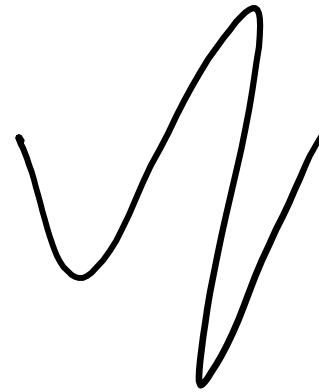


CS 412



JAN 28TH – LINEAR REGRESSION

HTF – CHAPTER 3.1-3.3

Administrivia

Office hours, my office (SEO931):

- Today: 12-1
- Thursday: 5-7

HW1 is due on Thursday

Lecture capture

- Only one section has access
- I have emailed the ACCC and hopefully they'll have it ready by Thursday.

HW1 Notes

Cross_val_score and cross_val_predict

- Predict returns a list of each of the fold errors, helpful for calculating mean and variance

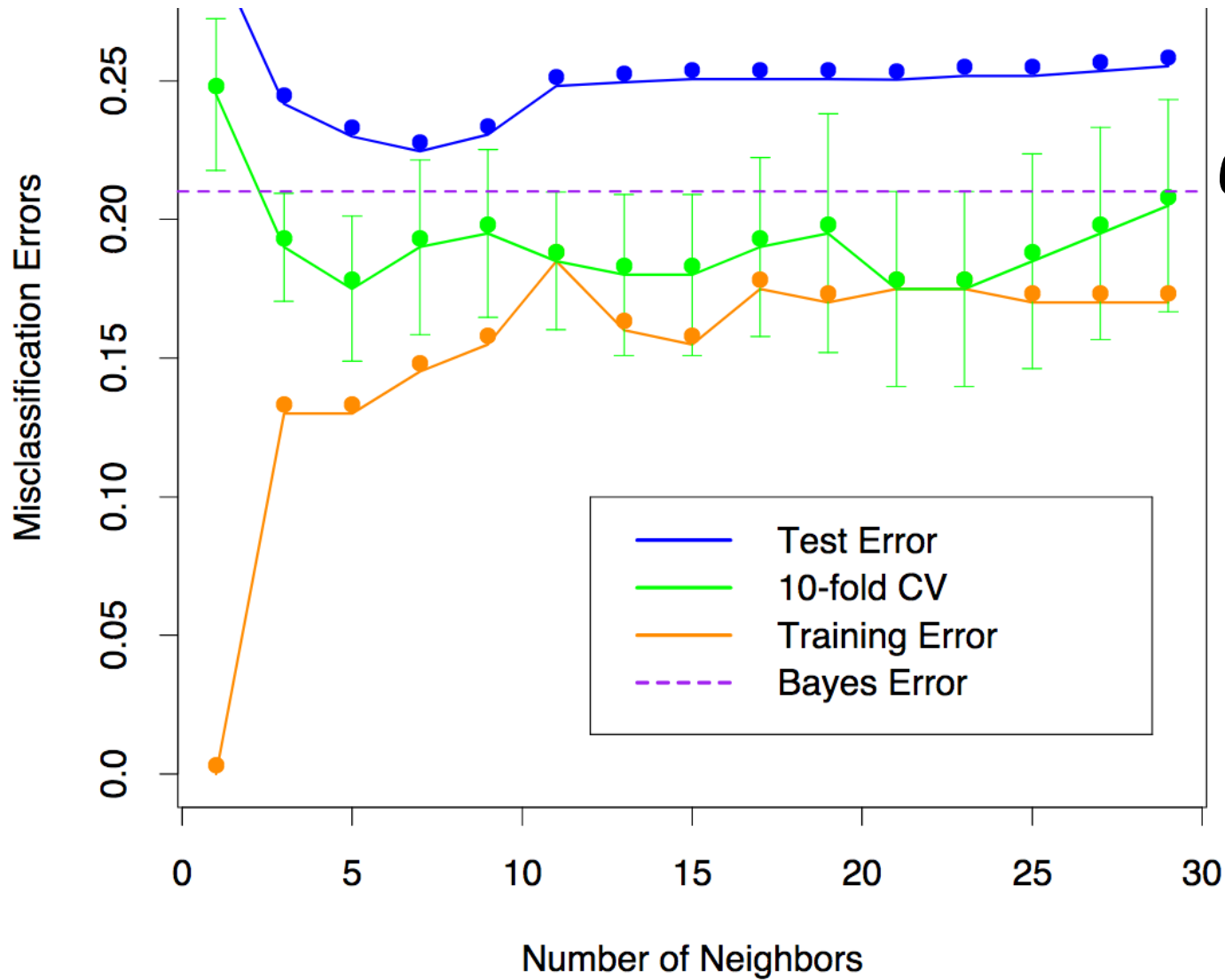
What steps are necessary to produce the graph (Figure 1.4) of E_{cv} ?

- Need to find the X and Y points to scatterplot
- What is X?

- What is Y?

- How are they calculated?

E_{cv}
for loop(k)
cross_val_predict
 E_{cv}



← HW

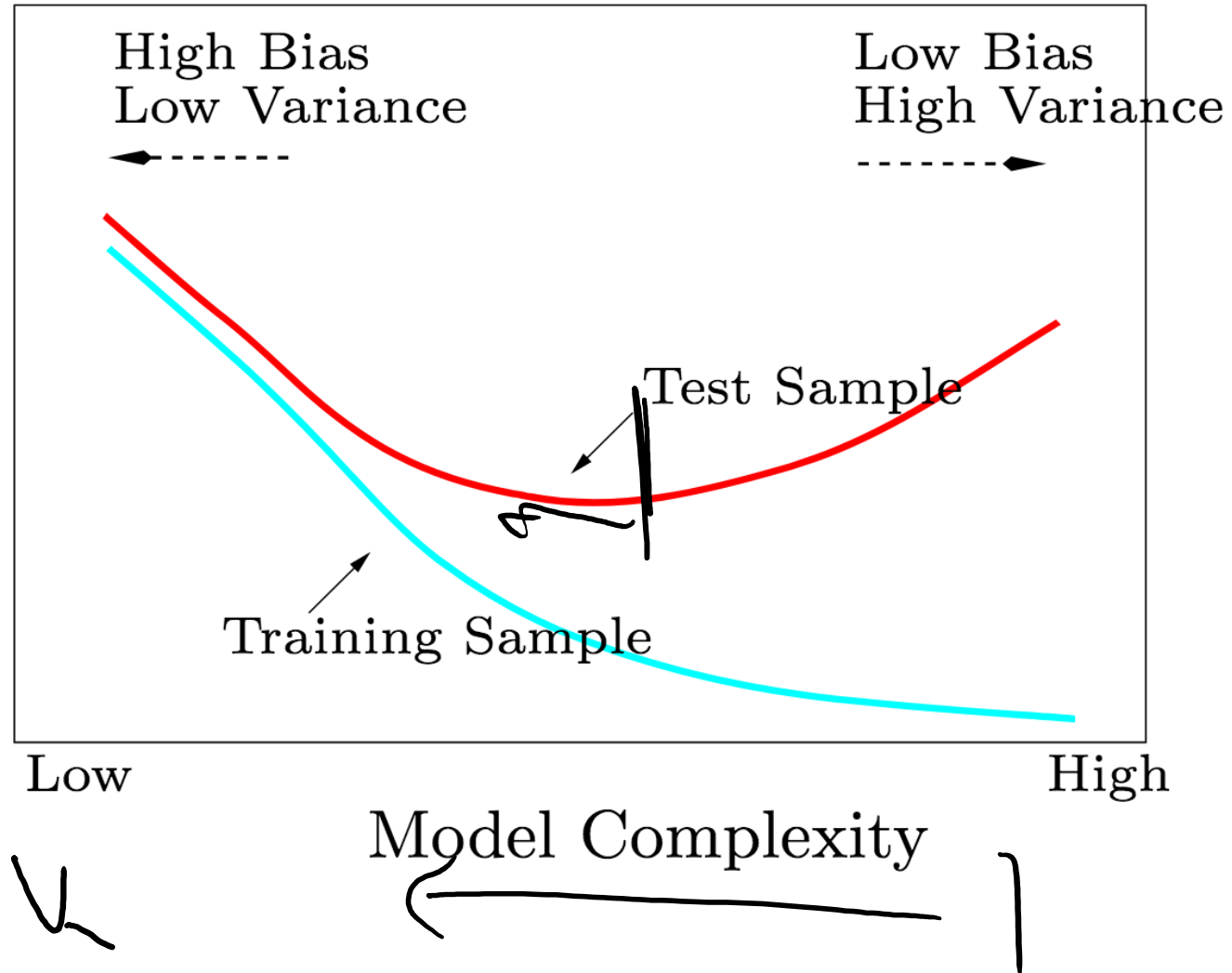
Results

This illustrates a common trade-off

Bias vs. Variance

The more we test (and the more complicated our model), the lower our bias is.

However, we introduce more variance, which is represented in the test data.



Results

This illustrates a common trade-off

Bias vs. Variance

The more we test (and the more complicated our model), the lower our bias is.

However, we introduce more variance, which is represented in the test data.

The Linear Model

The linear model assumes that $E(Y|X)$ is a linear combination of inputs X_1, X_2, \dots, X_p

not necessarily assuming
 $Y|X = \sum$

$+ \epsilon$
↑
mean zero

The Linear Model

The linear model assumes that $E(Y|X)$ is a linear combination of inputs X_1, X_2, \dots, X_p

Under this assumption, how do we produce our 'linear model'?

Choose the value that minimizes our EPE!

Estimated prediction error

The Linear Model

The linear model assumes that $E(Y|X)$ is a linear combination of inputs X_1, X_2, \dots, X_p

Under this assumption, how do we produce our 'linear model'?

Choose the value that minimizes our EPE!

For a linear model, what is the EPE?

- Recall: $EPE(f) = E(Y - f(X))^2$

This minimization problem is the least squares!

$$E(\epsilon) = 0$$

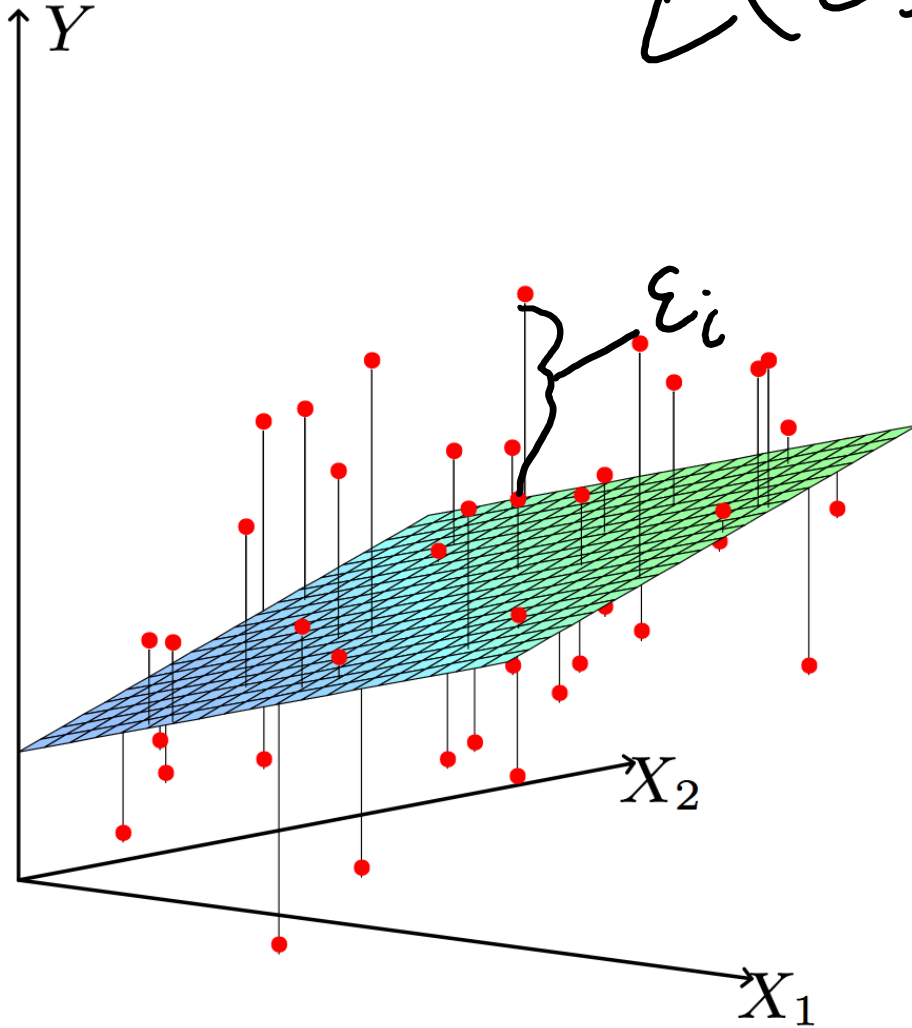
Linear Regression

Regression problems are trying to predict some output value ($Y \in \mathbb{R}$) which is a function of the input variables ($X \in \mathbb{R}^2$)

This is a different problem than *classification*

- kNN can be used for both
- So can linear models

Start with linear regression and move to logistic regression



Two input features

A note about preqreqs

A note about the linear algebra

- Not a prerequisite for the course
- Whatever linear algebra we'll need, I'll give in class
- Mostly helpful for notation

Linear Regression

Assume that the output variable is some linear combination of input variables

Linear Regression

Assume that the output variable is some linear combination of input variables

This is probably uncommon.

Why learn linear regression models then?

- Basis for kernel methods
- Can be a good baseline performance
- More expressive than you'd initially expect

Linear Regression

Assume that the output variable is some linear combination of input variables

This is probably uncommon.

Why learn linear regression models then?

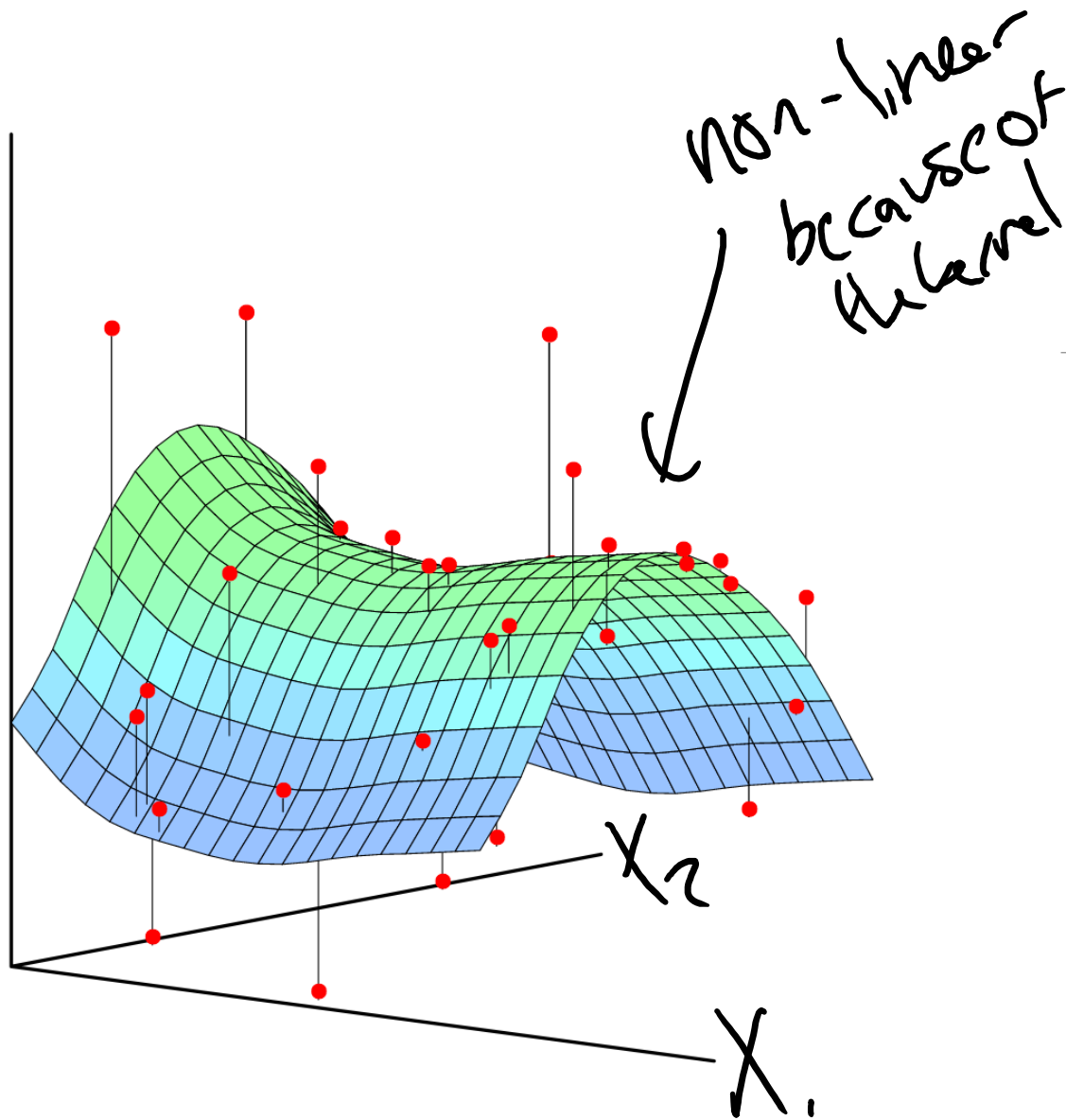
- Basis for kernel methods
- Can be a good baseline performance
- More expressive than you'd initially expect

Variables are not restricted to being just the inputs (X), they can also be:

- Transformations or interactions ($\log(X_1)$, X_1X_2)
- Dummy encodings
- Polynomial expansions

← need to convert categorical variables

kernel



Linear Regression

The more variables we include:

- higher our risk for overfitting
- higher our expected error
- more complex data can be modeled

Since this is a statistical approach, we can directly bound the error of the model

This is an easy approach for a more robust statistical (and interpretable) result

$$\varphi = \{x_1, x_2, x_1^2, x_2^2, x_1 x_2\}$$


Linear Regression

$$\beta_0 \rightarrow \beta_j$$

So what is the linear regression problem?

- For each of our p variables in X
- Apply some constant (not dependent on any X_i) multiple β

$f(X)$ is our approximation of the output


$$f(X) = \beta_0 + \sum_{j=1}^p X_j \beta_j.$$

this how
we ensure $E(y - f(X)) = 0$

Linear Regression

How accurate is the model?

- We need a good metric
- *Misclassification accuracy* is common, but this isn't a classification problem

Loss function

Linear Regression

How accurate is the model?

- We need a good metric
- *Misclassification accuracy* is common, but this isn't a classification problem

Use the residual sum of squares

Least squares

Why is β the input?

vector of
weights + bias
 $|B| = p + 1$

for the training set

$$\text{RSS}(\beta) = \sum_{i=1}^N (y_i - f(x_i))^2$$

substituting

$$= \sum_{i=1}^N \left(y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j \right)^2.$$

Linear Regression

How accurate is the model?

- We need a good metric
- *Misclassification accuracy* is common, but this isn't a classification problem

Use the residual sum of squares

$$\begin{aligned}\text{RSS}(\beta) &= \sum_{i=1}^N (y_i - f(x_i))^2 \\ &= \sum_{i=1}^N \left(y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j \right)^2.\end{aligned}$$

Why is β the input?

- β is the vector of weights
- *This is the only part of the model*

Finding the model

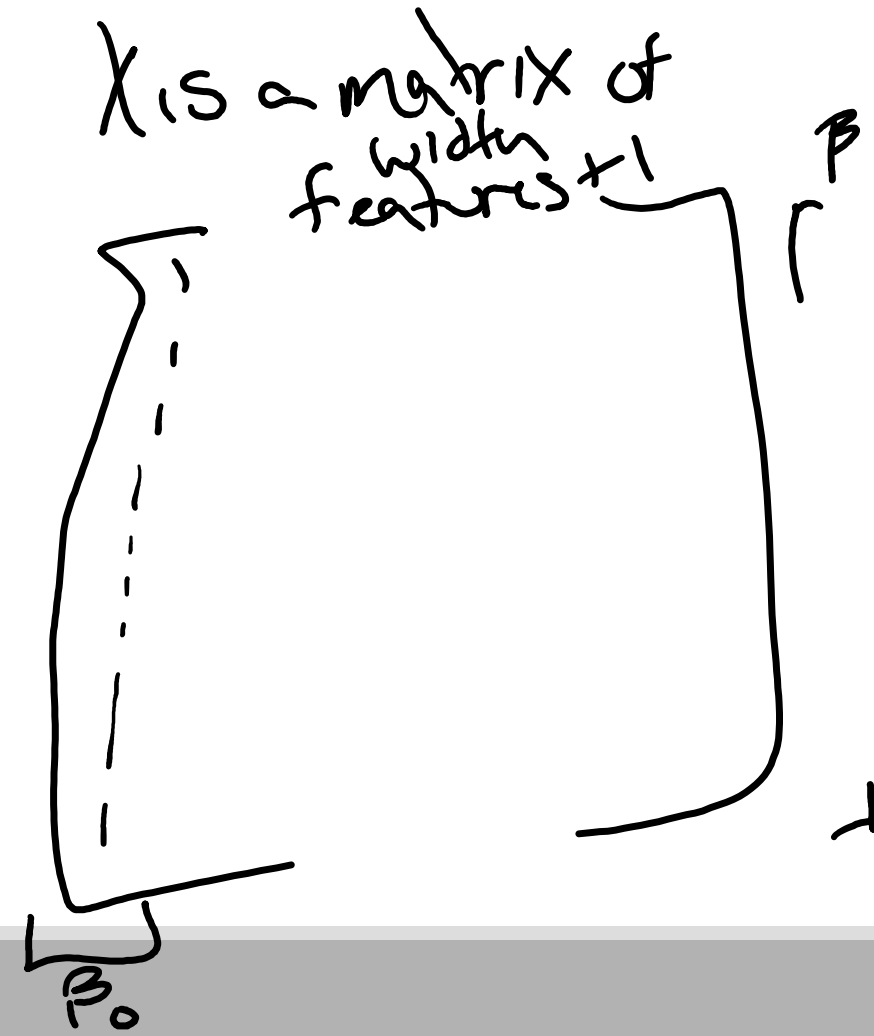
How do we minimize the residual sum of squares (RSS)?

- *Matrix calculus! Very cool stuff!*

$$\text{RSS}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta).$$

$$\mathbf{X}\beta = \hat{\mathbf{y}}$$

height =
 n
number
of rows



Finding the model

How do we minimize the residual sum of squares (RSS)?

- *Matrix calculus! Very cool stuff!*

$$\text{RSS}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta).$$

\mathbf{X} is the matrix of data, where each column is a variable
(including 1 as the first variable) and each row is a data point

β is the vector of linear weights from the model

\mathbf{y} is the vector of outputs

\mathbf{Z}^T is the transpose of \mathbf{Z} (the matrix is flipped over its diagonal)

Finding the model

How do we minimize the residual sum of squares (RSS)?

- *Matrix calculus! Very cool stuff!*

$$\text{RSS}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta).$$

How do we minimize it?

$$\frac{\partial \text{RSS}}{\partial \beta} = -2\mathbf{X}^T (\mathbf{y} - \mathbf{X}\beta) = 0$$

$$\frac{\partial^2 \text{RSS}}{\partial \beta \partial \beta^T} = 2\mathbf{X}^T \mathbf{X} \succ 0$$

Finding the model

How do we minimize the residual sum of squares (RSS)?

- *Matrix calculus! Very cool stuff!*

$$\text{RSS}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta).$$

How do we minimize it?

- *Don't worry about understanding this*
- *This is just to help you be less intimidated by the book*

$$\begin{aligned}\frac{\partial \text{RSS}}{\partial \beta} &= -2\mathbf{X}^T (\mathbf{y} - \mathbf{X}\beta) \\ \frac{\partial^2 \text{RSS}}{\partial \beta \partial \beta^T} &= 2\mathbf{X}^T \mathbf{X}.\end{aligned}$$

Finding the model

What's the optimum solution then?

Okay, great! That's what packages are for!

- *But also as CS people, it's worth knowing at some point*

Do we need anything else here?

- *We want to bound the likelihood that our weights are close to accurate*

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}.$$

↑
best/linear
model (minimize bias)

Bounding the expected error

This leads to our first way of quantifying error statistically!

- Notice that it depends on $N-p-1$. What is this? Degrees of freedom!
- Why do we need the constant $\frac{1}{N-p-1}$?

degrees of freedom

$$\hat{\sigma}^2 = \frac{1}{N-p-1} \sum_{i=1}^N (y_i - \hat{y}_i)^2.$$

Bounding the expected error

This leads to our first way of quantifying error statistically!

- *Notice that it depends on $N-p-1$. What is this? Degrees of freedom!*
- *Why do we need the constant $\frac{1}{N-p-1}$?*
 - This is to make our estimator unbiased
 - This is equivalent to degrees of freedom.
 - What if $N = 3$ and $P = 2$?

$$\hat{\sigma}^2 = \frac{1}{N - p - 1} \sum_{i=1}^N (y_i - \hat{y}_i)^2.$$

$X \in \mathbb{R}^p$

Bounding the expected error

This leads to our first way of quantifying error statistically!

- *Notice that it depends on $N-p-1$. What is this? Degrees of freedom!*
- *Why do we need the constant $\frac{1}{N-p-1}$?*
 - This is to make our estimator unbiased
 - This is equivalent to degrees of freedom.
 - What if $N = 3$ and $P = 2$?
 - There is only one “degree of freedom” only one point can vary from the line as drawn

$$\hat{\sigma}^2 = \frac{1}{N - p - 1} \sum_{i=1}^N (y_i - \hat{y}_i)^2.$$

makes this unbiased

Bounding the expected error

This leads to our first way of quantifying error statistically!

ε is the error between the value of y (from the model) and its reported value.

- We usually assume that this random error is normally distributed.
- What's an example of when it wouldn't be?

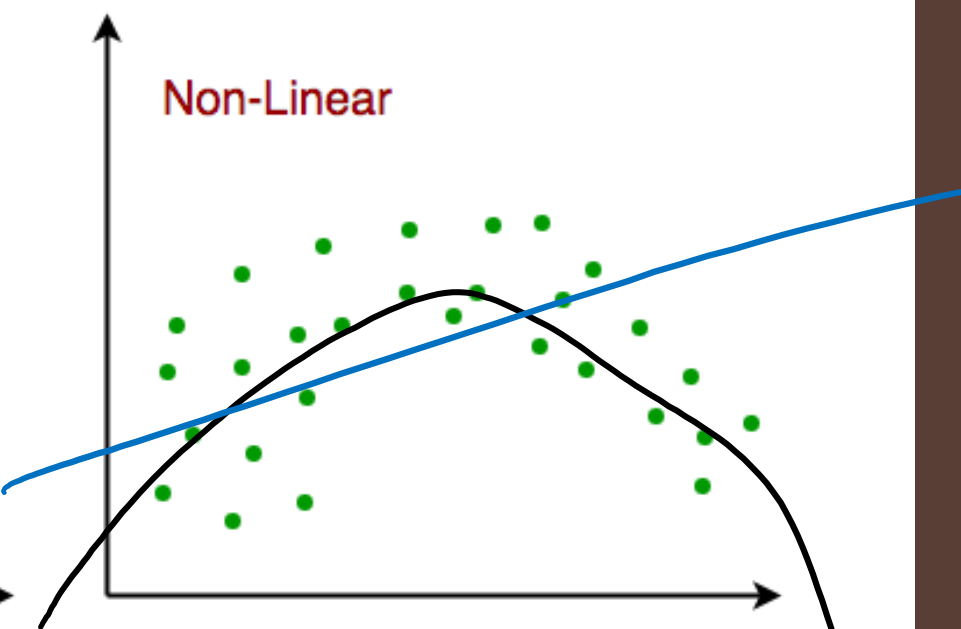
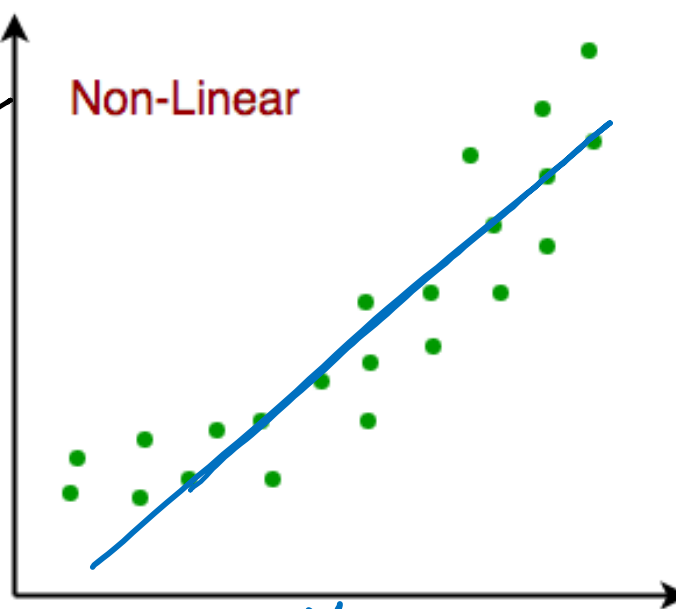
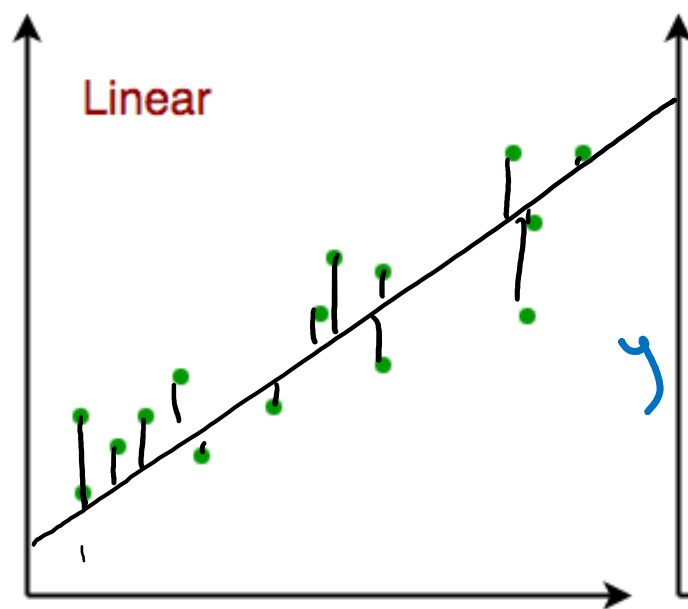
Because of β_0

$$\forall i \quad E(\varepsilon_i) = 0$$
$$E(\varepsilon_i^2) = \hat{\sigma}^2$$

$$Y = E(Y|X_1, \dots, X_p) + \varepsilon$$
$$= \beta_0 + \sum_{j=1}^p X_j \beta_j + \varepsilon,$$

$$\hat{\sigma}^2 = \frac{1}{N - p - 1} \sum_{i=1}^N (y_i - \hat{y}_i)^2.$$

the linear model assumes linear combination



distance from \hat{y}

ϵ is not independent
of X

Bounding the expected error

This leads to our first way of quantifying error statistically!

ε is the Gaussian error between the expected value of y (from the model) and its reported value

Each of these metrics can be calculated individually,
meaning you can use that to help determine
which variables you can eliminate from the model

Gauss-Markov Theorem

The model produced by the minimal RSS also has the lowest variance
(among unbiased estimators)

- A rare case of having our cake and eating it too!
- The tuning parameters here are the number of variables

Subset selection or ℓ_1 kernel expansion

Let $\tilde{\theta} = \mathbf{c}^T \mathbf{y}$ be some other linear estimator and consider the mean squared error (MSE)
per point

$$\begin{aligned}\text{MSE}(\tilde{\theta}) &= \mathbb{E}(\tilde{\theta} - \theta)^2 \\ &= \text{Var}(\tilde{\theta}) + [\mathbb{E}(\tilde{\theta}) - \theta]^2.\end{aligned}$$

$$E(\tilde{\theta} - \theta) = 0$$

Gauss-Markov Theorem

Let $\tilde{\theta} = \mathbf{c}^T \mathbf{y}$ be some other linear estimator and consider the mean squared error (MSE)

$$\begin{aligned} \text{MSE}(\tilde{\theta}) &= E[(\tilde{\theta} - \theta)^2] & E(Y_0 - \tilde{f}(x_0))^2 &= \sigma^2 + E(x_0^T \tilde{\beta} - f(x_0))^2 \\ &= \text{Var}(\tilde{\theta}) + [E(\tilde{\theta}) - \theta]^2. & &= \sigma^2 + \text{MSE}(\tilde{f}(x_0)). \end{aligned}$$

↑ which we calculated

↑ Squared error

What's going on here? Remember that σ^2 is the inherent variation of the output Y

- The right side equations here are for the variation at a particular point in y
- Notice that the variance at a particular point is a constant addition to the MSE
- Therefore, the function \tilde{f} has minimal variance at the point of minimal bias

*Variance is
some constant
distance from the
mean*

*↳ we can still introduce variance
with a more complex model
(kernel)*

Univariate model

All of this has been in terms of the arbitrary multivariate model

↔ matrix calculus

In the book, Section 3.2.3 p52-55 goes through this where $p = 1$,
or there is only one variable

You will not need to know this matrix algebra stuff,
but if you want to understand it completely,
this is a good vector-only introduction to the concepts

Subset selection

$$\beta_4 = 6$$

Since we can find the statistical impact of each variable
(or combination of variables using the F-statistic)

AND, since we pay a penalty (through degrees of freedom) for having more variables

We want to build our model on the most explanatory subset of variables

↳ helps w/ interpretability

also reduces chance for overfitting

Subset selection

Since we can find the statistical impact of each variable
(or combination of variables using the F-statistic)

AND, since we pay a penalty (through degrees of freedom) for having more variables

We want to build our model on the most explanatory value for our variables

- *This leaves us with two options*
 - *Best Subset*
 - *Shrinkage methods*

depend on how many features we want

Conclusion

This has been a really statistics heavy day, *don't expect to understand/memorize all this*

It's an important framework for how we'll be able to bound statistical regression models

The big takeaways here are:

- "Linear" models are not just straight lines
 - Least squares provides the best unbiased linear model
 - Statistical analysis of linear models is very well bounded
 - These models can still perform really well and they're **very** interpretable
- Handwritten notes:* $E(\hat{y}) = y$ (with an arrow pointing to the first bullet point) and t -tests (with an arrow pointing to the third bullet point).

Consider using linear models when:

- Data, either in points or in features is small
- Interpretability is important
- Want a tight statistical error bound (*this makes linear models a good benchmark when the Bayes decision boundary cannot be exactly calculated*)