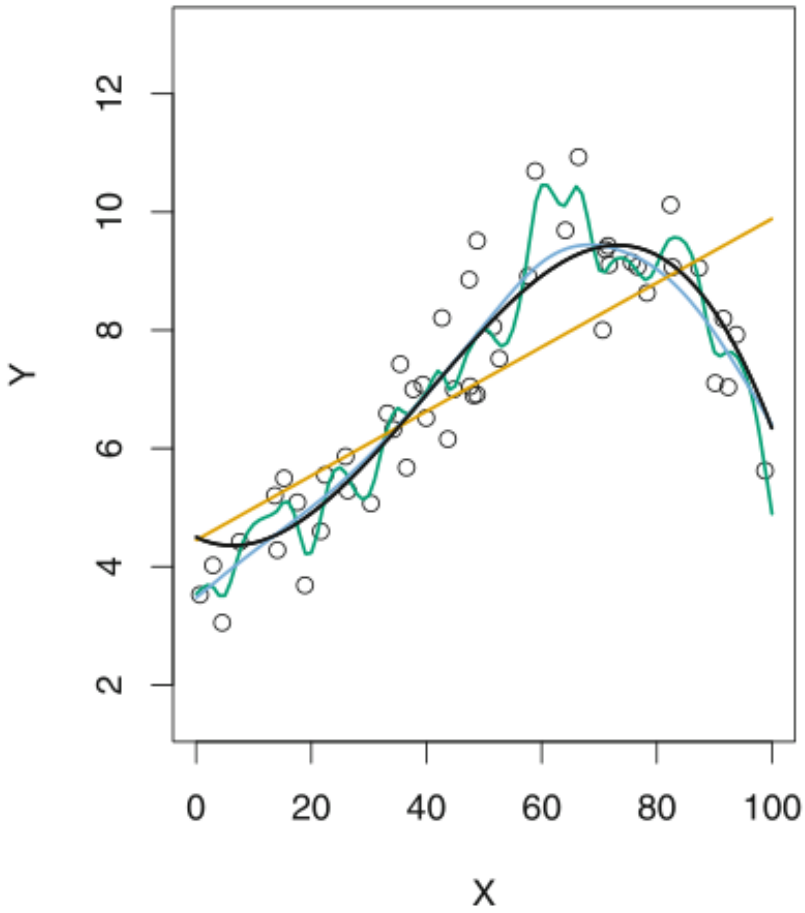
# Prediction and Classification

## Statistical Learning

Dr. Markus Heinrich

# Sample vs Population and True vs Estimated Predictor Function




- What is the optimal prediction given a loss function?
- How is the prediction error decomposed?
- How can I minimize the prediction error?

# Today's outline

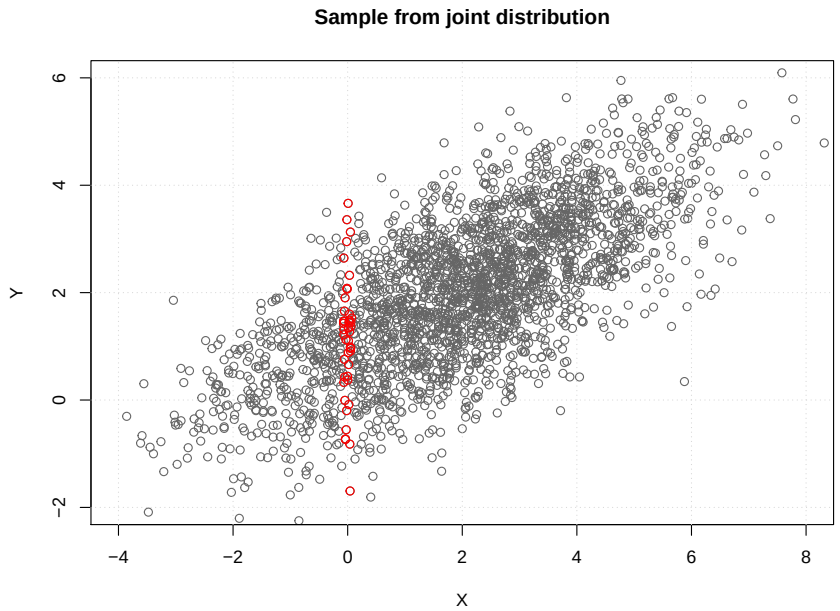
## Prediction and Classification

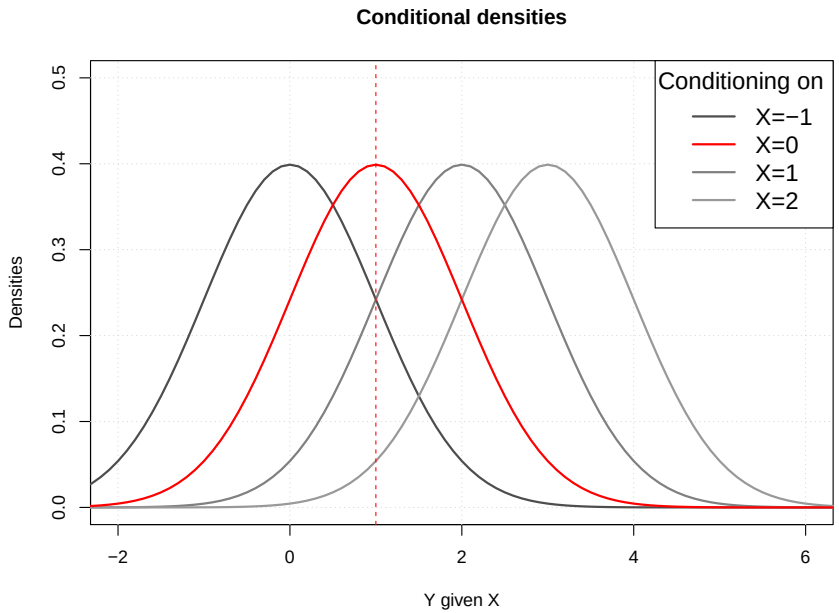
1. Prediction errors and quantification
2. Classification errors
3. Learning for prediction
4. Up next

# Outline

# There will be errors

Conditionally on $\boldsymbol{X}$, $Y$ has a distribution and its outcomes (continuous or discrete) are random.




At the same time, $\hat{Y}$ is a deterministic function of $\boldsymbol{X}$ and therefore prediction (or classification) errors will appear by construction.

To minimize the (impact of the) errors, we need to quantify them.

# Let us discuss prediction first

For **continuous** variables,

- The probability of $Y = \hat{Y}$ is zero.
- A good prediction lies as close as possible to $Y$ **on average**.
- Seek a *good* **predictor function** $f(\boldsymbol{X}) = \hat{Y}$ for predicting Y given values of the input X.

But what is good / close?

We shall measure this with the help of a loss function $\mathcal{L}(Y, f(\boldsymbol{X}), \ldots)$ for penalizing errors $u = (Y - \hat{Y})$ in prediction.

# Loss function

**Definition**

*A loss function $\mathcal{L}(\cdot)$ is a quasi-convex function minimized at $0$.*

In other words, $\mathcal{L}$ is increasing for positive arguments and decreasing for negative arguments.

- The most common loss function is the squared error loss (MSE).
- But asymmetric loss functions may also be encountered, say asymmetric linear or asymmetric quadratic.

Typically, we pick the (optimal) prediction that minimizes the expected loss!

$$\hat{Y}^* = \arg\min_y \mathrm{E}\left(\mathcal{L}\left(Y, y\right)\right).$$

# What is the optimal prediction?

**General approach:**

1. State your loss (objective) function $\mathcal{L}(Y, \hat{Y})$
2. Minimize expected loss $E\left[\mathcal{L}(Y, \hat{Y})\right]$:
   - First derivative:

$$\frac{d}{d\hat{Y}} E\left[\mathcal{L}(Y, \hat{Y})\right] = \frac{d}{d\hat{Y}} \int_y \mathcal{L}(Y, \hat{Y}) p(y) dy$$

   - Assuming interchangeability of differentiation and integration:

$$\int_y \frac{d\mathcal{L}(Y, \hat{Y})}{d\hat{Y}} p(y) dy = E\left[\mathcal{L}'(Y, \hat{Y})\right]$$

   - First order condition (FOC):

$$E\left[\mathcal{L}'(Y, \hat{Y})\right] = 0$$

# What is the optimal prediction under squared error loss?

**Proposition**

*For the squared error loss, the optimal unconditional prediction is the mean of* $Y$, $\hat{Y} = \mathrm{E}(Y)$.

1. Loss function $\mathcal{L}(Y, \hat{Y}) = \mathcal{L}(Y - \hat{Y}) = (Y - \hat{Y})^2$
2. Minimize expected loss:
   - First derivative:
   $$\mathcal{L}'(Y, \hat{Y}) = -2\left(Y - \hat{Y}\right)$$
   - First order condition (FOC):
   $$E\left[\mathcal{L}'(Y, \hat{Y})\right] = E\left[-2\left(Y - \hat{Y}\right)\right] = 0$$
   - Solution:
   $$\hat{Y}^* = E[Y]$$

# What is the optimal prediction under squared error loss?

This clearly extends to conditional distributions and expectations.

So for prediction we typically focus on **conditional means**, i.e. the regression function and take as predictor function

$$f(\boldsymbol{x}) = \mathrm{E}\left(Y | \boldsymbol{X} = \boldsymbol{x}\right).$$

Of course, for a new data point $\boldsymbol{X}$, we take $\hat{Y} = f(\boldsymbol{X})$.

# Other loss functions

Mean Absolute Deviation is sometimes an alternative to MSE.[1]

**Proposition**

*For the absolute deviation loss, the optimal predictor is the median (which we assume unique).*

Analogously, asymmetric linear losses leads to conditional quantiles.
Let

$$\mathcal{L}(u) = \begin{cases} -au & u < 0 \\ bu & u \geq 0 \end{cases} \quad \text{for } a, b > 0.$$

Then,

$$f(\boldsymbol{x}) = F^{-1}_{Y|\boldsymbol{X}=\boldsymbol{x}} \left( \frac{b}{a+b} \right)$$

Of course, $\hat{Y} = f(\boldsymbol{X})$ for new data $\boldsymbol{X}$.

[1] Usually because of robustness to outliers; but beware of the computational burden.

# Decision theory

Different losses imply different predictions for the same prediction problem!

Let $Y \sim N(0,1)$ and $I$ the indicator function:

$$\mathcal{L}(u,a,b,p) = \left(a * I_{(u<0)} + b * I_{(u\geq 0)}\right) * |u|^p$$

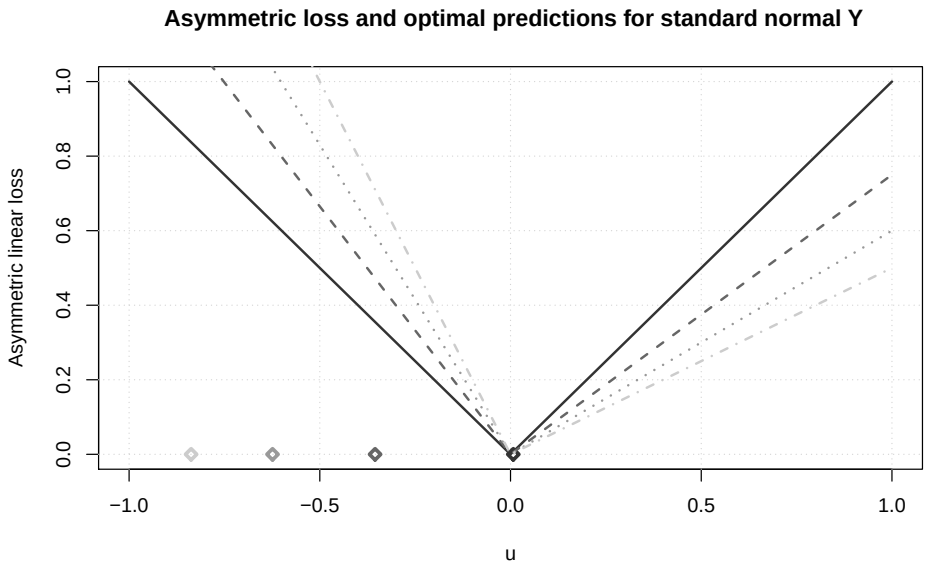



Figure: Linear loss function p = 1

# Decision theory

Different losses imply different predictions for the same prediction problem!

Let $Y \sim N(0,1)$ and $I$ the indicator function:

$$\mathcal{L}(u,a,b,p) = \left(a * I_{(u<0)} + b * I_{(u\geq 0)}\right) * |u|^p$$

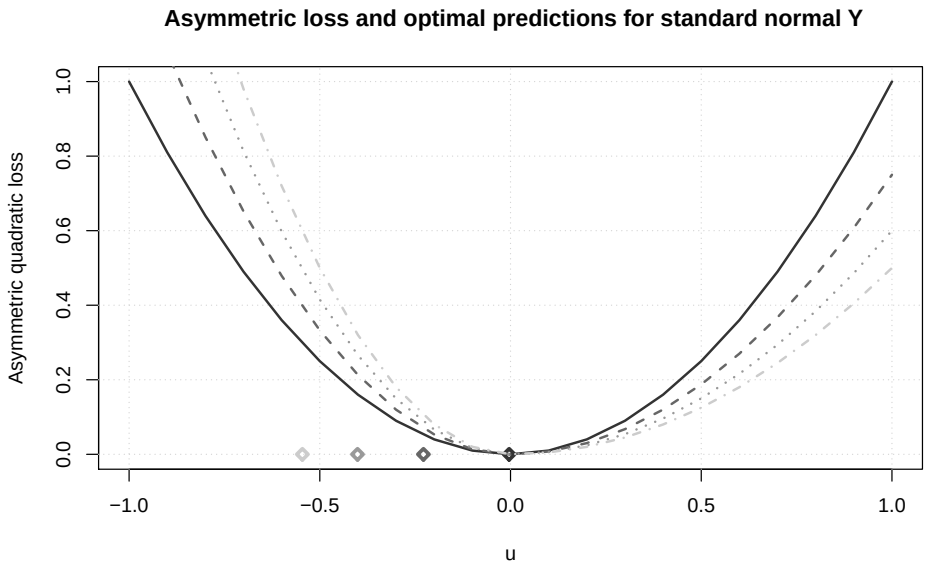



Figure: Quadratic loss function p = 2

# Measures of location and predictions

In general, we get a specific optimal prediction for each loss $\mathcal{L}$.

**Definition ($\mathcal{L}$-Measure of location)**

*The functional* $\mu_Y^{\mathcal{L}} = \arg\min_m \mathrm{E}\left(\mathcal{L}(Y - m)\right)$ *is the $\mathcal{L}$-measure of location of* $Y$.

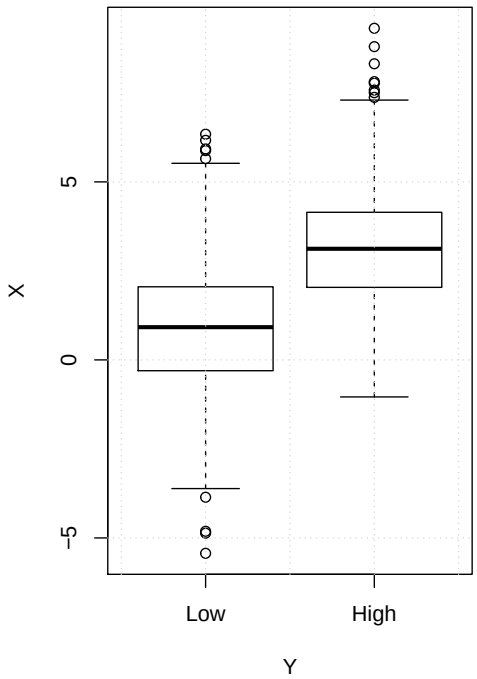
The optimal prediction $\hat{Y}$ given $\boldsymbol{X}$ under $\mathcal{L}$ is the conditional $\mathcal{L}$-measure of location of $Y$, $\mu_{Y|\boldsymbol{X}}^{\mathcal{L}}$, and the predictor function is

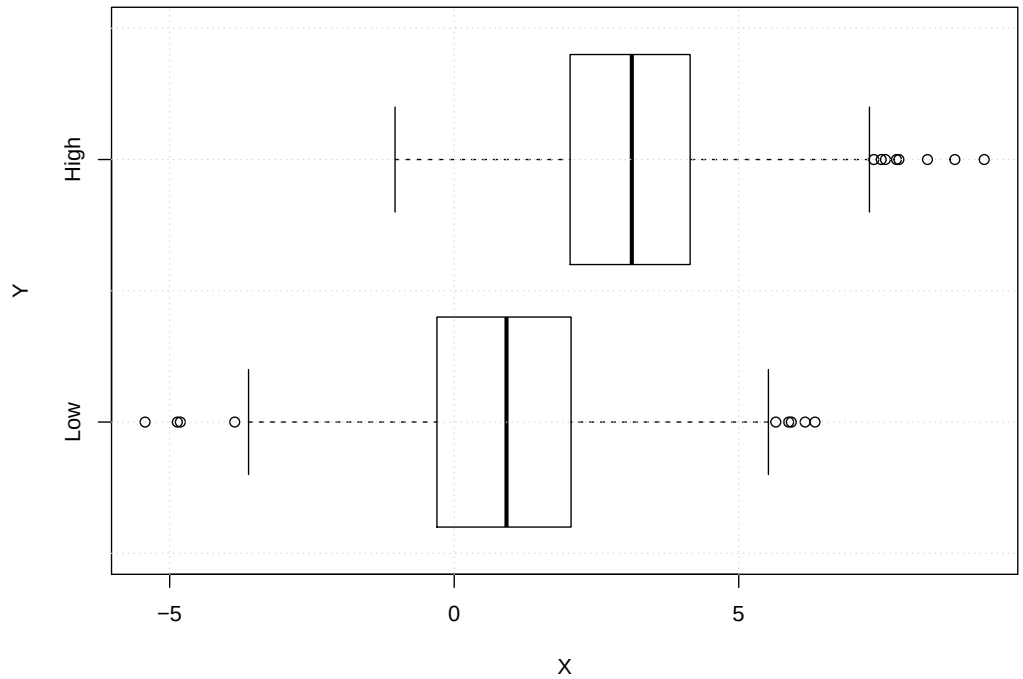
$$f(\boldsymbol{x}) = \mu_{Y|\boldsymbol{X}=\boldsymbol{x}}^{\mathcal{L}}.$$

Changing the loss function changes the prediction problem!

# Classification

Let's look at the binary case, with 2 possible outcomes, say $0$ and $1$.




Here the distribution of $Y$ is Bernoulli, so what matters is the conditional probability that $Y = 1$ given $\boldsymbol{X}$.

# Classification errors

The so-called **confusion matrix** summarizes the error possibilities

| | $\hat{Y}=0$ | $\hat{Y}=1$ |
|---|---|---|
| $Y=0$ | $\ell_{0,0}$ | $\ell_{0,1}$ |
| $Y=1$ | $\ell_{1,0}$ | $\ell_{1,1}$ |

Like for prediction, you could have any cost attached to errors.

The most common loss function for binary classification is the $0/1$ loss:

(We'll start with this one.)

| | $\hat{Y}=0$ | $\hat{Y}=1$ |
|---|---|---|
| $Y=0$ | $0$ | $1$ |
| $Y=1$ | $1$ | $0$ |

# Outline

# Optimal classifier

**Proposition**

*The optimal binary classifier under* $0/1$ *loss sets* $\hat{Y} = 1$ *if* $\mathrm{P}(Y = 1) > \mathrm{P}(Y = 0)$ *and* $\hat{Y} = 0$ *if* $\mathrm{P}(Y = 0) > \mathrm{P}(Y = 1)$.

I.e. the optimal classifier

- picks the class with the larger (conditional) probability of occurrence.
- This extends to several classes, as long as one uses the $0/1$ loss.

For binary classification, we may equivalently state

"Set $\hat{Y} = 1$ if conditional probability of $Y = 1$ exceeds $1/2$".

Other loss functions simply imply another threshold for the conditional class probability.

# Errors and the confusion matrix

Especially in machine learning, it is customary to work with error rates directly rather than losses.

We need some terminology to define them.

| | $\hat{Y} =$ No | $\hat{Y} =$ Yes | |
|---|---|---|---|
| $Y =$ No | True Negatives | False Positives | Actual No |
| $Y =$ Yes | False Negatives | True Positives | Actual Yes |
| | Predicted No | Predicted Yes | Total |

(These error rates are typically given in a yes/no classification situation, and we follow this convention.)

# Error rates

**Accuracy:** (True Positives + True Negatives)/Total

(How often is the classifier correct?)

**Misclassification/Error rate:** (False Positives + False Negatives)/Total

(How often is the classifier wrong?)

**True Positive Rate, Sensitivity or Recall:** True Positives/Actual Yes

(How often is the classifier correct when truth is yes?)

**True Negative Rate or Specificity:** True Negatives/Actual No

(How often is the classifier correct when truth is no?)

**False Positive Rate:** False Positives/Actual No

(How often is the classifier wrong when truth is no?)

**Precision:** True Positives/Predicted Yes

(How often is the classifier correct when prediction is yes?)

**Prevalence:** Actual Yes/Total

(How often does the yes occur in the population or the sample?)

# The Receiver Operating Curve

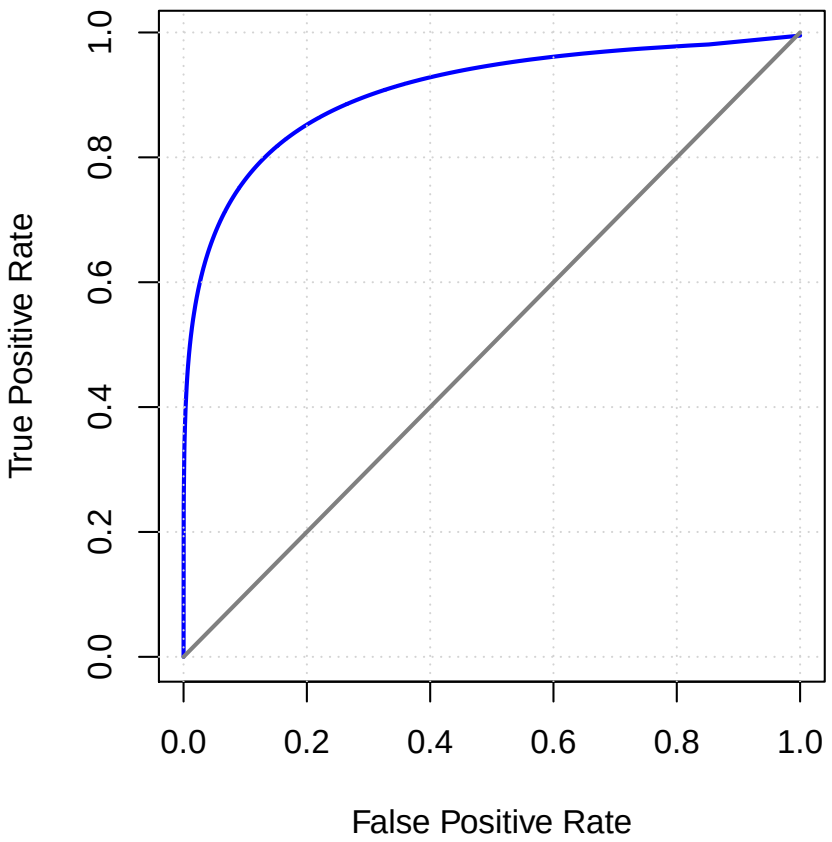
Accuracy is important, but can be misleading if prevalence is either very high or very low.

Sensitivity and specificity should be high, but there's a trade-off:

- This trade-off is controlled by the threshold (say $\tau$) for classification based on conditional class probability.
  - lower $\tau$ means more $\hat{Y} = 1$ (higher sensitivity)
    ... but more of them will be false positives (lower specificity);
  - higher $\tau$ means less $\hat{Y} = 1$ & less false positives (higher specificity)
    ... but at the same time lower sensitivity.
- The "optimal" $\tau$ depends on the loss function ($\tau = 1/2$ for $0/1$ loss).
- Plotting sensitivity (TPR) against specificity (TNR) for all $\tau \in (0, 1)$ gives the so-called Receiver Operating Curve.
- (In fact TPR is plotted for historical reasons against FPR=1-TNR.)

# ROC & AUC

The ROC is informative about the classification performance of a model (population or fitted) for the conditional class probability.




A pure chance classifier (i.e. $\hat{Y}$ independent of $Y$) has TPR = FPR.

So the further the ROC is away from the gray diagonal, the better!

The "Area Under the Curve" quantifies this:

- $0 <$ AUC $< 1$
- AUC $= 1/2$ for pure chance.

# Outline

# Recall classical statistics

Say you have a parametric regression model,

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

where $\mathrm{E}(\varepsilon|\boldsymbol{X}) = 0$.

Estimate using squared error loss,

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}^*} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - f(\boldsymbol{X}_i, \boldsymbol{\theta}^*)\right)^2.$$

**Proposition**

*Regularity conditions assumed,* $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}$.

Since $\mathrm{E}(Y|\boldsymbol{X}) = f(\boldsymbol{X}, \boldsymbol{\theta})$, plug in $\hat{\boldsymbol{\theta}}$ to obtain

$$\hat{f}(\boldsymbol{x}) = f\left(\boldsymbol{x}, \hat{\boldsymbol{\theta}}\right) \qquad \text{and} \qquad \hat{Y} = f\left(\boldsymbol{X}, \hat{\boldsymbol{\theta}}\right).$$

Consistency of $\hat{\boldsymbol{\theta}}$ implies that $\hat{f}(\boldsymbol{x}) \xrightarrow{p} f(\boldsymbol{x})$ and $\hat{Y} \xrightarrow{p} \mathrm{E}(Y|\boldsymbol{X})$.

# Prediction errors

The prediction errors at a (new) data point $(Y, \boldsymbol{X})$ are

$$Y - \hat{Y} = \varepsilon - \left( f\left(\boldsymbol{X}, \hat{\boldsymbol{\theta}}\right) - f\left(\boldsymbol{X}, \boldsymbol{\theta}\right) \right).$$

Since $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}$, we may linearize $f$ in $\boldsymbol{\theta}$ (at the given $\boldsymbol{X}$) and obtain

$$\hat{\varepsilon} = \varepsilon - \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)' \nabla f + O_p\left(\frac{1}{n}\right)$$

where $\nabla f$ is the gradient of $f$ at the population $\boldsymbol{\theta}$ and the given $\boldsymbol{X}$.

We thus have **estimation error** and **irreducible error**.

The estimation error can be made arbitrarily small if there are enough data.

# Misspecification

Say you want to fit the same parametric $f(\boldsymbol{x}, \boldsymbol{\theta})$, but the actual regression function is different, say $\mathrm{E}(Y|\boldsymbol{X}) = m(\boldsymbol{X})$.

**Proposition**

*Let* $\boldsymbol{\theta}_p = \underset{\boldsymbol{\theta}^*}{\arg\min}\ \mathrm{E}\left(\left(f\left(\boldsymbol{X}, \boldsymbol{\theta}^*\right) - m\left(\boldsymbol{X}\right)\right)^2\right)$ *(we call* $\boldsymbol{\theta}_p$ *pseudo-true value). Regularity conditions assumed,* $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_p$.

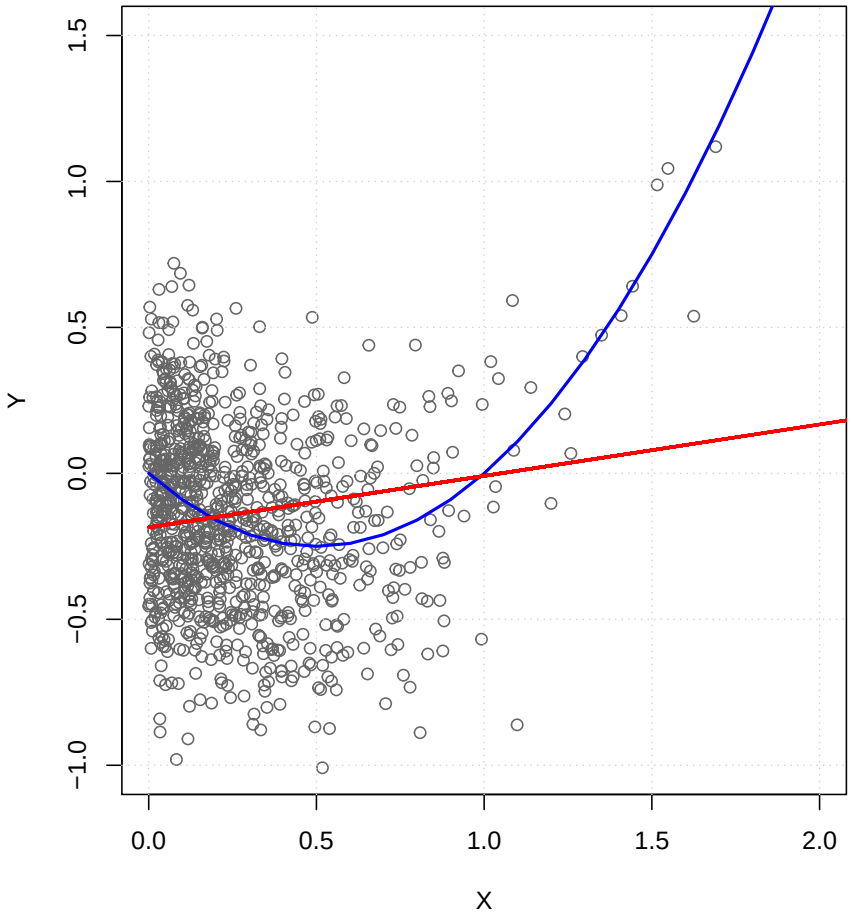
Regarding prediction errors,

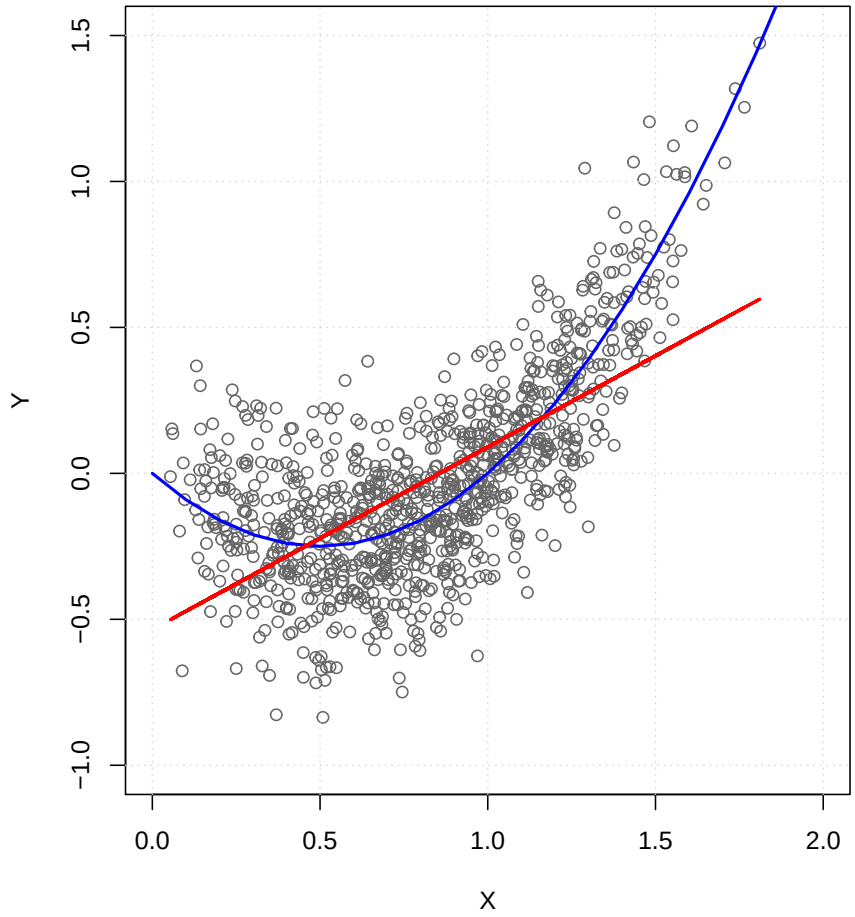
$$Y - \hat{Y} \approx \varepsilon + \left(m\left(\boldsymbol{X}\right) - f\left(\boldsymbol{X}, \boldsymbol{\theta}_p\right)\right) - \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_p\right)' \nabla f$$

where $\nabla f$ is the gradient of $f$ at $\boldsymbol{X}$ for the pseudo-true $\boldsymbol{\theta}_p$.

We thus have **model bias** as an additional source of error.

# Example: Fitting linear regression instead of quadratic




Note that the bias depends on the marginal distribution of $X$!

# Sideline remark: Empirical risk minimization

It is important to realize that you get what you fit!

Say you use a different loss function,

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}^*} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\left(Y_i - f\left(\boldsymbol{X}_i, \boldsymbol{\theta}^*\right)\right).$$

**Proposition**

*Regularity conditions assumed,* $f\left(\boldsymbol{x}, \hat{\boldsymbol{\theta}}\right) \xrightarrow{p} \mu^{\mathcal{L}}_{Y|\boldsymbol{X}=\boldsymbol{x}}$.

E.g. fitting under MAD gives the (fitted) conditional median as predictor.

So you should typically fit predictive models using the same loss function which is used to evaluate the errors.

# Flexibility I

Some parametric families of functions are **universal approximators**.
I.e., as $\dim \boldsymbol{\theta} \to \infty$,

- Polynomials, Fourier series, splines, and also
- regression trees and neural networks,

can approximate smooth functions arbitrarily well.

So choosing "large" $\dim \boldsymbol{\theta}$

- gives enough flexibility of the model such that it
- accommodates arbitrary unknown regression functions.
- The analogous findings holds for classifiers.

Free lunch anybody?

# Flexibility II

The problem is that

$$\operatorname{Var}(\hat{\varepsilon}) \approx \operatorname{Var}(\varepsilon) + \nabla f' \operatorname{Cov}\left(\hat{\boldsymbol{\theta}}\right) \nabla f$$

where

$$\nabla f' \operatorname{Cov}\left(\hat{\boldsymbol{\theta}}\right) \nabla f \leq ||\nabla f||^2 \left\|\operatorname{Cov}\left(\hat{\boldsymbol{\theta}}\right)\right\|.$$

In the worst-case scenario, $\left\|\operatorname{Cov}\left(\hat{\boldsymbol{\theta}}\right)\right\| = O(\dim \boldsymbol{\theta})$!

You buy bias reduction by increasing the variance!

The bias-variance trade-off[2] is really everywhere...

[2] Further readings: ESL Chapter 7.2 Bias, Variance and Model Complexity

# Bias-Variance trade-off

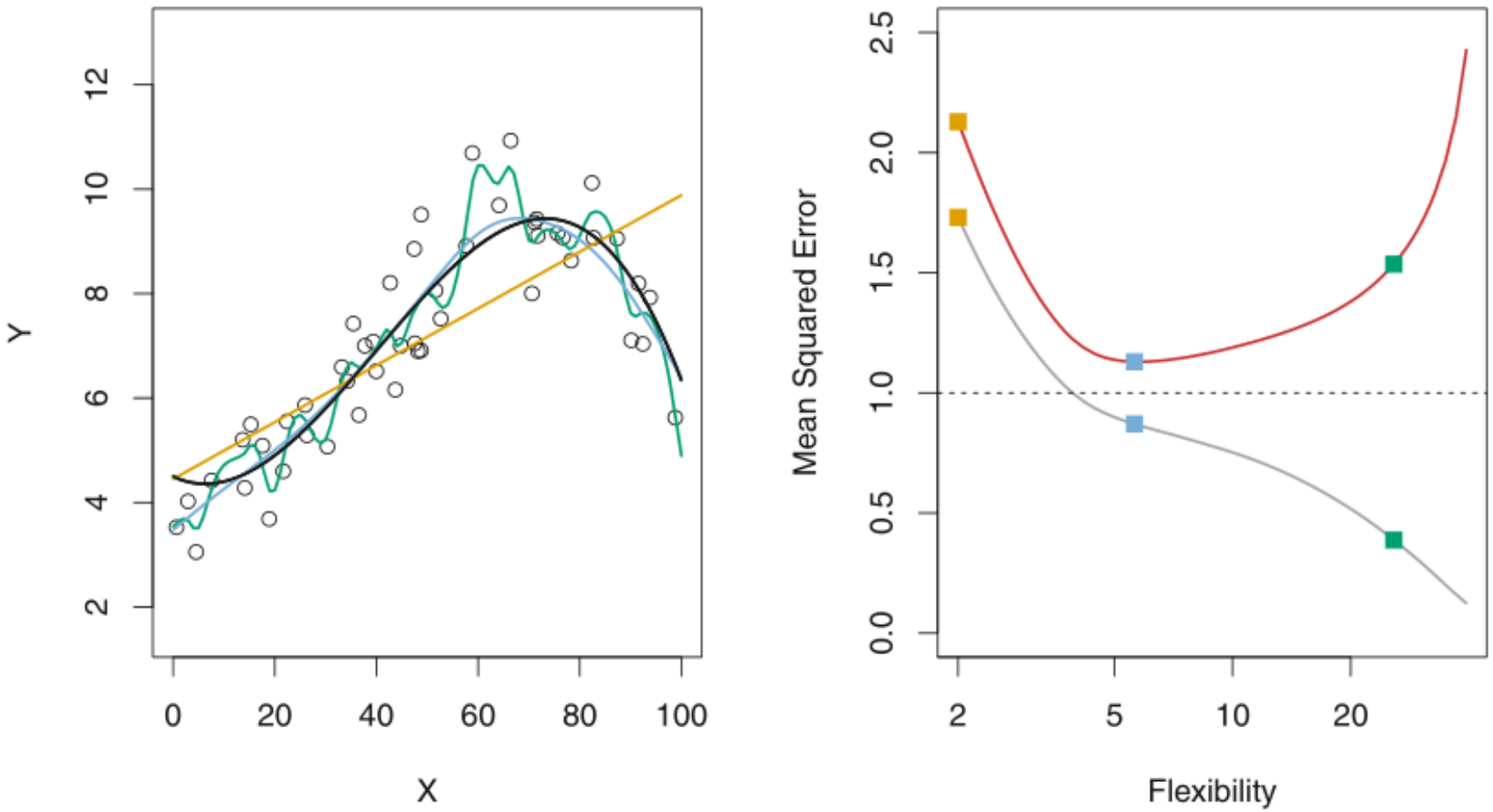



Figure: Left: Data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves). Right: Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line). Squares represent the training and test MSEs for the three fits shown in the left-hand panel. FIGURE 2.9. ISLR

# Towards model selection

One minimizes prediction MSE by balancing bias and estimation error.

- This is a critical step, since both influence prediction MSE,
- and can be seen as a model selection issue.

In statistical learning we have new names:

- Bias dominates: **underfitting** (oversmoothing)
- Variance dominates: **overfitting** (undersmoothing)

Choosing the right amount of smoothing is sometimes method-specific so we deal with it on suitable occasions.

# More information is a good thing, right?

Intuitively,

The more features you have the better!

E.g. chances increase that a powerful predictor is among them.

- You need however enough data to be able to incorporate the extra features in your model.
- Clearly, additional features (and correspondingly parameters) for the same amount of data increase estimation variance.

# The curse of dimensionality

The **curse of dimensionality** is a generic term used to describe situations where it is very costly to exploit additional features.

It originated in dynamic optimization, where the cost of adding one dimension to a grid search increases exponentially with the number of dimensions.

Its exact form depends on the model, but the source is the combination of high dimensionality with high flexibility.

# The curse is not only for prediction!

Classification models also suffer from the curse of dimensionality.

And unsupervised learning has its dimensionality issues as well.

- Imagine a clustering task with $n$ data points and $p = n$ features.
- Say for simplicity that the $n$ features are binary $0/1$,
- ... and $X_{i,j} = 1$ for $i = j$ only.
- Unless you pay attention, most algorithms will immediately find $n$ different clusters

For all, **dimensionality reduction** may help deal with the curse.
(And so does **model selection**.)

# Outline

# Coming up

Linear prediction