CSCE 633: Machine Learning

Lecture 6: Linear Regression

Texas A&M University
Bobak Mortazavi

Goals For This Lecture

- Motivate a simple supervised learning problem
- Introduce a linear machine learning method (Linear regression)
- Develop a Loss Function
- Ordinary Least Squares - Optimally solve the learning problem
- Interpret model
- Understanding Accuracy and Error
- Acknowledgements: example and figure sources: James, Witten, Hastie, Tibshirani (ISLR)

Least Absolutes

- The residual sum of absolutes

$$RSS = |e_1| + |e_2| \cdots |e_n|$$

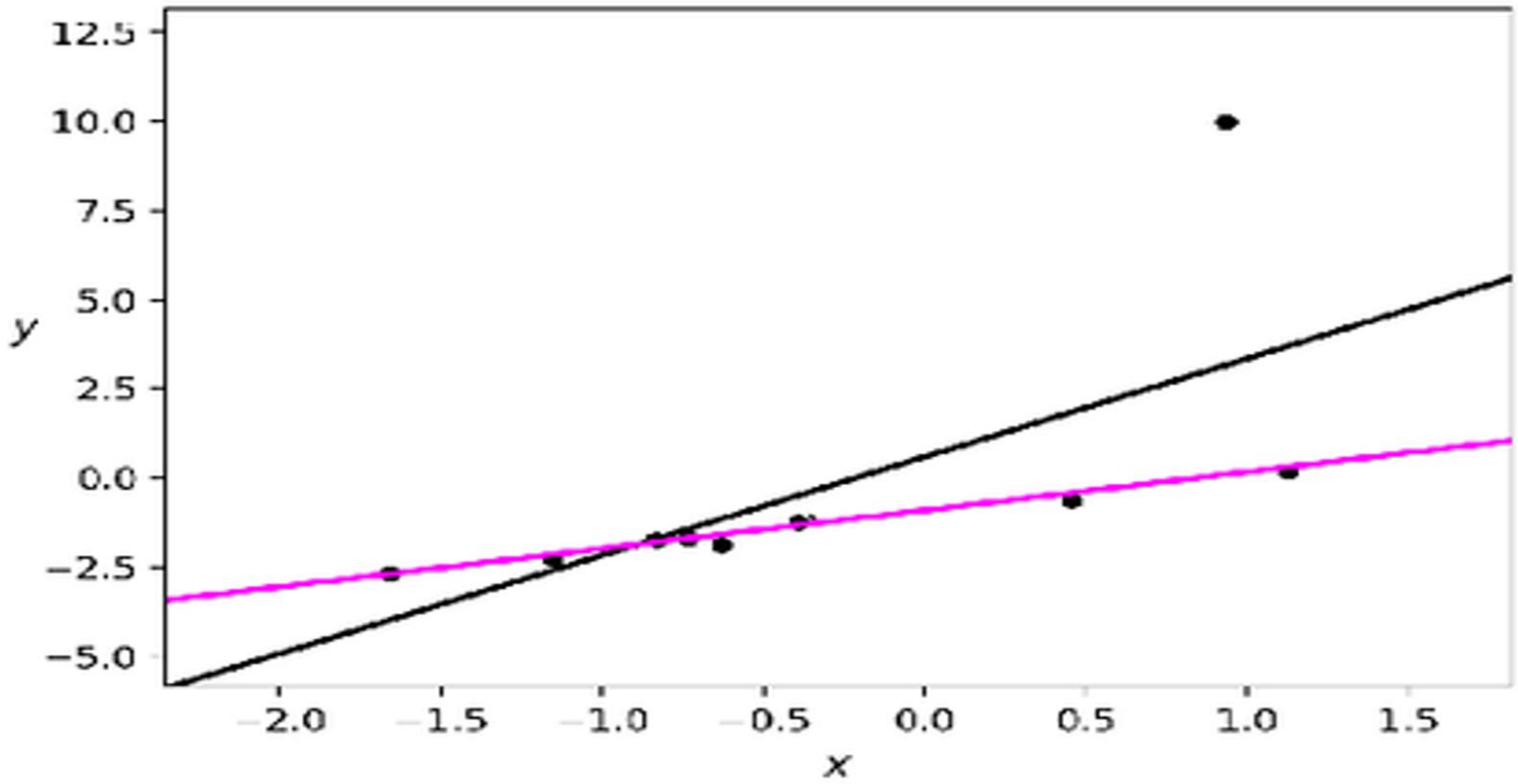
$$= |y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1| + |y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2| \cdots |y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n|$$

Least Absolutes

- Downside of least square cost:
  - Squaring errors larger than 1 emphasizes them
  - Forces the weights to minimize larger errors, typically those of outliers
  - Susceptible to overfitting to outliers
- Least absolute error partially addresses this problem

Least Absolutes

- Black line fitted using least squares
- Pink line fitted using least absolute



Accuracy of Coefficient Estimates

- Assume the true relationship is $\mathrm{Y} = f(X) + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$ (mean zero random error term)
- So, $\mathrm{Y} = \beta_0 + \beta_1 X + \epsilon$

Accuracy of Coefficient Estimates

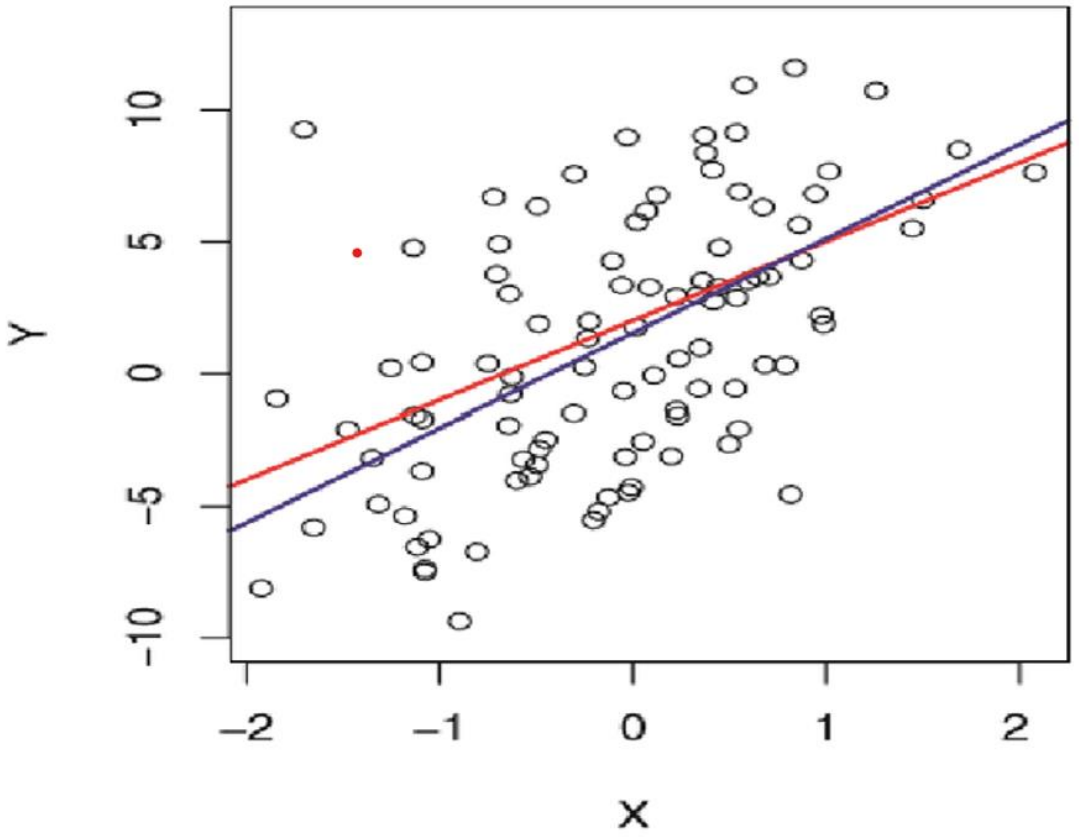
- Assume the true relationship is $\mathrm{Y} = f(X) + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$ (mean zero random error term)
- So, $\mathrm{Y} = \beta_0 + \beta_1 X + \epsilon$
- This is the population regression line which is the best linear approximation to the true relationship between X and Y.

Accuracy of Coefficient Estimates

- Assume the true relationship is $\mathrm{Y} = f(X) + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$ (mean zero random error term)
- So, $\mathrm{Y} = \beta_0 + \beta_1 X + \epsilon$
- This is the population regression line which is the best linear approximation to the true relationship between X and Y.
- Assume, for example $\mathrm{Y} = 2 + 3X + \epsilon$ and you sample this population with 100 random variables X to generate 100 Y.

Accuracy of Coefficient Estimates

- Assume, for example Y= $2 + 3X + \epsilon$ and you sample this population with 100 random variables X to generate 100 Y.



Accuracy of Coefficient Estimates

- Assume, for example Y= $2 + 3X + \epsilon$ and you sample this population with 100 random variables X to generate 100 Y.

- $\hat{\mu} = \bar{y}$ - sample mean from observations recorded is close with lots of sampling. Same $\hat{\beta}_0$ and $\hat{\beta}_1$ - is a good estimate with enough data.

- Linear regression versus estimation of the mean of a random variable leads to concept of bias.

Accuracy of Coefficient Estimates

- Assume, for example Y= $2 + 3X + \epsilon$ and you sample this population with 100 random variables X to generate 100 Y.

- $\hat{\mu} = \bar{y}$ - sample mean from observations recorded is close with lots of sampling. Same $\hat{\beta}_0$ and $\hat{\beta}_1$ - is a good estimate with enough data.

- Linear regression versus estimation of the mean of a random variable leads to concept of bias.

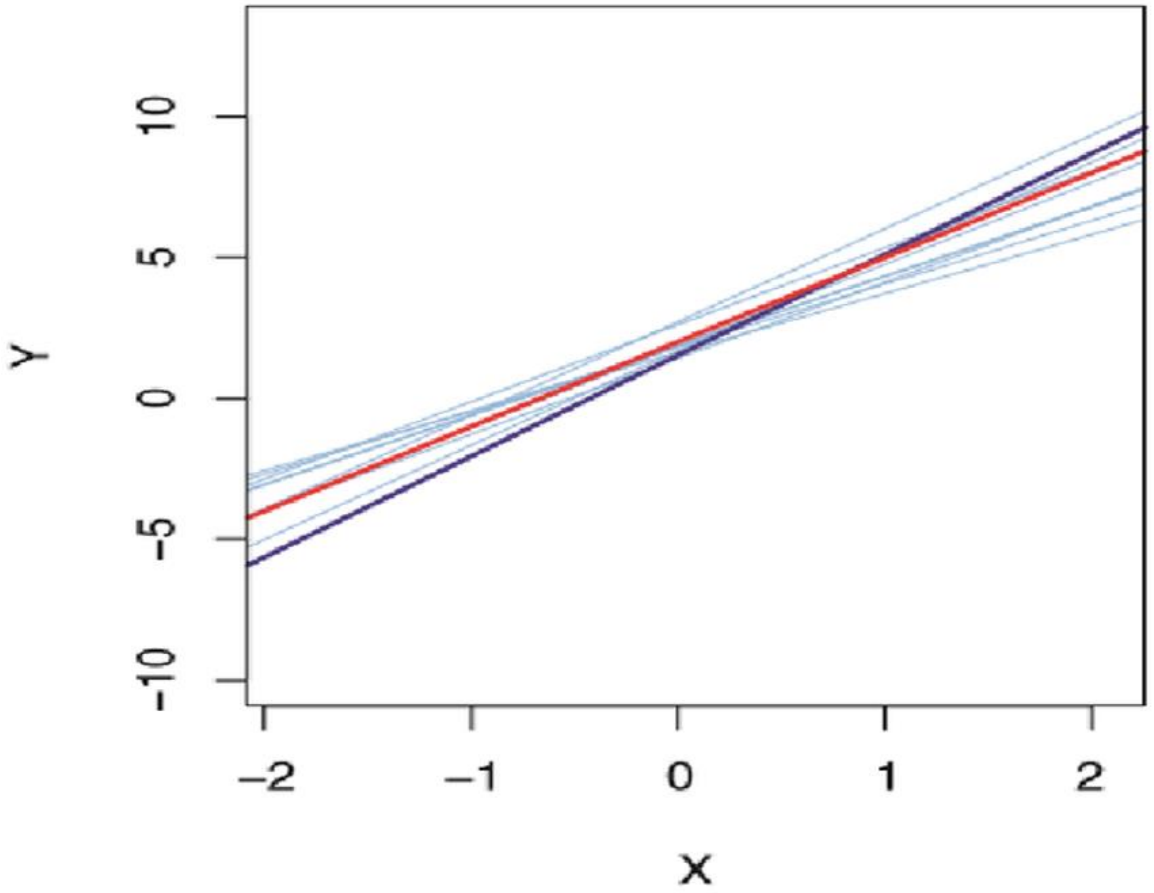
- If we use the sample mean $\hat{\mu}$ to estimate true μ, this is unbiased since, on average, we expect them to be the same.
  - One set of $y_1, y_2, \cdots, y_n$ might result in $\hat{\mu}$ that underestimates μ
  - Another that overestimates μ
  - etc

Accuracy of Coefficient Estimates

- Same with $\hat{\beta}_0$ and $\hat{\beta}_1$ - average enough samples and enough regressions to get to the true β_0 and β_1.

Accuracy of Coefficient Estimates

- Assume, for example Y= $2 + 3X + \epsilon$ and you sample this population with 100 random variables X to generate 100 Y – repeating the process



Accuracy of Coefficient Estimates

- Same with $\hat{\beta}_0$ and $\hat{\beta}_1$ - average enough samples and enough regressions to get to the true β_0 and β_1
- So, we ask, how accurate is the sample mean $\hat{\mu}$ from the estimate of μ – how far off is a single estimate?

Accuracy of Coefficient Estimates

- Same with $\hat{\beta}_0$ and $\hat{\beta}_1$ - average enough samples and enough regressions to get to the true β_0 and β_1
- So, we ask, how accurate is the sample mean $\hat{\mu}$ from the estimate of μ – how far off is a single estimate?
- We need to calculate the standard error of $\hat{\mu}$, $\mathrm{SE}(\hat{\mu})$

$$Var(\hat{\mu}) = \mathrm{SE}(\hat{\mu})^2 = \frac{\sigma^2}{n}$$

- Where σ^2 is the standard deviation of each of the realizations of y_i of Y (the n observations must be uncorrelated)
- Average amount $\hat{\mu}$ differs from μ – larger n, smaller error

Accuracy of Coefficient Estimates

- In the same vein – How close can we make $\hat{\beta}_0$ and $\hat{\beta}_1$ to β_0 and β_1?

Accuracy of Coefficient Estimates

- In the same vein – How close can we make $\hat{\beta}_0$ and $\hat{\beta}_1$ to β_0 and β_1?

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left(\frac{1}{N} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)$$

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \sigma^2 = Var(\epsilon)$$

- We assume ϵ_i are uncorrelated with common variance σ^2 (Often not true but a good approximation)

Accuracy of Coefficient Estimates

- In the same vein – How close can we make $\hat{\beta}_0$ and $\hat{\beta}_1$ to β_0 and β_1?

$$\mathrm{SE}\left(\hat{\beta}_0\right)^2 = \sigma^2 \left(\frac{1}{N} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right)$$

$$\mathrm{SE}\left(\hat{\beta}_1\right)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \sigma^2 = Var(\epsilon)$$

- We assume ϵ_i are uncorrelated with common variance σ^2 (Often not true but a good approximation)
- When x_i are spread out, and smaller, we have more leverage to estimate the slope, reducing $\mathrm{SE}\left(\hat{\beta}_1\right)$
- $SE\left(\hat{\beta}_0\right) = SE(\bar{\mu})$ if $\bar{x} = 0$

Accuracy of Coefficient Estimates

- In the same vein – How close can we make $\hat{\beta}_0$ and $\hat{\beta}_1$ to β_0 and β_1?

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left(\frac{1}{N} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right)$$

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \sigma^2 = Var(\epsilon)$$

- We assume ϵ_i are uncorrelated with common variance σ^2 (Often not true but a good approximation)
- When x_i are spread out, and smaller, we have more leverage to estimate the slope, reducing $\mathrm{SE}(\hat{\beta}_1)$
- $SE(\hat{\beta}_0) = SE(\bar{\mu})$ if $\bar{x} = 0$
- σ^2 is not known either but can be estimated from data. The estimate, σ is the residual standard error

$$RSE = \sqrt{\frac{RSS}{n-2}}$$

Coefficient Estimates: Confidence Intervals

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left(\frac{1}{N} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)$$

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \sigma^2 = Var(\epsilon)$$

$$\hat{\beta} \pm SE(\hat{\beta})$$

Hypothesis Testing

- Standard Errors let us hypothesis test
- Most common is the Null Hypothesis
- H_0: There is no relation between X and Y
- Alternatively, we have H_a: There is some relationship between X and Y
- Mathematically, this is like testing
  - $H_0: \beta_1 = 0$ Therefore Y= $\beta_0 + \epsilon$
  - $H_a: \beta_1 \neq 0$ therefore determine that $\hat{\beta}_1$ is sufficiently far from 0
- The important question becomes – how far is far enough?

T-Statistic

- T-statistic $t_\beta = \frac{\widehat{\beta}_1 - \beta}{SE(\widehat{\beta}_1)}$
- T-statistic $t_\beta = \frac{\widehat{\beta}_1 - 0}{SE(\widehat{\beta}_1)}$ for H_0

T-Statistic

- T-statistic $t_\beta = \frac{\hat{\beta}_1 - \beta}{SE(\hat{\beta}_1)}$
- T-statistic $t_\beta = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$ for H_0
- If no relationship between X and Y exists, we expect a t-distribution with n - 2 degrees of freedom
- Compute the probability of observing any number equal to t_β or larger in absolute value, assuming $\beta_1 = 0$
- This probability is called the p-value
- A small p-value – it is unlikely to observe a substantial association between predictor and response due to chance
- Therefore, a small p-value means there is an association between X and Y so we can reject the null hypothesis
- The cutoff is usually 5% or 1%

Advertising Example

- If n = 30

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

With n = 30 the t-statistic for the null hypothesis are around 2 and 2.75 respectively.

We conclude $\beta_0 \neq 0$ and $\beta_1 \neq 0$

Important Questions to Ask

- Is there a relationship between budget and sales?
- If there is a relationship, how strong is it?
- Which of the three media contribute to sales?
- How accurately can we estimate the effect of each medium on sales?
- Is the relationship linear?
- Is there synergy among the advertising media?

Accuracy of Simple Linear Regression

- Once we reject the null hypothesis for w_0 and w_1, it is natural to ask how well the model fits the data
- One measure is the residual standard error

$$RSE = \sqrt{\frac{RSS}{n-2}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- Measure of lack of fit, it is an absolute measure. It is not always clear what a good value of RSE is.
- Another possible measurement is the R^2 statistic

R^2 Statistic

- Proportion of variance explained, always between 0 and 1, independent of scale of Y
- Total sum of squares $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

R^2 Statistic

- Proportion of variance explained, always between 0 and 1, independent of scale of Y
- Total sum of squares $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- TSS measures the total variance in response Y (amount inherent in response before the regression is performed)
- RSS amount left unexplained after the regression

R^2 Statistic

- $R^2 = \frac{TSS-RSS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$

- R^2 is the proportion of variability in Y that can be explained using X.
- R^2 close to 1 – large proportion of variation explained by the regression
- R^2 close to 0 – regression does not explain the variation – perhaps because model is wrong. σ^2 is too high, or possibly both?
- R^2 is a measure of the linear relationship between X and Y
- Still. What is a good value for R^2

R^2 Statistic: Correlation

$$Cor(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- This is also a measure of the linear relationship between X and Y
- $r = Cor(X,Y)$
- In simple linear regression, $R^2 = r^2$. In multiple regression however r^2 does not extend.

Takeaways

- Understanding key notation
- Important questions to ask for supervised learning problem
- Ordinary Least Squares
- Simple Linear Regression
- Optimizing RSS
- Next Time: Multiple Linear Regression and Coding