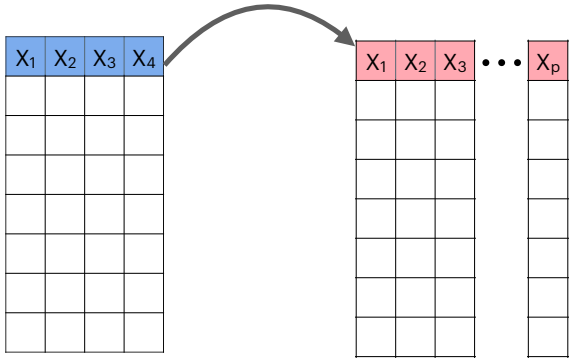
Linear Regression II

Lecture 3

Termeh Shafie

Feature Engineering



when do we do this and why?

Interactions

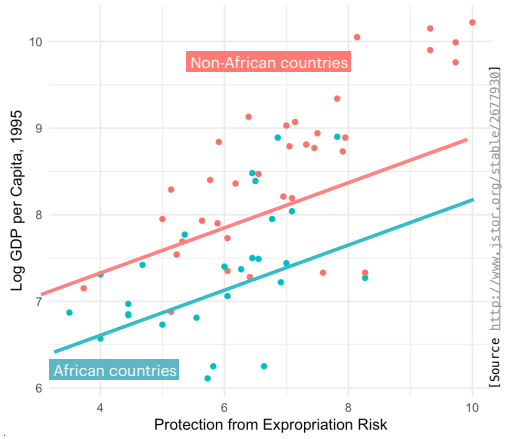
$$Y = \beta_0 + \beta_1 X + \beta_2 Z$$

log GDP | expropriation risk protection | African/Non-African countries

$$Z = 0 \implies Y = \beta_0 + \beta_1 X + \beta_2 Z = \beta_0 + \beta_1 X + \beta_2 \cdot 0 = (\beta_0) + (\beta_1) X$$

$$Z = 1 \implies Y = \beta_0 + \beta_1 x_1 + \beta_2 Z = \beta_0 + \beta_1 X + \beta_2 \cdot 1 = (\beta_0 + \beta_2) + (\beta_1) X$$

different intercepts | same slope



Interactions

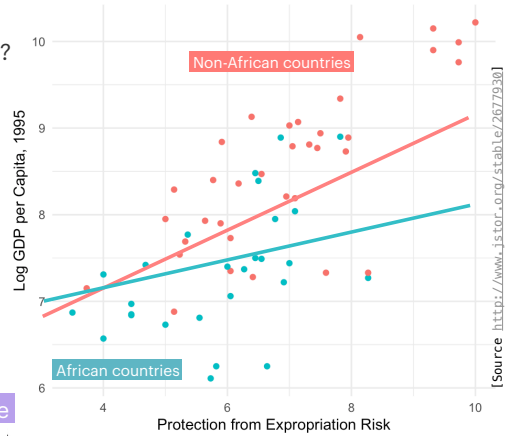
Is the relationship between **X** and **Y** different when you consider values of **Z**?

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ$$

$$Z = 0 \implies Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 ZX = \beta_0 + \beta_1 X + \beta_2 \cdot 0 + \beta_3 \cdot 0 = (\beta_0) + (\beta_1) X$$

$$Z = 1 \implies Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 ZX = \beta_0 + \beta_1 X + \beta_2 \cdot 1 + \beta_3 \cdot 1 \cdot X = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X$$

different intercepts | different slope



Interactions

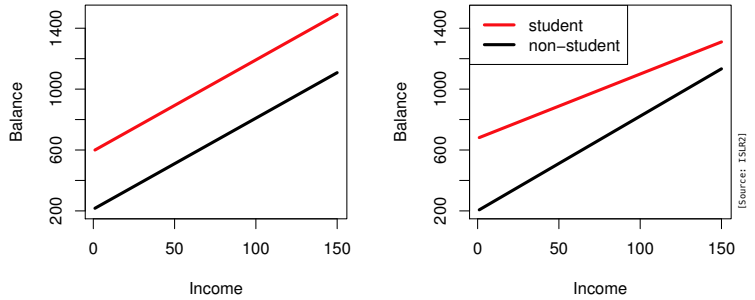
Income	Limit	Rating	Cards	Age	Education	Own	Student	Married	Region	Balance
14.891	3606	283	2	34	11	No	No	Yes	South	333
106.025	6645	483	3	82	15	Yes	Yes	Yes	West	903
104.593	7075	514	4	71	11	No	No	No	West	580
148.924	9504	681	3	36	11	Yes	No	No	West	964
55.882	4897	357	2	68	16	No	No	Yes	South	331
80.180	8047	569	4	77	10	No	No	No	South	1151

[Source: First six rows of dataset "Credit", ISLR2]

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \times X_2)$$

5

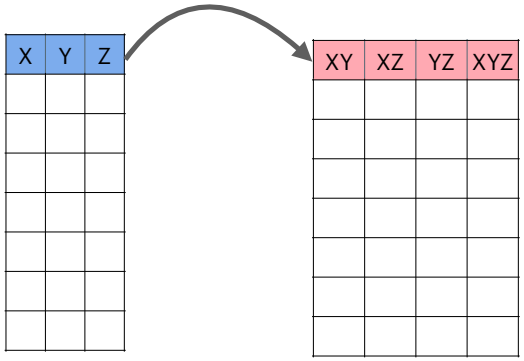
Interactions



$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \times X_2)$$

6

Interactions



7

Dummy Variables: Adding a Bivariate Covariate

"one-hot encoding"

Income	Limit	Rating	Cards	Age	Education	Own	Student	Married	Region	Balance
14.891	3606	283	2	34	11	No	No	Yes	South	333
106.025	6645	483	3	82	15	Yes	Yes	Yes	West	903
104.593	7075	514	4	71	11	No	No	No	West	580
148.924	9504	681	3	36	11	Yes	No	No	West	964
55.882	4897	357	2	68	16	No	No	Yes	South	331
80.180	8047	569	4	77	10	No	No	No	South	1151

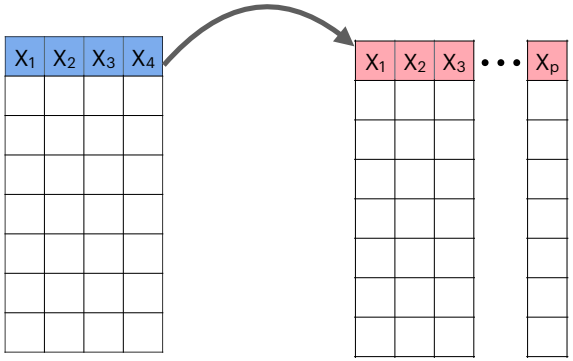
[Source: First six rows of dataset "Credit", ISLR2]

$$Y = \beta_0 + \beta_1 X$$

0/1 -1/1

8

Feature Engineering



we'll return to feature engineering in our lecture on "non-liner" linear regression

Estimate $\hat{f}$ = Learn $\hat{f}$

$$Y = f(\boldsymbol{X}) + \varepsilon$$

sources of error:
irreducible error ε
reducible error $\hat{f}$

the squared error for a given estimate $\hat{f}$ is

$$E(\text{actual} - \text{predicted})^2 = E(Y - \hat{Y})^2$$

which factors as

$$E[f(\boldsymbol{X}) + \varepsilon - \hat{f}(\boldsymbol{X})]^2$$

$$\underbrace{[f(\boldsymbol{X}) - \hat{f}(\boldsymbol{X})^2]}_{\text{reducible}} + \underbrace{\text{Var}(\varepsilon)}_{\text{irreducible}}$$

until now, training data was the only data we considered
we compute reducible error (e.g. MSE) on the same data used to learn $\hat{f}$
let's change that!

Training

training data set

$$\{(y_1, x_1,), \ldots, (y_n, x_n))\}$$

used to find function q that minimizes **Training MSE**

$$\hat{f} = \arg\min_q MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - q(x_i))^2$$

Testing

testing data sets (unseen)

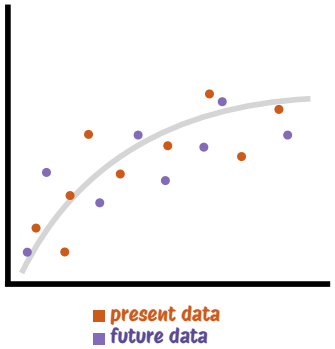
$$(y_0, x_0)$$

used to compute **Test MSE**

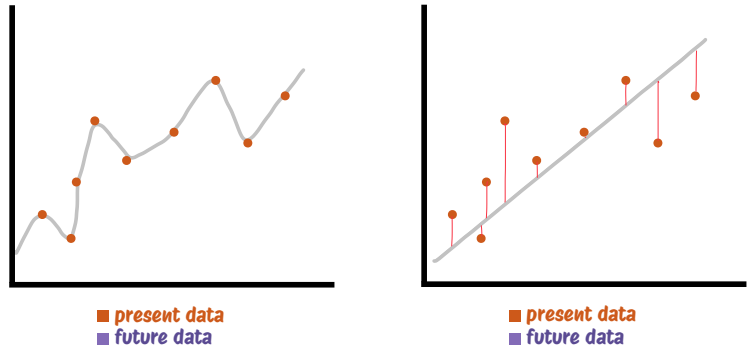
$$E[y_0 - \hat{f}(x_0)^2]$$

`often not so closely related`

Training and Testing

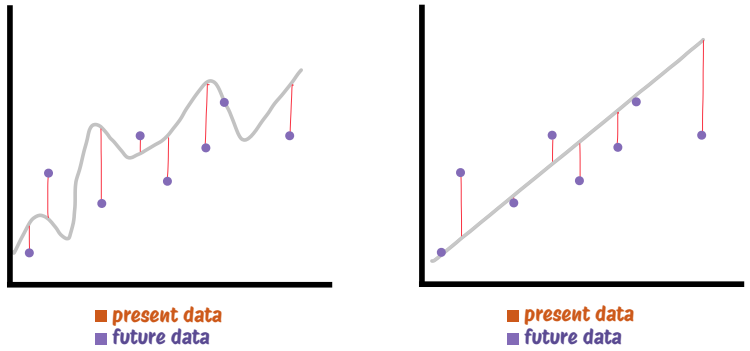


Training and Testing



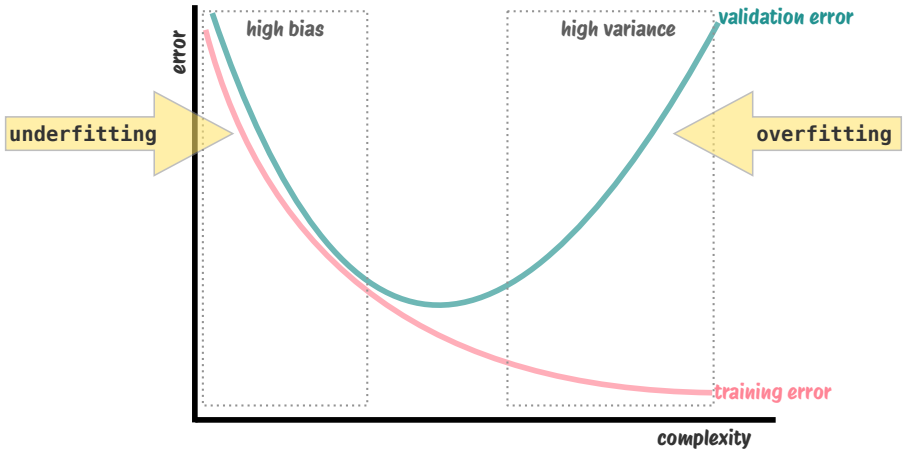
13

Training and Testing



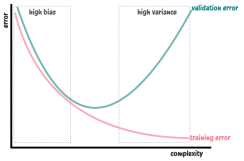
14

Bias Variance Trade-Off



15

Formalizing Bias Variance Trade-Off



$$Y = f(X) + \varepsilon$$

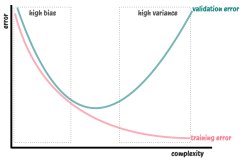
Sources of Error:

1. Irreducible: ε
2. Reducible: $\hat{f}(x)$ not similar to $f(x)$

16

Formalizing Bias Variance Trade-Off

Expected **test MSE**

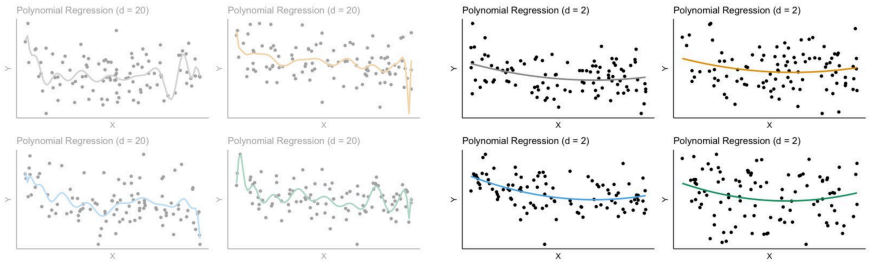


$$\underbrace{E\left(y_0 - \hat{f}(x_0)\right)^2}_{\text{expected MSE at } x_0 \text{ if we repeatedly estimated } f(x) \text{ with different training sets}} = \text{Var}(\hat{f}(x_0)) + \left[\text{bias}(\hat{f}(x_0))\right]^2 + \underbrace{\text{Var}(\varepsilon)}_{\text{irreducible error}}$$

Formalizing Bias Variance Trade-Off

Expected **test MSE**

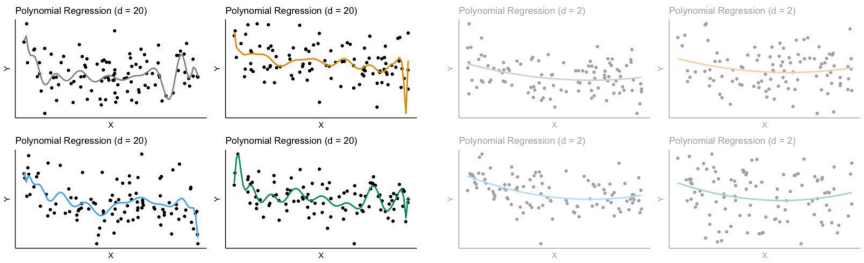
$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \text{Var}(\hat{f}(x_0)) + \left[\text{bias}(\hat{f}(x_0))\right]^2 + \text{Var}(\varepsilon)$$



Formalizing Bias Variance Trade-Off

Expected **test MSE**

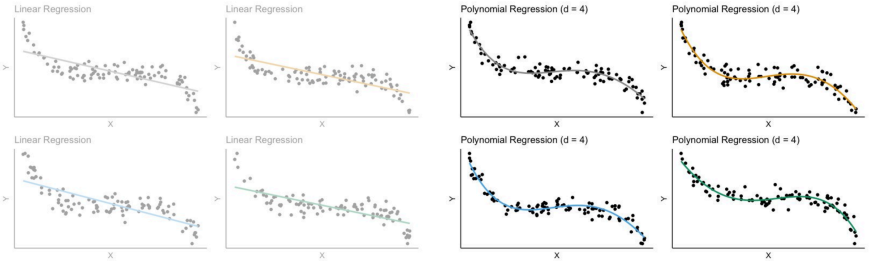
$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \text{Var}(\hat{f}(x_0)) + \left[\text{bias}(\hat{f}(x_0))\right]^2 + \text{Var}(\varepsilon)$$



Formalizing Bias Variance Trade-Off

Expected **test MSE**

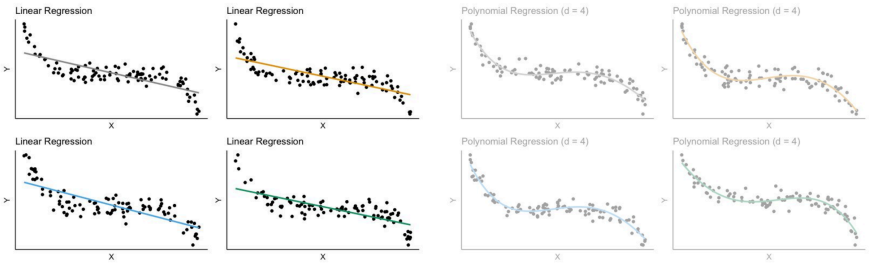
$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \text{Var}(\hat{f}(x_0)) + \left[\text{bias}(\hat{f}(x_0))\right]^2 + \text{Var}(\varepsilon)$$



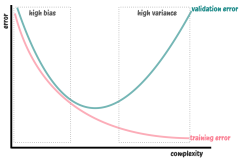
Formalizing Bias Variance Trade-Off

Expected **test MSE**

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \text{Var}(\hat{f}(x_0)) + \left[\text{bias}(\hat{f}(x_0))\right]^2 + \text{Var}(\varepsilon)$$



Formalizing Bias Variance Trade-Off

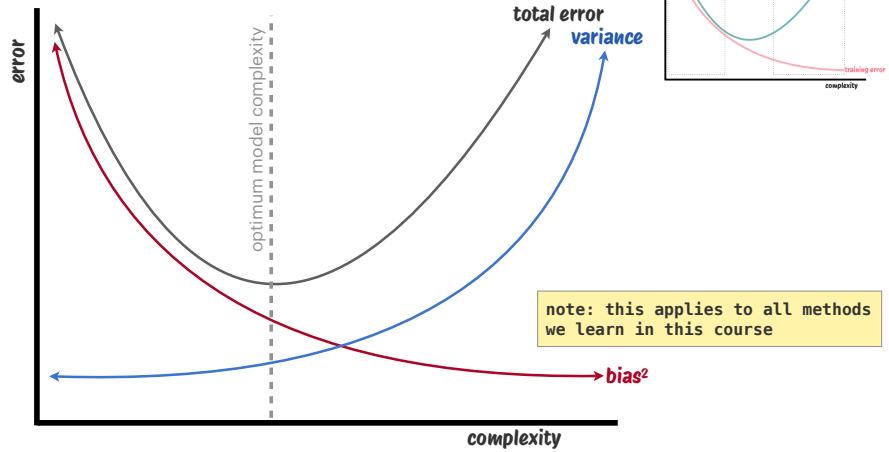


Expected **test MSE**

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \underbrace{\text{Var}(\hat{f}(x_0))}_{\text{variance increases with complexity}} + \underbrace{\left[\text{bias}(\hat{f}(x_0))\right]^2}_{\text{bias decreases with complexity}} + \text{Var}(\varepsilon)$$

[try it out: https://floswald.shinyapps.io/bias_variance/

Bias Variance Trade-Off



note: this applies to all methods we learn in this course

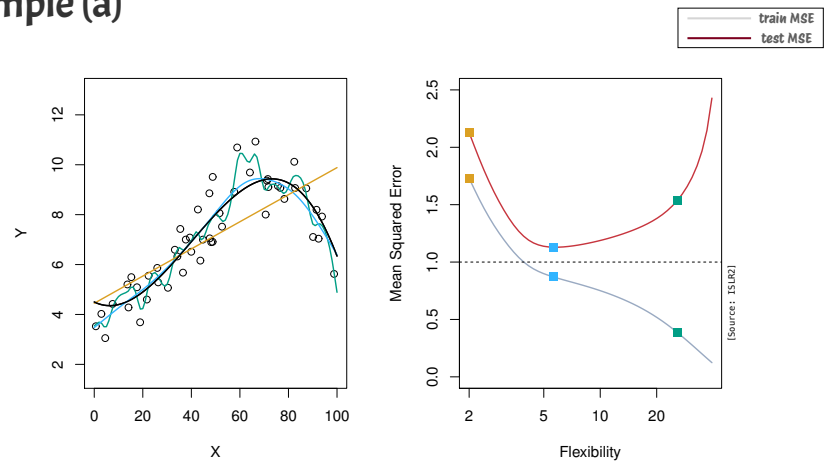
Bias Variance Trade-Off

There commonly used methods for finding the optimal model between simple and complicated are

- regularization
- boosting
- bagging

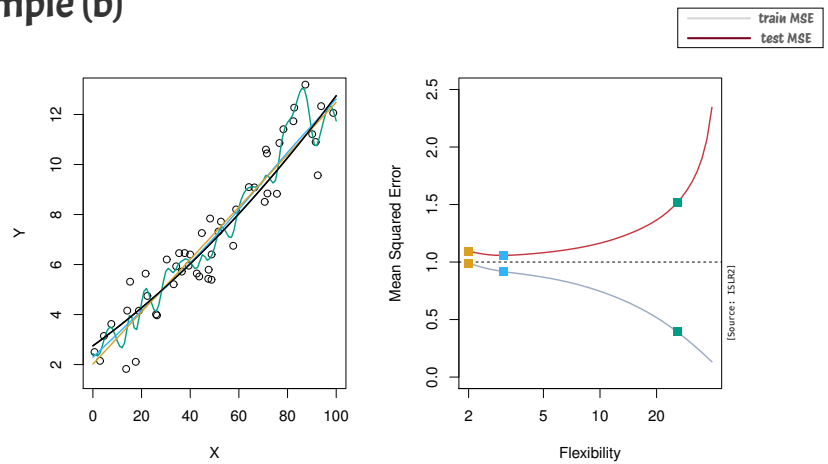
(the key here is **model validation**)

Example (a)



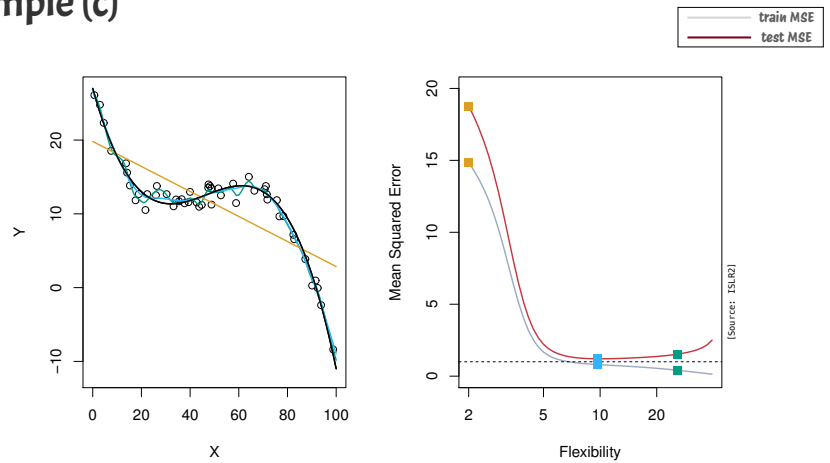
25

Example (b)



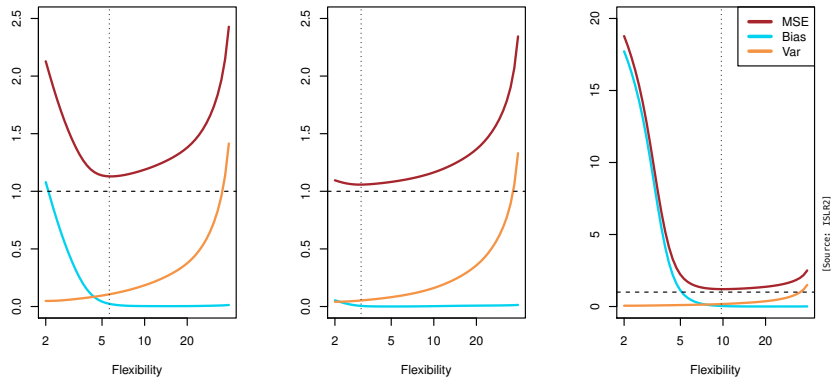
26

Example (c)



27

Example (a)-(c): Bias Variance Trade-Off



28

A Simulation Example

Estimate the conditional mean of Y given $\boldsymbol{X}$

$$Y = f(\boldsymbol{X}) + \varepsilon$$

Assume probability model:

$$Y = 1 - 2x - 3x^2 + 5x^3 + \varepsilon$$

where $\varepsilon \sim \mathrm{N}(0,\sigma^2)$

Alternatively:

$$Y \mid X \sim \mathrm{N}(1 - 2x - 3x^2 + 5x^3, \sigma^2) \quad \text{or}$$

$$\mu(x) = E[Y \mid \boldsymbol{X} = \boldsymbol{x}] = 1 - 2x - 3x^2 + 5x^3$$

conditional mean is a **linear combination of the feature variables**

note: the true probability model and thus also $\mu(x)$ are often not known!

A Simulation Example

1. Simulate data from assumed probability model

$$Y = 1 - 2x - 3x^2 + 5x^3 + \varepsilon$$

2. Fit three models to data:

I. Degree 1 Polynomial $\mu(x) = \beta_0 + \beta_1 x$

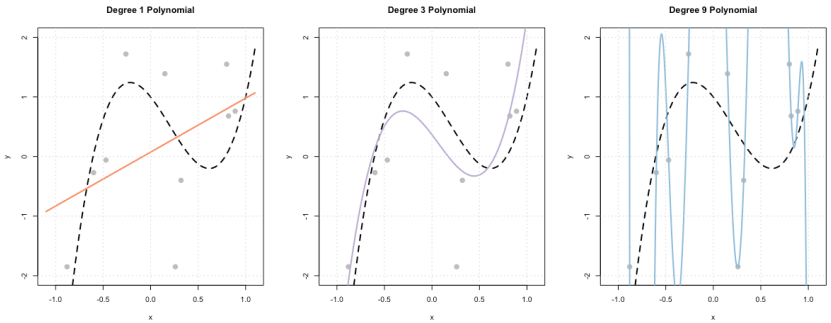
II. Degree 3 Polynomial $\mu(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$

III. Degree 9 Polynomial $\mu(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \ldots + \beta_9 x^9$

ENTER R

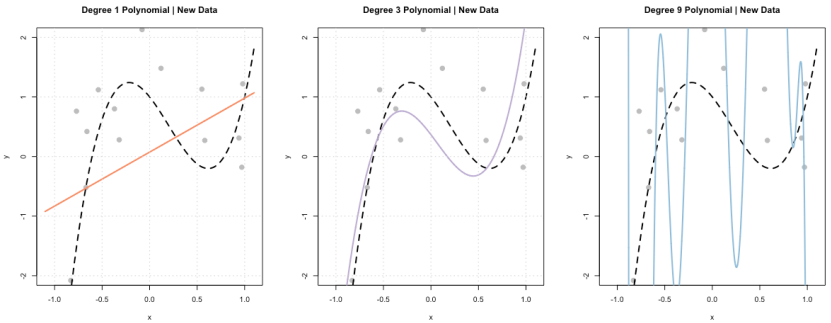
A Simulation Example

- How close is the estimated regression (mean) function to the data?
- How close is the estimated regression (mean) function to the true regression (mean) function?



A Simulation Example

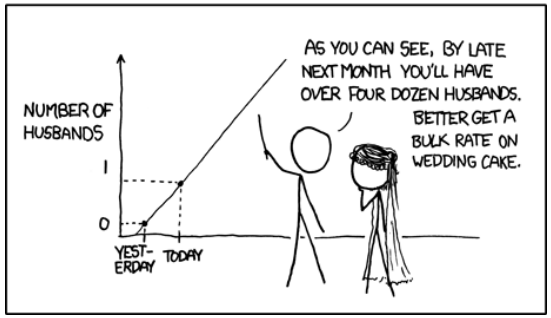
Generate new data and check...



Some Do Not's!

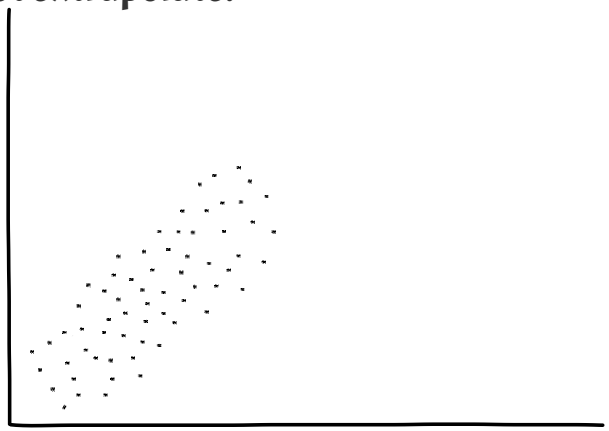
33

Extrapolation



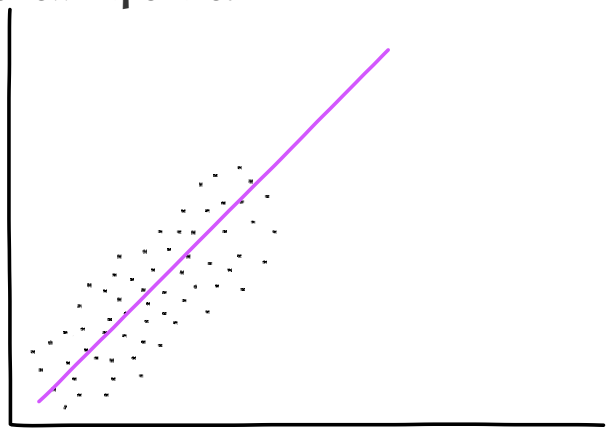
34

Do not extrapolate!



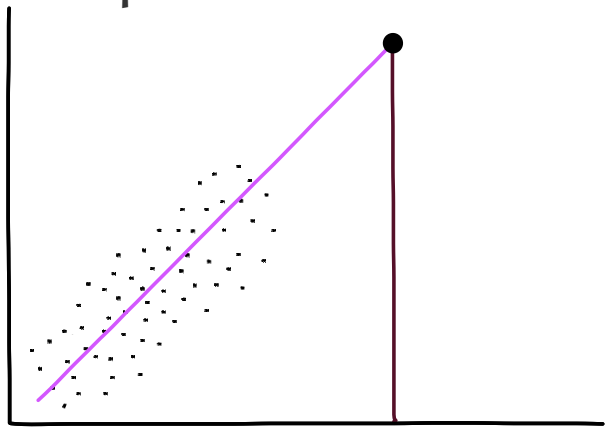
35

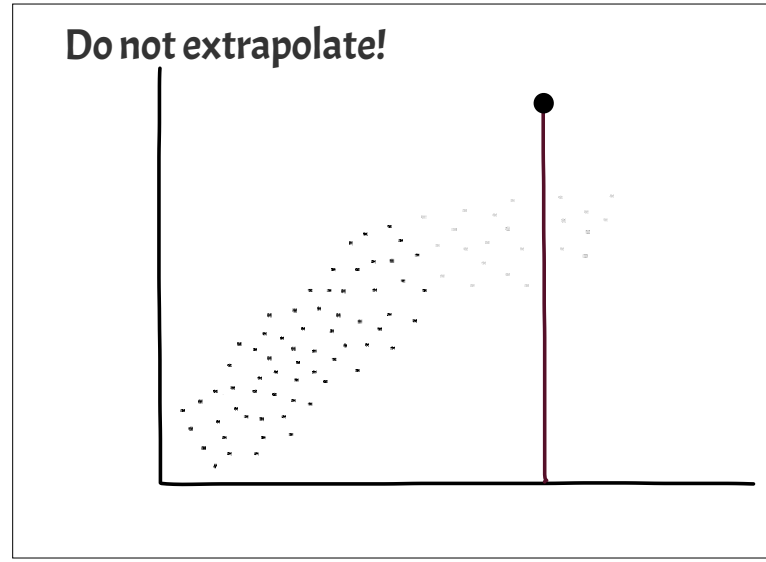
Do not extrapolate!



36

Do not extrapolate!



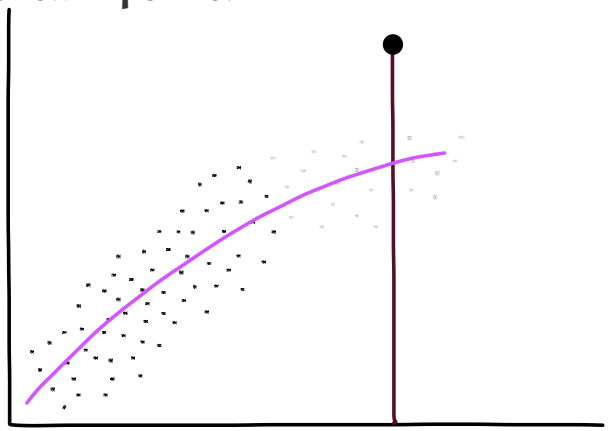


37

Do not extrapolate!

38

Do not extrapolate!



39

Do not fit model on test data!

In addition to the train-test split,
we will later split the data into validation set

40

This Week's Practical

Linear Regression: Evaluate Model Performance