Interpretation of regression coefficients[1]

STA 302 Fall 2020

[1] See last slide for copyright information.

Average response

The model says

$$E(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

- Can be viewed as a conditional expected value, given the values $x_1, \ldots, x_k$.
- Theoretically, there is a sub-population for each set of $x_1, \ldots, x_k$ values.
- $E(y|x_1, \ldots, x_k)$ is the sub-population mean (average response) for that sub-population.

$E(y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$

$$g(x_1, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

Examine $g(x_1, \ldots, x_k)$ as a mathematical function, to see what the regression coefficients mean.

Simple regression

$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

$$g(x) = \beta_0 + \beta_1 x$$

- The equation of a straight line.
- Say x is income and y is credit card debt.
- $\beta_1 > 0$ would mean that higher income tends to go with higher debt, on average.
- Call it a "positive (linear) relationship."
- $\beta_1 < 0$ would mean that higher income tends to go with lower debt, on average.
- Call it a "negative (linear) relationship."
- If the model is correct, $\beta_1 = 0$ would mean that there is no connection at all between income and average credit card debt.
- This is why testing $H_0 : \beta_1 = 0$ is so important.

$$t = \frac{\mathbf{a}'\widehat{\boldsymbol{\beta}} - t_0}{\sqrt{MSE\,\mathbf{a}'(X'X)^{-1}\mathbf{a}}} \stackrel{H_0}{\sim} t(n-k-1)$$

Estimated regression coefficients

$\widehat{E(y|x)} = \widehat{\beta}_0 + \widehat{\beta}_1 x$

- The same talk applies, with the addition of "estimated" or "predicted."
- *Estimated* average credit card debt is higher for consumers with higher incomes (if $\widehat{\beta}_1 > 0$).
- *Predicted* credit card debt is higher for consumers with higher incomes (if $\widehat{\beta}_1 > 0$).
- *Estimated* average credit card debt is lower for consumers with higher incomes (if $\widehat{\beta}_1 < 0$).
- *Predicted* credit card debt is lower for consumers with higher incomes (if $\widehat{\beta}_1 < 0$).
- Suppose annual income is in thousands of dollars. The question says: "When annual income is $1,000 higher, estimated average credit card debt is ________ higher. The answer is a number from your printout." Write the value of $\widehat{\beta}_1$.

Sometimes loose language is okay

- Technically, regression is about the connection between x and *expected*, or *average* y.
- But sometimes people (and my questions) speak just of the relationship between x and y.
- Like the relationship between High School GPA and University GPA.
- Yes, technically $g(x) = \beta_0 + \beta_1 x$ gives the relationship between High School GPA and *average* University GPA.
- But it's harmless – actually it's helpful. If necessary you can clarify.

Plain language is important

- If you can only be understood by mathematicians and statisticians, your knowledge is much less valuable.
- Often a question will say "Give the answer in plain, non-statistical language."
- This means if x is income and y is credit card debt, you make a statement about income and average or predicted credit card debt, like the ones on the preceding slides.
- If you use mathematical notation or words like null hypothesis, unbiased estimator, p-value or statistically significant, you will lose a lot of marks even if the statement is correct. Even avoid "positive relationship," and so on.
- If the study is about fish, talk about fish.
- If the study is about blood pressure, talk about blood pressure.
- If the study is about breaking strength of yarn, talk about breaking strength of yarn.
- Assume you are talking to your boss, who was a Commerce major and does not like to feel stupid.

We will be guided by hypothesis tests with $\alpha = 0.05$

For plain-language conclusions

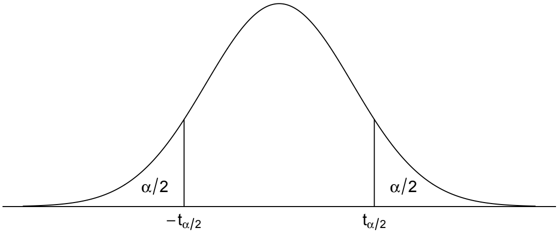
- If we do not reject a null hypothesis like $H_0 : \beta_1 = 0$, we will not draw a definite conclusion.
- Instead, say things like:
  - There is no evidence of a connection between blood sugar level and mood.
  - These results are not strong enough for us to conclude that attractiveness is related to mark in first-year Computer Science.
  - These results are consistent with no effect of dosage level on bone density.
- If the null hypothesis is not rejected, please do *not* claim that the drug has no effect, etc..
- In this we are taking Fisher's side in a historical fight between Fisher on one side and Neyman & Pearson on the other.
- Though we are guided by $\alpha = 0.05$, we *never* mention it when plain language is required.

A technical issue

- In this class we will avoid one-tailed tests.
- Why? Ask what would happen if the results were strong and in the opposite direction to what was predicted (dental example).
- But when H_0 is rejected, we still draw directional conclusions.
- For example, if x is income and y is credit card debt, we test $H_0 : \beta_1 = 0$ with a two-sided t-test.
- Say $p = 0.0021$ and $\widehat{\beta}_1 = 1.27$. We say "Consumers with higher incomes tend to have more credit card debt."
- Is this justified? We'd better hope so, or all we can say is "There is a connection between income and average credit card debt."
- Then they ask: "What's the connection? Do people with lower income have more debt?"
- And you have to say "Sorry, I don't know."
- It's a good way to get fired, or at least look silly.

The technical resolution

- Decompose the two-sided test into a set of two one-sided tests with significance level $\alpha/2$, equivalent to the two-sided test.



- In practice, just look at the sign of the regression coefficient.
- Under the surface you are decomposing the two-sided test, but you never mention it.
- *Marking rule*: If the question asks for plain language and you draw a non-directional conclusion when a directional conclusion is possible, you get half marks.

Multiple regression

$$g(x_1, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

- It's the equation of a hyper-plane, a k-dimensional surface in $k+1$ dimensions.
- Again, think of a sub-population at each combination of x values.
- $g(x_1, \ldots, x_k)$ is the average response at that set of values.

$g(x_1, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$

- Hold all the x values except x_j fixed.
- That is, do it in your mind. We are studying the function $g(\mathbf{x})$.

$$
\begin{aligned}
g(\mathbf{x}) &= \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \\
&= (\beta_0 + \sum_{i \neq j} \beta_i x_i) + \beta_j x_j \\
&= \alpha_0 + \beta_j x_j
\end{aligned}
$$

- Another straight line.
- The slope is unaffected by where you hold those other variables constant.
- The intercept is affected, but usually nobody cares.

How to talk about it

- With all other x values held constant as x_j varies, $E(y) = \alpha_0 + \beta_j x_j$.
- We talk about it as before, but say "controlling for" or "allowing for" or "taking into account" or "correcting for" the other variables.
- Controlling for parents' income, there is no evidence of a relationship between education and career success.
- Allowing for age, there is still a tendency for adults who exercise more to have lower blood pressure.
- These results are corrected for age, sex and severity of disease.
- Holding other variables constant, a student who studies one hour more per day is predicted to have a grade point average that is 0.47 higher.

Call it *model-based control*

- This is a big selling point for multiple regression of all kinds.
- To see what happens when variables are held constant at certain values, you don't literally have to hold them constant.
- Like "controlling for number of cigarettes smoked per day ..."

- It's valid provided that the model is approximately correct.
- It's risky outside the range of the data.

Correlation-causation

- In the model, the x values are literally producing y.
- For real data, this may be true, and it may not.
- A real (non-chance) connection between x and y does establish *why* the connection exists.
- People say "Correlation does not imply causation."
- By *correlation* they mean any kind of non-independence.

Examples

- Exercise and arthritis pain.
- The Mozart effect.
- Private music lessons, athletic training.
- Baldness and wearing a hat.
- Smoking and lung cancer.
- Vitamin B and spina bifida.

Solution?

- The best solution is random assignment,
- But this is not always possible.
- Be aware of the correlation-causation issue when making plain-language statements about the results of a statistical analysis.
- Watch out for going too far beyond what the data are actually telling you.

Copyright Information

This slide show was prepared by Jerry Brunner, Department of Statistical Sciences, University of Toronto. It is licensed under a Creative Commons Attribution - ShareAlike 3.0 Unported License. Use any part of it as you like and share the result freely. The LaTeX source code is available from the course website:

http://www.utstat.toronto.edu/~brunner/oldclass/302f20