Today

- Recap & questions from homework
- Pair programming: Q4 prediction intervals
- Frequentist inference algorithms
  - lm
  - prediction intervals
  - bootstrap

Miscellaneous

- 00_big_ideas_in_data_science.md
  - central ideas and theory
- 00_fundamental_algorithms.md
  - all the algorithms as we encounter them
- 00_portfolio_checklist.md
  - keep track of what you need to submit

Homework check in

- Please fill out the survey

Git skills

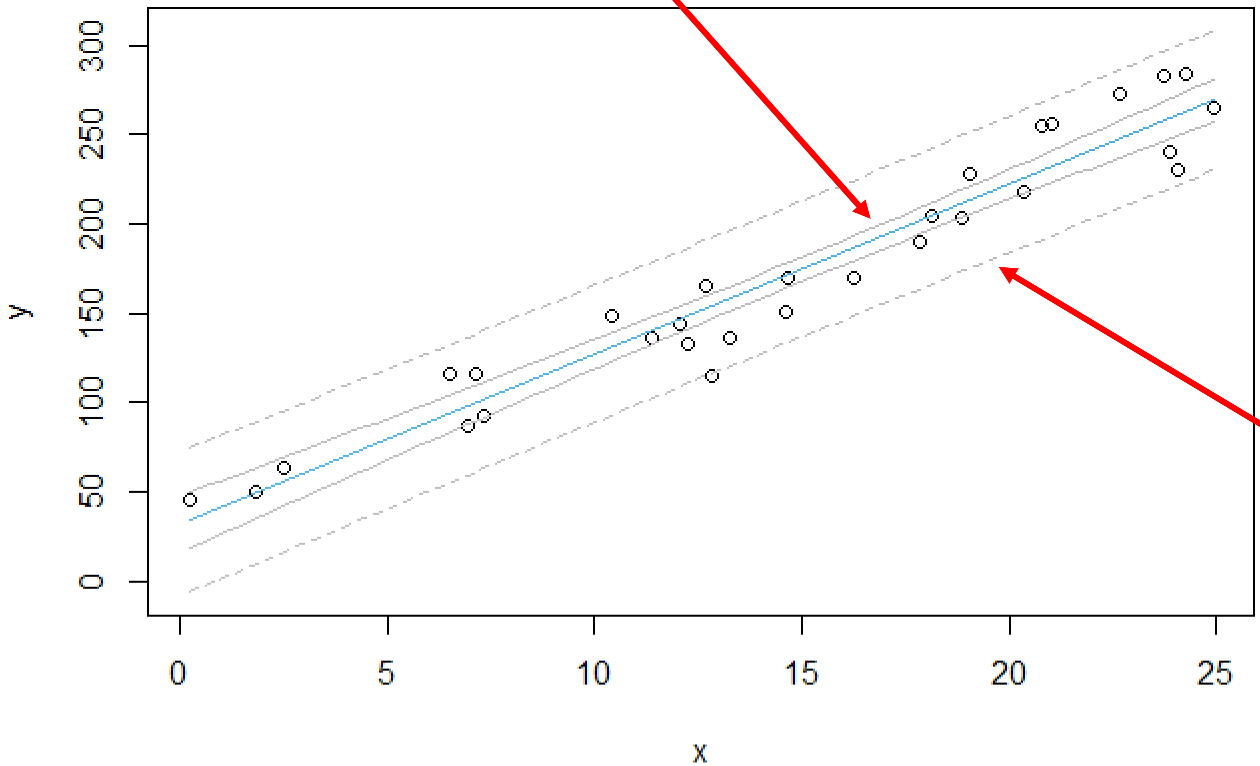
- Branching & merging
  - git branch branch_name
  - git checkout branch_name
  - git merge branch_name
- Trying something out
- Contribute to a collaborative project
  - work in a branch
  - changes merged after review

Frequentist inference algorithms

- Inference algorithms in lm are frequentist
- All frequentist inferences are based on the sampling distribution
- The sampling distribution is the frequentist approach to considering all the ways data could have happened (i.e. looking back)
- CI & p-values for parameters (betas)
- CI & prediction intervals for mean(y | x), aka “the line”

Confidence vs prediction intervals

CI: uncertainty in mean response (estimation uncertainty)



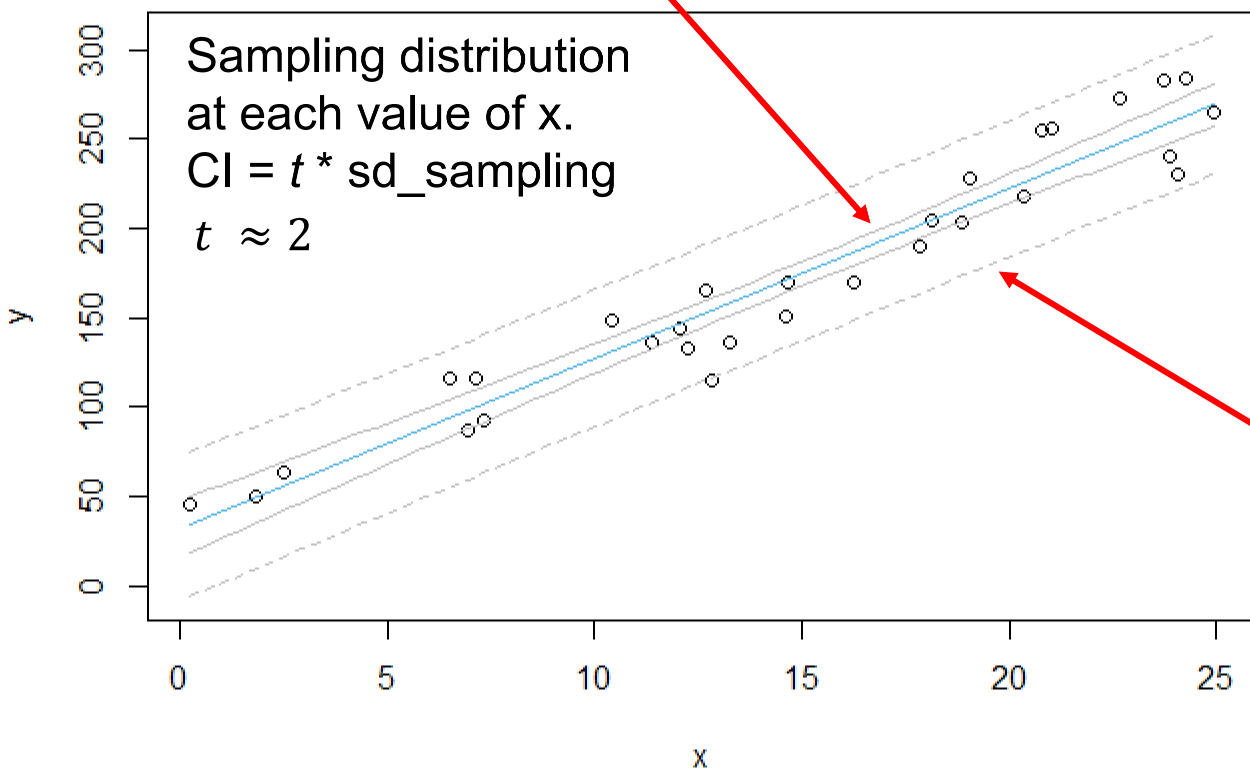
PI: uncertainty in individual response (estimation uncertainty + data generating process)

Pair programming

- Q4 prediction intervals

Confidence vs prediction intervals

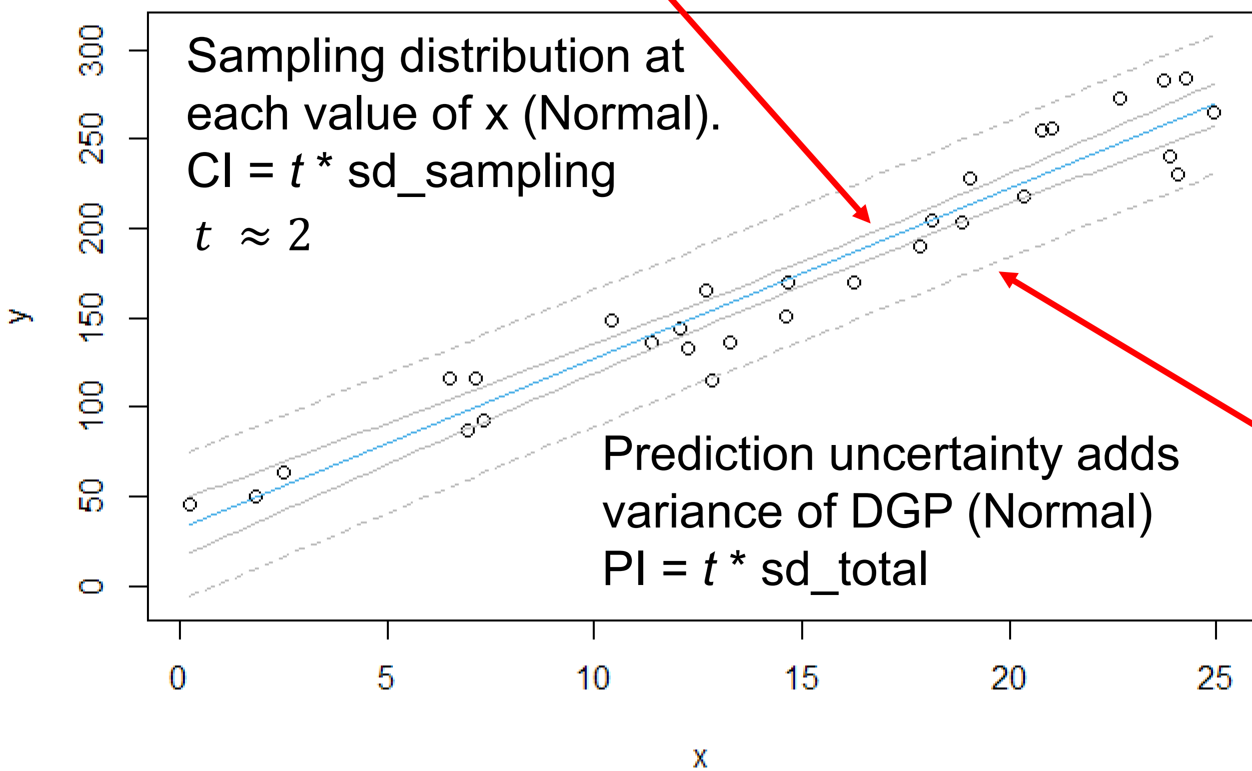
CI: uncertainty in mean response (estimation uncertainty)



PI: uncertainty in individual response (estimation uncertainty + data generating process)

Confidence vs prediction intervals

CI: uncertainty in mean response (estimation uncertainty)



PI: uncertainty in individual response (estimation uncertainty + data generating process)

sd_total = sqrt(var_sampling + var_DGP)
var_DGP estimated by residual variance

Sampling distribution algorithm

repeat very many times

  sample data from the population

  fit the model

  estimate the parameters

plot sampling distribution (histogram) of the parameter estimates

Bootstrap algorithm

repeat very many times

  generate data based on the sample ← plug in

  fit the model

  estimate the parameters

plot sampling distribution (histogram) of the parameter estimates

Bootstrap algorithms

- Non-parametric bootstrap
  - resample the data
- Empirical bootstrap
  - resample the residuals
- Parametric bootstrap
  - generate data from a distribution
  - use estimated parameters of the distribution

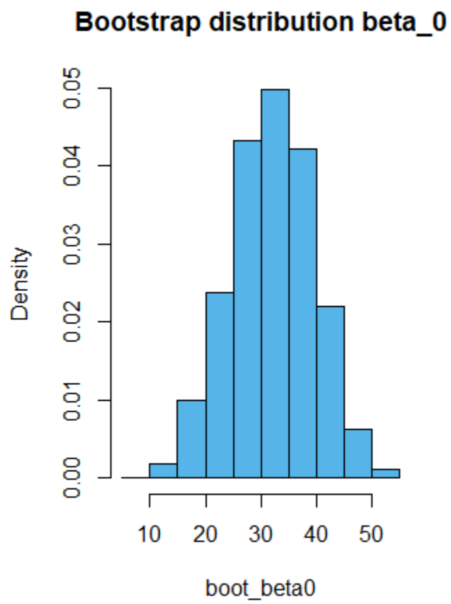
Huge advantage

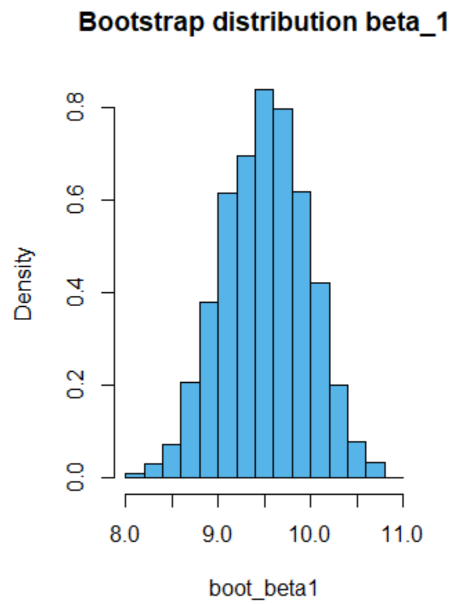
- Can obtain uncertainty for any quantity that can be calculated from any fitted model

Code (e.g. empirical bootstrap)

```
for ( i in 1:10000 ) {
    e_boot <- sample(e_fit, replace=TRUE)
    df_boot$y <- coef(fit)[1] + coef(fit)[2]*df_boot$x + e_boot
    fit_boot <- lm(y ~ x, data=df_boot)
    boot_beta0[i] <- coef(fit_boot)[1]
    boot_beta1[i] <- coef(fit_boot)[2]
}
```

plug in





Pseudocode
For many times
    Resample errors from model fit (with replacement)
    Create new y-values at original x values
    Fit the model
    Keep parameter estimates

Bootstrap: further reading

Brief exposition:

James G, Witten D, Hastie T, Tibshirani R (2021). An Introduction to Statistical Learning: With Applications in R, Second edition. Springer, New York. Chapter 5.2.

Definitive references:

Davison AC, Hinkley DV (1997). Bootstrap Methods and Their Application. Cambridge University Press, Cambridge ; New York, NY, USA.

Efron B, Tibshirani R (1993). An Introduction to the Bootstrap. Chapman & Hall, New York.

Bootstrapped CI

- Learning goals
  - Understand how bootstrap algorithms mimic the sampling distribution algorithm
  - Gain intuition for the plug in principle by using it directly
  - Using simulation, understand how the sampling distribution is the basis for frequentist inference
  - Develop algorithms from first principles that can be applied to any model

Code a bootstrapped CI

- Use your linear data
- CI for β_0, β_1
- CI for y|x (aka the line)

- Use empirical bootstrap
- Use percentile method