Multiple Linear Regression: Parameter Inference

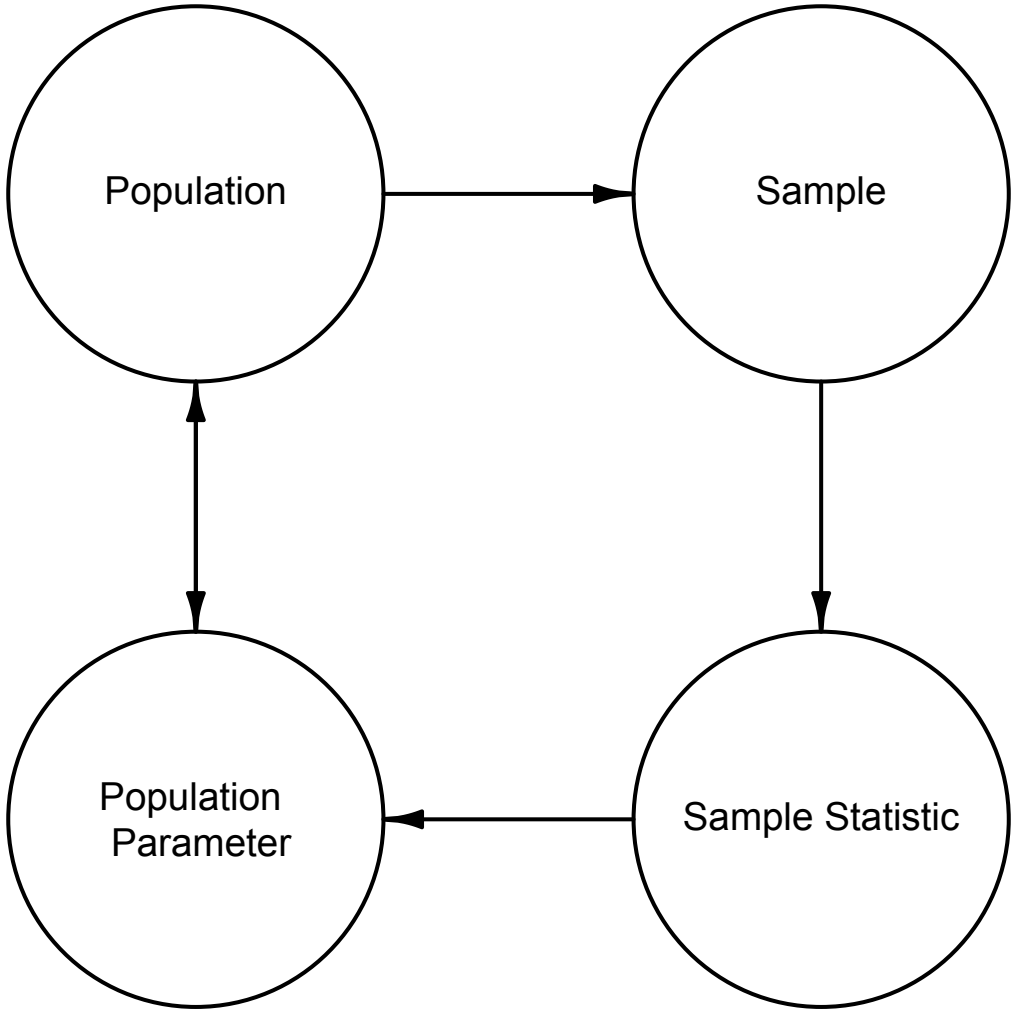
Author: Nicholas G Reich, Jeff Goldsmith

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

Today's Lecture

- Sampling distribution of $\hat{\boldsymbol{\beta}}$
- Confidence intervals
- Hypothesis tests for individual coefficients
- Multiple testing

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`?

Sampling distribution

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

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

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

- 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
  $\alpha = P(\text{falsely rejecting a true null hypothesis})$
- 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

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

- 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
##
## Residual standard error: 3.644 on 94 degrees of freedom
## Multiple R-squared:  0.8664, Adjusted R-squared:  0.8607
## F-statistic: 152.4 on 4 and 94 DF,  p-value: < 2.2e-16
```

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

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

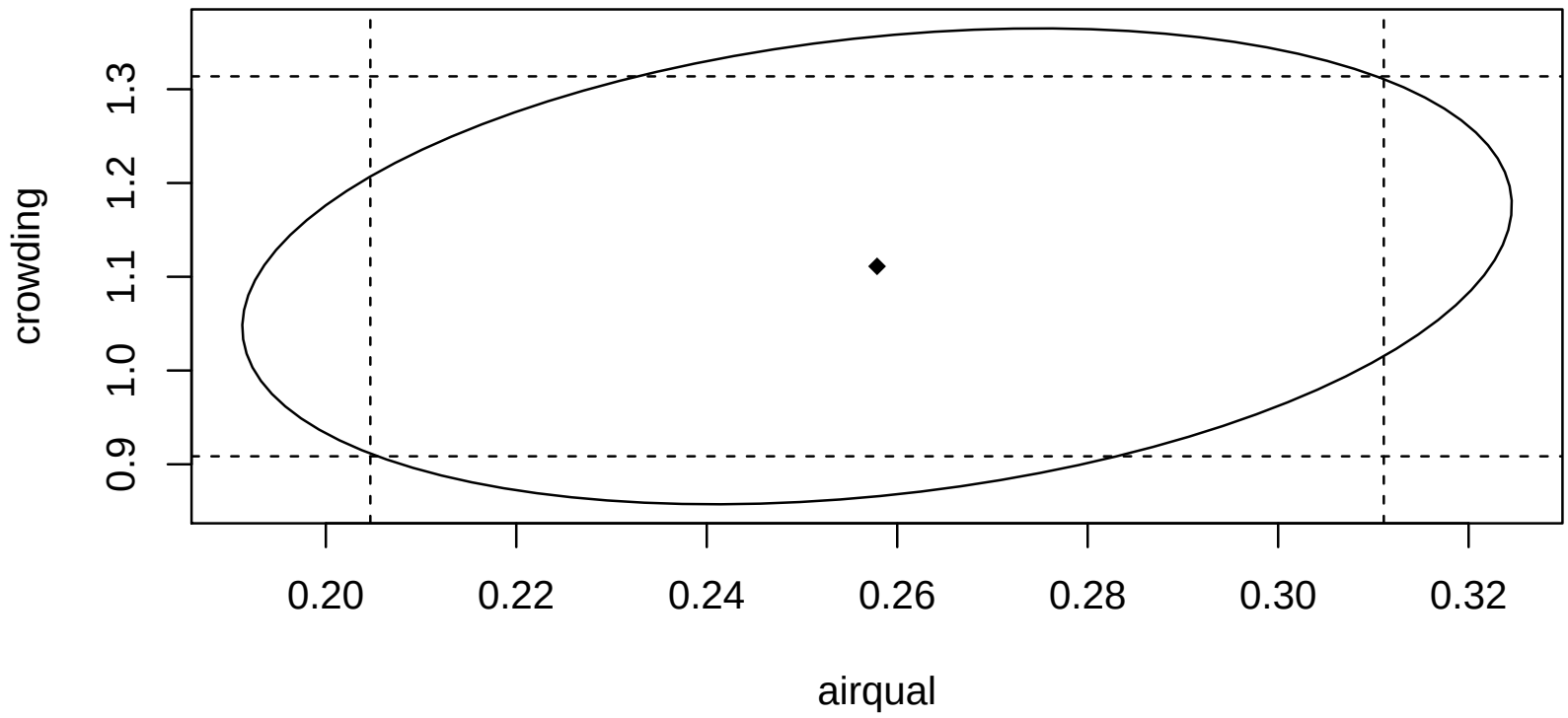
```
##                                2.5 %      97.5 %
## (Intercept) 11.86333314  6.74302724 16.98363903
## airqual      0.25788257  0.20467182  0.31109332
## crowding     1.11112603  0.90852947  1.31372260
## nutrition   -0.03278397 -0.04857606 -0.01699189
## smoking      4.96093131  2.80605790  7.11580472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

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

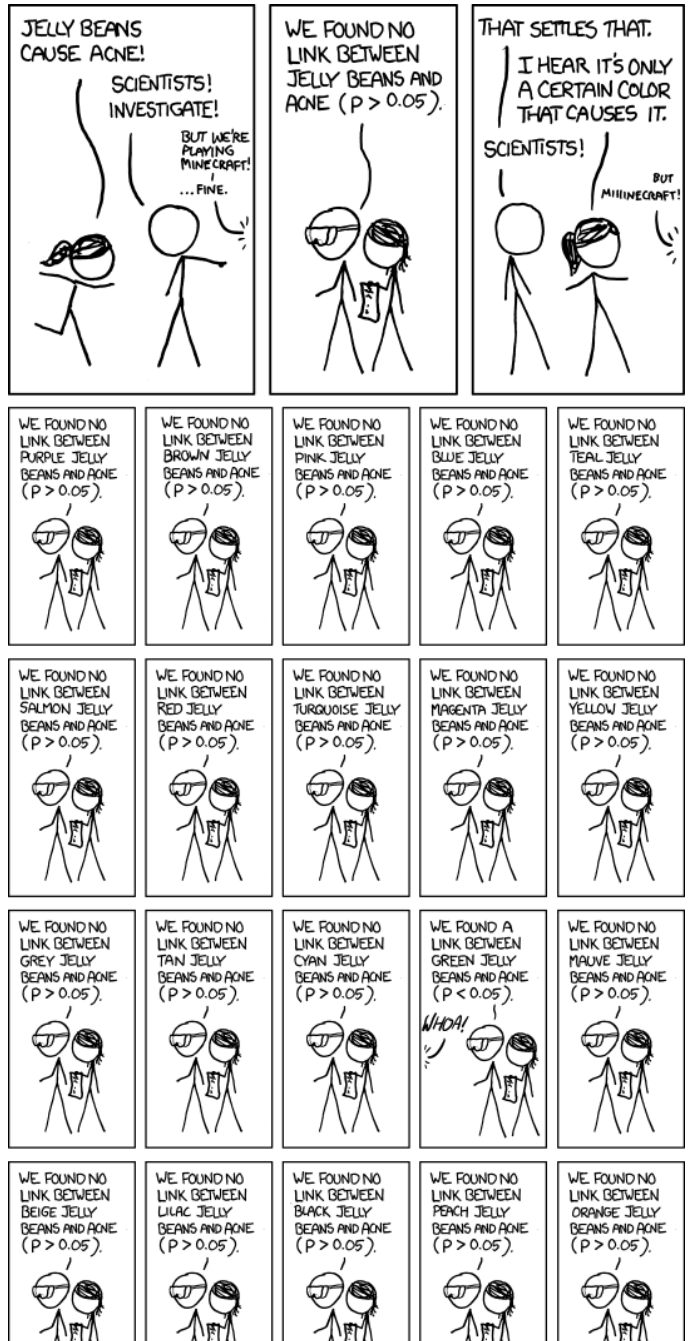
```
library(ellipse)
plot(ellipse(mlr1,c(2,3)),type="l")
points(coef(mlr1)[2],coef(mlr1)[3], pch=18)
abline(v=c(confint(mlr1)[2,1], confint(mlr1)[2,2]), lty=2)
abline(h=c(confint(mlr1)[3,1], confint(mlr1)[3,2]), lty=2)
```



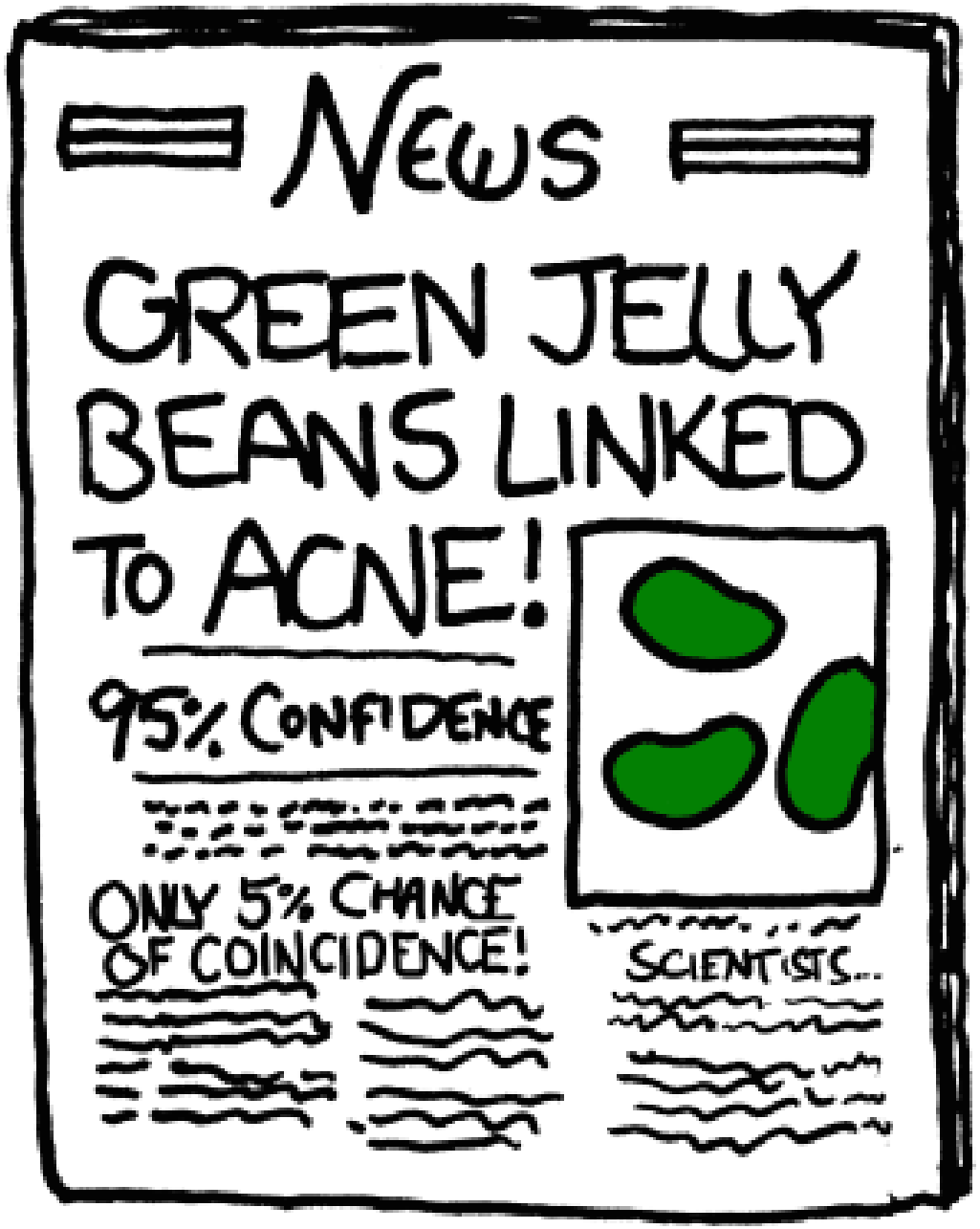
Progress report

- Sampling distribution of $\hat{\boldsymbol{\beta}}$
- Confidence intervals
- Hypothesis tests for individual coefficients
- Multiple testing

Multiple testing - preserving your Type I error rate



Multiple testing - preserving your Type I error rate



Inference about multiple coefficients

Our model contains multiple parameters; often we want ask a question about multiple coefficients simultaneously. I.e. "are any of these k coefficients significantly different from 0?" This is equivalent to performing multiple tests (or looking at confidence ellipses):

$$\begin{aligned} H_{01} : \beta_1 &= 0 \\ H_{02} : \beta_2 &= 0 \\ \vdots &= \vdots \\ H_{0k} : \beta_k &= 0 \end{aligned}$$

where each test has a size of α

- For any individual test, $P(\text{reject } H_{0i} | H_{0i}) = \alpha$

Inference about multiple coefficients

For any individual test, $P(\text{reject } H_{0i} | H_{0i}) = \alpha$.

But it DOES NOT FOLLOW that

$$P(\text{reject } \underline{\text{at least one}} \ H_{0i} | \text{all } H_{0i} \text{are true}) = \alpha.$$

This is called the Family-wise error rate (FWER). Ignore it at your own peril!

Family-wise error rate

To calculate the FWER

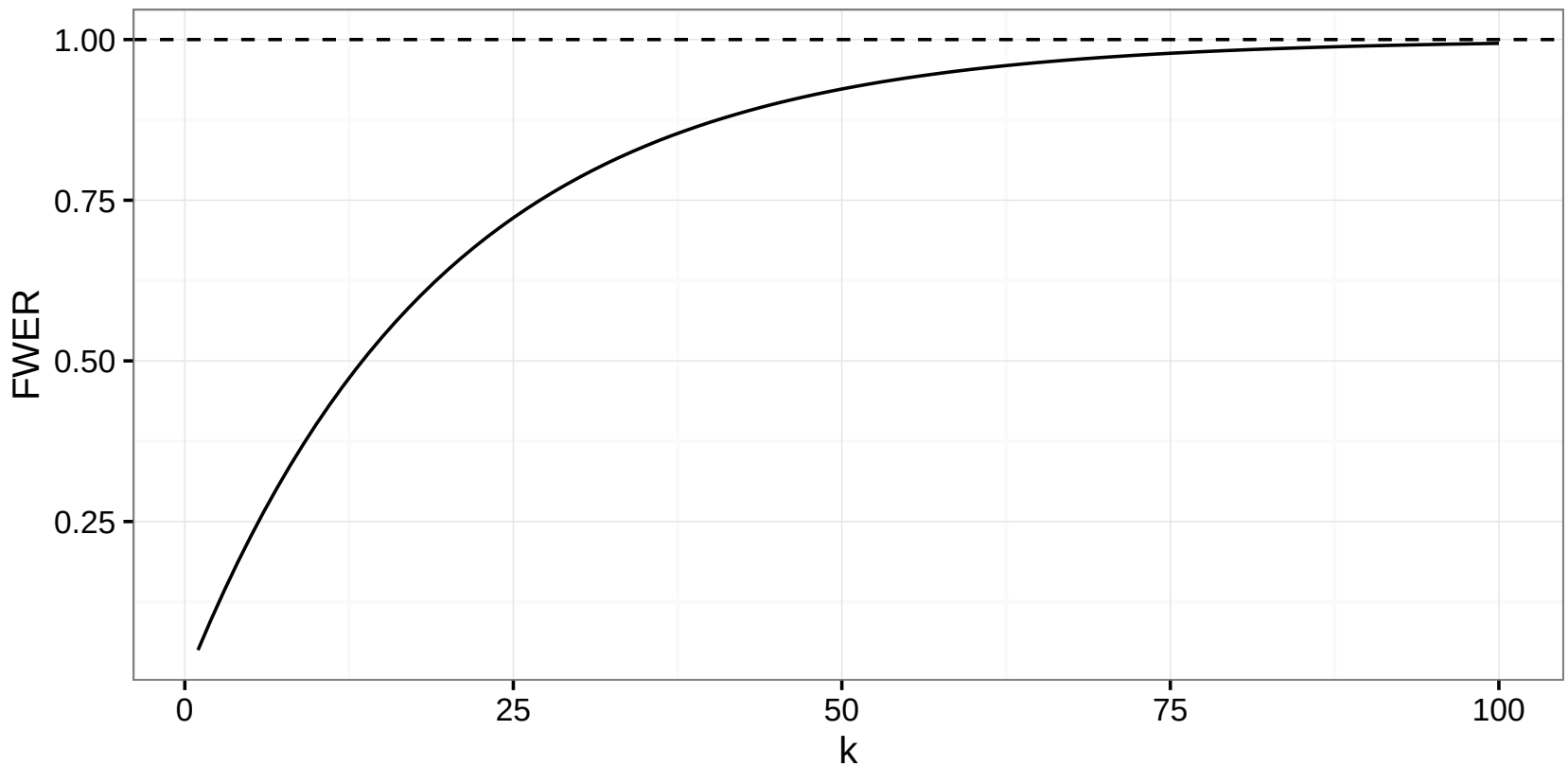
- First note $P(\text{no rejections}|\text{all } H_{0i} \text{are true}) = (1-\alpha)^k$
- It follows that

$$\begin{aligned} \text{FWER} &= P(\text{at least one rejection}|\text{all } H_{0i} \text{are true}) \\ &= 1-(1-\alpha)^k \end{aligned}$$

Family-wise error rate

$$\text{FWER} = 1 - (1 - \alpha)^k$$

```
alpha <- .05
k <- 1:100
FWER <- 1-(1-alpha)^k
qplot(k, FWER, geom="line") + geom_hline(yintercept = 1, lty=2)
```



Addressing multiple comparisons

Three general approaches

- Do nothing in a reasonable way
  - Don't trust scientifically implausible results
  - Don't over-emphasize isolated findings
- Correct for multiple comparisons
  - Often, use the Bonferroni correction and use $\alpha_i = \alpha/k$ for each test
  - Thanks to the Bonferroni inequality, this gives an overall $FWER \leq \alpha$
- Use a global test

Global tests

Compare a smaller "null" model to a larger "alternative" model

- Smaller model must be nested in the larger model
- That is, the smaller model must be a special case of the larger model
- For both models, the RSS gives a general idea about how well the model is fitting
- In particular, something like

$$\frac{RSS_S - RSS_L}{RSS_L}$$

compares the relative RSS of the models

Nested models

- These models are nested:

$$\begin{aligned} \text{Smaller} &= \text{Regression of } Y \text{ on } X_1 \\ \text{Larger} &= \text{Regression of } Y \text{ on } X_1, X_2, X_3, X_4 \end{aligned}$$

- These models are not:

$$\begin{aligned} \text{Smaller} &= \text{Regression of } Y \text{ on } X_2 \\ \text{Larger} &= \text{Regression of } Y \text{ on } X_1, X_3 \end{aligned}$$

Global F tests

- Compute the test statistic

$$F_{obs} = \frac{(RSS_S - RSS_L)/(df_S - df_L)}{RSS_L/df_L}$$

- If H_0 (the null model) is true, then $F_{obs} \sim F_{df_S - df_L, df_L}$
- Note $df_s = n - p_S - 1$ and $df_L = n - p_L - 1$
- We reject the null hypothesis if the p-value is above α, where

$$\text{p-value} = P(F_{df_S - df_L, df_L} > F_{obs})$$

Global F tests

There are a couple of important special cases for the F test

- The null model contains the intercept only
  - When people say ANOVA, this is often what they mean (although all F tests are based on an analysis of variance)
- The null model and the alternative model differ only by one term
  - Gives a way of testing for a single coefficient
  - Turns out to be equivalent to a two-sided t-test: $t^2_{df_L} \sim F_{1,df_L}$

Lung data: multiple coefficients simultaneously

You can test multiple coefficients simultaneously using the F test

```
mlr_null <- lm(disease ~ nutrition, data=dat)
mlr1 <- lm(disease ~ nutrition+ airqual + crowding + smoking, data=dat)
anova(mlr_null, mlr1)

## Analysis of Variance Table
##
## Model 1: disease ~ nutrition
## Model 2: disease ~ nutrition + airqual + crowding + smoking
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     97 9192.7
## 2     94 1248.0  3    7944.7 199.47 < 2.2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This test shows that airqual, crowding, and smoking together significantly improve the fit of our model (assuming model diagnostics look good). Further analyses may be warranted to determine which, if any, coefficients are not different from 0.

Lung data: single coefficient test

The F test is equivalent to the t test when there's only one parameter of interest

```
mlr_null <- lm(disease ~ nutrition, data=dat)
mlr1 <- lm(disease ~ nutrition + airqual, data=dat)
anova(mlr_null, mlr1)

## Analysis of Variance Table
##
## Model 1: disease ~ nutrition
## Model 2: disease ~ nutrition + airqual
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     97 9192.7
## 2     96 5969.5  1    3223.1 51.833 1.347e-10 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(mlr1)$coef

##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 37.62538251 2.43946243 15.423637 9.946294e-28
## nutrition   -0.03469855 0.01692446 -2.050202 4.307101e-02
## airqual      0.36114435 0.05016218  7.199535 1.346935e-10
```

Today's Big Ideas

Basic parameter inference for multiple linear regression models

- How to determine "significance" of your covariates
- F tests can control for multiple comparisons!

Multiple testing activity!