Lectures on Simple Linear Regression
Stat 431, Summer 2012

Hyunseung Kang

July 16 - 18, 2012

Last Updated: July 18, 2012 11:59PM

1 Introduction

Previously, we have been investigating various properties of the population F_θ where θ is the parameter of the population through sampling. All of our analysis were univariate. That is, we only studied one particular measurement, say price of a cereal box, the mean price of a showcase on "Price is Right", or the perceived age of the instructor.

However, little attention has been devoted to studying multivariate measurements. In particular, we have yet to develop tools to study *relationships* between measurements. Consider the following examples

Example 1.1. Suppose you want to study the relationship between the height of a boy and his father. Here, we take two measurements per child, the child's height and his/her father's height, and we are interested in whether there is a relationship between the heights.

Example 1.2. We are interested in studying the relationship between height and weight of UPenn students. Here, each Penn student has two measurements, height and weight.

Example 1.3. Geologists are interested in studying the relationship between seismic activity from different outposts and the distance from these outposts to a nearby active volcano. Here, each outpost collects two measurements, the seismic activity and the distance from it to the nearby active volcano.

These lectures will illustrate how we can study *linear* relationships between *two* measurements. We'll develop tools to derive linear relationships, provide tools to infer whether such a relationship exists, and finally consider various diagnostic tools to validate the derived linear relations.

2 "How do you draw the *best* line?"

Suppose we collect n pairs of measurements (X_i, Y_i) from individuals from a population. Any *linear relationship* between two variables, if it exists, can be summarized by the equation for a line

$$Y_i = \beta_0 + \beta_1 X_i \tag{1}$$

If there are two pairs of points (X_1, Y_1) and (X_2, Y_2), then finding β_0 and β_1 would be easy; simply use the slope and intercept formulas you learned in middle school. However, it is often the case that we have n pairs of (X_i, Y_i) and we want to fit *one* single line that captures the linear relationship between X and Y. So, how do we proceed to find this one line?

Amongst many candidates of lines, we can choose the one that is the "best" in some sense. In least squares regression, we define the "best" line to the line that minimizes the square of the *residuals.* Mathematically, we attempt to minimize the following quantity

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2 = \sum_{i=1}^{n} r_i^2, \text{ where } r_i = Y_i - (\beta_0 + \beta_1 X_i) \tag{2}$$

To minimize this, we take partial derivatives with respect to β_0 and β_1 and set both derivatives equal to zero[1],

$$\frac{\delta}{\delta \beta_0} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2 = 0 \tag{3}$$

$$\frac{\delta}{\delta \beta_1} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2 = 0 \tag{4}$$

The β_0 and β_1 we obtain from the derivatives are denoted as $\hat{\beta}_0$ and $\hat{\beta}_1$, which are

$$\hat{\beta}_0 = \bar{Y} - \rho \bar{X} \tag{5}$$

$$\hat{\beta}_1 = \rho_{x,y} \frac{\hat{\sigma}_y}{\hat{\sigma}_x} \text{ where } \rho_{x,y} \text{ is the correlation between } X \text{ and } Y \tag{6}$$

Because the minimization is done over a convex function on a convex set, we know that the solutions to equations, (3) and (4) are global minimizers of equation (2).

Once we fit the line, we can plug in X_i to get a prediction for Y_i at X_i, denoted as $\hat{Y}_i$. We can measure the deviation from the predicted Y_i, $\hat{Y}_i$, to the actual value Y_i as *residuals*, or

$$r_i = Y_i - \hat{Y}_i$$

r_i is different from ϵ_i because r_i is derived from $\hat{\beta}_0$ and $\hat{\beta}_1$ while ϵ_i is derived from β_0 and β_1.

3 Inference

Our sample of n pairs of measurements, (X_i, Y_i) helps us study the relationship between the two measurements. For example, by fitting a line based on the sample, we have an estimate of the linear relationship between the two variables. However, because this is a random sample of the population, there is some uncertainty as to whether the fitted line is the actual, true relationship between the two. Hence, the fitted line we constructed can help us *infer* about the underlying relationship between the two variables.

3.1 Assumptions

Similar to how we made assumptions about our population in the one-sample and two sample tests to derive inference (e.g. hypothesis testing, CIs,etc), there are standard assumptions we make about how our measurements are drawn. These are, in essence, assumptions about the populations of X_i and Y_i.

Assumption 1. A simple linear regression model assumes the following about X_i and Y_i

1. X_i are assumed to be fixed, non-random quantities. Y_i's are the only random quantities in (X_i, Y_i)

2. Y_i is related to X_i by the following *linear* relationship

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{7}$$

   where ϵ_i are i.i.d random variables, commonly referred to as errors. Here, β_0 and β_1 are the parameters that characterize the true underlying relationship between X_i and Y_i

[1] There are other (easier) ways to minimize this quantity, but you need to know some linear algebra

3. ϵ_i is Normally distributed

4. ϵ_i have the same variance, σ^2, for any value of X_i. Another way to say this is that Y_i's have *homoscedastic* errors.

These assumptions may seem a bit idealistic for real-world data. However, without these assumptions, classical inference on regression would be difficult. Modern techniques try to relax these assumptions by making weaker assumptions about the structure of the relationship (e.g. nonparametric regression) or incorporating different families of distributions for the errors (e.g. generalized linear models)

3.2 Inference about $\hat{\beta}_0$ and $\hat{\beta}_1$

From section 2, we learned how to obtain $\hat{\beta}_0$ and $\hat{\beta}_1$, estimates of the true underlying relationship based on the sample. Because these estimates are derived from a random sample, they must have distributions associated with them. These *sampling distributions* are listed below

Proposition 1. *Under assumption 1, $\hat{\beta}_0$ and $\hat{\beta}_1$ are bivariate normals. Specifically,*

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}}\right)\right)$$
$$\hat{\beta}_1 \sim N\left(\beta_1, \sigma^2\frac{1}{S_{xx}}\right)$$

where $S_{xx} = \sum_{i=1}^n (X_i - \bar{X})^2$. Note that $S_{xx} = (n-1)\hat{\sigma}_{xx}^2$

Based on proposition 1, we see that $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimates for β_0 and β_1, assuming that assumption 1 holds. Also, we have *sampling distributions* for our estimators $\hat{\beta}_0$ and $\hat{\beta}_1$. Specifically, under assumption 1,

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}}\right)}} \sim N(0,1) \quad , \quad \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2\left(\frac{1}{S_{xx}}\right)}} \sim N(0,1)$$

However, unless σ^2, the variance of the errors, are known, the above distributions are not true and we must estimate σ^2.

A natural estimator for σ^2 is the variation around the actual Y_i and the fitted line, or[2]

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^n (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2 = \frac{1}{n-2}\sum_{i=1}^n r_i^2 \tag{8}$$

Just like all of our estimators for variance in previous lectures, we can derive that

$$(n-2)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$$

Furthermore, in the case where we derived the sampling distribution for the sample mean with unknown variance, we have

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{X}^2}{(n-1)\hat{\sigma}_x^2}\right)}} \sim t_{n-2} \quad , \quad \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma}^2\left(\frac{1}{(n-1)\hat{\sigma}_x^2}\right)}} \sim t_{n-2} \tag{9}$$

With the sampling distributions for $\hat{\beta}_0$ and $\hat{\beta}_1$, we can ask a number of inference questions, for instance CIs of $\hat{\beta}_0$ and $\hat{\beta}_1$ or hypotheses regarding β_0 and β_1.

[2] $\hat{\sigma}^2$ can also be written as $\hat{\sigma}^2 = MSE$ from our ANOVA table in table 1

Example 3.1. Suppose we believe that there is no linear relationship between X_i and Y_i. This hypothesis can be written as

$$H_0 : \beta_1 = 0 \text{ vs. } H_a : \beta_1 \neq 0$$

To test this hypothesis, we can compute the Type I Error from the sampling distributions we obtained in equation

$$\begin{aligned} P(\text{ reject } H_0|H_0 \text{ is true}) &= P\left(\left|\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma}^2\left(\frac{1}{(n-1)\hat{\sigma}_x^2}\right)}}\right| > \left|\frac{\hat{\beta}_{1,obs} - \beta_1}{\sqrt{\hat{\sigma}^2_{obs}\left(\frac{1}{(n-1)\hat{\sigma}_x^2}\right)}}\right| |H_0 \text{ is true}\right) \\ &= P\left(|t_{n-2}| > \left|\frac{\hat{\beta}_{1,obs} - \beta_1}{\sqrt{\hat{\sigma}^2_{obs}\left(\frac{1}{(n-1)\hat{\sigma}_x^2}\right)}}\right| |H_0 \text{ is true}\right) \end{aligned}$$

To obtain the p-value, we maximize the Type I Error. B because H_0 is true only at β_1, we can replace our β_1 in the expression above with $\beta_1 = 0$ and get

$$\text{P-value } = P\left(|t_{n-2}| > \left|\frac{\hat{\beta}_{1,obs}}{\sqrt{\hat{\sigma}^2_{obs}\left(\frac{1}{(n-1)\hat{\sigma}_x^2}\right)}}\right|\right)$$

Example 3.2. Suppose we want to construct a 95% confidence interval for the intercept term β_0. We have the expression for the middle 95% quantiles of the t_{n-2} distribution.

$$P(t_{n-2,\alpha/2} < t_{n-2} < t_{n-2,1-\alpha/2}) = P(t_{n-2,0.025} < t_{n-2} < t_{n-2,0.975}) = 0.95 = 1 - \alpha$$

From the sampling distributions we obtained in equation , we can derive the following results

$$\begin{aligned} 0.95 &= P(t_{n-2,(0.025)} < t_{n-2} < t_{n-2,(0.975)}) \\ &= P\left(t_{n-2,(0.025)} < \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{X}^2}{(n-1)\hat{\sigma}_x^2}\right)}} < t_{n-2,(0.975)}\right) \\ &= P\left(\hat{\beta}_0 - t_{n-2,(0.975)}\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{X}^2}{(n-1)\hat{\sigma}_x^2}\right)} < \beta_0 < \hat{\beta}_0 - t_{n-2,(0.025)}\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{X}^2}{(n-1)\hat{\sigma}_x^2}\right)}\right) \quad \text{(basic algebra here)} \end{aligned}$$

Using the fact that $t_{n-2,(1-\alpha/2)} = -1 * t_{n-2,(\alpha/2)}$, we obtain the 95% CI , which is $\hat{\beta}_0 \pm t_{n-2,(0.975)}\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{X}^2}{(n-1)\hat{\sigma}_x^2}\right)}$

3.3 ANOVA Table and Goodness of Fit Test

In addition to using explicitly formulas for sampling distributions laid out in proposition 1, we can also summarize the inference behind regression with an *ANOVA* table. ANOVA tables, in a nutshell, explains the variance of the linear model. Using ANOVA tables 1[3] will be much easier for doing inference with regression, especially multiple regression. A couple of remarks about the ANOVA table

1. MST is the estimate for the variance of Y. This is not the same as σ^2! The variance of Y refers to just looking at $Y_1, ..., Y_n$ and computing the variance of $(Y_1, ..., Y_n)$ without using X_is.

2. SST always has $n - 1$ degrees of freedom. This is from our lectures on sampling distribution for the sample variance and how we lost a degree of freedom because we have to estimate one quantity, $\bar{Y}$, the center of Y, in order to estimate the variation of Y around its center.

[3] I don't know why people use SSR instead of SSM for ANOVA tables

Sum of Squares (SS)	Mean Sum of Squares (MS)	Degrees of Freedom (DF)
Sum of Square Errors: $SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = r_i^2$	$MSE = \frac{SSE}{DFE}$	$DFE = n - 2$
Sum of Square Mode: $SSR = \sum_{i=1}^{n}(\bar{Y} - \hat{Y}_i)^2$	$MSR = \frac{SSR}{DFR}$	$DFR = 1$
Sum of Square Total: $SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$	$MST = \frac{SST}{DFT}$	$DFT = n - 1$

Table 1: ANOVA Table for Simple Linear Regression

3. The number of degrees of freedom for each SS requires some thought about the number of estimates are in the expression. For example, for SSE, we see that we are estimating two things, β_0 and β_1 and hence, we lose two degrees of freedom from the sample, $n - 2$. For SSR, just take my word that it is the number of parameters being estimated in addition to the intercept. In our case, it' is the number of slope terms, or 1.

4. For any ANOVA table,

$$SST = SSE + SSR \ , \ DFE + DFR = DFT$$

The proof for $SST = SSE + SSR$ is trivial and requires basic manipulation of quadratic expressions. On the contrary, for the equality expression for the degrees of freedom, it's easier to think about them in terms of what's being estimated in the square expression in the sum to justify the equality expression.

5. We can define the *coefficient of determination*, or R^2. Intuitively, R^2 measures the following how well a line with a slope explains the relationship between X and Y, with the "wellness" measured in comparison to a horizontal $\bar{Y}$ line. Mathematically,

$$R^2 = SSR/SST \tag{10}$$

Because $SSR \leq SST$, $0 \leq R^2 \leq 1$. An $R^2 = 0$ would indicate that the slope is useless (or that there is no relationship between X and Y. An $R^2 = 1$ would indicate that $SSE = 0$ or that there is a perfect relationship between X and Y. Interestingly enough,

$$R^2 = \rho_{x,y}^2$$

From the ANOVA table, we can perform inference just like we did in the previous section. Take a look at example 3.1. Here, we wanted to test whether a linear relationship exists between X and Y. Based on the ANOVA table, R^2 measures this relationship by comparing how well a line does in comparison to a horizontal line. Indeed, our test statistic can also be based off of R^2. Specifically, the test statistic F in

$$F = \frac{\frac{SSR}{DFR}}{\frac{SSE}{DFE}} \sim F_{DFR,DFE} \tag{11}$$

is known as the *Goodness of Fit Test* for regression. The p-value for this test matches that of example 3.1. In fact, the value for F matches that of the t in previous regression set ups.

$$F = \left(\frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2 \left(\frac{1}{(n-1)\hat{\sigma}_x^2} \right)}} \right)^2$$

3.4 Inference about prediction

Given a value X_i, you want to predict Y_i based on the estimated relationship you obtained from your sample. The natural thing to do is plug in X_i to the fitted line to obtain $\hat{Y}_i$, the predicted value for X_i. $\hat{Y}_i$ is actually the estimated *mean* value of the regression at the point X_i. That is, on average, the value of Y at X_i would be$E(Y|X_i)$ and an estimate of that average is the $\hat{Y}_i$ we obtain from the sample.

Since $\hat{Y}_i$ is an estimate of the mean of Y at X_i, there is always some uncertainty in this estimate. And where there is uncertainty, we can create confidence intervals! The formula for the $1 - \alpha$ confidence interval is

$$\hat{Y}_i \pm t_{n-2,(1-\alpha/2)}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(X_i - \bar{X})^2}{S_{xx}}} \tag{12}$$

The interpretation of this CI is the same as the interpretation of the CI for the population mean.

In addition to CIs, we also have *prediction intervals*, or PIs. $1-\alpha$ Prediction intervals are random intervals where future observations will fall with $1-\alpha$ probability. PIs have very straightforward and intuitive definitions, in comparison to CIs because PIs claim that there is $1-\alpha$ probability that a new observation will fall in the range given by the interval. In contrast, CIs claim that the probability of the *interval* covering the parameter is $1-\alpha$; for CIs, the probability is attached to the interval while the probability is attached to the object being covered, the $\hat{Y}_i$.

The formula for $1-\alpha$ prediction interval is

$$\hat{Y} \pm t_{n-2,(1-\alpha/2)}\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(X_i-\bar{X})^2}{S_{xx}}} \tag{13}$$

Notice that prediction intervals are larger than confidence intervals because in PIs, there is the uncertainty in not only the interval, but also the object being covered, the $\hat{Y}_i$.

4 Diagnonstics

In this section, we'll discuss techniques to verify assumptions of regression stated in 1 and remedy them whenever possible. It is generally advised in applied statistics to check the assumptions in the following order

1. Homoscedasticity (i.e. common variance σ^2
2. Linearity
3. Normal errors
4. Leverage points, influential points and outliers

Most of these assumptions, for better or worse, rely on looking at the *residual plot*, which is an x-y plot where x represents the fitted values, $\hat{Y}_i$ and the y represents the residuals of that fitted value, r_i, $(\hat{Y}_i, r_i)$.

4.1 Homoscedasticity

Homescedasticity is the condition that all the variances of ϵ_i must be identical for every X_i. A violation of this assumption means that some X_is have more variability in Y_i measurements than other X_is. To check for violation of this, we look for...

1. ...the scatterplot of (X_i, Y_i) with the fitted line. From the perspective of the fitted line, if the points scatter away in a $\prec$ as x increases, this implies that there is more variability in Y as X increases. To fix this, transform your Y_i by log(), $\sqrt{()}$, or $1/y$. From the perspective of the fitted line, if the points scatter away in a $\succ$ as x decreases, this implies that there is more variability in Y as X decreases. To fix this, transform your Y_i by y^2 or e^y
2. ...the residual plot. From the perspective of the x axis, if the points scatter away in a $\prec$ as x increases, this implies that there is more variability in Y as X increases. To fix this, transform your Y_i by log(), $\sqrt{()}$, or $1/y$. From the perspective of the x axis, if the points scatter away in a $\succ$ as x decreases, this implies that there is more variability in Y as X decreases. To fix this, transform your Y_i by y^2 or e^y

4.2 Linearity

Linearity is the condition that there must exist an underlying linear relationship between X and Y, of which we can estimate from the sample we collected. To check for this violation, we look for

1. the scatterplot of (X_i, Y_i) with the fitted line. If the fit doesn't "look right" and you see any nonlinear relationships, there is reason to believe linearity of X and Y is violated. To fix this, transform your x based on the rule given in lecture.

2. the residual plot. If the residual plot has any nonlinear relationships with respect to the x axis, there is reason to believe linearity of X and Y is violated. To fix this, transform your x based on the rule given in lecture.

4.3 Normality

Regression assumes that ϵ_i are Normally distributed. To check for this, we resort to the residuals, which gives us an idea of ϵ_i. In particular, we use a Normal QQ plot to check to see whether r_is are Normally distributed or not. If there is reason to believe that r_is are not normally distributed from the QQ plot, we transform Y_i based on what type of deviation it is. These transformations are identical to those in the QQ plot lecture. That is, if there is a right-skew, use $\log()$, $\sqrt{()}, 1/r$ transformations. If there is a left-skew, use r^2 or $\exp()$.

4.4 Leverage and Influential Points, Outliers

Regression has three type of outliers we have to watch out for.

1. *Regression Outliers (i.e. Outliers in Y):* These are outliers in the vertical direction (i.e. y-axis direction). To check for presence of outliers, use a residual plot. In particular, check for large deviations in the y-direction.

2. *Leverage point (i.e. Outliers in X):* These are outliers in the horizontal direction (i.e. x-axis direction). To check for presence of leverage points, use the (X, Y) scatterplot and see which points are "far away" in the x-axis. Also, leverage can be measured by the value known as H_{ii}, which is a diagonal of a matrix known as the *Hat matrix*. Leverage points are potential influential points.

3. *Influential points:* Influential points are leverage points where their removal would cause the fitted line to change drastically. These changes are reflected in either drastically different $\hat{\beta}_j$, estimated variances of $\hat{\beta}_j$, p-values associated with testing the hypotheses $H_0 : \beta_j = 0$ vs. $H_a : \beta_j \neq 0$, and estimates of $\hat{\sigma}^2$. *Cook's Distance* measures how influential each point. Each point has a Cook's distance and it can be computed using the formula

$$D_i = \frac{\sum_{j=1}^n (\hat{Y}_j - \hat{Y}_{j,-i})^2}{2MSE} = \frac{(Y_i - \hat{Y}_i)^2}{2MSE} \frac{H_{ii}}{(1 - H_{ii})^2} \tag{14}$$

   where $\hat{Y}_{j,-i}$ is the predicted value for the regression where the ith point is removed. Another way to interpret this is that it is the squared difference in prediction between the regression with all the points included and the regression with all but the ith point included. MSE is the MSE of the regression with all the points included. Generally speaking, $D_i > 1$ is considered an influential point.

When we discover outliers of these kind, the general advice is to remove them from the regression and refit the line with the outliers removed.