Question_3_Project_3

Oluwanifemi

2024-10-04

Smallmouth bass (Data file: wblake)

2.15.1 Using the West Bearskin Lake smallmouth bass data in the file wblake, obtain 95% intervals for the mean length at ages 2, 4, and 6 years. 2.15.2 Obtain a 95% interval f or the mean length at age 9. Explain why this interval is likely to be untrustworthy.

```
#PART 1

model <- lm(Length ~ Age, data = file)
summary(model)
```

```
##
## Call:
## lm(formula = Length ~ Age, data = file)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -85.794 -19.499  -4.499  16.177  94.853
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  65.5272     3.1974   20.49   <2e-16 ***
## Age          30.3239     0.6877   44.09   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.65 on 437 degrees of freedom
## Multiple R-squared:  0.8165, Adjusted R-squared:  0.8161
## F-statistic:  1944 on 1 and 437 DF,  p-value: < 2.2e-16
```

```
ages <- data.frame(Age = c(2, 4, 6))

# Obtain 95% confidence intervals for the mean length at ages 2, 4, and 6
predict(model, newdata = ages, interval = "confidence", level = 0.95)
```

```
##        fit      lwr      upr
## 1 126.1749 122.1643 130.1856
## 2 186.8227 184.1217 189.5237
## 3 247.4705 243.8481 251.0929
```

```r
#PART 2
x_0 <- 9
y_hat_0 <- 65.5272 + (30.3239*9)
se <- summary(model)$sigma
n_2 <- 437
n <- nrow(file)
x_bar <- mean(file$Age)
se_2 <- se^2 #check how to raise power
x_sum_sq <- sum((file$Age - x_bar)^2)

se_y_hat_0 <- se * sqrt(1/n + (x_0 - x_bar)^2 / x_sum_sq)

#alpha <- 0.05

t_critical <- 1.966
t_critical
```

```
## [1] 1.966
```

```r
# Calculate the confidence interval
lower_bound <- y_hat_0 - t_critical * se_y_hat_0
upper_bound <- y_hat_0 + t_critical * se_y_hat_0

# Output the results
c(lower_bound, upper_bound)
```

```
## [1] 331.4212 345.4634
```

```r
#FOR CHECKING
ages1 <- data.frame(Age = c(9))

predict(model, newdata = ages1, interval = "confidence", level = 0.95)
```

```
##        fit      lwr      upr
## 1 338.4422 331.4231 345.4612
```

Why This Interval is Untrustworthy The confidence interval for age 9 is likely untrustworthy because it may be an extrapolation beyond the range of the observed ages in the dataset. For example, if the data contains fish aged only up to 6 years, predicting for age 9 is unreliable since the linear model might not accurately capture the trend for such an age.

Question 3

```r
file_1 <-  "C:/Users/modim/Downloads/Regression Methods/Regression-Method-Ass1/Health_Sleep_Statistics.c

data <- read.csv(file_1)

model_1 <- lm(formula = Daily.Steps~Age, data = data)
summary(model_1)
```

```
##
## Call:
## lm(formula = Daily.Steps ~ Age, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3090.7  -688.4   118.6   610.5  3700.0
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16125.50     537.53   30.00   <2e-16 ***
## Age          -258.14      14.54  -17.76   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1223 on 98 degrees of freedom
## Multiple R-squared:  0.7629, Adjusted R-squared:  0.7605
## F-statistic: 315.4 on 1 and 98 DF,  p-value: < 2.2e-16
```

Question 3(b)

1. Hypotheses:

$$H_0 : \beta_1 = 0 \quad \text{(No linear relationship between } \textit{Age} \text{ and } \textit{Daily Steps}\text{)}$$
$$H_A : \beta_1 \neq 0 \quad \text{(There is a linear relationship between } \textit{Age} \text{ and } \textit{Daily Steps}\text{)}$$

This is a two-sided test since we are testing whether β_1 is significantly different from zero.

I will be using t-statistic The t-statistic follows a t-distribution with $n - 2$ degrees of freedom, where n is the number of observations in your sample.

Let's make our LINEar assumptions, which are captured as follows:

Let Y_t be the number of Daily Steps in the t-th observation of the dataset, and let X_t be the corresponding Age for the same observation.

Assumptions:

The following assumptions must be satisfied to validate the test:

- **Linearity**: The relationship between the independent variable (*Age*) and the dependent variable (*Daily Steps*) is linear.
- **Independence**: The observations are independent of each other.
- **Homoscedasticity**: The variance of the residuals is constant across all levels of the independent variable.
- **Normality**: The residuals are approximately normally distributed.

```r
summary_model <- summary(model_1)

# Extract the slope estimate (beta_1) and standard error
beta1 <- summary_model$coefficients["Age", "Estimate"]
```

```r
se_beta1 <- summary_model$coefficients["Age", "Std. Error"]

# Calculate the t-statistic
t_statistic <- beta1 / se_beta1
t_statistic
```

```
## [1] -17.7587
```

```r
# Degrees of freedom
df <- summary_model$df[2]

# Critical value for two-tailed test at alpha = 0.1
alpha <- 0.1
t_critical <- qt(1 - alpha/2, df)
t_critical
```

```
## [1] 1.660551
```

```r
# Compute the p-value
p_value <- 2*pt(abs(t_statistic),df, lower.tail = FALSE)
p_value
```

```
## [1] 2.147977e-32
```

-17.76 in absolute from is greater than 1.660551, meaning our observed t_statistic is greater than the t_critical therefore we reject null hypothesis. Aside from this, p-value is lower than significance level of 0.05, further supporting the notion to reject null hypothesis. Therefore, we **reject** H_0 at the $\alpha = 0.1$ significance level, providing strong evidence that there is a significant linear relationship between *Age* and *Daily Steps*.That is, at least under our linear model with the linearity, independence, normality, and equal variance assumptions, we have sufficient evidence for an association between between *Age* and *Daily Steps*.

```r
beta1_hat <- coef(model_1)[2]
se_beta1_hat <- summary(model_1)$coefficients[2, 2]
n <- nrow(data)
alpha <- 0.10
df <- n - 2
t_value <- qt(1 - alpha / 2, df)
margin_error <- t_value * se_beta1_hat
lower_bound_beta1 <- beta1_hat - margin_error
upper_bound_beta1 <- beta1_hat + margin_error
print(c(lower_bound_beta1, upper_bound_beta1))
```

```
##       Age       Age
## -282.2740 -233.9992
```

```r
# CHECKING: Construct a 90% confidence interval using confint()
confint(model_1, level = 0.90)
```

```
##                    5 %       95 %
## (Intercept) 15232.907 17018.0896
## Age          -282.274  -233.9992
```

Relationship Between the Confidence Interval for β_1 and the Null Hypothesis

1. **Hypothesis Statement**:
   - We set up the null hypothesis $H_0 : \beta_1 = 0$, which states that there is no linear relationship between Age and Daily Steps.
   - The alternative hypothesis is $H_A : \beta_1 \neq 0$, indicating that a relationship exists.
2. **Confidence Interval Interpretation**:
   - The 90% confidence interval for β_1 provides a range of values that likely contain the true value of the slope coefficient.More specifically, we are "90% confident" that the true value of β_1 falls within this range.
   - The interval does **not include** 0, this suggests that it is unlikely that the true effect of Age on Daily Steps is zero. In this case, we have strong evidence against the null hypothesis H_0. Therefore, we reject the null hypothesis at the 0.10 significance level.