Outline

The Role of Binary Variables

Statistical Inference for Several Coefficients
- Sets of Regression Coefficients
- The General Linear Hypothesis
- Estimating and Predicting Several Coefficients

One Factor ANOVA Model
- One Factor ANOVA Model
- Matrix Expressions

Combining Categorical and Continuous Explanatory Variables

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.

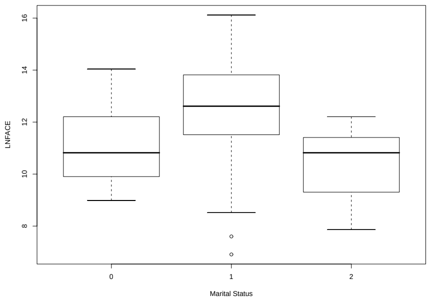
Example. Term Life Insurance

- 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	47	11.01	1.455
Married	1	155	12.50	1.794
Living together	2	10	10.51	1.314
Total		212	12.07	1.840



Example. Term Life Insurance

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

$$\begin{aligned}\widehat{y} &= 4.915 + 0.308 LNINCOME + 0.214 EDUCATION + 0.253 NUMHH \\ &\quad -0.518 MAR0 - 1.222 MAR2\end{aligned}$$

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

$$\widehat{y} = 4.915 + 0.308 LNINCOME + 0.214 EDUCATION + 0.253 NUMHH$$

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

$$\widehat{y} = 4.915 + 0.308 LNINCOME + 0.214 EDUCATION + 0.253 NUMHH - 1.222$$

  - The difference in these two equations is -1.222.
- 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 -0.518 to be the difference between the "other" category and the married category.
- $-0.518 - (-1.222) = 0.704$ 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 $Cor(MAR0, MAR1) = -0.88$, $Cor(MAR0, MAR2) = -0.12$ AND $Cor(MAR1, MAR2) = -0.37$.
- Any two out of the three produce the same model in terms of goodness of fit

Table: Term Life ANOVA Table

Source	Sum of Squares	df	Mean Square
Regression	237.78	5	47.56
Error	476.74	206	2.31
Total	714.52	211	

Residual standard error $s = 1.521$, $R^2 = 33.3\%$, $R_a^2 = 31.7\%$

Example. Term Life Insurance

Table: Term Life Regression Coefficients

	Model 1		Model 2		Model 3	
Explanatory Variable	Coefficient	*t*-ratio	Coefficient	*t*-ratio	Coefficient	*t*-ratio
LNINCOME	0.308	3.40	0.308	3.40	0.308	3.40
EDUCATION	0.214	4.61	0.214	4.61	0.214	4.61
NUMHH	0.253	2.97	0.253	2.97	0.253	2.97
Intercept	3.693	3.40	4.915	4.58	4.397	4.50
MAR0	0.703	1.31	-0.518	-1.71		
MAR1	1.222	2.41			0.518	1.71
MAR2			-1.222	-2.41	0.703	-1.31

- Model 2 *appears* the best in the sense that the *t*-ratios are larger (in absolute value). The *p*-values are close to statistically significant (0.088 for -1.71 and 0.017 for -2.41).
- Model 3 *appears* the worst in the sense that the *t*-ratios are smaller (in absolute value).
- In fact, Model 3 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.

Example. How does Cost-Sharing in Insurance Plans affect Expenditures in Healthcare?

- Rand Health Insurance Experiment (HIE) - Keeler and Rolph (1988)
- Cost-Sharing
  - 14 health insurance plans were grouped by the co-insurance rate (the percentage paid as out-of-pocket expenditures that varied by 0, 25, 50 and 95%).
  - One of the 95% plans limited annual out-of-pocket outpatient expenditures to $150 per person ($450 per family), providing in effect an individual outpatient deductible. This plan was analyzed as a separate group.
  - There were $c = 5$ categories of insurance plans.
- Adverse Selection
  - Individuals choose insurance plans making it difficult to assess cost-sharing effects.
  - Adverse selection can arise because individuals in poor chronic health are more likely to choose plans with less cost sharing, thus giving the appearance that less coverage leads to greater expenditures.
  - In the Rand HIE, individuals were randomly assigned to plans, thus removing this potential source of bias.

Example. Cost-Sharing in Insurance Plans

- Although families were randomly assigned to plans, Keeler and Rolph (1988) used regression methods to control for participant attributes and isolate the effects of plan cost-sharing.
  - Based on a sample of $n = 1,967$ episode expenditures.
  - Logarithmic expenditure was the dependent variable.
- Cost-sharing was decomposed into 5 binary variables.
  - These variables are "Co-ins25," "Co-ins50," and "Co-ins95," for coinsurance rates 25, 50 and 95%, respectively, and "Indiv Deductible" for the plan with individual deductibles.
  - The omitted variable is the free insurance plan with 0% coinsurance.
- The HIE was conducted in six cities; location was decomposed into 5 binary variables: Dayton, Fitchburg, Franklin, Charleston and Georgetown, with Seattle being the omitted variable.
- Age and sex was decomposed into 5 binary variables: "Age 0-2," "Age 3-5," "Age 6-17," "Woman age 18-65," and "Man age 46-65," the omitted category was "Man age 18-45."
- Other control variables included a health status scale, socioeconomic status, number of medical visits in the year prior to the experiment on a logarithmic scale and race.

Example. Cost-Sharing in Insurance Plans

Table: Coefficients of Episode Expenditures from the Rand HIE

Variable	Regression Coefficient	Variable	Regression Coefficient
Intercept	7.95		
Dayton	0.13*	Co-ins25	0.07
Fitchburg	0.12	Co-ins50	0.02
Franklin	-0.01	Co-ins95	-0.13*
Charleston	0.20*	Indiv Deductible	-0.03
Georgetown	-0.18*		
Health scale	-0.02*	Age 0-2	-0.63**
Socioeconomic status	0.03	Age 3-5	-0.64**
Medical visits	-0.03	Age 6-17	-0.30**
Examination	-0.10*	Woman age 18-65	0.11
Black	0.14*	Man age 46-65	0.26

Note: * significant at 5%, ** significant at 1%
Source: Keeler and Rolph (1988)

Example. Cost-Sharing in Insurance Plans. Findings

- As noted by Keeler and Rolph, there were large differences by site and age although the regression only served to summarize $R^2 = 11\%$ of the variability.
- For the cost-sharing variables, only "Co-ins95" was statistically significant, and this only at the 5% level, not the 1% level.
- The paper of Keeler and Rolph (1988) examines other types of episode expenditures, as well as the frequency of expenditures.
- They conclude that cost-sharing of health insurance plans has little effect on the amount of expenditures per episode although there are important differences in the frequency of episodes.
  - This is because an episode of treatment is composed of two decisions.
  - The amount of treatment is made jointly between the patient and the physician and is largely unaffected by the type of health insurance plan.
  - The decision to seek health care treatment is made by the patient; this decision-making process is more susceptible to economic incentives in cost-sharing aspects of health insurance plans.

Sets of Regression Coefficients

- Wish to consider the *joint* effect of regression coefficients.
  - For example, in the Rand HIE, is “location” important? This means examining all of the binary city variables at the same time.
- Recall the regression coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)'$, a $(k+1) \times 1$ vector.
- Introduce $\mathbf{C}$, a generic $p \times (k+1)$ matrix that is user-specified (and depends on the application)
- Thus, $\mathbf{C}\boldsymbol{\beta}$, will denote several linear combinations of regression coefficients
- Some applications involve testing whether $\mathbf{C}\boldsymbol{\beta}$ equals a specific known value (denoted as $\mathbf{d}$).
- We call $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{d}$ the *general linear hypothesis*.

Sets of Regression Coefficients. Special Cases

- A regression coefficient, say β_j. Choose $p = 1$ and $\mathbf{C}$ to be a $1 \times (k+1)$ vector with a one in the $(j+1)$*st* column and zeros otherwise. These choices result in

$$\mathbf{C}\boldsymbol{\beta} = (0\ \ldots\ 0\ 1\ 0\ \ldots\ 0) \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix} = \beta_j.$$

- A linear combination of regression coefficients. Choose $p = 1$ and $\mathbf{C} = \mathbf{c}' = (c_0, \ldots, c_k)'$, to get

$$\mathbf{C}\boldsymbol{\beta} = \mathbf{c}'\boldsymbol{\beta} = c_0\beta_0 + \ldots + c_k\beta_k.$$

  For example, if $c_0 = 1, c_1 = x_1, \ldots, c_k = x_k$, then $\mathbf{c}'\boldsymbol{\beta} = c_0\beta_0 + \ldots + c_k\beta_k$ = E y, the regression function.

Sets of Coefficients. Hypothesis Testing

- Testing equality of regression coefficients, say $H_0 : \beta_1 = \beta_2$.
  - For this purpose, choose $p = 1$, $\mathbf{c}' = (0, 1, -1, 0, \ldots, 0)$, and $\mathbf{d}$=0.

$$\mathbf{C}\boldsymbol{\beta} = \mathbf{c}'\boldsymbol{\beta} = (0, 1, -1, 0, \ldots, 0) \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix} = \beta_1 - \beta_2 = 0,$$

  so that $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{d}$ becomes $H_0 : \beta_1 = \beta_2$.
- Testing portions of the model. With a *full* regression function

$$\mathrm{E}\ y = \beta_0 + \beta_1 x_1 \ldots + \beta_k x_k + \beta_{k+1} x_{k+1} + \ldots + \beta_{k+p} x_{k+p}$$

  Test the null hypothesis $H_0 : \beta_{k+1} = \ldots = \beta_{k+p} = 0$
  - Choose $\mathbf{d}$ and $\mathbf{C}$ such that

$$\mathbf{C}\boldsymbol{\beta} = \begin{pmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \\ \beta_{k+1} \\ \vdots \\ \beta_{k+p} \end{pmatrix} = \begin{pmatrix} \beta_{k+1} \\ \vdots \\ \beta_{k+p} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} = \mathbf{d}.$$

The General Linear Hypothesis

- Wish to test $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{d}$.
- Do this by checking whether $\mathbf{Cb} - \mathbf{d}$ is close to zero
  - The expected value of $\mathbf{Cb} - \mathbf{d}$ is $\mathbf{C}\boldsymbol{\beta} - \mathbf{d}$.
  - The variance of $\mathbf{Cb} - \mathbf{d}$ is $\sigma^2 \mathbf{C}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{C}'$.
  - We use

$$F - ratio = \frac{(\mathbf{Cb} - \mathbf{d})' \left(\mathbf{C}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{C}' \right)^{-1} (\mathbf{Cb} - \mathbf{d})}{p s_{full}^2}.$$

  - Here, s_{full}^2 is the mean square error from the full regression model.
  - Under the null hypothesis, this statistic follows an F-distribution with numerator degrees of freedom $df_1 = p$ and denominator degrees of freedom $df_2 = n - (k+1)$

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

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

- 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.

Example. Term Life Insurance

- Our first (Chapter 3) regression

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

yielded $s = 1.541$, $R^2 = 30.9\%$, $R^2_a = 29.9\%$, $Error\ SS = 493.84$.

- A regression with the binary variables MAR0 and MAR2,

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

yielded $s = 1.521$, $R^2 = 33.3\%$, $R^2_a = 31.7\%$, $Error\ SS = 476.74$

- Comparing the two, we have

$$\begin{aligned} F-ratio &= \frac{(Error\ SS)_{reduced} - (Error\ SS)_{full}}{ps^2_{full}} \\ &= \frac{493.84 - 476.74}{2 \times 1.521^2} = 3.696. \end{aligned}$$

- The degrees of freedom are $df_1 = 2$ and $df_2 = 206$.
- At $\alpha = 5\%$, the F-value is $F_{2,206,0.95} = 3.039723$.
- Thus, we reject H_0.
- The p-value is $\Pr(F_{2,206} > 3.696) = 0.0265$.

Extra Sum of Squares

- Adding variables to a regression function reduces (never increases) the error sum of squares
  - This is because we are minimizing over additional parameters

$$\begin{aligned} (Error\ SS)_{full} &= min_{b_0^*,\ldots,b_{k+p}^*} SS(b_0^*,\ldots,b_{k+p}^*) \\ &= min_{b_0^*,\ldots,b_{k+p}^*} \sum_{i=1}^{n} \left(y_i - (b_0^* + \ldots + b_{k+p}^* x_{i,k+p})\right)^2 \\ &\leq min_{b_0^*,\ldots,b_k^*} \sum_{i=1}^{n} \left(y_i - (b_0^* + \ldots + b_k^* x_{i,k})\right)^2 \\ &= (Error\ SS)_{reduced}. \end{aligned}$$

  - Equivalently, the amount of variability explained increases (never decreases) because

$$(Error\ SS)_{reduced} - (Error\ SS)_{full} = (Regression\ SS)_{full} - (Regression\ SS)_{reduced}.$$

  - This difference is known as a *Type III Sum of Squares.*

F-ratio

- We can also write

$$\begin{aligned} F-ratio &= \frac{(Error\ SS)_{reduced} - (Error\ SS)_{full}}{ps^2_{full}} \\ &= \frac{(Regression\ SS)_{full} - (Regression\ SS)_{reduced}}{ps^2_{full}}. \end{aligned}$$

- Dividing the numerator and denominator by *Total SS* yields

$$F-ratio = \frac{\left(R^2_{full} - R^2_{reduced}\right)/p}{\left(1 - R^2_{full}\right)/(n-(k+1))}.$$

The F-ratio measures the statistical significance of the drop in the coefficient of determination, R^2.

- In the special case that $p = 1$, one can show that $(t-ratio)^2 = F-ratio$. That is, they are equivalent statistics.

Estimating Linear Combinations of Regression Coefficients

- Might wish to estimate $\beta_1 + \beta_2$. (Charitable contributions example, represented the expected giving rate per unit of income, for income in excess of the Social Security taxable wage base).
- More generally, estimate $\mathbf{c}'\boldsymbol{\beta} = c_0\beta_0 + \ldots + c_k\beta_k$.
- Use as a point estimate $\mathbf{c}'\mathbf{b} = c_0 b_0 + \ldots + c_k b_k$.
  - This is unbiased and has variance $\mathrm{Var}\,(\mathbf{c}'\mathbf{b}) = \sigma^2 \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}$.
  - Thus, the standard error is

$$se\,(\mathbf{c}'\mathbf{b}) = s\sqrt{\mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}}.$$

  - With this quantity, a $100(1-\alpha)\%$ confidence interval for $\mathbf{c}'\boldsymbol{\beta}$ is

$$\mathbf{c}'\mathbf{b} \pm t_{n-(k+1),1-\alpha/2}\; se(\mathbf{c}'\mathbf{b}).$$

Prediction Intervals

- Assume that the set of explanatory variables $\mathbf{x}^*$ is known and wish to predict the corresponding response y^*.
- This new response follows the same assumptions as the observed data.
- Specifically, the expected response is $\mathrm{E}\; y^* = \mathbf{x}^{*\prime}\boldsymbol{\beta}$, $\mathbf{x}^*$ is nonstochastic, $\mathrm{Var}\; y^* = \sigma^2$, y^* is independent of $\{y_1, ..., y_n\}$ and is normally distributed.
- Under these assumptions, a 100(1-α)% prediction interval for y^*

$$\mathbf{x}^{*\prime}\mathbf{b} \pm t_{n-(k+1),1-\alpha/2}\; s\sqrt{1 + \mathbf{x}^{*\prime}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}^*}.$$

This generalizes the prediction interval for introduced in Section 2.4.

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 special statistical theory is needed to establish its statistical inference capabilities.

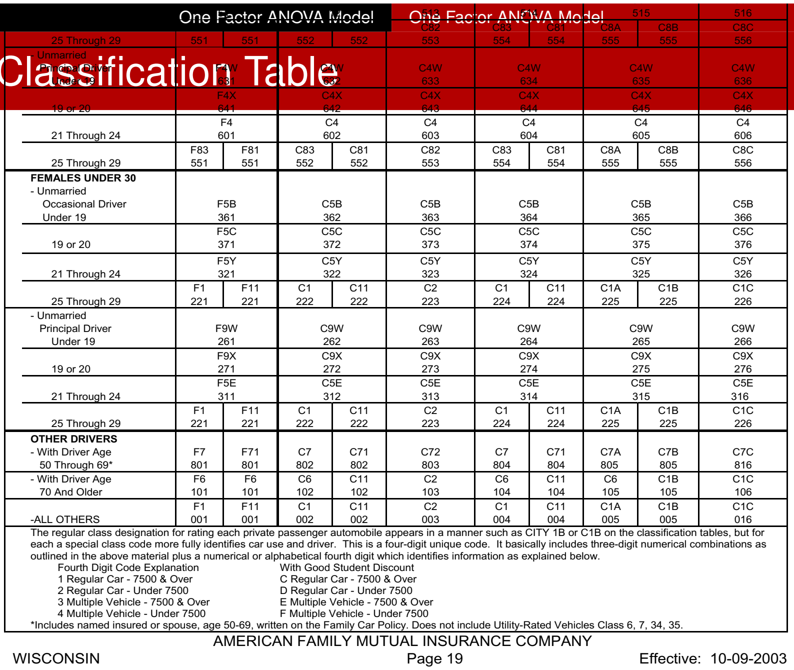
Example. Automobile Claims

- We examine claims experience from a large midwestern (US) property and casualty insurer for private passenger automobile experience.
- The dependent variable is the amount paid on a closed claim, in (US) dollars (claims that were not closed by year end are handled separately).
- Insurers categorize policyholders according to a risk classification system.
- The risk classification system is based on
  - Automobile operator characteristics (age, gender, marital status and whether the primary or occasional driver of a car)
  - Vehicle characteristics (city versus farm usage, if the vehicle is used to commute to school or work, used for business or pleasure, and if commuting, the approximate distance of the commute)
- These factors are summarized by the risk class categorical variable.
- Also available is the state in which the claims was filed (another categorical variable)

Example. Automobile Claims

- Insurers categorize policyholders according to a risk classification system.
  - The idea behind this is to create groups of policyholders with similar risk characteristics that will have similar claims experience.
  - These groups form the basis of the pricing of insurance, so that each policyholder is charged an amount that is appropriate to their risk category.
- We will examine only claims that are filed and paid by the company.
  - Not based on how frequently policyholders have claims.
  - For rating purposes, the frequency is often more important than the severity (amount)
  - Many insurers decompose their analyses into frequency and severity components.
  - In our analysis, we will not find many explanatory variables that are important in explaining claims amounts.
- However, the one factor ANOVA models can also be used to analyze claims experience at the policyholder level
  - Specifically, define y_i to be the amount paid for the ith policyholder
  - There will be many zeros here, for policyholders without a claim filed during the year.
  - This is known as the “pure premium” approach to rating.

Risk Classification Table



					513	514		515		516
25 Through 29	551	551	552	552	C82 553	C83 554	C81 554	C8A 555	C8B 555	C8C 556
- Unmarried Principal Driver Under 19	F4W 631		C4W 632		C4W 633	C4W 634		C4W 635		C4W 636
19 or 20	F4X 641		C4X 642		C4X 643	C4X 644		C4X 645		C4X 646
21 Through 24	F4 601		C4 602		C4 603	C4 604		C4 605		C4 606
25 Through 29	F83 551	F81 551	C83 552	C81 552	C82 553	C83 554	C81 554	C8A 555	C8B 555	C8C 556
FEMALES UNDER 30 - Unmarried Occasional Driver Under 19	F5B 361		C5B 362		C5B 363	C5B 364		C5B 365		C5B 366
19 or 20	F5C 371		C5C 372		C5C 373	C5C 374		C5C 375		C5C 376
21 Through 24	F5Y 321		C5Y 322		C5Y 323	C5Y 324		C5Y 325		C5Y 326
25 Through 29	F1 221	F11 221	C1 222	C11 222	C2 223	C1 224	C11 224	C1A 225	C1B 225	C1C 226
- Unmarried Principal Driver Under 19	F9W 261		C9W 262		C9W 263	C9W 264		C9W 265		C9W 266
19 or 20	F9X 271		C9X 272		C9X 273	C9X 274		C9X 275		C9X 276
21 Through 24	F5E 311		C5E 312		C5E 313	C5E 314		C5E 315		C5E 316
25 Through 29	F1 221	F11 221	C1 222	C11 222	C2 223	C1 224	C11 224	C1A 225	C1B 225	C1C 226
OTHER DRIVERS - With Driver Age 50 Through 69*	F7 801	F71 801	C7 802	C71 802	C72 803	C7 804	C71 804	C7A 805	C7B 805	C7C 816
- With Driver Age 70 And Older	F6 101	F6 101	C6 102	C11 102	C2 103	C6 104	C11 104	C6 105	C1B 105	C1C 106
-ALL OTHERS	F1 001	F11 001	C1 002	C11 002	C2 003	C1 004	C11 004	C1A 005	C1B 005	C1C 016

The regular class designation for rating each private passenger automobile appears in a manner such as CITY 1B or C1B on the classification tables, but for each a special class code more fully identifies car use and driver. This is a four-digit unique code. It basically includes three-digit numerical combinations as outlined in the above material plus a numerical or alphabetical fourth digit which identifies information as explained below.

Fourth Digit Code Explanation	With Good Student Discount
1 Regular Car - 7500 & Over	C Regular Car - 7500 & Over
2 Regular Car - Under 7500	D Regular Car - Under 7500
3 Multiple Vehicle - 7500 & Over	E Multiple Vehicle - 7500 & Over
4 Multiple Vehicle - Under 7500	F Multiple Vehicle - Under 7500

*Includes named insured or spouse, age 50-69, written on the Family Car Policy. Does not include Utility-Rated Vehicles Class 6, 7, 34, 35.

AMERICAN FAMILY MUTUAL INSURANCE COMPANY

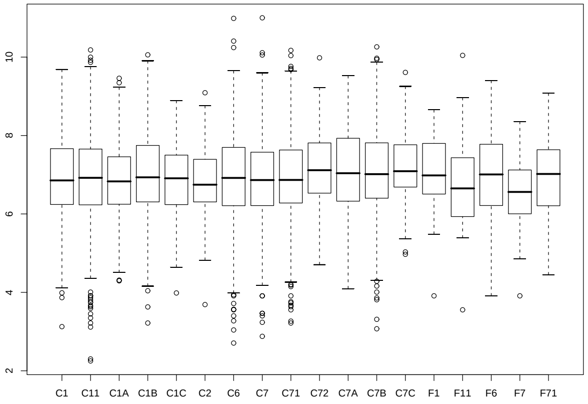
WISCONSIN Page 19 Effective: 10-09-2003

Example. Automobile Claims

- We consider $n = 6,773$ claims from drivers aged 50 and above
- The distribution of claims is very skewed, so we consider y = logarithmic claims.
- These are from 14 different states
- There are 18 risk classes in the table below. This table shows little difference in claim amount by risk class.

Risk Class	C1	C11	C1A	C1B	C1C	C2	C6	C7	C71
Number	726	1151	77	424	38	61	911	913	1129
Average	6.941	6.952	6.866	6.998	6.786	6.801	6.926	6.901	6.954
Standard Deviation	1.064	1.074	1.072	1.068	1.110	0.948	1.115	1.058	1.038
Risk Class	**C72**	**C7A**	**C7B**	**C7C**	**F1**	**F11**	**F6**	**F7**	**F71**
Number	85	113	686	81	29	40	157	59	93
Average	7.183	7.064	7.072	7.244	7.004	6.804	6.910	6.577	6.935
Standard Deviation	0.988	1.021	1.103	0.944	0.996	1.212	1.193	0.897	0.983

Box Plots of Logarithmic Claims by Risk Class



One Factor ANOVA Model

- Assume that there are c levels of the factor.
- At the jth level, there are n_j observations. In total, there are $n = n_1 + n_2 + \ldots + n_c$ observations.
- The data are:

Data for Level 1	y_{11}	y_{21}	$\cdots$	$y_{n_1,1}$
Data for Level 2	y_{12}	y_{22}	$\cdots$	$y_{n_2,1}$
.	.	.	...	.
Data for Level c	y_{1c}	y_{2c}	$\cdots$	$y_{n_c,c}$

- At the jth level, the sample average is

$$\bar{y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij}.$$

- The one factor ANOVA model is

$$y_{ij} = \mu_j + \varepsilon_{ij} \qquad i = 1, \ldots, n_j, \qquad j = 1, \ldots, c.$$

- With this, use mean μ_j to be the mean at the jth level.
- Let $\mathrm{Var}\, y_{ij} = \sigma^2$ be the variance term.

Method of Least Squares

- Estimate the parameters $\mu_j,\ j = 1, \ldots, c$ using least squares.
  - For candidate estimates μ_j^* of μ_j, define the sum of squares

$$SS(\mu_1^*, \ldots, \mu_c^*) = \sum_{j=1}^{c}\sum_{i=1}^{n_j}(y_{ij} - \mu_j^*)^2$$

  - Take a derivative with respect to each μ_j^*, set = 0 and solve. For example,

$$\frac{\partial}{\partial \mu_1^*} SS(\hat{\mu}_1, \ldots, \hat{\mu}_c) = \frac{\partial}{\partial \mu_1^*}\sum_{i=1}^{n_1}(y_{i,1} - \mu_1^*)^2 = \sum_{i=1}^{n_1}(-2)(y_{i,1} - \mu_1^*) = 0.$$

  - Thus, $\bar{y}_1$ is the *least squares estimate* of μ_1.
  - In general, $\bar{y}_j$ is the *least squares estimate* of μ_j.
- This yields least squares estimates *without* matrix computations
  - Note that here, the dimension of $\mathbf{X}'\mathbf{X}$ is $c \times c$; this is 18×18 for our auto claims data.
  - For large auto insurers, common to have a risk classification system where c is in the thousands, making the usual estimation methods tedious.

ANOVA Table for One Factor ANOVA Model

- Begin with the *error sum of squares*

$$\text{Error SS} = SS(\bar{y}_1, \ldots, \bar{y}_c) = \sum_{j=1}^{c}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2$$

- Define the "Factor" (Regression) Sum of Squares
  Factor SS = Total SS – Error SS
- To get the variability decomposition is summarized in the following analysis of variance (ANOVA) table.

Table: ANOVA Table for One Factor Model

Source	Sum of Square	*df*	Mean Square
Factor	Factor SS	$c - 1$	Factor MS
Error	Error SS	$n - c$	Error MS
Total	Total SS	$n - 1$	

- The usual (regression) definitions of R^2, s, and so forth, hold.

Regression and the One Factor ANOVA Model

- To link the linear (regression) model to the one factor ANOVA model:
  - For a categorical variable with c levels, define c binary variables, $x_1, x_2, \ldots, x_c$.
  - Here, x_j indicates whether or not an observation falls in the jth level.
  - Rewrite the one factor ANOVA model $y = \mu_j + \varepsilon$ as

$$y = \mu_1 x_1 + \mu_2 x_2 + \ldots + \mu_c x_c + \varepsilon.$$

- To include an intercept term, define $\tau_j = \mu_j - \mu$, where μ is an, as yet, unspecified parameter.
- The Greek "t", τ is for "treatment" effects.
- With $x_1 + x_2 + \ldots + x_c = 1$, we have

$$y = \mu + \tau_1 x_1 + \tau_2 x_2 + \ldots + \tau_c x_c + \varepsilon.$$

- A simpler expression is

$$y_{ij} = \mu + \tau_j + \varepsilon_{ij}.$$

Reparameterizing the One Factor ANOVA Model

- The one factor ANOVA Model can be expressed as

$$y_{ij} = \mu + \tau_j + \varepsilon_{ij}, \quad i = 1, \ldots, n + j, \quad j = 1, \ldots, c.$$

- There are $1 + c$ parameters, too many (starting with c means). “Overparameterized.”
- Two ways to restrict the parameters
  - Drop one of the binary variables, for example,

$$y = \mu + \tau_1 x_1 + \tau_2 x_2 + \ldots + \tau_{c-1} x_{c-1} + \varepsilon.$$

  - Minimize the sum of squares subject to the constraint that the sum of taus equals zero. More formally, we use the weighted sum $\sum_{j=1}^{c} n_j \tau_j = 0$.

The One Factor ANOVA Model - Matrix Expressions

- This model can be written as

$$y = \mu_1 x_1 + \mu_2 x_2 + \ldots + \mu_c x_c + \varepsilon.$$

- Using matrix notation, we have

$$\mathbf{y} = \begin{bmatrix} y_{1,1} \\ \cdot \\ \cdot \\ \cdot \\ y_{n_1,1} \\ \cdot \\ \cdot \\ \cdot \\ y_{1,c} \\ \cdot \\ \cdot \\ \cdot \\ y_{n_c,c} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ 1 & 0 & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & 1 \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \cdot \\ \cdot \\ \mu_c \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,1} \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_{n_1,1} \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_{1,c} \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_{n_c,c} \end{bmatrix} = \mathbf{X}\,\beta + \varepsilon$$

The One Factor ANOVA Model - Matrix Expressions

- We write $\mathbf{0}$ and $\mathbf{1}$ for a column of zeros and ones, respectively. Thus,

$$\mathbf{y} = \begin{bmatrix} \mathbf{1}_1 & \mathbf{0}_1 & \cdots & \mathbf{0}_1 \\ \mathbf{0}_2 & \mathbf{1}_2 & \cdots & \mathbf{0}_2 \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \mathbf{0}_c & \mathbf{0}_c & \cdots & \mathbf{1}_c \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \cdot \\ \cdot \\ \cdot \\ \mu_c \end{bmatrix} + \varepsilon = \mathbf{X}\beta + \varepsilon$$

- Here, $\mathbf{0}_1$ and $\mathbf{1}_1$ stand for vector columns of length n_1 of zeros and ones, respectively.

The One Factor ANOVA Model - Matrix Expressions

$$\begin{aligned} (\mathbf{X}'\mathbf{X})^{-1} &= \left(\begin{bmatrix} \mathbf{1}_1 & \mathbf{0}_2 & \cdots & \mathbf{0}_c \\ \mathbf{0}_1 & \mathbf{1}_2 & \cdots & \mathbf{0}_c \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \mathbf{0}_1 & \mathbf{0}_2 & \cdots & \mathbf{1}_c \end{bmatrix}' \begin{bmatrix} \mathbf{1}_1 & \mathbf{0}_1 & \cdots & \mathbf{0}_1 \\ \mathbf{0}_2 & \mathbf{1}_2 & \cdots & \mathbf{0}_2 \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \mathbf{0}_c & \mathbf{0}_c & \cdots & \mathbf{1}_c \end{bmatrix} \right)^{-1} \\ &= \begin{bmatrix} n_1 & 0 & \cdots & 0 \\ 0 & n_2 & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & n_c \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{n_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{n_2} & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & \frac{1}{n_c} \end{bmatrix}. \end{aligned}$$

The One Factor ANOVA Model - Matrix Expressions

- We don't need to calculate least square parameter estimates using $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, but we can...

$$\mathbf{b} = \begin{bmatrix}\hat\mu_1\\ \cdot \\ \cdot\\ \cdot\\ \hat\mu_c\end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix} \frac{1}{n_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{n_2} & \cdots & 0\\ \cdot&\cdot&\cdots&\cdot \\ \cdot&\cdot&\cdots&\cdot \\ \cdot&\cdot&\cdots&\cdot \\ 0&0&\cdots&\frac{1}{n_c}\end{bmatrix}\begin{bmatrix}\mathbf{1}_1 & \mathbf{0}_2 & \cdots & \mathbf{0}_c\\ \mathbf{0}_1 & \mathbf{1}_2 & \cdots & \mathbf{0}_c \\ \cdot&\cdot&\cdots&\cdot \\ \cdot&\cdot&\cdots&\cdot \\ \cdot&\cdot&\cdots&\cdot \\ \mathbf{0}_1 & \mathbf{0}_2 & \cdots & \mathbf{1}_c\end{bmatrix}' \begin{bmatrix} y_{1,1}\\ \cdot\\ \cdot\\ \cdot\\ y_{n_1,1}\\ \cdot\\ \cdot\\ \cdot\\ y_{1,c}\\ \cdot\\ \cdot\\ y_{n_c,c}\end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{n_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{n_2} & \cdots & 0\\ \cdot&\cdot&\cdots&\cdot \\ \cdot&\cdot&\cdots&\cdot \\ \cdot&\cdot&\cdots&\cdot \\ 0&0&\cdots&\frac{1}{n_c}\end{bmatrix} \begin{bmatrix}\sum_{i=1}^{n_1} y_{i1}\\ \cdot\\ \cdot\\ \cdot\\ \sum_{i=1}^{n_c} y_{ic}\end{bmatrix} = \begin{bmatrix}\bar{y}_1\\ \cdot\\ \cdot\\ \cdot\\ \bar{y}_c\end{bmatrix}$$

Combining a Factor and Covariate

- When combining, we use the terminology *factor* for the categorical variable and *covariate* for the continuous variable.
- Here are several different approaches for combining these variables

Table: Combinations of One Factor and One Covariate

Model Description	Notation
One factor ANOVA (no covariate model)	$y_{ij} = \mu_j + \varepsilon_{ij}$
Regression with constant intercept and slope (no factor model)	$y_{ij} = \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij}$
Regression with variable intercept and constant slope (analysis of covariance model)	$y_{ij} = \beta_{0j} + \beta_1 x_{ij} + \varepsilon_{ij}$
Regression with constant intercept and variable slope	$y_{ij} = \beta_0 + \beta_{1j} x_{ij} + \varepsilon_{ij}$
Regression with variable intercept and slope	$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + \varepsilon_{ij}$

Combining a Factor and Covariate

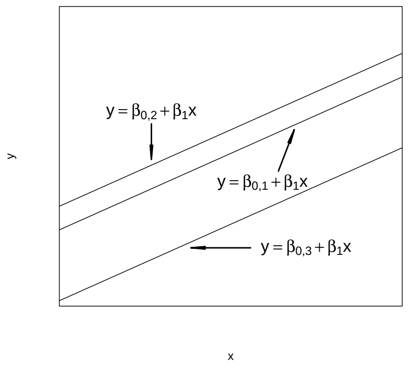


Figure: Plot of the expected response versus the covariate for the regression model with variable intercept and constant slope.

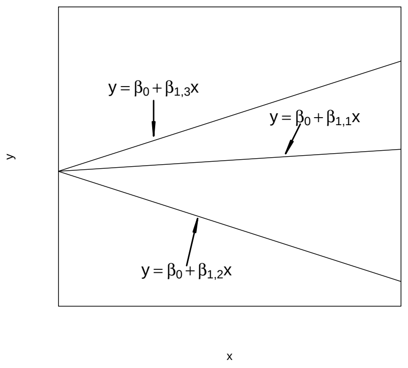


Figure: Plot of the expected response versus the covariate for the regression model with constant intercept and variable slope.

Combining a Factor and Covariate

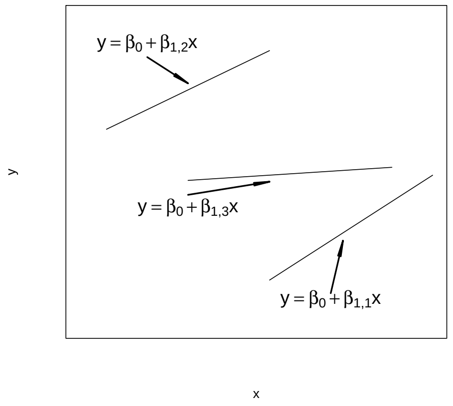


Figure: Plot of the expected response versus the covariate for the regression model with variable intercept and variable slope.

General Linear Model

- In the general linear model $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon$,
  - each explanatory variable may be continuous or categorical
  - The categorical variables are decomposed into binary variables (typically dropping one for identifiability)
  - Interactions between continuous and categorical variables are interpreted to be slopes that vary by the level of the category.
  - Interactions between categorical variables are interpreted to be as creating a new categorical variable
- Example, take gender [male versus female] and categorical age [young, middle, old], for six combinations of gender and age.
  - The binary variables $x_1 =$(gender=female), $x_2 =$(age=young) and $x_3 =$(age=middle) are known as "main effects."
  - We can incorporate two interaction variables $x_4 =$(gender=female)*(age=young) and $x_5 =$(gender=female)*(age=middle)
  - This gives five binary variables needed to represent age and gender.

Gender	Age	x_1	x_2	x_3	x_4	x_5	$\mathrm{E}\ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$
Male	Young	0	1	0	0	0	$\beta_0 + \beta_2$
Male	Middle	0	0	1	0	0	$\beta_0 + \beta_3$
Male	Old	0	0	0	0	0	β_0
Female	Young	1	1	0	1	0	$\beta_0 + \beta_1 + \beta_2 + \beta_4$
Female	Middle	1	0	1	0	1	$\beta_0 + \beta_1 + \beta_3 + \beta_5$
Female	Old	1	0	0	0	0	$\beta_0 + \beta_1$