Let's take as our running example the data about sales and advertising revenue on three different media: tv, radio and newspaper. Thus, my model is:

$$\begin{aligned} sales &= h_\beta(TV, Radio, Newspaper) \\ &= \beta_0 + \beta_{TV} \times TV + \beta_{Radio} \times Radio + \beta_{Newspaper} \times Newspaper \end{aligned}$$

Suppose I fit a linear regression model and get estimates for each coefficient. Now what?

Goodness of fit

The total sum of squares (TSS) is the error I would get if I fit a linear regression model in which every β is forced to be zero except for β_0. In other words, my model always predicts $\bar{y}$, the average value of y.

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

where $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ is the average value of y.

Total sum of squares measures the variability in Y.

Residual sum of squares is the residual error of the model that I fit:

$$RSS = \sum_{i=1}^{n} (y_i - h_\beta(x_i))^2$$

One measure of goodness of fit is R^2,

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

In other words, it *measures the proportion of variability in Y that can be explained using X.* An R^2 statistic can range from 0 to 1 with 1 indicating that the linear regression model perfectly predicts all y values.

What do the coefficients mean?

We interpret β_j as the effect that X_j has on Y "all else being equal". Example:

If $\beta_{Radio} = 0.189$, then this means that if we hold the rest of our advertising budgets fixed, for every additional thousand dollars spent on radio advertising leads to the additional sale of 189 units. (The Sales variable is in thousands of units and Radio is in thousands of dollars.)

Which predictor variables are "important"?

We cannot simply look at the size of the coefficent to see which coefficient is the most "important." (Why not?)

We can use hypothesis testing to help us infer the significance of each coefficient.

Background assumptions: we assume that $Y = h_\beta(X) + \epsilon$ and that ϵ is a zero-mean Gaussian (aka normal) random variable with some fixed (but unknown variance) which we'll denote as σ^2. These assumptions are *not* being tested!

Null hypothesis: $\beta_j = 0$; alternative hypothesis $\beta_j \neq 0$.

Our usual approach for hypothesis testing is this:

- Compute the test statistic on our data. Let's call this the *observed* value.
- Simulate a random sample of data assuming the null hypothesis is true.
- Calculate the test statistic for each sample.
- Calculate the p-value as the fraction of samples on which we observe a test statistic value larger than the observed value.

Problem #1: What should our test statistic be? β_j isn't meaningful by itself because, for instance, if we changed the scale of the data (dollars instead of thousands of dollars), we would get a different value of β_j.

Our test statistic is based on the following fact: if our background assumptions are true, then this quantity

$$\frac{\beta_j - \beta_j^{TRUE}}{SD(\beta_j)}$$

is approximately normally distributed with mean zero and standard deviation of 1. Note: $SD(\beta_j)$ is the standard deviation of β_j, i.e., how much the estimated β_j deviates, on average, from its "true" value β_j^{TRUE}. (Technically, it should be standard error not standard deviation and it follows a Student's t distribution with $n-2$ degrees of freedom, but these details are not essential.)

The value of SD depends on σ^2, the variance of the error term ϵ. (Why might this be so?) It also depends on the distribution X: in particular, if the x_i values are spread out, then SD tends to be lower. (Why does this make sense?)

If the null hypothesis is true, then $\beta_j^{TRUE} = 0$. So our **test statistic** could be $\frac{\beta_j - 0}{SD(\beta_j)}$... if we knew what SD was!

If we knew what SD was, then we could compare this test statistic from simulated random samples from normal distribution (or just look up the p-value).

Problem #2: how do we estimate SD? If we could generate fresh new datasets X and Y, we could run linear regression on each one, get β_j and then simply measure the standard

deviation across samples. But we since we don't know the true distribution that generates X, the true β coefficients, nor the value of σ^2, we cannot do this.

Let's use **bootstrap sampling**. Bootstrap sampling is described in the book. The basic idea is to treat the data we have as an estimate of the "true" distribution. One bootstrap sample is n records sampled with replacement from our dataset (which has n records). We take M bootstrap samples generated M datasets $D^{(1)}, \ldots, D^{(M)}$.

For each bootstrap sample dataset $D^{(i)}$, we run linear regression and get some β vector of coeffcients. Let $\beta_j^{(i)}$ denote the β_j coefficient obtained from running linear regression on bootstrap sample i.

Our estimate of $SD(\beta_j)$ is $\sqrt{\frac{1}{M-1}\sum_{i=1}^{M-1}(\beta_j^{(i)} - \bar{\beta}_j)^2}$ where $\bar{\beta}_j$ is the mean value across bootstrap samples $\bar{\beta}_j = \frac{1}{M}\sum_{i=1}^{M}\beta_j^{(i)}$.