




WEEK 7: MACHINE LEARNING & REGRESSION

Dr. Kai Li

School of Information Sciences
University of Tennessee, Knoxville
Spring 2025



Overview

- Announcements
- Machine learning
- Regression

Mid-term course evaluation

- This is a device I used in all my classes in the past few years, to collect feedback from you before it's too late to address them.
- If possible, please try to post your feedback before the next class and I will address all your comments in the next class.
- https://utk.co1.qualtrics.com/jfe/form/SV_eLhTryhle1F6I9Q
- The link is also posted on the Week 7 folder.

Discussing your final project idea

- In week 10 (March 27), I am going to ask you to share your idea for the final project.
- Please prepare 1-2 slides (using around 2-5 minutes) to introduce (1) your dataset, (2) research questions, and (3) what methods or type of analysis you plan to do.
 - *We will also make some adjustments to our topics in W10-11, which I will announce in the next week.*

Review

- Correlation:
- Different methods:
 - *T-test*
 - *ANOVA*
- Any questions so far?
 - *I will talk about one specific question later: when to choose a model?*

Chi-square

- Another method to understand the differences between groups.
 - *Chi-square is used to understand **how to categorical variables are influencing each other.***
 - In this method, there is no DV and IV.
 - Null hypothesis: the two variables are independent from each other.
 - For example, if a medicine can cure the patients (cured vs. not cured).
 - *We can use a matrix (or contingency table) of two categorical variables to do the analysis.*
 - <https://www.sthda.com/english/wiki/chi-square-test-of-independence-in-r>

Assumptions of chi-square

- Most of the methods we discussed last week have a strong requirement for the sample size.
 - *I want you to check the minimum sample size for each method before using it.*
- Chi-square, as a nonparametric method, has very few data assumptions.
 - *As compared, we *generally expect t-test and ANOVA data to be normally distributed.*

Difference

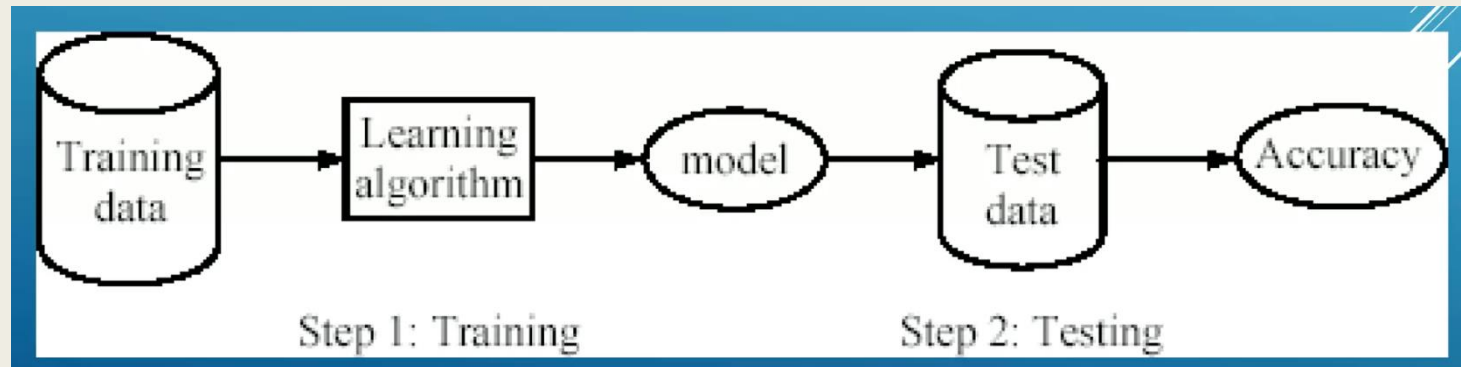
- We can use various methods to **compare the difference between two or multiple groups**.
 - *T-test: a set of methods to compare the **difference (in a numeric DV) between the means of two groups**.*
 - *ANOVA (analysis of variance): it is similar with t-test (also **numeric DV**) but can be used to **compare more than two groups**.*
 - *Chi-square test: it is used to compare the **differences between categorical variables**.*

What is machine learning?

- Machine learning (ML) is to program an algorithm to learn from data, so that the algorithm will solve the problems.
 - *Most the classic machine learning methods are based on statistical methods, i.e., we train statistical models on a selection of data, which will be used to predict a larger dataset.*
 - *ML is an important application of AI.*

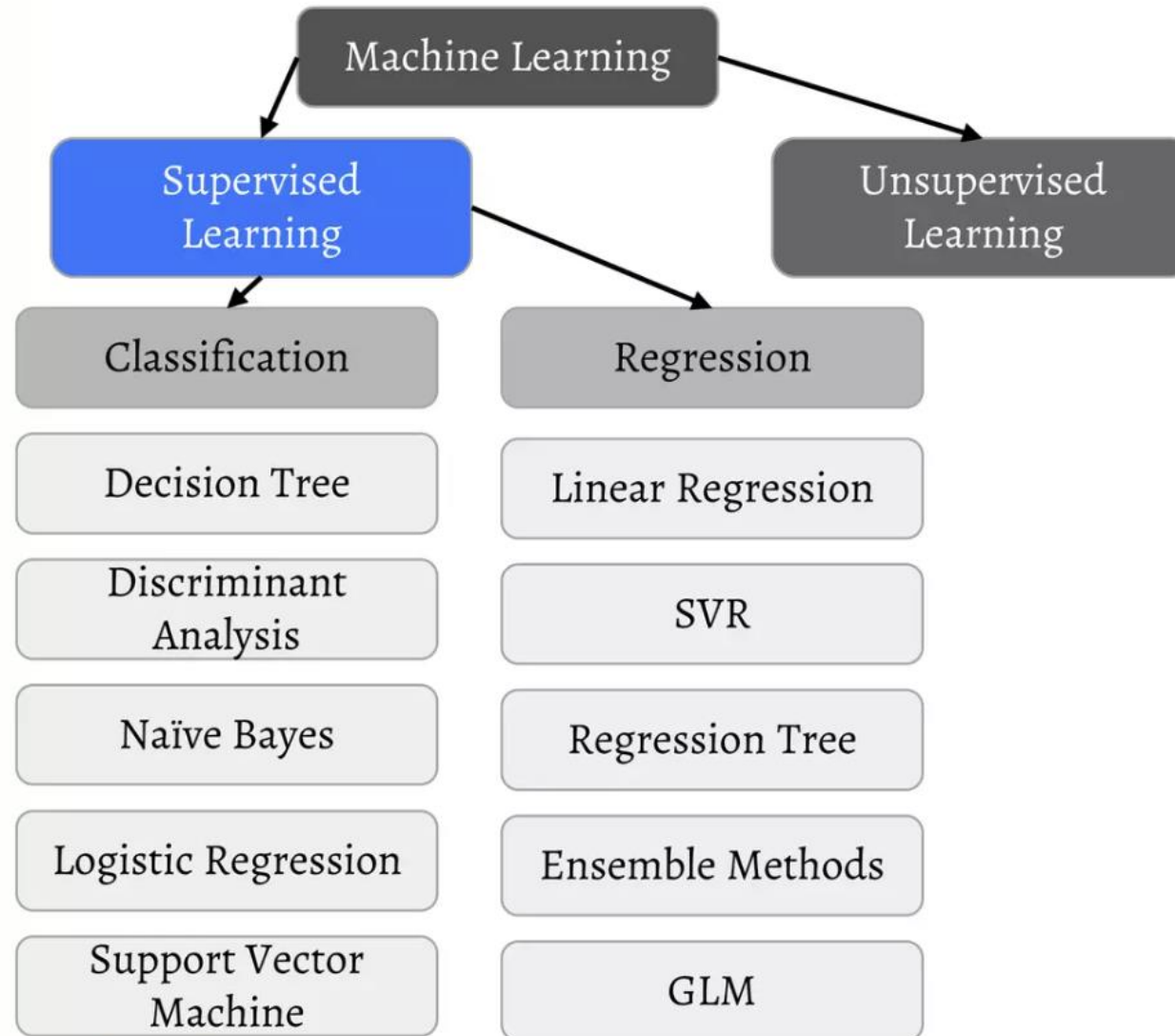
Supervised vs. unsupervised learning

- **Supervised learning** used pre-labelled **training set** (for *training the model*) and apply the model on the **testing set** (for *prediction*).
 - We will need to have **ground truth** to teach the algorithm and to evaluate the results from prediction.



Supervised vs. unsupervised learning

- Two common tasks in supervised learning:
 - *Regression: when the target variable is numeric*
 - *Classification: when the target variable is categorical*
 - *The major difference between the two is the **class of the dependent variable**.*



Spotle.ai Study Material

<https://www.slideshare.net/SpotleAI/supervised-and-unsupervised-machine-learning>

Supervised vs. unsupervised learning

- **Unsupervised learning** does not require training or testing data, but we can apply a model on our data directly.
 - *In other words, we don't teach machines and don't even have correct answer for the question.*
 - *It is rather a way for the algorithm to identify the data structure.*
- Two typical methods:
 - *Clustering: how are individuals grouped in our data*
 - *Association: "diaper and beer"*

Inferential statistical methods

- So these statistical methods are different from each other in terms of the questions they can answer!
 - *Correlation: how are two variables correlated with each other?*
 - *Difference: how are some groups of observations different from each other?*
 - *Regression: how do some variables influence another variable?*

Regression

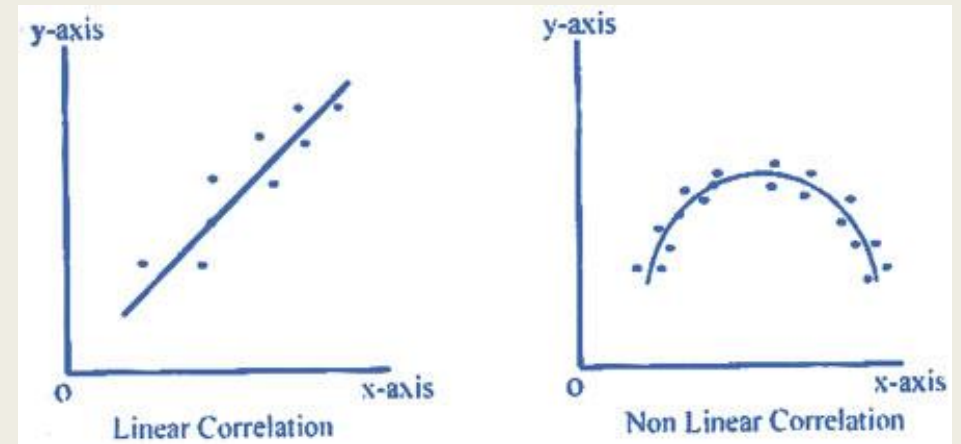
- Regression is a subset of statistical methods that examines the nature and extent of the relationship between two or more variables.
 - *In most cases, we have just one DV and multiple IVs.*
- Typical question:
 - *Inference: how does one (or more) factor(s) influence another factor?*
 - *Prediction: how accurate can we predict one variable based on other variables?*

Regression vs. correlation

- While correlation can present the relationship between two variables (if two variables have similar patterns), it does not show as one variable changes, how will the other one changes as well.
 - *For example, if we earn more money, how much happiness we will get?*
- The correlation relationship is symmetrical, whereas regression is not.
- The results from regression are still **correlational (not causal)**.
 - *We cannot prove that more money definitely caused happiness.*

Linear regression

- Linear regression assumes that there is a linear relationship between DV and all IVs.
 - *We can normalize the data or move to a different type of regression model.*
- Linear regression is the most popular model in the regression family.



Assumptions of linear regression models

- Linearity - the Y variable is linearly related to the value of the X variable (for non-linear regression, we need to test if the shape will work).
- Independence of Error - the error (residual) is independent for each value of X.
- Homoscedasticity - the variation around the line of regression be constant for all values of X.
- Normality - the values of Y be normally distributed at each value of X.

Sample size

- **All inferential statistical methods** (including the ones we discussed last week) have requirements for the minimum sample size!
- While you can probably find a lot of answers to this question for regression model, we need to pay attention to the sample size on two levels:
 - *The total sample size: > 50 – 500*
 - *Sample size in each category: > 10 – 25*
- We also need to consider the number of IVs in our model.

Simple linear regression formula

The formula for a simple linear regression is:

$$y = \beta_0 + \beta_1 X + \epsilon$$

- **y** is the predicted value of the dependent variable (**y**) for any given value of the independent variable (**x**).
- **B₀** is the **intercept**, the predicted value of **y** when the **x** is 0.
- **B₁** is the regression coefficient – how much we expect **y** to change as **x** increases.
- **x** is the independent variable (the variable we expect is influencing **y**).
- **e** is the **error** of the estimate, or how much variation there is in our estimate of the regression coefficient.

Interpretation of results

- R-squared value: % of variation in the data that can be explained by the model (the higher, the better)
- If the income is 0, the mean happiness value will be 0.204.
- As income increases by one unit, happiness will increase by 0.714 unit on average and this increasing trend is statistically significant.

Call:

```
lm(formula = happiness ~ income, data = income.data)
```

Residuals:

Min	1Q	Median	3Q	Max
-2.02479	-0.48526	0.04078	0.45898	2.37805

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.20427	0.08884	2.299	0.0219 *
income	0.71383	0.01854	38.505	<2e-16 ***

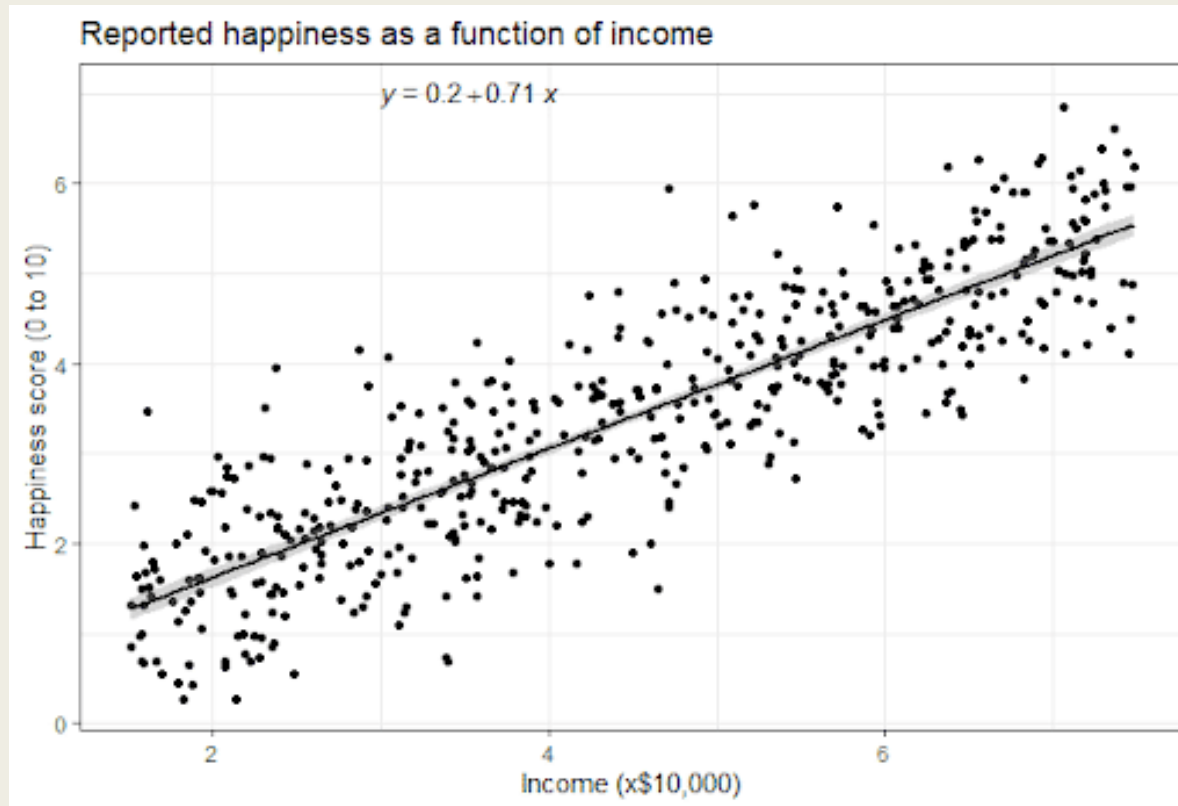
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7181 on 496 degrees of freedom

Multiple R-squared: 0.7493, Adjusted R-squared: 0.7488

F-statistic: 1483 on 1 and 496 DF, p-value: < 2.2e-16

$$\text{happiness} = 0.20 + 0.71 \cdot \text{income}$$



VISUALIZING RESULTS

We can use scatterplots to visualize the relationship between two variables in the data with a **trend line**.

Report your regression results

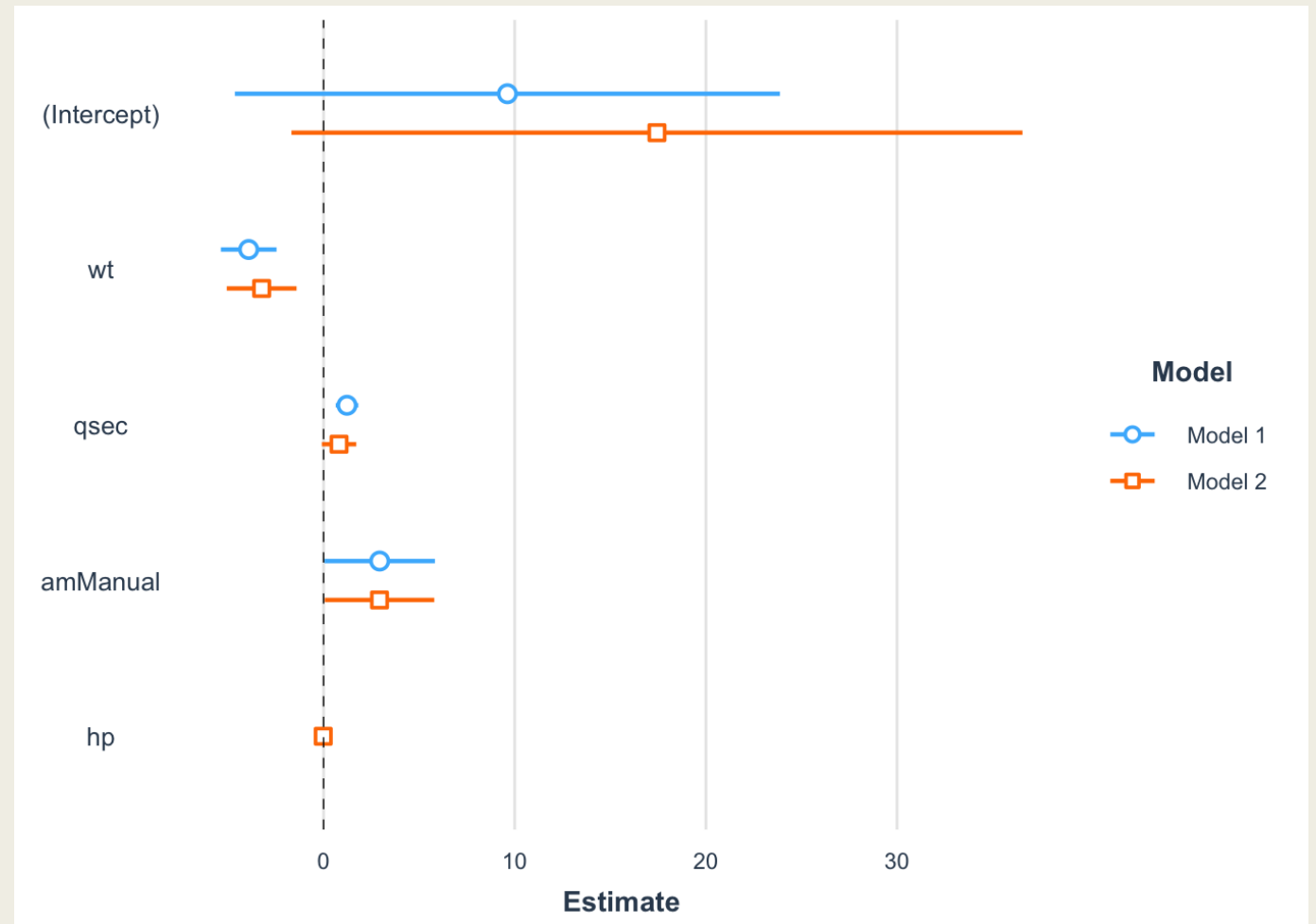
- One example is here:
- <https://personalpages.bradley.edu/~rob/EHC310/Results/Regression>
- *Tables are very useful to report results on the variable-level.*

Table 1: Insert Title Here

	Simple OLS Model 1	Simple OLS Model 2	NL Least Squares
A	0.2141** (0.0275)	0.2066** (0.0271)	0.2245** (0.0271)
B	-0.1505** (0.0335)	-0.1532** (0.0328)	-0.1451** (0.0335)
C	-0.1443** (0.0423)	-0.1438** (0.0416)	-0.1346** (0.0426)
D_1		-0.1721** (0.0327)	-0.1573** (0.0329)
D_2		-0.0004** (0.0002)	-0.0575 (0.0417)
D_3		0.1659 (0.2218)	0.002 (0.0121)
R^2	0.1557	0.1736	0.1548
λ			0.675
N	432	450	601

Report your regression results

- Using the segment chart to show the estimate of each IV is another useful approach to visualize the results.
- The p-value is implied by whether or not the line passes the 0 point.
- We can further compare multiple models using this graph.



Predicting the testing set

- After we have the model, if we want to do prediction, rather than just making inference, we should apply our model to the testing set and evaluate how well the prediction is.
 - *The model for prediction should be trained just on the training set!*
 - *High goodness-of-fit of the model does not mean it will create accurate prediction!*
- We can use various measurement to evaluate the prediction.

Multiple regression

- It is super rare that we only want to examine the relationship between ONE independent variable and ONE dependent variable.
 - *There are many **compounding factors** behind any relationship: education or personality could also affect the relationship between income and happiness!*
- So, we may want to consider the effect of multiple independent variables on the outcome.
 - *It is not very different from the model we discussed but each IV will have its own estimates and p-values.*

Dependent Variable
(Response Variable)

Independent Variables
(Predictors)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \varepsilon$$

Y intercept

Slope
Coefficient

Error Term

In multiple regression, each IV has its own independent coefficient.

Coefficients ^a								
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	87.830	6.385		13.756	.000	75.155	100.506
	age	-.165	.063	-.176	-2.633	.010	-.290	-.041
	weight	-.385	.043	-.677	-8.877	.000	-.471	-.299
	heart_rate	-.118	.032	-.252	-3.667	.000	-.182	-.054
	gender	13.208	1.344	.748	9.824	.000	10.539	15.877

a. Dependent Variable: VO2max

This is the easiest way to “control” the effect of other variables on the outcome, so that we can **evaluate the independent effect of IVs on the outcome**.

Here, we can say after controlling one’s weight, heart rate, and gender, age has a significant negative effect (at 0.05 level) on the outcome (maximal oxygen consumption).

Categorical independent variables

- We can have categorical IVs in linear regression model, as long as the model still stands against all assumptions.
 - *For example, happiness ~ gender*
- We will have different results from the model.

Coefficients^a

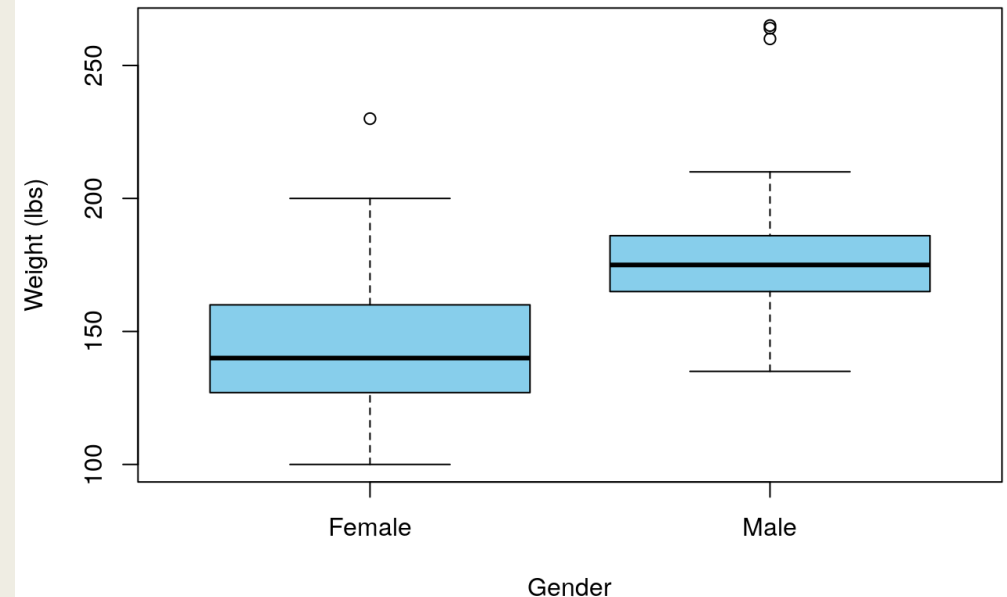
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	51.678	.982		52.619	.000
	X1	-7.597	1.989	-.261	-3.820	.000
	X2	3.945	2.823	.095	1.398	.164
	X3	-5.855	2.153	-.186	-2.720	.007

a. Dependent Variable: WRITE

The results show that comparing to the baseline category (which is NOT shown in the table), X2 group has insignificant positive impact on the outcome and X1 and X3 has significant negative impact on the outcome.

The full comparison can be visualized using boxplot.

Weight of College Students



The family of linear regression

- You may see different names for regression models, such as:
 - *Ordinary least squares (OLS) → the ordinary model in R*
 - *Ridge regression*
 - *Lasso regression*
- As a beginner, you don't need to understand all their statistical differences.
 - *OLS model can be useful in most cases.*
 - *The latter two models can be more useful for prediction, rather than inference.*

Demonstration

- Regression methods