Topic 16: Multicollinearity and Polynomial Regression
Outline

• Multicollinearity
• Polynomial regression
An example (KNNL p256)

- The P-value for ANOVA F-test is < .0001
- The P values for the individual regression coefficients are 0.1699, 0.2849, and 0.1896
- None of these are near our standard significance level of 0.05
- What is the explanation?

Multicollinearity!!!
Multicollinearity

• **Numerical analysis problem** in that the matrix $X'X$ is close to singular and is therefore difficult to invert accurately

• **Statistical problem** in that there is too much correlation **among the explanatory variables** and it is therefore difficult to determine the regression coefficients
Multicollinearity

- Solve the statistical problem and the numerical problem will also be solved
  - We want to refine a model that currently has redundancy in the explanatory variables
  - Do this regardless if $X^\prime X$ can be inverted without difficulty
Multicollinearity

• Extreme cases can help us understand the problems caused by multicollinearity
  – Assume columns in X matrix were uncorrelated
    • Type I and Type II SS will be the same
    • The contribution of each explanatory variable to the model is the same whether or not the other explanatory variables are in the model
Multicollinearity

- Suppose a linear combination of the explanatory variables is a constant
  - The Type II SS for the X’s involved will all be zero
    - Example: $X_1 = X_2 \rightarrow X_1 - X_2 = 0$
    - Example: $3X_1 - X_2 = 5$
    - Example: SAT total = SATV + SATM
An example: Part I

Data a1;
    infile '../data/csdata.dat';
    input id gpa hsm hss hse satm satv genderm1;

Data a1; set a1;
    hs = (hsm+hse+hss)/3;
Proc reg data=a1;
    model gpa= hsm hss hse hs;
run;
Output

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of 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>

• Something is wrong
• $df_M = 3$ but there are 4 X’s
Output

NOTE: Model is not full rank. Least-squares solutions for the parameters are not unique. Some statistics will be misleading. A reported DF of 0 or B means that the estimate is biased.
Output

NOTE: The following parameters have been set to 0, since the variables are a linear combination of other variables as shown.

hs = 0.33333*hsm + 0.33333*hss + 0.33333*hse
### Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| | Type I SS | Type II SS |
|----------|----|--------------------|----------------|---------|-------|----------------|------------|------------|
| Intercept| 1  | 0.58988            | 0.29424        | 2.00    | 0.0462| 1555.5459      | 1.96837    |
| hsm      | B  | 0.16857            | 0.03549        | 4.75    | <.0001| 25.80989       | 11.04779   |
| hss      | B  | 0.03432            | 0.03756        | 0.91    | 0.3619| 1.23708        | 0.40884    |
| hse      | B  | 0.04510            | 0.03870        | 1.17    | 0.2451| 0.66536        | 0.66536    |
| hs       | 0  | 0                  | .              | .       | .     | .              | .          |

In this extreme case, SAS does not consider hs in the model.
Extent of multicollinearity

• This example had one explanatory variable equal to a linear combination of other explanatory variables

• This is the most extreme case of multicollinearity and is detected by statistical software because \((X'X)\) does not have an inverse

• We are concerned with cases less extreme
An example: Part II

*add a little noise to break up perfect linear association;
Data a1; set a1;
hs1 = hs + normal(612)*.05;

Proc reg data=a1;
  model gpa= hsm hss hse hs1;
run;
Output

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>4</td>
<td>27.81586</td>
<td>6.95396</td>
<td>14.15</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>219</td>
<td>107.64693</td>
<td>0.49154</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>

Model seems to be good here
### Output

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| | Type I SS | Type II SS |
|----------|----|--------------------|----------------|---------|------|--------|------------|-------------|
| Intercept| 1  | 0.56271            | 0.30066        | 1.87    | 0.0626 | 1555.5459 | 1.72182    |
| hsm      | 1  | 0.02411            | 0.31677        | 0.08    | 0.9394 | 25.80989  | 0.00285    |
| hss      | 1  | -0.11093           | 0.31872        | -0.35   | 0.7281 | 1.23708   | 0.05954    |
| hse      | 1  | -0.10038           | 0.31937        | -0.31   | 0.7536 | 0.66536   | 0.04856    |
| hs1      | 1  | 0.43805            | 0.95451        | 0.46    | 0.6467 | 0.10352   | 0.10352    |

- None of the predictors significant.
- Much larger SEs.
- Look at the differences in Type I and II SS
- Sign of each coefficient make sense?
Effects of multicollinearity

- Regression coefficients are not well estimated and may be meaningless
- Similarly for standard errors of these estimates
- Type I SS and Type II SS will differ
- $R^2$ and predicted values are usually ok in these situations
Pairwise Correlations

- Pairwise correlations can be used to check for “pairwise” collinearity
- Recall KNNL p256

```plaintext
proc reg data=a1 corr;
  model fat=skinfold thigh midarm;
  model midarm = skinfold thigh;
run;
```
Pairwise Correlations

- \( \text{Cor(skinfold, thigh)} = 0.9238 \)
- \( \text{Cor(skinfold, midarm)} = 0.4578 \)
- \( \text{Cor(thigh, midarm)} = 0.0847 \)

Multicollinearity may involve multiple X’s
- \( \text{Cor(midarm, skinfold+thigh)} = 0.9952!!! \)

See p 284 for change in coeff values of skinfold and thigh depending on what variables are in the model

None of these appear too high
Polynomial regression

• We can fit a quadratic, cubic, etc. relationship by defining squares, cubes, etc., of a single X in a data step and using them as additional explanatory variables

• We can do this with more than one explanatory variable if needed

• Issue: When we do this we generally create a multicollinearity problem
KNNL Example p300

- Response variable is the life (in cycles) of a power cell
- Explanatory variables are
  - Charge rate (3 levels)
  - Temperature (3 levels)
- This is a designed experiment!
Input and check the data

Data a1;
   infile '../data/ch08ta01.txt';
   input cycles chrate temp;
run;

Proc print data=a1;
run;
### Output

<table>
<thead>
<tr>
<th>Obs</th>
<th>cycles</th>
<th>chrate</th>
<th>temp</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>150</td>
<td>0.6</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>86</td>
<td>1.0</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>49</td>
<td>1.4</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>288</td>
<td>0.6</td>
<td>20</td>
</tr>
<tr>
<td>5</td>
<td>157</td>
<td>1.0</td>
<td>20</td>
</tr>
<tr>
<td>6</td>
<td>131</td>
<td>1.0</td>
<td>20</td>
</tr>
<tr>
<td>7</td>
<td>184</td>
<td>1.0</td>
<td>20</td>
</tr>
<tr>
<td>8</td>
<td>109</td>
<td>1.4</td>
<td>20</td>
</tr>
<tr>
<td>9</td>
<td>279</td>
<td>0.6</td>
<td>30</td>
</tr>
<tr>
<td>10</td>
<td>235</td>
<td>1.0</td>
<td>30</td>
</tr>
<tr>
<td>11</td>
<td>224</td>
<td>1.4</td>
<td>30</td>
</tr>
</tbody>
</table>

 Known as center points
Design Layout
Create new variables and run the regression

Data a1; set a1;
  chrate2=chrate*chrate;
  temp2=temp*temp;
  ct=chrate*temp;

Proc reg data=a1;
  model cycles=
    chrate temp chrate2 temp2 ct;
run;
## Output

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>5</td>
<td>55366</td>
<td>11073</td>
<td>10.57</td>
<td>0.0109</td>
</tr>
<tr>
<td>Error</td>
<td>5</td>
<td>5240.4386</td>
<td>1048.0877</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>10</td>
<td>60606</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
|----------|----|--------------------|----------------|---------|-------|---|
| Intercept| 1  | 337.72149          | 149.96163      | 2.25    | 0.0741|
| chrate   | 1  | -539.51754         | 268.86033      | -2.01   | 0.1011|
| temp     | 1  | 8.91711            | 9.18249        | 0.97    | 0.3761|
| chrate2  | 1  | 171.21711          | 127.12550      | 1.35    | 0.2359|
| temp2    | 1  | -0.10605           | 0.20340        | -0.52   | 0.6244|
| ct       | 1  | 2.87500            | 4.04677        | 0.71    | 0.5092|
Conclusion

• Overall F significant, individual t’s not significant → multicollinearity problem
• Look at the correlations (proc corr)
• There are some very high correlations
  – r(chrate,chrate2) = 0.99103
  – r(temp,temp2) = 0.98609
• Common to have correlation between powers of a variable
A remedy

• We can often remove the correlation between explanatory variables and their powers by centering.

• Centering means that you subtract off the mean before squaring etc.

• KNNL rescaled by standardizing (subtract the mean and divide by the standard deviation) but subtracting the mean is key here because you get positive and negative values of X.
A remedy

- Use Proc Standard to center the explanatory variables
- Recompute the squares, cubes, etc., using the centered variables
- Rerun the regression analysis
Proc standard

Data a2; set a1;
    schrate=chrate; stemp=temp;
    keep cycles schrate stemp;

Proc standard data=a2
    out=a3 mean=0 std=1;
    var schrate stemp;

Proc print data=a3;
run;
<table>
<thead>
<tr>
<th>Obs</th>
<th>cycles</th>
<th>schrate</th>
<th>stemp</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>150</td>
<td>-1.29099</td>
<td>-1.29099</td>
</tr>
<tr>
<td>2</td>
<td>86</td>
<td>0.00000</td>
<td>-1.29099</td>
</tr>
<tr>
<td>3</td>
<td>49</td>
<td>1.29099</td>
<td>-1.29099</td>
</tr>
<tr>
<td>4</td>
<td>288</td>
<td>-1.29099</td>
<td>0.00000</td>
</tr>
<tr>
<td>5</td>
<td>157</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>6</td>
<td>131</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>7</td>
<td>184</td>
<td>0.00000</td>
<td>0.00000</td>
</tr>
<tr>
<td>8</td>
<td>109</td>
<td>1.29099</td>
<td>0.00000</td>
</tr>
<tr>
<td>9</td>
<td>279</td>
<td>-1.29099</td>
<td>1.29099</td>
</tr>
<tr>
<td>10</td>
<td>235</td>
<td>0.00000</td>
<td>1.29099</td>
</tr>
<tr>
<td>11</td>
<td>224</td>
<td>1.29099</td>
<td>1.29099</td>
</tr>
</tbody>
</table>
Recompute squares and cross product

Data a3; set a3;
   schrate2=schrate*schrate;
   stemp2=stemp*stemp;
   sct=schrate*stemp;
Rerun regression

Proc reg data=a3;
    model cycles=schrate stemp
    schrate2 stemp22 sct;
run;
Output

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>5</td>
<td>55366</td>
<td>11073</td>
<td>10.57</td>
<td>0.0109</td>
</tr>
<tr>
<td>Error</td>
<td>5</td>
<td>5240.4386</td>
<td>1048.0877</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>10</td>
<td>60606</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Exact same ANOVA table as before!!
## Output

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
|----------|----|--------------------|----------------|---------|-------|---|
| Intercept | 1  | 162.84211          | 16.60761       | 9.81    | 0.0002|
| schrate  | 1  | -43.24831          | 10.23762       | -4.22   | 0.0083|
| stemp    | 1  | 58.48205           | 10.23762       | 5.71    | 0.0023|
| schrate2 | 1  | 16.43684           | 12.20405       | 1.35    | 0.2359|
| stemp2   | 1  | -6.36316           | 12.20405       | -0.52   | 0.6244|
| sct      | 1  | 6.90000            | 9.71225        | 0.71    | 0.5092|
Conclusion

• Overall F significant
• Individual t’s significant for chrate and temp
• Appears linear model will suffice
• Could do formal general linear test to assess this. (P-value is 0.5527)
Last slide

• We went over KNNL 7.6 and 8.1.
• We used programs Topic16.sas to generate the output for today