SCHOOL OF PUBLIC HEALTH · UNIVERSITY of WASHINGTON
Department of Biostatistics

REGRESSION AND ANALYSIS OF VARIANCE

Motivation

- Objective: Investigate associations between two or more variables
- What tools do you already have?
  - T-test
    - Comparison of means 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
    - Comparison of a continuous outcome over a fixed number of groups

REGRESSION 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 versus Influential observations

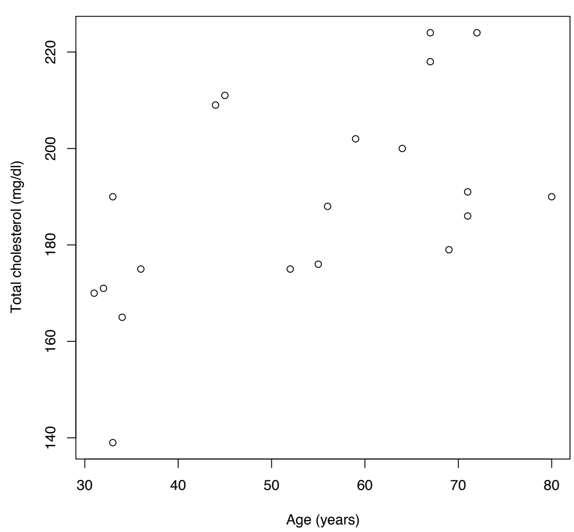
Motivation: Cholesterol Example

- Data: Factors affecting serum total cholesterol

	sex	**age**	**chol**	BMI	TG	APOE	rs174548	rs4775401
1	1	74	215	26.2	367	4	1	2
2	1	51	204	24.7	150	4	2	1
3	0	64	205	24.2	213	4	0	1
4	0	34	182	23.8	111	1	1	1
5	1	52	175	34.1	328	1	0	0
6	1	39	176	22.7	53	4	0	2

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

Motivation: Cholesterol Example

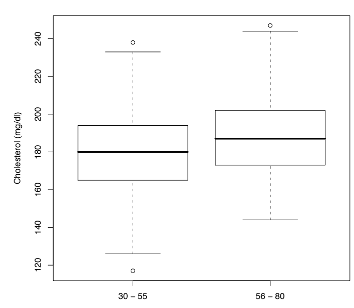


Motivation: Cholesterol Example

- Is serum cholesterol associated with age?
  - You could dichotomize age and compare the mean cholesterol between two groups: t-test

Motivation: Cholesterol Example

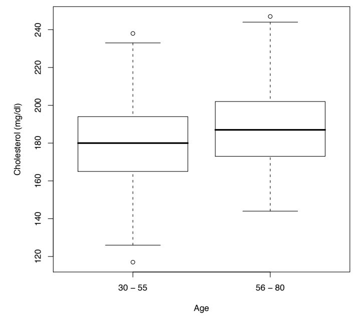
- Is cholesterol associated with age?
  - You could dichotomize age and compare the mean systolic between two groups: t-test




```
> group = 1*(age > 55)
> 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 means is not equal to 0
95 percent confidence interval:
 -12.200209  -3.638487
sample estimates:
mean in group 0 mean in group 1
       179.9751        187.8945
```

Motivation: Cholesterol Example

- Question: What does this plot and t-test tell us about the relationship between age and cholesterol?



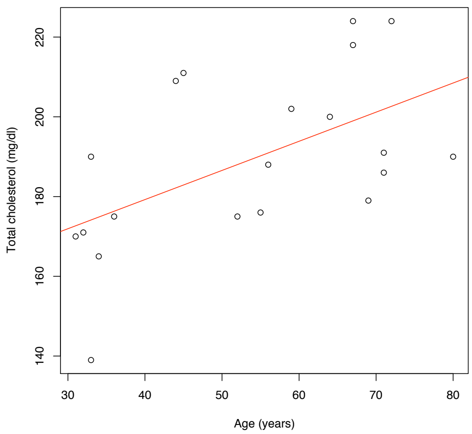

```
> group = 1*(age > 55)
> t.test(chol ~ group)

        Welch Two Sample t-test

data:  chol by group
t = -3.637, df = 393.477, p-value = 0.0003125
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -12.200209  -3.638487
sample estimates:
mean in group 0 mean in group 1
       179.9751        187.8945
```


Motivation: Cholesterol Example

- Using 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?

Motivation: Cholesterol Example

- Can we find the equation for a straight line



that best fits these data?

Linear Regression

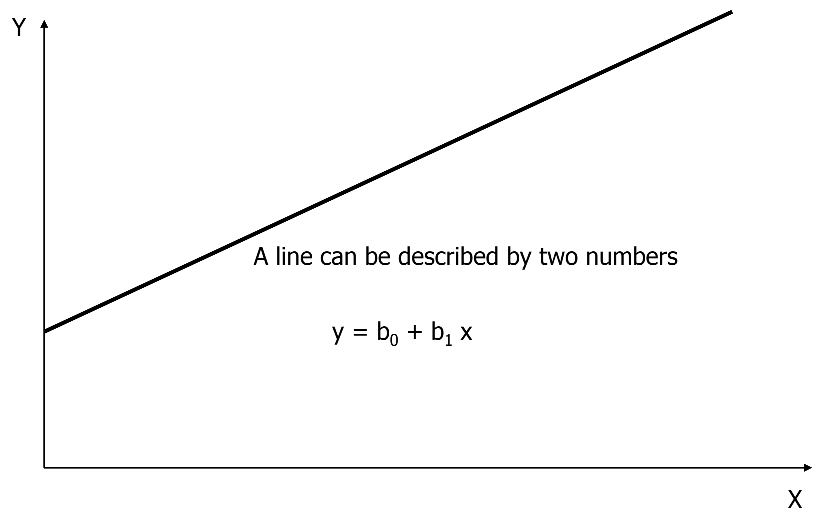
- Statistical method for modeling the relationship between a continuous variable [response/outcome/dependent] and other variables [predictors/exposure/independent]
  - Most commonly used statistical model
  - Flexible
  - Well-developed and understood properties
  - Easy interpretation
  - Building block for more general models

- Goals of analysis:
  - Study 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

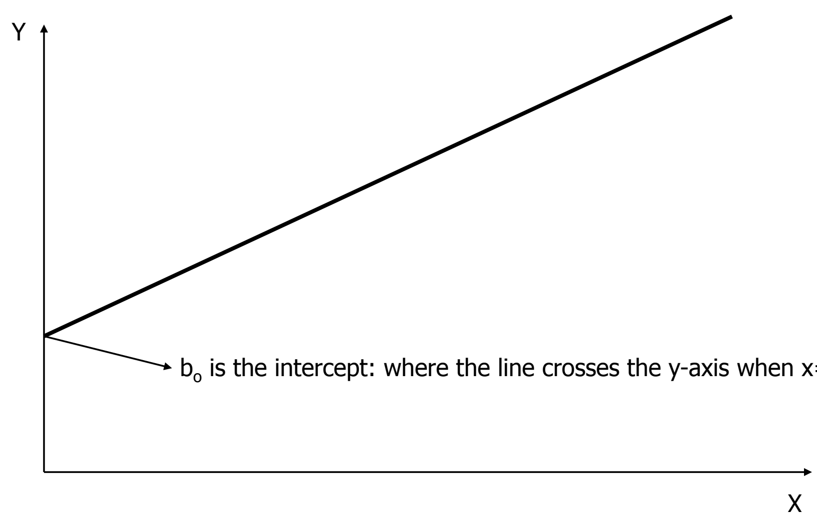
The straight line equation



A line can be described by two numbers

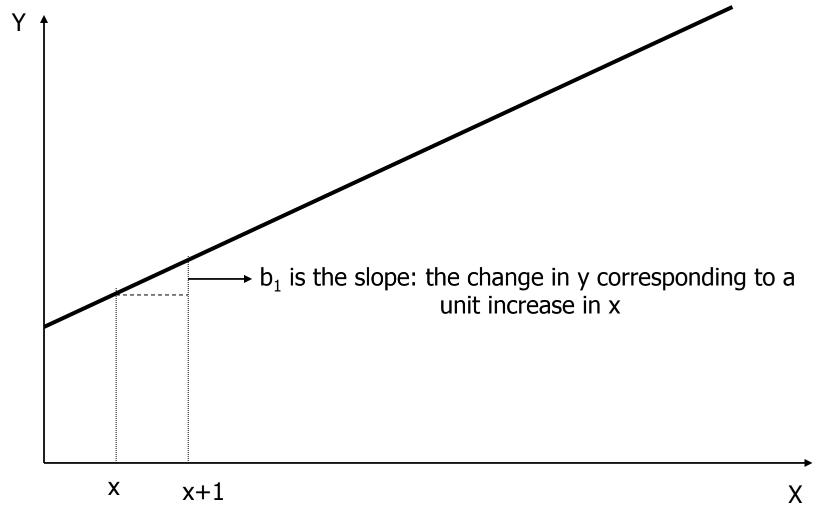
$$y = b_0 + b_1 x$$

The straight line equation



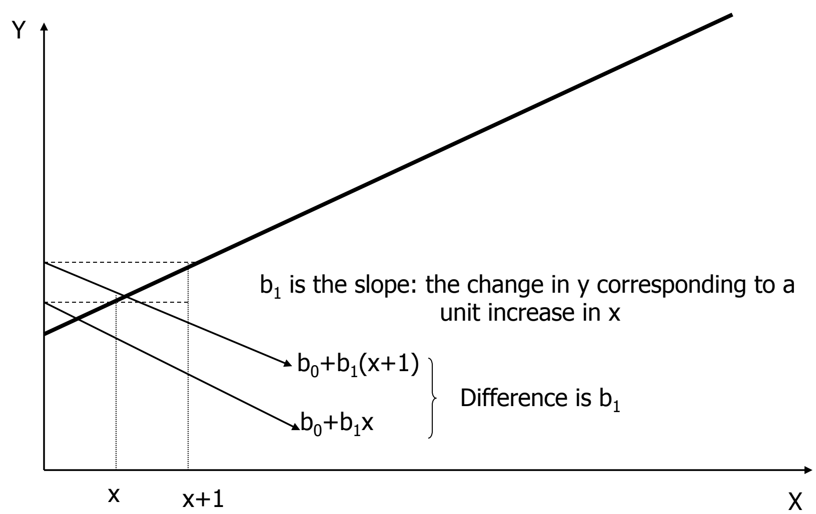
b_o is the intercept: where the line crosses the y-axis when x=0

The straight line equation



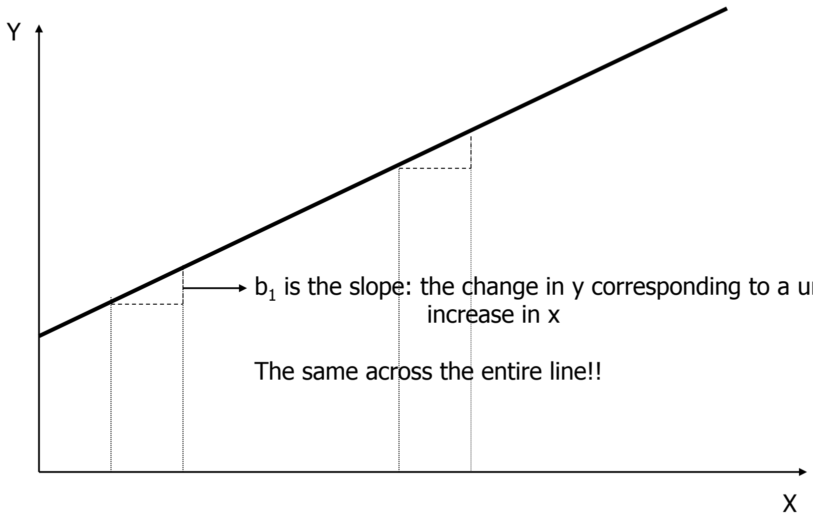
15

The straight line equation



16

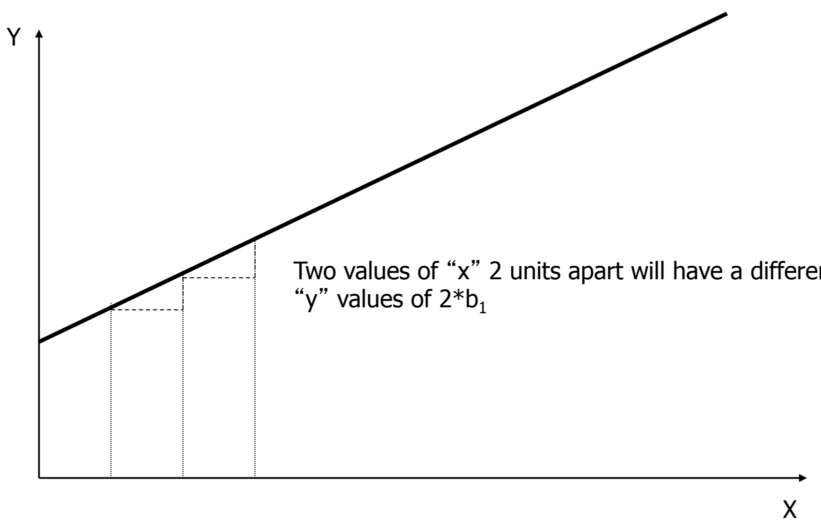
The straight line equation



b_1 is the slope: the change in y corresponding to a unit increase in x

The same across the entire line!!

The straight line equation

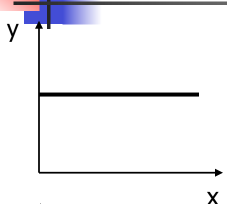


Two values of “x” 2 units apart will have a difference in “y” values of $2*b_1$

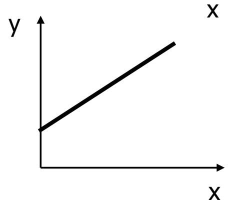
The straight line equation

- Slope b_1 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

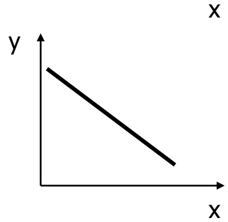
The straight line equation



(b_1=0) No association between x and y
(values of y are the same regardless of x)



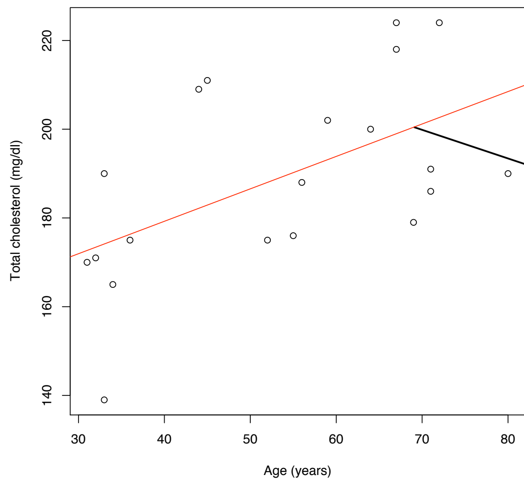
(b_1 > 0) Positive association between x and y
(values of y increase as values of x increase)



(b_1 < 0) Negative association between x and y
(values of y decrease as values of x increase)

Simple Linear Regression

- Dealing with situations where points don't fit exactly to the straight line



We estimate a straight line describing trends in the **mean** of an outcome Y as a function of predictor X

Simple Linear Regression

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

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

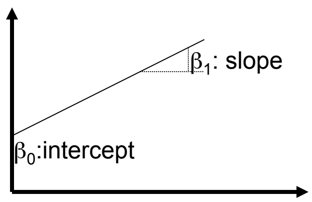
Simple Linear Regression

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

Model consists of two components:

- Systematic component: $E[Y \mid X = x] = \beta_0 + \beta_1 x$

  Mean population value of Y at X=x



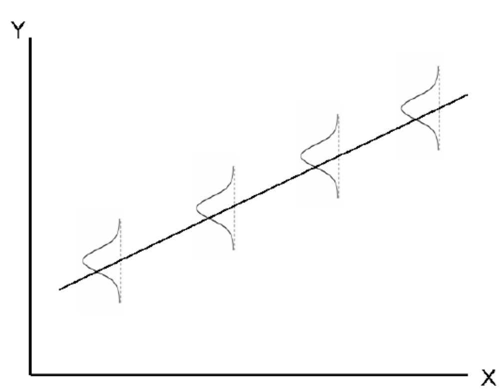
- Random component: $Var[Y \mid X = x] = \sigma^2$

  Variance does not depend on x

Simple Linear Regression: Assumptions

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

Distribution of Y at different x values:



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 β_0?
- Answer:
  - $\beta_0 = E[Y|x=0]$, that is, the mean response when x=0

Your turn: interpret β_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 β_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. β_1 is the difference in the mean response associated with a one unit positive difference in 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 β_1?
- Answer:

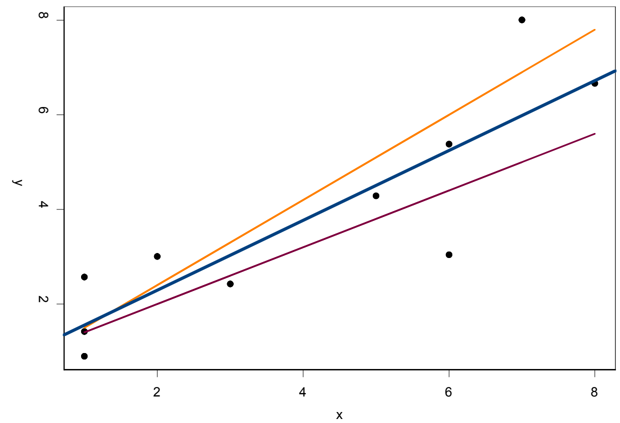
Example: Cholesterol and age

- 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 β_1?
- Answer:

  β_1 is the difference in mean serum cholesterol associated with a one year increase in age

Least Squares Estimation

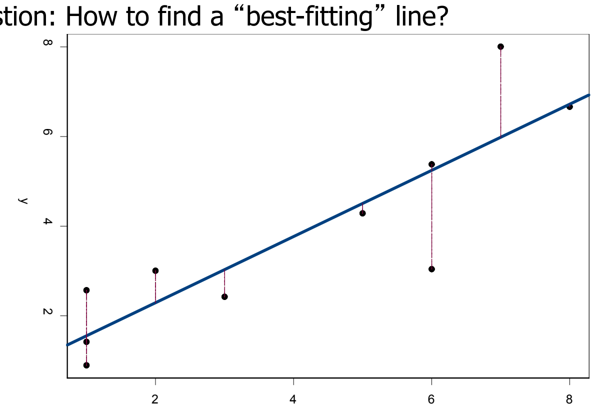
- Question: How to find a "best-fitting" line?



29

Least Squares Estimation

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



- Method: Least Squares Estimation
  - Idea: minimizes the sum of squares of the vertical distances from the observed points to the least squares regression line.

30

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 β_0 and β_1 by finding the values that minimize D

Least Squares Estimation

- These values are:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

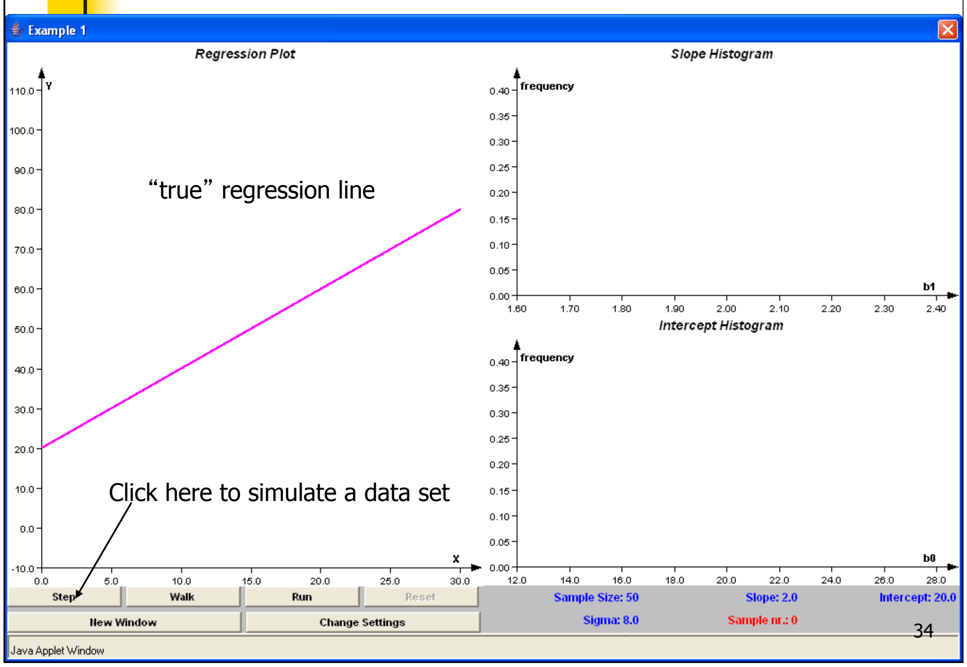
- We estimate the variance as

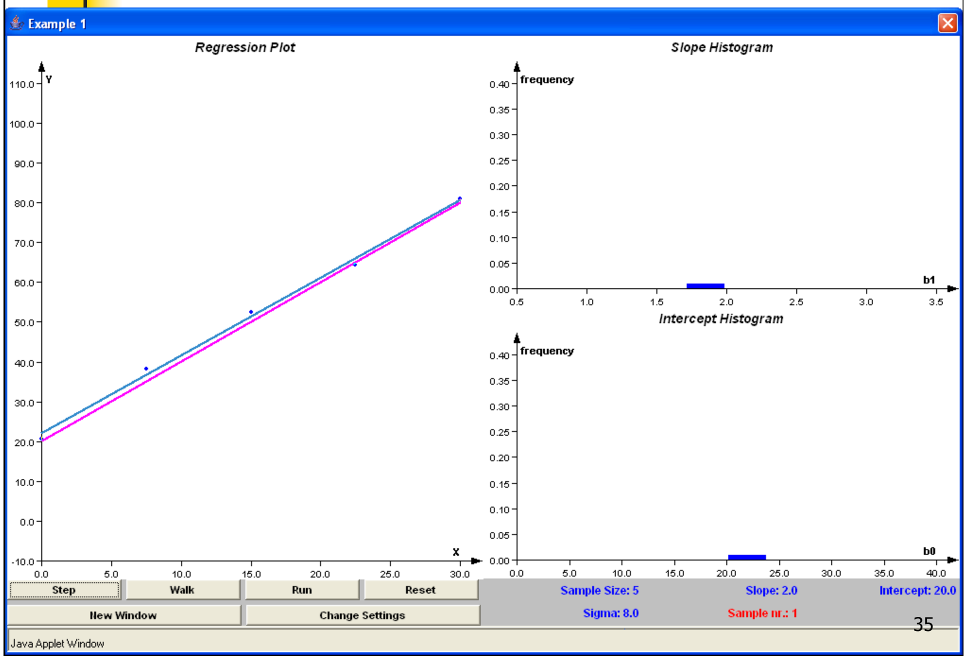
$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} r_i^2}{n-2} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n-2}$$

Estimated Standard Errors

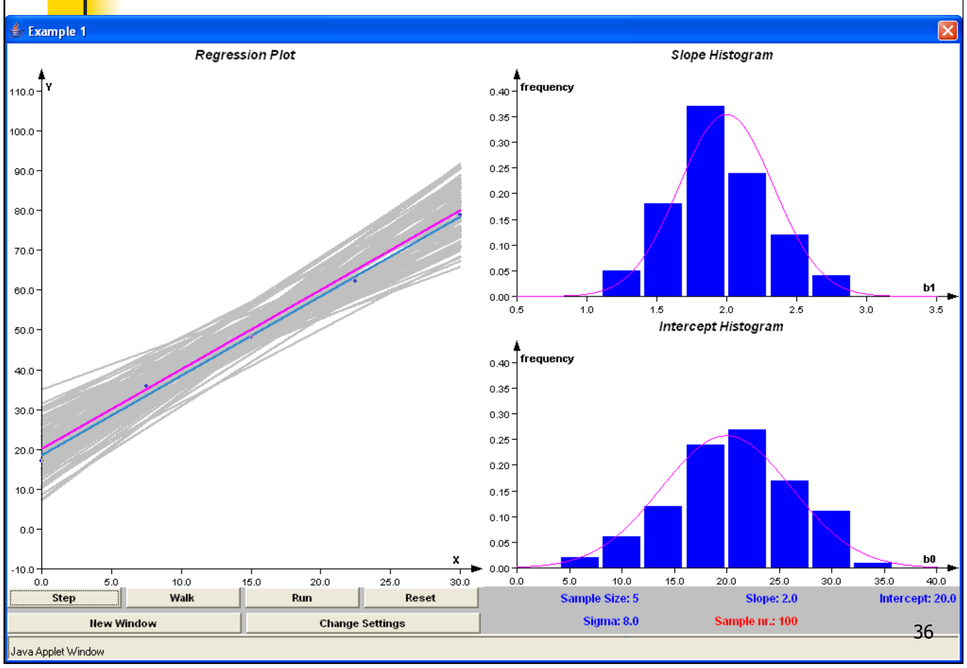
- Recall that when estimating parameters, sampling variability exists in our estimates
- Same is true for regression parameter estimates
- Looking at the formulas for $\hat{\beta}_0$ and $\hat{\beta}_1$, we can see that these are just complicated means
- In repeated sampling we would get different estimates
- Knowledge of sampling distribution of parameter estimates can help us make inference about the line

http://lstat.kuleuven.be/java/version2.0/Applet003.html

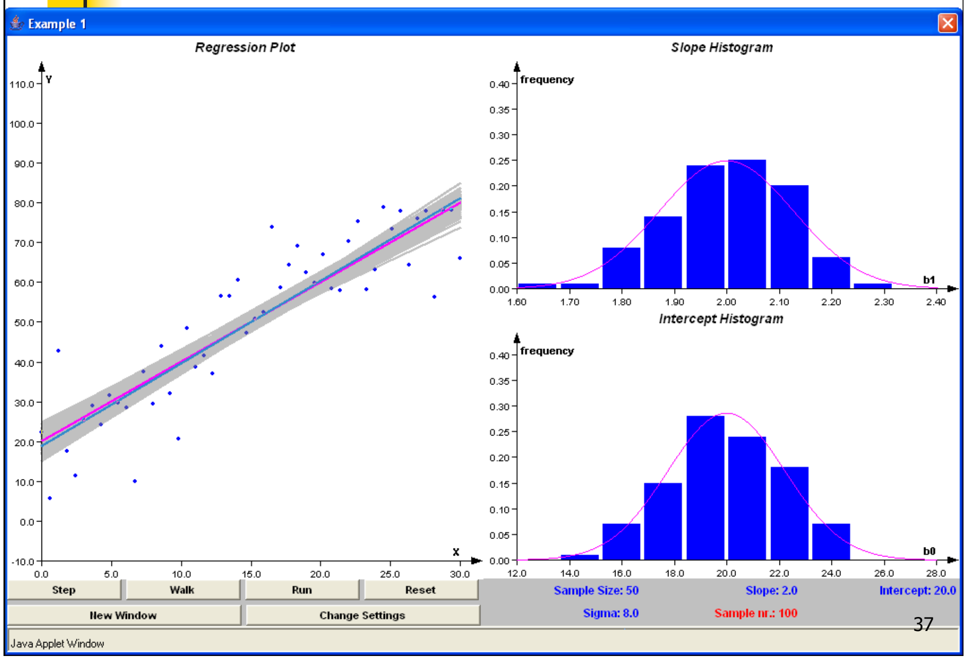


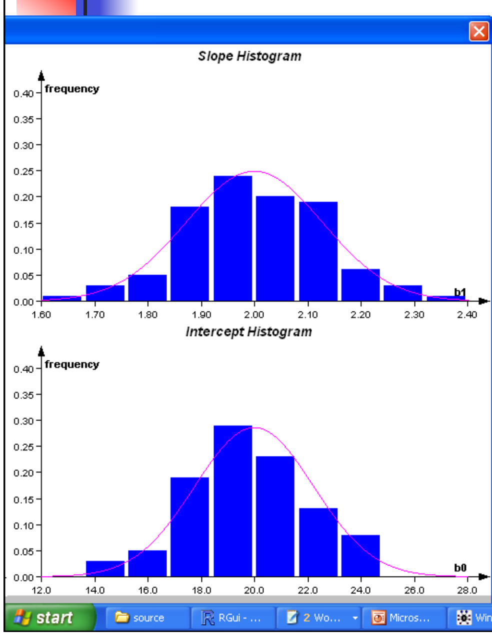


Sampling Distribution



Sampling Distribution





- From statistical theory

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}\right)\right)$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \sigma^2\left(\frac{1}{(n-1)s_x^2}\right)\right)$$

(in the figure, the theoretical distribution is shown in pink)

Estimated Standard Errors

- Estimate the variability of $\hat{\beta}_0,\ \hat{\beta}_1$ in repeated sampling

$$SE(\hat{\beta}_0)=\hat{\sigma}\sqrt{\frac{1}{n}+\frac{\overline{x}^2}{(n-1)s_x^2}}$$

$$SE(\hat{\beta}_1)=\hat{\sigma}\sqrt{\frac{1}{(n-1)s_x^2}}$$

Inference

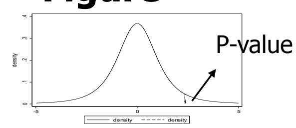
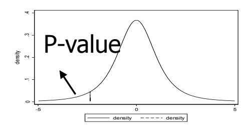
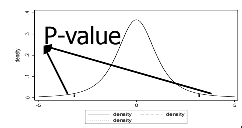
- About regression model parameters
  - Hypothesis testing: H_0: $\beta_j=0$
    - Test Statistic:
      - Large Samples: $\dfrac{\hat{\beta}_j-(null\ \ hyp)}{se(\hat{\beta}_j)} \sim N(0,1)$
      - Small Samples: $\dfrac{\hat{\beta}_j-(null\ \ hyp)}{se(\hat{\beta}_j)} \sim T_{n-2}$
  - Confidence Intervals:

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

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

Inference: Hypothesis Testing

Null Hypothesis: $\beta_j = 0$

Alternative	**P-Value**	**Figure**
$\beta_j > 0$	$P(T_{n-2} > T)$	
$\beta_j < 0$	$P(T_{n-2} < T)$	
$\beta_j \neq 0$	$2P(T_{n-2} > \lvert T \rvert)$	

Inference: Confidence Intervals

100 (1-α)% Confidence Interval for β_j (j=0,1)

$$\hat{\beta}_j \pm t_{n-2,\alpha/2} SE(\hat{\beta}_j)$$

Gives intervals that (1- α)100% of the time will cover the true parameter value (β_0 or β_1).

We say we are "(1- α)100% confident" the interval covers β_j.

Example:
Scientific Question: Is cholesterol associated with age?

```
> fit = lm(chol ~ age)
> summary(fit)

Call:
lm(formula = chol ~ age)

Residuals:
      Min        1Q    Median        3Q       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 ***
age           0.31033    0.07524   4.125 4.52e-05 ***
---
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       0.1624211    0.4582481
```


Example:
Scientific Question: Is cholesterol associated with age?

```
> fit = lm(chol ~ age)
> summary(fit)

Call:
lm(formula = chol ~ age)

Residuals:
      Min        1Q    Median        3Q       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 ***
age           0.31033    0.07524   4.125 4.52e-05 ***
---
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
```

Estimates of the model parameters and standard errors

$\hat{\beta}_0 = 166.90$; $se(\hat{\beta}_0) = 4.26$

$\hat{\beta}_1 = 0.31$; $se(\hat{\beta}_1) = 0.08$

```
> confint(fit)
                  2.5 %      97.5 %
(Intercept) 158.5171656 175.2861949
age             0.1624211    0.4582481
```

Example:
Scientific Question: Is cholesterol associated with age?

```
> fit = lm(chol ~ age)
> summary(fit)

Call:
lm(formula = chol ~ age)

Residuals:
      Min        1Q    Median        3Q       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 ***
age           0.31033    0.07524   4.125 4.52e-05 ***
---
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
```

95% Confidence intervals

```
> confint(fit)
                  2.5 %      97.5 %
(Intercept) 158.5171656 175.2861949
age           0.1624211   0.4582481
```


Example:
Scientific Question: Is cholesterol associated with age?

- 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: $\hat{\beta}_0$: The estimated average serum cholesterol for someone of age = 0 is 166.9
  - Your turn: What about $\hat{\beta}_1$?

Example:
Scientific Question: Is cholesterol associated with age?

- What do these models results mean in terms of our scientific question?
  - Parameter estimates and confidence intervals:

$$\hat{\beta}_0 = 166.90 \quad \text{95\% CI: (158.5, 175.3)}$$

$$\hat{\beta}_1 = 0.31 \quad \text{95\% CI: (0.16, 0.46)}$$

  - Answer: $\hat{\beta}_1$: mean cholesterol is estimated to differ by 0.31 mg/dl for each one year difference in age.
  - Question: What about the confidence intervals?

Example:
Scientific Question: Is cholesterol associated with age?

- What do these models results mean in terms of our scientific question?
  - Parameter estimates and confidence intervals:

$$\hat{\beta}_0 = 166.90 \quad \text{95\% CI: (158.5, 175.3)}$$

$$\hat{\beta}_1 = 0.31 \quad \text{95\% CI: (0.16, 0.46)}$$

  - 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

Example:
Scientific Question: Is cholesterol associated with age?

- Presentation of the results?
  - The mean serum total cholesterol is significantly higher in older individuals ($p < 0.001$). 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).
    - Note:
      - Emphasis on slope parameter (sign and magnitude)
      - Confidence interval
      - Units for predictor and response

Inference for predictions

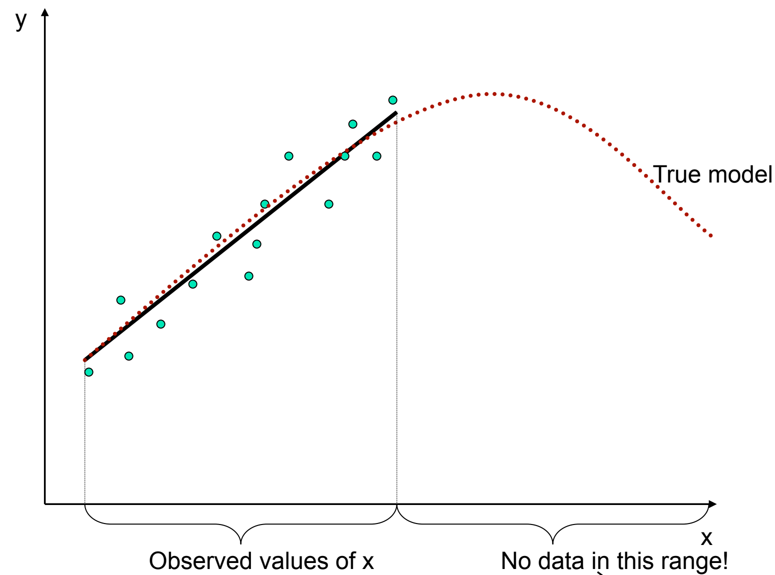
- Given estimates $\hat{\beta}_0$, $\hat{\beta}_1$ we can find the **predicted value**, $\hat{y}_i$ for any value of x_i as

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

- Interpretation of $\hat{y}_i$:
  - Estimated mean value of Y at $X = x_i$.

Be Cautious: It 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!!



Prediction

- Prediction of the mean E[Y|X=x]:
  - Point Estimate: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
  - Standard Error: $se(\hat{y}) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$

    Note that as x diverges from $\bar{x}$, variance increases!

  - 100 (1-α)% confidence interval for E[Y|X=x]:

    $$\hat{y} \pm t_{n-2,1-\alpha/2}\, se(\hat{y})$$

Prediction

- Prediction of a new future observation, y*, at X=x:
  - Point Estimate: $\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x$
  - Standard Error: $se(\hat{y}^*) = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$
  - 100 (1-α)% 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 the mean

```
> predict.lm(fit, newdata=data.frame(age=c(46,47,48)), interval="confidence")
       fit      lwr      upr
1 181.1771 178.6776 183.6765
2 181.4874 179.0619 183.9129
3 181.7977 179.4392 184.1563

> predict.lm(fit, newdata=data.frame(age=c(46,47,48)), interval="prediction")
       fit      lwr      upr
1 181.1771 138.4687 223.8854
2 181.4874 138.7833 224.1915
3 181.7977 139.0974 224.4981
```

Prediction of a new observation

Example:
Scientific Question: Is cholesterol associated with age?

- Let's interpret these predictions
  - For $x = 46$

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

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

  - Question: How do our interpretations for $\hat{y}$ and $\hat{y}^*$ differ?

Example:
Scientific Question: Is cholesterol associated with age?

- Let's interpret these predictions
  - For $x = 46$

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

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

  - 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 ($\hat{y}$) and for a single new individual of age 46 ($\hat{y}^*$)

Example:
Scientific Question: Is cholesterol associated with age?

- Let's interpret these predictions
  - For $x = 46$

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

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

  - Question: Why are the confidence intervals for $\hat{y}$ and $\hat{y}^*$ of differing widths?

Example:
Scientific Question: Is cholesterol associated with age?

- Let's interpret these predictions
  - For $x = 46$

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

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

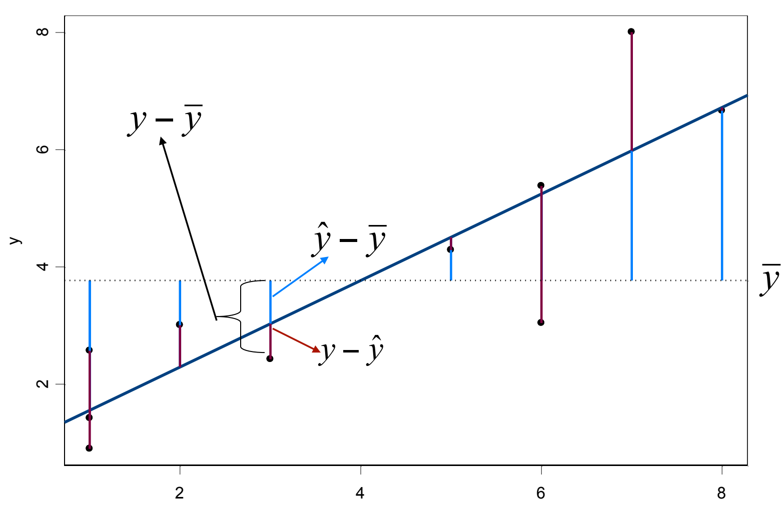
  - Question: Why are the confidence intervals for $\hat{y}$ and $\hat{y}^*$ of differing widths?
  - Answer: The interval is broader when we make a prediction for a single individual because it must incorporate random variability around the mean.

Simple Linear Regression: R^2

- 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.
- 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: $y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$

It can be shown that:

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

$$SST \quad = \quad SSE \quad + \quad SSR$$

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*.

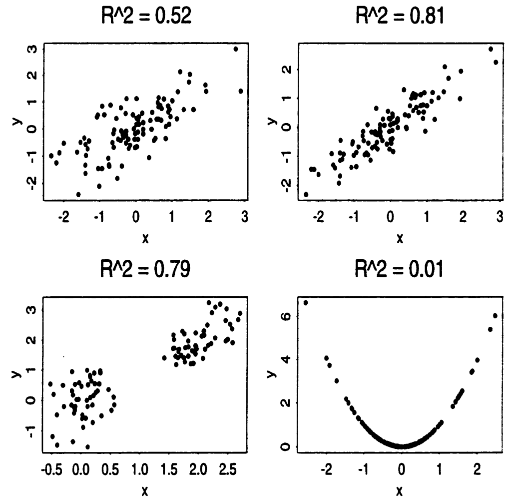
Simple Linear Regression: R^2

- 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.

Examples of R^2



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:
      Min        1Q    Median        3Q       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 ***
age           0.31033    0.07524   4.125 4.52e-05 ***
---
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           0.1624211   0.4582481
```

Cholesterol Example:

Scientific Question: Can we predict cholesterol based on age?

- $R^2=0.04$
- What does R^2 tell us about our model for cholesterol?

Cholesterol Example:

Scientific Question: Can we predict cholesterol based on age?

- $R^2=0.04$
- What does R^2 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

Degrees of freedom
Decomposition of the Sum of Squares
Mean Squares: SS/df
F-statistic: MSR/MSE

```
> anova(fit)
Analysis of Variance Table

Response: chol
              Df Sum Sq Mean Sq F value    Pr(>F)
SSR=age        1   8002  8001.7  17.013 4.522e-05 ***
SSE=Residuals 398 187187   470.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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_i) between the observed response and the predicted response, that is,

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

The residual captures the component of the measurement y_i that cannot be "explained" by x_i.

Model Checking: Residuals

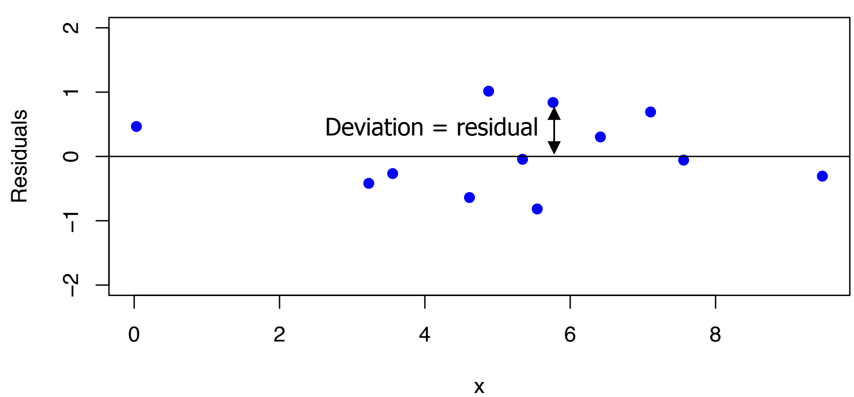
- Residuals can be used to
  - Identify poorly fit data points
  - Identify unequal variance (heteroscedasticity)
  - Identify nonlinear relationships
  - Identify additional variables
  - Examine normality assumption

Model Checking: Residuals

Linearity	Residual vs X or vs $\hat{Y}$ Q: Is there any trend?
Independence	 Q: Any scientific concerns?
Normality	Residual histogram or qq-plot Q: Symmetric? Normal?
Equal variance	Residual vs X Q: Is there any pattern?

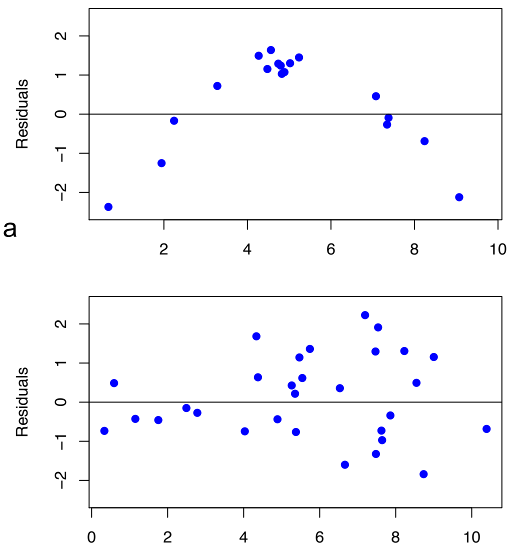
Model Checking: Residuals

- 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 model does not provide a good fit in these cases!



Violations of the model assumptions? How?

Simple Linear Regression: Residual Analysis: Non-normality of errors

- QQ-plot
  - Graphical technique that allows us to assess whether or not 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.

Simple Linear Regression:
Residual Analysis: Non-normality of errors

- Construction of QQ-Plot: (an example)

residuals
0.30
-0.25
-0.91
0.56
-0.79
-1.45
-0.42
-0.80
-0.39
-1.09
0.37
-0.56
1.15
-1.14
0.06
0.60
0.11
0.51
-0.02
-0.24

Sort

sorted	index(i)
-1.45	1
-1.14	2
-1.09	3
-0.91	4
-0.80	5
-0.79	6
-0.56	7
-0.42	8
-0.39	9
-0.25	10
-0.24	11
-0.02	12
0.06	13
0.11	14
0.30	15
0.37	16
0.51	17
0.56	18
0.60	19
1.15	20

Empirical Probab.

i/n (i-0.5)/n

pr1	pr2
0.05	0.025
0.10	0.075
0.15	0.125
0.20	0.175
0.25	0.225
0.30	0.275
0.35	0.325
0.40	0.375
0.45	0.425
0.50	0.475
0.55	0.525
0.60	0.575
0.65	0.625
0.70	0.675
0.75	0.725
0.80	0.775
0.85	0.825
0.90	0.875
0.95	0.925
1.00	0.975

Get z-values (quantiles from Normal distr.)

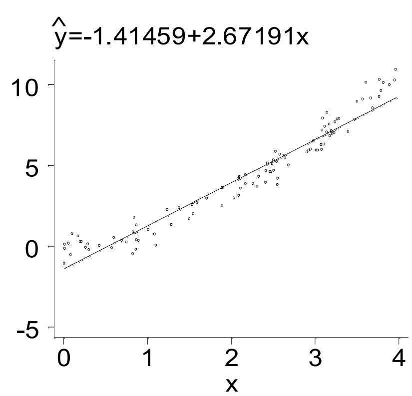
Z-quantile
-1.96
-1.44
-1.15
-0.93
-0.76
-0.60
-0.45
-0.32
-0.19
-0.06
0.06
0.19
0.32
0.45
0.60
0.76
0.93
1.15
1.44
1.96

$\underbrace{\text{Plot of sorted residuals (sample quantiles) versus z-quantile (theoretical quantiles)}}$
= QQ-plot

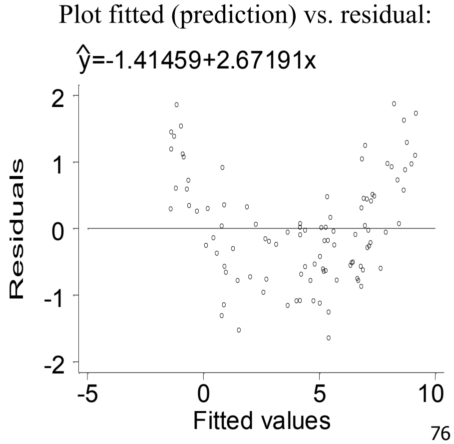
Simple Linear Regression:
Residual Analysis: Nonlinear Association

True model: $y = x$^1.7

Plot of Fitted Model:



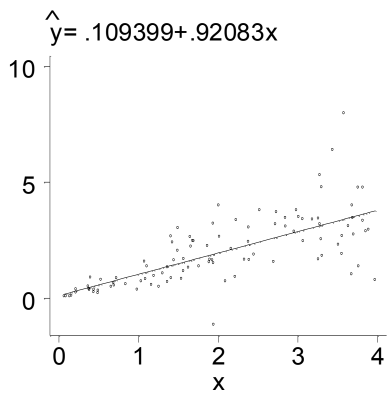
Plot fitted (prediction) vs. residual:



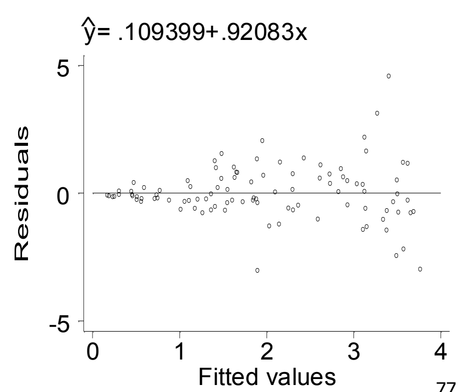
Simple Linear Regression: Residual Analysis: Non Constant Variance

True model: $y = x +$ errors increasing with x

Plot of Fitted Model:



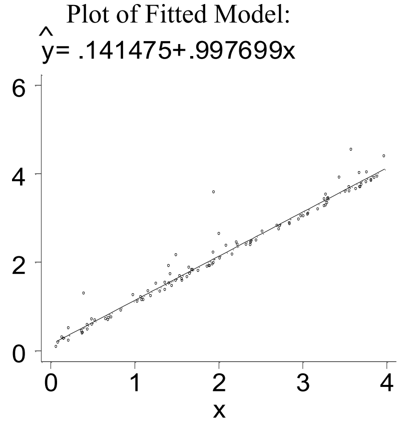
Plot fitted (prediction) vs. residual:



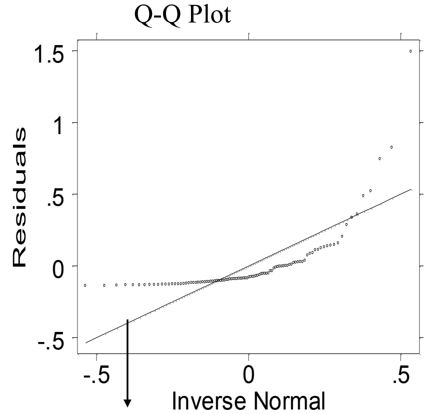
Simple Linear Regression: Residual Analysis: Non-normality of errors

True model: $y = x +$ chi-squared errors

Plot of Fitted Model:



Q-Q Plot

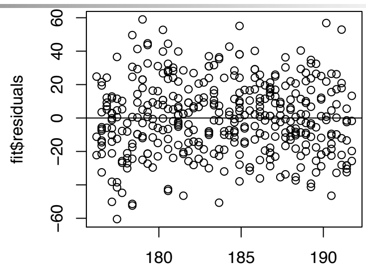


Under normality, residuals should fall on the straight line!

Cholesterol Example: Residuals

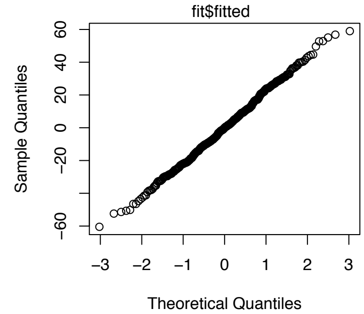
Plot of residuals versus fitted values
  Curvature?
  Heteroscedasticity?

R COMMANDS:
plot(fit$fitted, fit$residuals)



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

R COMMANDS:
qqnorm(fit$residuals)



Non-constant variance

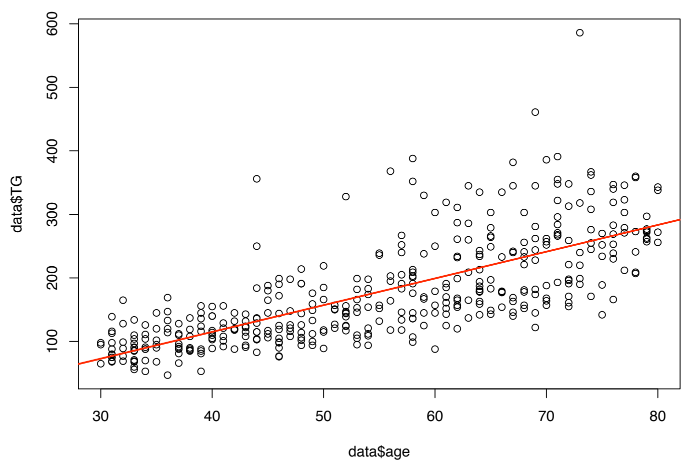
- Sometimes variance of y is not constant across the range of x (heteroscedasticity)
- Little effect on point estimates but variance estimates will be incorrect
- This affects confidence intervals and p-values
- To account for heteroscedasticity we can
  - Use robust standard errors
  - Transform the 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 non-constant variance
- Regression point estimates will be unchanged
- Robust or empirical standard errors will give correct confidence intervals and p-values

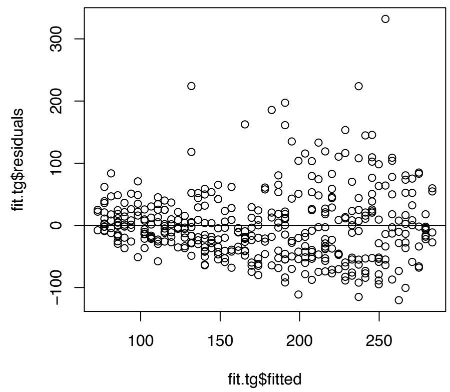
Cholesterol example: Robust standard errors

- Linear regression for association between age and triglycerides



Cholesterol example: Robust standard errors

- Residuals analysis suggests mean-variance relationship
- Use robust standard errors to get correct variance estimates



Cholesterol example: Robust standard errors

- Linear regression results:

```
> summary(fit.tg)

Call:
lm(formula = TG ~ age, data = data)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -53.3059    11.1339  -4.788 2.38e-06 ***
age           4.2090     0.1964  21.429  < 2e-16 ***
```

Point estimates are unchanged

- Results incorporating robust SEs:

```
> summary(fit.tg.ese)

Call:
gee(formula = TG ~ age, id = seq(1, length(age)), data = data)

Coefficients:
              Estimate Naive S.E.   Naive z Robust S.E.   Robust z
(Intercept) -53.305930 11.1339178 -4.787706   8.7387366 -6.099958
age           4.208964  0.1964165 21.428771   0.1813358 23.210880
```

Cholesterol example: Robust standard errors

- Linear regression results:

```
> summary(fit.tg)

Call:
lm(formula = TG ~ age, data = data)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -53.3059    11.1339  -4.788 2.38e-06 ***
age           4.2090     0.1964  21.429  < 2e-16 ***
```

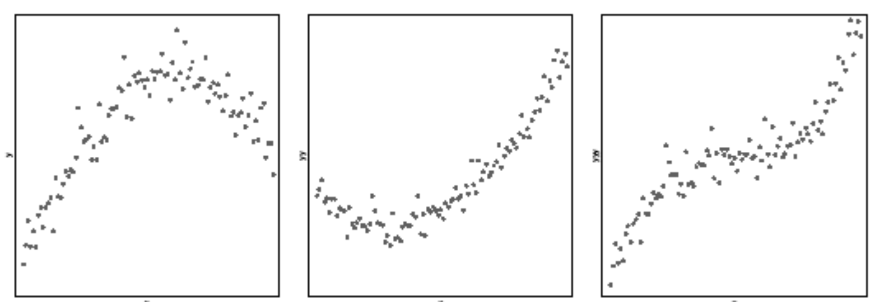
Standard errors are corrected

- Results incorporating robust SEs:

```
> summary(fit.tg.ese)

Call:
gee(formula = TG ~ age, id = seq(1, length(age)), data = data)

Coefficients:
              Estimate Naive S.E.   Naive z Robust S.E.  Robust z
(Intercept) -53.305930 11.1339178 -4.787706   8.7387366 -6.099958
age           4.208964  0.1964165 21.428771   0.1813358 23.210880
```


Transformations

- Sometimes the relationship between Y and X is not linear



To model "curvilinear relationships" one can look at transformations in X or Y [or both]

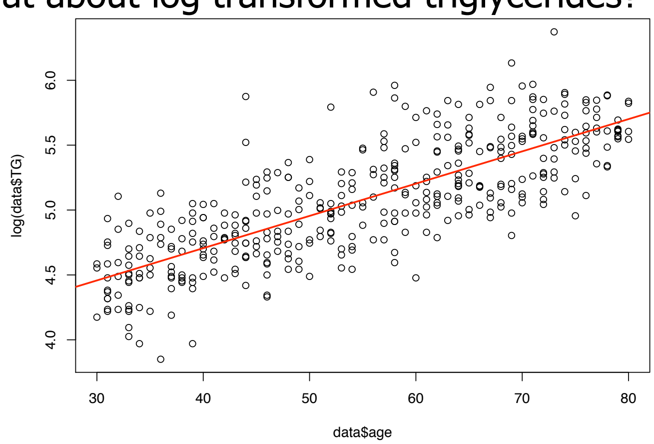
Transformations

- Some reasons for using data transformations
  - 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

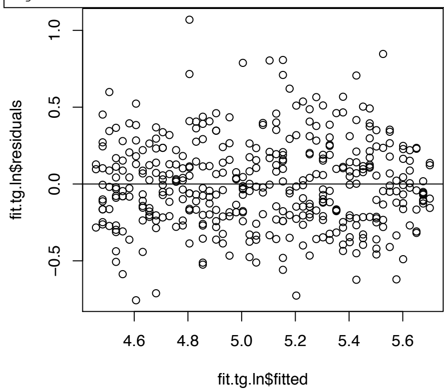
Cholesterol example: Transformations

- We have seen that triglycerides are associated with age but display non-constant variance
- What about log transformed triglycerides?



Cholesterol example: Transformations

```
> summary(fit.tg.ln)

Call:
lm(formula = log(TG) ~ age)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.7115803  0.0559237   66.37   <2e-16 ***
age         0.0248646  0.0009866   25.20   <2e-16 ***
```



- Heteroscedasticity is corrected
- But interpretation of model is more complicated

Transformations

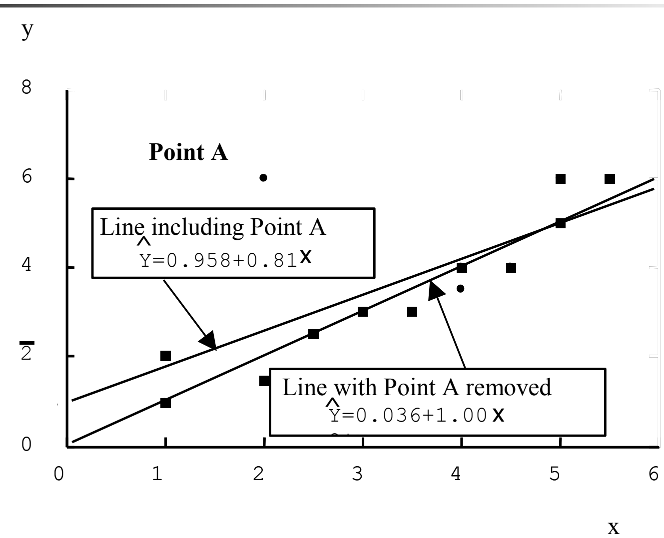
- Rarely do we know which transformation of the predictor provides best "linear" fit
  - As always, there is a danger in using the data to estimate the best transformation to use
    - If there is no association of any kind between the response and the predictor, a "linear" fit (with a zero slope) is the correct one
    - Trying to detect a transformation is thus an informal test for an association
      - Multiple testing procedures inflate the type I error
- It is best to choose the transformation of the predictor on scientific grounds
  - However, sometimes it doesn't matter – it is often the case that many functions are well approximated by a straight line over a small range of the data

Model Checking: Outlier vs Influential observations

- **Outlier:** an observation with a residual that is unusually large (positive or negative) as compared to the other residuals.

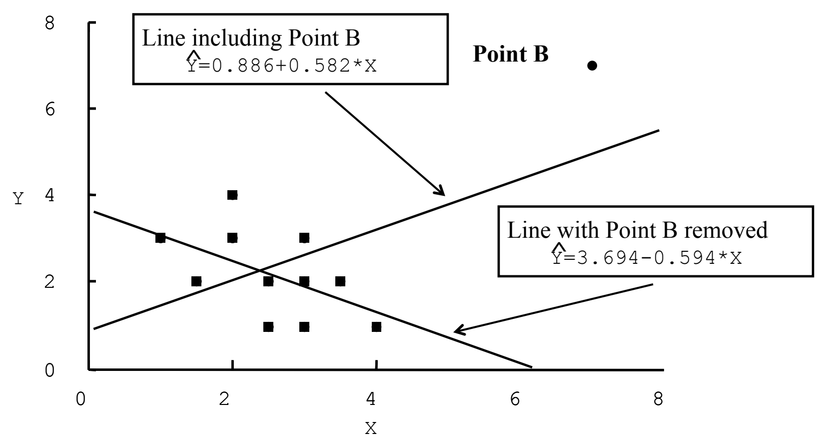
- **Influential point:** an observation that has a great deal of influence in determining the regression equation.
  - Removing such a point would markedly change the position of the regression line.
  - Observations that are somewhat extreme for the value of x are often influential.

Outlier vs Influential observations



Point A is an *outlier, but is not influential*.

Outlier vs Influential observations



Point B *is influential, but not an outlier*.

Model Checking: Deletion diagnostics

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

$$\frac{\Delta\beta_{(i)}}{se(\hat{\beta})}$$:Standardized delta-beta

Delta-beta : tells how much the regression coefficient changed by including the i^{th} observation

Standardized delta-beta : approximates how much the t-statistic for a coefficient changed by adding the i^{th} observation

Cholesterol Example: Deletion diagnostics

```
> dfb = dfbeta(fit)
> dfb[order(abs(dfb[,2]), decreasing = T)[1:15],]
    (Intercept)          age
114  -0.9893663  0.015268514
166  -0.6827966  0.014888475
255  -0.6190643  0.013902713
186  -0.8544144  0.013279531
113   0.5376293 -0.011943495
325  -0.7517511  0.011308451
365   0.7676508 -0.011297278
257  -0.7374003  0.011092575
290  -0.7024787  0.010757541
144   0.7120264 -0.010710881
197  -0.6784150  0.010469720
296  -0.6499386  0.010101515
231  -0.6293174  0.009712016
7     0.4403297 -0.009524470
252  -0.5981020  0.009412761
```

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

Model Checking: Deletion diagnostics

- What to do if you find an 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 to Model Assumptions

	Non Linearity	Non Normality	Unequal Variances	Dependence
Estimates	Rubbish	Minimal for most departures. Outliers can be a disaster.	Minimal impact.	Often the estimates are unbiased.
Tests/CIs	Rubbish	Minimal for most departures. CIs for correlation are sensitive.	Variance estimates are wrong, but the effect is usually not dramatic.	Variance estimates are wrong (overestimate the precision and inflate test)
Correction	Transform or Choose a nonlinear model.	Delete outliers (if warranted) or Use robust regression	Transform or Use robust standard error.	Regression for dependent data.

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 simultaneous influence of several variables on the response
- It may reveal relationships that are completely hidden in univariate regression models

Motivation

- Why not fit multiple separate simple linear regressions?
  - A confounder can make the observed association between the predictor of interest and the response variable look
    - stronger than the true association,
    - weaker than the true association, or
    - even the reverse of the true association
- What could we do?
  - We can adjust for the effects of the confounder by adding a corresponding term to our linear regression! (more details later)

Motivation: Cholesterol Example

- Data

	sex	age	chol	BMI	TG	APOE	rs174548	rs4775401
1	1	74	215	26.2	367	4	1	2
2	1	51	204	24.7	150	4	2	1
3	0	64	205	24.2	213	4	0	1
4	0	34	182	23.8	111	1	1	1
5	1	52	175	34.1	328	1	0	0
6	1	39	176	22.7	53	4	0	2

- 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:

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

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

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:

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

  - Random component:

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

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

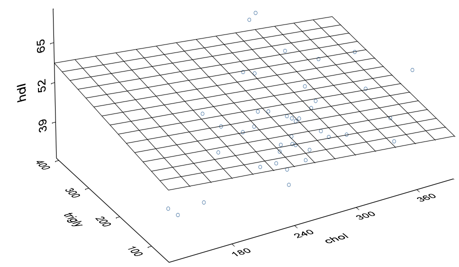
$E[Y|x_1=1, x_2=0]-E[Y|x_1=0,x_2=0] = \beta_1$
$E[Y|x_1=1, x_2=1]-E[Y|x_1=0,x_2=1] = \beta_1$
$E[Y|x_1=0, x_2=1]-E[Y|x_1=0,x_2=0] = \beta_2$
$E[Y|x_1=1, x_2=1]-E[Y|x_1=1,x_2=0] = \beta_2$

Estimation

- Least Squares Estimation:
  - minimizes the residual sum of squares

$$\sum_i (y_i - \hat{y}_i)^2$$

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



Estimation and Inference

- Inference
  - About regression model parameters
    - **Hypothesis Testing** H_0: $\beta_j=0$

<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= $\sum(\hat{y}_i - \bar{y})^2$	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= $\sum(y_i - \bar{y})^2$		

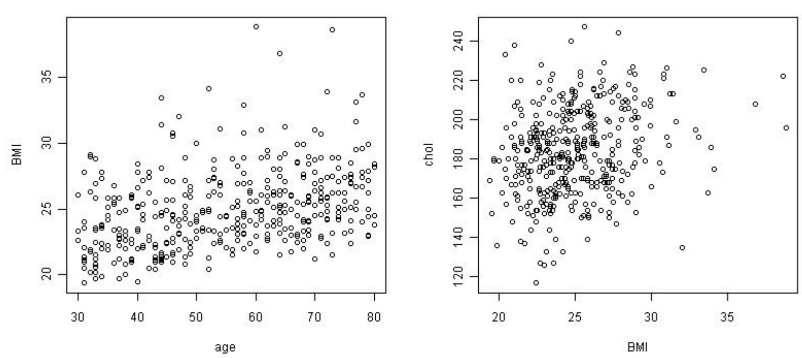
Estimation and Inference

- 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 age
- And cholesterol increases with 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

```
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
$$\hat{y} = 137.16 + 0.20 Age + 1.43 BMI$$
- Question: How do we interpret the age coefficient?

Scientific example: Modeling cholesterol using age and BMI

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

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.

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 this model compare to a model that contains only the mean?

```
> 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

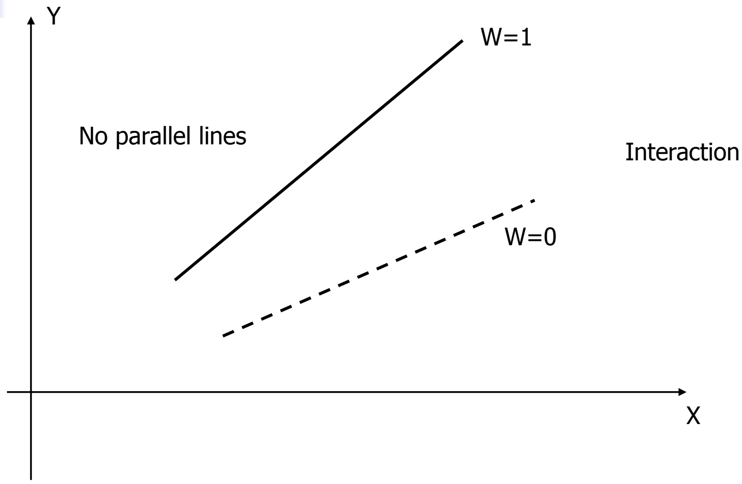
Discriminating between different classifications

- It is often very difficult to decide whether a new variable should be treated as a confounding or effect modification variable
- Data and scientific assessments help discriminate 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 vary across strata of the new variable

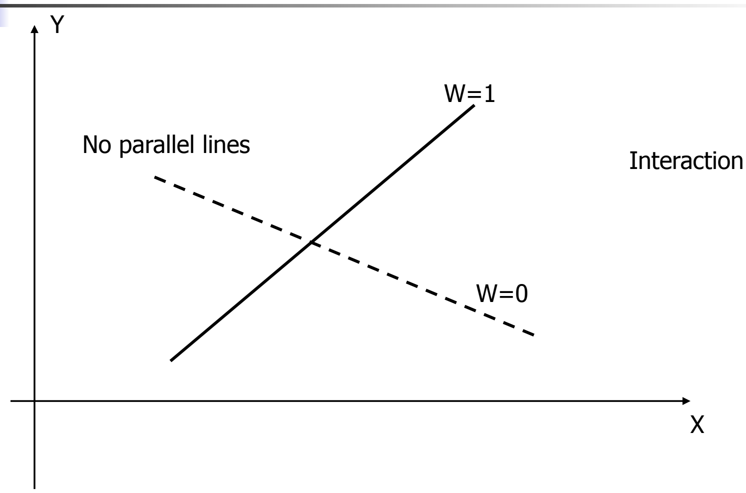
Confounding vs. Interaction/Effect Modification

- 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 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

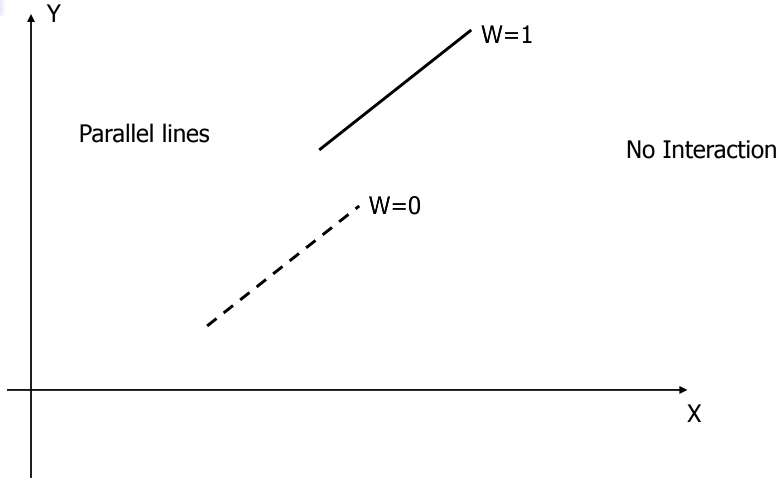
Graphical Representation



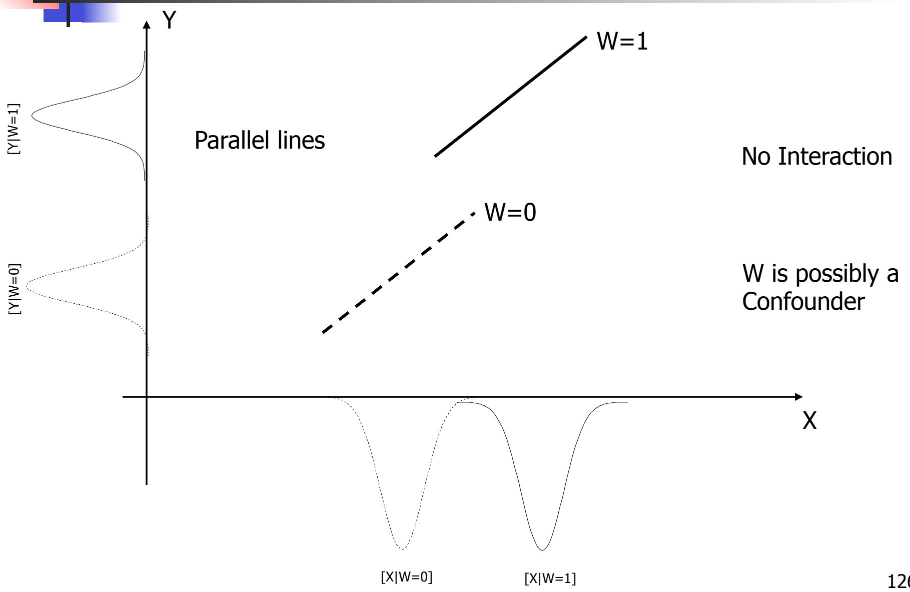
Graphical Representation



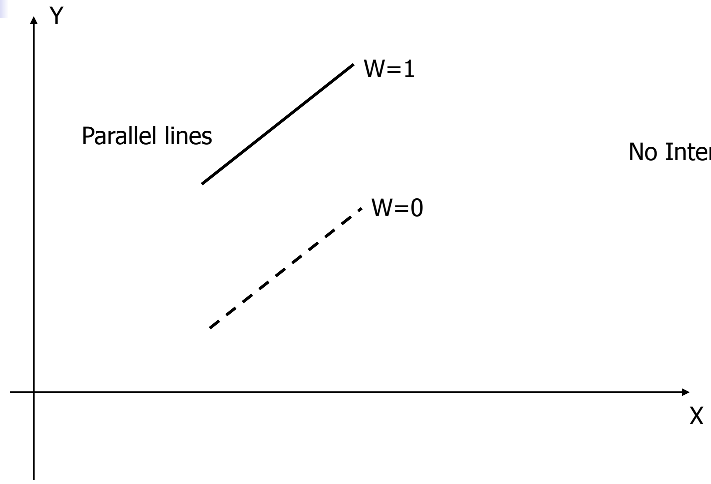
Graphical Representation



Graphical Representation



Graphical Representation



No Interaction

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 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 now 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!]

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

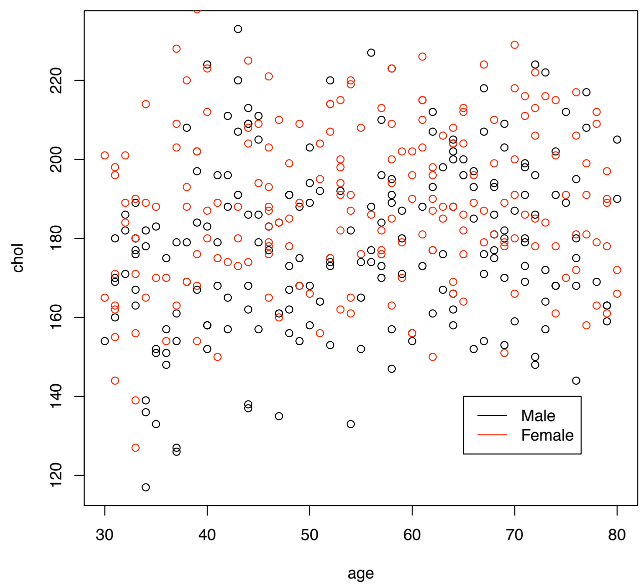
- Factor terms involving x_1:

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

Slope of x_1 changes with x_2 =

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 gender affect the age – cholesterol relationship?



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

We first fit the model with age and sex terms only

```
> fit2 = lm(chol ~ age+sex)
> summary(fit2)

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

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 ***
sex          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 gender affect the age – cholesterol relationship?

- 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

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

Model with age and sex main effects, plus interaction effect

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

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 ** 
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
```

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

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

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 **
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
```

Mean cholesterol for males at age 0

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

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

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 **
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

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

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

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 ** 
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

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

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

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 ** 
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

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

- Interpretation?
  - Estimated model:

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

Subject 1: Age = a+1, sex = b
Subject 2: Age = a, sex = b
Difference in the estimated cholesterol:

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

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

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

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

```
> anova(fit2,fit3)
Analysis of Variance Table

Model 1: chol ~ age + sex
Model 2: chol ~ age * sex
  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 gender affect the age – cholesterol relationship?