

# Unit 6: Simple Linear Regression

Statistics 102 Teaching Team

March 30, 2020

Introduction

Examining scatterplots

Least squares regression

Interpreting a linear model

Statistical inference in regression

# Introduction

## THE MAIN IDEAS

Linear regression provides methods for examining the association between a quantitative response variable and a set of possible predictor variables.

- Linear regression should only be used with data that exhibit linear or approximately linear relationships.

**Simple linear regression** is used to estimate the linear relationship between a response variable  $y$  and a single predictor  $x$ .

- The response variable  $y$  can be referred to as the *dependent* variable, and the predictor variable  $x$  the *independent* variable.
- The statistical model for simple linear regression is based on the straight line relationship

$$y = b_0 + b_1x$$

## THE MAIN IDEAS . . .

**Multiple linear regression** is used to estimate the linear relationship between a response variable  $y$  and several predictors  $x_1, x_2, \dots, x_p$ .

- The statistical model for multiple linear regression is based on

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p$$

Multiple regression will be covered in Unit 7.

## Examining scatterplots

# THE PREVEND STUDY

As adults age, cognitive function changes over time; largely due to various cerebrovascular and neurodegenerative changes.

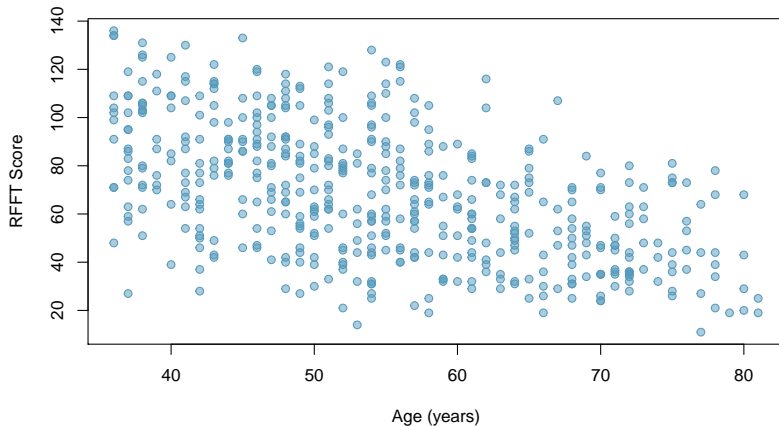
The Prevention of REnal and Vascular END-stage Disease (PREVEND) study measured various clinical and demographic data for participants in a series of surveys between 1997 - 2006.

- Data from 4,095 participants are in the `prevend` dataset in the `oibiostat` package.
- Cognitive function was assessed with the Ruff Figural Fluency Test (RFFT), which provides information about cognitive abilities such as planning and the ability to switch between different tasks.
  - Scores range from 0 to 175; higher scores indicate better cognitive function.

We will work with a random sample of 500 participants.

# AGE VS RFFT IN prevend.samp

**Association of RFFT Score with Age in the PREVENT data (n = 500)**





## AGE VS RFFT IN prevend.samp...

The relationship between age and RFFT score appears linear. A line might provide a useful summary of this association.

Lab 1 steps through fitting and interpreting a line as well as evaluating whether the assumptions for linear regression are satisfied.

# ASSUMPTIONS FOR LINEAR REGRESSION

There are 4 assumptions that should be satisfied for a line to be considered a reasonable approximation for a relationship shown in a scatterplot.

1. Linearity: the data show a linear trend.
2. Constant variability: the variability of the response variable about the line remains roughly constant as the predictor variable changes.
3. Independent observations: the  $(x, y)$  pairs are independent; i.e., values of one pair provide no information about values of other pairs.
4. Approximate normality of *residuals*: definition coming later...

Special plots for formally evaluating these assumptions are discussed in the next section.

## Least squares regression

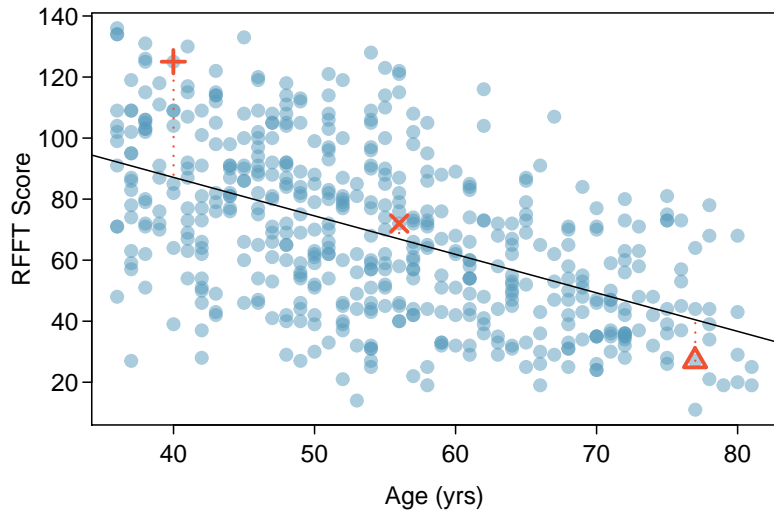
## RESIDUALS IN LINEAR REGRESSION

The vertical distance between a point in the scatterplot and the predicted value on the regression line is the **residual** for the point.

For an observation  $(x_i, y_i)$ , where  $\hat{y}_i$  is the predicted value according to the line  $\hat{y} = b_0 + b_1x$ , the residual is the value

$$e_i = y_i - \hat{y}_i$$

## RESIDUALS IN LINEAR REGRESSION...



## ESTIMATING A LINE USING LEAST SQUARES

The least squares regression line is the line which minimizes the sum of the squared residuals for all the points in the plot.

In other words, the least squares line is the line with coefficients  $b_0$  and  $b_1$  such that the quantity

$$e_1^2 + e_2^2 + \cdots + e_n^2$$

is the smallest, where  $n$  is the number of data points.

# STATISTICAL MODEL FOR LEAST SQUARES REGRESSION

For a general population of ordered pairs  $(x, y)$ , the **population regression model** is

$$y = \beta_0 + \beta_1 x + \epsilon,$$

where  $\epsilon \sim N(0, \sigma)$ .

- The error term  $\epsilon$  can be thought of as a population parameter for the residuals ( $e$ ).

Since the mean of  $\epsilon$  is 0, the population model can also be written as

$$E(Y|x) = \beta_0 + \beta_1 x,$$

where  $E(Y|x)$  denotes the expected value of  $Y$  when the predictor variable has value  $x$ .

## COEFFICIENTS OF THE LINE IN LEAST SQUARES REGRESSION

The terms  $\beta_0$  and  $\beta_1$  are parameters with estimates  $b_0$  and  $b_1$ . These estimates can be calculated from summary statistics.

$$b_1 = r \frac{s_y}{s_x} \quad b_0 = \bar{y} - b_1 \bar{x}$$

$\bar{x}$ ,  $\bar{y}$ : sample means of  $x$  and  $y$ .

$s_x$ ,  $s_y$ : sample standard deviations of  $x$  and  $y$ .

$r$ : correlation between  $x$  and  $y$ .



## USING R TO CALCULATE A LEAST SQUARES LINE

```
lm(prevend.samp$RFFT ~ prevend.samp$Age)$coef
```

```
##      (Intercept) prevend.samp$Age  
##      137.549716      -1.261359
```

The least squares line can be written as

$$\widehat{\text{RFFT}} = 137.55 - (1.26)(\text{Age})$$

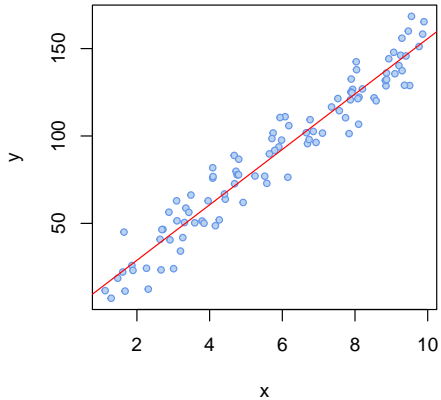
## CHECKING ASSUMPTIONS WITH RESIDUAL PLOTS

The assumptions in linear regression are linearity, constant variability, independent observations, and approximate normality of residuals.

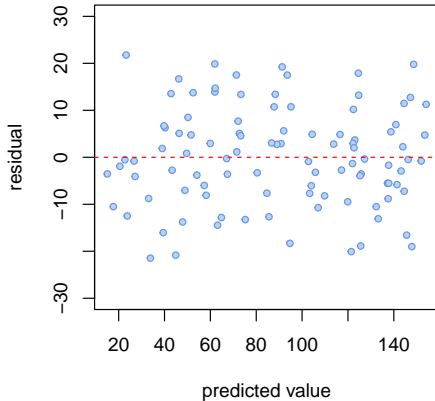
- The independence assumption has to be checked by considering study design.
- The other assumptions can be examined using *residual plots* and *normal probability plots*.
  - Residual plots: scatterplots in which predicted values are on the x-axis and residuals are on the y-axis
  - Normal probability plots: theoretical quantiles for a normal versus observed quantiles

# CHECKING LINEARITY AND CONSTANT VARIABILITY

**Y versus X**

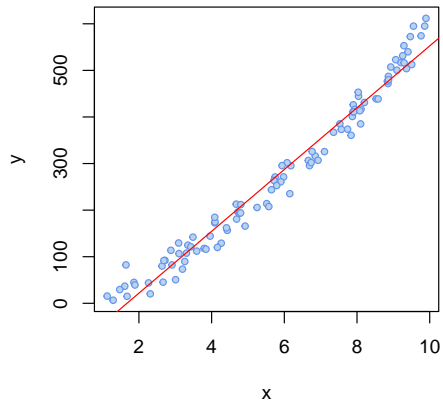


**Residual Plot of Y versus X**

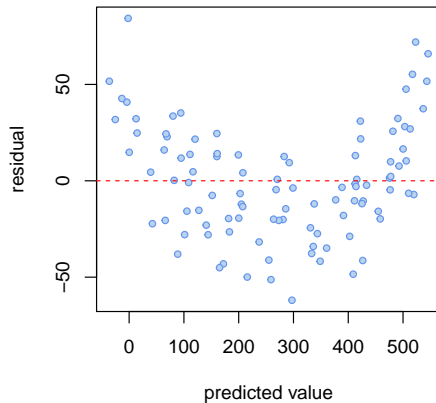


# CHECKING LINEARITY AND CONSTANT VARIABILITY...

**Y versus X**



**Residual Plot of Y versus X**



# CHECKING NORMALITY OF THE RESIDUALS

Normality of the residuals is fundamental to the underlying model

$$y = \beta_0 + \beta_1 x + \epsilon,$$

since  $\epsilon$  is assumed to be normally distributed with mean 0 and standard deviation  $\sigma$ .

Normality of the residuals is checked using normal probability plots.

- These plots were used to check the normality assumption in ANOVA.
- Normal probability plots are discussed in *Ol Biostat* Section 3.3.7.

## Interpreting a linear model

Categorical predictors with two levels

# CATEGORICAL PREDICTORS WITH TWO LEVELS

Although the response variable in linear regression is necessarily numerical, the predictor may be either numerical or categorical.

Simple linear regression only allows for categorical predictor variables with two levels.

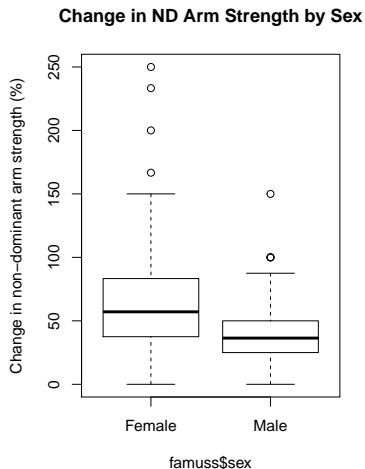
- Examining categorical predictors with more than two levels requires multiple linear regression.

Fitting a simple linear regression model with a two-level categorical predictor is analogous to comparing the means of two groups, where the groups are defined by the categorical variable.



## FAMuSS: COMPARING NDRM.CH BY SEX

Let's re-examine the association between change in non-dominant arm strength after resistance training and sex in the FAMuSS data.



## FAMUSS: COMPARING NDRM.CH BY SEX...

```
#calculate mean ndrm.ch in each group  
tapply(famuss$ndrm.ch, famuss$sex, mean)
```

```
##      Female      Male  
## 62.92720 39.23512
```

```
#fit a linear model of ndrm.ch by sex  
lm(famuss$ndrm.ch ~ famuss$sex)$coef
```

```
##      (Intercept) famuss$sexMale  
##      62.92720      -23.69207
```

$$\widehat{ndrm.ch} = 62.93 - 23.59(sexMale)$$

- The intercept is the mean of one category, the baseline category.
- The slope is the difference between the means.

Using  $R^2$  to describe the strength of a fit

## THE QUANTITY $R^2$

The correlation coefficient  $r$  measures the strength of the linear relationship between two variables.

- It is more common to use  $r^2$  to measure the strength of a linear fit, which is written as  $R^2$  in the context of regression.

$R^2$  describes the amount of variation in the response that is explained by the least squares line.

$$R^2 = \frac{\text{variance of predicted } y\text{-values}}{\text{variance of observed } y\text{-values}} = \frac{\text{Var}(\hat{y}_i)}{\text{Var}(y_i)}.$$

If a linear model perfectly captured the variability in the observed data, then  $\text{Var}(\hat{y}_i)$  would equal  $\text{Var}(y_i)$  and  $R^2$  would be 1.

## THE QUANTITY $R^2$ ...

$R^2$  can also be calculated using the following formula:

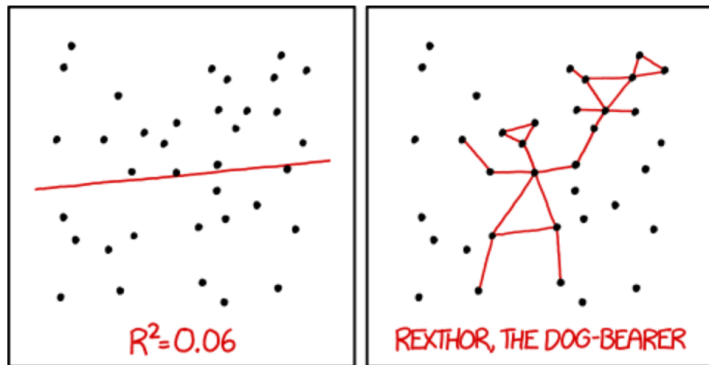
$$R^2 = \frac{\text{variance of observed } y\text{-values} - \text{variance of residuals}}{\text{variance of observed } y\text{-values}} = \frac{\text{Var}(y_i) - \text{Var}(e_i)}{\text{Var}(y_i)}$$

The variability of the residuals about the line represents the remaining variability after the model is fit.

- In other words,  $\text{Var}(e_i)$  is the variability unexplained by the model.

Lab 4 explores the idea behind  $R^2$  and provides an example of using  $R^2$  to assess the strength of a model fit.

## XKCD DOES REGRESSION



I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER  
TO GUESS THE DIRECTION OF THE CORRELATION FROM THE  
SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

## Outliers in regression

## OUTLIERS IN REGRESSION

Depending on their position, data points in a scatterplot have varying degrees of contribution to the estimated coefficients of a regression line.

- Points with particularly high or low values of the predictor are said to have high **leverage**, and have a large effect on  $b_0$  and  $b_1$ .

A data point is considered a **regression outlier** if its value for the response variable does not follow the general linear trend in the data.

If an observation has a strong effect on the estimated coefficients, such that the estimates change substantially when the observation is omitted, it is considered **influential**.

- Outliers with high leverage may be influential.

Advanced courses in regression have formal definitions for these terms.

For the purposes of Stat 102, it is important to simply be able to identify influential points and comment on whether it might be best to exclude them from the data.



# INFANT MORTALITY AND NUMBER OF DOCTORS

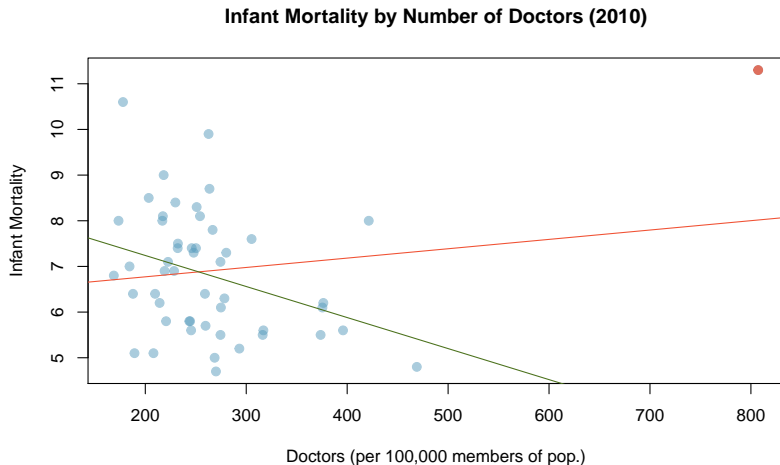
The following slide shows a scatterplot of infant mortality versus number of doctors, where the observations are for each state and the District of Columbia.<sup>1</sup>

- Infant mortality is measured as number of infant deaths in the first year of life per 1,000 births.
- Number of doctors is recorded as number of doctors per 100,000 members of the population.

---

<sup>1</sup>Data from US Census Records, 2010. Available in the `oibiostat` package as the `census.2010` dataset.

# INFANT MORTALITY AND NUMBER OF DOCTORS...



The **red** line is the model fit to all 51 observations.

The **green** line is the model fit to 50 observations, excluding the red point.

## INFANT MORTALITY AND NUMBER OF DOCTORS...

```
#identify the influential point
```

```
census.2010$state[census.2010$doctors > 700]
```

```
## [1] "District of Columbia"
```

The point marked in red corresponds to the District of Columbia, where there were approximately 800 doctors per 100,000 members of the population, and the infant mortality rate was 11.3 per 1,000 live births.

- This observation has high leverage, since 800 is a high value for the predictor variable, number of doctors.
- This observation is an outlier, since it has an unusually high  $y$ -value paired with a high  $x$ -value; the other points show a negative association where higher  $x$ -values tend to have low  $y$ -values.
- The observation is influential; its inclusion substantially affects the estimated coefficients, reversing the sign of the slope.

## Statistical inference in regression

# THE MODEL FOR STATISTICAL INFERENCE

The observed data  $(x_i, y_i)$  are assumed to have been randomly sampled from a population where the explanatory variable  $X$  and the response variable  $Y$  follow a population model

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

where  $\epsilon \sim N(0, \sigma)$ .

Under this assumption, the slope and intercept of the regression line,  $b_0$  and  $b_1$ , are estimates of the population parameters  $\beta_0$  and  $\beta_1$ .

# HYPOTHESIS TESTING IN REGRESSION

Inference in a regression context is usually about the slope parameter,  $\beta_1$ .

The null hypothesis is most commonly a hypothesis of 'no association':

- $H_0 : \beta_1 = 0$ , the  $X$  and  $Y$  variables are not associated
- $H_A : \beta_1 \neq 0$ , the  $X$  and  $Y$  variables are associated

The  $t$ -statistic has degrees of freedom  $n - 2$ , where  $n$  is the number of ordered pairs in the dataset.

$$t = \frac{b_1 - \beta_1^0}{\text{s.e.}(b_1)} = \frac{b_1}{\text{s.e.}(b_1)}$$

The value  $\beta_1^0$  equals 0 when the null hypothesis is one of no association.

## CONFIDENCE INTERVALS IN REGRESSION

A 95% confidence interval for  $\beta_1$  has the following formula

$$b_1 \pm (t^* \times \text{s.e.}(b_1)),$$

where  $t^*$  is the point on a  $t$ -distribution with  $n - 2$  degrees of freedom and  $\alpha/2$  area to the right.

## THE STANDARD ERRORS OF $b_0$ AND $b_1$

Formulas for calculating the standard errors of the model coefficients are in Section 6.4 of *OI Biostat*.

In practice, statistical software like R is used to obtain  $t$ -statistics and  $p$ -values for linear models.