Regression Modeling with Actuarial and Financial Applications

by Edward Frees

Ch. 3 notes

- A scatterplot matrix is basically one large graph that is subdivided into every combination of scatterplots that you can have between every pair of variables. This is useful to look at to quickly detect any problems that need to be corrected for in any variable’s list, such as nonlinearity or outliers.
- Correlations are great because numbers are hard to lie with, they are concrete values of what you are seeing graphically. However, they only measure *linear* relationships, your variables may have a strong relationship which is not linear, this would not be reflected in a correlation.

Some goodness of fit measures

- The proportion of variability (still, just like the simple linear regression model) explained by the regression model is

$$R^2 = \frac{\text{Regression SS}}{\text{Total SS}} = \frac{\sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}$$

- This is also called the coefficient of determination.
- When an explanatory variable is added to the regression model, unfortunately, this R^2 never decreases.
- The adjusted R^2 defined by

$$R_a^2 = 1 - \frac{\text{Error SS}/(n-(k+1))}{\text{Total SS}/(n-1)} = 1 - \frac{s^2}{s_Y^2},$$

provides for the proportion of the variation explained by the regression, but adjusted for the number of predictor variables (or degrees of freedom).

Example. Demand for Term Life Insurance

- "Who buys insurance and how much do they buy?"
  - Companies have data on current customers
  - How do get info on potential (new) customers?
- To understand demand of insurance, we look at the Survey of Consumer Finances (SCF)
  - This is a nationally representative sample that contains extensive information on assets, liabilities, income, and demographic characteristics of those sampled (potential U.S. customers).
  - We study a random sample of 500 of the 4,519 households with positive income that were interviewed in the 2004 survey.
  - We now focus on $n = 275$ households that purchased term life insurance
  - Chapter 11 will consider models to predict whether or not someone purchases insurance.

Example. Term Life Insurance

- We study y = FACE, the amount that the company will pay in the event of the death of the named insured.
- We focus on $k = 3$ explanatory variables
  - annual INCOME,
  - the number of years of EDUCATION of the survey respondent and
  - the number of household members, NUMHH.
- The statistics suggest that INCOME and FACE are skewed.

Table: Term Life Summary Statistics

Variable	Mean	Median	Standard Deviation	Minimum	Maximum
FACE	747,581	150,000	1,674,362	800	14,000,000
INCOME	208,975	65,000	824,010	260	10,000,000
EDUCATION	14.524	16.000	2.549	2.000	17.000
NUMHH	2.960	3.000	1.493	1.000	9.000
LNFACE	11.990	11.918	1.871	6.685	16.455
LNINCOME	11.149	11.082	1.295	5.561	16.118

Example. Term Life Insurance

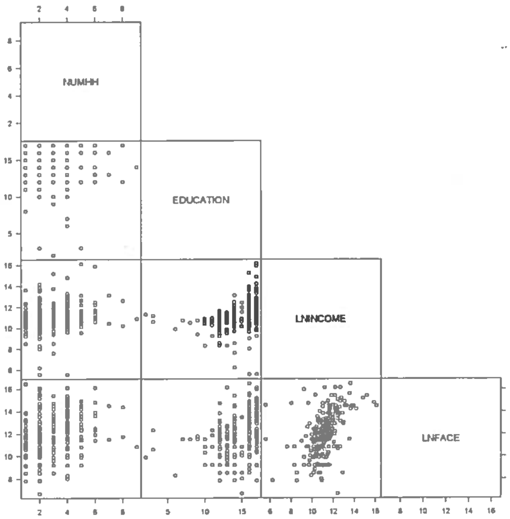
- We first examine pairs of correlations

Table: Term Life Correlations

	NUMHH	EDUCATION	LNINCOME
EDUCATION	-0.064		
LNINCOME	0.179	0.343	
LNFACE	0.288	0.383	0.482

- The following graph, called a *scatterplot matrix*, gives the same type of information graphically

Scatterplot Matrix



Regression Coefficient Interpretation

- For the Term Life example, recall

$$\widehat{y} = 2.584 + 0.206 EDUCATION + 0.306 NUMHH + 0.494 LNINCOME.$$

- Begin by interpreting the *sign* of each coefficient.
  - For example, the positive sign associated with EDUCATION ($b_1 = 0.206$) is reasonable, more education suggests that respondents are more aware of their insurance needs, other things being equal.
- Also interpret the *amount* of the regression coefficient.
  - From the regression equation, we might say that if EDUCATION increases by .1 years, then we anticipate y to increase by 0.0206 logarithmic dollars.

General Procedures to Test a Single Variable

Table: Decision-Making Procedures for Testing $H_0 : \beta_j = d$

Alternative Hypothesis (H_a)	Procedure: Reject H_0 in favor of H_a if
$\beta_j > d$	$t - ratio > t_{n-(k+1),1-\alpha}$.
$\beta_j < d$	$t - ratio < -t_{n-(k+1),1-\alpha}$.
$\beta_j \neq d$	$\lvert t - ratio \rvert > t_{n-(k+1),1-\alpha/2}$.

Notes: The significance level is α. Here, $t_{n-(k+1),1-\alpha}$ is the (1-α)th percentile from the t-distribution using $df = n - (k+1)$ degrees of freedom.

Table: Probability Values for Testing $H_0 : \beta_j = d$

Alternative Hypothesis (H_a)	$\beta_j > d$	$\beta_j < d$	$\beta_j \neq d$
p-value	$\Pr(t_{n-(k+1)} > t\text{-ratio})$	$\Pr(t_{n-(k+1)} < t\text{-ratio})$	$\Pr(\lvert t_{n-(k+1)} \rvert > \lvert t\text{-ratio} \rvert)$

Example. Refrigerator Prices

Table: Summary Statistics for each variable for 37 Refrigerators

Variable	Mean	Median	Standard Deviation	Minimum	Maximum
ECOST	70.51	68.00	9.14	60.00	94.00
RSIZE	13.400	13.200	0.600	12.600	14.700
FSIZE	5.184	5.100	0.938	4.100	7.400
SHELVES	2.514	2.000	1.121	1.000	5.000
FEATURES	3.459	3.000	2.512	1.000	12.000
PRICE	626.4	590.0	139.8	460.0	1200.0

Source: *Consumer Reports* "Refrigerators: A Comprehensive Guide to the Big White Box."

- We wish to understand the PRICE of a refrigerator.
  - ECOST - energy cost - the average amount of money spent per year to operate the refrigerator
  - including the size of the refrigerator in cubic feet (RSIZE), the size of the freezer compartment in cubic feet (FSIZE), the number of shelves in the refrigerator and freezer doors (SHELVES), and the number of features (FEATURES).

Example. Refrigerator Prices

- Surprisingly, there is a positive correlation between ECOST and PRICE.

Table: Matrix of Correlation Coefficients

	ECOST	RSIZE	FSIZE	SHELVES	FEATURES
RSIZE	0.333				
FSIZE	0.855	0.235			
SHELVES	0.188	0.363	0.251		
FEATURES	0.334	0.096	0.439	0.160	
PRICE	0.522	0.024	0.720	0.400	0.697

- The fitted regression equation is:

$\widehat{PRICE}$ =	798	-6.96 ECOST	+ 76.5 RSIZE	+ 137 FSIZE	+ 37.9 SHELVES
std errors	(271.4)	(2.275)	(19.44)	(23.76)	(9.886)
t-ratios	[-2.9]	[-3.1]	[3.9]	[5.8]	[3.8]

with $s = 60.65$ and $R^2 = 83.8$ percent.

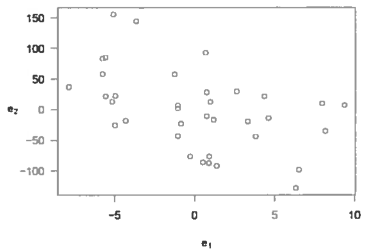
- The regression shows a negative coefficient for ECOST!!!

Added Variable Plot

- The added variable plot (partial regression plot) is a plot of PRICE versus ECOST, *controlling for* the effects of other explanatory variables.
- For this example, "size" is driving both PRICE and ECOST. Larger refrigerators have larger ECOST and PRICE. We wish to control for this.
- *Procedure for producing an added variable plot.*
  - Run a regression of PRICE on RSIZE, FSIZE, SHELVES and FEATURES, omitting ECOST. Compute the residuals from this regression, which we label $\widehat{e}_1$.
  - Run a regression of ECOST on RSIZE, FSIZE, SHELVES and FEATURES. Compute the residuals from this regression, which we label $\widehat{e}_2$.
  - Plot $\widehat{e}_1$ versus $\widehat{e}_2$. This is the added variable plot of PRICE versus ECOST, controlling for the effects of the RSIZE, FSIZE, SHELVES and FEATURES.

Added Variable Plot

- Vertical axis - the residuals from the regression of PRICE on the explanatory variables, omitting ECOST.
- Horizontal axis - the residuals from the regression fit of ECOST on the other explanatory variables.
- The correlation coefficient is -0.48.



Partial Correlation Coefficients

- Variables left out of a regression are called *omitted variables*.
  - This omission could cause a serious problem in regression model fit.
- The correlation for the added variable plot is called a *partial correlation coefficient*.
- The partial correlation coefficient can also be calculated using

$$r(y, x_j | x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k) = \frac{t(b_j)}{\sqrt{t(b_j)^2 + n - (k+1)}}.$$

  - Here, $t(b_j)$ is the t-ratio for b_j from a regression of y on $x_1, \ldots, x_k$ (including the variable x_j).
  - This allows us to calculate partial correlation coefficients running only one regression.
  - For example, the partial correlation between PRICE and ECOST in the presence of the other explanatory variables is $(-3.1)/\sqrt{(-3.1)^2 + 37 - (5+1)} \approx -0.48$.