ST5202: Applied Regression Analysis

Department of Statistics and Applied Probability
National University of Singapore

05-March-2018
Lecture 7

Announcement

- Assignment #3 due today
- Midterm on **12 March** from 7:00pm to 9:00pm at **LT28**.
  - NON-PROGRAMMABLE calculator is allowed.
  - ONE A4-sized help sheet is allowed. You can write or print anything on both sides.

Lecture 7

Multiple Regression II (Chapter 7 continued) &
Regression Models for Quantitative and Qualitative Predictors
(Chapter 8)

Outline

- Multiple Regression II (Ch. 6)
  - Correlated predictor variables
- Polynomial Regression Models
- Interaction Terms
- Qualitative Variables
- Interactions with Qualitative Variables

Review

- Extra sum of squares
  - Decompose SSR to measure marginal reduction in error sum of squares when an extra variable is added to the model.

    e.g., $SSR(X_2|X_1)$, $SSR(X_3|X_1, X_2)$, ...
  - For two sets S and R for predictor variables:
    $SSR(X_S|X_R) = SSR(X_S, X_R) - SSR(X_R)$

    e.g., $SSR(X_2, X_3|X_1) = SSR(X_1, X_2, X_3) - SSR(X_1)$
    $SSR(X_3|X_1, X_2) = SSR(X_1, X_2, X_3) - SSR(X_1, X_2)$
  - $SSR(X_1, X_2, \cdots, X_{p-1}) =$
    $SSR(X_1) + SSR(X_2|X_1) + \cdots + SSR(X_{p-1}|X_1, \cdots, X_{p-2})$

Review

- Partial F-test
  - for single predictor variable
    - Test $\beta_k = 0$ with general linear test approach.
      Reduced model: $E\{Y\} = \beta_0 + \sum_{j \neq k} \beta_j X_j$ versus full model:$E\{Y\} = \beta_0 + \sum_j \beta_j X_j$
    - Partial F-statistic:

$$
\begin{aligned}
F^* &= \frac{SSE(R) - SSE(F)}{1} \Big/ \frac{SSE(F)}{n-p} \\
&= \frac{SSR(X_k | X_{-k})}{SSE(X_1, \cdots, X_{p-1})/(n-p)} \\
&\sim F(1, n-p) \text{ under } H_0
\end{aligned}
$$

    - Equivalent to t-test for testing $\beta_k = 0$: $F^* = t^{*2}$

Review

- Partial F-tests, for a subset of predictor variables
  - Test if several regression coefficients are zero:
    Test $H_0 : \beta_k = 0$ for any $k \in S$,
    (with S a set of indices, e.g., $S = \{3, 4, 5\}$)
    versus $H_a : \exists k \in S$, with $\beta_k \neq 0$,
  - Partial F-statistic (with $\tilde{S}$ the number of elements in S):

$$
\begin{aligned}
F^* &= \frac{SSR(X_S | X_{-S})/\tilde{S}}{SSE(X_1, \cdots, X_{p-1})/(n-p)} \\
&\sim F(\tilde{S}, n-p) \text{ under } H_0
\end{aligned}
$$

Correlated predictor variables

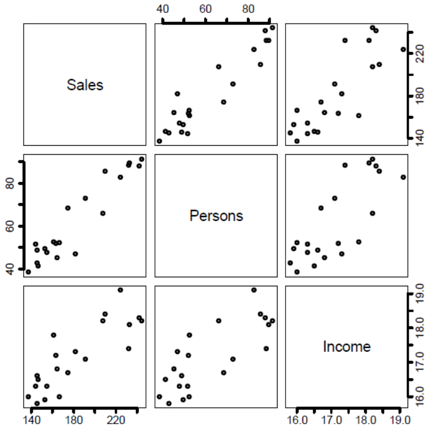
- Portrait studio example: the regression coefficient for X_2 (income in city) differs between models I and II:

```
---------------------------------------------------------
Model I: lm(formula = Y ~ X2)
            Estimate Std. Error t value Pr(>|t|)
X2            31.173      4.698   6.636 2.39e-06 ***
Model II: lm(formula = Y ~ X1 + X2)
X2            9.3655     4.0640   2.305   0.0333 *
---------------------------------------------------------
```

Correlated predictor variables–continued

- Why? Interpretation of regression coeff. for X_2:
  - in model I: average increase in expected sales when X_2 increases by 1 unit (regardless of what's happening with X_1)
  - in model II: average increase in expected sales when X_2 increases by 1 unit, **when holding X_1 constant**
- If X_1 and X_2 are correlated (if there is an empirical relation between X_1 and X_2), the coefficient for X_2 will change when X_1 is included, because the relation of X_1 with Y is now "controlled for".

Empirical relation between Y, X_1, and X_2



make a scatter plot matrix (R code: pair($\cdot$))

Correlated predictor variables

- Regression coefficients of correlated predictor variables depend on whether the other predictor variable is included in the model ("has been controlled for").
- Regression coefficients of uncorrelated predictor variables do NOT depend on whether the other predictor variable is included in the model ("has been controlled for").
- What happens with standard error $s\{b_k\}$, confidence intervals for β_k, test statistics related to β_k when adding/removing predictor variables?
  - Removing any predictor variables with $\beta_k \neq 0$ will change the degrees of freedom for SSE and the residuals e_i
  - Result: (most likely) MSE changes and $s\{b_k\}$'s change, degrees of freedom in $t-$distribution for $\frac{b_k - \beta_k}{s\{b_k\}}$ change, confidence intervals and test statistics change

Correlated transformation

- In linear regreesion model, center all the variables at zero and rescale:

$$
\begin{aligned}
Y_i^* &= \frac{1}{\sqrt{n-1}} \frac{Y_i - \bar{Y}}{s_Y} \\
X_{ik}^* &= \frac{1}{\sqrt{n-1}} \frac{X_{ik} - \bar{X}_k}{s_k}, \quad k = 1, \cdots, p-1,
\end{aligned}
$$

with $S_Y = \sqrt{\frac{\sum (Y_i - \bar{Y})^2}{n-1}}$, $s_k = \sqrt{\frac{\sum_i (X_{ik} - \bar{X}_k)^2}{n-1}}$

- Why? To avoid rounding errors in $(\mathbf{X}'\mathbf{X})^{-1}$, to make regression coefficients comparable between predictors, and because it's helpful in thinking about the effect of correlation between the predictor variables on inference

Correlation transformation–continued

- Lack of comparability of regression coefficients
  - Suppose we have a model with two predictors.
    X_1-trees per 10^4 square meters in the range of $0 - 5000$, and
    X_2-tree diameter with a range of $0 - 50$ cm.

    The regression coefficients b_1 and b_2 are likely to have very different magnitude, with the result that increase of one unit in X_1 will have an entirely different effect on the response than from a unit change in X_2.

Standardized regression model

- Define matrix consisting of the transformed X variables

$$\underbrace{\mathbf{X}}_{n\times(p-1)} = \begin{pmatrix} X_{11}^* & \cdots & X_{1,p-1}^* \\ X_{21}^* & \cdots & X_{2,p-1}^* \\ \vdots & \vdots & \vdots \\ X_{n1}^* & \cdots & X_{n,p-1}^* \end{pmatrix}$$

- Recall the correlation matrix of the X variables

$$\underbrace{\mathbf{r}_{XX}}_{(p-1)\times(p-1)} = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1,p-1} \\ r_{12} & 1 & \cdots & r_{2,p-1} \\ \vdots & \vdots & & \vdots \\ r_{p-1,1} & r_{p-1,2} & \cdots & 1 \end{pmatrix}$$

Standardized regression model

- $Y_i^* = \beta_0^* + \beta_1^* X_{i1}^* + \cdots + \beta_{p-1}^* X_{i,p-1}^* + \epsilon_i^*$
  - $b_0^* = \bar{Y} - b_1^* \bar{X}_1^* - \cdots - b_{p-1}^* \bar{X}_{p-1}^* = 0$
- For the standardized model:

$$\mathbf{X}'\mathbf{X} = \mathbf{r}_{XX}$$

  with $\mathbf{r}_{XX}$ being the correlation matrix of the X_k's:

$$(\mathbf{r}_{XX})_{[k,s]} = \frac{\sum (X_{ik} - \bar{X}_k)(X_{is} - \bar{X}_s)}{\sqrt{\sum (X_{ik} - \bar{X}_k)^2 \sum (X_{is} - \bar{X}_s)^2}}$$

  thus all elements in $\mathbf{X}'\mathbf{X}$ are between -1 and 1
- $\mathbf{X}'\mathbf{Y} = \mathbf{r}_{YX}$, with $\mathbf{r}_{YX}$ the correlation vector with the correlation between Y and the X_k's.
- Then $\mathbf{b}^* = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{r}_{XX}^{-1}\mathbf{r}_{YX}$
- What happens if the X_k's are correlated?

Standardized regression model

- Employing the relations,

$$\underbrace{\boldsymbol{b}^*}_{(p-1)\times 1} = \begin{pmatrix} b_1^* \\ b_2^* \\ \vdots \\ b_{p-1}^* \end{pmatrix}, \underbrace{\boldsymbol{b}}_{(p)\times 1} = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_{p-1} \end{pmatrix}$$

$$\begin{aligned} b_k &= \frac{s_Y}{s_{X_k}} b_k^* (k = 1, \cdots, p-1) \\ b_0 &= \bar{Y} - b_1 \bar{X}_1 - \cdots - b_{p-1} \bar{X}_{p-1} \end{aligned}$$

Two extreme examples for $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

- X_1 and X_2 are uncorrelated:
  - $\mathbf{r}_{xx} = \mathbf{I}$
  - In the standardized model $\mathbf{b}^* = \mathbf{r}_{XX}^{-1}\mathbf{r}_{YX} = \mathbf{r}_{YX}$ thus $b_k^* = r_{YX_k}$
  - Then $b_k^{(ordinary)} = \frac{s_Y}{s_{X_k}} r_{YX_k}$, the same b_k's as in a simple linear regression with just one X_k (recall $b_2 = r_{YX}\frac{s_Y}{s_{X_2}}$)
  - The information contained in X_1 and X_2 "do not overlap"
- Correlation between X_1 and X_2 is 1:
  - $\mathbf{r}_{XX} = \mathbf{J}$
  - The inverse of $\mathbf{r}_{XX}$ does not exist (since determinant is zero), so there is NO solution to the normal equations:
    there is no unique solution for (b_1, b_2)
  - Similarly if $X_1 = X_2$, then $(b_1 + c, b_2 - c)$ are estimates for β_1 and β_2 for any constant c (thus no unique solution)

What if X's are highly correlated?

- X_1 and X_2 are highly correlated (multicollinearity):
  - SE's of the b_k's are very large, because $[(\mathbf{r}_{XX})^{-1}]_{[kk]}$ are very large
  - You can get a wide range of solutions for b_1 and b_2, depending on the random errors in Y (e.g., sign is unexpected)
- What does that mean for inference?
  - Inference about mean response and for new observations are still okay (because the b_k's are used jointly)
  - Inference about the β_k;s based on b_k's and their standard errors is unstable, which causes problems when the goal is:
    - to estimate the effect of a given predictor X_k on Y
    - to choose "important" variables that are associated with Y
- Diagnostics in Chapter 10, remedial measures in Chapter 11.

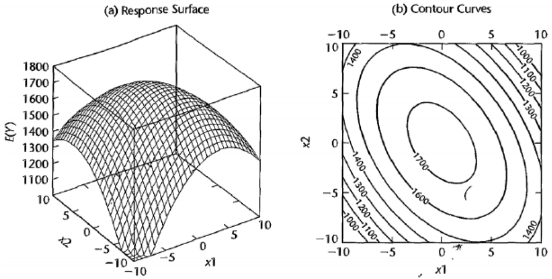
Polynomial regression models

- Include higher order terms (e.g., X_1^2) in the regression model
- Used when:
  - response function is truly a polynomial
  - response function can be approximated by a polynomial
- Approach:
  - Center the predictor variables to reduce correlation, use $(X_1 - \bar{X}_1)$
  - Hierarchical approach: don't exclude lower order terms
- Disadvantage: Hard to interpret the coefficients, and extrapolation risky

Representation for the response surface

- Response surface and contour curves for a second-order response function

$$E\{Y\} = 1740 - 4x_1^2 - 3x_2^2 - 3x_1x_2$$



Hierarchical approach

- Idea: often fit a second-order or third-order model and then explore whether a lower-order is adequate

$$Y_i = \beta_0 + \beta_1 x_i + \beta_{11} x_i^2 + \beta_{111} x_i^3 + \epsilon_i$$

- To test $\beta_{111} = 0$, or test $\beta_{11} = \beta_{111} = 0$
- To check $SSR(x^3|x, x^2)$, or check $SSR(x^2, x^3|x)$ (note that $SSR(x^2, x^3|x) = SSR(x^2|x) + SSR(x^3|x, x^2)$)
- Provides more basic information about the shape of the response function (cubic term is only refinement compared with lower order)
- Would not drop the lower order (e.g., quadratic term) but retain the higher order (e.g., cubic term)

Hierarchical approach

- Researcher studied the effects of the charge rate and temperature on the life a new type of power cell in a preliminary small-scale experiment
- The charge rate (X_1) was controlled at three levels and the ambient temperature (X_2) was controlled at three levels. Factors pertaining to the discharge of the power cell were held at fixed levels. The life of the power cell (Y) was measured in terms of the number of discharge-charge cycles that a power cell underwent before it failed.
- The regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{11} x_{i1}^2 + \beta_{22} x_{i2}^2 + \beta_{12} x_{i1} x_{i2} + \epsilon_i$$

Power cell example-continued

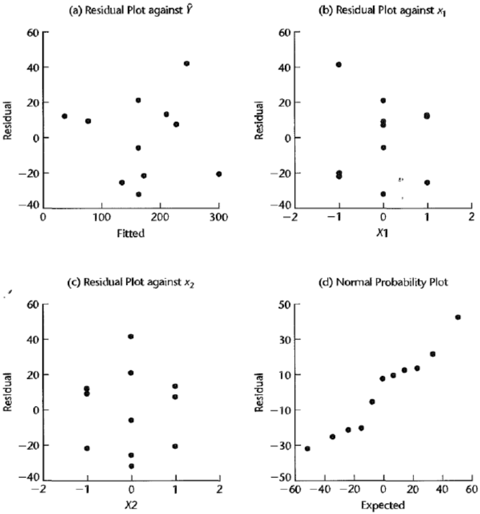
- Centering and scaling X_1 and X_2

$$x_{i1} = \frac{X_{i1} - \bar{X}_1}{.4} = \frac{X_{i1} - 1.0}{.4}$$

$$x_{i2} = \frac{X_{i2} - \bar{X}_2}{10} = \frac{X_{i2} - 20}{10}$$

Cell i	(1) Number of Cycles Y_i	(2) Charge Rate X_{i1}	(3) Temperature X_{i2}	(4) Coded Values x_{i1}	(5) x_{i2}	(6) x_{i1}^2	(7) x_{i2}^2	(8) $x_{i1}x_{i2}$
1	150	.6	10	−1	−1	1	1	1
2	86	1.0	10	0	−1	0	1	0
3	49	1.4	10	1	−1	1	1	−1
4	288	.6	20	−1	0	1	0	0
5	157	1.0	20	0	0	0	0	0
6	131	1.0	20	0	0	0	0	0
7	184	1.0	20	0	0	0	0	0
8	109	1.4	20	1	0	1	0	0
9	279	.6	30	−1	1	1	1	−1
10	235	1.0	30	0	1	0	1	0
11	224	1.4	30	1	1	1	1	1
		$\bar{X}_1 = 1.0$	$\bar{X}_2 = 20$					

Power cell example-continued



Power cell example-continued

Model: MODEL1
Dependent Variable: Y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	5	55365.56140	11073.11228	10.565	0.0109
Error	5	5240.43860	1048.08772		
C Total	10	60606.00000			

Root MSE	32.37418	R-square	0.9135
Dep Mean	172.00000	Adj R-sq	0.8271
C.V.	18.82220		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	162.842105	16.60760542	9.805	0.0002
X1	1	-55.833333	13.21670483	-4.224	0.0083
X2	1	75.500000	13.21670483	5.712	0.0023
X1SQ	1	27.394737	20.34007956	1.347	0.2359
X2SQ	1	-10.605263	20.34007956	-0.521	0.6244
X1X2	1	11.500000	16.18709146	0.710	0.5092

Variable	DF	Type I SS
INTERCEP	1	325424
X1	1	18704
X2	1	34202
X1SQ	1	1645.966667
X2SQ	1	284.928070
X1X2	1	529.000000

Power cell example-continued

- Lack of fit: $c = 9$ distinct combinations of levels of the X variables.
- $SSPE = (157 - 157.33)^2 + (131 - 157.33)^2 + (184 - 157.33)^2 = 1404.67$ (only three replications at $x_1 = 0$, and $x_2 = 0$)
- $SSLF = SSE - SSPE = 5240.44 - 1404.67 = 3835.77$
- $F^* = \frac{SSLF}{c-p} \times \frac{n-c}{SSPE} = \frac{3835.77}{9-6} \times \frac{11-9}{1404.67} = 1.82 \leq F(0.95; 3, 2) = 19.2$
- The second-order polynomial regression function

$$E\{Y\} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2$$

is a good fit

Power cell example-continued

- Consider whether a first-order model would be sufficient

$$\begin{aligned} H_0 &: \beta_{11} = \beta_{22} = \beta_{12} = 0 \\ H_a &: \text{not all } \beta\text{'s in } H_0 \text{ equal zero} \end{aligned}$$

The partial F test statistic is:

$$\begin{aligned} F^* &= \frac{SSR(x_1^2, x_2^2, x_1x_2|x_1, x_2)}{3} / MSE \\ &= SSR(x_1^2|x_1, x_2) + SSR(x_2^2|x_1, x_2, x_1^2) + SSR(x_1x_2|x_1, x_2, x_1^2, x_2^2) \\ &= \frac{2459.9}{3} / 1048.1 = .78 < F(.95; 3, 5) = 5.41^2 \end{aligned}$$

(Note $SSR(x_1^2|x_1, x_2) + SSR(x_2^2|x_1, x_2, x_1^2) + SSR(x_1x_2|x_1, x_2, x_1^2, x_2^2) = 1646.0 + 284.9 + 529.0$)

- Conclude H_0 that no curvature and interaction effects are needed.

Power cell example-continued

- Fit a first-order model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

- The estimated function

$$\hat{Y} = 172.00 - \underbrace{55.83}_{s\{b_1\}=12.67} x_1 + \underbrace{75.50}_{s\{b_2\}=12.67} x_2$$

- The coefficients b_1 and b_2 are the same for the fitted second-order model (this is a result of the choices of the X_1 and X_2 levels)
- Transforming the regression function back to the original variable (how?)

Power cell example-continued

- The estimated function

$$\hat{Y} = 160.58 - \underbrace{139.58}_{s\{b_1'\}=31.68} X_1 + \underbrace{7.55}_{s\{b_2'\}=1.267} X_2$$

- The standard deviations of b_1' and b_2'

$$\begin{aligned} s\{b_1'\} &= \left(\frac{1}{.4}\right) s\{b_1\} = \frac{12.67}{.4} = 31.68 \\ s\{b_2'\} &= \left(\frac{1}{10}\right) s\{b_2\} = \frac{12.67}{10} = 1.267 \end{aligned}$$

- Statistical inference applies such as the Bonferroni confidence limits for β_1 and β_2

Interaction terms

- Two variables interact (in determining a dependent variable) if the partial effect of one depends on the value/level/outcome of the other
- Example:
  - Expected sales is predicted using expenditure on local news paper advertisement (X_1) and TV commercials (X_2)
  - X_1 and X_2 interact if the association between newspaper advertisement and sales depends on how much is spent on TV commercials (and v.v.)
- Model with an interaction term:

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

such that the association between Y and X_1 depends on the level of X_2:

$$E\{Y|X_2 = x_2\} = (\beta_0 + \underbrace{\beta_2 \cdot x_2}_{\textit{fixed}}) + (\beta_1 + \underbrace{\beta_3 \cdot x_2}_{\textit{fixed}})X_1,$$

and v.v

$$E\{Y|X_1 = x_1\} = (\beta_0 + \beta_1 \cdot x_1) + (\beta_2 + \beta_3 \cdot x_1)X_2$$

Interaction terms: example

- Suppose

$$E\{Y\} = 10 + 2X_1 + 5X_2 + \beta_3 X_1 X_2$$

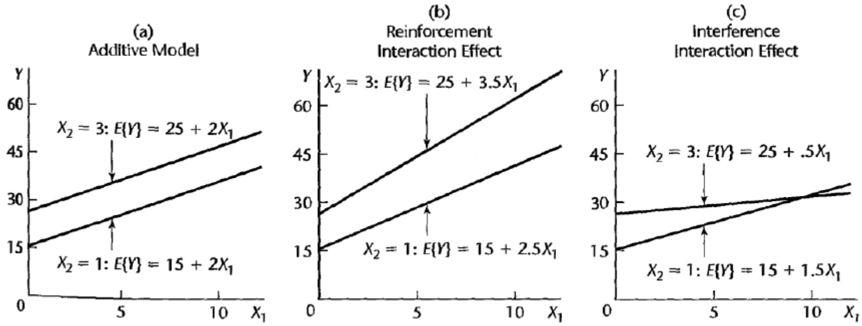
draw $E\{Y\}$ as a function of X_1, for outcome $X_2 = 1$ and $X_2 = 3$ for:

1. $\beta_3 = 0$ (additive model, no interaction)
2. $\beta_3 = 0.5$ (reinforcement interaction)
3. $\beta_3 = -0.5$ (interference interaction)

- Note that

$$\begin{aligned} E\{Y|X_2 = x_2\} &= (\beta_0 + \beta_2 \cdot x_2) + (\beta_1 + \beta_3 \cdot x_2)X_1 \\ &= (10 + 5 \cdot x_2) + (2 + \beta_3 \cdot x_2)X_1 \end{aligned}$$

Illustration of reinforcement and interference interaction effects



$$
\begin{aligned}
E\{Y|X_2 = x_2\} &= (\beta_0 + \beta_2 \cdot x_2) + (\beta_1 + \beta_3 \cdot x_2)X_1 \\
&= (10 + 5 \cdot x_2) + (2 + \beta_3 \cdot x_2)X_1
\end{aligned}
$$

Interaction terms: interpretation

- Model for two pred. variables with interaction:

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

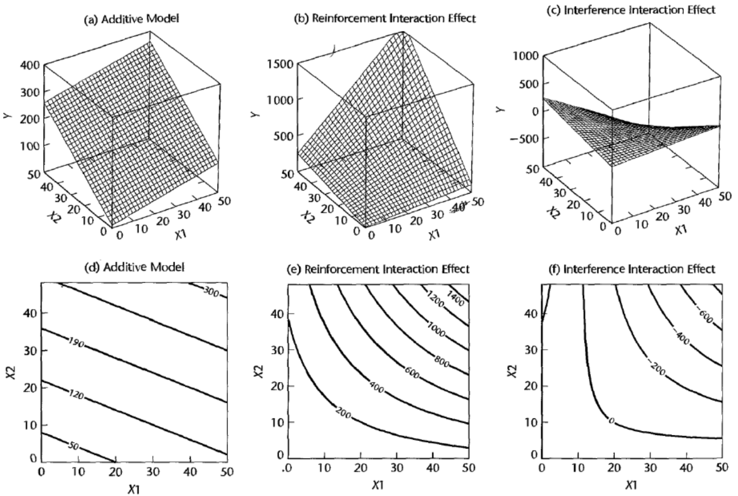
- Interpretation of the parameters:

$$\frac{\partial E\{Y\}}{\partial X_1} = \beta_1 + \beta_3 X_2,$$

  thus one unit change in X_1, at a certain lvel for X_2, is associated with $\beta_1 + X_2\beta_3$ change in $E\{Y\}$

- Note that interacting predictors are different from correlated predictors:
  X_1 and X_2 can interact *only if/only if not/whether or not* X_1 and X_2 are correlated?

Response surfaces and contour for additive and interaction regression models

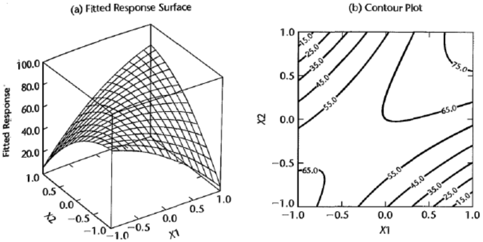


Interaction terms: models with curvilinear effects

- Model for two pred. variables with curvilinear terms and interactions:

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{12} X_1 X_2$$

- Lack of parallelism in the contour curves



Qualitative variables

- Qualitative variable = categorical variable, e.g., male/female, region (A, B, or C)
- include a qualitative variable in a regression model
  - to examine its effect on/association with Y
  - to better predict Y
  - to get more accurate estimates of the effects of other predictor variables
- How to include a qualitative variable in regression model?
  - example: model mean income (Y) by region (A,B, or C)
  - We can use "dummy variables" X_1 and X_2:

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

$$X_1 = \begin{cases} 1, & \text{region A} \\ 0, & \text{otherwise} \end{cases}, \quad X_2 = \begin{cases} 1, & \text{region B} \\ 0, & \text{otherwise} \end{cases}$$

Qualitative variables

- Interpretation?

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

$$X_1 = \begin{cases} 1, & \text{region A} \\ 0, & \text{otherwise} \end{cases}, \quad X_2 = \begin{cases} 1, & \text{region B} \\ 0, & \text{otherwise} \end{cases}$$

  - β_0 is the mean response in the left-out category, thus in region C
  - β_1 is the difference in the mean response between A and C
  - β_2 is the difference in the mean response between region B and C
  - Or: $E\{Y\}$ in region $A = \beta_0 + \beta_1$ etc

Qualitative variables

- Note that we used $c-1$ dummie variables to model a qual. variables with c categories:
$$E\{Y\} = \beta_0 \quad + \quad \beta_1 X_1 + \beta_2 X_2$$
$$X_1 = \begin{cases} 1, & \text{region } A \\ 0, & \text{otherwise} \end{cases} \quad , \quad X_2 = \begin{cases} 1, & \text{region } B \\ 0, & \text{otherwise} \end{cases}$$
- Why not include c dummie variables, adding $X_3 = \begin{cases} 1, & \text{region } C \\ 0, & \text{otherwise} \end{cases}$
$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$
    - Then $X_1 + X_2 + X_3 = 1$ for each observation: perfect collinearity between the intercept and the predictors (unlimited set of estimates $b_0 + conts, b_1 - const, b_2 - const, b_3 - const$)
    - Dropping the intercept would work

Qualitative variables... "Why not" continued

- Why not include a categorical variables as a quantitative predictor? For example, region as a quantitative predictor X_1 with outcome $1, 2, 3$ for region A, B, and C:

$$E\{Y\} = \beta_0 + \beta_1 X_1$$

- This implies differences between the means in the different categories depends on the value of coding which doesn't have to hold true
- V.v., Quantitative predictors are sometimes "recoded" into categorical variables
  - example: age groups

Qualitative variables AND quantitative variables

- Example: model personal income (Y) with years of education (X_1), and gender (male/female, use $X_2 = 1$ for males):

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

- Then

$$E\{Y\} = \begin{cases} (\beta_0 + \beta_2) + \beta_1 X_1, & \text{for males} \\ \beta_0 + \beta_1 X_1, & \text{for females} \end{cases}$$

- Interpretation:
  - $\beta_2 = E\{Y|X_2 = 1\} - E\{Y|X_2 = 0\}$;
    the average difference in income between men and women
  - the slope β_1 is the same for men and women
- Draw $E\{Y\}$ as a function of X_1, for men and women

Example: data and indicator coding

- An insurance innovation example

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

where X_{i1} =size of firm, $X_{12} = 1$ if stock company or 0 if mutual company

Firm i	(1) Number of Months Elapsed Y_i	(2) Size of Firm (million dollars) X_{i1}	(3) Type of Firm	(4) Indicator Code X_{i2}	(5) $X_{i1}X_{i2}$
1	17	151	Mutual	0	0
2	26	92	Mutual	0	0
3	21	175	Mutual	0	0
4	30	31	Mutual	0	0
5	22	104	Mutual	0	0
6	0	277	Mutual	0	0
7	12	210	Mutual	0	0
8	19	120	Mutual	0	0
9	4	290	Mutual	0	0
10	16	238	Mutual	0	0
11	28	164	Stock	1	164
12	15	272	Stock	1	272
13	11	295	Stock	1	295
14	38	68	Stock	1	68
15	31	85	Stock	1	85
16	21	224	Stock	1	224
17	20	166	Stock	1	166
18	13	305	Stock	1	305
19	30	124	Stock	1	124
20	14	246	Stock	1	246

Example: data and indicator coding-continued

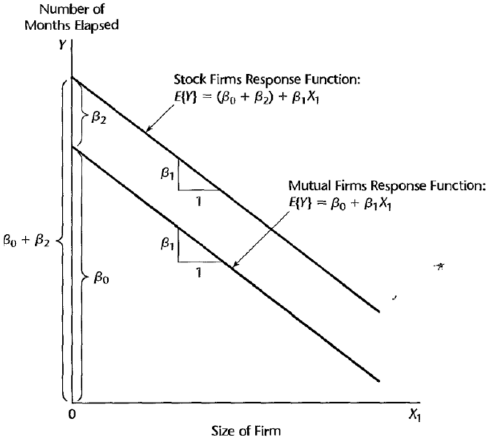
(a) Regression Coefficients

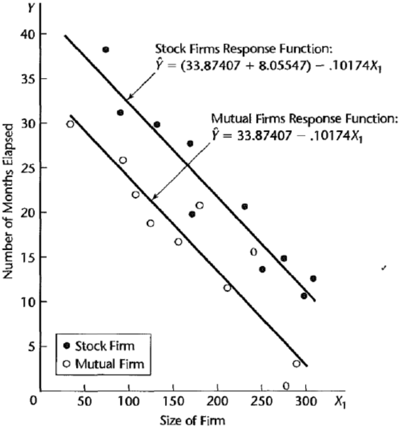
Regression Coefficient	Estimated Regression Coefficient	Estimated Standard Deviation	t^*
β_0	33.87407	1.81386	18.68
β_1	−.10174	.00889	−11.44
β_2	8.05547	1.45911	5.52

(b) Analysis of Variance

Source of Variation	SS	df	MS
Regression	1,504.41	2	752.20
Error	176.39	17	10.38
Total	1,680.80	19	

Example: data and indicator coding-continued





Example: indicator coding of more than two classes

- A tool wear example

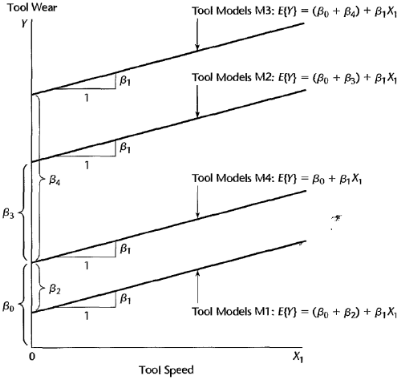
$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \epsilon_i$$

X_1: tool speed
$X_2 = 1$ if tool model M_1 or 0 otherwise,
$X_3 = 1$ if tool model M_2 or 0 otherwise,
$X_4 = 1$ if tool model M_3 or 0 otherwise.

Tool Model	X_1	X_2	X_3	X_4
M1	X_{i1}	1	0	0
M2	X_{i1}	0	1	0
M3	X_{i1}	0	0	1
M4	X_{i1}	0	0	0

Example: indicator coding of more than two classes–continued

- An arrangement of the response functions



Interaction between quantitative and qualitative variable

- What if the effect of education on income is stronger for women?
- Add an interaction term to the model:

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

- Then

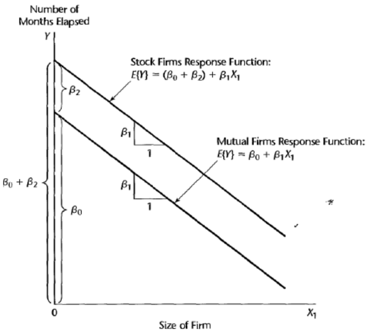
$$E\{Y\} = \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1, & \text{if } X_2 = 1 \\ \beta_0 + \beta_1 X_1, & \text{if } X_2 = 0 \end{cases}$$

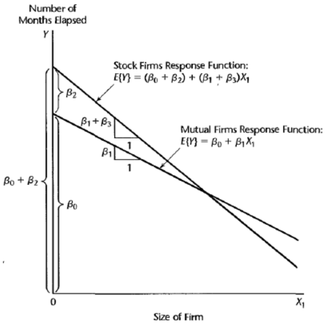
- Interpretation:
    - β_2 is the difference in intercept between men and women
    - β_3 is difference in slope between men and women; the difference in mean income, when education differs by 1 unit, is β_3 higher for men compared to women
- Draw $E\{Y\}$ as a function of X_1, for men and women

Interaction between quantitative and qualitative variable

- The insurance innovation example (left) + interaction term $\beta_3 X_{i1} X_{i2}$
- The model becomes

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 (right)$$





Interaction between quantitative and qualitative variable-continued

- Test whether $\beta_3 = 0$ (how?)
- Test whether $\beta_2 = \beta_3 = 0$ (how?)

(a) Regression Coefficients

Regression Coefficient	Estimated Regression Coefficient	Estimated Standard Deviation	t^*
β_0	33.83837	2.44065	13.86
β_1	−.10153	.01305	−7.78
β_2	8.13125	3.65405	2.23
β_3	−.0004171	.01833	−.02

(b) Analysis of Variance

Source of Variation	SS	df	MS
Regression	1,504.42	3	501.47
Error	176.38	16	11.02
Total	1,680.80	19	

Why not fit separate models for the different groups?

- The estimated regression functions will be the same if we fit separate models
- Disadvantage of separate models:
  - If it is reasonable to assume that the error variance is the same for men and women, it is more efficient to use all the data to estimate the parameters which implies more observations (degree of freedom) to estimate the variance parameter
  - Easy to do various test to examine if intercepts and/or slopes are different between groups:
    E.g. test the effect of gender and education on income

Quick questions

For income (Y) versus education (X_1) and gender ($X_2 = 1$ for males):

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

1. Does the level of income and/or the association between income and education differ between men and women?
2. Does the association between income and education differ between men and women?
3. If the association between income and education is the same between men and women, is there a difference in the level of mean income?

Quick questions

Use $F-$test to answer these three questions:

1. Does the level of income and/or the association between income and education differ between men and women?
   $H_0 : E\{Y\} = \beta_0 + \beta_1 X_1$ versus
   $H_a : E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$
2. Does the association between income and education differ between men and women?
   $H_0 : E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_1 X_2$ versus
   $H_a : E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$
3. If the association between income and education is the same between men and women, is there a difference in the level of mean income? $H_0 : E\{Y\} = \beta_0 + \beta_1 X_1$ versus
   $H_a : E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
   The F-statistics are given by the extra sum of squares with the full model under H_a and the reduced model under H_0

GPA example

- Response variable is GPA (Y), predictor variable are ACT test score (X_1) and concentration chosen ($X_2 = 1$ if yes)
- Fit model:

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

```
------------------------------------------------------
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.226318   0.549428   5.872 4.18e-08 ***
X1          -0.002757   0.021405  -0.129   0.8977
X2          -1.649577   0.672197  -2.454   0.0156 *
X1:X2        0.062245   0.026487   2.350   0.0205 *
-------------------------------------------------------
Residual standard error: 0.6124 on 116 degrees of fr.
Multiple R-squared: 0.1194, Adjusted R-squared: 0.09664
F-statistic: 5.244 on 3 and 116 DF,  p-value: 0.001982
```

F-test

- F-test for $H_0 : E\{Y\} = \beta_0 + \beta_1 X_1$ versus full model $H_a : E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$

```
-----------------------------------------------------------
> mod_reduced = lm(Y ~ X1)
> mod_full = lm(Y ~ X1+X2+X1*X2)
> anova(mod_reduced, mod_full)
-----------------------------------------------------------
Model 1: Y ~ X1
Model 2: Y ~ X1 + X2 + X1 * X2
  Res.Df    RSS  Df Sum of Sq      F  Pr(>F)
1    118 45.818
2    116 43.506   2     2.312 3.0822 0.04963 *
-----------------------------------------------------------
```

- Creating dummy variables by hand:
    - $D1 = (X2 == \text{"}male\text{"})$
    - lm(Y $\sim$ X1+D1)
- Let R do things automatically:
    - mod=lm(Y $\sim$ X1 + factor(X2))
- The use of "factor()":
    - factor() is not needed if the categorical variable is already coded in words
    - but it is essential if the categories are coded numerically
    - to be safe, you can always use "factor"

Estimated mean GPA for the two groups