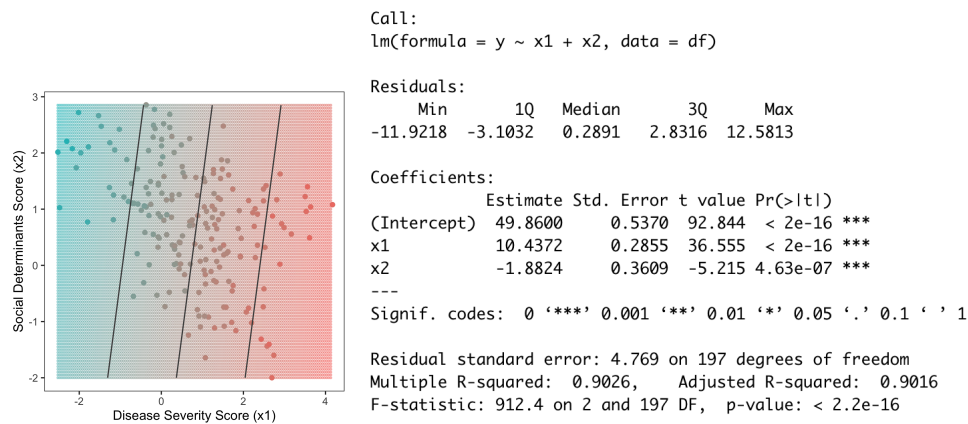# Chapter 7

Interpreting a Linear Regression Model

This chapter is devoted to understanding the structure of linear regression models. We first encountered them in Chapter 2 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 output produced by model fitting software.

7.1   Biomarker Example from Chapter 2

In Chapter 2, 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 level of a disease recurrence biomarker. One of the three supervised learning algorithms we tried was a **linear regression** model (Section 2.2.1). 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
```

Question 7.1

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

But what do all these numbers *mean*?

7.2 Understanding the Model Summary

A linear regression model looks like this (see also Chapter 2, Section 2.2.1):

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

where we assume that the error, ε, is normally distributed, $\mathcal{N}(0, \sigma)$. Another way of saying this is that we are assuming the outcome, y, is normally distributed with mean $\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ and standard deviation σ.

7.2.1 The Call

The first line of the output is just repeating the call you made to the `lm` function in R to fit the model. The `lm` package fits linear regression models using ordinary least squares. However, linear regression models are also a type of generalized linear model (see Chapter 12) and can be fit using maximum likelihood. In R, if you use the `glm` package with `family = "gaussian"`

you should get identical coefficients and error estimates to what you get using the `lm` package.
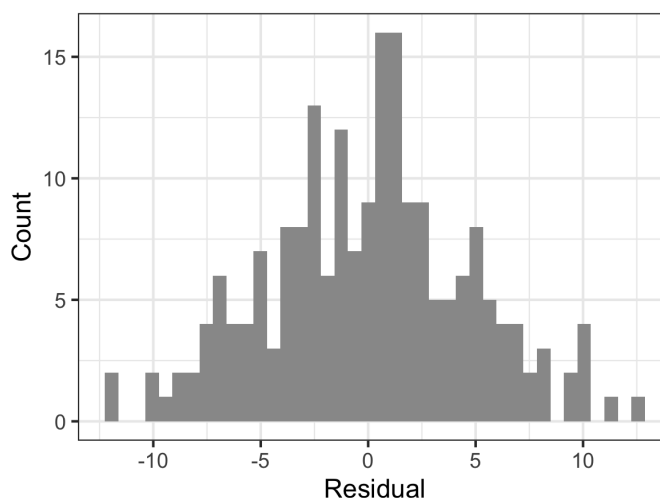
7.2.2 The Residuals

A **residual** is a measure of how much the true outcome value (y) of one datapoint differs from the model's prediction. For a linear regression model, the residual of training point i is:

$$\text{residual}^{(i)} = y^{(i)} - \hat{y}^{(i)}$$

where $\hat{y}^{(i)}$ is the model's prediction:

$$\hat{y}^{(i)} = \beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)}.$$

Here is a histogram of the residuals for this model:


Question 7.2

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

The residuals play an important role in linear regression models because they are what allow us to estimate σ. They also play an important role in

model diagnostics because they enable us to check one of the core model assumptions: the assumption that σ is constant. We can check this assumption by making a plot of the residuals vs. $\hat{y}$:
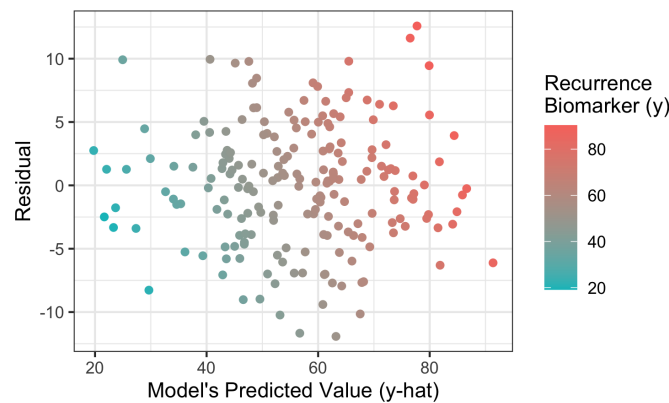
We can also check the normality of the residuals by making a plot of their values vs. what we would expect if the residuals were perfectly normally distributed with the same mean and standard deviation:
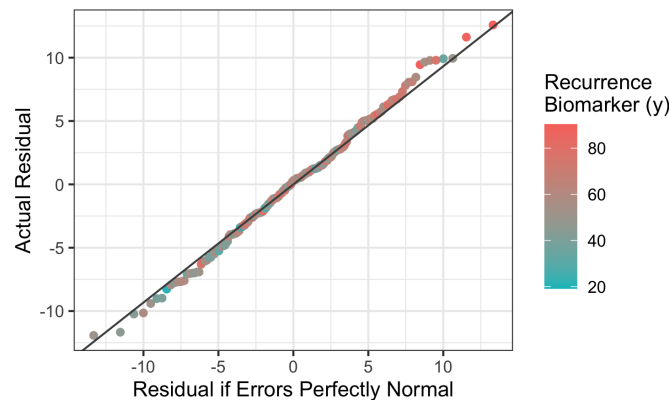

This is not quite a **QQ-plot**, one of the standard diagnostic plots of linear regression, because it's plotting the actual values of the residuals instead of their quantiles. But aside from the axis scales, it's exactly the same. We'll see formal QQ-plots later.

The fact that the points in the second plot lie close to a line is a good indication that the residuals are normally distributed. Another thing to notice
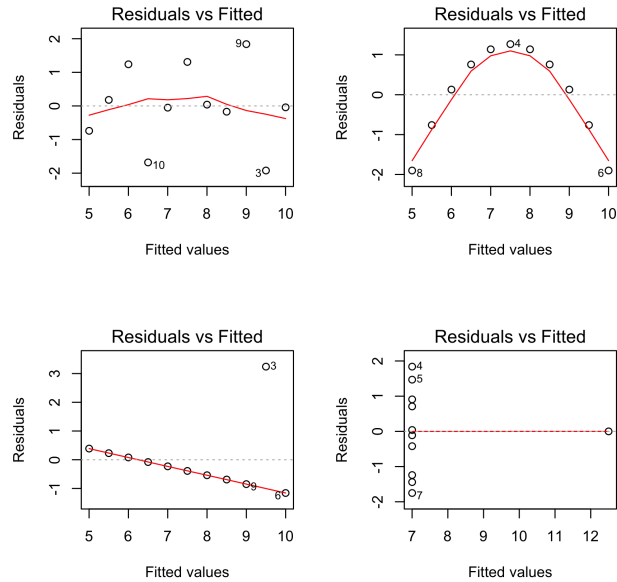
from this plot is that the colors of the points (their y values) are unrelated to their residuals.

Question 7.3

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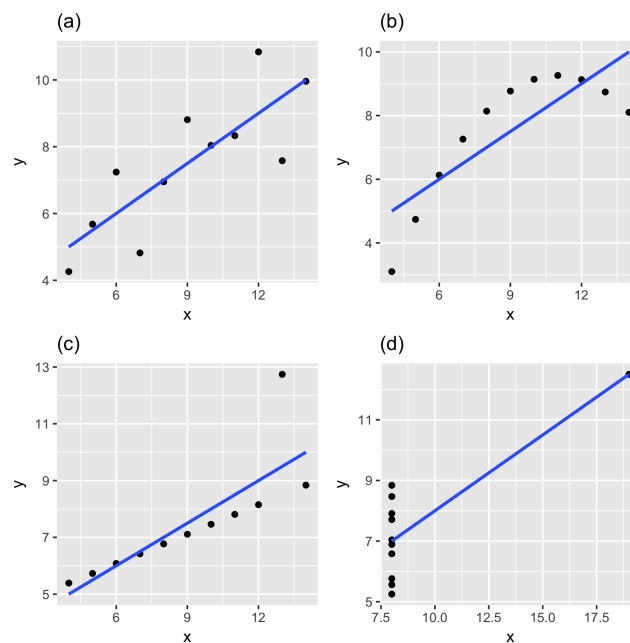
We can check these assumptions by examining plots of the residuals vs. fitted values of the model (here the "fitted value" of point i means $\hat{y}^{(i)}$).

Residuals vs Fitted

Residuals

Fitted values

Residuals vs Fitted

Residuals

Fitted values

Residuals vs Fitted

Residuals

Fitted values

Residuals vs Fitted

Residuals

Fitted values

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

7.2.3 Coefficients and Standard Errors

Although residuals are important, the parts of the model output that will be scrutinized, reported in papers, etc. are the coefficients, along with their standard errors and hypothesis tests. The coefficients are what create the regression surface, which captures the model's prediction for every point in the feature space (see Section 2.2.1).

There are actually a few different ways to derive the coefficients in a linear regression model. The most common way uses **least squares**, which adjusts the coefficients $\beta_0, \beta_1, \ldots, \beta_p$ until the sum of the squared residuals is

minimized:

$$\text{SSE} = \sum_{i=1}^{n} (y^{(i)} - \hat{y}^{(i)})^2$$

$$= \sum_{i=1}^{n} (y^{(i)} - \beta_0 - \beta_1 x_1^{(i)} - \beta_2 x_2^{(i)} - \cdots - \beta_p x_p^{(i)})^2$$

It turns out that you can find the optimal values of the βs analytically by taking derivatives of this thing and setting them equal to zero. Alas, this requires some matrix multiplication and the taking of matrix inverses, so we will save it for a later chapter. Suffice it to say that the βs are adjusted to minimize the SSE, and the values in the model output are the optimal values.

Question 7.4

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?

The model's overall estimate of σ, the standard deviation of the error term, is obtained very naturally by averaging over the residuals, taking into account the number of predictors, p:

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

Question 7.5

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?

The standard errors of the model coefficients likewise require matrix multiplication to fully understand, but they are related to three factors: (1) the value of $\hat{\sigma}$, (2) the spread of the values of the corresponding covariate about

its mean (more spread will decrease the standard error), and (3) correlations between that covariate and other covariates in the model (tighter correlations will increase the standard error).

Question 7.6

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?

7.2.4 Hypothesis Tests of Coefficients

Armed with our coefficients and standard errors, we can perform a hypothesis test on each regression coefficient. Our null hypothesis in each case will be that the true value of that coefficient is zero: that is, it has no effect on the outcome. Under the null hypothesis that $\beta_j = 0$ and assuming n, our number of samples, is large enough, the quantity $\hat{\beta}_j / \text{se}(\hat{\beta}_j)$ will be distributed according to a Student's T distribution (see Chapter 3, Section 3.9) with $n - p - 1$ degrees of freedom.

Question 7.7

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?

7.2.5   Other Model Output

The software also provides some other output. The quantity **R-squared** is defined as:

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

$$= 1 - \frac{\sum_{i=1}^{n} (y^{(i)} - \hat{\beta}^T x^{(i)})^2}{\sum_{i=1}^{n} (y^{(i)} - \bar{y})^2}$$

or rather, the proportion of total variance in y explained by the model. The **adjusted R-squared** is almost exactly the same except it fixes a source of bias in R^2, namely that R^2 will favor models with more parameters. Adjusted R^2 penalizes models with more parameters. It is defined as:

$$R^2_{\text{adj}} = 1 - \frac{n-1}{n-p-1} \frac{\sum_{i=1}^{n} (y^{(i)} - \hat{y}^{(i)})^2}{\sum_{i=1}^{n} (y^{(i)} - \bar{y})^2}$$

$$= 1 - (1 - R^2) \frac{n-1}{n-p-1}$$

The **F-statistic** is the ratio of two variances: the variance in the outcome, y, that is explained by the model parameters ("sum of squares of regression", or SSR) and the residual, or unexplained variance ("sum of squares of error", or SSE). The corresponding **F-test** tests the null hypothesis that a model with no independent variables (that is, an intercept-only model with $\beta_1 = 0$ and $\beta_2 = 0$) fits the data as well as our model. The F-statistic follows an F distribution with p and $n - p - 1$ degrees of freedom (see Chapter 3, Section 3.10). The p-value reported in the model output is the p-value for this hypothesis test.

7.3   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 output for a fitted linear regression model:

```
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 7.8

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?

Question 7.9

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 7.10

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 3 if you need a refresher.