# Class 4 – Regression Continue – 2017-09-01

## Going from Qualitative Data to Quantitative Data – Dummy Variables

Reading 3 – (p258-263) – Regression with Categorical Independent Variables

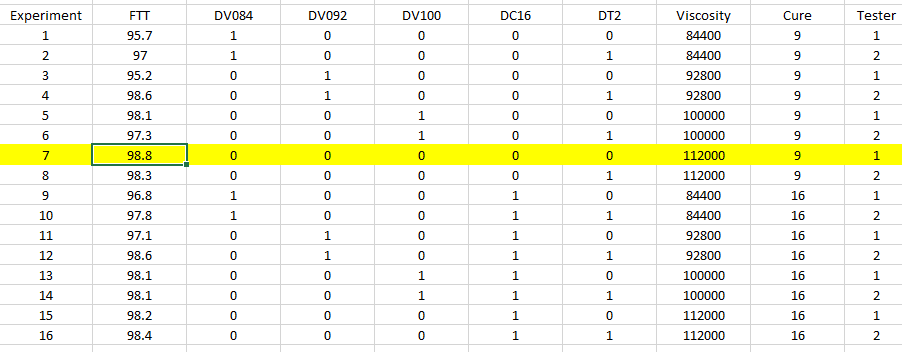
* **Dummy Variables:** Some data may be ordinal or normal – because regression requires numerical data, include these coding by coding the variables (ie Yes: 1 No: 0)
* **Steps for Dummy Variables**

1. **Look at the Data; Look at the Data; Look at the Data**
   * Determine if the data is non-linear (ie ordinal or nominal) and must be coded to enable regression analysis
     + *Ex Visteon Case indicated no linear fit for Viscosity in terms of values 🡪 go with LowV, MedV, MedHighV, and HighV*
2. **Identify the Base Case** – what you want to test **AGAINST** 
   * Option 1: The current situation –want to see if there’s something that can be improved
   * Option 2: The situation that appears to be the best (see Visteon)
     + *Ex Visteon Case: Viscosity = 112,000 (HighV); Cure = 9sec; Test Machine # 1*
3. **Determine # of DV:** The number of Dummy Variables per Independent Quantitative Variable is the number of options, k, less one: k-1. There can be multiple Independent Quantitative Variables (IQV), each with k-1 Dummy Variables within the model.
   * *Ex Visteon Case (see Excel)*

|  |  |  |  |
| --- | --- | --- | --- |
| ***IQV*** | ***# of Options*** | ***# Dummy Variables (k-1)*** | ***Variables*** |
| *Viscosity* | *4* | *3* | *DV084, DV092, DV100* |
| *Cure Time* | *2* | *1* | *DC16* |
| *Test Machine* | *2* | *1* | *DT2* |

* + *Ex Book: Tools A, B,C,D (see Excel)*
  + *Ex Book: Model Salary based on X1 Age and X2 MBA (Yes/No)*
    - *Model for Salary = Bo + B1Age + B2DVMBA (Yes=1/No=0) + e*
    - *Where B2 is a DV for MBA where Yes = 1*

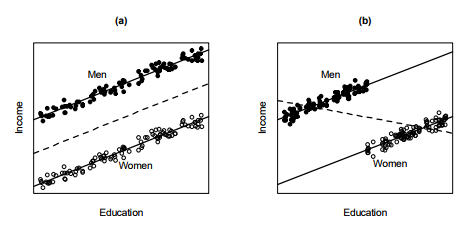
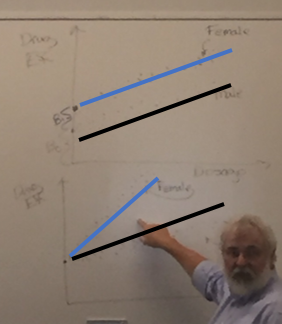
1. **Re-arrange Data with Dependent Variable and Dummy Variables & Run Regression**
   * *Ex Visteon Case:*



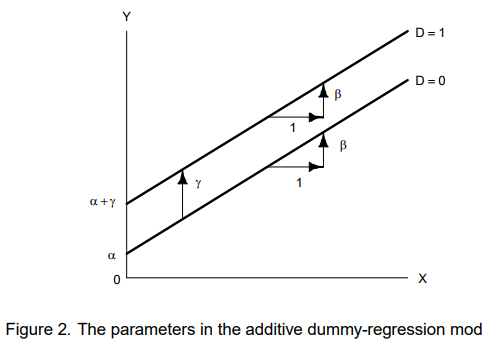
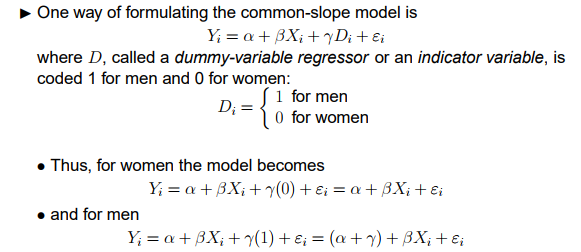
1. **Analyze Regression Model** – Make Observation on….
   * **R2**: How much % of the relationship does this model explain
   * **Standard Error** vs Mean
   * **ANOVA F Stat**: How much % Significance? Ie How much evidence is there to reject the Null Hypothesis and conclude there is a linear relationship with at least 1 of the B
   * **Intercept Value**: expected value of the BASE case
   * **Each variables’ (B) slope** (impact on the model) **and P-value** (evidence to reject the Null Hypothesis and conclude there is a relationship)
   * **Logical observations relative to the case**
     + *Ex Visteon: Use the 100 viscosity because it is cheaper, less risk, and no evidence there is an impact on the model*

**3 Ways to Use Dummy Variables**

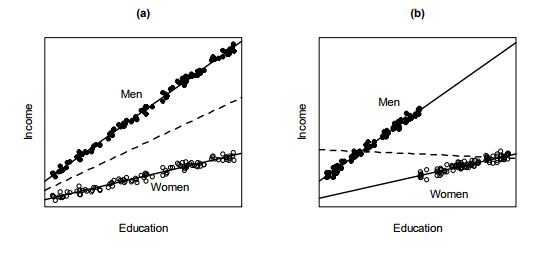
1. **To model specific points in a function that is not straight**
   * *Ex Visteon Case: Viscosities were not linear – use quantitative dummy variable*
2. **To shift the intercept for Independent Variables (IV) with categories**
   * Relationship are additive – partial effect of each dependent variable (DV) is same, regardless of specific value at which the other is held constant.
   * *Best fit lines from (a) and (b) do not truly model the data…*



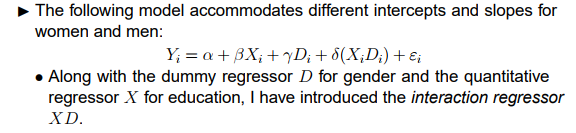
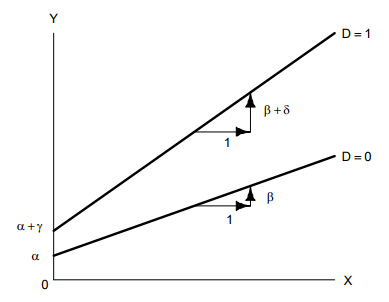
* + *So we introduce the dummy variable that when activated (ie Male=1) it will provide a provide the constant addition for* ***γ*** *in the regression model*

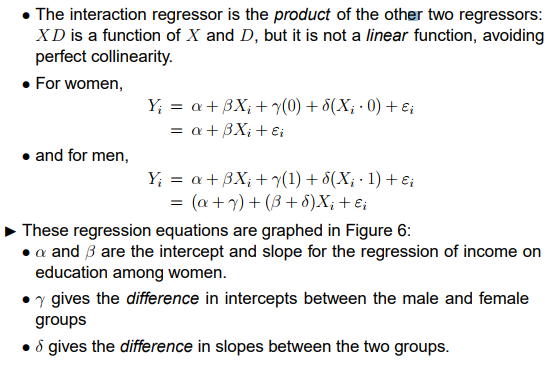


1. **To Model of Interaction**
   * Independent Variable (X1, X2,…) interact in determining a Dependent Variable (Y) when partial effect of one depends on the other
   * If the regression in different categories (ie Male and Female) is not parallel, then the qualitative IV interacts with one or more of the other IV. Modify regression model!!
     + *Example: With regards to income 🡪 the effect of education varies by gender 🡪 education and gender interact affecting income.*



* + - *Interaction goes both ways: effect of education varies by gender and the effect of gender varies by education*
  + Interaction refers to the way the IVs combine to affect the DV – not the relationship between the IVs
  + We could model separately; however a combined model (1) facilitates the test and (2) we can produce this model… so why not.
  + **We do this by introducing an additional IV which is the product of both: IV1\*IV2 = IV’**
  + Add this IV’ to your model 🡪 Run regression 🡪 Determine if IV’ is significant in new model
    - *See Example 8.16 on page 259 (MBA vs Age)*





## Testing

Reading 1 (P205-228) – Statistical Inference

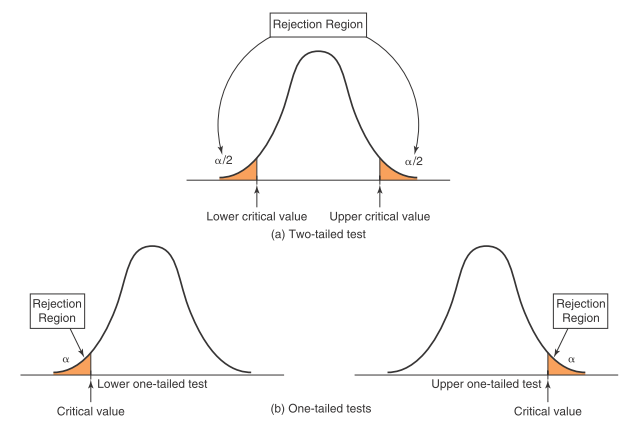
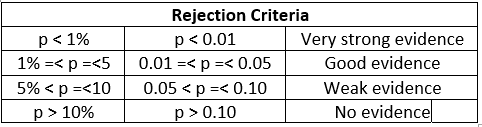
Reading 2 – (p244-247) – Regression Coefficients (See Running Notes Class 3 Regression)

**Introduction**

* Need to know if decisions will be effective.
  + Did advertising increase sales; will grocery store placement make a difference; did assembly method improve productivity etc.
* Statistical inference focuses on drawing conclusions about populations and samples. Estimation of population parameters and hypothesis testing 🡪 drawing conclusions about the value of the parameters of one or more population based on sample data. Done through hypothesis testing 🡪 a technique that allows you to draw valid statistical conclusions about value of population parameters or differences among them

**Hypothesis Testing**

* **Ho**: Null Hypothesis (Bo=B1=B2=O)
  + Initial assumption is variables don’t impact (ie no linear relationship with any variables) 🡪 need evidence to prove that is not true.
    - 1. Reject and conclude statistical evidence to support alternative
    - 2. Fail to reject null and conclude sample does not support alternative
      * If we fail to reject, we accept but cannot prove the null
* **Ha**: Alternative Hypothesis
* **Outcomes of Hypothesis Test**
  + 1. Ho is true & test fails to reject
    - Probability = **Confidence coefficient (1-α)**; controlled in advance
  + 2. Ho is false & test rejects
    - Probability = **Power of the Test (1-β)**
  + 3. Ho true & incorrectly rejects – Type I error
    - Probability = **Level of significance α**
  + 4. Ho is false & incorrectly fails to reject
    - Probability = **β -** cannot predict before; sensitive to sample size ؞ higher level of significance should have a larger sample size



* **T-distribution**: Tails are slightly larger than normal
* **P-value:** Probability of achieving the value you’re looking for based on null hypothesis B=0
* **T-stat**: # of standard deviations the critical value is away from the Null hypothesis
* **Story of the Hypothesis Test**: What’s the probability of achieving the value you’re looking for based on the hypothesis that B=0 .The smaller the P value, the farther away we are from the “null hypothesis”, the better evidence we have to reject the null assumption.
* **Economic Significance:** Even if there is a big difference but the cost impact is not there…. It’s not worth it! With enough data things will look different.
* **Bias vs Efficiency:** with more sample you take and the larger samples you take the more similar they will appear 🡪 they will begin to appear normal
  + Eg: population of dice roll
* **Central Limit Theorem:** 
  + If a truly random sample is independently drawn from any population, the sampling distribution of a sample mean is approximately normal (for significantly sized n)
  + Larger the sample, more closely sampling distribution of sample mean will resemble normal
* **Hendricks Skipped: 1 and 2 tail test; z stat with population mean known**
* **Items in Readings Not Included in these notes**
  + One Sample Hypothesis Test: Selecting Test Statistic
  + One and Two Tailed Tests: One-Sample Test for Proportions and Confidence Internals and Hypothesis Test
  + Two Sample Hypothesis Test: Two-Sample Tests for Differences in Means, with aired Samples, and Test for Equality of Variances
  + ANOVA: All
  + Chi-Square Test for Independence: All