COMMON ~~MISTEAKS~~ MISTAKES IN USING STATISTICS: Spotting and Avoiding Them

---

Common Mistakes in Interpretation of Regression Coefficients

1. **Interpreting a coefficient as a rate of change in Y instead of as a rate of change in the conditional mean of Y.**

2. Not taking confidence intervals for coefficients into account.

Even when a regression coefficient is (correctly) interpreted as a rate of change of a conditional mean (rather than a rate of change of the response variable), it is important to take into account the uncertainty in the estimation of the regression coefficient. To illustrate, in the example used in item 1 above, the computed regression line has equation $\hat{y} = 0.56 + 2.18x$. However, a 95% confidence interval for the slope is (1.80, 2.56). So saying, "The rate of change of the conditional mean of Y with respect to x is estimated to be between 1.80 and 2.56" is usually[1] preferable to saying, "The rate of change of the conditional mean Y with respect to x is about 2.18."

3. Interpreting a coefficient that is not statistically significant.[2]

Interpretations of results that are not statistically significant are made surprisingly often. If the t-test for a regression coefficient is not statistically significant, it is not appropriate to interpret the coefficient. A better alternative might be to say, "No statistically significant linear dependence of the mean of Y on x was detected.

4. Interpreting coefficients in multiple regression with the same language used for a slope in simple linear regression.

Even when there is an exact linear dependence of one variable on two others, the interpretation of coefficients is not as simple as for a slope with one dependent variable.

Example: If $y = 1 + 2x_1 + 3x_2$, it is *not accurate* to say "For each change of 1 unit in x_1, y changes 2 units". What *is* correct is to say, "If x_2 is fixed, then for each change of 1 unit in x_1, y changes 2 units."

Similarly, if the computed regression line is $\hat{y} = 1 + 2x_1 + 3x_2$, with confidence interval (1.5, 2.5), then a correct interpretation would be, "The estimated rate of change of the conditional mean of Y with respect to x_1, when x_2 is fixed, is between 1.5 and 2.5 units."

For more on interpreting coefficients in multiple regression, see Section 4.3 (pp 161-175) of Ryan[3].

5. Multiple inference on coefficients.

When interpreting more than one coefficient in a regression equation, it is important to use appropriate methods for multiple inference, rather than using just the individual confidence intervals that are automatically given by most software. One technique for multiple inference in regression is using *confidence regions*. [4]

Notes:
1. The decision needs to be made on the basis of what difference is practically important. For example, if the width of the confidence interval is less than the precision of measurement, there is no harm in neglecting the range. Another factor that is also important in deciding what level of accuracy to use is what level of accuracy your audience can handle; this, however, needs to be balanced with the possible consequences of not communicating the uncertainty in the results of the analysis.

In fact, this is just a special case of the more general problem not taking confidence intervals into account, as well stated by Good and Hardin:

> "Point estimates are seldom satisfactory in and of themselves. First, if the observations are continuous, the probability is zero that a point estimate will be correct and will equal the estimated

> parameter. Second, we still require some estimate of the precision of the point estimate."
> Philip I. Good and James W Hardin (2009), Common Errors In Statistics (and How to Avoid Them), 3rd ed, Wiley, p. 61.

2. This is really just a special case of the mistake in item 2. However, it is frequent enough to deserve explicit mention. If you'd like a reference, here's one from a very good introductory statistics textbook:

> "**If a coefficient's t-statistic is not significant, don't interpret it at all.** You can't be sure that the value of the corresponding parameter in the underlying regression model isn't really zero." (Boldface theirs)
> DeVeaux, Velleman, and Bock (2012), Stats: Data and Models, 3rd edition, Addison-Wesley
> p. 801 (in Chapter 10: Multiple Regression, under the heading "What Can Go Wrong?")

3. T. Ryan (2009), *Modern Regression Methods*, Wiley

4. Many texts on regression discuss confidence regions. See, for example, S. Weisberg (2005) *Applied Linear Regression*, Wiley, Section 5.5 (pp. 108 - 110), or R. D. Cook and S. Weisberg (1999), *Applied Regression Including Computing and Graphics*, Wiley, Section 10.8 (pp. 250 - 255).

Last updated March 7, 2014