# Selecting the “best” model

# Readings for today

- Chapter 6: Linear model selection and regularization. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning: with applications in R (Vol. 6). New York: Springer

# Topics

1. How to compare models
2. Full subset selection
3. Stepwise selection

# How to compare models

# Curse of dimensionality

Dimensions of a model

**n:** number of observations (i.e., rows)

**p:** number of features/independent variables (i.e., columns)

Dimensionality of a model: n x p

As n →p, dimensionality increases & model variance increases

$$\begin{pmatrix} x_{1,1} \\ \cdots \\ x_{n,1} \end{pmatrix} \rightarrow \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,15} \\ \cdots & & & \\ x_{n,1} & x_{n,2} & \cdots & x_{n,15} \end{pmatrix} \rightarrow \uparrow \text{ model flexibility}$$



(James et al. 2013)

# Problems from high dimensionality

**1. Prediction accuracy:** $\uparrow$ variance $=$ $\downarrow$ test accuracy

- $p > n$= no <u>unique</u> solution

**2. Model interpretability:** $\uparrow$ variance $=$ $\downarrow$ ability to interpret specific $X \rightarrow Y$ relationships

- Feature selection: determine which predictor variables matter



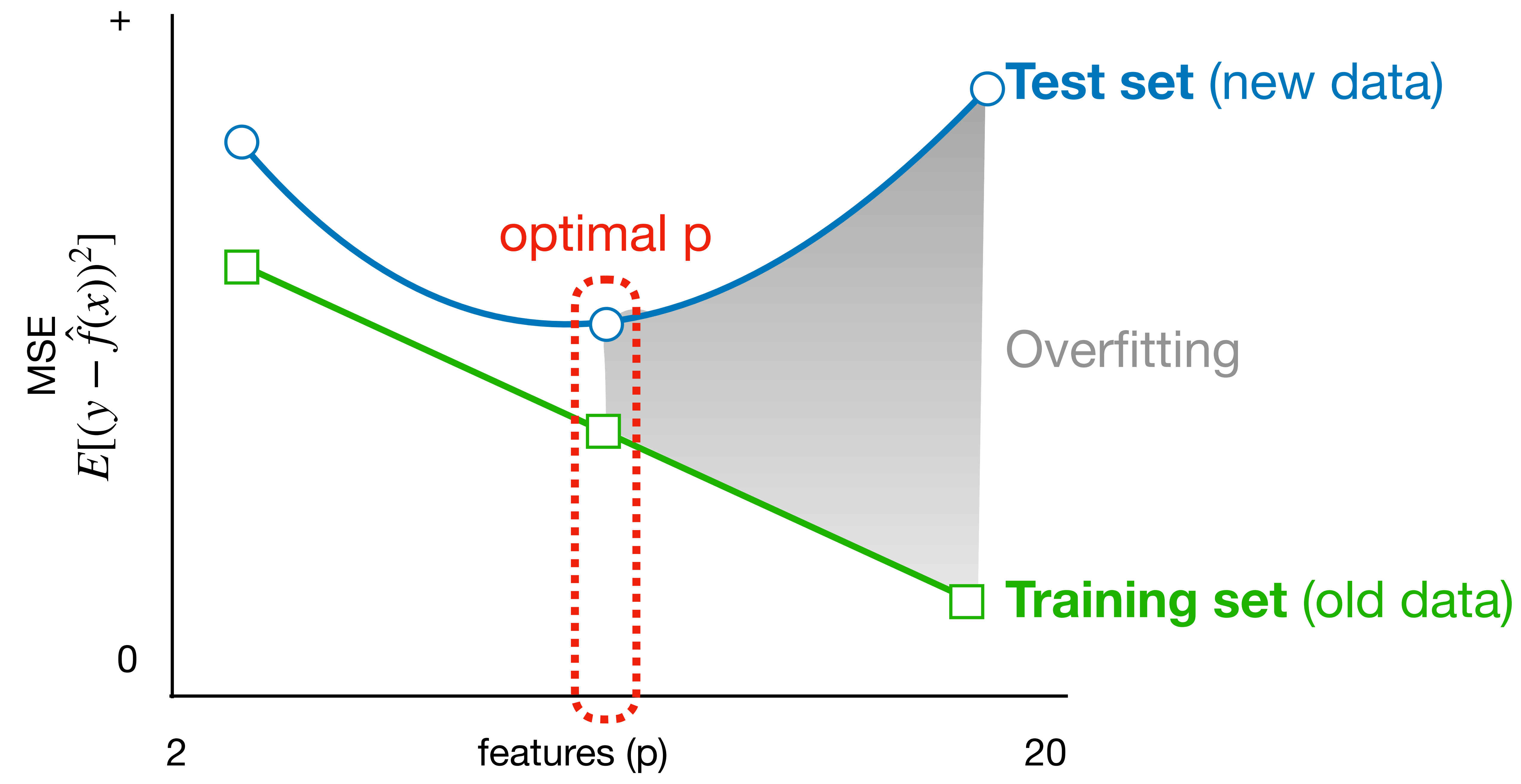
(James et al. 2013)

# Comparing models

Problem: How do you find the most parsimonious model?

Solutions

**1. Validation:** Use test error to guide choice of predictors






(James et al. 2013)

# Comparing models

Problem: How do you find the most parsimonious model?

Solutions

1. **Validation:** Use test error to guide choice of predictors

2. **Bias adjustment:** Adjust the *training* error estimate to account for differences in p (i.e., dimensionality)



(James et al. 2013)

# Bias adjusted criteria

**1. Mallow's Cp**

$$C_p = \frac{1}{n}(RSS + 2p\sigma_y^2)$$

**2. Akaike Information Criterion (AIC)**

$$AIC = \frac{1}{n\sigma_y^2}(RSS + 2p\sigma_y^2)$$
$$= 2p - 2\log(L)$$

$$L = P(Y|X, \beta, \sigma)$$

**3. Bayesian Information Criterion (BIC)**

$$BIC = \frac{1}{n}(RSS + \log(n)p\sigma_y^2)$$
$$= p\log(n) - 2\log(L)$$

**4. Adjusted $r^2$**

$$adj\ r^2 = 1 - \frac{\frac{1}{n-p-1}RSS}{\frac{1}{n-1}TSS}$$



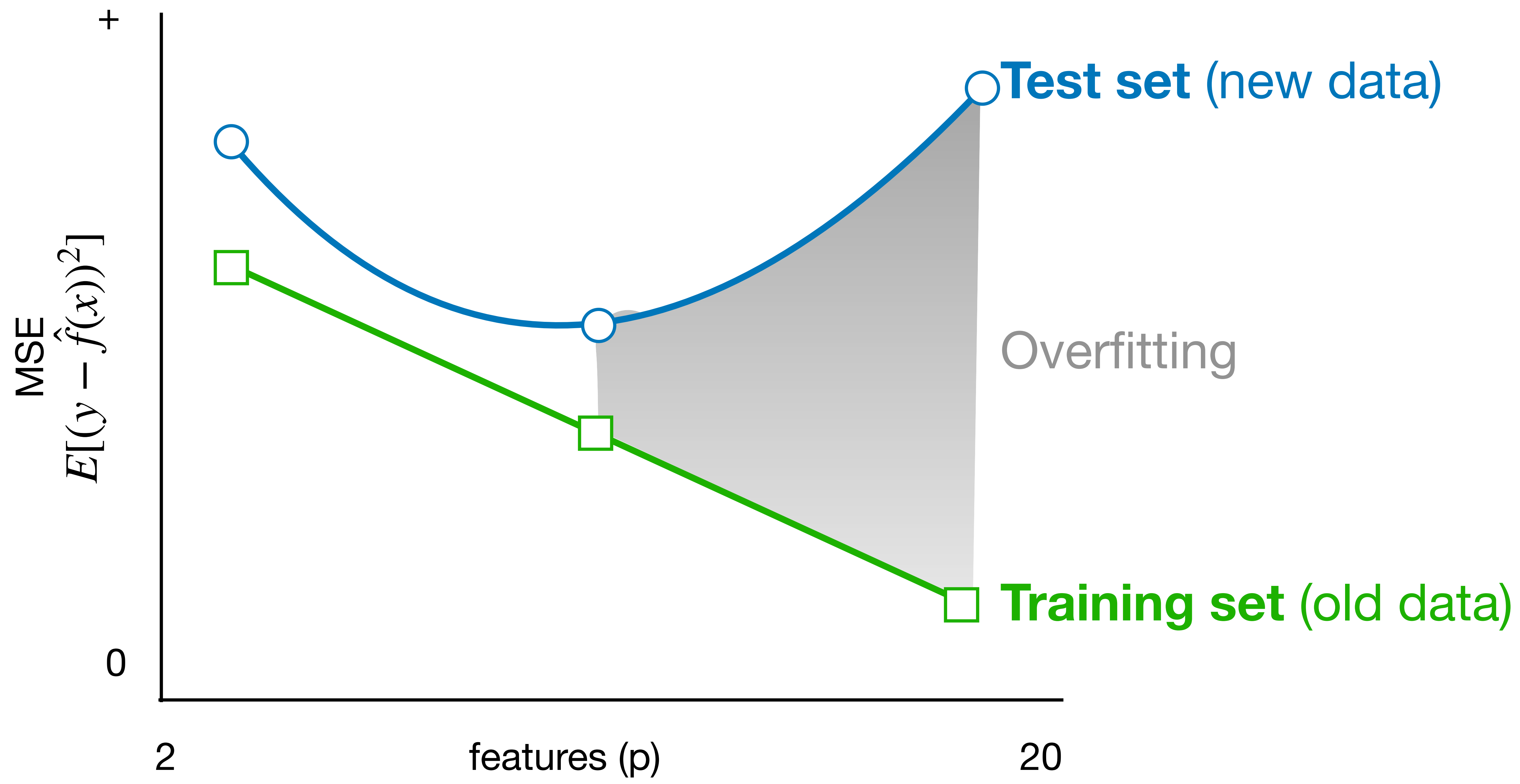
(James et al. 2013)

# Full subset selection

# What factors are relevant?

Problem: How do you find the most parsimonious model?

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p Xp + \epsilon$$




What is the best combination of variables that provides the best explanation for your data *after accounting for model complexity?*



(James et al. 2013)

# Full subset selection

Goal: Try all possible permutations of your model

eg: $Y = \beta_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3$

Model variants: $2^p$

$Y = \beta_0$

$Y = \beta_0 + \hat{\beta}_1 X_1$

$Y = \beta_0 + \hat{\beta}_2 X_2$

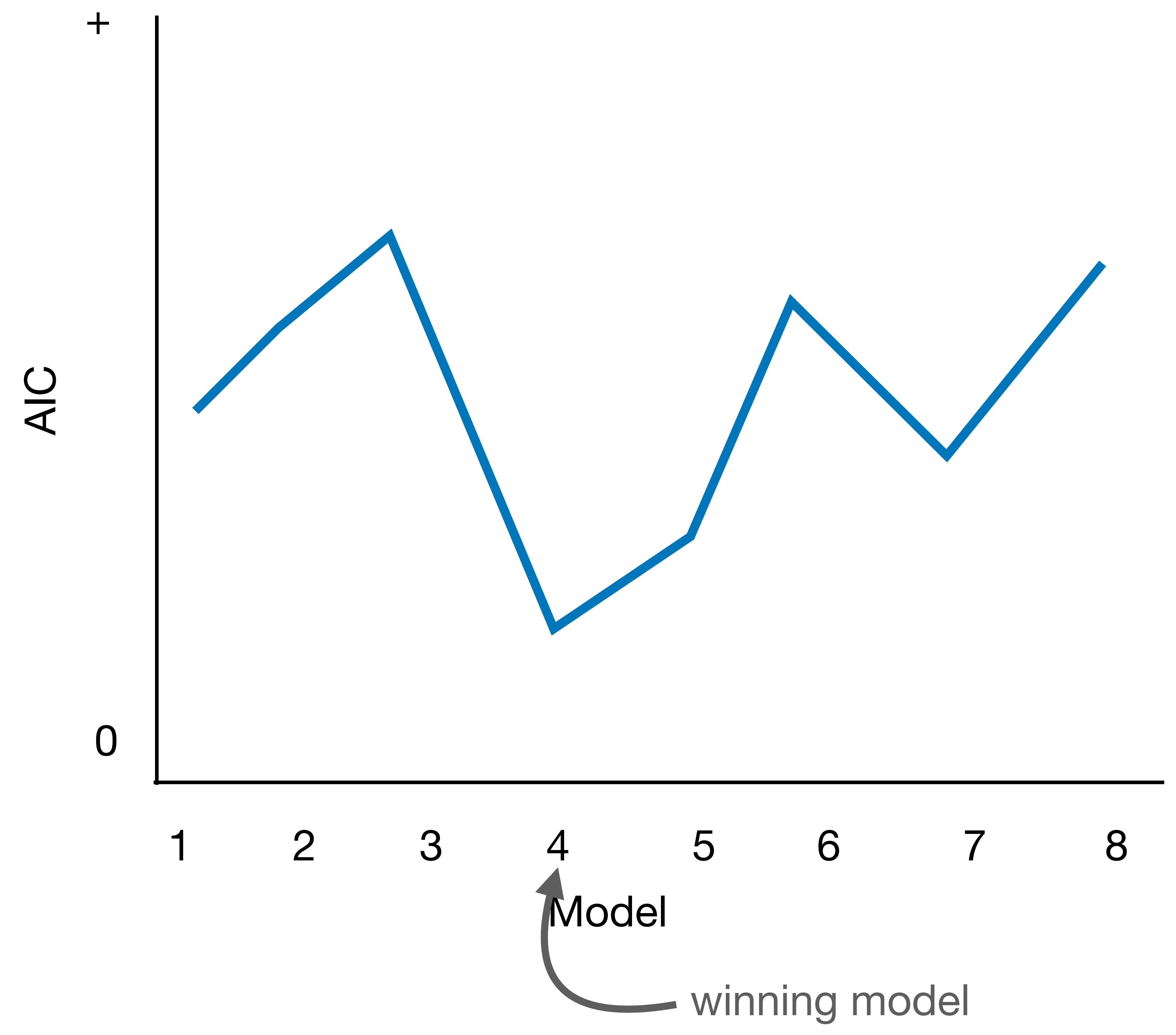
$\boxed{Y = \beta_0 + \hat{\beta}_3 X_3}$

$Y = \beta_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$

$Y = \beta_0 + \hat{\beta}_1 X_1 + \hat{\beta}_3 X_3$

$Y = \beta_0 + \hat{\beta}_2 X2 + \hat{\beta}_3 X_3$

$Y = \beta_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3$






(James et al. 2013)

# Stepwise selection

# Forward stepwise selection

Goal: Build up from the simplest model. Evaluate in stages of same $p$.

eg: $Y = \beta_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3$

Model variants: $1 + \dfrac{p(p+1)}{2}$

**1.** $Y = \beta_0$

**2.** $\boxed{Y = \beta_0 + \hat{\beta}_1 X_1}$

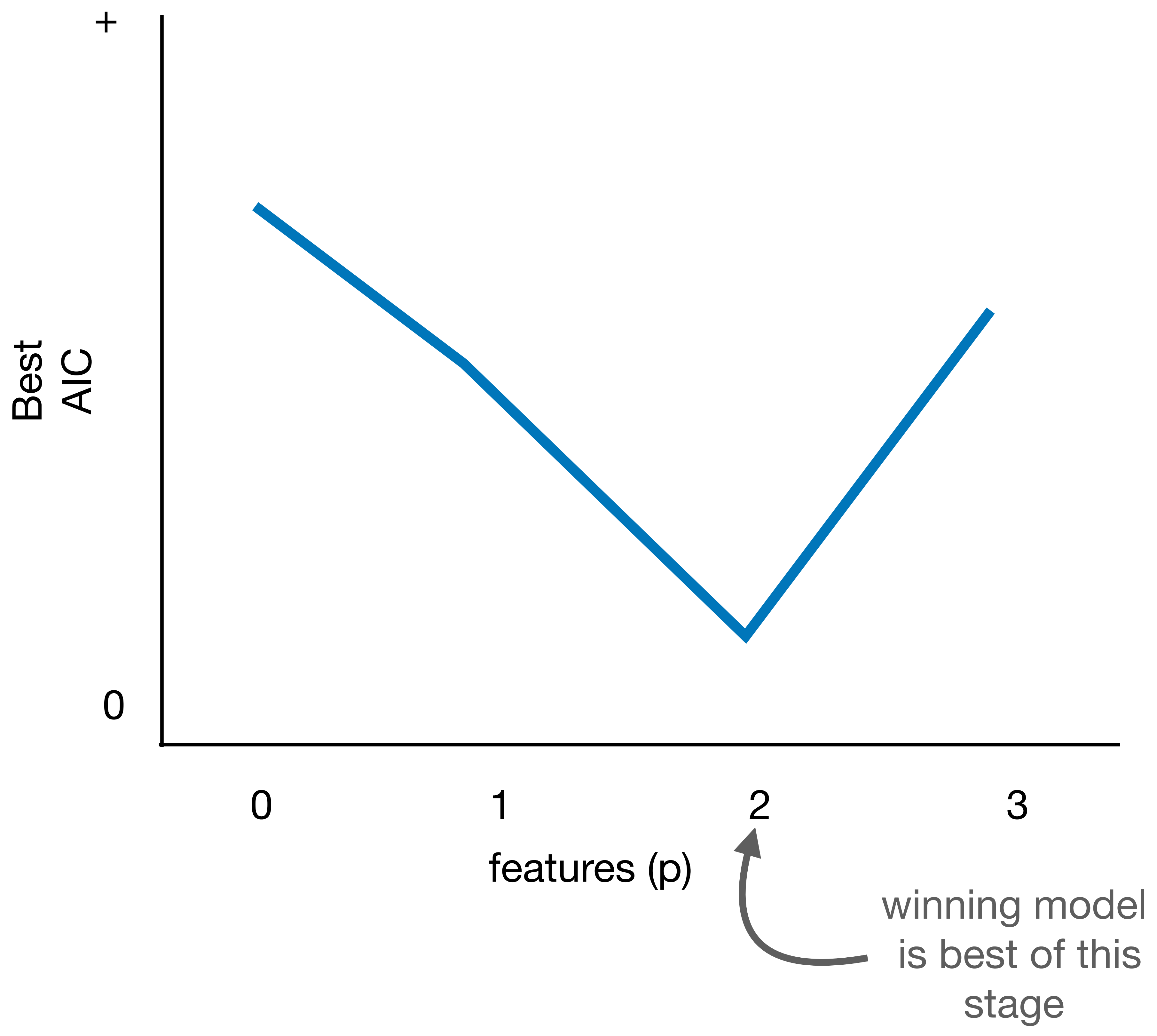
$Y = \beta_0 + \hat{\beta}_2 X_2$

$Y = \beta_0 + \hat{\beta}_3 X_3$

**3.** $Y = \beta_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$

$\boxed{Y = \beta_0 + \hat{\beta}_1 X_1 + \hat{\beta}_3 X_3}$

**4.** $Y = \beta_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3$






(James et al. 2013)

# Backwards stepwise selection

Goal: Build down from most complex model. Evaluate in stages of same $p$.

eg: $Y = \beta_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3$

Model variants: $1 + \dfrac{p(p+1)}{2}$

**1.** $Y = \beta_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3$

**2.** $Y = \beta_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$

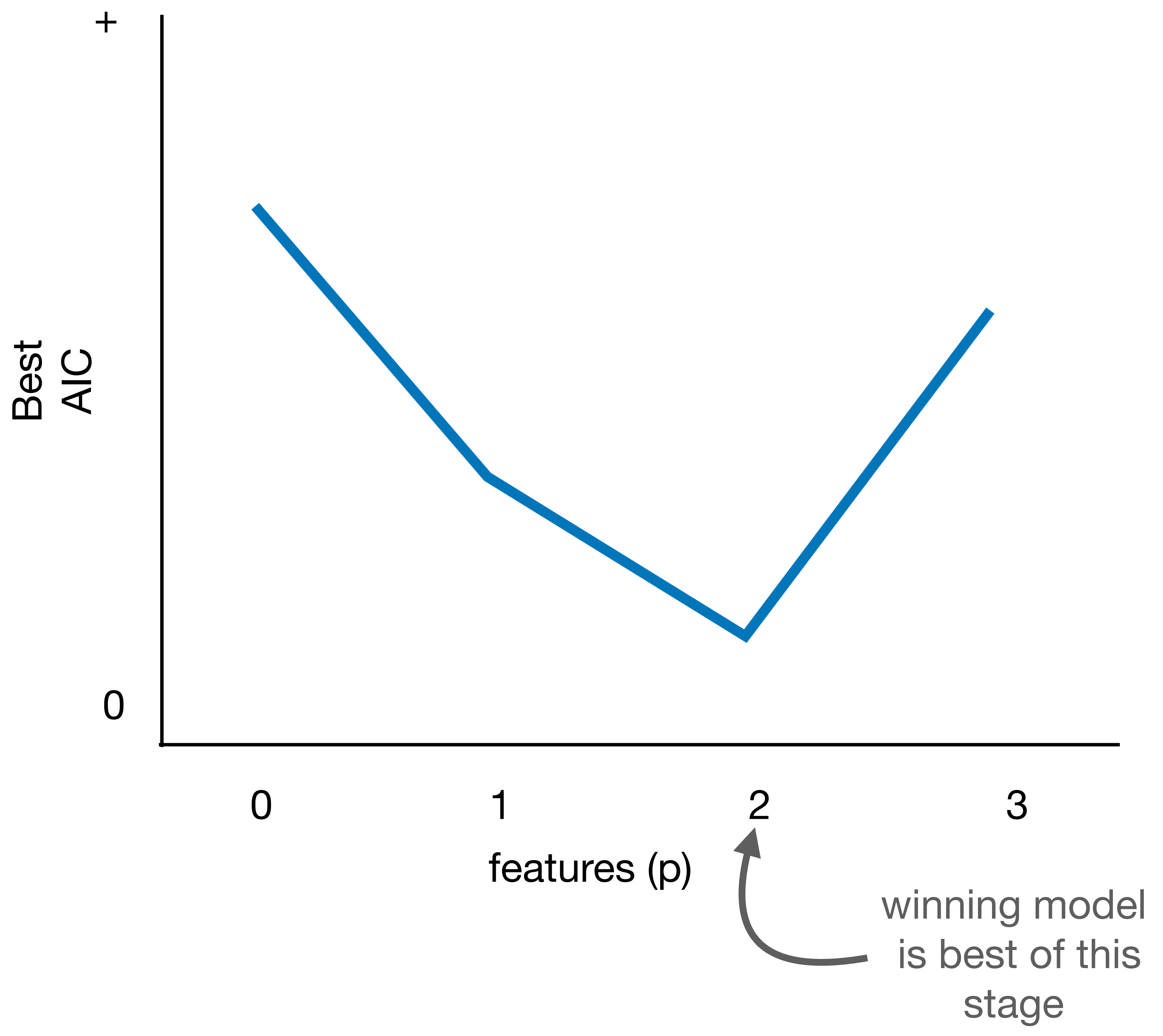
$\boxed{Y = \beta_0 + \hat{\beta}_1 X_1 + \hat{\beta}_3 X_3}$

$Y = \beta_0 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3$

**3.** $\boxed{Y = \beta_0 + \hat{\beta}_1 X_1}$

$Y = \beta_0 + \hat{\beta}_3 X_3$

**4.** $Y = \beta_0$



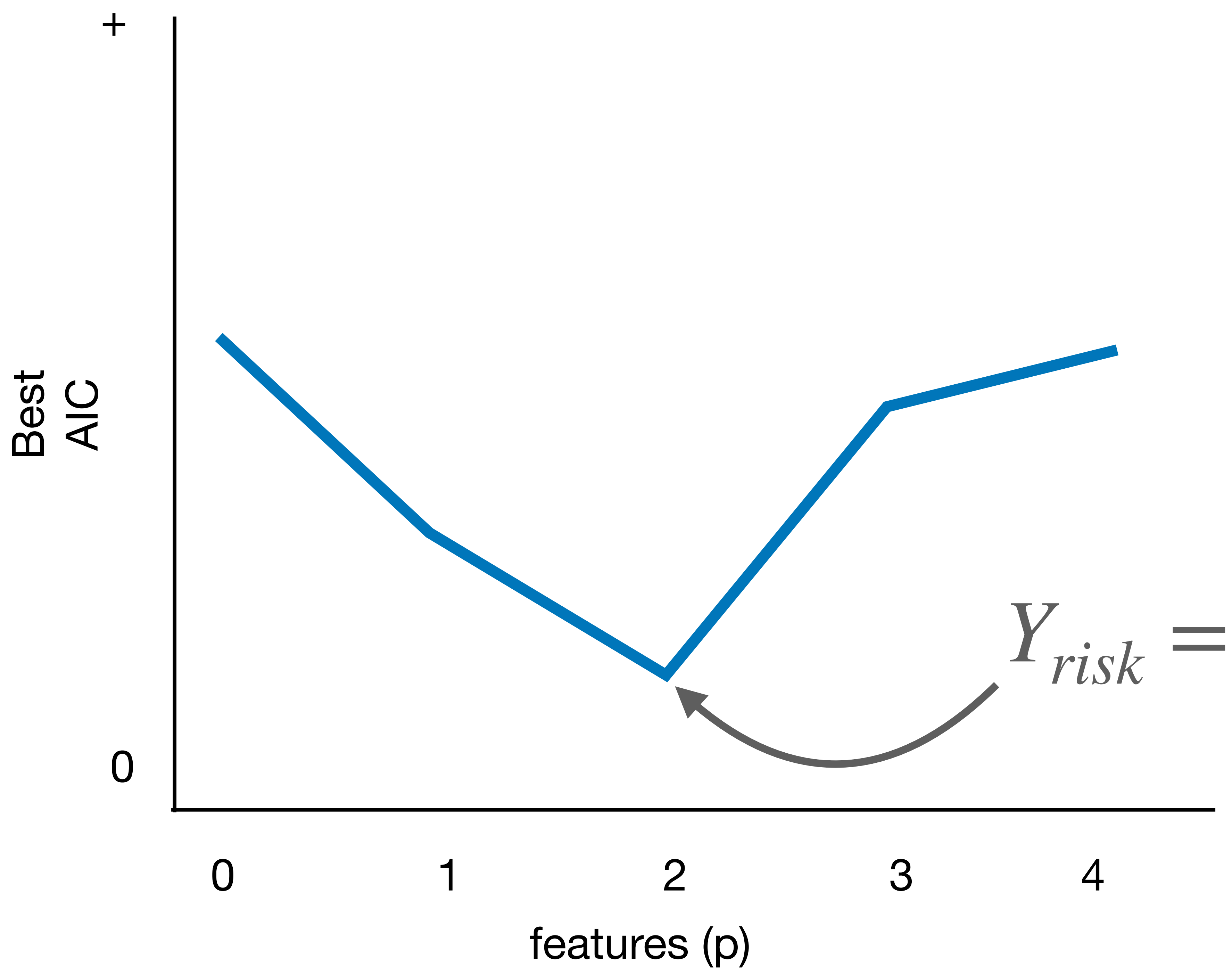



(James et al. 2013)

# Inference via model selection

Q: Is risk aversion associated with childhood trauma, parental income, psychiatric risk, and/or social network size?

$$Y_{risk} = \hat{\beta}_0 + \hat{\beta}_1 X_{CT} + \hat{\beta}_2 X_{income} + \hat{\beta}_3 X_{psych} + \hat{\beta}_4 X_{social}$$




$$Y_{risk} = \hat{\beta}_0 + \hat{\beta}_1 X_{CT} + \hat{\beta}_3 X_{psych}$$

Conclusion: Only childhood trauma & psychiatric risk are relevant for understanding variation in risk aversion.



(James et al. 2013)

# Speed vs. completeness

Full subset: $2^p$

Stepwise: $1 + \dfrac{p(p+1)}{2}$

Stepwise methods improve speed dramatically with more complex models.

$p$

| | 5 | 10 | 15 | 20 |
|---|---|---|---|---|
| **Full Subset** | 32 | 1,024 | 32,768 | 1,048,576 |
| **Stepwise** | 16 | 56 | 121 | 211 |

$\}$ models compared



(James et al. 2013)

# Take home message

- Model selection approaches allow for you to find the most parsimonious at explaining variance of your response variable, $Y$, while accounting for model complexity.