QM 11/11/2019

8 Regression Analysis

Purpose: To predict values of Y

Model:

$$Y \approx b_0 + b_1 x_1 + b_2 x_2 + \dots + b_p x_p$$

Y → dependent variable

$x_1, x_2, \dots, x_p$ → independent variables (Do not mistake for stochastically independent variables)

8.1 Multiple Linear Regression

$Y, X_1, X_2, \ldots, X_p$ scale leveled variables

Example

Miles-Per-Gallon.sav

Y = MPG = miles per gallon

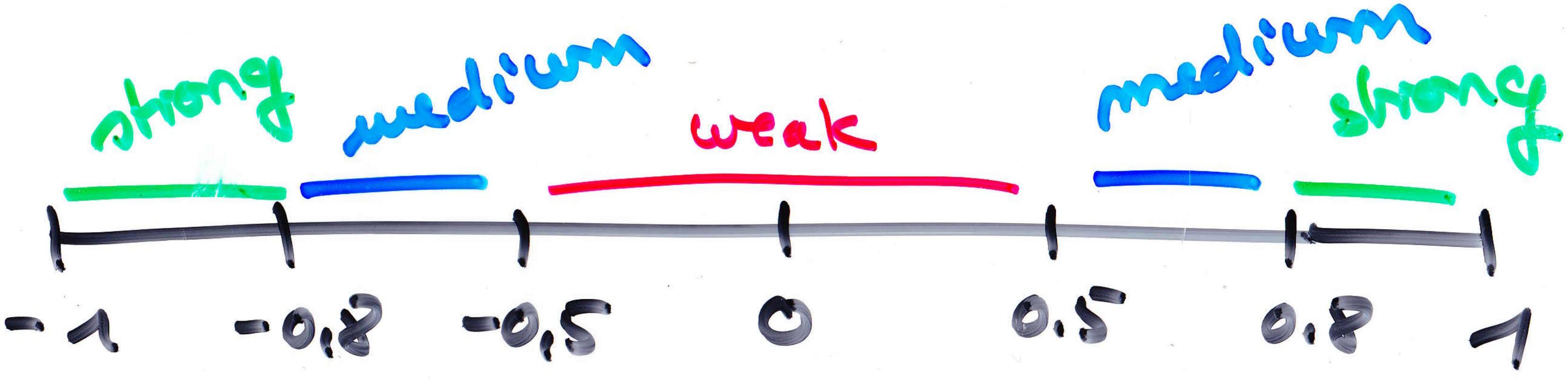
X_1 = weight (in pounds)

X_2 = horsepower

8.1.1 One Independent Variable

r(MPG, weight) = − 0.825 stronger than

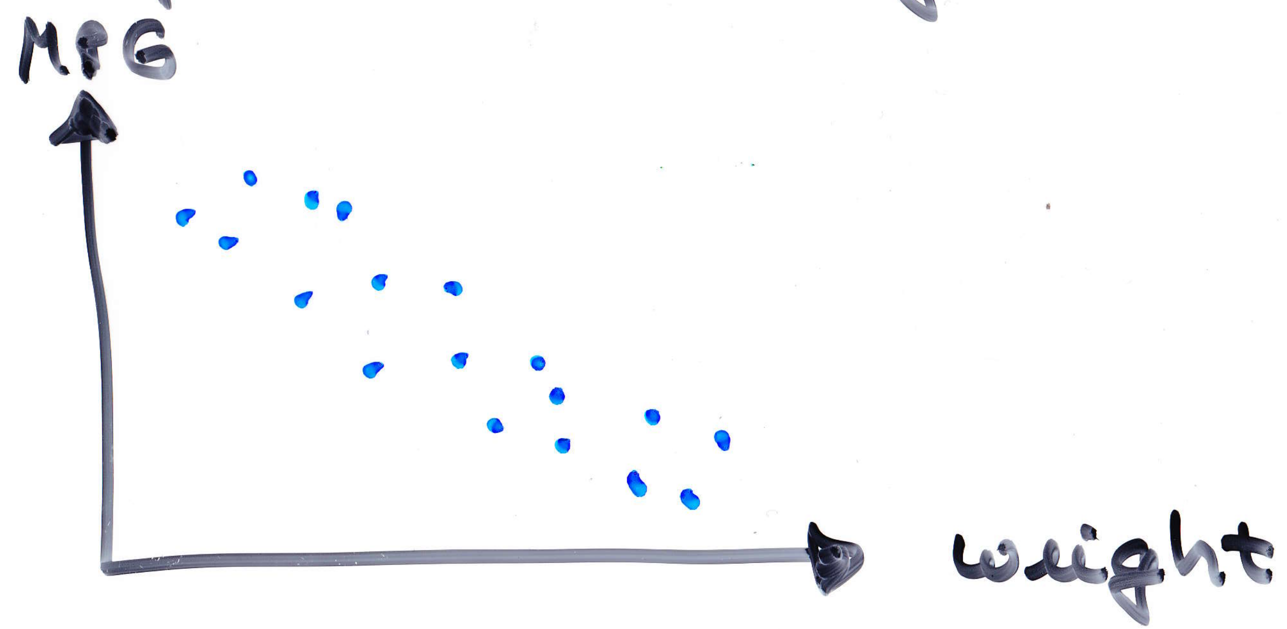
r(MPG, horsepower) = − 0.788



Model:

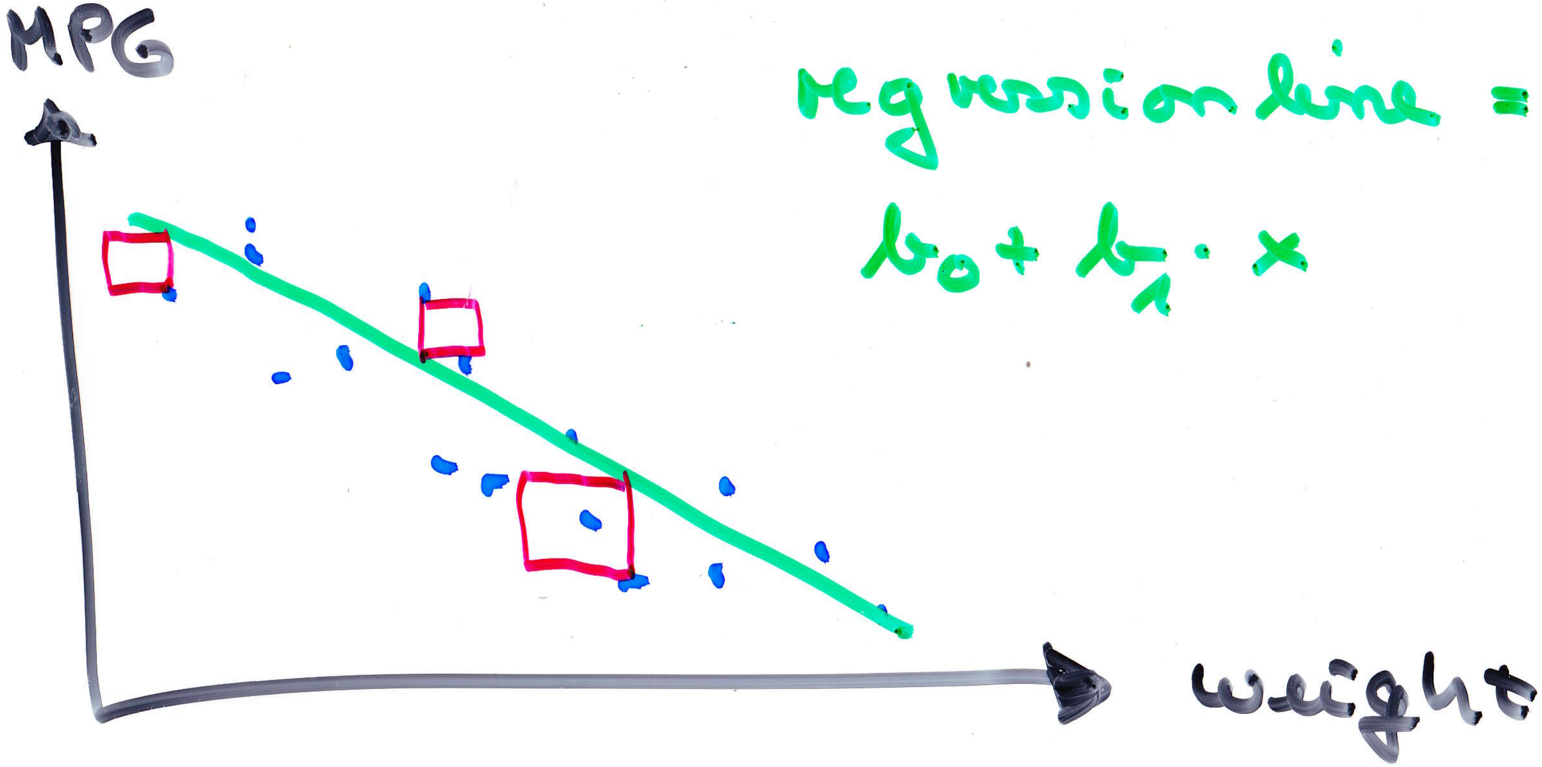
$$\boxed{MPG} \approx b_0 + b_1 \cdot \boxed{weight}$$

a) Scatter plot without regression line



SPSS Commands 8.1.10

with increasing weight the miles per gallon are decreasing b)



regression line = $b_0 + b_1 \cdot x$

Method of the least squares = the sum of the squared distances of the points from the line is minimum

$\hat{b}_0 = 57.797$ estimated value of b_0

$\hat{b}_1 = -0.011$ estimated value of b_1

scatterplot with regression line

→ regression coefficient: if the weight increases by one pound the miles per gallon will decrease by 0.011 miles

→ if a car has a weight of 0 pounds the car will drive 57.797 miles per gallon = nonsense!

c) linear regression model

Commands 8.1.10

d) predict values

4500 pounds weight

Miles per gallon = ?

$57.797 - 0.011 \cdot 4500$ = predicted

miles due to the model = 10.04 miles

= PRE = predicted values

e) 2250 pounds weight

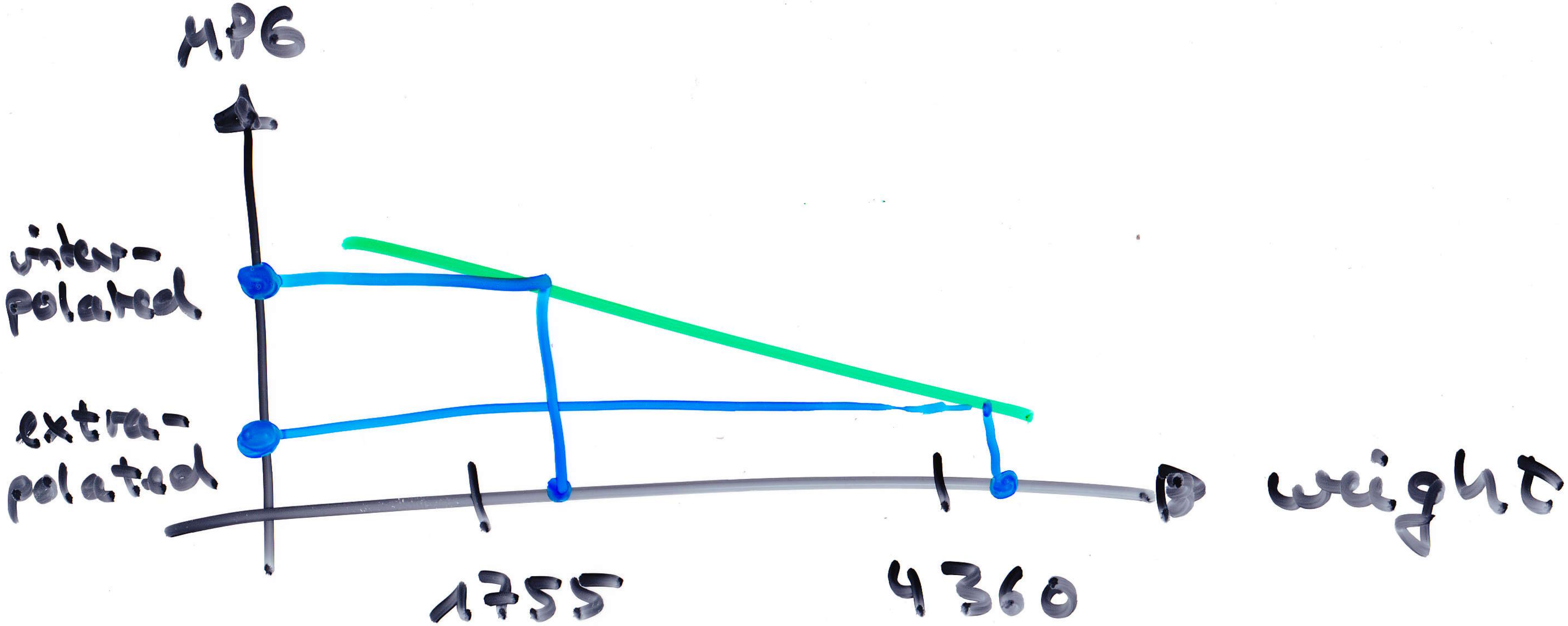
miles per gallon = ?

$57.797 - 0.11 \cdot 2250 = 33.92$ miles

f) Are the predicted values reliable ?

$x_{min} = 1755$ weight

$x_{max} = 4360$ weight



33.92 miles is an interpolated value

because $2250 \in [1755; 4360]$

10.04 miles is an extrapolated value

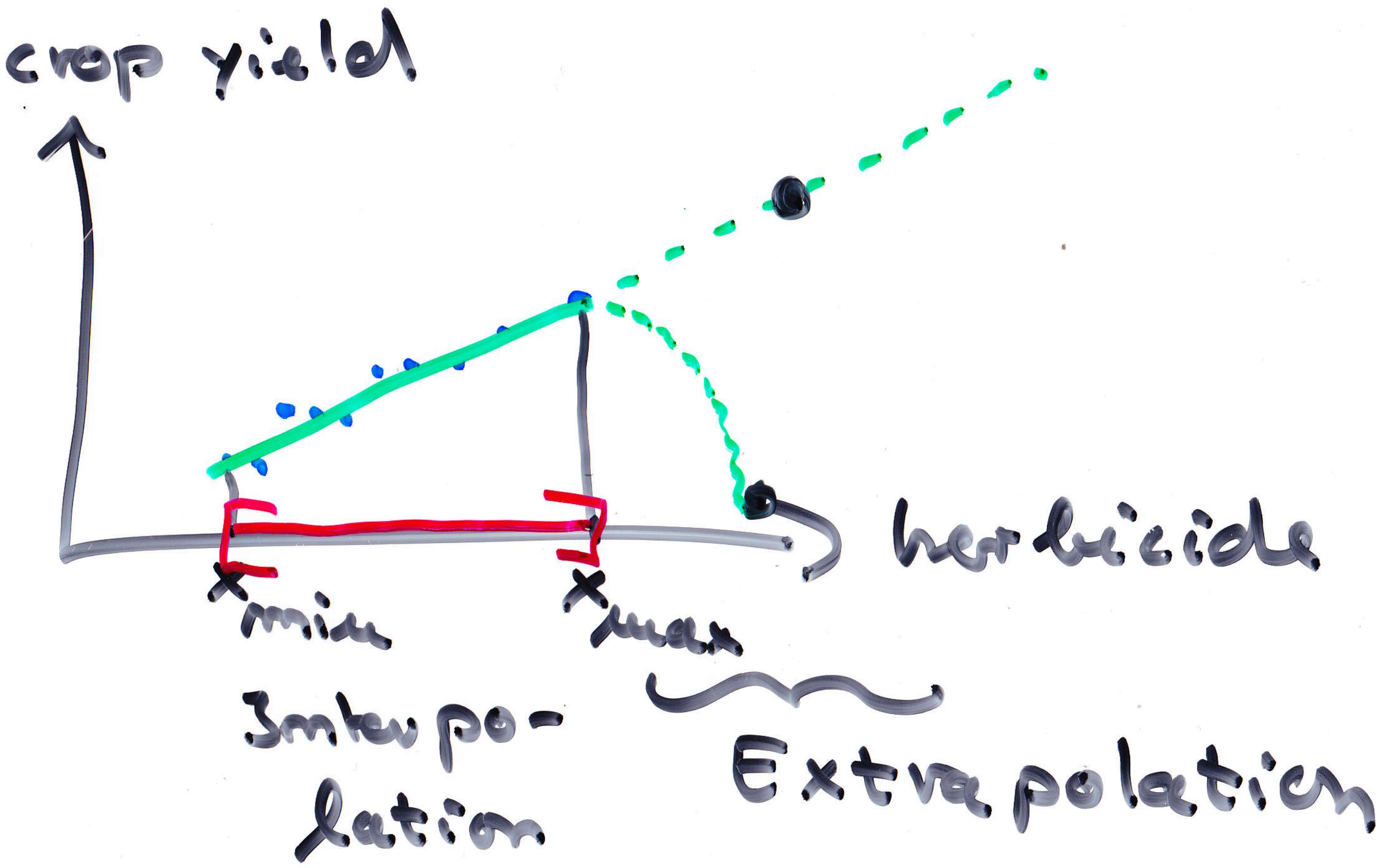
because $4500 \notin [1755; 4360]$

	Correlation		
	weak	medium	strong
interpolation	no	no	yes
extrapolation	no	no	no

The predicted value of 33.92 miles is reliable, because this value is an interpolated value and the correlation of -0.825 is strong.

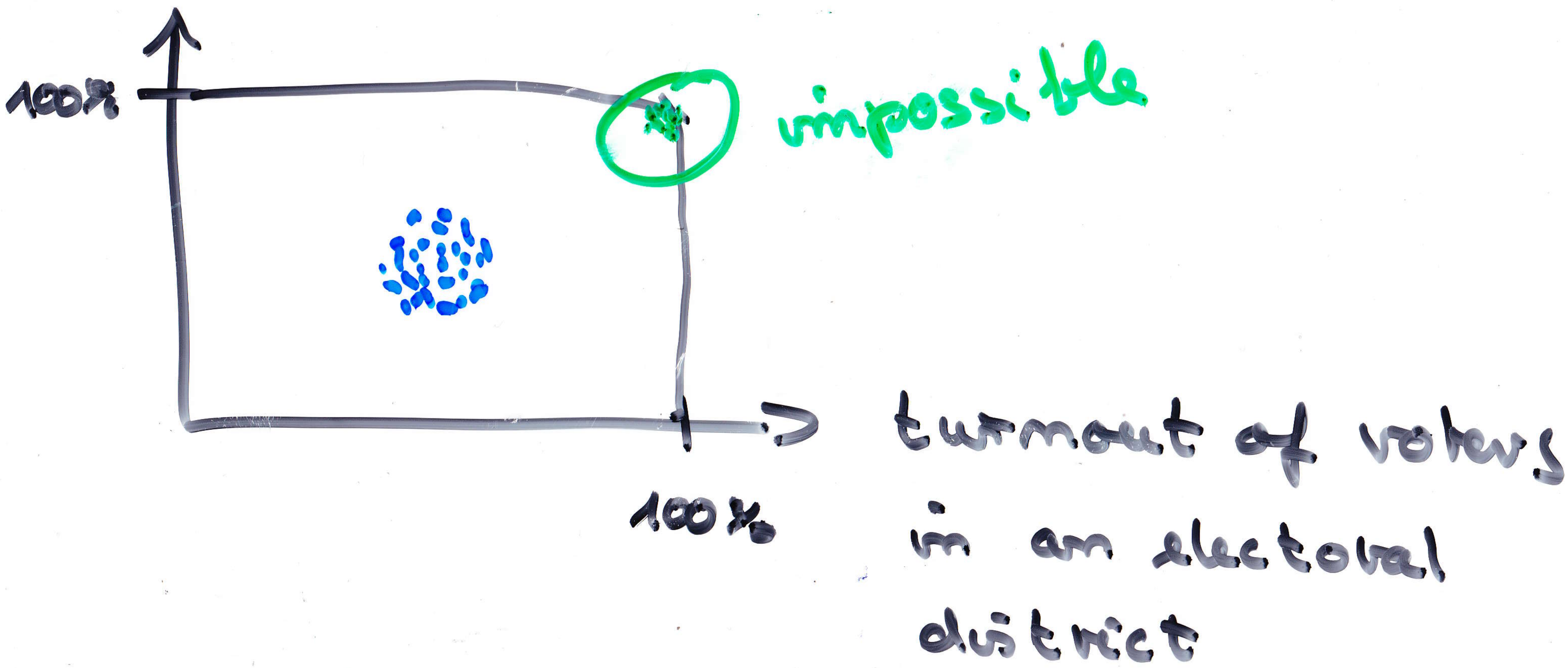
The predicted value of 10.04 miles is not reliable, because this value is an extrapolated value.

Example

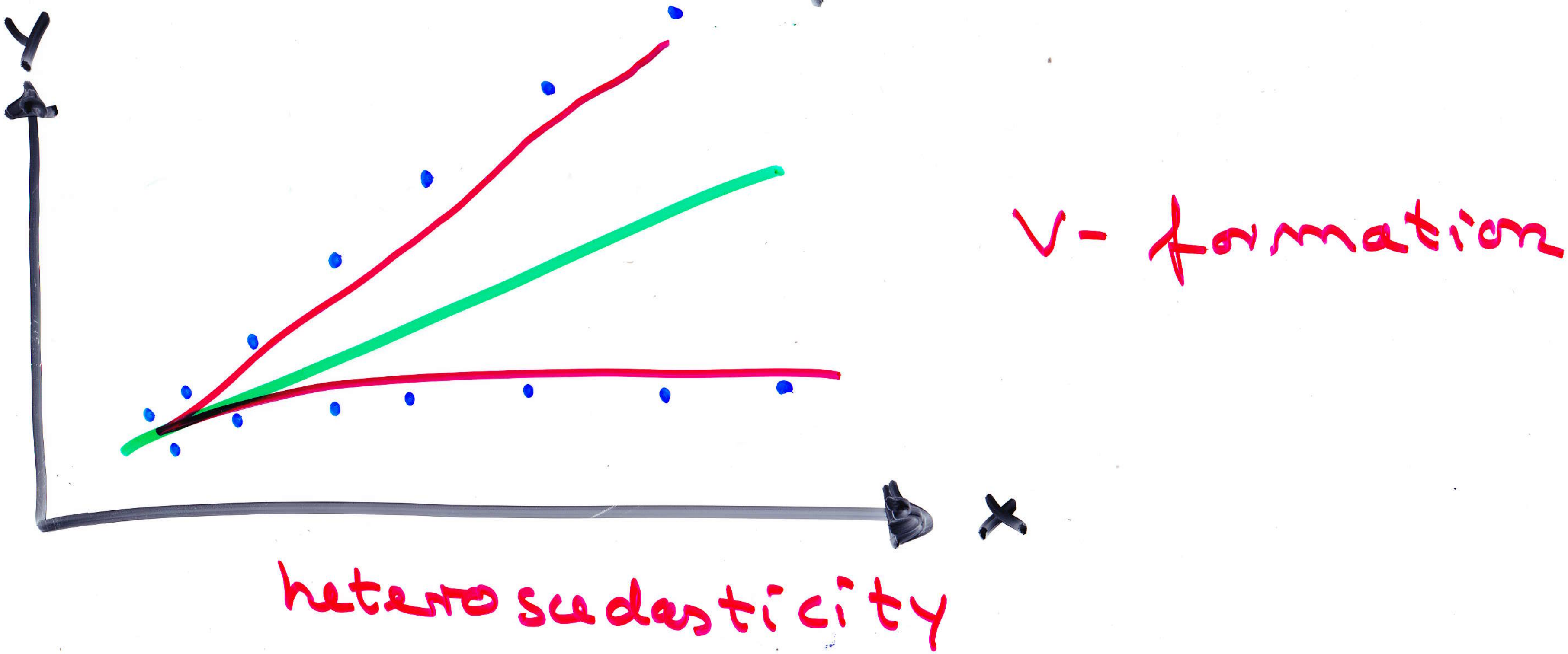


Example: **Election fraud**

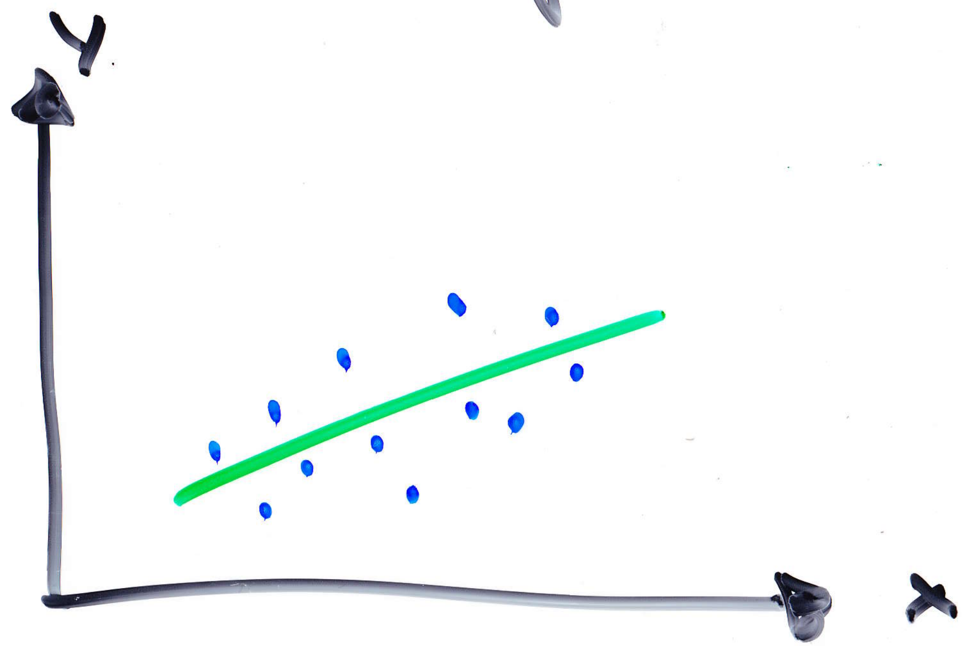
percentage of votes of the winner


100%
100%
impossible
turnout of voters
in an electoral
district

Hetero scedasticity



In case of heteroscedasticity the values of b_0 and b_1 are inappropri-ate if they were calculated due to the method of least squares.

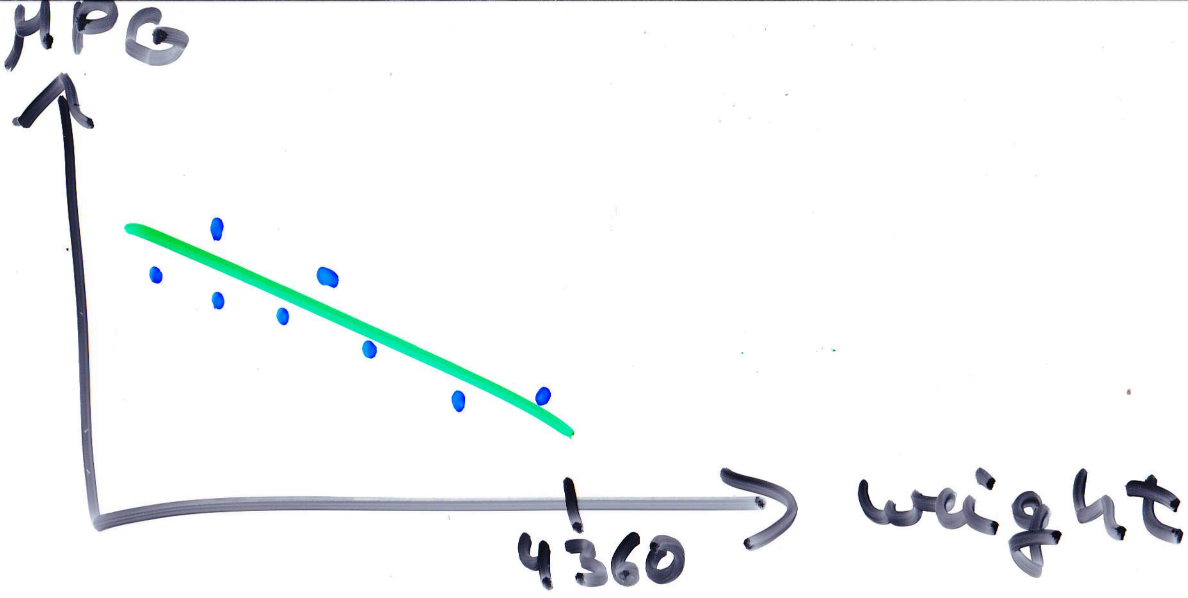


homo scedasticity

Example

Miles_Per_Gallon.sav

$\boxed{MPG} \approx b_0 + b_1 \cdot \boxed{weight}$



homoscedasticity

LEV = leverage points

Exercise 8.9

Leverage value of car with 4360 pounds

weight = 0.13011 largest effect on

the shape of the regression line

8.1.2 Two or more independent variables

Example

Miles_Per_Gallon.sav

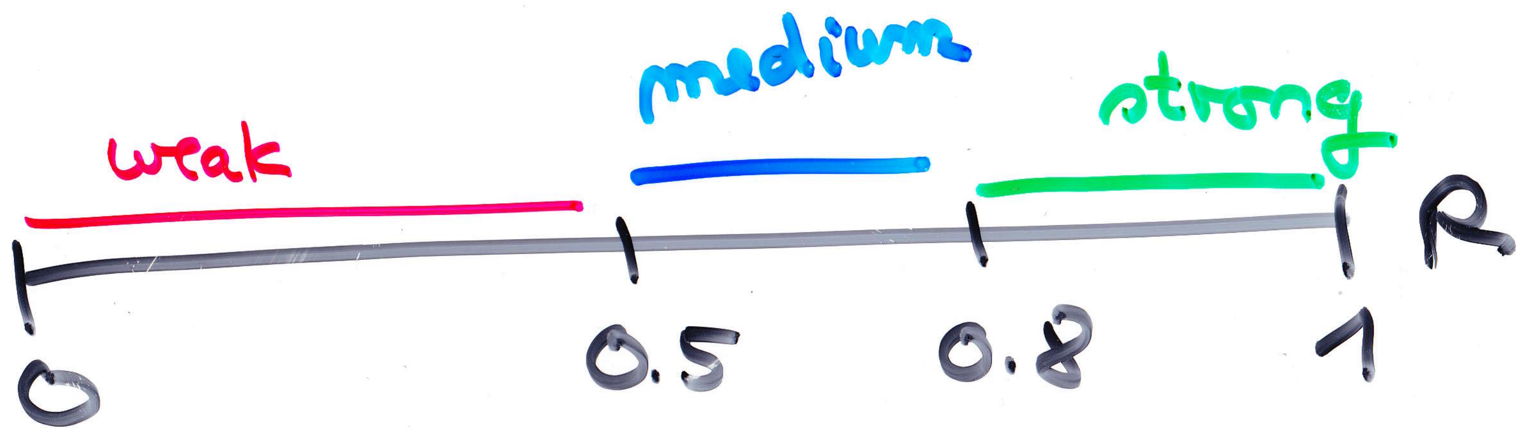
$$\boxed{MPG} \approx b_0 + b_1 \cdot \boxed{weight} + b_2 \cdot \boxed{Horse\text{-}power}$$

$r(MPG, weight) = -0.825$
$r(MPG, Horsepower) = -0.788$

} column zero-order

8.1.3 Multiple correlation coefficient

$R \in [0; 1]$



R(MPG, weight, Horsepower) = 0.866

strong correlation

R^2 = coefficient of determination $\in [0; 1]$

$R^2 = 0.866^2 = 0.749$ strong correlation

8.1.4 Adjusted R-Square

1. Model

$\boxed{MPG} \approx b_0 + b_1 \cdot \boxed{weight}$ $R_a^2 = 0.674$

2. Model

$$\boxed{MPG} \approx b_0 + b_1 \cdot \boxed{Weight} + b_2 \cdot \boxed{Horsepower}$$

$$R_a^2 = 0.739$$

Aim of the linear regression model: <u>only a few</u> independent variables should be in the model

Problem: If we add another indepen-dent variable into the model, the value of R will never decrease.

R_a^2 is an <u>indicator</u> whether it is a good or bad idea, to add another independent variable into the model: If the value of R_a^2 is increasing, the additional variable should remain in the model. If the value of R_a^2 is decreasing, the additional variable should be cancelled from the model.

The increase of R_a^2 indicates that the horsepower should be added into the model.

8.1.5 Multicollinearity

multicollinearity = if two or more independent variables are depending on each other multiple correlation coefficient of the independent variables:

$$R \geq 0.95$$

$$R^2 \geq 0.9025 \approx 0.9 \qquad | \cdot (-1)$$

$$-R^2 \leq -0.9 \qquad | +1$$

$$1 - R^2 \leq 0.1 \qquad | \text{ reciprocal}$$

$$\boxed{\frac{1}{1-R^2}} \geq \frac{1}{0.1} = 10$$

variance inflation factor = VIF

multicollinearity $\Leftrightarrow$ $VIF \geq 10$

In case of multicollinearity the estimators of the regression coefficients are incorrect.

$VIF = 2.224 < 10$

no multicollinearity of weight and horsepower

8.1.6 Forecasting

$\hat{b}_0 = 58.157$

$\hat{b}_1 = -0.007$

$\hat{b}_2 = -0.118$

2500 pounds

80 horsepower

MPG = ?

$\in [1755; 4360]$

$\in [48; 165]$

$$58.157 - 0.007 \cdot 2500 - 0.118 \cdot 80 = 31.6$$

miles

Is the predicted value reliable ?

31.6 is an interpolated value and

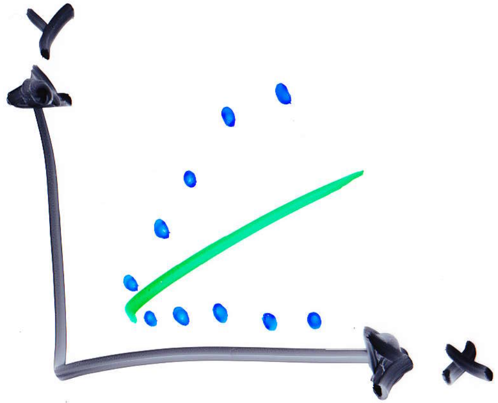
$R = 0.866$ strong correlation, thus the

predicted value of 31.6 miles is reliable.

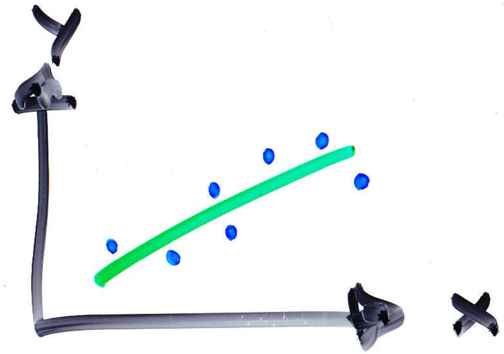
$\hat{b}_2 = -0.118$ if the horsepower increases by one unit and the weight is unchanged the miles per gallon will decrease by 0.118 miles.

8.1.7 Heteroscedasticity

Model with only one independent variable



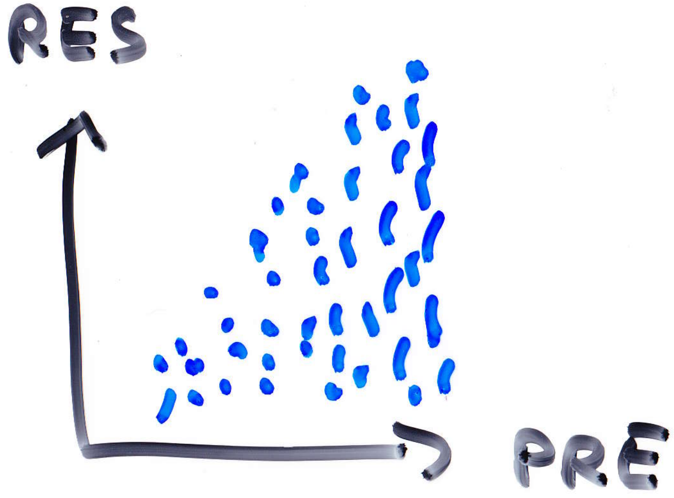
heteroscedasticity



homoscedasticity

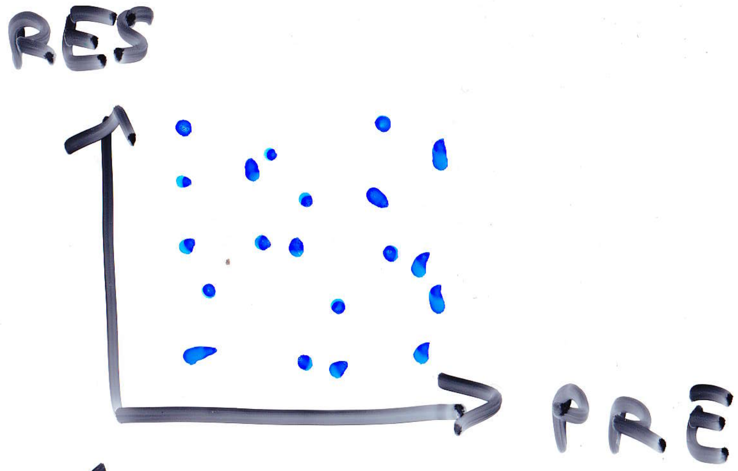
Model with two or more indepent variables

RES = residuals = observed value minus predicted value



(v-formation or a pattern)

heteroscedasticity

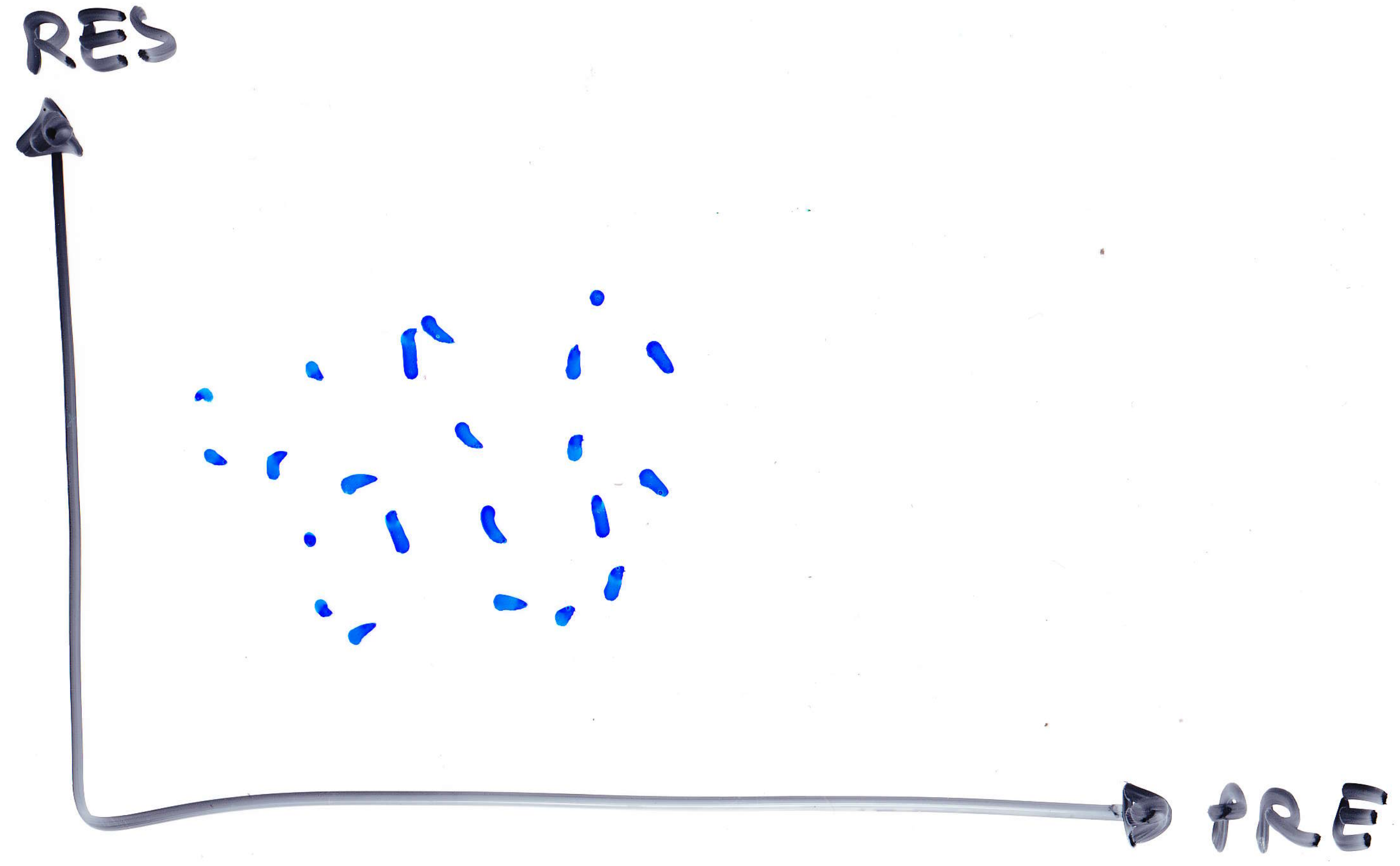


(points are plotted by random)

homoscedasticity

Example

Miles_Per_Gallon.sav



homoscedasticity

Homework : 6.1 and 6.2