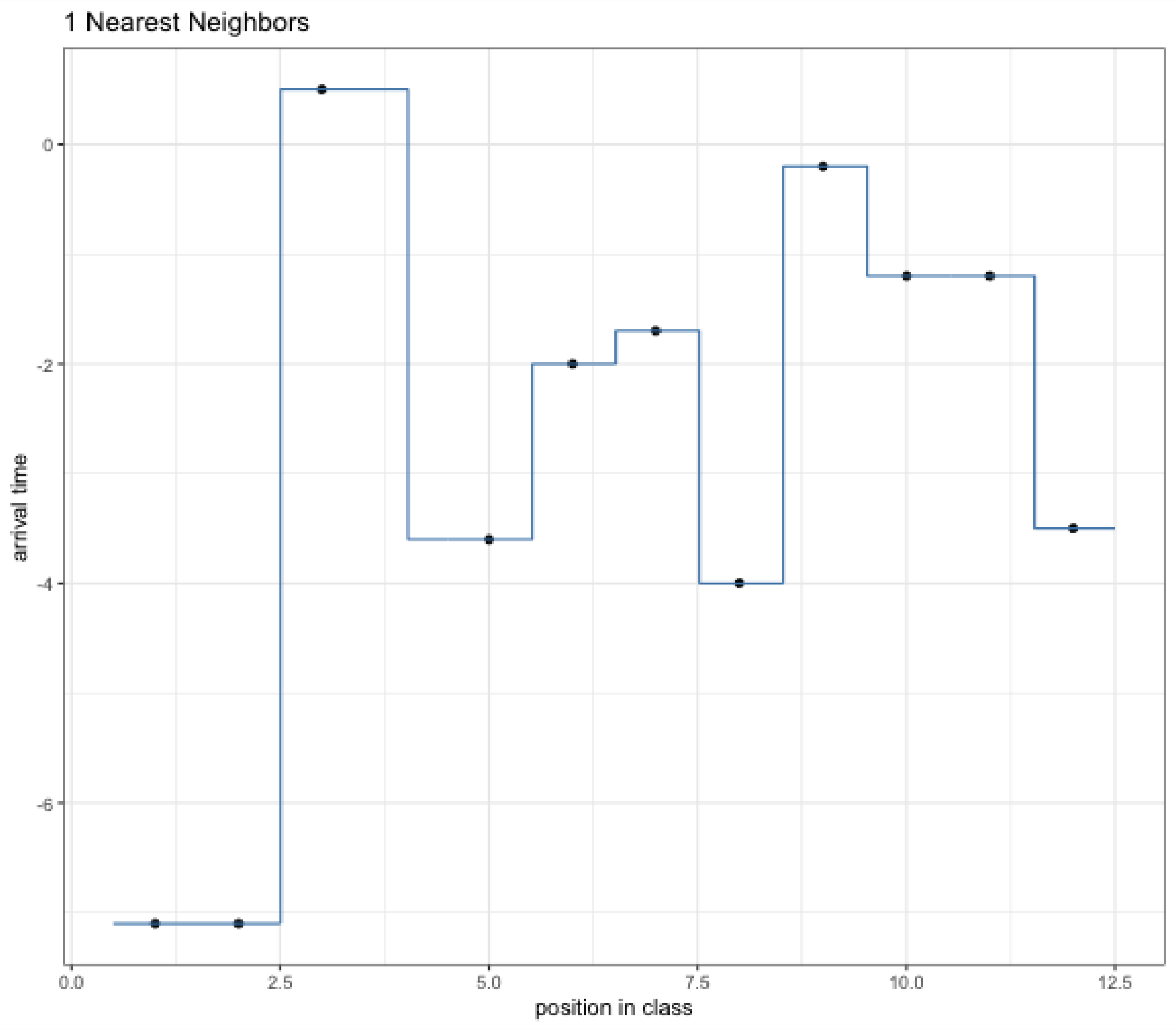
Bias-Variance Tradeoff: KNN



Linear Regression (on board, live coding)

Plato's Allegory of the Cave

Statistical Inference

Goal: use *statistics* calculated from data to makes inferences about the nature of *parameters*.

In regression,

- statistics: $\hat{\beta}_0$, $\hat{\beta}_1$
- parameters: β_0, β_1

Classical tools of inference:

- Confidence Intervals
- Hypothesis Tests

Quick Review (start the timer)

Confidence Intervals

A confidence interval expresses the amount of uncertainly we have in our estimate of a particular parameter. A general 1 - α confidence interval takes the form

$$\hat{\theta} \pm t^* * SE(\hat{\theta})$$

- α: is the confidence level, often .05
- $\hat{\theta}$: a statistic (point estimate)
- t^* is the $100(1 - \alpha/2)$ quantile of the sampling distribution of $\hat{\theta}$
- SE is the standard error of $\hat{\theta}$, i.e. the standard deviation of its sampling distribution.

Common Regression Assumptions

1. Y is related to x by a simple linear regression model.

$$E(Y|X) = \beta_0 + \beta_1 * x$$

2. The errors $e_1, e_2, \ldots, e_n$ are independent of one another.

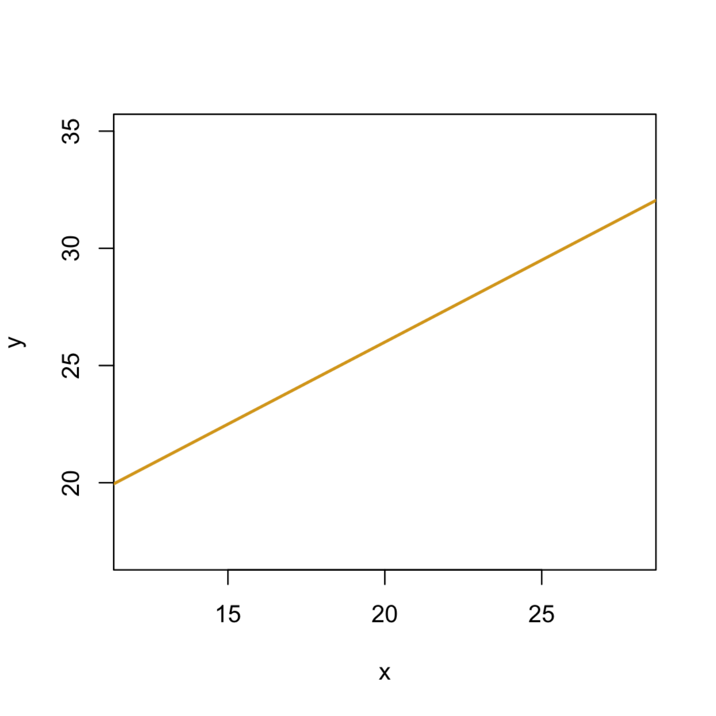
3. The errors have a common variance σ^2.

4. The errors are normally distributed: $\epsilon \sim N(0, \sigma^2)$

The Sampling Distribution of $\hat{\beta}_1$

Assume the following true model:

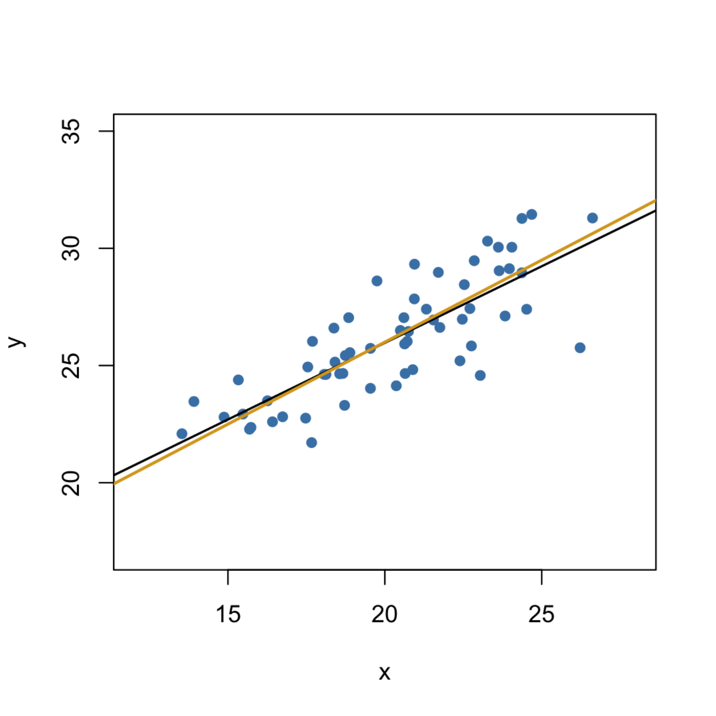
$$E(Y|X) = 12 + .7 * x; \epsilon \sim N(0, 4)$$



The Sampling Distribution of $\hat{\beta}_1$

Assume the following true model:

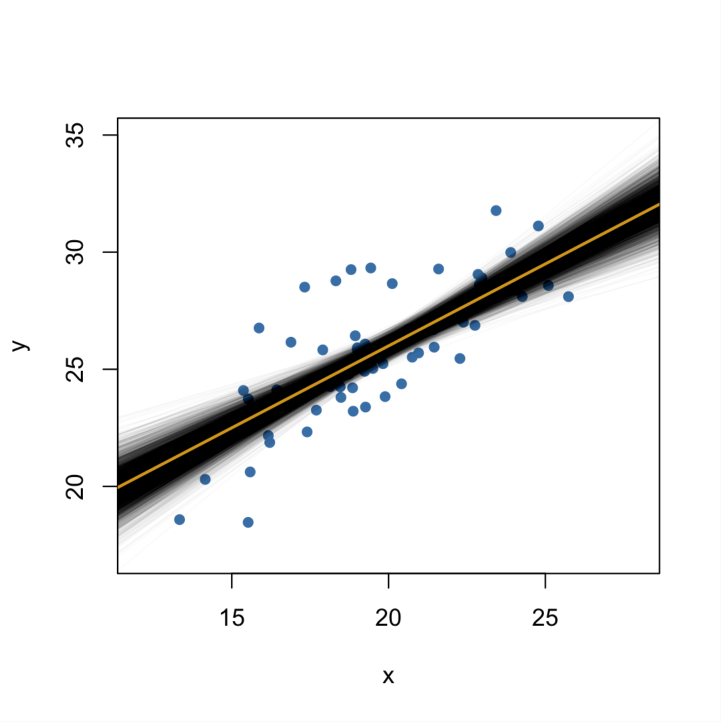
$$E(Y|X) = 12 + .7 * x; \epsilon \sim N(0, 4)$$



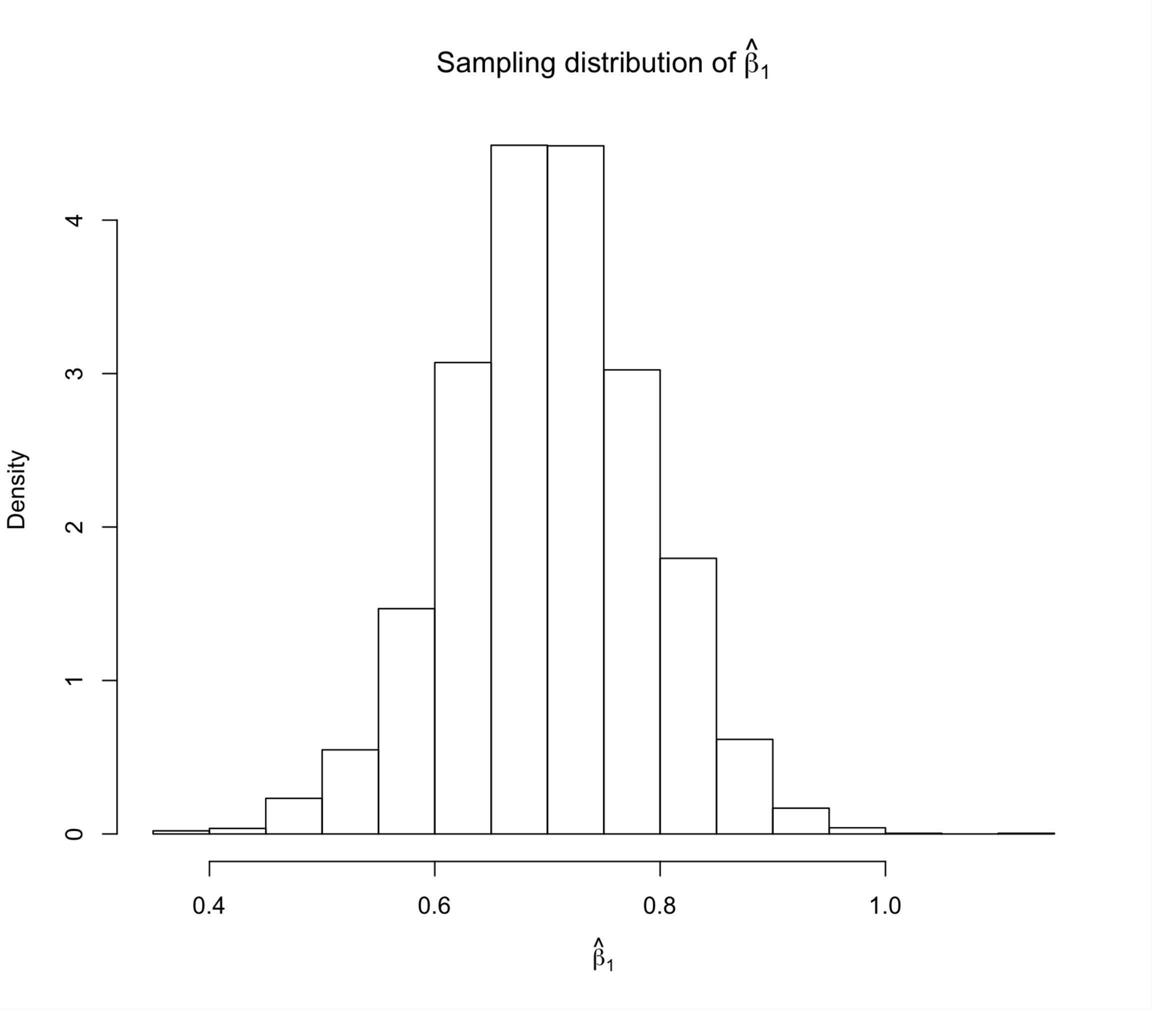
The Sampling Distribution of $\hat{\beta}_1$

Assume the following true model:

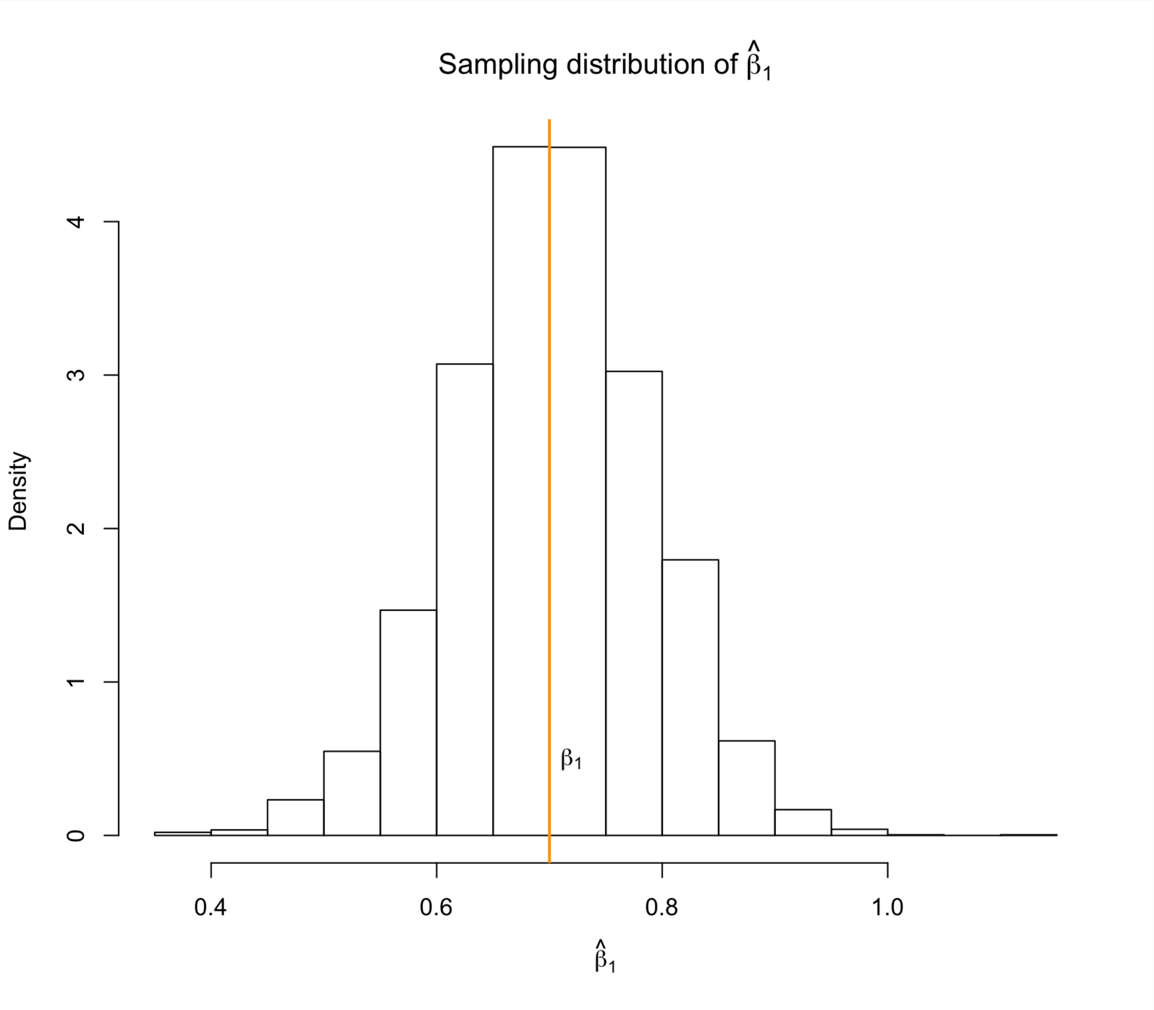
$$E(Y|X) = 12 + .7 * x; \epsilon \sim N(0, 4)$$



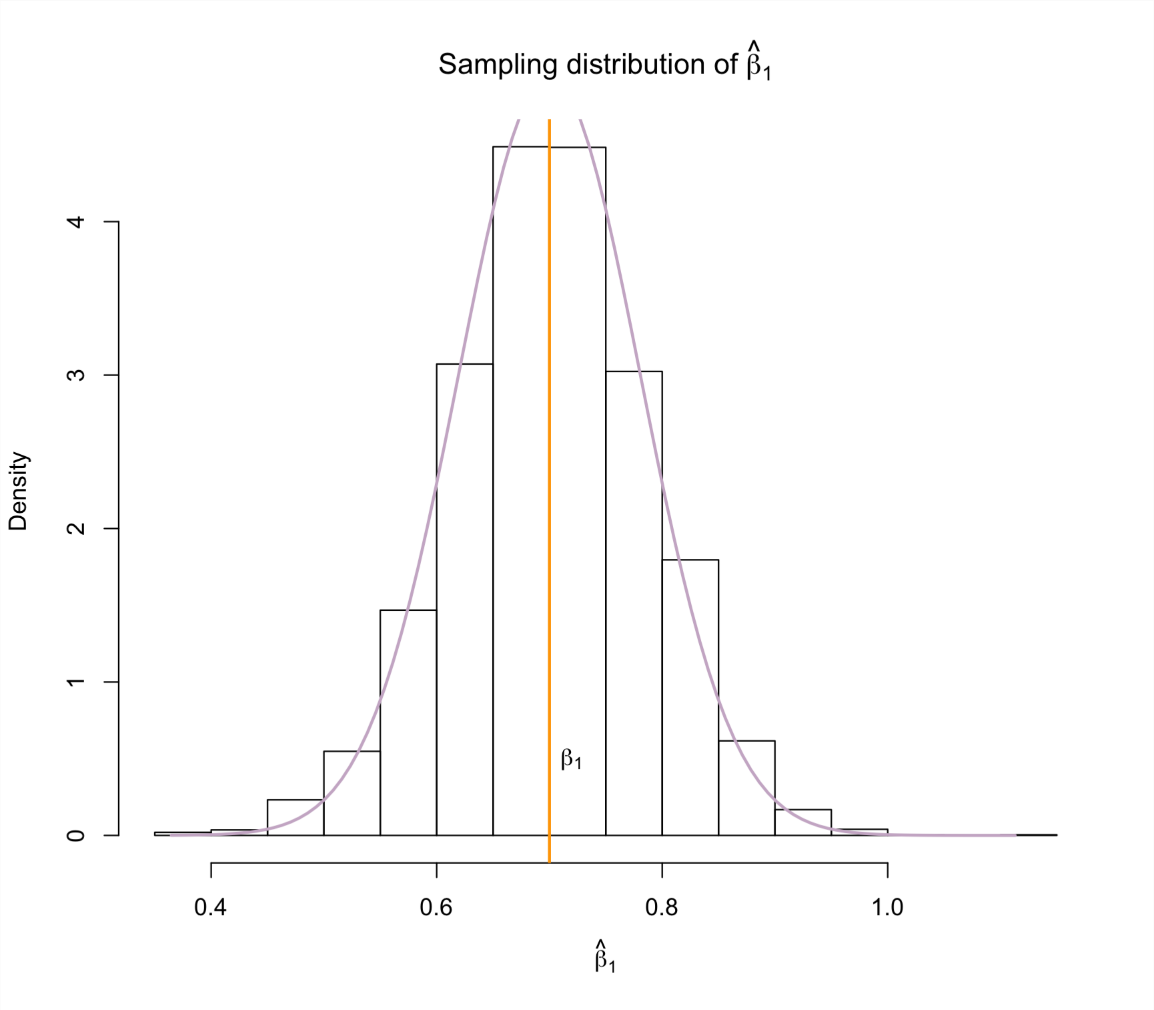
The Sampling Distribution of $\hat{\beta}_1$



The Sampling Distribution of $\hat{\beta}_1$



The Sampling Distribution of $\hat{\beta}_1$



The Sampling Distribution of $\hat{\beta}_1$

Characteristics:

1. Centered at β_1, i.e. $E(\hat{\beta}_1) = \beta$.
2. $Var(\hat{\beta}_1) = \frac{\sigma^2}{SXX}$.
    - where $SXX = \sum_{i=1}^{n}(x_i - \bar{x})^2$
3. $\hat{\beta}_1 | X \sim N(\beta_1, \frac{\sigma^2}{SXX})$.

Approximating the Sampling Dist. of $\hat{\beta}_1$

Our best guess of β_1 is $\hat{\beta}_1$. And since we have to estimate σ with $\hat{\sigma}^2 = RSS/n - 2$, the distribution isn't normal, but...

T with n - 2 degrees of freedom.

And we summarize that approximate sampling distribution using a CI:

$$\hat{\beta}_1 \pm t_{\alpha/2, n-2} * SE(\hat{\beta}_1)$$

where

$$SE(\hat{\beta}_1) = s/\sqrt(SXX)$$

Interpreting a CI for $\hat{\beta}_1$

We are *95% confident* that the true slope between x and y lies between LB and UB.

Hypothesis test for $\hat{\beta}_1$

Suppose we are interested in testing the claim that the slope is zero.

$$H_0 : \beta_1^0 = 0 \; H_A : \beta_1^0 \neq 0$$

We know that

$$T = \frac{\hat{\beta}_1 - \beta_1^0}{SE(\hat{\beta}_1)}$$

T will be t distributed with $n - 2$ degrees of freedom and with $SE(\hat{\beta}_1)$ calculated the same as in the CI.

Inference for $\hat{\beta}_0$

Often less interesting (but not always!). You use the t-distribution again but with a different SE.