Statistics One

Lecture 7
Introduction to Regression

1

Three segments

- Overview
- Calculation of regression coefficients
- Assumptions

2

Lecture 7 ~ Segment 1

Regression: Overview

3

Regression: Overview

- Important concepts & topics
  - Simple regression vs. multiple regression
  - Regression equation
  - Regression model

4

Regression: Overview

- *Regression*: a statistical analysis used to predict scores on an outcome variable, based on scores on one or multiple predictor variables
  - *Simple regression*: one predictor variable
  - *Multiple regression*: multiple predictors

5

Regression: Overview

- Example: IMPACT (see Lab 2)
  - An online assessment tool to investigate the effects of sports-related concussion
    - http://www.impacttest.com

6

IMPACT example

- IMPACT provides data on 6 variables
  - Verbal memory
  - Visual memory
  - Visual motor speed
  - Reaction time
  - Impulse control
  - Symptom score

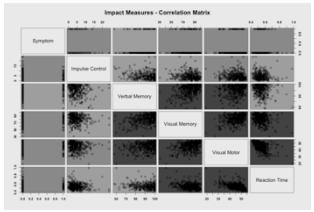
7

IMPACT: Correlations pre-injury

```
> cor(impact)
                Verbal Memory Visual Memory Visual Motor Reaction Time Impulse Control     Symptom
Verbal Memory      1.00000000    0.41549808   0.24573123   -0.15638810     -0.18184017 -0.09333058
Visual Memory      0.41549808    1.00000000   0.34044313   -0.25796852     -0.10059464 -0.06243145
Visual Motor       0.24573123    0.34044313   1.00000000   -0.50452903     -0.07151656 -0.09090637
Reaction Time     -0.15638810   -0.25796852  -0.50452903    1.00000000     -0.10547302  0.02403135
Impulse Control   -0.18184017   -0.10059464  -0.07151656   -0.10547302      1.00000000  0.02908636
Symptom           -0.09333058   -0.06243145  -0.09090637    0.02403135      0.02908636  1.00000000
```

8

IMPACT: Correlations pre-injury



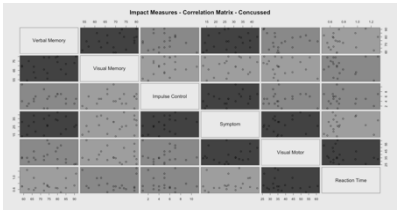
9

IMPACT: Correlations post-injury

	Verbal_Memory	Visual_Memory	Visual_Motor	Reaction_Time	Impulse_Control	Symptom
Verbal_Memory	1.00000000	0.3469979	-0.059072912	0.1205644	-0.050532262	0.2161228
Visual_Memory	0.34699791	1.0000000	-0.155630881	-0.0975997	0.183486515	0.2067446
Visual_Motor	-0.05907291	-0.1556309	1.000000000	-0.2863136	0.004794629	0.2281889
Reaction_Time	0.12056437	-0.0975997	-0.286313634	1.0000000	-0.042379705	0.1477275
Impulse_Control	-0.05053226	0.1834865	0.004794629	-0.0423797	1.000000000	0.4008124
Symptom	0.21612276	0.2067446	0.228188865	0.1477275	0.400812421	1.0000000

10

IMPACT: Correlations post-injury



11

IMPACT example

- For this example, assume:
  - Symptom Score is the outcome variable
  - *Simple regression* example:
    - Predict Symptom Score from just one variable
  - *Multiple regression* example:
    - Predict Symptom Score from two variables

12

Regression equation

- $Y = m + bX + e$
  - Y is a linear function of X
  - m = intercept
  - b = slope
  - e = error (residual)

13

Regression equation

- $Y = B_0 + B_1X_1 + e$
  - Y is a linear function of X_1
  - B_0 = intercept = regression constant
  - B_1 = slope = regression coefficient
  - e = error (residual)

14

Model R and R^2

- R = multiple correlation coefficient
  - $R = r_{\hat{Y}Y}$
  - The correlation between the predicted scores and the observed scores
- R^2
  - The percentage of variance in Y explained by the model

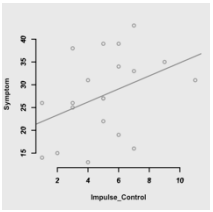
15

IMPACT example

- $Y = B_0 + B_1X_1 + e$
  - Let Y = Symptom Score
  - Let X_1 = Impulse Control
  - Solve for B_0 and B_1
  - In R, function lm

16

IMPACT example



$\hat{Y} = 20.48 + 1.43(X)$

$r = .40$

$R^2 = 16\%$

17

IMPACT example

```
> model1 <- lm(Symptom ~ Impulse_Control)
> summary(model1)

Call:
lm(formula = Symptom ~ Impulse_Control)

Residuals:
    Min      1Q  Median      3Q     Max
-14.5189  -6.2156   0.7189   4.8172  13.2189

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    20.4779     4.3098   4.752 0.000159 ***
Impulse_Control  1.4344     0.7728   1.856 0.079884 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.536 on 18 degrees of freedom
Multiple R-squared:  0.1607, Adjusted R-squared:  0.114
F-statistic: 3.445 on 1 and 18 DF,  p-value: 0.07988
```

18

Regression model

- The regression model is used to model or predict future behavior
  - The model is just the regression equation
  - Later in the course we will discuss more complex models that consist of a set of regression equations

19

Regression: It gets better

- The goal is to produce better models so we can generate more accurate predictions
  - Add more predictor variables, and/or…
  - Develop better predictor variables

20

IMPACT example

- $Y = B_0 + B_1X_1 + B_2X_2 + e$
  - Let Y = Symptom Score
  - Let X_1 = Impulse Control
  - Let X_2 = Verbal Memory
  - Solve for B_0 and B_1 and B_2
  - In R, function lm

21

IMPACT example

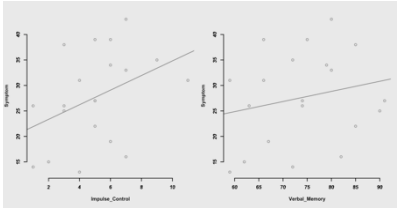
```
> model2 <- lm(Symptom ~ Impulse_Control + Verbal_Memory)
> summary(model2)

Call:
lm(formula = Symptom ~ Impulse_Control + Verbal_Memory)

Residuals:
    Min      1Q  Median      3Q     Max
-16.337  -6.018   0.230   4.848  13.104

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.1313    15.4322   0.268   0.7921
Impulse_Control  1.4773     0.7692   1.921   0.0717
Verbal_Memory    0.2179     0.1976   1.103   0.2856
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.485 on 17 degrees of freedom
Multiple R-squared:  0.2167, Adjusted R-squared:  0.1245
F-statistic: 2.351 on 2 and 17 DF,  p-value: 0.1255
```

$\hat{Y} = 4.13 + 1.48(X_1) + 0.22(X_2)$

$R^2 = 22\%$

22

IMPACT example



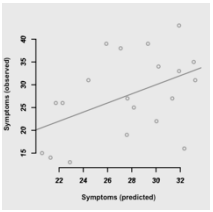
23

Model R and R^2

- R = multiple correlation coefficient
  - $R = r_{\hat{Y}Y}$
  - The correlation between the predicted scores and the observed scores
- R^2
  - The percentage of variance in Y explained by the model

24

IMPACT example



$R^2 = 22\%$

$r_{\hat{Y}Y} = .47$

25

Segment summary

- Important concepts & topics
  - Simple regression vs. multiple regression
  - Regression equation
  - Regression model

26

END SEGMENT

27

Lecture 7 ~ Segment 2

Calculation of regression coefficients

28

Estimation of coefficients

- Regression equation:
  - $Y = B_0 + B_1X_1 + e$
  - $\hat{Y} = B_0 + B_1X_1$
    - $(Y - \hat{Y}) = e$ (residual)

29

Estimation of coefficients

- The values of the coefficients (e.g., B_1) are estimated such that the regression model yields optimal predictions
  - Minimize the residuals!

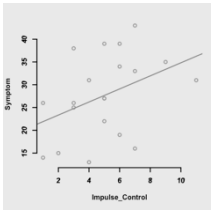
30

Estimation of coefficients

- *Ordinary Least Squares* estimation
  - Minimize the sum of the squared (SS) residuals
  - $SS.RESIDUAL = \Sigma(Y - \hat{Y})^2$

31

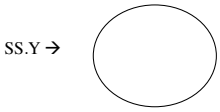
IMPACT example



32

Estimation of coefficients

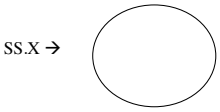
- Sum of Squared deviation scores (SS) in variable Y
  - SS.Y



33

Estimation of coefficients

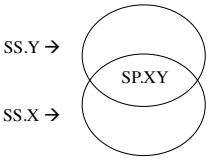
- Sum of Squared deviation scores (SS) in variable X
  - SS.X



34

Estimation of coefficients

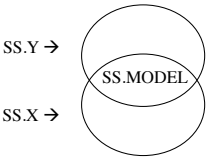
- Sum of Cross Products
  - SP.XY



35

Estimation of coefficients

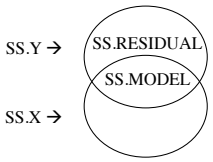
- Sum of Cross Products = SS of the Model
  - SP.XY = SS.MODEL



36

Estimation of coefficients

- SS.RESIDUAL = (SS.Y – SS.MODEL)



37

Estimation of coefficients

- Formula for the unstandardized coefficient
  - $B_1 = r \times (SD_y / SD_x)$

38

Estimation of coefficients

- Formula for the standardized coefficient
  - If X and Y are standardized then
    - $SD_y = SD_x = 1$
    - $B = r \times (SD_y / SD_x)$
    - $\beta = r$

39

Segment summary

- Important concepts
  - Regression equation and model
  - Ordinary least squares estimation
  - Unstandardized regression coefficients
  - Standardized regression coefficients

40

END SEGMENT

41

Lecture 7 ~ Segment 3

Assumptions

42

Assumptions

- Assumptions of linear regression
  - Normal distribution for Y
  - Linear relationship between X and Y
  - Homoscedasticity

43

Assumptions

- Assumptions of linear regression
  - Reliability of X and Y
  - Validity of X and Y
  - Random and representative sampling

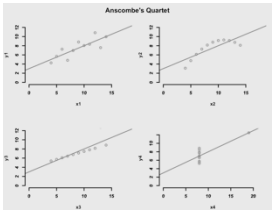
44

Assumptions

- Assumptions of linear regression
  - Normal distribution for Y
  - Linear relationship between X and Y
  - Homoscedasticity

45

Anscombe's quartet



46

Anscombe's quartet

- Regression equation for all 4 examples:
  - $\hat{Y} = 3.00 + 0.50(X_1)$

47

Anscombe's quartet

- To test assumptions, save residuals
  - $Y = B_0 + B_1X_1 + e$
  - $e = (Y - \hat{Y})$

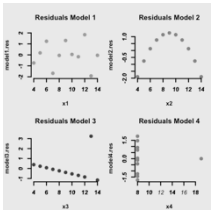
48

Anscombe's quartet

- Then examine a scatterplot with
  - X on the X-axis
  - Residuals on the Y-axis

49

Anscombe's quartet



50

Segment summary

- Assumptions when interpreting *r*
  - Normal distributions for Y
  - Linear relationship between X and Y
  - Homoscedasticity
  - Examine residuals to evaluate assumptions

51

END SEGMENT

52

END LECTURE 7

53