Regression and Other Stories: Regression Background

Regression methods have two purposes: prediction and comparison. Regression allows prediction of the distribution of the outcome or distributions of comparisons (sometimes called contrasts).

This section will focus on simulating fake data to understand models before fitting them to "non-fake" data.

Regression Models

As we have seen, a regression model can be written as

$$* y = \beta_0 + \beta_1 x + \epsilon *, \tag{1}$$

where y is the outcome variable, x is a predictor, β_0 and β_1 are coefficients that correspond to the linear relationship between y and x (slope and intercept, respectively), and ϵ is the error term.

Note ROS suggests interpreting β_0 and β_1 as a comparison not an **effect.** They suggest **effect** is reserved for causal inference.

While the linear model (and the functional form between x and y) may seem overly restrictive, there are many extensions.

- more predictors: $y = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \epsilon$
- non-linear model: $\log y = \beta_0 + \beta_1 \log x_1 + \epsilon$
- non-additive models: interaction $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$
  - Q assume x_2 is binary and sketch the model fit for this interaction model
- generalized-linear models: *allow non-normal probability distributions for discrete (or other) responses*
- non-parametric models: fit curves for the relationships between y and x. *power basis function or Gaussian process (534)*
- Multilevel (hierarchical) models: coefficients can vary by group (506)

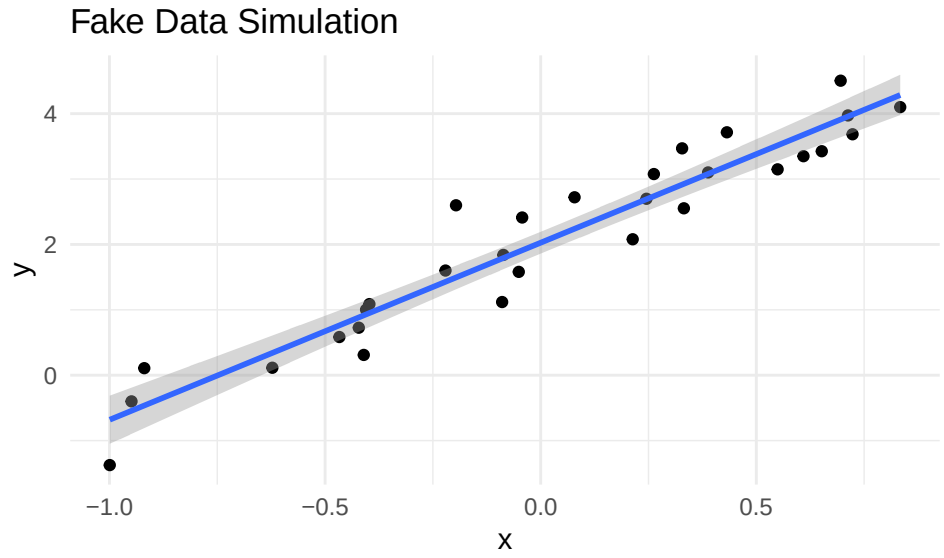

Figure 1: Fake data with n = 30, beta = (3, 3), sigma = .5

Regression Models for Fake Data

Activity: Simulate fake data and fit a regression model to the data.

1. Simulate Fake Data

```
n <- 30
x <- runif(n, -1, 1)
beta <- c(2,3)
X <- cbind(rep(1,n), x)
X_beta <- X %*% beta
sigma <- .5
y <- rnorm(n, mean = X_beta, sd  = sigma)
data_tibble <- tibble(y = y, x = x)
```

2. Visualize Fake Data

```
data_tibble %>%
  ggplot(aes(y=y,x=x)) +
  geom_point() +
  labs(title = 'Fake Data Simulation') +
  geom_smooth(method = 'lm', formula = "y ~ x") +
  theme_minimal()
```

3. Fit Model

- `lm`

```
lm_fit <- lm(y ~ x, data = data_tibble)
display(lm_fit)
```

```
## lm(formula = y ~ x, data = data_tibble)
##             coef.est coef.se
## (Intercept) 2.03     0.08
## x           2.71     0.16
## ---
## n = 30, k = 2
## residual sd = 0.45, R-Squared = 0.91
```

- `stan_glm()`

```
stan_fit <- stan_glm(y ~ x, data = data_tibble, refresh = 0)
print(stan_fit)
```

```
## stan_glm
##  family:       gaussian [identity]
##  formula:      y ~ x
##  observations: 30
##  predictors:   2
## ------
##             Median MAD_SD
## (Intercept) 2.0    0.1
## x           2.7    0.2
##
## Auxiliary parameter(s):
##       Median MAD_SD
## sigma 0.5    0.1
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

4. Estimate comparison

- How to the estimates compare to the true values: *Close but not exact*

- What happens as n gets larger or smaller?

History of Regression

‘ “Regression” is defined in the dictionary as “the process or an instance of regressing, as to a less perfect or less developed state.” ’ The term was first used in the statistical context by Francis Galton, who used linear models to understand the effect of heredity on human height. *In a related note and a darker side of statistical history, Galton also put a lot of work into eugenics.*

Predicting a child’s height form parents height he noticed that tall parents tended to have children that were taller than average, but shorter than the parent. Similarly, short parents tended to have children that were shorter than average, but taller than the parent. Thus the people’s heights **regressed** to the average heights.

The model from this data can be written as

$$y = 30 + .54x + error \tag{2}$$
$$y = 65.1 + .54(x - 65) + error \tag{3}$$

So for an average height woman (65 inches), the daughter would also be predicted to be about 65 inches. Then for each additional inch the daughter would be expected to deviate by $\pm .54$ inches.

It might seem that the heights of the daughters would eventually collapse to 65 inches, but there is enough variability in the prediction (distribution of the predicted outcome) that the overall variation has stayed fairly constant. In other words, it is possible that a tall mother can have an even taller daughter.

This phenemenon (regression to the mean) can sometimes be confused with causality. The textbook talks about student test scores on a midterm and a final exam.