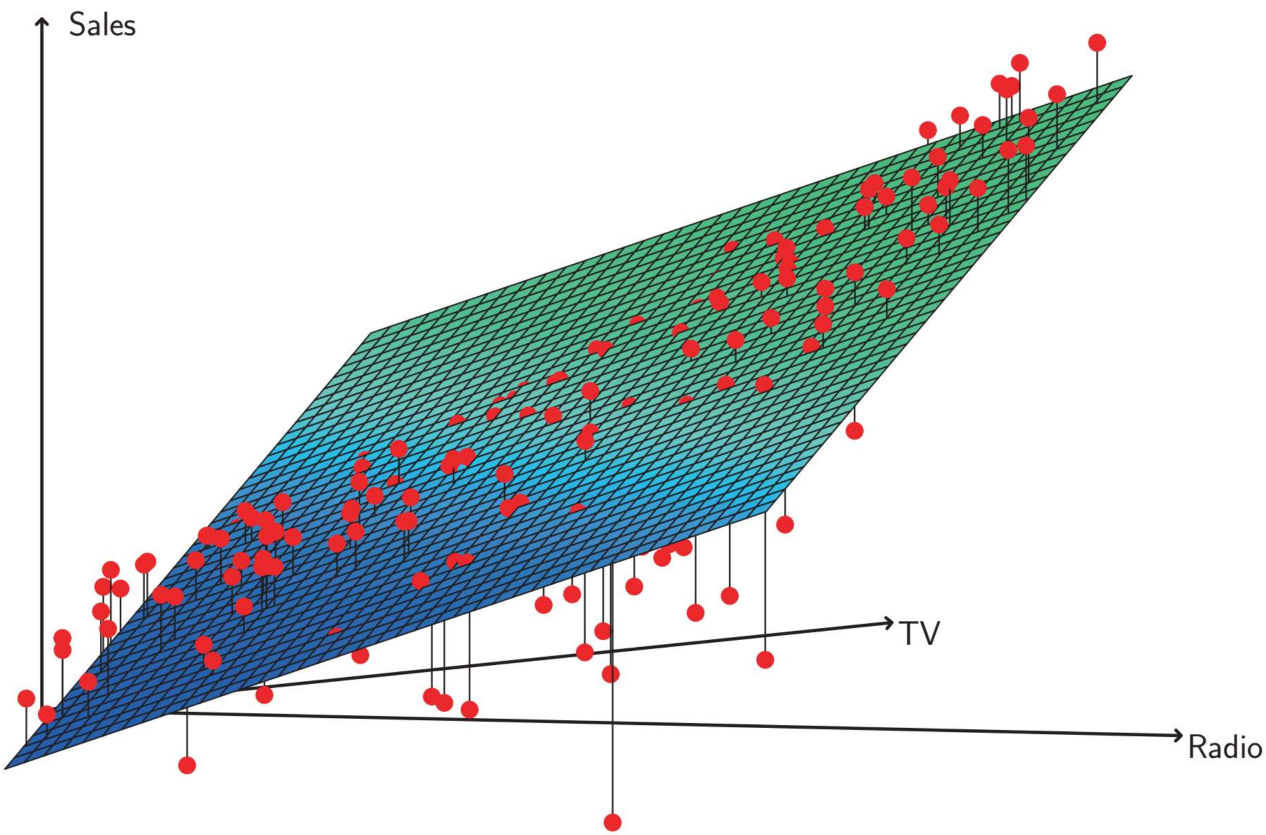


There seems to be a sales increase whenever the budget is split between TV and radio advertising.

This is referred to as **synergy effect in marketing**, and **interaction effect in statistics**.

Removing the additive assumption

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

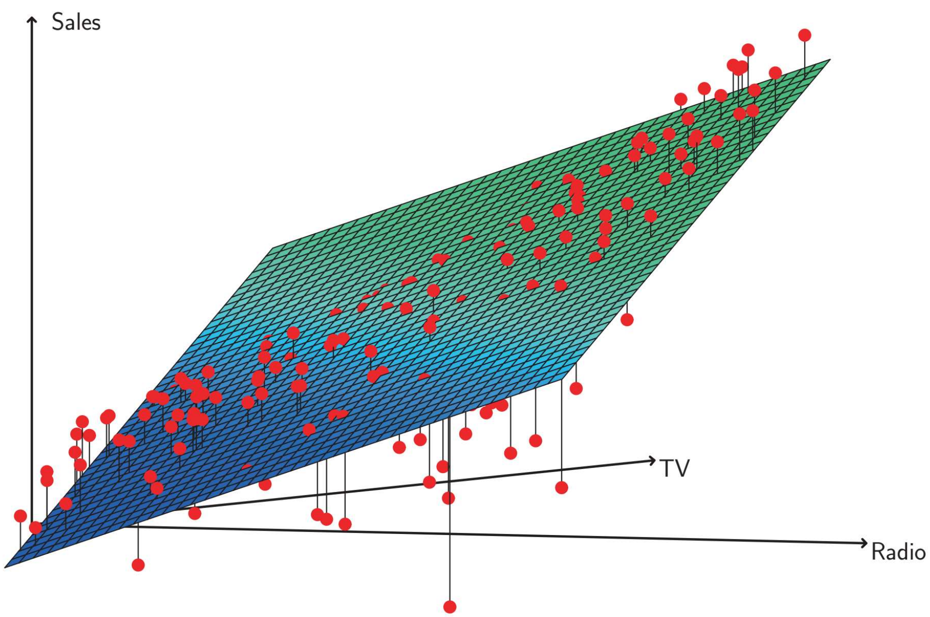


Let us modify the additive assumption for the advertising model

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

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

	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



The true relationship is not additive.

When the value of radio changes, the coefficient of TV will thus change.

Let us modify the additive assumption for the advertising model

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

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

	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

Let us modify the additive assumption for the advertising model

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

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

	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

$R^2 = 96.8\%$ for this model, it was 89.7% for the previous model in which we only considered an additive effect among predictors.

Recall that R^2 measures the proportion of variability in Y that can be explained using X.

$\frac{96.8-89.7}{100-89.7} \approx 69\%$ of the variability in sales that remains after fitting the additive model has been explained by the interaction term.

Let us modify the additive assumption for the advertising model

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

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

	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

What is the effect of a $1000 increase in TV advertising?

$\underbrace{19.1}+\underbrace{1.1 \times \text{radio}}$ units.

Let us modify the additive assumption for the advertising model

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

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

	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

What is the effect of a $\$1000$ increase in TV advertising?

sales increase of $(\hat{\beta}_1 + \hat{\beta}_3 \times radio) \times 1000 = 19.1 + 1.1 \times radio$ units

Let us modify the additive assumption for the advertising model

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

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

	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

What is the effect of a $\$1000$ increase in radio advertising?

Let us modify the additive assumption for the advertising model

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

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

	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

What is the effect of a $\$1000$ increase in radio advertising?

sales increase of $(\hat{\beta}_2 + \hat{\beta}_3 \times \mathrm{TV}) \times 1000 = 28.9 + 1.1 \times TV$ units

Comments

	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~~ 0.65
TV×radio	0.0011	0.000	20.73	< 0.0001

- In this example, all p-values are statistically significant -> all three should be included in the model.
- In some cases, we may have a significant p-value for the interaction term but insignificant p-values for the main effects.
- Based on the hierarchical principle**, if we include the interaction in the model** -> we should **also include the main effects**, **even if they have insignificant p-values**.

What to do when there are both quantitative and qualitative attributes?

- The same applies to qualitative variables and to combination of both types.
- Considering the **credit data** which we discussed earlier, let's assume that **we want to predict balance using income and student variables**.
- Without considering interaction:

$$\text{balance} \approx \beta_0 + \beta_1 \times \text{income} + \beta_2 \times \text{student} + \epsilon$$

$\beta_0 + \beta_1 \times \text{income} + \beta_2$ if student

$\beta_0 + \beta_1 \times \text{income}$ if not student

What to do when there are both quantitative and qualitative attributes?

- With interaction:

$$bal = \beta_0 + \beta_1 \times inc + \beta_2 \times st + \beta_3 \times inc \times st$$

$$\beta_0 + \beta_1 \times inc + \beta_2 + \beta_3\, inc$$

$$\beta_0 + \beta_2 + (\beta_1 + \beta_3)\, inc$$

if student

$$\beta_0 + \beta_1 \times inc$$

if not student

What to do when there are both quantitative and qualitative attributes?

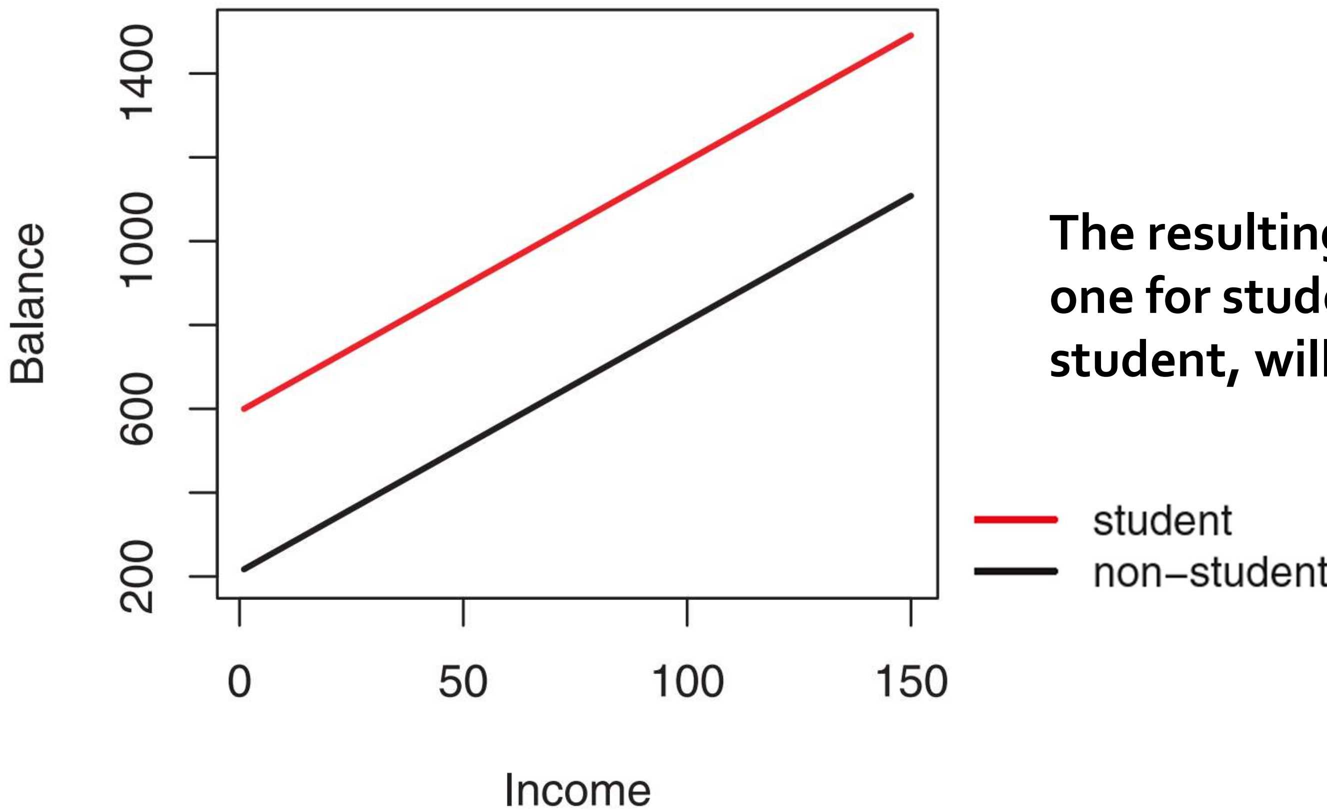
- The same applies to qualitative variables and to combination of both types.
- Considering the **credit data** which we discussed earlier, let's assume that **we want to predict balance using income and student variables**.
- With **no interaction term**, the model would be:

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

How would the least squares lines look like?

What to do when there are both quantitative and qualitative attributes?

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



The resulting least squares lines, one for student and one for non-student, will be parallel.

By adding an **interaction term**, the model would be:

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

How would the least squares lines look like?

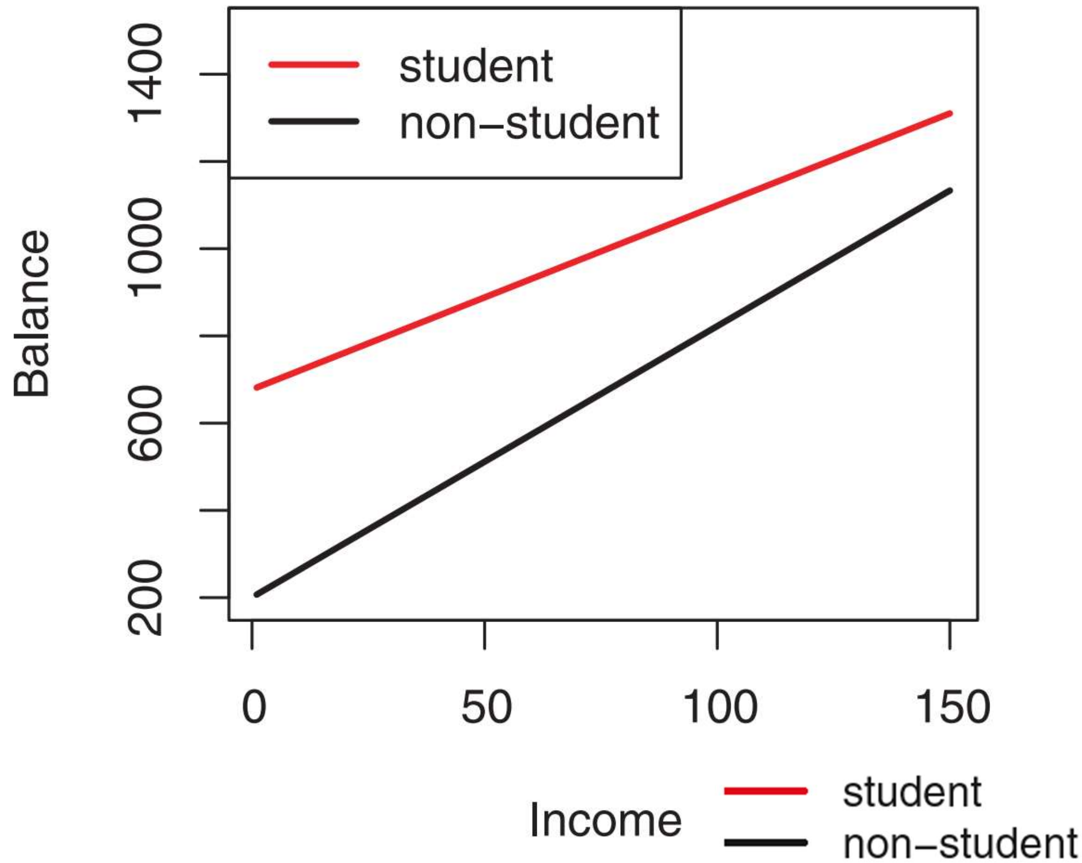
By adding an **interaction term**, the model would be:

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

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

We can observe that the slope for student is different than the slope for non-student.

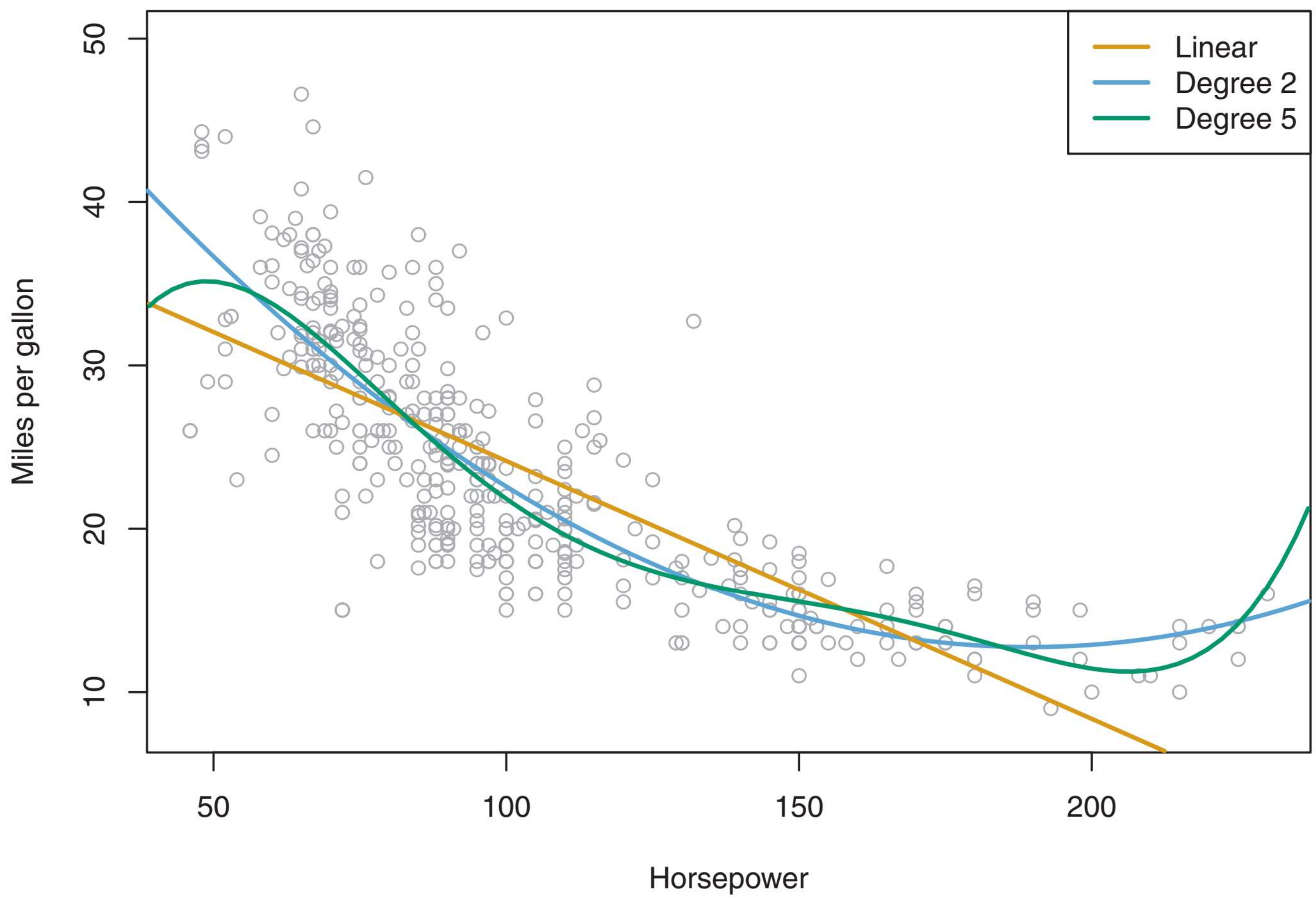
-> for student, smaller changes in credit balance when income is increased.



Non-linear relationships – polynomial regression

- The true relationship between predictors and response may not be linear in some cases.

- One simple way is to use **polynomial regression** to account for such non-linear relationships.

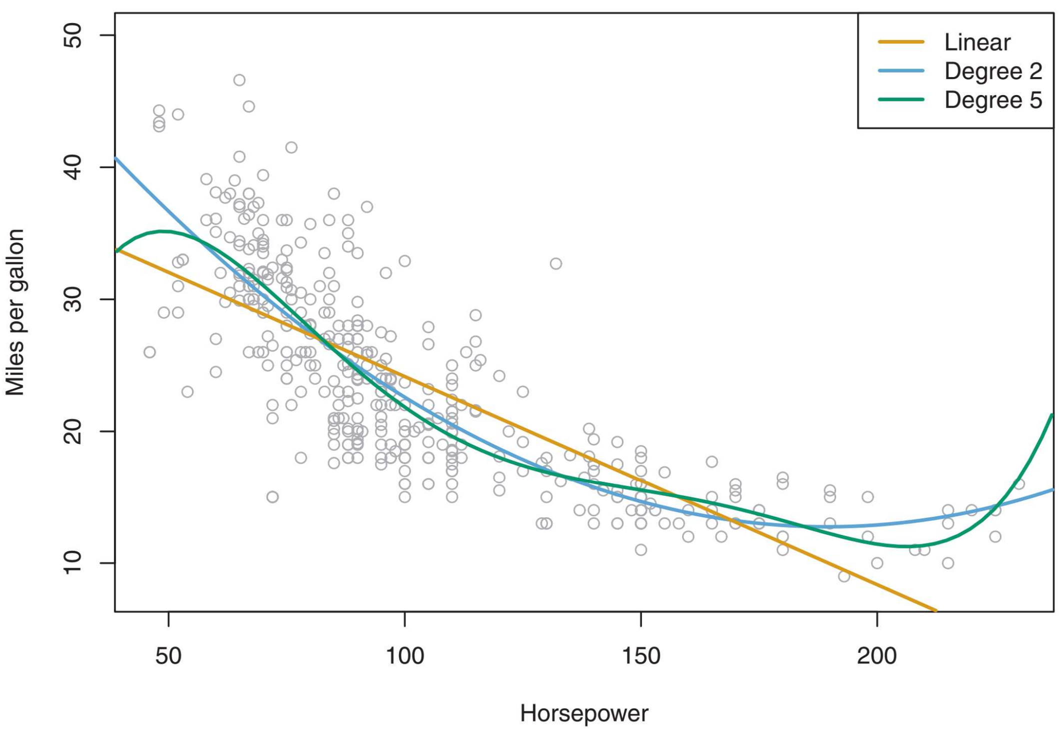


Non-linear relationships

- A simple way to approach non-linear associations in a linear model is to **add transformed versions of the predictors**.

- For instance, points in this graph show a **quadratic shape**.

$$\texttt{mpg} = \beta_0 + \beta_1 \times \texttt{horsepower} + \beta_2 \times \texttt{horsepower}^2 + \epsilon$$



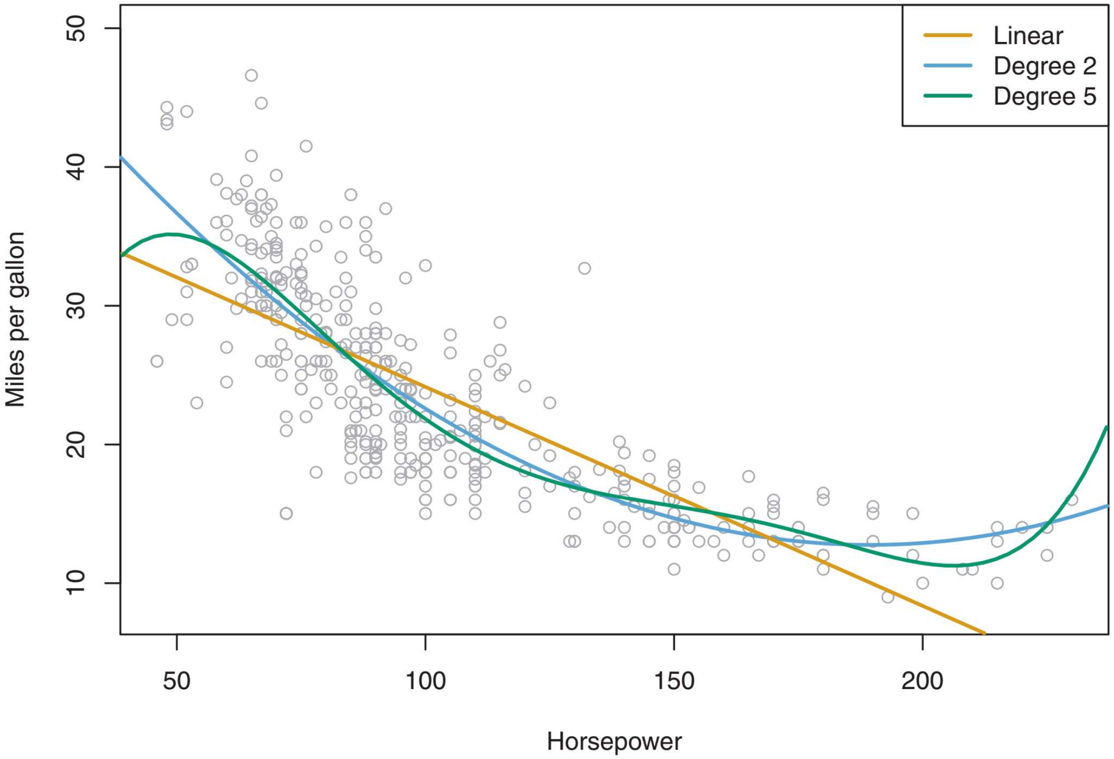
- Such model may provide a better fit of the data.

- Note that it will predict *mpg* based on a non-linear function of *horsepower*, but it is still linear!

- In fact, it is a multiple linear regression model with
  $X_1 = \text{horsepower}$
  and $X_2 = \text{horsepower}^2$

Non-linear relationships

$$\texttt{mpg} = \beta_0 + \beta_1 \times \texttt{horsepower} + \beta_2 \times \texttt{horsepower}^2 + \epsilon$$

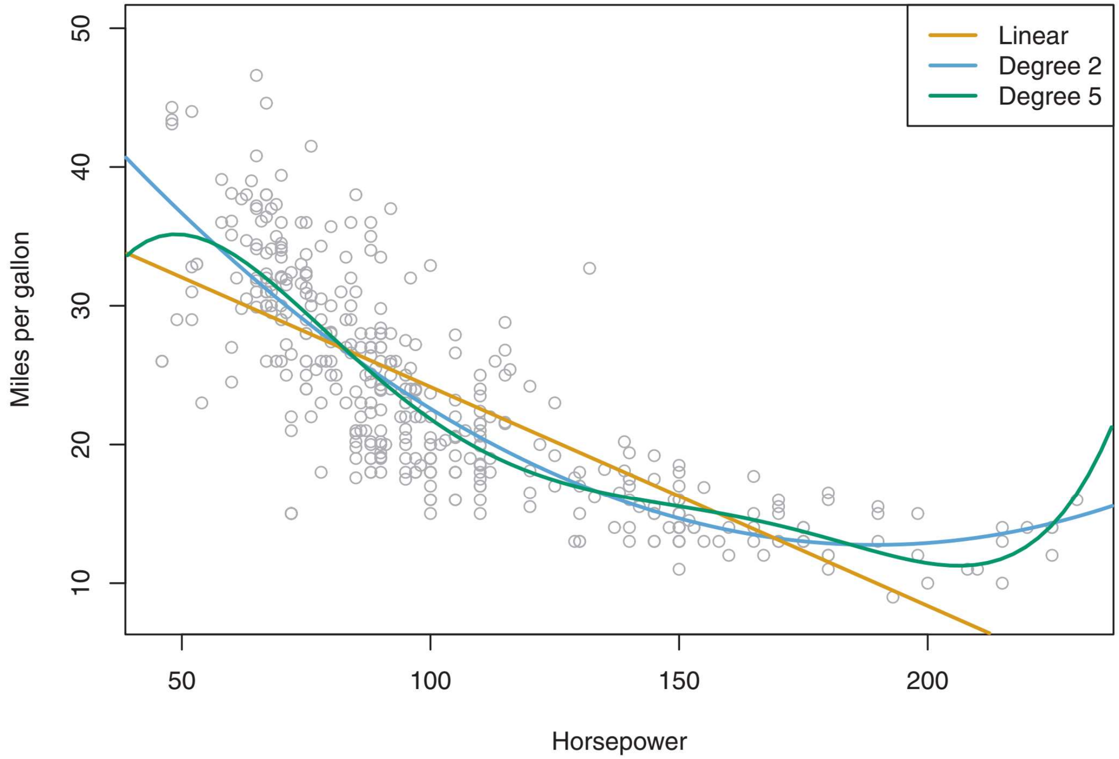


So, it is like **using a standard linear regression** software **to generate a non-linear fit by estimating the coefficients**.

	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
$\texttt{horsepower}^2$	0.0012	0.0001	10.1	< 0.0001

Non-linear relationships

$$\texttt{mpg} = \beta_0 + \beta_1 \times \texttt{horsepower} + \beta_2 \times \texttt{horsepower}^2 + \epsilon$$



What if we increase the degree of polynomials in the model?

The curve tends to become unnecessarily wiggly…

	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
$\texttt{horsepower}^2$	0.0012	0.0001	10.1	< 0.0001

Linear regression – common 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**

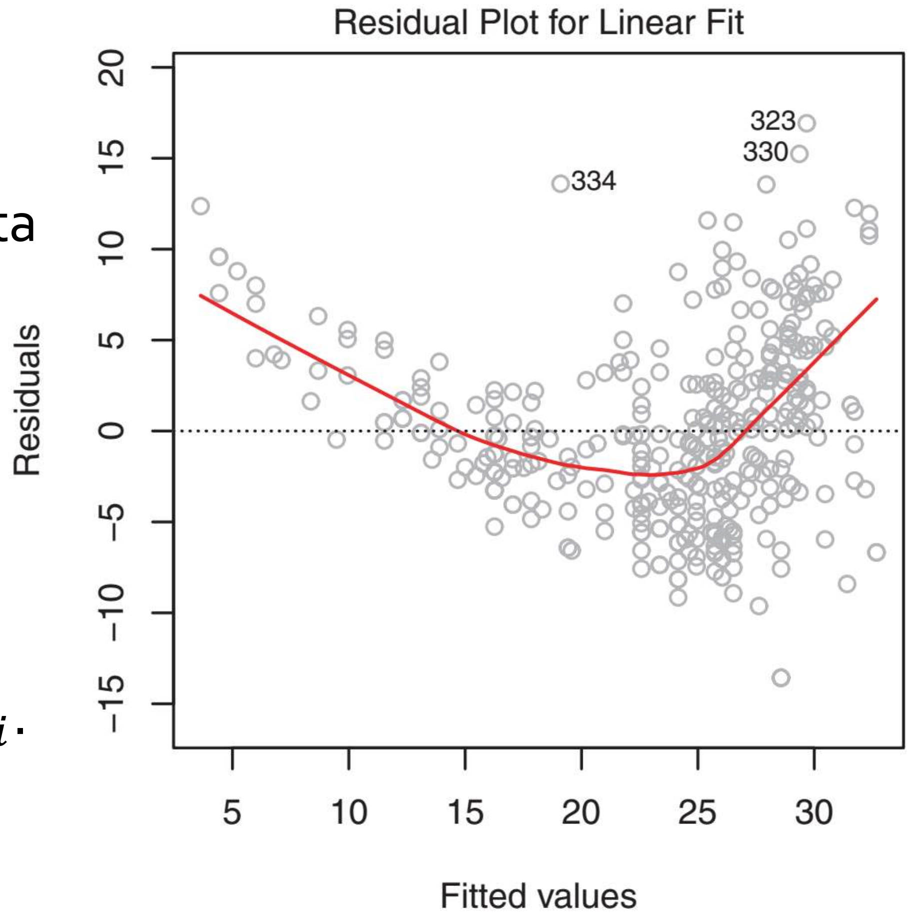
Non-linearity of the response-predictor relationships

- If the data that we are trying to fit is far from linear, using a linear regression would lead to erroneous conclusions as well as low prediction accuracy.

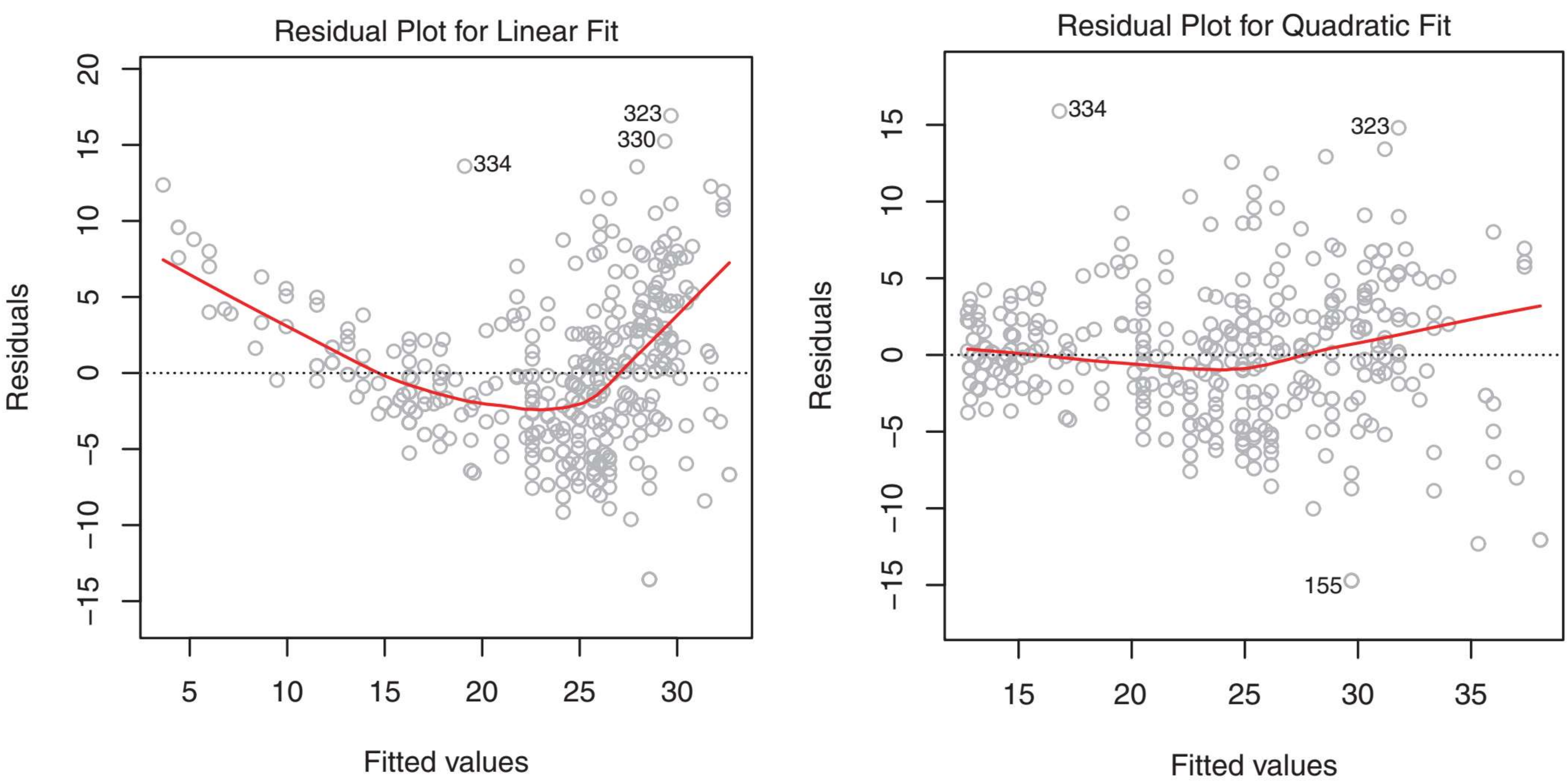
- One way to identify non-linearity of data in simple linear regression is to use **residual plots**.

$$e_i = y_i - \hat{y}_i \text{ vs the predictor } x_i$$

  - In multiple linear regression, plot residuals against predicted values $\hat{y}_i$.
    - If you spot a **pattern -> there** may be a **problem** with some aspect of the linear model.



Non-linearity of the response-predictor relationships



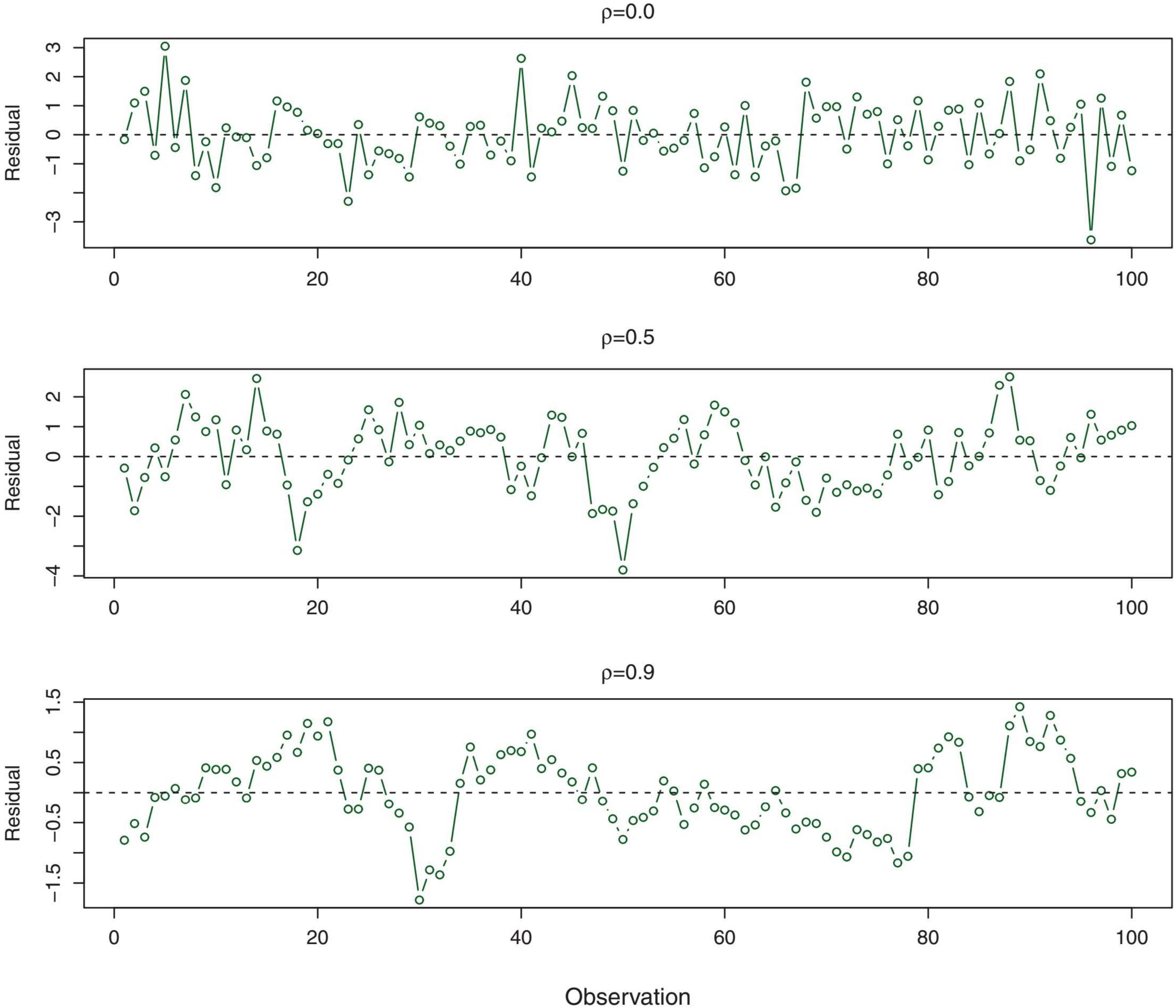
Correlation of error terms

- **It is assumed in a linear regression that the error terms $\in_1, \in_2, \ldots, \in_n$ are uncorrelated.**
- The standard errors are also computed based on this assumption.
- **If** error terms are **correlated** -> the **estimated standard errors** will tend to **underestimate** the **true standard errors.**
  - In such cases, the **prediction intervals will be narrower** than they should be.
  - One consequence could be that **a 95% confidence interval may have a much lower probability than 0.95** of containing true value of a parameter.
  - **p-values will be lower than they should be** -> may incorrectly conclude that a parameter is statistically significant.

Correlation of error terms

- **How is it possible to have correlations among the error terms?**
  - Think about **time series data**,
    - i.e., observations with measurements obtained as **discrete points in time.**
      - mostly end up with **correlated errors between adjacent observations.**
- So, we need a way to **determine if we have such correlations in our data!**
  - One way is to **plot residuals from the model against time.**
    - **If no pattern observed -> errors are uncorrelated.**
    - **If they are positively correlated, we say that there is a tracking in the residuals.**

Correlation of error terms

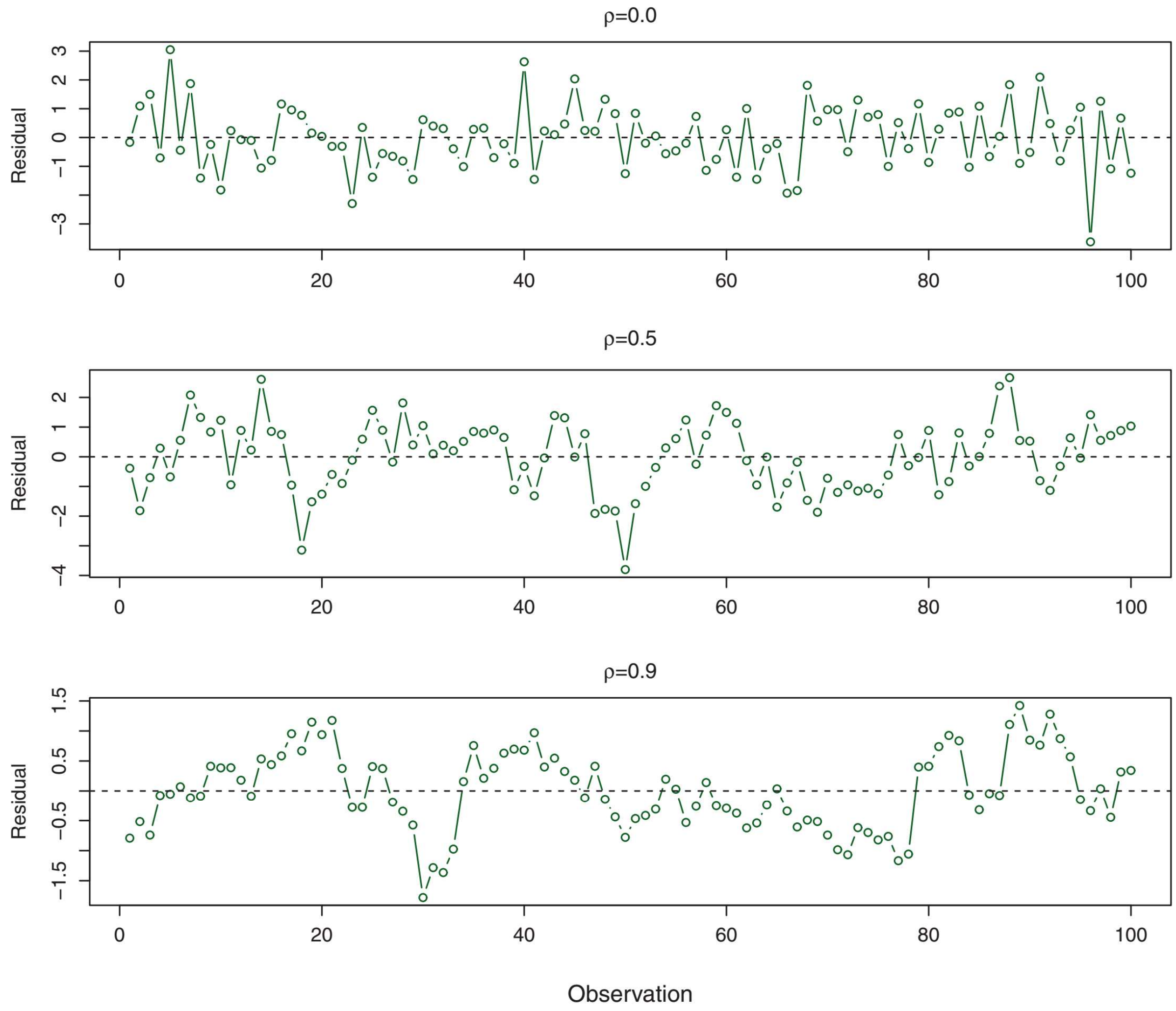


uncorrelated

0.5

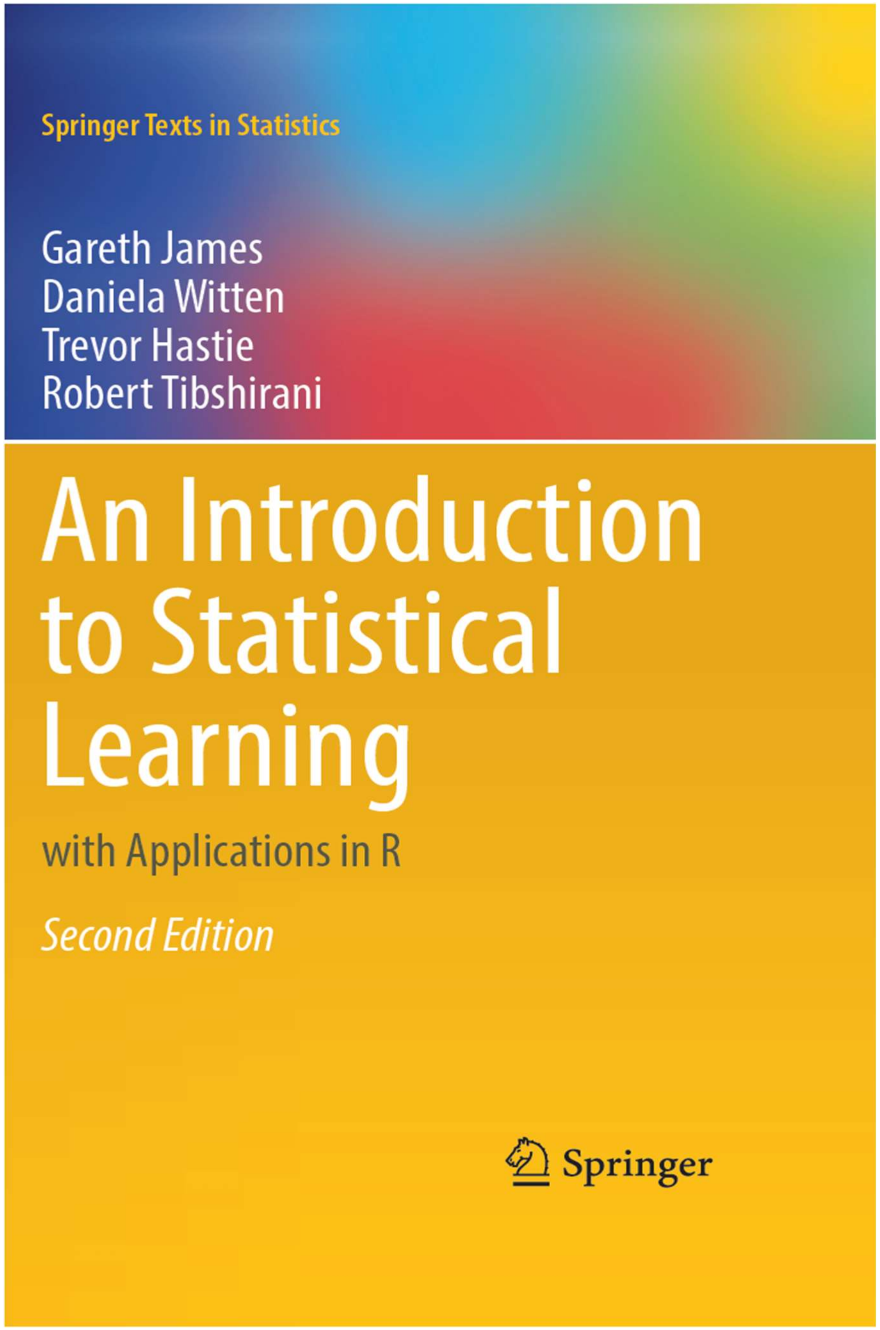
0.9

Correlation of error terms



- Such correlations could result from factors **other than time series, e.g.?**
- In general, a good statistical design seeks to **ensure that errors are uncorrelated, starting from data collection**.

Reference


Springer Texts in Statistics
Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani
An Introduction to Statistical Learning
with Applications in R
Second Edition
Springer