## Problem Description

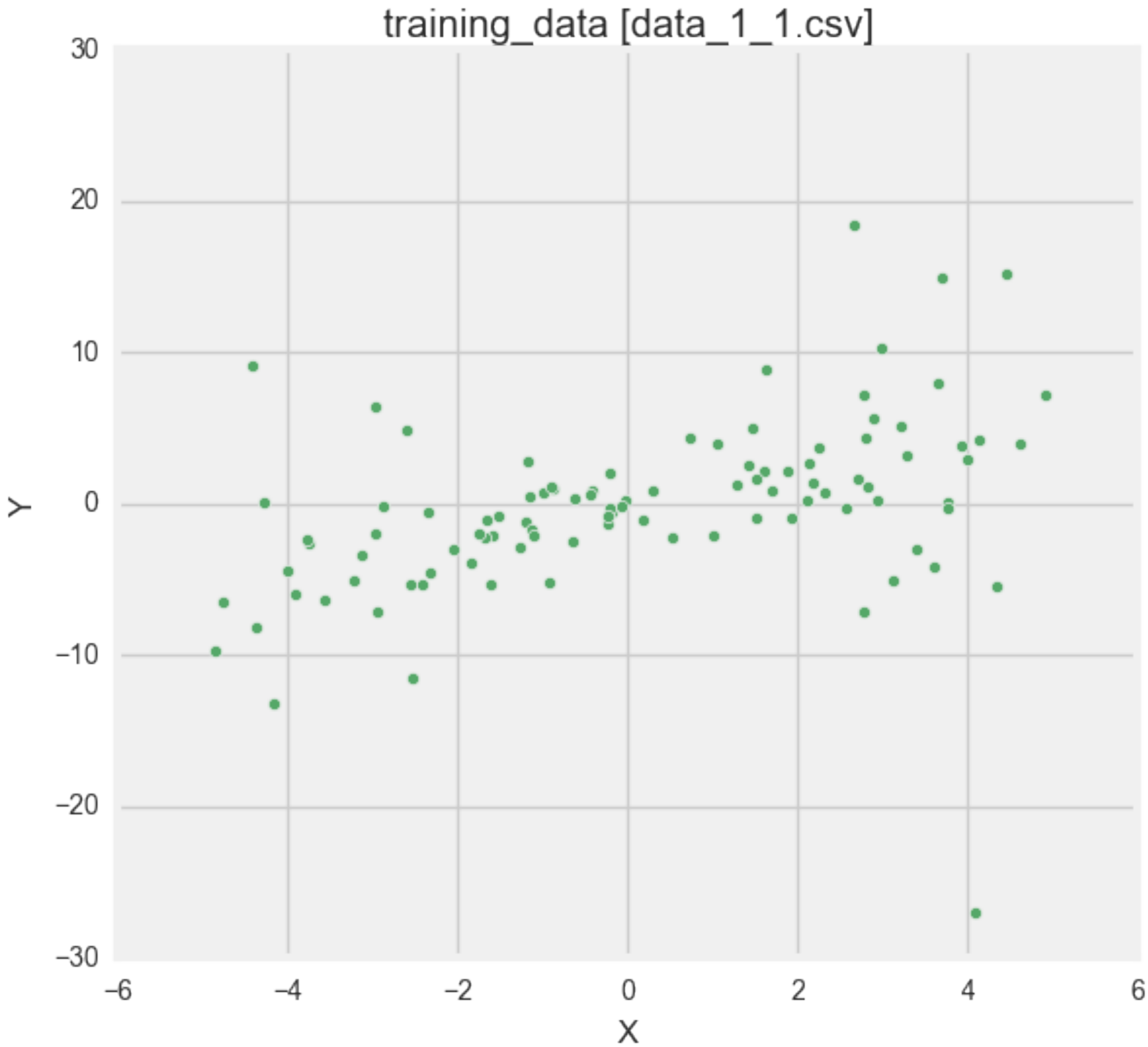
The goal of this project is to build a linear model for $Y$ given $X$ to estimate a pair of (a, b) values for each of the dataset $D_k$ where $k = 1..5$ with $(x_i, y_i) \in D_k$ where $i = 1..100$. In addition estimate the global parameters for the random variable $e$ given $E[e|X] = 0$

$y = ax + b + e$

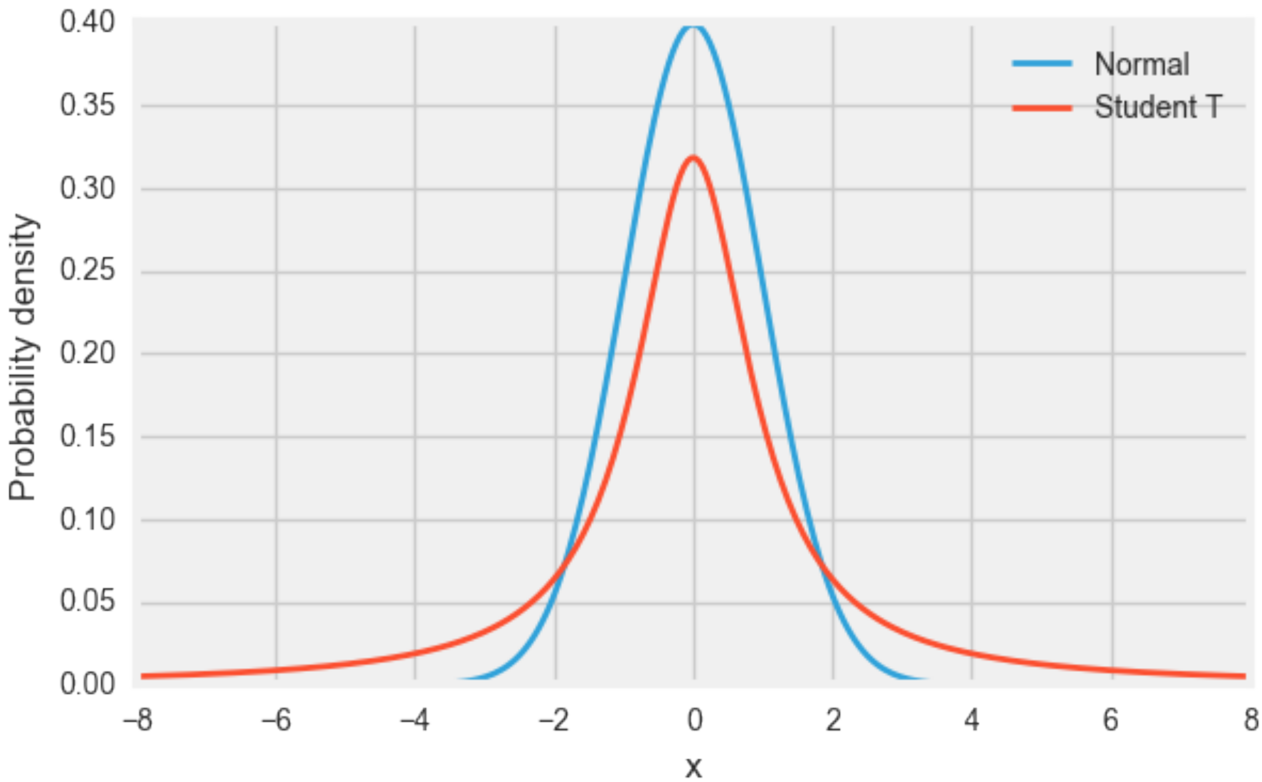
```
y ~ Observed Dependent Variable

x ~ Independent Variable

a ~ Unknown Parameter or Slope

b ~ Unknown Parameter or Intercept

e ~ Unknown Stochastic Error Variable
```

## Error Distribution

It is stated that the error variable $e$ has a distribution with a heavy tail. By plotting the $x_i$, $y_i$ values on a scatter plot it is noticable that the variance of the observed $y_i$ variable is large.




A distribution with heavier tails than normal distribution is the Student T distribution. In linear regression problems, a Student T distribution makes the parameter estimates robust to outliers in the training dataset. In this model we will assume that $e$ is sampled from such a Student-T distribution.




## Understanding the relationship between $e$ and $X$

Given conditional expectation $E[e|X] = 0$ :

- $E[Y|X] = a * x + b$ or $E[y_i|x_i] = a * x_i + b$

Using law of iterated expectations $E[E[e|X]] = E[e]$ and $E[eX|X] = 0$

- $E[e] = 0$
- $Cov(e, X) = E[(e - E[e]) * (x - E[x])] = E[eX] - E[x]E[e] = E[eX] = E[E[eX|X]] = 0$
- $\rho_{e,X} = 0$

$Cov(e, X) = 0$ and $E[e|X] = 0$ implies there is no linear relationship between $e$ and $X$.

Here, an assumption of homoskasdicity or constant variance may be made which will exclude any non-linear relationship between $e$ and $X$. This would be a convenience assumption, since it makes $e$ and $X$ independent random variables. Even if this assumption is not true, the estimates of slope and intercept will be un-biased. However the error distribution will be very far off from the training data.

If the error variable $e$ is allowed to be heteroskasdicity then $Var(e)$ must be modeled as a function of $X$.

In this work, we will compare 2 models. One of them will be based on a homoskasdicity assumption

Model (1) - $Var(e)$ is independent of x

Model (2) - $Var(e) \propto x^2$

The two dependencies have been hypothesized based on intuition by looking at the scatter plot of the datasets.

## Model assumptions and its implications

Given samples $(x_i, y_i)$ from a training dataset $D_k$ where $k = 1..5$

1. $E[Y|X] = aX + b$
2. The conditional mean of $e$ given $x_i$ is 0 for all values $x_i \in X$
3. The conditional variance of $e$ given $x_i$ is proportional to $f(x)$ where $f(x) \in [1, x^2]$ and $w > 0$

## Model

A single generative model is built for each training data file.

$$a \sim Normal(\mu = 0, \sigma = 100^2)$$

$$b \sim Normal(\mu = 0, \sigma = 100^2)$$

$$w \sim Uniform(lower = 0.0, higher = 0.5)$$

$$\nu \sim Uniform(lower = 1.0, higher = 10.0)$$

$e \sim StudentT(\mu = 0.0, \lambda = \frac{w}{f(x)}, \nu = \nu)$, where $Var(e) \propto \frac{1}{\lambda}$

$$Y_{obs} \sim aX + b + e$$

The prior distribution used for a, b are weakly informative. The priors for $\nu$ are restrictive to values > 1 in order for the gamma function in the student-t distribution to be defined. A uniform prior is selected for $w$ and it is restrictive to positive values.

## Bayesian Inference

The goal is to compute the full posterior probability distribution for each of the unknown parameters, $\theta = [a,\ b,\ w,\ \nu]$. The distribution for $\theta$ is computed using Metropolis Hasting sampling technique.

Here is the algorithm :

1. $\theta^{(0)} = \theta^{(MAP)}$
2. for i = 1 to N

   Propose $\theta^{cand} \sim q(\theta^{(cand)}|\theta^{(i-1)})$

   Acceptance Probability:

   $$\alpha(\theta^{(cand)}|\theta^{(i-1)}) = min\{1, \frac{q(\theta^{(i-1)}|\theta^{(cand)})\pi(\theta^{(cand)})}{q(\theta^{cand}|\theta^{(i-1)})\pi(\theta^{(i-1)})}\}$$

   $u \sim Uniform(u; 0, 1)$

   if $u < \alpha$ then

   Accept the proposal: $\theta^{(i)} \leftarrow \theta^{(cand)}$

   else:

   Reject the proposal: : $\theta^{(i)} \leftarrow \theta^{(i-1)}$

   end if

   end for

The proposal distribution used for every parameter is $q(\theta^i|\theta^{i-1}) = \mathcal{N}(\mu = \theta^{i-1}, 1)$. An additional scaling parameter is used to tune the candidate values, to be further or closer to the $\theta^{i-1}$, if the acceptance rate is higher or lower respectively during the last tuning interval. The tune interval is set to 100. The parameter N or the number of samples generated is 10000.

The distribution is computed by the MCMC sampling technique by using the maximum aposteriori value of $\theta$ for initialization.

$$\theta^{MAP} = argmax_\theta \sum_{i=1}^{N} log\, p(y_i|x_i, \theta)\ for\ (x_i, y_i) \in\ TrainingData$$

where

- $p(y_i \mid \nu, \mu = ax_i + b, \sigma^2) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu\sigma^2}}\left(1 + \frac{1}{\nu}\frac{(y_i - \mu)^2}{\sigma^2}\right)^{-\frac{\nu+1}{2}}$

The minimum of the function given the 4-dimensional $theta$ variable is calculated by Powell's method. It is implemented in the scipy python package as $scipy.optimize.fmin\_powell$.

The point estimate for each of the unknown parameters given a file is calculated as the mean of the posterior probability distribution generated by the MCMC sampling technique.

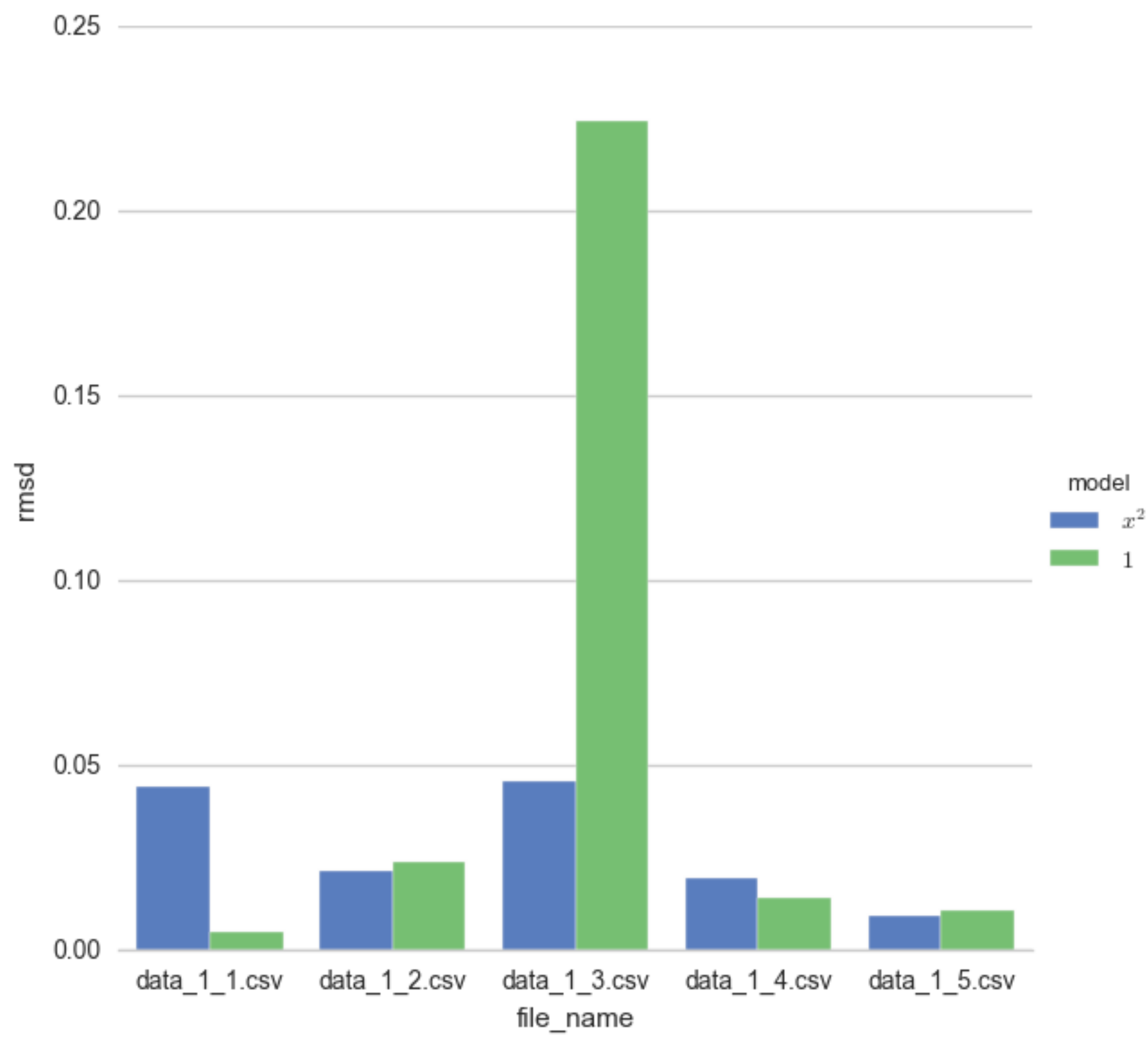
$$\theta^{est} = average(\theta^{i...N})$$

## Posterior Predictive Check

There are 2 possible model each with different f(x)$. The test statistic for the posterior predictive check is the RMSD or Root Mean Square Deviation to measure the goodness of fit.

$$RMSD(y, \hat{y}) = \sqrt{\sum_{i=1..N} \frac{(\hat{y}_i^{ppc} - y_i^{obs})^2}{N}}$$

1. $f_1(x) \ = \ x^2$ or $Var(e) \propto x^2$
2. $f_2(x) \ = \ 1$ or $Var(e)$ independent of x




RMSD values on the 5 files for $f_1(x) = x^2$ and $f_2(x) = 1$

It is not obvious whether a homoskedastic or a heteroskastic model is a better fit for the error values. Since there are different number of samples in each file, a weighted average of the rmsd value will be calculated which can be interpreted as rmsd per sample across all the training datasets.

| $f(x)$ | **weighted_rmsd** |
| --- | --- |
| $x^2$ | 0.0295 |
| 1 | 0.0438 |

## Results

| files | $\hat{a}$ | $\hat{b}$ |
|---|---|---|
| data_1_1.csv | 1.033639 | 0.005087 |
| data_1_2.csv | 1.250717 | -0.554186 |
| data_1_3.csv | -0.665783 | 0.135664 |
| data_1_4.csv | 1.002800 | 0.522109 |
| data_1_5.csv | -0.964518 | -0.087334 |

| files | $w$ | $\nu$ |
|---|---|---|
| data_1_1.csv | 0.243841 | 4.088898 |
| data_1_2.csv | 0.157571 | 2.656124 |
| data_1_3.csv | 0.159151 | 2.971437 |
| data_1_4.csv | 0.187357 | 6.328168 |
| data_1_5.csv | 0.200360 | 3.781542 |
| weighted_mean | 0.192813 | 3.79588 |

## Generative Model

$$y_{obs} \sim Student - T(\mu = \hat{a}x + \hat{b}, \lambda = \frac{\mathbf{0.19283}}{x^2}, \nu = \mathbf{3.79588})$$

where $\hat{a}$, $\hat{b}$ are the estimated parameters