Computational Statistics & Probability

Lab 2 - Linear Models

Fall 2025

Learning Objectives

By the end of this lab, you will be able to:

- Specify and justify priors for linear regression parameters
- Use prior predictive simulation to check if priors are reasonable
- Fit linear models using `quap()`
- Interpret regression coefficients in terms of associations
- Make predictions with uncertainty for new observations
- Use posterior predictive checks to assess model fit

Introduction: Predicting Height from Weight

In this lab, we'll build a linear model to predict adult height from weight using the !Kung San census data.

The Model:

$$\begin{aligned} \text{height}_i &\sim \text{Normal}(\mu_i, \sigma) \\ \mu_i &= \alpha + \beta \cdot (\text{weight}_i - \bar{\text{weight}}) \\ \alpha &\sim \text{Normal}(178, 20) \\ \beta &\sim \text{Log-Normal}(0, 1) \\ \sigma &\sim \text{Uniform}(0, 50) \end{aligned}$$

Key concepts:

- α (alpha): Average height when weight = mean weight
- β (beta): Change in height per 1 kg increase in weight
- σ (sigma): Standard deviation around the line
- Centering weight makes α interpretable and improves computation

```r
# be sure `rethinking` is loaded
data(Howell1)
d <- Howell1
d2 <- d[ d$age >= 18 , ] # adults only  # Adults only

# Summary
cat("Number of adults:", nrow(d2), "\n")
```

```
## Number of adults: 352
```

```r
cat("Weight range:", min(d2$weight), "to", max(d2$weight), "kg\n")
```

```
## Weight range: 31.07105 to 62.99259 kg
```

```r
cat("Height range:", min(d2$height), "to", max(d2$height), "cm\n")
```

```
## Height range: 136.525 to 179.07 cm
```

1. Prior Predictive Simulation

Before fitting, check if priors make sense.

a) Understanding the Priors

We choose: - $\alpha \sim$ Normal(178, 20): Average height $\approx$ 178 cm, $\pm$20 cm uncertainty - $\beta \sim$ Log-Normal(0, 1): Positive relationship (can't shrink by gaining weight!) - $\sigma \sim$ Uniform(0, 50): Residual variation up to 50 cm

Why use Log-Normal for β instead of Normal? (1-2 sentences)

```r
# YOUR ANSWER:
```

b) Simulate from Priors

Simulate N = 100 prior regression lines:

```r
set.seed(212)
N <- 100

# Sample from priors
# alpha <- # TODO #YOUR CODE HERE
# beta <- # TODO #YOUR CODE HERE

# Plot
# plot(NULL,
#      xlim = c(30, 70), ylim = c(50, 250),
#      xlab = "Weight (kg)", ylab = "Height (cm)",
#      main = "Prior Predictive Simulation")
#
# abline(h = 0, lty = 2, col = "gray")
# abline(h = 272, lty = 2, col = "gray")
#
# xbar <- mean(d2$weight)
# for (i in 1:N) {
#   curve(alpha[i] + beta[i] * (x - xbar),
#         from = 30, to = 70, add = TRUE,
#         col = col.alpha("black", 0.2))
# }
```

c) Check Prior Predictions For 50 kg adult, what heights does prior predict?

```r
# predicted_heights <- # TODO (use alpha, beta from part b)
# mean(predicted_heights)
# PI(predicted_heights, prob = 0.89)
```

2. Fit the Model

```r
m4.3 <- quap(
  alist(
    height ~ dnorm(mu, sigma),
    mu <- a + b * (weight - xbar),
```

```r
    a ~ dnorm(178, 20),
    b ~ dlnorm(0, 1),
    sigma ~ dunif(0, 50)
  ),
  data = d2
)

precis(m4.3)
```

```
##              mean         sd      5.5%      94.5%
## a     154.6013710 0.27030803 154.1693666 155.0333754
## b       0.9032803 0.04192369   0.8362781   0.9702824
## sigma   5.0718878 0.19115544   4.7663845   5.3773912
```

a) Interpret Coefficients

```r
# 1. What is posterior mean for intercept (a)? What does it mean?
# YOUR ANSWER:

# 2. What is posterior mean for slope (b)? What does it mean?
# YOUR ANSWER:

# 3. Has data updated beliefs about beta?
#    Prior Log-Normal(0,1) has mean approx. 1.65
# YOUR ANSWER:
```

b) Prior vs. Posterior Comparison

3. Make Predictions

Five new adults - predict their heights:

Individual	Weight (kg)	Predicted Height	89% Interval
1	47		
2	60		
3	37		
4	51		
5	43		

```r
# new_data <- data.frame(
#   individual = 1:5,
#   weight = c(47, 60, 37, 51, 43)
# )

# height_sim <- # TODO: use sim()
# Exp_height <- # TODO: mean for each individual
# height_CI <- # TODO: 89% CI for each
```

Interpret Predictions

```r
# For Individual 2 (60 kg):
# 1. Predicted height?
```

```r
# 2. What does 89% interval tell you?
# 3. Why uncertainty even though we know weight?
```

4. Visualize Regression Line

```r
# weight_seq <- # TODO
# mu <- # TODO: use link()
# mu_mean <- # TODO
# mu_PI <- # TODO

# plot(height ~ weight, data = d2, col = col.alpha("black", 0.5))
# lines(weight_seq, mu_mean, col = "steelblue", lwd = 3)
# shade(mu_PI, weight_seq, col = col.alpha("steelblue", 0.3))
```

5. Posterior Predictive Checks

a) Simulate for Existing Data

```r
# height_post_pred <- # TODO: sim() for all d2
# pred_mean <- # TODO
# pred_PI <- # TODO
```

b) Visual Checks

```r
par(mfrow = c(1, 2))

# Observed vs. Predicted
plot(d2$height, pred_mean,
     xlab = "Observed (cm)", ylab = "Predicted (cm)",
     main = "Observed vs. Predicted",
     col = col.alpha("black", 0.5), pch = 16)
abline(0, 1, col = "red", lwd = 2, lty = 2)

# Residuals
plot(d2$weight, residuals,
     xlab = "Weight (kg)", ylab = "Residual (cm)",
     main = "Residuals vs. Weight",
     col = col.alpha("black", 0.5), pch = 16)
abline(h = 0, col = "red", lwd = 2, lty = 2)

par(mfrow = c(1, 1))
```

c) Check Coverage

```r
in_interval <- (d2$height >= pred_PI[1, ]) &
                (d2$height <= pred_PI[2, ])
coverage <- mean(in_interval)
cat("Coverage:", round(coverage, 3), "\n")
cat("Expected: 0.89\n")
if (abs(coverage - 0.89) < 0.05) {
  cat("Well-calibrated!\n")
} else {
  cat(" Coverage off\n")
```

```
}
```

6. Reflection Questions

a) Prior vs. Posterior Predictive

What's the difference? Why do both?

```
# YOUR ANSWER:
```

```
# Prior predictive: Check priors BEFORE data
# Shows what we'd expect if priors were true
# Posterior predictive: Validate model AFTER fitting
# Shows if model reproduces observed patterns
# Both needed: assumptions reasonable (prior) + model adequate (posterior)
```

b) Uncertainty in Predictions

Why are individual predictions (Q3) wider than regression line (Q4)?

```
# YOUR ANSWER:
```

```
# Regression line (link): Only parameter uncertainty
# Individual predictions (sim): Parameter uncertainty + individual variation
# Even if we knew true line, people vary around it (genetics, etc.)
```

c) Association vs. Causation

Does gaining weight CAUSE increased height?

```
# YOUR ANSWER:
```

Summary

Complete Bayesian workflow for linear regression:

1. Specified priors
2. Prior predictive simulation
3. Fit model with quap()
4. Interpreted coefficients
5. Made predictions with uncertainty
6. Posterior predictive checks

Key: Always check before (prior) and after (posterior)!