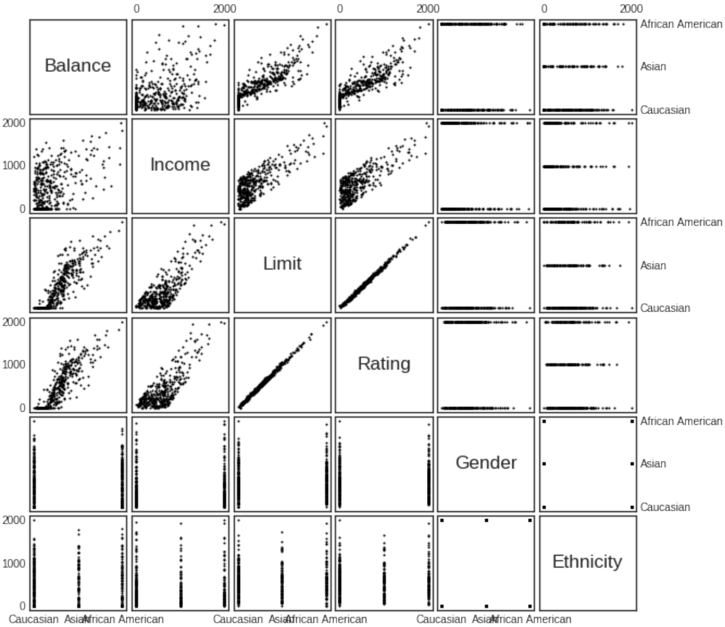
Linear Regression 3
DS351

Credit balance data



Predictor with two levels

Find the difference in credit card balance (y_i) between **male** and **female**.

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

Predictor with two levels

Find the difference in credit card balance (y_i) between **male** and **female**.

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

Create a **dummy variable** x_i:

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

Using x_i, the regression can be written as

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Estimates of coefficients

	$\hat{\beta}_i$	SE($\hat{\beta}_i$)	t-statistic	p-value
Intercept	509.80	33.13	15.389	<0.0001
gender(Female)	19.73	46.05	0.429	0.6690

$$\hat{y}_i = 509.80 + 19.73x_i.$$

Main takeaway:

- Male has credit card debt of 509.80 **on average**.
- Female has credit card debt of 509.80+19.73 = 529.53 **on average**.
- The difference in credit card debt is $\hat{\beta}_1 = 19.73$ **on average**.

Estimates of coefficients

	$\hat{\beta}_i$	SE($\hat{\beta}_i$)	t-statistic	p-value
Intercept	509.80	33.13	15.389	<0.0001
gender(Female)	19.73	46.05	0.429	0.6690

$$\hat{y}_i = 509.80 + 19.73x_i.$$

Main takeaway:

- Male has credit card debt of 509.80 **on average**.
- Female has credit card debt of 509.80+19.73 = 529.53 **on average**.
- The difference in credit card debt is $\hat{\beta}_1 = 19.73$ **on average**.

Question: Can we conclude that females have more credit debt on average than males?

Predictor with more than two levels

Find the difference in credit card balance (y_i) between **Asian**, **Caucasian** and **African American**.

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

Predictor with more than two levels

Find the difference in credit card balance (y_i) between **Asian**, **Caucasian** and **African American**.

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

Create two **dummy variables** x_{i1} and x_{i2} :

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

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

Predictor with more than two levels

Find the difference in credit card balance (y_i) between **Asian**, **Caucasian** and **African American**.

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

Create two **dummy variables** x_{i1} and x_{i2} :

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

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

Using x_{i1} and x_{i2}, the regression can be written as

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

Estimates of coefficients

	$\hat{\beta}_i$	SE($\hat{\beta}_i$)	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

Main takeaway: **On average,**

- African American has credit debt of 531.00 .
- Asian has 18.69 less debt than the African American.
- Caucasian has 12.50 less debt than the African American.
- Asian has ________________ less debt than Caucasian.

Estimates of coefficients

	$\hat{\beta}_i$	SE($\hat{\beta}_i$)	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

Main takeaway: **On average,**

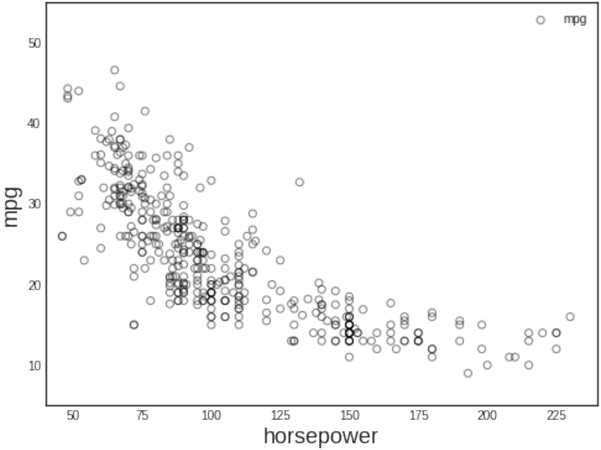
- African American has credit debt of 531.00 .
- Asian has 18.69 less debt than the African American.
- Caucasian has 12.50 less debt than the African American.
- Asian has ________________ less debt than Caucasian.

Question: How can we decide if there is any difference in credit card balance between the ethnicities?

Linear model diagnosis

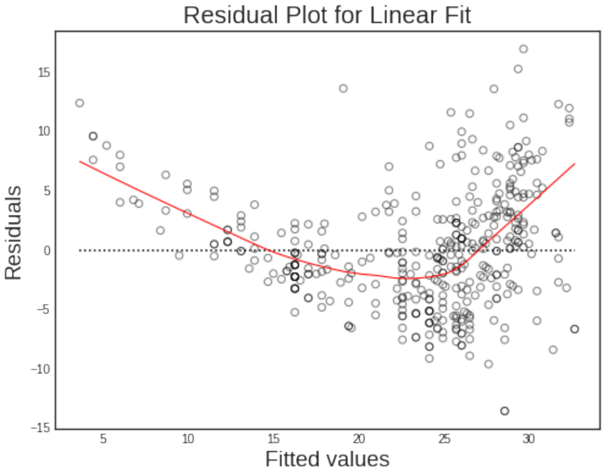
1. Non-linearity of the data

- Maybe the relationship between the predictors and the response is non-linear.



Residual plot

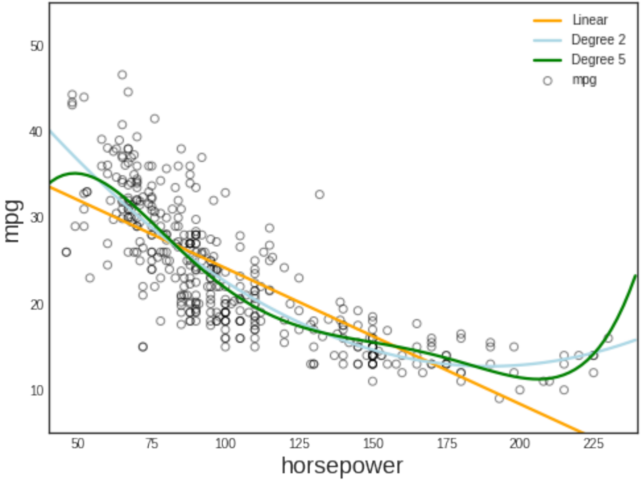
- Plot between the **fitted values** $\hat{y}_i$ and the **residuals** $y_i - \hat{y}_i$.



Non-linear regression

Try a polynomial function of the horsepower:

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



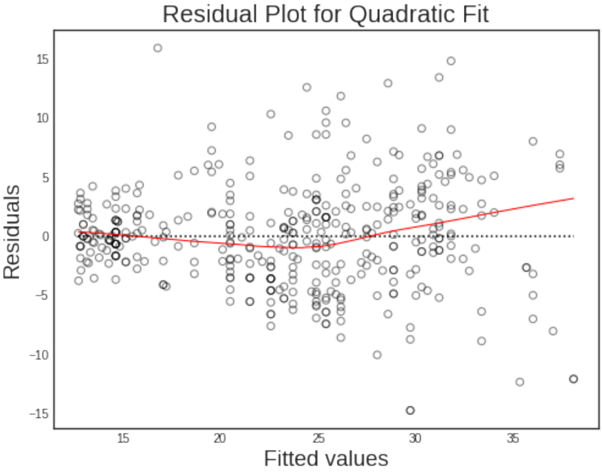
Estimates of coefficients

	$\hat{\beta}_i$	$\mathrm{SE}(\hat{\beta}_i)$	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

Two things indicate that the quadratic fit is better:

- The p-value of `horsepower`^{2} is significant.
- The R^2 of this model is 0.688 compared to 0.606 of the linear model.

Residual plot of non-linear regression



The pattern disappears

2. Correlation of error terms

- We assumed that the error terms

$$\epsilon_1, \epsilon_2, \ldots, \epsilon_n$$

  are independent to each other. This is an important assumption!
- What happens if this is not the case?

Example: Suppose we accidentally doubled the data

$$(x_1, y_1), (x_1, y_1), (x_2, y_2), (x_2, y_2), \ldots$$

and train the simple linear model

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \epsilon_i.$$

2. Correlation of error terms

Recall that the standard error of a coefficient is

$$\text{Model 2:} \qquad \mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{2n}(x_i - \bar{x})^2} \qquad (2n \text{ points})$$

compared to

$$\text{Model 1:} \qquad \mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (n \text{ points})$$

2. Correlation of error terms

Recall that the standard error of a coefficient is

$$\text{Model 2:} \qquad \mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{2n}(x_i - \bar{x})^2} \quad (2n \text{ points})$$

compared to

$$\text{Model 1:} \qquad \mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad (n \text{ points})$$

- The standard error of Model 2 is $\sqrt{2}$ times small than that of Model 1.

2. Correlation of error terms

Recall that the standard error of a coefficient is

$$\text{Model 2:} \qquad \text{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{2n}(x_i - \bar{x})^2} \qquad (2n \text{ points})$$

compared to

$$\text{Model 1:} \qquad \text{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad (n \text{ points})$$

- The standard error of Model 2 is $\sqrt{2}$ times small than that of Model 1.
- The confidence interval

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

is $\sqrt{2}$ times narrower.

2. Correlation of error terms

- From previous example, we learn that **correlated errors cause the confidence interval to be narrower.**
- As a result, we could mistakenly conclude that the coefficients are significant.

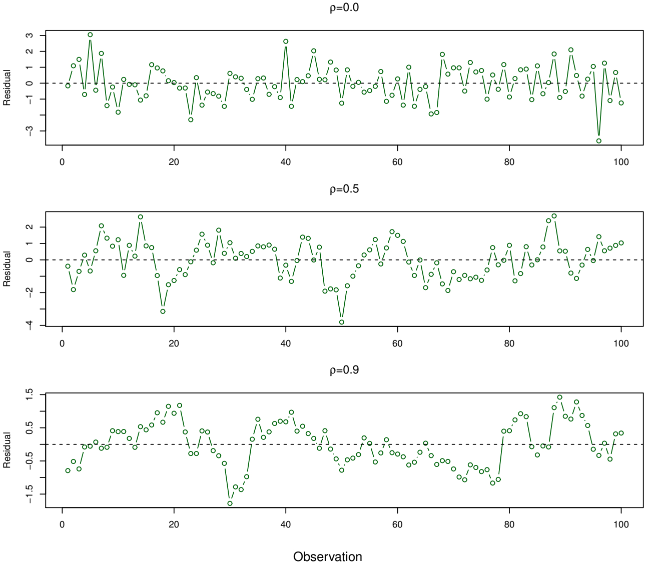
2. Correlation of error terms

- From previous example, we learn that **correlated errors cause the confidence interval to be narrower.**
- As a result, we could mistakenly conclude that the coefficients are significant.
- How can we detect correlated errors?
- Hard to detect in general, easier if we are studying **time series**.

2. Correlation of error terms

- From previous example, we learn that **correlated errors cause the confidence interval to be narrower.**
- As a result, we could mistakenly conclude that the coefficients are significant.
- How can we detect correlated errors?
- Hard to detect in general, easier if we are studying **time series**.
- Detect by looking at the **time vs residual** plot.

Time vs residual plot



3. Non-constant variance of error terms

- We also assumed that the variance of $\text{Var}(\epsilon_i) = \sigma^2$ for all i.
- The formula for standard error, hypothesis test and confidence interval are all derived **under this assumption**.

3. Non-constant variance of error terms

- We also assumed that the variance of $\text{Var}(\epsilon_i) = \sigma^2$ for all i.
- The formula for standard error, hypothesis test and confidence interval are all derived **under this assumption**.
- For example, the formula

$$\text{Cov}\hat{\boldsymbol{\beta}} = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

  holds because we assumed that ϵ_i's share the same variance σ^2.

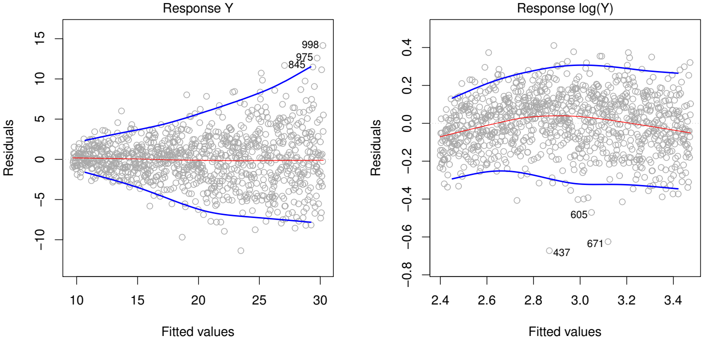
3. Non-constant variance of error terms

- We also assumed that the variance of $\mathrm{Var}(\epsilon_i) = \sigma^2$ for all i.
- The formula for standard error, hypothesis test and confidence interval are all derived **under this assumption**.
- For example, the formula

  $$\mathrm{Cov}\hat{\boldsymbol{\beta}} = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$$

  holds because we assumed that ϵ_i's share the same variance σ^2.
- Again, detect non-constant variance using **fitted value vs residual plot**.

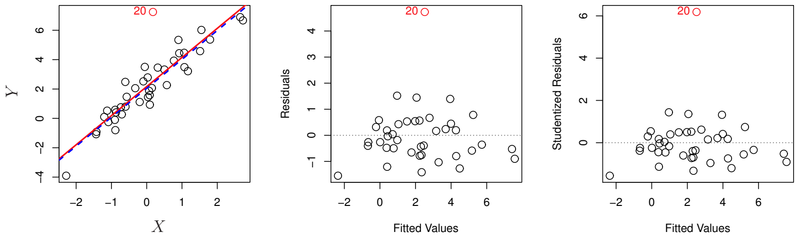
Fitted value vs residual plot



- The variance increases as the fitted value increases.
- Try transformation $Y \to \log(Y)$ or $Y \to \sqrt{Y}$ before training the model.

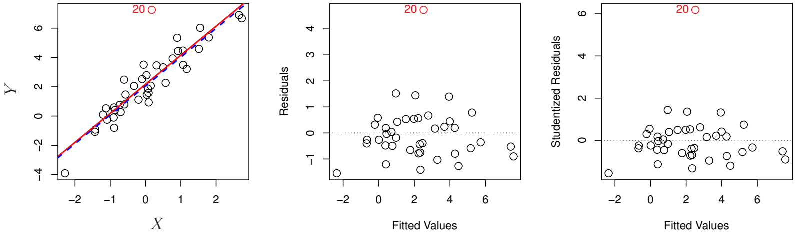
4. Outliers

We can detect outliers with actual data plot (single variable) or the residual plot (multiple variables).



4. Outliers

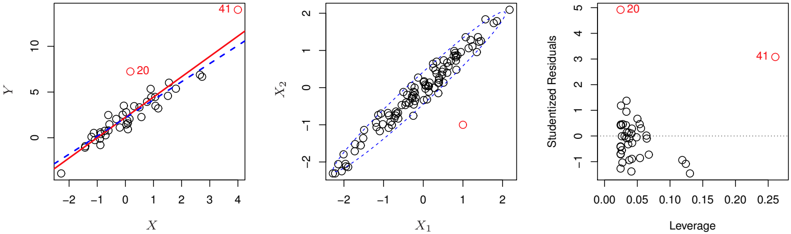
A single point can heavily influence the RSE and R^2 of the model.



	RSE	R^2
Model with outlier	1.09	0.805
Model without outlier	0.77	0.892
Improvement	29%	11%

5. High leverage points

- **High leverage point** is a point with an unusual value of x_i.
- Detect high leverage points using the **leverage statistic**.



6. Collinearity

- **collinearity problem** happens when two predictors are highly correlated to each other.
- Highly correlated variables cause problems when training the model.

6. Collinearity

- **collinearity problem** happens when two predictors are highly correlated to each other.
- Highly correlated variables cause problems when training the model.

Example: Suppose we have data with two predictors x and z.

$$(y_1, x_1, z_1), (y_2, x_2, z_2), \ldots$$

where $z_i = 2x_i$.

6. Collinearity

Suppose that we have a solution $(0, 1, 1)$

$$\hat{y}_i = x_i + z_i$$

6. Collinearity

Suppose that we have a solution $(0, 1, 1)$

$$\hat{y}_i = x_i + z_i$$

Since $z_i = 2x_i$

$$\begin{aligned} \hat{y}_i &= x_i + 2x_i \\ &= 3x_i \end{aligned}$$

In other words, $(0, 3, 0)$ is also a solution.

6. Collinearity

Suppose that we have a solution $(0, 1, 1)$

$$\hat{y}_i = x_i + z_i$$

Since $z_i = 2x_i$

$$\begin{aligned} \hat{y}_i &= x_i + 2x_i \\ &= 3x_i \end{aligned}$$

In other words, $(0, 3, 0)$ is also a solution.
In fact, any $(0, a, b)$ where $a + b = 3$ is also a solution. This causes confusion when implemented by a computer program!

6. Collinearity

Suppose that we have a solution $(0, 1, 1)$

$$\hat{y}_i = x_i + z_i$$

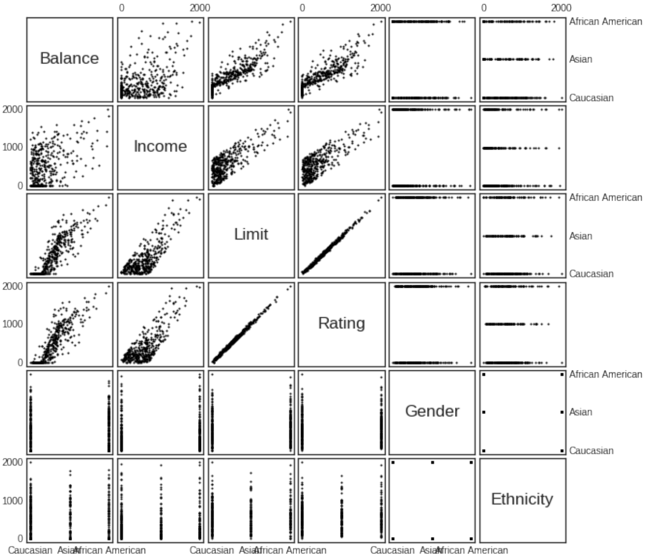
Since $z_i = 2x_i$

$$\begin{aligned}\hat{y}_i &= x_i + 2x_i \\ &= 3x_i\end{aligned}$$

In other words, $(0, 3, 0)$ is also a solution.
In fact, any $(0, a, b)$ where $a + b = 3$ is also a solution. This causes confusion when implemented by a computer program!

Detect collinearity using **correlation matrix**. Remove a variable if the correlation is close to -1 or 1.

Credit balance data



Multicollinearity

Multicollinearity happens when a predictor is a linear combination of other predictors.

Multicollinearity

Multicollinearity happens when a predictor is a linear combination of other predictors.

Example: Predictors x_i, z_i and w_i where $x_i = z_i + 2w_i$.

Multicollinearity

Multicollinearity happens when a predictor is a linear combination of other predictors.

Example: Predictors x_i, z_i and w_i where $x_i = z_i + 2w_i$.

Cannot be detected with correlation matrix. Instead, we use **variance inflation factor**

$$VIF(\hat{\beta}_i) = \frac{1}{1 - R^2_{X_i|X_{-i}}},$$

where $R^2_{X_i|X_{-i}}$ is the R^2 from a regression of X_i onto all other predictors.

Variance inflation factor

$$VIF(\hat{\beta}_i) = \frac{1}{1 - R^2_{X_i|X_{-i}}}.$$

[High multicol. in X_i] $\rightarrow$ [$R^2_{X_i|X_{-i}}$ is close to 1] $\rightarrow$ [high $VIF(\hat{\beta}_i)$]

Variance inflation factor

$$VIF(\hat{\beta}_i) = \frac{1}{1 - R^2_{X_i|X_{-i}}}.$$

[High multicol. in X_i] $\rightarrow$ [$R^2_{X_i|X_{-i}}$ is close to 1] $\rightarrow$ [high $VIF(\hat{\beta}_i)$]

General rule: There is multicollinearity if VIF is higher than 5 or 10

Solution: Drop the variable (in this case, X_i).

Acknowledgement

Some of the figures in this presentation are taken from ”An Introduction to Statistical Learning, with applications in R” (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani