STAT 224 Lecture 5
Multiple Linear Regression, Part 2

Yibi Huang
Department of Statistics
University of Chicago

Outline

- Example: The Trees Data
- Standard Errors & Distributions of Least Squares Estimates for Coefficients
- Hypothesis Tests & Confidence Intervals for Coefficients

Example: The Trees Data

Example: The Trees Data

The `trees` data are measurements of the diameter, height and volume of timber in 31 felled black cherry trees. The variables are

- `Girth`: Tree diameter (rather than girth, actually) in inches measured at 4 ft 6 in above the ground
- `Height`: Height in ft
- `Volume`: Volume of timber in cubic ft

The `trees` data are build-in in R. One can load the the data by the command

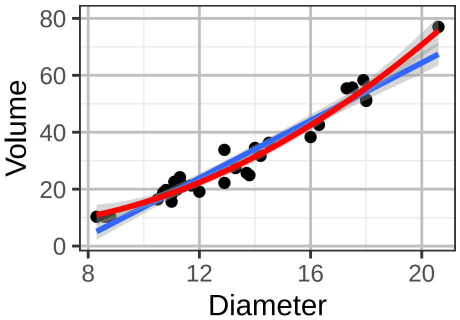
```
data("trees")
```

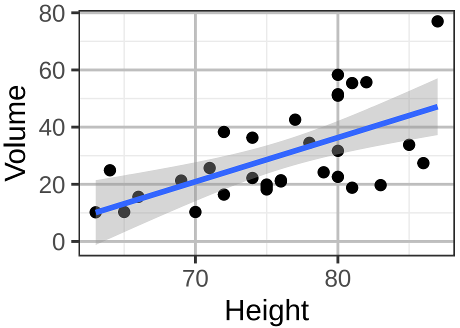
Let's rename the misleading `Girth` variable as `Diameter`

```
trees$Diameter = trees$Girth
```

Pairwise Scatter Plots

```r
library(ggplot2)
ggplot(trees, aes(x=Diameter, y=Volume)) + geom_point() +
  geom_smooth(method='lm', formula='y~x') +
  geom_smooth(method='lm', formula='y~x+I(x^2)', col="red")
ggplot(trees, aes(x=Height, y=Volume)) + geom_point() +
  geom_smooth(method='lm', formula='y~x')
```

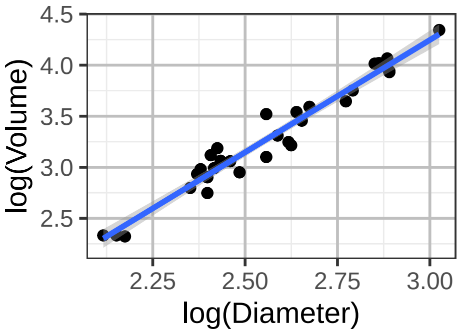


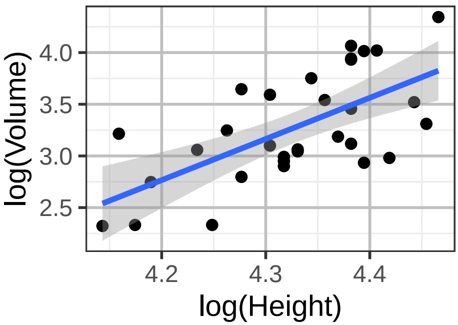


- slight non-linearity between `Diameter` & `Volume`
- Variability of `Volume` increases w/ `Height`

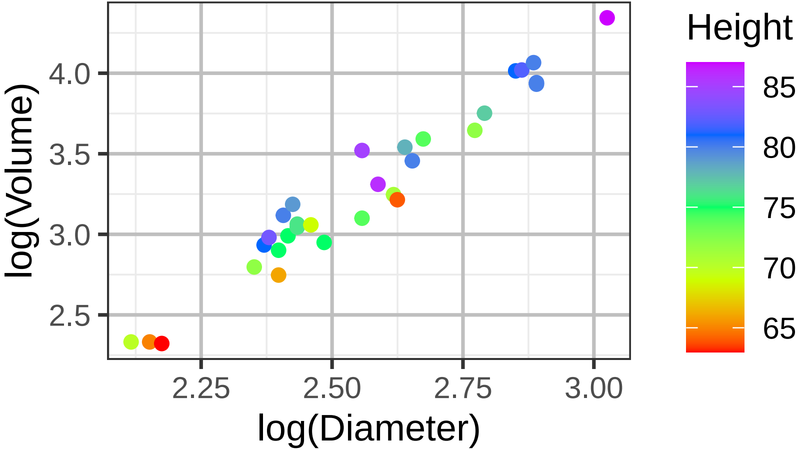
After Log-Transformation ...

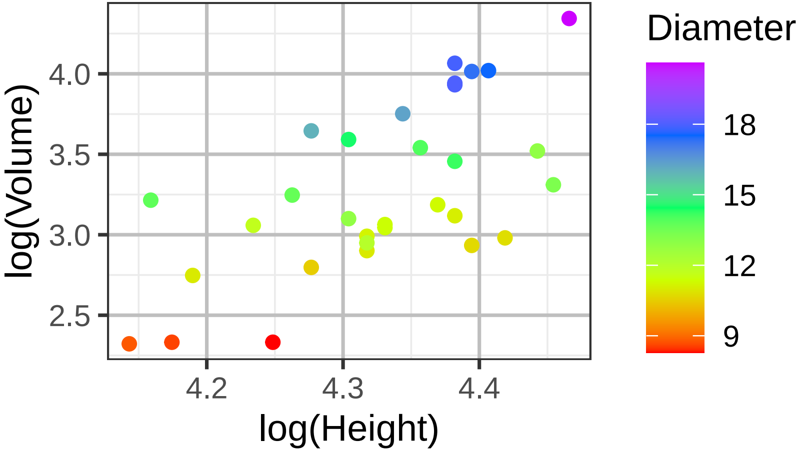
```
ggplot(trees, aes(x=log(Diameter), y=log(Volume))) + geom_point() +
  geom_smooth(method='lm', formula='y~x')
ggplot(trees, aes(x=log(Height), y=log(Volume))) + geom_point() +
  geom_smooth(method='lm', formula='y~x')
```





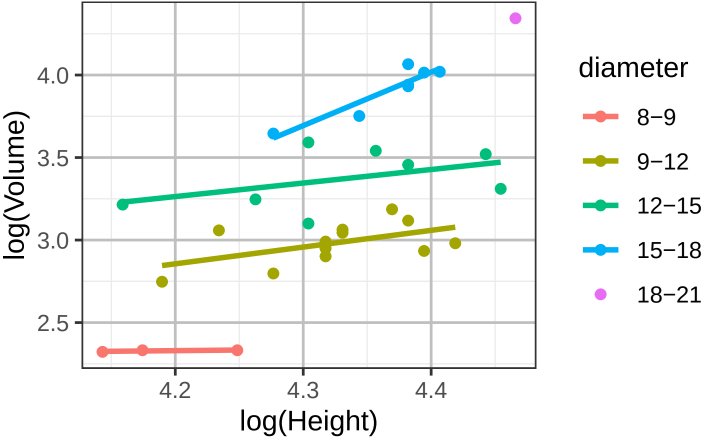
```r
ggplot(trees, aes(x=log(Diameter), y=log(Volume), color=Height)) +
  geom_point() + scale_color_gradientn(colours = rainbow(5))
```





4.0
3.5
3.0
2.5
log(Volume)
4.2
4.3
4.4
log(Height)
Diameter
18
15
12
9

```r
trees$diameter = cut(trees$Diameter, breaks=c(8,9,12,15,18,21),
                     labels=c("8-9","9-12", "12-15", "15-18","18-21"))
ggplot(trees, aes(x=log(Height), y=log(Volume), color=diameter)) +
  geom_point() + geom_smooth(method='lm', formula='y~x', se=F)
```



Model for the Timber Volume of Trees

Recall in the previous lecture we argued that

$$\text{Timber Volume} \approx (\text{constant})(\text{Diameter})^2(\text{Height})$$

Taking logarithm on both sides, we consider the model

$$\log(\text{Volume}) = \beta_0 + \beta_1 \log(\text{Diameter}) + \beta_2 \log(\text{Height}) + \varepsilon$$

and we expect $\beta_1 = 2$ and $\beta_2 = 1$. We thus fit the model

```
lmtrees = lm(log(Volume) ~ log(Diameter) + log(Height), data=trees)
lmtrees$coef
  (Intercept) log(Diameter)   log(Height)
       -6.632         1.983         1.117
```

We get $\widehat{\beta}_1 = 1.983$ and $\widehat{\beta}_2 = 1.117$.
Are they close to $\beta_1 = 2$ and $\beta_2 = 1$?
Need to know the variability of the LS estimates.

Standard Errors & Distributions of Least Squares Estimates

Least Squares Estimates Are Unbiased

Recall in L04, we said the LS estimate $\hat{\beta}$ in matrix notation is

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Based on the MLR model in matrix notation $Y = X\beta + \varepsilon$, the expected value of Y is

$$\mathrm{E}[Y] = \mathrm{E}[X\beta] + \underbrace{\mathrm{E}[\varepsilon]}_{=0} = X\beta$$

Recall in MLR, X are regard as fixed numbers, no randomness. The expected value of $\hat{\beta}$ is hence

$$\mathrm{E}[\hat{\beta}] = \mathrm{E}[(X^T X)^{-1} X^T Y] = (X^T X)^{-1} X^T \underbrace{\mathrm{E}[Y]}_{=X\beta} = (X^T X)^{-1} X^T X\beta = \beta$$

The LS estimate $\hat{\beta}$ is hence an unbiased estimate for β.

Variance of the LS Estimate $\hat{\beta}$

It can be shown that the variance of $\hat{\beta}$ in matrix notation is

$$\mathrm{Var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1},$$

where

$$\mathrm{Var}(\hat{\beta}) = \begin{bmatrix} \mathrm{Var}(\hat{\beta}_0) & \mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) & \mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_2) & \cdots & \mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_p) \\ \mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_0) & \mathrm{Var}(\hat{\beta}_1) & \mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2) & \cdots & \mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_p) \\ \mathrm{Cov}(\hat{\beta}_2, \hat{\beta}_0) & \mathrm{Cov}(\hat{\beta}_2, \hat{\beta}_1) & \mathrm{Var}(\hat{\beta}_2) & \cdots & \mathrm{Cov}(\hat{\beta}_2, \hat{\beta}_p) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(\hat{\beta}_p, \hat{\beta}_0) & \mathrm{Cov}(\hat{\beta}_p, \hat{\beta}_1) & \mathrm{Cov}(\hat{\beta}_p, \hat{\beta}_2) & \cdots & \mathrm{Var}(\hat{\beta}_p) \end{bmatrix}$$

and

$$(X^T X)^{-1} = \begin{bmatrix} n & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} & \cdots & \sum_{i=1}^n x_{ip} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1} x_{i2} & \cdots & \sum_{i=1}^n x_{i1} x_{ip} \\ \sum_{i=1}^n x_{i2} & \sum_{i=1}^n x_{i2} x_{i1} & \sum_{i=1}^n x_{i2}^2 & \cdots & \sum_{i=1}^n x_{i2} x_{ip} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_{ip} & \sum_{i=1}^n x_{ip} x_{i1} & \sum_{i=1}^n x_{ip} x_{i2} & \cdots & \sum_{i=1}^n x_{ip}^2 \end{bmatrix}^{-1}$$

Variance of the LS Estimates for SLR

For SLR, $(X^T X)^{-1}$ equals

$$\begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix}^{-1} = \begin{bmatrix} n & n\bar{x} \\ n\bar{x} & \sum_{i=1}^n x_i^2 \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i-\bar{x})^2} & -\frac{\bar{x}}{\sum_{i=1}^n (x_i-\bar{x})^2} \\ -\frac{\bar{x}}{\sum_{i=1}^n (x_i-\bar{x})^2} & \frac{1}{\sum_{i=1}^n (x_i-\bar{x})^2} \end{bmatrix}$$

Hence,

$$\begin{bmatrix} \mathrm{Var}(\hat{\beta}_0) & \mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) \\ \mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_0) & \mathrm{Var}(\hat{\beta}_1) \end{bmatrix} = \sigma^2 (X^T X)^{-1} = \sigma^2 \begin{bmatrix} \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i-\bar{x})^2} & -\frac{\bar{x}}{\sum_{i=1}^n (x_i-\bar{x})^2} \\ -\frac{\bar{x}}{\sum_{i=1}^n (x_i-\bar{x})^2} & \frac{1}{\sum_{i=1}^n (x_i-\bar{x})^2} \end{bmatrix}$$

We get that

$$\mathrm{Var}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right), \quad \mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

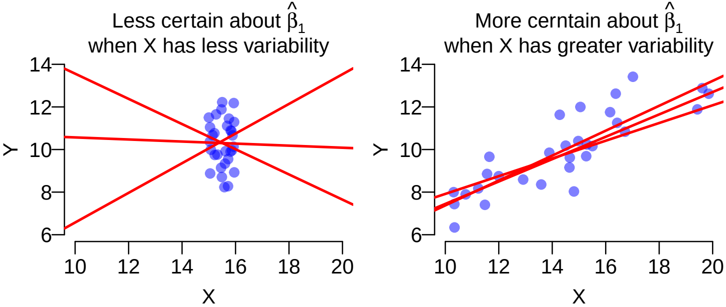
$$\mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_0) = -\frac{\sigma^2 \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Greater Variability in X, Better Estimate for Slope (SLR)

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{\sigma^2}{(n-1)s_x^2}, \quad \text{where } s_x^2 = \frac{\sum_{i=1}^n (x_i - \overline{x})^2}{n-1} = \left(\begin{array}{c}\text{sample}\\ \text{variance of } x\end{array}\right)$$

$\hat{\beta}_1$ will be closer to β_1 if

1) the sample size n is larger, or 2) X has greater variability



Remark 1: Another way to derive $\mathrm{Var}(\hat{\beta}_0)$:

$$\begin{aligned}\mathrm{Var}(\hat{\beta}_0) = \mathrm{Var}(\bar{y} - \hat{\beta}_1 \bar{x}) &= \mathrm{Var}(\bar{y}) - 2\bar{x} \overbrace{\mathrm{Cov}(\bar{y}, \hat{\beta}_1)}^{=0} + \bar{x}^2 \mathrm{Var}(\hat{\beta}_1) \\ &= \frac{\sigma^2}{n} + 0 + \bar{x}^2 \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\end{aligned}$$

- $\bar{y}$ and $\hat{\beta}_1$ are uncorrelated because the slope ($\hat{\beta}_1$) is invariant if you shift the response up or down ($\bar{y}$).

Remark 2: The LS estimates for the slope and the intercept are negatively correlated

$$\mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_0) = \mathrm{E}[(\hat{\beta}_1 - \beta_1)(\hat{\beta}_0 - \beta_0)] = \frac{-\sigma^2 \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

- If $\bar{x} > 0$, when the slope is high, the intercept tends to be low

Standard Error (s.e.) of the LS Estimate

For MLR,

- $\mathrm{Var}(\hat{\beta}_j) = \sigma^2 \times$(the jth diagonal element of $(X^T X)^{-1}$), $j = 0, 1, \ldots, p$
- $\mathrm{Cov}(\hat{\beta}_j, \hat{\beta}_k) = \sigma^2 \times$(the (j,k) entry of $(X^T X)^{-1}$), $j, k = 0, 1, \ldots, p$
- $s.e.(\hat{\beta}_j) = \sqrt{\mathrm{Var}(\hat{\beta}_j)}$ but the unknown σ^2 is replaced by its estimate

$$\text{MSE} = \frac{\sum_{i=1}^n e_i^2}{n-p-1}.$$

For SLR,

$$s.e.(\hat{\beta}_1) = \sqrt{\frac{\text{MSE}}{\sum_{i=1}^n (x_i - \bar{x})^2}}, \quad s.e.(\hat{\beta}_0) = \sqrt{\text{MSE}}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

R Can Compute the Standard Errors!

Don't worry about the computation of $s.e.(\hat{\beta}_j)$.

R can compute the **standard errors**!

```
lmtrees = lm(log(Volume) ~ log(Diameter) + log(Height), data=trees)
summary(lmtrees)$coef
              Estimate Std. Error t value  Pr(>|t|)
(Intercept)     -6.632    0.79979  -8.292 5.057e-09
log(Diameter)    1.983    0.07501  26.432 2.423e-21
log(Height)      1.117    0.20444   5.464 7.805e-06
```

- The column "**estimate**" shows the LS estimates

$$\hat{\beta}_0 = -6.6316, \quad \hat{\beta}_1 = 1.9826, \quad \text{and} \quad \hat{\beta}_2 = 1.1171$$

- The column "**std. error**" gives the standard errors:

$$s.e.(\hat{\beta}_0) = 0.7998, \quad s.e.(\hat{\beta}_1) = 0.075 \quad \text{and} \quad s.e.(\hat{\beta}_2) = 0.2044$$

```
lmtrees = lm(log(Volume) ~ log(Diameter) + log(Height), data=trees)
summary(lmtrees)

Call:
lm(formula = log(Volume) ~ log(Diameter) + log(Height), data = trees)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.16856 -0.04849  0.00243  0.06364  0.12922 

Coefficients:
              Estimate Std. Error t value     Pr(>|t|)
(Intercept)     -6.632      0.800   -8.29 0.0000000051
log(Diameter)    1.983      0.075   26.43      < 2e-16
log(Height)      1.117      0.204    5.46 0.0000078053

Residual standard error: 0.0814 on 28 degrees of freedom
Multiple R-squared:  0.978, Adjusted R-squared:  0.976 
F-statistic:  613 on 2 and 28 DF,  p-value: <2e-16
```

Hypothesis Tests & Confidence Intervals for Coefficients

t-Test of a Single β_j

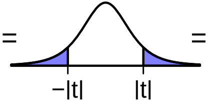
For an MLR model $Y_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip} + \varepsilon_i$, the t-statistic for testing H$_0$: $\beta_j = \beta_j^0$ is

$$t = \frac{\hat{\beta}_j - \beta_j^0}{s.e.(\hat{\beta}_j)} \quad \text{which has a } t\text{-distribution with df} = n - p - 1$$

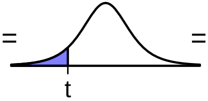
where $s.e.(\hat{\beta}_j)$ is given on the previous slide.

The P-value can be calculated using `pt()` based on the alternative hypothesis H_1.

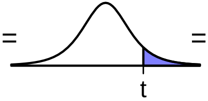
If H_a: $\beta_j \neq \beta_j^0$ (two-sided),

P−Value =  = 2*pt(abs(t), df, lower.tail=FALSE)

If H_a: $\beta_j < \beta_j^0$ (lower one-sided),

P−Value =  = pt(t, df)

If H_a: $\beta_j > \beta_j^0$ (upper one-sided),

P−Value =  = pt(t, df, lower.tail=FALSE)

The bell curve above is the t-curve with $\text{df} = N - p - 1$.

```
summary(lmtrees)$coef
              Estimate Std. Error t value  Pr(>|t|)
(Intercept)     -6.632    0.79979  -8.292 5.057e-09
log(Diameter)    1.983    0.07501  26.432 2.423e-21
log(Height)      1.117    0.20444   5.464 7.805e-06
```

As we believe $\beta_1 = 2$ and $\beta_2 = 1$, can test them w/ the t-statistics

$$t_1 = \frac{\hat{\beta}_1 - 2}{s.e.(\hat{\beta}_1)} = \frac{1.9826 - 2}{0.075} \approx -0.2313 \text{ with df } = 31 - 2 - 1 = 28$$

$$t_2 = \frac{\hat{\beta}_2 - 1}{s.e.(\hat{\beta}_2)} = \frac{1.1171 - 1}{0.2044} \approx 0.5729 \text{ with df } = 31 - 2 - 1 = 28.$$

The two-sided p-values are about 0.82 and 0.57

```
2*pt(0.2313, df = 28, lower.tail=F)
[1] 0.8188
2*pt(0.5729, df = 28, lower.tail=F)
[1] 0.5713
```

That is, "Volume $\approx$ (Diameter)^2(Height)" seems reasonable.

```r
summary(lmtrees)$coef
              Estimate Std. Error t value  Pr(>|t|)
(Intercept)     -6.632    0.79979  -8.292 5.057e-09
log(Diameter)    1.983    0.07501  26.432 2.423e-21
log(Height)      1.117    0.20444   5.464 7.805e-06
```

- The column "**statistic** shows the t-statistic for testing H$_0$: $\beta_j = 0$ against H$_1$: $\beta_j \neq 0$,

$$t_0 = \frac{\hat{\beta}_0 - 0}{s.e.(\hat{\beta}_0)} = \frac{-6.632 - 0}{0.79979} = -8.292,$$

$$t_1 = \frac{\hat{\beta}_1 - 0}{s.e.(\hat{\beta}_1)} = \frac{1.983 - 0}{0.07501} = 26.432,$$

$$t_2 = \frac{\hat{\beta}_2 - 0}{s.e.(\hat{\beta}_2)} = \frac{1.117 - 0}{0.20444} = 5.464$$

  which are simply **the ratios of the first two columns**
- The column Pr($>|t|$) shows the **2-sided** P-values for testing H$_0$: $\beta_0 = 0$ and H$_0$: $\beta_1 = 0$.

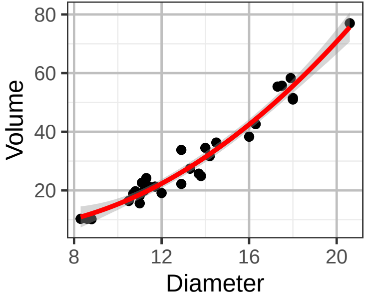
Digression: Checking Non-Linearity

Recall we said earlier that the relation between Volume and Diameter is slightly nonlinear.

We can check **nonlinearity** by fitting the polynomial model

$$\text{Volume} = \beta_0 + \beta_1 \text{Diameter} + \beta_2 (\text{Diameter})^2 + \varepsilon$$

and test H_0: $\beta_2 = 0$



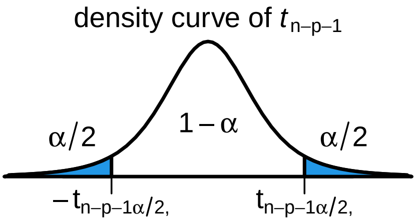

```
lm2 = lm(Volume ~ Diameter + I(Diameter^2), data=trees)
summary(lm2)$coef
              Estimate Std. Error t value  Pr(>|t|)
(Intercept)    10.7863   11.22282  0.9611 0.3447282
Diameter       -2.0921    1.64734 -1.2700 0.2145344
I(Diameter^2)   0.2545    0.05817  4.3756 0.0001524
```

Confidence Intervals For Coefficients

The $100(1-\alpha)\%$ confidence interval for β_j is

$$\hat{\beta}_j \pm t_{(n-p-1,\alpha/2)} s.e.(\hat{\beta}_j)$$

where $t_{(n-p-1,\alpha/2)} = t^*$ is the critical value for the t_{n-p-1} distribution at confidence level $1-\alpha$, i.e.,



which can be found using either of the R commands

```
qt(alpha/2, df=n-p-1, lower.tail=FALSE)
qt(1-alpha/2, df=n-p-1)
```

Example: CI for β_1

```
summary(lmtrees)$coef
               Estimate Std. Error t value  Pr(>|t|)
(Intercept)      -6.632    0.79979  -8.292 5.057e-09
log(Diameter)     1.983    0.07501  26.432 2.423e-21
log(Height)       1.117    0.20444   5.464 7.805e-06
```

A 95% confidence interval for β_1 is

$$\begin{aligned}\hat{\beta}_1 \pm t_{0.05/2,28}\mathsf{SE}(\hat{\beta}_1) &\approx 1.983 \pm 2.048 \times 0.07501 \\ &\approx 1.983 \pm 0.1536 \approx (1.829, 2.137)\end{aligned}$$

where $t_{28,0.05/2} \approx 2.048$ can be found using either R commands below

```
qt(0.05/2, df=28, lower.tail=F)
[1] 2.048
qt(0.975, df=28)
[1] 2.048
```

Finding CIs for Coefficients Using `confint()`

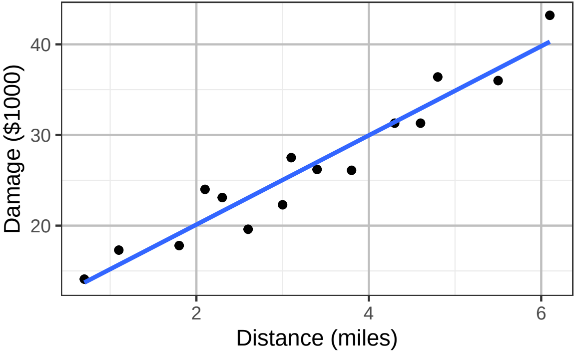
The R command `confint()` can produce confidence intervals for coefficients β_i's.

```
confint(lmtrees)
                2.5 % 97.5 %
(Intercept)   -8.2699 -4.993
log(Diameter)  1.8290  2.136
log(Height)    0.6984  1.536
confint(lmtrees, level = 0.9)  # changing the confidence level to 90%
                  5 %   95 %
(Intercept)   -7.9922 -5.271
log(Diameter)  1.8550  2.110
log(Height)    0.7693  1.465
confint(lmtrees, level = 0.95, "log(Diameter)")
              2.5 % 97.5 %
log(Diameter) 1.829  2.136
confint(lmtrees, level = 0.95, "(Intercept)")
            2.5 % 97.5 %
(Intercept) -8.27 -4.993
```

Example: Fire Damage Data

Distance (mile)	Damage ($1000)
0.7	14.1
1.1	17.3
1.8	17.8
2.1	24.0
2.3	23.1
2.6	19.6
3.0	22.3
3.1	27.5
3.4	26.2
3.8	26.1
4.3	31.3
4.6	31.3
4.8	36.4
5.5	36.0
6.1	43.2

An insurance company wanted to relate the amount of fire damage in major residential fires to the distance between the burning house and the nearest fire station. A sample of 15 recent fires was selected in a large suburb of a major city.



Fire Damage Data

```
fire = data.frame(
  dist=c(0.7,1.1,1.8,2.1,2.3,2.6,3.0,3.1,3.4,3.8,4.3,4.6,4.8,5.5,6.1),
  damage=c(14.1,17.3,17.8,24.0,23.1,19.6,22.3,27.5,26.2,26.1,31.3,31.3,36.4,36.0,43.2)
)
lmfire = lm(damage ~ dist, data=fire)
lmfire$coef
(Intercept)        dist
     10.278       4.919
```

$$\widehat{\text{fire damage in \$1000's}} = 10.28 + 4.92 \times (\text{distance to the nearest fire dept})$$

- The intercept 10.278 means the predicted fire damage for houses located right next to a fire station is $10,278.
- The slope 4.919 means every extra mile to the nearest fire station increases the fire damage by $4,919 on average

Example: Test for the Slope β_1

	Estimate	Std.Error	t value	Pr(>|t|)
(Intercept)	10.2779	1.4203	7.237	6.59e-06
dist	4.9193	0.3927	12.525	1.25e-08

To test H$_0$: $\beta_1 = 4$ v.s. H$_A$: $\beta_1 > 4$, the t-statistic is

$$t = \frac{b_1 - 4}{\text{SE}(b_1)} = \frac{4.9193 - 4}{0.3927} = 2.3409, \quad \text{df} = n - 2 = 15 - 2 = 13.$$

The upper one-sided P-value can be found in R to be ≈ 0.018.

```
pt(2.3409, df=13, lower.tail=F)
[1] 0.01791
```

Conclusion: At 5% level, the extra amount of damage for every extra mile from the nearest fire station is significantly higher than $4000 on average.

Example: CI for β_1

	Estimate	Std.Error	t value	Pr(>|t|)
(Intercept)	10.2779	1.4203	7.237	6.59e-06
dist	4.9193	0.3927	12.525	1.25e-08

A 95% confidence interval for β_1 is

$$\hat{\beta}_1 \pm t_{13,0.05/2}\text{SE}(\hat{\beta}_1) \approx 4.9193 \pm 2.16 \times 0.3927 \approx 4.919 \pm 0.848 = (4.071, 5.767)$$

where $t_{13,0.05/2} \approx 2.16$ can be found by he R command

```
qt(0.05/2, df=13, lower.tail=F)
[1] 2.16
confint(lmfire, "dist")
     2.5 % 97.5 %
dist 4.071  5.768
```

Interpretation: We have 95% confidence that every extra mile from the nearest fire station increases the amount of damage by \$4071 to \$5767 on average

Coming Up Next

- Confidence Intervals and Prediction Intervals for Prediction
- Sum of Squares
- Model Comparison