VII. Coefficient of Determination

0

- R^2 will be derived
  - basic measure of a model quality?
    $\rightarrow$ prediction quality

DATA $(Y_i, X_i^T)^T$, $i = 1, \dots, n$

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad X = \begin{pmatrix} X_1^T \\ \vdots \\ X_n^T \end{pmatrix}$$ assumed to follow LM $Y | X \sim (X\beta, \sigma^2 I)$

(no distributional assumpt.)

7.1 Intercept only model

1

(Usual) notation $\quad \overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i = \frac{1}{n} Y^T \mathbb{1}$

Def. 7.1 Regression and total sums of squares in a linear model

2

Consider a linear model $Y | X \sim (X\beta, \sigma^2 I_n)$, $\text{rank}(X_{n \times k}) = r \leq k$. The following expressions define the following quantities

(i) Regression sum of squares and corresponding degrees of freedom: $SS_R = \| \hat{Y} - \overline{Y} \mathbb{1}_n \|^2 = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2$,

$\nu_R = r - 1$

(ii) Total sum of squares and corresponding degrees of freedom: $SS_T = \| Y - \overline{Y} \mathbb{1}_n \|^2 = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$

$\nu_T = n - 1$.

REMARK: $Y - \bar{Y}\mathbb{1} = Y - \hat{Y} + \hat{Y} - \bar{Y}\mathbb{1}$

$$SS_T \overset{?}{=} SS_e + SS_R$$

$$n-1 = n-r + r-1$$

Lemma 7.1 Model with intercept only

3

Let $Y \sim (\mathbb{1}_n \mu, \sigma^2 I_n)$ ($EY_i = \mu$, $\operatorname{var} Y_i = \sigma^2$ $\forall i$, $Y_1, \dots, Y_n$ uncorrelated)

Then (i) $\hat{Y} = \bar{Y}\mathbb{1}_n = (\bar{Y}, \dots, \bar{Y})^T$.

(ii) $SS_e = SS_T$.

Proof: It is a full-rank model with $X = \mathbb{1}$

$$X^TX = \mathbb{1}^T\mathbb{1} = n \quad , \quad (X^TX)^{-1} = \frac{1}{n}$$

$$X^TY = \mathbb{1}^TY = \sum_{i=1}^{n} Y_i$$

$$\Rightarrow \hat{\mu} = (X^TX)^{-1}(X^TY) = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}$$

$$\Rightarrow \hat{Y} = X\hat{\mu} = \mathbb{1}\bar{Y} = \bar{Y}\mathbb{1}$$

$$SS_e = \|Y - \hat{Y}\|^2 = \|Y - \bar{Y}\mathbb{1}\|^2 = SS_T$$

$\blacksquare$

7.2 Models with intercept

$\equiv$ model where $\mathbb{1} \in \mathcal{M}(\mathbb{X})$

(not only explicitely, e.g. splines as well)

4

Lemma 7.2 Identity in a linear model with intercept

Let $\mathbf{Y} \mid \mathbb{X} \sim (\mathbb{X}\beta, \sigma^2 I_n)$ where $\mathbb{1}_n \in \mathcal{M}(\mathbb{X})$. Then

$$\mathbb{1}_n^\top \mathbf{Y} = \sum_{i=1}^n Y_i = \sum_{i=1}^n \hat{Y}_i = \mathbb{1}_n^\top \hat{\mathbf{Y}}.$$

5

Proof: If $\mathbb{X} = (\mathbb{1} \ * \cdots *)$, the statement of Lemma follows directly from normal equations:

$\hat{\beta}$ Minimizer $\mathbf{b}$ to sum of squares satisfies:
(even if $r<k$)

$$\mathbb{X}^\top \mathbb{X} \mathbf{b} = \mathbb{X}^\top \mathbf{Y}$$

$$\mathbb{X}^\top(\underbrace{\mathbb{X}\mathbf{b}}_{\hat{\mathbf{Y}}} - \mathbf{Y}) = 0, \quad \text{that is } \mathbb{X}^\top(\hat{\mathbf{Y}} - \mathbf{Y}) = 0,$$

with $\mathbb{X} = (\mathbb{1} \ * \cdots *)$, the first equation is $\sum \hat{Y}_i = \sum Y_i$.

In general ($\mathbb{1} \in \mathcal{M}(\mathbb{X})$ but not necessarily the first column in the model matrix)

$$\mathbb{1}^\top \hat{\mathbf{Y}} = \mathbb{1}^\top \mathbf{H} \mathbf{Y} = \mathbf{Y}^\top \underbrace{\mathbf{H}\mathbb{1}}_{=\mathbb{1}, \text{ since } \mathbb{1} \in \mathcal{M}(\mathbb{X})} = \mathbf{Y}^\top \mathbb{1} = \mathbb{1}^\top \mathbf{Y}.$$

$\blacksquare$

Consequence: $\sum_{i=1}^n U_i = \sum_{i=1}^n (Y_i - \hat{Y}_i) = 0$

One-sample t-test on U_i's to test $\mathsf{E}\varepsilon_i = 0$? Useful?

Lemma 7.3 Breakdown of the total sum of squares in a linear model with intercept

Let $Y|X \sim (X\beta; \sigma^2 I_n)$, where $1_n \in \mathcal{M}(X)$. Then

$$\begin{array}{ccccc} SS_T & = & SS_e & + & SS_R \\ \sum(Y_i-\bar{Y})^2 & = & \sum(Y_i-\hat{Y}_i)^2 & + & \sum(\hat{Y}_i-\bar{Y})^2. \end{array}$$

Proof:

(a) $r = \text{rank}(X) = 1 \Rightarrow \mathcal{M}(X) = \mathcal{M}(1) \overset{\text{Lemma 7.1}}{\Rightarrow} \hat{Y} = \bar{Y}1$

$\Rightarrow SS_T = SS_e, \quad SS_R = 0.$

not needed

(b) $r = \text{rank}(X) \geq 1$ (i.e. arbitrary)

$$SS_T = \sum_i (Y_i - \bar{Y})^2 = \sum_i (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2 =$$

$$= \underbrace{\sum_{i=1}^n (Y_i-\hat{Y}_i)^2}_{SS_e} + \underbrace{\sum_{i=1}^n (\hat{Y}_i-\bar{Y})^2}_{SS_R} + 2\underbrace{\sum_{i=1}^n (Y_i-\hat{Y}_i)(\hat{Y}_i-\bar{Y})}_{* = 0}$$

$$* = \sum Y_i\hat{Y}_i - \bar{Y}\sum Y_i - \sum \hat{Y}_i^2 + \bar{Y}\underbrace{\sum \hat{Y}_i}_{= \sum Y_i \text{ (Lemma 7.2)}} = 0$$

because

$$\sum Y_i \hat{Y}_i = Y^T\hat{Y} = Y^T H Y = Y^T H H Y = \hat{Y}^T\hat{Y} =$$

$$= \sum \hat{Y}_i^2$$

7.3 Theoretical evaluation of a prediction quality of the model

7

Frequent aim of regression modelling: prediction

- $(Y_i, X_i^T)^T$: data to build/develop a model for $Y_i | X_i$
- $X_{new} = x_{new}$, $Y_{new} | X_{new}$ assumed to follow the same model (= distribution etc.)
- $\hat{Y}_{new} \equiv$ prediction of $Y_{new} \sim Y_{new} | X_{new} = x_{new}$
- Quality of prediction based on the model?

Let us now assume iid data

8

$(Y_i, X_i^T)^T \overset{iid}{\sim} \underbrace{(Y, X^T)^T}$, $X \in \mathbb{R}^k$ follow some (joint) distribution which satisfies:

(a) conditional distrib. $Y|X$

$\mathsf{E}(Y|X) = X^T\beta$, $\mathrm{var}(Y|X) = \sigma^2$ for some $\beta \in \mathbb{R}^k$, $\sigma^2 > 0$

that is $\quad \mathbb{Y}|\mathbb{X} \sim (\mathbb{X}\beta, \sigma^2 I_n)$

(LM for $\mathbb{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$, $\mathbb{X} = \begin{pmatrix} X_1^T \\ \vdots \\ X_n^T \end{pmatrix}$)

(b) marginal distribution of Y:

$\mathsf{E}(Y) = \mu$, $\mathrm{var}(Y) = \xi^2$ for some $\mu \in \mathbb{R}$, $\xi^2 > 0$

that is $\quad \mathbb{Y} \sim (\mathbb{1}\mu, \xi^2 I_n)$

First: suppose that $\beta, \sigma^2, \mu, \xi^2$ are known

AIM: predict Y_{new} from $(Y_{new}, X_{new}^T)^T \sim (Y, X^T)^T$

i.e. predict Y component of a random vector with distribution of $(Y, X^T)^T$

Simplify notation: Y used instead of Y_{new}

ASSUMPTION: $(Y, X^T)^T$ defined on $(\Omega, \mathcal{A}, P)$

NOTATION (see also TP1):

$\sigma(X) \subset \mathcal{A}$: σ-algebra generated by X

$P_X := P|_{\sigma(X)}$: probab. measure restricted to $\sigma(X)$

$L_2(X) := L_2(\Omega, \sigma(X), P_X)$

$\sigma(\emptyset) = \{\emptyset, \Omega\} \subset \mathcal{A}$

$P_0 := P|_{\sigma(\emptyset)}$

$L_2(\emptyset) := L_2(\Omega, \sigma(\emptyset), P_0)$

Which random variables Y live here?

- $\forall B \in \mathcal{B}$ $Y^{-1}(B) = \begin{cases} \emptyset \\ \Omega \end{cases}$ $\quad P(Y \in B) = \begin{cases} 0 \\ 1 \end{cases}$

- $Y = c$ almost surely

Prediction of Y

$\equiv$ look for $\hat{Y}$ which minimizes (how?)

$$MSEP(\hat{Y}) = \mathbb{E}(\hat{Y}-Y)^2$$

(a) no information on X available

$$\hat{Y} = \underset{\tilde{Y} \in L_2(\emptyset)}{\operatorname{argmin}} \mathbb{E}(\tilde{Y}-Y)^2 = \underset{\tilde{Y} \in \mathbb{R}}{\operatorname{argmin}} \mathbb{E}_Y(\tilde{Y}-Y)^2$$

$$= \mathbb{E}Y = \mu =: \hat{Y}^M \quad (\underline{\text{marginal}} \text{ prediction})$$

$$MSEP(\hat{Y}^M) = \mathbb{E}(\mu - Y)^2 = \operatorname{var} Y = \xi^2$$

(b) information on X available $\equiv \sigma(X)$ and P_X known

$$\hat{Y} = \underset{\tilde{Y} \in L_2(X)}{\operatorname{argmin}} \mathbb{E}(\tilde{Y}-Y)^2 \overset{TP1}{=} \mathbb{E}(Y|X) \overset{\text{ASSUMED MODEL}}{=}$$

$$= X^T\beta =: \hat{Y}^C$$

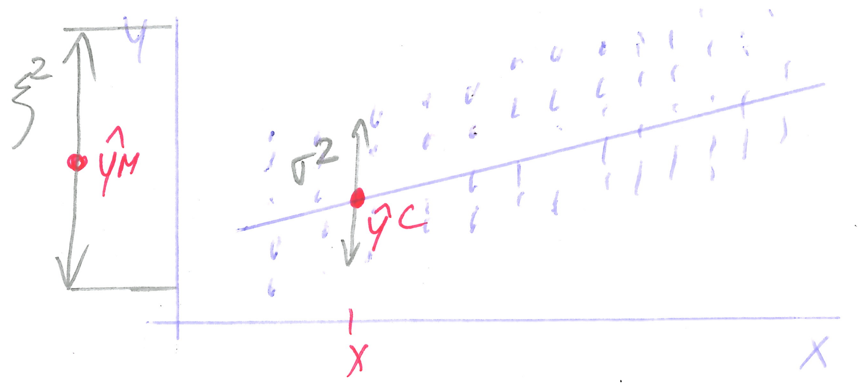
$(\underline{\text{conditional}}$ prediction$)$

$$MSEP(\hat{Y}^C) = \mathbb{E}(X^T\beta - Y)^2 = \mathbb{E}\big(\underbrace{\mathbb{E}((X^T\beta - Y)^2 | X)}_{\operatorname{var}(Y|X)}\big) =$$

$$\overset{\text{ASSUM. MODEL}}{=} \mathbb{E}(\sigma^2) = \sigma^2$$

Comparison of the two predictions

$$\frac{MSEP(\hat{Y}^C)}{MSEP(\hat{Y}^M)} = \frac{\sigma^2}{\xi^2} = \frac{var(Y|X)}{var(Y)}$$



7.4 Coefficient of determination

9

From previous page: $\dfrac{MSEP(\hat{Y}^C)}{MSEP(\hat{Y}^M)} = \dfrac{\sigma^2}{\xi^2} = \dfrac{var(Y|X)}{var\,Y}$

Unbiased estimators of σ^2, ξ^2

10

$$\hat{\sigma}^2 = \frac{1}{n-r} SS_e = \frac{1}{n-r} \sum_i (Y_i - \hat{Y}_i)^2$$

$$\hat{\xi}^2 = \frac{1}{n-1} SS_T = \frac{1}{n-1} \sum_i (Y_i - \bar{Y})^2$$

$$\longrightarrow \quad \frac{\hat{\sigma}^2}{\hat{\xi}^2} = \frac{n-1}{n-r} \frac{SS_e}{SS_T}$$

MLE's of σ^2, ξ^2 under normality (≡ EHW/SDC)

$$\hat{\sigma}^2_{ML} = \frac{1}{n} SS_e = \frac{1}{n} \sum_i (Y_i - \hat{Y}_i)^2$$

$$\hat{\xi}^2_{ML} = \frac{1}{n} SS_T = \frac{1}{n} \sum_i (Y_i - \bar{Y})^2$$

$$\longrightarrow \quad \frac{\hat{\sigma}^2_{ML}}{\hat{\xi}^2_{ML}} = \frac{SS_e}{SS_T}$$

Def 7.2 Coefficient of determination

11

Consider a linear model $Y|X \sim (X\beta, \sigma^2 I_n)$, $\operatorname{rank}(X) = r$, $1 \in \mathcal{M}(X)$. A value $R^2 = 1 - \dfrac{SS_e}{SS_T}$ is called the <u>coefficient of determination</u> of the linear model.

A value $R^2_{adj} = 1 - \dfrac{n-1}{n-r} \dfrac{SS_e}{SS_T}$ is called the <u>adjusted coefficient of determination</u> of the linear model.

Remarks: $\mathbb{1} \in \mathcal{M}(\mathbb{X}) \Rightarrow SS_T = SS_e + SS_R$

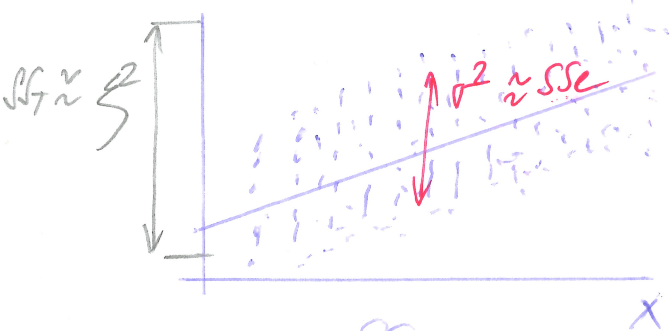
$$R^2 = 1 - \frac{SS_e}{SS_T} = \frac{SS_R}{SS_T}$$

$$\underset{\substack{\approx \xi^2 \\ \equiv \text{total variability of response}}}{SS_T} = \underset{\substack{\approx \sigma^2 \\ \equiv \text{residual variability}}}{SS_e} + SS_R$$

(in the whole population)

(≡ variability not explained by X)

SS_R ≡ variability explained by X



$$R^2 = \frac{SS_R}{SS_T} = \frac{\text{variability of Y explained by X}}{\text{total variability of Y}}$$

often $R^2 \cdot 100\%$ is reported

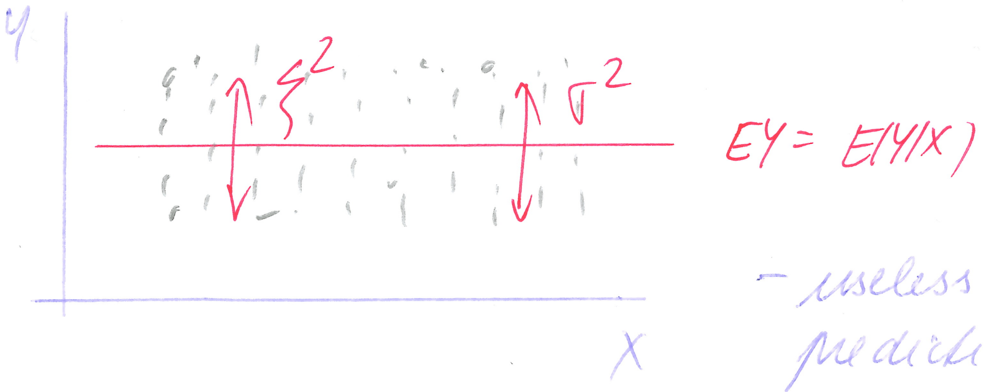
– clearly:
- $0 \leq R^2 \leq 1$
- $R^2 = 1 \Leftrightarrow SS_e = 0$

  (≡ all data points lying on a fitted regression function)

REMARKS:

- R^2 only evaluates quality of possible PREDICTION.
- R^2 says <u>nothing</u> concerning a quality of a model for $\mathbb{E}(Y|X)$!
- For some/many problems high value of R^2 cannot be reached (with a given/available set of covariates. Still, chosen ~~mo~~ LM $X^T\beta$ can be OK as a model for $\mathbb{E}(Y|X)$. It's only high $var(Y|X)$ as compared to $var(Y)$ which makes the model <u>useless for PREDICTION</u>.

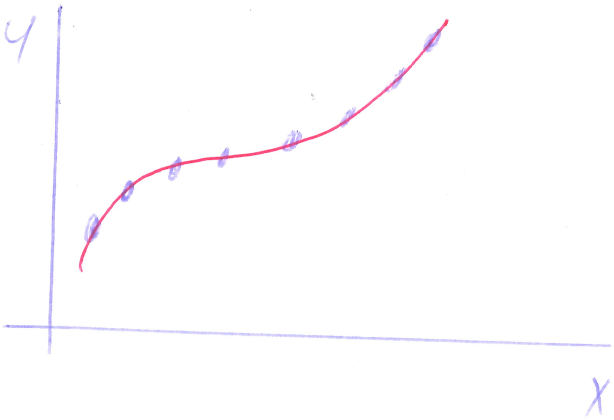
Extreme 1



- useless for prediction

but model $E(Y|X) = \beta_0$ is OK

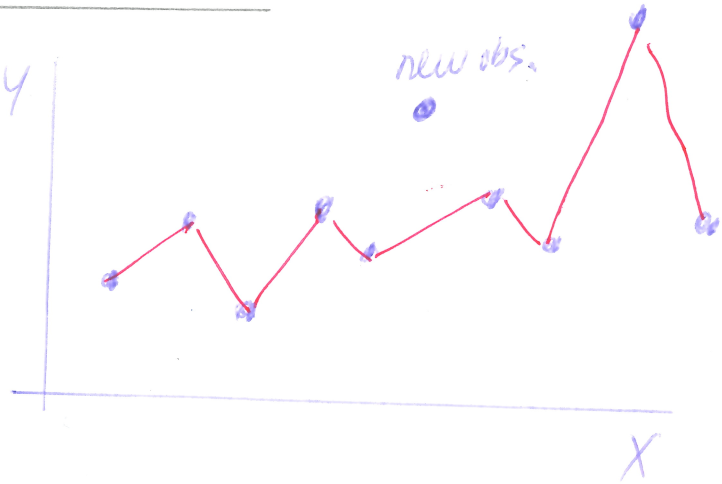
Extreme 2a



$R^2 = 1$

- probably fine here

Extreme 2b



assumed (model) $E(Y|X)$

$R^2 = 1$

- useful model
- will it provide a good prediction of a new observation?

- good prediction models require validation using external data or at least cross-validation
- R^2 provides just self-evaluation of prediction quality, as such, it is over optimistic!

(samochwała ...)