# Lesson 38: Inference for Regression Predictions: CI and PI

# Role-Type Classification Table

| | | RESPONSE | |
|---|---|---|---|
| | | CATEGORICAL | QUANTITATIVE |
| EXPLANATORY | CATEGORICAL | **C → C** | **C → Q** |
| | QUANTITATIVE | **Q → C** | **Q → Q** |

# Confidence and Prediction Intervals for $\mu_y$

# Inference Based on the Least-Squares Regression Line

In this lesson we will use the least-squares regression line to make predictions and discuss two interpretations of these predictions with their associated errors.

# Inference Based on the Least-Squares Regression Line

Let $x^*$ denote a particular value of the explanatory variable $X$.

The predicted y-value obtained by using $x^*$ in the regression equation has two different interpretations:

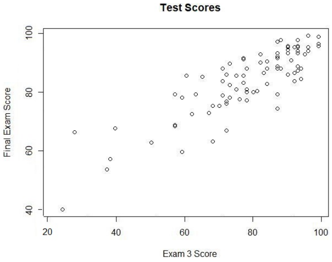
- an estimate of the mean y-value when $X = x^*$
- an estimate of an observed y-value when $X = x^*$

# Example 1: Stat 121 Exam 3 and Final Exam Score

Research question: Can Exam 3 scores of Statistics 121 students be used to predict their Final exam scores?

A random sample of 107 Statistics 121 students was obtained and their exam 3 and Final exam scores were recorded.

A scatterplot and the least-squares regression equation for the sample data are given below.




$\widehat{\text{Final}} = 0.11 + 0.8 \times$ Exam 3

In addition, $r^2$ was calculated to be 0.671.

# Example 1: Final $= 0.11 + 0.8\times$ Exam 3

For an exam 3 score of 70% (.70), the predicted Final exam score is 0.67.

Interpretation 1: 67% estimates the mean Final exam score of all students who obtained a score of 70% in Exam 3. **A confidence interval estimates the mean y-value at $\mathbf{x} = \mathbf{x}^*$.**

Interpretation 2: 67% estimates the Final exam score for one particular student who obtained a score of 70% in Exam 3. **A prediction interval estimates an individual response y at $\mathbf{x} = \mathbf{x}^*$.**

# Computer Output for an Exam 3 score of 70%

| $x^*$ | Predicted Y | Std. Error of predicted Y | 95% C.I. or Confidence Interval | 95% P.I. or Prediction Interval |
|---|---|---|---|---|
| 0.7 | 0.67 | 0.0801 | (0.649, 0.691) | (0.581, 0.759) |

How do we interpret these intervals in context?

Interpretation of C.I.: We are 95% confident that the mean Final exam score of all students who obtained 70% on exam 3 will be between 65 and 69%.

Interpretation of P.I.: We are 95% confident that the Final exam score of a student who obtained 70% on exam 3 will be between 58 and 76%.

# Why are prediction intervals wider than confidence intervals for predictions?

If we use the sample regression line to obtain a predicted y-value, this prediction will most likely not be equal to the true y-value for two reasons:

- The sample regression line will not be equal to the population regression line.
- Even if we knew the population regression line, observed y-values will not fall exactly on the population regression line. There is an additional source of error–the deviation from the population line.

This means that there is more uncertainty (more variability) associated with predicting a single y-value than with estimating the mean y-value at $x = x^*$.

This extra variability is reflected in the width of the corresponding intervals.

# Inference for $\mu_y$ at $x = x^\times$

Confidence Interval:

$$(a + bx^*) \pm t^* SE_{\hat{\mu}}$$

$SE_{\hat{\mu}}$ takes into account uncertainty in estimated regression line

# Inference for $y$ at $x = x^\times$

Prediction Interval:

$$(a + bx^*) \pm t^* SE_{\hat{y}}$$

$SE_{\hat{y}}$ is much larger than $SE_{\hat{\mu}}$

Takes into account

- uncertainty in estimated regression line
- variability around regression line

# Example 2: Schooling and Life Expectancy

1. State:
    1. What is the mean life expectancy for countries that have an average of 16.3 years of schooling?
    2. The United States has an average of 16.3 years of schooling.
       What is the average life expectancy for the United States?

# Example 2: Schooling and Life Expectancy

2. Plan:

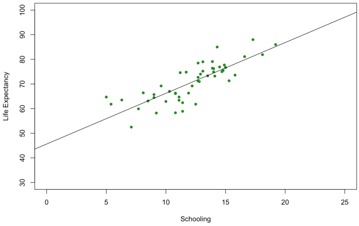
- Data are available for a random sample of 50 countries.
- Use least squares regression to calculate *a*, *b*, and $\hat{\sigma}$ for straight line relationship.

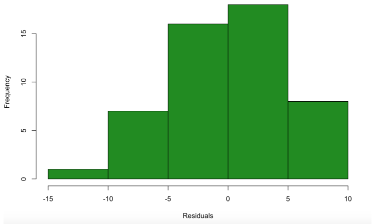
1. Compute 95% confidence interval for $\mu_y$ at $x = 16.3$.
2. Compute 95% prediction interval for *y* at $x = 16.3$.

# Example 2: Schooling and Life Expectancy

3. Solve: Check conditions:

| |
|---|
| Linearity |
| Independence |
| Normality of residuals |
| Equal population standard deviation |



# Example 2: Schooling and Life Expectancy

3. Solve (continued): Fit model using regression software.

| Estimates | Estimate | Std Error | t Statistic | Pr(>t) | CI Lower Bound | CI Upper Bound |
|---|---|---|---|---|---|---|
| Intercept | 45.627 | 2.737 | 16.67 | $<.0001$ | 40.095 | 51.159 |
| Slope | 2.060 | 0.219 | 9.40 | $<.0001$ | 1.617 | 2.503 |

| | |
|---|---|
| $n$ | 50 |
| $r$-Squared | 0.65 |

| $x$ | P.I. | predict.y | C.I. |
|---|---|---|---|
| 16.3 | (69.446, 88.957) | 79.201 | (76.941, 81.462) |

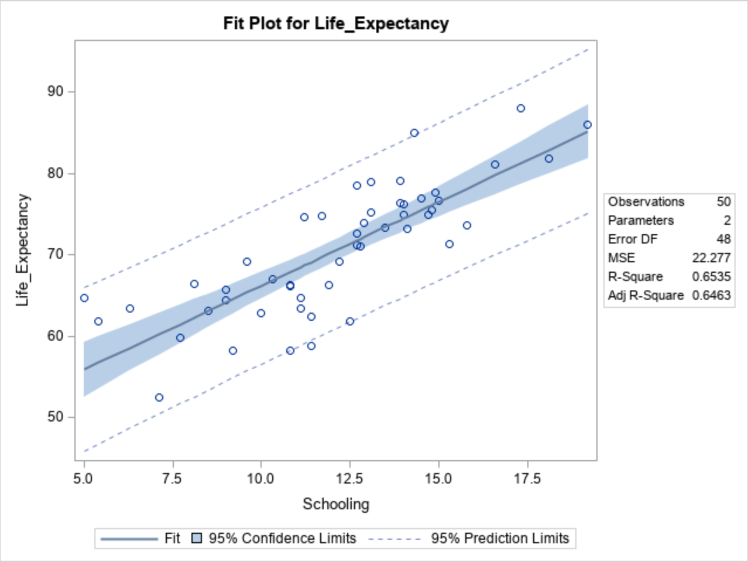
# Example 2: Schooling and Life Expectancy

| $x$ | P.I. | predict.y | C.I. |
|---|---|---|---|
| 16.3 | (69.446, 88.957) | 79.201 | (76.941, 81.462) |

4. Conclude:

    1. With 95% confidence, mean life expectancy for countries with an average of 16.3 years of schooling is between 76.941 and 81.462 years.
    2. With 95% confidence, average life expectancy for the United States of America is between 69.446 and 88.957 years.

# Confidence and Prediction Intervals for all years of schooling

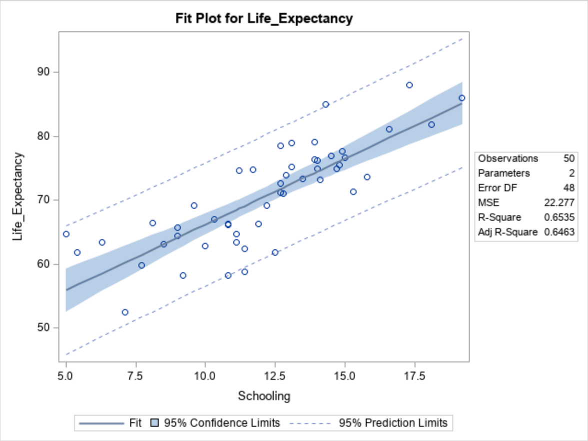


# Self-check

Canada's average years of schooling is 11.6 years. With 95% confidence, Canada's average life expectancy is between ______ and ______.

(a) 68, 71

(b) 59, 79

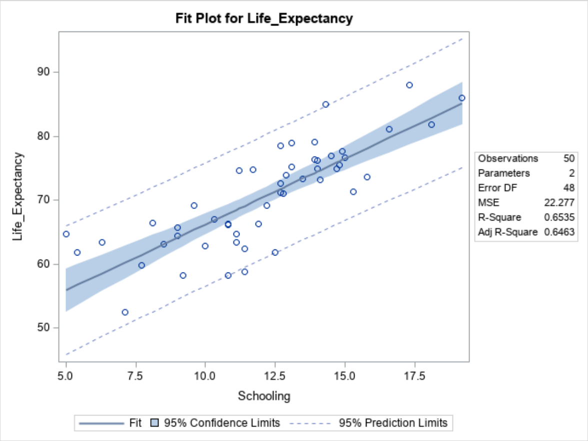


# Self-check

With 95% confidence, the mean life expectancy for all countries with the same average years of schooling as Canada (11.6 years) is between ______ and ______ .

(a) 68, 71

(b) 59, 79

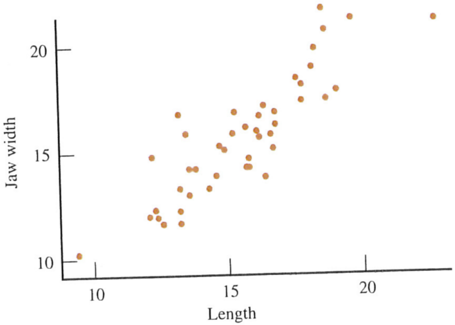


# Example 3: Shark Length and Jaw Width

Physical characteristics of sharks are of interest to surfers and scuba divers as well as to marine researchers. Because it is difficult to measure jaw width in living sharks, researchers would like to determine whether it is possible to estimate jaw width from body length, which is more easily measured. Data on X=length (in feet) and Y=jaw width (in inches) for 44 sharks were obtained from the magazine *Skin Diver and Scuba News*.

# Example 3: Shark Length and Jaw Width

A scatterplot of the data shows a linear pattern which allows for the use of a linear regression model.




The regression equation obtained from software is:
$\widehat{\text{Jaw width}} = 0.69 + 0.963 \times$ Length with $r^2 = 0.766$

# Example 3: Shark Length and Jaw Width

Given: $\text{Jaw } \hat{\text{width}} = 0.69 + 0.963 \times$ Length

What is the predicted jaw width for 15-foot long sharks?
$0.69 + 0.963(15) = 15.135$

The 95% confidence interval for $x = 15$ was calculated to be 14.7882 and 15.498.

How would you interpret this confidence interval in context?

We are 95% confident that the mean jaw width for all 15-foot long sharks is between 14.79 and 15.50 inches.

# Example 3: Shark Length and Jaw Width

The 95% prediction interval for $x = 15$ was calculated to be 12.80 and 17.48 inches.

How would you interpret this prediction interval in context?

With 95% confidence, a shark that is 15 feet long will have a jaw width between 12.80 and 17.48 inches.

# R Shiny App example

1. Open R Shiny app: http://shinyserver.byu.edu/users/stat121res/shinyApp/
2. Regression
3. Suggested data:
   - Dataset: **Windmill**
   - Variables: **CSpd** and **RSpd**
4. Compare prediction and confidence intervals in **Prediction** section

# Self-check

The linear regression equation used to predict the first semester college GPA for a new college freshmen based on their high school GPA is:

$\widehat{\text{College GPA}} = 1.42 + 0.62 \times$ High school GPA

**Fill in the blank**: Joe Taylor has a high school GPA of 2.80; his 95% prediction interval for his first semester college GPA will be _______ the 95% confidence interval for the average first semester college GPA for all new freshmen with a high school GPA of 2.80.

(a) wider than

(b) the same as

(c) narrower than

(d) not comparable with

# Self-check

The linear regression equation used to predict the first semester college GPA for a new college freshmen based on their high school GPA is:

$\widehat{\text{College GPA}} = 1.42 + 0.62 \times$ High school GPA

**True or False**: Joe's prediction interval estimates the mean first semester college GPA for all freshmen with a 2.80 high school GPA.

(a) True

(b) False

# Vocabulary

- confidence interval for $\mu$ at $x = x^*$
- prediction interval for $y$ at $x = x^*$