

## 9. Additional topics in linear modeling

- Fitting polynomial relationships and other nonlinear trend models using linear models.
- The  $R^2$  statistic to assess model fit.
- Multicollinearity: What happens when two or more explanatory variables are highly correlated. How to notice it, and what to do about it.
- More on the linear model formula notation in R: Interactions between explanatory variables.
- Model selection from a large number of possible models.
- More on causation, observational studies and designed experiments.

# Using linear models to fit polynomial relationships

- Recall the basic linear trend model from Chapter 1 for data  $y_1, \dots, y_n$  with  $y_i$  measured at time  $t_i$ ,

$$[\text{M1}] \quad y_i = b_0 + b_1 t_i + e_i, \quad i = 1, \dots, n$$

- What if the data have a trend that is not linear?
- The next thing we might consider is a quadratic trend model,

$$[\text{M2}] \quad y_i = b_0 + b_1 t_i + b_2 t_i^2 + e_i, \quad i = 1, \dots, n$$

- M1 and M2 are both linear models, with respective design matrices

$$\mathbb{X}^{[1]} = \begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_n \end{bmatrix} \quad \mathbb{X}^{[2]} = \begin{bmatrix} 1 & t_1 & t_1^2 \\ 1 & t_2 & t_2^2 \\ \vdots & \vdots & \vdots \\ 1 & t_n & t_n^2 \end{bmatrix}$$

# The order $p$ polynomial smoothing model

- When the explanatory variable for  $y_i$  is the time of measurement,  $t_i$ , then we call the linear model a trend.
- When we fit  $y_i$  using a function of an arbitrary explanatory variable  $x_i$  we say we are **smoothing**.
- We can choose any  $p$  in the general order  $p$  polynomial smoothing model,

$$[\text{M3}] \quad y_i = b_0 + b_1x_i + b_2x_i^2 + b_3x_i^3 + \cdots + b_px_i^p + e_i, \quad i = 1, \dots, n$$

- This is a linear model with design matrix

$$\mathbb{X}^{[3]} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^p \\ 1 & x_2 & x_2^2 & \cdots & x_2^p \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^p \end{bmatrix}$$

**Question 9.1.** How would you decide what order  $p$  to use when applying the polynomial smoothing model,

$$y_i = b_0 + b_1x_i + b_2x_i^2 + b_3x_i^3 + \cdots + b_px_i^p + e_i, \quad i = 1, \dots, n$$

Ideas: (i) look for which coefficients are statistically significant in the probability model

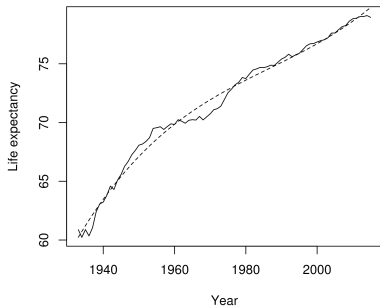
$$Y_i = \beta_0 + \beta_1x_i + \beta_2x_i^2 + \beta_3x_i^3 + \cdots + \beta_px_i^p + \epsilon_i,$$

with  $\epsilon_i \sim \text{iid normal}(0, \sigma)$  and  $i = 1, \dots, n$ . (ii) make F tests to compare models with two different values of  $p$  under this probability model. (iii) look at the fitted line and see if you like it.

# Cubic polynomial smoothing of life expectancy

```
L_poly3 <- lm(Total~Year+I(Year^2)+I(Year^3),data=L)
```

```
plot(L$Year,L$Total,  
     type="line",  
     xlab="Year",  
     ylab="Life expectancy")  
  
lines(L$Year,fitted(L_poly3),  
      lty="dashed")
```



**Question 9.2.** Why do we need to write  $I(\text{Year}^2)$  not just  $\text{Year}^2$  to fit a polynomial smoothing model in the R formula notation?

This is a technical consideration.  $\text{Year}^2$  happens to denote an interaction term in R's model formula language, so we have to tell R we just want the squared variable.

# Checking the cubic smoothing calculation

**Question 9.3.** How would you check that the R model formula we wrote is correct for the cubic polynomial we intend to fit?

Look at the design matrix!

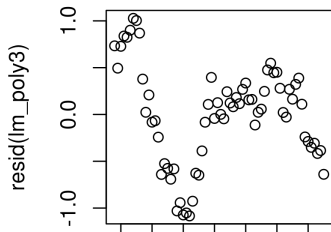
**Question 9.4.** If we have done a good job of modeling the trend, we might hope that the residuals look like independent measurement errors. How would you check if this is the case?

make a timeplot and/or a plot of the residuals against the lagged residuals.

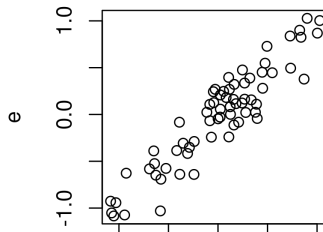
# Repeating diagnostic tests for life expectancy vs unemployment using cubic detrending

```
L_detrended <- L_poly3$residuals
U_annual <- apply(U[,2:13],1,mean)
U_detrended <- lm(U_annual~Year+I(Year^2)+I(Year^3),
  data=U)$residuals
L_detrended <- subset(L_detrended,L$Year %in% U$Year)
lm_poly3 <- lm(L_detrended~U_detrended)
n <- length(resid(lm_poly3))
e <- resid(lm_poly3)[2:n] ; lag_e <- resid(lm_poly3)[1:(n-1)]
```

```
plot(U$Year,resid(lm_poly3))
```



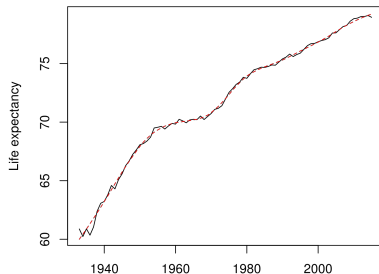
```
plot(lag_e,e)
```



# Local linear smoothing of life expectancy

```
L_loess <- loess(Total~Year,data=L,span=0.3)
```

```
plot(L$Year,L$Total,  
     type="line",  
     xlab="Year",  
     ylab="Life expectancy")  
  
lines(L$Year,fitted(L_loess),  
      lty="dashed",col="red")
```



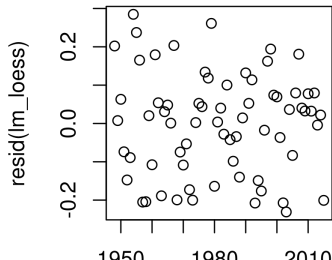
- `loess()` is a **smoother** that fits a local linear model. This means that, at each point  $x_j$ , the smoother predicts  $y_i$  fitting a linear model that ignores all the data except for points close to  $x_i$ .
- Setting `span=0.3` means that the closest 30% of the points are used.



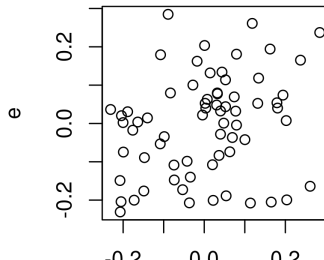
# Repeating diagnostic tests for life expectancy vs unemployment using a smoother

```
L_detrended <- resid(L_loess)
U_annual <- apply(U[,2:13],1,mean)
U_detrended <- resid(loess(U_annual~Year,data=U,span=0.3))
L_detrended <- subset(L_detrended,L$Year %in% U$Year)
lm_loess <- lm(L_detrended~U_detrended)
n <- length(resid(lm_loess))
e <- resid(lm_loess)[2:n] ; lag_e <- resid(lm_loess)[1:(n-1)]
```

```
plot(U$Year,resid(lm_loess))
```



```
plot(lag_e,e)
```



# Revisiting the evidence for pro-cyclical mortality

```
coef(summary(lm_loess))
```

```
##              Estimate Std. Error   t value    Pr(>|t|)
## (Intercept) 0.007138079 0.01613621 0.4423641 0.6596720450
## U_detrended 0.067235405 0.01628394 4.1289394 0.0001045733
```

- Recall that linear detrending gave a significant association between life expectancy and unemployment.
- This suggested that mortality is **pro-cyclical**, meaning it increases when the business cycles is in economic expansion and unemployment is low.
- In Chapter 7, we found the residuals in this regression had a strong pattern, casting doubt on the validity of our linear model and its unintuitive conclusion.

**Question 9.5.** Re-assess the evidence based on this new analysis.

# The R-squared statistics to assess goodness of fit

- $R^2$  is the square of the correlation between the data and the fitted values.
- It can also be computed as

$$R^2 = 1 - \frac{\text{RSS}}{\text{SST}} = \frac{\text{SST} - \text{RSS}}{\text{SST}}$$

where  $\text{RSS}$  is the residual sum of squares and  $\text{SST}$  is the total sum of squares, defined as

$$\text{SST} = \sum_{i=1}^n (y_i - \bar{y})^2, \quad \text{where } \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i.$$

- $R^2$  is sometimes described as the fraction of the variation in the data explained by the linear model.
- $1 - R^2$  is the fraction of the variation in the data left unexplained by the model.

# Uses and abuses of R-squared

- A low  $R^2$  sends a clear signal: the fitted model doesn't describe the data much better than the sample mean.
- Sometimes a small, but statistically significant, correlation is of interest. If you are monitoring data on the operation of an aircraft jet engine, you want to know about evidence suggesting a malfunction as soon as it is statistically significant. **Interpretation of R-squared depends on context.**
- The  $R^2$  statistic compares the residual sum of squares under the full model with the residual sum of squares under a model with a constant mean. By contrast, the F test compares the full model with a model that omits specific selected explanatory variables. The F test is more appropriate for assessing whether a variable, or group of variables, should be included in the model.

# A relationship between the F statistic and R-squared

- Recall that in a regression setting, the F statistic is expressed in the following way.

$$f = \frac{(\text{RSS}_0 - \text{RSS}_a)/d}{\text{RSS}_a/(n - q)}.$$

- $q$  is the dimension of the alternative hypothesis.
- $d$  is the difference in dimension between the null and alternative hypotheses.

**Question 9.6.** Write the hypotheses  $H_0$  and  $H_a$  to match the  $R^2$  statistic in a linear model with  $p$  explanatory variables (including the intercept). Set up these hypotheses so that  $\text{RSS}_0$  is  $\text{SST}$ , and  $\text{RSS}_a$  is  $\text{RSS}$ .

todo

- In this context, we see that  $q = p$  and  $d = p - 1$ .

# Writing $R^2$ in terms of an F statistic

- From last slide, we have

$$f = \frac{(\text{SST} - \text{RSS})/(p - 1)}{\text{RSS}/(n - p)}$$

- Recall that  $R^2 = 1 - \text{RSS}/\text{SST}$ .

**Question 9.7.** Check by algebra that  $R^2 = 1 - \frac{1}{1 + f \times (p - 1)/(n - p)}$

todo

**Question 9.8.** What is  $R^2$  when  $F$  is very large? or close to zero?

todo

**Question 9.9.** Explain why  $R^2$  cannot decrease when you add an extra explanatory variable into a linear model. (Explanations for questions like this should involve some math notation, not just words.)

todo

- Simplicity in a model is a good thing. The fact that any added model complexity makes  $R^2$  seem “better” requires caution in interpretation.

# Adjusted R-squared

- One approach to penalize  $R^2$  for a more complex model is to divide each sum of squares by its degrees of freedom. This gives the **adjusted R-squared**,

$$R_{\text{adj}}^2 = 1 - \frac{\text{RSS}/(n - p)}{\text{SST}/(n - 1)}.$$

- Dividing by the degrees of freedom in  $R_{\text{adj}}^2$  is like what we do in the F statistic.
- The F statistic takes advantage of the nice mathematical property that  $\text{SST} - \text{RSS}$  and  $\text{RSS}$  are independent random variables for the probability model with normally distributed measurement error.
- For comparing two **nested** models (when the larger model consists of adding variables to the smaller model) an F test is a clearer statistical argument than comparing  $R_{\text{adj}}^2$ .
- When the models are not nested, the F test is not applicable. Comparing  $R_{\text{adj}}^2$  values gives one way to assess the models, though not a formal test.
- Now we've studied  $R_{\text{adj}}^2$ , we understand everything in `summary(lm())`.



```
##
## Call:
## lm(formula = GPA ~ ACT + High_School, data = gpa)
##
## Residuals:
```

	Min	1Q	Median	3Q	Max
	-2.10265	-0.29862	0.07311	0.40355	1.31336

```
##
## Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.292793	0.136725	9.455	< 2e-16 ***
ACT	0.037210	0.005939	6.266	6.48e-10 ***
High_School	0.010022	0.001279	7.835	1.74e-14 ***

```
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5672 on 702 degrees of freedom
## Multiple R-squared: 0.2033, Adjusted R-squared: 0.2011
## F-statistic: 89.59 on 2 and 702 DF, p-value: < 2.2e-16
```

# Collinear explanatory variables in a linear model

- Let  $\mathbb{X} = [x_{ij}]_{n \times p}$  be an  $n \times p$  design matrix.
- If there is a nonzero vector  $\alpha = (\alpha_1, \dots, \alpha_p)$  such that  $\mathbb{X}\alpha = \mathbf{0}$  then the columns of  $\mathbb{X}$  are **collinear**.
- Here,  $\mathbf{0}$  is the zero vector,  $(0, 0, \dots, 0)$ .
- We can write  $\mathbf{x}_j = (x_{1j}, x_{2j}, \dots, x_{nj})$  for the  $j$ th column of  $\mathbb{X}$ . Then,

$$\mathbb{X}\alpha = \alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + \dots + \alpha_p\mathbf{x}_p.$$

We see that  $\mathbb{X}\alpha$  can be thought of as a **linear combination of the columns of  $\mathbb{X}$** .

- Collinearity of explanatory variables has important consequences for fitting a linear model to data.
- It can also be useful to notice whether the variables are close to collinear, meaning that  $\mathbb{X}\alpha$  is small but nonzero.

## Example: an intercept with a coefficient for each factor

- Recall the mouse weight dataset. Consider a sample linear model,

$$y_{ij} = \mu + \mu_j + e_{ij}.$$

- Suppose that we don't set the  $\mu_1 = 0$  so we try to estimate both  $\mu_1$  and  $\mu_2$  at the same time as the intercept,  $\mu$ .
- Let's work with just 3 mice in each treatment group, so  $i = 1, 2, 3$  and  $j = 1, 2$ . The design matrix is therefore

```
X <- cbind(rep(1,6),rep(c(1,0),each=3),rep(c(0,1),each=3)) ; X
##      [,1] [,2] [,3]
## [1,]    1    1    0
## [2,]    1    1    0
## [3,]    1    1    0
## [4,]    1    0    1
## [5,]    1    0    1
## [6,]    1    0    1
```

- For  $\alpha = (1, -1, -1)$ , we have  $\mathbb{X}\alpha = 0$

# The least squares fit with collinear predictors

- Suppose that  $\mathbf{b}$  is a least squares coefficient vector, so that the fitted value vector  $\hat{\mathbf{y}} = \mathbb{X}\mathbf{b}$  minimizes  $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ .
- Suppose that  $\mathbb{X}$  is collinear, with  $\mathbb{X}\boldsymbol{\alpha} = \mathbf{0}$ .
- Since

$$\mathbb{X}(\mathbf{b} + \boldsymbol{\alpha}) = \mathbb{X}\mathbf{b} + \mathbb{X}\boldsymbol{\alpha} = \mathbb{X}\mathbf{b} + \mathbf{0} = \mathbb{X}\mathbf{b},$$

we see that  $\mathbf{b} + \boldsymbol{\alpha}$  is also a least squares coefficient vector.

- **When  $\mathbb{X}$  is collinear, a least squares coefficient still exists, but it is not unique.**

**Question 9.10.** Let  $c$  be any number. Recall multiplication of a vector by a scalar:  $c\boldsymbol{\alpha} = (c\alpha_1, \dots, c\alpha_p)$ . Show that  $\mathbf{b} + c\boldsymbol{\alpha}$  is also a least squares fit.

## Standard errors for collinear variables

**Question 9.11.** Any variable that is part of a collinear combination of variables has infinite standard error. Why?

todo

# What does R do if give it collinear variables?

```
mice <- read.table("femaleMiceWeights.csv",header=T,sep=",")
chow=rep(c(1,0),each=12)
hf=rep(c(0,1),each=12)
lm1 <- lm(Bodyweight~chow+hf,data=mice)
coef(summary(lm1))
```

	Estimate	Std. Error	t value	Pr(> t )
## (Intercept)	26.834167	1.039353	25.818139	6.045435e-18
## chow	-3.020833	1.469867	-2.055174	5.192480e-02

- R noticed that the three explanatory variables are collinear, and refused to fit the third

```
model.matrix(lm1)
```

##	(Intercept)	chow	hf
## 1	1	1	0
## 2	1	1	0
## 3	1	1	0
## 4	1	1	0
## 5	1	1	0
## 6	1	1	0
## 7	1	1	0
## 8	1	1	0
## 9	1	1	0
## 10	1	1	0
## 11	1	1	0
## 12	1	1	0
## 13	1	0	1
## 14	1	0	1
## 15	1	0	1
## 16	1	0	1
## 17	1	0	1
## 18	1	0	1
## 19	1	0	1





# Linearly independent vectors and matrix rank

- Columns of a matrix that are not collinear are said to be **linearly independent**.
- The **rank** of  $\mathbf{X}$  is the number of linearly independent columns.
- $\mathbf{X}$  has **full rank** if all the columns are linearly independent. In this case, we expect the least squares coefficient to be uniquely defined and so  $\mathbf{X}^T \mathbf{X}$  has non-zero determinant and is invertible.
- If  $\mathbf{X}$  does not have full rank, we can drop **linearly dependent** columns until the remaining columns are linearly independent. This is a practical approach to handling collinearity.

## Example: reducing a design matrix to full rank

```
X <- model.matrix(lm1)
```

```
det(t(X)%*%X)
```

```
## [1] 0
```

```
X2 <- X[,1:2]
```

```
det(t(X2)%*%X2)
```

```
## [1] 144
```

- Dropping the third column of  $X$  has given us a full-rank design matrix.

**Question 9.12.** The least squares fitted values are the same using the predictor matrix  $X_2$  as  $X$ . Why does dropping the last column not change the fitted values?

todo

# Almost collinear variables

- If the determinant of  $\mathbf{X}^T \mathbf{X}$  is close to zero, the variance of the model-generated least squares coefficient vector becomes large.
- This can happen when multiple explanatory variables are included in a model which all model similar things.

**Question 9.13.** Recall our data analysis using unemployment to explain life expectancy. What would happen if we added total employment as an additional explanatory variable? (Being unemployed is not the only alternative to being employed, since only adults currently looking for work are counted as unemployed.)

todo

## More on the R model formula notation

- A **model formula** in `lm()` is something that looks like  $y \sim x$ .
- The R formula notation has various conventions that are designed to make it easy to specify useful models.
- `?formula` tells you everything you need to know, and more.
- You can think of the R formula for `lm()` is a way of constructing a design matrix.
- Inspect the resulting design matrix using `model.matrix()` and check you understand what R has produced. If you can do this, you can safely use the power of the formula notation.

**Question 9.14.** In a report, the model should be written in mathematical notation, not as an R formula. Why?

todo

# Experimenting with the R formula notation

- Consider the freshman GPA data

```
gpa <- read.table("gpa.txt",header=T); head(gpa,3)
##    ID  GPA High_School ACT Year
## 1   1 0.98          61  20 1996
## 2   2 1.13          84  20 1996
## 3   3 1.25          74  19 1996
```

- We can play the game of trying out various things in R formula notation, inspecting the resulting design matrix, and figuring out how to write the model efficiently in mathematical notation.
- You can also think about whether the different models give any new insights into the data.

```
lm1 <- lm(GPA~ACT+High_School*Year,data=gpa)
coef(summary(lm1))[,1:2]
```

##	Estimate	Std. Error
## (Intercept)	-4.722613e+01	1.350854e+02
## ACT	3.708961e-02	5.946966e-03
## High_School	3.460100e-01	1.702035e+00
## Year	2.428369e-02	6.760800e-02
## High_School:Year	-1.681424e-04	8.518297e-04

- The \* here denotes inclusion of an **interaction** between High\_School and Year, written in the R output as High\_School:Year.

**Question 9.15.** Conceptually, what do you think an interaction between two variables is, and why might it be needed?

todo

- To find out exactly what R thinks an interaction is, we can check the design matrix.

```
head(model.matrix(lm1))
```

```
##      (Intercept) ACT High_School Year High_School:Year
## 1             1  20             61 1996             121756
## 2             1  20             84 1996             167664
## 3             1  19             74 1996             147704
## 4             1  23             95 1996             189620
## 5             1  28             77 1996             153692
## 6             1  23             47 1996              93812
```

**Question 9.16.** Write out the sample model that R has computed in `lm1` using subscript notation.

todo

# Interactions and additivity

```
lm2 <- lm(GPA~ACT+High_School+Year+High_School:Year,data=gpa)
head(model.matrix(lm2),4)
```

##	(Intercept)	ACT	High_School	Year	High_School:Year
## 1	1	20	61	1996	121756
## 2	1	20	84	1996	167664
## 3	1	19	74	1996	147704
## 4	1	23	95	1996	189620

- `lm2` has the same design matrix as `lm1`.
- We see that, in R formula notation,  $y \sim u * v$  is the same as  $y \sim u + v + u : v$ .
- In the model  $y \sim u + v$  the effects of the variables are said to be **additive**.
- In a causal interpretation of an additive model, the result of changing  $u$  to  $u_2$  and  $v$  to  $v_2$  is the sum of the marginal effect of changing  $u$  to  $u_2$  plus the marginal effect of changing  $v$  to  $v_2$ .
- The interaction term  $u : v$  breaks additivity. In this case, we can't know the consequence for the fitted value of changing  $u$  to  $u_2$  unless we know



# The interaction between ACT and high school percentile

- We have not (yet) found any interesting effect of year. Let's drop year out of the model and look for whether there is an interaction between ACT and high school percentile for predicting freshman GPA.

```
lm3 <- lm(GPA~ACT*High_School, data=gpa)
```

**Question 9.17.** Write out the fitted sample linear model in subscript form, letting  $y_i$ ,  $a_i$ ,  $h_i$  and  $e_i$  be the freshman GPA, ACT score, high school percentile and residual error respectively for the  $i$ th student.

todo

# Interpreting a discovered interaction

```
coef(summary(lm3))[,1:2]
```

##	Estimate	Std. Error
## (Intercept)	3.157679842	0.4788067771
## ACT	-0.046067744	0.0213355076
## High_School	-0.014405030	0.0061479608
## ACT:High_School	0.001071326	0.0002638611

**Question 9.18.** Explain in words to the admissions director what you have found about the interaction under investigation here.

todo

# Marginal effects when there is an interaction

- Notice in 'lm3' that the coefficients for ACT score and high school percentile are negative. That is surprising!

```
ACT_centered <- gpa$ACT - mean(gpa$ACT)
HS_centered <- gpa$Hi - mean(gpa$Hi)
lm3b <- lm(GPA ~ ACT_centered * HS_centered, data = gpa)
signif(coef(summary(lm3b))[, c(1, 2, 4)], 3)
```

##	Estimate	Std. Error	Pr(> t )
## (Intercept)	2.94000	0.022900	0.00e+00
## ACT_centered	0.03640	0.005880	1.04e-09
## HS_centered	0.01190	0.001350	8.23e-18
## ACT_centered:HS_centered	0.00107	0.000264	5.46e-05

**Question 9.19.** After centering the variables, the interaction effect stays the same, but the marginal effects change sign. What is happening? Why?

todo

# Quantifying the improvement in the model

```
s3 <- summary(lm3)$sigma
lm4 <- lm(GPA~ACT+High_School,data=gpa)
s4 <- summary(lm4)$sigma
lm5 <- lm(GPA~1,data=gpa)
s5 <- summary(lm5)$sigma
cat("s3 =",s3,"; s4 =",s4,"; s5 =",s5)

## s3 = 0.5610067 ; s4 = 0.5671605 ; s5 = 0.6345278
```

**Question 9.20.** Comment on both **statistical significance** and **practical significance** of the interaction between a prediction of freshman GPA.

todo

# An interaction involving a factor

- Let's go back to the football field goal data.

```
goals <- read.table("FieldGoals2003to2006.csv",header=T,sep=",")
goals[1,c("Name","Teamt","FGt","FGtM1")]
```

```
##           Name Teamt  FGt FGtM1
## 1 Adam Vinatieri    NE 73.5    90
```

```
lm6 <- lm(FGt~FGtM1*Name,data=goals)
```

**Question 9.21.** What model do you think is being fitted here? Write it in subscript form, where  $y_{ij}$  is the field goal average for the  $j$ th year of kicker  $i$ , with  $i = 1, \dots, 19$  and  $j = 1, 2, 3, 4$ . Let  $e_{ij}$  be the residual error, and let  $x_{ij}$  be the previous year's average. Check your answer against the design matrix shown on the next slide.

todo

```
X<-model.matrix(lm6) ; colnames(X)<-1:38 ; X[1:17,c(1:8,21:26)]
```

##		1	2	3	4	5	6	7	8	21	22	23	24	25	26
## 1	1	90.0	0	0	0	0	0	0	0	0.0	0.0	0.0	0	0	0
## 2	1	73.5	0	0	0	0	0	0	0	0.0	0.0	0.0	0	0	0
## 3	1	93.9	0	0	0	0	0	0	0	0.0	0.0	0.0	0	0	0
## 4	1	80.0	0	0	0	0	0	0	0	0.0	0.0	0.0	0	0	0
## 5	1	88.2	1	0	0	0	0	0	0	88.2	0.0	0.0	0	0	0
## 6	1	82.7	1	0	0	0	0	0	0	82.7	0.0	0.0	0	0	0
## 7	1	84.3	1	0	0	0	0	0	0	84.3	0.0	0.0	0	0	0
## 8	1	72.7	1	0	0	0	0	0	0	72.7	0.0	0.0	0	0	0
## 9	1	72.2	0	1	0	0	0	0	0	0.0	72.2	0.0	0	0	0
## 10	1	87.0	0	1	0	0	0	0	0	0.0	87.0	0.0	0	0	0
## 11	1	85.2	0	1	0	0	0	0	0	0.0	85.2	0.0	0	0	0
## 12	1	75.0	0	1	0	0	0	0	0	0.0	75.0	0.0	0	0	0
## 13	1	82.1	0	0	1	0	0	0	0	0.0	0.0	82.1	0	0	0
## 14	1	95.6	0	0	1	0	0	0	0	0.0	0.0	95.6	0	0	0
## 15	1	85.7	0	0	1	0	0	0	0	0.0	0.0	85.7	0	0	0
## 16	1	79.1	0	0	1	0	0	0	0	0.0	0.0	79.1	0	0	0
## 17	1	80.0	0	0	0	1	0	0	0	0.0	0.0	0.0	80	0	0

**Question 9.22.** Interpret the ANOVA table below.

todo

```
anova(lm6)

## Analysis of Variance Table
##
## Response: FGt
##              Df  Sum Sq Mean Sq F value    Pr(>F)
## FGtM1          1    87.20   87.199    1.9008 0.176047
## Name          18 2252.47  125.137    2.7279 0.004565 **
## FGtM1:Name     18  417.75   23.209    0.5059 0.938592
## Residuals     38 1743.20   45.874
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# The causal interpretation of observational studies

- Consider a simple least-squares linear model  $y_i = ax_i + b + e_i$  for  $i = 1, \dots, n$ . The usual corresponding probability model is  $Y_i = \alpha x_i + \beta + \epsilon_i$  with  $\epsilon_1, \dots, \epsilon_n$  being independent  $N[0, \sigma]$  random variables.
- The coefficient  $\alpha$  for  $x_i$ ,  $i = 1, \dots, n$  is commonly called the **effect** of  $x_i$  on  $y_i$ .
- Sometimes  $a$  is called the effect, but it is more properly an **estimated effect**.
- The **causal interpretation** of the linear model is that, if we manipulated  $x_i$  to increase it by one unit for individual  $i$ , keeping everything else fixed, we would expect  $y_i$  to increase by  $a$  units.
- The use of the word “effect” has a causal meaning in common usage.
- We should think carefully about when this meaning is justified.



# Does coffee cause heart attacks?

- Coffee has relatively high levels of caffeine, a commonly consumed drug. Many studies have been done to see if it has adverse (or positive) health effects.
- A typical observational study will model a health outcome (say, a measure of heart health) and investigate linear models based on available explanatory variables.
- If higher levels of coffee consumption are associated with lower heart health scores, beyond what can be explained by chance variation in our sample, we will be suspicious about drinking coffee.

**Question 9.23.** Suggest important confounding variable(s) in the causal interpretation of this model. What would you do to help make a convincing argument for or against coffee?

todo

# Which surgeon do you choose?

- Cost effectiveness of medical treatment is a major current issue. You are advising a health insurance program, and your boss gives you data on success rates for a certain heart surgery, together with the salary of the surgeon performing the operation.

**Question 9.24.** Suppose you find the estimated effect is negative and statistically significant: higher salaries are associated with lower success rates. How would you interpret this result? What are possible confounding factors?

todo

# When can we infer causation from observational data?

The following considerations may add weight to the causal interpretation of an association

- There is a plausible mechanism.
- There are no un-measured variables considered plausible mechanisms.
- The effect is consistent across population subgroups.
- For data collected though time, the proposed cause precedes the consequence.
- Consistency with available experimental evidence.
- A consistent gradient between increases in the proposed cause and its consequence.

The ideas were developed in the 1950s while tracking down the case against cigarettes (Wikipedia: Bradford Hill criteria) and continue to be debated.

# How did the observations get into the study?

- There is risk of **selection bias** if the individuals are not selected randomly from the population they are supposed to represent.
- Selection bias is a type of confounding. The confounder is a variable that explains the selection process.

**Question 9.25.** In World War II, the US Airforce was suffering heavy losses in bombing raids over Germany. To decide where to add extra armor, engineers studied bullet holes on returning planes to see which parts were exposed to most gunfire. A prominent statistician, Abraham Wald, provided a different interpretation. What was it?

todo

## Revisiting the fieldgoal kicker data

- Any observations study can and should be examined for confounding and selection bias issues.

**Question 9.26.** Consider the field goal percentage data. Recall that we analyzed the 19 NFL kickers who made at least ten field goal attempts in each of 2002, 2003, 2004, 2005 and 2006 seasons. We found a slope of  $-0.504$  when predicting field goal percentage in year  $t$  using field goal percentage in year  $t-1$ , with a separate intercept for each kicker. Comment on the possible roles of selection bias and/or confounding for interpreting this result.

todo

# Randomized experiments and random samples

- The huge difficulties interpreting observational studies motivate avoiding them whenever possible
- Random assignment to treatment in a controlled experiment removes the possibility of confounding, and ensures that any statistically significant effect can legitimately be given a causal interpretation.
- A **randomized experiment** occurs when individual  $i$  is randomly assigned a **treatment**. A treatment is a set of explanatory variables corresponding to a row of  $\mathbf{X}$ .
- In a randomized experiment the independence assumption on the errors is reasonable: we can view the errors as coming from differences between individuals drawn independently from a large population.
- STATS 470 investigates design and analysis of experiments, which is an extension of linear statistical modeling.
- Random sampling removes selection bias, apart from missing data.
- STATS 480 investigates sample survey analysis, which also builds on linear statistical modeling.

# Selecting from many possible models

- Suppose we have a large number  $\ell$  of potential explanatory variables in our dataset.
- The total number of possible linear models is  $2^\ell$  since each of the  $\ell$  variables can be either in or out of the model.
- If we allow for the possibility of interactions, things are even worse.
- For two variables  $x_{i1}$  and  $x_{i2}$  on each individual  $i = 1, 2, \dots, n$ , modeling an **interaction** can be viewed as including a new variable  $x_{i3} = x_{i1}x_{i2}$ .

**Question 9.27.** If there are  $\ell$  explanatory variables, considered as **main effects**, and any pair of them could give rise to an **interaction effect**, how many possible models are there? For simplicity, allow for the possibility of including interactions without the main effects.

# Practical considerations for model selection

- Sometimes, you build models based on specific hypotheses about the system you are investigating.
- In this case, our tools for hypothesis testing work well. You work through a process of starting with a basic model and considering a relatively small sequence of alternative hypotheses to build up an understanding of the data.
- A different scenario occurs when you explore a very large number of different models.
- If you consider 1000 alternative models and each one is tested at significance level 0.01 then you expect to find 10 models that would formally let you reject the null hypothesis at a “high” level of significance for random variables generated under the null model.
- Similar issues arise if you consider many variables in a single linear model and look to identify significant ones.



# The expected number of false discoveries

**Question 9.28.** Suppose that you consider  $\ell = 100$  variables by placing them all in a linear model and reporting the variables whose t statistic is significant at the 0.05 level. How many “significant” variables would you expect to report under a null probability model where all the coefficients are zero?

todo

## Confidence intervals after model selection

**Question 9.29.** Suppose you have  $\ell = 100$  explanatory variables and you consider  $\ell = 100$  different models, each with only one of the explanatory variables in the model. You pick as your favorite model the one with the highest  $R^2$  statistic, which is equivalent to picking the one with the smallest p-value for its t statistic. You report a 95% confidence interval for the coefficient in this linear model. What is the chance that this confidence interval will cover the truth, under the null probability model where all the coefficients for all the explanatory variables are zero?

todo

# Dealing with multiple testing

- The difficulty of properly evaluating statistical significance when investigating very many hypotheses is called the **multiple testing** situation.
- Dealing with multiple testing is a current scientific concern. It is related to the so-called crisis in scientific reproducibility.
- Advances in data acquisition and computation increasingly lead to large datasets to be investigated.
- One principle: report all the tests you make, not just the nominally significant ones. This lets the reader assess the hazard of multiple testing bias.
- Another principle: any result not yet confirmed by an independent experiment is suspicious.