The ordinary least squares solution

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. Maximum likelihood estimation

2. Model evaluation

Maximum likelihood estimation

Structure of a linear model

Fundamental form:

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

Solution:

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

$$= \bar{y} - \sum_{j=1}^{p} \hat{\beta}_j \bar{x}_p$$

$$\hat{\beta}_j = \frac{\sum_{i=1}^{n} (x_{i,j} - \bar{x}_j)(y_i - \bar{y})}{\sum_{i=1}^{n} (x_{i,j} - \bar{x}_j)^2}$$

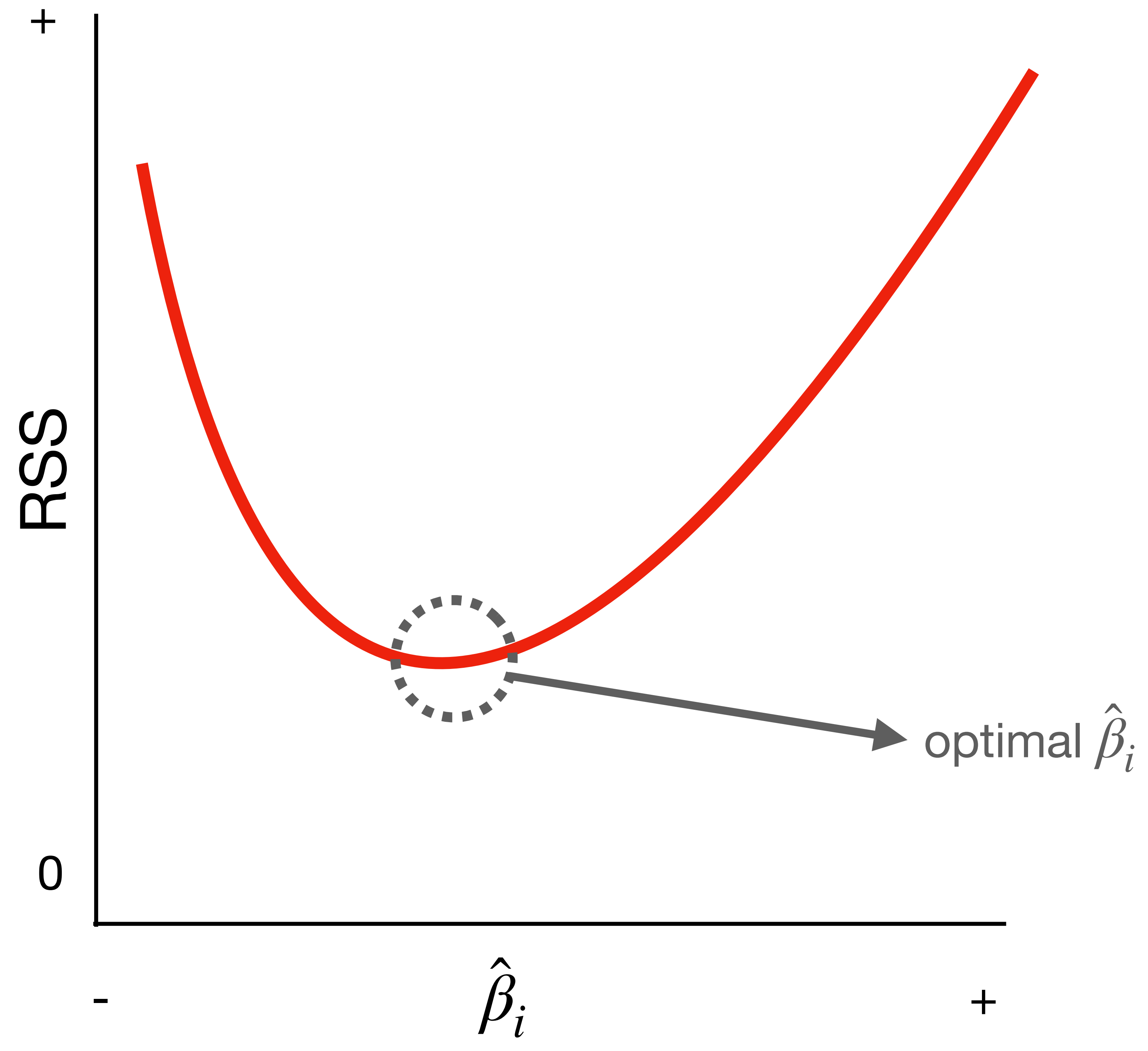
Assumptions:

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.

(James et al. 2013)

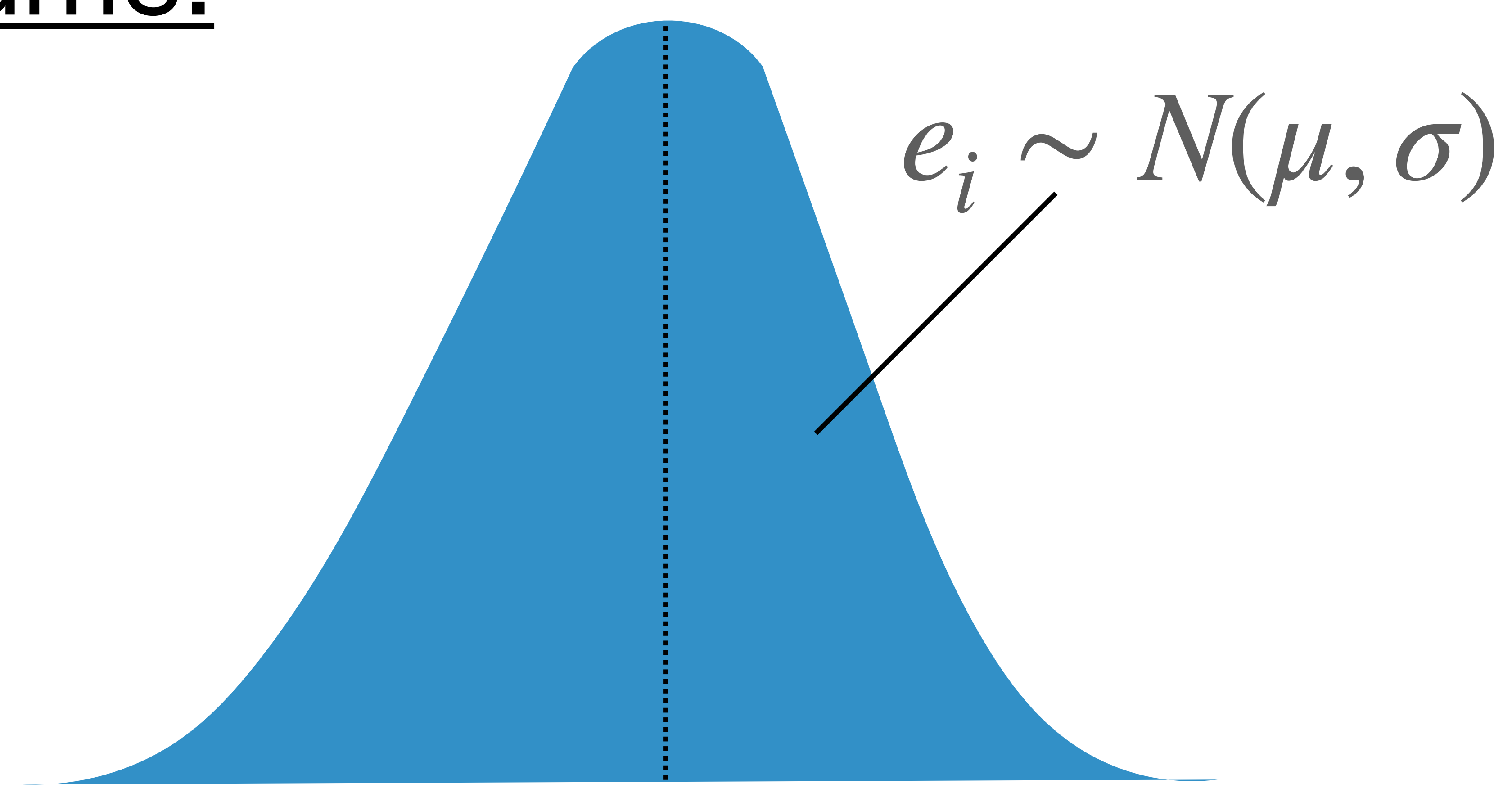
Residuals



residuals: $(e_1^2, \ldots, e_n^2)$

$: ((y_1 - \hat{f}(x_1))^2, \ldots, (y_n - \hat{f}(x_n))^2)$

<u>Assume:</u>



(James et al. 2013)

Likelihood function

The probability that the data you observe arises from a specific probability distribution with a specific set of parameters.

$$f(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Gaussian likelihood:

(James et al. 2013)

Returning to residuals

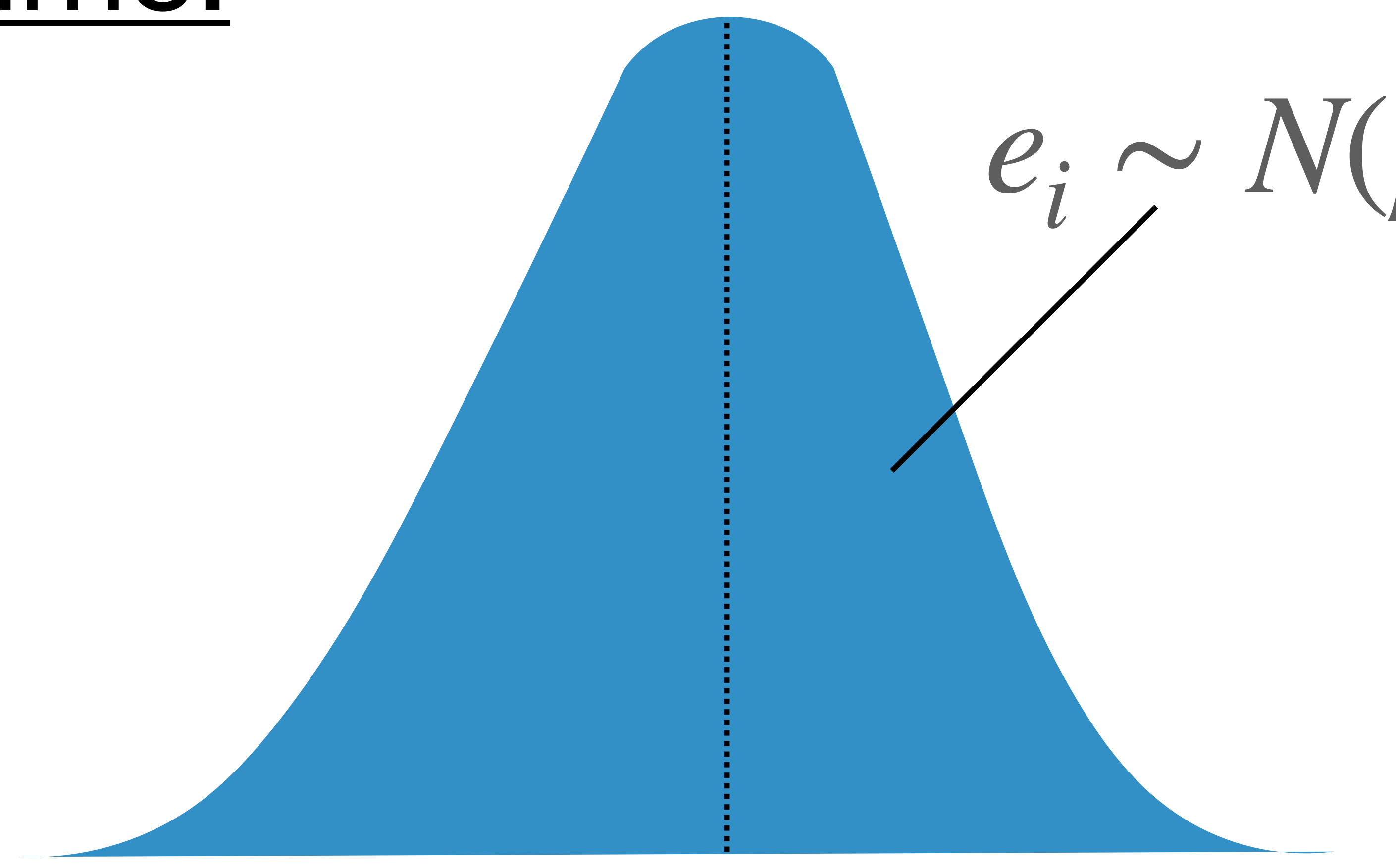
<u>Likelihood of residuals:</u>

$$L(\hat{\beta}_0, \ldots, \hat{\beta}_p, \sigma) = \prod_{i=1}^{n} p(y_i \,|\, x_i; \hat{\beta}_0, \ldots, \hat{\beta}_p, \sigma)$$

$$= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i - \sum_{j=0}^{p} \beta_j x_i)^2}{2\sigma^2}}$$

product

<u>Assume:</u>



$e_i \sim N(\mu, \sigma)$

If Y is normally distributed, then the residuals of the model $\hat{f}(X)$ fit to Y should also be normally distributed.

(James et al. 2013)

Log likelihood of residuals

Likelihood of residuals:

$$L(\hat{\beta}_0, \ldots, \hat{\beta}_p, \sigma) = \prod_{i=1}^{n} p(y_i \mid x_i; \hat{\beta}_0, \ldots, \hat{\beta}_p, \sigma)$$

$$= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i - \sum_{j=0}^{p} \beta_j x_i)^2}{2\sigma^2}}$$

Log likelihood of residuals:

$$ln(L(\hat{\beta}_0, \ldots, \hat{\beta}_p, \sigma)) = ln(\prod_{i=1}^{n} p(y_i \mid x_i; \hat{\beta}_0, \ldots, \hat{\beta}_p, \sigma))$$

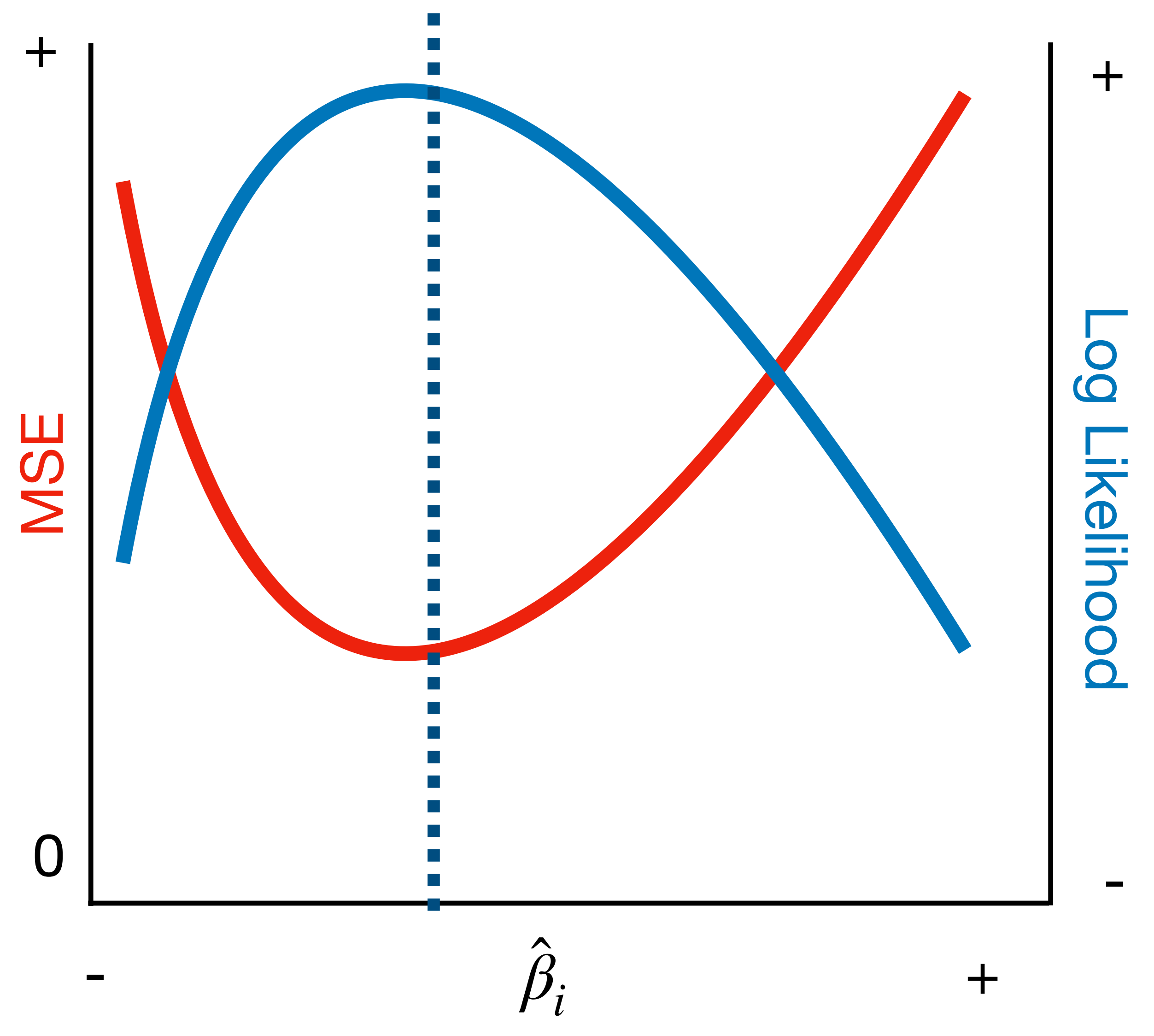
$$= \sum_{i=1}^{n} ln(p(y_i \mid x_i; \hat{\beta}_0, \ldots, \hat{\beta}_p, \sigma)$$

$$= -\frac{n}{2} ln(2\pi) - n\ ln(\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \sum_{j=0}^{p} \hat{\beta}_j x_{i,j})^2$$

(James et al. 2013)

Maximum likelihood estimation

Maximizing the likelihood of your data given a particular model means finding the parameters that minimize your error (cost) function.



<u>Mean Squared Error (MSE):</u>

$$MSE(\hat{\beta}_0, \hat{\beta}_1) = E[(Y - (\hat{\beta}_0 + \hat{\beta}_1 X))^2]$$

$$= E[Y^2] - 2\hat{\beta}_0 E[Y] - 2\hat{\beta}_1 E[XY] + E[(\hat{\beta}_0 + \hat{\beta}_1 X)^2]$$

$$= E[Y^2] - 2\hat{\beta}_0 E[Y] - 2\hat{\beta}_1 (Cov[XY] + E[X]E[Y])$$
$$+ \hat{\beta}_0^2 + 2\hat{\beta}_0 \hat{\beta}_1 E[X] + \hat{\beta}_1^2 E[X^2]$$

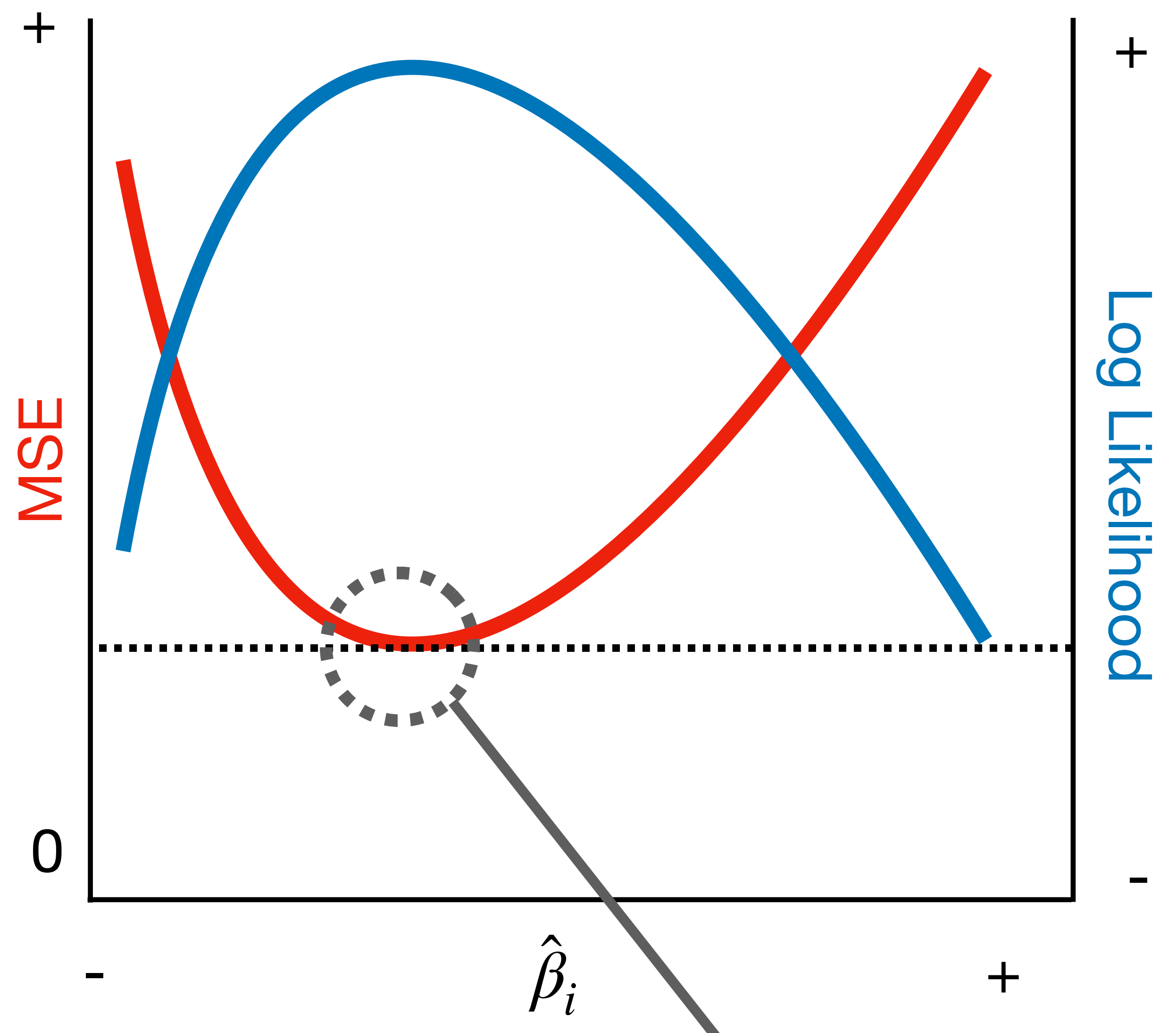
$$= E[Y^2] - 2\hat{\beta}_0 E[Y] - 2\hat{\beta}_1 Cov[XY] - 2\hat{\beta}_1 E[X]E[Y] + \hat{\beta}_0^2$$
$$+ 2\hat{\beta}_0 \hat{\beta}_1 E[X] + \hat{\beta}_1^2 Var[X] + \hat{\beta}_1^2 (E[X])^2$$

<u>Task:</u> Solve for $\hat{\beta}_0$ and any $\hat{\beta}_i$ where $i > 0$.

(James et al. 2013)

Maximum likelihood estimation

$$MSE(\hat{\beta}_0, \hat{\beta}_1) = E[Y^2] - 2\hat{\beta}_0 E[Y] - 2\hat{\beta}_1 Cov[XY] - 2\hat{\beta}_1 E[X]E[Y]$$
$$+\hat{\beta}_0^2 + 2\hat{\beta}_0\hat{\beta}_1 E[X] + \hat{\beta}_1^2 Var[X] + \hat{\beta}_1^2 (E[X])^2$$

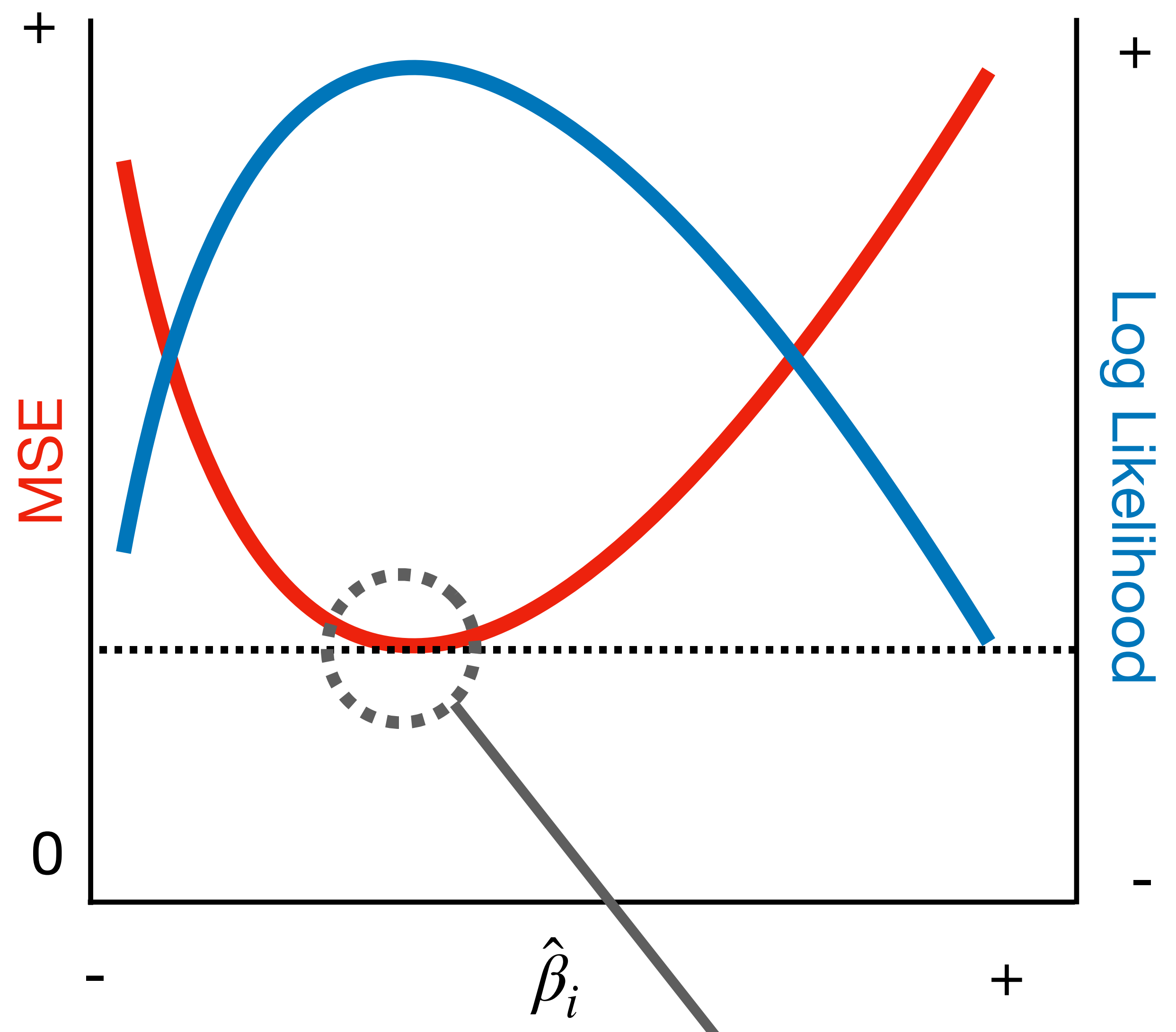


$$\frac{\delta RSS}{\beta_j} = 0$$

$$\hat{\beta}_0: \quad \frac{\partial E[(Y - (\hat{\beta}_0 + \hat{\beta}_1 X))^2]}{\partial \hat{\beta}_0} = -2E[Y] + 2\hat{\beta}_0 + 2\hat{\beta}_1 E[X]$$

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

Maximum likelihood estimation



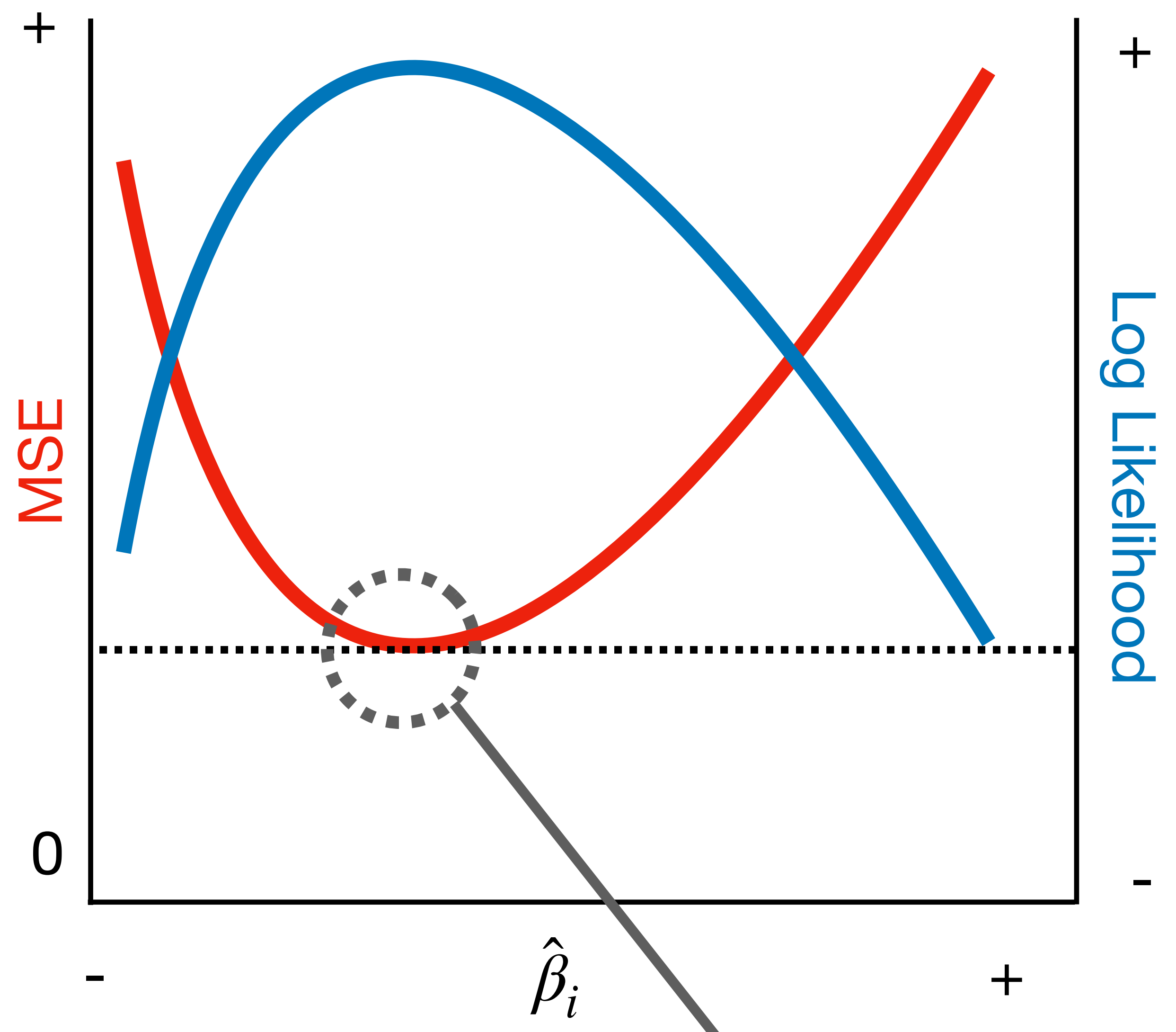
$$\frac{\delta RSS}{\beta_j} = 0$$

$$MSE(\hat{\beta}_0, \hat{\beta}_1) = E[Y^2] - 2\hat{\beta}_0 E[Y] - 2\hat{\beta}_1 Cov[XY] - 2\hat{\beta}_1 E[X]E[Y]$$
$$+\hat{\beta}_0^2 + 2\hat{\beta}_0\hat{\beta}_1 E[X] + \hat{\beta}_1^2 Var[X] + \hat{\beta}_1^2 (E[X])^2$$

$$\hat{\beta}_1: \quad \frac{\partial E[(Y - (\hat{\beta}_0 + \hat{\beta}_1 X))^2]}{\partial \hat{\beta}_1} = -2Cov[XY] - 2E[X]E[Y]$$
$$+2\hat{\beta}_0 E[X] + 2\hat{\beta}_1 Var[X] + 2\hat{\beta}_1 (E[X])^2$$

$$\hat{\beta}_1 = \frac{Cov[X,Y]}{Var[X]} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Maximum likelihood estimation



$$\frac{\delta RSS}{\beta_j} = 0$$

$$MSE(\hat{\beta}_0, \hat{\beta}_1) = E[Y^2] - 2\hat{\beta}_0 E[Y] - 2\hat{\beta}_1 Cov[XY] - 2\hat{\beta}_1 E[X]E[Y] \\ + \hat{\beta}_0^2 + 2\hat{\beta}_0\hat{\beta}_1 E[X] + \hat{\beta}_1^2 Var[X] + \hat{\beta}_1^2 (E[X])^2$$

$$\hat{\beta}_0 = E[Y] - \hat{\beta}_1 E[X] = \bar{y} - \hat{\beta}_1 \bar{x} \\ = \bar{y} - \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \bar{x}$$

Assumptions make sense

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.

Advantages for MLE:

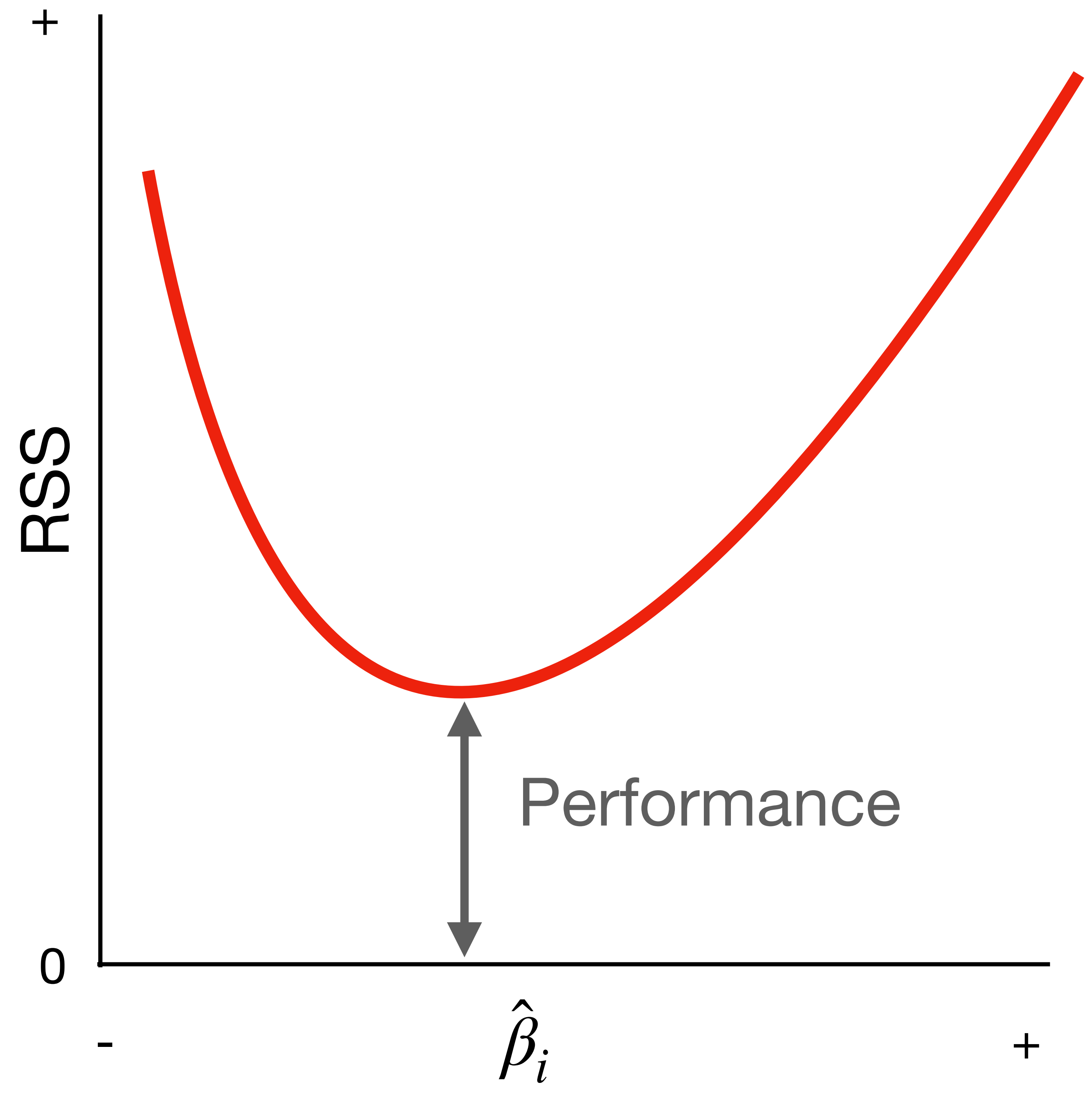
Well defined probability distribution & convex error gradient.

Allows for solving each $\hat{\beta}_i$ separately.

Only one function to solve for.

Model evaluation

Best you can do?



Ideal model: One that accounts for all of the variance in your data.

$$
\begin{aligned}
RSS &= e_1^2 + \ldots + e_n^2 \\
&= \sum_{i=1}^{n} (y_i - \hat{\beta}_1 x_i)^2 \\
&= 0
\end{aligned}
$$

(James et al. 2013)

Residual square error (RSE)

$$RSE = \sigma^2_{model} = \sqrt{\frac{RSS}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2}}$$

- Bounded at 0
- Simplest to calculate
- ↓ RSE = ↑ model fit
- Closes evaluation of $\hat{y}$ itself

(James et al. 2013)

Coefficient of determination (r^2)

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

- Direct measure of percent variance in Y explained by $\hat{f}$
- $\uparrow r^2 = \uparrow$ model fit
- Meaningful units ($0\,\% \rightarrow 100\,\%$)

(James et al. 2013)

F-test (F)

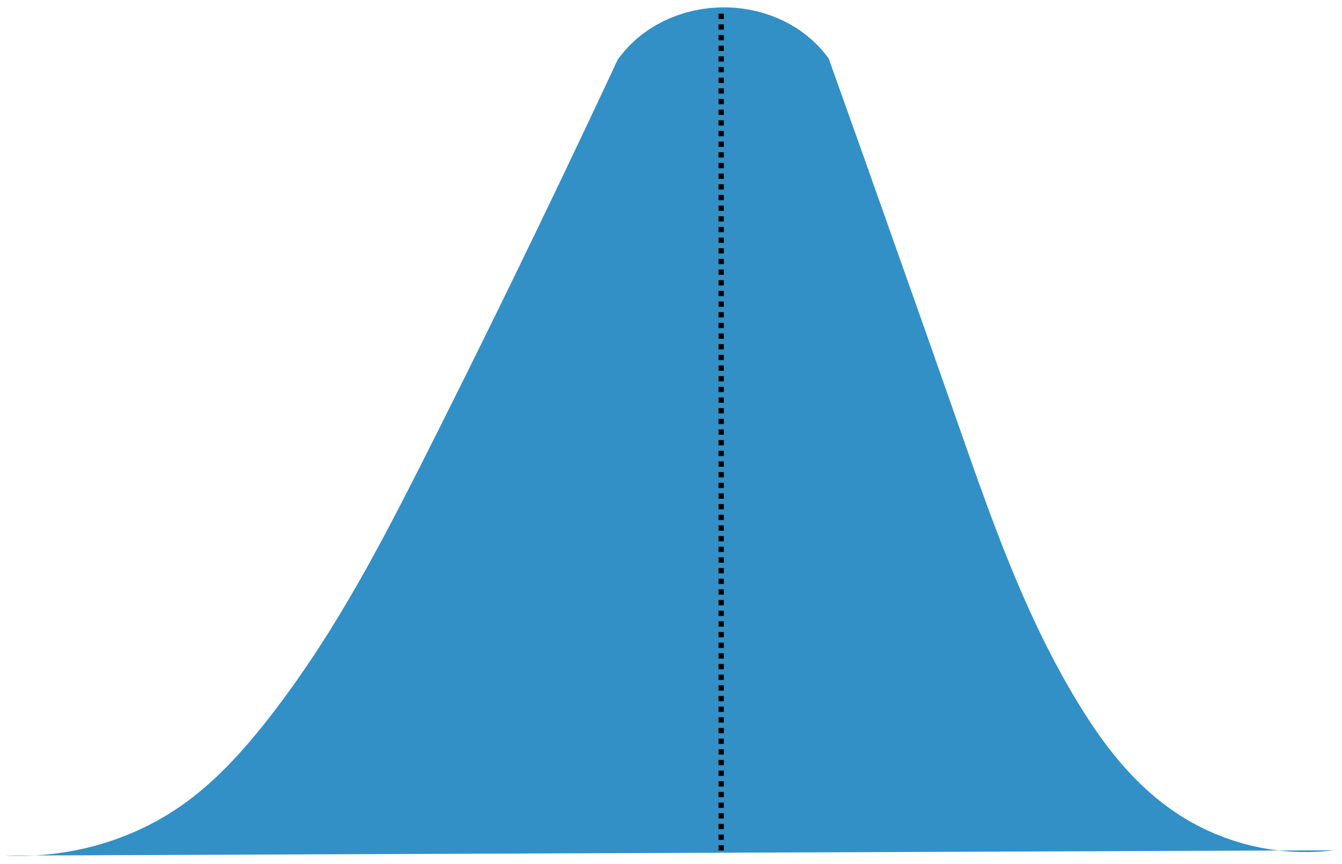
$$F = \frac{\frac{1}{p}(TSS - RSS)}{\frac{1}{n-p-1}RSS} = \frac{\frac{1}{p}(\sum_{i=1}^{n}(y_i - \bar{y})^2 - \sum_{i=1}^{n}(y_i - \hat{y})^2}{\frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \hat{y})^2}$$

- Bounded at 0
- $\uparrow F = \uparrow$ model fit
- Ratio of variances (variance explained by $\hat{f}$ over variance of y)

(James et al. 2013)

Common theme

All model evaluation measures how well the prediction, $\hat{y}$, explains y against compared to the mean, $\bar{y}$.



$\bar{y}$

Best prediction of y with no other information available.

$$\underbrace{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}_{\text{RSS}} \ll \underbrace{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2}_{\text{TSS}}$$

$$\sim \sigma_y^2$$

(James et al. 2013)

Take home message

- Ordinary least squares regression provides a simple, closed form solution to the model fitting problem… so long as the data meet the assumptions.