Biostat705 course expectations:

- Class attendance is required for this course. Participation in-class discussions is encouraged, if something is not clear, ask questions, don't fall behind on class materials!
- Working together on HW assignments and class projects is OK, but the report write-up should be yours alone.
- There is no official textbook for this class, however materials posted on Sakai as class lecture slides, hand-written notes, write-up on the class white-board as well as Microsoft Teams discussions are expected topics/questions showing-up on the quizzes and exams.
- All questions related to 705 outside the class-room need to be posted on Microsoft Teams, and will be addressed in a timely manner by the instructor or the TAs (and sometimes by your classmates).
- There is no make-up quiz or exam in this class.
- All quizzes and exams are closed notes, in-class and proctored, no questions are allowed during the quizzes or exams.
- You will need to bring a laptop or iPad (not BOTH) to take a quiz or an exam via Sakai. You can bring a calculator or use R as a calculator.
- ALL quizzes and exams has to be uploaded to Sakai in a PDF format file (no JPG or TIFF etc.) and must be in the following PDF filename, quizx_Lastname_Firstname.pdf, for example quiz1_Smith_joe.pdf, midterm_Smith_joe.pdf and final_Smith_joe.pdf.
- All quizzes/exams are graded uniformly among the students, ie same points are taken for similar wrong answers. Thus, detected points will not be discussed, unless the final score is not summed correctly.

Simple Linear Regression: Review

Biostat 705

Hussein R. Al-Khalidi, PhD

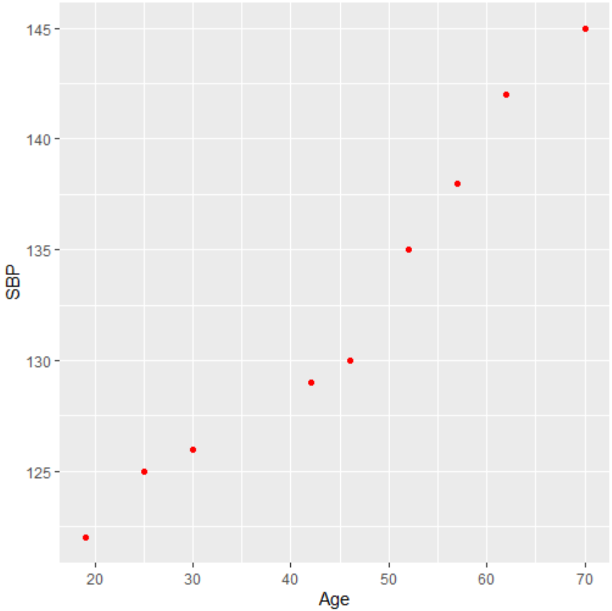
Department of Biostatistics and Bioinformatics,
Duke University

January 10, 2024

Below data (age and sbp) are measured on 9 subjects so that relationship between SBP (response) and age (independent variable) can be studied:

Obs	Age	SBP
1	19	122
2	25	125
3	30	126
4	42	129
5	46	130
6	52	135
7	57	138
8	62	142
9	70	145

Example1: SBP and Age

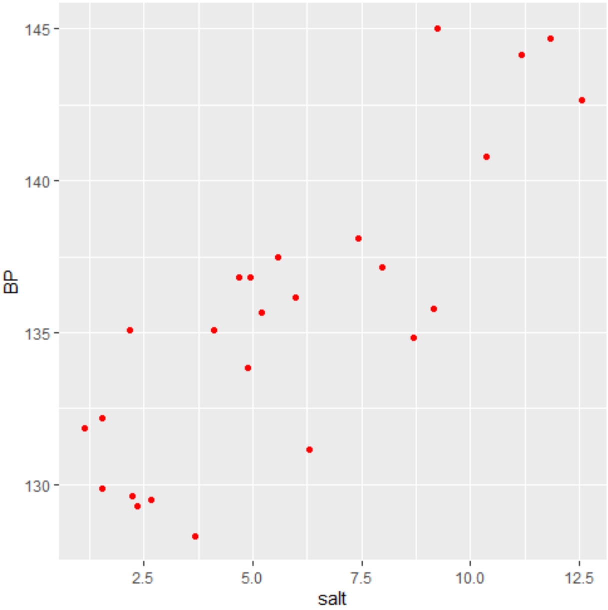


Below data (SBP, salt, saltLevel [sodium chloride diet (<6 vs. ≥ 6 grams of salt per day)]) are measured on 25 elderly subjects so that relationship between SBP as a response and salt (or saltLevel) as an independent variable can be studied:

SBP	salt	saltLevel
132.19	1.55	0
131.84	1.13	0
133.86	4.88	0
135.08	4.11	0
129.85	1.55	0
136.84	4.69	0
135.10	2.16	0
129.61	2.23	0
129.51	2.65	0
128.30	3.68	0
129.29	2.34	0
136.14	5.98	0
137.50	5.59	0
135.65	5.21	0
136.83	4.95	0
135.79	9.15	1
138.12	7.43	1
144.67	11.84	1
131.13	6.31	1
140.78	10.36	1
144.13	11.18	1
137.17	7.98	1
145.02	9.24	1
142.64	12.57	1
134.84	8.68	1

DukeHealth

Example2: SBP and Salt



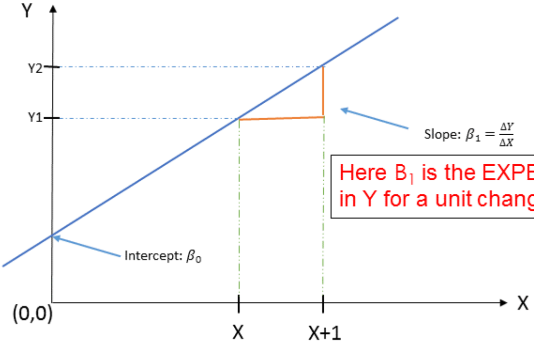
DukeHealth

Simple linear regression model

- Simple linear regression model:
  $Y = \beta_0 + \beta_1 X + \epsilon$, (population regression line)
  where β_0 and β_1 are called *parameters* (unknown) and are fixed.
- Linear regression model is linear in parameters; for example
  $log(Y) = \beta_0 + \beta_1 \sqrt{X} + \epsilon$, is still a linear model, but
  $Y = \beta_0 + \sqrt{\beta_1} X + \epsilon$, is not linear in parameters, thus it's not a linear model.
- β_0 = intercept; value of Y when X=0,
- β_1 = slope; change in Y for every unit change in X,
- Y is the dependent (response) variable,
- X is the independent (predictor) variable.

DukeHealth

Simple Linear Model



Here B_1 is the EXPECTED change in Y for a unit change in X.

Simple Linear Model with Error

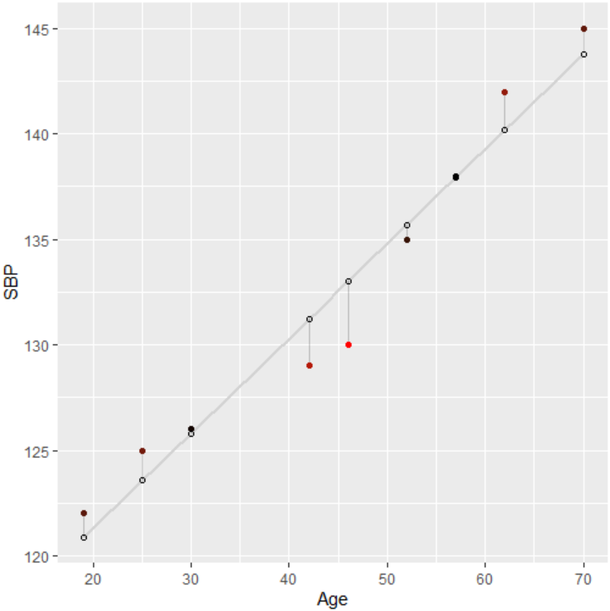
- Linear regression

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

$$\epsilon = Y - \beta_0 - \beta_1 X$$

- ϵ equals vertical distance from Y to line defined by $\beta_0 + \beta_1 X$
- Residual e equals vertical distance from y (observed) to line defined by $\hat{\beta}_0 + \hat{\beta}_1 x$, ie $e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$; where $\hat{\beta}_0$ and $\hat{\beta}_1$ are called *statistics* (known) and are <u>variable</u>.

DukeHealth



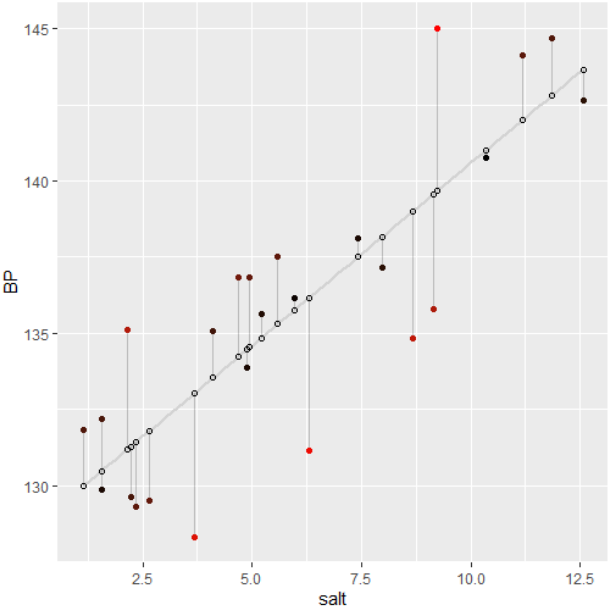
DukeHealth

```
library(MASS)
library(ggplot2)
sbp <- read.table("C:\\FILE PATH\\sbp.txt", header=T)

fit2 <- lm(SBP ~ Age, data = sbp)
sbp$predicted <- predict(fit2)   # Save the predicted values
sbp$residuals <- residuals(fit2) # Save the residual values

ggplot(sbp, aes(x = Age, y = SBP)) +
  geom_smooth(method = "lm", se = FALSE, color = "lightgrey") +
  geom_segment(aes(xend = Age, yend = predicted), alpha = .2) +

  # > Color adjustments made here...
  geom_point(aes(color = abs(residuals))) + # Color mapped to abs(residuals)
  scale_color_continuous(low = "black", high = "red") +  # Colors to use here
  guides(color = FALSE) +  # Color legend removed
  geom_point(aes(y = predicted), shape = 1)
```



DukeHealth

Model Assumptions

- Data are (y_i, x_i); $i = 1, 2, \ldots, n$
- Assume:
  1. Linearity: $Y = \beta_0 + \beta_1 X + \epsilon$
  2. $X^{'}$s are fixed constants
  3. ϵ_i iid $N(0, \sigma^2)$
  4. Constant variance, $\mathrm{Var}(\epsilon_i) = \sigma^2$
  5. Note: Normality assumption is not needed to estimate linear regression parameters. However, normality is needed to make inference about the model parameters, such as testing hypotheses, confidence intervals, etc.

DukeHealth

Under simple regression model, we would like to minimize the residual sum of squares (RSS) (ie., $\min \sum_{i=1}^{n} e_i^2$). This is called least-square estimation and based on a random sample of data points this yields the least-square estimates (LSE) $\hat{\beta}_0$ and $\hat{\beta}_1$. That is,

$$\text{RSS} = e_1^2 + e_2^2 + \ldots + e_n^2,$$

or equivalently

$$\text{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

DukeHealth

Least square estimates

The least squares approach leads to below $\hat{\beta}_0$ and $\hat{\beta}_1$, that minimize the RSS,

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

and,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

DukeHealth

Estimated regression line

- Estimated simple linear regression line:

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

- Estimate error variance:

$$\hat{\sigma}^2 = S^2_{y.x} = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- Observed residuals:

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

Linear regression inference

- Estimated slope variance: $\widehat{\text{var}}(\hat{\beta}_1) = \dfrac{S_{y.x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
- Estimated intercept variance:
  $\widehat{\text{var}}(\hat{\beta}_0) = S_{y.x}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$

DukeHealth

CI and Hypotheses Tests

- A key question when fitting regression lines is whether or not the slope of the line is really different (statistically different) from a flat line ($\beta_1 = 0$)

In inferences on parameters β_1 and β_0, we must assume normality [ie., $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$] to perform testing hypotheses on β_1 and β_0.

$$\mathrm{H}_0 : \beta_1 = \beta_{01} (\beta_{01} \text{ can be any value, in most cases is 0})$$
$$\mathrm{H}_\mathrm{a} : \beta_1 \neq \beta_{01}$$

$t = \frac{(\hat{\beta}_1 - \beta_{01})}{\mathrm{SE}(\hat{\beta}_1)} \sim t_{n-2}$; where $\mathrm{SE}(\hat{\beta}_1) = \sqrt{\widehat{\mathrm{var}(\hat{\beta}_1)}}$
note: $t^2_{n-2} = F_{1,n-2}$
A 95% CI on β_1 is given as : $\hat{\beta}_1 \pm t_{0.975,n-2}\mathrm{SE}(\hat{\beta}_1)$.

DukeHealth

Example 1 in R

```
> fit=lm(SBP ~ Age)
> summary(fit)

Call:
lm(formula = SBP ~ Age)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9934 -0.6884  0.1933  1.2265  1.8199

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 112.33169    1.73773   64.64 5.57e-11 ***
Age           0.44917    0.03644   12.32 5.31e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.792 on 7 degrees of freedom
Multiple R-squared:  0.9559,    Adjusted R-squared:  0.9497
F-statistic: 151.9 on 1 and 7 DF,  p-value: 5.313e-06
```

DukeHealth

Example 1 in SAS

```
proc reg data=example1;
   model sbp=age;
run;
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	487.74667	487.74667	151.91	<.0001
Error	7	22.47555	3.21079		
Corrected Total	8	510.22222			

Root MSE	1.79187	R-Square	0.9559
Dependent Mean	132.44444	Adj R-Sq	0.9497
Coeff Var	1.35292		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	112.33169	1.73773	64.64	<.0001
Age	1	0.44917	0.03644	12.33	<.0001

DukeHealth

Example 1: Interpretation

- $\hat{\beta}_1 = 0.45 \rightarrow$ Assuming the model is either correct or reasonably close, this implies that if you compared two people, one with an age that was one year older than the other, we would expect, on average, that the person with the older age would have a systolic blood pressure that is 0.45 (mmHg) higher.
- $\hat{\beta}_0 = 112.3 \rightarrow ?$ depends on whether it is measuring something meaningful or not.
- $\hat{\beta}_0$ is the predicted $\hat{y}$ at $x = 0$.
  a) Is $x = 0$ within the range of your data? If so, then $\hat{\beta}_0$ is interpreted as that predicted value.
  b) If $x = 0$ is NOT within the range of the data, then $\hat{\beta}_0$ is merely a centering constant. Its role is to shift the line vertically (up and down) so that it falls in the midst of all the data.

DukeHealth

Example 2 in R

```
> fit=lm(BP ~ salt)
> summary(fit)

Call:
lm(formula = BP ~ salt)

Residuals:
    Min      1Q  Median      3Q     Max
-5.0388 -1.6755  0.3662  1.8824  5.3443

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  128.616      1.102 116.723  < 2e-16 ***
salt           1.197      0.162   7.389 1.63e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.745 on 23 degrees of freedom
Multiple R-squared:  0.7036,    Adjusted R-squared:  0.6907
F-statistic: 54.59 on 1 and 23 DF,  p-value: 1.631e-07
```

Example 2 in SAS

```
proc reg data=example2;
   model BP=salt;
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: BP

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	411.47701	411.47701	54.59	<.0001
Error	23	173.35282	7.53708		
Corrected Total	24	584.82982			

Root MSE	2.74537	R-Square	0.7036
Dependent Mean	135.67520	Adj R-Sq	0.6907
Coeff Var	2.02349		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	128.61640	1.10189	116.72	<.0001
salt	1	1.19689	0.16199	7.39	<.0001

DukeHealth

Example 2: Interpretation

- $\hat{\beta}_1 = 1.2 \rightarrow$, which means the systolic blood pressure (BP) increases by 1.2 (mmHg) for every 1 gram increase in daily sodium chloride intake (salt).
- $\hat{\beta}_0 = 128.6 \rightarrow ?$ depends on whether it is measuring something meaningful or not.
- $\hat{\beta}_0$ is the predicted $\hat{y}$ at $x = 0$.

DukeHealth

Total variability & Sum of Squares Partition

- The total variability $Y_i - \overline{Y}$ can be partitioned as: $\underbrace{(\widehat{Y}_i - \overline{Y})}_{1} + \underbrace{(Y_i - \widehat{Y}_i)}_{2}$. The first term is due to regression and the 2nd term is due to error, thus

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2 + \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$$

$$SST = SSreg + SSE$$

- Total sample variance of the Y's (ignoring the X's, ie no regression model)

$$s_y^2 = \frac{SST}{n-1} = \frac{\sum_i (Y_i - \bar{Y})^2}{n-1}$$

- Similarly, one can partition the total degrees of freedom (df_T), that is $df_T(n-1) = df_{reg}(1) + df_{error}(n-2)$. where n is total observations; 2 is number of parameters in the regression model.

R^2

- Coefficient of determination (R^2) is defined as:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- R^2 explains the total variability due to regression model
- In example 1,

$$R^2 = \frac{487.75}{510.22} = 0.9559$$

model, explains about 96% of total variability.

- In example 2,

$$R^2 = \frac{411.48}{584.83} = 0.7036$$

Having salt in the model, explains 70% of total variability.

Adjusted R^2

- In example 1, the sample variance of the Y's is s_y^2= 63.78 while $s_{y.x}^2$=3.21
- Thus X "explains"

$$\frac{63.78 - 3.21}{63.78} = 0.9497$$

  proportion of the variance in Y.
- In example 2:

$$\frac{24.37 - 7.54}{24.37} = 0.6906$$

- This quantity is called the adjusted R^2

$$R_a^2 = \frac{s_y^2 - s_{y.x}^2}{s_y^2} = 1 - \frac{s_{y.x}^2}{s_y^2} = 1 - \frac{SSE/(n-2)}{SST/(n-1)}$$

DukeHealth

R^2 and adjusted R^2

- Note
$$R_a^2 = 1 - \frac{SSE/(n-2)}{SST/(n-1)}$$
and
$$R^2 = 1 - \frac{SSE}{SST}$$
- Implying
$$R_a^2 = 1 - \frac{n-1}{n-2}(1 - R^2)$$
- Thus $R^2 \approx R_a^2$ for large n

- Proportion of total variation attributable to regression
- Degree of linear association
- Ranges between 0 and 1
- $R^2 = 0 \rightarrow$ no linear association between X and Y; However, a non-linear association may still exist!
- $R^2 = 1 \rightarrow$ indicates perfect fit

Least square estimates

In a matrix from:

$$\overbrace{\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}}^{Y} = \overbrace{\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}}^{X} \overbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}}^{\boldsymbol{\beta}} + \overbrace{\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{bmatrix}}^{\epsilon}$$

Thus, the estimated β's are given as:

$$\hat{\beta} = (X^{'}X)^{-1}X^{'}Y$$

Least square estimates

- $\hat{\beta} = (X^{'}X)^{-1}X^{'}Y$
- $\mathrm{Var}(\hat{\beta}) = \sigma^2(\mathrm{X}^{'}\mathrm{X})^{-1}$ and $\widehat{\mathrm{Var}(\hat{\beta})} = S^2_{y.x}(X^{'}X)^{-1}$
  thus, $\widehat{\mathrm{Var}(\hat{\beta})} = \mathrm{MSE}(\mathrm{X}^{'}\mathrm{X})^{-1}$