Linear models

Readings for today

- Chapter 3: Linear regression. 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. Ordinary least squares regression

2. Polynomial models

3. Isomorphisms with other statistical tests

Ordinary least squares regression

Structure of a linear regression model

Fundamental form:

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

$$= \underbrace{\beta_0 + \beta_1 X_1 + \ldots + \beta_p Xp + \epsilon}_{\text{slope} * X + \text{intercept}}$$

Assumptions:

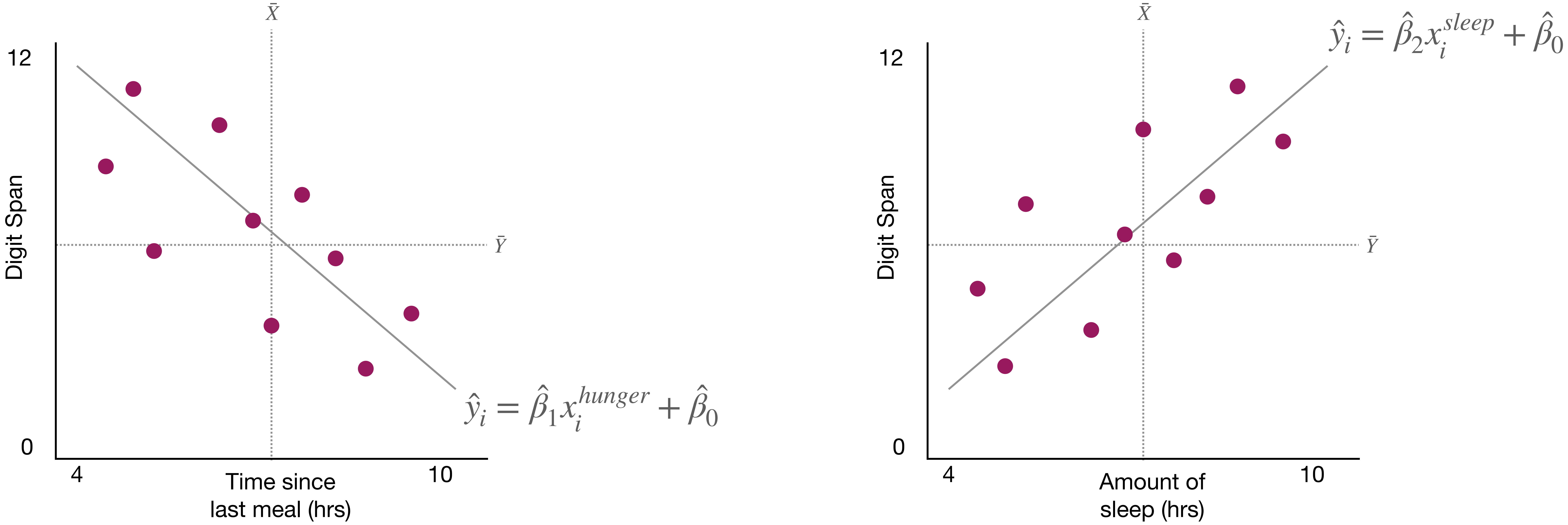
1. $f(X)$ describes a linear relationship between X and Y.
2. Y is normally distributed.
3. There is no collinearity between features in X.
4. $f(X)$ is stationary.

i.i.d.

"independent and identically distributed."

(James et al. 2013)

Structure of a linear regression model



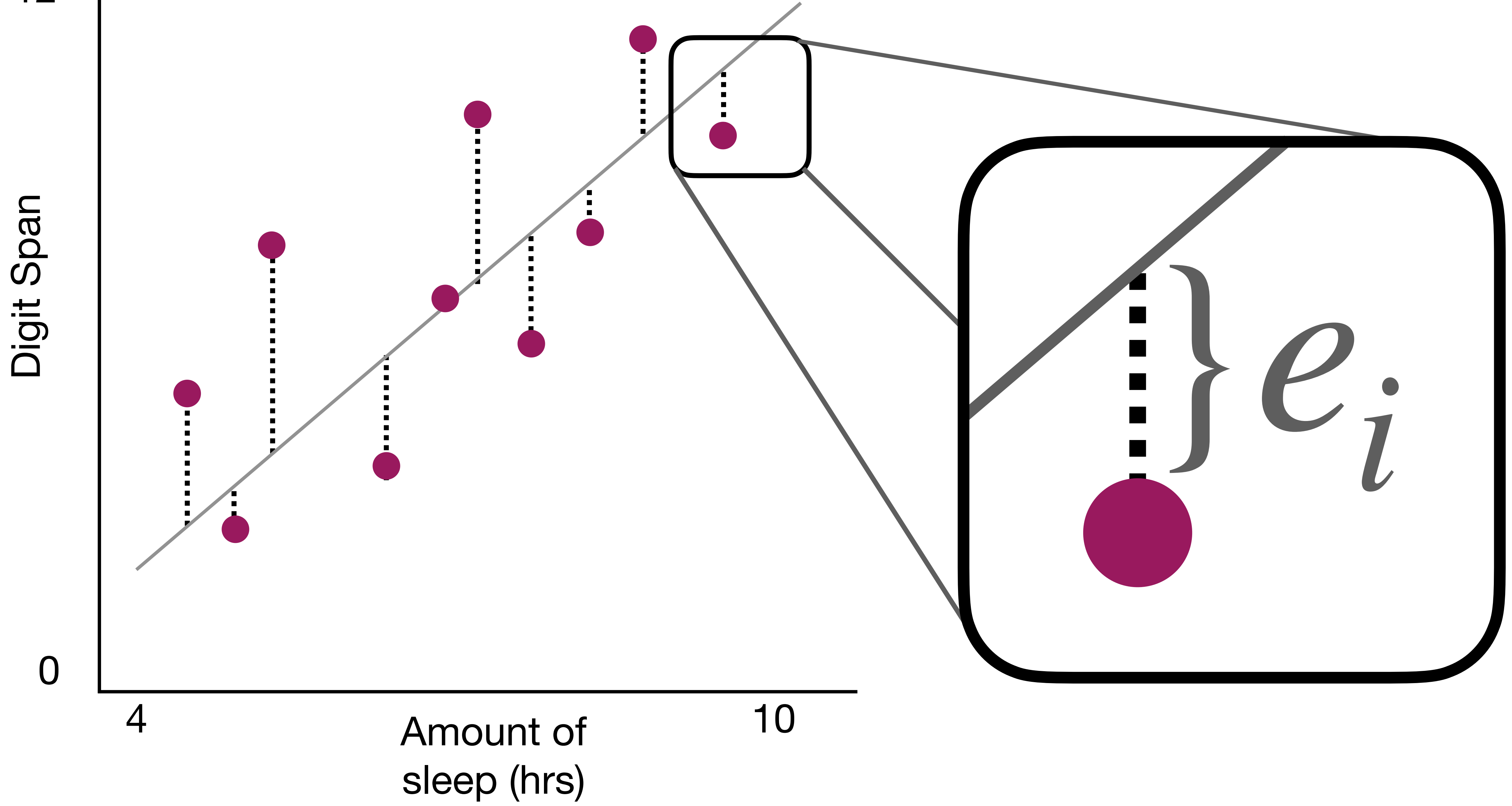
Model: $\hat{y}_i = \hat{\beta}_1 x_i^{hunger} + \hat{\beta}_2 x_i^{sleep} + \hat{\beta}_0$

(James et al. 2013)

Ordinary least squares (OLS)

Error function: $e_i = y_i - \hat{y}_i$

$\hat{y}_i = \hat{\beta}_1 x_i^{sleep} + \hat{\beta}_0$

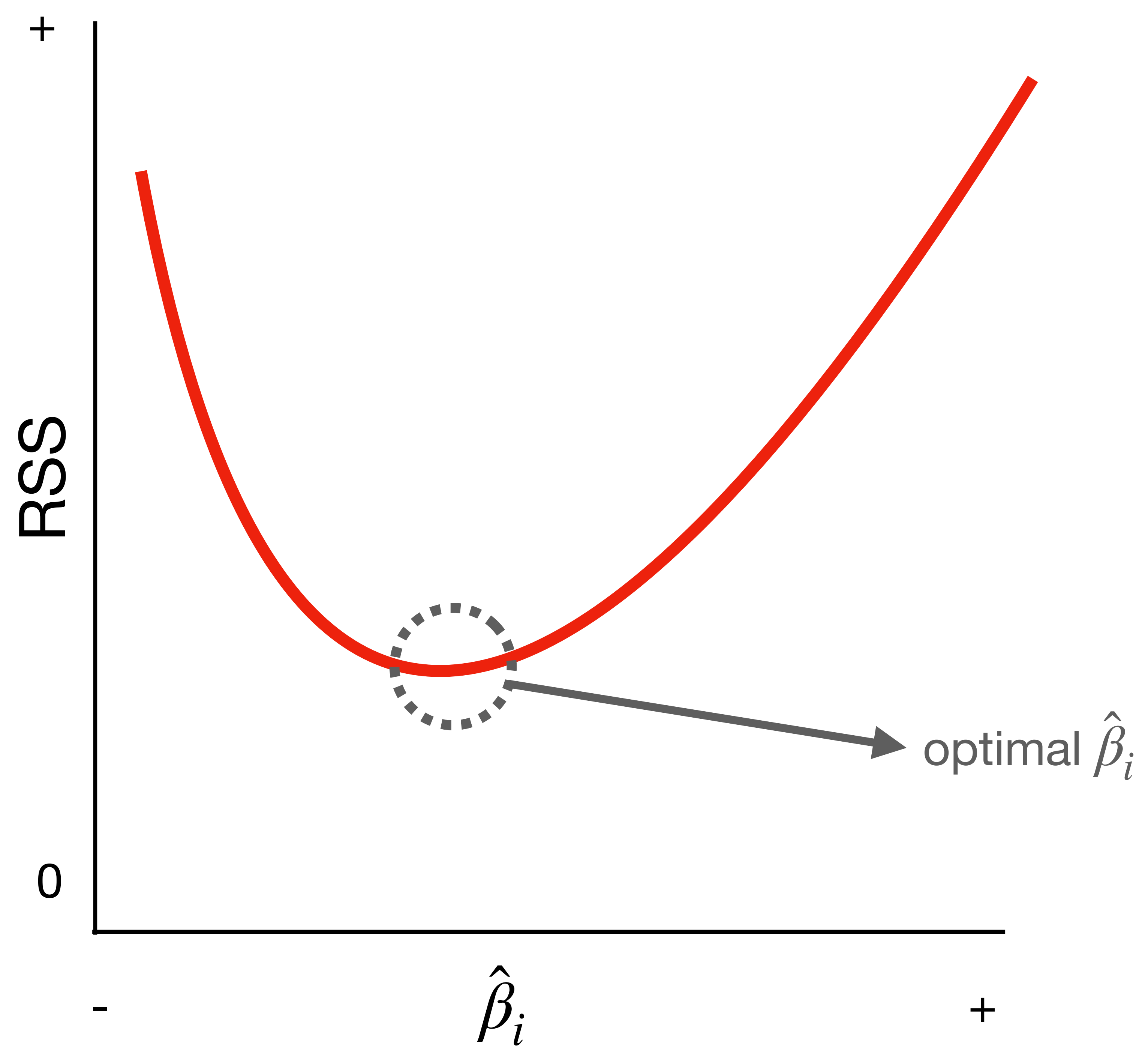


Residual Sums of Squares (RSS):

$$
\begin{aligned}
RSS &= e_1^2 + \ldots + e_n^2 \\
&= (y_1 - \hat{\beta}_1 x_1)^2 \ldots (y_n - \hat{\beta}_1 x_n)^2 \\
&= \sum_{i=1}^{n} (y_i - \hat{\beta}_1 x_i)^2
\end{aligned}
$$

(James et al. 2013)

Convexity & optimal solution



Ordinary Least Squares (OLS):

$$\hat{\beta}_0 = E[Y] - \hat{\beta}_1 E[X]$$
$$= \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

(James et al. 2013)

Confidence of estimates

<u>Types of error:</u> $Y = \underbrace{f(X)}_{\text{reducible}} + \underbrace{\epsilon}_{\text{irreducible}}$

1. Residual square error: $RSE = \sigma^2_{model} = \sqrt{\frac{RSS}{n-2}}$

2. Standard error of estimate: $SE(\hat{\beta}_i) = \sigma^2_{model}(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2})$

$$= \frac{\sigma^2_{model}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

(James et al. 2013)

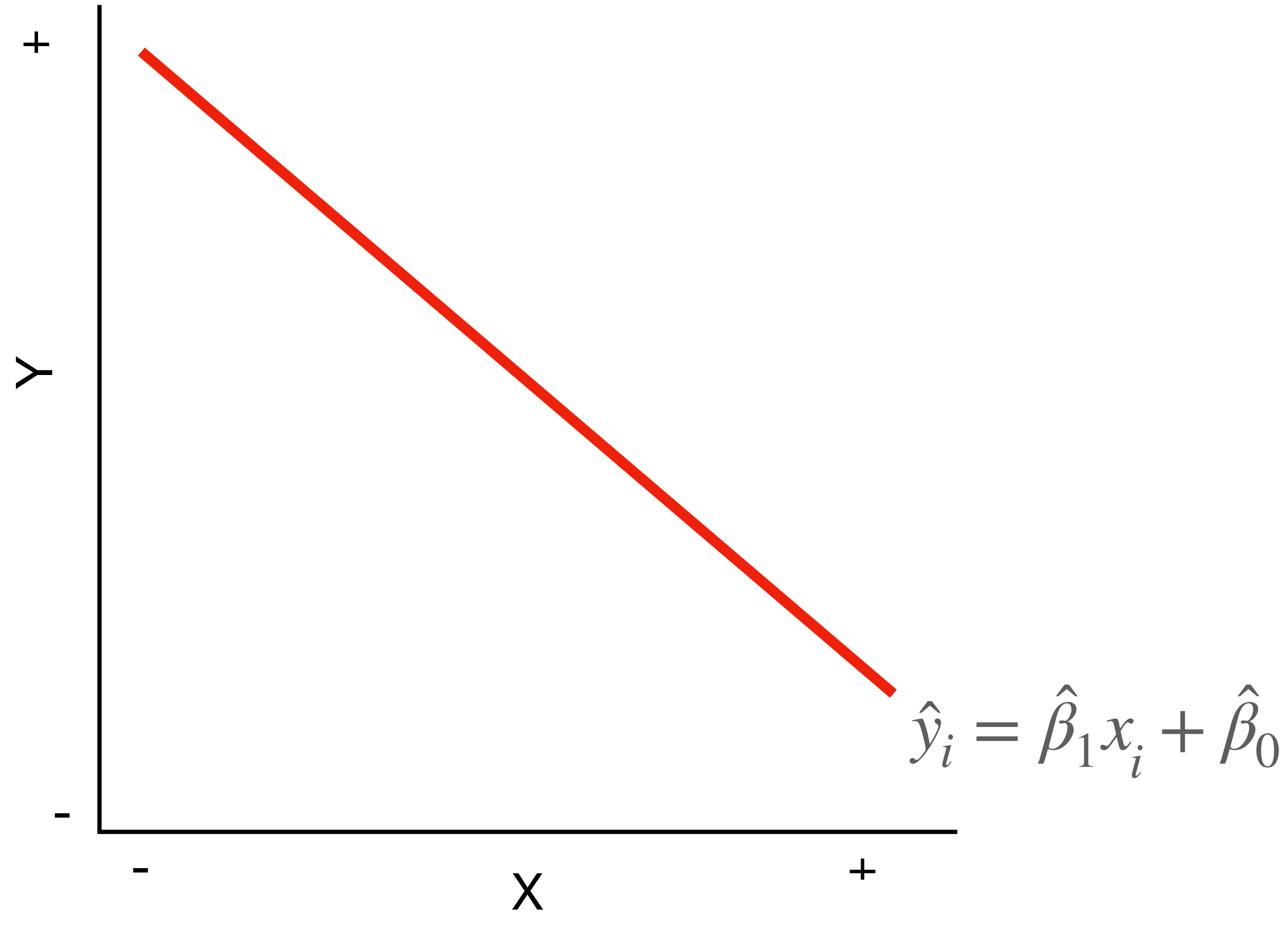
Polynomial models

OLS solution works for all linear terms

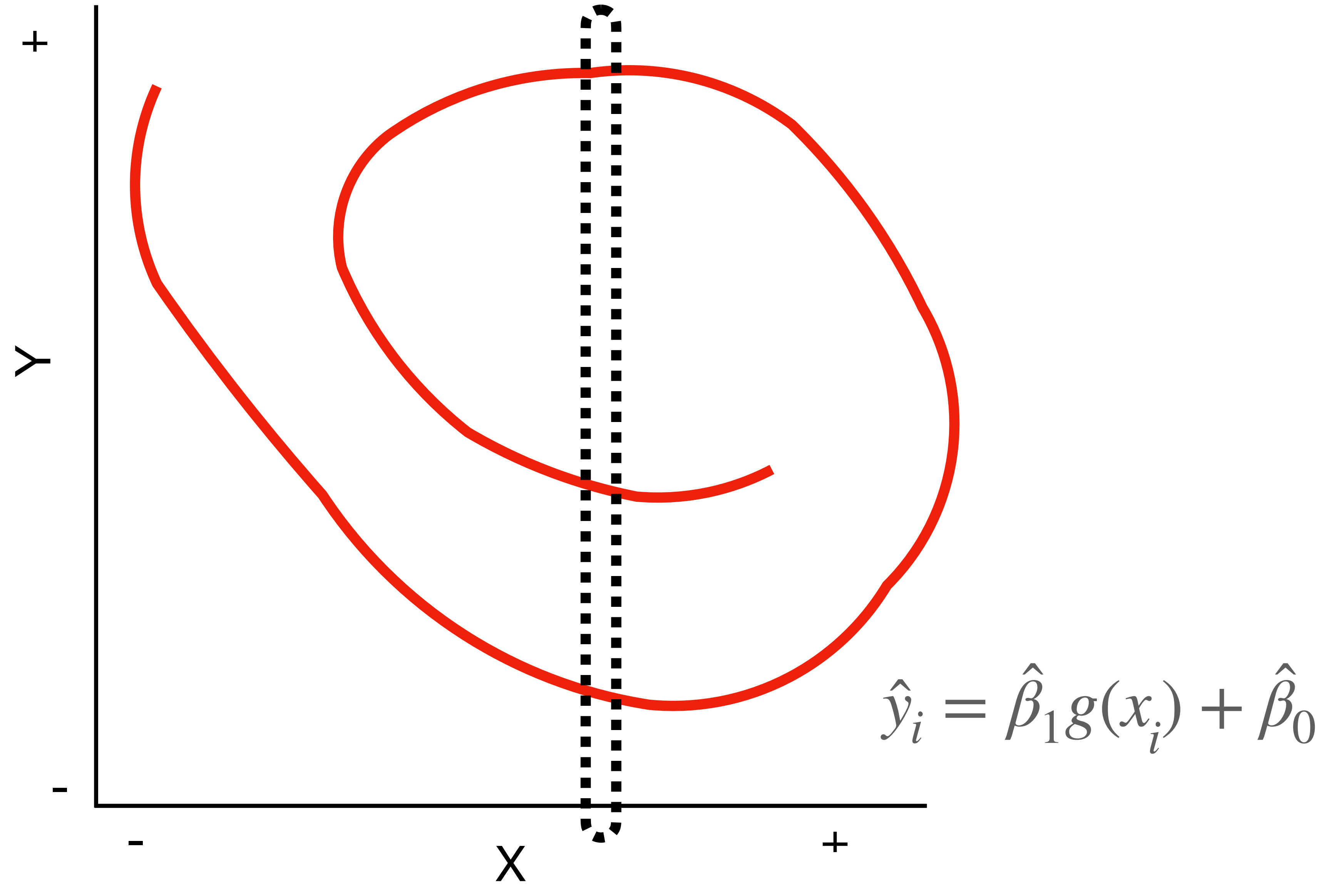
Linear system:

- Successive effects are *additive.*
- For every unique x_i there exists *only one* possible y_i (stationary)

Linear



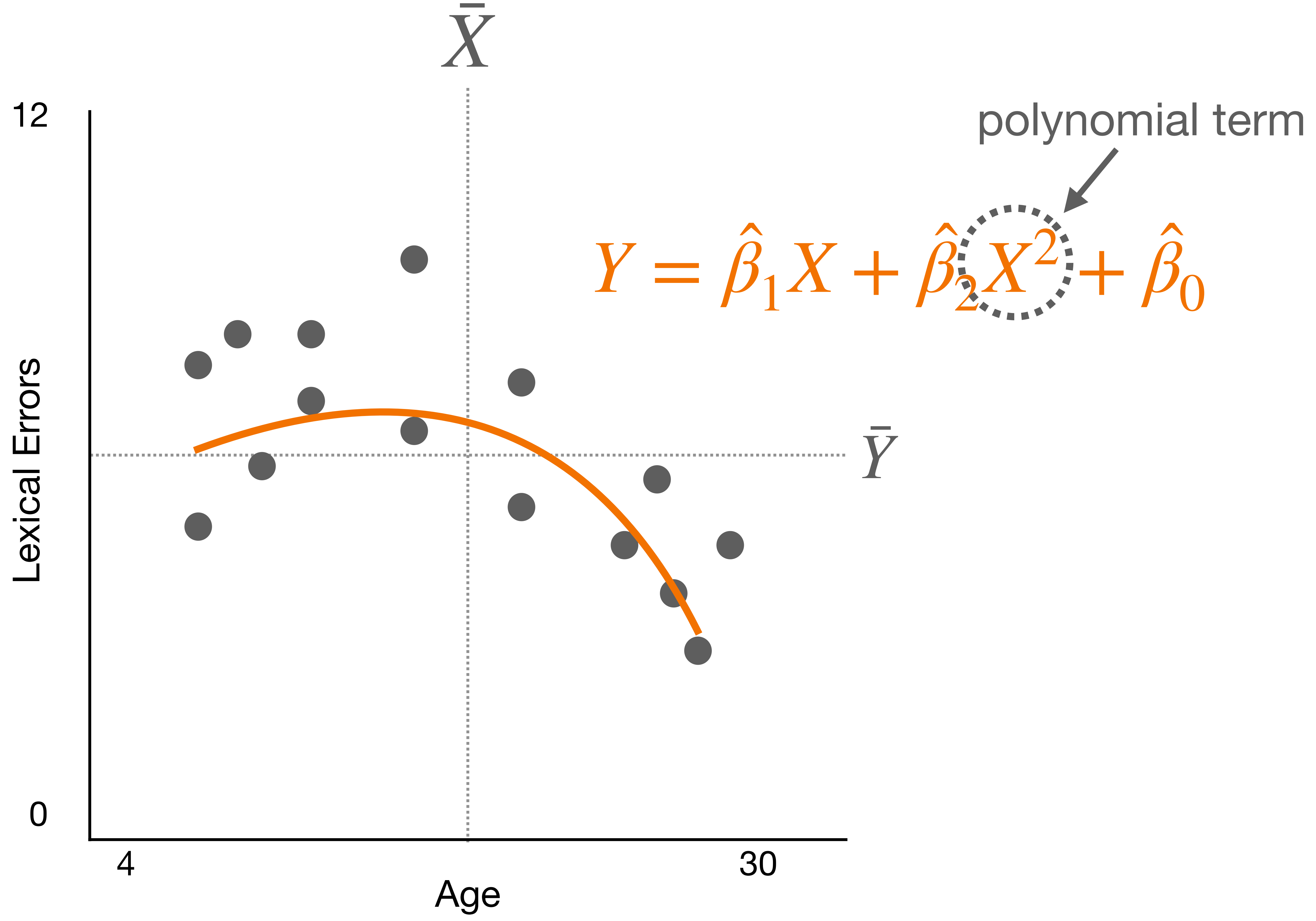
Non-linear

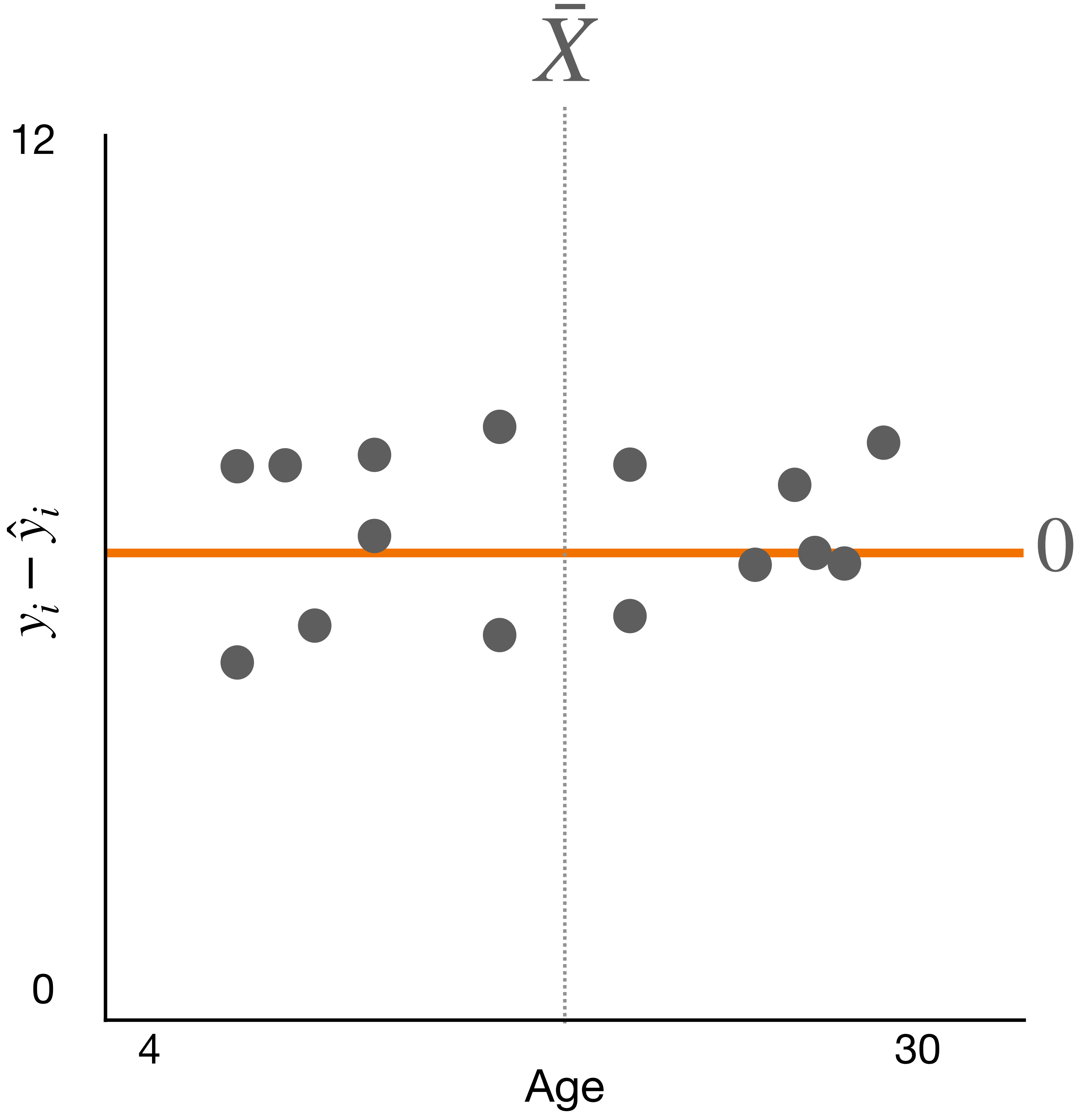


(James et al. 2013)

Polynomial regression

Example





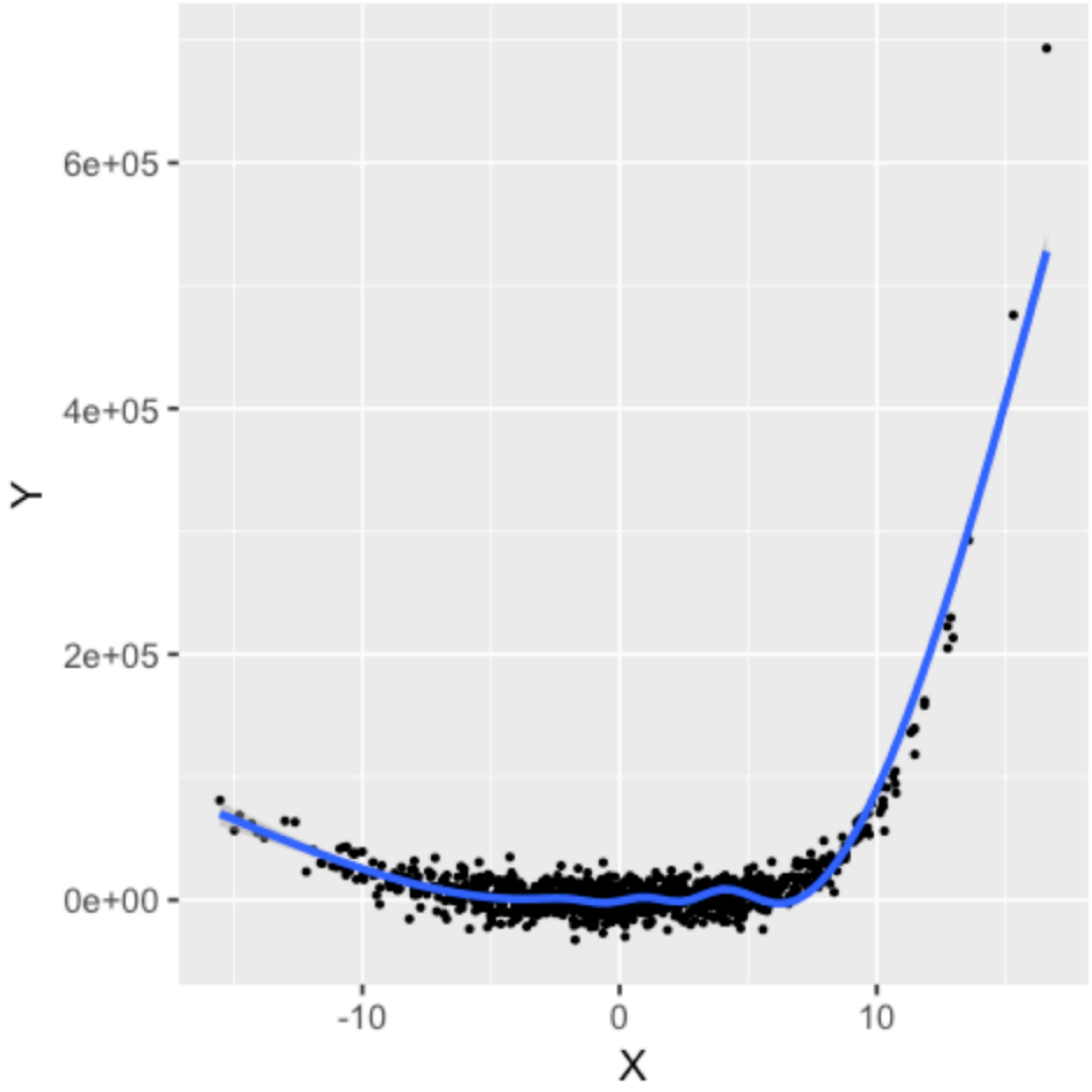
(James et al. 2013)

Kth degree polynomial models

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \beta_5 X^5$$

- Expand x out to the power k.
- Include *all* terms up to k.

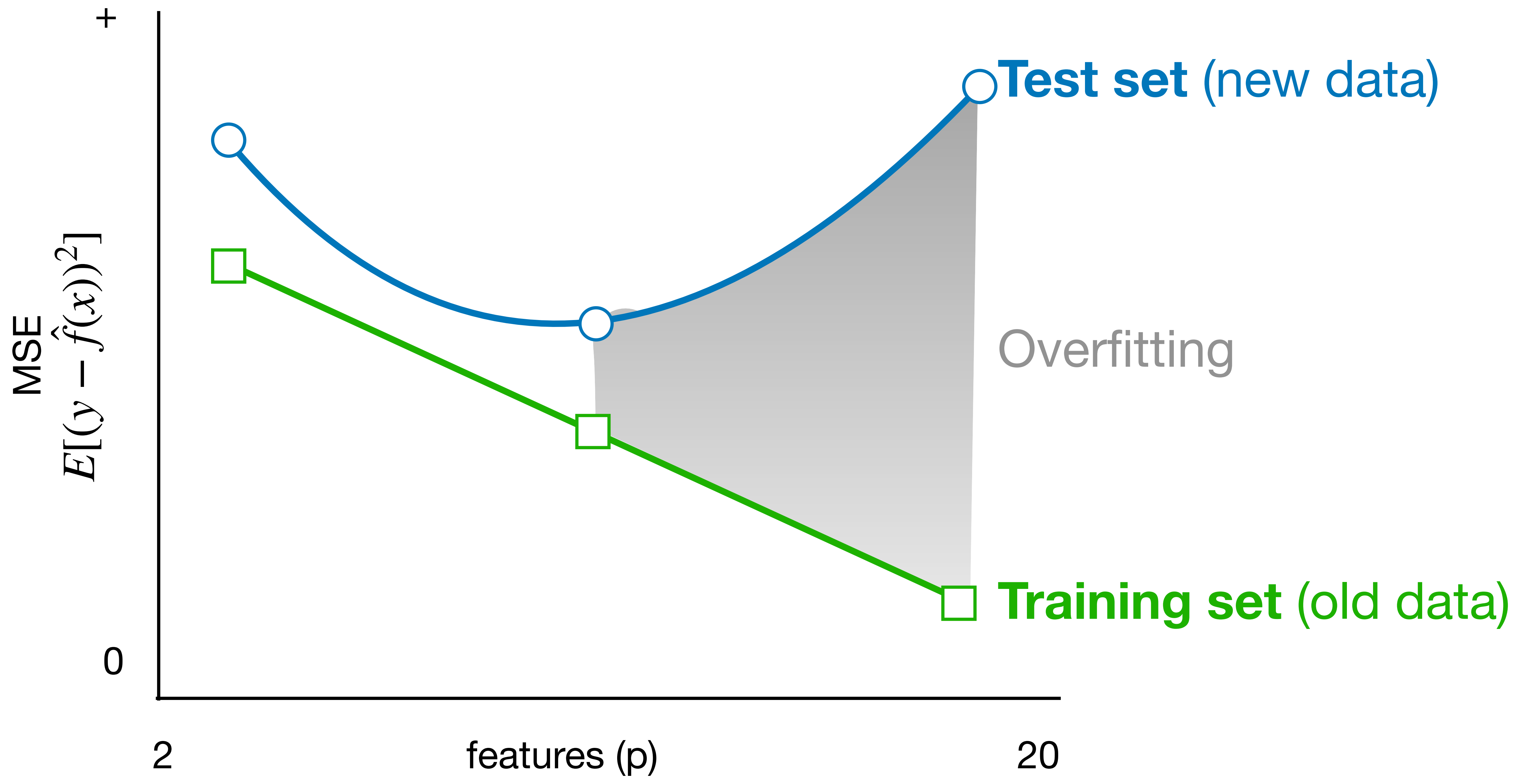
$$Y = \hat{\beta}_0 + \sum_{j=1}^{K} \hat{\beta}_j X^j$$
$$= poly(x, k)$$



(James et al. 2013)

Complexity

- p is the number of features (i.e., predictor variables) in X.
- $n \times p$ defines the flexibility/variance of a model.
- $k = p$ for k[th] order polynomial models.

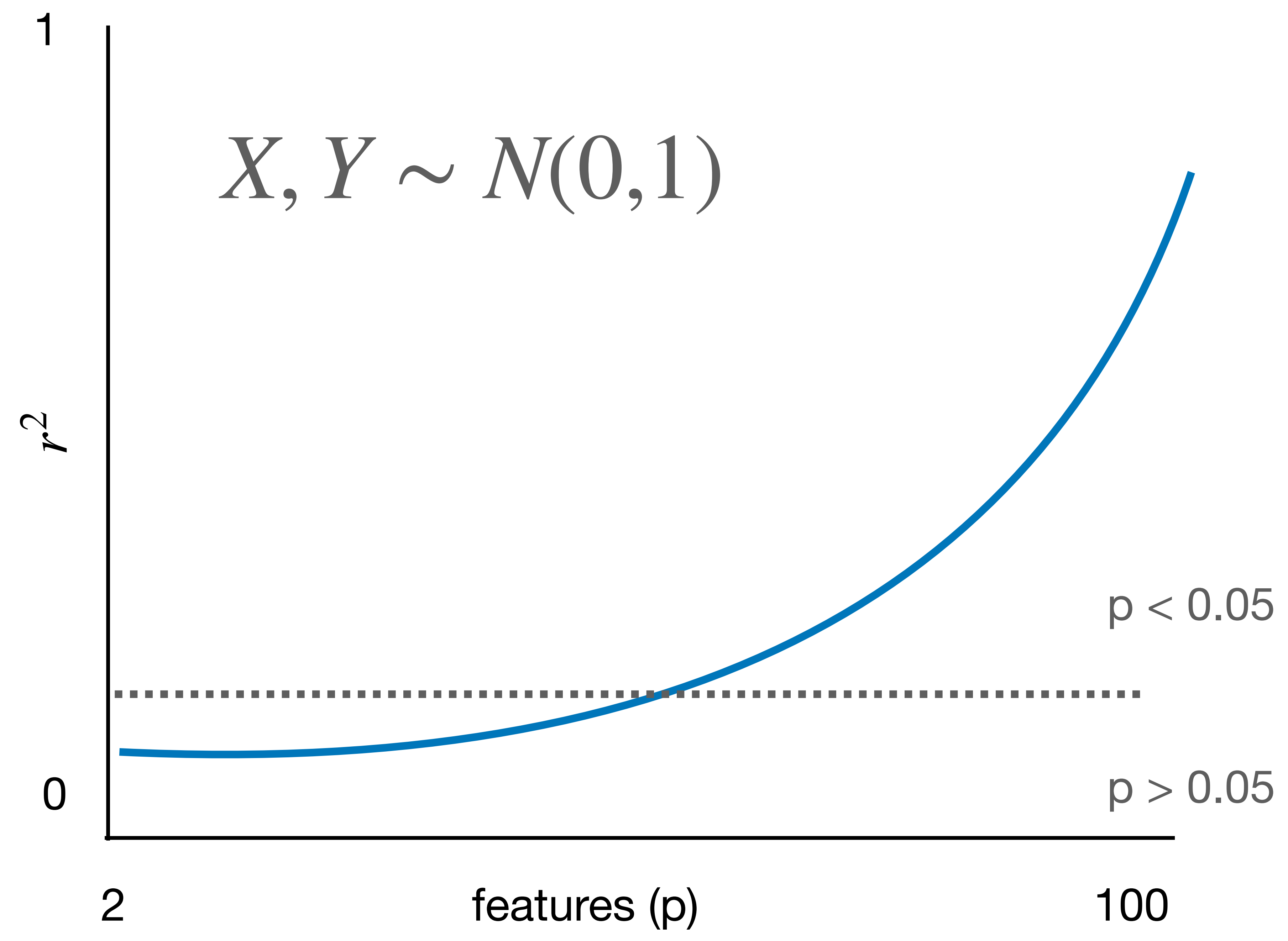


(James et al. 2013)

Risk to inference

- Overfitting means that increasing p risks explaining a "statistically significant" portion of variance by chance.

$$Y = \hat{\beta}_0 + \sum_{j=1}^{K} \hat{\beta}_j X^j$$
$$= poly(x, k)$$



(James et al. 2013)

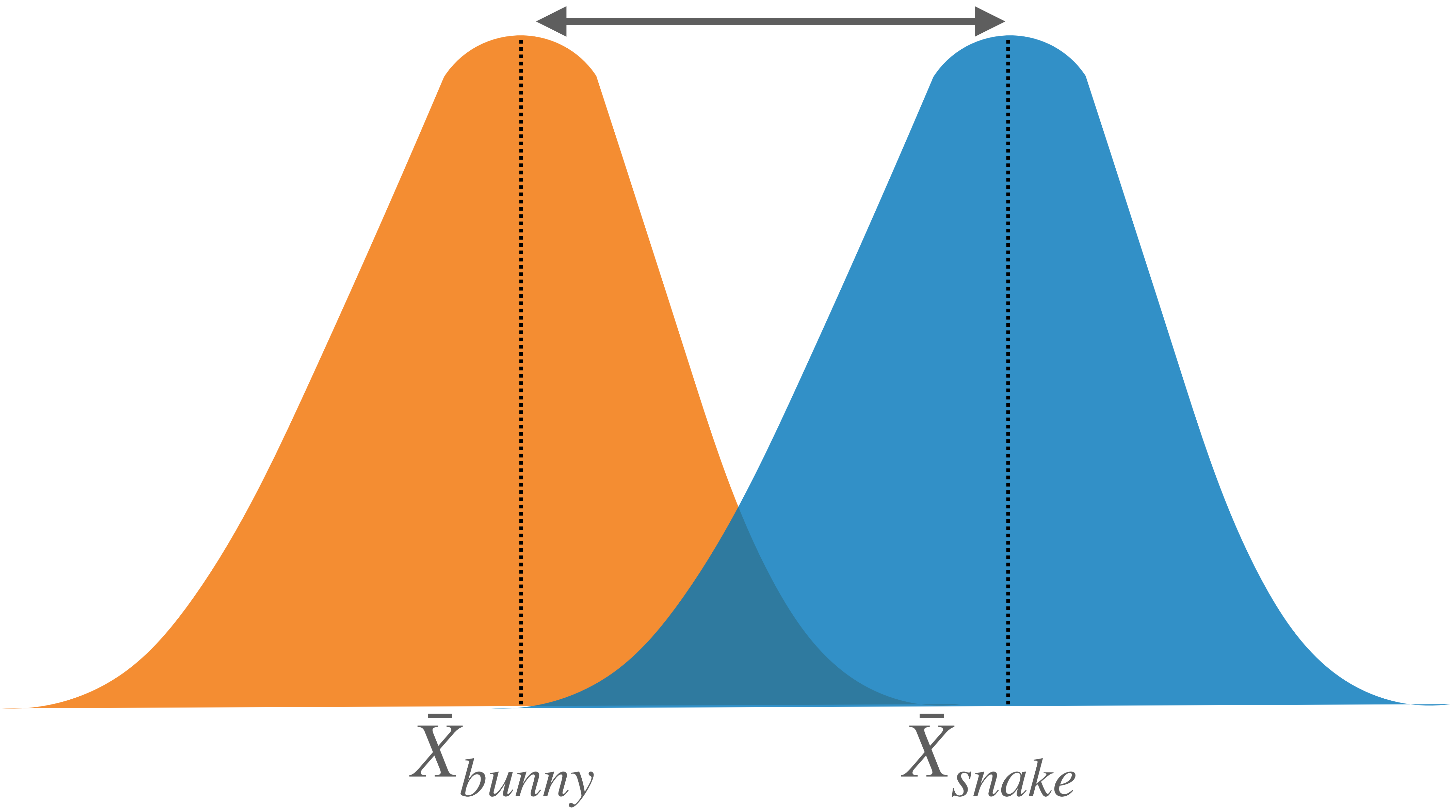
Isomorphisms with other statistical tests

The t-test

Q: Do infants attend more to pictures of snakes or bunnies?

Trial	Look time (ms)	Snake picture	Bunny picture
1	1600	1	0
2	120	0	1
3	874	1	0
4	333	0	1
5	740	1	0
6	201	0	1

$$t = \frac{\bar{X}_{snake} - \bar{X}_{bunny}}{\sigma_{snake,bunny}}$$



(James et al. 2013)

The t-test

Q: Do infants attend more to pictures of snakes or bunnies?

Trial	Look time (ms)	Snake picture	Bunny picture
1	1600	1	0
2	120	0	1
3	874	1	0
4	333	0	1
5	740	1	0
6	201	0	1

<u>Regression form:</u>

$$Y_{time} = \hat{\beta}_0 + \hat{\beta}_1 X_{snake} + \hat{\beta}_2 X_{bunny}$$

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{SE(\hat{\beta}_1 - \hat{\beta}_2)}$$

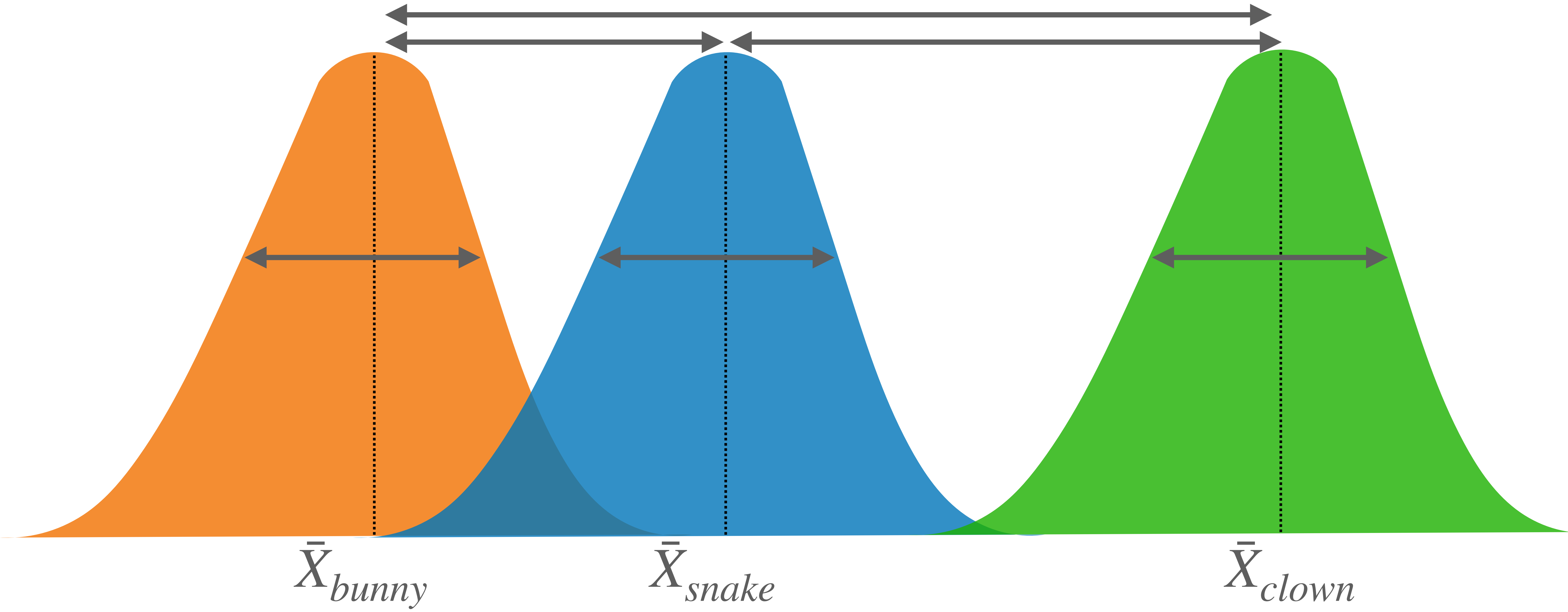
(James et al. 2013)

ANOVA

Q: Do infants attend more to pictures of snakes, bunnies, or clowns?

Trial	Look time	Snake picture	Bunny picture	Clown picture
1	1600	1	0	0
2	120	0	1	0
3	3010	0	0	1
4	333	0	1	0
5	740	1	0	0
6	2237	0	0	1

$$F = \frac{\sigma^2_{between}}{\sigma^2_{within}} = \frac{\dfrac{\sum_{i=1}^{p} n_i (y_i - \bar{y}_i)^2}{p-1}}{\dfrac{\sum_{i=1}^{p} \sum_{j=1}^{n_i} (y_{i,j} - \bar{y}_i)^2}{n-p}}$$



(James et al. 2013)

ANOVA

Q: Do infants attend more to pictures of snakes, bunnies, or clowns?

Trial	Look time	Snake picture	Bunny picture	Clown picture
1	1600	1	0	0
2	120	0	1	0
3	3010	0	0	1
4	333	0	1	0
5	740	1	0	0
6	2237	0	0	1

<u>Regression form:</u>

$$Y_{time} = \hat{\beta}_0 + \hat{\beta}_1 X_{snake} + \hat{\beta}_2 X_{bunny} + \hat{\beta}_3 X_{clown}$$

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$F = \frac{\frac{1}{p}(TSS - RSS)}{\frac{RSS}{n-p-1}}$$

(James et al. 2013)

Take home message

- Linear regression models are extensible enough to test many questions typically evaluated using more narrow statistical methods.