Statistics 305/605: Introduction to Biostatistical Methods for Health Sciences

Chapter 18, part 1 (including Demo): Simple Linear-Regression Models

Jinko Graham

Response and Explanatory Variables

- In simple linear regression,
  - The **response** variable, Y, measures the outcome.
  - The **explanatory** variable(s), X, are there to explain the outcome.

Example

- Recall the study of head circumference in 100 infants with birth weight less than 1500g.
  - Variables included head circumference (cm) and gestational age (weeks), among others.

```
uu <- url("http://people.stat.sfu.ca/~jgraham/Teaching/S305_17/Data/lbwt.csv")
lbwt <- read.csv(uu)
head(lbwt)
```

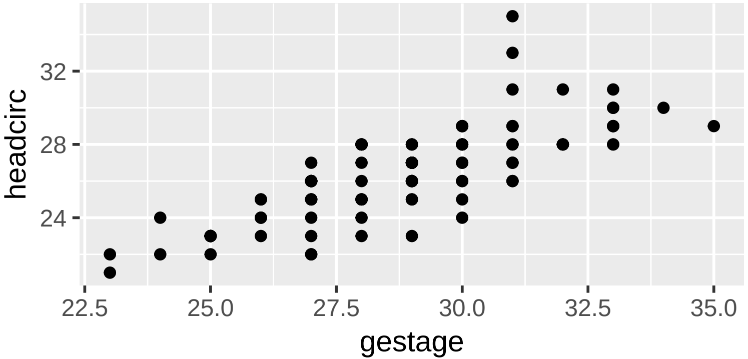
```
##   headcirc length gestage birthwt momage toxemia
## 1       27     41      29    1360     37       0
## 2       29     40      31    1490     34       0
## 3       30     38      33    1490     32       0
## 4       28     38      31    1180     37       0
## 5       29     38      30    1200     29       1
## 6       23     32      25     680     19       0
```

- Let's view **head circumference** (`headcirc`) as the response variable, Y, with observed measurements denoted y.
- Use **gestational age** (`gestage`) as an explanatory variable, X, with observed values x.

Scatterplot of the Low Birthweight Data

- There appears to be a linear relationship between Y and X:

```
library(ggplot2)
ggplot(lbwt,aes(x=gestage,y=headcirc)) + geom_point()
```

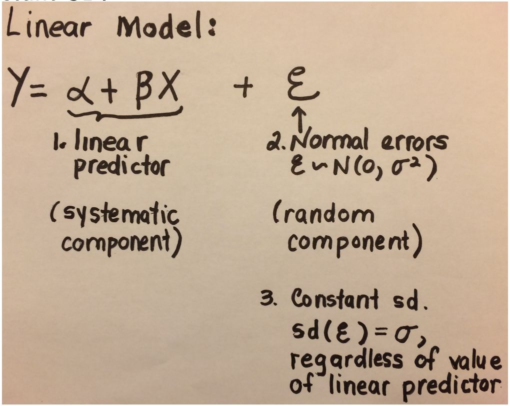


Linear Regression

- If we have response and explanatory variables, we may summarize a linear relationship by a **regression line** through the scatterplot.
- The regression line describes how the average value of Y changes as X changes.
  - Specifically, the line models the **population mean** of Y given that $X = x$.
- We use the method of least squares to fit or estimate the line from our sample of data.
- Under modelling assumptions, we can:
  - infer the slope of the regression line in the population from the slope fitted in our sample, and
  - make predictions from the model we have fitted to our data.
- Model assumptions are checked *after* the model is fit to our sample of data.

Model Overview

- The components of the statistical model are:
  1. the linear predictor,
  2. normal error terms,
  3. constant SD.



- Will discuss each component.
- In addition, we assume that the observations are **independent**.

Linear Predictor

- When there is a linear relationship between Y and X, the conditional mean of Y given $X = x$ in the population, denoted $\mu_{y|x}$, is modelled by a <u>line</u>:

$$\mu_{y|x} = \alpha + \beta x,$$

- Think of $\mu_{y|x}$ as the population mean value of Y for all data with $X = x$.
- β is the change in $\mu_{y|x}$ for a one-unit increase in x.

Normal Errors, Constant SD

- Observed values of y will not fall perfectly along a line.
- Deviations of the y's from the line are called errors.
- Write $y = \alpha + \beta x + \epsilon$ where ϵ is the error term.
- Errors are assumed to be normally distributed with mean zero and SD $\sigma_{y|x}$.
- The SD of the error terms is assumed to be constant for all x; i.e. $\sigma_{y|x} = \sigma_y$

Model Summary

- We can summarize the model assumptions by saying that:
  1. the (X, Y) pairs are independent;
     - i.e., for individual i with measurements (X_i, Y_i) and a different individual j with measurements (X_j, Y_j), knowing i's measurements tells us nothing about what j's are, and vice versa.
  2. conditional on $X = x$, the outcome Y has a normal distribution $N(\mu_{y|x}, \sigma_{y|x})$, with
     - mean $\mu_{y|x} = \alpha + \beta x$, and
     - SD $\sigma_{y|x}$ being the same for all x, so that $\sigma_{y|x} = \sigma_y$.

Fitting the Model

- Goal: Let's use the observed data on the n individuals — $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ — to fit the model

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i,$$

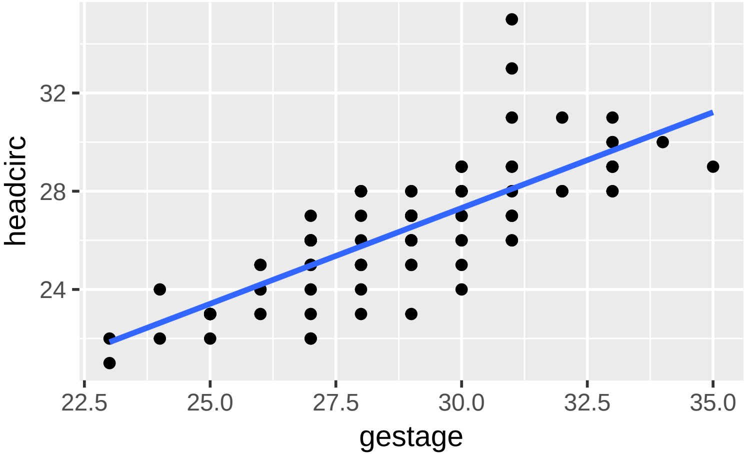
  where $\hat{y}_i$ is the **predicted** or **fitted value** of Y for $X = x_i$.

- Idea: Try all possible $\hat{\alpha}$ and $\hat{\beta}$, until we find the line that fits the data the "best" in the sense that the $\hat{y}$'s are as close to the y's as possible.

- Need to explore the criteria for "best" ...

Vertical Distance

- Here is a plot of the data from the low-birth-weight babies study:



- By comparing y to $\hat{y}$, we are measuring the vertical distance between points in the scatterplot and the regression line.

Vertical Distance

- Question: How should we summarize vertical distances between the points, y, and the regression line, $\hat{y}$?
- We will discuss the method that minimizes the sum of squared distances, or least squares.
- There are many visual demonstrations of the least squares idea on the internet; e.g., http://www.dangoldstein.com/regression.html
  - The sum of squared distances between the y's and their $\hat{y}$'s is summarized by the blue square in this demo.
  - To minimize the sum of squared distances, try clicking the buttons for
    - $-$ slope, $+$ slope,
    - $-$ intercept, $+$ intercept.
  - Then click "Fit and lock" to see the line that minimizes the sum of squares.

Least-Squares Regression

- We choose the regression line to minimize the squares of the discrepancies $y - \hat{y}$; i.e, to

$$\text{minimize} \quad Q = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

- The line that minimizes Q has

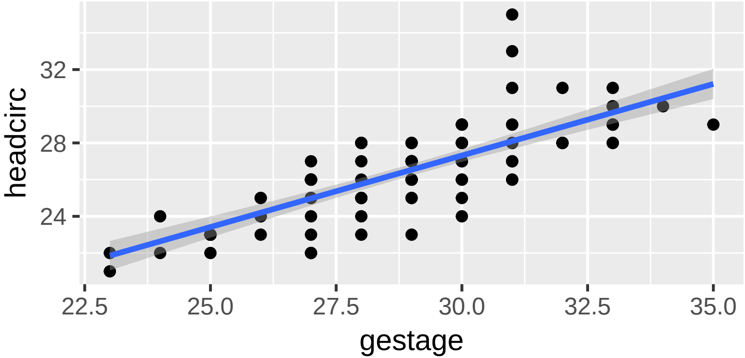
$$\begin{aligned} \hat{\beta} &= r\,\frac{s_y}{s_x} \\ \hat{\alpha} &= \overline{y} - \hat{\beta}\,\overline{x}, \end{aligned}$$

  where $r, s_y, s_x, \overline{y}$ and $\overline{x}$ are, respectively:
    - the sample correlation (see your prereq. Stats class), the sample SD of y, the sample SD of x, the sample mean of y and the sample mean of x.
- However, we'll use computer software to get the least-squares estimates of the parameters α and β.

Example

- We can superpose the least-squares regression line onto our initial scatterplot of head circumference vs. gestational age, as follows:

```
ggplot(lbwt,aes(x=gestage,y=headcirc)) + geom_point() +
  geom_smooth(method="lm")
```



Software Notes

- overlaying `geom_smooth()` adds a curve to the plot that summarizes the trends and is called a *scatterplot smoother*
  - the argument `method=lm` specifies that the smoother should be the least squares regression line.
- The grey shaded region around the regression line is a *point-wise confidence interval* for the population means $\mu_{y|x}$: more on these later.

Fitted Model and Coefficients

- To fit the model in R, we will use the `lm()` function and put the resulting fitted-model into an R object called `lfit`.

```
lfit <- lm(headcirc ~ gestage,data=lbwt)
names(lfit)
```

```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
```

- Let's see what the fitted coefficients are that estimate the population intercept α and the population slope β.

```
coefficients(lfit)
```

```
## (Intercept)     gestage
##   3.9142641   0.7800532
```

- The estimated intercept and slope are $\hat{\alpha} = 3.9$ and $\hat{\beta} = 0.78$.
    - A one week increase in gestational age is associated with an estimated 0.78cm increase in head circumference.

Software Notes

- `lm()` is the R function that fits linear models to data by the least-squares method of minimizing the sum of squared vertical distances between the y's and their $\hat{y}$'s.
- `lm()` uses formulas to specify the response and explanatory variables.
  - e.g., in the call to `lm()`, we specify
    `lfit <- lm(headcirc ~ gestage,data=lbwt)`
    and the formula being used is `headcirc ~ gestage`
  - the response variable, `headcirc` is on the left-hand side of the formula, to the left of `~`.
  - the explanatory variable, `gestage` is on the right-hand side of the formula, to the right of `~`.
- Extract the fitted coefficients with the `coefficients()` function; i.e.

```
coefficients(lfit)
```

```
## (Intercept)     gestage
##   3.9142641   0.7800532
```