Regression Modeling with Actuarial and Financial Applications

by Edward Frees

Ch. 4 notes

Categorical Variables

- *Categorical variables* provide labels for observations to denote membership in distinct groups, or categories.
- A *binary variable* is a special case of a categorical variable.
    - To illustrate, a binary variable may tell us whether or not someone has health insurance.
    - A categorical variable could tell us whether someone has (i) private individual health insurance, (ii) private group insurance, (iii) public insurance or (iv) no health insurance.
- For categorical variables, there may or may not be an ordering of the groups.
    - For health insurance, it is difficult to say which is "larger," private individual versus public health insurance (such as Medicare).
    - However, for education, we may group individuals from a dataset into "low," "intermediate" and "high" years of education.
- *Factor* is another term used for a (unordered) categorical explanatory variable.

Categorical Variables

- A categorical variable with c levels can be represented using c binary variables, one for each category.
    - For example, from a categorical education variable, we could code c=3 binary variables: (1) a variable to indicate low education, (2) one to indicate intermediate education and (3) one to indicate high education.
    - These binary variables are often known as *dummy variables*.
    - In regression analysis with an intercept term, we use only c-1 of these binary variables. The remaining variable enters implicitly through the intercept term.
- Through the use of binary variables, we do not make use of the ordering of categories within a factor.
    - Because no assumption is made regarding the ordering of the categories, for the model fit it does not matter which variable is dropped with regard to the fit of the model.
    - However, it does matter for the interpretation of the regression coefficients.

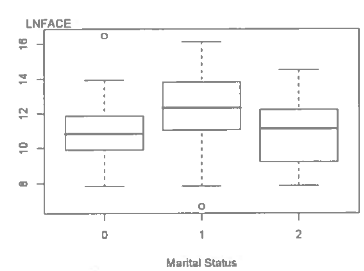
Example. Term Life Insurance

- We studied y = LNFACE, the amount that the company will pay in the event of the death of the named insured (in logarithmic dollars), focusing on the explanatory variables
    - annual income of the family (LNINCOME, in logarithmic dollars),
    - the number of years of EDUCATION of the survey respondent and
    - the number of household members, NUMHH.
- We now supplement this by including the categorical variable, MARSTAT, that is the marital status of the survey respondent. This may be:
    - 1, for married
    - 2, for living with partner
    - 0, for other (SCF actually breaks this category into separated, divorced, widowed, never married and inapplicable, for persons age 17 or less or no further persons)

Example. Term Life Insurance

Table: Summary Statistics of Logarithmic Face By Marital Status

	MARSTAT	Number	Mean	Standard deviation
Other	0	57	10.958	1.566
Married	1	208	12.329	1.822
Living together	2	10	10.825	2.001
Total		275	11.990	1.871



- For a binary variable with more than 2 categories, you only enter in c-1 of the binary variables and then the one that is omitted is automatically the baseline and the other variables' slopes(or betas) are changes from that baseline.
- The baseline is also known as the reference level. Most statistical software programs, if you include all variables, it will choose one for you automatically as a baseline. But, it may not be the one that you want or the one that makes the best sense.

Example. Term Life Insurance

- If we run a regression with the binary variables MAR0 and MAR2, then

$$\begin{aligned}\hat{y} &= 2.605 + 0.452\text{LNINCOME} + 0.205\text{EDUCATION} + 0.248\text{NUMHH} \\ &\quad -0.557\text{MAR0} - 0.789\text{MAR2}.\end{aligned}$$

  - If you are married, then $MAR0 = 0$, $MAR1 = 1$ and $MAR2 = 0$, and

$$\hat{y}_m = 2.605 + 0.452\text{LNINCOME} + 0.205\text{EDUCATION} + 0.248\text{NUMHH}.$$

  - If living together, then $MAR0 = 0$, $MAR1 = 0$ and $MAR2 = 1$, and

$$\hat{y}_{lt} = 2.605 + 0.452\text{LNINCOME} + 0.205\text{EDUCATION} + 0.248\text{NUMHH} - 0.789.$$

  - The difference in these two equations is 0.789.
- Interpret the regression coefficient associated with $MAR2$ to be the difference in fitted value for someone living together, compared to a similar person who is married (the omitted category).
- Similarly, interpret -0.557 to be the difference between the "other" category and the married category.
- $-0.557 - (-0.789) = 0.232$ is the difference between the other and the living together category.

Example. Term Life Insurance

- Note that $MAR0 + MAR1 + MAR2 = 1$ - there is a *perfect* linear dependency among the three.
- However, there is not a perfect dependency among any two of the three. It turns out that $Cor(MAR0, MAR1) = -0.90$, $Cor(MAR0, MAR2) = -0.10$ AND $Cor(MAR1, MAR2) = -0.34$.
- Any two out of the three produce the same model in terms of goodness of fit

Table: Term Life with Marital Status ANOVA Table

Source	Sum of Squares	df	Mean Square
Regression	343.28	5	68.66
Error	615.62	269	2.29
Total	948.90	274	

Residual standard error $s = 1.513$, $R^2 = 35.8\%$, $R_a^2 = 34.6\%$

Example. Term Life Insurance

Table: Term Life Regression Coefficients with Marital Status

Explanatory Variable	Model 1 Coefficient	Model 1 t-ratio	Model 2 Coefficient	Model 2 t-ratio	Model 3 Coefficient	Model 3 t-ratio
LNINCOME	0.452	5.74	0.452	5.74	0.452	5.74
EDUCATION	0.205	5.30	0.205	5.30	0.205	5.30
NUMHH	0.248	3.57	0.248	3.57	0.248	3.57
Intercept	3.395	3.77	2.605	2.74	2.838	3.34
MAR0	-0.557	-2.15	0.232	0.44		
MAR1			0.789	1.59	0.557	2.15
MAR2	-0.789	-1.59			-0.232	-0.44

- Model 1 *appears* the best in the sense that the t-ratios are larger (in absolute value). The p-values are close to statistically significant (0.113 for -1.59 and 0.032 for -2.15).
- Model 2 *appears* the worst in the sense that the t-ratios are smaller (in absolute value).
- Model 2 suggests that marital status is not statistically significant!!
- The three models are equivalent - same estimates, same fitted values, as long as you keep your interpretations straight.

Procedure for Testing the General Linear Hypothesis

- Run the full regression and get the error sum of squares and mean square error, which we label as $(Error\ SS)_{full}$ and s^2_{full}, respectively.
- Consider the model assuming the null hypothesis is true. Run a regression with this model and get the error sum of squares, which we label $(Error\ SS)_{reduced}$.
- Calculate

$$F-ratio = \frac{(Error\ SS)_{reduced} - (Error\ SS)_{full}}{p s^2_{full}}.$$

- Reject the null hypothesis in favor of the alternative if the F -ratio exceeds an F-value.
  - The F-value is a percentile from the F-distribution with $df_1 = p$ and $df_2 = n - (k+1)$ degrees of freedom.
  - Following our notation with the t-distribution, we denote this percentile as $F_{p,n-(k+1),1-\alpha}$, where α is the significance level.

Sets of Regression Coefficients

- Consider the joint effect of regression coefficients.
  - For example, is marital status (MARSTAT) important? This examines all of the binary variables at the same time.
- Introduce C, a generic matrix, where $C\beta$ will denote a linear combination of regression coefficients.
  - Test Ho: $C\beta=d$ (a known value, often 0).
  - For MARSTAT, I would test Ho:

Example. Term Life Insurance

- Our first (Chapter 3) regression

$$\mathrm{E}y = \beta_0 + \beta_1 LNINCOME + \beta_2 EDUCATION + \beta_3 NUMHH$$

yielded $s = 1.525$, $R^2 = 34.3\%$, $R_a^2 = 33.6\%$., $Error\ SS = 630.43$.
- A regression with the binary variables MAR0 and MAR2,

$$\mathrm{E}y = \beta_0 + \beta_1 LNINCOME + \beta_2 EDUCATION + \beta_3 NUMHH + \beta_4 MAR0 + \beta_5 MAR2$$

yielded $s = 1.513$, $R^2 = 35.8\%$, $R_a^2 = 34.6\%$, $Error\ SS = 615.62$.
- Comparing the two, we have

$$F\text{-}ratio = \frac{(Error\ SS)_{reduced} - (Error\ SS)_{full}}{p s_{full}^2} = \frac{630.43 - 615.62}{2 \times 1.513^2} = 3.235.$$

  - Degrees of freedom are $df_1 = p = 2$ and $df_2 = n - (k + p + 1) = 269$.
  - At $\alpha = 5\%$, the F-value is $F_{0.95,2,269} = 3.029$.
  - Thus, we reject H_0.
  - The p-value is $\Pr(F_{2,269} > 3.235) = 0.0409$.

One Factor ANOVA Model

- Recall that *factor* is another term used for a (unordered) categorical explanatory variable.
- Although factors may be represented as binary variables in a linear regression model, we study one factor models as a separate unit because
  - The method of least squares is much simpler, obviating the need to take inverses of high dimensional matrices
  - The resulting interpretations of coefficients are more straightforward
- The one factor model is still a special case of the linear regression model. Hence, no additional statistical theory is needed to establish its statistical inference capabilities.

An illustration using data on age and memory

- Consider the data from an experiment conducted regarding "memory".
- Eysenck, M.W. (1974), Age differences in incidental learning, *Developmental Psychology*, Vol 10, pp. 936-941.
- Response variable will be the number of words recalled by the subject (words).
- A theory regarding memory is that verbal material is remembered as a function of the degree to which it was processed when initially presented.
- 50 Younger subjects and 50 Older (between 55 and 65 yrs old) were randomly assigned to one of five learning groups: Counting, Rhyming, Adjective, Imagery, Intentional.
- None of the first 4 groups was told they will be later asked to recall words; after experiment, subjects were asked to list the words they could remember.

Description of the groups

- Counting group - asked to read a list of words and count the number of letters in each word.
- Rhyming group - asked to reach each word and think of word that rhymed with it.
- Adjective group - asked to give adjective to describe each word in a list.
- Imagery group - asked to form vivid images of each word.
- Intentional group - asked to memorize words for later recall.

The one-factor ANOVA model

- We decompose the response as

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \quad \text{for } j = 1, \ldots, J_i, \ i = 1, \ldots, I.$$

- The random errors follow usual assumptions:

$$\mathrm{E}(\varepsilon_{ij}) = 0 \quad \textit{and} \quad \mathrm{Var}(\varepsilon_{ij}) = \sigma^2.$$

- It is common to further decompose the parameter μ_i as

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}.$$

Possible restriction - drop a reference level

- Force one τ to be zero, say $\tau_1 = 0$.
- This corresponds to *dropping* one of the indicator variables - the level dropped is referred to as the baseline or reference level.
- The model equation becomes

$$Y_{ij} = \mu_1 + \tau_2 X_2 + \cdots + \tau_I X_I + \varepsilon_{ij},$$

  so that the intercept term $\mu = \mu_1$ becomes the mean of the level dropped and each regression coefficient becomes

$$\tau_i = \mu_i - \mu_1, \quad \textit{for } i = 2, 3, \ldots, I.$$

- Can drop any level, but the interpretation of the parameters depends on the variable dropped.

Confidence intervals/pairwise comparisons

- Mean of level i, μ_i, has $\overline{Y}_i$ for point estimate with the following corresponding confidence interval:

$$\overline{Y}_i \pm t_{\alpha/2,n-I}\frac{s}{\sqrt{J_i}} = \overline{Y}_i \pm 2\frac{s}{\sqrt{J_i}},$$

  where the degrees of freedom is $n - I$, with n, the total number of observations and I, the number of levels.

- A pairwise comparison of levels i and j can be made using a CI for $\tau_i - \tau_j$ with:

$$(\hat{\tau}_i - \hat{\tau}_j) \pm t_{\alpha/2,n-I}\frac{s}{\sqrt{1/J_i + 1/J_j}} = (\hat{\tau}_i - \hat{\tau}_j) \pm 2\frac{s}{\sqrt{1/J_i + 1/J_j}}.$$

- Testing for $\tau_i = \tau_j$ is equivalent to seeing whether zero lies in the CI or not.
- However, because for multiple pairwise comparisons, this result may be too conservative: either make a Bonferroni adjustment or do the Tukey's honest significant difference (HSD).

Tukey's honest significant difference test

- If you have a random sample $X_1, X_2, \ldots, X_n$ from $N(\mu, \sigma^2)$, then $R/\hat{\sigma}$ has a studentized range distribution, where $R = \max_i X_i - \min_i X_i$ is the range.
- The studentized range distribution $q_{n,\nu}$ where ν is the degrees of freedom used in estimating σ.
- The Tukey's confidence interval becomes

$$(\hat{\tau}_i - \hat{\tau}_j) \pm \frac{q_{I,n-I}}{\sqrt{2}} \cdot \frac{s}{\sqrt{1/J_i + 1/J_j}},$$

  where $q_{I,n-I}$ is the $(1 - \alpha)^{\text{th}}$ quantile of the studentized range distribution.