Seoul National University
Dept. of Civil and Environmental Engineering

Instructor: Junho Song
junhosong@snu.ac.kr

457.212 Statistics for Civil & Environmental Engineers

In-Class Material: Class 25

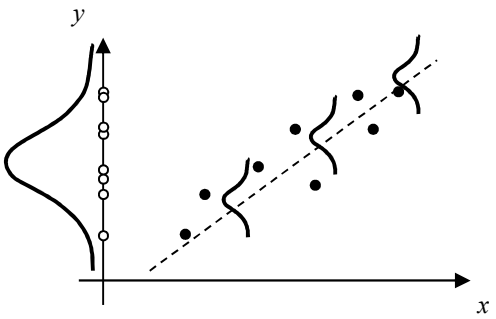
Regression Analysis (A&T: 8.2-8.7)

Given: Sample data set $\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$

Question: The functional relation between two random variables X and Y? $Y = f(X)$

→ "Regression" Analysis

1. Regression & Conditional Mean



(a) Marginal and conditional standard deviation of Y:

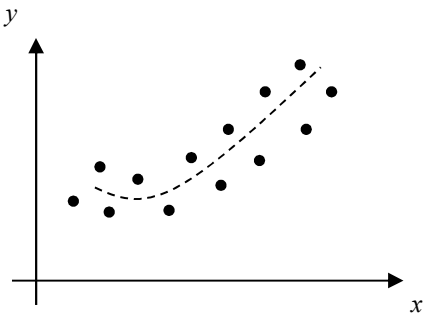
σ_Y $\sigma_{Y|x}$

(b) Marginal and conditional mean of Y:

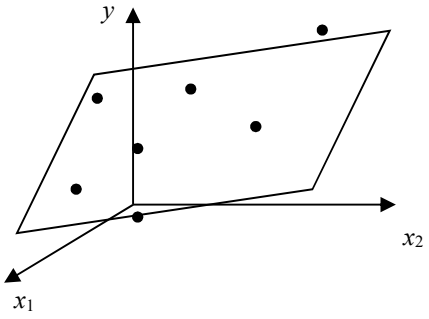
$\mathrm{E}[Y] =$ constant.
$\mathrm{E}[Y \mid x] = f(x)$

→ Conditional mean predicts the outcome of Y more accurately (i.e. smaller variation).
→ Regression analysis aims at finding the functional relationship for the conditional mean to describe the hidden relation between X and Y.

(c) Linear vs. nonlinear regression

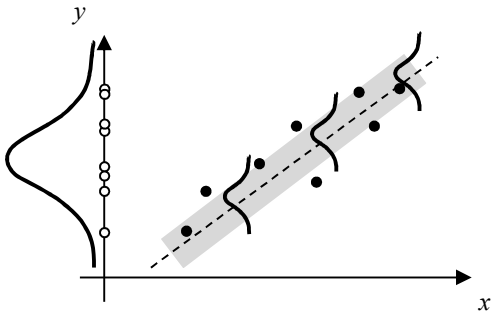


(d) Single vs. multiple regression



Instructor: Junho Song
junhosong@snu.ac.kr

2. Single Linear Regression with Constant (Conditional) Variance



(a) Assumption: the conditional mean is a linear function of x and the conditional variance is constant, i.e.

$\mathrm{E}[Y \mid x] = \alpha + \beta x$ and $\sigma^2_{Y|x} = \text{const.}$

"Linear regression of Y on X"

(b) Estimation of α and β

"Best" estimates on α and β: $\hat{\alpha}$ and $\hat{\beta}$ ~ the values minimizing the sum of squared errors between the prediction by the linear relationship ($y_i' = \alpha + \beta x_i$) and the given data point y_i (least square estimators)

Sum of Squared Errors (SSE):

$$\Delta^2 = \sum_{i=1}^{n} (y_i - y_i')^2$$
$$= \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

Note: The same weight is given to each data point because the conditional variance is assumed to be constant.

Find α and β that minimize SSE → Solve the following equations for α and β:

$$\frac{\partial \Delta^2}{\partial \alpha} = 2\sum_{i=1}^{n} (y_i - \alpha - \beta x_i)(-1) =$$
$$\frac{\partial \Delta^2}{\partial \beta} = 2\sum_{i=1}^{n} (y_i - \alpha - \beta x_i)(-x_i) =$$

Instructor: Junho Song
junhosong@snu.ac.kr

As a result,

$$\hat{\beta} = \frac{\sum (x_i \cdot y_i) - n \cdot \bar{x} \cdot \bar{y}}{\sum x_i^2 - n \cdot \bar{x}^2}$$
$$\hat{\alpha} = \bar{y} - \hat{\beta} \cdot \bar{x}$$

Need: $\sum x_i y_i$, $\sum x_i$, $\sum y_i$ and $\sum x_i^2$

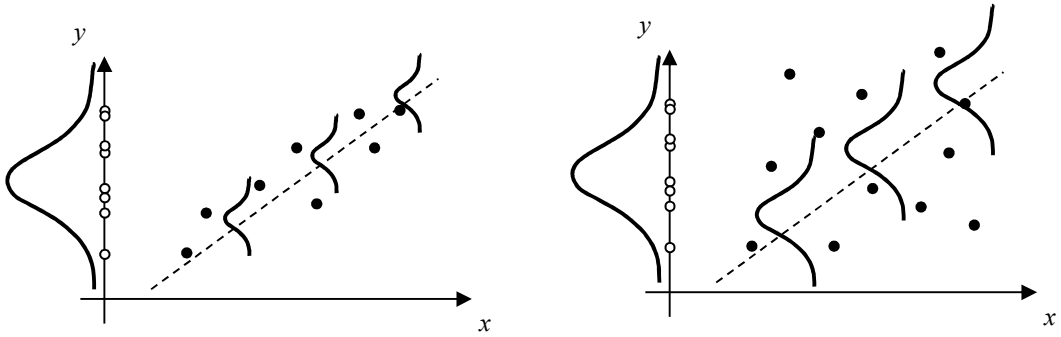
(c) $\sigma_{Y|x}^2$?

Estimated as

$$s_{Y|x}^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - y_i')^2$$
$$= \frac{\Delta^2}{n-2}$$

(d) Reduction of variance: from marginal $\sigma_Y^2 (s_Y^2)$ to conditional variance $\sigma_{Y|x}^2 (s_{Y|x}^2)$?

→ A measure of the strength of the linear relationship



$$r^2 = \frac{s_Y^2 - s_{Y|x}^2}{s_Y^2} = 1 - \frac{s_{Y|x}^2}{s_Y^2}$$

$r^2 \cong 0$: No reduction (weak linear relationship)

$r^2 \cong 1$: Large reduction (strong linear relationship)

Note: $r^2 \cong \rho_{XY}$ as $n \to \infty$

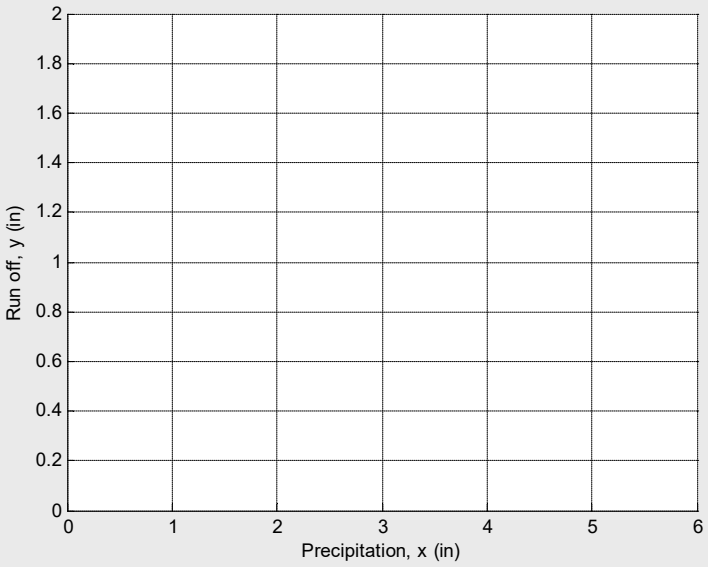
Seoul National University
Dept. of Civil and Environmental Engineering

Instructor: Junho Song
junhosong@snu.ac.kr

Example 1: Regression analysis of Runoff (*Y*) on Precipitation (*X*)

	x_i (in.)	y_i (in)	$x_i y_i$	x_i^2	y_i^2	y_i'	$(y_i - y_i')^2$
	1.01	0.30	0.303	1.0201	0.09		
	2.09	0.95	1.9855	4.3681	0.9025		
	3.57	1.59	5.6763	12.7449	2.5281		
	5.11	1.74	8.8914	26.1121	3.0276		
	2.93	1.12	3.2816	8.5849	1.2544		
Sum							
Avg							

(a) Scatter plot?



(b) Linear regression of Y on X
(i.e. Find $\hat{\alpha}$ and $\hat{\beta}$)?

$$\hat{\beta} = \frac{\sum (x_i \cdot y_i) - n \cdot \bar{x} \cdot \bar{y}}{\sum x_i^2 - n \cdot \bar{x}^2}$$

$=$

$$\hat{\alpha} = \bar{y} - \hat{\beta} \cdot \bar{x} =$$

Thus, $\mathrm{E}[Y \mid x] =$

Show it in the plot.

(c) Estimate on the conditional variance, $s_{Y|x}^2 = \frac{\Delta^2}{n-2} =$

(d) Estimate on the marginal variance, $s_Y^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right) =$

(e) Reduction ratio, $r^2 = 1 - \frac{s_{Y|x}^2}{s_Y^2} =$

(f) Suppose the precipitation is 4.0 in.
What is the mean run off?

Probability that the run-off exceeds 2 in.?

Instructor: Junho Song
junhosong@snu.ac.kr

(e) **Confidence interval** on the estimated conditional mean $y'(x) = \mathrm{E}[y|x] = \hat{\alpha} + \hat{\beta}x$

Due the statistical uncertainty in the parameters $\hat{\alpha}$ and $\hat{\beta}$, it is known that

$$y'(x) \sim N\left(\hat{\alpha} + \hat{\beta}x, \sigma^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)\right)$$

Since σ, i.e. the (constant) standard deviation of the residual $(y' - y_i)$ is unknown, y' should follow *t*-distribution. Thus $(1-\alpha)\times 100(\%)$ confidence interval on the conditional mean is given as

$$y'(x) \pm t_{\frac{\alpha}{2},n-2} \cdot s\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

where s is the "standard error," which is defined as $s = \sqrt{s_{y|x}^2} = \sqrt{\frac{\Delta^2}{n-2}}$

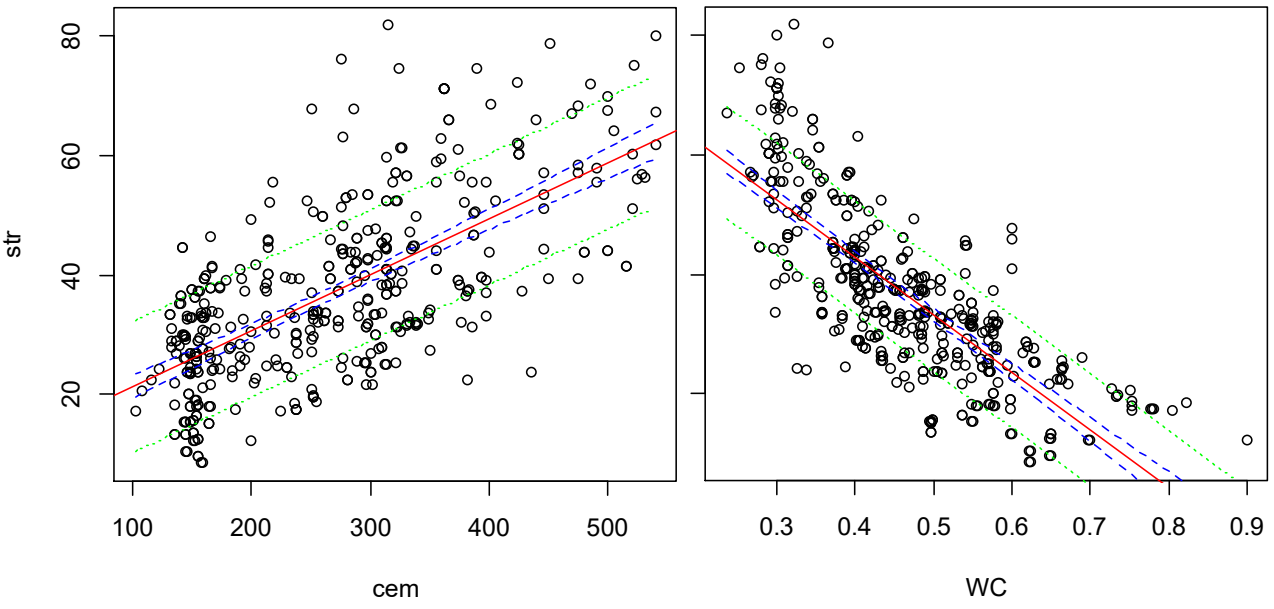
(f) **Prediction interval** on $y(x) = \alpha + \beta x + \epsilon$

Due to the uncertainty of the residual ϵ, the confidence interval on the dependent variable y at x is derived as

$$y'(x) \pm t_{\frac{\alpha}{2},n-1} \cdot s\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

Example 2: Using the dataset 'concrete.txt' and R, perform linear regression analysis of the concrete strength on the cement amount, and on the water-cement ratio, respectively.

(Data provided by Prof. Juhyuk Moon at Multi-scale Structural Materials Lab)



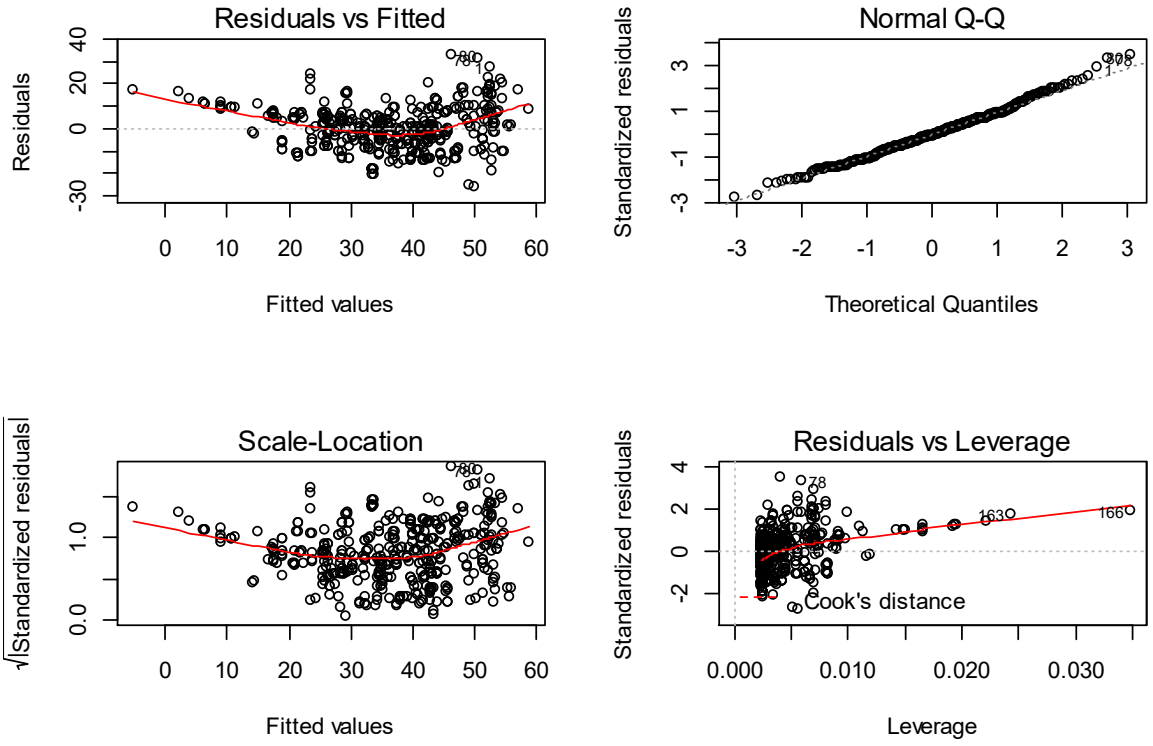
Instructor: Junho Song
junhosong@snu.ac.kr

```
rm(list=ls())
concrete=read.table("concrete.txt", header=TRUE)

str = concrete$Strength # Concrete strength
cem = concrete$Cement # Cement amount

# Calculate Water-Cement Ratio
concrete[,6]=concrete[,4]/(concrete[,1]+concrete[,2]+concrete[,3])
colnames(concrete)[6] = "WCratio"
WC = concrete$WCratio

# [Regression 1]
# Linear regression of strength (y) on cement amount (x)
reg1 = lm(str~cem)

# Plot
par(mfrow=c(1,1))
plot(str~cem)
abline(reg1,col='red')
# or alternatively by... lines(cem, fitted(reg1), col='red')

# Summary of regression results
coef(reg1)
summary(reg1)

# Diagnosis
par(mfrow = c(2, 2))
# or alternatively by ... layout(matrix(1:4, 2, 2))
plot(reg1)

# Confidence interval and prediction interval
cem_v = seq(min(cem),max(cem),5) # create a vector for cement amount
# Confidence interval (95%) by uncertaint regression coeff
str_ci = predict(reg1, data.frame(cem = cem_v),interval =
c("confidence"),level=0.95)
# Prediction interval (68%) by SSE, i.e. mean+/-SD
str_pi = predict(reg1, data.frame(cem = cem_v),interval =
c("prediction"),level=0.68)

# Plot with CI and PI
par(mfrow=c(1,1))
plot(str~cem)
abline(reg1,col='red')
lines(cem_v,str_ci[,2],col='blue',lty=2)
lines(cem_v,str_ci[,3],col='blue',lty=2) # blue curves for CI
lines(cem_v,str_pi[,2],col='green',lty=3)
lines(cem_v,str_pi[,3],col='green',lty=3) # green curves for PI

# Other results from 'lm' object:
# fitted(reg1): fitted results at the data points
# resid(reg1): residuals at the data points
# predict(reg1,...): predict using the regression model
predict(reg1,data.frame(cem=c(100,200,300,400,500))) # use the same name
'cem'

# [Regression 2]
# Linear regression of strength (y) on WC ratio (x)
reg2 = lm(str~WC) # linear regression

# Plot
par(mfrow=c(1,1))
plot(str~WC)
abline(reg2,col='red')
```

Instructor: Junho Song
junhosong@snu.ac.kr

```
# Summary of regression results
coef(reg2)
summary(reg2)

# Diagnosis
par(mfrow = c(2, 2))
plot(reg2)

# Confidence interval and prediction interval
WC_v = seq(min(WC),max(WC),0.05)
str_ci = predict(reg2, data.frame(WC = WC_v),interval =
c("confidence"),level=0.95)
str_pi = predict(reg2, data.frame(WC = WC_v),interval =
c("prediction"),level=0.68)

# Plot with CI and PI
par(mfrow=c(1,1))
plot(str~WC)
abline(reg2,col='red')
lines(WC_v,str_ci[,2],col='blue',lty=2)
lines(WC_v,str_ci[,3],col='blue',lty=2)
lines(WC_v,str_pi[,2],col='green',lty=3)
lines(WC_v,str_pi[,3],col='green',lty=3)

# [Regression 3]
# Multiple linear regression of strength (y) on cement amount (x1) and WC
ratio (x2)
reg3 = lm(str~cem+WC) # linear regression of str on cem and WC
summary(reg3)
```

*** Diagnosis plots for Regression 2**



Instructor: Junho Song
junhosong@snu.ac.kr

3. Multiple Linear Regression

"Linear regression of Y on $X_1, \ldots, X_m$"

(a) Define Δ^2 by assuming $\sigma_{Y|x}^2 = \sigma^2$ (constant) or $\sigma_{Y|x}^2 = \sigma^2 g^2(x_1,...,x_m)$ (non-constant)

(b) Find

$$\mathrm{E}[Y \mid x_1,...,x_m] = \beta_0 + \beta_1 x_1 + ... + \beta_m x_m$$

(c) Estimate $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_m$ by solving

$$\frac{\partial \Delta^2}{\partial \beta_0} = \frac{\partial \Delta^2}{\partial \beta_1} = \cdots = \frac{\partial \Delta^2}{\partial \beta_m} = 0$$

(d) $s_{Y|x_1,...,x_m}^2 = \dfrac{\Delta^2}{n-m-1}$

(**Note**: $m = 1$ for single linear regression)

4. Nonlinear Regression & Applications of Regression Analysis (Read A&T 8.6-8.7)

5. Correlation Analysis

(a) (True or theoretical) correlation coefficient

$$\rho_{XY} = \frac{Cov[X,Y]}{\sigma_X \sigma_Y} = \frac{\mathrm{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} = \frac{\mathrm{E}[XY] - \mu_X \mu_Y}{\sigma_X \sigma_Y}$$

(a) Unbiased estimator of ρ_{XY}, $\hat{\rho}$

$$\hat{\rho} = \frac{1}{n-1} \frac{\sum_{i=1}^{n} x_i y_i - n\bar{X}\bar{Y}}{s_x s_y}$$

(b) $\hat{\rho}$ and $\hat{\beta}$

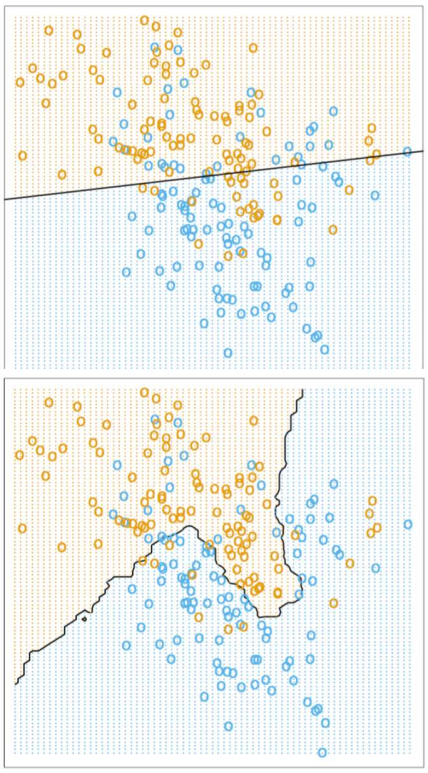
$$\hat{\rho} = \hat{\beta} \frac{s_X}{s_X} = \frac{\Sigma(x_i - \bar{X})(y_i - \bar{Y})}{(n-1)s_X^2} \frac{s_X}{s_Y} = \frac{\Sigma(x_i - \bar{X})(y_i - \bar{Y})}{\Sigma(x_i - \bar{X})^2} \frac{s_X}{s_Y} = \hat{\beta} \frac{s_X}{s_Y}$$

(d) $\hat{\rho}^2$ and $r^2 = 1 - s_{Y|x}^2 / s_Y^2$

$$\hat{\rho}^2 = 1 - \frac{n-2}{n-1} \frac{s_{Y|x}^2}{s_Y^2}. \quad \text{As } n \to \infty,\ \hat{\rho}^2 \to 1 - \frac{s_{Y|x}^2}{s_Y^2} = r^2$$

Instructor: Junho Song
junhosong@snu.ac.kr

6. "Model-based" vs "Data-based" prediction



(a) Model-based prediction: assumes a smooth model and fit

e.g. Linear regression $Y = \beta_0 + \sum_{i=1}^{N} \beta_i x_i$

- maybe inaccurate, but stable

(b) Data-based prediction (interpolation): does not assume a model, just interpolate from adjacent data points

e.g. *k*-nearest neighbor model $Y = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i$

- Accurate, but may be unstable

7. Statistical/Machine Learning

Build a prediction model by (1) Clustering, (2) Classification, and (3) Regression

➔ Will be covered by the new undergraduate course "**457.310 Information Engineering for CEE Engineers**" from Spring 2019) along with other machine learning techniques.