Statistics Recap

M. Fuat Kına

Statistics recap

- Descriptive statistics
    - Collecting, presenting, and describing data

- Inferential statistics
    - Drawing conclusions and/or making decisions concerning a population based only on sample data

- A **Population** is the set of all items or individuals of interest.
- A **Sample** is a subset of the population.

Statistics recap

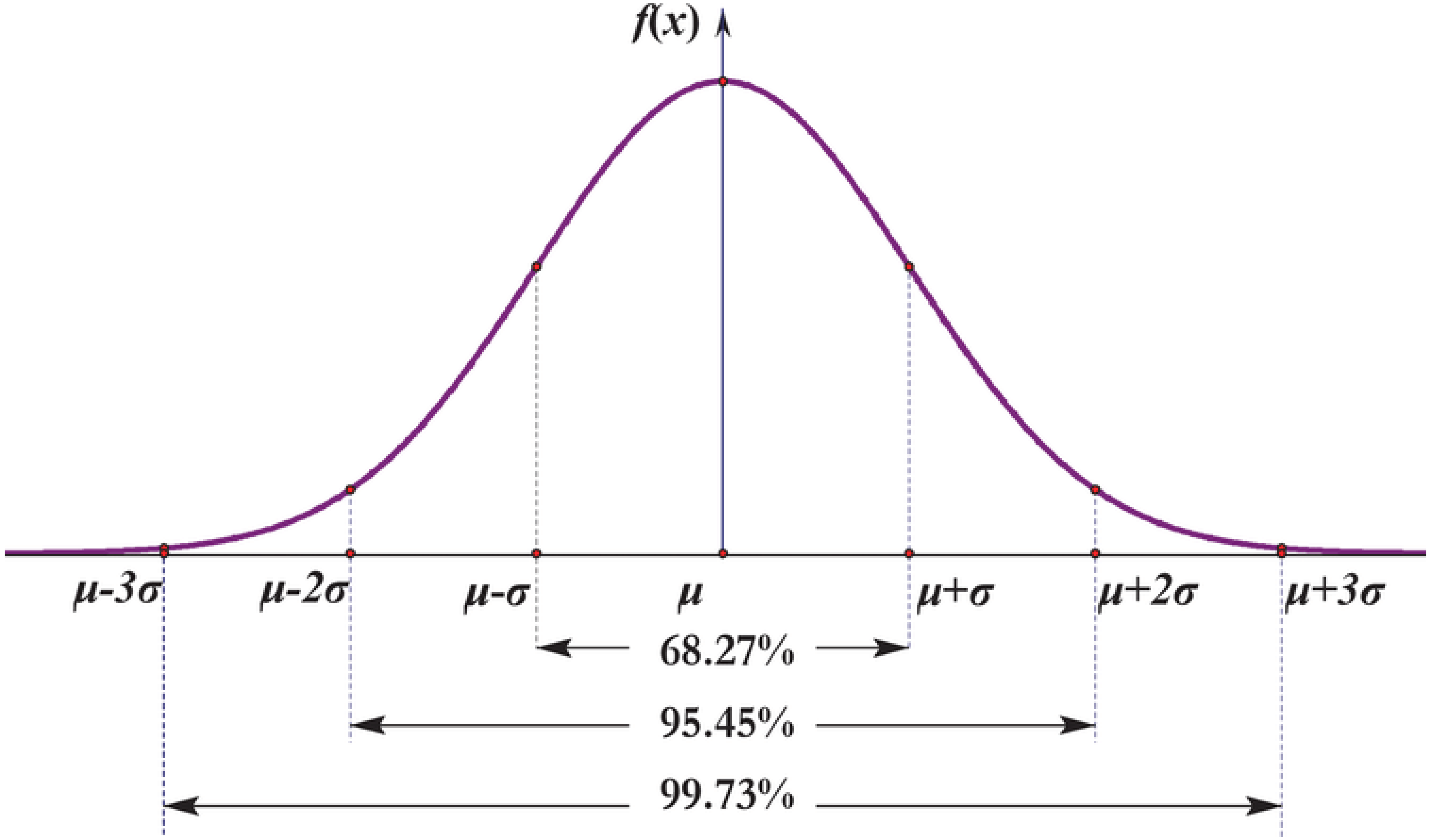
- **Univariate**: mean, mode, median, quartiles, range, standard deviation, variance, frequency, etc.
- **Bivariate**: crosstabs, t-test, anova, correlations, etc.
- **Multivariate**: regressions, etc.

- Data types matter:
  - Categorical, nominal
  - Ordinal, hierarchical
  - Continuous, numeric

Bivariate

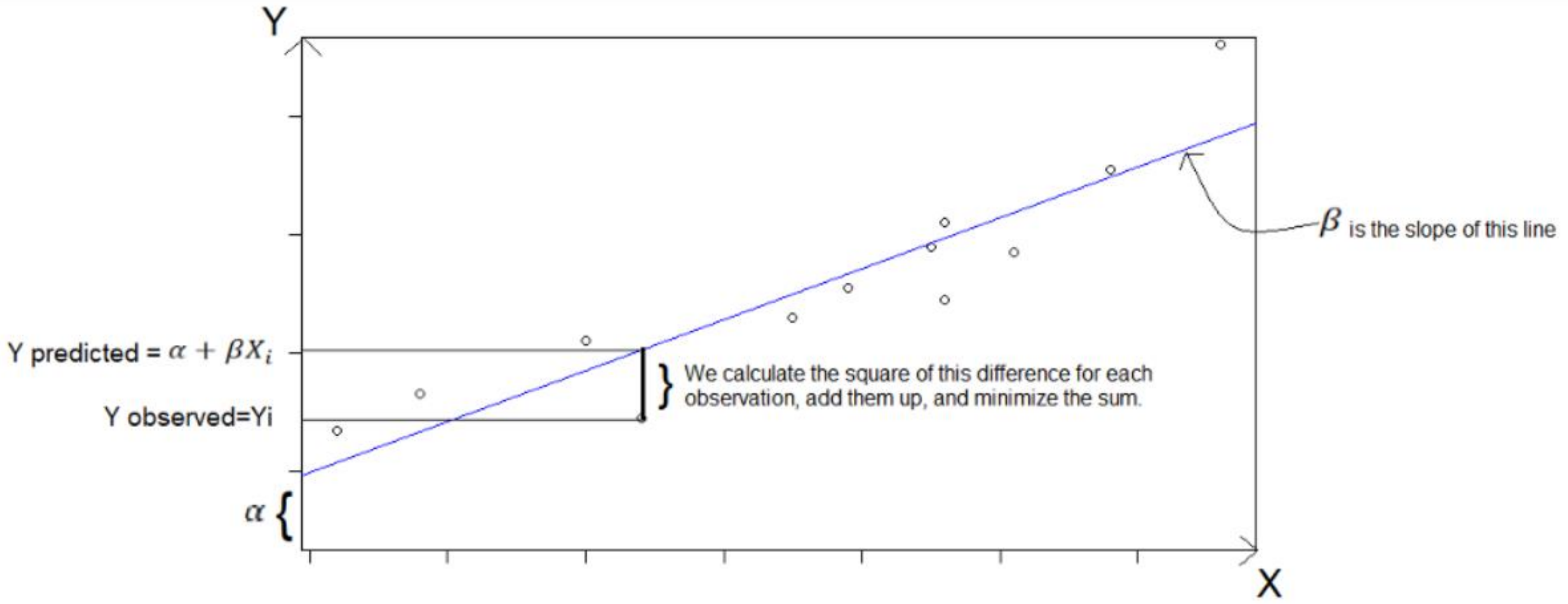
- **Cross-tabulation** analysis (crosstab) is the initial level for bivariate analyses.
  - It presents a descriptive relationship between two categorical (nominal or ordinal) variables. Note that it is not common to use crosstabs for numeric (interval or ratio) variables.
- **T-test** analysis is an inferential statistical test that determines whether there is a statistically significant difference between the means in two unrelated groups.
  - For a t-test analysis we need one categorical (nominal or ordinal) variable and one continuous viariable. However note that categorical variable should have only two values, which we will compare.
- **One-way ANOVA** analysis is an inferential statistical test that determines whether there is a statistically significant difference between the means in more than two unrelated groups. It is very much similar to T-test analysis. The only difference is the number of values.
  - For a one-way ANOVA analysis, we need one categorical (nominal or ordinal) variable and one continuous variable. However, note that the categorical variable should have more than two values, which we will compare.
- **Correlation** basically refers "a mutual relationship or connection between two or more things." It is a statistical measure that indicates the extent to which two or more variables fluctuate together. A positive correlation indicates the extent to which those variables increase or decrease in parallel; a negative correlation indicates the extent to which one variable increases as the other decreases.
  - Positive Correlation – when the value of one variable increases with respect to another.
  - Negative Correlation – when the value of one variable decreases with respect to another.
  - No Correlation – when there is no linear dependence or no relation between the two variables.

Reminder: Confidence intervals for normal distribution



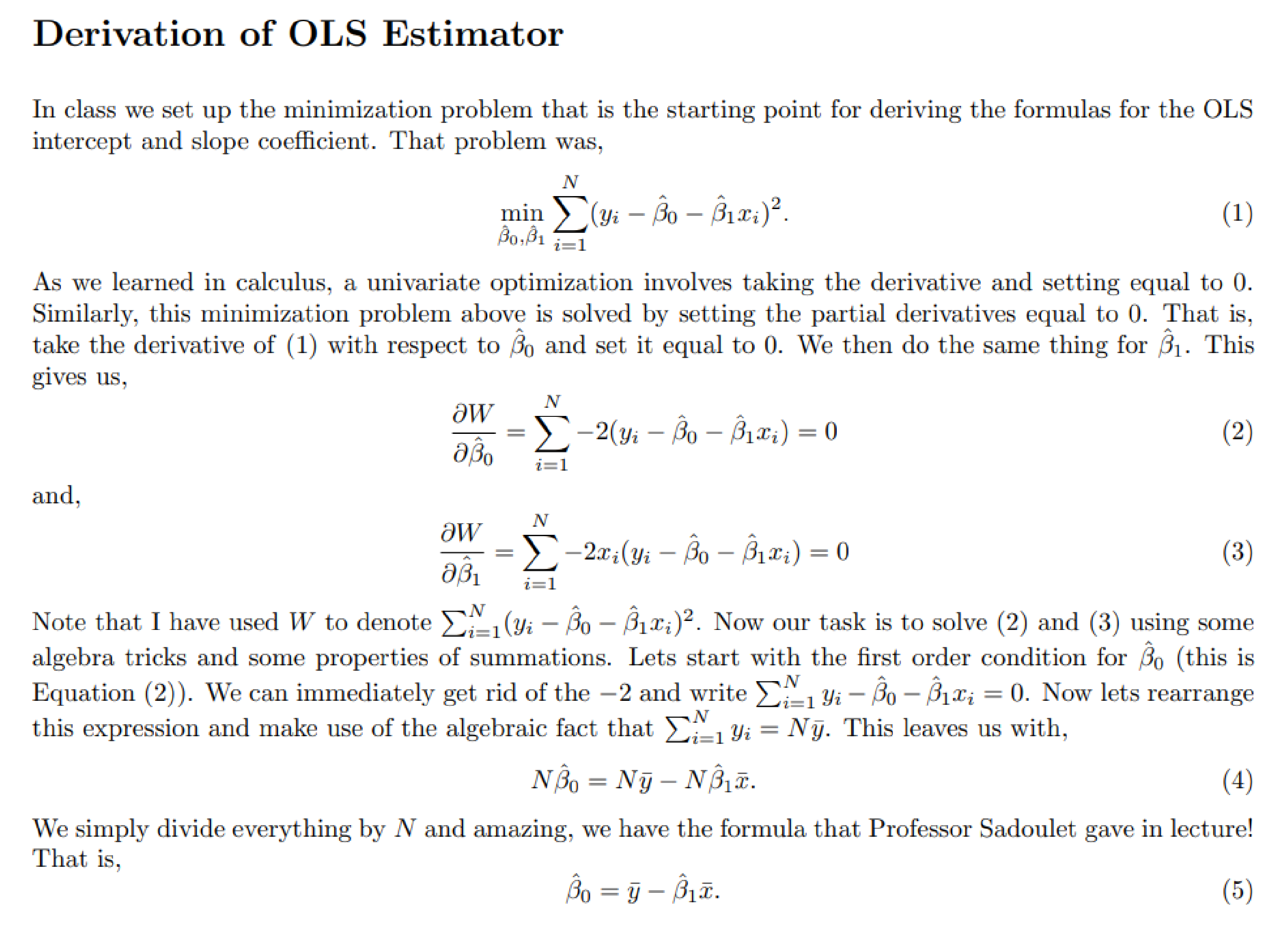
Linear regression (OLS)

- OLS is a good use-case, and will be important while learning basic prediction mechanism behind Machine Learning models
- Derivation of OLS



Y
Y predicted = $\alpha + \beta X_i$
Y observed=Yi
α
X
We calculate the square of this difference for each observation, add them up, and minimize the sum.
β is the slope of this line

Linear regression (OLS)



Derivation of OLS Estimator

In class we set up the minimization problem that is the starting point for deriving the formulas for the OLS intercept and slope coefficient. That problem was,

$$\min_{\hat{\beta}_0,\hat{\beta}_1} \sum_{i=1}^{N} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2. \tag{1}$$

As we learned in calculus, a univariate optimization involves taking the derivative and setting equal to 0. Similarly, this minimization problem above is solved by setting the partial derivatives equal to 0. That is, take the derivative of (1) with respect to $\hat{\beta}_0$ and set it equal to 0. We then do the same thing for $\hat{\beta}_1$. This gives us,

$$\frac{\partial W}{\partial \hat{\beta}_0} = \sum_{i=1}^{N} -2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{2}$$

and,

$$\frac{\partial W}{\partial \hat{\beta}_1} = \sum_{i=1}^{N} -2x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \tag{3}$$

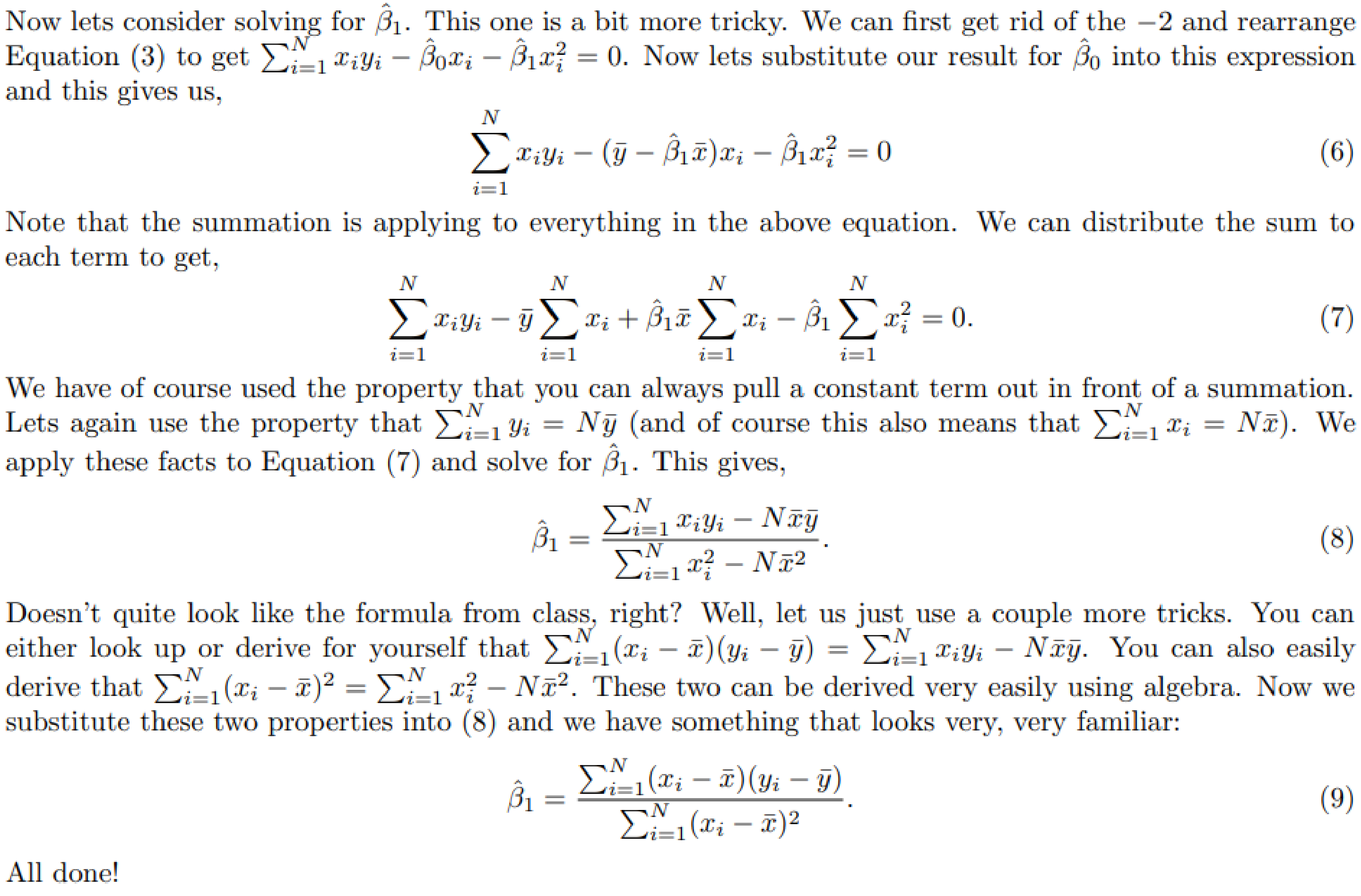
Note that I have used W to denote $\sum_{i=1}^{N}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$. Now our task is to solve (2) and (3) using some algebra tricks and some properties of summations. Lets start with the first order condition for $\hat{\beta}_0$ (this is Equation (2)). We can immediately get rid of the -2 and write $\sum_{i=1}^{N} y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i = 0$. Now lets rearrange this expression and make use of the algebraic fact that $\sum_{i=1}^{N} y_i = N\bar{y}$. This leaves us with,

$$N\hat{\beta}_0 = N\bar{y} - N\hat{\beta}_1\bar{x}. \tag{4}$$

We simply divide everything by N and amazing, we have the formula that Professor Sadoulet gave in lecture! That is,

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}. \tag{5}$$

Linear regression (OLS)



Now lets consider solving for $\hat{\beta}_1$. This one is a bit more tricky. We can first get rid of the -2 and rearrange Equation (3) to get $\sum_{i=1}^{N} x_i y_i - \hat{\beta}_0 x_i - \hat{\beta}_1 x_i^2 = 0$. Now lets substitute our result for $\hat{\beta}_0$ into this expression and this gives us,

$$\sum_{i=1}^{N} x_i y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) x_i - \hat{\beta}_1 x_i^2 = 0 \tag{6}$$

Note that the summation is applying to everything in the above equation. We can distribute the sum to each term to get,

$$\sum_{i=1}^{N} x_i y_i - \bar{y} \sum_{i=1}^{N} x_i + \hat{\beta}_1 \bar{x} \sum_{i=1}^{N} x_i - \hat{\beta}_1 \sum_{i=1}^{N} x_i^2 = 0. \tag{7}$$

We have of course used the property that you can always pull a constant term out in front of a summation. Lets again use the property that $\sum_{i=1}^{N} y_i = N\bar{y}$ (and of course this also means that $\sum_{i=1}^{N} x_i = N\bar{x}$). We apply these facts to Equation (7) and solve for $\hat{\beta}_1$. This gives,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} x_i y_i - N\bar{x}\bar{y}}{\sum_{i=1}^{N} x_i^2 - N\bar{x}^2}. \tag{8}$$

Doesn't quite look like the formula from class, right? Well, let us just use a couple more tricks. You can either look up or derive for yourself that $\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{N} x_i y_i - N\bar{x}\bar{y}$. You can also easily derive that $\sum_{i=1}^{N} (x_i - \bar{x})^2 = \sum_{i=1}^{N} x_i^2 - N\bar{x}^2$. These two can be derived very easily using algebra. Now we substitute these two properties into (8) and we have something that looks very, very familiar:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2}. \tag{9}$$

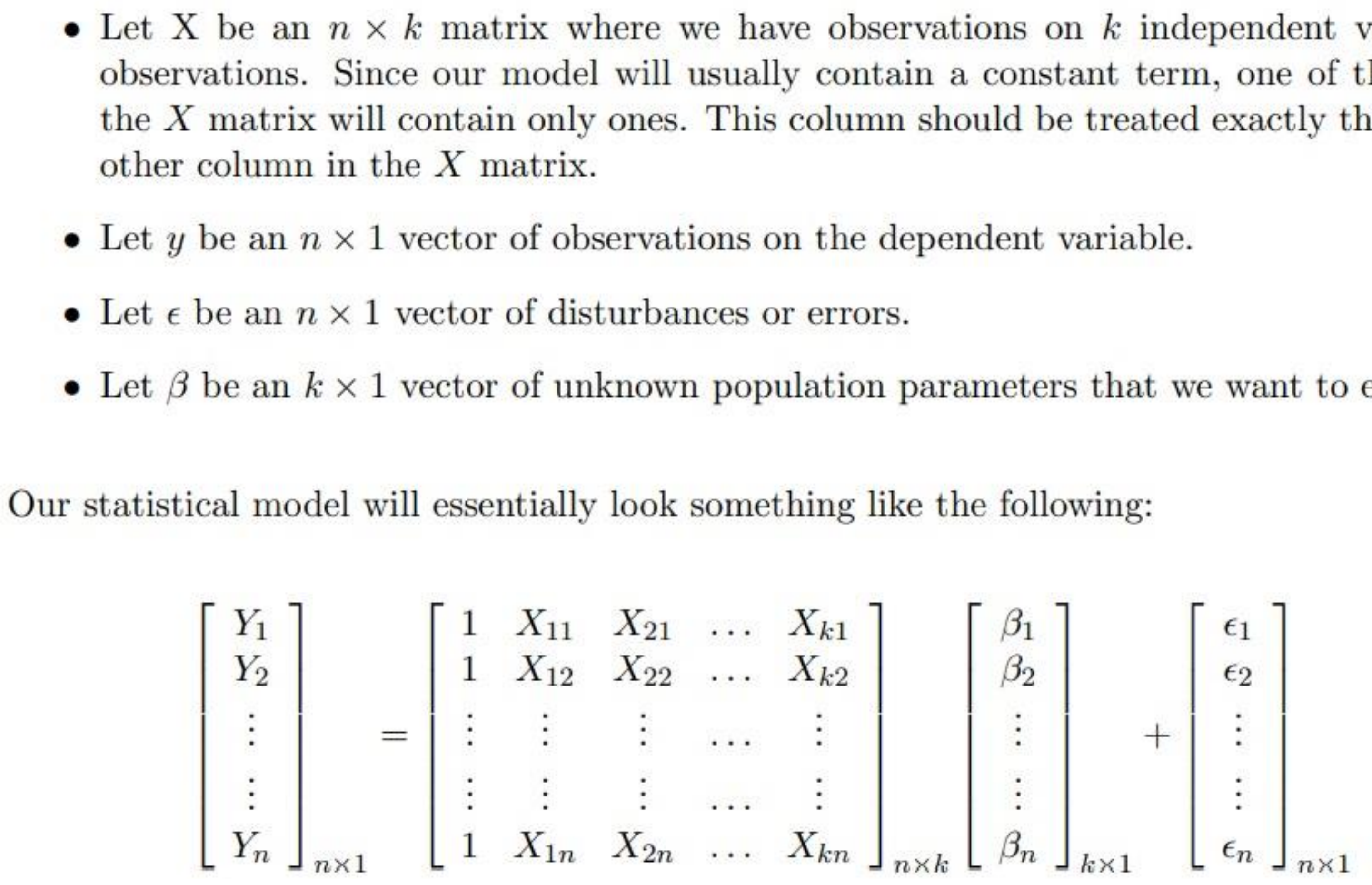
All done!

Predicting the DV in real life

- According to DV structure
  - Logistic regression: Binary
  - Mutinomial regression: Ordinal/categorical (including more than two values)
  - Poisson regression: Right skewed
  - Negative binomial regression: Right skewed + count variable

- Multilevel regressions
- Structural equation models
- Working with Panel Data structures (time series cross sectional data)
- Spatial and temporal autocorrelations
- Clustered standard errors, robust standard errors

Homework hints



- Let X be an $n \times k$ matrix where we have observations on k independent variables for n observations. Since our model will usually contain a constant term, one of the columns in the X matrix will contain only ones. This column should be treated exactly the same as any other column in the X matrix.
- Let y be an $n \times 1$ vector of observations on the dependent variable.
- Let ϵ be an $n \times 1$ vector of disturbances or errors.
- Let β be an $k \times 1$ vector of unknown population parameters that we want to estimate.

Our statistical model will essentially look something like the following:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ \vdots \\ Y_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} 1 & X_{11} & X_{21} & \dots & X_{k1} \\ 1 & X_{12} & X_{22} & \dots & X_{k2} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & X_{1n} & X_{2n} & \dots & X_{kn} \end{bmatrix}_{n\times k} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \vdots \\ \beta_n \end{bmatrix}_{k\times 1} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \vdots \\ \epsilon_n \end{bmatrix}_{n\times 1}$$

Homework hints

- Calculate n, k
- $\hat{\beta} \quad = \quad (X'X)^{-1}X'y$
- Covariance matrix $\quad = \quad \sigma^2(X'X)^{-1}$
- $\hat{\sigma}^2 = \dfrac{e'e}{n-k}$
- X’ = X.T
- @ is matrix multiplication
- numpy.linalg

Practical session for basic statistics