Lecture 3
Inference for Simple Linear Regression (part 2):
Prediction Intervals
STAT 632, Spring 2020

Given a new value for the explanatory variable x^*, the prediction for the response is

$$\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$$

We are also interested in quantifying the uncertainty in this prediction. That is, we are interested in constructing a prediction interval.

Example

- 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

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

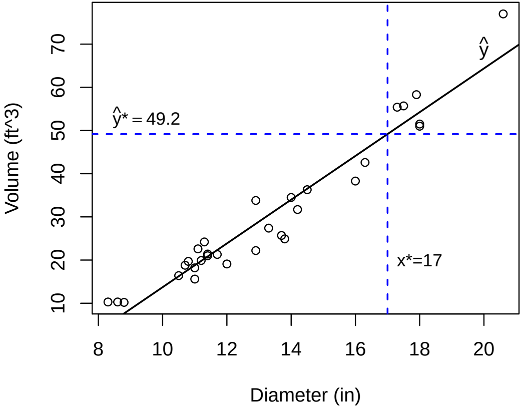
```
> lm1 <- lm(Volume ~ Girth, data=trees)
> summary(lm1)

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
```

Given a new diameter measurement, $x^* = 17$ inches, the prediction for timber volume is

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



When quantifying uncertainty, we need to distinguish between predicting the mean response and a new, actual value of the response.

- The mean response:

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

  For example, this represents the average volume for cherry trees that have an $x^* = 17$ inch diameter. Note that the mean response is fixed (non-random) since β_0 and β_1 are population parameters.

- A new, actual value of the response:

$$Y^* = \beta_0 + \beta_1 x^* + e, \text{ where } e \sim N(0, \sigma^2)$$

  For example, this represents the volume for a single cherry tree that has an $x^* = 17$ inch diameter. Note that Y^* is defined here as a random variable.

Confidence interval for the mean response

When constructing a confidence interval for the mean response there is only one source of variability: the estimation of the population parameters (β_0 and β_1).

$$Var(\hat{y}^*) = Var(\hat{\beta}_0 + \hat{\beta}_1 x^*) = \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\text{SXX}} \right],$$

where SXX $= \sum_{i=1}^{n} (x_i - \bar{x})^2$.

Confidence interval for the mean response

A 1-α confidence interval for the mean response:

$$\hat{y}^* \pm t_{\alpha/2;n-2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\mathrm{SXX}}},$$

where $\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$, and $\hat{\sigma}$ is the residual standard error.

The interpretation is "We are 95% confident that the mean response is between ..."

Example - R Computation

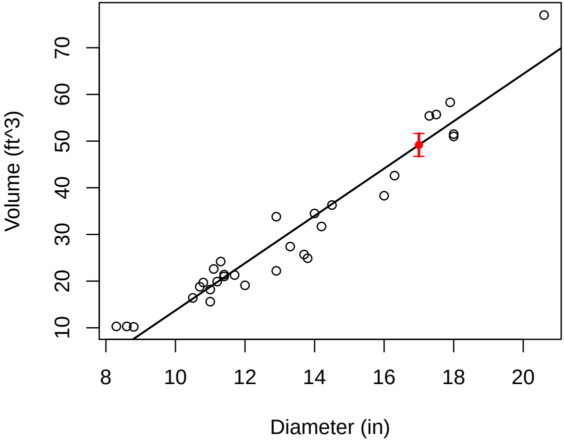
Use R to calculate a 95% confidence interval for the mean volume of cherry trees that have diameter $x^* = 17$ inches.

```
> 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 the predicted mean volume, for cherry trees that have a 17 inch diameter, is 49.18 cubic feet. Additionally, we are 95% confident that the population mean volume, for cherry trees that have a 17 inch diameter, is between 46.72 and 51.63 cubic feet.

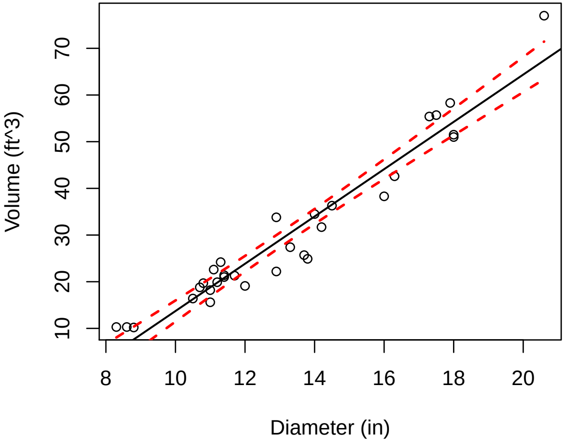
Example - Illustration

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



Example - Illustration

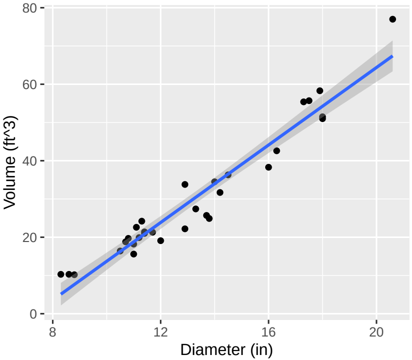
A 95% confidence band (or envelope) for mean timber volume. We can also think of this as a confidence band for the population regression line.



Example - Illustration

A 95% confidence band using ggplot2.

```
ggplot(trees, aes(Girth, Volume)) +
  geom_point() + stat_smooth(method = "lm", se = TRUE) +
  xlab("Diameter (in)") + ylab("Volume (ft^3)")
```



Prediction interval for an actual response value

When constructing a prediction interval for a new, actual value of the response there are two sources of variability: the estimation of the population parameters (β_0 and β_1), and the random error e.

$$
\begin{aligned}
Var(\hat{y}^* + e) &= Var(\hat{\beta}_0 + \hat{\beta}_1 x^*) + Var(e) \\
&= \sigma^2 \left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\mathrm{SXX}} \right],
\end{aligned}
$$

where $\mathrm{SXX} = \sum_{i=1}^{n} (x_i - \bar{x})^2$.

Prediction interval for an actual response value

A 1-α prediction interval for a new, actual value of the response:

$$\hat{y}^* \pm t_{\alpha/2;n-2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\text{SXX}}},$$

where $\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$, and $\hat{\sigma}$ is the residual standard error.

The interpretation is "A 95% prediction interval for the response is ..."

Example - R Computation

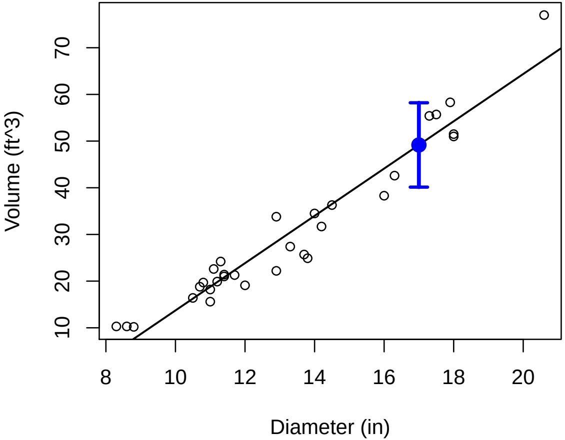
Use R to construct a 95% prediction interval for the volume of a single cherry tree that has diameter $x^* = 17$ inches.

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

The interpretation is that the predicted volume, for a cherry tree that has a 17 inch diameter, is 49.18 cubic feet. Additionally, the 95% prediction interval is between 40.14 and 58.21. This means that the actual volume of a cherry tree, with a 17 inch diameter, is likely to be between 40.14 and 58.21 cubic feet.

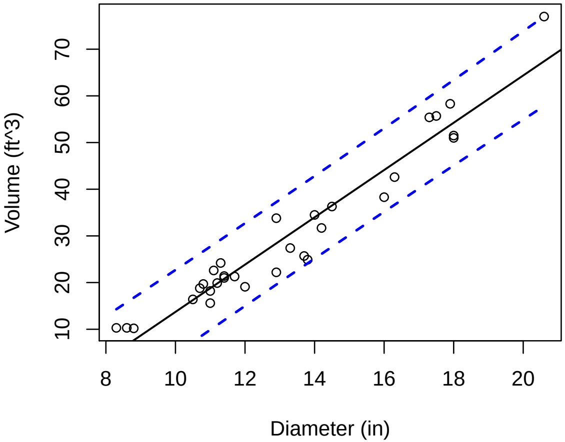
Example - Illustration

95% prediction interval for the volume of a cherry tree with diameter $x^* = 17$ in.



Example - Illustration

95% prediction interval band.



Here is the R code for the previous figure:

```
> lm1 <- lm(Volume ~ Girth, data=trees)
> plot(Volume ~ Girth, xlab='Diameter (in)',
       ylab = 'Volume (ft^3)', data=trees, cex=0.9)
> abline(lm1, lwd=1.5)
> min_x <- min(trees$Girth)
> max_x <- max(trees$Girth)
> grd_x <- seq(min_x, max_x, by=0.1)
> new_x <- data.frame(Girth = grd_x)
> PI <- predict(lm1, newdata = new_x, interval="prediction")
> PI <- as.data.frame(PI)
> lines(grd_x, PI$lwr, lty=2, lwd=2, col="blue")
> lines(grd_x, PI$upr, lty=2, lwd=2, col="blue")
```

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
```

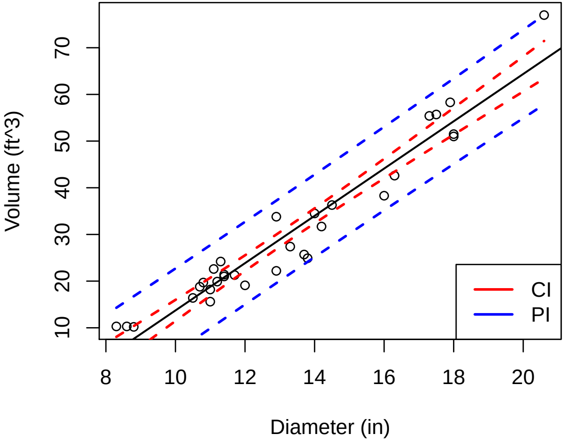
Comparing PIs and CIs

```
> 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
```

- The point predictions for the mean response and an actual value of the response are the same ($\hat{y}^* = 49.176$ when $x^* = 17$).
- The prediction interval for the actual response is substantially wider than the confidence interval for the mean response.

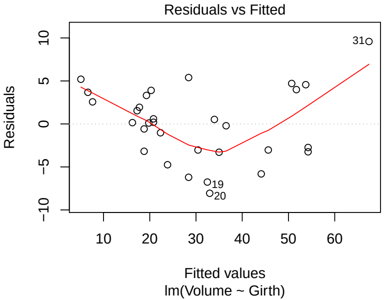
Comparing PIs and CIs

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



Diagnostics?

When making inferences we should also check that the conditions for SLR are satisfied (linearity, constant variance, independence, normality). One useful diagnostic is a plot of the residuals versus the fitted values.



There is obvious curvature in the residuals. Transformations or incorporating quadratic effects might improve the model (topics for future lectures).

Summary

- In addition to using SLR 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 confidence interval for the mean response and a prediction interval for the actual response.
- Prediction intervals are more useful and common in practice.