# Multivariable Visualization

Tools:

- Scatterplot Array

- Rotating 3-D Plots

Let's try these out. Each of the data sets sasdata.eg10_2a sasdata.eg10_2b, sasdata.eg10_2c and sasdata.eg10_2d contains data generated by one of four models shown on the next page. Using only the display of the data set itself and a scatterplot array, you are to tell which data set was generated by which model.

The models are:

1. $Y = -1 + 7x_1 + 6x_2 - 3x_1^2 + 2x_2^2 + 7x_1x_2 + \epsilon,$
2. $Y = 5 + 7x_1 + 6x_2 - 3x_1^2 + 2x_2^2 + \epsilon,$
3. $Y = 5 + 7x_1 + 6x_2 - 3x_1^2 + 2x_2^2 + 7x_1x_2 + \epsilon,$
4. $Y = -1 + 7x_1 + 6x_2 - 3x_1^2 + 2x_2^2 + \epsilon,$

where $\epsilon \sim N(0,1)$. Be sure to write down your answers.

Now use the rotating 3-D plot to view the data. Does this change your guesses?

# The MLR Model

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1(Z_1, Z_2, \ldots, Z_p) + \beta_2 X_2(Z_1, Z_2, \ldots, Z_p) \\ &\quad + \ldots + \beta_q X_q(Z_1, Z_2, \ldots, Z_p) + \epsilon, \end{aligned}$$

where the $Z$s are the predictor variables and $\epsilon$ is a random error. Examples are

$$\begin{aligned} Y &= \beta_0 + \beta_1 Z_1 + \beta_2 Z_1^2 + \epsilon, \\ Y &= \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_1^2 + \beta_4 Z_1 Z_2 + \beta_5 Z_2^2 + \epsilon, \\ Y &= \beta_0 + \beta_1 \log(Z_2) + \beta_3 \sqrt{Z_1 Z_2} + \epsilon. \end{aligned}$$

We will write these models generically as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_q X_q + \epsilon.$$

**Fitting the MLR Model** As we did for SLR model, we use least squares to fit the MLR model. This means finding estmators of the model parameters $\beta_0, \beta_1, \ldots, \beta_q$ and $\sigma^2$. The LSEs of the $\beta$s are those values, of $b_0, b_1, \ldots, b_q$, denoted $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_q$, which minimize

$$\text{SSE}(b_0, b_1, \ldots, b_q) =$$

$$\sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_{i1} + b_2 X_{i2} + \cdots + b_q X_{iq})]^2.$$

The **fitted values** are

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_{i1} + \widehat{\beta}_2 X_{i2} + \cdots + \widehat{\beta}_q X_{iq},$$

and the residuals are

$$e_i = Y_i - \widehat{Y}_i.$$

Let's see what happens when we fit models to sasdata.eg10_2a and sasdata.eg10_2c.

**Assessing Model Fit** Residuals and studentized residuals are the primary tools to analyze model fit. We look for outliers and other deviations from model assumptions. Let's look at the residuals from some fits to sasdata.eg10_2c.

# Interpretation of the Fitted Model

The intercept has the interpretation "expected response when the $X_i$ all equal 0". The coefficient $\widehat{\beta}_i$ is interpreted as the change in expected response per unit change in $X_i$ when the other $X$s are held fixed (if that is possible).

Otherwise can interpret the model using multivariate calculus: change in expected response per unit change in $Z_i$ (with the other predictors held fixed) is

$$\frac{\partial}{\partial Z_i}(\widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \ldots + \widehat{\beta}_q X_q).$$

So, for example, if the fitted model is

$$\widehat{\beta}_0 + \widehat{\beta}_1 Z_1 + \widehat{\beta}_2 Z_2 + \widehat{\beta}_3 Z_1 Z_2,$$

$$\frac{\partial}{\partial Z_1}(\widehat{\beta}_0 + \widehat{\beta}_1 Z_1 + \widehat{\beta}_2 Z_2 + \widehat{\beta}_3 Z_1 Z_2) = \widehat{\beta}_1 + \widehat{\beta}_3 Z_2.$$

# Theory-Based Modeling

Two ways of building models:

- Empirical modeling
- Theoretical modeling

# Comparison of Fitted Models

- Residual analysis
- Principle of parsimony (simplicity of description)
- Coefficient of multiple determination, and its adjusted cousin.

# ANOVA

Idea:

- Total variation in the response (about its mean) is measured by

$$\text{SSTO} = \sum_{i=1}^{n} (Y_i - \overline{Y})^2.$$

  This is the variation or uncertainty of prediciton if no predictor variables are used.

- SSTO can be broken down into two pieces: SSR, the regression sum of squares, and SSE, the error sum of squares, so that SSTO=SSR+SSE.

- SSE $= \sum_i^n e_i^2$ is the total sum of the squared residuals. It measures the variation of the response unaccounted for by the fitted model or the uncertainty of predicting the response using the fitted model.

- SSR $= SSTO - SSR$ is the variability explained by the fitted model or the reduction in uncertainty of prediction due to using the fitted model.

**Degrees of Freedom** The degrees of freedom for a SS is the number of independent pieces of data making up the SS. For SSTO, SSE and SSR the degrees of freedom are $n-1$, $n-q-1$ and $q$. These add just as the SSs do. A SS divided by its degrees of freedom is called a Mean Square.

**The ANOVA Table** This is a table which summarizes the SSs, degrees of freedom and mean squares.

| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| Source | DF | SS | MS | F Stat | Prob > F |
| Model | $q$ | SSR | MSR | F=MSR/MSE | $p$-value |
| Error | $n-q-1$ | SSE | MSE | | |
| C Total | $n-1$ | SSTO | | | |

# Inference for the MLR Model: The F Test

- **The Hypotheses:**

$$\begin{aligned} H_0: &\quad \beta_1 = \beta_2 = \cdots = \beta_q = 0 \\ H_a: &\quad \text{Not } H_0 \end{aligned}$$

- **The Test Statistic:** F=MSR/MSE

- **The P-Value:** $P(F_{q,n-q-1} > F^*)$, where $F_{q,n-q-1}$ is a random variable from an $F_{q,n-q-1}$ distribution and $F^*$ is the observed value of the test statistic.

# T Tests for Individual Predictors

- **The Hypotheses:**

$$\begin{aligned} H_0: &\quad \beta_i = 0 \\ H_a: &\quad \beta_i \neq 0 \end{aligned}$$

- **The Test Statistic:** $t = \frac{\widehat{\beta}_i}{\widehat{\sigma}(\widehat{\beta}_i)}$

- **The P-Value:** $P(|t_{n-q-1}| > |t^*|)$, where $t_{n-q-1}$ is a random variable from a $t_{n-q-1}$ distribution and $t^*$ is the observed value of the test statistic.

# Summary of Intervals for MLR Model

- **Confidence Interval for Model Coefficients:** A level $L$ confidence interval for $\beta_i$ is

  $$(\widehat{\beta}_i - \widehat{\sigma}(\widehat{\beta}_i)t_{n-q-1,(1+L)/2}, \widehat{\beta}_i + \widehat{\sigma}(\widehat{\beta}_i)t_{n-q-1,(1+L)/2}).$$

- **Confidence Interval for Mean Response:** A level $L$ confidence interval for the mean response at at predictor values $X_{10}, X_{20}, \ldots, X_{q0}$ is

  $$(\widehat{Y}_0 - \widehat{\sigma}(\widehat{Y}_0)t_{n-q-1,(1+L)/2}, \widehat{Y}_0 + \widehat{\sigma}(\widehat{Y}_0)t_{n-q-1,(1+L)/2}),$$

  where

  $$\widehat{Y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 X_{10} + \cdots + \widehat{\beta}_q X_{q0},$$

  and $\widehat{\sigma}(\widehat{Y}_0)$ is the estimated standard error of the response.

- **Prediction Interval for a Future Observation:** A level $L$ prediction interval for a new response at predictor values $X_{10}, X_{20}, \ldots, X_{q0}$ is

$$(\hat{Y}_{new} - \hat{\sigma}(Y_{new} - \hat{Y}_{new})t_{n-q-1,(1+L)/2},$$

$$\hat{Y}_{new} + \hat{\sigma}(Y_{new} - \hat{Y}_{new})t_{n-q-1,(1+L)/2}),$$

where

$$\hat{Y}_{new} = \hat{\beta}_0 + \hat{\beta}_1 X_{10} + \cdots + \hat{\beta}_q X_{q0},$$

and

$$\hat{\sigma}(Y_{new} - \hat{Y}_{new}) = \sqrt{\text{MSE} + \hat{\sigma}^2(\hat{Y}_0)}.$$

Multicollinearity Multicollinearity is correlation among the predictors.

- **Consequences**
  - o Large sampling variability for $\widehat{\beta}_i$
  - o Questionable interpretation of $\widehat{\beta}_i$ as change in expected response per unit change in $X_i$.

- **Detection** $R_i^2$, the coefficient of multiple determination obtained from regressing $X_i$ on the other $X$s, is a measure of how highly correlated $X_i$ is with the other $X$s. This leads to two related measures of multicollinearity.

  - **Tolerance** $\text{TOL}_i = 1 - R_i^2$ Small $\text{TOL}_i$ indicates $X_i$ is highly correlated with other $X$s. We should begin getting concerned if $\text{TOL}_i < 0.1$.

  - **VIF** VIF stands for variance inflation factor. $\text{VIF}_i = 1/\text{TOL}_i$. Large $\text{VIF}_i$ indicates $X_i$ is highly correlated with other $X$s. We should begin getting concerned if $\text{VIF}_i > 10$.

- **Remedial Measures**

  - Center the $X_i$ (or sometimes the $Z_i$)

  - Drop offending $X_i$

Empirical Model Building Selection of variables in empirical model building is an important task. We consider only one of many possible methods: **backward elimination**, which consists of starting with all possible $X_i$ in the model and eliminating the non-significant ones one at at time, until we are satisfied with the remaining model.