

Limits to Simple Linear Regression

Only one predictor is used.

The predictor variable must be quantitative (not categorical).

The relationship between response and predictor must be linear.

The errors must be normally distributed.

Multiple Linear Regression

For predictor variables x_1, x_2, \dots, x_k , use the linear model

$$y = b_0 + b_1x_1 + b_2x_2 + \dots + b_kx_k + e.$$

The b_i s are fixed parameters, e is the error term.

In matrix terms, $y = \mathbf{Xb} + e$

Of key interest is finding values for the b_i s. This is (usually) an over-constrained set of equations. See Jain's formula for a method of getting values.

Next questions are:

1. Is the mean significant?
2. Are the parameter values significant?

Multiple Regression Example (15.1)

Table 15.1 shows observations of CPU time, Disk I/O, and Memory Size.

The b values are found to be such that we get:

$$\text{CPU time} = -0.1614 + 0.1182(\# \text{ of disk I/O}) + 0.0265(\text{memory size})$$

We can compute the error terms from this (difference between the regression formula and the actual observations)

We can compute the coefficient of determination SSR/SST to find that the regression explains 97% of the variation of y .

The 90% confidence intervals for the parameters are $(-2.11, 1.79)$, $(-0.29, 0.53)$, and $(-0.06, 0.11)$, respectively.

- What does this mean?

Yet, the model seems to give good predictive ability. For example, what is the predicted CPU time for 100 disk I/Os and a memory size of 550? It is 26.23 with confidence interval $(19.55, 32.93)$ at 90% confidence. What does this mean?

Analysis of Variance (ANOVA)

ANOVA lets us determine if explanation of variance given by the model is “good” or not.

Specifically, we want to know if the following null hypothesis is correct:

- y does not depend upon any predictor x_j , that is, all of the b_j s are indistinguishable (with confidence) from zero.

To do this, we want to compare SSR (sum of squares explained by regression) to SSE (sum of squared errors) and see if the ratio is “good enough”.

An important part of calculating the ratio is the number of degrees of freedom for each term.

- SSR has k degrees of freedom.
- SSE has $n - k - 1$ degrees of freedom.

Thus, the actual ratio calculated is $(SSR/k)/(SSE/n - k - 1)$.

ANOVA (cont)

The other thing we need is to know what is “good enough” on the ratio.

- This depends on the amount of information we have.
- Use the F table appropriate for the desired confidence.
- Look at position $F[k, n - k - 1]$ in the table.

For the Disk I/O example, the computed ratio is 75.40, and the F table gives 4.32 as the minimum acceptable ratio.

Thus, we have very high confidence that the regression model has predictive ability.

Multicollinearity

Dilemma: None of our parameters are significant, yet the model is!!

The problem is that the correlation between the two predictors (memory and disk I/O) is correlated ($R = .9947$).

Next we test if the two parameters each give significant regression on their own.

- We already did this for the Disk I/O regression model, and found that it alone accounted for about 97% of the variance.
- We get the same result for memory size.

Conclusion: Each predictor alone gives as much predictive power as the two together!

Moral: Adding more predictors is not necessarily better in terms of predictive ability (aside from cost considerations).

ANOVA for Categorical Variables

A common problem is to determine if groups are different.

- Do plumbers make more than electricians?
- Is system A, B, or C better on a performance metric?

Now the question becomes: Are the between-group sums of squares (BSS) more or less important than the within-group variances sums of squares (WSS)?

Again, DOF is important.

- BSS has $k - 1$ DOF
- WSS has $n - k$ DOF

Calculate, and compare the ratio to the F table to determine if the differences are significant.

Curvilinear Regression

A model is a model. You can do anything you want, then measure the errors.

What is natural?

Do a scatterplot of response vs. predictor to see if its linear.

Often you can convert to a linear model and use the standard linear regression.

- Take the log when the curve looks like

$$y = bx^a$$

Example: Amdahls law says I/O rate is proportional to the processor speed.

- $I/O \text{ rate} = \alpha(\text{CPU rate})^{b_1}$
- Taking logs we get $\log(I/O \text{ rate}) = \log \alpha + b_1 \log(\text{MIPS rate})$.
- Using standard linear regression, we find that the regression explains 84% of the variatoin.

Outliers

“Any observation that is atypical of the remaining observations *may* be considered an outlier.”

The key question is whether the outlier represents a correct observation of system behavior.

An outlier might make a big change in the analysis, even to the extent of changing the conclusion (i.e., significance at a given confidence level).

To identify outliers, look at a scatterplot of the data.

If the outlier is not clearly an erroneous observation, then might want to do analysis with and without the outlier(s) and report both.

Common Mistakes with Regression

Verify that the relationship is linear. Look at a scatter diagram.

Don't worry about absolute value of parameters. They are totally dependent on an arbitrary decision regarding what dimensions to use.

Always specify confidence intervals for parameters and coefficient of determination.

Test for correlation between predictor variables, and eliminate redundancy. Test to see what subset of the possible predictors is "best" depending on cost vs. performance.

Don't make predictions too far out of the measured range.