Section 8: Linear Regression

TF: Reagan Rose (rrose@g.harvard.edu)

November 4, 2015

Motivation

So far in this course, we've been talking a lot about **significant differences** and **associations**. These are important, but can be hard to interpret practically. Now we want to be able to find stronger relationships between variables. In particular, we're interested in the question: *does the response variable vary deterministically as a function of the independent variable*?

In a vacuum, this would be easy. If you can draw a straight line (with non-zero slope) through the points, then you have a significant relationship. The problem (and purpose of statistics) is that there is a bunch of noise in the world, and so its possible to just observe things randomly. Thus we have to determine: *is the evidence strong enough to convince us that there is a relationship, above and beyond that resulting from random noise?*.

Model

Assume we have a dataset with $i = 1, \ldots, n$ observations, each with one response variable Y_i, and p predictor variables $X_{i1}, \ldots, X_{ip}$. In the case of **simple linear regression**, we assume we have only one predictor X_i. In this case, our regression equation is

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. In **multiple linear regression**, we expand this model to include multiple predictors. This model can include as many predictors as you want, but no more than the number of observations you have (i.e. $p < n$). The regression equation in this case is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \epsilon_i.$$

Assumptions

The regression framework relies on the following assumptions

1. **Linearity**: $E[Y_i|X_i]$ is linear (i.e. not curved/polynomial)
2. **Homoskedasticity**: Also known as *constant variance*, this requires that the variance of the error term ϵ_i is independent of X_i
3. **Normality**: For fixed X_i, $Y_i|X_i \sim \mathcal{N}(\beta_0 + \beta_1 X_i, \sigma^2)$
4. **Independence**: Y_i is independent of Y_j for all $i \neq j$

To check these assumptions, every time you run a regression you should include

1. **Scatterplot**: to check linearity, homoskedasticity
2. **Histogram**: look at the histogram of the *residuals* to check for normality and independence

Estimation by Least Squares

Two terms you should get very comfortable with are

- **Fitted value**: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_i$ is the predicted value of Y_i using the regression model
- **Residual**: $e_i = Y_i - \hat{Y}_i$ is the "prediction error", and is **not** the same as the error term ϵ_i! Residuals give us a way of understanding how good our regression model is, since if it is perfect at predicting our values Y, then all the residuals will be 0

With that in mind, the overall goal of the **least squares method of estimation** is to *minimize the overall prediction error*. We do this by minimizing the sum of the residuals, that is:

$$\hat{\beta}_0, \hat{\beta}_1 = \text{ arg min} \sum_{i=1}^{n} e_i^2 = \text{ arg min} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2.$$

This yields the **least squares estimators**:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \sim \mathcal{N}\left(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{(n-1)S_X^2}\right)\right),$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \sim \mathcal{N}\left(\beta_1, \sigma^2 \frac{1}{(n-1)S_X^2}\right).$$

We can estimate the assumed constant variance of the observations, σ^2, by

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2} \sim \frac{\sigma^2}{n-2}\chi^2_{n-2}.$$

Inference by t-Tests

Recall that the question we are trying to answer with linear regression is: *do we have sufficient evidence to convince us that there is a relationship between the variables in question, above and beyond that resulting from random noise*? Now that we have a **reference distribution** for $\hat{\beta}_0$ and $\hat{\beta}_1$, we can test this! Unfortunately, the reference distribution of $\hat{\beta}_1$ relies on σ^2, which is (almost) always unknown in real-life settings. Therefore, we must use the *t-test* to evaluate our evidence.

Observe that to test whether there is a significant relationship between X and Y, the only coefficient we care about is $\hat{\beta}_1$. Moreover, our null hypothesis is that there is no relationship, namely:

$$H_0 : \beta_1 = 0,$$

and under this assumption we have

$$\hat{\beta}_1 \sim \mathcal{N}(0, \frac{\sigma^2}{n-1} S_X^2),$$

so the t-statistic is

$$T = \frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{(n-1)S_X^2}} \sim t_{n-2}.$$

Knowing the standard errors for the estimators, we can form confidence intervals using the formula:

$$\hat{\beta}_1 \pm t_{n-2,1-\alpha/2} \frac{\hat{\sigma}}{\sqrt{(n-1)S_X^2}}.$$

Exercises

Question 1. Mathematics of Regression
Assume we're working with a standard linear regression model, $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ with $i = 1, \ldots, n$ observations where $\epsilon_i \sim N(0, \sigma^2)$.

(a) Derive the distribution of $\bar{Y}$.

(b) Use the fact that $Cov(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\sigma^2 \bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$ to prove that $Cov(\bar{Y}, \hat{\beta}_1) = 0$.

(c) Let $\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 X_0$ represent the predicted mean value of the response, Y, given some value of the predictor X_0. Derive the distribution of $\hat{\mu}$.

(d) Derive the test statistic for for the two-sided hypothesis $H_0 : \hat{\mu} = \mu_0$ at the $(1 - \alpha)\%$ confidence level. Under what conditions will you reject H_0?
Note: You can flip this test statistic around to form a *confidence interval* - you'll have to do this on your HW!

Question 2. Toy Example
Come up with a "toy example" of how regression should work and run some simulations in R to convince yourself.

Question 3. Data of Dating Example
Download the "eharmony.csv" dataset from Canvas or Github and load it into R. Imagine you work for eHarmony and want to better understand what type of person is successful at finding matches on your site.

(a) Run some exploratory data analysis to see what's going on in the dataset. What variables do you think might be related to number of matches?

(b) After some initial exploration, we're interested in better understanding how age affects the number of matches received at eHarmony. Run a regression to answer this question. Make sure to check the assumptions of regression and state whether you think they hold for this data.

(c) What is the effect of age on number of matches? Provide a 95% confidence interval for this estimate.

(d) Find a 95% confidence interval for the average number of matches received by a 60 year old individual.

(e) Find a 95% prediction interval for the number of matches received by a 60 year old.