

# What Do We Mean by “Fitting Models to Data”?

*Brady T. West*

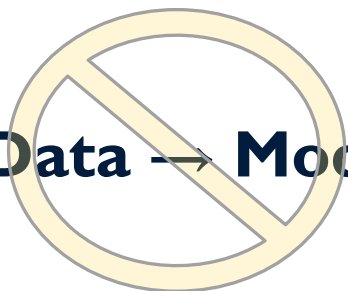
# Fitting Models to Data

**Goal:** How to fit **statistical models** to **data**  
to help answer research questions

# Fitting Models to Data

**Goal:** How to fit **statistical models** to **data**  
to help answer research questions

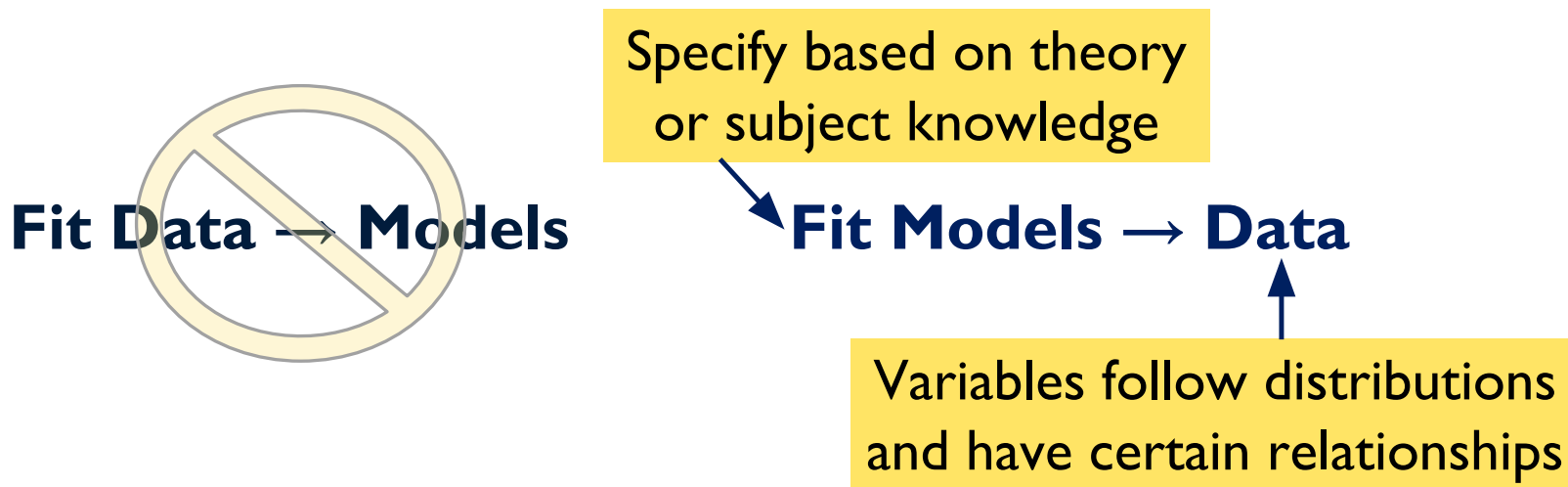
**Fit Data → Models**



**Fit Models → Data**

# Fitting Models to Data

**Goal:** How to fit **statistical models** to **data**  
to help answer research questions



# Fitting Models to Data

## Why do we fit models to data?

- **Estimate** distributional properties of variables, potentially conditional on other variables
- Concisely **summarize relationships** between variables, and make inferential statements about those relationships
- **Predict** values of variables of interest conditional on values of other predictor variables, and characterize prediction uncertainty

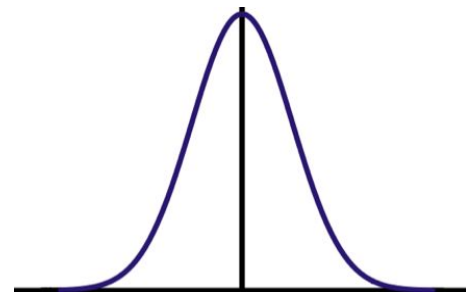
means  
variances  
quantiles

# Fitting Models to Data

- Focus on **parametric models** → estimating **parameters** that describe the distributions of variables

# Fitting Models to Data

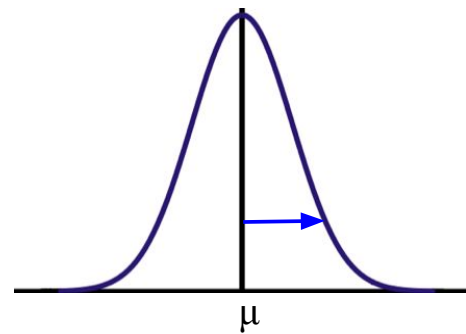
- Focus on **parametric models** → estimating **parameters** that describe the distributions of variables
- Given data, suggest **variable of interest** follows certain **probability model**



e.g. normal distribution

# Fitting Models to Data

- Focus on **parametric models** → estimating **parameters** that describe the distributions of variables
- Given data, suggest **variable of interest** follows certain **probability model**  
→ estimate parameters that define that model



e.g. normal distribution

e.g. mean and variance



# Fitting Models to Data

Estimate model parameters + sampling variance  
= **make inference about parameters**  
by testing hypotheses or generating confidence intervals

**Course 2!**

# Fitting Models to Data

Estimate model parameters + sampling variance  
= **make inference about parameters**  
by testing hypotheses or generating confidence intervals


**Course 2!**

**Up next ...**

**Example** of specifying probability model (given a research question)  
and estimating the parameters of that model

Idea of **assessing model fit**: *Does model seem to fit observed data well?*

# Example: Test Performance and Age



**Variable of Interest**  
**Test Performance**  
(0 – 8 points)

# Example: Test Performance and Age


**Possible Predictor**  
**Age**  
(Standardized)



**Variable of Interest**  
**Test Performance**  
(0 – 8 points)

# Example: Test Performance and Age

**Possible Predictor**  
**Age**  
(Standardized)


  
Believe age has **curvilinear**  
relationship with  
performance

**Variable of Interest**  
**Test Performance**  
(0 – 8 points)

moderate values of age → performance best  
smaller or larger values of age → performance tends to be worse

# Example: Test Performance and Age

**Possible Predictor**  
**Age**  
**(Standardized)**

  
Believe age has **curvilinear**  
relationship with  
performance

**Variable of Interest**  
**Test Performance**  
**(0 – 8 points)**

moderate values of age → performance best  
smaller or larger values of age → performance tends to be worse

- Goals:**
- 1) estimate **marginal mean** of performance across all ages
  - 2) estimate mean performance **conditional** on age

# Example: Test Performance and Age

Performance follows **normal distribution** overall  
defined by mean and variance (two parameters)

**“mean-only” model**  
for performance

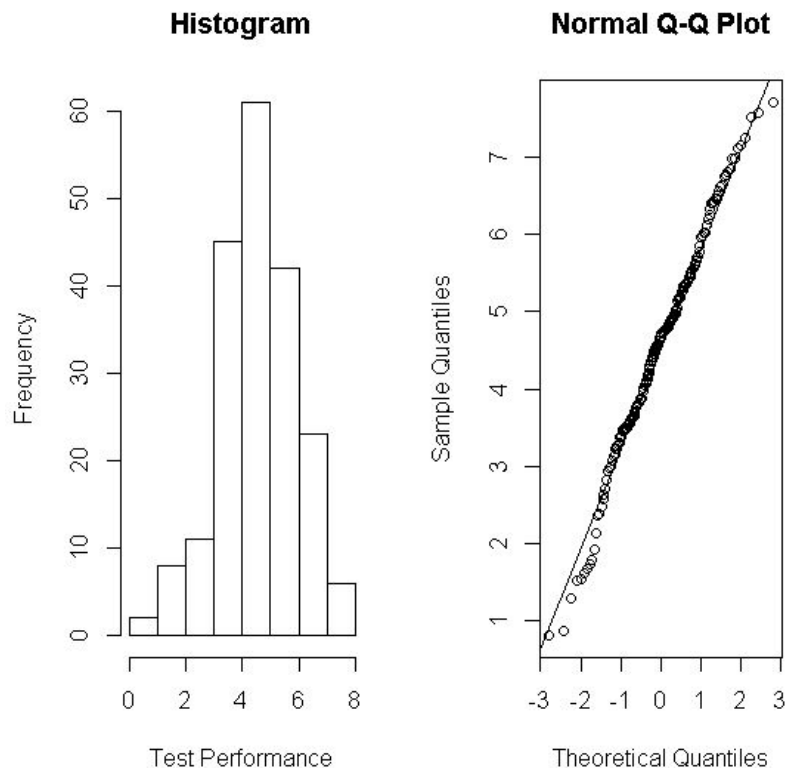
**Conditional on age**, performance follows **normal distribution**  
mean defined by a **quadratic function** of age  
 $a + b \cdot \text{age} + c \cdot \text{age}^2$  (three parameters:  $a$ ,  $b$ , and  $c$ )  
and variance  $\sigma^2$  (one parameter)

**Conditional model**  
for performance

# The Data: Performance

**Examine marginal distribution  
of performance ( $n = 200$ )  
via Histogram and Normal Q-Q plot**

Does the **normal distribution**  
seem like a **reasonable** model  
for performance?

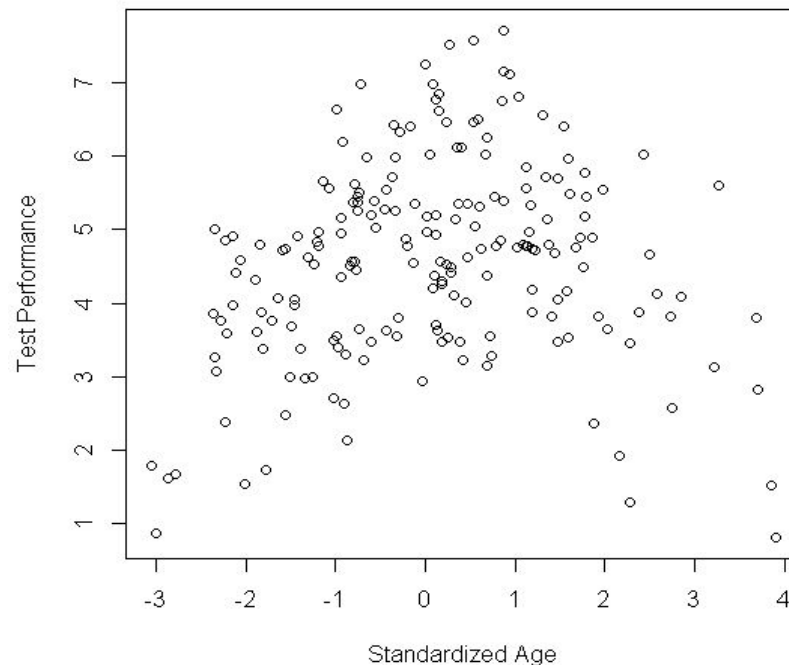




# The Data: Performance and Age

**Visualize relationship**  
between age and test performance  
via scatter plot

Support for our theory regarding  
**curvilinear** relationship?



# Fit the “Mean-Only” Model

Fit **regression model** to performance data

$$\text{perf} = m + e$$

1<sup>st</sup> parameter  
(*unknown constant*)

**m = marginal mean**

# Fit the “Mean-Only” Model

Fit **regression model** to performance data

$$\text{perf} = m + e$$

1<sup>st</sup> parameter  
(*unknown constant*)

**m = marginal mean**

**e = random error** defining each observation's  
deviation from the overall mean m,

Errors are **normally distributed** with mean 0 and variance  $\sigma^2$

where  $e \sim N(0, \sigma^2)$

2<sup>nd</sup> parameter

# Fit the “Mean-Only” Model

Fit **regression model** to performance data

$$\text{perf} = m + e$$

where  $e \sim N(0, \sigma^2)$

## Parameter Estimates

Estimate of overall (marginal) mean  $m = 4.57$  points

with  $SE = 0.10$  points

→ support overall mean non-zero!

Estimate of  $\sigma^2 = 1.82$  (points<sup>2</sup>)

# Fit the “Mean-Only” Model

Fit **regression model** to performance data

$$\text{perf} = m + e$$

where  $e \sim N(0, \sigma^2)$

*Need to check this!*

## Parameter Estimates

Estimate of overall (marginal) mean  $m = 4.57$  points

with  $SE = 0.10$  points

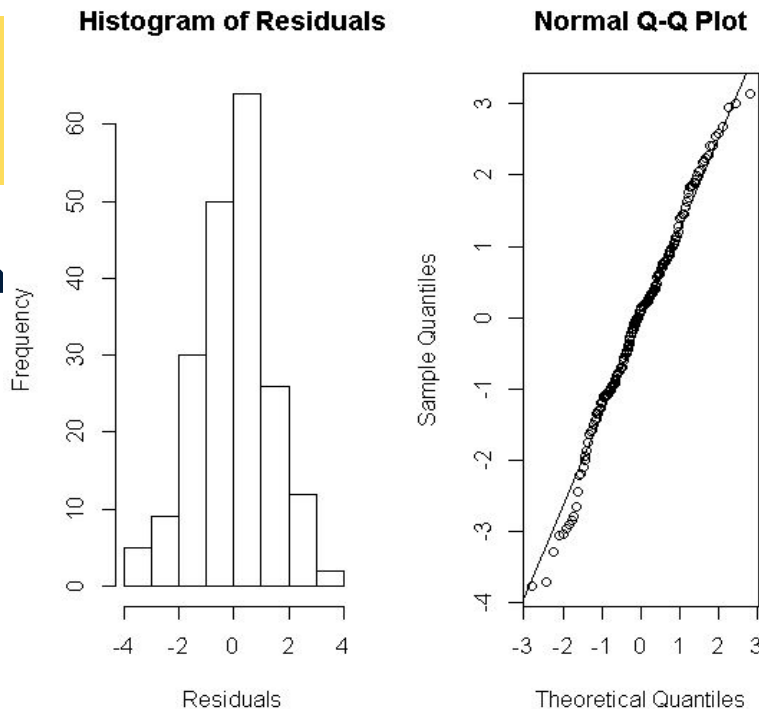
→ support overall mean non-zero!

Estimate of  $\sigma^2 = 1.82$  (points<sup>2</sup>)

# Assess Fit of “Mean-Only” Model

**Residuals** = realized values of random errors  
= observed performance – estimated mean  $m$

- **Examine realized residuals via histogram and normal Q-Q plot**  
to see if normal model is good fit for data
- **If normal model was not good fit**, would see large deviations from normality in realized residuals.



# Fit the Conditional Model

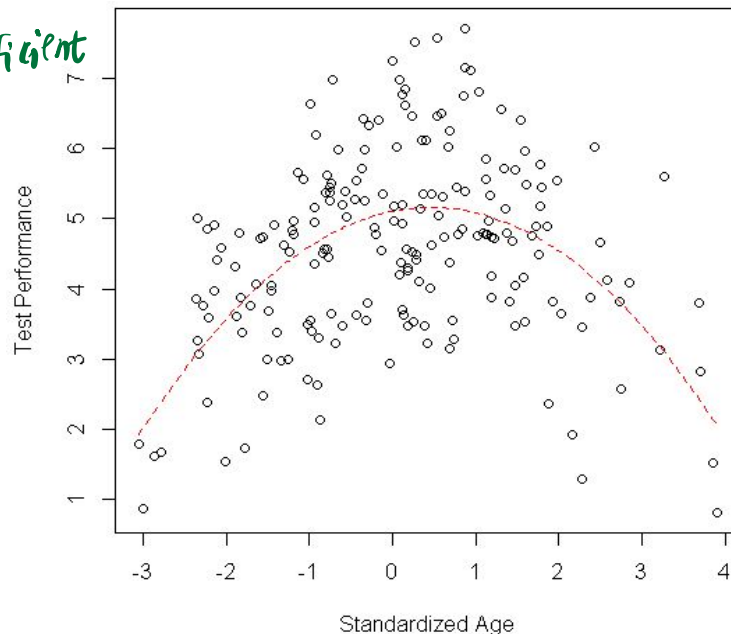
Fit **regression model**: regress performance on age and age<sup>2</sup>

$$\text{perf} = a + b \cdot \text{age} + c \cdot \text{age}^2 + e, \text{ where } \underline{e \sim N(0, \sigma^2)}$$

a, b, and c = three parameters *regression coefficient*

e = random error

Errors are normally distributed



# Fit the Conditional Model

Fit **regression model**: regress performance on age and age<sup>2</sup>

$$\text{perf} = a + b \cdot \text{age} + c \cdot \text{age}^2 + e, \quad \text{where } e \sim N(0, \sigma^2)$$

## Parameter Estimates

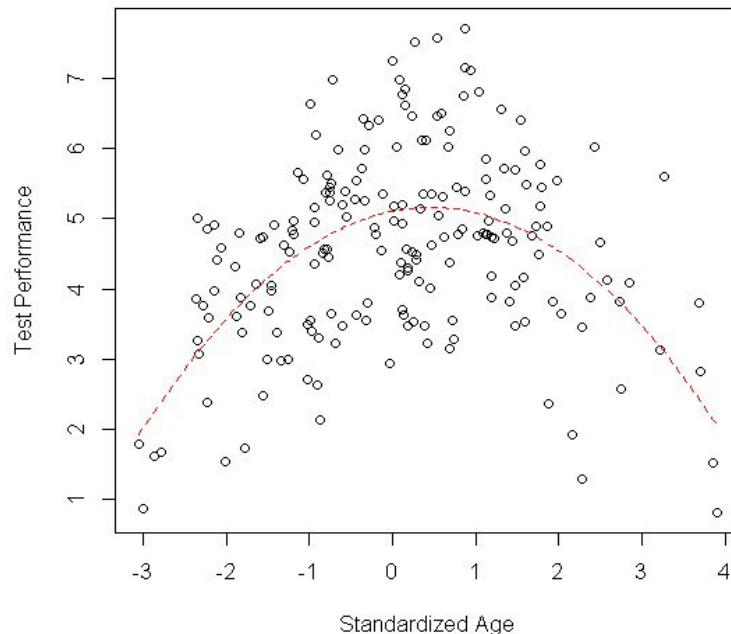
Estimate of  $a = 5.11$  (SE = 0.10)

Estimate of  $b = 0.24$  (SE = 0.06)

Estimate of  $c = -0.26$  (SE = 0.03)

→ support  $a, b, c$  non-zero!

Estimate of  $\sigma^2 = 1.29$  (points<sup>2</sup>)





# Fit the Conditional Model

Fit **regression model**: regress performance on age and age<sup>2</sup>

$$\text{perf} = a + b \cdot \text{age} + c \cdot \text{age}^2 + e, \quad \text{where } e \sim N(0, \sigma^2)$$

## Parameter Estimates

Estimate of  $a = 5.11$  (SE = 0.10)

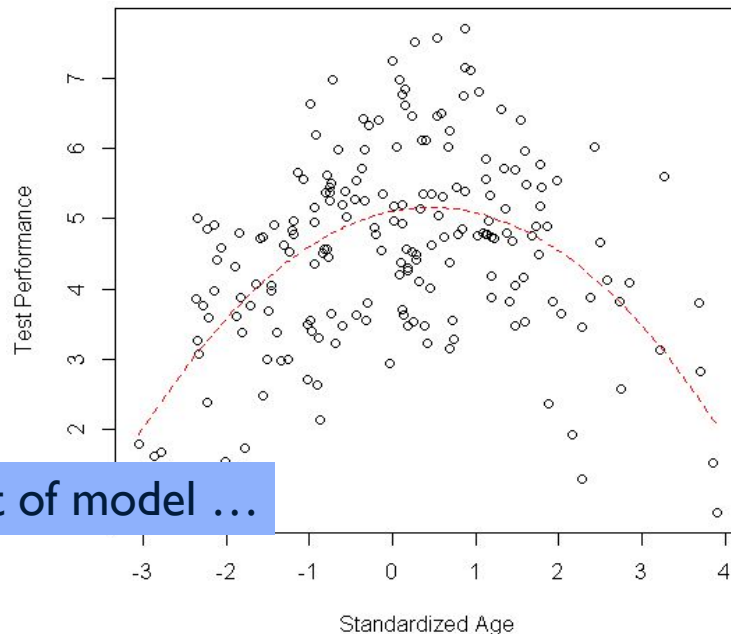
Estimate of  $b = 0.24$  (SE = 0.06)

Estimate of  $c = -0.26$  (SE = 0.03)

→ support  $a, b, c$  non-zero!

Estimate of  $\sigma^2 = 1.29$  (points<sup>2</sup>)

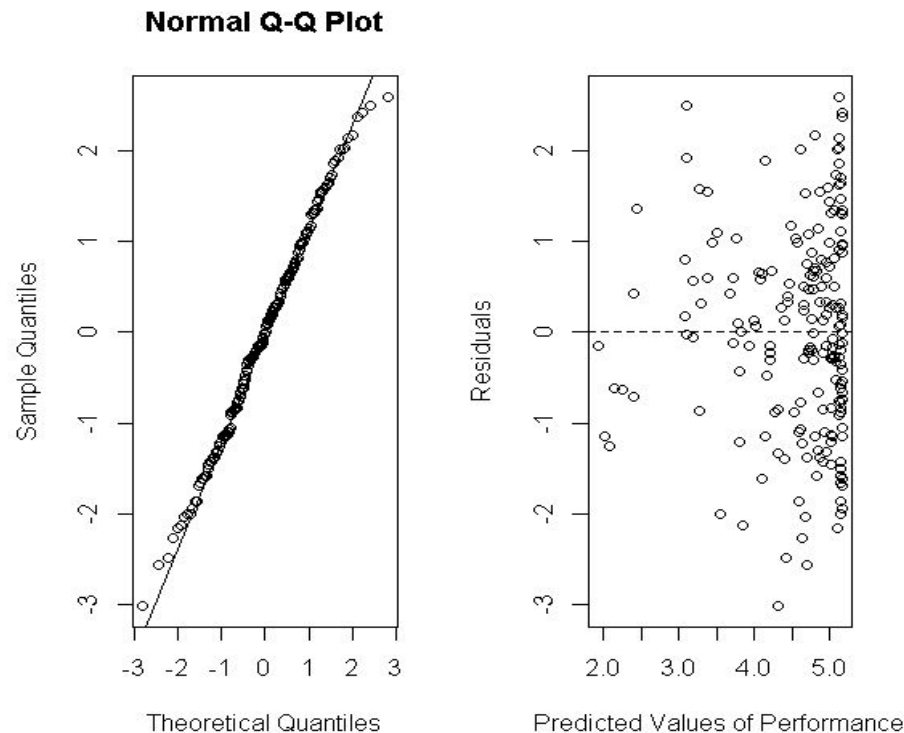
Assess fit of model ...



# Assess Fit of Conditional Model

See if residuals (realized values of  $e$ ):

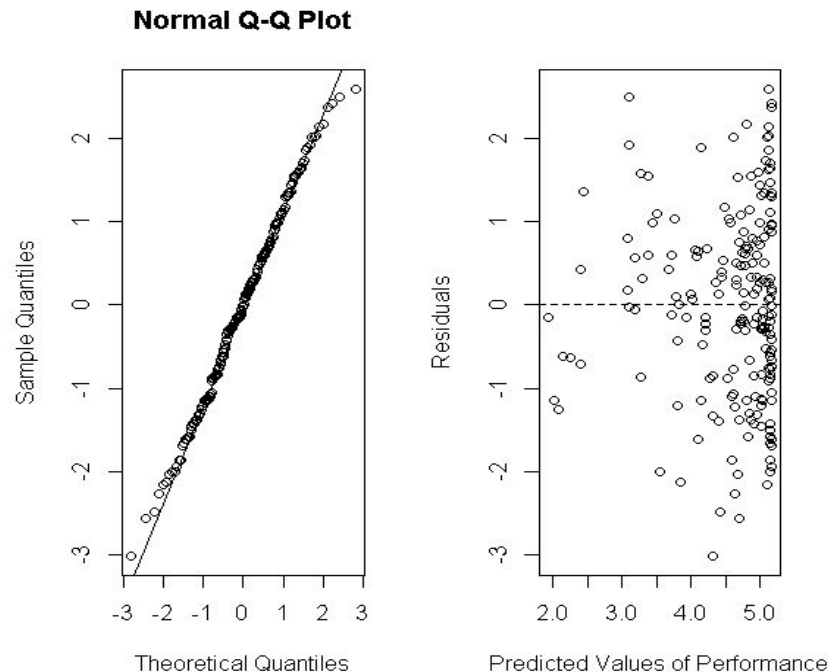
- appear to be normally distributed



# Assess Fit of Conditional Model

See if residuals (realized values of  $e$ ):

- appear to be normally distributed
- are symmetrically distributed around zero with constant variance  
(as function of **predicted values** of performance, given estimates of parameters  $a$ ,  $b$ , and  $c$ )



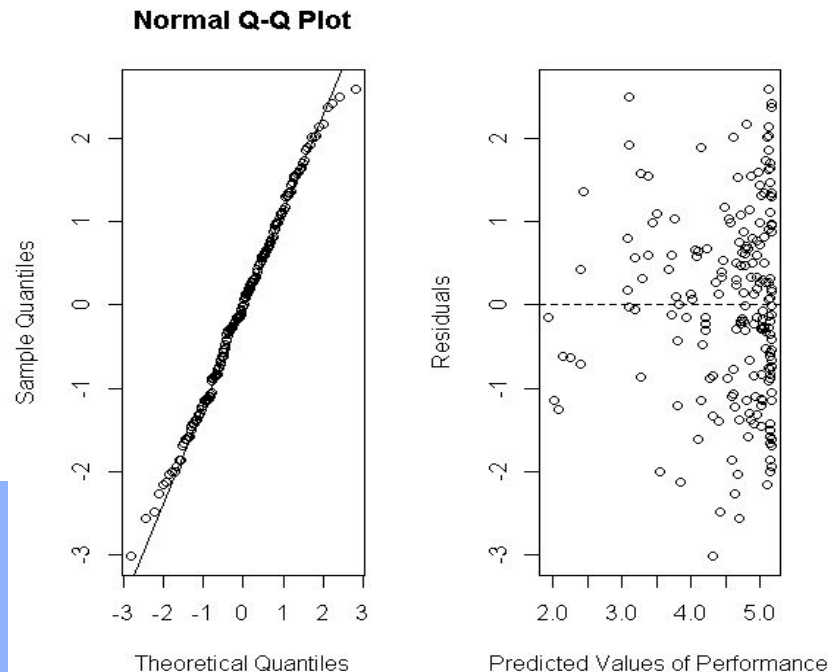
# Assess Fit of Conditional Model

See if residuals (realized values of  $e$ ):

- appear to be normally distributed
- are symmetrically distributed around zero with **constant variance**  
(as function of **predicted values** of performance, given estimates of parameters  $a$ ,  $b$ , and  $c$ )

**Model fit looks good!**

Could predict performance well,  
given standardized age ... can we do better?



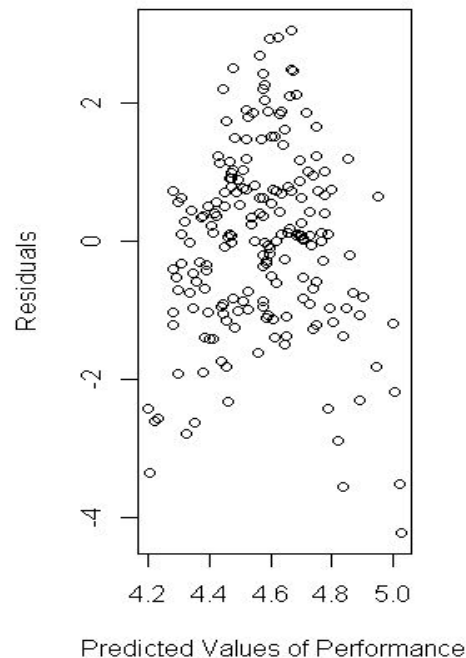
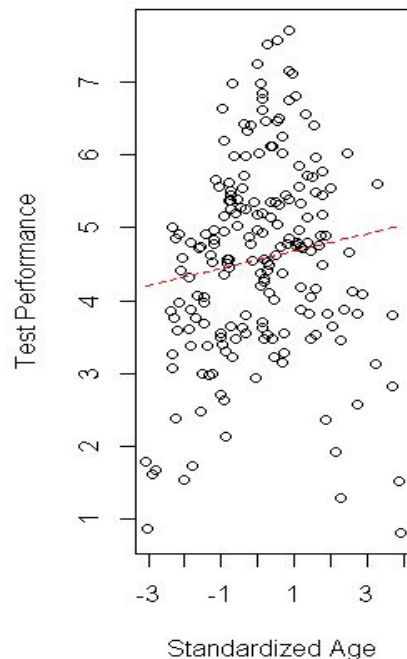
# A model that does not fit well...

## What if ...?

Fit **misspecified model** to data, assuming a **linear** relationship of performance and age?

## Model fit looks poor ...

residuals are **not** symmetrically scattered around zero;  
poor predictions at low and high values of age, higher  $\sigma^2$



# What did we see and what is next?

- **We have** ... introduced idea of fitting parametric models to data and assessing model fit
- **We will** ... talk about different types of variables interested in modeling, different types of data sets depending on study design (and implications for modeling), and different approaches to estimation and inference when fitting models
- **We will** ... discuss specific examples of modeling in detail!