



REGRESSION METHODS: CONCEPTS & APPLICATIONS

LECTURE 1: SIMPLE LINEAR REGRESSION



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

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



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



Motivation: Cholesterol Example

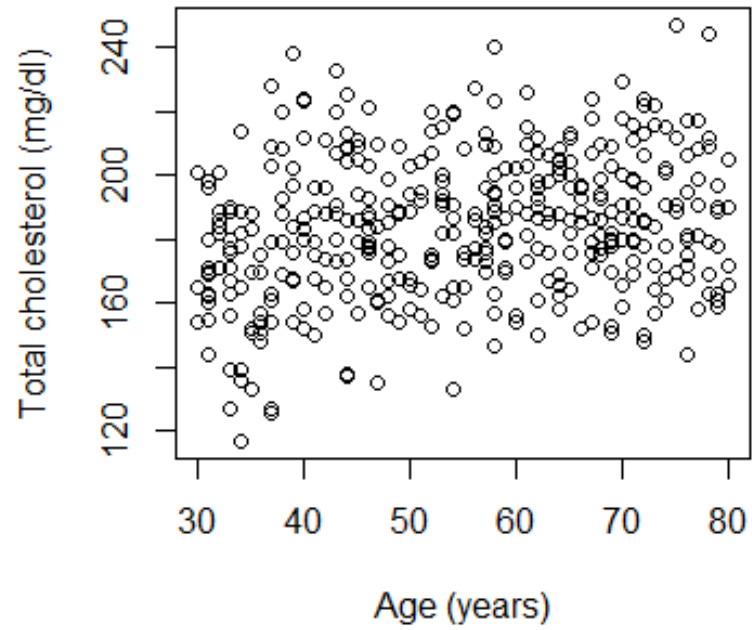
- Linear regression is concerned with a **continuous** outcome
- Data: Factors related to serum total cholesterol (continuous outcome), 400 individuals, 11 variables

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



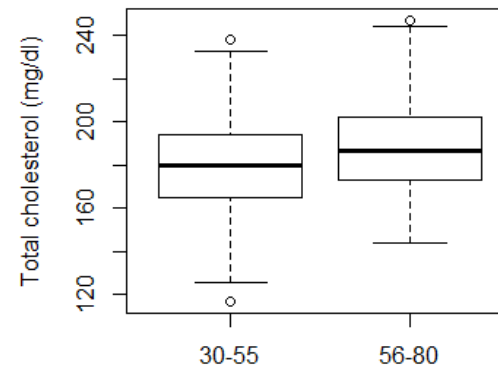
Motivation: Cholesterol Example



Motivation: Cholesterol Example

- 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)")
```





Motivation: Cholesterol Example

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

```
> 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 30-55 mean in group 56-80
      179.9751      187.8945
```

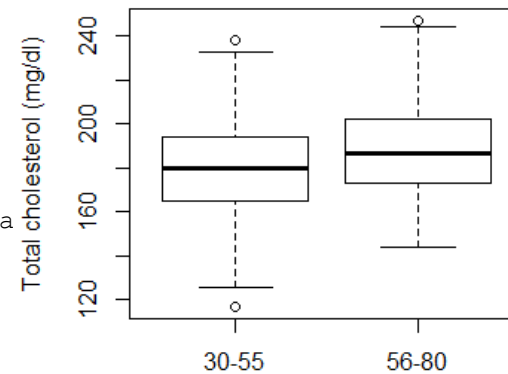

Motivation: Cholesterol Example

- **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
```





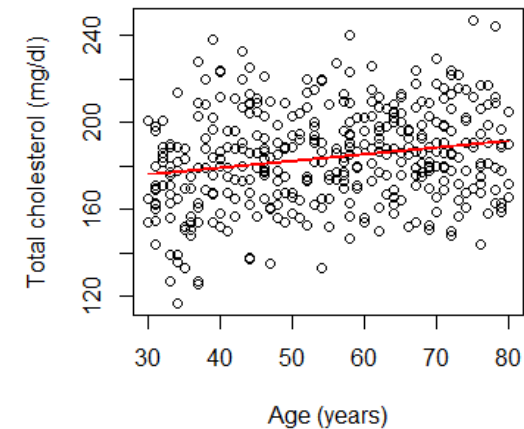
Motivation: Cholesterol Example

- Using the t-test:
 - There is a statistically significant 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?

Motivation: Cholesterol Example

- We might assume that mean cholesterol changes linearly with age:

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



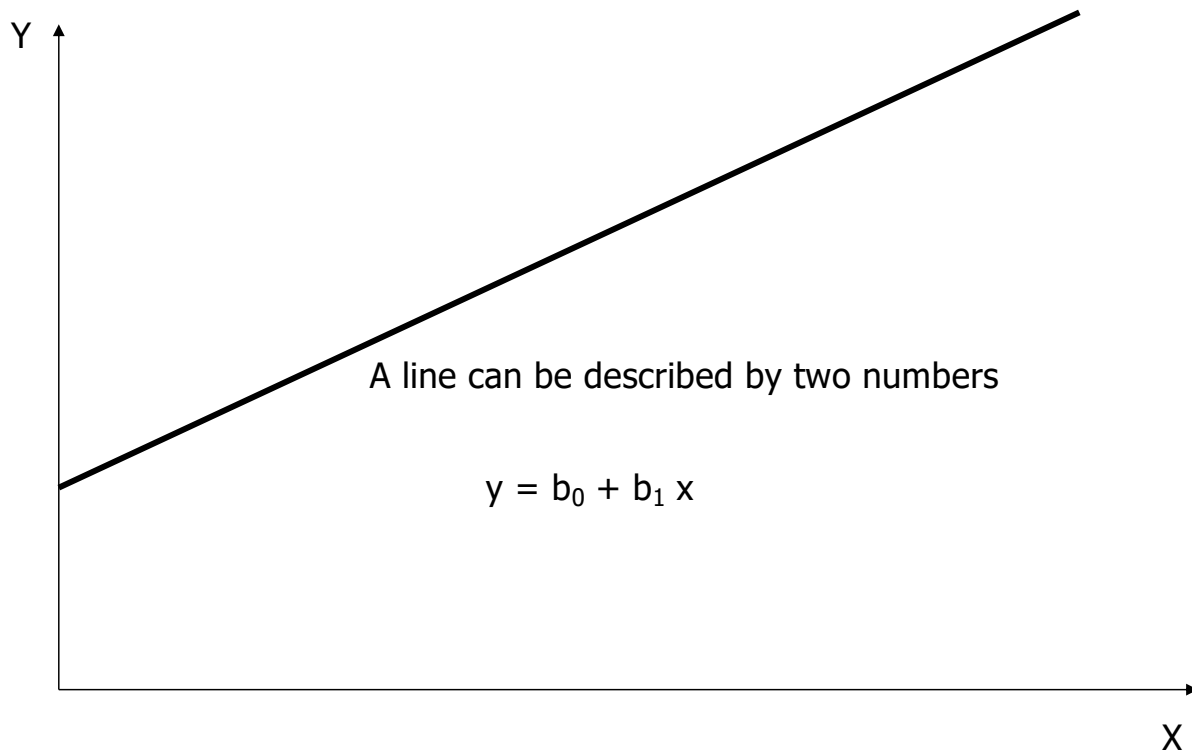


Linear Regression

- 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 a single predictor
 - Simple linear regression model

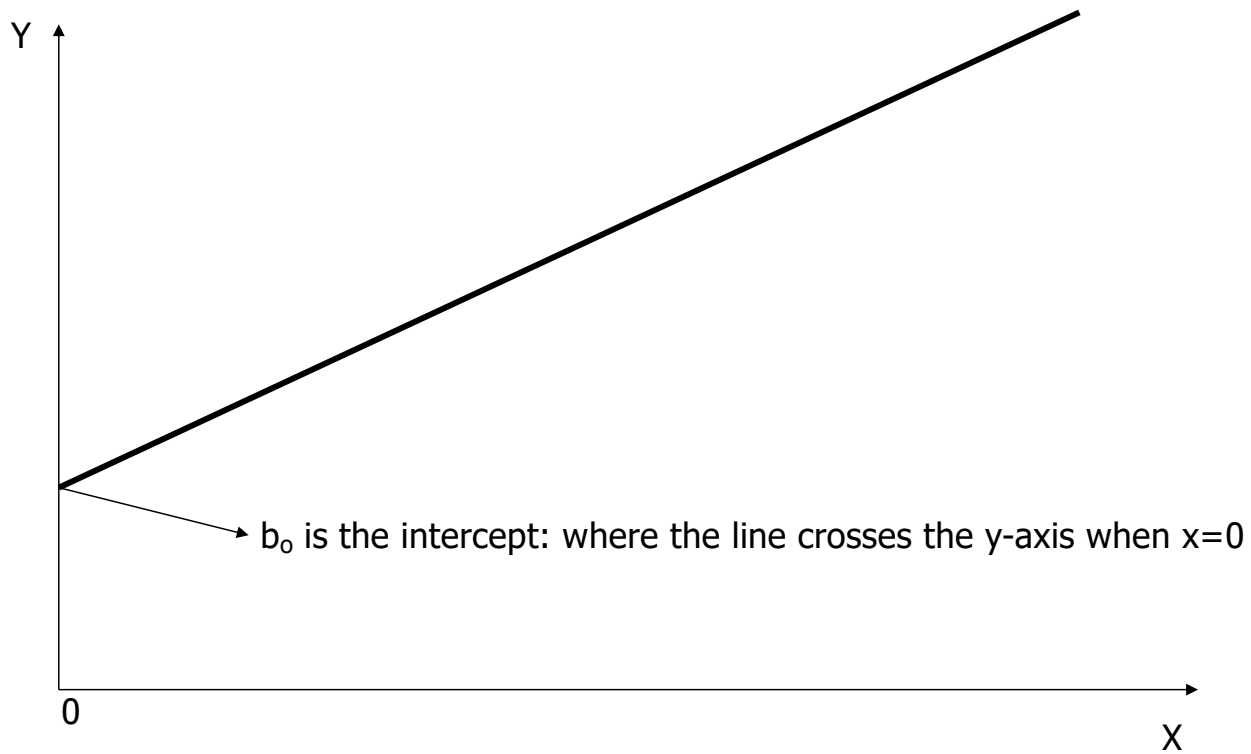


The straight line equation



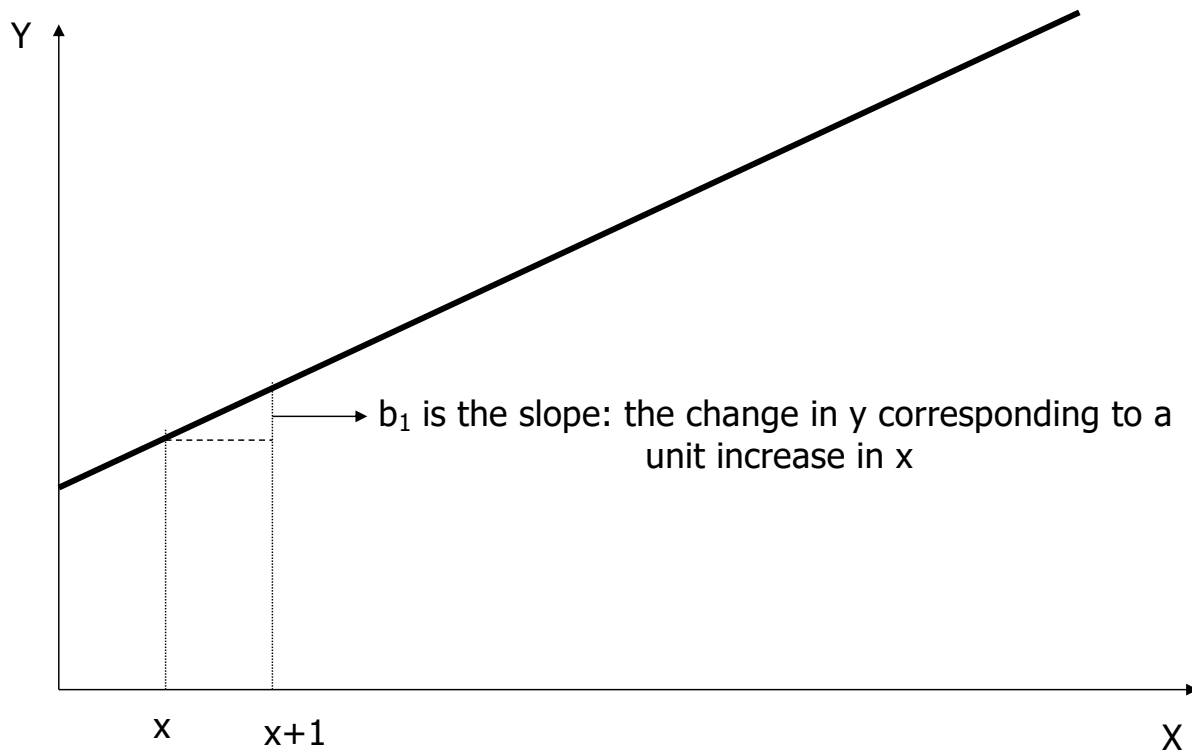


The straight line equation

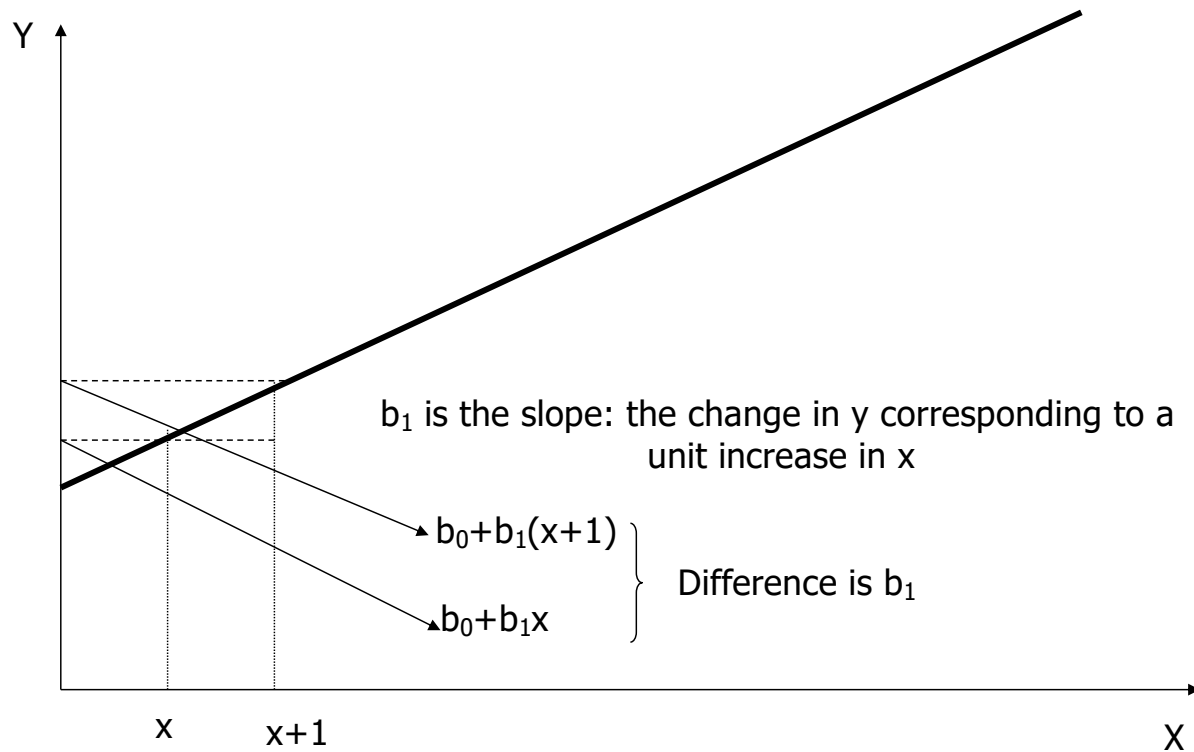




The straight line equation

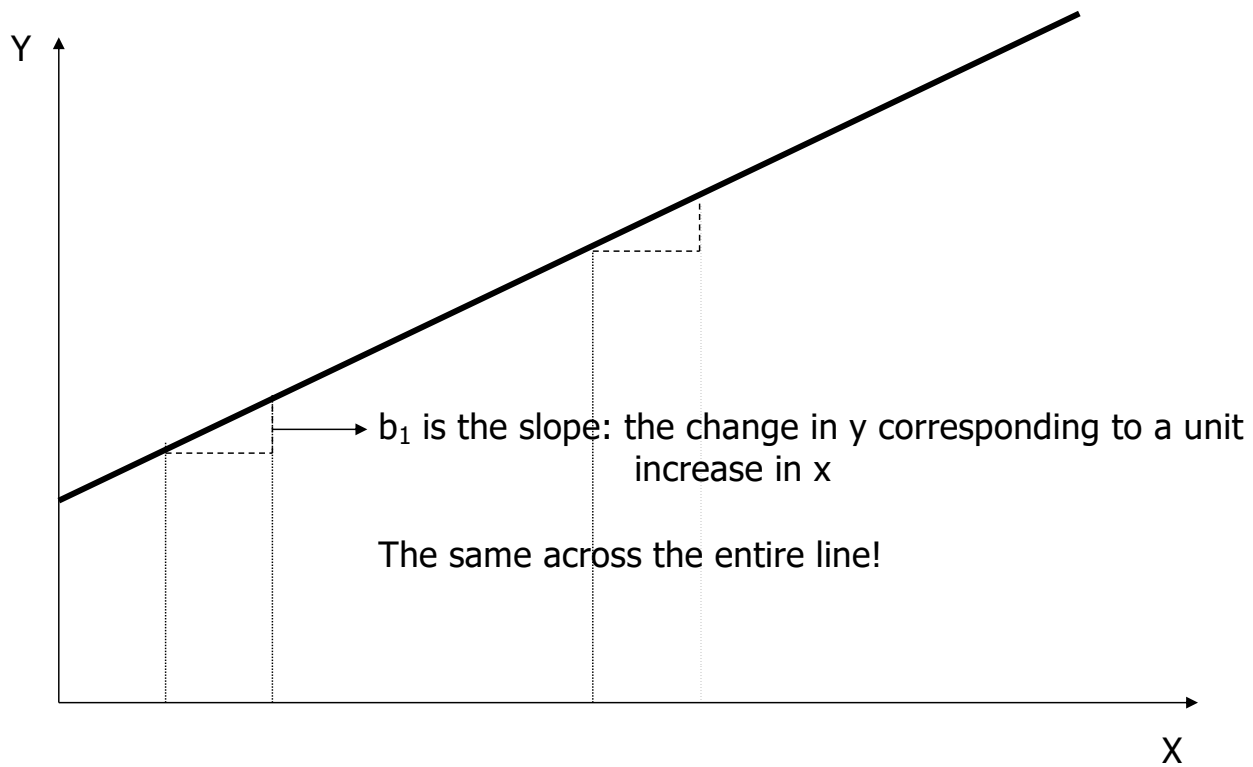


The straight line equation



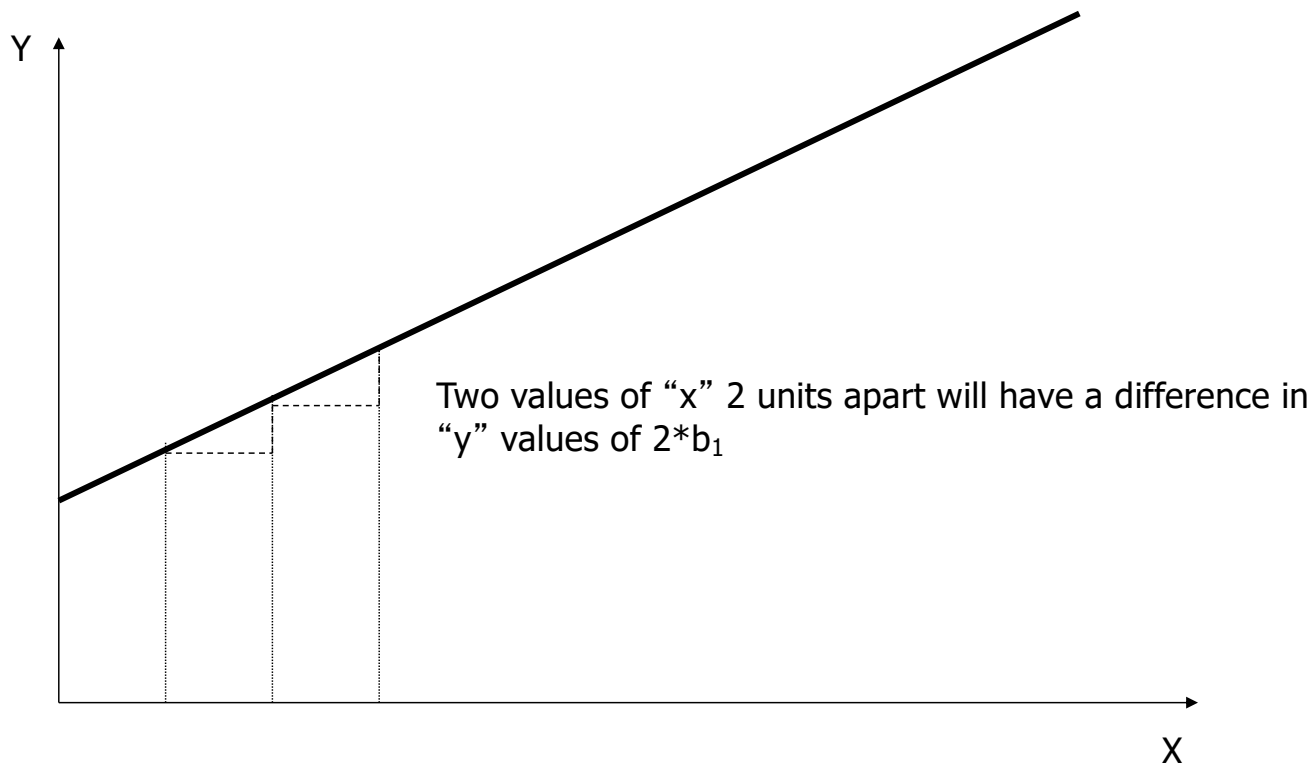


The straight line equation





The straight line equation



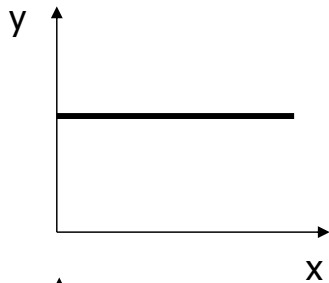


The straight line equation

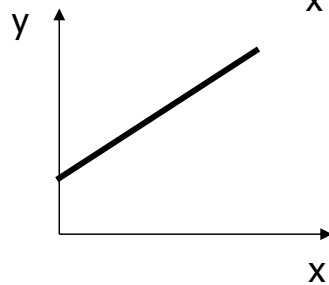
- Slope b_1 is the change in y corresponding to a one unit increase in x
- Slope gives information about magnitude and direction of the association between x and y



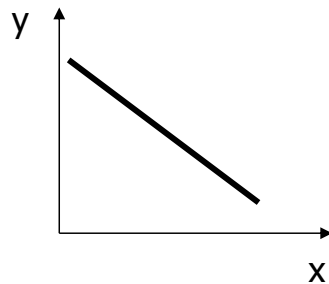
The straight line equation



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



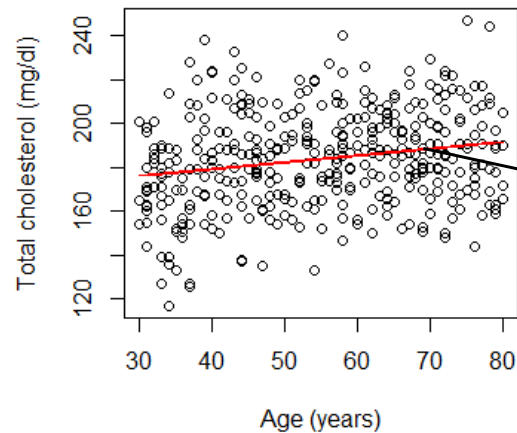
($b_1 > 0$) Positive association between x and y
(values of y increase as values of x increase)



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

Simple Linear Regression

- 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



We estimate a straight line describing trend 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):
 - continuous variable we want to predict or explain
- **Explanatory** or **independent** or **predictor** variable (X):
 - attempts to explain the response
- **Simple Linear Regression Model:**

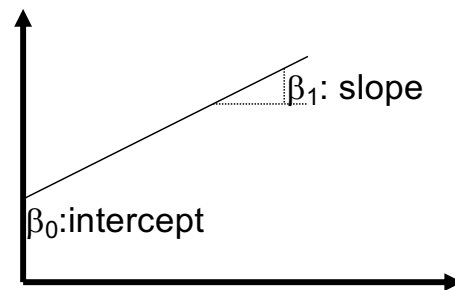
$$y = \beta_0 + \beta_1 x + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

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

- Random component:

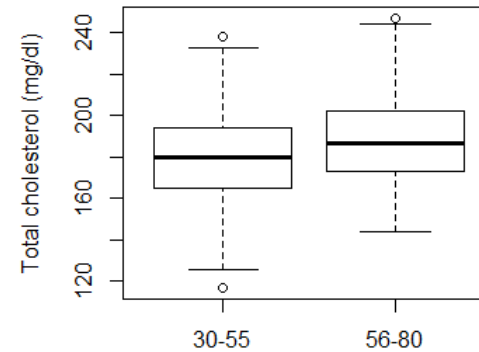
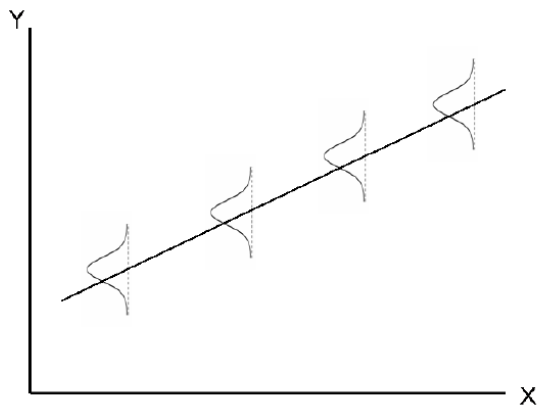
$$\text{Var}[Y | X = x] = \sigma^2$$

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:** $E[Y|x] = \beta_0 + \beta_1 x$ $\text{Var}[Y|x] = \sigma^2$
- **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:** $E[Y|x] = \beta_0 + \beta_1 x$ $\text{Var}[Y|x] = \sigma^2$
- **Question:** How do you interpret β_1 ?
- **Answer:**

$$\begin{aligned} E[Y|x] &= \beta_0 + \beta_1 x \\ E[Y|x+1] &= \beta_0 + \beta_1(x+1) = \beta_0 + \beta_1 x + \beta_1 \end{aligned}$$

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

i.e. β_1 is the difference in the mean response associated with a one unit positive difference in x



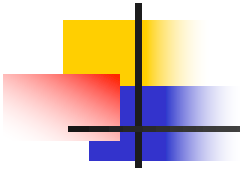
Example: Cholesterol and age

- **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:**



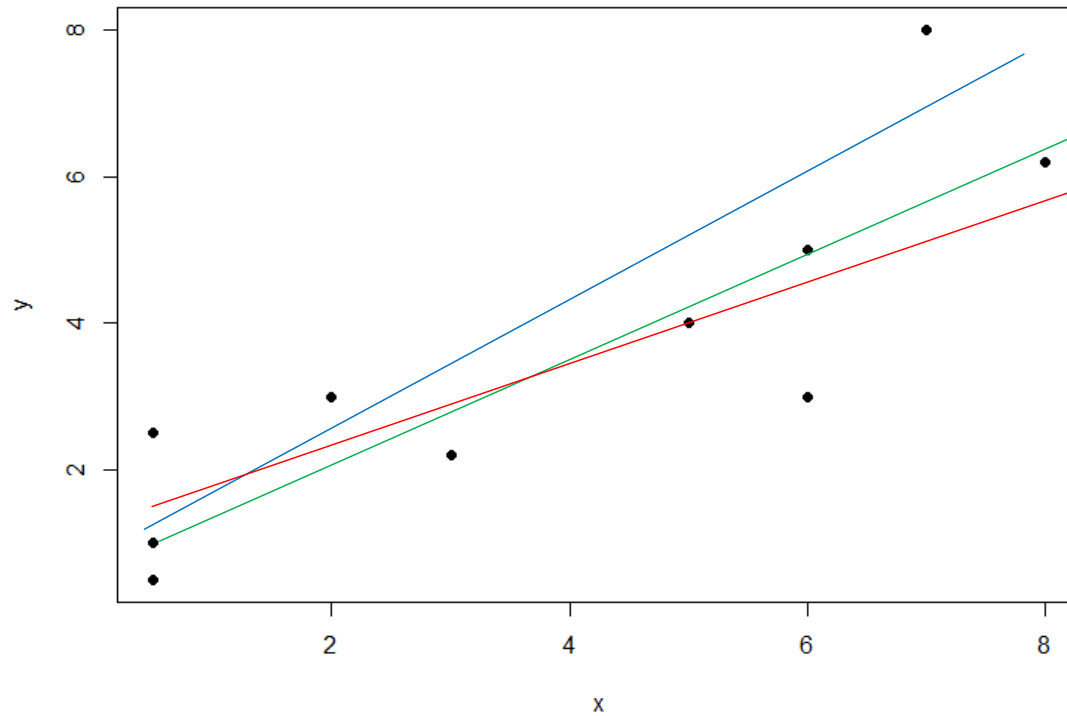
Example: Cholesterol and age

- **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 cholesterol associated with a one year increase in age



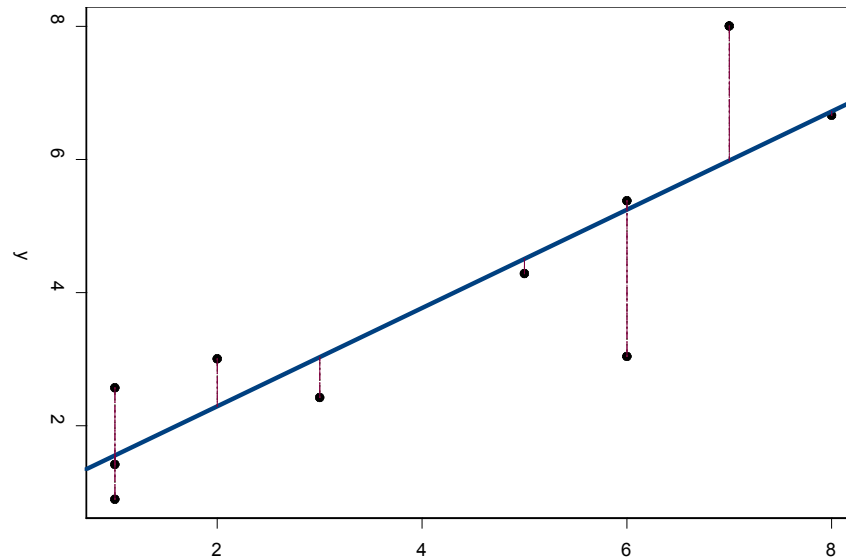
Least Squares Estimation

- Question: How to find a “best-fitting” line?



Least Squares Estimation

- Question: How to find a “best-fitting” line?



- Method: Least Squares Estimation

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 β_0 and β_1 by finding the values that minimize D
- We can use these estimates to get an estimate of the variance about the line (σ^2)



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:

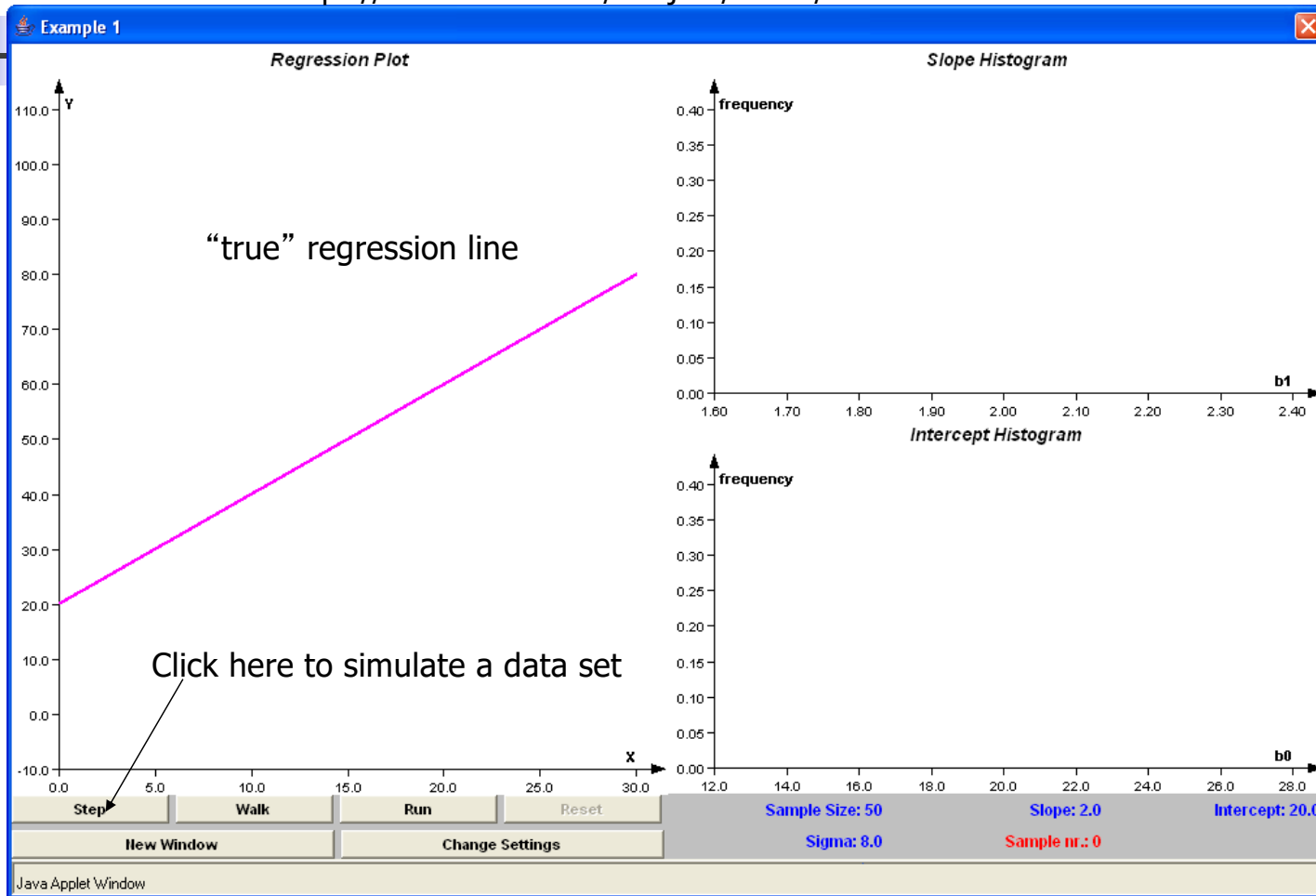
$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n r_i^2}{n-2} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n-2}$$

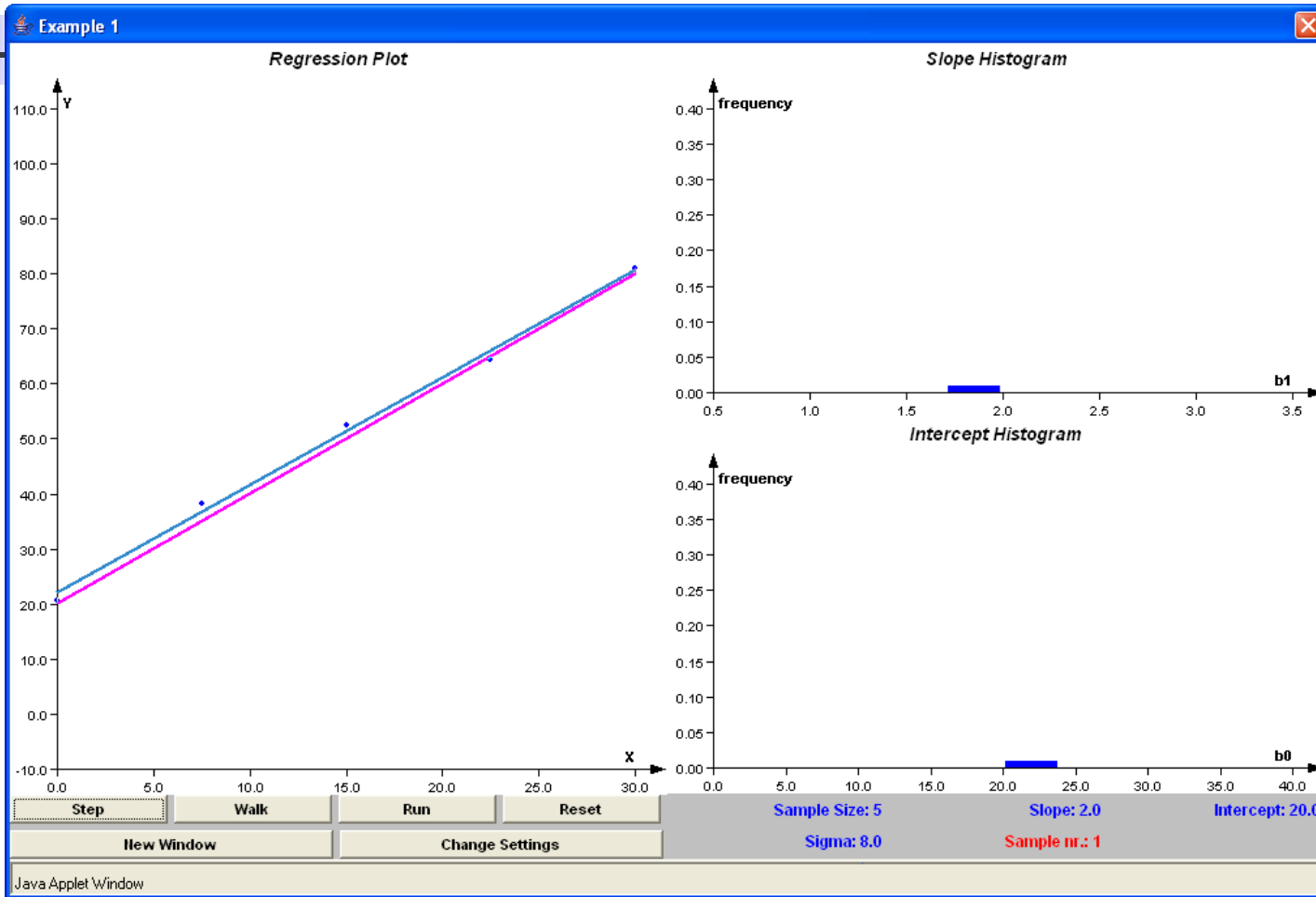
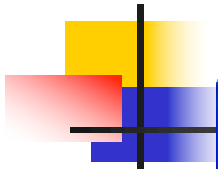


Estimated Standard Errors

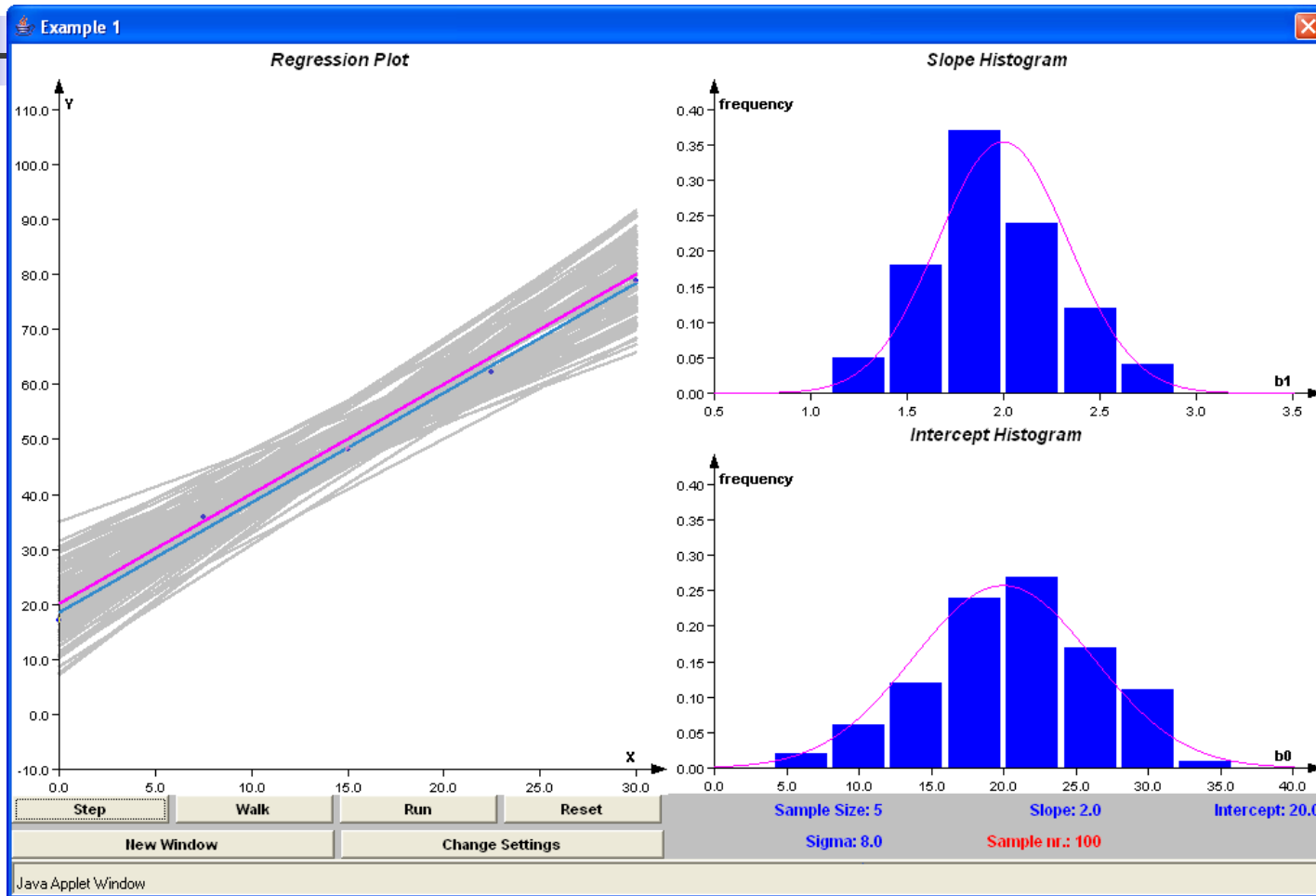
- Recall that, when estimating parameters from a sample, there will be **sampling variability** in the estimates
- This is true for regression parameter estimates
- Looking at the formulas for $\hat{\beta}_0$ and $\hat{\beta}_1$, we can see that they are just complicated means
- In repeated sampling 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/>

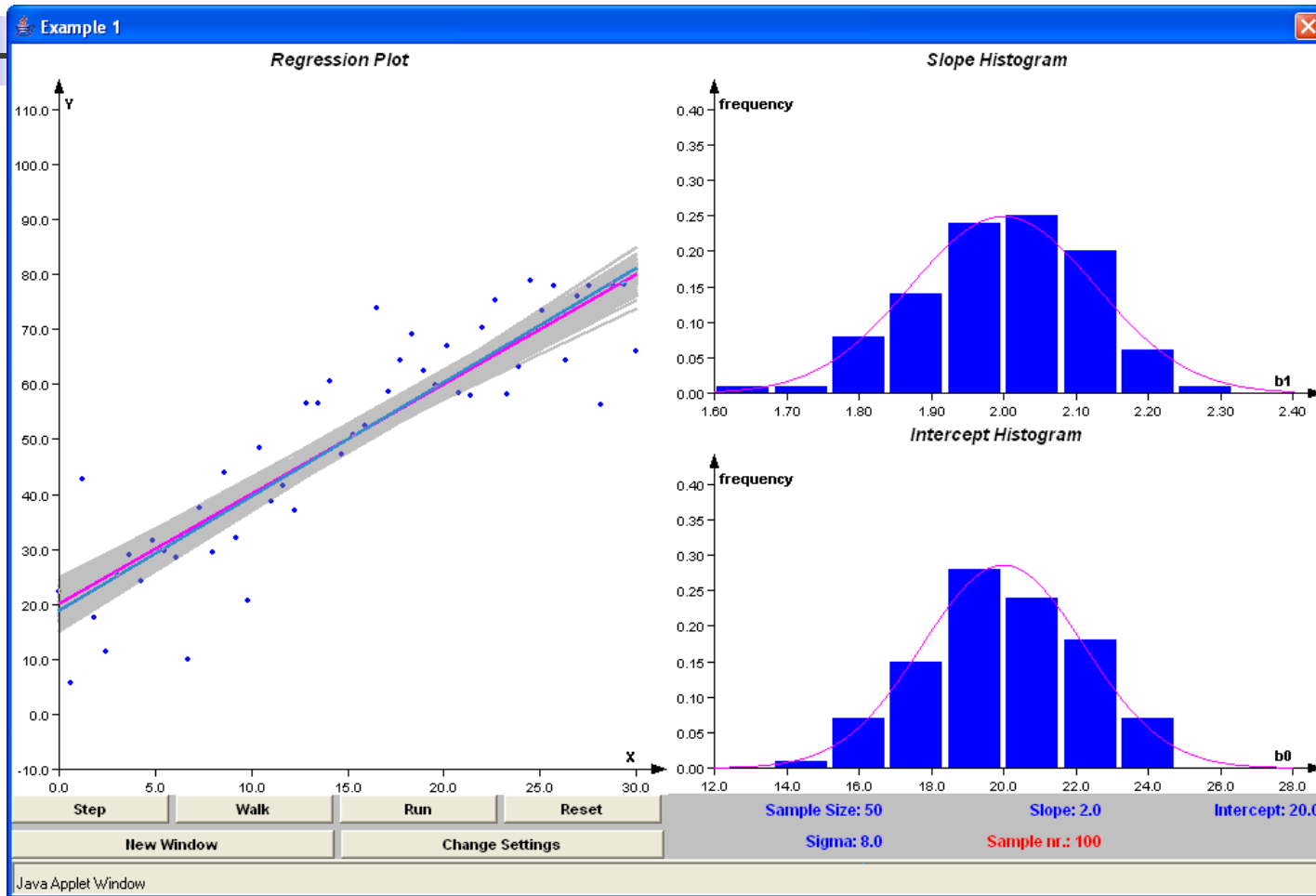




Sampling Distribution



Sampling Distribution





Inference

- About regression model parameters

- Hypothesis testing: $H_0: \beta_j=0$ ($j=0,1$)

- Test Statistic:

- Large Samples:

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

- Small Samples:

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

- Confidence Intervals:

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

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

Inference: Hypothesis Testing

Null Hypothesis: $\beta_j = 0$

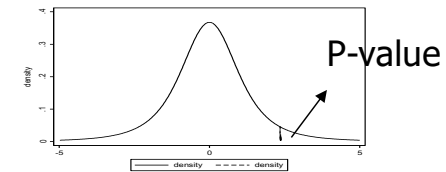
T=test statistic

Alternative

P-Value

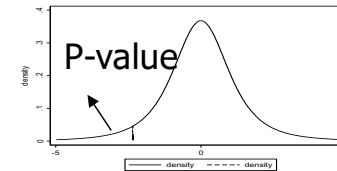
$$\beta_j > 0$$

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



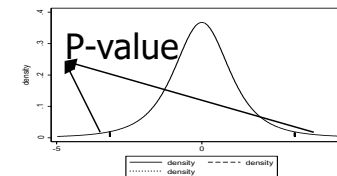
$$\beta_j < 0$$

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



$$\beta_j \neq 0$$

$$2P(t_{n-2} > |T|)$$





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 model results mean in terms of our scientific question?
 - Parameter estimates and confidence intervals:

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

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

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

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



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:

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

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

- Answer: $\hat{\beta}_1$: mean cholesterol is estimated to be 0.31 mg/dl higher for each additional year of 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:

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

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

- 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). 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
 - Units for predictor and response. Scale matters!



Inference for predictions

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

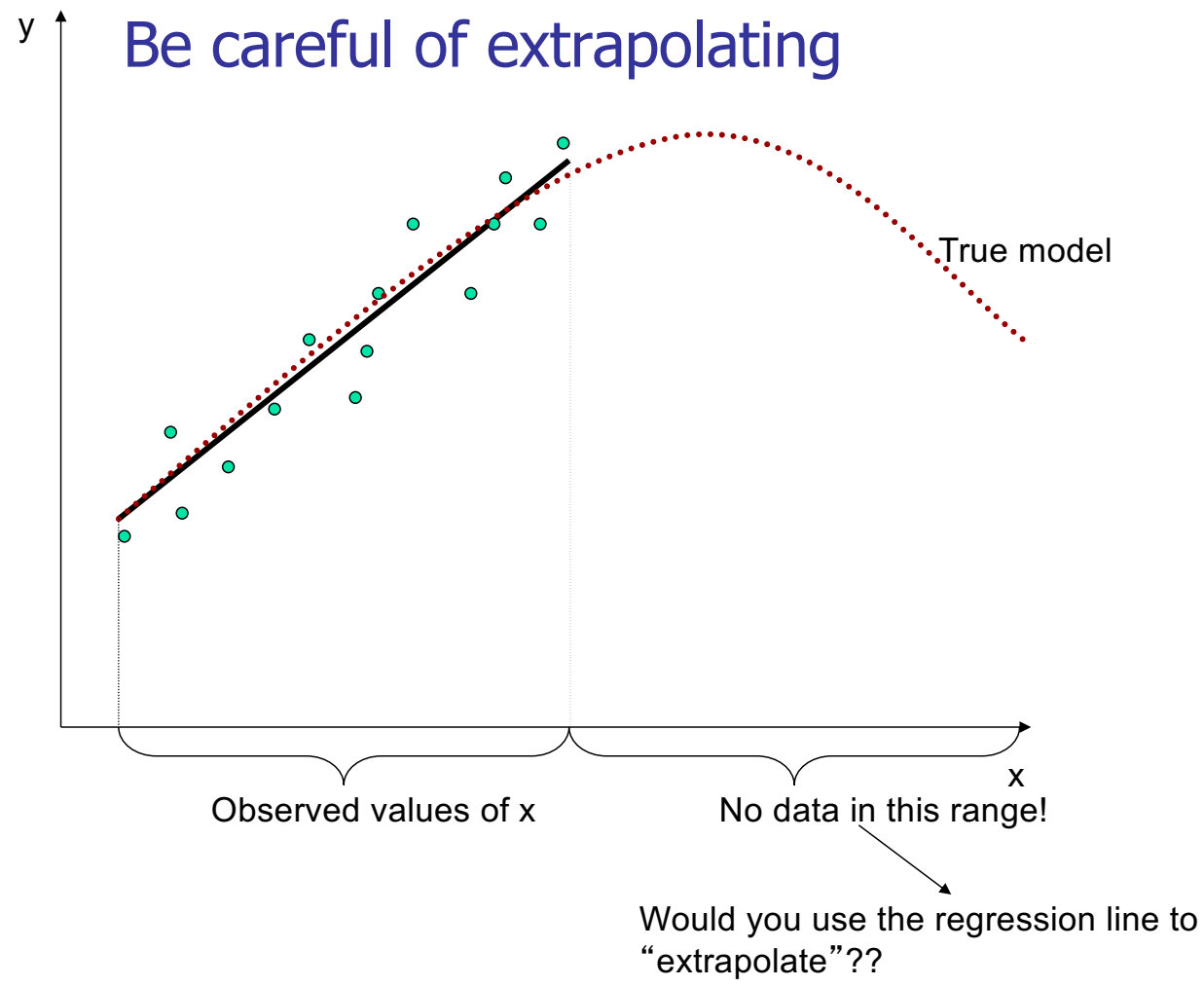
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

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

Be Cautious: This assumes the model is true.

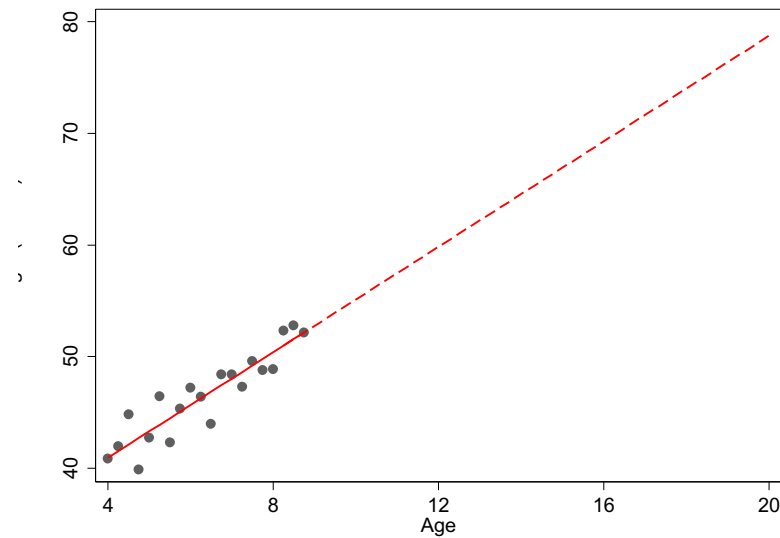
- May be a reasonable assumption within the range of your data.
- It may not be true outside the range of your data!

Be careful of extrapolating





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 $E[Y|X=x]$:

- 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 gets further 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 $x = 46$

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

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

- **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 $x = 46$

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

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

- **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 $x = 46$

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

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

- **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 $x = 46$

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

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

- **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.
- Note: Unlike confidence intervals, the formula for the prediction interval depends on the normality assumption regardless of sample size.



Lab

- Let's put some of the concepts we have been discussing into practice
- Open up the Labs file (2023_SISG_5_Labs.html) and RStudio and follow the directions to load the class data set and install the R packages you will need for this module
- For our first lab we will work on **Exercises 1-3**



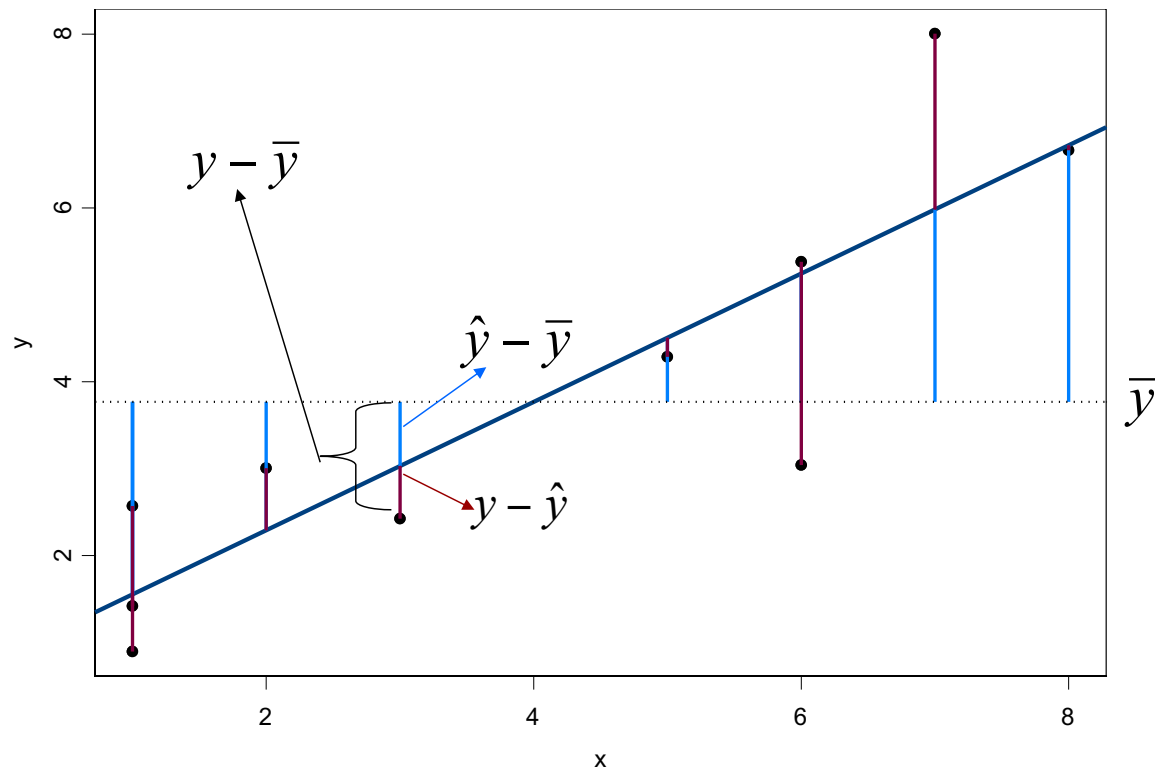
Simple Linear Regression: R^2

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

It can be shown that:

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

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 .

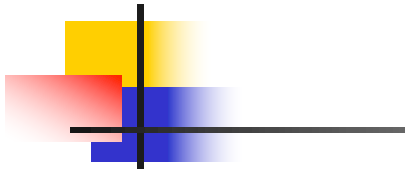


Simple Linear Regression: R^2

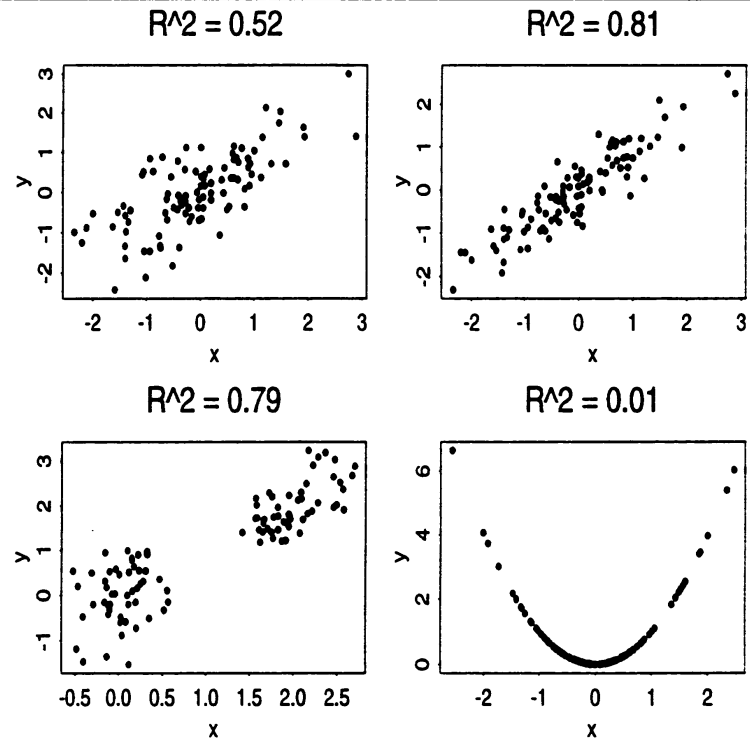
- Given no linear association:
 - We could simply use the sample mean to predict $E(Y)$. The variability between the data and this simple prediction is given as SST.
- Given a linear association:
 - The use of X permits a potentially better prediction of Y by using $E(Y | X)$.
 - **Question:** What did 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

```
> 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

Annotations:

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

In simple linear regression:

$$F\text{-statistic} = (t\text{-statistic for slope})^2$$

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. $\text{Var}[Y|x]$ does not depend on x

L inearity
I ndependence
N ormality
E qual 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

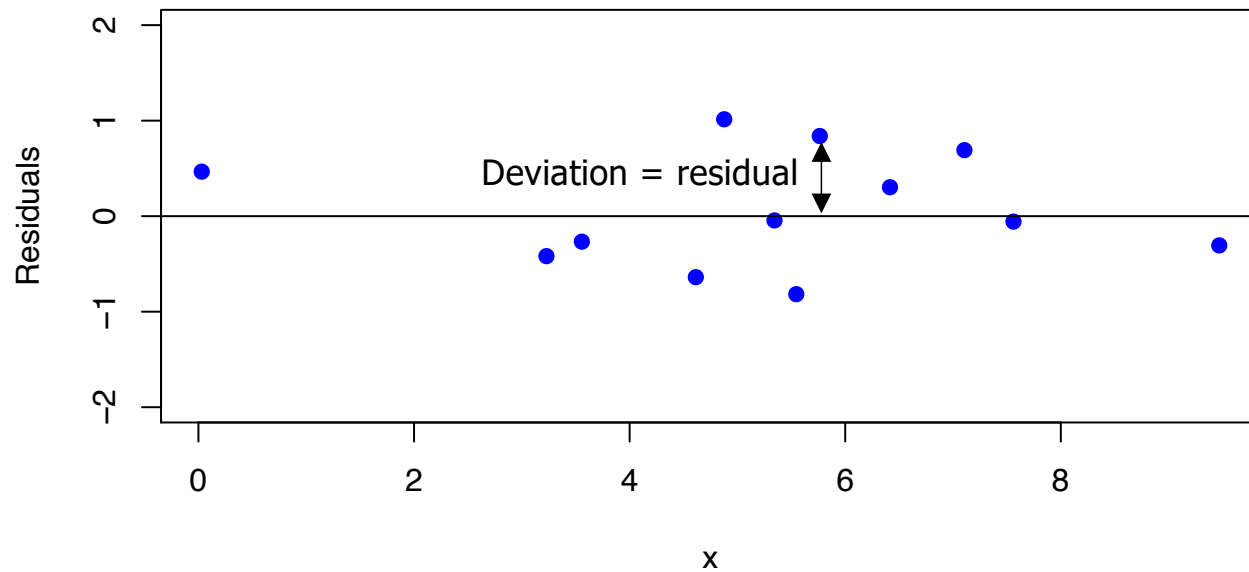


Model Checking: Residuals

L inearity	Plot residual vs X or vs \hat{Y} Q: Is there any structure?
I ndependence	Q: Any scientific concerns?
N ormality	Residual histogram or qq-plot Q: Symmetric? Normal?
E qual variance	Plot residual vs X Q: Is there any structure?

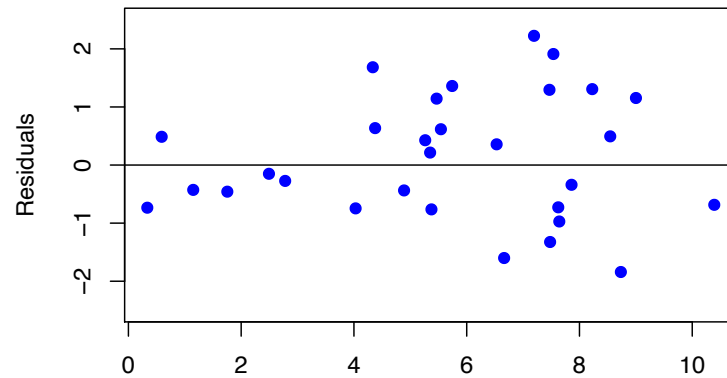
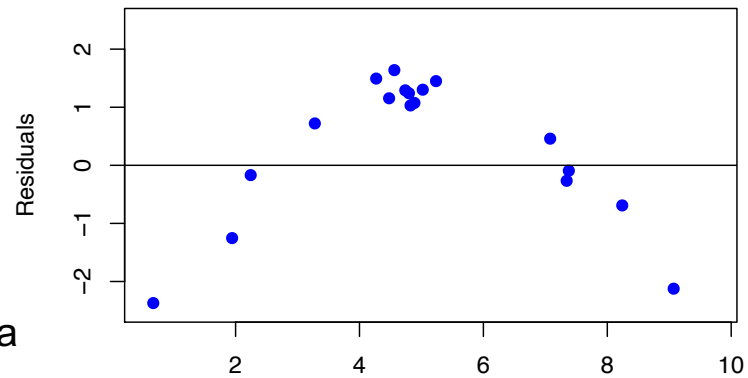
Model Checking: Residuals

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



Violations of the model assumptions? How?



Linearity

- 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



Independence

- 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

See, e.g., 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 non-constant 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?

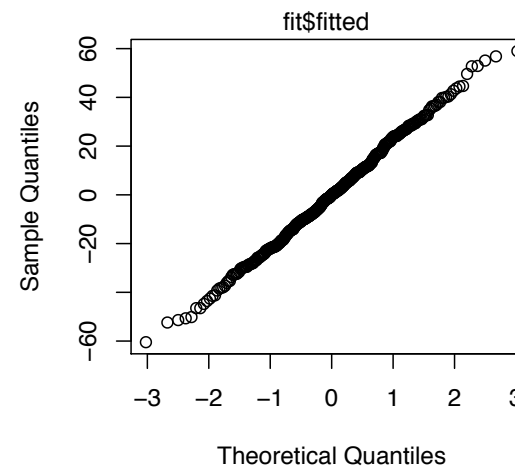
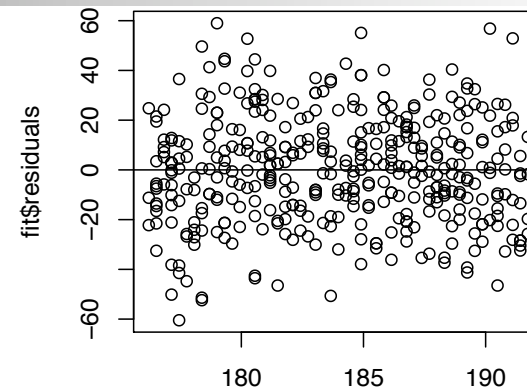
R COMMAND:

```
plot(fit$fitted, fit$residuals)
```

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

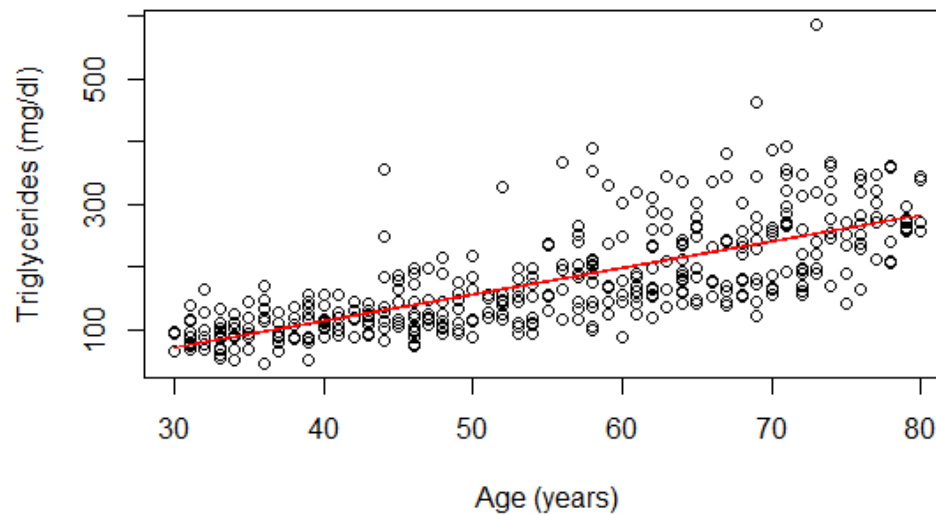
R COMMAND:

```
qqnorm(fit$residuals)
```



Another example

- Linear regression for association between age and triglycerides

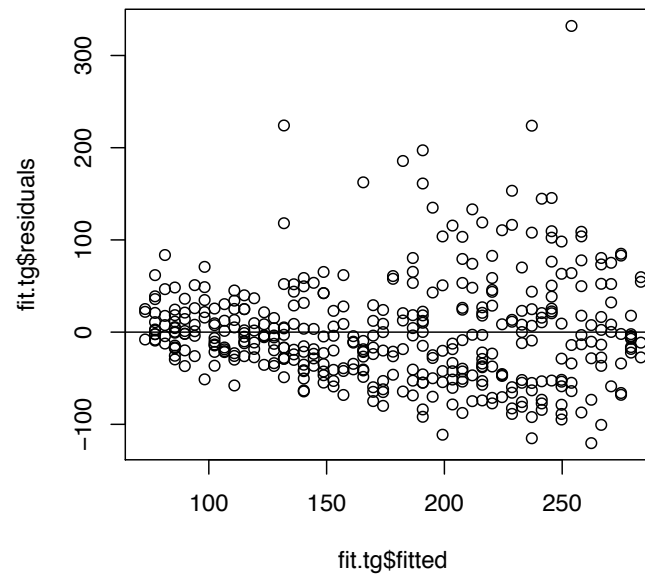


```
> fit.tg=lm(TG~age)
```




Robust standard errors

- Residual 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)

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:

```
> fit.tg.robust = coeftest(fit.tg, vcov = sandwich)
> fit.tg.robust

t test of coefficients:

```

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-53.30593	8.73874	-6.100	2.515e-09	***
age	4.20896	0.18134	23.211	< 2.2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Cholesterol example: Robust standard errors

- Linear regression results:

```
> summary(fit.tg)

Call:
lm(formula = TG ~ age)

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:

```
> fit.tg.robust = coeftest(fit.tg, vcov = sandwich)
> fit.tg.robust

t test of coefficients:
      Estimate Std. Error t value Pr(>|t|)
(Intercept) -53.30593     8.73874  -6.100 2.515e-09 ***
age           4.20896     0.18134  23.211 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Transformations

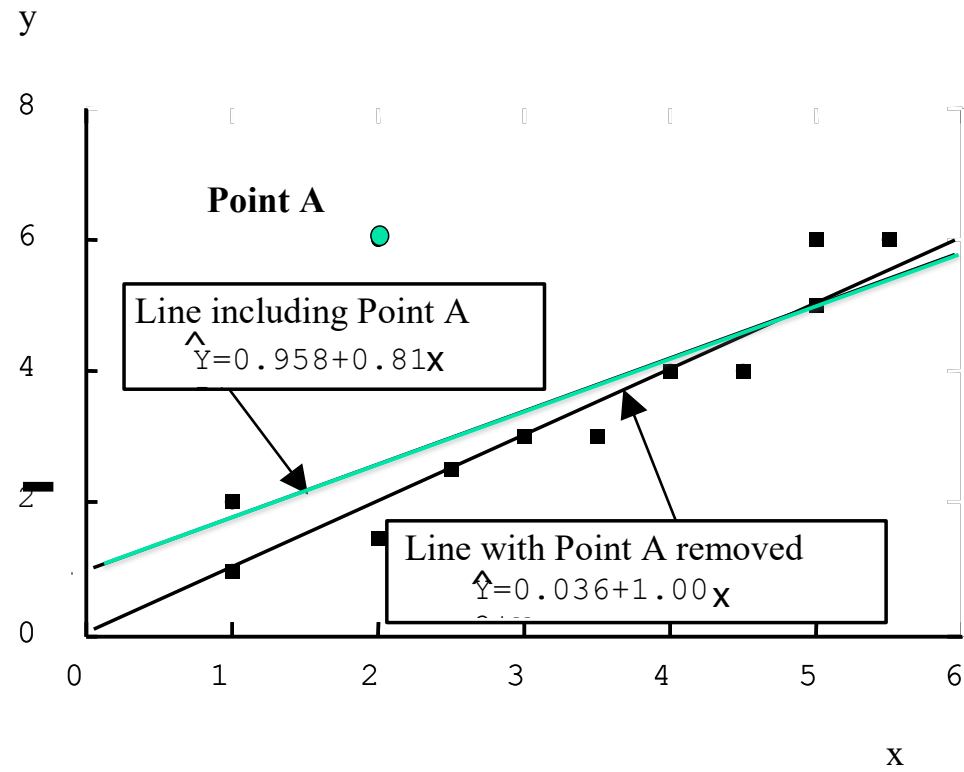
- Some reasons for using data transformations
 - Content area knowledge suggests nonlinearity
 - Original data suggest nonlinearity
 - Equal variance assumption violated
 - Normality assumption violated
- Transformations may be applied to the response, predictor or both
 - Be careful with the interpretation 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

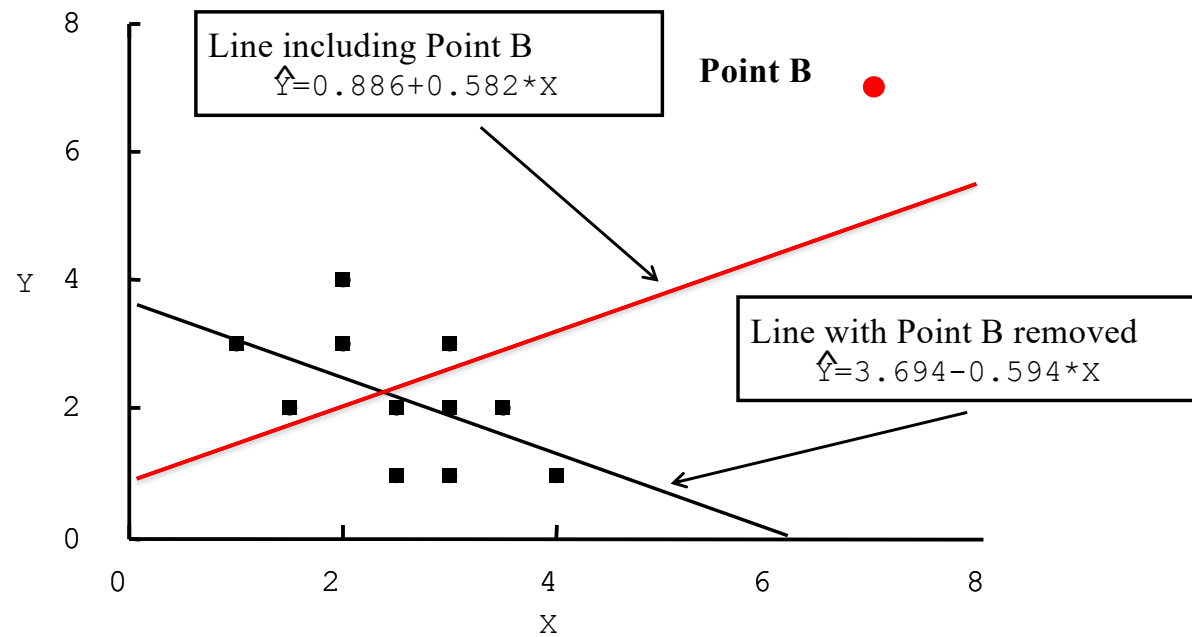
- **Outlier:** an observation with a residual that is unusually large (positive or negative) as compared to the other residuals.
- **Influential point:** an observation that has a notable influence in determining the regression equation.
 - Removing such a point would markedly change the position of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

Outlier vs Influential observations



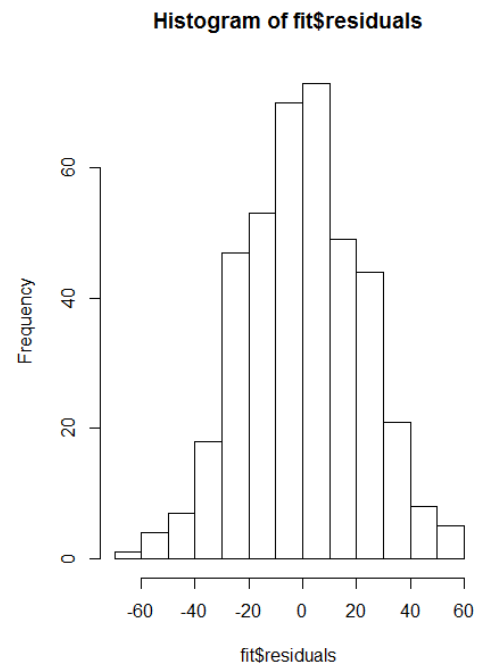
Point A is an outlier, but is not influential.

Outlier vs Influential observations



Point B is influential, but not an outlier.

Cholesterol-Age Example: Residuals



No extreme outliers



Model Checking: Deletion diagnostics

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

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

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

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



Cholesterol-Age Example: Deletion diagnostics

```
> dfb = dfbeta(fit)
> index=order(abs(dfb[,2]),decreasing=T)
> cbind(dfb[index[1:15],],age[index[1:15]])
```

	(Intercept)	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 Linearity	Non Normality	Unequal Variances	Dependence
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 regression framework)	Mostly no correction needed. Delete outliers (if warranted) or use robust regression	Use robust standard errors	Regression for dependent data