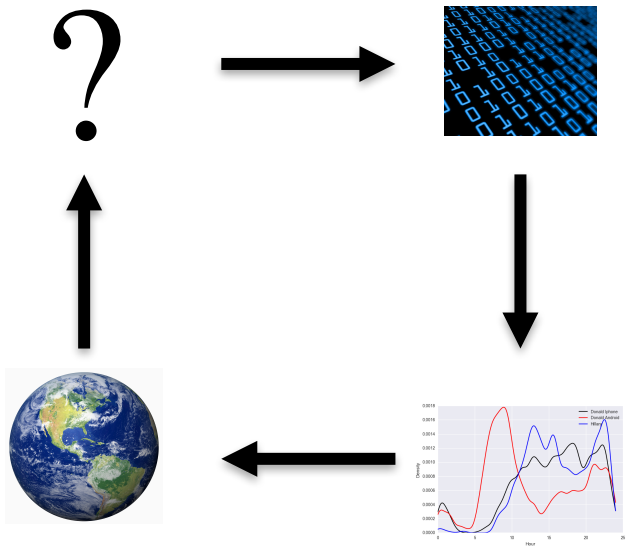
Data Science 100
Lec 18: Inference using LS and Linear Regression Model

DS100
Spring
2017

Slides by:
Bin Yu
binyu@stat.berkeley.edu
Joey Gonzalez
jegonzal@berkeley.edu
Thanks to Andrew Do for his assistance on data analysis



Recall from last lecture

- Represent data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ as:

Covariate (Design) Matrix

$$X = \begin{bmatrix} \text{—} & x_1 & \text{—} \\ \text{—} & x_2 & \text{—} \\ & \cdots & \\ \text{—} & x_n & \text{—} \end{bmatrix} \in \mathbb{R}^{np}$$

n

p

Response Vector

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^{n}$$

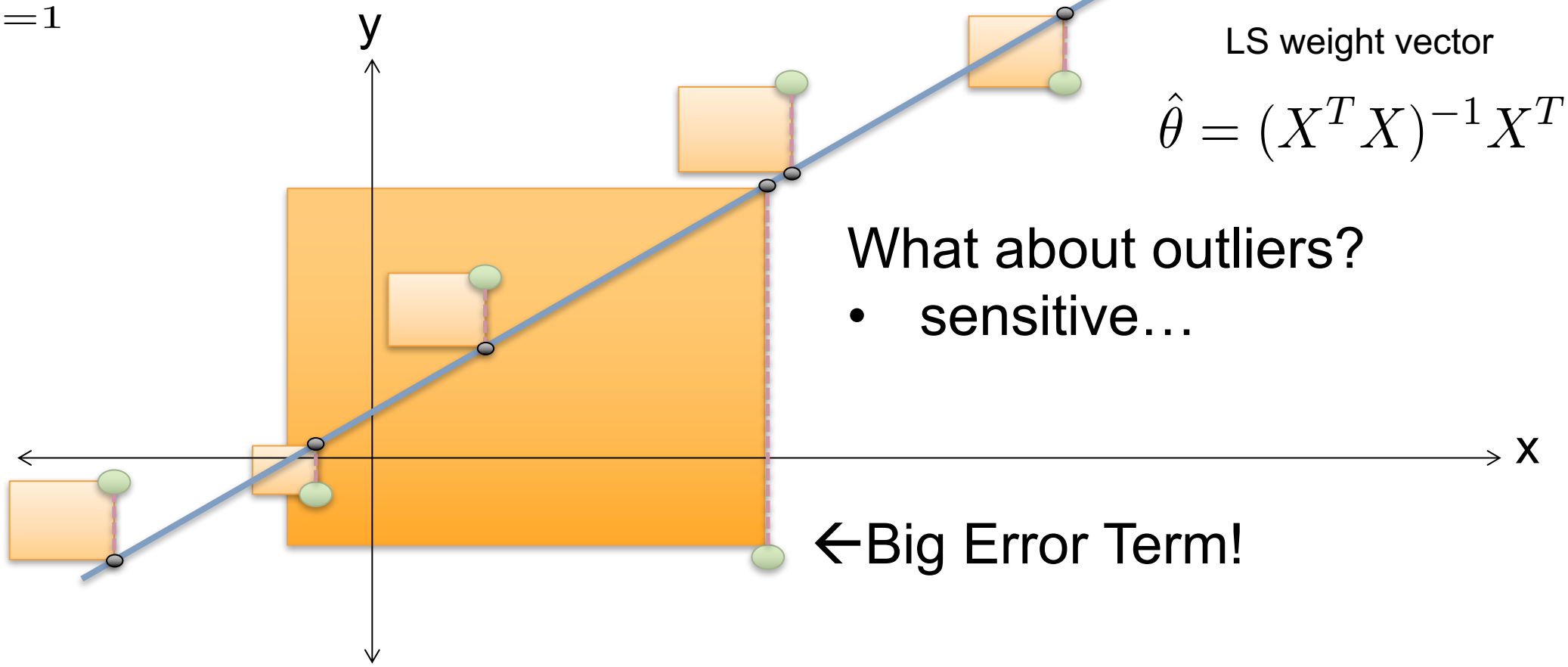
n

1

Least Squares (LS) predictor: $\hat{\theta}^T x$

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \theta^T x_i)^2$$

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta^T x_i)^2$$



LS weight vector

$$\hat{\theta} = (X^T X)^{-1} X^T Y$$

What about outliers?

- sensitive…

←Big Error Term!

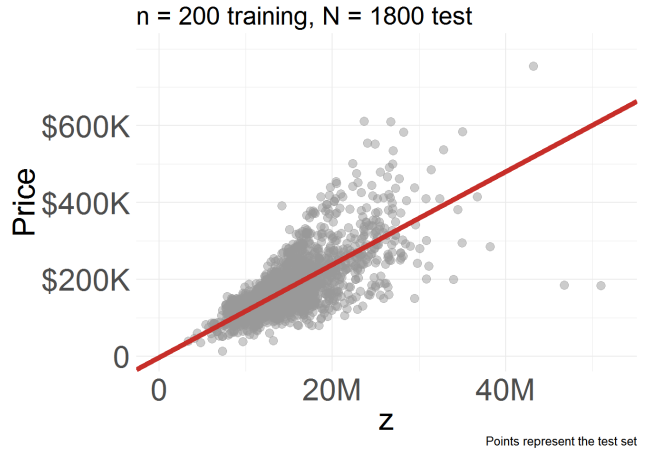
New training data with expensive houses added

LS fit with a new predictor z and intercept (p=2)

LS fit: y = - 2749 + 0.0121 z ($)

y= house price

Q: Do you think z is important for predicting house price? Or is 0.0121 really different from 0?



RMS = Root Mean Square
= $ 53.8K

$$RMS = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{\theta}^T x_i)^2}{n-p}}$$

n=200 training data size
p=2

Assessment of “importance” via pseudo replicates

- Ideally, we want to replicate or random sample from the population again and see how the slope 0.0121 in the LS weight vector changes

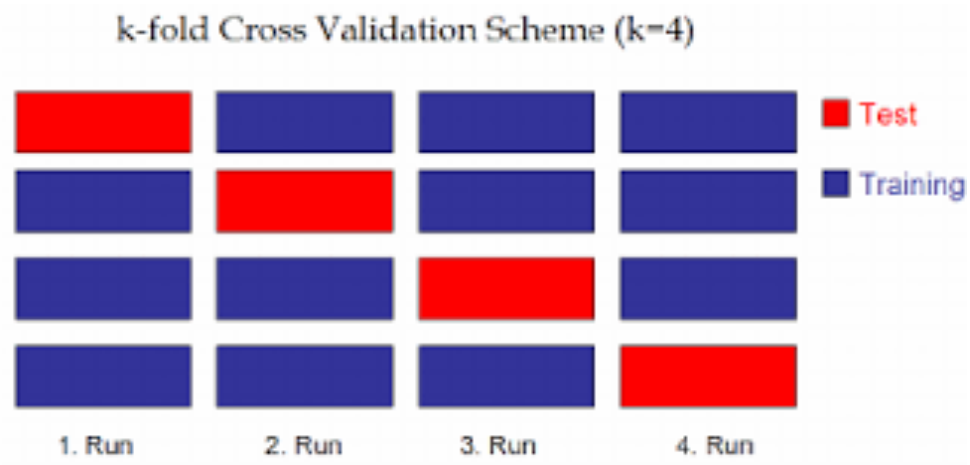
- Most of time, we have to content with what we have, and use pseudo replicates by “perturbing data” appropriately so that the pseudo-replicates captures the essential characteristics of the original data, or they are “representative” of the original data

Stability pseudo replicates –

intuitive and simple, and also one focus of Bin’s research group

k-fold cross-validation pseudo-replicates

- Cross-validation (CV) idea: create k prediction problems within a data set



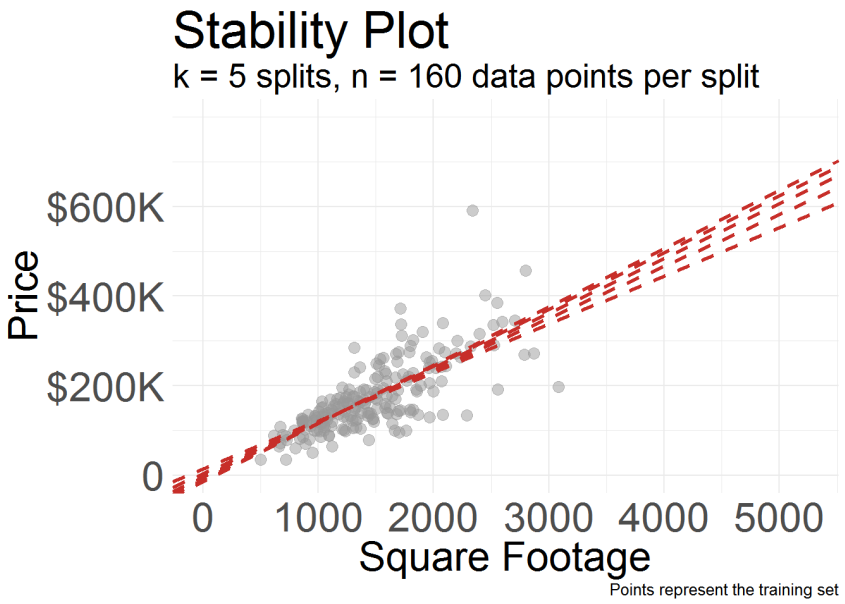
- Pseudo-replication of k prediction problems: they should be similar to each other and to the original problem – “representativeness”

- When are CV pseudo-replicates are not “representative” of the original data?

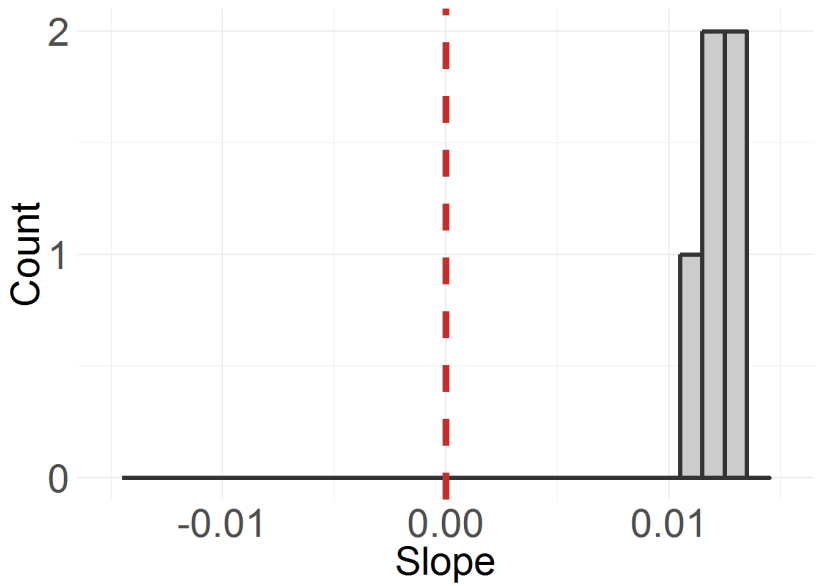
Results from 5-fold CV pseudo-replicates

- Cross-validation (CV) idea: create k prediction problems within a data set
- k=5, we got 0.0122170, 0.0130000, 0.0116000, 0.0126993, 0.0107874

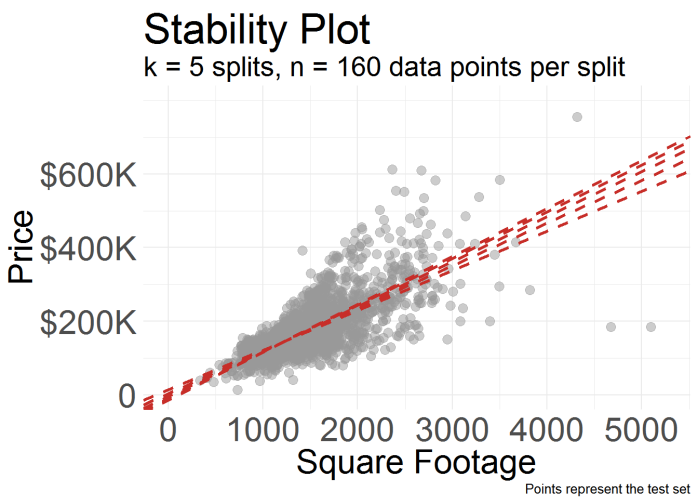
Is the slope really zero?



Training set



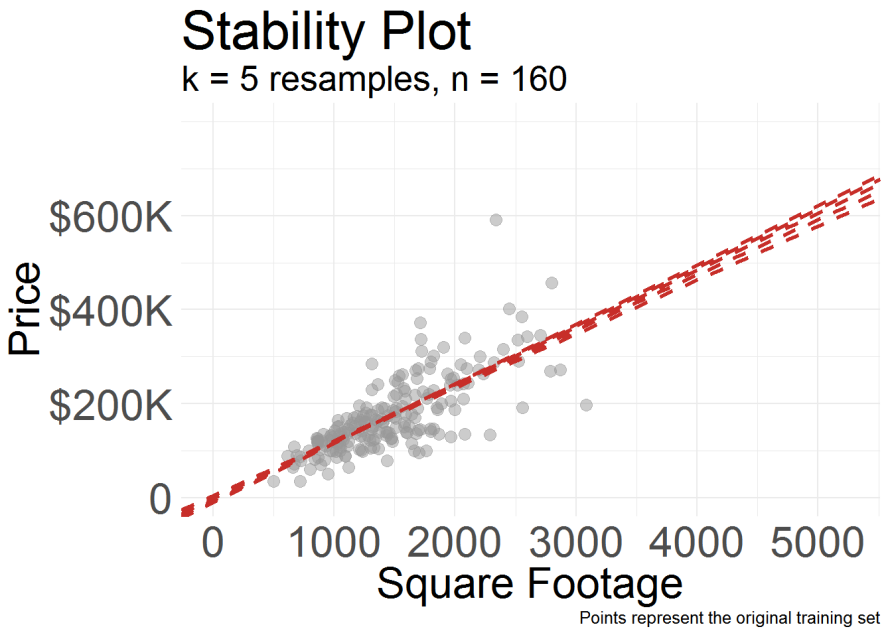
From training

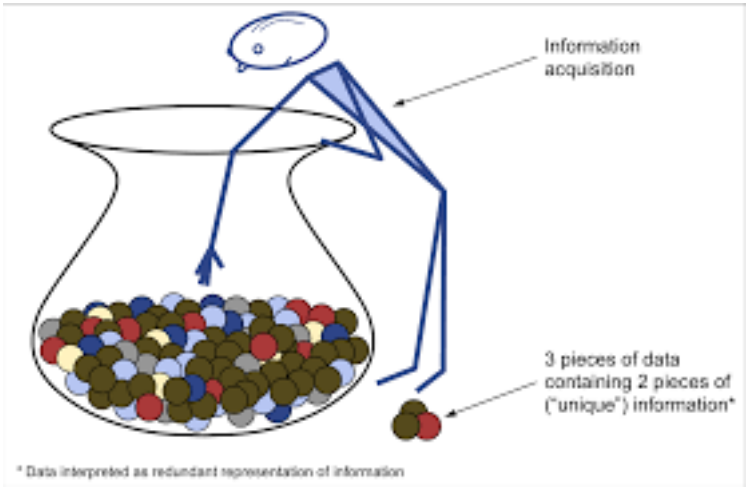


Test set

Another kind of pseudo-replicates: random splits or random sampling of 80% without replacement

- k=5, random splits

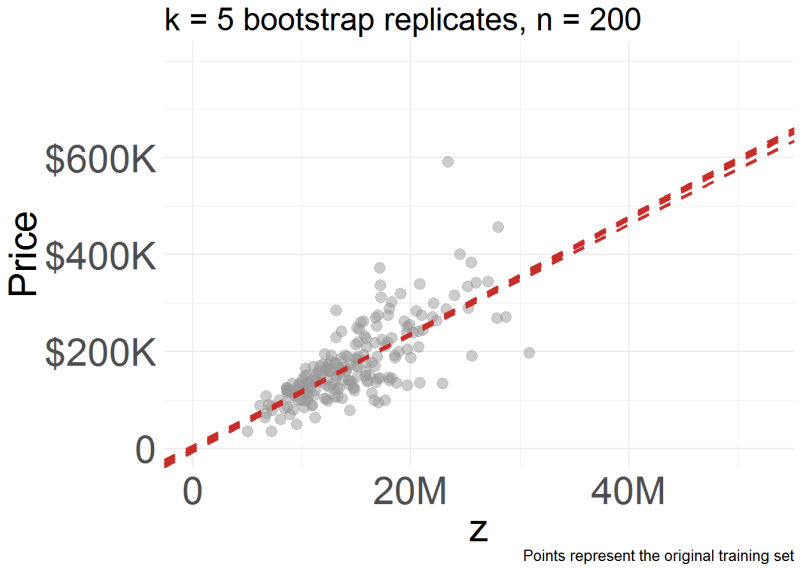


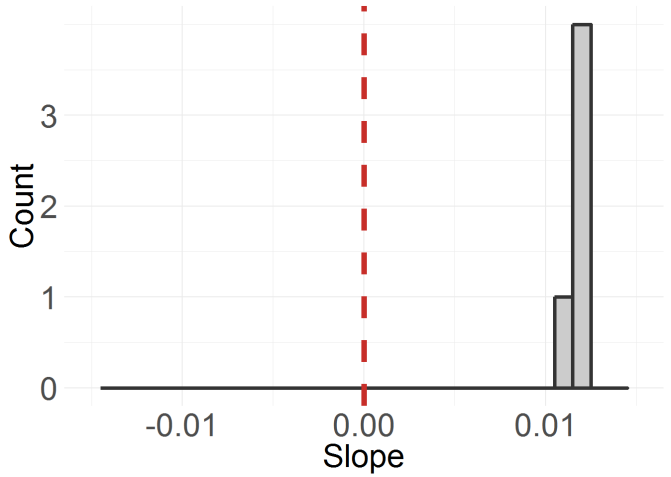


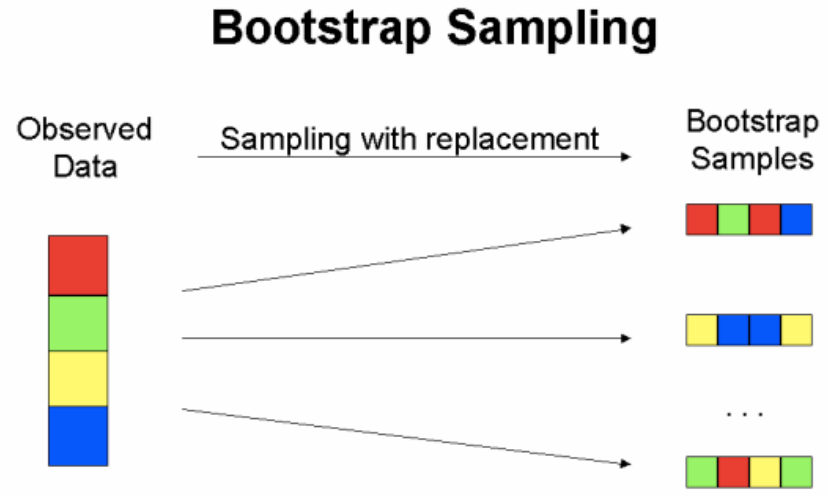
Is the slope really zero?

Yet another kind of pseudo-replicates: bootstrap or random sampling of size n with replacement

- k=5, bootstrap





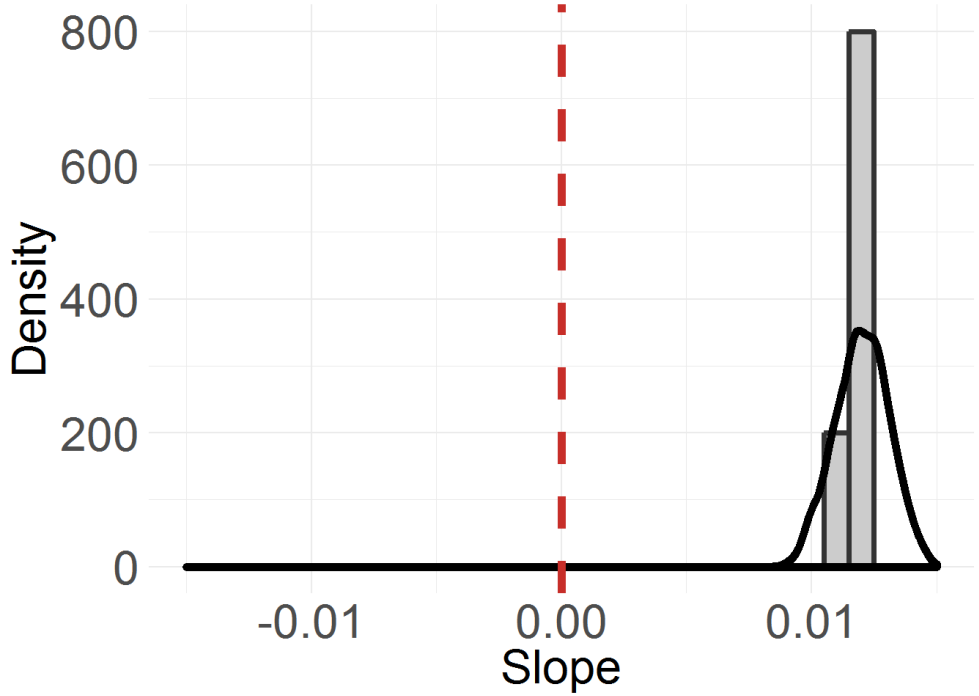


Is the slope really zero?

(Full-blown) Bootstrap replicates

- The above three pseudo-replicate methods provide useful info. on sampling variability in a concise manner

- If we increase the bootstrap replication number to 100, we get a fuller picture on the sampling variability at much higher computation cost (20 times)

- 

kernel density (local smoothing of histogram) of bootstrapped 100 slopes, imposed on the 5 bootstrapped slopes earlier.

Do think think the real slope is zero?

No… what do we mean by real slope?

How was training data collected?

- We sampled 200 out of 3000 houses to make the samples more or less independent and identically distributed (iid) because every time, the same is drawn from approximately the same population since 200/3000 =1/15 is small.

Defining “real” or “true” slope

- The real slope is about the population data:

  all houses in Ames from 2006-2010 with house prices from the tax office
  (is this population clearly defined?)

  The “real” slope is the LS slope based on the population data.

LS is often called (mistakenly, in my view) a Linear Regression method

LS is simply a prediction method and it can evaluated by prediction performance (the data generating mechanism could be non-linear, non-indep. noise term, etc) – if it works, it works… But

How does LS relate to normal linear regression model?
What is Normal Linear Regression Model?

Normal Linear Regression Model – idealized but useful

All models are wrong, but some are useful – George Box

Vector of Parameters

Vector of Features

Real Valued Observations

$$y_i = \theta^T x_i + \epsilon_i$$

Real Value Noise

Linear Combination of Covariates

$$\sum_{j=1}^{p} \theta_j x_{ij}$$

$$\theta, x \in \mathbb{R}^p$$

ϵ are independent, and independent of X, and with dist. $N(0, \sigma^2)$

$$\epsilon = (\epsilon_1, ..., \epsilon_n)^T$$

Maximum likelihood Est. under the Normal Linear Regression Model is the same as LS

- Since $y_i \sim N(\theta^T x_i, \sigma^2)$ and they are indep, the likelihood function is

$$\prod_{i=1}^{n} \{ \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-(y_i - \theta^T x_i)^2/(2\sigma^2)\right] \}$$

- The log likelihood function is

$$-\frac{n}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{n}(y_i - \theta^T x_i)^2/(2\sigma^2)$$

- Maximizing the above for θ is the same as minimizing the squared sum term or LS regardless of the value of variance σ^2

Estimating variance σ^2

- HW: Prove that the maximum likelihood estimator (MLE) of σ^2 is

$$\hat{\sigma}_1^2 = \sum_{i=1}^{n} (y_i - \hat{\theta}^T x_i)^2 / n$$

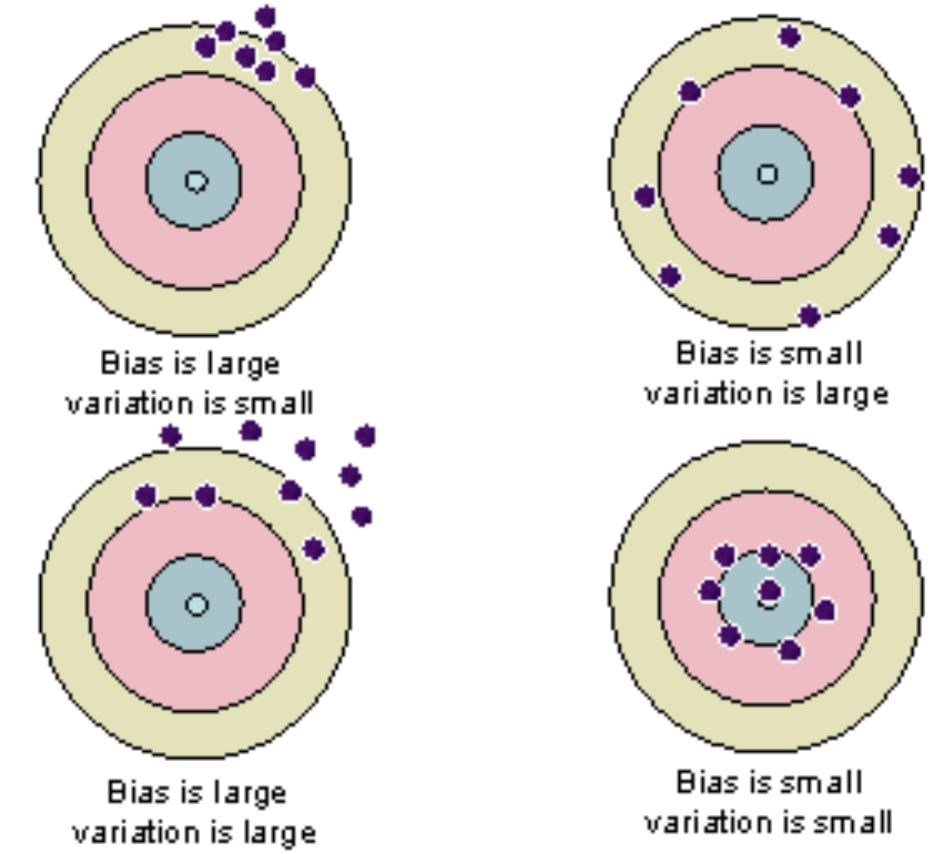
where $\hat{\theta} = (X^T X)^{-1} X^T Y$ is the MLE or LS (or OLS) estimator of θ

- In fact, we always use another variance estimator

$$\hat{\sigma}^2 = \sum_{i=1}^{n} (y_i - \hat{\theta}^T x_i)^2 / (n - p)$$

Unbiased estimators under linear model

- $\hat{\sigma}^2$ is an unbiased estimator of σ^2
- $\hat{\theta}$ is an unbiased estimator of θ



Accuracy versus Quality of an Estimator Using Bias and Variation as Measurable Quantities Respectively

Estimating variance σ^2

- Fact 1: under the Normal Linear Regression model, the maximum likelihood estimator (MLE) of σ^2 is

$$\hat{\sigma}_1^2 = \sum_{i=1}^{n} (y_i - \hat{\theta}^T x_i)^2 / n$$

where $\hat{\theta} = (X^T X)^{-1} X^T Y$ is the MLE or LS (or OLS) estimator of θ

- Fact 2: under the Gaussian (or Normal) Linear Regression model, LS has a distribution $N(\theta, \sigma^2 (X^T X)^{-1})$

Using $\hat{\sigma}^2$ in place of σ^2, we get confidence intervals and regions for θ

Linear Regression Model – idealized but useful

All models are wrong, but some are useful – George Box

Vector of Parameters

Vector of Features

Real Valued Observations

$$y_i = \theta^T x_i + \epsilon_i$$

Real Value Noise

Linear Combination of Covariates

$$\sum_{j=1}^{p} \theta_j x_{ij}$$

$$\theta, x \in \mathbb{R}^p$$

ϵ are independent, and independent of X, and with mean 0 and var. σ^2

$$\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$$

(ϵ 's distribution's tail is not too heavy)

Facts under Linear Regression Model

- LS estimator or weight vector is still unbiased and so is $\hat{\sigma}^2$

- When **n is large**, LS estimator has an **approximate normal distribution**

$$N(\theta, \sigma^2(X^T X)^{-1})$$

provided that $X^T X/n$ **is approximately a positive definite matrix.**

- No maximum likelihood estimation for this model, which does not specify enough information for us to write down a likelihood function

What does a 95% confidence interval for θ_1 mean?

- Given 100 sets of n random samples (or 100 replicates) following the Normal Linear Regression Model, we get 100 confidence interval for θ_1 as describe earlier, about 95 of them should cover the true θ_1 in the model.

- For the particular interval that you have on a particular set of data, it might not cover the true θ_1

- Confidence interval is one form of statistical inference since it provides an uncertainty measure about a data result or a point estimate

For one data set, bootstrap comes to rescue for confidence interval construction

- We need 100 or more bootstrap replicates to do the job!
- The concise 3 stability versions will NOT do…

Hypothesis testing

- Null hypothesis H_0 : slope θ_1 =0
- Alternative hypothesis H_1: $\theta_1 \neq 0$ (or $\theta_1 > 0$)
- Rationale: null hypothesis is favored unless there is strong evidence to refute it. This asymmetry makes a lot of sense when the null corresponds to an established scientific theory. Often in today's use of statistics, it is not the case. We need to keep in mind that **accepting the null is not the same as proving null**.
- Recall the example of testing the null that a coin is a fair coin based on only one toss. Null will be always accepted, but we know we can not prove that the coin is fair with one toss.

Terminologies in hypothesis testing

- Type I error = Probability that the null rejected when that null is true
- Type II error = Probability that the alternative is rejected when it is true

- Power = 1- Type II error = Prob. that the alternative is corrected accepted

In medical sciences (these terms below sound better or more positive)

- Sensitivity = 1- Type II error = Power
- Specificity (or selectivity) = 1- Type I error

- Setting type I error at 5% -- statistically significant
- Setting type I error at 1% -- highly statistically significant

Testing hypothesis via confidence interval construction

- Use a 95% confidence interval construction to test the null:

  if the interval contains the 0, the null is accepted;
  otherwise rejected – all at level 5%

Two confidence interval constructions

- (Full-blown) Bootstrap

- Approximate normality

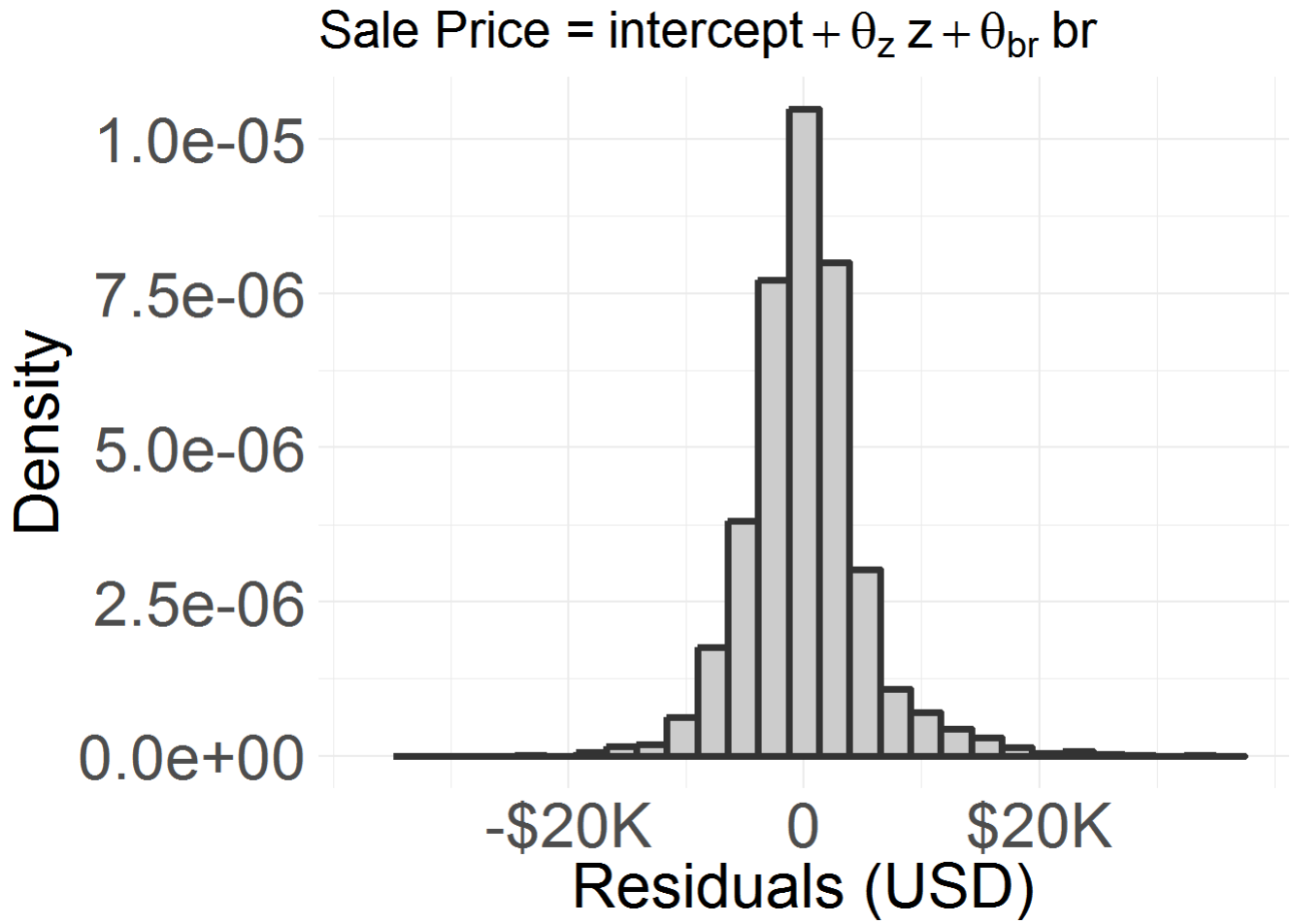
Does our housing data follow a normal linear reg. model? A linear reg. model?

- Use “true” weights from the population to investigate to get

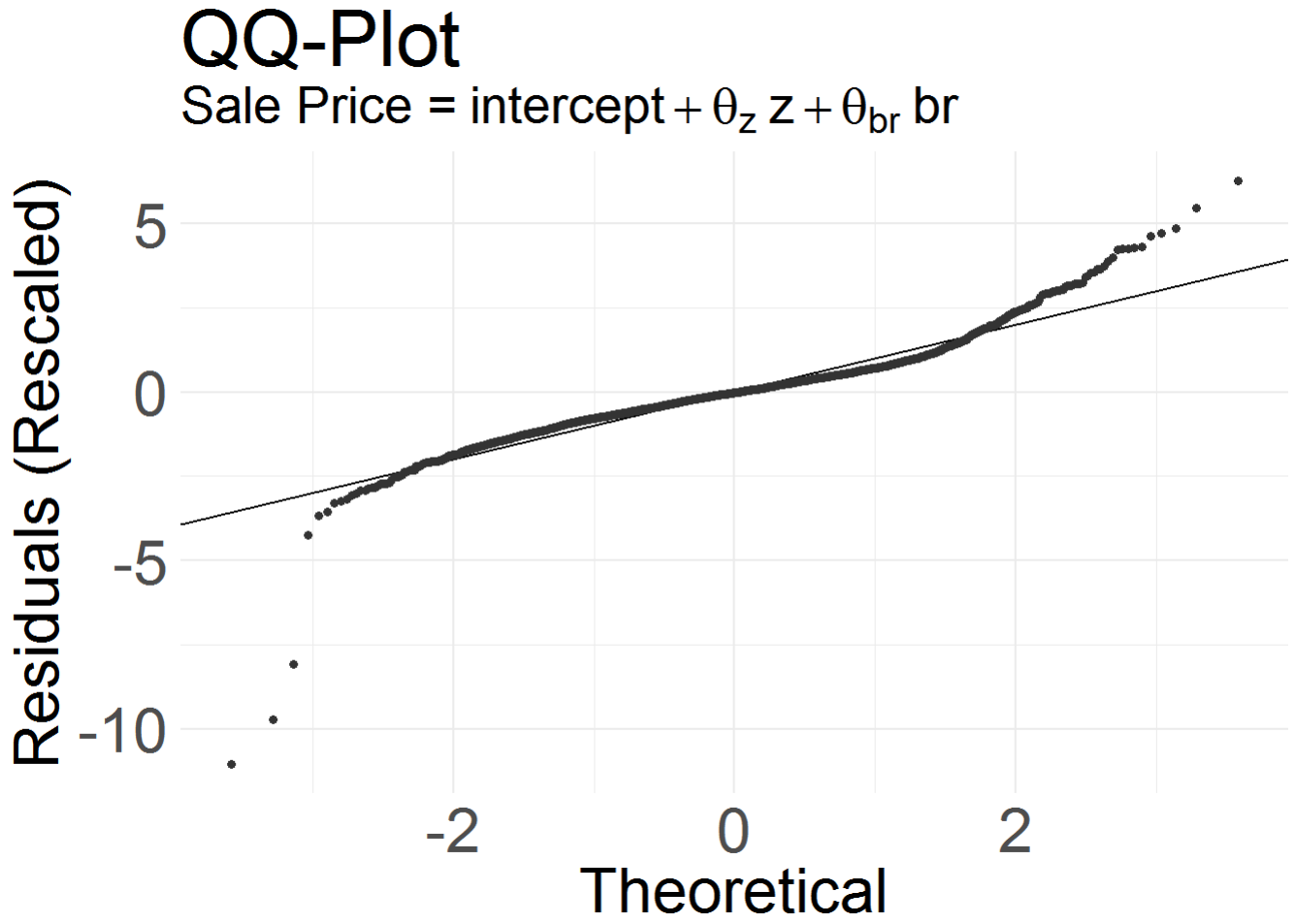
$$u_i = y_i - \theta^T x_i$$

- Q1: Are u_i normally distributed with the same variance?
- Q2: Are u_i independent of x_i?

Histogram of oracle noise terms u's: true weight for z=0.01, true weight for br = -29K



Q-Q plot of oracle noise terms or u's – look at the tails, not normal



Two predictor case with z: confidence intervals true weights =(0.0136, -0.29)

- **Bootstrap** 95% confidence interval:

  (0.012, 0.017) – covers the true weight for z ($/(0.01 ft)^2)

  (-0.33K, -0.13K) – covers the true weight for br. (unit: $/br)

- **Normal approximation** interval (oracle noise terms are not normal, but not too heavy tails. Linear Reg. model holds approximately).

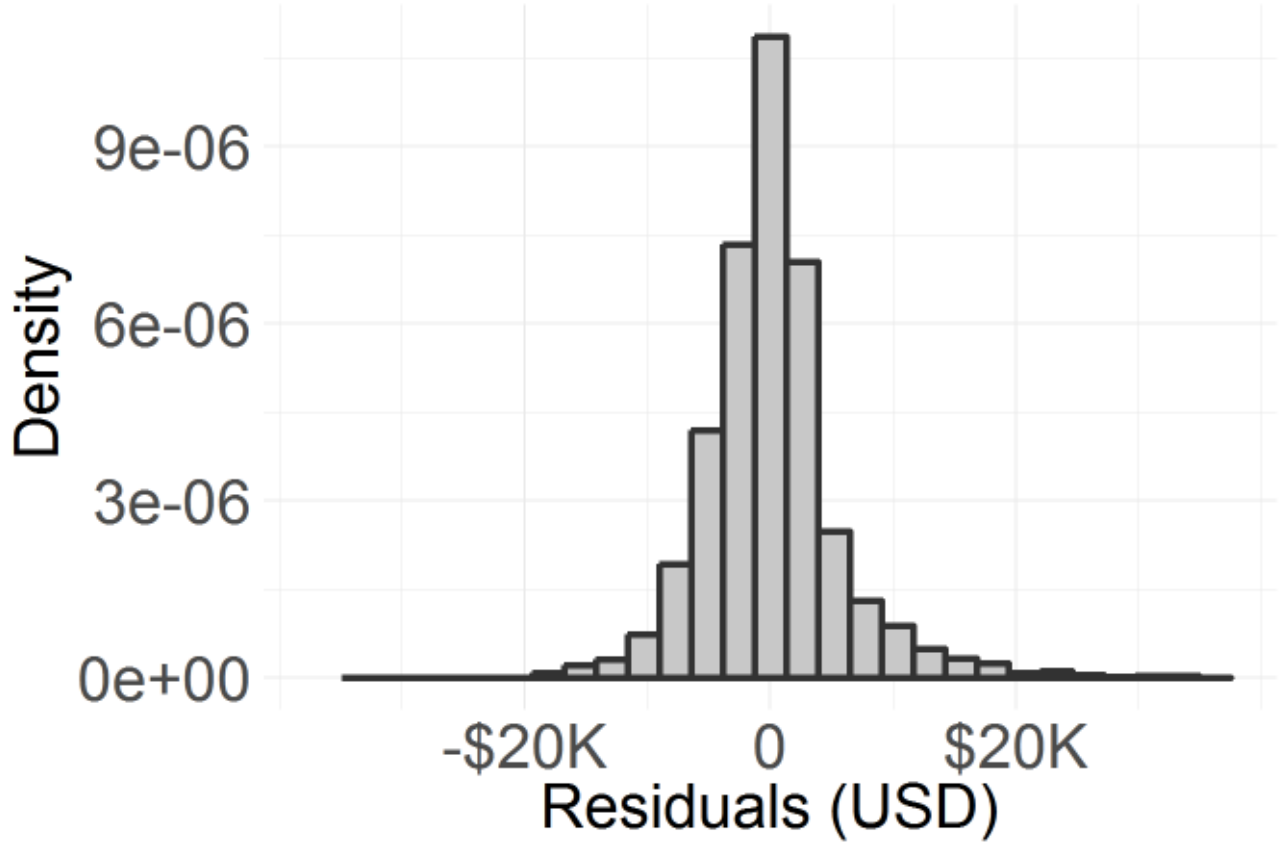
  95% confidence interval:

  (0.13, 0.016), covers the true slope barely

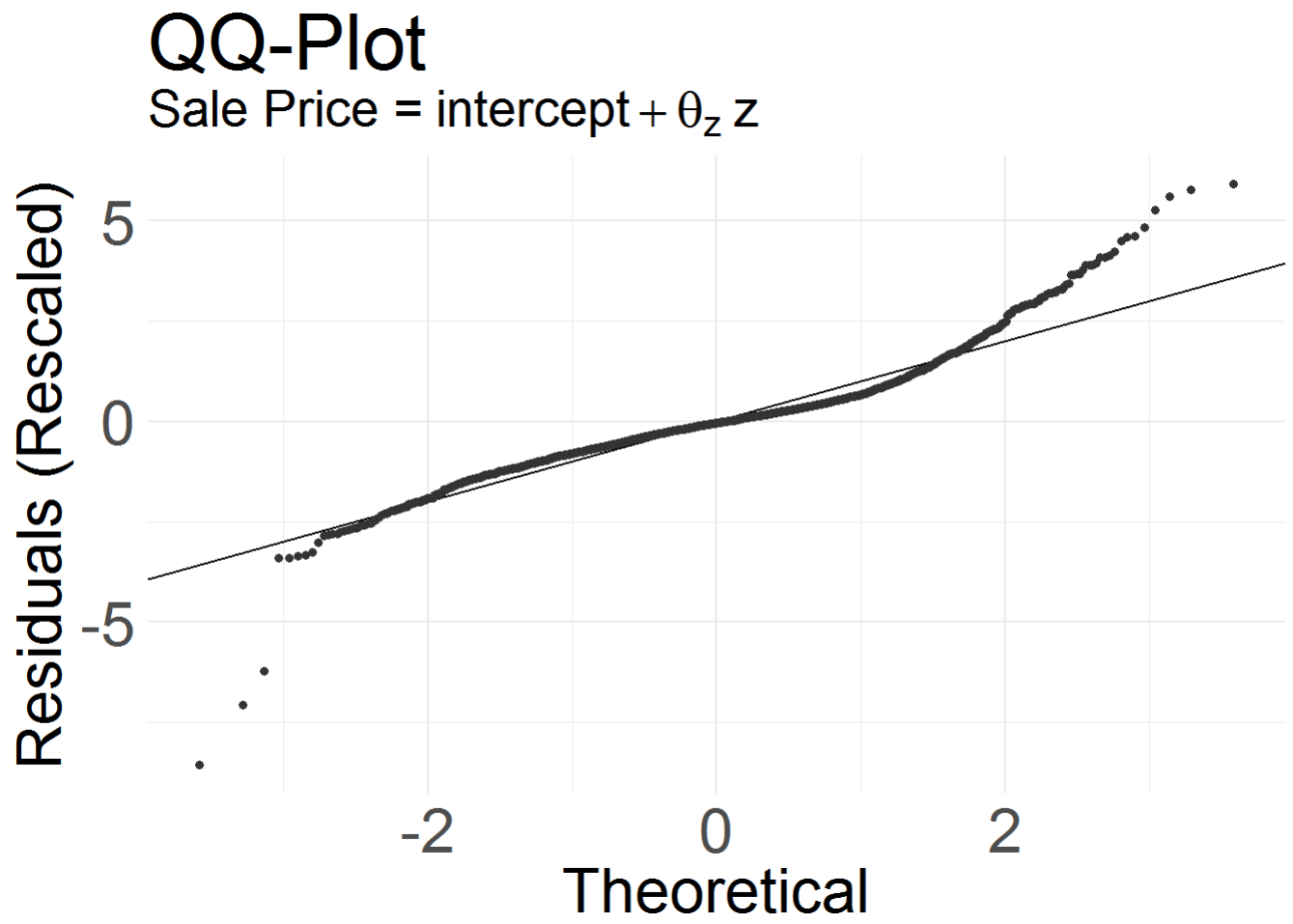
  (-0.34K, -0.12K) – covers the true weight for br

  LS estimate based on training: 0.014, -0.23K

With one z only, histogram of oracle noise terms

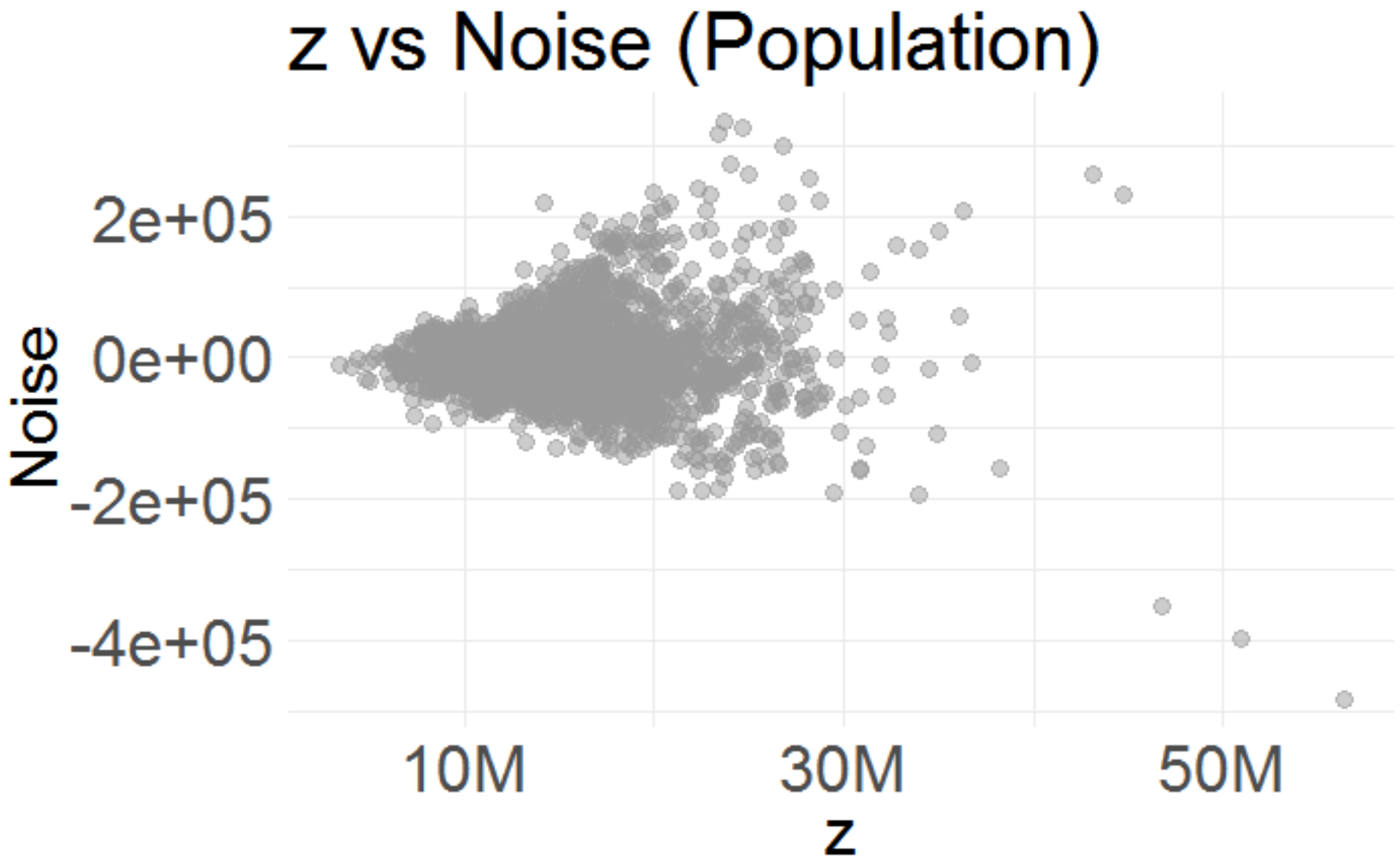


QQ plot of u's with one z



Checking independence of oracle noise and z:

indep seem ok, but same variance for the noise term does not hold



One predictor case with z: true slope=0.011

- Bootstrap 95% confidence interval:

  (0.009, 0.013) – covers the true slope

- Normal approximation interval (oracle noise terms are not normal, but not too heavy tails. Linear Reg. model holds approximately):

  95% confidence interval: (0.011, 0.014), covers the true slope barely

  LS estimate based on training: 0.012

Back to two predictors -- sq.ft. and number of bedrooms – why the LS weight for br. is negative?

To make the math cleaner, let's center y, and the two predictors (meaning subtract their means). Then we don't need an intercept term. Let's also make their centered variables Euclidean norm 1 (normalized) – Use the same notations: y, x1, x2 for the centered and normalized variables.

$$\rho = corr(x1, x2)$$

$$r1 = corr(y, x1), r2 = corr(y, x2)$$

Note that, they are the same for the un-centered and un-normalized variables since the correlation calculations do the centering and normalizing in the process.

Why is the LS slope or weight for Br. negative?

Recall that for a 2 by 2 (invertible) matrix A, $A^{-1} = \frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$

$\rho = corr(x1, x2)$

$r1 = corr(y, x1), r2 = corr(y, x2)$

$$A = X^T X = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \quad A^{-1} = \frac{1}{1-\rho^2}\begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix} \quad X^T Y = \begin{bmatrix} r1 \\ r2 \end{bmatrix}$$

$$\hat{\theta} = (X^T X)^{-1} X^T Y = \frac{1}{1-\rho^2}\begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}\begin{bmatrix} r1 \\ r2 \end{bmatrix}$$

$$= \frac{1}{1-\rho^2}\begin{bmatrix} r1 - \rho\, r2 \\ -\rho\, r1 + r2 \end{bmatrix}$$

ρ=0.62, r1= 0.71, r2=0.3; so the LS slope for Br. = [-0.62*0.71+0.3]/[1-0.62*0.62]= - 0.23 <0

Note true weight= - 0.29

Most important question: what is z?

- Z= sq. ft. *10000

Summary

- Stability pseudo-replicates for quick and concise assessments of sampling variability: CV, random split, small number of bootstrap replicates

- (Full blown) bootstrap for further assessment, e.g. confidence intervals

- Normal Linear Regression model, Linear Regression model – ASSUMPTIONS, ASSUMPTIONS, ASSUMPTIONS

  Check using data (use residuals instead of oracle noise terms)

- All models are wrong, but some are useful –

  the key is to figure out when and why

Taking the same size of data 100 times WITH replacement: Bootstrap

- Get 100 weight vectors

- Scatter plot: answer the same question again
- Translating “importance” to the “real” weight is not zero:
- Histogram: