Multiple Linear Regression: Inference for multiple linear regression

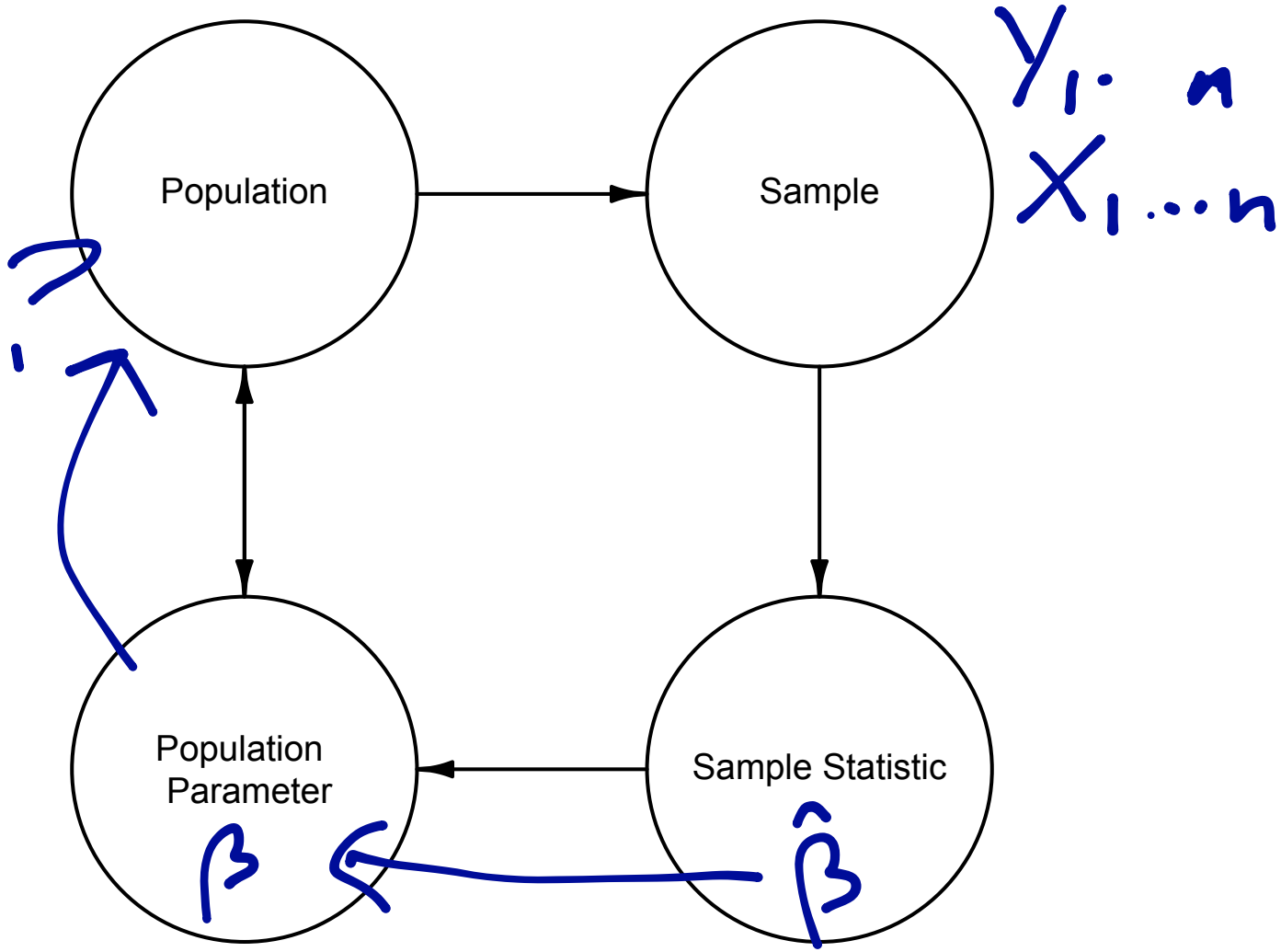
Author: Nicholas G Reich, Jeff Goldsmith

This material is part of the **statsTeachR** *project*

Today's Lecture

- Sampling distribution of $\hat{\beta}$
- Confidence intervals
- Hypothesis tests for individual coefficients
- Global tests (next week!)

Circle of Life



Statistical inference

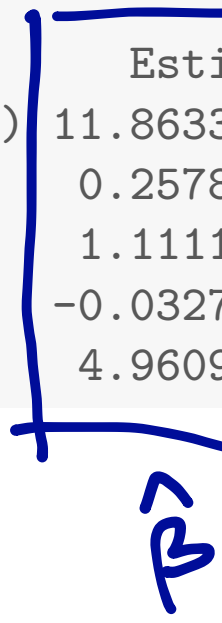
- We have LSEs $\hat{\beta}_0, \hat{\beta}_1, \ldots$; we want to know what this tells us about $\beta_0, \beta_1, \ldots$.
- Two basic tools are confidence intervals and hypothesis tests
  - Confidence intervals provide a plausible range of values for the parameter of interest based on the observed data
  - Hypothesis tests ask how probable are the data we gathered under a null hypothesis about the data generating distribution

Motivation

How can we draw **inference** about each of these parameters and relationships that our model is encoding?

```
mlr1 <- lm(disease ~ airqual + crowding + nutrition + smoking,
           data=dat)
summary(mlr1)$coef
```

```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 11.86333314 2.578819159  4.600297 1.315919e-05
## airqual      0.25788257 0.026799356  9.622715 1.165263e-15
## crowding     1.11112603 0.102036855 10.889458 2.403742e-18
## nutrition   -0.03278397 0.007953614 -4.121896 8.094957e-05
## smoking      4.96093131 1.085292354  4.571055 1.475259e-05
```

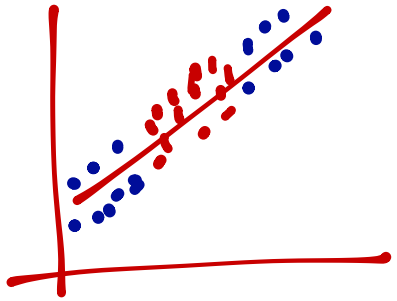


Motivation

- Can we say anything about whether the effect of `airquality` is "significant" after adjusting for other variables?
- Can we say whether adding `airquality` improves the fit of our model?
- Can we compare this model to a model with `crowding`, `nutrition` and `smoking`?

model selection

Sampling distribution



If our usual assumptions are satisfied and $\epsilon \overset{iid}{\sim} \mathrm{N}\left[0, \sigma^2\right]$ then

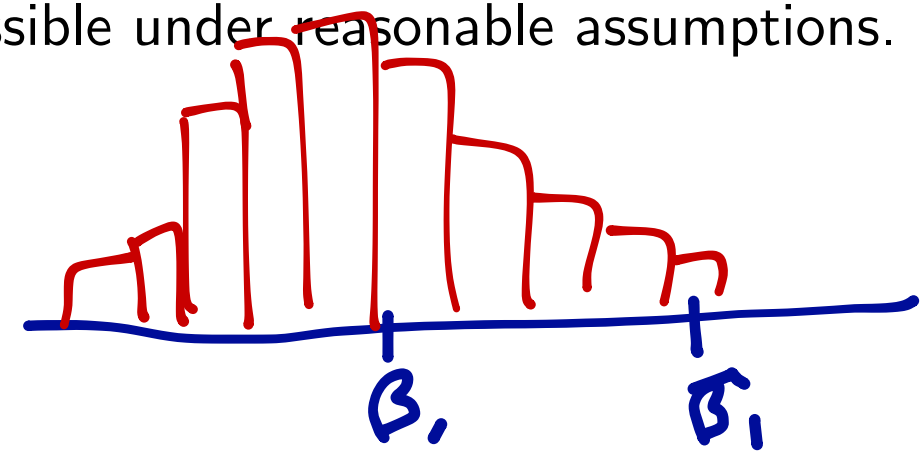
$$\hat{\beta} \sim \mathrm{N}\left[\beta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\right].$$

multivariate

$$\hat{\beta}_j \sim \mathrm{N}\left[\beta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}_{jj}\right].$$

univariate

- This will be used later for inference.
- Even without Normal errors, asymptotic Normality of LSEs is possible under reasonable assumptions.



Sampling distribution

For real data we have to estimate σ^2 as well as $\boldsymbol{\beta}$.

- Recall our estimate of the error variance is

$$\hat{\sigma^2} = \frac{RSS}{n-p-1} = \frac{\sum_i (y_i - \hat{y}_i)^2}{n-p-1}$$

- With Normally distributed errors, it can be shown that

$$(n-p-1)\frac{\hat{\sigma^2}}{\sigma^2} \sim \chi^2_{n-p-1}$$

Testing procedure

Calculate the probability of the observed data (or more extreme data) under a null hypothesis.

- Often $H_0 : \beta_j = 0$ and $H_a : \beta_j \neq 0$
- Set type I error rate
  $\alpha = P(\text{falsely rejecting a true null hypothesis})$ = .05
- Calculate a test statistic assuming the null hypothesis is true
- Compute a p-value =

$$P(\hat{\beta}_j \text{ as or more extreme as observed} | H_0)$$

- Reject or fail to reject H_0

Individual coefficients

For individual coefficients

- We can use the test statistic

$$T = \frac{\hat{\beta}_j - \beta_j}{\widehat{se}(\hat{\beta}_j)} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}_{jj}}} \sim t_{n-p-1}$$

if $H_0: \beta_j = 0$, then $\beta_j = 0$

- For a two-sided test of size α, we reject if

$$|T| > t_{1-\alpha/2, n-p-1}$$

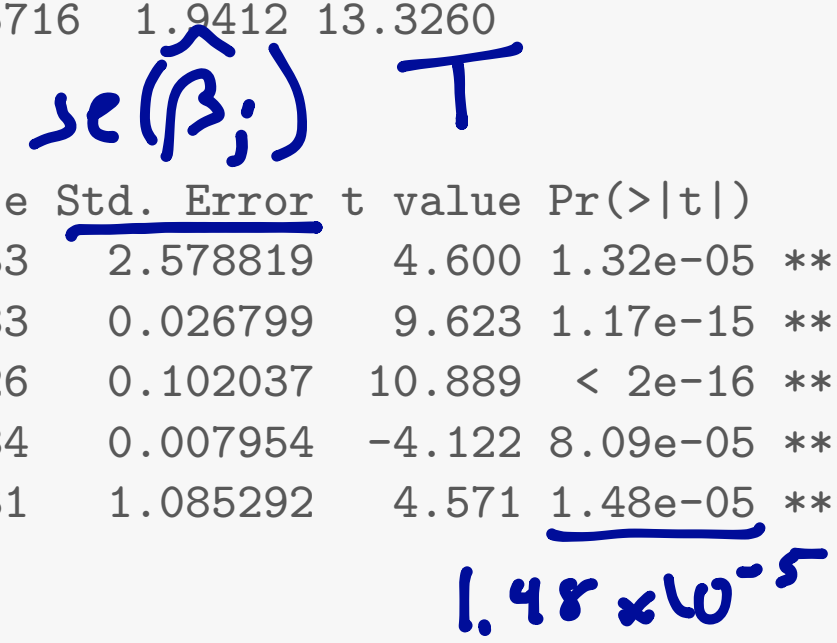
- The p-value gives $P(t_{n-p-1} > T_{obs}|H_0)$

Note that t is a symmetric distribution that converges to a Normal as $n - p - 1$ increses.

Back to the example

```
summary(mlr1)


##
## Call:
## lm(formula = disease ~ airqual + crowding + nutrition + smoking,
##     data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.1297 -2.1834 -0.5716  1.9412 13.3260
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.863333   2.578819   4.600 1.32e-05 ***
## airqual      0.257883   0.026799   9.623 1.17e-15 ***
## crowding     1.111126   0.102037  10.889  < 2e-16 ***
## nutrition   -0.032784   0.007954  -4.122 8.09e-05 ***
## smoking      4.960931   1.085292   4.571 1.48e-05 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```



Individual coefficients: CIs

Alternatively, we can construct a confidence interval for β_j

- A confidence interval with coverage $(1-\alpha)$ is given by

$$\hat{\beta}_j \pm t_{1-\alpha/2,n-p-1}\widehat{se}(\hat{\beta}_j)$$

- Assuming all the standard assumptions hold,

$$(1-\alpha) = P(LB < \beta_j < UB)$$

Detour: confidence interval interpretations

The semantics of confidence intervals are tricky!

The technically correct interpretation of a (frequentist) confidence interval is:
if the current experiment were repeated under similar conditions, we expect that $1 - \alpha\%$ *of the time the confidence interval for a parameter would cover the true value of the parameter.*

Detour: confidence interval interpretations

Possible interpretations

- "There is a 95% probability that this confidence interval contains the true value of the parameter."
  WRONG!
- "We are 95% confident that this interval contains the truth."
  NOT VERY TECHNICALLY SPECIFIC, BUT NOT INCORRECT EITHER.
- "The 95% confidence interval for this parameter is (a, b)."
  COMMONLY USED, ASSUMES THE READER KNOWS HOW TO INTERPRET.
- "With confidence coefficient .95, we estimate that the average change in Y per 1 unit increase of X lies somewhere between (a and b)."
  TECHNICALLY CORRECT, BUT NOT CLEAR WHAT CONF COEF IS.

Back to the example

$\alpha = .05$

$1 - \alpha/2$

```
cbind(coef(mlr1), confint(mlr1))
```

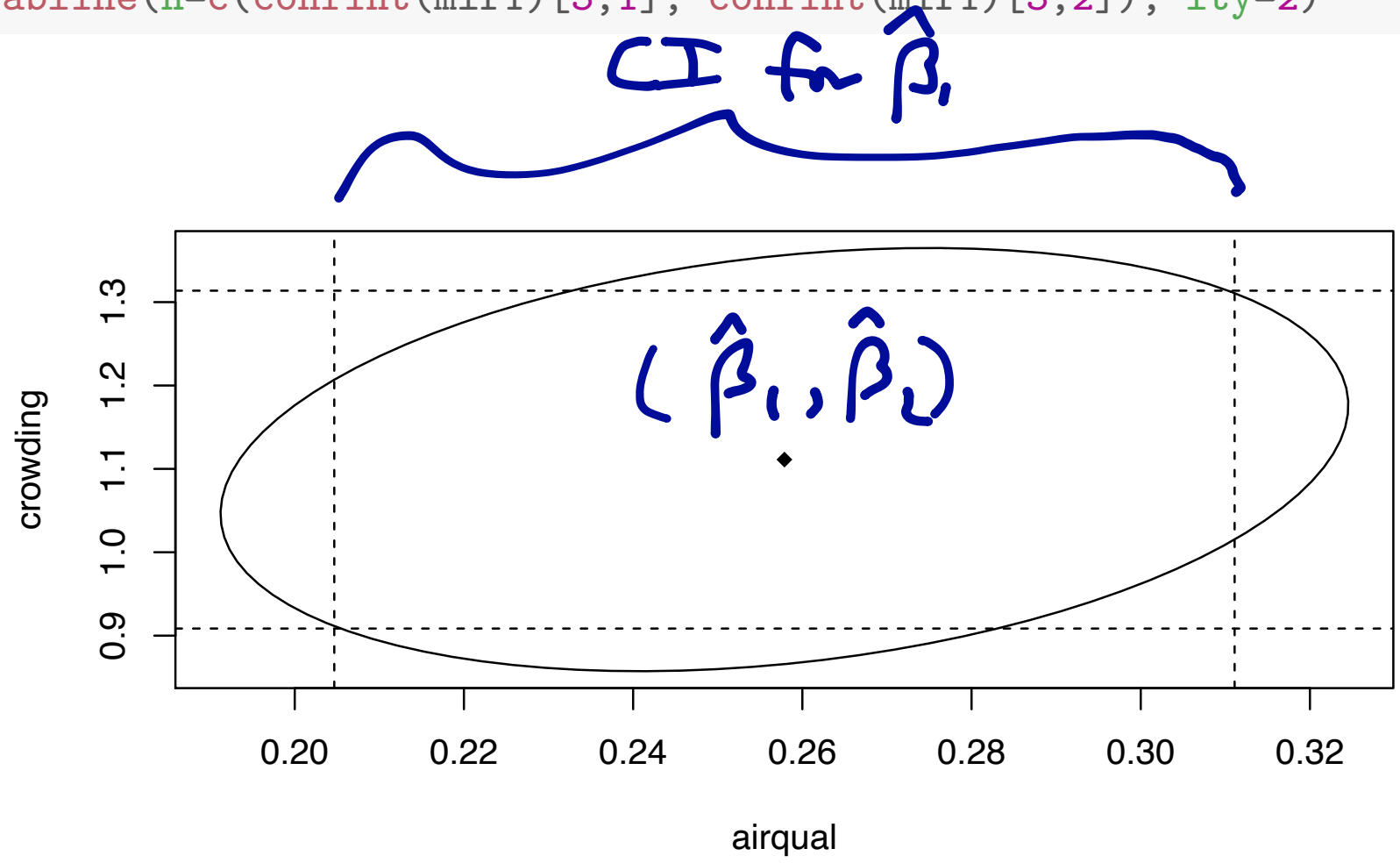
```
##                                  2.5 %     97.5 %
## (Intercept) 11.86333314  6.74302724 16.98363903
## airqual      0.25788257  0.20467182  0.31109332
## crowding     1.11112603  0.90852947  1.31372260
## nutrition   -0.03278397 -0.04857606 -0.01699189
## smoking      4.96093131  2.80605790  7.11580472
```

Confidence regions for multiple parameters

If you want to draw inference about multiple parameters, it is better to look at them simultaneously.

Plotting 2D confidence regions

```
library(ellipse)
plot(ellipse(mlr1,c(2,3)),type="l")
points(coef(mlr1)[2],coef(mlr1)[3], pch=18)
abline(v=c(confint(mlr1)[2,1], confint(mlr1)[2,2]), lty=2)
abline(h=c(confint(mlr1)[3,1], confint(mlr1)[3,2]), lty=2)
```



Today's Big Ideas

- Basic parameter inference for multiple linear regression models

Lab on regression inference

Run the code for today's class (on the website), and modify it to answer the following questions:

- Compute the 95% confidence interval coverage for β_1. What is it and is it what you would expect?
- Given the constant values defined at the top of the file, determine what the sampling distribution for β_1 should be. Using the estimated values of the $\hat{\beta}_1$, calulate summary metrics and or use appropriate visualizations to determine whether these your simulated distribution of $\hat{\beta}_1$ matches with the theoretical distribution.
- Adapt the simulation to simulate data for two covariates, x_1 and x_2, using `mvrnorm()`. Define x_1 and x_2 so that you may modify the degree of correlation between them. Run the simulation again for two scenarios, one with low and one with high correlation. For each of these scenarios, does the 95% confidence interval coverage change?