565_HW3

Matthew Stoebe

2025-04-09

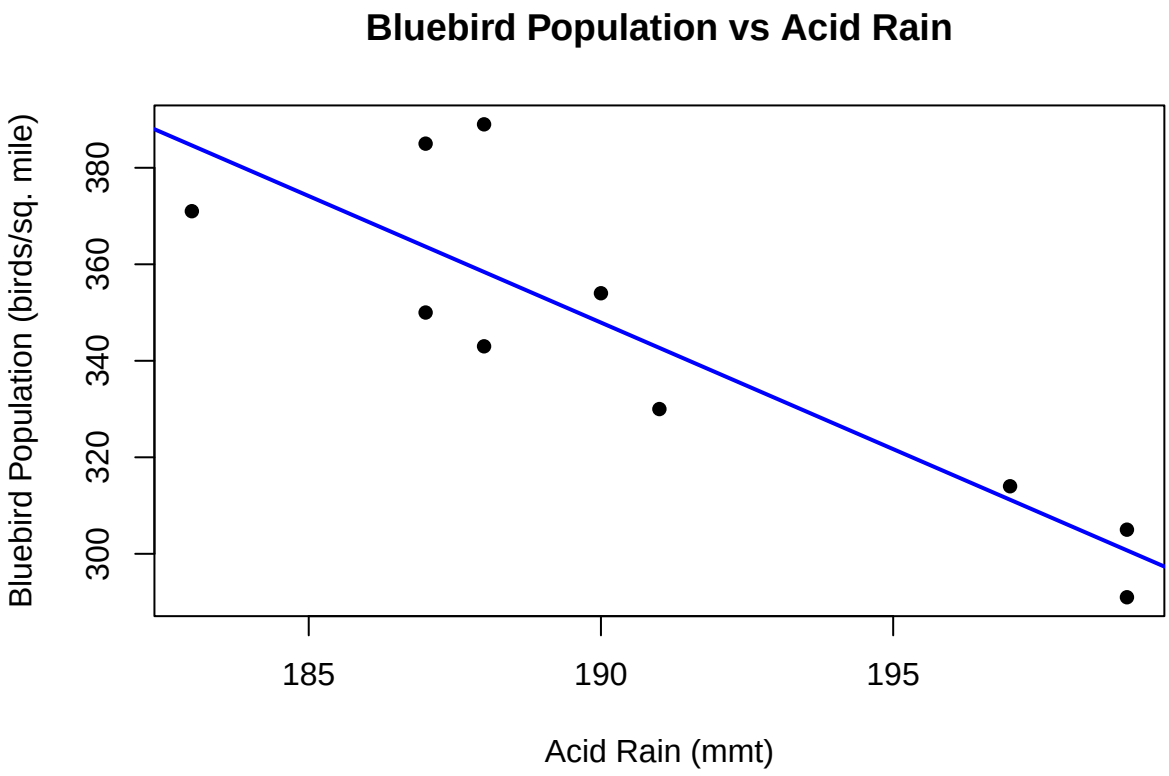
#Question 1

a)

```
years <- 2011:2020
acid_rain <- c(188, 183, 187, 188, 190, 187, 191, 197, 199, 199)
bluebirds <- c(389, 371, 385, 343, 354, 350, 330, 314, 305, 291)

plot(acid_rain, bluebirds,
     xlab = "Acid Rain (mmt)",
     ylab = "Bluebird Population (birds/sq. mile)",
     main = "Bluebird Population vs Acid Rain",
     pch = 16)

fit <- lm(bluebirds ~ acid_rain)
abline(fit, col = "blue", lwd = 2)
```




```
summary(fit)
```

```
##
## Call:
## lm(formula = bluebirds ~ acid_rain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.410 -13.395  -3.461   5.631  30.590
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1344.450    194.710   6.905 0.000124 ***
## acid_rain     -5.245      1.020  -5.144 0.000881 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.03 on 8 degrees of freedom
## Multiple R-squared:  0.7679, Adjusted R-squared:  0.7389
## F-statistic: 26.46 on 1 and 8 DF,  p-value: 0.0008805
```

b) Slope: For every 1mt increase in acid rain, the bluebird population density is expected to decrease by 5.24 Intercept: when no acid rain is present, the expected bluebird population density is 1344. however, this is outside the range of our observed data and may not be as valuable as other points.

c)

```r
coef_summary <- summary(fit)$coefficients

t_value <- coef_summary["acid_rain", "t value"]
p_value_two_sided <- coef_summary["acid_rain", "Pr(>|t|)"]

p_value_one_sided <- p_value_two_sided / 2
cat("t-value =", t_value, "\nOne-sided p-value =", p_value_one_sided, "\n")
```

```
## t-value = -5.144229
## One-sided p-value = 0.0004402716
```

Since the one sided P value is below .05, we can reject our null hypothesis and state that there is a statistically signifficant evidence that an increase in acid rain is associated with a decrease iin bluebird population

d) This is an observational analysis and as such cannot be used for causal conclusions. We can only identify the association but without randomization, case controll, or more advanced causal methods, we cannot draw the conclusion that a "reduction in acid rain will resultin increased bluebird population

#Question 2

a)

```r
set.seed(123)

true_ferritin <- scan("Data/ferritin.txt", sep = " ")

measured_1 <- true_ferritin + rnorm(1000, mean = 0, sd = 20)

num_below_40 <- sum(measured_1 < 40)
cat("Number of patients with measured ferritin < 40:", num_below_40, "\n")
```

```
## Number of patients with measured ferritin < 40: 12
```

b)

```r
recruited_indices <- which(measured_1 < 40)
print(recruited_indices)
```

```
##  [1] 108 135 194 195 212 248 280 572 573 585 722 952
```

```r
measured_2 <- true_ferritin[recruited_indices] + rnorm(length(recruited_indices), mean = 0, sd = 20)
num_higher <- sum(measured_2 > measured_1[recruited_indices])
cat("Number of recruited patients with higher ferritin after 4 months:", num_higher, "\n")
```

```
## Number of recruited patients with higher ferritin after 4 months: 11
```

```r
t_test_result <- t.test(measured_2, measured_1[recruited_indices],
                        paired = TRUE, alternative = "greater")
cat("Paired t-test p-value:", t_test_result$p.value, "\n")
```

```
## Paired t-test p-value: 0.01238386
```

d) While this appears significant, it actually may represent regression to the mean. Patients selected for low ferritin values tend to have a second measurement closer to their true level leading to a positive impact.

#Question 3

a)

```
accident_counts <- 0:5
probs <- c(0.22, 0.33, 0.22, 0.13, 0.08, 0.02)
expected_value <- sum(accident_counts * probs)
cat("Expected number of accidents per intersection:", expected_value, "\n")
```

```
## Expected number of accidents per intersection: 1.58
```

b)

```
first_year <- sample(accident_counts, size = 250, replace = TRUE, prob = probs)
indices_worst <- order(first_year, decreasing = TRUE)[1:25]

second_year <- sample(accident_counts, size = 250, replace = TRUE, prob = probs)
fewer_count <- sum(second_year[indices_worst] < first_year[indices_worst])
prop_fewer <- fewer_count / 25
cat("Proportion of worst intersections with fewer accidents in second year:", prop_fewer, "\n")
```

```
## Proportion of worst intersections with fewer accidents in second year: 0.96
```

Question 4

```
N <- 48
group <- factor(rep(1:6, each = 8))
y <- rnorm(N, mean = 50, sd = 10)
data <- data.frame(group, y)

anova_res <- anova(lm(y ~ group, data = data))
p_anova <- anova_res$`Pr(>F)`[1]
cat("One-way ANOVA p-value:", p_anova, "\n")
```

```
## One-way ANOVA p-value: 0.1497299
```

```
group_means <- tapply(y, group, mean)

min_group <- as.numeric(names(group_means)[which.min(group_means)])
max_group <- as.numeric(names(group_means)[which.max(group_means)])

group_min_data <- data$y[data$group == min_group]
```

```r
group_max_data <- data$y[data$group == max_group]

t_res <- t.test(group_min_data, group_max_data, var.equal = TRUE)
cat("Extreme-groups t-test p-value:", t_res$p.value, "\n")
```

```
## Extreme-groups t-test p-value: 0.01766979
```

```r
set.seed(123)
num_iterations <- 1000
p_values <- numeric(num_iterations)

for (i in 1:num_iterations) {
  y <- rnorm(48, mean = 50, sd = 10)
  group <- factor(rep(1:6, each = 8))
  data <- data.frame(group, y)

  anova_res <- anova(lm(y ~ group, data = data))
  p_anova <- anova_res$`Pr(>F)`[1]

  group_means <- tapply(y, group, mean)
  min_group <- as.numeric(names(group_means)[which.min(group_means)])
  max_group <- as.numeric(names(group_means)[which.max(group_means)])

  group_min_data <- data$y[data$group == min_group]
  group_max_data <- data$y[data$group == max_group]

  t_res <- t.test(group_min_data, group_max_data, var.equal = TRUE)
  p_values[i] <- t_res$p.value
}

num_significant <- sum(p_values < 0.05)
cat("Number of times p-value < 0.05 out of", num_iterations, "simulations:", num_significant, "\n")
```

```
## Number of times p-value < 0.05 out of 1000 simulations: 332
```