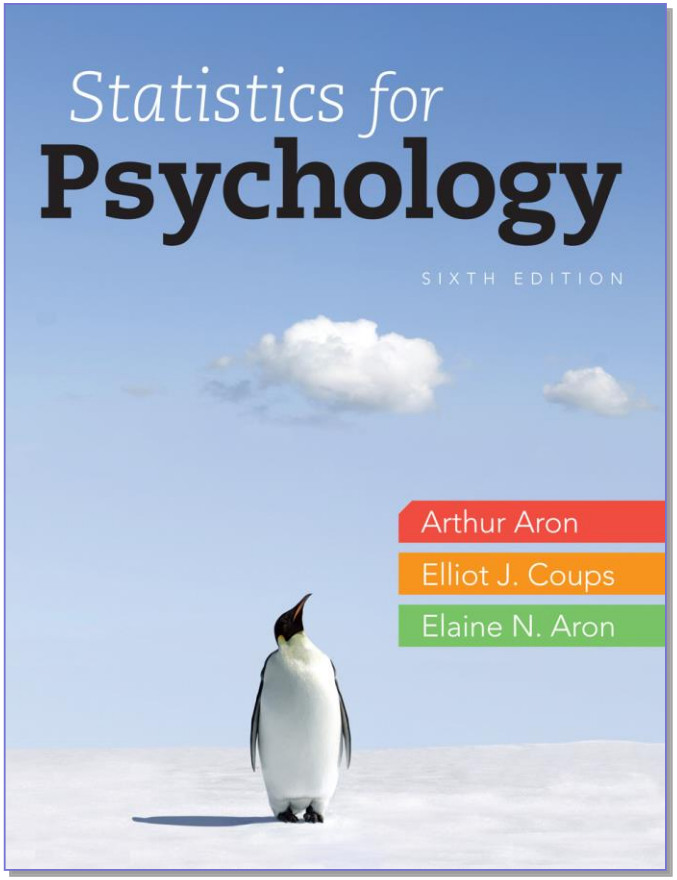
# Statistics for **Psychology**

SIXTH EDITION




## CHAPTER 12

## Prediction

# Prediction

- a major practical application of statistical methods: making predictions
- make informed (and precise) guesses about such things as
  - how well a particular job applicant is likely to perform if hired,
  - how much a reading program is likely to help a particular third grader,
  - how likely a particular patient is to commit suicide, or
  - what factors in people who marry predict whether they will be happy and together 10 years later
  - what are the factors in childhood that predict depression and anxiety in adulthood

# Prediction

- If two variables are correlated it means that you can predict one from the other.
  - So if sleep the night before is correlated with happiness the next day, this means that you should be able, to some extent, to predict how happy a person will be the next day from knowing how much sleep the person got the night before.
- But if two variables are not correlated, then knowing about one does not help you predict the other.
  - So if shoe size and income have a zero correlation, knowing a person’s shoe size does not allow you to predict anything about the person’s income.

# Prediction

- Psychologists construct mathematical rules to predict people's scores on one variable from knowledge of their score on another variable (predicting college grades from SAT scores)
- Prediction is also called “regression”

# Prediction

- In correlation, it did not matter much which variable was which.
- But with prediction we have to decide which variable is *being predicted from (yordayan değişken, yordayıcı) and which variable is being predicted (yordanan değişken).*
- *Think of them like IV and DV, are they alike?*
- The variable **being predicted from** is called the **predictor variable → X.**
- The variable **being predicted** is called the **criterion variable → Y.**

**Table 12-1** Predictor and Criterion Variables

| | Variable Predicted from | Variable Predicted to |
|---|---|---|
| | **Predictor Variable** | **Criterion Variable** |
| Symbol | $X$ | $Y$ |
| Example | SAT scores | College GPA |

**Table 12-1** Predictor and Criterion Variables

# The Linear Prediction Rule

- Using raw scores, the prediction model is
  - Predicted raw score (on the criterion variable) = regression constant plus the product of the raw score regression coefficient times the raw score on the predictor variable
  - Formula

$$\hat{Y} = a + (b)(X)$$

# An example

- The relationship between “the number of hours of sleep (Y)” for six students.
- The regression constant (*a) in this example is -3 and the* regression coefficient (*b) is 1.*
- $\hat{Y} = a + (b)(X) = -3 + (1)(X)$
- Predicted mood = -3 + (1)(hours of sleep).

- Predicting mood after having 9 hours of sleep,
- Predicted mood = -3 + (1)(9) = -3 + 9 = 6.

# The Linear Prediction Rule

$$\hat{Y} = a + (b)(X)$$

- For 0 hours sleep, predicted mood will be **-3 = -3 + 1*(0)**
- 1 hours sleep -2 = -3 + 1*(1)
- 2 hours sleep -1 = -3 + 1*(2)
- 3 hours sleep 0 = -3 + 1*(3)
- 4 hours sleep 1 = -3 + 1*(4)
- 5 hours sleep 2 = -3 + 1*(5)
- 6 hours sleep 3 = -3 + 1*(6)
- 7 hours sleep 4 = -3 + 1*(7)
- 8 hours sleep 5 = -3 + 1*(8)

# The Linear Prediction Rule

- **Regression constant (*a*)**
  - Predicted raw score on criterion variable ($\hat{Y}$) when raw score on predictor variable (X) is 0 $\hat{Y} = a + (b)(X)$
    - **-3 = -3 + 1*(0)**
- **Regression coefficient (*b*)**
  - How much the predicted criterion variable ($\hat{Y}$) increases for every increase of 1 on the predictor variable (X)

# The Linear Prediction Rule

1. Figure the regression constant (*a*)
2. Figure the raw-score regression coefficient (*b*)
3. Find predicted raw score on the criterion variable

- Again, the formula is

$$\hat{Y} = a + (b)(X)$$

# The Regression Line

- Indicates the relation between predictor variable and predicted values of the criterion variable
- Slope of the regression line
  - The steepness of the angle of the regression line, called its slope, is the amount the line moves up for every unit it moves across.
  - Equals ***b***, the raw-score regression coefficient

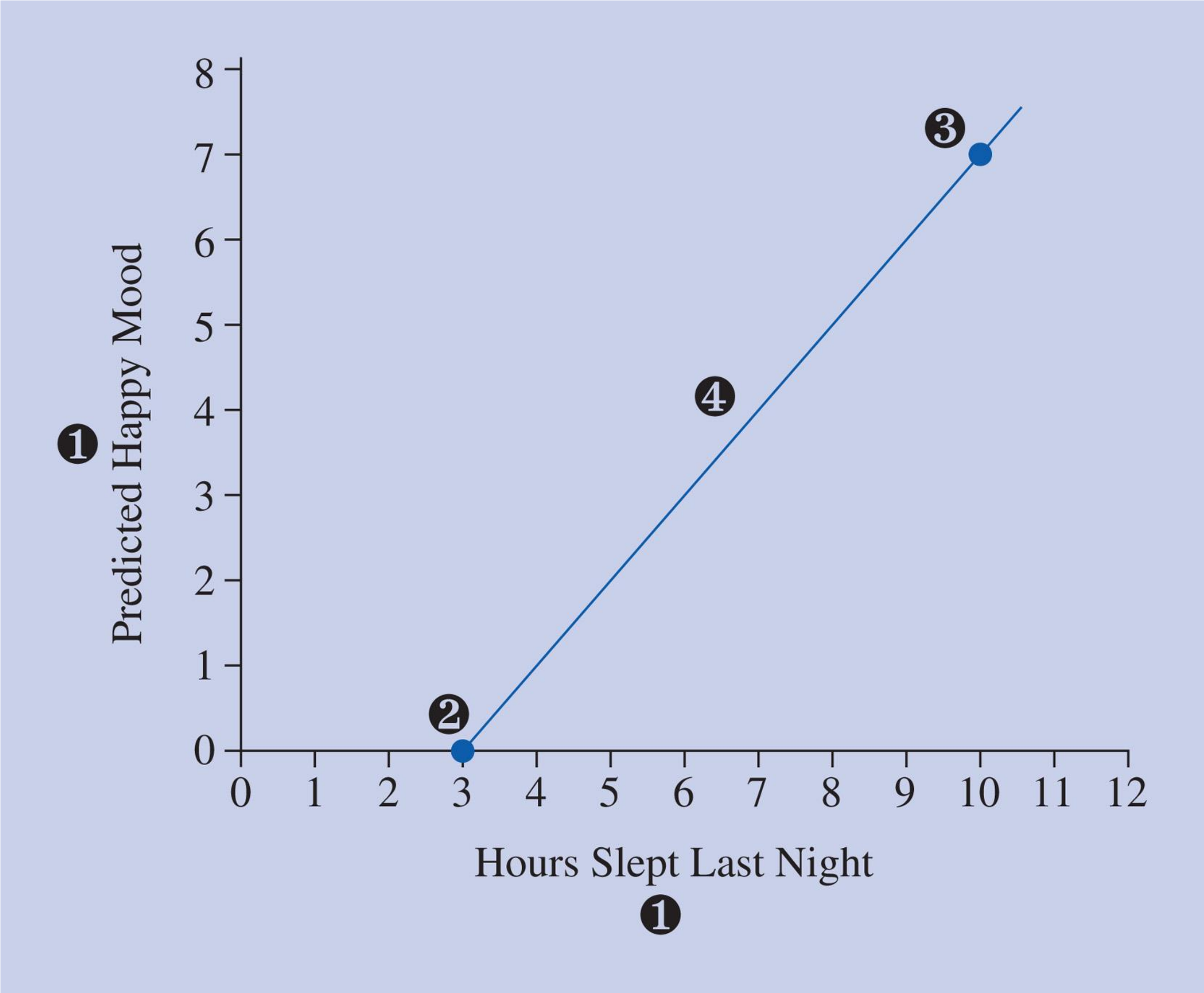
# The Regression Line

- The point at which the regression line crosses (or "intercepts") the vertical axis is called the intercept (or sometimes the *Y intercept).*
- Intercept of the regression line
  - Equals ***a***, the regression constant
- The intercept is the predicted score on the criterion variable ($\hat{Y}$) when the score on the predictor variable ($X$) *is 0.*

# Drawing the Regression Line

1. Draw and label the axes for a scatter diagram (put the predictor variable (X) on the horizontal axis)
2. Figure predicted value on criterion variable for a low value of the predictor variable and mark the point on graph
3. Repeat step 2. with a high value on predictor variable
4. Draw a line passing through the two marks

**Figure 12-4** Steps in drawing a regression line for the hours of sleep and happy mood example. (1) Draw and label the axes, (2) mark a point for a low value of the predictor variable and its predicted value on the criterion variable, (3) mark a point for a high value of the predictor variable and its predicted value on the criterion variable, and (4) draw the regression line.



# Finding the Regression Line

- You can figure out many linear regression lines
  - By trying different X and Y scores in a group of data
- But you need one which gives closest results to actual Y scores
  - *$\hat{Y}$ scores should be close to Y scores*
- Therefore the best regression line should be chosen to predict Y as accurate as possible

# Least Squared Error Principle

- Used to determine the one best prediction rule
- Error $(\text{error} = Y - \hat{Y})$
  - Actual score minus the predicted score
  - Attention! (–) errors and (+) errors can cancel out
  - So, we get **the square of the errors**
  - Then we get **the sum of them**
  - **Sum of the squared errors → SS**

Ta… d Errors for Rules 1, 2, … 12-3), Showing the Figuring for Rule 1

| | | Rule … | | | Rule 4 | | |
|---|---|---|---|---|---|---|---|
| Hours Slept | Actual Mood | Rule 1 Predicted Mood $\hat{Y} = 8 - (.18)(X)$ | Rule 1 Error | Rule 1 Squared Error | Rule 4 Predicted Mood $\hat{Y} = -3 + (1)(X)$ | Rule 4 Error | Rule 4 Squared Error |
| $X$ | $Y$ | $\hat{Y}$ | $(Y - \hat{Y})$ | $(Y - \hat{Y})^2$ | $\hat{Y}$ | $(Y - \hat{Y})$ | $(Y - \hat{Y})^2$ |
| 5 | 2 | 7.10 | −5.10 | 26.01 | 2.00 | .00 | .00 |
| 6 | 2 | 6.92 | −4.92 | 24.21 | 3.00 | −1.00 | 1.00 |
| 6 | 3 | 6.92 | −3.92 | 15.37 | 3.00 | .00 | .00 |
| 7 | 4 | 6.74 | −2.74 | 7.51 | 4.00 | .00 | .00 |
| 8 | 7 | 6.56 | .44 | .19 | 5.00 | 2.00 | 4.00 |
| 10 | 6 | 6.20 | −.20 | .04 | 7.00 | −1.00 | 1.00 |
| | | | | $\sum = 73.33$ | | | $\sum = 6.00$ |

Rule 1 sum of squared errors = 73.33

Rule 2 sum of squared errors = 22.00

Rule 3 sum of squared errors = 7.50

Rule 4 sum of squared errors = 6.00

First regression line
$\hat{Y}$ *scores*

Second regression line
$\hat{Y}$ *scores*

Actual DV scores
Actual Y scores

$(Y - \hat{Y})^2$
*For the first regression line*

The sum of $(Y - \hat{Y})^2$
*For the second regression line*

# Least Squared Error Principle

- **The best prediction rule has the smallest sum of squared errors**
- **Least squares criterion:** A statistical term
- We found the regression line that gave the lowest sum of squared errors between the *actual scores on the criterion variable (Y) and the predicted scores on the criterion variable ($\hat{Y}$)*

# Finding *a* and *b* for the Least Squares Linear Prediction Rule

- These formulas give you the linear prediction rule guaranteed to produce less squared error than any other possible prediction rule.
- To find the regression coefficient (*b*)

$$b = \frac{\Sigma[(X - M_X)(Y - M_Y)]}{SS_X}$$

- *b is the regression coefficient.*
- *$(X - M_X)$ is the deviation score* for each person on the *X (predictor) variable;*
- *$(Y - M_Y)$ is the deviation score* for each person on the *Y (criterion) variable;*
- *$(X - M_X)(Y - M_Y)$ is the product* of deviation scores for each person;
- $\Sigma[(X - M_X)(Y - M_Y)]$ *is the sum* the products of deviation scores over all the people in the study;
- *$SS_X$ is the sum of* squared deviations for the *X variable.*

# Finding *a* and *b* for the Least Squares Linear Prediction Rule

- To find the regression constant (*a*)

$$a = M_Y - (b)(M_X)$$

- *a* is the regression constant;
- ($M_Y$) is the mean of the criterion variable;
- ($b$) is the result of multiplying the regression coefficient;
- ($M_x$) is the mean of the predictor variable.
- Notice that you need to know the value of *b in order to figure the value of a!* ☺

**Table 12-5** Figuring the Regression Coefficient (*b*), Regression Constant (*a*), and the Linear Prediction Rule for Predicting Happy Mood from Hours of Sleep (Fictional Data)

| Number of Hours Slept ($X$) | | | Happy Mood ($Y$) | | |
|---|---|---|---|---|---|
| | Deviation ❶ | Deviation Squared ❹ | | Deviation ❶ | Products of ❷ Deviation Scores |
| $X$ | $X - M_X$ | $(X - M_X)^2$ | $Y$ | $Y - M_Y$ | $(X - M_X)(Y - M_Y)$ |
| 5 | −2 | 4 | 2 | −2 | 4 |
| 7 | 0 | 0 | 4 | 0 | 0 |
| 8 | 1 | 1 | 7 | 3 | 3 |
| 6 | −1 | 1 | 2 | −2 | 2 |
| 6 | −1 | 1 | 3 | −1 | 1 |
| 10 | 3 | 9 | 6 | 2 | 6 |
| $\Sigma = 42$ | | $\Sigma = SS_X = 16$ | $\Sigma = 24$ | | $\Sigma = 16$ |
| $M = 7$ | | ❺ | $M = 4$ | | ❸ |

❻

$$b = \frac{\Sigma[(X - M_X)(Y - M_Y)]}{SS_X} = \frac{16}{16} = 1$$

$$a = M_Y - (b)(M_X) = 4 - (1)(7) = 4 - 7 = -3$$

Linear prediction rule: Using formula $\hat{Y} = a + (b)(X)$, $\hat{Y} = -3 + (1)(X)$

*Note*: The circled numbers refer to the steps for figuring the regression coefficient, *b*.

# Another example…

| X | Y |
|---|---|
| 4 | 6 |
| 6 | 8 |
| 7 | 3 |
| 3 | 7 |

- Find the linear prediction rule for this example.

# Issues in Prediction

# Standardized Regression Coefficient

- 2 different studies a/ sleep and happy mood
- $\hat{Y} = -3 + (1).\ X$ (first study)
- $\hat{Y} = -3 + (2).\ X$ (second study)
- *b* is the regression coefficient and it gives the slope. In other words, it is the predicted amount of increase in units for the criterion variable when the predictor variable increases by one unit
- *b,* IV bir ünite arttığında, DV'de tahmin edilen ünite artışının miktarını gösterir

# Standardized Regression Coefficient

- 2 different studies a/ sleep and happy mood
- $\hat{Y} = -3 + (1). X$ (first study)
- $\hat{Y} = -3 + (2). X$ (second study)
- First study's b is 1; the second's b is 2
- The first study shows a 1-unit increase in sleep is associated with a **1-unit** increase in mood.
- The second study, however, shows a 1-unit increase in sleep is associated with a **2-unit** increase in mood.

# Standardized Regression Coefficient

- One important difference between the two studies of sleep and happy mood: In the original study, happy mood was measured on a scale going from 0 to 8. However, the researchers in the new study used a scale of 0 to 20. (METHODS SECTION)
- Thus, with the first study, a 1-hour increase in sleep predicts a 1-point increase in mood on a scale from 0 to 8,
- But in the second study a 1-hour increase in sleep predicts a 2-point increase in mood but on a scale of 0 to 20
- **The scale used for the predictor and criterion variables will affect the value of *b (the regression coefficient) in the linear prediction rule.***

# Standardized Regression Coefficient

- You need a **standardized score** to compare two different regression line
- Could this remind you smt?
- Z score (Z = (X – M) / SD)
  - w/ z score we can compare two different scores on different scales
- So, we need smt like Z scores to compare two regression lines measured on different scales
- There is a formula for changing a regression coefficient into what is known as a **standardized regression coefficient**.

# Standardized Regression Coefficient

- **A standardized regression coefficient,** which is referred to using the Greek symbol β (beta), shows the predicted amount of change in standard deviation units of the criterion variable if the value of the predicted variable increases by one standard deviation.
- Regression coefficient expressed in standard deviation units

$$\beta = (b)\frac{\sqrt{SS_X}}{\sqrt{SS_Y}}$$

# Example

- Sleep and happy mood example
- $\hat{Y} = -3 + (1)\ X$; so *b* is 1
- $SS_x = 16$
- $SS_y = 22$
- $\beta = b\ (\sqrt{SS_x}\ /\ \sqrt{SS_y}\ ) = 1\ (\sqrt{16}/\sqrt{22})$

= $(4/4.69) = .85$
- This means that for every standard deviation increase in sleep, the predicted level of mood increases by .85 standard deviations.

# Standardized Regression Coefficient ($\beta$) and Regression Coefficient ($r$)

- Up to this point, the examples included only one predictor variable (IV)
- Bivariate prediction
- With one predictor variable, the standardized regression coefficient is equal to the correlation coefficient
  - If 1 IV, $\beta = r$

# Multiple Regression

- To predict the happy mood, the number of hours slept and the number of dreams were needed → 2 predictors
- Multiple regression prediction models
  - Each predictor variable has its own regression coefficient
  - Multiple regression formula with three predictor variables:

$$\hat{Y} = a + (b_1)(X_1) + (b_2)(X_2) + (b_3)(X_3)$$

# Important difference between bivariate prediction and multiple regression

- Bivariate prediction the standardized regression coefficient is the same as the correlation coefficient.
- The standardized regression coefficient ($\beta$) for each predictor variable in multiple regression is *not the same as the ordinary correlation coefficient (r) of that predictor* with the criterion variable.
- Usually, a $\beta$ will be closer to 0 than *r*.

# Multiple Regression

- In multiple regression, the correlation between the criterion variable and all the predictor variables taken together is called the **multiple correlation coefficient and** is symbolized as ***R.***
- *R* = correlation between the criterion variable and all the predictor variables taken together

# Assumptions of Prediction

- The significance test for prediction assumes:
  - An equal distribution of each variable at each point of the other variable
  - A linear relationship between variables
  - That the cases are independent
  - That the error scores are normally distributed

# Limitations of Regression

- Regression is inaccurate if
  - Correlation is curvilinear
  - Restriction in range is present
  - Unreliable measures are used

# Controversies and Limitations

- Controversy about how to judge the relative importance of each predictor variable in predicting the dependent variable
- Consider both the *r*'s and the β's

# Prediction in Research Articles

- Bivariate prediction models rarely reported
- Multiple regression results commonly reported

**Table 12-8** Multiple Regression Analysis Predicting Average Intragroup Effect Size at Postassessment

| Independent Variable | *r* | β |
|---|---|---|
| BDI | .30*** | .30*** |
| Age | −.21*** | −.20** |
| No. of sessions | .12* | .08 |
| Duration of disorder | −.13* | −.02 |

*Note*: $R = .36$; $R^2 = .13$. BDI

*$p < .05$. **$p < .01$. ***$p < .000$.

*Source*: Hahlweg, K., Fiegenbaum, W., Frank, M., Schroeder, B., & von Witzleben,
of an empirically supported treatment of agoraphobia. *Journal of Consulting and Clinic*
by the American Psychological Association. Reprinted with permission.

only 13% of the overall variation in treatment outcome was predicted by these four variables

**Table 12-8** Multiple Regression Analysis Predicting Average Intragroup Effect Size at Postassessment

# Proportionate Reduction in Error

- The most common way to think about the accuracy of a prediction rule is to compare the amount of squared error
- using your prediction rule to the amount of squared error you would have without the prediction rule.

# Proportionate Reduction in Error

- Error
  - Actual score minus the predicted score

    $(\text{error} = Y - \hat{Y})$
- Proportionate reduction in error
  - Squared error using prediction model = $SS_{\text{Error}}$
  - Total squared error when predicting from the mean = $SS_{\text{Total}}$

# Proportionate Reduction in Error

- Formula for proportionate reduction in error:

$$\text{Proportionate reduction in error} = \frac{SS_{\text{Total}} - SS_{\text{Error}}}{SS_{\text{Total}}}$$

- Proportionate reduction in error =
  - $r^2$ for bivariate predicton
  - $R^2$ for multiple regression
- Proportion of variance accounted for

# End of Chapter 12