Advanced Statistics
DS2003 (BDS-4A)
Lecture 20

Instructor: Dr. Syed Mohammad Irteza

Assistant Professor, Department of Computer Science, FAST

28 April, 2022

Previous Lecture

- Use of Python for linear regression analysis:
  - Example with real sample data
  - Example with randomly generated values
- Multiple Linear Regression

Today

- Revision of Linear Regression
  - Sum of squares for x and y
  - Sum of products for x, y
- Least squares regression model
  - Residuals sum to zero, and the line always passes through $(\bar{x}, \bar{y})$

Weights of books

	weight (g)	volume (cm^3)	cover
1	800	885	hc
2	950	1016	hc
3	1050	1125	hc
4	350	239	hc
5	750	701	hc
6	600	641	hc
7	1075	1228	hc
8	250	412	pb
9	700	953	pb
10	650	929	pb
11	975	1492	pb
12	350	419	pb
13	950	1010	pb
14	425	595	pb
15	725	1034	pb

Modeling weights of books using volume

somewhat abbreviated output...

```
Coefficients:
             Estimate  Std. Error t value  Pr(>|t|)
(Intercept) 107.67931    88.37758   1.218     0.245
Volume        0.70864     0.09746   7.271  6.26e-06

Residual standard error: 123.9 on 13 degrees of freedom
Multiple R-squared:  0.8026,Adjusted  R-squared: 0.7875
F-statistic: 52.87 on 1 and 13  DF,  p-value: 6.262e-06
```

Modeling weights of books using volume and cover type

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	197.96284	59.19274	3.344	0.005841	**
volume	0.71795	0.06153	11.669	6.6e-08	***
cover:pb	-184.04727	40.49420	-4.545	0.000672	***

Residual standard error: 78.2 on 12 degrees of freedom
Multiple R-squared: 0.9275,Adjusted R-squared: 0.9154 F-statistic: 76.73 on 2 and 12 DF, p-value: 1.455e-07

28/4/2022
001

$\hat{\beta}_0$ & $\hat{\beta}_1$ are chosen to minimize

$$\sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 \quad \text{\} sum of squared residuals.}$$

This is called the method of least squares.

SS_{xx} = "sum of squares for X" = $\sum (x_i - \bar{x})^2$ $\qquad s_x^2 = \frac{SS_{xx}}{n-1} \rightarrow$ variance, sample

SS_{yy} = "sum of squares for y" = $\sum (y_i - \bar{y})^2$ $\qquad s_y^2 = \frac{SS_{yy}}{n-1} \rightarrow$ variance of y, sample

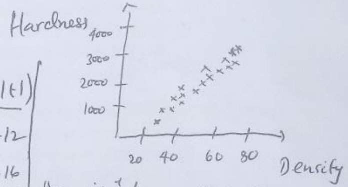
SP_{xy} = "sum of products" = $\sum (x_i - \bar{x})(y_i - \bar{y})$

Sample covariance $\rightarrow cov(x,y) = \frac{SP_{xy}}{n-1}$

$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$

$\hat{\beta}_1 = \frac{SP_{xy}}{SS_{xx}} = \frac{cov(x,y)}{Var(x)}$

Hardness vs Density for 26 Australian Tree Species:



Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
(intercept)	-1160.50	108.58	-10.69	2.07e-12
Density	57.507	2.279	25.24	<2e-16

use "Density" to predict "Hardness"

$$\hat{y} = \underbrace{-1160.50}_{\hat{\beta}_0} + \underbrace{57.507}_{\hat{\beta}_1} x$$

For least squares regression:

→ The residuals sum to 0 $(\sum e_i = 0)$.

→ The line passes through $(\bar{x}, \bar{y})$

$SE_{b_0} = \frac{s}{\sqrt{n}} * \sqrt{1 + \frac{(\bar{x})^2}{Var(x)}}$ $\qquad SE_{b_1} = \frac{s}{\sqrt{n}} * \frac{1}{stdev(x)}$

s (st. error of the regression) $= \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} e_i^2} = stdev(errors) \times \sqrt{\frac{(n-1)}{(n-2)}}$

$R^2 = 1 - \frac{Var(errors)}{Var(Y)}$

28/4/2022
002

A study investigated a possible relationship btw eggshell thickness & environmental contaminants in brown pelican eggs.

pesticide, banned

The figure shows a scatterplot of shell thickness vs. DDT level in a sample of 65 brown pelican eggs on Anacapa Island, California.

We use a computer to fit the least squares regression line.

	Estimate	Std-Error	t value	Pr(>|t|)
Intercept	4.231e-01	2.128e-02	19.880	<2e-16
DDT (ppm)	-8.732e-05	1.603e-05	-5.448	9e-07

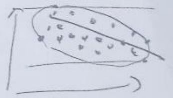
$$\hat{Y} = 0.4231 - 0.00008732x$$

egg with DDT = 2000 → prediction of thickness?

$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X = 0.4231 - (0.00008732 * 2000)$

$= 0.24846$ mm

What proportion of the variability in eggshell thickness can be explained by the linear relationship w/ DDT?



2 reasons → decreasing trend

→ random variability about the line..

proportion due to the "decreasing trend" is r^2 (r = correlation coefficient)

"coefficient of determination"

$r^2 \Rightarrow$ proportion of the variability in response variable Y that is attributable to the linear relationship w/ X.

Multiple R-squared = 0.3202 $\Rightarrow r^2 = 0.3202$.

"32% of the variability in eggshell thickness can be explained by the linear relationship with DDT".

$\Rightarrow$ Thus, the other 68% of variability $\Rightarrow$ random variation...

28/1/2022

003

Before doing any statistical inference,
we should check the residual plots.

$$\left(\text{Plots of } e_i = \underbrace{Y_i}_{\text{obs.}} - \underbrace{\hat{Y}_i}_{\text{pred.}}\right)$$

Simple reg. model ...
→ errors follow normal distribution, ... etc.
→ Constant variance ...

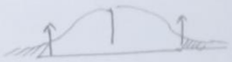
Plot looks fine..
Normal? → plot theoretical quantiles

Test the H_0: no linear relationship btw Y & X,

$H_0: \beta_1 = 0$

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \frac{-8.732 \times 10^{-5} - 0}{1.603 \times 10^{-5}}$$

$= -5.448$



But, we don't conclude "causation"!

p-value = 9e-07 = 0.0000009
"strong evidence against H_0".

Weights of books (g)

		volume (cm^3)
1	800	885
2	950	1016
3	1050	1125
4	350	239
5	750	701
6	600	641
7	1075	1228
8	250	412
9	700	953
10	650	929
11	975	1492
12	350	419
13	950	1010
14	425	595
15	725	1034

Coefficients

	Est^	SE	t-val
(Intercept)	107.679	88.37	1.218
Volume	0.70864	0.09746	7.271

$R^2 = 0.8026$

Jbstatistics (Youtube)

- **Simple Linear Regression: The Least Squares Regression Line**
  - https://www.youtube.com/watch?v=coQAAN4eY5s
- **Simple Linear Regression: An Example**
  - https://www.youtube.com/watch?v=xIDjj6ZyFuw

Sources

- openintro.org/os (Chapter 9, Section 9.1)

Linear Regression using Excel:

https://1drv.ms/x/s!Apc0G8okxWJ1zCUKXCGBs8TgfywO?e=I69d5e