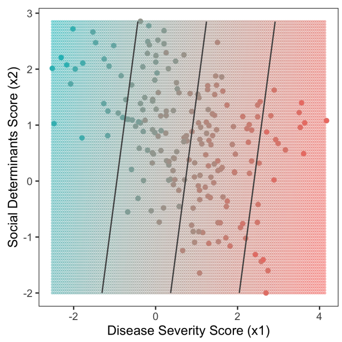
Chapter 8

Interpreting a Linear Regression Model

This chapter is devoted to understanding the structure of linear regression models. We first encountered them in Chapter 3 as "just one example" of a regression model. However, linear regression's overwhelming popularity in the clinical domain means that one cannot do clinical data science without fully understanding these models' structure and how to interpret the software output.

8.1 Biomarker Example from Chapter 3

In Chapter 3, we saw an example where information about two predictors – a disease severity score (x_1) and a social determinants score (x_2) – was used to predict the numeric level of a disease recurrence biomarker. One of the three supervised learning algorithms we tried was a **linear regression** model (Section 3.2.1). The output from that model is repeated below.




```
Call:
lm(formula = y ~ x1 + x2, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-11.9218  -3.1032   0.2891   2.8316  12.5813

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  49.8600     0.5370  92.844  < 2e-16 ***
x1           10.4372     0.2855  36.555  < 2e-16 ***
x2           -1.8824     0.3609  -5.215 4.63e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.769 on 197 degrees of freedom
Multiple R-squared:  0.9026,    Adjusted R-squared:  0.9016
F-statistic: 912.4 on 2 and 197 DF,  p-value: < 2.2e-16
```

8.2 Example: Small Cities Pollution Dataset

The following data come from an early study that examined the possible link between air pollution and mortality. The authors examined 60 cities throughout the United States and recorded the following data:

MORT	Total age-adjusted mortality from all causes, in deaths per 100,000 population
PRECIP	Mean annual precipitation (in inches)
EDUC	Median number of school years completed for persons of age 25 years or older
NONWHITE	Percentage of the 1960 population that is nonwhite
NOX	Relative pollution potential of oxides of nitrogen
SO2	Relative pollution potential of sulfur dioxide

Note: “Relative pollution potential” refers to the product of the tons emitted per day per square kilometer and a factor correcting the SMSA dimensions and exposure.

We want to predict the value of MORT (y) using the predictors PRECIP, EDUC, NONWHITE, NOX, and SO2 (x_1, x_2, x_3, x_4 and x_5). Here is the GLM output for this model in R:

```
Call:
```

```
glm(formula = MORT ~ PRECIP + EDUC + NONWHITE + NOX + SO2,
    family = "gaussian", data = d)

Deviance Residuals:
   Min       1Q   Median       3Q      Max
-91.38   -18.97    -3.56    16.00    91.83

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 995.63646   91.64099  10.865 3.35e-15 ***
PRECIP        1.40734    0.68914   2.042 0.046032 *
EDUC        -14.80139    7.02747  -2.106 0.039849 *
NONWHITE      3.19909    0.62231   5.141 3.89e-06 ***
NOX          -0.10797    0.13502  -0.800 0.427426
SO2           0.35518    0.09096   3.905 0.000264 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 1375.723)

    Null deviance: 228275  on 59  degrees of freedom
Residual deviance:  74289  on 54  degrees of freedom
AIC: 611.56

Number of Fisher Scoring iterations: 2
```

Side note: Most models can be fit multiple ways. Linear regression models are normally fit using **ordinary least squares** and the `lm` package, as opposed to maximum likelihood and the `glm` package. The coefficients and most of the output are exactly the same:

```
Call:
lm(formula = MORT ~ PRECIP + EDUC + NONWHITE + NOX + SO2,
   data = d)

Residuals:
   Min      1Q Median      3Q     Max
-91.38 -18.97  -3.56  16.00  91.83

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) 995.63646   91.64099  10.865 3.35e-15 ***
PRECIP        1.40734    0.68914   2.042 0.046032 *
EDUC        -14.80139    7.02747  -2.106 0.039849 *
NONWHITE      3.19909    0.62231   5.141 3.89e-06 ***
NOX          -0.10797    0.13502  -0.800 0.427426
SO2           0.35518    0.09096   3.905 0.000264 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 37.09 on 54 degrees of freedom
Multiple R-squared:  0.6746,  Adjusted R-squared:  0.6444
F-statistic: 22.39 on 5 and 54 DF,  p-value: 4.407e-12
```

Question 8.1

Interpret the values of each of these coefficients. Based on the coefficient values and their standard errors, which predictor(s) do you think have the greatest impact on mortality?

Question 8.2

In this model, is the effect of one predictor (say, `PRECIP`) impacted by the value(s) of any of the other predictor(s)? How does this differ from the other regression algorithms we've seen (KNN and decision trees)? What are the advantages and disadvantages of this choice?