Outline

What the Modeling Procedure Tells Us

The Importance of Variable Selection

The Importance of Data Collection

Missing Data Models

Application: Risk Managers Cost Effectiveness

Interpreting Individual Effects

- Substantive Significance
  - Does a 1 unit change in x imply an economically meaningful change in y?
  - Example: Looking at urban and rural claims experience, is there a big enough difference to warrant differentiating prices by location?
- Statistical Significance
  - We have standards for deciding whether or not a variable is statistically significant.
  - A "statistically significant effect" is the result of a regression coefficient that is large relative to its standard error, $se(b_j) = s\frac{\sqrt{VIF_j}}{s_{x_j}\sqrt{n-1}}$.
  - Statistical significance is driven by
    - precision of s,
    - collinearity (VIF) and
    - sample size
- Causal Effects
  - If we change x, would y change?
  - Three necessary conditions for causality
    - statistical association between variables,
    - appropriate time order and
    - the elimination of alternative hypotheses or establishment of a formal causal mechanism.

Other Interpretations

- Regression function and pricing
  - The regression function is $\mathrm{E}\ y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$.
  - Think about expected claims as our baseline price for short-term insurance coverages.
- Benchmarking studies
  - In studies of CEO's salaries, who is making a lot (or a little), controlled for industry, years of experience and so forth?
  - In studies of medical claims, who are the high-cost patients?
- Prediction
  - A new patient comes in with a given set of characteristics, $\mathbf{x}_* = (1, x_{*1}, \ldots, x_{*k})'$
  - What can I say about her future medical claims?

Prediction

- The new response is $y_* = \beta_0 + \beta_1 x_{*1} + ... + \beta_k x_{*k} + \varepsilon_*$.
- We use as our point predictor $\hat{y}_* = b_0 + b_1 x_{*1} + ... + b_k x_{*k}$.
- As in Chapter 2, we can decompose the prediction error as

$$\underbrace{y^* - \widehat{y}^*}_{\substack{\text{prediction}\\ \text{error}}} = \underbrace{\beta_0 - b_0 + ... + (\beta_k - b_k)x_{*k}}_{\substack{\text{error in estimating the}\\ \text{regression function at } x_{*1}, \ldots, x_{*k}}} + \underbrace{\varepsilon^*}_{\substack{\text{additional}\\ \text{deviation}}}$$

- We summarize this distribution using a prediction interval

$$\hat{y}_* \pm t_{n-(k+1),1-\alpha/2}\ se(pred),$$

where

$$se(pred) = s\sqrt{1 + \mathbf{x}_*'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_*}.$$

The Importance of Variable Selection

- With too many or too few variables, s is too large an estimate of σ.
  - Prediction intervals are too large
  - Standard errors for the partial slopes are too large
- With too few or incorrect variables, we produce biased estimates of the slopes β. Thus, our predictions are biased and hence inaccurate.

Principle of Parsimony

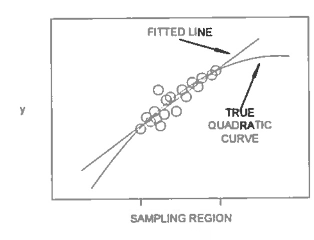
- The principle of parsimony, also known as Occam's Razor, states that when there are several possible explanations for a phenomenon, use the simplest.
  - A simpler explanation is easier to interpret.
  - Simpler models, also known as "more parsimonious" models, often do well on fitting out-of-sample data
  - Extraneous variables can cause problems of collinearity, leading to difficulty in interpreting individual coefficients.
- In contrast, in a quote often attributed to Albert Einstein, we should use "the simplest model possible, but no simpler."
  - Omitting important variables can lead to biased results, a potentially serious error.
  - Including extraneous variables decreases the degrees of freedom and increases the estimate of variability, typically of less concern in actuarial applications.

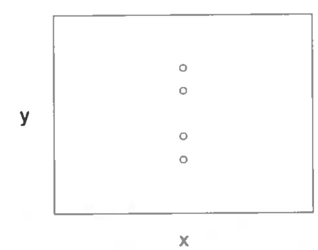
Bias Due to Sampling Frame Error

- Sampling frame error occurs when the sampling frame, the list from which the sample is drawn, is not an adequate approximation of the population of interest.
- Example: Literary Digest Poll - 1936 US presidential elections
  - Democrat Franklin D. Roosevelt versus Republican Alfred Landon
  - *Literary Digest*, a prominent magazine at the time, conducted a survey of *ten million* voters.
  - *2.4 million* responded: Landon 57% to 43%.
  - The actual election results: Roosevelt 62% to 38%!
  - What went wrong?
- Many things; among them, the wrong sampling frame
  - Literary Digest drew their sample from telephone books
  - Heavily skewed towards the wealthier
  - Economic problems were important in 1936.
  - Sampling frame did not represent the (poorer) Democrats
- Insurance company experience may not be representative of the overall population
  - May be due to underwriting, sometimes to self-selection in purchase of insurance
  - Example: the annuitant population typically has better mortality than the overall population
  - This is important when a company moves to a new market.

Bias Due to Limited Sampling Region

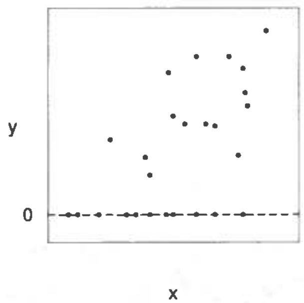
- A small spread of a variable, other things equal, means a less reliable estimate of the slope coefficient associated with that variable.
  - Left-hand panel: The lack of variation in x means that we cannot fit a unique line relating x and y. (Recall $se(b_j) = s\frac{\sqrt{VIF_j}}{s_{x_j}\sqrt{n-1}}$.)
- A potential bias can arise when we try to extrapolate outside of the sampling region.
  - Right-hand panel: Extrapolation outside of the sampling region may always be biased

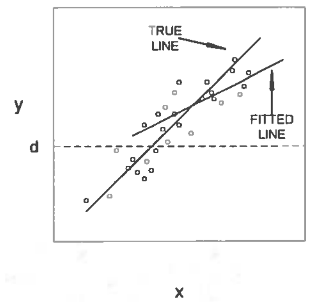




Bias Due to Limited Dependent Variables

- In some applications, the dependent variable is constrained to fall with certain regions.
- This means that our assumption of normal errors is not strictly correct, and may not even be a good approximation.
    - Left-hand panel: When individuals do not purchase anything, they are recorded as $y = 0$ sales. (Censored)
    - Right-hand panel: If the responses below the horizontal line at $y = d$ are omitted, then the fitted regression line is very different from the true regression line. (Truncated)





Bias Due to Omitted and Endogenous Variables

- Bias Due to Endogenous Explanatory Variables
    - An exogenous variable is one that can be taken as "given" for the purposes at hand. An endogenous variable is one that fails the exogeneity requirement.
    - Intuitively, an endogenous explanatory variable x is one where the dependent variable y affects the x
    - Up to now, the explanatory variables have been treated as non-stochastic.
    - For many social science applications, it is more intuitive to consider X's to be stochastic, and perform statistical conditional on their realizations.
        - For example, under common sampling schemes, we can estimate the conditional regression function

$$\mathrm{E}\ (y|x_1, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k.$$

        - Known as a "sampling-based" model.

Bias Due to Missing Data

- It is common in the social sciences for data to be unavailable for analysis, or *missing*.
- When the reason for the lack of availability of data is unrelated to actual data values, the data are said to be *missing at random*.
- Many ways to handle missing at random data, none clearly superior to the others.
    - One technique is to simply ignore the problem. Hence, missing at random is sometimes called the *ignorable case* of missing data.
    - If there are only a few missing data, compared to the total number available, a widely employed strategy is to delete the observations corresponding to the missing data.
    - If the missing data are primarily from one variable, we can consider omitting this variable.
    - Another strategy (many variations) is to fill in, or *impute*, missing data.

Survey Data

- Risk management practices are activities undertaken by a firm to minimize the potential cost of future contingent losses, such as the event of a fire in a warehouse or an accident that injures employees.
- What is the effect of risk management practices on firm costs?
- Schmit and Roth (1990) conducted a survey of risk managers of large U.S.-based organizations.
    - Questionnaire sent to 374 managers - 162 returned completed surveys. Response rate of 43%.
    - Only 73 forms were complete. Response rate of 20%.
    - Why such a dramatic difference?
        - Managers, like most people, typically do not mind responding to queries about their attitudes, or opinions, about various issues.
        - For hard financial information, they are less likely to respond because of effort involved in collecting the information or the proprietary nature of it.

Hypotheses

- The variables analyzed are:
  - FIRMCOST - total property and casualty premiums and uninsured losses as a percentage of total assets
  - SIZELOG is the logarithm of total assets
  - ASSUME is the per occurrence retention amount as a percentage of total assets
  - CAP indicates whether the company owns a captive insurance company
  - INDCOST - a measure of the firm's industry risk
  - CENTRAL - a measure of the importance of the local managers in choosing the amount of risk to be retained
  - SOPH - a measure of the degree of importance in using analytical tools, such as regression, in making risk management decisions
- The hypotheses are:
  - Larger retention amounts (ASSUME) means lower expenses to a firm, resulting in lower costs (FIRMCOST).
  - The use of a captive insurance company (CAP) results in lower costs.
  - There exists an inverse relationship between the measure of centralization (CENTRAL) and cost (FIRMCOST).
  - More sophisticated analytical tools (SOPH) help firms to manage risk better, resulting in lower costs (FIRMCOST).

Preliminary Results

Table: Regression Results from a Preliminary Model Fit

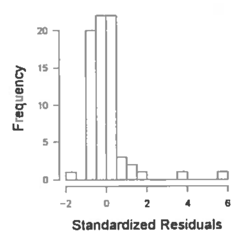
Variable	Coefficient	Standard Error	t-statistic
INTERCEPT	59.76	19.1	3.13
ASSUME	-0.300	0.222	-1.35
CAP	5.50	3.85	1.43
SIZELOG	-6.84	1.92	-3.56
INDCOST	23.08	8.30	2.78
CENTRAL	0.133	1.44	0.89
SOPH	-0.137	0.347	-0.39

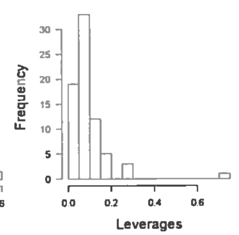
$R_a^2 = 18.8\%$, the $F - ratio = 3.78$ and $s = 14.56$

- The two risk measure variables are statistically insignificant.
- The p-value on ASSUME is 9%.
- The coefficient of CAP has the wrong sign!!

Critiques of the Preliminary Model Fit

- Histograms of standardized residuals and leverages from a preliminary regression model fit.
- The largest residual turns out to be $e_{15} = 83.73$.
- $Error\ SS = (n-(k+1))s^2 = (73-7)(14.56)^2 = 13,987$.
- This observation represents 50.1% of the error sum of squares $(= 83.73^2/13,987)$, suggesting that this 1 observation out of 73 has a dominant impact on the model fit.





Back to the Basics

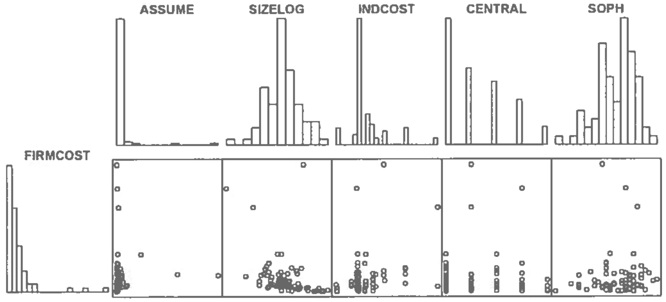
- The largest value of FIRMCOST is 97.55 is more than five standard deviations above the mean $[10.97 + 5(16.16) = 91.77]$.
- The largest value of ASSUME is more than 7 standard deviations above the mean.

Table: Summary Statistics of $n = 73$ Risk Management Surveys

	Mean	Median	Standard Deviation	Minimum	Maximum
FIRMCOST	10.97	6.08	16.16	0.20	97.55
ASSUME	2.574	0.510	8.445	0.000	61.820
CAP	0.342	0.000	0.478	0.000	1.000
SIZELOG	8.332	8.270	0.963	5.270	10.600
INDCOST	0.418	0.340	0.216	0.090	1.220
CENTRAL	2.247	2.200	1.256	1.000	5.000
SOPH	21.192	23.00	5.304	5.000	31.000

Histograms and Scatter Plots

- Distributions of FIRMCOST and ASSUME are skewed
- Negative relationship between FIRMCOST and SIZELOG.



Correlations

- Define $COSTLOG = \ln(FIRMCOST)$

Table: Correlation Matrix

	COST LOG	FIRM COST	ASSUME	CAP	SIZE LOG	IND COST	CENTRAL
FIRMCOST	0.713						
ASSUME	0.165	0.039					
CAP	-0.088	0.088	0.231				
SIZELOG	-0.637	-0.366	-0.209	0.196			
INDCOST	0.395	0.326	0.249	0.122	-0.102		
CENTRAL	-0.054	0.014	-0.068	-0.004	-0.080	-0.085	
SOPH	0.144	0.048	0.062	-0.087	-0.209	0.093	0.283

Revised Regression

Table: Regression Results - COSTLOG as Dependent Variable

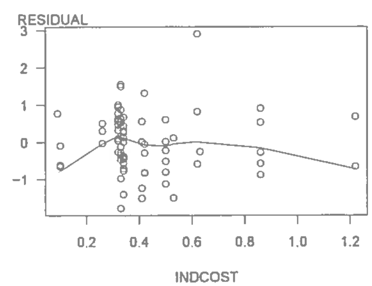
Variable	Coefficient	Standard Error	t-statistic
INTERCEPT	7.64	1.16	6.62
ASSUME	-0.008	0.013	-0.61
CAP	0.015	0.233	0.06
SIZELOG	-0.787	0.117	-6.75
INDCOST	1.90	0.503	3.79
CENTRAL	-0.080	0.087	-0.92
SOPH	0.002	0.021	0.12

$R_a^2 = 48\%$, the $F-ratio = 12.1$ and $s = 0.882$

- Still, the two risk measure variables are statistically insignificant.
- The leverages have not changed (why?).
- Four of six variables are statistically insignificant

Improving the Model

- Stepwise regression suggests that only SIZELOG and INDCOST are important
- This regression was run, producing residuals.
- The nonparametric fitted curve (using lowess) suggests a quadratic term in INDCOST.



Quadratic Model Fit

- The quadratic term appears to be statistically significant

Table: Regression Results with a Quadratic term in INDCOST

Variable	Coefficient	Standard Error	*t*-statistic
INTERCEPT	6.35	0.953	6.67
SIZELOG	-0.773	0.101	-7.63
INDCOST	6.26	1.61	3.89
$INDCOST^2$	-3.58	1.27	-2.83

$R_a^2 = 54.7\%$, the $F - ratio = 29.9$ and $s = 0.823$