Assignment 2 Guide

J Lacasa

2024-09-20

Exercise 1

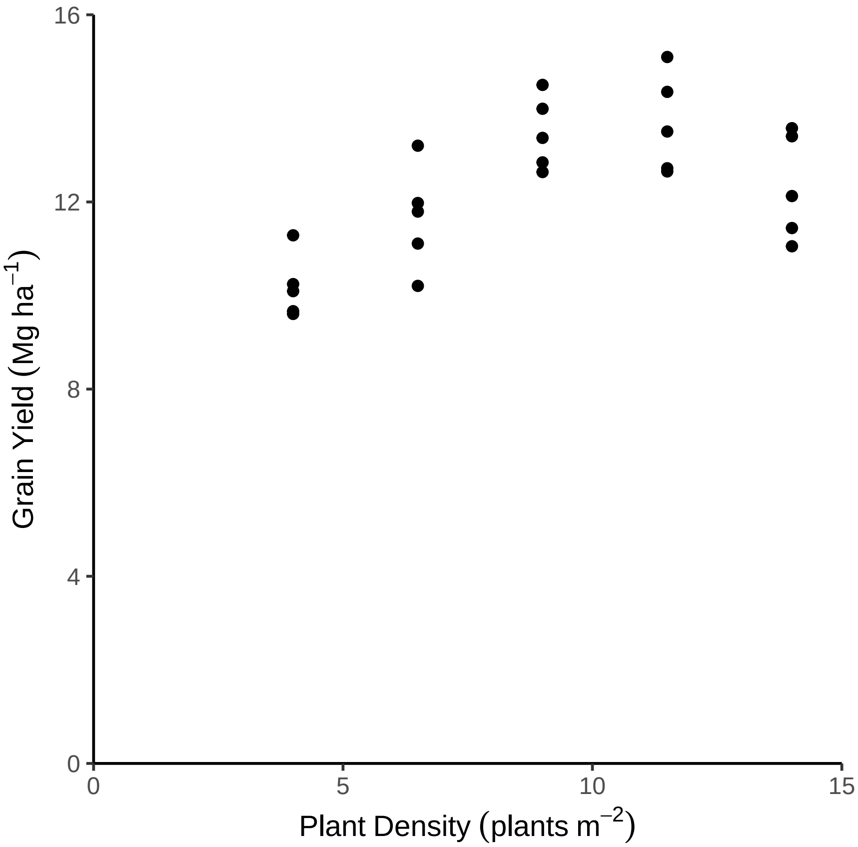
Read the data in the chunk below. Propose a statistical model (using mathematical notation) to describe the relationship between corn yield and plant density, and fit that model to the data.

$$y_i \sim N(\mu_i, \sigma^2),$$

$$\mu_i = \beta_0 + \beta_1 \cdot x_i + \beta_2 \cdot x_i^2,$$

where y_i is the yield of the ith observation, μ_i is its mean, σ^2 is the variance, x_i is the plant density for the ith observation, β_0 is the intercept (i.e., the yield at zero plant density), and β_1 and $beta_2$ give the shape to the response curve.

```
library(tidyverse)
url <- "https://raw.githubusercontent.com/jlacasa/stat705_fall2024/main/classes/data/corn_example2.csv"
data <- read.csv(url)
data %>%
  ggplot(aes(plant_density, yield_Mgha))+
  geom_point()+
  labs(x = expression(Plant~Density~(plants~m^{-2})),
       y = expression(Grain~Yield~(Mg~ha^{-1})))+
  theme_classic()+
  coord_cartesian(xlim = c(0, 15),
                  ylim = c(0, 16),
                  expand = F)+
  theme(aspect.ratio = 1)
```



Answer the following questions:

a. What is the plant density that maximizes grain yield? Provide a point estimate and some measure of uncertainty.

The plant density that maximizes grain yield (opd) is the value of plant density when the first derivative of the model is zero. That means $\widehat{opd} = -\frac{\hat{\beta}_1}{2\hat{\beta}_2}$.

```
# fit the model
m <- lm(yield_Mgha ~ plant_density + I(plant_density^2),
        data = data)
coef(m)
```

```
##        (Intercept)      plant_density I(plant_density^2)
##         4.30878773         1.74562330        -0.08300439
```

```
opd_hat <- -coef(m)[2]/(2*coef(m)[3])
opd_hat
```

```
## plant_density
##      10.51525
```

Then, the $\widehat{opd}$ is 10.52 plants per m2. Some uncertainty estimates that can be used to characterize this estimate are the standard error, or a 95% confidence interval. We can estimate that using the delta method.

```
covariance <- vcov(m)
opd_se_hat <- msm::deltamethod(g = ~ -x2/(2*x3), mean = coef(m), cov = covariance)
## 1.96 is typically used for approximate confidence intervals
lower <- opd_hat - 1.96*opd_se_hat
```

```r
upper <- opd_hat + 1.96*opd_se_hat
```

Then, $s.e.(\widehat{opd})$ is 0.47 plants per m2, and the 95% confidence interval is (9.58, 11.45).

b. How much yield do you expect the crop to yield, on average, with 8 plants per m2? What is a good 95% confidence interval for that value?

```r
X_new <- data.frame(plant_density = 8)
predict_a <- predict(m, newdata = X_new, interval = "confidence")
predict_a
```

```
##        fit     lwr      upr
## 1 12.96149 12.3505 13.57249
```

I expect the crop to yield on average with 8 plants per m2, with a 95% confidence interval of (12.35, 13.57).

c. What is a reasonable 95% confidence interval for observable yields at 8 plants per m2?

```r
predict_b <- predict(m, newdata = X_new, interval = "prediction")
```

The 95% prediction interval is (10.85, 15.07).

d. Which confidence interval would be affected most is the sample size was increased twofold?
As sample size increases, the confidence interval that is most affected is the estimation interval because as sample size increases, the accuracy for the $\hat{\beta}$ increases. The variance in the data σ^2, however, represents the variability in the data that often times cannot be reduced any more. The prediction interval thus represents also the uncertainty about observing new data.