Inference for Regression Coefficients

Math 430, Winter 2017

Climate data

- Measurements on CO_2 in the atmosphere and global temperature anomaly (deviation from the mean temperature from 1961 to 1990)

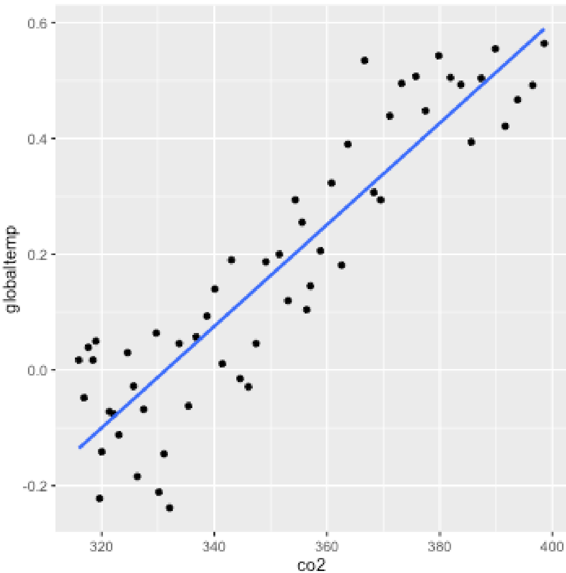
```
climate <- read.csv("https://github.com/math430-lu/data/raw/master/cl
head(climate)
```

```
##   year    co2 globaltemp
## 1 1959 315.97      0.017
## 2 1960 316.91     -0.048
## 3 1961 317.64      0.039
## 4 1962 318.45      0.017
## 5 1963 318.99      0.050
## 6 1964 319.62     -0.222
```

- Goals
    - understand the relationship between CO_2 and global temperatures
    - make predictions

Climate data

```
ggplot(data = climate, mapping = aes(x = co2, y = globaltemp)) +
  geom_point() +
  geom_smooth(method="lm", se = FALSE)
```



Fitting the SLR model

```
climate.lm <- lm(globaltemp ~ co2, data = climate)
summary(climate.lm)
```

```
##
## Call:
## lm(formula = globaltemp ~ co2, data = climate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.24377 -0.08048  0.01431  0.07905  0.22558
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.9083486  0.1943286  -14.97   <2e-16 ***
## co2          0.0087761  0.0005527   15.88   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1016 on 54 degrees of freedom
## Multiple R-squared:  0.8236,    Adjusted R-squared:  0.8204
## F-statistic: 252.2 on 1 and 54 DF,  p-value: < 2.2e-16
```

Inference for the slope

Statistical inference

Goal: use statistics calculated from data to makes inferences about the nature of parameters.

- Statistics: $\widehat{\beta}_0, \widehat{\beta}_1$
- Parameters: β_0, β_1

Tools:

- Confidence intervals
- Hypothesis tests

Overview of statistical inference

Confidence intervals

Idea: A confidence interval expresses the amount of uncertainly we have in our estimate of a particular parameter.

To find such a range of plausible values for the parameter of interest, θ, so that we know the long-run properties of the intervals.

$$P(\widehat{\theta}_L < \theta < \widehat{\theta}_U) = 1 - \alpha$$

- The endpoints are random variables **before** observing the data
- θ is fixed but unknown

Confidence intervals

A general $1 - \alpha$ confidence interval takes the form

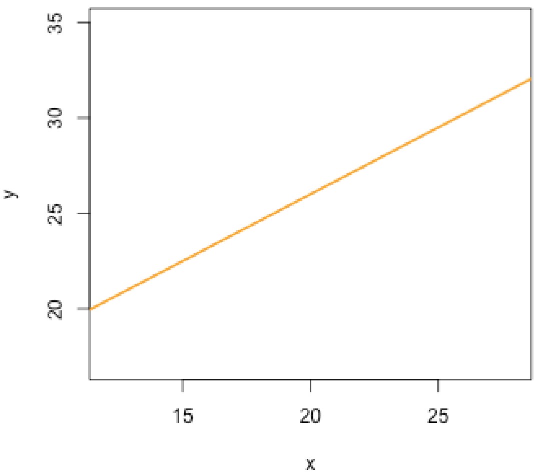
$$\widehat{\theta} \pm t^* \cdot SE(\widehat{\theta})$$

- α: the confidence level
- $\widehat{\theta}$: a statistic (i.e. point estimate)
- t^*: the $1 - \alpha/2$ quantile of a reference distribution
- $SE(\widehat{\theta})$: the standard error of $\widehat{\theta}$; i.e. the standard deviation of the sampling distribution

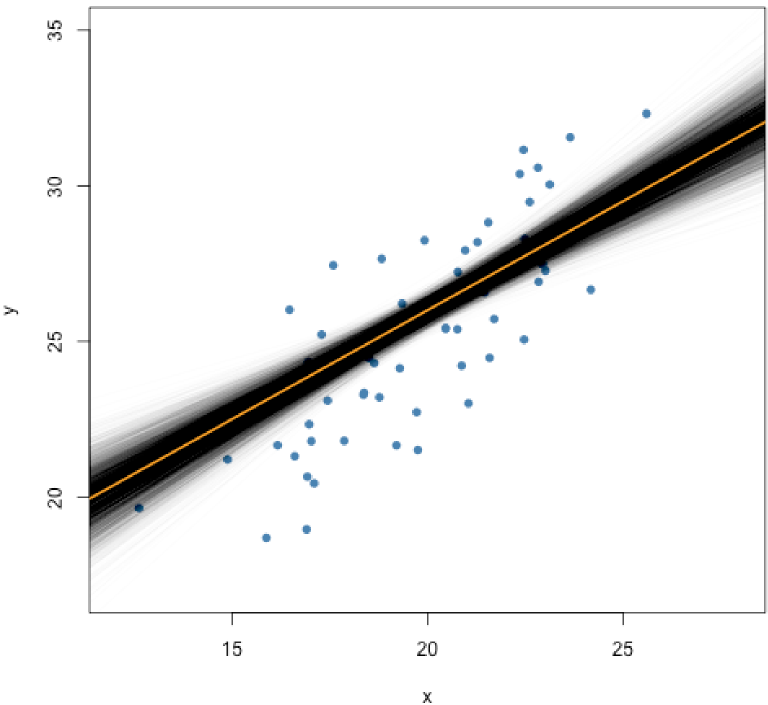
Sampling distribution of the slope

Assume that the true model is

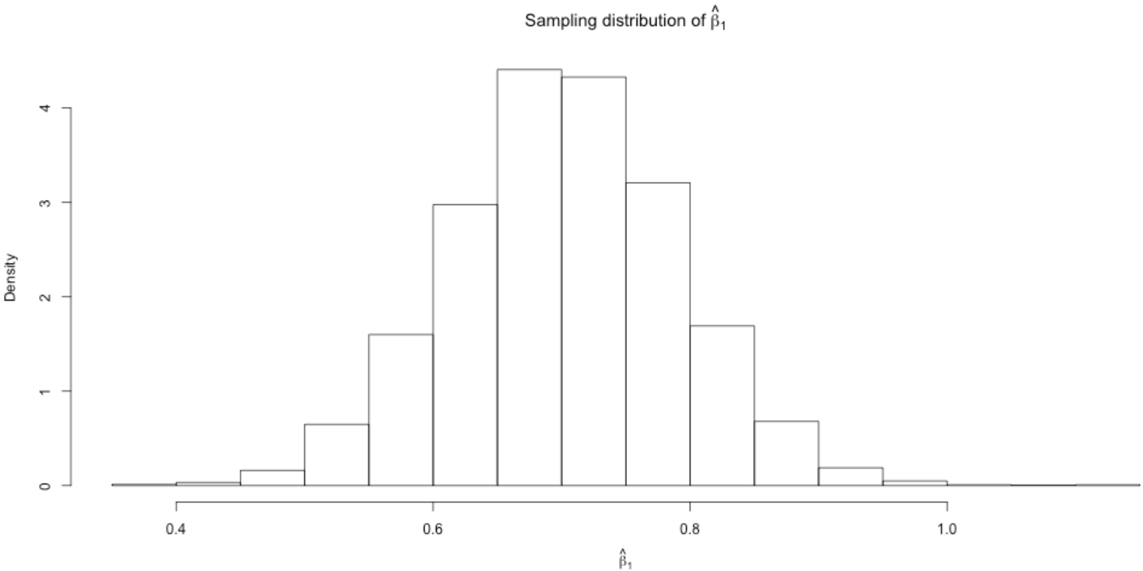
$$E(Y|X = x) = 12 + .7x, \quad e \sim \mathcal{N}(0, 4)$$



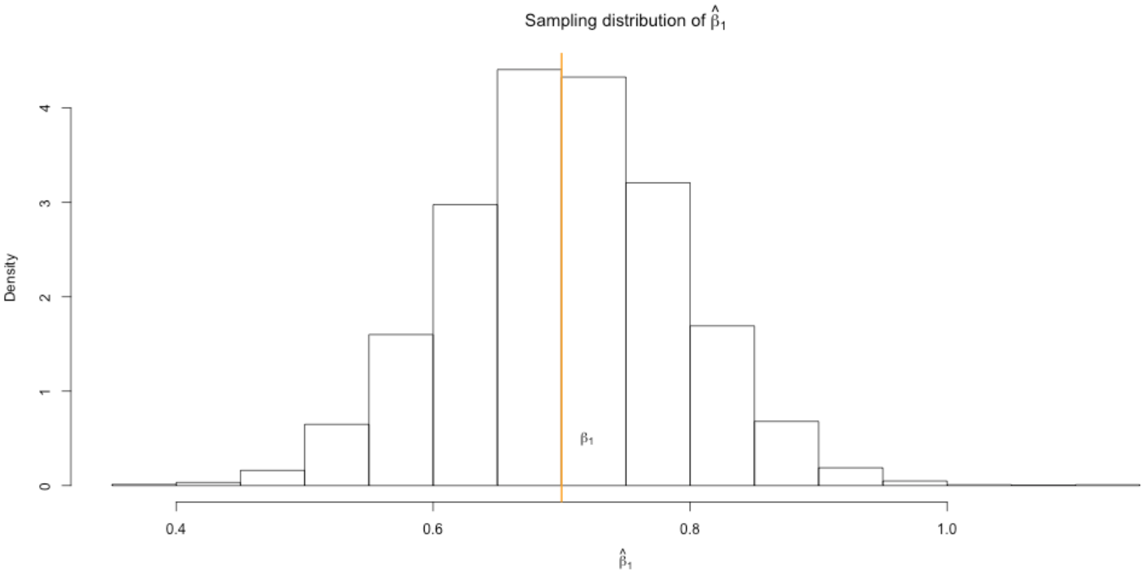
Sampling distribution of the slope



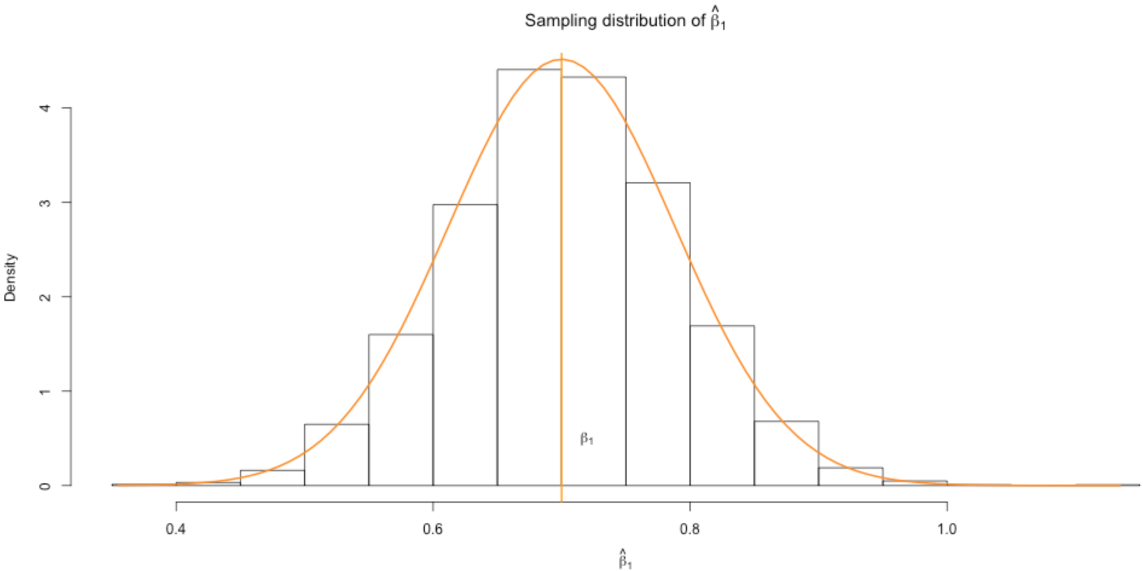
Sampling distribution of the slope



Sampling distribution of the slope



Sampling distribution of the slope



Properties

1. $E(\widehat{\beta}_1|X) = \beta_1$

2. $Var(\widehat{\beta}_1|X) = \dfrac{\sigma^2}{SXX}$

3. $\widehat{\beta}_1|X \sim \mathcal{N}\left(\beta_1, \dfrac{\sigma^2}{SXX}\right)$

Approximating the sampling distribution

We don't know σ^2, so we have to plug in our best guess at it's value, $S^2 = RSS/(n-2)$.

- The distribution is no longer normal due to the added uncertainty (heavier tails)
- Use the t distribution with $n-2$ degrees of freedom (d.f.)
- Use R to find the quantiles

```
qt(1 - alpha/2, df = n-2)
```

CI for the slope

$$\widehat{\beta}_1 \pm t^*_{\alpha/2,n-2} \frac{S}{\sqrt{SXX}}$$

```r
summary(climate.lm)
```

```
##
## Call:
## lm(formula = globaltemp ~ co2, data = climate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.24377 -0.08048  0.01431  0.07905  0.22558
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.9083486  0.1943286  -14.97   <2e-16 ***
## co2          0.0087761  0.0005527   15.88   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1016 on 54 degrees of freedom
## Multiple R-squared:  0.8236,    Adjusted R-squared:  0.8204
## F-statistic: 252.2 on 1 and 54 DF,  p-value: < 2.2e-16
```

An easier way in R

```r
confint(climate.lm, level = 0.95)
```

```
##                    2.5 %       97.5 %
## (Intercept) -3.297953969 -2.518743249
## co2          0.007668089  0.009884172
```

Interpreting CIs

We are 95% confident that the true slope between x and y lies between LB and UB.

For our climate example:

Hypothesis testing framework

1. Formulate two competing hypotheses: H_0 and H_A.

2. Choose a test statistic that characterizes the information in the sample relevant to H_0.

3. Determine the sampling distribution of the chosen statistic when H_0 is true.

4. Compare the calculated test statistic to the sampling distribution to determine whether it is "extreme."

Tests for the slope

Competing Claims: $H_0 : \beta_1 = \beta_1^0$ vs. $H_a : \beta_1 \neq \beta_1^0$

(R assumes that $\beta_1^0 = 0$)

Test statistic: $T = \dfrac{\widehat{\beta}_1 - \beta_1^0}{SE(\widehat{\beta}_1)}$

Reference distribution: $T \sim t_{n-2}$ when H_0 is true

In R

```
summary(climate.lm)
```

```
##
## Call:
## lm(formula = globaltemp ~ co2, data = climate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.24377 -0.08048  0.01431  0.07905  0.22558
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.9083486  0.1943286  -14.97   <2e-16 ***
## co2          0.0087761  0.0005527   15.88   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1016 on 54 degrees of freedom
## Multiple R-squared:  0.8236,    Adjusted R-squared:  0.8204
## F-statistic: 252.2 on 1 and 54 DF,  p-value: < 2.2e-16
```

p-values

Inference for the intercept

Sampling distribution of the intercept

$$\widehat{\beta}_0 | X \sim \mathcal{N}\left(\beta_0,\ \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{SXX}\right)\right)$$

Again, we can only estimate σ^2 using $S^2 = RSS/(n-2)$

Inference for the intercept

Test statistic:

$$T = \frac{\widehat{\beta}_0 - \beta_0^0}{se(\widehat{\beta}_0)}, \text{ where } se(\widehat{\beta}_0) = S\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SXX}}$$

$(1-\alpha)100$**% CI:**

$$\widehat{\beta}_0 \pm t^*_{\alpha/2,n-2} \cdot se(\widehat{\beta}_0)$$

Centered SLR Model

Issue: β_0 is usually not interpretable

Solution: Center the predictor variable

$$x_i^* = x_i - \overline{x}$$

and fit the model

$$E(Y_i|X) = \beta_0^* + \beta_1^* x_i^*$$

Advantages:

- Intercept is now the average/predicted value of y when $x_i = \overline{x}$
- Slope and residual standard deviation stay the same

Centering a variable in R

```r
library(dplyr)
climate <- mutate(climate, co2.center = co2 - mean(co2))
head(climate)
```

```
##   year    co2 globaltemp co2.center
## 1 1959 315.97      0.017  -34.78768
## 2 1960 316.91     -0.048  -33.84768
## 3 1961 317.64      0.039  -33.11768
## 4 1962 318.45      0.017  -32.30768
## 5 1963 318.99      0.050  -31.76768
## 6 1964 319.62     -0.222  -31.13768
```

Inference for β_0 in Centered SLR Model

Test statistic:

$t = \dfrac{\widehat{\beta}_0^* - \beta_0^*}{se(\widehat{\beta}_0^*)}$, where $se(\widehat{\beta}_0^*) = S/\sqrt{n}$

$(1-\alpha)100$**% CI:**

$$\widehat{\beta}_0^* \pm t^*_{\alpha/2,n-2} \cdot se(\widehat{\beta}_0^*)$$

Centered SLR Model in R

```
centered.lm <- lm(globaltemp ~ co2.center, data = climate)
summary(centered.lm)
```

```
##
## Call:
## lm(formula = globaltemp ~ co2.center, data = climate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.24377 -0.08048  0.01431  0.07905  0.22558
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.1699464  0.0135717   12.52   <2e-16 ***
## co2.center  0.0087761  0.0005527   15.88   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1016 on 54 degrees of freedom
## Multiple R-squared:  0.8236,    Adjusted R-squared:  0.8204
## F-statistic: 252.2 on 1 and 54 DF,  p-value: < 2.2e-16
```