Simple Linear Regression: Bayesian Inference

Lecture 11B handout

Name: _____

Capital Bikeshare

The `bikes` data in the {bayesrules} package has 500 days of Capital Bikeshare data from Washington, D.C. (2011–2012): `rides` (number of rides that day), `temp_feel` (what the temperature felt like, in °F), and `weekend` (`TRUE` for Saturday or Sunday).

We fit Bayesian regression models with `stan_glm()` from the {rstanarm} package.

Example from Johnson, Ott, & Dogucu (2022), *Bayes Rules!*, Chapter 9.

Same model, different view. On Monday, β_0 and β_1 were _____. Today they are _____ with a prior and a posterior:

$$Y_i \mid \beta_0, \beta_1, \sigma \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$$

$\beta_0 + \beta_1 X_i$ is the _____ on days with temperature X_i, and σ is how far a single day typically is _____.

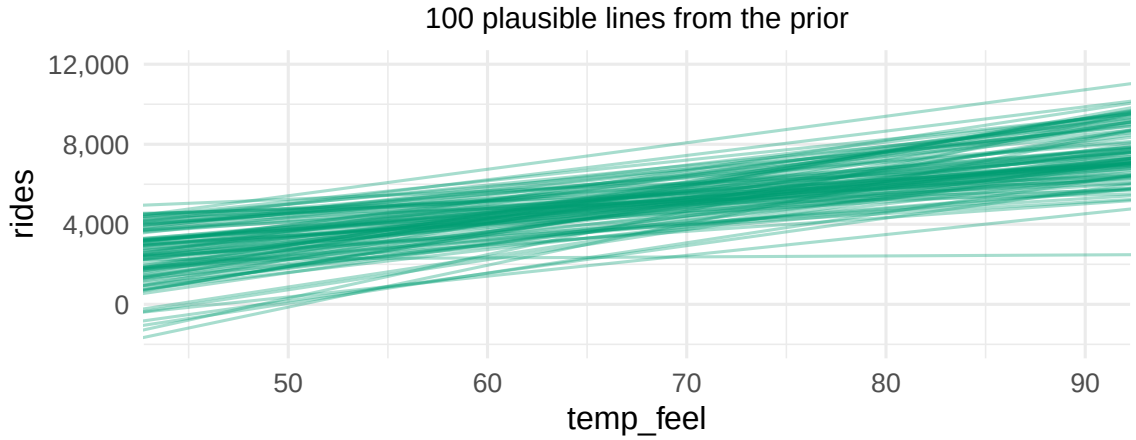
Step 1: prior

Table 1: Prior understanding of bikeshare ridership

Parameter	Prior understanding	Prior
Centered intercept	On an **average temperature** day (69.1 °F), typically about 5,000 rides (between 3,000 and 7,000).	$N(5000, 1000^2)$
β_1	Each extra degree, about 100 more rides (between 20 and 180).	$N(100, 40^2)$

We set priors for the centered intercept and the slope. For σ, we let `stan_glm()` choose the prior.[1]

Centered intercept. Remember centering from Monday! In `stan_glm()`, `prior_intercept` is a prior for the **centered** intercept: the typical number of rides on an average temperature day, not on a 0 °F day. Why is it easier to set a prior for the centered intercept?



[1] It is usually not a good idea to leave a prior unspecified: we should say what we believe about every parameter. We do it here to keep our first Bayesian regression model simple. σ can only be positive, so a normal prior does not fit it well. In more advanced classes, you will see other probability distributions that are useful for setting a prior for σ.

Steps 2 and 3: data and posterior

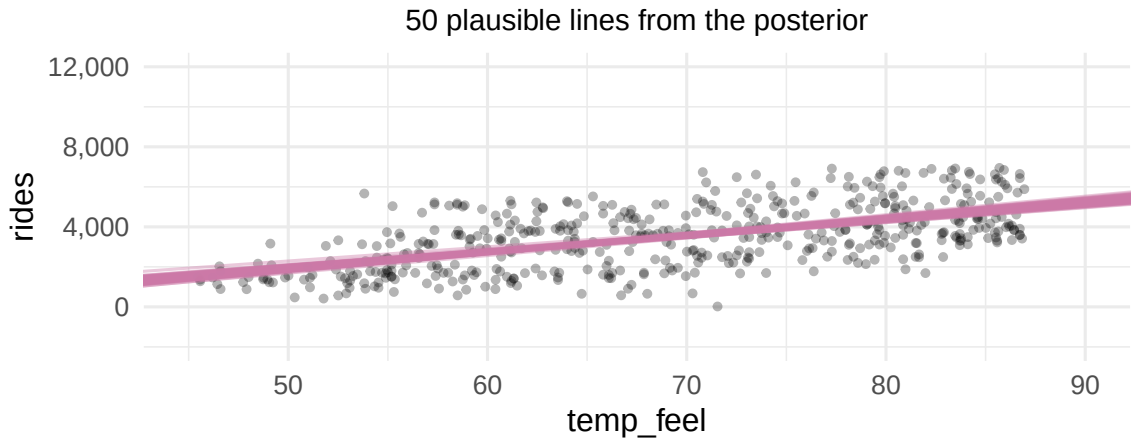
```
library(rstanarm)
library(bayesrules)

bike_model <- stan_glm(rides ~ temp_feel, data = bikes,
                       prior_intercept = normal(5000, 1000),
                       prior = normal(100, 40),
                       seed = 84735, refresh = 0)

bike_draws <- as.data.frame(bike_model)
head(bike_draws, 3)
```

```
  (Intercept) temp_feel    sigma
1   -1828.195  76.14610 1277.580
2   -2029.840  80.28227 1276.993
3   -1676.518  73.69067 1220.550
```

The posterior has no simple formula, so `stan_glm()` **simulates** 4,000 plausible values of $(\beta_0, \beta_1, \sigma)$ from it. Each _____ of `bike_draws` is one plausible line.



Compared to the prior lines, the posterior lines are _____.

Using the posterior

Two intercepts. Our prior was for the centered intercept, but `bike_draws` has only one intercept column:

- `(Intercept)`: the usual _____ intercept, β_0, the typical number of rides on a _____ day. That is why it is negative.
- The _____ intercept, the typical number of rides on an average temperature day, is not a column. Each draw's line gives it: $\beta_0 + \beta_1 \times$ _____.

```
bike_draws |>
  mutate(centered = `(Intercept)` + temp_feel * mean(bikes$temp_feel)) |>
  summarize(posterior_mean_centered = mean(centered),
            posterior_mean_uncentered = mean(`(Intercept)`))
```

```
  posterior_mean_centered posterior_mean_uncentered
1                  3486.5                 -2198.588
```

Our prior said about 5,000 rides on an average temperature day. What does the posterior say?

_____

```
bike_draws |>
  summarize(posterior_mean_b1 = mean(temp_feel),
            posterior_lower_b1 = quantile(temp_feel, 0.025),
            posterior_upper_b1 = quantile(temp_feel, 0.975))
```

```
  posterior_mean_b1 posterior_lower_b1 posterior_upper_b1
1          82.22134           72.82252           92.06594
```

a. Interpret the posterior mean of β_1 in context.

b. The 95% **posterior credible interval** for β_1 is the middle 95% of the posterior draws. Interpret it in context.

c. $H_0 : \beta_1 \leq 0$ and $H_A : \beta_1 > 0$. The **posterior probability** of a hypothesis is the _____ where it is true.

```
bike_draws |>
  summarize(posterior_prob_positive = mean(temp_feel > 0))
```

```
  posterior_prob_positive
1                       1
```

$P(H_A \mid \text{data}) \approx$ _____. Conclusion:

d. **Prediction.** Using the posterior means, the typical number of rides on a 75 °F day is _____. A single 75 °F day could be quite different from this: individual days vary around the line by about $\sigma \approx$ _____ rides (the posterior mean of σ).

A binary explanatory variable: weekends

$X_i = 1$ on weekends and 0 on weekdays. Typical rides on a weekday: _____. On a weekend day: _____. So β_1 is the _____.

We keep the same prior for the centered intercept (a typical day). We are not sure whether weekends have more or fewer rides, so the prior for β_1 is $N(0, 1000^2)$. Prior probability that $\beta_1 < 0$: _____

```
week_model <- stan_glm(rides ~ weekend, data = bikes,
                       prior_intercept = normal(5000, 1000),
                       prior = normal(0, 1000),
                       seed = 84735, refresh = 0)
week_draws <- as.data.frame(week_model)
```

```
week_draws |>
  summarize(posterior_mean_b0 = mean(`(Intercept)`),
            posterior_mean_b1 = mean(weekendTRUE))
```

```
  posterior_mean_b0 posterior_mean_b1
1          3713.116         -794.5523
```

a. `(Intercept)` is the uncentered intercept, so it is the typical number of rides when $X = 0$. Posterior mean of the typical rides on a weekday: _____. On a weekend day: _____

```
week_draws |>
  summarize(posterior_lower_b1 = quantile(weekendTRUE, 0.025),
            posterior_upper_b1 = quantile(weekendTRUE, 0.975),
            posterior_prob_negative = mean(weekendTRUE < 0))
```

```
  posterior_lower_b1 posterior_upper_b1 posterior_prob_negative
1          -1081.287          -504.9607                       1
```

b. Interpret the 95% posterior credible interval for β_1 in context.

c. $H_A : \beta_1 < 0$ (fewer rides on weekends). Prior probability: _____. Posterior probability: _____. What do you conclude? Can we say weekends **cause** fewer rides?

Frequentist vs. Bayesian

`lm()` and `confint()` are from Monday (the {stats} package, which loads automatically).

```
confint(lm(rides ~ temp_feel, data = bikes))
```

```
                  2.5 %      97.5 %
(Intercept) -2883.88257 -1474.65468
temp_feel      71.82028    91.93961
```

Table 2: Frequentist and Bayesian intervals for the slope

	Frequentist (Monday)	Bayesian (today)
R function (package)	_____ ({stats})	_____ ({rstanarm})
95% interval for β_1	_____	_____
"95% probability that β_1 is in the interval"	_____	_____

Why are the two intervals so similar? _____

Summary

- In R: _____ from the {rstanarm} package, with a prior for the intercept (`prior_intercept`) and the slope (`prior`); here `stan_glm()` chose the prior for `sigma`. `prior_intercept` is for the _____ intercept. In the draws, `(Intercept)` is the _____ intercept. The posterior is a set of simulated _____: `as.data.frame()`.
- From the draws: the **posterior mean** (`mean()`), the 95% **posterior credible interval** (_____), and the **posterior probability** of a hypothesis (_____).
- With a lot of data, the posterior credible interval and the confidence interval are _____, but only the posterior credible interval is a probability statement about β_1.