Today

- Algorithms:
- Derived quantities
- Bootstrapped confidence bands
- Bootstrapped prediction bands
- Bootstrapped p-value

General principles

- Illustrated with linear model
- But generalizes to any model

Derived quantities

- Any quantity that is a function of the parameters
- e.g. $y|x=10$ in the linear model

  Value of y given x = 10:

  $y = fn(\beta_0, \beta_1, x=10) = \beta_0 + 10\beta_1$
- Very common that interesting scientific questions are addressed by a derived quantity

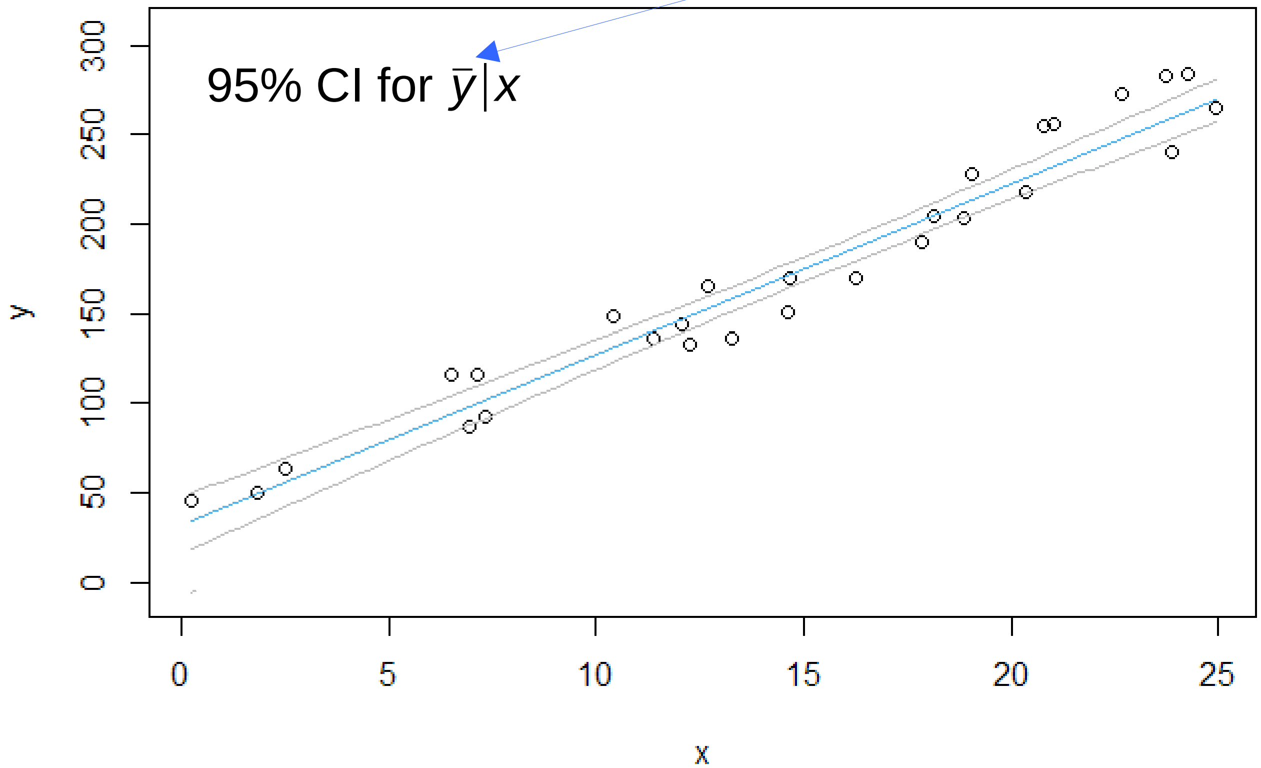
Derived quantities

- To do inference:
- Derived quantity is the sample statistic
- Bootstrap its sampling distribution
  - already have bootstrapped samples of parameter values. Reuse them!
  - derived quantity sampling distribution = fn(parameter bootstrap samples)

Example: uncertainty of line

- A set of derived quantities
- e.g. y|x for x in (0, 25)

“y bar” is mean(y) or expected value of y

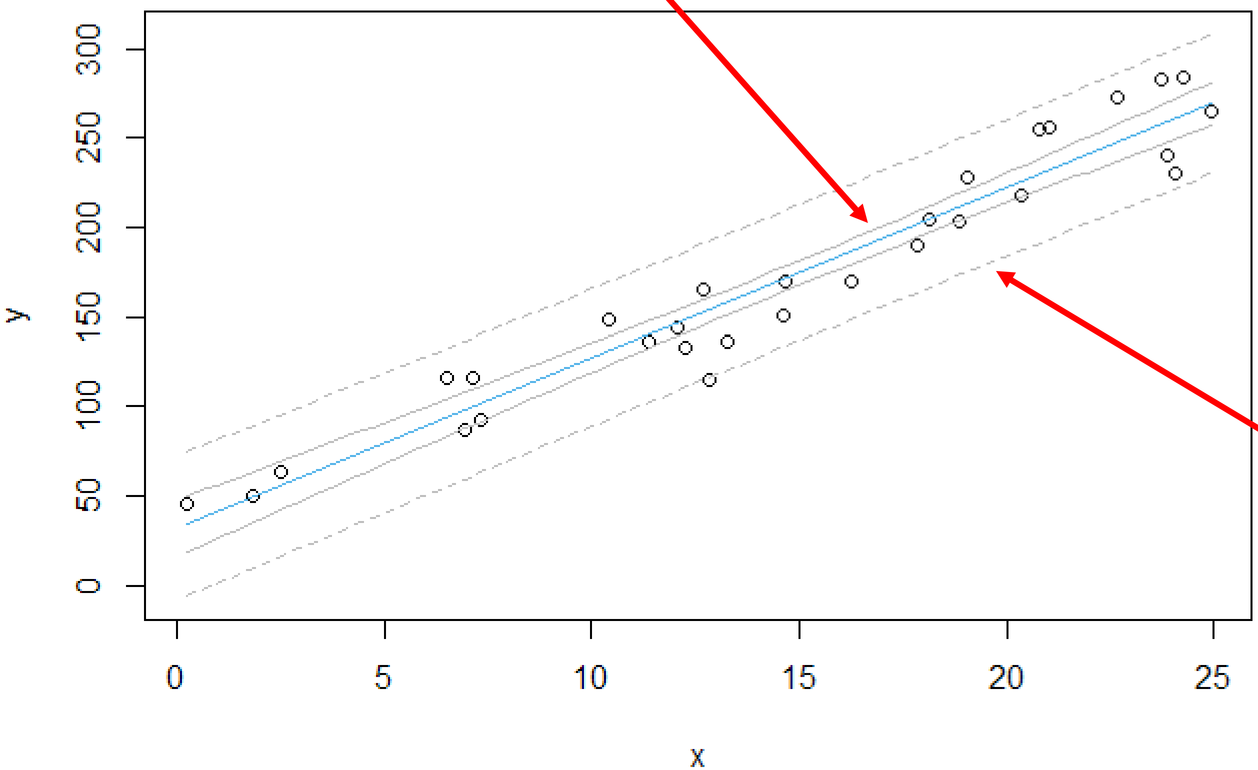


Prediction interval

- Uncertainty of a predicted new data point
- Need to propagate uncertainty, 2 components:
- 1) Estimation uncertainty
- 2) Uncertainty from data generating process

Confidence vs prediction intervals

CI: uncertainty in mean response (estimation uncertainty)



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

Bootstrap prediction interval

- Prediction uncertainty for new y
- bootstrap_prediction_interval.md
- Powerful idea: estimate uncertainty by
  - repeatedly
  - simulate training the model on a sample (parameter uncertainty)
  - simulate generating data from the trained model (data generating process)

Bootstrap prediction interval

e.g. prediction band for y = fn(x)

Algorithm

define a grid of new x values to predict y
repeat very many times

- sample from the error distribution of DGP
- simulate new y-values from original estimated parameters of model
- train the model (estimate parameters: beta_0, beta_1, sigma_e)
- keep: simulate new data y|x using estimated parameters

calculate quantiles of the generated data distributions
plot quantiles

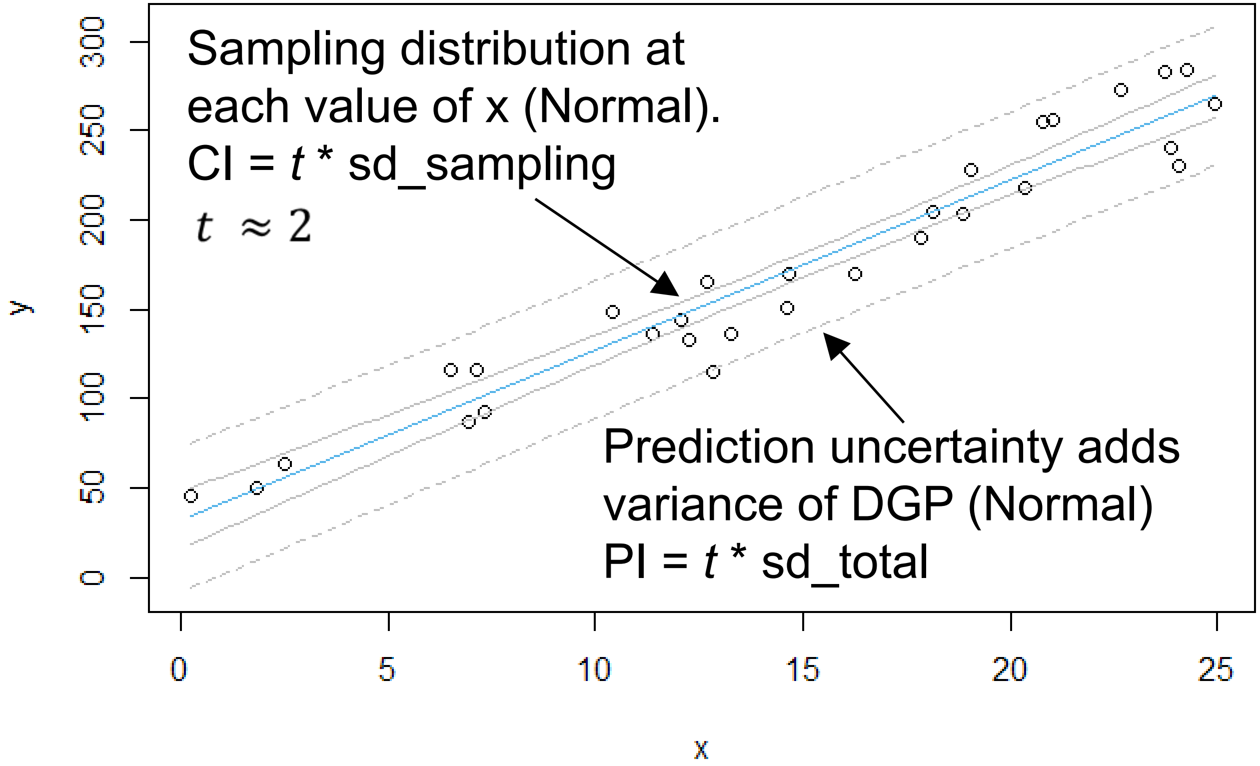
parametric

simulate generating data from the fitted model

simulate training the model on a sample

Classical prediction intervals

Special case: linear model



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

Bootstrapped p-value

- Learning goals
  - Understand p-values by understanding the underlying sampling algorithm
  - Further understand how the sampling distribution is the basis for frequentist inference
  - Understand how bootstrap algorithms mimic the sampling distribution algorithm
  - Formulate a bootstrap algorithm and translate it to code

Parametric bootstrap for H_0: $\beta_1 = 0$

Combine these concepts (pseudocode first):

Definition of a p-value

The probability of a sample statistic as large or larger than the one observed given that some hypothesis is true

Basic parametric bootstrap algorithm

```
repeat very many times
    sample from the error distribution
    create new y-values from the estimated parameters and errors
    train the linear model to estimate the parameters
plot sampling distribution (histogram) of the parameter estimates
```

plug in: create simulated data from model