# Regression I: Mean Squared Error and Measuring Quality of Fit

## -Applied Multivariate Analysis-

Lecturer: Darren Homrighausen, PhD

# The Setup

Suppose there is a scientific problem we are interested in solving

(This could be estimating a relationship between height and weight in humans)

The perspective of this class is to define these quantities as random, say

- $X =$ height
- $Y =$ weight

We want to know about the joint distribution of $X$ and $Y$

This joint distribution is unknown, and hence we

1. GATHER DATA
2. ESTIMATE IT

# The Setup

Now we have data

$$D_n = \{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\},$$

where

- $X_i \in \mathbb{R}^p$ are the covariates, explanatory variables, or predictors
  (NOT INDEPENDENT VARIABLES!)
- $Y_i \in \mathbb{R}$ are the response or supervisor variables.
  (NOT DEPENDENT VARIABLE!)

Finding the joint distribution of $X$ and $Y$ is usually too ambitious

A good first start is to try and get at the mean, say

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

where $\epsilon$ describes the random fluctuation of $Y$ around its mean

# Parameterizing this relationship

Even estimating $\mu(X)$ is often too much

A good simplification is to assume that it is linear.

This means we suppose that there is a $\beta \in \mathbb{R}^p$ such that:

$$Y = \underbrace{\mu(X) + \epsilon = X^\top \beta + \epsilon}_{\text{simplification}} \in \mathbb{R}$$

(The notation $\in$ indicates 'in' and if I say $x \in \mathbb{R}^q$, that means that $x$ is a vector with $q$ entries)

# Parameterizing this relationship

Translated into using our data, we get

$$Y = \mathbb{X}\beta + \epsilon \in \mathbb{R}^n$$

where

$$\mathbb{X} = \begin{bmatrix} X_1^\top \\ X_2^\top \\ \vdots \\ X_n^\top \end{bmatrix}$$

Commonly, $\mathbb{X}_{i1} = 1$, which encodes an intercept term in the model.

$\mathbb{X}$ is known as the design or feature matrix

# Parameterizing this relationship: In detail

Back to the height/weight example:

$$X_1^\top = [1, 62], X_2^\top = [1, 68], \ldots, X_{10}^\top = [1, 65]$$

Estimating

$$\hat\beta = \underset{\beta}{\operatorname{argmin}} \, ||\mathbb{X}\beta - Y||_2^2 = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \left( Y_i - X_i^\top \beta \right)^2$$

finds the usual simple linear regression estimators: $\hat\beta^\top = [\hat\beta_0, \hat\beta]$

# Some terminology and overarching concepts

# Core terminology

Statistics is fundamentally about two GOALS

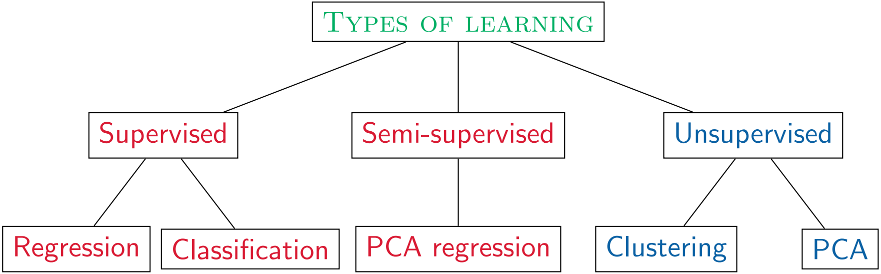
- Formulating an estimator of unknown, but desired, quantities

  (This is known as (statistical) learning)

- Answering the question: How good is that estimator?

  (For this class, we will focus on if the estimator is sensible and if it can make 'good' predictions)

Let's address some relevant terminology for each of these GOALS



Types of learning

Supervised

Semi-supervised

Unsupervised

Regression

Classification

PCA regression

Clustering

PCA

Some comments:

Comparing to the response (aka supervisor) $Y$ gives a natural notion of prediction accuracy

Much more heuristic, unclear what a good solution would be. We'll return to this later in the semester.

# How good is that estimator?

Suppose we are attempting to estimate a quantity $\beta$ with an estimator $\hat{\beta}$.

(We don't need to be too careful as to what this means in this class. Think of it as a procedure that takes data and produces an answer)

How good is this estimator?

We can look at how far $\hat{\beta}$ is from $\beta$ through some function $\ell$

Any distance function[1] will do...

[1] Or even topology...

# Risky (and lossy) business

We refer to the quantity $\ell(\hat{\beta}, \beta)$ as the loss function.

As $\ell$ is random (it is a function of the data, after all), we usually want to average it over the probability distribution of the data:

This produces

$$R(\hat{\beta}, \beta) = \mathbb{E}\ell(\hat{\beta}, \beta)$$

which is called the risk function.

# Risky business

To be concrete, however, let's go through an important example:

$$\ell(\hat{\beta}, \beta) = \left|\left|\hat{\beta} - \beta\right|\right|_2^2$$

(Note that we tend to square the 2-norm to get rid of the pesky square root.)

Also,

$$R(\hat{\beta}, \beta) = \mathbb{E}\left|\left|\hat{\beta} - \beta\right|\right|_2^2$$

# Risky business

Back to the original question: What Makes a Good Estimator of $\beta$?

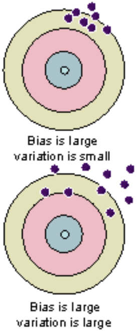
Answer: One that has small risk!

It turns out that we can form the following

$$\begin{aligned}
R(\hat{\beta}, \beta) &= \mathbb{E} \left|\left|\hat{\beta} - \beta\right|\right|_2^2 \\
&= \mathbb{E} \left|\left|\hat{\beta} - \mathbb{E}\hat{\beta} + \mathbb{E}\hat{\beta} - \beta\right|\right|_2^2 \\
&= \mathbb{E} \left|\left|\hat{\beta} - \mathbb{E}\hat{\beta}\right|\right|_2^2 + \mathbb{E} \left|\left|\mathbb{E}\hat{\beta} - \beta\right|\right|_2^2 + 2\mathbb{E}(\hat{\beta} - \mathbb{E}\hat{\beta})^\top(\mathbb{E}\hat{\beta} - \beta) \\
&= \mathbb{E} \left|\left|\hat{\beta} - \mathbb{E}\hat{\beta}\right|\right|_2^2 + \left|\left|\mathbb{E}\hat{\beta} - \beta\right|\right|_2^2 + 0 \\
&= \mathrm{trace}\mathbb{V}\hat{\beta} + \left|\left|\mathbb{E}\hat{\beta} - \beta\right|\right|_2^2 \\
&= \mathrm{Variance} \ + \mathrm{Bias}
\end{aligned}$$

# Bias and variance

Two important concepts in statistics: Bias and Variance



Accuracy versus Quality of an Estimator Using Bias and Variation as Measurable Quantities Respectively

$$R(\hat{\beta}, \beta) = \text{Variance } + \text{Bias}$$

# Bias and variance

So, what makes a good estimator?

If...

1. $R(\hat{\beta}, \beta) = \text{Variance } + \text{Bias}$
2. we want $R(\hat{\beta}, \beta)$ to be small

$\Rightarrow$ We want a $\hat{\beta}$ that optimally trades off bias and variance

(Note, crucially, that this implies that biased estimators are generally better)

This is great but...

# Bias and variance

So, what makes a good estimator?

If...

1. $R(\hat{\beta}, \beta) = \text{Variance } + \text{Bias}$
2. we want $R(\hat{\beta}, \beta)$ to be small

$\Rightarrow$ We want a $\hat{\beta}$ that optimally trades off bias and variance

(Note, crucially, that this implies that biased estimators are generally better)

This is great but...

We don't know $\mathbb{E}$ and hence we don't know $R(\hat{\beta}, \beta)$!

# We don't know the risk

Since the risk is unknown, we need to estimate it

The risk is by definition an average, so perhaps we should use the data...

This means translating

- risk in terms of just $\beta$ into risk in terms of $\mathbb{X}\beta$

  (This might seem strange. I'm omitting some details to limit complexity)

$$\mathbb{E}\left|\left|\hat{\beta} - \beta\right|\right|_2^2 \Rightarrow \mathbb{E}\left|\left|\mathbb{X}\hat{\beta} - \mathbb{X}\beta\right|\right|_2^2$$

- "$\mathbb{E}$" into "$\sum$"

  (i.e.: using the data to compute an expectation. You've done this before!)

# What Makes a Good Estimator of $\beta$?

An intuitive and well-explored criterion is known variously as

- Mean Squared Error (MSE)
- Residual Sums of Squares (RSS)
- Training error

  (We'll get back to this last one)

which for an arbitrary estimator $\hat{\beta}$ has the form:

$$\mathrm{MSE}(\hat{\beta}) = \sum_{i=1}^{n} \left( Y_i - X_i^\top \hat{\beta} \right)^2 \overset{??}{\approx} \mathbb{E} \left| \left| \mathbb{X}\hat{\beta} - \mathbb{X}\beta \right| \right|_2^2$$

Here, we see that if $\hat{\beta}$ is such that $X_i^\top \hat{\beta} \approx Y_i$ for all $i$, then $\mathrm{MSE}(\hat{\beta}) \approx 0$.

But, there's a problem... we can make MSE arbitrarily small...

# What Makes a Good Estimator of $\beta$?

Here's an example:

Let's suppose we have 20 observations with one explanatory variable and one response.

Now, let's fit some polynomials to this data.

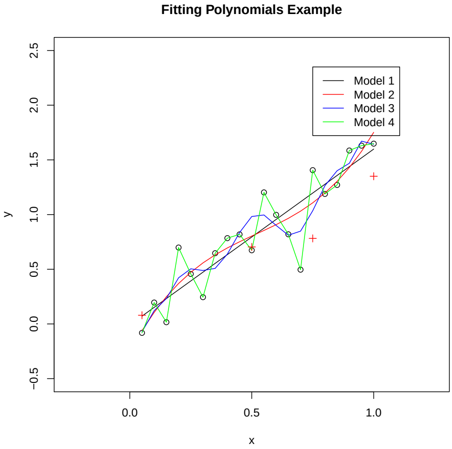
Let $\mu_i$ be the conditional mean of the response $Y_i$.
(That is $\mu_i = \mu(X_i)$ )

We consider the following models:

- Model 1: $\mu_i = \beta_0 + \beta_1 X_i$
- Model 2: $\mu_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3$
- Model 3: $\mu_i = \beta_0 + \sum_{k=1}^{10} \beta_k X_i^k$
- Model 4: $\mu_i = \beta_0 + \sum_{k=1}^{100} \beta_k X_i^k$

Let's look at what happens...

# What Makes a Good Estimator of $\beta$?



The MSE's are:

$$\text{MSE(Model 1)} = 0.98$$

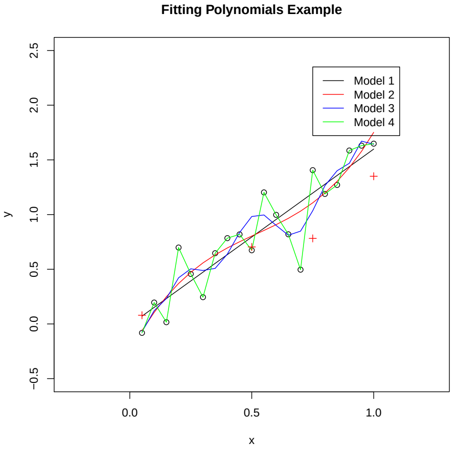
$$\text{MSE(Model 2)} = 0.86$$

$$\text{MSE(Model 3)} = 0.68$$

$$\text{MSE(Model 4)} = 0$$

What about predicting new observations (red crosses)?

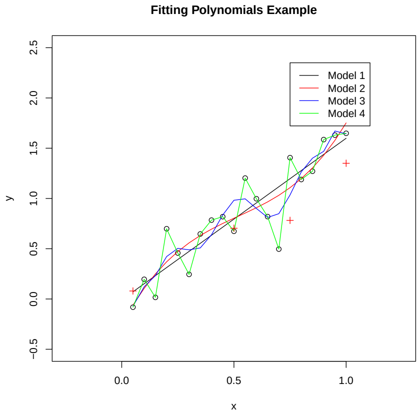
# Example of this Inverse Relationship

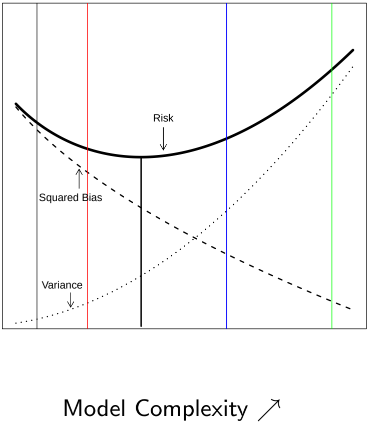


- Black model has low variance, high bias
- Green model has low bias, but high variance
- Red model and Blue model have intermediate bias and variance.

We want to balance these two quantities.

# Bias vs. Variance





The best estimator is at the vertical black line (minimum of Risk)

# Training and test error

# A better notion of risk

What we are missing is that the same data that goes into training $\hat{\beta}$ goes into testing $\hat{\beta}$ via $\mathrm{MSE}(\hat{\beta})$

What we really want is to be able to predict a new observation well

Let $(X_0, Y_0)$ be a new observation that has the same properties as our original sample $D_n$, but is independent of it.

# A better notion of risk

It turns out that

$$\mathbb{E}\left|\left|\mathbb{X}\hat{\beta} - \mathbb{X}\beta\right|\right|_2^2 \text{ is the "same" as } \mathbb{E}(Y_0 - X_0^\top \hat{\beta})^2$$

("same" means that it isn't equal, but behaves the same)

Of course,

$$\mathrm{pred}(\hat{\beta}) = \mathbb{E}(Y_0 - X_0^\top \hat{\beta})^2$$

still depends on information that is unavailable to the data analyst

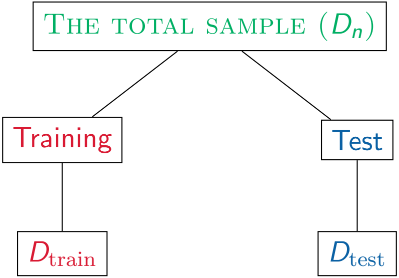
(in particular, the joint distribution of $X_0$ and $Y_0$)

# Training and test split

We can mimic

$$\mathrm{pred}(\hat{\beta}) = \mathbb{E}(Y_0 - X_0^\top \hat{\beta})^2$$

by splitting the sample $D_n$ into two parts



The total sample ($D_n$)

Training

Test

$D_{\text{train}}$

$D_{\text{test}}$

Where

- Every observation from $D_n$ is in $D_{\text{train}}$ or $D_{\text{test}}$
- No observation is in both

# Training and test split

Now, instead of trying to compute

$$\mathrm{pred}(\hat{\beta}) = \mathbb{E}(Y_0 - X_0^\top \hat{\beta})^2,$$

we can instead

- train $\hat{\beta}$ on the observations in $D_{\text{train}}$
- compute the $\mathrm{MSE}$ using observations in $D_{\text{test}}$ to test

Example: Commonly, this might be 90% of the data in $D_{\text{train}}$ and 10% of the data in $D_{\text{test}}$

Question: Where does the terminology training error come from?

# Training and test split

This approach has major pros and cons

- Pro: This is a fair estimator of risk as the training and test observations are independent
- Con: We are sacrificing power by using only a subset of the data for training

# Other estimates of risk

There are many candidates for estimating $\text{pred}$

- AIC (Akaike information criterion)
- AICc (AIC with a correction)
- BIC (Bayesian information criterion)
- Mallows Cp

Don't worry overly much about the differences between each of these criteria.

# What Makes a Good Estimator of $\beta$?

For example:

$$\text{pred}(\hat{\beta}) \approx \text{AIC}(\hat{\beta}) = -2\ell(\hat{\beta}) + 2|\hat{\beta}|.$$

Here,

- $\ell$ is the log likelihood under Gaussian errors
  (This term is effectively $\text{MSE}$)
- $|\hat{\beta}|$ is the length of the estimator (the number of parameters)

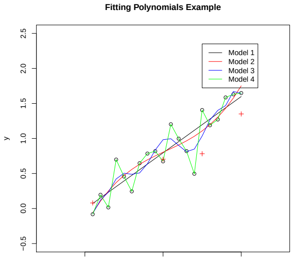
For the polynomial example:

$$|\hat{\beta}| = \begin{cases} 2 & \text{for Model 1} \\ 4 & \text{for Model 2} \\ 11 & \text{for Model 3} \\ 101 & \text{for Model 4} \end{cases}$$

# What Makes a Good Estimator of $\beta$?

$$\begin{aligned} \text{MSE(Model 1)} &= 0.98 \\ \text{MSE(Model 2)} &= 0.86 \\ \text{MSE(Model 3)} &= 0.68 \\ \text{MSE(Model 4)} &= 0 \end{aligned} \quad \text{and} \quad |\hat{\beta}| = \begin{cases} 2 & \text{for Model 1} \\ 4 & \text{for Model 2} \\ 11 & \text{for Model 3} \\ 101 & \text{for Model 4} \end{cases}$$

$$\begin{aligned} \text{AIC(Model 1)} &= 0.98 + 2 * 2 &&= 4.98 \\ \text{AIC(Model 2)} &= 0.86 + 2 * 4 &&= 8.86 \\ \text{AIC(Model 3)} &= 0.68 + 2 * 11 &&= 22.68 \\ \text{AIC(Model 4)} &= 0 + 2 * 101 &&= 202.000 \end{aligned}$$



# What Makes a Good Estimator of $\beta$?

I'll include the form of each of the criteria for future reference, formulated for regression:

$$\begin{aligned}
\mathrm{AIC}(\hat{\beta}) &= \mathrm{MSE}(\hat{\beta}) + 2|\hat{\beta}| \\
\mathrm{AICc}(\hat{\beta}) &= \mathrm{AIC} + \frac{2|\hat{\beta}|(|\hat{\beta}|+1)}{n-|\hat{\beta}|-1} \\
\mathrm{BIC}(\hat{\beta}) &= \mathrm{MSE}(\hat{\beta}) + |\hat{\beta}|\log n \\
\mathrm{Cp}(\hat{\beta}) &= \frac{\mathrm{MSE}}{\hat{\sigma}^2} - n + 2|\hat{\beta}|.
\end{aligned}$$

Note:

- As long as $\log n \geq 2$ (which is effectively always), BIC picks a smaller model than AIC.
- As $n \geq |\hat{\beta}| + 1$, AICc always picks a smaller model than AIC.
- AICc is a correction for AIC. You should use AICc in practice/research if available. In this class we'll just use AIC so as not to get bogged down in details.

# Finally, we can answer: What Makes a Good Estimator of $\beta$?

A good estimator is one that minimizes one of those criteria. For example:

$$\hat{\beta}_{\text{good}} = \underset{\hat{\beta}}{\text{argmin}}\, AIC(\hat{\beta}).$$

Algorithmically, how can we compute $\hat{\beta}_{\text{good}}$? One way is to compute all possible models and their associated AIC scores.

However, this is a lot of models...

# Caveat

In order to avoid some complications about AIC and related measures, I've over-simplified a bit

Sometimes you'll see AIC as written in three ways:

- $\text{AIC} = \log(\text{MSE}/n) + 2|\hat{\beta}|$
- $\text{AIC} = \text{MSE} + 2|\hat{\beta}|$
- $\text{AIC} = \text{MSE} + 2|\hat{\beta}|\text{MSE}/n$

The reasons to prefer one over the other is too much of a distraction for this class

If you're curious, I can set up a special discussion outside of class to discuss this.

(The following few slides are optional and get at the motivation behind 'information criteria' based methods)

# Comparing probability measures

Information criteria come from many different places

Suppose we have data $Y$ that comes from the probability density function $f$.

What happens if we use the probability density function $g$ instead?

One central idea is Kullback-Leibler discrepancy[2]

$$\begin{aligned} KL(f,g) &= \int \log\left(\frac{f(y)}{g(y)}\right) f(y)dy \\ &\propto -\int \log(g(y))f(y)dy \qquad \text{(ignore term without } g\text{)} \\ &= -\mathbb{E}[\log(g(Y))] \end{aligned}$$

This gives us a sense of the loss incurred by using $g$ instead of $f$.

[2] This has many features of a distance, but is not a true distance as $KL(f,g) \neq KL(g,f)$.

# KULLBACK-LEIBLER DISCREPANCY

Usually, $g$ will depend on some parameters, call them $\theta$, and write $g(y;\theta)$.

**Example:** In regression, we can specify $f \sim N(X^\top\beta, \sigma^2)$ for a fixed (true)[3] $\beta$, and let $g \sim N(X^\top\theta, \sigma^2)$ over all $\theta \in \mathbb{R}^p$

As $KL(f,g) = -\mathbb{E}[\log(g(Y;\theta))]$, we want to minimize this over $\theta$.

Again, $f$ is unknown, so we minimize $-\log(g(y;\theta))$ instead. This is the maximum likelihood value

$$\hat{\theta}_{ML} = \arg\max_\theta g(y;\theta)$$

[3] We actually don't need to assume things about a true model nor have it be nested in the alternative models.

# Kullback-Leibler discrepancy

Now, to get an operational characterization of the KL divergence at the ML solution

$$-\mathbb{E}[\log(g(Y;\hat{\theta}_{ML}))]$$

we need an approximation (don't know $f$, still)

This approximation is exactly AIC:

$$\mathrm{AIC} = -2\log(g(Y;\hat{\theta}_{ML})) + 2|\hat{\theta}_{ML}|$$

**Example:** If $\log(g(y;\theta)) = \frac{n}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}||y - \mathbb{X}\theta||_2^2$, as in regression, and $\sigma^2$ is known, Then using $\hat{\theta} = (\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top y$,

$$\mathrm{AIC} \propto MSE/\sigma^2 + 2p$$