Regression

- Objective: You have a bunch of data points with true label attached,

You want to fit a model that can be used to predict the label of the new data points.

In linear regression, we fit a line:

$$\mathrm{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_n x_n + \epsilon$$

Example: you want to predict the house price in Toronto based on some features such as, location, number of bedrooms, age of the house, etc.

- We have different kinds of regression, such as polynomial regression, where you just fit a polynomial (instead of line) to your data set; and logistic regression, which is used when you have binary labels (i.e. y can only take on 0 or 1). You will learn about them more in future courses.

Linear regression

$$\mathrm{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_n x_n + \epsilon$$

y here means the prediction of our model.

Residual = true label/response – prediction

Loss function

- Loss function defines how good your regression model fit to your data set. We want to find the $\boldsymbol{\beta}$ that minimize the loss.
- Mean squares error(MSE): Most commonly used loss function.
- Sum over all the residual: L($\boldsymbol{\beta}$) = $\Sigma_i (y_i - \hat{y}_i)^2$
- **Least square method**:
- taking the partial derivative with respect to each model parameter $\boldsymbol{\beta}$, and set them to zero to solve for the least square estimates: $\hat{\boldsymbol{\beta}}$
- Example: homework question 1.
- R's lm( ) function does the job of fitting the parameter for us!

In class presentation exercise

- Pick one of the following 4 topics as the subject of presentations.
- Every member of each group has to present

1. Explain the difference between a linear regression model with both horsepower and rear axle ratio as predictors for gas mileage (2e) and two simple linear regression models (2d). Which model explains more of the variability in gas mileage?

2. Based on the model you fit in 2e, what would be the predicted gas mileage for a car with a horsepower of 25 and rear axle ratio of 6. Do you think this prediction is reliable? Explain why or why not.

3. Based on the linear regression model of median value of homes (medev) and percentage of lower status of the population that lives in the census tract (lstat) (3d-ii), is there any evidence of overfitting? (hint: use RMSE and coefficient of determination)

4. How accurate is a multiple regression model to predict medev where the following covariates (inputs) are used in addition to lstat: crim, zn, rm, nox, age, tax, ptratio (3d-iii)? Does this model provide more accurate predictions compared to the model that just include lstat? Explain the differences.