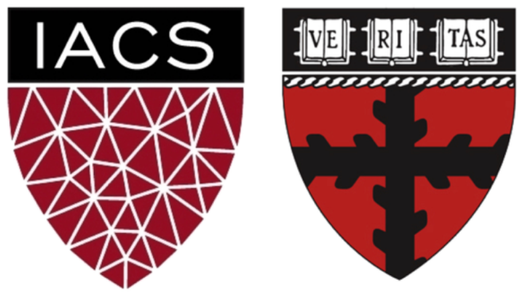
Lecture #1: Course Overview

AM 207: Advanced Scientific Computing

Stochastic Methods for Data Analysis, Inference and Optimization

Fall, 2020


IACS
VE RI TAS

Outline

1. What is this course about?
2. How is this course structured?
3. How do I get help for the course?

What is this course about?

How Do We Model Patterns in Data?

This is a scatter plot of home prices vs square footage of some homes in southern California.

Can you see any patterns or trends?

How Do We Model Patterns in Data?

We see that as **square footage** increases, so does **price**.

But what is a precise, mathematical description of this relationship?

What is a Model?

Building a model to capture a hypothesized relationship means we predict the value of one group of attributes using another group.

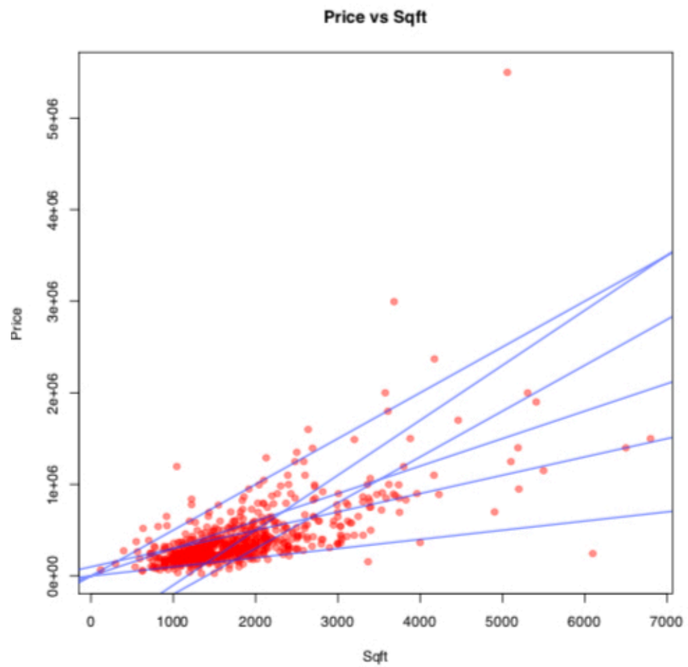
This prediction problem is called ***regression***, the attribute we are trying to predict (e.g.price) is called the ***outcome*** or the ***target***, denoted by y.

The group of attributes (e.g. square footage) we use to make the prediction is called the ***covariates***, denoted by x.

A ***regression model*** is a mathematical function, $f(x)$, that predicts the target. We denote our prediction by $\hat{y} = f(x)$.

What is a Model?

We conjectured that the model for this data is a line: $\hat{y} = f(x) = w_1 x + w_0$.

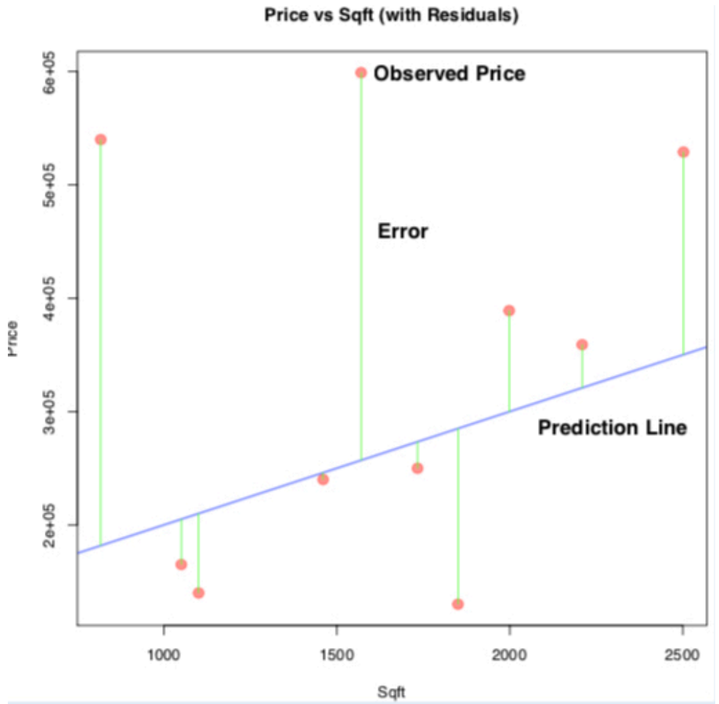


But which line fits the data best?

A Notion of Error

An ***absolute residual*** is the absolute difference between the actual price of a home and the price predicted by the line for a given square footage:

$$\texttt{Residual}_n = y_n - \hat{y}_n$$



How do we quantify the overall error?

1. **(Max absolute deviation)** Count only the biggest "error"
$$\max_n |y_n - \hat{y}_n|$$
2. **(Sum of absolute deviations)** Add up all the "errors"
$$\sum_n |y_n - \hat{y}_n|$$
3. **(Sum of squared errors)** Add up the squares of the "errors"
$$\sum_n |y_n - \hat{y}_n|^2$$
4. **(Mean squared errors)** We can also average the squared "errors".
$$\frac{1}{N} \sum_{n=1}^{N} |y_n - \hat{y}_n|^2$$

Again, y_n is the observed target, $\hat{y}_n$ is the predicted target.

Model Fitting

Question: What do we mean by choosing "best" line, $\hat{y} = w_1 x + w_0$?

The ***model fitting*** process:

1. *Choose* an overall error metric. This metric is called the ***loss function*** or ***training objective***:

$$\mathcal{L}(w_1, w_0) = \frac{1}{N} \sum_{n=1}^{N} |y_n - (w_1 x_n + w_0)|^2, \qquad \text{(Mean Squared Error Loss)}$$

2. Set up the problem of finding coefficients or ***parameters***, w_0, w_1, such that the loss function is **minimized**:

$$\underset{w_0, w_1}{\text{argmin}} \mathcal{L}(w_1, w_0) = \underset{w_0, w_1}{\text{argmin}} \frac{1}{N} \sum_{n=1}^{N} |y_n - (w_1 x_n + w_0)|^2$$

3. Choose a method of minimizing the loss function.

Note: For linear regression, we can minimize $\mathcal{L}$ analytically. We cannot do this for every model!

What is a Statistical Model?

Perhaps our **choice** of an overall error can be less arbitrary if we explain how, we believe, the residual arise.

Belief: The theoretical relationship between price and square footage (x) is given by $f(x)$. But, in real-life, due to unpredictable circumstances observed prices (y) differ from $f(x)$ by some random amount, ϵ, called ***noise***:

$$y = f(x) + \epsilon, \qquad \epsilon \sim p(\epsilon)$$

A ***statistical model*** is one that explicitly accounts for uncertainty or randomness.

A Statistical Model for Regression

Let us *assume* that (1) the underlying relationship between price and square footage x is given by $f(x) = w_1 x + w_0$; (2) that the observed price y deviates from $f(x)$ by a random amount that is independent from x and is distributed as $\mathcal{N}(0, 1)$:

$$y = f(x) + \epsilon, \qquad \epsilon \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$$

Note that y is now a random variable with distribution $\mathcal{N}(f(x), 1)$, denoted by $p(y|x, w_1, w_0)$.

How Do We Quantify Fitness?

Given our statistical model, a natural way for quantifying how well $f(x) = w_1 x + w_0$ fits the data is by checking how likely our choice of w_0 and w_1 makes the observed data, i.e. compute

$$\mathcal{L}(w_1, w_0) = \prod_{n=1}^{N} p(y_n | x_n, w_1, w_0).$$

The function $\mathcal{L}(w_1, w_0)$ is called the ***likelihood function***.

Exercise: suppose we have two models, $f(x) = 3x + 2$ and $f(x) = 10 - x$. Suppose that $\mathcal{L}(w_1 = 3, w_0 = 2) = 10.2$ and $\mathcal{L}(w_1 = -1, w_0 = 10) = 0.002$. Which model is a better fit for the data and why?

Model Fitting

Question: What do we mean by choosing "best" line, $\hat{y} = f(x) = w_1 x + w_0$?

The ***model fitting*** process:

1. *Choose* a method of estimation for statistical models. For example, set up the problem of finding coefficients or ***parameters***, w_0, w_1, such that the likelihood of the data is **maximized**:

$$\text{argmax}_{w_0,w_1} \mathcal{L}(w_1, w_0) = \text{argmax}_{w_0,w_1} \prod_{n=1}^{N} p(y_n | x_n, w_1, w_0)$$

2. Choose a method of computing the estimate. For example, choose a way to maximize the likelihood.

Maximimum Likelihood and Minimum Mean Square Error

Given our statistical model

$$y = f(x) + \epsilon, \qquad \epsilon \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$$

Maximimizing the likelihood is equivalent to minimizing the mean squared error:

$$\text{argmax}_{w_0, w_1} \prod_{n=1}^{N} p(y_n | x_n, w_1, w_0) \equiv \text{argmin}_{w_0, w_1} \frac{1}{N} \sum_{i=1}^{N} |y_i - (w_1 x_1 + w_0)|^2$$

Hint: note that

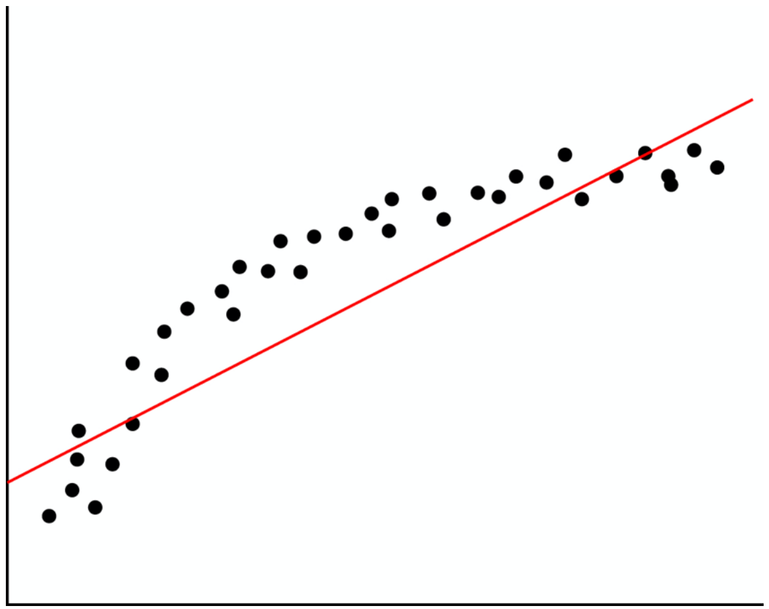
$$\prod_{n=1}^{N} p(y_n | x_n, w_1, w_0) = \frac{1}{\sqrt{2\pi 1}^N} \exp\left\{ -\frac{\sum_{i=1}^{N} (y_n - (w_1 x_n + w_0))^2}{2 * 1} \right\}$$

and that

$$\log p(y | x, w_1, w_0) = N \log\left(\frac{1}{\sqrt{2\pi 1}} \right) - \frac{\sum_{i=1}^{N} (y_n - (w_1 x_n + w_0))^2}{2 * 1}$$

Model Evaluation

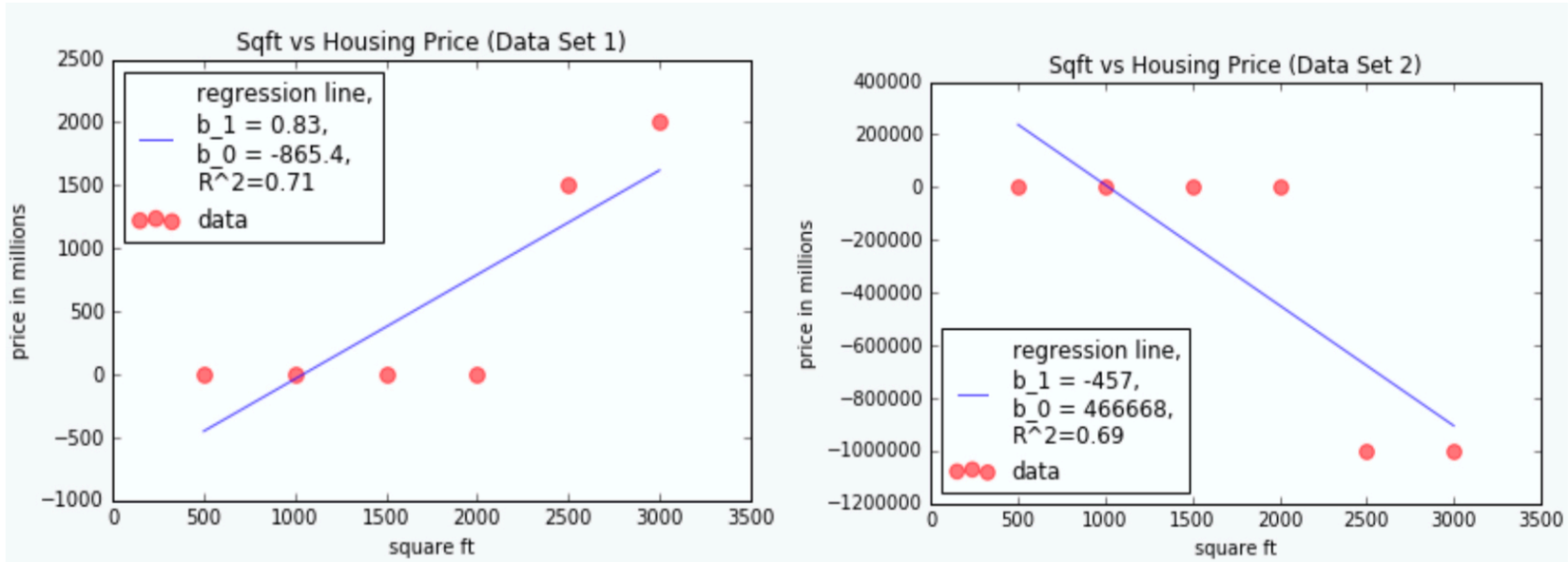
After fitting the model (finding coefficients that maximizes the likelihood or that minimizes the loss function), we need to **check the error or residuals of the model**. Why?



Working with statistical models gives us an advantage in model evaluation, can you see why?

Model Interpretation

In addition to evaluating our model on training and testing data, we must also examine the coefficients themselves. Why?



What is a Bayesian Model?

In addition to a statistical model that explains trends $f(x)$ and observation noise ϵ, we also want to incorporate our **prior beliefs** about the model. Finally, we want to obtain a measure of **uncertainty** for our parameter estimates as well as our predictions.

Our Bayesian model for linear regression:

$$
\begin{aligned}
y &= w_1 x + w_0 + \epsilon \\
\epsilon &\overset{\text{iid}}{\sim} \mathcal{N}(0, 1) \\
w_1 &\sim p(w_1) \\
w_0 &\sim p(w_0)
\end{aligned}
$$

where the prior $p(w_i)$ may express that we want w_i to be non-negative and not too large.

Model Inference

How do we "learn" the parameters in a Bayesian model?

Baye's Rule gives us a way to obtain a distribution over w_1, w_0 given the data $(x_1, y_1), \ldots, (x_N, y_N)$:

$$p(w_1, w_0 | x_1, \ldots, x_N, y_1, \ldots, y_N) \propto \underbrace{\left(\prod_{n=1}^{N} p(y_n | x_n, w_1, w_0) \right)}_{\text{How well params fit the data}} \underbrace{p(w_1)p(w_0)}_{\text{How well the params fit priors}}$$

The distirbution $p(w_1, w_0 | x_1, \ldots, x_N, y_1, \ldots, y_N)$ is called the ***poseterior*** and gives the likelihood of a pair of parameters w_1, w_0 given the observed data.

We see that the likelihood score of the parameters under the posterior is influence both by how well the parameters fit the data and how well the parameters fit our prior beliefs.

Bayesian Linear Regression

When we choose normal priors for the parameters in a linear regression model, for example,

$$
\begin{aligned}
y &= w_1 x + w_0 + \epsilon \\
\epsilon &\overset{\text{iid}}{\sim} \mathcal{N}(0, 1) \\
w_1 &\sim \mathcal{N}(0, 0.5) \\
w_0 &\sim \mathcal{N}(0, 1)
\end{aligned}
$$

The posterior $p(w_1, w_0 | x_1, \ldots, x_N, y_1, \ldots, y_N)$ is again a (multivariate) normal distribution, $\mathcal{N}(\mu, \Sigma)$, and we can derive closed form solutions for μ and Σ.

Why is this observation important?

Bayesian versus Frequentist Uncertainty

The main advantage of the Bayesian approach is that rather than obtaining a single "best" estimate of the model parameters, the posterior gives us a distribution over a set of plausible model parameters (with some models being more likely than others).

The spread of this distribution over plausible models naturally gives us a way to quantify our **uncertainty** over which is the "best" model. When the spread is wide (when many very different models are equally very likely), our uncertainty is high. When the spread is narrow (when all likely models look very similar), our uncertainty is low.

We can also obtain a sense of uncertainty over models using the non-Bayesian probabilistic model. Typically, we randomly sample sets of training data point from the training data, on each set, we compute the MLE of the model parameters. This process is called **bootstrapping**.

In HW#0 you will compare the uncertainties from the posteriors of Bayesian models with those from bootstrapping maximum likelihood models.

Model Evaluation

With a Bayesian model we get a distribution $p(w_1, w_0|\text{Data})$ over likely functions rather than a single function $f(x) = w_1 x + w_0$. How then do we evaluate the "error" of model?

In the Maximum Likelihood model, we can explicitly check the correctness of our assumptions by checking the distribution of the residuals. How do we criticize a Bayesian model?

Why is This Hard?

1. Stating that our goal is to maximize likelihood or minimize MSE is easy. Finding the optimal parameters is often very hard (especially if $f(x)$ is not linear, but rather, a complex function represented by a neural network).
2. If we choose more "interesting" or "expressive" priors, or if we choose more complex $f(x)$, then it is often the case that the posterior cannot be computed in closed form.

Both model fitting and inference requires sophisticated algorithms derived from deep theoretical understanding of the models.

What is AM207?

1. Build statistical (Bayesian and non-Bayesian) models for: continuous, ordinal and categorical data
2. Study algorithms for model fitting and inference
3. Study paradigms for model evaluation and critique

Goal: students become familiar with standard statistical models and modern techniques of inference. At the end of the course you should be able to productively read current machine learning research papers.

Focus:

- **Why:** theory should serve a concrete purpose.
- **How:** emphasize computational aspects of inference.

Related Courses: Bayesian Inference (Stats), Advanced Machine Learning (CS), Computational Statistics (Stats)

What technologies do you need for this class?

For Virtual Meetings

- Class meetings will happen over Zoom.
- Office hours will happen in a gather.town room.
- All students should connect to class meetings and office hours with an iPad (note that gather.town does not load on an iPad) or drawing pad, if you do not have access to one, contact the instructor as soon as possible.
- For collaborative white-boarding please familiarize yourselves with either ZiteBoard or Miro.

For Completing Assignments

Homework will be completed in Jupyter Notebooks.

You have one of two options

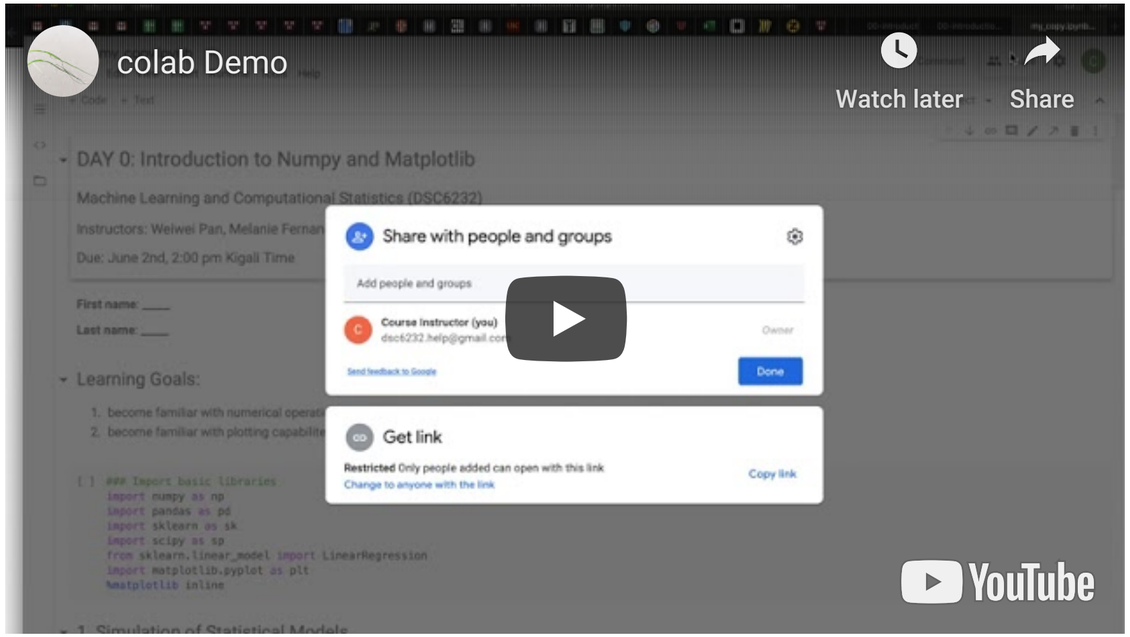
1. Download the latest Anaconda `python` 3.x distribution on your personal machine
2. Complete homework using Google Colab - a free cloud computing service that comes with pre-installed machine learning tools. Colab is built on Jupyter Notebooks, an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text.

For In-Class Exercises

In class exercises are completed in Google Colab. Save a copy of the starter code to your Google Drive. Turn on "Sharing" for your notebook and submit the share link to the Course Canvas. Each group submits a single notebook.

In [3]:
```
HTML('<iframe width="560" height="315" src="https://www.youtube.com/embed/Dvv-gEKd_Jc?rel=0&amp;controls=0&amp;showinfo=0" frameborder="0" allowfullscreen></iframe>')
```

Out[3]:



How is this course structured?

Graded Components

1. In-class group exercises
2. 10 equally weighted weekly homework
3. 1 group project

Each homework will be a combination of derivations/proofs (theory) and programming (implementation).

The group project involves choosing one pre-approved research paper and producing a tutorial in Jupyter Notebook to demonstrate the concepts and methodologies in the paper.

Policies

Grading: Unreadable formatting or code with syntactic or runtime errrors will not be graded. "Right" answer without a (brief) justification will not receive full score. You can drop your lowest HW grade.

Late HW: Each student has **three** late days that can be applied to any one or two homework. Outside of late days, late submissions will not be accepted.

Collaboration: Collaboration is strongly encouraged, but copying is strictly not allowed (see policy on Syllabus).

Attendance: Attendance of class meetings is required. Attendance waivers can be obtained by contacting the instructor.

How do I get help for the course?

Teaching Staff

Instructor: Weiwei Pan

TFs:

- Ruby Zhang: Saturday 9:30am - 11:00am
- Rylan Schaeffer: Saturday 4:00pm - 5:30pm
- Jonathan Chu: Saturday 8:00pm - 9:30pm
- Qiuyang Yin: Sunday 9:30am - 11:00am
- Lin Zhu: Sunday 3:30pm - 5:00pm
- Théo Guenais: Monday 10:30am-12:00pm
- Dimitris Vamvourellis: Monday 3:30pm-5:00pm
- Cooper Lorsung: Tuesday 10:30am-12:00pm
- Hari Kothapalli: Tuesday 3:30pm-5:00pm
- Jiayu Yao: Wednesday 9:00am-10:30am
- Michael Downs (Grading TF)
- Yaniv Yacoby (Project TF)
- Kela Roberts (Extension School)

TF Office Hours

Students need to submit their questions as comments on the appropriate Piazza post ***prior to each TF office hour***. During each OH, similar questions will be consolidated and answered in the order they were submitted.

Each OH has a specific focus, questions that are not aligned with the focus of the session will be given lower priority. For example, on Monday, questions about how to get started on homework problems will be prioritized lower than trouble-shooting questions on solutions in progress.

- **(Saturday)** Focus: background concepts and homework problem setup
- **(Sunday)** Focus: homework problem setup and trouble-shooting
- **(Monday)** Focus: trouble-shooting and interpretation
- **(Tuesday)** Focus: interpretation, trouble shooting
- (Wednesday) Focus: interpretation

Instructor Office Hours

The focus of TF office hours is on providing support for homework assignments. The focus on instructor office hours is to support general understanding of the class material.

While you are welcome to bring questions about homework assignments to instructor office hours, you may get more out of your total face-time with staff if you can make use of both sets of office hours.

Suggested workflow:

- **(Friday)** clarify concepts covered during the week during instructor office hour
- **(Saturday and/or Sunday)** setup all homework problems during TF office hours
- **(Monday and/or Tuesday)** trouble-shoot implementation issues during TF office hours
- **(Wednsday)** discuss interpretation during instructor office hour

Piazza

There is a course Piazza to faciliate collaboration amongst students.

Teaching staff moderate the discussions but are **not responsible for answering questions**!

If you want help from the staff come to an office hour or schedule a meeting.

Extension Students

The dedicated TF for extension school students is Kela Roberts, who will answer questions via email (see email policy on Syllabus).

Extension students are welcome to attend scheduled FAS TF office hours and work with FAS students in teams. However, since FAS TFs are allotted based on FAS enrollment numbers, FAS TFs will prioritize answering questions from FAS students.

Final Words of Advice

How We've Changed AM207

When we learned that AM207 will be remote, we made a number of fundamental changes to maximize the time you have to interact with each other and staff (this has always been the main way learning happens in this class):

1. We've doubled the sections to keep enorllment in each section small
2. We've doubled OHs
3. We've flipped the classroom, so that class time is interactive and active
4. We've adopted a number of new tech to facilitate collaboration and team work

What We are Asking From You

We are asking you also put in work in order to make this a successful learning experience:

1. Come to office hours
2. Ask questions:
    - **Ask questions to understand**. There is no such thing as an obvious fact or a trivial question. Don't let shyness of intimidation prevent you from asking for help to understand something.
    - **Ask questions to dig deeper**. Every single concept in this course serves a purpose and has a justification. Don't settle for knowing facts, there's always a questions you can ask about something you already know that will show you something new and something deep.
3. Focus on creating connections, relation between and syntheses of concept. Don't worry about memorizing lines of math.

How to Work in Teams

Most of the work you do in AM207 you will do in teams. Since this is a traditionally diverse class in terms of backgrounds, your teammates will likely not have the same outlook and expertise as you.

1. **When you're the one in the know** If you find a section of the material easy, don't settle for just doing the work for your team! Your challenge in this case is to teach, find a way to bring your teammates to your level of understanding.
2. **When you're the one in the dark** If you find yourself lost on a task and someone else seems to be taking the lead, don't settle into the "back-seat"! Your challenge in this case is to ask good critical questions and interrogate the validity of every "answer" being proposed.
3. **When you disagree** Solicit everyone's opinion, take time to understand what they are saying, be open to discussions (be able to suspend your own skepticism). When you can't reconcile difference, come to us and we can continue the discussion.