Lecture 16

Regression

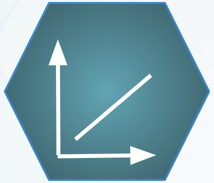
agenda

study guide	discuss midterm 2 study guide
lecture	introduction to simple linear regression (SLR)
R activity	practice SLR in R
review	review for midterm 2 - any questions?

Examples

- Do students with a higher college GPAs have higher paying jobs after graduation?
- Does increased exercise reduce blood pressure?
- Do Tik Tok creators with more followers produce more content during a week?

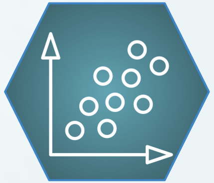
Linear relationships



Deterministic

knowing the value of X tells you the *exact* value of Y

Ex: Fahrenheit vs. Celsius
F = 32 + 1.8C



Statistical

knowing the value of X tell you the *approximate* value of Y (i.e., there is variation in the possible values of Y for each value of X)

Ex: weight vs. height

Example

<u>Question</u>: Do college students with higher **IQ's** have higher **GPA's**, on average?

<u>Setting</u>: We have a random sample of 84 Islanders aged 18 to 25 from across three islands. We have them take an IQ test and obtained their college GPA.

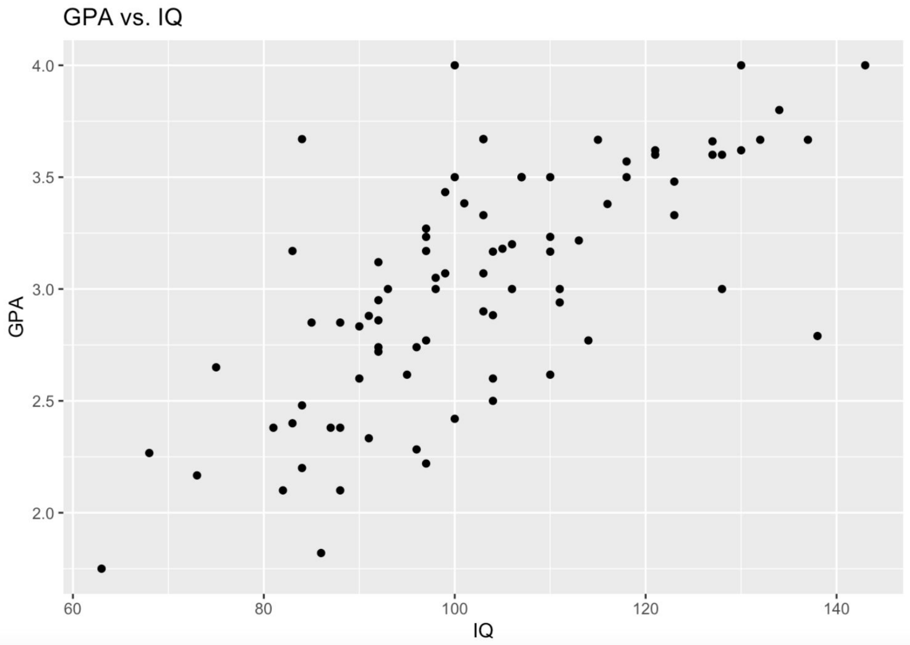
```
iq_gpa <- read_csv("islander-data.csv")

glimpse(iq_gpa)
```

```
## Rows: 84
## Columns: 3
## $ name <chr> "GUNNAR BLOMGREN ", "JULIAN BLOMGREN ", "AIDAN COLLINS ", "ALLAN…
## $ iq   <dbl> 138, 96, 115, 118, 91, 137, 88, 128, 95, 68, 104, 101, 73, 132, …
## $ gpa  <dbl> 2.790, 2.283, 3.667, 3.500, 2.333, 3.667, 2.100, 3.600, 2.617, 2…
```

Example

```r
ggplot(iq_gpa, aes(x = iq, y = gpa)) +
  geom_point() +
  labs(title = "GPA vs. IQ",
       x = "IQ",
       y = "GPA")
```



Relationships between *two quantitative variables*

response (dependent) variable

the variable of interest that we are trying to model

e.g., GPA

explanatory (independent) variable

the variable(s) that may explain the differences in the response variable

e.g., IQ

Goals:

- **estimate** the association between an explanatory variable and the response variable
- **predict** the response for a given value of the explanatory variable

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Response Variable

Intercept Coefficient

Slope Coefficient

Explanatory Variable

Random Error

Parameters

Population model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

islander example

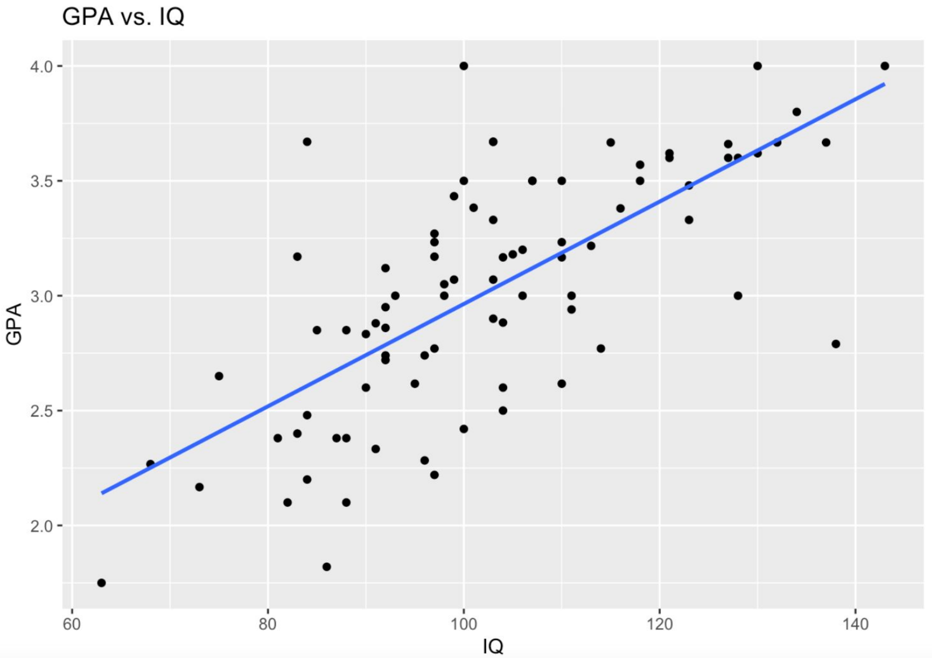
$$GPA_i = \beta_0 + \beta_1 IQ_i + \varepsilon_i$$

interpretations:

- β_0: the average GPA for students with IQ = 0
- β_1: the average difference in GPA for students whose IQ differs by one unit

Islander example - scatterplot

```
ggplot(iq_gpa, aes(x = iq, y = gpa)) +
  geom_point() +
  geom_smooth(method = 'lm', se = FALSE) +
  labs(title = "GPA vs. IQ",
       x = "IQ",
       y = "GPA")
```



Line of "best" fit

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

islander example

$$\widehat{GPA}_i = \hat{\beta}_0 + \hat{\beta}_1 IQ_i$$

interpretations:

- $\hat{y}_i$: the *estimated* average value of *y*
- $\hat{\beta}_0$: the *estimated* average GPA for students with IQ = 0
- $\hat{\beta}_1$: the *estimated* average difference in GPA for students whose IQ differs by one unit

Error

Residual = observed y - predicted y

$$e_i = y_i - \hat{y}_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)$$

e_i is called the **residual** for subject i.

```
iq_gpa <- read_csv("islander-data.csv")

glimpse(iq_gpa)
```

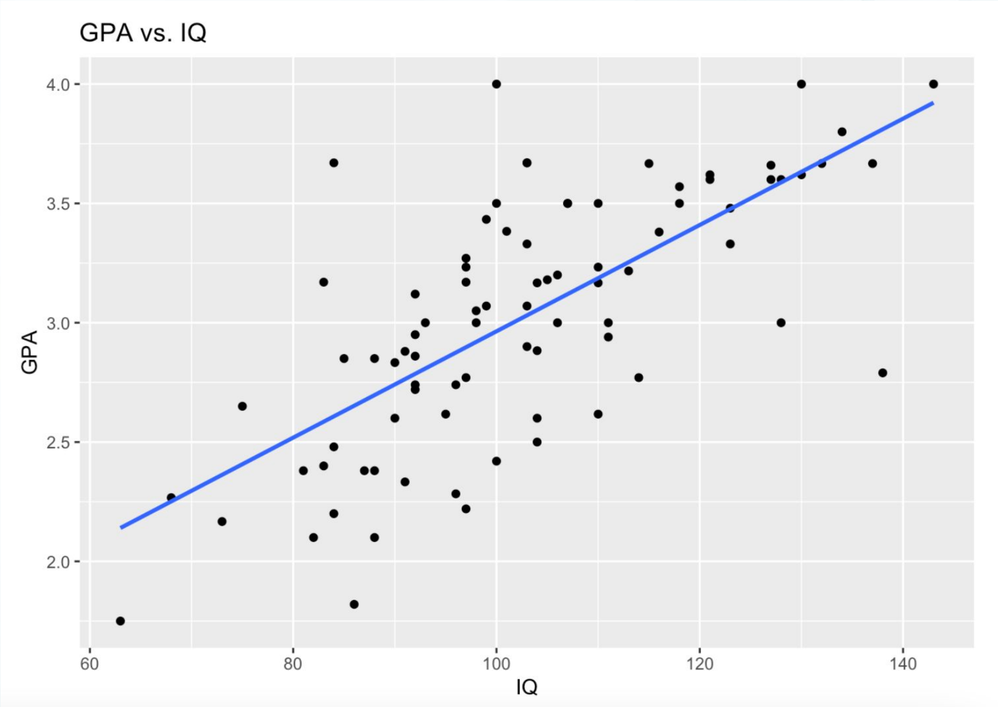
```
## Rows: 84
## Columns: 3
## $ name <chr> "GUNNAR BLOMGREN ", "JULIAN BLOMGREN ", "AIDAN COLLINS ", "ALLAN…
## $ iq   <dbl> 138, 96, 115, 118, 91, 137, 88, 128, 95, 68, 104, 101, 73, 132, …
## $ gpa  <dbl> 2.790, 2.283, 3.667, 3.500, 2.333, 3.667, 2.100, 3.600, 2.617, 2…
```

Least squares

Minimize the **sum of squared errors**:

$$SSE = \sum_{i=1}^{n} \left[y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i \right) \right]^2$$

$$= \sum_{i=1}^{n} e_i^2$$



Modeling in R

`lm(formula = y ~ x, data = data)`

Note: by default, lm() assumes the model has an intercept

```
mod <- lm(gpa ~ iq, data = iq_gpa)
summary(mod)
```

```
##
## Call:
## lm(formula = gpa ~ iq, data = iq_gpa)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.02062 -0.18794  0.03675  0.19511  1.06234
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.736399   0.258061   2.854  0.00547 **
## iq          0.022277   0.002481   8.977 7.98e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3771 on 82 degrees of freedom
## Multiple R-squared:  0.4957, Adjusted R-squared:  0.4895
## F-statistic:  80.6 on 1 and 82 DF,  p-value: 7.979e-14
```

Interpreting the results

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.736399   0.258061   2.854  0.00547 **
iq          0.022277   0.002481   8.977 7.98e-14 ***
```

estimated model

$$\widehat{GPA} = 0.7364 + 0.0223IQ$$

Intercept ($\hat{\beta}_0$): the estimated average GPA for college students with an IQ of 0 is 0.7364.

Slope ($\hat{\beta}_1$): the average difference in GPA for two subpopulations of college students with an IQ 1 point higher than another is 0.0223.

Correlation & R-Squared

Correlation (r)

Def: Strength and direction of a linear relationship

- Ranges from -1 to 1 (where 0 represents no association)

$$r = \frac{cov(X,Y)}{s_X s_Y}$$

R-Squared (coefficient of determination)

Def: The amount of variability in the response explain by the line (x)

- Ranges from 0-1 (literally r^2)
- R^2 closer to 1 indicates model is a good fit

$$R^2 = 1 - \frac{RSS}{SST} = 1 - \frac{\sum_i e_i^2}{\sum_i (y_i - \bar{y})^2}$$

Correlation

Guess the correlation