Diagnostics

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

Scatterplots

- Scatterplot matrix summarizes bivariate relationships between $Y$ and $X_j$ as well as between $X_j$ and $X_k$ ($j, k = 1, 2, ..., p - 1$):
  - Nature of bivariate relationships
  - Strength of bivariate relationships
  - Detection of outliers
  - Range spanned by $X$'s
- Scatterplot matrix combines many scatterplots.
- Examples presented later in this topic.
Correlation Matrix

• Complementary summary
• Displays all pairwise correlations
• When interpreting, be wary of
  – Nonlinear relationships
  – Outliers
  – Influential observations

Residual Plots

• Used for similar assessment of assumptions
  – Model is “correct”
  – Errors are Normally distributed
  – Errors have constant variance
  – Errors are independent
• Plot $e$ vs $\hat{Y}$ (overall)
• Plot $e$ vs $X_j$ (with respect to $X_j$)
• Plot $e$ vs non-included variable (e.g., $X_jX_k$)

Tests

• Univariate graphical summaries of $e$ still preferred
• NORMAL option in UNIVARIATE test normality
• Modified Levene’s and Breusch-Pagan for constant variance
• Lack of fit test: But need repeat observations where all $X$ fixed at same levels or can be comfortably grouped together….this hinders its applicability

Example I - Dwaine Studios (pg 236)

• Company that specializes in portraits of children. It has studios in 21 medium-sized cities nationwide and is considering expansion into other cities.
• Goal: To investigate whether sales are associated with certain characteristics of the city. If so, this could help in determining where to expand.
• Variables:
  – Annual sales ($Y$) - expressed in thousands of $\$\$
  – Persons aged 16 and younger ($X_1$) - expressed in thousands
  – Per capita disposable income ($X_2$) - expressed in thousands of $\$\$
Correlations

```sas
proc corr data=a1; 
    var young income sales; 

Pearson Correlation Coefficients, N = 21
    Prob > |r| under HO: Rho=0

young     income    sales
young     1.00000   0.78130   0.94455 <.0001 <.0001
income    0.78130   1.00000   0.83580 <.0001 <.0001
sales     0.94455   0.83580   1.00000 <.0001 <.0001
```
Example II - Predict Success?

- Goal: To find entry-level predictors of academic success
- Define academic success as high GPA after 3 semesters
- Predictors include
  - GPA after three semesters
  - HS math grades
  - HS science grades
  - HS english grades
  - SAT Math
  - SAT Verbal
- Data available on $n = 224$ students

Descriptive Statistics

- Using Proc MEANS or Proc UNIVARIATE

<table>
<thead>
<tr>
<th>Var</th>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>gpa</td>
<td>224</td>
<td>2.64</td>
<td>0.78</td>
</tr>
<tr>
<td>hsm</td>
<td>224</td>
<td>8.32</td>
<td>1.64</td>
</tr>
<tr>
<td>hss</td>
<td>224</td>
<td>8.09</td>
<td>1.70</td>
</tr>
<tr>
<td>hse</td>
<td>224</td>
<td>8.09</td>
<td>1.51</td>
</tr>
<tr>
<td>satm</td>
<td>224</td>
<td>595.29</td>
<td>86.40</td>
</tr>
<tr>
<td>satv</td>
<td>224</td>
<td>504.55</td>
<td>92.61</td>
</tr>
</tbody>
</table>
Correlations

[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;
   var satm satv;
   satm  satv
   satm 1.00 0.46 <.0001
   satv 0.46 1.00 <.0001

Regression Models

• Will now investigate:

  Model 1: GPA = HSM HSS HSE
  Model 2: GPA = HSM HSE
  Model 3: GPA = HSM
  Model 4: GPA = SATM SATV
  Model 5: GPA = HSM HSS HSE SATM SATV

• Should check residuals prior to any inference

Model 1

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3</td>
<td>27.71233</td>
<td>9.23744</td>
<td>18.86</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>220</td>
<td>107.75046</td>
<td>0.48977</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>223</td>
<td>135.46279</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 | Standard | Estimate | Error | t Value | Pr > |t| |
|----------|----|-----------|----------|----------|-------|---------|-------|---|
| Intercept| 1  | 0.58988   | 0.29424  | 2.00     | 0.0462|
| hsm      | 1  | 0.16857   | 0.03649  | 4.75     | <.0001|
| hss      | 1  | 0.03432   | 0.03756  | 0.91     | 0.3619|
| hse      | 1  | 0.04510   | 0.03870  | 1.17     | 0.2451|
### Model 2

#### Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Squares</th>
<th>Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2</td>
<td>27.30349</td>
<td>13.65175</td>
<td>27.89</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>221</td>
<td>108.15930</td>
<td>0.48941</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>223</td>
<td>135.46279</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE = 0.69958
R-Square = 0.2016
Dependent Mean = 2.63522
Coeff Var = 26.54718

#### Parameter Estimates

| Parameter | DF | Estimate | Error | t Value | Pr > |t| |
|-----------|----|----------|-------|---------|------|---|
| Intercept | 1  | 0.62423  | 0.29172| 2.14    | 0.0335|
| hsm       | 1  | 0.18265  | 0.03196| 5.72    | <.0001|
| hse       | 1  | 0.06067  | 0.03473| 1.75    | 0.0820|

### Model 3

#### Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Squares</th>
<th>Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>25.80989</td>
<td>25.80989</td>
<td>52.25</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>222</td>
<td>109.65290</td>
<td>0.49393</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>223</td>
<td>135.46279</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE = 0.70280
R-Square = 0.1905
Dependent Mean = 2.63522
Coeff Var = 26.66958

#### Parameter Estimates

| Parameter | DF | Estimate | Error | t Value | Pr > |t| |
|-----------|----|----------|-------|---------|------|---|
| Intercept | 1  | 0.90768  | 0.24355| 3.73    | 0.0002|
| hsm       | 1  | 0.20760  | 0.02872| 7.23    | <.0001|

### Model 4

#### Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Squares</th>
<th>Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2</td>
<td>8.58384</td>
<td>4.29192</td>
<td>7.48</td>
<td>0.0007</td>
</tr>
<tr>
<td>Error</td>
<td>221</td>
<td>126.87895</td>
<td>0.57411</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>223</td>
<td>135.46279</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE = 0.75770
R-Square = 0.0634
Dependent Mean = 2.63522
Coeff Var = 28.75287

#### Parameter Estimates

| Variable | DF | Estimate | Error | t Value | Pr > |t| |
|----------|----|----------|-------|---------|------|---|
| Intercept| 1  | 1.28868  | 0.37604| 3.43    | 0.0007|
| satm     | 1  | 0.00228  | 0.00066291| 3.44    | 0.0007|
| satv     | 1  | -0.00002456 | 0.00061847| -0.04   | 0.9684|

### Model 5

#### Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Squares</th>
<th>Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>5</td>
<td>28.64364</td>
<td>5.72873</td>
<td>11.69</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>218</td>
<td>106.81914</td>
<td>0.49000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>223</td>
<td>135.46279</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE = 0.70000
R-Square = 0.2115
Dependent Mean = 2.63522
Coeff Var = 26.56311

#### Parameter Estimates

| Variable | DF | Estimate | Error | t Value | Pr > |t| |
|----------|----|----------|-------|---------|------|---|
| Intercept| 1  | 0.32672  | 0.40000| 0.82    | 0.4149|
| satm     | 1  | 0.00094359| 0.00068566| 1.38    | 0.1702|
| satv     | 1  | -0.00040785 | 0.00059189| -0.69   | 0.4915|
| hsm      | 1  | 0.14596  | 0.03926| 3.72    | 0.0003|
| hss      | 1  | 0.03591  | 0.03780| 0.95    | 0.3432|
| hse      | 1  | 0.05529  | 0.03957| 1.40    | 0.1637|
General Linear Test

- Can use TEST statement in SAS
  ```
  proc reg data=a1;
  model gpa=satm satv hsm hss hse;
  sat: test satm, satv;
  hs: test hsm, hss, hse;
  ```

Test sat
  Results for Dep Var gpa
  Mean
  Source  DF  Square  F  Pr > F
  Num 2  0.46566  0.95  0.3882
  Den 218  0.49000

Test hs
  Results for Dep Var gpa
  Mean
  Source  DF  Square  F  P
  Num 3  6.68660  13.65 <.0001
  Den 218  0.49000

What’s the Best Model?

- Will discuss selection approaches in Chpts 8, 9, and 10
- Appears HSM only is best model
- Should also be looking at diagnostics

  - Important:
    - Look at variables one at a time
    - Look at all pairwise relationships
    - PLOT! PLOT! PLOT!

Model Fit Diagnostics

Key Results

- The relationship between $Y$ and $X_j$ depends on the other predictors in the model
- A predictor may be significant alone but not significant when other variables are in the model
- Similarly, coefficients and standard errors depend on the variables that are in the model
Background Reading

- KNNL Sections 6.8-6.9
- KNNL Sections 7.1-7.3