

SCHOOL OF PUBLIC HEALTH · UNIVERSITY of WASHINGTON

Department of Biostatistics

## **REGRESSION METHODS: CONCEPTS AND APPLICATIONS**

## **Motivation**

- Objective: Investigate associations between two or more variables
- What tools do you already have?
  - t-test
    - Comparison of means in two populations
  - Chi-squared test
    - Comparison of proportions in two populations
- What will we cover in this module?
  - Linear Regression
    - Association of a continuous outcome with one or more predictors (categorical or continuous)
  - Analysis of Variance (as a special case of linear regression)
    - Comparison of a continuous outcome over a fixed number of groups
  - Logistic and Relative Risk Regression
    - Association of a binary outcome with one or more predictors (categorical or continuous)

#### Module structure

- 10 sessions over 2.5 days
- Lectures and hand-on exercises in R
- Day 1
  - Simple linear regression
- Day 2
  - Model checking
  - Multiple linear regression
  - ANOVA
- Day 3
  - Logistic regression
  - Generalized linear models



#### SIMPLE LINEAR REGRESSION

## **Outline: Simple Linear Regression**

#### Motivation

- The equation of a straight line
- Least Squares Estimation
- Inference
  - About regression coefficients
  - About predictions
- Model Checking
  - Residual analysis
  - Outliers & Influential observations

- Linear regression is concerned with a continuous outcome
- Data: Factors related to serum total cholesterol (continuous outcome), 400 individuals, 11 variables

>	· head(cholesterol)											
	ID	sex	age	chol	BMI	ΤG	APOE	rs174548	rs4775401	HTN	chd	
	1	1	74	215	26.2	367	4	1	2	1	1	
	2	1	51	204	24.7	150	4	2	1	1	1	
	3	0	64	205	24.2	213	4	0	1	1	1	
	4	0	34	182	23.8	111	2	1	1	1	0	
	5	1	52	175	34.1	328	2	0	0	1	0	
	6	1	39	176	22.7	53	4	0	2	0	0	

- Our first goal:
  - Investigate the relationship between cholesterol (mg/dl) and age in adults



#### Is cholesterol associated with age?

 You could dichotomize age and compare cholesterol between two age groups

```
> group = 1*(age > 55)
> group=factor(group,levels=c(0,1), labels=c("30-55","56-80"))
> table(group)
group
30-55 56-80
201 199
> boxplot(chol~group,ylab="Total cholesterol(mg/dl)")
```



Is cholesterol associated with age?

• You could compare mean cholesterol between two groups: t-test

Question: What do the boxplot and the t-test tell us about the relationship between age and cholesterol?

```
> t.test(chol ~ group)
Welch Two Sample t-test
data: chol by group
t = -3.637, df = 393.477, p-value = 0.0003125
alternative hypothesis: true difference in mea
95 percent confidence interval:
-12.200209 -3.638487
sample estimates:
mean in group 30-55 mean in group 56-80
179.9751 187.8945
```

#### Using the t-test:

- There is a statistical association between cholesterol and age
- There appears to be a positive association between cholesterol and age
  - Is there any way we could estimate the magnitude of this association without breaking the "continuous" measure of age into subgroups?
- With the t-test, we compared mean cholesterol in two age groups, could we compare mean cholesterol across "continuous" age?

We might assume that mean cholesterol changes linearly with age:



Can we find the equation for a straight line that best fits these data?

## Linear Regression

- A statistical method for modeling the relationship between a continuous variable [response/outcome/dependent] and other variables [predictors/exposure/independent]
  - Most commonly used statistical model
  - Flexible
  - Well-developed and understood properties
  - Easy interpretation
  - Building block for more general models
- Goals of analysis:
  - Estimate the association between response and predictors or,
  - Predict response values given the values of the predictors.
- We will start our discussion studying the relationship between a response and <u>a single predictor</u>
  - Simple linear regression model









Х x+1



Х





- Slope b<sub>1</sub> is the change in y corresponding to a unit increase in x
- Slope gives information about magnitude and direction of the association between x and y

## The straight line equation



# Simple Linear Regression

- We can use linear regression to model how the mean of an outcome Y changes with the level of a predictor, X
- The individual Y observations will be scattered about the mean



## Simple Linear Regression

#### In regression:

- X is used to predict or explain outcome Y.
- Response or dependent variable (Y):
  - continuous variable we want to predict or explain
- Explanatory or independent or predictor variable (X):
  - attempts to explain the response
- Simple Linear Regression Model:

$$y = \beta_0 + \beta_1 x + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

Simple Linear Regression  

$$y = \beta_0 + \beta_1 x + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

The model consists of two components:

- Systematic component:  $E[Y | X = x] = \beta_0 + \beta_1 x$ Mean population value of Y at X=x  $\beta_0$ :intercept
- Random component:

$$Var[Y | X = x] = \sigma^2$$

Variance does not depend on x

#### Simple Linear Regression: Assumptions

**MODEL:** 
$$E[Y | X = x] = \beta_0 + \beta_1 x$$
  $Var[Y | X = x] = \sigma^2$ 

Distribution of Y at different x values:





Compare with the boxplots for two age groups

Simple Linear Regression: Interpreting model coefficients

• Model:  $E[Y|x] = \beta_0 + \beta_1 x$   $Var[Y|x] = \sigma^2$ 

• Question: How do you interpret  $\beta_0$ ?

#### Answer:

 $\beta_0 = E[Y|x=0]$ , that is, the mean response when x=0

Your turn: interpret  $\beta_1$ !

Simple Linear Regression: Interpreting model coefficients

• Model:  $E[Y|x] = \beta_0 + \beta_1 x$   $Var[Y|x] = \sigma^2$ 

• Question: How do you interpret  $\beta_1$ ?

Answer:

$$E[Y|x] = \beta_0 + \beta_1 x$$
  

$$E[Y|x+1] = \beta_0 + \beta_1(x+1) = \beta_0 + \beta_1 x + \beta_1$$

 $E[Y|x+1] - E[Y|x] = \beta_1$  independent of x (linearity)

i.e.  $\beta_1$  is the difference in the mean response associated with a one unit positive difference in  $\boldsymbol{x}$ 

Example: Cholesterol and age

- Recall: Our motivating example was to determine if there is an association between age (a continuous predictor) and cholesterol (a continuous outcome)
- Suppose: We believe they are associated via the linear relationship  $E[Y|x] = \beta_0 + \beta_1 x$
- Question: How would you interpret  $\beta_1$ ?
- Answer:

Example: Cholesterol and age

- Recall: Our motivating example was to determine if there is an association between age (a continuous predictor) and cholesterol (a continuous outcome)
- Suppose: We believe they are associated via the linear relationship  $E[Y|x] = \beta_0 + \beta_1 x$
- Question: How do you interpret  $\beta_1$ ?
- Answer:

 $\beta_1$  is the difference in mean cholesterol associated with a one year increase in age



Question: How to find a "best-fitting" line?



#### Least Squares Estimation

• Question: How to find a "best-fitting" line?



Idea: chooses the line that minimizes the sum of squares of the vertical distances from the observed points to the line.

#### Least Squares Estimation

The least squares regression line is given by

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

 So the (squared) distance between the data (y) and the least squares regression line is

$$D = \sum_{i} (y_i - \hat{y}_i)^2$$

- We estimate  $\beta_0$  and  $\beta_1$  by finding the values that minimize D
- We can use these estimates to get an estimate of the variance about the line (σ<sup>2</sup>)

### Least Squares Estimation

• These values are:

$$\hat{\beta}_{0} = \overline{y} - \hat{\beta}_{1}\overline{x}$$

$$\hat{\beta}_{1} = \frac{\sum (x_{i} - \overline{x})(y_{i} - \overline{y})}{\sum (x_{i} - \overline{x})^{2}}$$

• We estimate the variance as:

$$\hat{\sigma}^{2} = \frac{\sum_{i=1}^{n} r_{i}^{2}}{n-2} = \frac{\sum_{i=1}^{n} (y_{i} - \hat{y}_{i})^{2}}{n-2} = \frac{\sum_{i=1}^{n} (y_{i} - \hat{\beta}_{0} - \hat{\beta}_{1} x_{i})^{2}}{n-2}$$

### **Estimated Standard Errors**

- Recall that, when estimating parameters from a sample, there will be sampling variability in the estimates
- This is true for regression parameter estimates
- Looking at the formulas for  $\hat{\beta}_0$  and  $\hat{\beta}_1$ , we can see that they are just complicated means
- In repeated sampling we would get different estimates
- Knowledge of the sampling distribution of parameter estimates can help us make inference about the line
- Statistical theory shows that the sampling distributions are Normal and provides expressions for the mean and standard error of the estimates over repeated samples

"Regression" -> "Histograms on Simple Linear Regression" at https://lstat.kuleuven.be/newjava/vestac/




#### **Sampling Distribution**



#### **Sampling Distribution**



Java Applet Window

# Inference

### About regression model parameters

- Hypothesis testing: H<sub>0</sub>: β<sub>j</sub>=0 (j=0,1)
  - Test Statistic:
    - Large Samples:

$$\frac{\hat{\beta}_j - (null \ hyp)}{se(\hat{\beta}_j)} \sim N(0,1)$$

Small Samples:

$$\frac{\hat{\beta}_j - (null \ hyp)}{se(\hat{\beta}_j)} \sim t_{n-2}$$

Confidence Intervals:

 $\hat{\beta}_{j} \pm (critical \ value) \times se(\hat{\beta}_{j})$ 

[Don't worry about these formulae: we will use R to fit the models!]

Inference: Hypothesis Testing

# **Null Hypothesis:** $\beta_j = 0$

T=test statistic

**Alternative** 

P-Value

 $\beta_j > 0$   $P(t_{n-2} > T)$ 

 $\beta_j < 0$   $P(t_{n-2} < T)$ 

 $\beta_{j} \neq 0 \qquad \qquad 2P(t_{n-2} > |T|)$ 



## Inference: Confidence Intervals

100 (1- $\alpha$ )% Confidence Interval for  $\beta_j$  (j=0,1)

$$\hat{\beta}_{j} \pm t_{n-2,\frac{\alpha}{2}} SE(\hat{\beta}_{j})$$

Gives intervals that  $(1 - \alpha)100\%$  of the time will cover the true parameter value ( $\beta_0$  or  $\beta_1$ ).

We say we are " $(1-\alpha)100\%$  confident" the interval covers  $\beta_i$ .

```
> fit = lm(chol ~ age)
> summary(fit)
Call:
lm(formula = chol ~ age)
Residuals:
     Min
         10 Median 30 Max
-60.45306 -14.64250 -0.02191 14.65925 58.99527
Coefficients:
          Estimate Std. Error t value Pr(>|t|)
0.31033 0.07524 4.125 4.52e-05 ***
age
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1
Residual standard error: 21.69 on 398 degrees of freedom
Multiple R-squared: 0.04099, Adjusted R-squared: 0.03858
F-statistic: 17.01 on 1 and 398 DF, p-value: 4.522e-05
                                   > confint(fit)
                                                  2.5 % 97.5 %
                                   (Intercept) 158.5171656 175.2861949
```

age

42

0.1624211 0.4582481





What do these models results mean in terms of our scientific question?

Parameter estimates and confidence intervals:

 $\hat{\beta}_0 = 166.90$  95% CI: (158.5, 175.3)  $\hat{\beta}_1 = 0.31$  95% CI: (0.16, 0.46)

 $\hat{\beta}_0$ : The estimated average serum cholesterol for someone of age = 0 is 166.9 !?

Your turn: What about  $\hat{\beta}_1$  ?

What do these models results mean in terms of our scientific question?

Parameter estimates and confidence intervals:

$$\hat{\beta}_0 = 166.90$$
 95% CI: (158.5, 175.3)  
 $\hat{\beta}_1 = 0.31$  95% CI: (0.16, 0.46)

• Answer:  $\hat{\beta}_i$  mean cholesterol is estimated to differ by 0.31 mg/dl for each one year difference in age.

Question: What about the confidence intervals?

What do these models results mean in terms of our scientific question?

Parameter estimates and confidence intervals:

 $\hat{\beta}_0 = 166.90$  95% CI: (158.5, 175.3)  $\hat{\beta}_1 = 0.31$  95% CI: (0.16, 0.46)

- Answer: 95% CIs give us a range of values that will cover the true intercept and slope 95% of the time
  - For instance, we can be 95% confident that the true difference in mean cholesterol associated with a one year difference in age lies between 0.16 and 0.46 mg/dl

### Presentation of the results?

- The mean serum total cholesterol is significantly higher in older individuals (p < 0.001).</li>
- For each additional year of age, we estimate that the mean total cholesterol differs by approximately 0.31 mg/dl (95% CI: 0.16, 0.46). Or:
- For each additional 10 years of age, we estimate that the mean total cholesterol differs by approximately 3.10 mg/dl (95% CI: 1.62, 4.58).
  - Note:
    - Emphasis on slope parameter (sign and magnitude)
    - Confidence interval
    - <u>Units</u> for predictor and response. Scale matters!

# Inference for predictions

Given estimates  $\hat{\beta}_0$ ,  $\hat{\beta}_1$  we can find the **predicted**  $\hat{y}_i$ **value**, for any value of  $x_i$  as

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

• Interpretation of  $\hat{y}_i$  :

• Estimated mean value of Y at  $X = x_i$ 

Be Cautious: This assumes the model is true.

- May be a reasonable assumption within the range of your data.
- It may not be true outside the range of your data!



# Be careful of extrapolating



 It would not make sense to extrapolate height at age 20 from a study of girls aged 4-9 years!

### Prediction

Prediction of the mean <u>E[Y|X=x]</u>:
 Point Estimate:  $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ 

• Standard Error: 
$$se(\hat{y}) = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}}$$

Note that as x gets further from  $\overline{x}$ , variance increases!

• 100 (1- $\alpha$ )% confidence interval for E[Y|X=x]:  $\hat{y} \pm t_{n-2,1-\alpha/2} se(\hat{y})$ 

### Prediction

Prediction of a <u>new future observation</u>, y\*, at X=x:
 Point Estimate: ŷ<sup>\*</sup> = β̂₀ + β̂₁x

• Standard Error: 
$$se(\hat{y}^*) = \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}}$$

• 100 (1- $\alpha$ )% prediction interval for a new future observation:  $\hat{y}^* \pm t_{n-2,1-\alpha/2} se(\hat{y}^*)$ 

Standard error for the prediction of a future observation is bigger:It depends not only on the precision of the estimated mean, but also on the amount of variability in Y around the line.

#### Cholesterol Example: Prediction



Prediction of a new observation

Let's interpret these predictions
For x = 46

- $\hat{y} = 181.2$  95% CI: (178.7, 183.7)
- $\hat{y}^* = 181.2$  95% CI: (138.5, 223.9)
- Question: How do our interpretations for  $\hat{y}$  and  $\hat{y}^*$  differ?

Let's interpret these predictions
For x = 46

 $\hat{y} = 181.2$  95% CI: (178.7, 183.7)

 $\hat{y}^* = 181.2$  95% CI: (138.5, 223.9)

- Question: How do our interpretations for  $\hat{y}$  and  $\hat{y}^*$  differ?
- Answer: The point estimates represent our predictions for the mean serum cholesterol for individuals age 46 (ŷ) and for a single new individual of age 46 (ŷ\*)

Let's interpret these predictions
 For x = 46

- $\hat{y} = 181.2$  95% CI: (178.7, 183.7)
- $\hat{y}^* = 181.2$  95% CI: (138.5, 223.9)
- Question: Why are the confidence intervals for ŷ and ŷ\* of differing widths?

Let's interpret these predictions
For x = 46

 $\hat{y} = 181.2$  95% CI: (178.7, 183.7)

 $\hat{y}^* = 181.2$  95% CI: (138.5, 223.9)

- Question: Why are the confidence intervals for  $\hat{y}$  and  $\hat{y}^*$  of differing widths?
- Answer: The interval is broader when we make a prediction for a cholesterol level for a single individual because it must incorporate random variability around the mean.



- Let's put some of the concepts we have been discussing into practice
- Open up the Labs file and R Studio and follow the directions to load the class data set and install the R packages you will need for this module
- Work on Exercises 1-3
  - Try each exercise on your own
  - Once you have had a change to work on each question, form a group of 3 or 4 and share your ideas with your group
  - Each group will have an opportunity to present their ideas to the class

# Simple Linear Regression: R<sup>2</sup>

- Given no linear association:
  - We could simply use the sample mean to predict E(Y). The variability using this simple prediction is given by SST (to be defined shortly).
- Given a linear association:
  - The use of X permits a potentially better prediction of Y by using E(Y|X).
  - **Question:** What did we gain by using *X*?

Let's examine this question with the following figure

# Decomposition of sum of squares



# Decomposition of sum of squares

It is always true that: 
$$y_i - \overline{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \overline{y})$$

It can be shown that:

$$\sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$$

$$SST = SSE + SSR$$

- **SST:** describes the total variation of the  $Y_{i}$ .
- **SSE:** describes the variation of the  $Y_i$  around the regression line.
- **SSR:** describes the structural variation; how much of the variation is due to the regression relationship.
- This decomposition allows a characterization of the usefulness of the covariate X in predicting the response variable Y.

# Simple Linear Regression: R<sup>2</sup>

#### Given no linear association:

- We could simply use the sample mean to predict E(Y). The variability between the data and this simple prediction is given as SST.
- Given a linear association:
  - The use of X permits a potentially better prediction of Y by using E(Y | X).
  - **Question:** What did we gain by using X?
  - Answer: We can answer this by computing the proportion of the total variation that can be explained by the regression on X

$$R^{2} = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$$

This  $R^2$  is, in fact, the correlation coefficient squared.





Low values of  $R^2$  indicate that the model is not adequate. However, high values of  $R^2$  do not mean that the model is adequate!!

### **Cholesterol Example:**

Scientific Question: Can we predict cholesterol based on age?

```
> fit = lm(chol ~ age)
> summary(fit)
Call:
lm(formula = chol ~ age)
Residuals:
          10 Median
     Min
                                  30
                                           Max
-60.45306 -14.64250 -0.02191 14.65925 58.99527
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 166.90168 4.26488 39.134 < 2e-16 ***
      0.31033 0.07524 4.125 4.52e-05 ***
age
___
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1
Residual standard error: 21,69 on 398 degrees of freedom
Multiple R-squared: 0.04099,
                             Adjusted R-squared: 0.03858
F-statistic: 17.01 on 1 and 398 DF, p-value: 4.522e-05
```

#### 

65

### Cholesterol Example: Scientific Question: Can we predict cholesterol based on age?

- R<sup>2</sup>=0.04
- What does R<sup>2</sup> tell us about our model for cholesterol?

## Cholesterol Example:

Scientific Question: Can we predict cholesterol based on age?

- R<sup>2</sup>=0.04
- What does R<sup>2</sup> tell us about our model for cholesterol?
- Answer: 4% of the variability in cholesterol is explained by age. Although mean cholesterol increases with age, there is much more variability in cholesterol than age alone can explain

### Cholesterol Example: Scientific Question: Can we predict cholesterol based on age?

Decomposition of Sum of Squares and the F-statistic



In simple linear regression:

F-statistic =  $(t-statistic for slope)^2$ 

Hypothesis being tested:  $H_0$ :  $\beta_1=0$ ,  $H_1$ :  $\beta_1\neq 0$ .

# Simple Linear Regression: Assumptions

- 1. E[Y|x] is related linearly to x
- 2. Y's are independent of each other
- 3. Distribution of [Y|x] is normal
- 4. Var[Y|x] does not depend on x

Linearity Independence Normality Equal variance

Can we assess if these assumptions are valid?

Model Checking: Residuals

 (Raw or unstandardized) Residual: difference (r<sub>i</sub>) between the observed response and the predicted response, that is,

$$r_i = y_i - \hat{y}_i$$
$$= y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

The residual captures the component of the measurement y<sub>i</sub> that cannot be "explained" by x<sub>i</sub>.

# Model Checking: Residuals

- Residuals can be used to
  - Identify poorly fit data points
  - Identify unequal variance (heteroscedasticity)
  - Identify nonlinear relationships
  - Identify additional variables
  - Examine normality assumption

# Model Checking: Residuals

Linearity	Plot residual vs X or vs Ŷ
	Q: Is there any structure?
Independence	
	Q: Any scientific concerns?
Normality	Residual histogram or qq-plot
	Q: Symmetric? Normal?
Equal variance	Plot residual vs X
	Q: Is there any structure?
# Model Checking: Residuals

 If the linear model is appropriate we should see an unstructured horizontal band of points centered at zero as seen in the figure below



# Model Checking: Residuals





- The linearity assumption is important: interpretation of the slope estimate depends on the assumption of the same rate of change in E(Y|X) over the range of X
- Preliminary Y-X scatter plots and residual plots can help identify non-linearity
- If linearity cannot be assumed, consider alternatives such as polynomials, fractional polynomials, splines or categorizing X



- The independence assumption is also important: whether observations are independent will be known from the study design
- There are statistical approaches to accommodate dependence, e.g. dependence that arises from cluster designs

# Normality

- The Normality assumption can be visually assessed by a histogram of the residuals or a normal QQ-plot of the residuals
- A QQ-plot is a graphical technique that allows us to assess whether a data set follows a given distribution (such as the Normal distribution)
  - The data are plotted against a given theoretical distribution
    - Points should approximately fall in a straight line
    - Departures from the straight line indicate departures from the specified distribution.
- However, for moderate to large samples, the Normality assumption can be relaxed

Lumley T et al. The importance of the normality assumption in large public health data sets. Annu Rev Public Health 2002; 23: 151-169.

# Equal variance

- Sometimes variance of Y is not constant across the range of X (heteroscedasticity)
- Little effect on point estimates but variance estimates may be incorrect
- This may affect confidence intervals and p-values
- To account for heteroscedasticity we can
  - Use robust standard errors
  - Transform the data
  - Fit a model that does not assume constant variance (GLM)

# Robust standard errors

- Robust standard errors correctly estimate variability of parameter estimates even under nonconstant variance
  - These standard errors use empirical estimates of the variance in y at each x value rather than assuming this variance is the same for all x values
- Regression point estimates will be unchanged
- Robust or empirical standard errors will give correct confidence intervals and p-values

#### Cholesterol-Age example: Residuals

Plot of residuals versus fitted values Structure? Heteroscedasticity?

**R COMMAND:** plot(fit\$fitted, fit\$residuals)

Plot of residuals versus quantiles of a normal distribution(for n > 30) Normality?

**R COMMAND:** qqnorm(fit\$residuals)





# Linear regression for association between age and triglycerides



> fit.tg=lm(TG~age)

Robust standard errors

- Residual analysis suggests meanvariance relationship
- Use robust standard errors to get correct variance estimates



### Cholesterol example: Robust standard errors

### Linear regression results:



### Cholesterol example: Robust standard errors

### Linear regression results:



## Transformations

- Some reasons for using data transformations
  - Content area knowledge suggests nonlinearity
  - Original data suggest nonlinearity
  - Equal variance assumption violated
  - Normality assumption violated
- Transformations may be applied to the response, predictor or both
  - Be careful with the interpretation of the results
- Rarely do we know which transformation of the predictor provides best "linear" fit – best to choose transformation on scientific grounds
  - As always, there is a danger in using the data to estimate the best transformation to use
    - If there is no association of any kind between the response and the predictor, a "linear" fit (with a zero slope) is the correct one
    - Trying to detect a transformation is thus an informal test for an association
      - Multiple testing procedures inflate the Type I error

Model Checking: Outliers vs Influential observations

- Outlier: an observation with a residual that is unusually large (positive or negative) as compared to the other residuals.
- Influential point: an observation that has a notable influence in determining the regression equation.
  - Removing such a point would markedly change the position of the regression line.
  - Observations that are somewhat extreme for the value of x can be influential.

# **Outlier vs Influential observations**

У



Point A is an outlier, but is not influential.

Χ

# **Outlier vs Influential observations**



Point B is influential, but not an outlier.

# Cholesterol-Age Example: Residuals

Histogram of fit\$residuals



# Model Checking: Deletion diagnostics

$$\Delta \beta_{(i)} = \hat{\beta} - \hat{\beta}_{(-i)}$$
 : Delta-beta

: Standardized Delta-beta

Delta-beta

- : tells how much the regression coefficient changed by excluding the i<sup>th</sup> observation
- Standardized delta-beta

 $se(\hat{\beta})$ 

: approximates how much the t-statistic for a coefficient changed by excluding the i<sup>th</sup> observation

### **Cholesterol-Age Example: Deletion diagnostics**

	> aī.	p = dipeta()	ELT)						
	<pre>&gt; index=order(abs(dfb[,2]),decreasing=T)</pre>								
<pre>&gt; cbind(dfb[index[1:15],],age[index[1;1])</pre>									
	(In	tercept)	age						
	114	-0.9893663	0.015268514	34					
	166	-0.6827966	0.014888475	78					
	255	-0.6190643	0.013902713	75					
	186	-0.8544144	0.013279531	33					
	113	0.5376293	-0.011943495	76					
	325	-0.7517511	0.011308451	37					
	365	0.7676508	-0.011297278	39					
	257	-0.7374003	0.011092575	37					
	290	-0.7024787	0.010757541	35					
	144	0.7120264	-0.010710881	37					
	197	-0.6784150	0.010469720	34					
	296	-0.6499386	0.010101515	33					
	231	-0.6293174	0.009712016	34					
	7	0.4403297	-0.009524470	79					
	252	-0.5981020	0.009412761	31					

(

No evidence of influential points. The largest (in absolute value) delta beta is 0.015 compared to the estimate of 0.31 for the regression coefficient.

Model Checking

What to do if you find an outlier and/or influential observation:

Check it for accuracy

- Decide (based on scientific judgment) whether it is best to keep it or omit it
  - If you think it is representative, and likely would have appeared in a larger sample, keep it
  - If you think it is very unusual and unlikely to occur again in a larger sample, omit it
  - Report its existence [whether or not it is omitted]

### Simple Linear Regression: Impact of Violations of Model Assumptions

	Non	Non	Unequal	Dependence	
	Linearity	Normality	Variances		
Estimates	Problematic	Little impact for most departures. Extreme outliers can be a problem.	Little impact	Mostly little impact	
Tests/CIs	Problematic	Little impact for most departures. CIs for correlation are sensitive.	Variance estimates may be wrong, but the impact is usually not dramatic	Variance estimates may be wrong	
Correction	Choose a nonlinear approach (possible within the linear	Mostly no correction needed. Delete outliers (if warranted) or	Use robust standard errors	Regression for dependent data	
	regression framework)	use robust regression		93	

### Exercise

- Work on Exercises 4-6
  - Try each exercise on your own
  - In groups of 3 or 4, share your ideas
  - Each group will have an opportunity to present their ideas to the class



UW School of Public Health and Community Medicine Department of Biostatistics

# **REGRESSION MODELS**

### MULTIPLE LINEAR REGRESSION

# **Outline: Multiple Linear Regression**

- Motivation
- Model and Interpretation
- Estimation and Inference
- Interaction

### **Motivation**

The response or dependent variable, Y, may depend on several predictors not just one!

- Multiple regression is an attempt to consider the <u>simultaneous</u> influence of several variables on the response
- This may be with the goal of an unbiased estimate of association or for better prediction

### **Motivation**

• Why not fit multiple separate simple linear regressions?

- If the goal is to estimate the *association* between the response and a predictor of interest, a confounder can make the observed association appear
  - stronger than the true association,
  - weaker than the true association, or
  - even the reverse of the true association
- How can we address this:
  - We can adjust for the effects of the confounder by adding a corresponding term to our linear regression
- If the goal is *prediction* of the response, we may be able to improve prediction by including additional variables in the regression model

# Motivation: Cholesterol Example

### Data

>	> head(cholesterol)										
	ID	sex	age	chol	BMI	ΤG	APOE	rs174548	rs4775401	HTN	chd
	1	1	74	215	26.2	367	4	1	2	1	1
	2	1	51	204	24.7	150	4	2	1	1	1
	3	0	64	205	24.2	213	4	0	1	1	1
	4	0	34	182	23.8	111	2	1	1	1	0
	5	1	52	175	34.1	328	2	0	0	1	0
	6	1	39	176	22.7	53	4	0	2	0	0

### • Our goal:

 Investigate the relationship between age (years), BMI (kg/m<sup>2</sup>) and serum total cholesterol (mg/dl)



In general, the multiple regression equation can be written as follows:

$$\mathsf{E}[\mathsf{Y} | \mathbf{x}_{1}, \mathbf{x}_{2}, \dots, \mathbf{x}_{p}] = \beta_{0} + \beta_{1} \mathbf{x}_{1} + \beta_{2} \mathbf{x}_{2} + \dots + \beta_{p} \mathbf{x}_{p}$$

- We use multiple variables when:
  - The predictor variable is categorical with more than two groups
  - We need polynomials, splines or other functions to model the shape of the relationship(s) accurately
- Estimating association:
  - We want to adjust for confounding by other variables
  - We want to allow the association to differ for different values of other variables (interaction)
- Prediction: we use multiple variables if we think more than one variable will be useful in predicting future outcomes accurately

# Model and Interpretation

• Model: 
$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \varepsilon$$
  
where we assume  $\varepsilon \sim N(0, \sigma^2)$ 

Extension of simple linear regression!

• Systematic component:

$$E[Y|x_1,...,x_p] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$$

Random component:

 $Var[Y \mid x_1, ..., x_p] = \sigma^2$ 

# Model and Interpretation

 For example, let us assume that there are two predictors in the model and so

```
\mathsf{E}[\mathsf{Y}|\mathsf{x}_{1}, \mathsf{x}_{2}] = \beta_{0} + \beta_{1} \mathsf{x}_{1} + \beta_{2} \mathsf{x}_{2}
```

Consider two observations with the same value for  $x_2$ , but one observation has  $x_1$  one unit higher, that is,

Obs 1: 
$$E[Y|x_1=k+1, x_2=c] = \beta_0 + \beta_1 (k+1) + \beta_2 c$$
  
Obs 2:  $E[Y|x_1=k, x_2=c] = \beta_0 + \beta_1 (k) + \beta_2 c$ 

Thus, 
$$E[Y|x_1=k+1, x_2=c] - E[Y|x_1=k, x_2=c] = \beta_1$$

That is,  $\beta_1$  is the expected mean change in y per unit change in  $x_1$  if  $x_2$  is held constant (adjusted/controlling for  $x_2$ )

Similar interpretation applies to  $\beta_2$ 

# Model and Interpretation

- To facilitate our discussion let's assume we have two predictors with binary values
- Model:

$$E[Y | x_1, x_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Mean of Y	X <sub>2</sub> =0	X <sub>2</sub> =1
X <sub>1</sub> =0	β <sub>0</sub>	$\beta_0 + \beta_2$
X <sub>1</sub> =1	$\beta_0 + \beta_1$	$\beta_0 + \beta_1 + \beta_2$

$$E[Y|x_1=1, x_2=0] - E[Y|x_1=0, x_2=0] = \beta_1$$
  

$$E[Y|x_1=1, x_2=1] - E[Y|x_1=0, x_2=1] = \beta_1$$
  

$$E[Y|x_1=0, x_2=1] - E[Y|x_1=0, x_2=0] = \beta_2$$
  

$$E[Y|x_1=1, x_2=1] - E[Y|x_1=1, x_2=0] = \beta_2$$

### **Estimation**

#### Least Squares Estimation:

 As in linear regression, chooses the coefficient estimates that minimize the residual sum of squares

$$D = \sum_{i} (y_i - \hat{y}_i)^2$$

 Computation more difficult, but statistical software (R) will do that for you!



# **Estimation and Inference**

Inference

About regression model parameters

Hypothesis Testing H<sub>0</sub>: β<sub>j</sub>=0 (j=0,1,2,...,p)

<u>Interpretation</u>: Is there a statistically significant relationship between the response y and x<sub>j</sub> after adjusting for all other factors (predictors) in the model?

Test Statistic:

$$\frac{\hat{\beta}_j - (null \ hyp)}{se(\hat{\beta}_j)} \sim t_{n-p-1}$$

Note: The square of the t-statistic gives the F-statistic and the test is known as the **partial F-Test** 

Confidence Intervals

 $\hat{\beta}_j \pm (critical \ value) \times se(\hat{\beta}_j)$ 

# **Estimation and Inference**

- About the full model
  - Hypotheses

H<sub>0</sub>:  $\beta_1 = \beta_2 = ... = \beta_p = 0$  vs. H<sub>1</sub>: At least one  $\beta_j$  is not null

Analysis of variance table

Source	df	SS	MS	F
Regression	р	$SSR = \sum (\hat{y}_i - \overline{y}_i)^2$	MSR= SSR/p	MSR/MSE
Residual	n-p-1	$SSE=\sum (y_i - \hat{y}_i)^2$	MSE= SSE/(n-p-1)	
Total	n-1	$SST = \sum (y_i - \overline{y})^2$		

# **Estimation and Inference**

- The F-value is tested against a F-distribution with p, n-p-1 degrees of freedom
  - If we reject the null hypothesis, then the predictors do aid in predicting Y [in this analysis we do not know which ones are important!]
  - Failing to reject the null hypothesis does not mean that none of the covariates are important, since the effect of one or more covariates may be "masked" by others. The hard part is choosing which covariates to include or exclude.
- This is known as the global (multiple) F-test

Scientific example: Modeling cholesterol using age and BMI

- We have seen that there is a significant relationship between age and cholesterol
- Can we better understand variability in cholesterol by incorporating additional covariates?


- It appears that BMI increases with increasing age
- And cholesterol increases with increasing BMI
- What if we want to estimate the association between age and cholesterol while holding BMI constant?
- Multiple regression!

```
> fit2=lm(chol~age+BMI)
> summary(fit2)
Call:
lm(formula = chol ~ age + BMI)
Residuals:
   Min 10 Median 30 Max
-58.994 -15.793 0.571 14.159 62.992
Coefficients:
           Estimate Std. Error t value Pr(>|t|)
(Intercept) 137.1612 9.0061 15.230 < 2e-16 ***
      0.2023 0.0795 2.544 0.011327 *
age
            1.4266 0.3822 3.732 0.000217 ***
BMT
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 21.34 on 397 degrees of freedom
Multiple R-squared: 0.07351, Adjusted R-squared: 0.06884
F-statistic: 15.75 on 2 and 397 DF, p-value: 2.62e-07
```

# • Our estimated regression equation is $\hat{y} = 137.16 + 0.20Age + 1.43BMI$

#### Question: How do we interpret the age coefficient?

• Our estimated regression equation is  $\hat{y} = 137.16 + 0.20Age + 1.43BMI$ 

- Question: How do we interpret the age coefficient?
- Answer: This is the estimated average difference in cholesterol associated with a one year difference in age for two subjects with the same BMI.

- Our estimated regression equation is  $\hat{y} = 137.16 + 0.20Age + 1.43BMI$
- The age coefficient from our simple linear regression model was 0.31.
- Question: Why do the estimates from the two models differ?

- Our estimated regression equation is  $\hat{y} = 137.16 + 0.20Age + 1.43BMI$
- The age coefficient from our simple linear regression model was 0.31.
- Question: Why do the estimates from the two models differ?
- Answer: We are now conditioning on or controlling for BMI so our estimate of the age association is among subjects with the same BMI.

```
Call:
lm(formula = chol ~ age + BMI)
Residuals:
   Min 10 Median 30
                                 Max
-58.994 -15.793 0.571 14.159 62.992
Coefficients:
           Estimate Std. Error t value Pr(>|t|)
(Intercept) 137.1612 9.0061 15.230 < 2e-16 ***
      0.2023 0.0795 2.544 0.011327 *
age
            1.4266 0.3822 3.732 0.000217 ***
BMT
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 21.34 on 397 degrees of freedom
Multiple R-squared: 0.07351, Adjusted R-squared: 0.06884
F-statistic: 15.75 on 2 and 397 DF, p-value: 2.62e-07
```

### **Cholesterol Example:**

Did adding BMI improve our model?

```
> anova(fit,fit2)
Analysis of Variance Table
Model 1: chol ~ age
Model 2: chol ~ age + BMI
Res.Df RSS Df Sum of Sq F Pr(>F)
1 398 187187
2 397 1 80842 1 6345.8 13.931 0.0002174 ***
--- Signif. codes: 0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

How does the model with age and BMI compare to a model that contains only the mean?

#### Exercise

- Work on Exercise 7
  - Try each exercise on your own
  - In groups of 3 or 4, share your ideas
  - Each group will have an opportunity to present their ideas to the class

### Interaction and Linear Regression

- Statistical interaction (aka effect modification) occurs when the relationship between an outcome variable and one predictor is different depending on the levels of a second predictor
- Interactions are usually investigated because of a priori assumptions/hypotheses on the part of the researchers
- Linear regression models allow for the inclusion of interactions with cross-product terms

### Confounding vs. Interaction/Effect Modification

- Data and scientific understanding help distinguish between confounding and effect modifying variables:
  - Confounder: Associated with predictor and response; Association between response and predictor constant across strata of the new variable
  - Effect modifier/interaction: Association between response and the predictor varies across strata of the new variable

#### Confounding vs. Interaction/Effect Modification

- Confounding: Estimates of association from unadjusted analysis are markedly different from estimates of association from adjusted analysis
  - Association within each stratum is similar, but different from the "crude" association in the combined data (ignoring the strata)
  - In linear regression, these symptoms are diagnostic of confounding
- Effect modification would show differences between adjusted analysis and unadjusted analysis, but would also show different associations in the different strata

- Even if present, effect modification may not always be of interest in summarizing the effect of a predictor.
- For example, pleconaril, an antiviral drug, reduced the mean duration of symptoms in subjects with a common cold due to rhinoviruses but had no effect in subjects whose cold was due to some other agent.
- In the case of the pleconaril, effect modification was important in checking that the drug did actually work by inhibiting rhinovirus. However, in clinical use of the drug, it would typically not be possible to determine the infectious agent (the tests are expensive and take longer than just recovering from the cold), and so the average effectiveness of the drug across all colds would be a more important quantity.

### **Graphical Representation**



### **Graphical Representation** Y W=1No parallel lines Interaction W=0Х

### **Graphical Representation** Y W=1No parallel lines Interaction W=0Х

### Graphical Representation





### Model and Interpretation: interaction

Assume that there are two predictors in the model  $E[Y|x_1, x_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$ 

Consider two observations with the same value, c, for  $x_2$ , but one observation has  $x_1$  one unit higher

Obs 1:  $E[Y|x_1=k+1, x_2=c] = \beta_0 + \beta_1 (k+1) + \beta_2 c + \beta_3 (k+1)c$ Obs 2:  $E[Y|x_1=k, x_2=c] = \beta_0 + \beta_1 (k) + \beta_2 c + \beta_3 kc$ 

Thus, 
$$E[Y|x_1=k+1, x_2=c] - E[Y|x_1=k, x_2=c] = \beta_1 + \beta_3 c$$

That is, the difference in means depends now on the value of x<sub>2</sub>!

Model and Interpretation: interaction

• Model:  $E[Y|x_1, x_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$ 

Difference in Means:
 E[Y|x<sub>1</sub>=k+1, x<sub>2</sub>=c] - E[Y|x<sub>1</sub>=k, x<sub>2</sub>=c] = β<sub>1</sub> + β<sub>3</sub> c

The difference in means depends on the value of  $x_2$ 

- The difference in means is  $\beta_1$  if c=0.
- The difference in means is  $\beta_1 + \beta_3$  if c=1
- The difference in means changes by  $\beta_3$  for each unit difference in c (that is, in  $x_2$ ) [that is,  $\beta_3$  is the difference of differences!]
- H<sub>0</sub>: β<sub>3</sub>=0 tests for interaction

Model and Interpretation: interaction

• Model: E[Y| $x_1$ ,  $x_2$ ] =  $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$ 

Another way to look at this

### • Factor terms involving $x_1$ : $E[Y|x_1, x_2] = \beta_0 + (\beta_1 + \beta_3 x_2)x_1 + \beta_2 x_2$

#### Slope of x<sub>1</sub> changes with x<sub>2</sub>, i.e.

Difference in means for each unit difference in  $x_1$  changes with  $x_2$  (for each one unit difference in  $x_2$ , the difference in means changes by  $\beta_3$ )



We first fit the model with age and sex terms only (Male: sex=0, Female: sex=1)

```
> fit3 = lm(chol ~ age+sex)
> summary(fit3)
Call:
lm(formula = chol ~ age + sex)
Residuals:
        10 Median 30
   Min
                                  Max
-55.662 -14.482 -1.411 14.682 57.876
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 162.35445 4.24184 38.275 < 2e-16 ***
        0.29697 0.07313 4.061 5.89e-05 ***
age
           10.50728 2.10794 4.985 9.29e-07 ***
sex
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1
Residual standard error: 21.06 on 397 degrees of freedom
Multiple R-squared: 0.09748, Adjusted R-squared: 0.09293
F-statistic: 21.44 on 2 and 397 DF, p-value: 1.440e-09
```



Age (years)

- This model indicates that, after controlling for the effect of sex, the average cholesterol differs by 0.30 for each additional year of age
- The age effect in this model is very similar to the effect from our simple linear regression (0.31)
- However, this does not mean that the age/cholesterol relationship is the same in males and females
- To answer this question we must add the interaction term

Model with age and sex main effects, plus interaction effect

```
> fit4=lm(chol~age*sex)
> summary(fit4)
Call:
lm(formula = chol ~ age * sex)
Residuals:
   Min 10 Median 30
                                Max
-56.474 -14.377 -1.215 14.764 58.301
Coefficients:
           Estimate Std. Error t value Pr(>|t|)
(Intercept) 160.31151 5.86268 27.344 < 2e-16 ***
age 0.33460 0.10442 3.204 0.00146 **
sex 14.56271 8.29802 1.755 0.08004.
age:sex -0.07399 0.14642 -0.505 0.61361
____
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1
Residual standard error: 21.08 on 396 degrees of freedom
Multiple R-squared: 0.09806, Adjusted R-squared: 0.09123
F-statistic: 14.35 on 3 and 396 DF, p-value: 6.795e-09
```

Call: lm(formula = chol ~ age * sex)						
Residuals: Min -56.474 -14	1Q Median 377 -1.215 14.	3Q Max 764 58.301				
Coefficients:						
	Estimate Std. E	Error t value	Pr(> t )			
(Intercept)	160.31151 5.8	36268 27.344	< 2e-16 ***			
age	0.33460 0.1	.0442 3.204	0.00146 **			
sex	14.56271 8.2	.9802 1.755	0.08004 .			
age:sex	-0.07399 0.1	4642 -0.505	0.61361			
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						
Residual standard error: 21.08 on 396 degrees of freedom Multiple R-squared: 0.09806, Adjusted R-squared: 0.09123 F-statistic: 14.35 on 3 and 396 DF, p-value: 6.795e-09						

Mean cholesterol for males at age 0

Call: lm(formula = chol ~ age * sex)						
Residuals:						
Min 1Q Median 3Q Max						
-50.474 -14.577 -1.215 14.764 58.501						
Coefficients:						
Estimate Std. Error t value Pr(> t )						
(Intercept) 160.31151 5.86268 27.344 < 2e-16 ***						
age 0.33460 0.10442 3.204 0.00146 **						
sex 14.56271 8.29802 1.755 0.08004 .						
age:sex -0.07399 0.14642 -0.505 0.61361						
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						
Residual standard error: 21.08 on 396 degrees of freedom Multiple R-squared: 0.09806, Adjusted R-squared: 0.09123 F-statistic: 14.35 on 3 and 396 DF, p-value: 6.795e-09						

Difference in mean cholesterol between males and females at age 0

Call:						
lm(formula =	chol ~ age * sex)					
Residuals:	X					
Min	10 Median 30 Max					
-56.474 -14.3	377 -1.215 14.764 58.301					
Coefficients:	:					
	Estimate Std. Error t value Pr(> t )					
(Intercept) 1	160.31151 5.86268 27.344 < 2e-16 ***					
age	0.33460 0.10442 3.204 0.00146 **					
sex	14.56271 8.29802 1.755 0.08004 .					
age:sex	-0.07399 0.14642 -0.505 0.61361					
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						
Residual standard error: 21.08 on 396 degrees of freedom Multiple R-squared: 0.09806, Adjusted R-squared: 0.09123						
F-statistic: 14.35 on 3 and 396 DF, p-value: 6.795e-09						

Difference in mean cholesterol associated with each one year change in age for males

Call: lm(formula = chol ~ age * sex)						
Residuals: Min -56.474 -14.	1Q Median 377 -1.215	3Q 14.764	Max 58.301			
Coefficients:						
	Estimate St	d. Error	t value	Pr(> t )		
(Intercept)	160.31151	5.86268	27.344	< 2e-16 ***		
age	0.33460	0.10442	3.204	0.00146 **		
sex	14.56271	8.29802	1.755	0.08004 .		
age:sex	-0.07399	0.14642	-0.505	0.61361		
 Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						
Residual standard error: 21.08 on 396 degrees of freedom Multiple R-squared: 0.09806, Adjusted R-squared: 0.09123 F-statistic: 14.35 on 3 and 396 DF, p-value: 6.795e-09						

Difference in change in mean cholesterol associated with each one year change in age for females compared to males

### Interpretation?

Estimated model:

160.3 + 0.33 Age + 14.56 Sex- 0.07 Age × Sex

Subject 1: Age = a+1, sex = bSubject 2: Age = a, sex = bDifference in the estimated cholesterol: [160.3 + 0.33(a+1) + 14.56(b) - 0.07(a+1)(b)] - [160.3 + 0.33(a) + 14.56(b) - 0.07(a)(b)] = 0.33-0.07b

 Sex exerts a small (not statistically significant) effect on the age/cholesterol relationship In males: 160.3+0.33 Age In females: 174.9+0.26 Age

 We can also test the significance of interaction terms using an F-test

 Adding the interaction term did not significantly improve model fit





We have considered:

- Simple linear regression
  - Interpretation
  - Estimation
  - Model checking
- Multiple linear regression
  - Confounding
  - Interpretation
  - Estimation
  - Interaction

#### Exercise

- Work on Exercise 8
  - Try each exercise on your own
  - In groups of 3 or 4, share your ideas
  - Each group will have an opportunity to present their ideas to the class