Fall 2023

BIF524/CSC463 Data Mining

Linear Regression
Logistic Regression

Eileen Marie Hanna, *PhD*

05/10/2023

Correlation of error terms

- **It is assumed in a linear regression that the error terms $\in_1, \in_2, \dots, \in_n$ 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.**

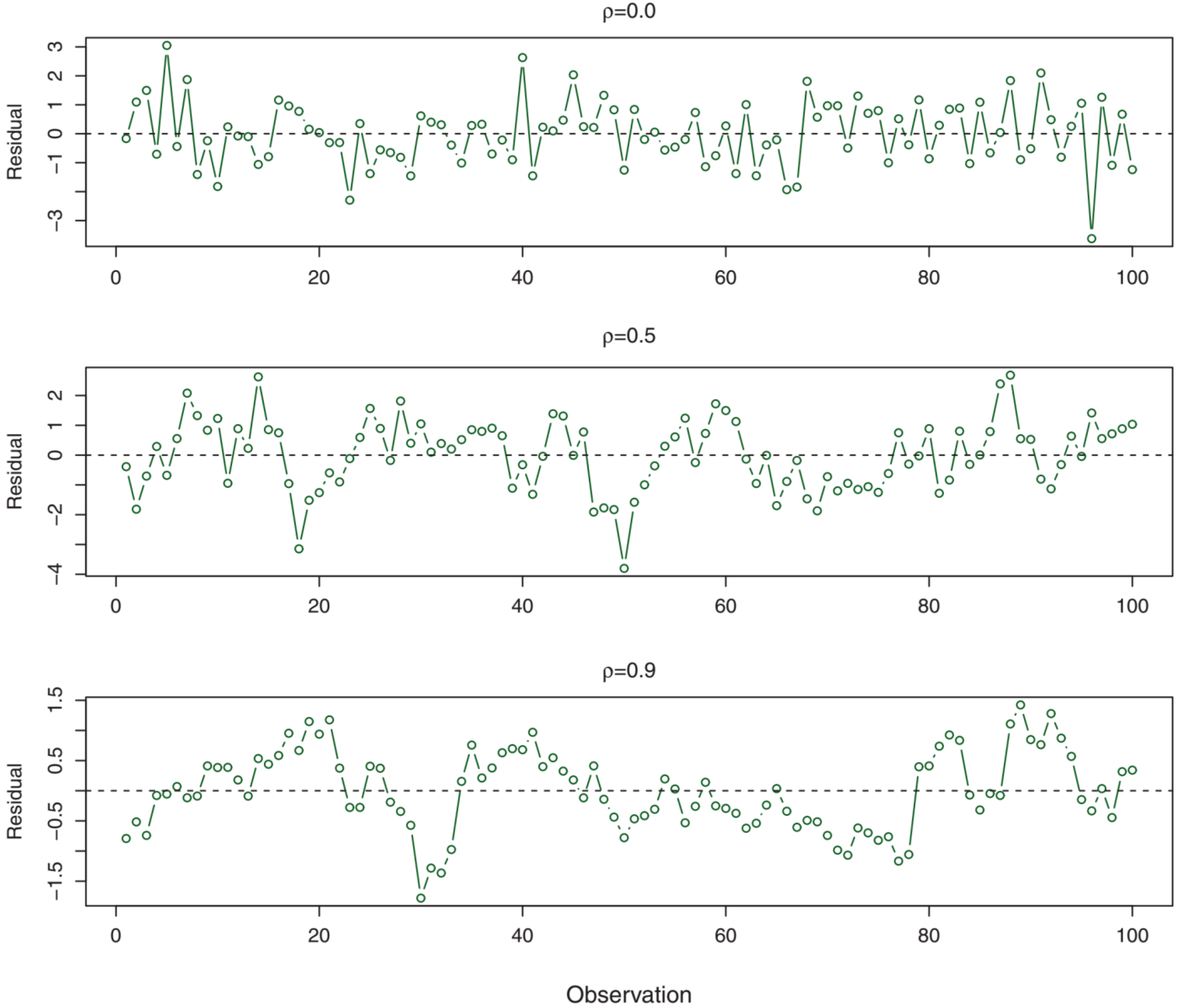
- So, we need a way to **determine if we have such correlations in our data!**

  - 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

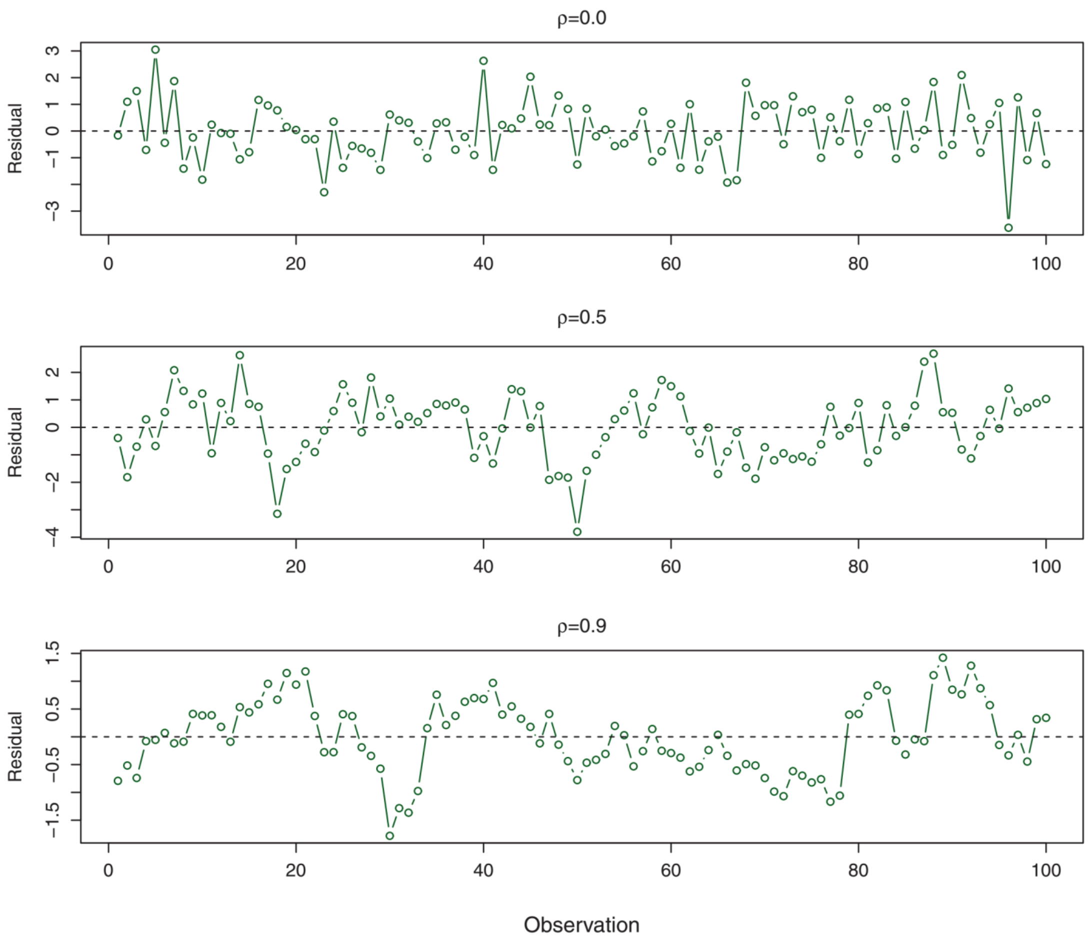


uncorrelated

0.5

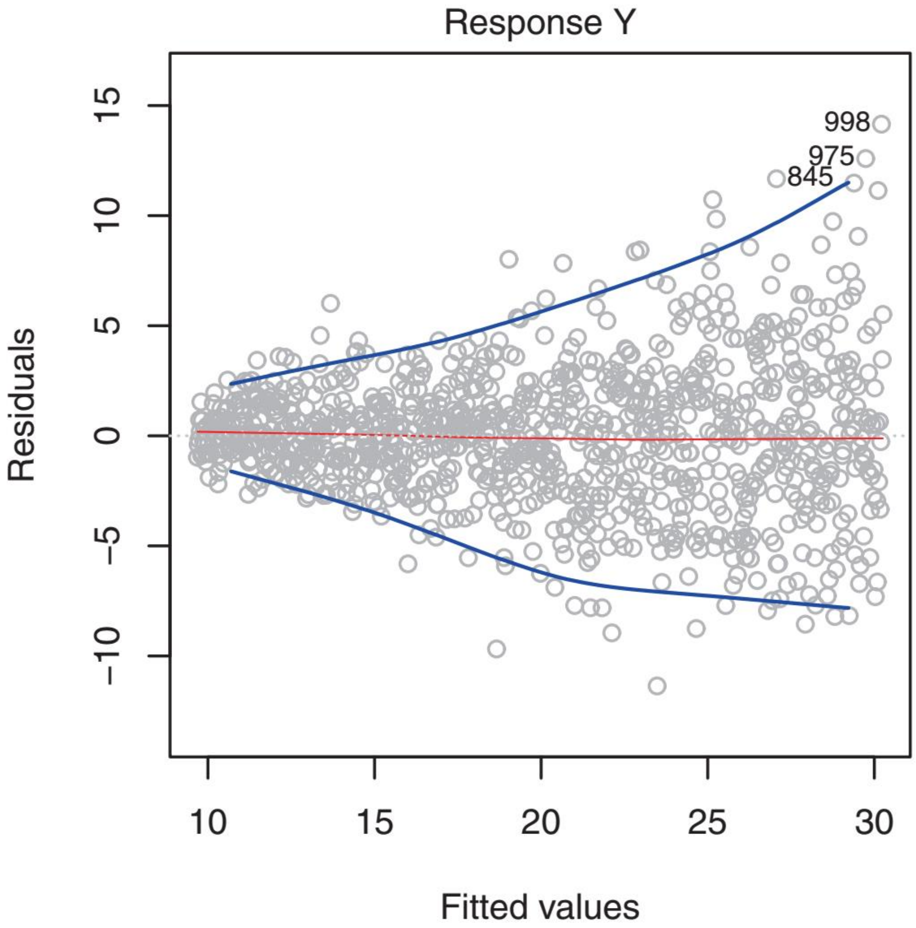
0.9

Correlation of error terms



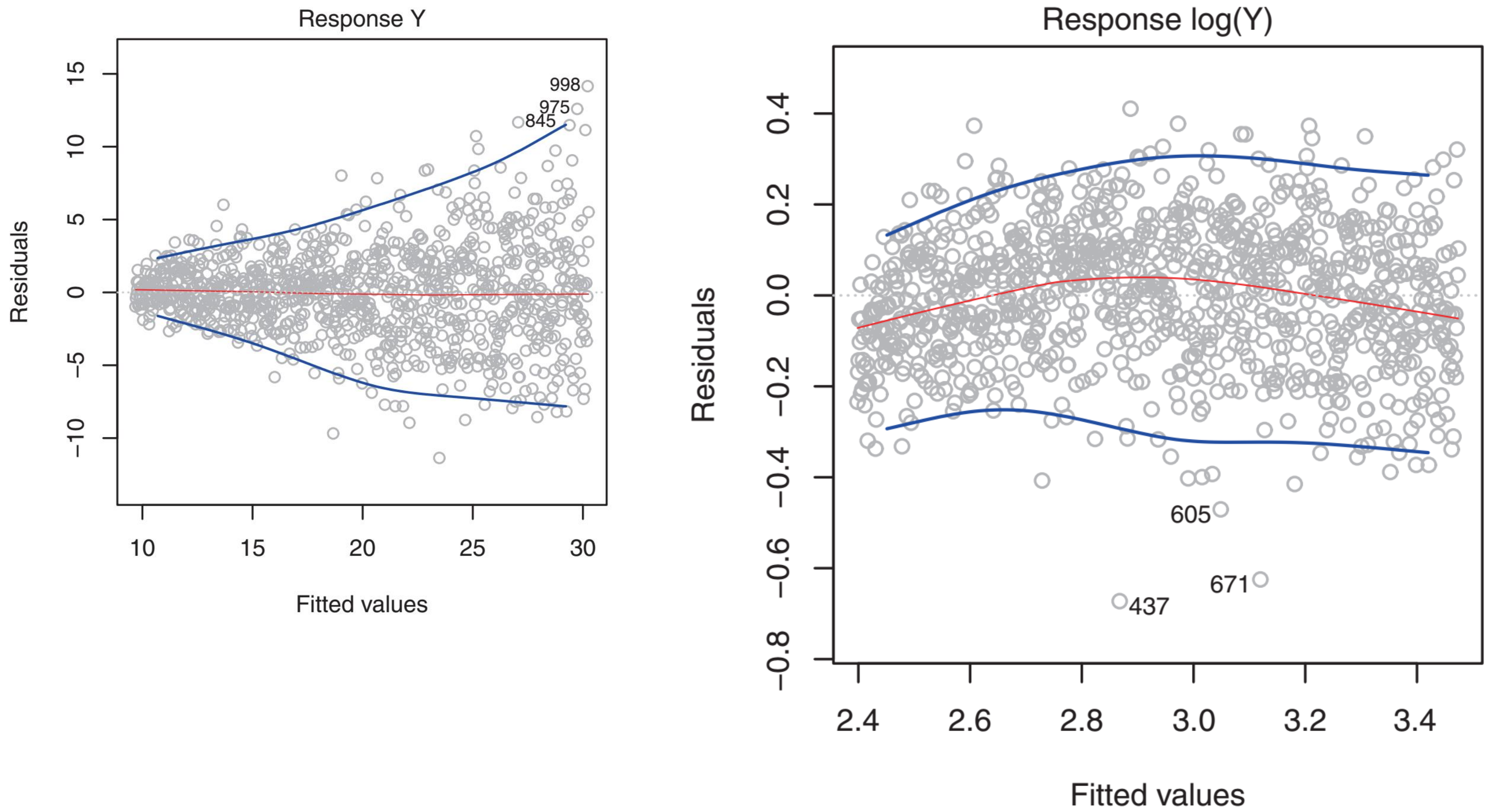
- Such correlations could result from factors **other than time series, e.g.?**
- In general, a good statistical design seeks to **ensure that errors are uncorrelated**, **starting from data collection**.

Non-constant variance of error terms



- **A linear model also assumes that the errors have a constant variance**, $Var(\epsilon_i) = \sigma^2$.
- However, variances of errors terms tend to **often** be **non-constant**.
- This leads to **heteroscedasticity** from the presence of a **funnel shape** in the residual plot.
- Here, the magnitude of the **residuals tend to increase with the fitted values**.
- One solution is to **transform Y to a concave function**, e.g., $logY$ or $\sqrt{Y}$.

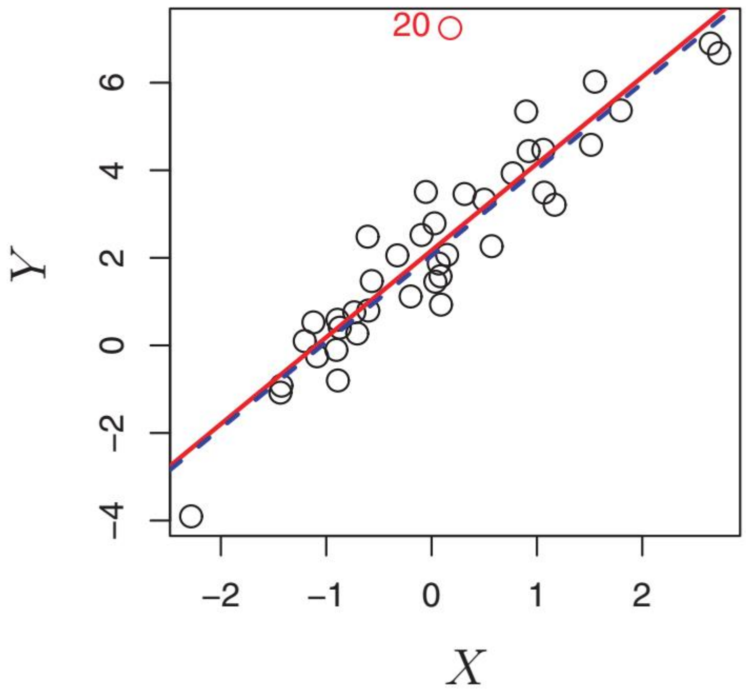
Non-constant variance of error terms



Constant variance with slight evidence of non-linear relationship
The residuals now appear to have constant variance, though there is some evidence of a slight non-linear relationship in the data.

Outliers

- Points that are very far from the predicted value by the model.



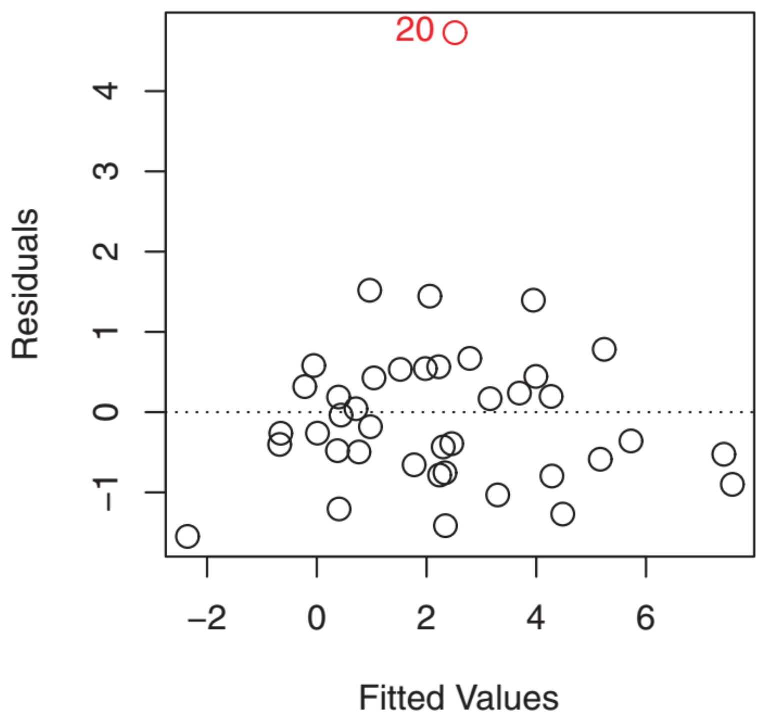
least squares regression fit

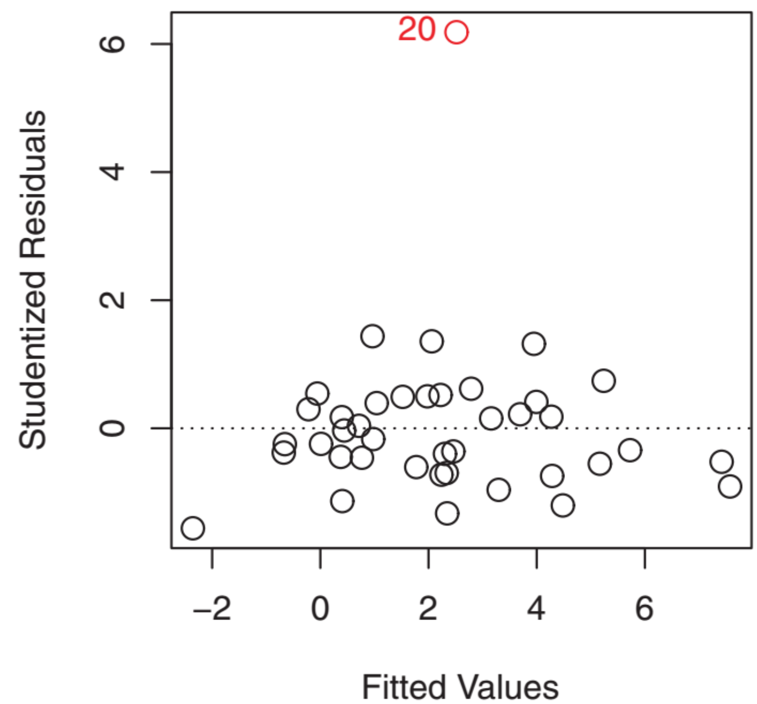
least squares regression fit after removing the outlier

In this case, it has a small effect on the fit, but an effect is shown in RSE (1.09 vs 0.77) and R^2 (0.892 vs 0.805).

Outliers

- A residual plot can help spot outliers.
- But how far is enough to consider a point as outlier?

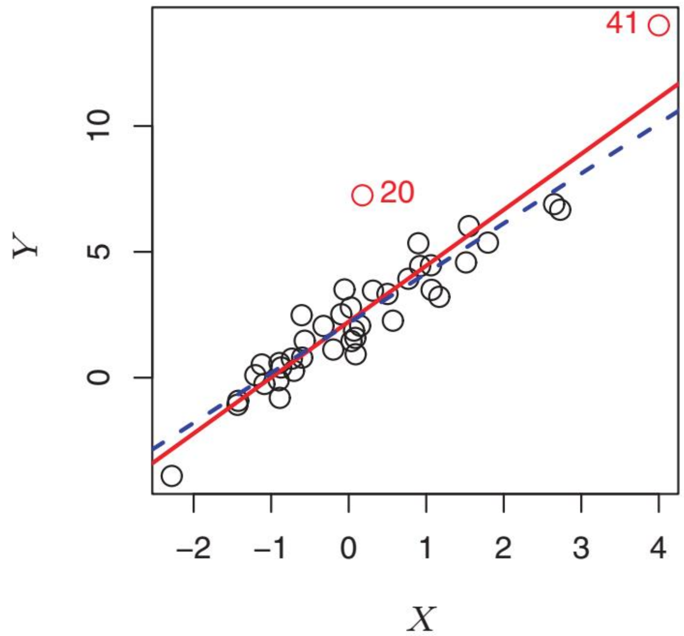




A **studentized residuals** (**residual divided by an estimate of its standard deviation**) plot where **each residual divided by its estimated standard error**.

Values with studentized residuals great than 3 in absolute value -> outliers.

High leverage points



regression fit

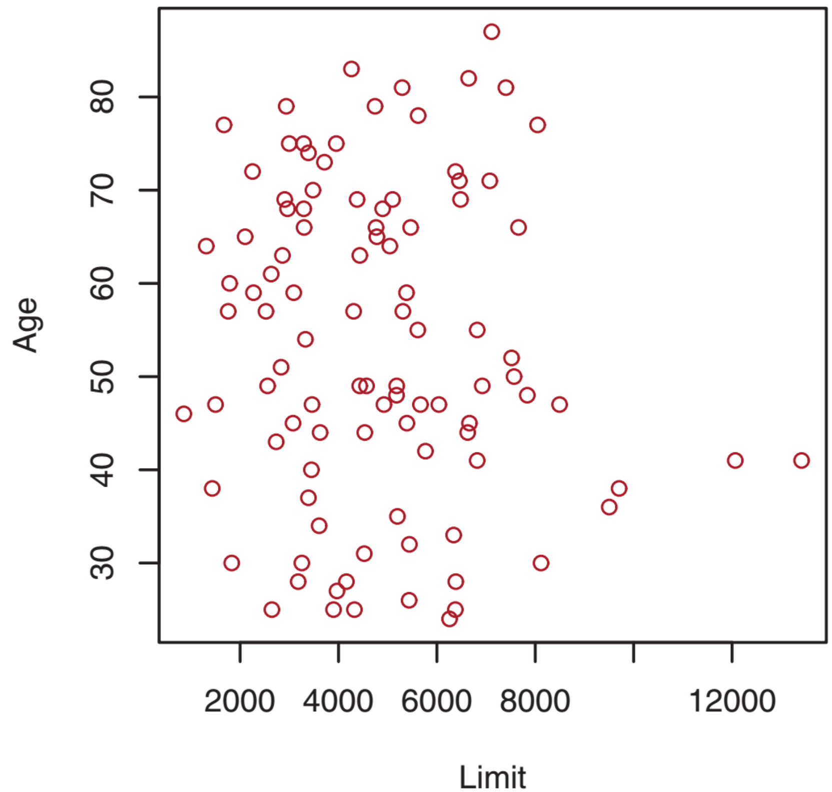
regression fit after removing the obs. 41

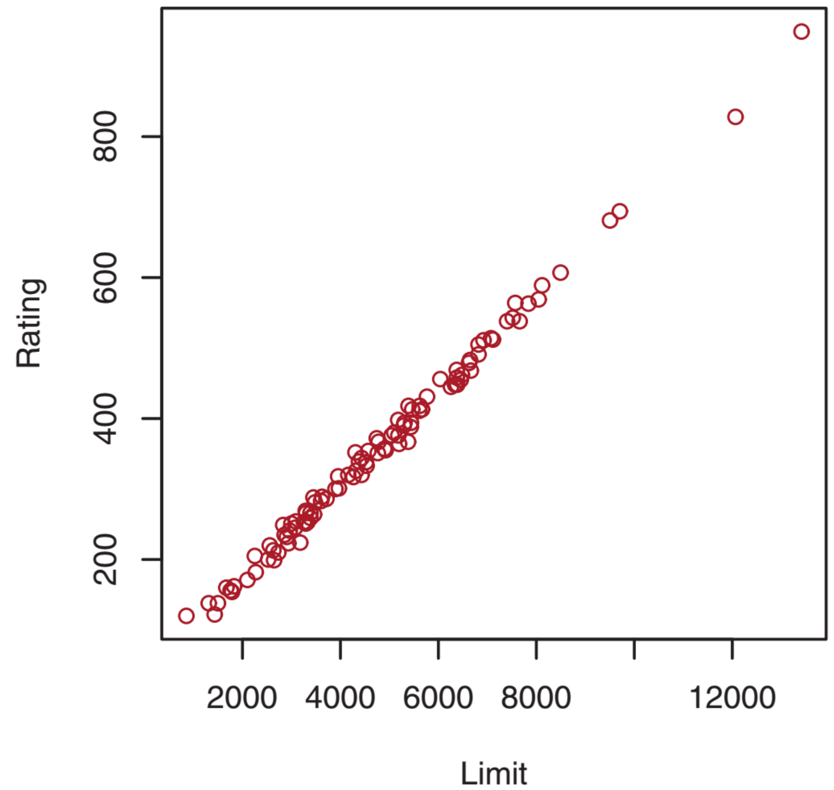
- Observations with high leverages often have high impact on the fitted line.
- A certain observation could be either an outlier or of high leverage, or both.
- One measure is the leverage statistic, here for simple regression:

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

large value -> high leverage

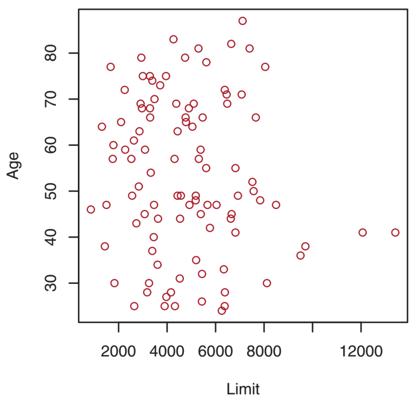
Collinearity

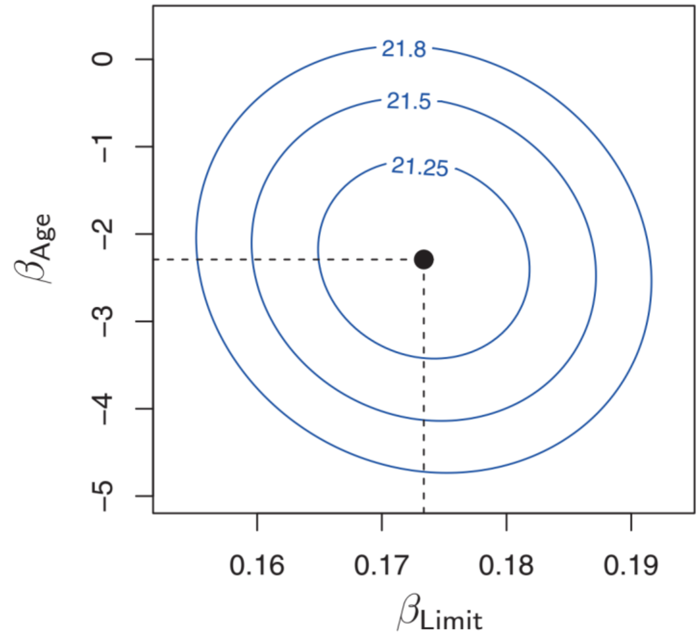




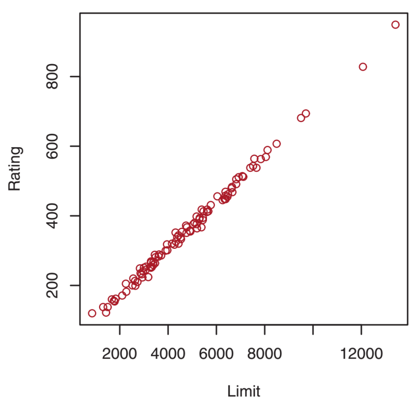
results from predictors that are highly correlated -> may lead to difficulties in differentiating the effect of each predictor on the response.

Collinearity

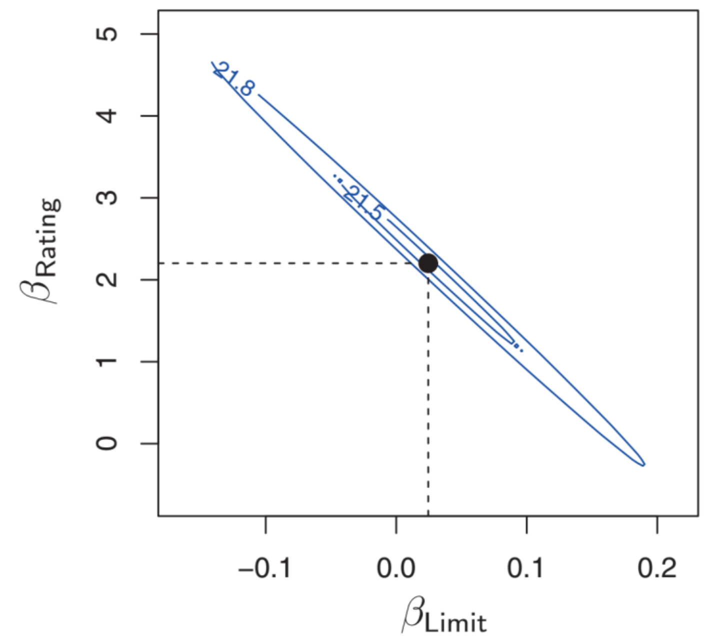




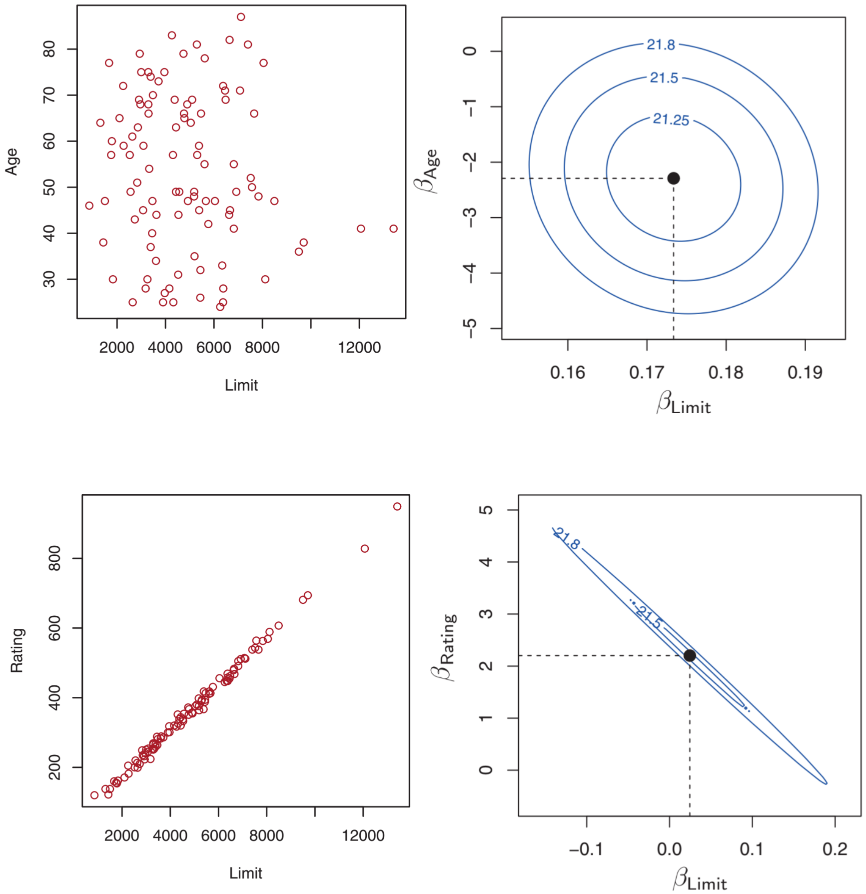
Balance against age and limit coefficients. black dots correspond to lowest RSS.



Multiple points may correspond to same RSS for correlated predictors.



Collinearity



- One way to identify such cases is to examine the **correlation matrix** of the predictors.
- But **sometimes multiple variables can be correlated** (multicollinearity) even if they show no pairwise correlation.
- Variance inflation factor (VIF):

$$\mathrm{VIF}(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j | X_{-j}}}$$

VIF has a minimum value of 1 indictating no collinearity. values greater than 5 indicate a problematic collinearity.

Linear regression belongs to the category of **parametric methods**.

\+ easy to fit – relatively small number of coefficients to predict
simple interpretation
statistical measures/tests

\- strict assumptions on the form of $f(X)$. If far from the true trend -> low prediction accuracy -> erroneous conclusions

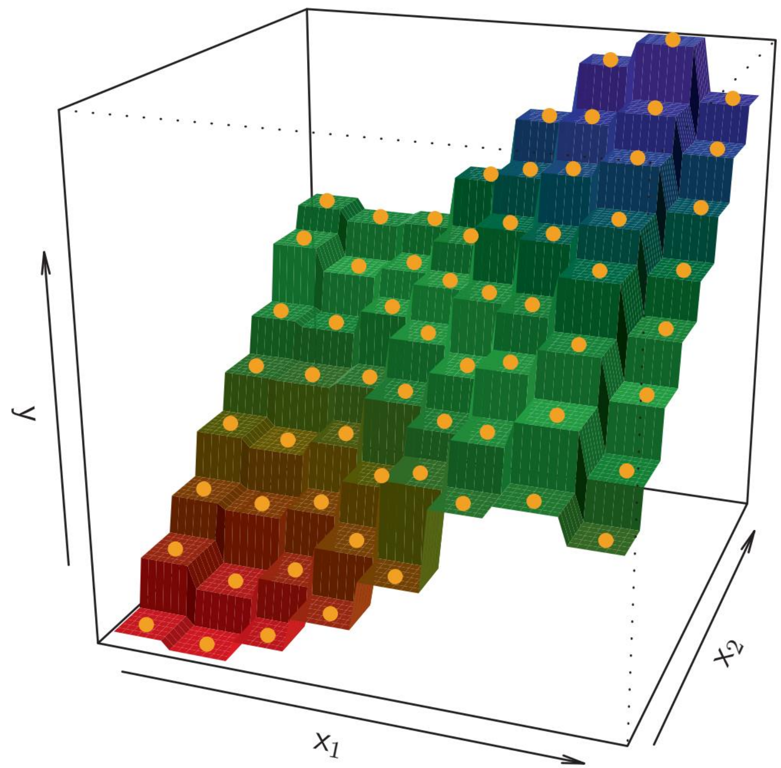
Non-parametric methods do not assume a parametric form of $f(X)$ -> more flexible in performing regression. e.g.,

K-nearest neighbors regression (KNN)

K-nearest neighbors regression (KNN)

- Similar to the concept of the K-nearest neighbors classifier.
- Given a value of K and x_0, KNN regression:
  - first identifies the K training observations that are closest to x_0 (forming set $\mathcal{N}_0$).
  - then estimates $f(x_0)$ as the average of training responses in $\mathcal{N}_0$:

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



$K = 1$

Task

- Predict the medical condition of a patient admitted to the emergency room, based on symptoms.

- Suppose that there are three possibilities:

$$Y = \begin{cases} 1 & \text{if } \texttt{stroke}; \\ 2 & \text{if } \texttt{drug overdose}; \\ 3 & \text{if } \texttt{epileptic seizure}. \end{cases}$$

Linear regression?

$$Y = \begin{cases} 1 & \text{if } \texttt{stroke}; \\ 2 & \text{if } \texttt{drug overdose}; \\ 3 & \text{if } \texttt{epileptic seizure}. \end{cases}$$

- With this quantitative encoding of the response, a linear model depicts the relationship between Y and the set of predictors (here symptoms) $X_1, X_2, \ldots, X_p$.

Linear regression?

$$Y = \begin{cases} 1 & \text{if } \texttt{stroke}; \\ 2 & \text{if } \texttt{drug overdose}; \\ 3 & \text{if } \texttt{epileptic seizure}. \end{cases}$$

- Why not

$$Y = \begin{cases} 1 & \text{if } \texttt{epileptic seizure}; \\ 2 & \text{if } \texttt{stroke}; \\ 3 & \text{if } \texttt{drug overdose}. \end{cases}$$

- What is the **ordering** of those responses?
  - Are the **differences** between these values meaningful?
- In case of ordered categories, can the difference between categories be always **quantified?**
- Fundamentally different linear models will be generated from such encodings!

Linear regression?

Generally, we cannot convert a qualitative response with more than two levels into a quantitative response that is ready for linear regression!

For a two-level qualitative response

- More applicable
- We can use the dummy variable approach to code the response, e.g.,

$$Y = \begin{cases} 0 & \text{if } \texttt{stroke}; \\ 1 & \text{if } \texttt{drug overdose}. \end{cases}$$

- Linear regression can thus predict drug overdose if $\hat{Y} > 0.5$ and stroke otherwise.

- $X\hat{B}$ is actually equivalent to:

$$\Pr(\texttt{drug overdose}|X)$$

- Inverting the encoding will eventually lead to the same predictions.

“Default” dataset

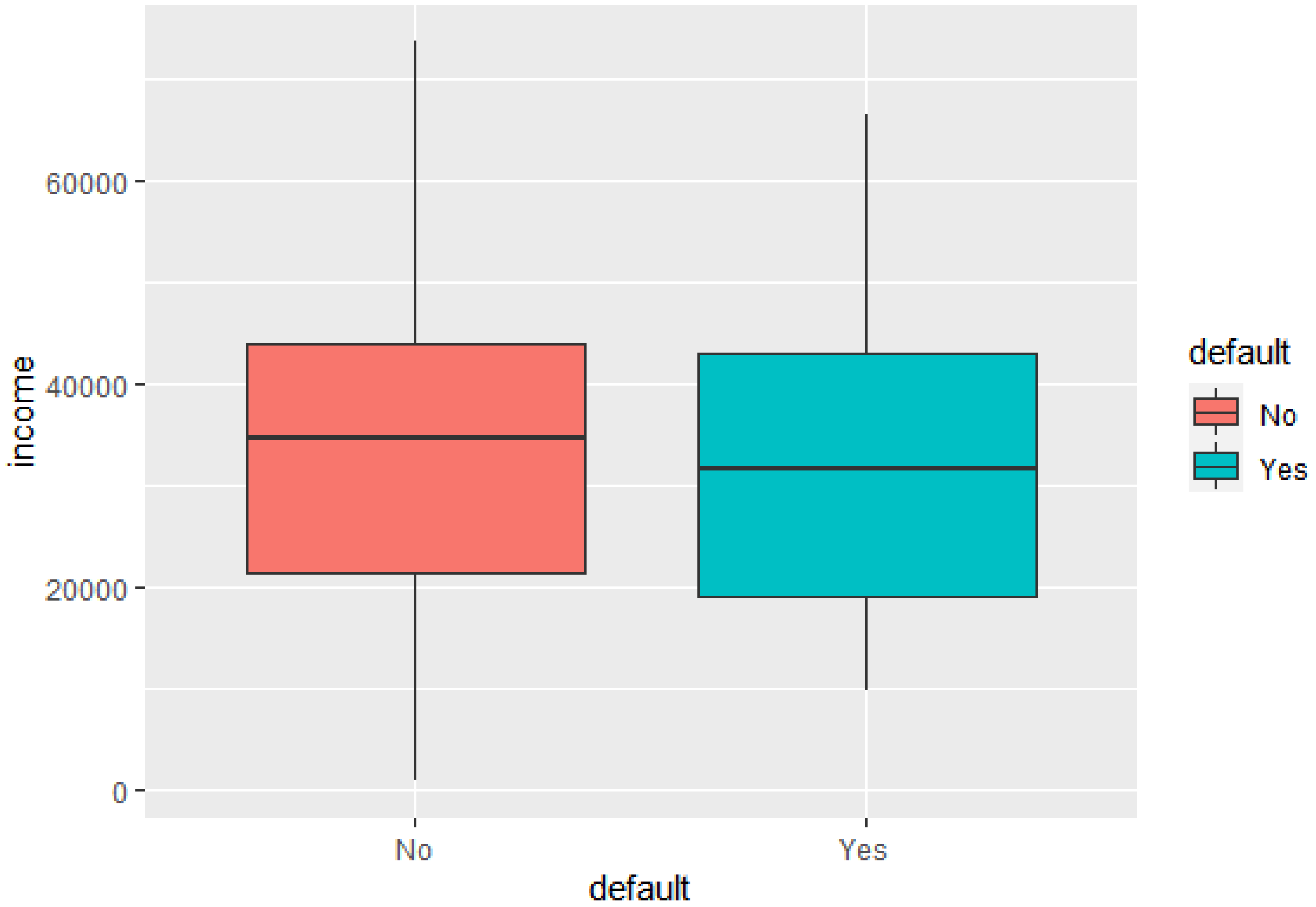
```
> glimpse(df)
Rows: 10,000
Columns: 4
$ default <fct> No, No, No, No, No, No, No, No, No, No, No, No, No, No, No, No…
$ student <fct> No, Yes, No, No, No, Yes, No, Yes, No, No, Yes, Yes, No, No, N…
$ balance <dbl> 729.5265, 817.1804, 1073.5492, 529.2506, 785.6559, 919.5885, 8…
$ income  <dbl> 44361.625, 12106.135, 31767.139, 35704.494, 38463.496, 7491.55…
```

“Default” dataset

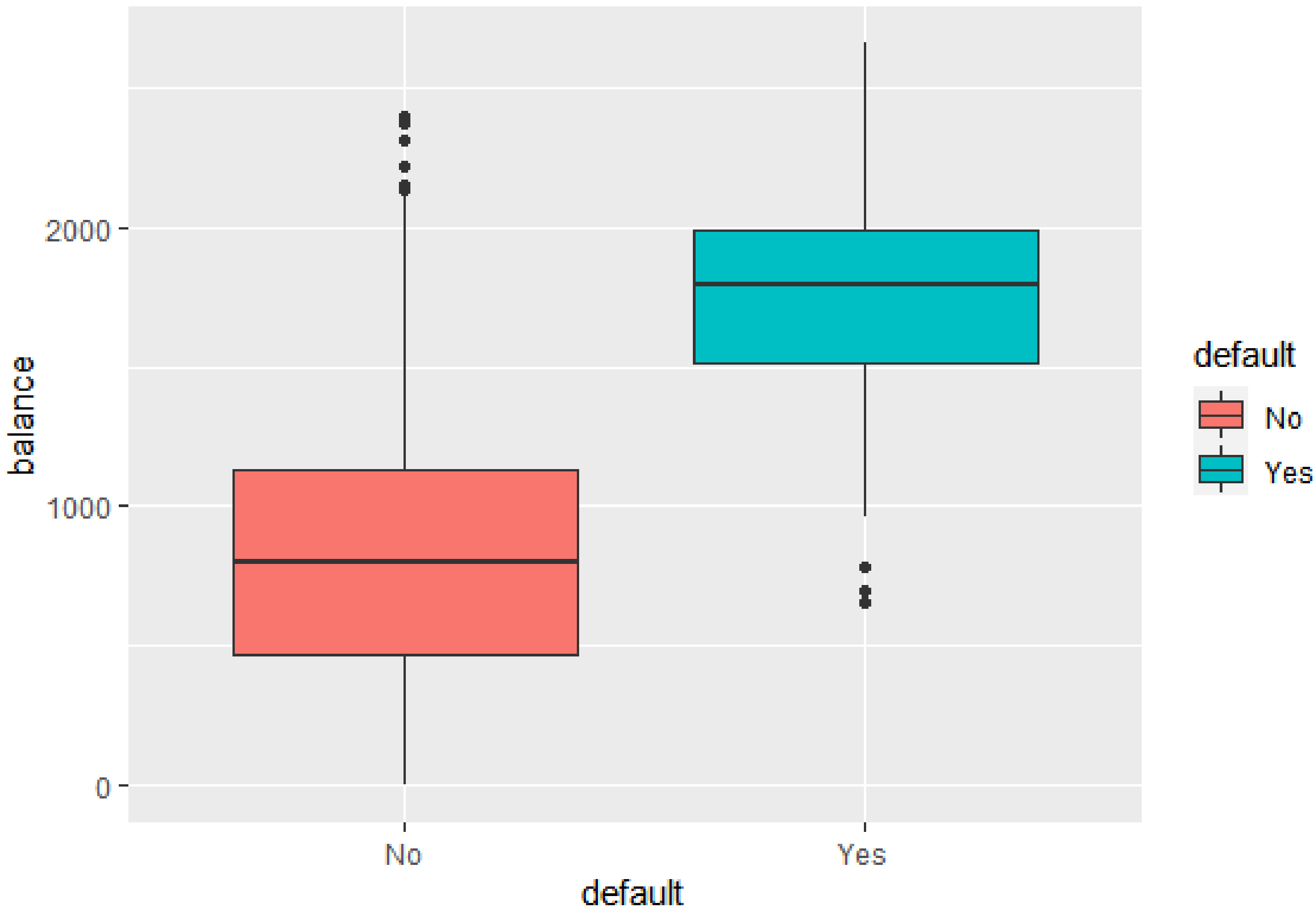
Will a person default on his/her credit card payment, based on annual income and monthly credit card balance?

What does this plot tell us?

"Default" dataset



“Default” dataset



Logistic regression

- Models the probability that Y belongs to a particular category.

$$p(\texttt{balance}) \equiv \Pr(\texttt{default} = \texttt{Yes}|\texttt{balance})$$

- The values of p(balance) will fall between $\mathbf{0}$ and $\mathbf{1}$.

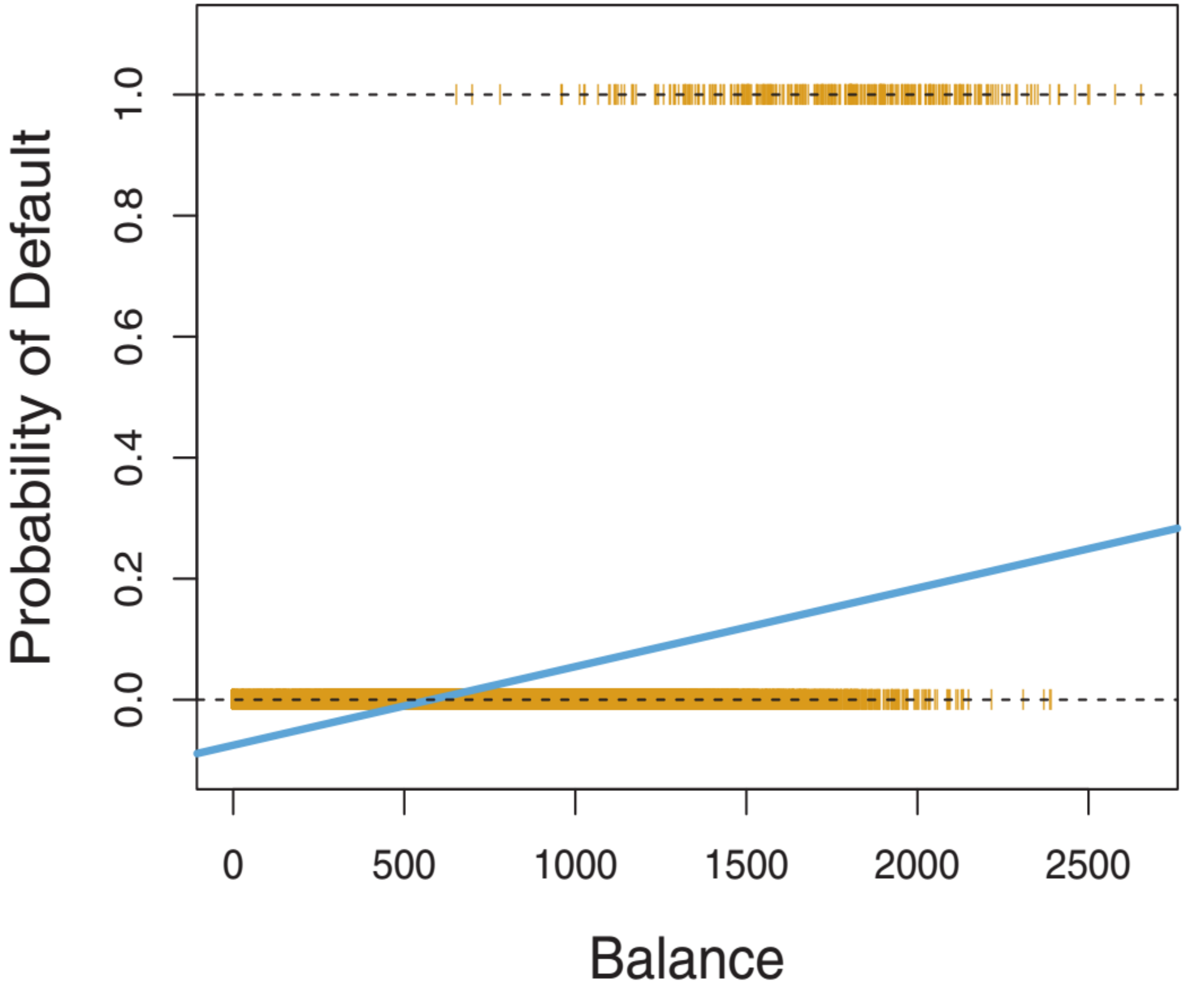
- We might thus predict that a person will default if the corresponding $\boldsymbol{p}(\mathbf{balance})>\mathbf{0.5}$.

- Stricter thresholds could be assigned.

Logistic regression

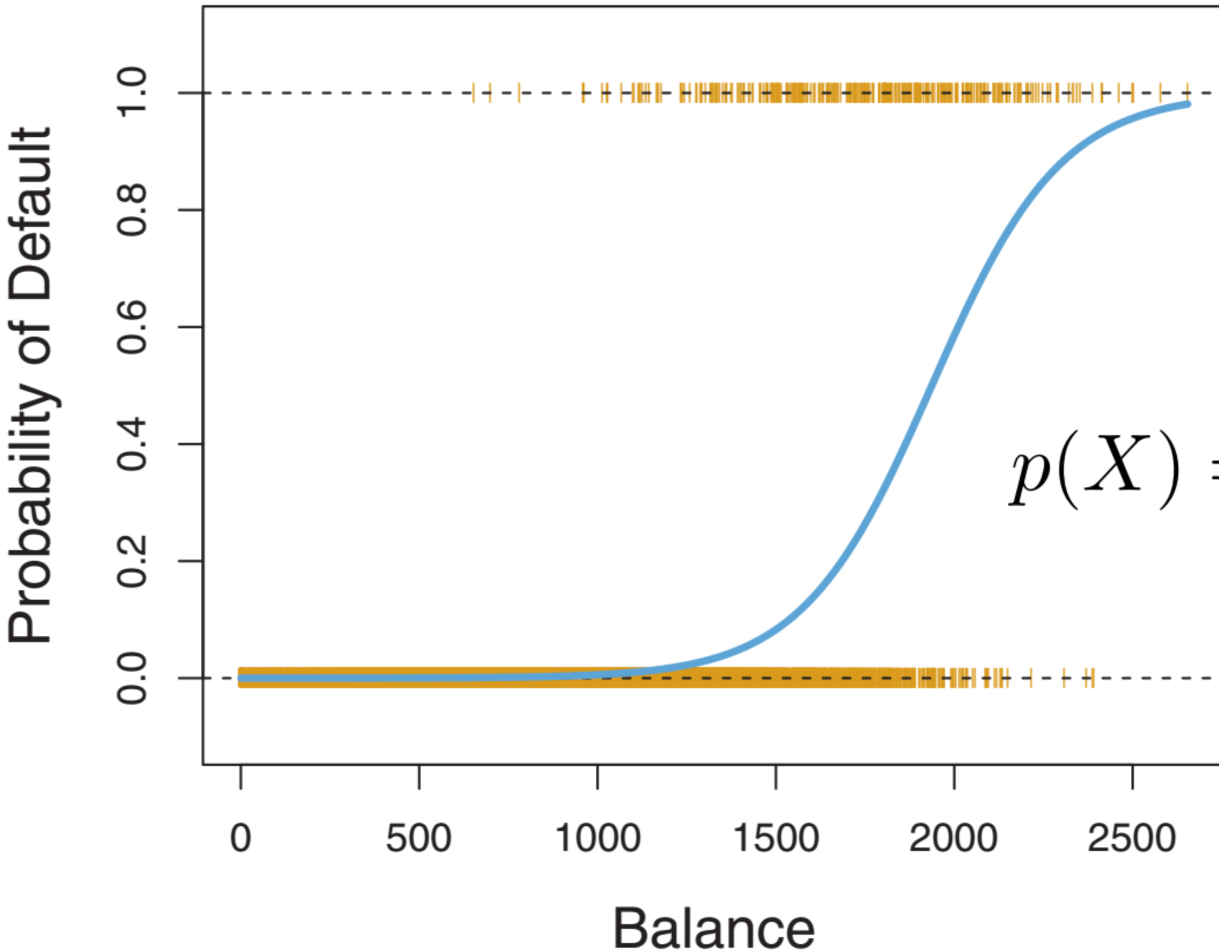
- A linear regression model in this context would be:

$$p(X) = \beta_0 + \beta_1 X$$



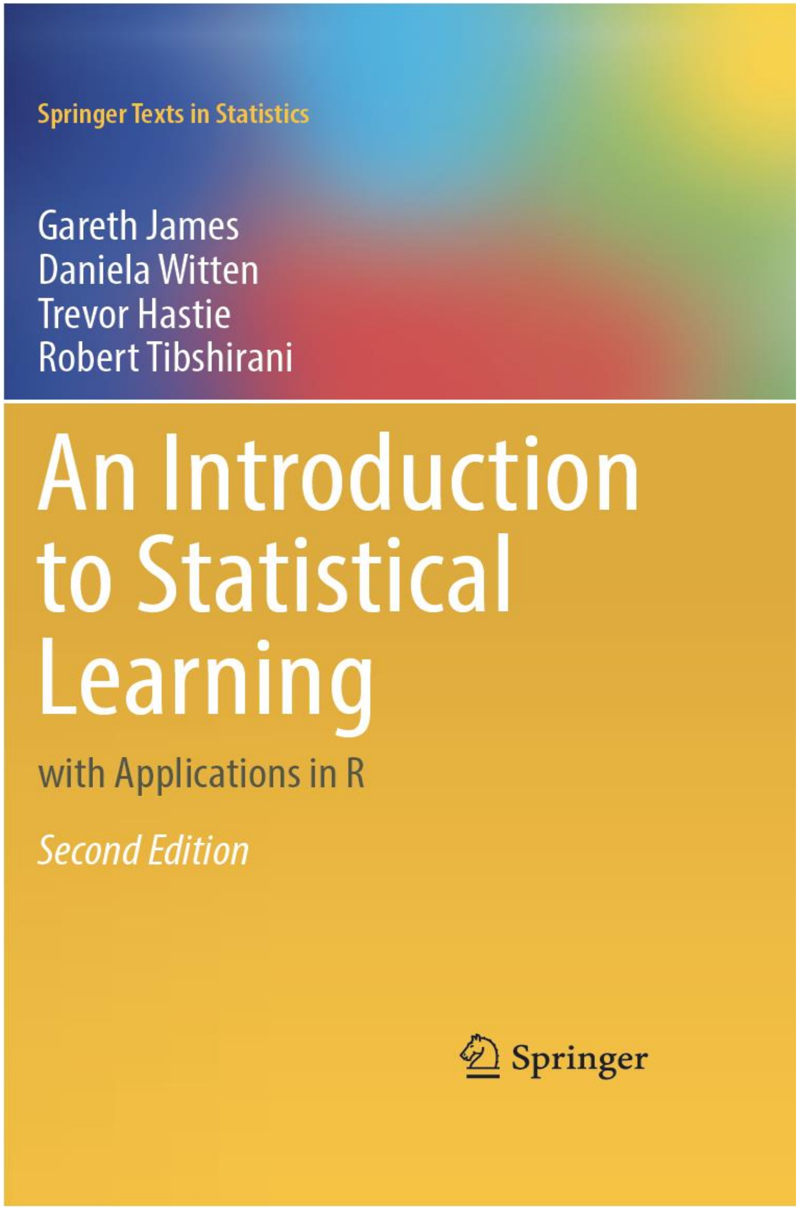
Logistic regression

- Logistic regression uses the **logistic function** which output values of $\boldsymbol{p}(\boldsymbol{X})$ between $\mathbf{0}$ and $\mathbf{1}$, unlike linear regression.



$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

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