Exercise 4: Linear regression; Generalized Linear regression

Benjamin Smith

July 8, 2026

Part 1: Linear regression in R

```
library(tidyverse) #ggplot2, dplyr, etc.
```

```
## Warning: package 'ggplot2' was built under R version 4.4.3
```

```
library(reshape2) #need this for melt()
library(knitr) #need this for kable
library(MASS) #contains dataset
```

```
## Warning: package 'MASS' was built under R version 4.4.3
```

Load the birthwt data. This data contains 189 observations, 9 predictors, and an outcome, birthweight, available both as a continuous measure and a binary indicator for low birth weight.

```
data(birthwt)
head(birthwt)
```

```
##    low age lwt race smoke ptl ht ui ftv  bwt
## 85   0  19 182    2     0   0  0  1   0 2523
## 86   0  33 155    3     0   0  0  0   3 2551
## 87   0  20 105    1     1   0  0  0   1 2557
## 88   0  21 108    1     1   0  0  1   2 2594
## 89   0  18 107    1     1   0  0  1   0 2600
## 91   0  21 124    3     0   0  0  0   0 2622
```

1. Plot a scatterplot of birthweight (bwt) and mother's weight (lwt).
2. Use OLS to fit the regression of birthweight on mother's weight.
3. Extract the following: estimated coefficients, standard errors, variance-covariance matrix, and confidence intervals.
4. Plot the regression line and interpret the intercept and slope
5. Does the interpretation of the intercept make sense? How might we change this?
6. Now, we want to fit a model that includes race, mother's age, and smoking status in the model. Race takes on value 1 for white, 2 for black, and 3 for other. Mother's age is continuous. Smoking status is binary. Use OLS to calculate the coefficient estimates in this model.
7. Interpret all the coefficient estimates.
8. Print the results in Rmarkdown using kable().

Part 2: Generalized linear model

Suppose that 2500 pregnant women are enrolled in a study and the outcome is the occurrence of preterm birth. Possible predictors of preterm birth include age of the woman, smoking, socioeconomic status, body mass index, bleeding during pregnancy, serum level of dde, and several dietary factors.

1. Formulate the problem of selecting the important predictors of preterm birth in a generalized linear model (GLM) framework.
2. Show the components of the GLM, including the link function and distribution (in exponential family form).
3. Describe (briefly) how estimation and inference could proceed via a frequentist approach.

Part 3: GLMs in R (Logistic regression)

Consider the space shuttle data in the MASS library. Consider modeling the use of the autolander as the outcome (variable name use).

1. Fit a logistic regression model with autolander (variable auto) use (labeled as "auto" 1) versus not (0) as predicted by wind sign (variable wind).
2. Give the estimated odds ratio for autolander use comparing head winds, labeled as "head" in the variable headwind (numerator) to tail winds (denominator).
3. Give the estimated odds ratio for autolander use comparing head winds (numerator) to tail winds (denominator) adjusting for wind strength from the variable magn.
4. If you fit a logistic regression model to a binary variable, for example use of the autolander, then fit a logistic regression model for one minus the outcome (not using the autolander) what happens to the coefficients?

```
library(MASS)
?shuttle
data(shuttle)
head(shuttle)
```

```
##   stability error sign wind   magn vis  use
## 1     xstab    LX   pp head  Light  no auto
## 2     xstab    LX   pp head Medium  no auto
## 3     xstab    LX   pp head Strong  no auto
## 4     xstab    LX   pp tail  Light  no auto
## 5     xstab    LX   pp tail Medium  no auto
## 6     xstab    LX   pp tail Strong  no auto
```

Part 4: GLMs in R (Poisson regression)

Consider the insect spray data InsectSprays. Fit a Poisson model using spray as a factor level.

1. Report the estimated relative rate comapring spray A (numerator) to spray B (denominator).

2. Consider a Poisson glm with an offset, t. So, for example, a model of the form glm(count $\sim x+$ offset(t), family = poisson) where x is a factor variable comparing a treatment (1) to a control (0) and t is the natural log of a monitoring time. What is impact of the coefficient for x if we fit the model glm(count $\sim x+$ offset(t2), family = poisson) where $t2 < -\log(10) + t$? In other words, what happens to the coefficients if we change the units of the offset variable. (Note, adding log(10) on the log scale is multiplying by 10 on the original scale.)

```
data("InsectSprays")
head(InsectSprays)
```

```
##   count spray
## 1    10     A
## 2     7     A
## 3    20     A
## 4    14     A
## 5    14     A
## 6    12     A
```