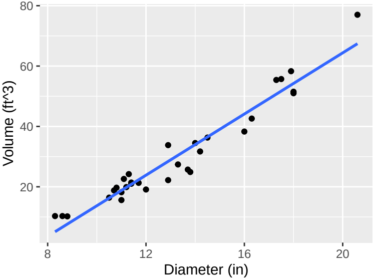
Lecture 8:
Prediction Intervals
STAT 432, Spring 2021

Example: Cherry Tree Data Set

- The R data set `trees` contains measurements of the diameter (girth), height, and volume of timber in 31 felled black cherry trees.
- **Question**: Can the diameter of a cherry tree be used to predict its volume? If so, what is the uncertainty associated with that prediction?

```
> head(trees)
  Girth Height Volume
1   8.3     70   10.3
2   8.6     65   10.3
3   8.8     63   10.2
4  10.5     72   16.4
5  10.7     81   18.8
6  10.8     83   19.7

> dim(trees)
[1] 31  3
```



Based on the regression summary below, the equation of the least squares line is

$$\hat{y} = -36.9435 + 5.0659x$$

```
> lm1 <- lm(Volume ~ Girth, data=trees)
> summary(lm1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -36.9435     3.3651  -10.98 7.62e-12 ***
Girth         5.0659     0.2474   20.48  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.252 on 29 degrees of freedom
Multiple R-squared:  0.9353,Adjusted R-squared:  0.9331
F-statistic: 419.4 on 1 and 29 DF,  p-value: < 2.2e-16
```

When the diameter measurement $x = 17$ inches, the prediction for timber volume is

$$\hat{y} = -36.9435 + 5.0659(17) = 49.18 \text{ ft}^3$$

There are two intpertations:

- 49.18 ft^3 is the prediction for the timber volume for a *single* cherry tree with a 17 inch diameter.
- 49.18 ft^3 is the prediction for the *mean* volume of *all* cherry trees with 17 inch diameters.

The uncertainty (margin of error) associated with this prediction depends on the interpreation.

Prediction versus Confidence Interval

Use a **prediction interval** to:
Calculate an interval of plausible values for the volume of a single cherry tree that has a 17 inch diameter.

Use a **confidence interval** to:
Calculate an interval of plausible values for the population mean volume of all cherry trees that have 17 inch diameters.

Prediction Interval: R Computation

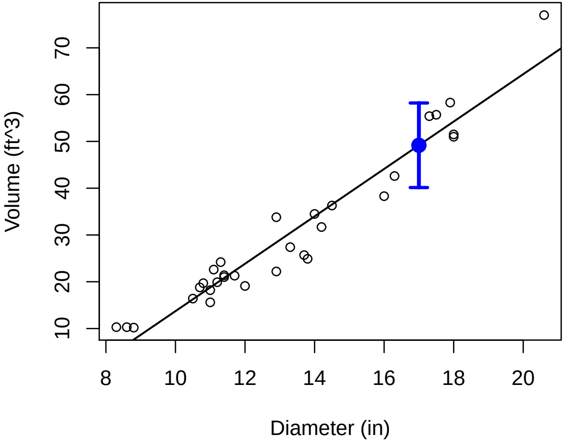
Use R to construct a 95% prediction interval for the volume of a cherry tree that has diameter $x = 17$ inches:

```
> new_x <- data.frame(Girth = 17)
> predict(lm1, newdata = new_x, interval = "prediction")
      fit      lwr      upr
1 49.1761 40.13908 58.21312
```

The interpreation is that if we measured a cherry tree with a 17 inch diameter, then the actual volume of that tree is likely to be between 40.14 and 58.21 cubic feet.

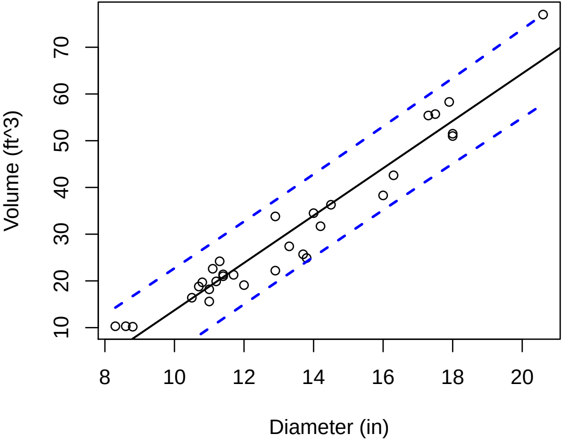
Prediction Interval: Illustration

95% prediction interval for the volume of a cherry tree with a 17 inch diameter.



Prediction Interval: Illustration

95% prediction interval band.



Prediction Interval: R Computation

We can also change the confidence level. Note that 95% is the default.

```
> new_x <- data.frame(Girth = 17)
> predict(lm1, newdata = new_x,
    interval = "prediction", level = 0.99)
      fit      lwr      upr
1 49.1761 36.99677 61.35543
```

Confidence Interval: R Computation

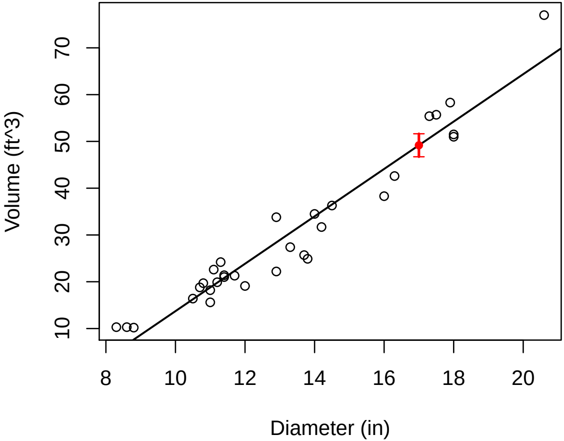
Use R to calculate a 95% confidence interval for the mean volume of cherry trees that have diameter $x = 17$ inches:

```
> new_x <- data.frame(Girth = 17)
> predict(lm1, newdata = new_x, interval="confidence")
      fit      lwr      upr
1 49.1761 46.71799 51.63421
```

The interpretation is that we are 95% confident that the mean volume of all cherry trees that have 17 inch diameters is between 46.72 and 51.63 cubic feet.

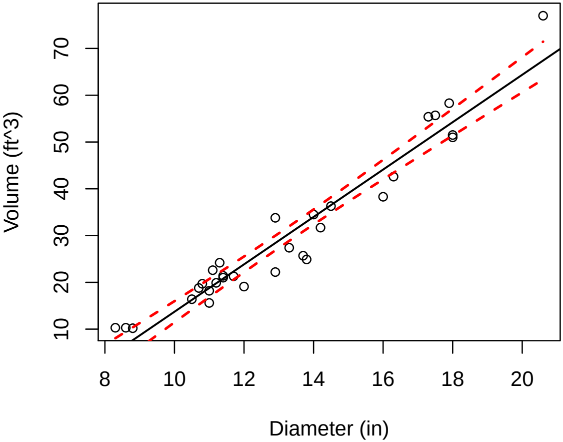
Confidence Interval: Illustration

A 95% confidence interval for the mean volume of cherry trees that have 17 inch diameters.



Confidence Interval: Illustration

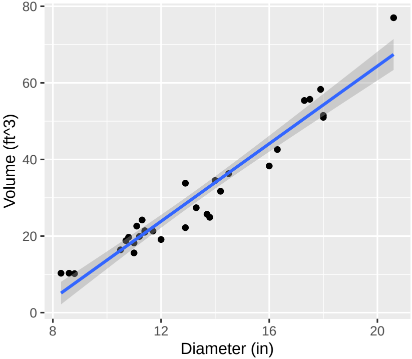
A 95% confidence band for mean timber volume. We can also think of this as a confidence band for the population regression line.



Confidence Interval: Illustration

A 95% confidence band using ggplot2:

```
ggplot(trees, aes(Girth, Volume)) + geom_point() +
  geom_smooth(method = "lm", se = TRUE) +
  labs(x = "Diameter (in)", y = "Volume (ft^3)")
```



Comparing PIs and CIs

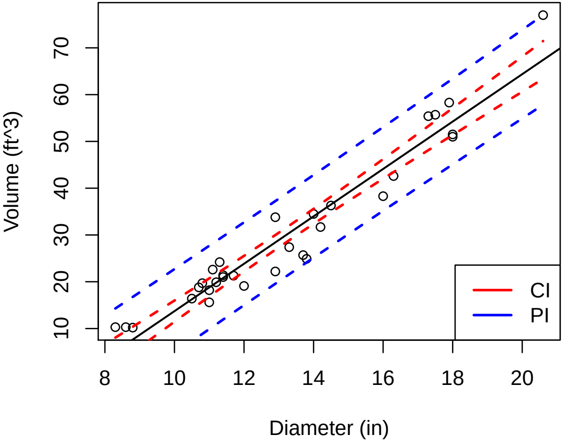
```
> new_x <- data.frame(Girth = 17)

> predict(lm1, newdata = new_x, interval = "confidence")
      fit      lwr      upr
1 49.1761 46.71799 51.63421

> predict(lm1, newdata = new_x, interval = "prediction")
      fit      lwr      upr
1 49.1761 40.13908 58.21312
```

Comparing PIs and CIs

The 95% prediction interval band is wider than the confidence interval band.



Summary

- In addition to using simple linear regression to make a prediction for the response variable, we can also construct a prediction interval that quantifies the uncertainty in that prediction.

- It is important to distinguish between a prediction interval for a single value of the response and a confidence interval for the mean response.

- Prediction intervals are more useful and common in practice.