REGRESSION MODELS

MULTIPLE LINEAR REGRESSION

Outline: Multiple Linear Regression

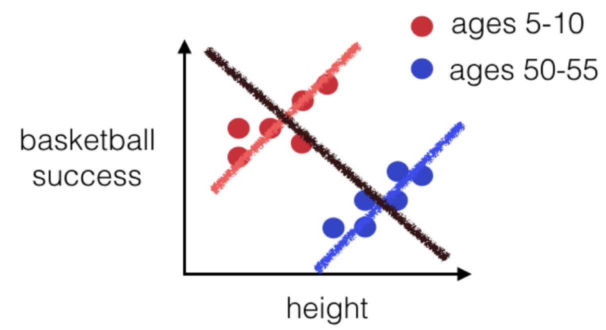
- Motivation
- Model and Interpretation
- Estimation and Inference
- Interaction

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  DM age chol  BMI  TG 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^2) and serum total cholesterol (mg/dl)

Motivation

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

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

- 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: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \varepsilon$

  where we assume $\varepsilon \overset{iid}{\sim} N(0, \sigma^2)$

  Extension of simple linear regression!

  - Systematic component:

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

  - Random component:

    $$Var[Y | x_1, \ldots, x_p] = \sigma^2$$

Model and Interpretation

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

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

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, β_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 β_2

Model and Interpretation

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

- Model:

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

Mean of Y	$X_2=0$	$X_2=1$
$X_1=0$	β_0	$\beta_0+\beta_2$
$X_1=1$	$\beta_0+\beta_1$	$\beta_0+\beta_1+\beta_2$

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

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

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

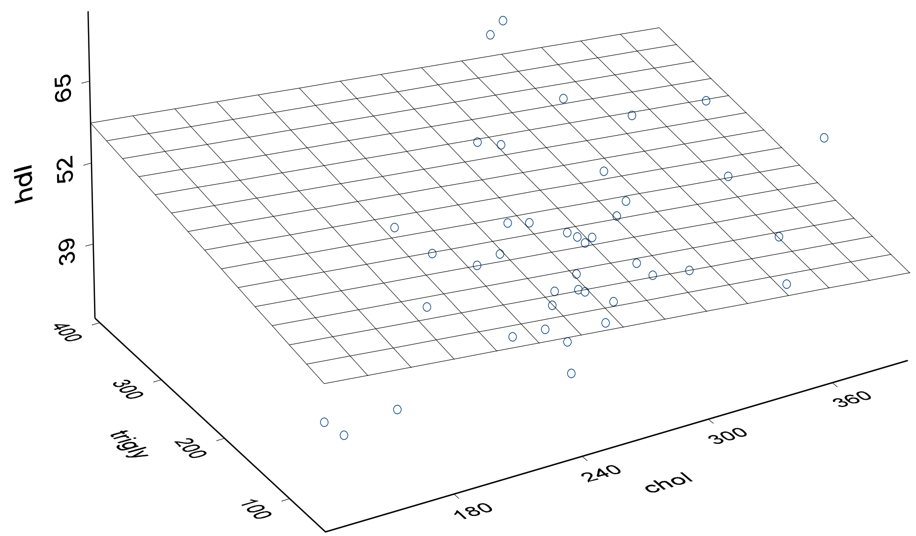
$E[Y|x_1=1, x_2=1] - E[Y|x_1=1, x_2=0] = \beta_2$

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_0: $\beta_j=0$ $(j=0,1,2,\ldots,p)$

<u>Interpretation:</u> Is there a statistically significant relationship between the response y and x_j 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 **<u>partial F-Test</u>**

    - **Confidence Intervals**

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

Estimation and Inference

- About the full model
  - Hypotheses

    H_0: $\beta_1 = \beta_2 = \ldots = \beta_p = 0$ vs. H_1: At least one β_j is not null

  - Analysis of variance table

Source	df	SS	MS	F
Regression	p	SSR= $\sum(\hat{y}_i - \bar{y}_i)^2$	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= $\sum(y_i - \bar{y})^2$		

Estimation and Inference

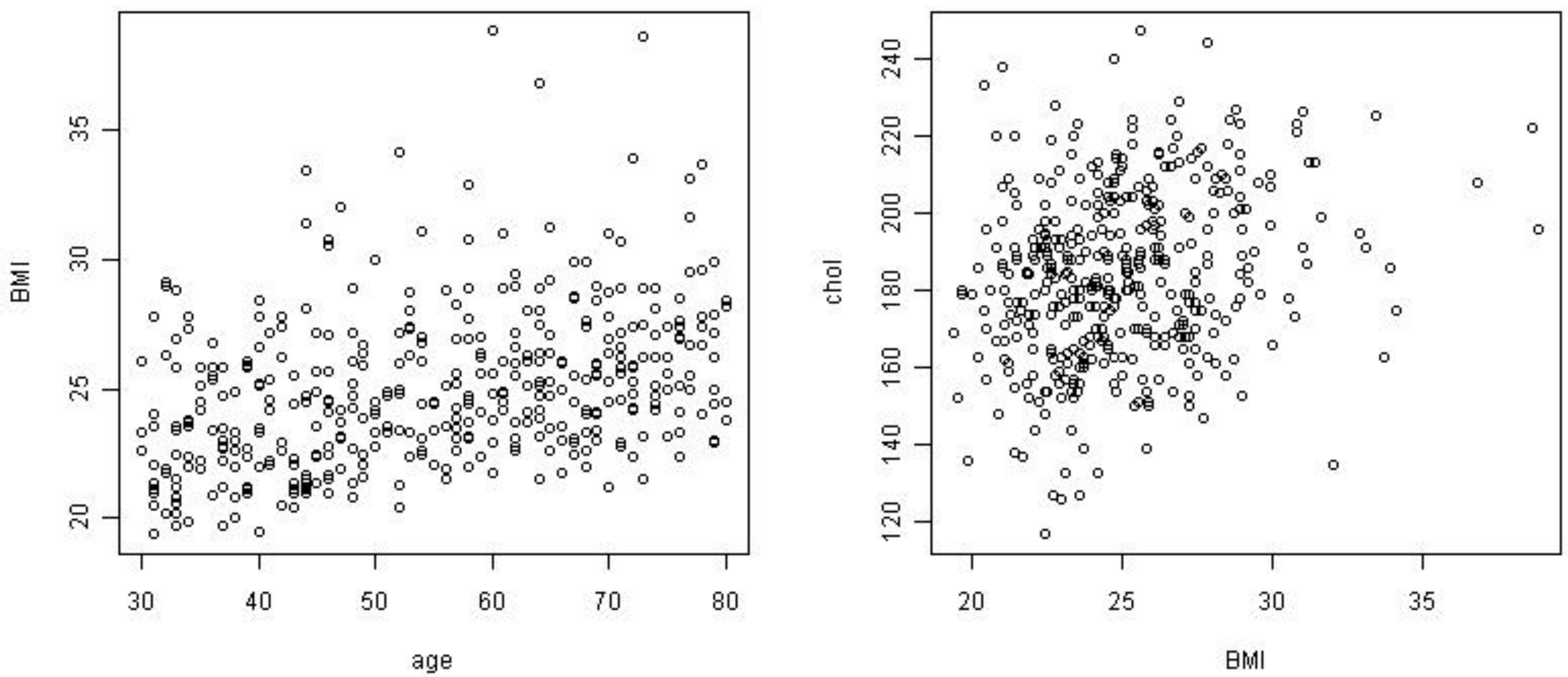
- The F-value is tested against a F-distribution with p, n-p-1 degrees of freedom
  - If we reject the null hypothesis, then the predictors do aid in predicting Y [in this analysis we do not know which ones are important!]
  - Failing to reject the null hypothesis does not mean that none of the covariates 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?

Scientific example: Modeling cholesterol using age and BMI



Scientific example: Modeling cholesterol using age and BMI

- 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!

Scientific example: Modeling cholesterol using age and BMI

```
> fit2=lm(chol~age+BMI)
> summary(fit2)
Call:
lm(formula = chol ~ age + BMI)

Residuals:
    Min      1Q  Median      3Q     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 ***
age           0.2023     0.0795   2.544 0.011327 *  
BMI           1.4266     0.3822   3.732 0.000217 ***
---
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
```

Scientific example: Modeling cholesterol using age and BMI

- Our estimated regression equation is

$$\hat{y} = 137.16 + 0.20Age + 1.43BMI$$

- Question: How do we interpret the age coefficient?

Scientific example: Modeling cholesterol using age and BMI

- 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.

Scientific example: Modeling cholesterol using age and BMI

- Our estimated regression equation is

$$\hat{y} = 137.16 + 0.20 Age + 1.43 BMI$$

- The age coefficient from our simple linear regression model was 0.31.
- Question: Why do the estimates from the two models differ?

Scientific example: Modeling cholesterol using age and 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?
- Answer: We are now conditioning on or controlling for BMI so our estimate of the age association is among subjects with the same BMI.

Scientific example: Modeling cholesterol using age and BMI

```
Call:
lm(formula = chol ~ age + BMI)

Residuals:
    Min      1Q  Median      3Q     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 ***
age           0.2023     0.0795   2.544 0.011327 *  
BMI           1.4266     0.3822   3.732 0.000217 ***
---
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?

```
> fit0=lm(chol~1)
> anova(fit0,fit2)
Analysis of Variance Table

Model 1: chol ~ 1
Model 2: chol ~ age + BMI
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1    399 195189
2    397 180842  2     14347 15.748 2.62e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

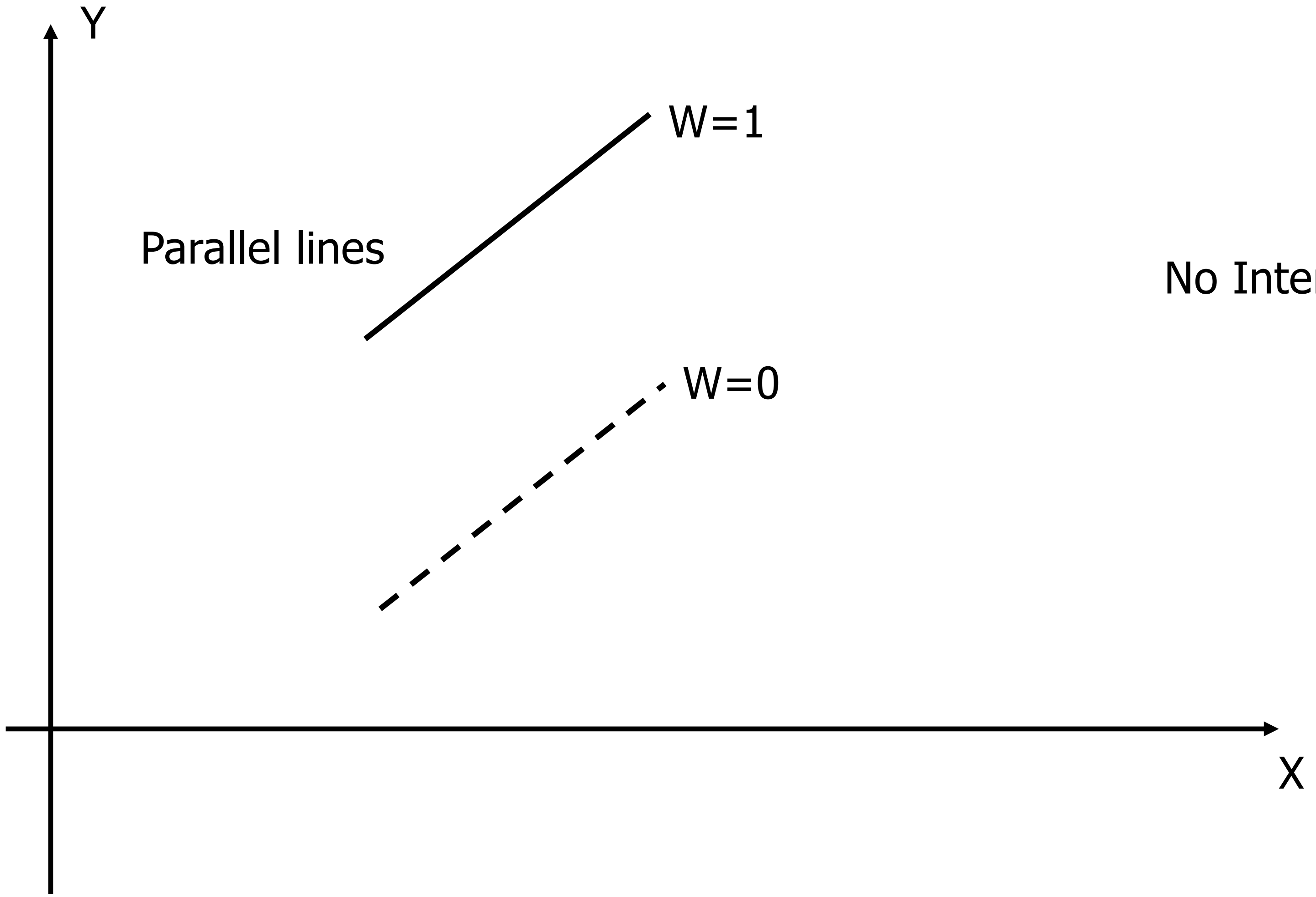
- Confounding: Estimates of association from unadjusted analysis are markedly different from estimates of association from adjusted analysis
  - Association within each stratum is similar, but different from the "crude" association in the combined data (ignoring the strata)
  - In linear regression, 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

Effect Modification /Interaction

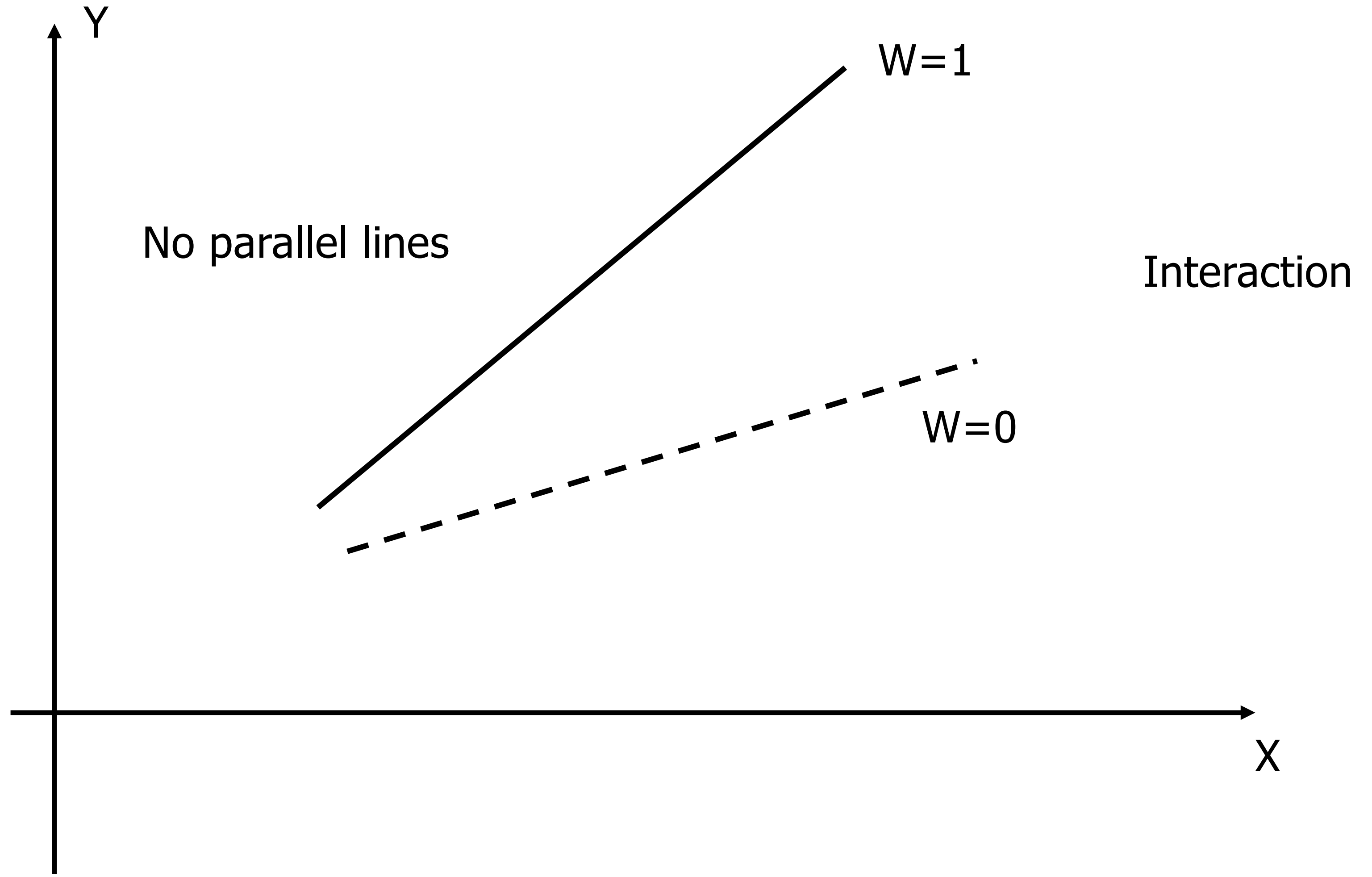
- Even if present, effect modification may not always be of interest in summarizing the effect of a predictor.
- For example, pleconaril, an antiviral drug, reduced the mean duration of symptoms in subjects with a common cold due to rhinoviruses but had no effect in subjects whose cold was due to some 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

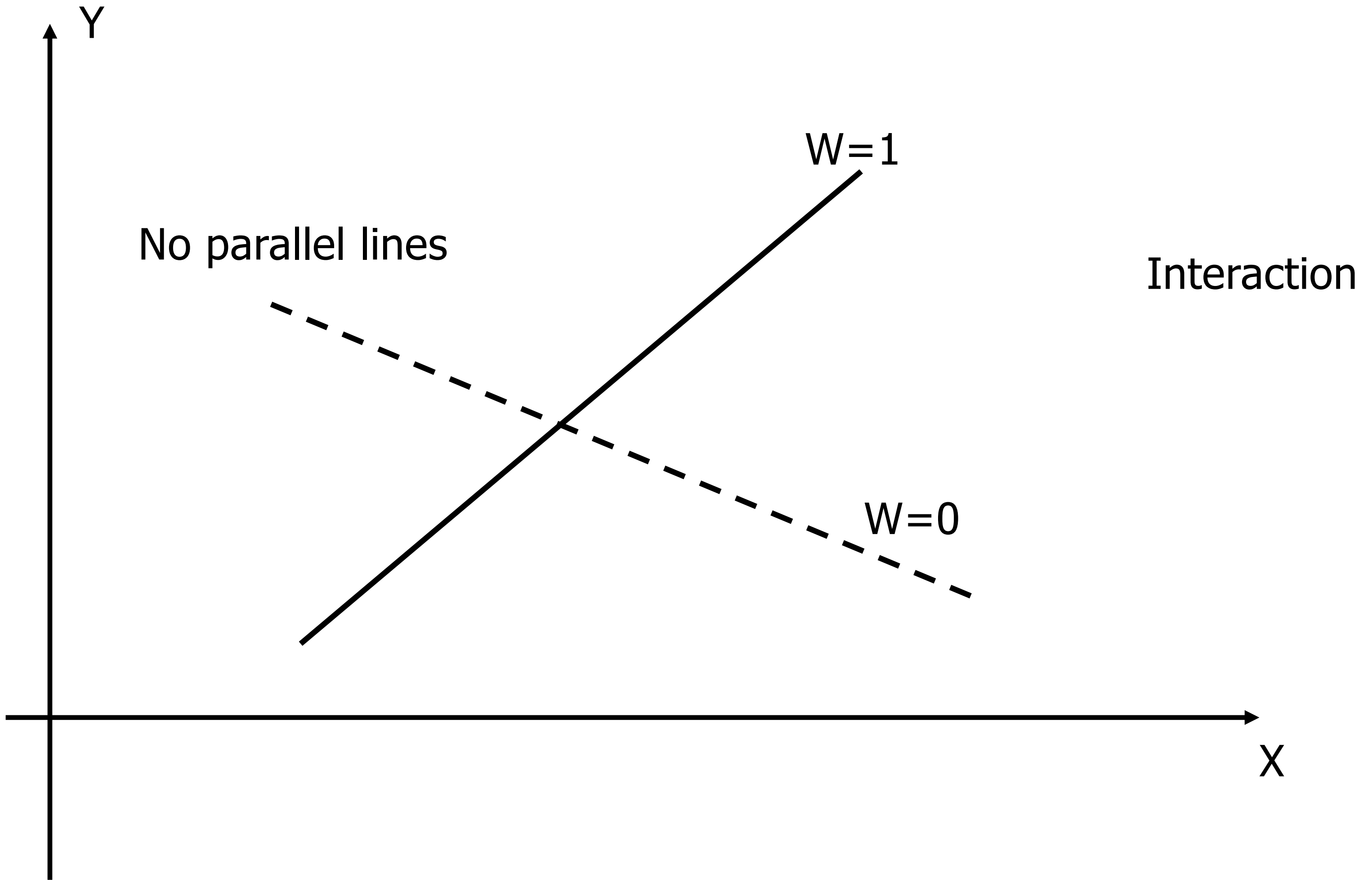


No Interaction

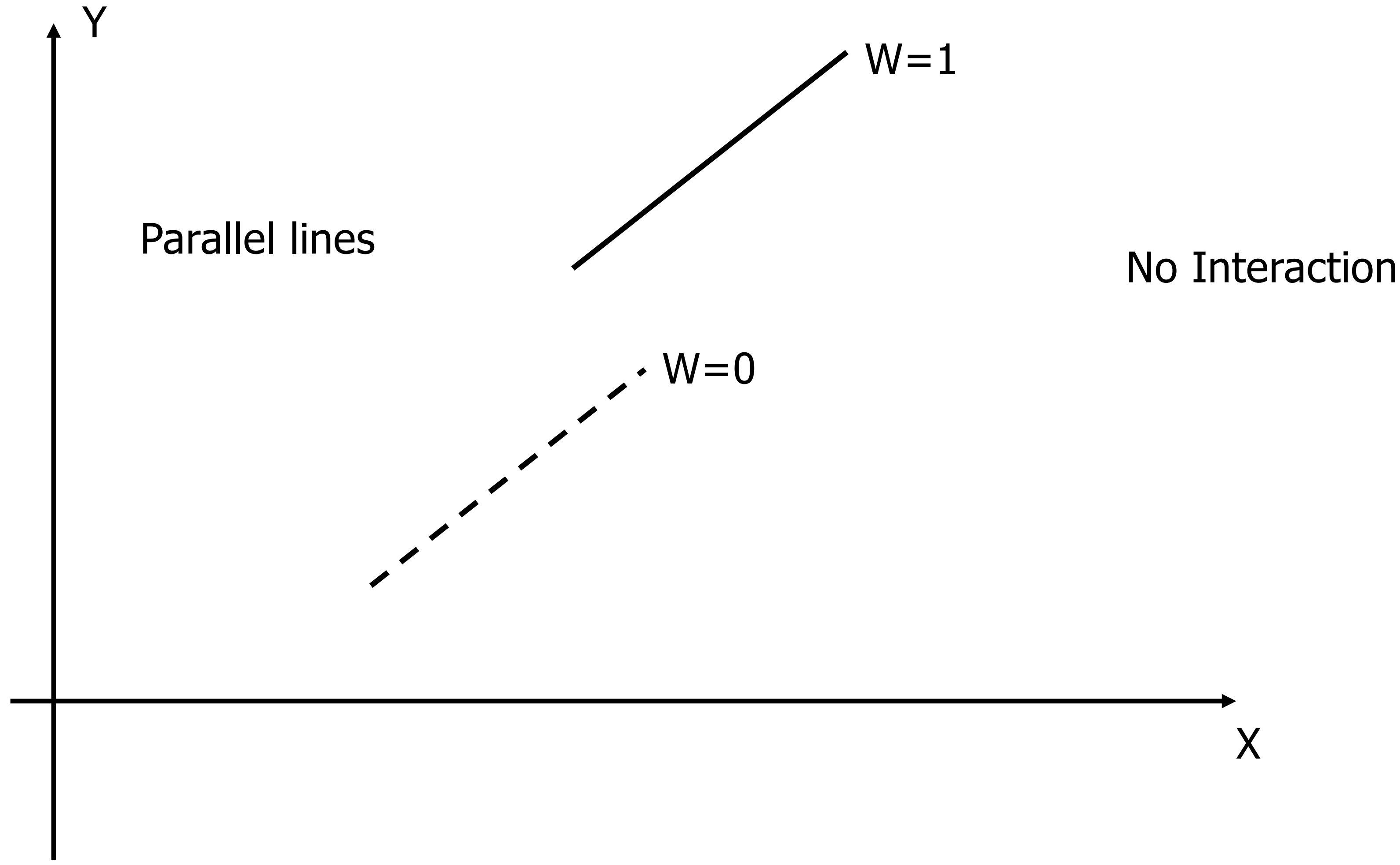
Graphical Representation



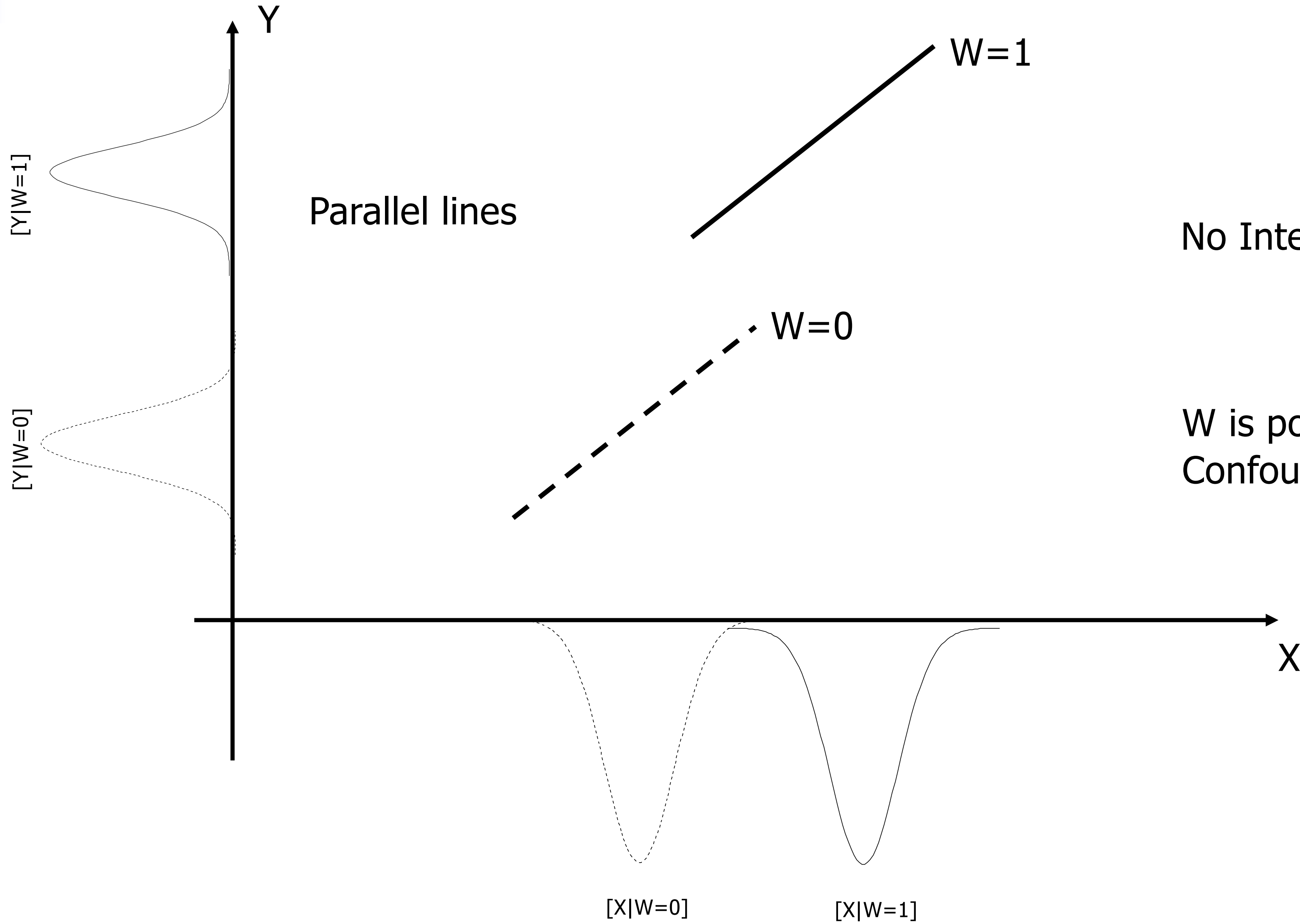
Graphical Representation



Graphical Representation



Graphical Representation



No Interaction

W is possibly a Confounder

Model and Interpretation: interaction

- Assume that there are two predictors in the model

$$E[Y|x_1, x_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \boxed{\beta_3 x_1 x_2}$$

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

That is, the difference in means depends now on the value of x_2!

Model and Interpretation: interaction

- Model: $E[Y|x_1, x_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \boxed{\beta_3 x_1 x_2}$

- Difference in Means:

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

  The difference in means depends on the value of x_2
  - The difference in means is β_1 if c=0.
  - The difference in means is $\beta_1 + \beta_3$ if c=1
  - The difference in means changes by β_3 for each unit difference in c (that is, in x_2) [that is, β_3 is the difference of differences!]

- H_0: $\beta_3=0$ tests for interaction

Model and Interpretation: interaction

- Model: $E[Y|x_1, x_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \boxed{\beta_3 x_1 x_2}$

- Another way to look at this

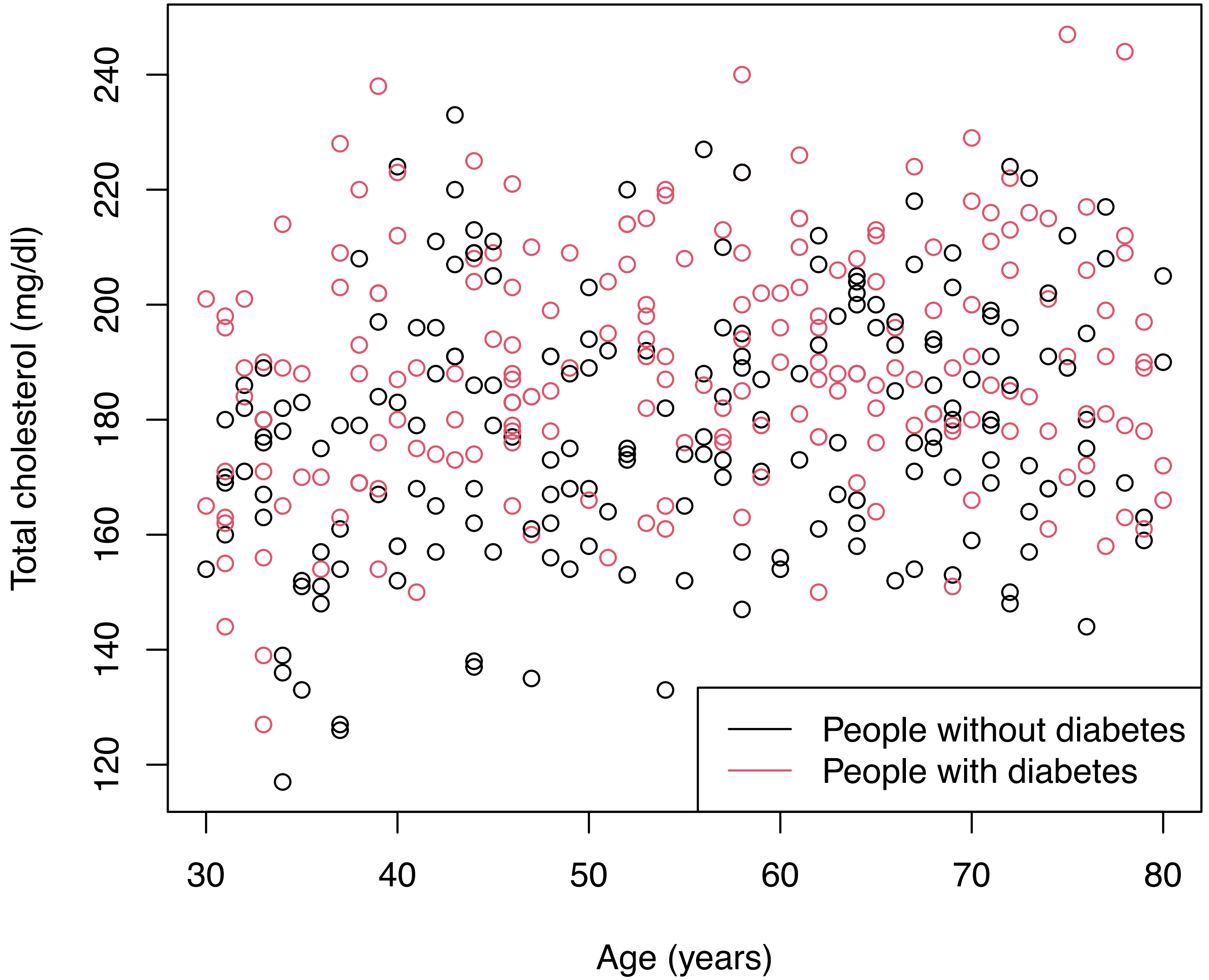
- Factor terms involving x_1:

  $E[Y|x_1, x_2] = \beta_0 + \underline{(\beta_1 + \beta_3 x_2)}x_1 + \beta_2 x_2$

Slope of x_1 changes with x_2 , 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 β_3)

Cholesterol Example: Does diabetes affect the age – cholesterol relationship?

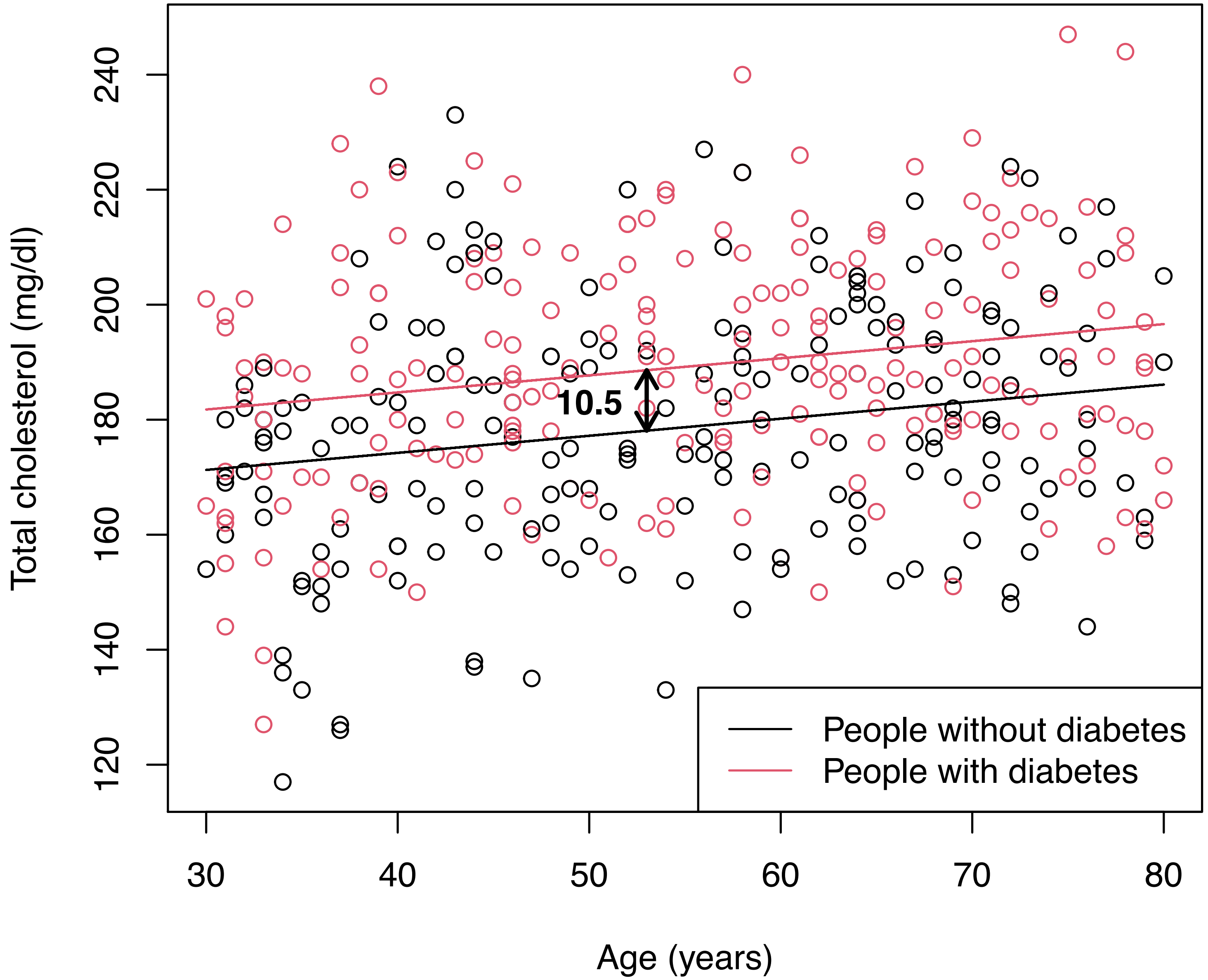


Cholesterol Example: Does diabetes affect the age – cholesterol relationship?

We first fit the model with age and DM terms only
(No diabetes: DM=0, With diabetes: DM=1)

```
> fit3 = lm(chol ~ age+DM)
> summary(fit3)

Call:
lm(formula = chol ~ age + DM)

Residuals:
    Min      1Q  Median      3Q     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 ***
age           0.29697    0.07313   4.061 5.89e-05 ***
DM           10.50728    2.10794   4.985 9.29e-07 ***
---
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
```

Cholesterol Example: Does diabetes affect the age – cholesterol relationship?



Cholesterol Example: Does diabetes affect the age – cholesterol relationship?

- This model indicates that, after controlling for the effect of diabetes, 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 people with and without diabetes
- To answer this question we must add the interaction term

Cholesterol Example: Does diabetes affect the age – cholesterol relationship?

Model with age and DM main effects, plus interaction effect

```
> fit4=lm(chol~age*DM)
> summary(fit4)
Call:
lm(formula = chol ~ age * DM)

Residuals:
    Min       1Q   Median       3Q      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 **
DM           14.56271    8.29802   1.755  0.08004 .
age:DM       -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
```

Cholesterol Example: Does diabetes affect the age – cholesterol relationship?

```
Call:
lm(formula = chol ~ age * DM)

Residuals:
    Min      1Q  Median      3Q     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 ** 
DM           14.56271    8.29802   1.755  0.08004 .  
age:DM       -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
```

Mean cholesterol for people without diabetes (DM = 0) at age 0

Cholesterol Example: Does diabetes affect the age – cholesterol relationship?

```
Call:
lm(formula = chol ~ age * DM)

Residuals:
    Min      1Q  Median      3Q     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 ** 
DM           14.56271    8.29802   1.755  0.08004 .  
age:DM       -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 people with and without diabetes at age 0

Cholesterol Example: Does diabetes affect the age – cholesterol relationship?

```
Call:
lm(formula = chol ~ age * DM)

Residuals:
    Min      1Q  Median      3Q     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 **
DM           14.56271    8.29802   1.755  0.08004 .
age:DM       -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 people without diabetes

Cholesterol Example: Does diabetes affect the age – cholesterol relationship?

```
Call:
lm(formula = chol ~ age * DM)

Residuals:
    Min      1Q  Median      3Q     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 ** 
DM           14.56271    8.29802   1.755  0.08004 .  
age:DM       -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 comparing people with and without diabetes

Cholesterol Example: Does diabetes affect the age – cholesterol relationship?

- Interpretation?
  - Estimated model:

$$160.3 + 0.33\ \text{Age} + 14.56\ \text{Diabetes} - 0.07\ \text{Age} \times \text{Diabetes}$$

Subject 1: Age = a+1, diabetes = b

Subject 2: Age = a, diabetes = b

Difference in the estimated cholesterol:

$$[160.3 + 0.33(a+1) + 14.56(b) - 0.07\,(a+1)(b)] - [160.3 + 0.33(a) + 14.56\,(b) - 0.07\,(a)(b)] = 0.33 - 0.07b$$

  - Diabetes exerts a small (not statistically significant) effect on the age/cholesterol relationship

In people without diabetes: 160.3+0.33 Age

In people with diabetes : 174.9+0.26 Age

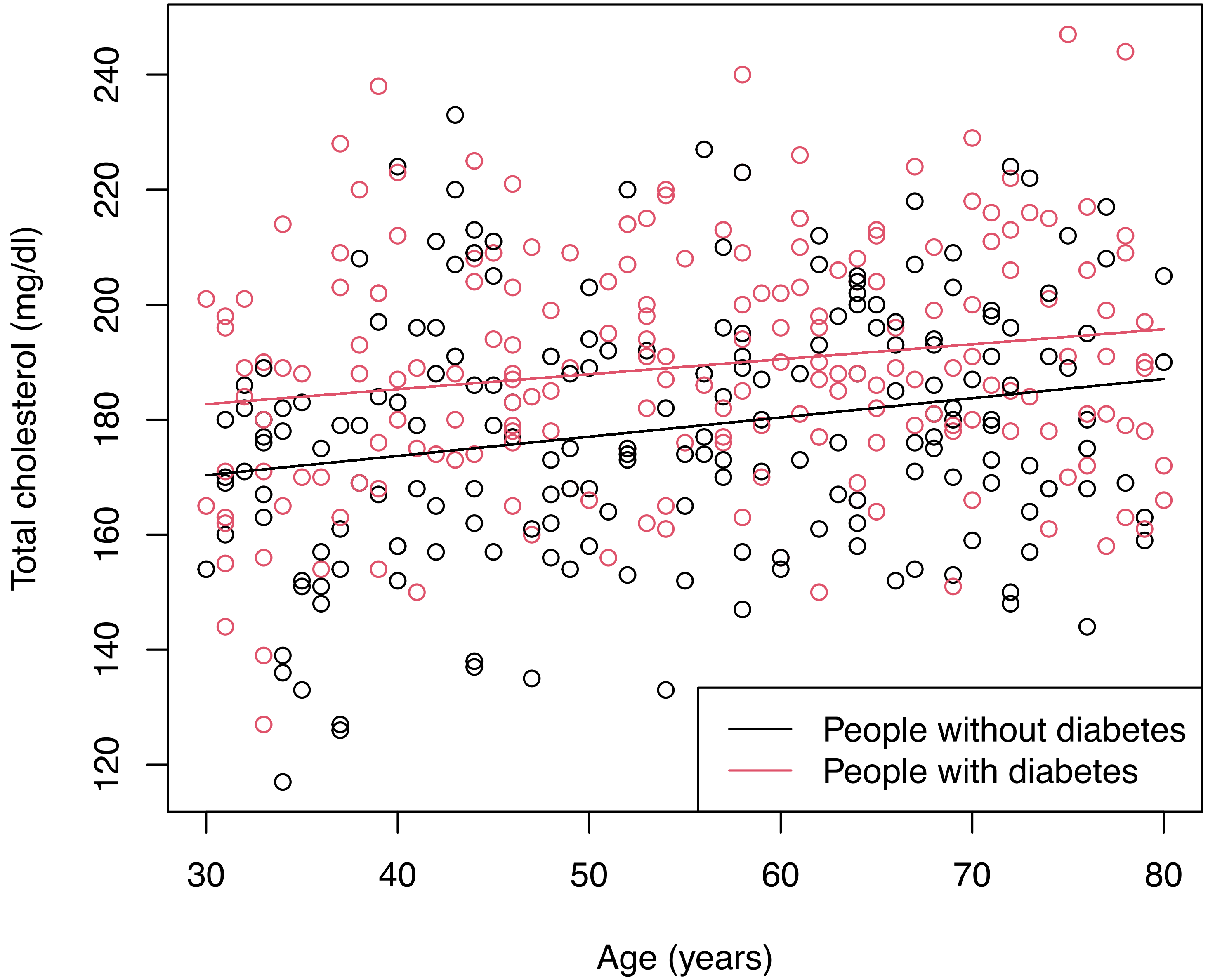
Cholesterol Example: Does diabetes affect the age – cholesterol relationship?

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

```
> anova(fit3,fit4)
Analysis of Variance Table

Model 1: chol ~ age + DM
Model 2: chol ~ age * DM
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    397 176162
2    396 176049  1    113.52 0.2554 0.6136
```

- Adding the interaction term did not significantly improve model fit

Cholesterol Example: Does diabetes affect the age – cholesterol relationship?



Summary

We have considered:

- Simple linear regression
  - Interpretation
  - Estimation
  - Model checking

- Multiple linear regression
  - Confounding
  - Interpretation
  - Estimation
  - Interaction

Exercise

- Work on Exercise 7-8
  - Try each exercise on your own
  - Make note of any questions or difficulties you have
  - At **1:15PT** we will meet as a group to go over the solutions and discuss your questions