STAT 525

Chapter 6
Multiple Regression

Dr. Qifan Song

The Data and Model

- Still have single response variable Y
- Now have multiple explanatory variables
- Examples:
  - Blood Pressure vs Age, Weight, Diet, Fitness Level
  - Traffic Count vs Time, Location, Population, Month
- Goal: There is a total amount of variation in Y (SSTO). We want to explain as much of this variation as possible using a linear model and our explanatory variables

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

- Have $p-1$ predictors $\longrightarrow p$ coefficients

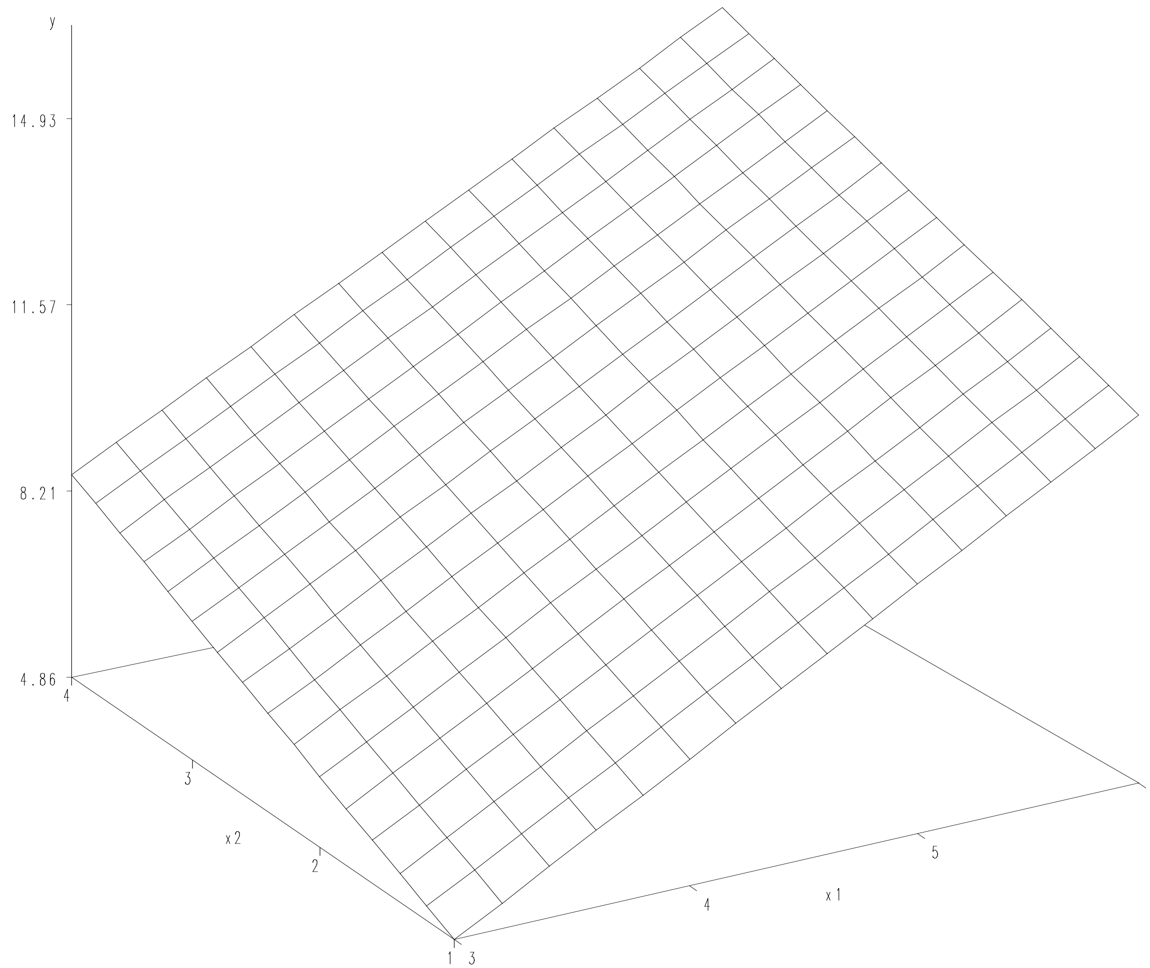
First Order Model with Two Predictors

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i; \qquad i = 1, \ldots, n$$

- β_0 is the intercept and β_1 and β_2 are the regression coefficients
- Meaning of regression coefficients
  - β_1 describes change in mean response per unit increase in X_1 when X_2 is held constant
  - β_2 describes change in mean response per unit increase in X_2 when X_1 is held constant
- Variables X_1 and X_2 are **additive**. Value of X_1 does not affect the change due to X_2. There is no **interaction**.
- The response surface is a plane.

Additive Response Surface

$$\widehat{Y}_i = -2.79 + 2.14X_{i1} + 1.21X_{i2}$$



Interaction Model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i$$

- Meaning of parameters:
  - Change in X_1 when $X_2 = x_2$

$$\begin{aligned}\Delta E[Y] &= \{\beta_0 + \beta_1(X_1+1) + \beta_2 x_2 + \beta_3(X_1+1)x_2\} \\ &\quad -\{\beta_0 + \beta_1 X_1 + \beta_2 x_2 + \beta_3 X_1 x_2\} \\ &= \beta_1 + \beta_3 x_2\end{aligned}$$

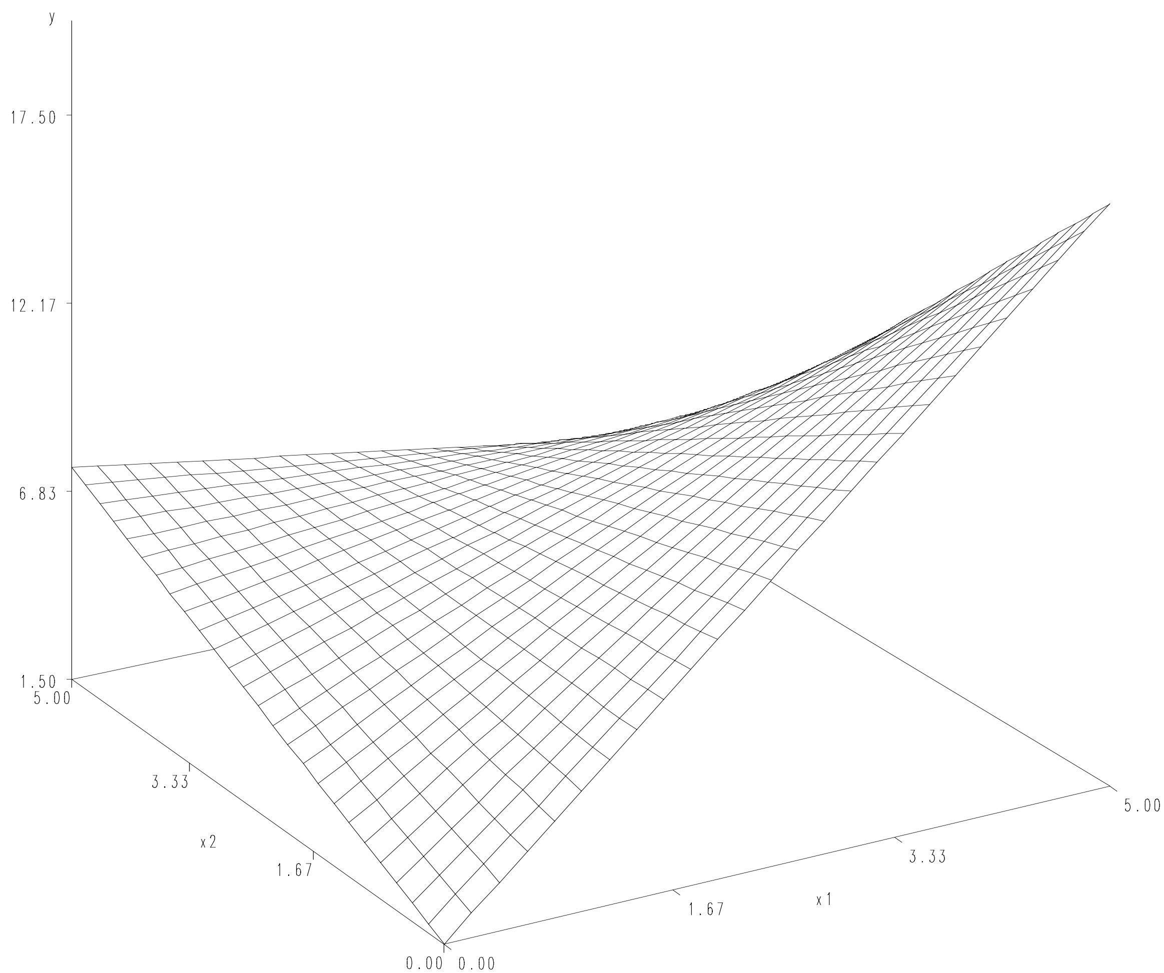
  - Change in X_2 when $X_1 = x_1$

$$\Delta E[Y] = \beta_2 + \beta_3 x_1$$

- Rate of change due to one variable affected by the other

Interaction Response Surface

$$\widehat{Y}_i = 1.5 + 3.2X_{i1} + 1.2X_{i2} - .75X_{i1}X_{i2}$$



Qualitative Predictors

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i$$

where Y is a senior student's GPA, X_1 is the SAT score.

- Let $X_2 = 1$ if case from Purdue, and $X_2 = 0$ if from IU
- Meaning of parameters:
  - Case from Purdue ($X_2 = 1$):

$$\begin{aligned} E[Y] &= \beta_0 + \beta_1 X_1 + \beta_2 1 + \beta_3 X_1(1) \\ &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1 \end{aligned}$$

  - Case from other location ($X_2 = 0$)

$$E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 0 + \beta_3 X_1(0) = \beta_0 + \beta_1 X_1$$

- Have <u>two</u> regression lines
  - β_2 quantify the difference between intercepts
  - β_3 quantify the difference between slops

Polynomial Regression and Transformations

- Polynomial regression:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i \\ &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \end{aligned}$$

  where $X_{i2} = X_i^2$.

  – this is a linear model because it is a linear function of parameters β

- Transformations

$$\begin{aligned} Y_i &= \frac{1}{\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i} \\ \Longleftrightarrow \frac{1}{Y_i} &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \\ log(Y_i) &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \end{aligned}$$

  – the last one is a linear model on the $log(Y_i)$ scale

General Linear Regression In Matrix Terms

- After transformation and re-organization, a linear model ("linear" w.r.t. unknown coefficient, not to actual predictors) is obtained

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

- As an array

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1\ p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2\ p-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n\ p-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

- In matrix notation

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

- Distributional assumptions:

$$\boldsymbol{\varepsilon} \sim N(0, \sigma^2 I) \longrightarrow \mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I)$$

Estimation of Regression Coefficients

- Least squares estimates
  - find $\mathbf{b}$ to minimize $(\mathbf{Y}-\mathbf{X}\mathbf{b})'(\mathbf{Y}-\mathbf{X}\mathbf{b})$
  - $\mathbf{b}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$
- Fitted values define a (hyper)plane
  - $\hat{\mathbf{Y}}=\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}=\mathbf{H}\mathbf{Y}$
  - $\mathbf{H}\mathbf{Y}$ forms a response surface
- Residuals
  - $\mathbf{e}=\mathbf{Y}-\hat{\mathbf{Y}}=(\mathbf{I}-\mathbf{H})\mathbf{Y}$

The Distribution of Residuals

- $\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$
  - $\mathbf{I} - \mathbf{H}$ is symmetric and idempotent
- Expected value $E(\mathbf{e}) = \mathbf{0}$
- Covariance Matrix

$$
\begin{aligned}
\boldsymbol{\sigma}^2(\mathbf{e}) &= \sigma^2(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})' \\
&= \sigma^2(\mathbf{I} - \mathbf{H})
\end{aligned}
$$

  - $\mathrm{Var}(e_i) = \sigma^2(1 - h_{ii})$ where $h_{ii} = \mathbf{X}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_i$
  - Residuals are usually correlated, i.e., $cov(e_i, e_j) = -\sigma^2 h_{ij}$, $i \neq j$
- Will use this information to look for outliers

Estimation of σ^2

- Similar approach as before
- Estimate it from $\mathbf{e}$, since $\mathbf{e}$ has nothing to do with β_i's.
- Now p model parameters

$$
\begin{aligned}
s^2 &= \frac{\mathbf{e}'\mathbf{e}}{n-p} \\
&= \frac{(\mathbf{Y}-\mathbf{X}\mathbf{b})'(\mathbf{Y}-\mathbf{X}\mathbf{b})}{n-p} \\
&= \frac{\text{SSE}}{n-p} \\
&= \text{MSE}
\end{aligned}
$$

ANOVA TABLE

Source of Variation	df	SS	MS	F Value
Regression (Model)	$p-1$	SSR	MSR=SSR/$(p-1)$	MSR/MSE
Error	$n-p$	SSE	MSE=SSE/$(n-p)$	
Total	$n-1$	SSTO		

- F Test: Tests if the predictors *collectively* help explain the variation in Y
  - $H_0 : \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$
  - H_a : at least one $\beta_k \neq 0$, $1 \leq k \leq p-1$
  - $F^* = \frac{SSR/(p-1)}{SSE/(n-p)} \overset{H_0}{\sim} F(p-1, n-p)$
  - Reject H_0 if $F^* > F(1-\alpha, p-1, n-p)$
- No conclusions possible regarding individual predictors

Testing Individual Predictor

t-Test

- Have already shown that $\mathbf{b} \sim N\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right)$
  - This implies $b_k \sim N(\beta_k, \sigma^2(b_k))$
- Perform t test
  - $H_0: \ \beta_k = 0$ vs $H_a: \beta_k \neq 0$
  - $\frac{b_k - \beta_k}{s(b_k)} \ \sim \ t_{n-p}$ so $t^* = \frac{b_k}{s(b_k)} \ \sim \ t_{n-p}$ under H_0
  - Reject H_0 if $|t^*| \ > \ t(1 - \alpha/2, n-p)$
- Confidence interval for β_k
  - $b_k \pm t(1 - \alpha/2, n-p)s\{b_k\}$

General Linear Test

- $H_0: \ \beta_k = 0$ vs $H_a : \beta_k \neq 0$
  - Full Model :
$$Y_i = \beta_0 + \sum_{j=1}^{p-1} \beta_j X_{ji} + \varepsilon_i$$
  - Reduced Model :
$$Y_i = \beta_0 + \sum_{j=1}^{k-1} \beta_j X_{ji} + \sum_{j=k+1}^{p-1} \beta_j X_{ji} + \varepsilon_i$$
  - $F^\star = \frac{(\text{SSE(R)} - \text{SSE(F)})/1}{\text{SSE(F)}/(n-p)}$
  - Reject H_0 if $F^* > F(1-\alpha, 1, n-p)$

Equivalence of t-Test and General Linear Test

- Can show that $F^* = (t^*)^2$
  - both tests result in the same conclusion
- Both tests investigate significance of a predictor *given the other variables are already in the model*
  - i.e. significance of the variable which is fitted last

Coefficient of Multiple Determination

- Coefficient of Determination R^2 describes proportionate reduction in total variation associated with the **full set** of X variables

  $R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$, $0 \leq R^2 \leq 1$

- R^2 usually increases with the increasing p
  - Adjusted R_a^2 attempts to account for p

    $R_a^2 = 1 - \frac{SSE/(n-p)}{SSTO/(n-1)}$, $0 \leq R_a^2 \leq 1$
  - The adustment is often insufficient

Example: Purdue Computer Science Students

- Computer Science majors at Purdue have a large drop-out rate
- Goal: Find predictors of success (defined as high GPA)
- Predictors must be available at time of entry into program. These are:
  - GPA: grade points average after three semesters
  - HSM: high-school math grades
  - HSS: high-school science grades
  - HSE: high-school english grades
  - SATM: SAT Math
  - SATV: SAT Verbal
- Data available on $n = 224$ students

- Now investigate the model: GPA = HSM HSS HSE

```
options nocenter linesize=72;
goptions colors=('none');

data a1;
    infile 'U:\.www\datasets525\csdata.dat';
    input id gpa hsm hss hse satm satv;

proc reg data=a1;
    model gpa=hsm hss hse;
run;
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	27.71233	9.23744	18.86	<.0001
Error	220	107.75046	0.48977		
Corrected Total	223	135.46279			

Root MSE	0.69984	R-Square	0.2046
Dependent Mean	2.63522	Adj R-Sq	0.1937
Coeff Var	26.55711		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.58988	0.29424	2.00	0.0462
hsm	1	0.16857	0.03549	4.75	<.0001
hss	1	0.03432	0.03756	0.91	0.3619
hse	1	0.04510	0.03870	1.17	0.2451

Estimation of Mean Response $E(Y_h)$

- We are interested in predictors $\mathbf{X}_h$
  - Can show $\widehat{Y}_h \sim N\left(\mathbf{X}_h'\boldsymbol{\beta}, \sigma^2\mathbf{X}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h\right)$
- Individual CI for $\mathbf{X}_h$
  - $\widehat{Y}_h \ \pm \ t(1-\alpha/2, n-p)s\{\widehat{Y}_h\}$
- Bonferroni CI for g vectors $\mathbf{X}_h$
  - $\widehat{Y}_h \ \pm \ t(1-\alpha/(2g), n-p)s\{\widehat{Y}_h\}$
- Working-Hotelling confidence band for the whole regression line
  - $\widehat{Y}_h \ \pm \ \sqrt{pF(1-\alpha, p, n-p)}\ s\{\widehat{Y}_h\}$
- Be careful to be only in range of X's

Predict New Observation

- $Y_{h(new)} = E(Y_h) + \varepsilon$
  - $s^2(pred) = s^2(\hat{Y}_h) + \text{MSE}$
  - $\hat{Y}_h + \varepsilon \sim N\left(\mathbf{X}_h'\boldsymbol{\beta}, \sigma^2(1 + \mathbf{X}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h)\right)$
- Individual CI of $Y_{h(new)}$
  - $\hat{Y}_h \ \pm\ t(1-\alpha/2, n-p)s\{pred\}$
- Bonferroni CI for g vectors $\mathbf{X}_h$
  - $\hat{Y}_h \ \pm\ t(1-\alpha/(2g), n-p)s\{pred\}$
- Simultaneous Scheffé prediction limits for g vectors $\mathbf{X}_h$
  - $\hat{Y}_h \ \pm\ \sqrt{gF(1-\alpha, g, n-p)}\ s\{pred\}$

Diagnostics

- Diagnostics play a key role in both the **development and assessment** of multiple regression models
- Most previous diagnostics carry over to multiple regression
- Given more than one predictor, must also consider relationship between predictors
- Specialized diagnostics discussed later in Chapters 9 and 10

Scatterplot Matrix

- Scatterplot matrix organizes all bivariate scatterplot, between Y and X_j as well as between X_j and X_k $(j, k = 1, 2, ..., p-1)$, in a matrix.
  - Nature of bivariate relationships
  - Strength of bivariate relationships
  - Detection of outliers
  - Range spanned by X's

- Can be generated within SAS

```
proc sgscatter data=cs;
  matrix gpa hsm hss hse;
run;
```

Correlation Matrix

- Complementary summary
- Displays all numerical pairwise correlations
- Must be wary of
  - Nonlinear relationships
  - Outliers
  - Influential observations

Example: Purdue Computer Science Student

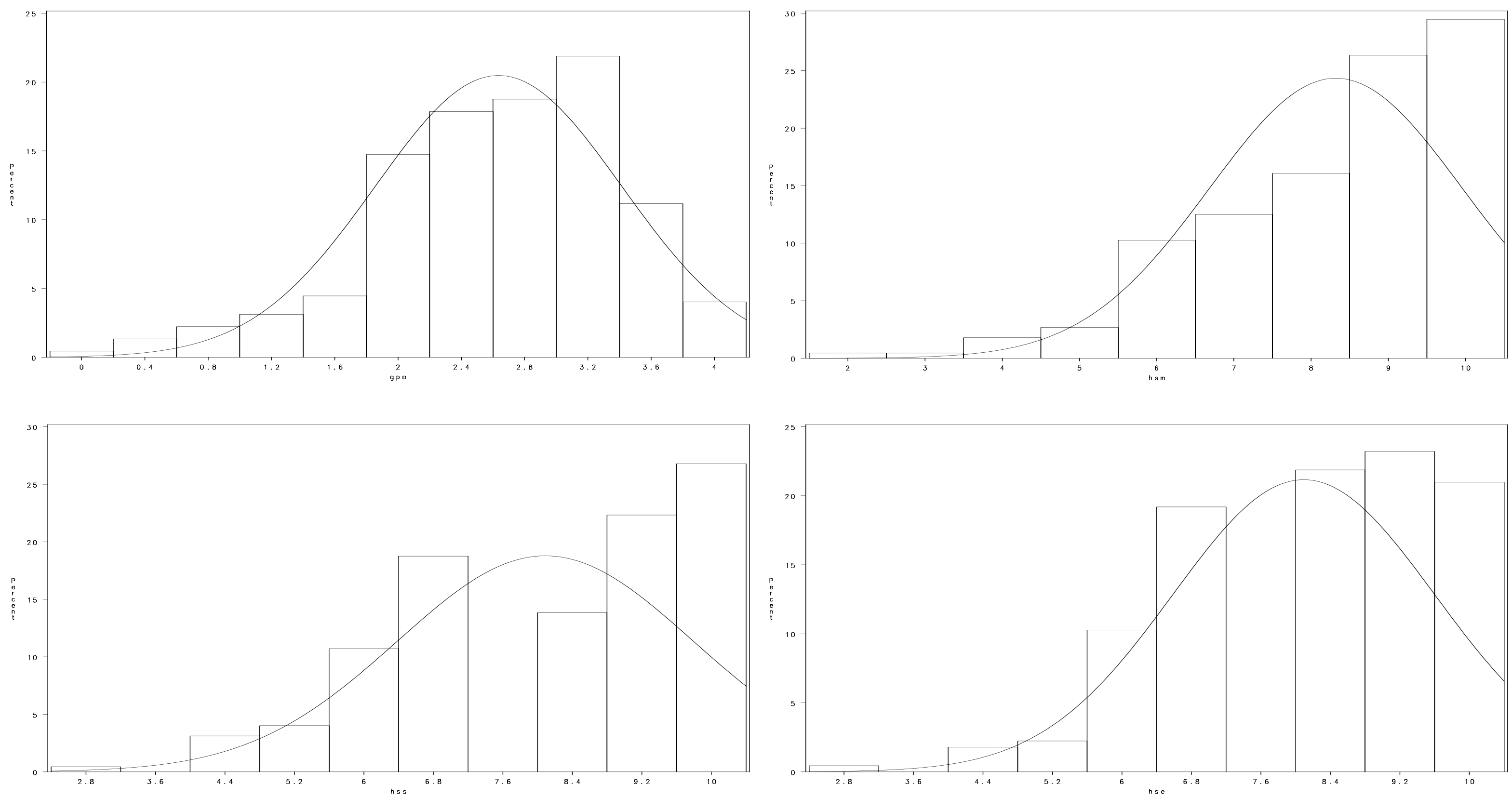
- Univaraite Descriptive Statistics (e.g., `PROC MEANS` or `PROC UNIVARIATE`): preliminary check for outliers/unusual observations.

```
proc means data=cs maxdec=2;
    var gpa hsm hss hse satm satv;
run;
```

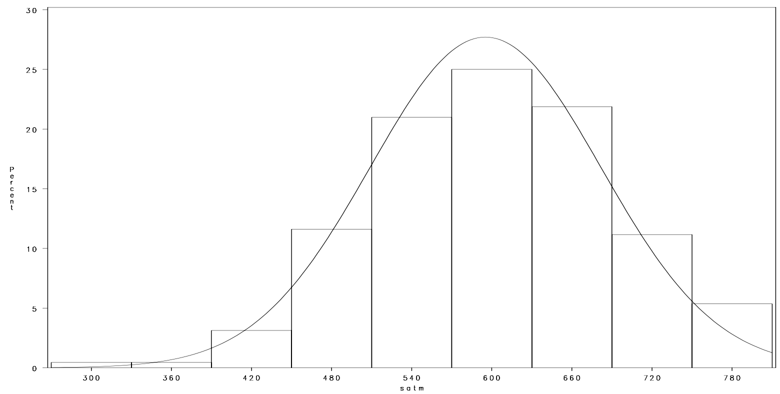
```
                     The MEANS Procedure
Variable    N         Mean      Std Dev      Minimum      Maximum
-------------------------------------------------------------
gpa       224         2.64         0.78         0.12         4.00
hsm       224         8.32         1.64         2.00        10.00
hss       224         8.09         1.70         3.00        10.00
hse       224         8.09         1.51         3.00        10.00
satm      224       595.29        86.40       300.00       800.00
satv      224       504.55        92.61       285.00       760.00
-------------------------------------------------------------
```

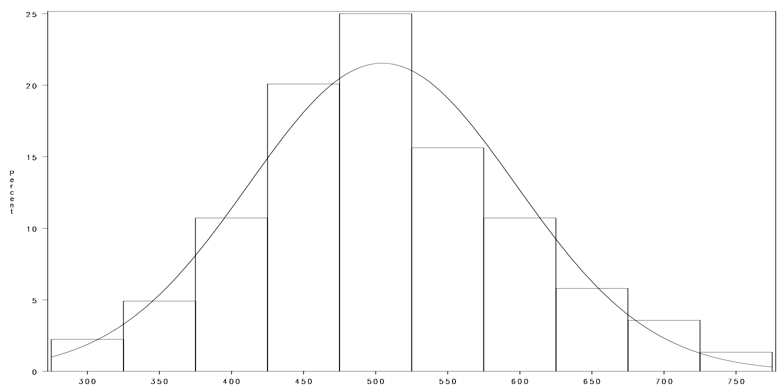
- `maxdec = 2` sets the number of decimal places in the output to 2.

```
proc univariate data=cs noprint;
    var gpa hsm hss hse satm satv;
    histogram gpa hsm hss hse satm satv /normal;
run;
```

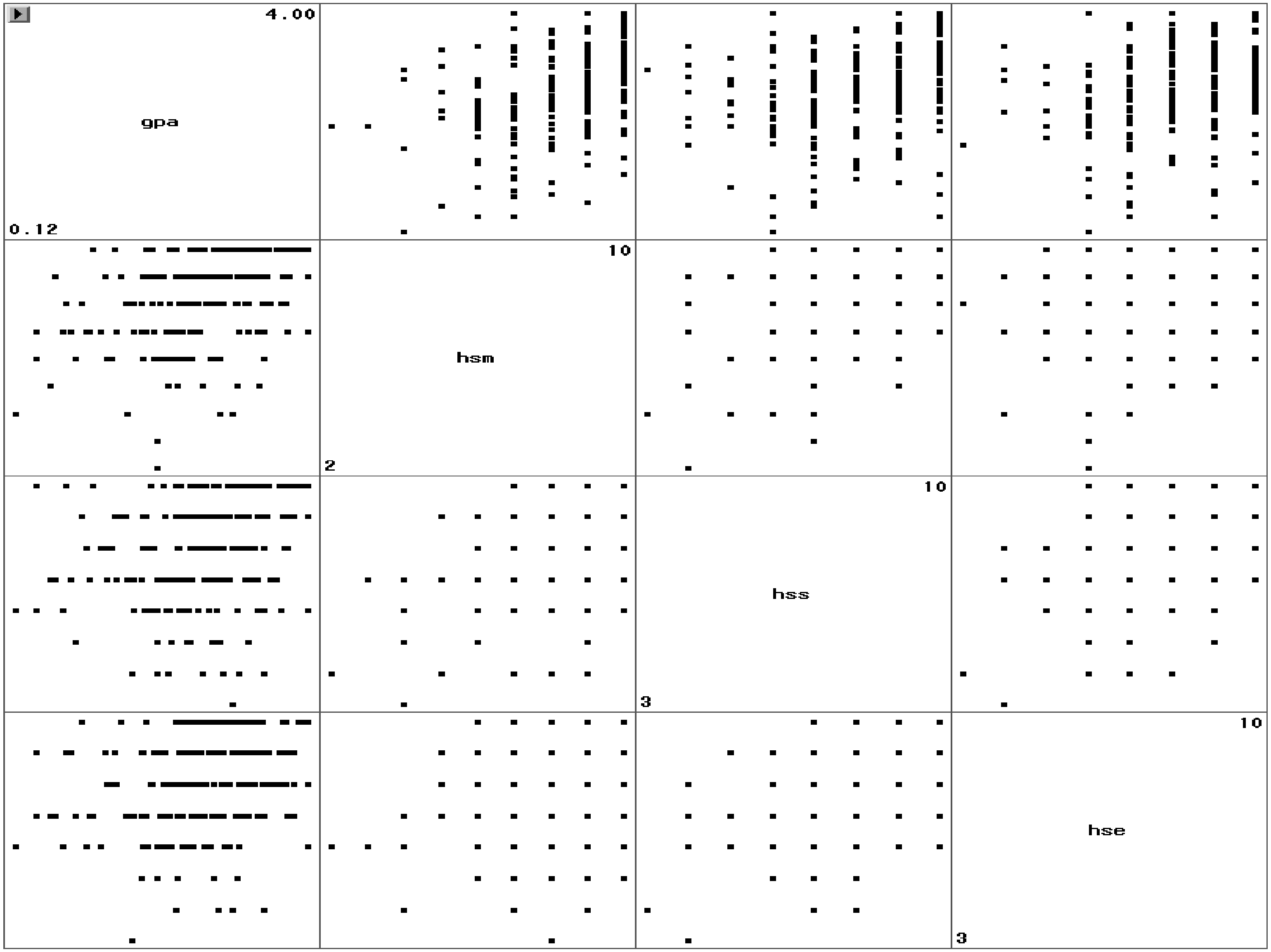


Top Left: GPA   Top Right: HSM   Bottom Left: HSS   Bottom Right: HSE





Upper – SATM Lower – SATV


4.00
gpa
0.12
10
hsm
2
10
hss
3
10
hse
3

- Uses `PROC CORR` to report correlation values;
- Use `NOPROB` statement to get rid of those p-values.

```
[1] proc corr data=a1;
    var hsm hss hse;

          hsm    hss    hse
    hsm  1.00   0.57    0.44
               <.0001  <.0001
    hss  0.57   1.00    0.57
       <.0001          <.0001
    hse  0.44   0.57   1.00
       <.0001 <.0001

[2] proc corr data=a1 noprob;
    var satm satv;

          satm     satv
    satm  1.00     0.46
    satv  0.46     1.00
```

```
[3] proc corr data=a1;
    var hsm hss hse satm satv;
    with gpa;

              hsm    hss     hse
     gpa   0.43   0.32    0.28
         <.0001 <.0001  <.0001

             satm    satv
    gpa    0.25    0.11
         0.0001  0.0873
```

Residual Plots

- Used for similar assessment of assumptions
  - Model is correct
  - Normality
  - Constant Variance
  - Independence
- Plot e vs $\hat{Y}$ (overall)
- Plot e vs X_j (with respect to X_j)
- Plot e vs missing variable (e.g., $X_j X_k$)

Tests

- Univariate graphical summaries of e are still preferred
- `NORMAL` option in `PROC UNIVARIATE` test normality
- Modified Levene's and Breusch-Pagan for constant variance
- Lack of fit test: need repeated observations where all X fixed at same levels

Lack of Fit Test

- Compare
  - (reduced) linear model

    $H_0 : E(Y_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1}$

  - (full) model where Y has c means (i.e. c combinations of X_i)

    $H_a : E(Y_i) \neq \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1}$

- $F^* = \frac{\{SSE(R)-SSE(F)\}/\{(n-p)-(n-c)\}}{SSE(F)/(n-c)} \sim F(c-p, n-c)$ under H_0

- Reject H_0 if $F^* > F(1-\alpha, c-p, n-c)$

- If reject H_0, conclude that a more complex relationship between Y and $X_1, \ldots, X_{p-1}$ is needed

Calculate SSE(F) in SAS

```
*  Analysis of Variance - Full Model
proc glm;
    class x1 x2;
    model y=x1*x2;
run;
```

- `CLASS` and `MODEL` specify that every combination of levels of X_1 and X_2 has their own mean
- Plug the SSE of this model into the lack of fit test

Example: Purdue Computer Science Student

- Investigate the model: GPA = HSM HSS HSE
  - `PROC REG` reports SSE(R) = 107.75046 with df(SSE) = 220.

```
proc glm data=a1;
    class hsm hss hse;
    model gpa=hsm*hss*hse;
run; quit;
```

```
         Class Level Information

Class         Levels    Values
hsm                9    2 3 4 5 6 7 8 9 10
hss                8    3 4 5 6 7 8 9 10
hse                8    3 4 5 6 7 8 9 10
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	99	85.0469697	0.8590603	2.11	<.0001
Error	124	50.4158191	0.4065792		
Corrected Total	223	135.4627888			

- $F^* = \frac{\{SSE(R)-SSE(F)\}/\{(n-p)-(n-c)\}}{SSE(F)/(n-c)} = \frac{\{107.75046-50.4158191\}/\{220-124\}}{50.4158191/124} = 1.4689 > 1.3688 = F^{-1}_{96,124}(0.95)$.

Chapter Review

- Data and Notation
- Model in Matrix Terms
- Parameter Estimation
- ANOVA F-test
- Estimation of Mean Responses
- Prediction of New Observations
- Diagnostics