I. Review of Basic Hypothesis Testing: Regression output example

A. An example of regression output and interpretation

**Table 1**

Y=% lecture portion = student activities

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Regression Statistics* |  |  |  |  |  |
| Multiple R | 0.082873 |  |  |  |  |  |
| R Square | 0.006868 |  |  |  |  |  |
| Adjusted R Square | 0.006389 |  |  |  |  |  |
| Standard Error | 22.35994 |  |  |  |  |  |
| Observations | 2076 |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |
| Regression | 1 | 7170.837 | 7170.837 | 14.34262 | 0.000157 |  |
| Residual | 2074 | 1036931 | 499.9668 |  |  |  |
| Total | 2075 | 1044102 |   |   |   |  |
|  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
| Intercept | 36.80395 | 1.645811 | 22.3622 | 2.1E-99 | 33.57634 | 40.03156 |
| #times taught | -2.42367 | 0.639968 | -3.78717 | 0.000157 | -3.67871 | -1.16862 |

**Table 2**

|  |  |  |  |
| --- | --- | --- | --- |
|  | % lecture portion = student activities | #times taught | year of highest degree |
| Mean | 30.85 | 2.45 | 1996.12 |
| St Dev | 22.43 | 0.77 | 11.67 |

B. Why a “t test?” Isn’t the t distribution appropriate for analyzing the mean of a normally distributed variable? Where’s the mean here?

1. Test of equality of means. Two means of normally distributed variables. The difference in means is also a mean of a (different, hybrid) normally distributed variable.

2. OLS coefs can be shown to be means of normally distributed variables if epsilon is distributed normally

3. Normality assumption? Not key. CLT says mean of variable of ANY distribution is normally distributed if n = infinity. Infinity practically kicks in around n=30 or so. In large samples, t equivalent to z. Should technically use z stat, but there’s no difference and bad habits die hard!

II. Model Selection

A. Part I: What variables to include? (Multivariate analysis where Y is continuous)

* t stats/F stats
* Theory (use example above), think of credible stories of confounding variables and then include controls.
	+ Ex. Effect of experience teaching a course on % time devoted to active learning. I see negative effect. But maybe age (or the environment in which they were educated) also matters. Maybe older generation more prone to stand-and-deliver. How would that affect my estimate? The older you are, the more times you have a chance to teach a course. So, the correlation between age and #times teaching a course creates a downward bias in my estimate. To correct for this, year of highest degree.

**Table 3**

Y=% lecture portion = student activities

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Regression Statistics* |  |  |  |  |  |
| Multiple R | 0.135999 |  |  |  |  |  |
| R Square | 0.018496 |  |  |  |  |  |
| Adjusted R Square | 0.017549 |  |  |  |  |  |
| Standard Error | 22.23402 |  |  |  |  |  |
| Observations | 2076 |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |
| Regression | 2 | 19311.36 | 9655.682 | 19.53202 | 3.95E-09 |  |
| Residual | 2073 | 1024791 | 494.3515 |  |  |  |
| Total | 2075 | 1044102 |   |   |   |  |
|  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
| Intercept | -408.46 | 89.86463 | -4.54528 | 5.8E-06 | -584.694 | -232.226 |
| #times taught | -1.23068 | 0.680376 | -1.80882 | 0.070624 | -2.56497 | 0.103613 |
| year of highest degree | 0.221598 | 0.044716 | 4.955653 | 7.79E-07 | 0.133905 | 0.309291 |

B. Model Selection Part II: What functional form?

* Examining data (plot)
* t stats (Ex. age vs. age and age2)
* Theory (Ex. Y=0/1 -> linear model a poor fit)

III. Multivariate analysis where Y is binomial

(Note: Below works for ordinal X & Y if you simplify to only 2 categories in Y)

**Table 4**

Y=% lecture portion = student activities <50% or >=50%

|  |  |  |  |
| --- | --- | --- | --- |
|  | **%<50** | **%>=50** | **Grand Total** |
| Not primarily AA | 1314 | 384 | 1698 |
| Primarily AA | 183 | 99 | 282 |
| **Grand Total** | **1497** | **483** | **1980** |

**Table 5**

|  |  |  |
| --- | --- | --- |
|  | **%<50** | **%>=50** |
| Not primarily AA | 77% | 23% |
| Primarily AA | 65% | 35% |

|  |  |
| --- | --- |
| Chi^2 | 20.4604 |
| p | <0.00001 |

Note: z=4.5233 = sqrt(20.4604). Same p. Same test, different presentation.Logit

**Table 6**

Y=% lecture portion = student activities <50% (0) or >=50% (1)

|  |  |  |  |
| --- | --- | --- | --- |
| Logistic regression | Number of obs | = | 1980 |
|  | LR chi2(1) | = | 19.13 |
|  | Prob > chi2 | = | 0 |
| Log likelihood = -1090.4824 | Pseudo R2 | = | 0.0087 |
|  |  |  |  |
|  | Coef. | Std. Err. | z | P>z | [95% Conf. | Interval] |
|  |  |  |  |  |  |  |
| Primarily AA | 0.6158224 | 0.13759 | 4.48 | 0 | 0.346153 | 0.885492 |
| \_cons | -1.230189 | 0.05801 | -21.21 | 0 | -1.34389 | -1.11649 |

Note: Implied p(heavy lecture) = .23, .77---just like above! z almost the same as above. Not quite same test, but we see agreement across the methods.

Logit adds a functional form assumption that isn’t needed in this univariate context...but it will come in handy when we move to multivariate.

Concern: Maybe being at a 2YC is correlated with experience in the field/when you got your degree...which also determines heavy student activities focus. So, we need to control for year of highest degree.

**Table 7**

|  |  |  |  |
| --- | --- | --- | --- |
| Logistic regression | Number of obs | = | 1980 |
|  | LR chi2(2) | = | 48.7 |
|  | Prob > chi2 | = | 0 |
| Log likelihood = -1075.6939 | Pseudo R2 | = | 0.0221 |
|  |  |  |  |  |  |  |
|  | Coef. | Std. Err. | z | P>z | [95% Conf. | Interval] |
|  |  |  |  |  |  |  |
| Primarily AA | 0.617758 | 0.138711 | 4.45 | 0 | 0.345889 | 0.889627 |
| Year highest degree | 0.025366 | 0.004763 | 5.33 | 0 | 0.016032 | 0.0347 |
| \_cons | -51.8861 | 9.515351 | -5.45 | 0 | -70.5359 | -33.2364 |

Yes, year of degree matters (newer -> more activities), but only marginally. And it doesn’t alter effect of 2YC institution.

How would you handle if X has 3 categories?

**Table 8**

|  |  |  |  |
| --- | --- | --- | --- |
| Logistic regression | Number of obs | = | 1978 |
|  | LR chi2(3) | = | 58.39 |
|  | Prob > chi2 | = | 0 |
| Log likelihood = -1069.1601 | Pseudo R2 | = | 0.0266 |
|  |  |  |  |  |  |  |
|  | Coef. | Std. Err. | z | P>z | [95% Conf. | Interval] |
|  |  |  |  |  |  |  |
| yr\_highest~e | 0.02518 | 0.004781 | 5.27 | 0 | 0.015809 | 0.034551 |
|  |  |  |  |  |  |  |
| basic2010 |  |  |  |  |  |  |
| Research Univ & Masters | -0.69198 | 0.141065 | -4.91 | 0 | -0.96846 | -0.4155 |
| Other | -0.18192 | 0.194866 | -0.93 | 0.351 | -0.56385 | 0.200008 |
|  |  |  |  |  |  |  |
| \_cons | -50.8963 | 9.550601 | -5.33 | 0 | -69.6151 | -32.1774 |
|  |  |  |  |  |  |  |

IV. When both X & Y are ordinal (and have >2 categories for Y)

A. Treat Y as continuous variable

1. Pros: easy

2. Cons: You’re making strong assumptions that might not be accurate (but if you find something...) and it is hard to interpret meaning

B. Collapse Y into binary ->logit

1. Pros: Easy to interpret a less-complicated Y variable, (relatively) easy to interpret logit

2. Cons: Lost richness of Y data

C. Ordered Logit

Y\*=a+bX is a score variable

Y=0 if Y\*<=mu1

Y=1 if mu1< Y\*<= mu2

Y=2 if mu2< Y\*<= mu3

...

Y=N if muN< Y\*

Assumption: proportional odds (not probabilities)

Ex. "poor", "fair", "good", "very good", and "excellent"

Then Prob(poor or worse) / Prob(NOT poor or worse) = Prob(fair or worse) / Prob(NOT fair or worse) = Prob(good or worse) / Prob(NOT good or worse) =....

D. Nonparametric

1. Spearman Rank Correlation (also rho, but NOT = Pearson’s Correlation Coefficient)

You change scores into ranks, and then compute Pearson’s correlation for the rank data. t or z test.

**Figure 1**



2. Kendal’s tau

Let (*x*1, *y*1), (*x*2, *y*2), …, (*xn*, *yn*) be a set of n observations

observations {\displaystyle (x\_{i},y\_{i})} (xi, yi) and (xj, yj) {\displaystyle (x\_{j},y\_{j})} are said to be *concordant* if the ranks for both agree: that is, if both {\displaystyle x\_{i}>x\_{j}}xi>xj and yi>yj{\displaystyle y\_{i}>y\_{j}}; or if both {\displaystyle x\_{i}<x\_{j}}xi<xj and {\displaystyle y\_{i}<y\_{j}}yi<yj

