Linear Regression

Reading

Ch. 8.1-8.2

Motivation

One of the most important models in statistics is the linear regression model. The idea behind this model is that **we can use a line to describe the relationship between a response variable and one (or more) explanatory variables**. In other words, this is an appropriate model to consider when the relationship between a response variable and explanatory variables is **linear** - hence the name linear regression! As with other topics we have seen in this course, **there are some assumptions that we have to be comfortable making if we want to use this model** - more on those later.

Anatomy of a Simple Linear Regression Model

In a linear regression model, we assume that the relationship between two variables, x (explanatory/predictor) and y (response) can be modeled with a straight line plus some error:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \ i = 1, ..., n$$

- y_i: observed value of the response variable for observation i
- x_i: observed value of the predictor for observation i
- ϵ_i (pronounced epsilon): error
- β_0: y-intercept parameter - unknown, we need to estimate this
- β_1: slope parameter - unknown, we need to estimate this

Once the model has been fit, we get estimates of the parameters β_0 and β_1, which are $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively. The fitted model has the form

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \ i = 1, ..., n$$

Correlation

- Describes the strength of the **linear** relationship between two variables.
- It is always **between -1 and 1**
- It is **unitless**
- Can be computed using
  - the formula:

$$R = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

  - R: `cor(data$y, data$x)` ***This is more important for this course***

Least Squares Regression

- Conditions for least squares (linear) regression
  1. **Linearity.** The data should show a linear trend (linear relationship between x and y).
     - Check condition using residual plot
  2. **Nearly normal residuals.** Residuals must be nearly normal
     - Check condition using Q-Q plot or histogram of residuals (see Ch.4 R lab for a refresher)
  3. **Constant variability.** The spread of points around the least squares line is roughly constant.
     - Check condition using residual plot
  4. **Independent observations.** Observations should be independent of each other (this is all we will work with in this class). Examples of when this is violated include time series data, e.g., daily stock price or daily temperature - in these cases there is an underlying structure to how these observations are related that has to be taken into account.
     - Check condition using residual plot

  *** See pg. 319, Figure 8.12 for nice examples of when these conditions are violated.

- Finding the least squares line
  - The basic idea is the we find a line (i.e. estimate β_0 and β_1) such that the **residuals** are as "small as possible".
    * **Residuals**: what is left over after accounting for the model fit
    * Data=Fit+Residual
    * The **residual** ($\hat{\epsilon}_i$) of the i^{th} observation (x_i, y_i) is the difference of the observed response (y_i) and the response we would predict based on the model fit ($\hat{y}_i$).

$$\begin{aligned}\hat{\epsilon}_i &= y_i - \hat{y}_i \\ &= y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\end{aligned}$$

    * What values can the residuals take on?
    * Where is the distribution of residuals centered?

Least Squares Line - Summary Statistics

- Estimating parameters from summary statistics:

	% high school grad	% in poverty
mean	$\bar{x} = 83.11\%$	$\bar{y} = 15.5\%$
sd	$s_x = 7.36\%$	$s_y = 6.38\%$
		$R = -0.68$

$\hat{\beta}_1 = \frac{s_y}{s_x} R$ – has units: units of y/units of x

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ – has units: what will they be?

- Extracting parameter estimates from `lm()` output
  - We can fit the linear model in R using the `lm()` function. The general form is `lm(data=data, y~x)`, where data is the name of your dataset, y is the name of your response variable, and x is the name of your predictor
  - You should assign a name when you run the `lm()` function so you can reference it later (e.g., lm_data)
  - Use the `summary()` function to extract information about the model fit (e.g., `summary(lm_data)`). This is a portion of what you get:

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	64.59437	0.94619	68.27	<2e-16
family_income	-0.59075	0.01134	-52.09	<2e-16

** We will talk about the Std. Error, t value, and Pr(> |t|) pieces of this when we get to inference (Ch. 8.4).

Interpretation

- Intercept ($\hat{\beta}_0$): describes the estimated outcome of y if $x = 0$ *and* the linear model is valid all the way to $x = 0$ (often not the case - be careful!)
- Slope ($\hat{\beta}_1$): describes the estimated difference in the response variable y if the predictor x for a case happened to be one unit larger
- Residual ($\hat{\epsilon}_i$): the amount by which the model overestimates/underestimates the response for observation i.
  - If the model overestimates an observation, will the residual be negative or positive?
  - If the model underestimates an observation, will the residual be negative or positive?

Extrapolation

- Applying a model estimate to values outside of the range of the original data
- Involves assuming that the approximate linear relationship will be valid in places where it has not been analyzed
- Sometimes interpretting the intercept is extrapolation

Model Fit Assessment

- R^2: "coefficient of determination"
  - Commonly used to evaluate strength of fit of a **linear** model
  - Calculated as the square of the correlation coefficient; range between 0 and 1

  - Available as output from model summary
  - Tells us what percent of variability in the response variable is explained by the model

- Adjusted R^2
  - Also available as output from model summary
  - This is important in multiple regression (when we have more than one predictor variable in the linear model)