Chapter 8: Interpreting a Linear Regression Model

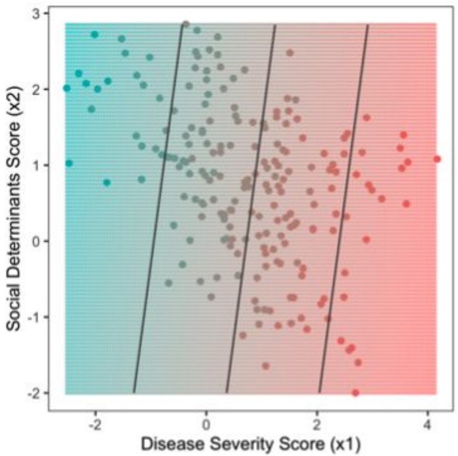
Modern Clinical Data Science
Chapter Guides
Bethany Percha, Instructor

How to Use this Guide

- Read the corresponding notes chapter first
- Try to answer the discussion questions on your own
- Listen to the chapter guide (should be 15 min, max) while following along in the notes

Question 8.1

What are the number of samples, n, and the number of predictors, p, for this dataset?



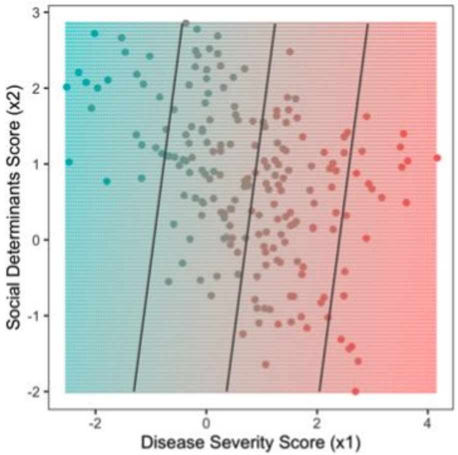

```
Call:
lm(formula = y ~ x1 + x2, data = df)

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
```

Question 8.4

Looking at the form of the linear regression model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$$

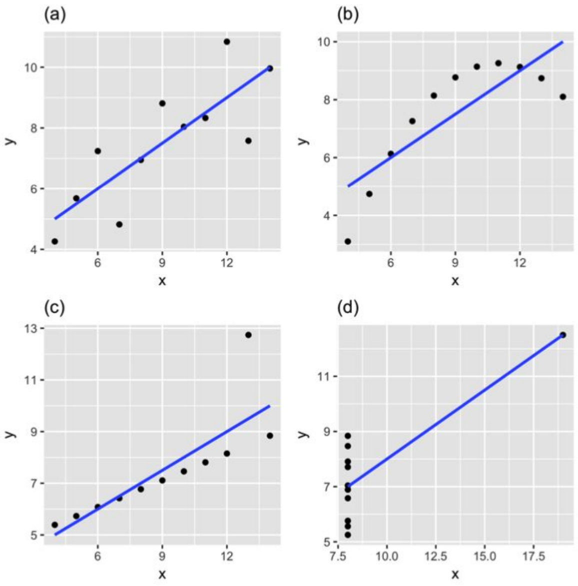
what does the value of each of the βs mean? What is β_j telling us about how y varies with the predictor j, all else being equal?



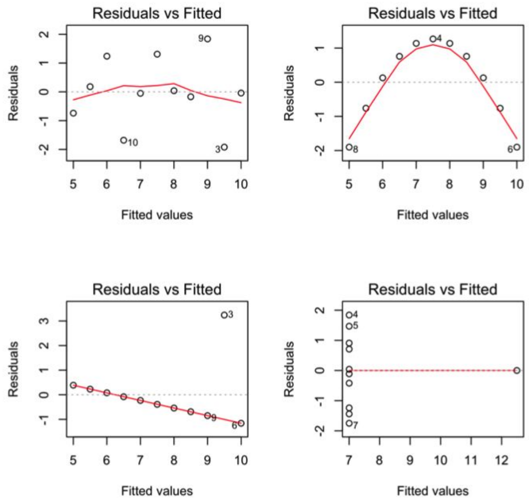

```
Call:
lm(formula = y ~ x1 + x2, data = df)

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
```

Question 8.3

The four plots below show a famous dataset called **Anscombe's quartet**. The regression lines produced by fitting a linear regression model to each dataset are identical, but only one dataset actually fulfills the assumptions of a linear regression model.



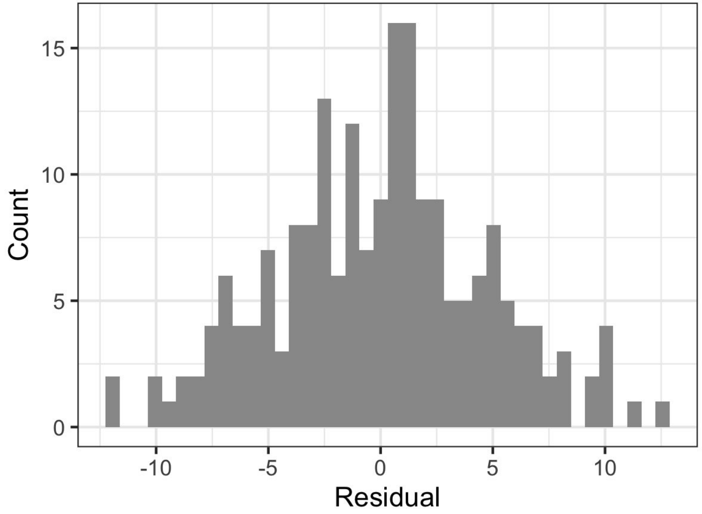
We can check these assumptions by examining plots of the residuals vs. fitted values of the model (here the "fitted value" of point i means $\hat{y}^{(i)}$).



Which of the four datasets fulfills the assumption of a linear regression model that the error has constant variance? How can you tell?

Question 8.2

Estimate the maximum, minimum, and mean residuals from this graph. Do they match what is in the model output?



$$\hat{\sigma}^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)})^2$$

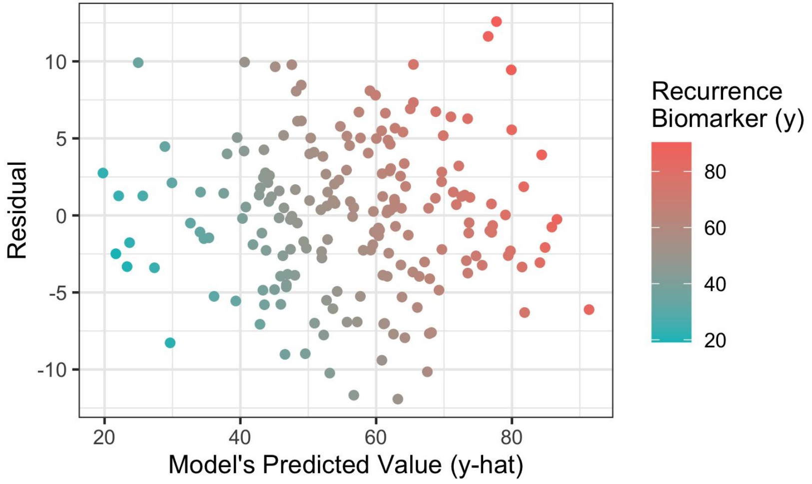
Question 8.5

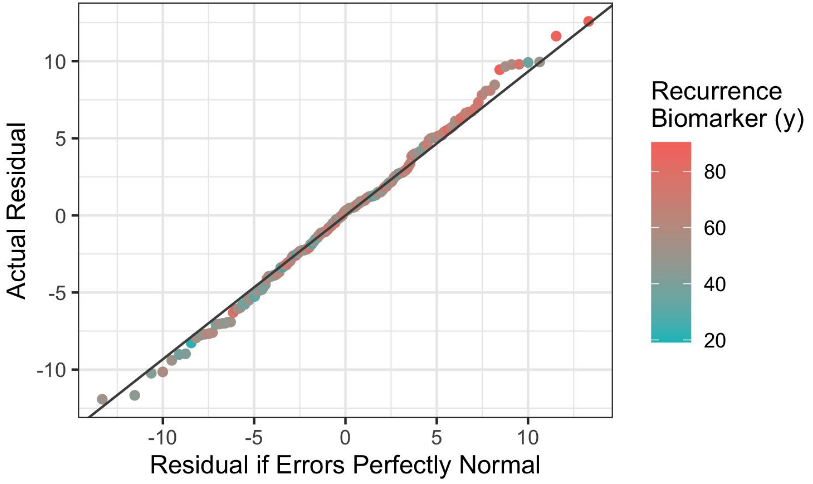
The sum of the squared residuals for our model is 4480.678. There are $n = 200$ datapoints, and the number of predictors, p, is 2. Calculate $\hat{\sigma}$ for this model. Do you see this number anywhere in the model output? What is it called?

```
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
```

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$$

Errors have constant variance (assume normal).





Question 8.6

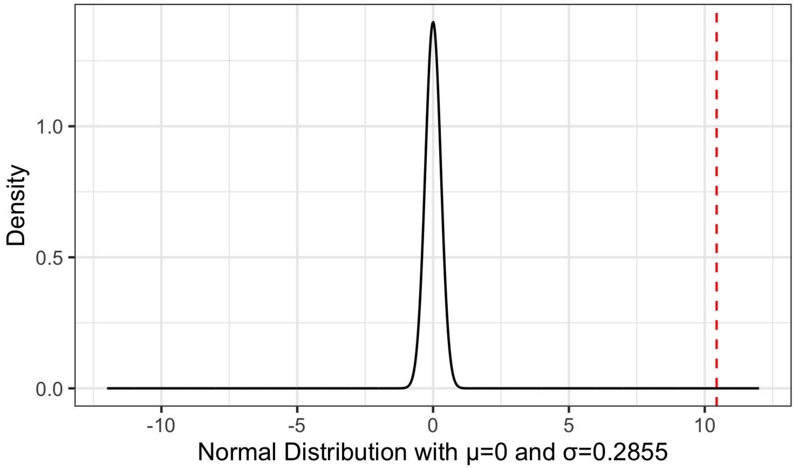
The standard errors attempt to capture how much we expect our estimates of the model coefficients to vary if we were to refit the model using a different dataset, provided that the new dataset is similar to (i.e., sampled from the same population distribution as) the one used to fit the model. On average, approximately how much would we expect β_0 (the intercept) to deviate from its fitted value of 49.8600? How much would we expect β_1 and β_2 to deviate from their fitted values?

```
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
```

Question 8.7

Sketch the null distributions for the hypothesis tests of our three regression coefficients, β_0, β_1, and β_2. Do you see why the *p*-values for these tests are so low?

```
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
```



Question 8.8

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?

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

```
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 8.9

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?

Question 8.10

Is a normal distribution the right distribution to model an outcome of age-adjusted mortality (MORT)? Why or why not? Look back at our discussion of the normal distribution in Chapter 4 if you need a refresher.