Regression Refresher

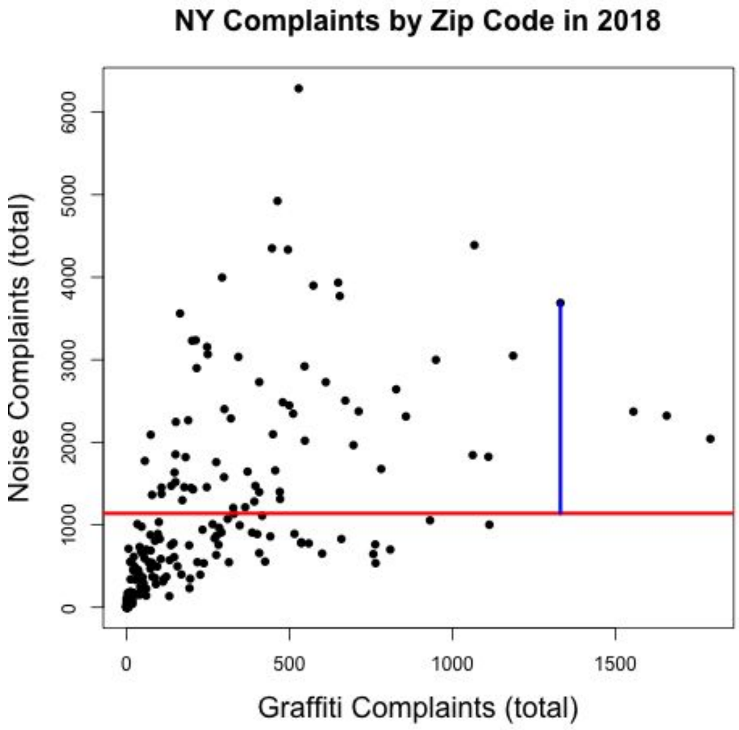
Lecture 5

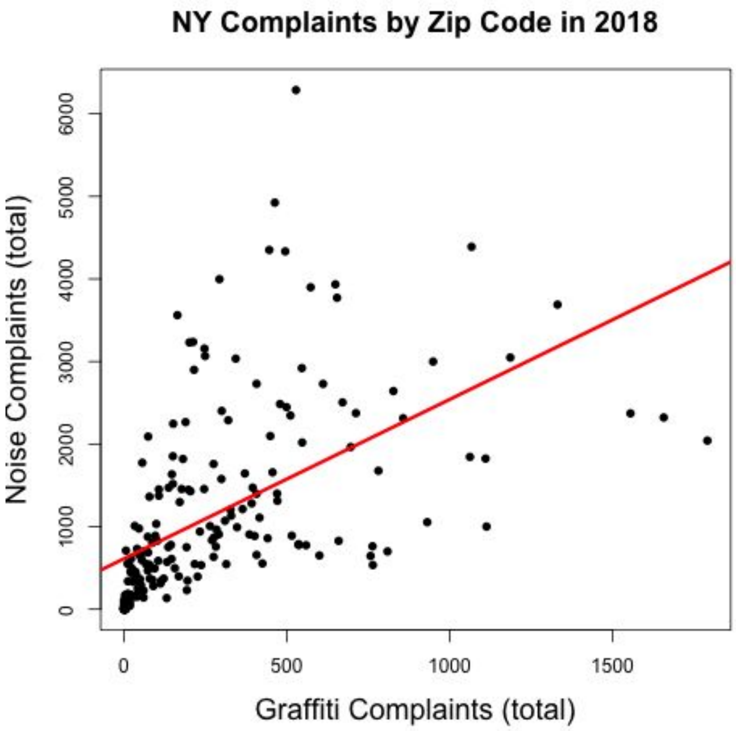
Xiaofei Shi

Learning objective

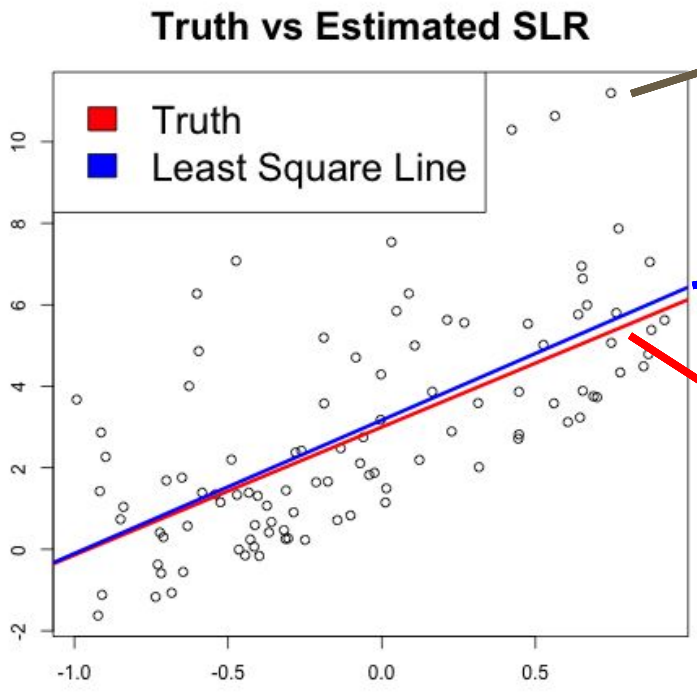
- Crash course on regression with data
  - Prediction
  - Estimation
  - Why multivariate regression is different

Regression is the “best-fit” line





Regression is a model for data



$$(x_i, y_i)$$

$$\hat{\beta}_0 + \hat{\beta}_1 * X$$

- Estimated model
- Fitted model
- Regression function

$$\beta_0 + \beta_1 * X$$

- Truth
- True model
- (unobservable!)

Regression is the first example for statistical inference

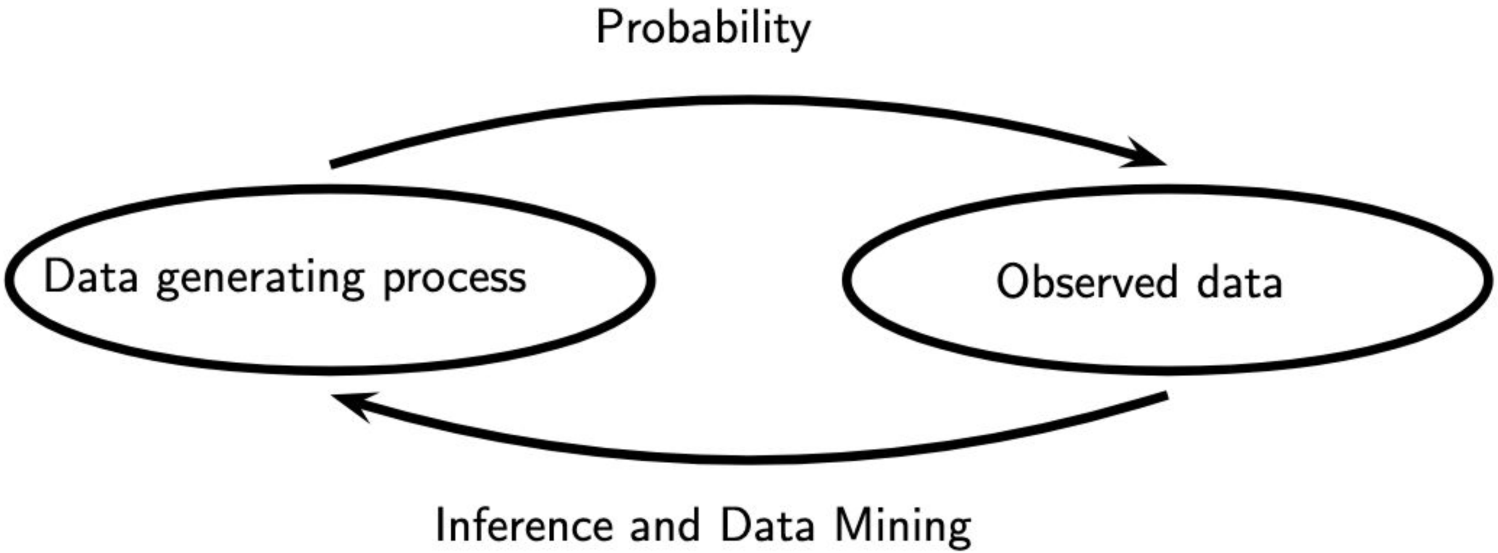


Image from “All of Statistics” by Larry Wasserman

Regression is used for prediction and inference

Prediction

Fitting a line to arbitrary point cloud only requires an objective to minimize

$$\arg\min_{a,b} \sum_i |Y_i - (a + bX_i)|^2$$

Inference

Using regression to estimate parameters for a statistical model for the data:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$$\hat{\beta}_0 \approx \beta_0$$

$$\hat{\beta}_1 \approx \beta_1$$

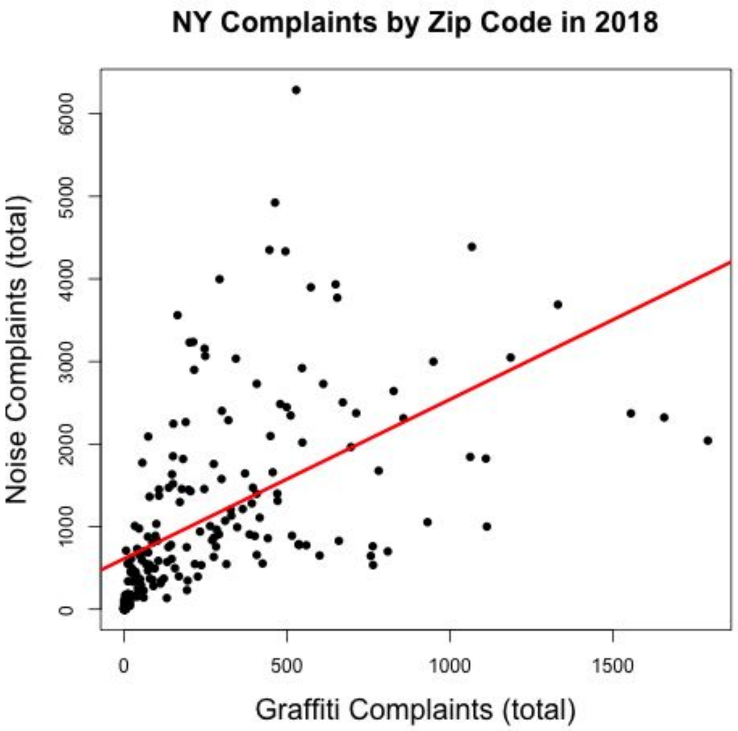
Relationship between inference and prediction

- Good inference would imply good prediction

- Good prediction can happen with bad inference

Assumptions for prediction?

$$\arg\min_{a,b} \sum_i |Y_i - (a + bX_i)|^2$$



Assumptions for unbiased estimates?

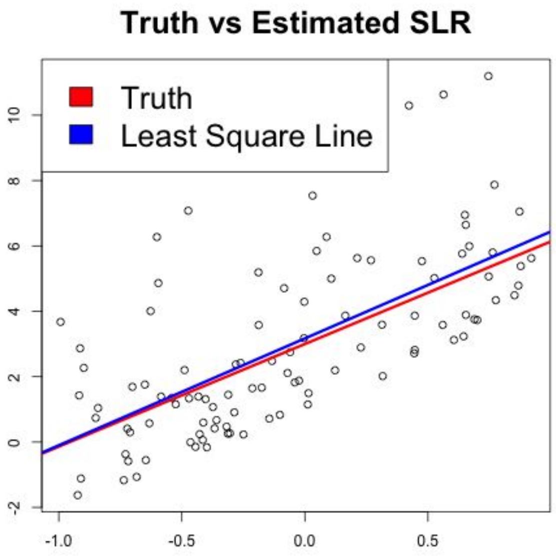
If we have the following assumptions

- Linearity: $y_i = \beta_0 + \beta_1 * x_i + \epsilon_i$
- $E(\epsilon_i|X) = 0$ for every i

Then we have

- $E(\hat{\beta}_0|X) = \beta_0$
- $E(\hat{\beta}_1|X) = \beta_1$

where $\hat{\beta}_i$ are the estimates derived using regression.



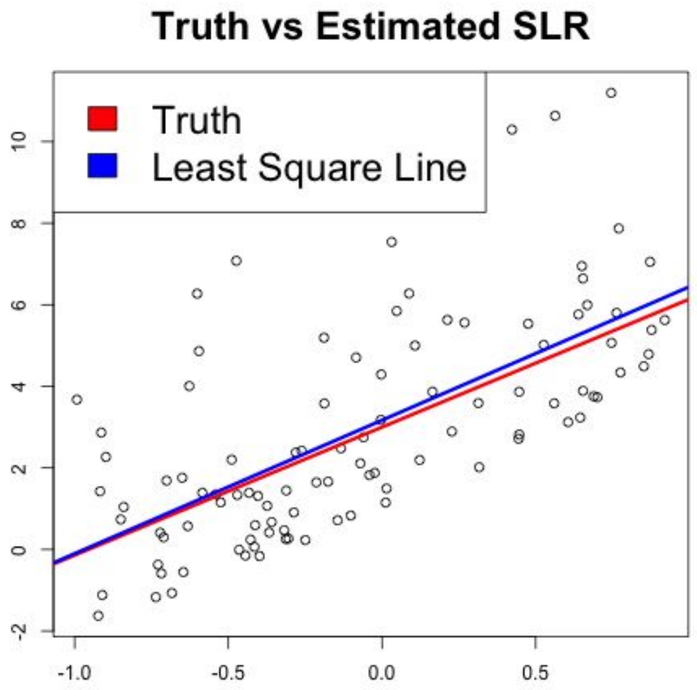
Assumptions for standard errors?

If we, in addition, have the assumption that:

- ϵ are independent from one another
- $Var(\epsilon_i|X) = \sigma^2$ for every i

Then we also have analytical solutions for the variance:

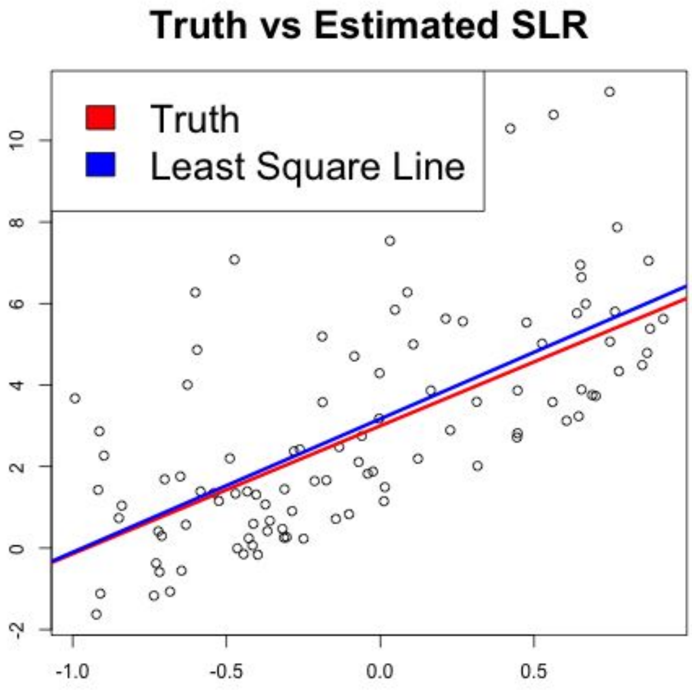
- $Var(\hat{\beta}_0|X) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{(\sum(x_i - \bar{X})^2}\right]$
- $Var(\hat{\beta}_1|X) = \frac{\sigma^2}{\sum(x_i - \bar{X})^2}$



Assumptions for p-values?

For p-values or confidence intervals, we need one of:

- Large sample
- Normal errors (Normally distributed data)



R Code for regression

```r
reg_model <- lm(y ~ x)
summary(reg_model)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.9078     1.0358  10.531  < 2e-16 ***
x            -1.6248     0.1746  -9.305 3.91e-15 ***
```

Meaning of coefficients?

```r
reg_model <- lm(y ~ x)
summary(reg_model)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.9078     1.0358  10.531  < 2e-16 ***
x            -1.6248     0.1746  -9.305 3.91e-15 ***
```

Requirements drive model validation efforts

To validate each assumption, what would you ask for?

Exercise - Linking problems to the fitted regression

Understand altitude and temperature drop?

Exercise - Linking problems to the fitted regression

Understand the impact of having a Chinese name on your resume?

Summary

- The problem determines how to use the model
- Trust in the model depends on the validation
    - Validation can also suggest improvements with the model!
- Validation should be driven by mathematical guarantees