Lecture 11: Computation Diagnostics

Reading Assignment:

- Muller and Fetterman, Chapter 8: "GLM Computation Diagnostics" (Required)

Naive acceptance of computer output may lead to major errors, even though all assumptions prove valid.

The culprit is finite precision arithmetic, which is typically 7-15 decimal digits of accuracy for a single computation. Regression computations accumulate sums and cross-products so that numerical inaccuracy accumulates.

In the ideal situation, you are interested in a specific number of predictors and know the model beforehand. In many observational studies, investigators have a large number of measurements of predictors or exposures and want to evaluate them all in a model (e.g., positive anxiety (promotion at work, engagement), negative anxiety (fired, illness of family member), impact of positive anxiety, impact of negative anxiety, and potential interactions of these with two stress hormones!).

Including a lot of variables in a model often results in multicollinearity, a high degree of correlation among several predictors. This happens when too many variables have been put into the model, and a number of these variables measure similar phenomena. Collinearity does not cause a violation of the HILE Gauss assumptions. However, it

1. tends to inflate the variances of predicted values, and
2. tends to inflate the variances of parameter estimates (bad!).

Single Variable Problems and Solutions

Disparity in location and scale (mean and variance) can create substantial inaccuracy. Examine summary statistics (mean, variance, minimum, and maximum) to find potential location or scale problems. The easiest remedy is to scale predictors to have range $\in \{-10, +10\}$, a common scientific practice.

Examples of location or scale problems:

- Consider children's birth year, ranging from 1980 to 1990. This creates a location problem (and *collinearity* with the intercept), but replacing birth year with age avoids the location problem.
- Consider children's spirometry performance measured by forced vital capacity (FVC, the maximum amount of air that can be forcibly expired). Recording values in mL may give a range of $\{1000, 4000\}$ mL, leading to very large elements in $\mathbf{X}'\mathbf{X}$. Using liters (dividing FVC by 1000) eliminates the scaling problem.

Reducing location and scale disparities never hurts and may substantially improve accuracy. In extreme cases, creating means of zero (centering) and sums of squares or variances of one (normalizing) may be necessary, though using convenient center points and scales often suffices. For example, use human body temperatures as $T - 37$ (C).

Collinearity

Predictors in a model are *collinear* whenever the columns of $\mathbf{X}$ contain some amount of redundancy.

Mathematically, collinearity corresponds to linear dependence among columns of $\mathbf{X}$. Collinearity exists along a continuum (from nonexistent to moderate to severe to disastrous). Often, collinearity involves more than two variables.

Classic Example: Socioeconomic Status

Outcome: anything
Predictors: gender, race, age, education, socioeconomic status

Collinearity diagnostics identify variables containing little or no information above and beyond that in the other predictors.

Recall that the # of linearly independent columns in $\mathbf{X}$ equals $r =$rank$(\mathbf{X})$, and a full rank model has rank$(\mathbf{X}) = r = p$.

Some ANOVA (Chap. 12) designs use purposefully less than full rank $\mathbf{X}$, but for now the focus centers on unintentionally less than full rank designs (i.e.,problems).

Uncontrolled collinearity creates computational difficulties and confuses testing and interpretation.

Loosely, consider

- Rank: the number of distinct dimensions of information.
- Eigenanalysis: creating sets of regression coefficients (eigenvectors) needed to produce new variables which are uncorrelated, with successively maximum variances (eigenvalues).

Consider the columns of $\mathbf{X}$ as a collection of vectors:

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_0 & \mathbf{X}_1 & \dots & \mathbf{X}_{p-1} \end{bmatrix}.$$

Here $\mathbf{X}_0 = \mathbf{J}_n$ whenever the model contains an intercept.

Recall that the columns of a matrix $\mathbf{X}$ are linearly dependent if there exists a vector $\boldsymbol{\delta} \neq 0$ such that $\mathbf{X}\boldsymbol{\delta} = \mathbf{0}$. So if $\exists \{\delta_0, \delta_1, \dots \delta_{p-1}\}$, not all zero, such that $\sum_{j=0}^{p-1} \delta_j \mathbf{X}_j = \mathbf{0}$, then the columns of $\mathbf{X}$ are linearly dependent. Otherwise, the columns of $\mathbf{X}$ are linearly independent. If the columns of $\mathbf{X}$ are linearly independent, then $\mathbf{X}$ is full rank.

Also recall that an inner product matrix, $\mathbf{X}'\mathbf{X}$, has non-negative eigenvalues.

The # of non-zero (strictly positive) eigenvalues of $\mathbf{X}'\mathbf{X}$ equals r, and if $r < p$ ($\mathbf{X}$ less than full rank), then $\mathbf{X}$ contains an exact collinearity.

Only predictor properties determine collinearity.

Exact collinearity rarely occurs, but too much collinearity causes a loss

of numerical, statistical, and scientific accuracy.

Matrices Providing Information about Collinearity

Difficulties with $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y}$ arise from computing cross-products. Four versions of the inner product matrix $\mathbf{X}'\mathbf{X}$ provide information about collinearity and accuracy.

SSCP Matrix

The $p \times p$ *sums-of-squares and cross-products (SSCP) matrix* is given by

$$\mathbf{X}'\mathbf{X} = \left\{ \sum_{i=1}^{n} x_{ij} x_{ij'} \right\}.$$

Dividing by n yields $\mathbf{X}'\mathbf{X}/n$, the average cross-products matrix.

Scaled SSCP Matrix

Consider the $p \times p$ matrix

$$\boldsymbol{D}_s = \begin{bmatrix} \sum x_{i0}^2 & & & 0 \\ & \sum x_{i1}^2 & & \\ & & \ddots & \\ 0 & & & \sum x_{i,p-1}^2 \end{bmatrix},$$

which contains the diagonal elements of the SSCP matrix.

Define the $p \times p$ scaled SSCP matrix of average scaled (but not centered) cross-products as

$$(\mathbf{X}'\mathbf{X})_s = \boldsymbol{D}_s^{-0.5}(\mathbf{X}'\mathbf{X})\boldsymbol{D}_s^{-0.5}$$

This matrix has 1's on the diagonal. Examining the eigenvalues and eigenvectors of the scaled SSCP matrix is a good way to detect collinearity with the intercept.

Covariance Matrix

Partition the covariate matrix $\mathbf{X}$ into $\begin{bmatrix} \mathbf{J}_n & \mathbf{X}^* \end{bmatrix}$, where $\mathbf{X}^*$ is the $n \times (p-1)$ matrix excluding the intercept.

The $(p-1) \times (p-1)$ covariance matrix adjusts for predictor means.

Recall that the column means of $\mathbf{X}^*$ are defined by

$$\overline{\mathbf{X}^*} = \mathbf{X}^{*'}\mathbf{J}_n/n = \left\{ \sum_{i=1}^{n} x_{ij}^*/n \right\} .$$

We define the covariance matrix $\mathbf{C}$ by

$$\begin{aligned}
\boldsymbol{C} = \{c_{jj'}\} &= \left\{ \sum_{i=1}^{n} (x^*_{ij} - \overline{x^*}_j)(x^*_{ij'} - \overline{x^*}_{j'}) \right\} /n \\
&= \mathbf{X}^{*\prime}(\boldsymbol{I}_n - \mathbf{J}_n \mathbf{J}_n'/n)\mathbf{X}^*/n \\
&= \mathbf{X}^{*\prime}\mathbf{X}^*/n - \overline{\mathbf{X}^*}\,\overline{\mathbf{X}^*}' \\
&= \mathbf{X}^{*\prime}_c \mathbf{X}^*{}_c/n,
\end{aligned}$$

where $\mathbf{X}^*_c = \{(x^*_{ij} - \overline{x^*}_j)\}$ contains the $p-1$ columns of centered data. Note that the diagonal of $\mathbf{C}$ contains sample variances of the predictors.

The covariance matrix contains average centered cross-products and excludes the intercept. It is also an inner product matrix as shown above.

If $\mathbf{X}$ is full rank, then $\mathbf{C}$ has rank $p-1$.

Correlation Matrix

Extract the diagonal elements of the covariance matrix $\mathbf{C}$:

$$\boldsymbol{D}_c = \begin{bmatrix} c_{11} & & & 0 \\ & c_{22} & & \\ & & \ddots & \\ 0 & & & c_{p-1,p-1} \end{bmatrix}.$$

Define

$$\begin{aligned} \boldsymbol{R} &= \boldsymbol{D}_c^{-0.5} \boldsymbol{C} \boldsymbol{D}_c^{-0.5} \\ &= \boldsymbol{D}_c^{-0.5} \mathbf{X}_{c'}^{*} \mathbf{X}_c^{*} \boldsymbol{D}_c^{-0.5} / n \\ &= \{r_{jj'}\} = \left\{ \frac{c_{jj'}}{\sqrt{c_j c_{j'}}} \right\} = \left\{ \frac{c_{jj'}}{s_j s_{j'}} \right\}. \end{aligned}$$

Note that $c_{jj} = s_j^2$, the sample variance.

$\boldsymbol{R}$ contains average centered and scaled cross-products and is an inner

product matrix.

Also, $\mathbf{X}^*{}_c \boldsymbol{D}_c^{-0.5} = \boldsymbol{Z} = \{(x^*_{ij} - \overline{x^*}_j)/s_j\}$ equals centered and scaled data.

Examining the eigenvalues and eigenvectors of the correlation matrix is a good way to detect collinearity among predictors other than the intercept.

Eigenanalysis!

An eigenanalysis is the best way to detect and describe collinearity.

Recall that the eigenvalues of the square matrix $\boldsymbol{A}$ are the roots the characteristic equation:

$$|\boldsymbol{A} - \lambda \boldsymbol{I}| = 0.$$

Equivalently, write $A\mathbf{X} = \lambda \mathbf{X}$.

Also recall that the rank of a (square) matrix equals the # of non-zero eigenvalues, so a matrix is full rank if and only if the matrix has no zero eigenvalues.

For the special case of a symmetric $\boldsymbol{A}$, we define the spectral decomposition:

$$\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}',$$

where $\boldsymbol{\Lambda} = \text{diag}(\boldsymbol{\lambda})$ represents the diagonal matrix of eigenvalues, ordered so that $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_p$. The kth column of $\boldsymbol{V}$ has an eigenvector corresponding to kth eigenvalue, with the eigenvectors and eigenvalues in the same order.

$\boldsymbol{V}$ is a full rank orthogonal matrix, regardless of the rank of $\boldsymbol{A}$.

Since we usually scale eigenvectors to unit length, $\boldsymbol{V}\boldsymbol{V}' = \boldsymbol{V}'\boldsymbol{V} = \boldsymbol{I}$, $\boldsymbol{V}^{-1} = \boldsymbol{V}'$ and $(\boldsymbol{V}')^{-1} = \boldsymbol{V}$.

We will consider eigenanalysis of inner-product matrices like $\mathbf{X}'\mathbf{X}$, $\mathbf{C}$, and $\mathbf{R}$, which are all symmetric and positive semi-definite (so that all eigenvalues are ≥ 0). The rank of these matrices equals the number of nonzero eigenvalues, and the number of zero eigenvalues equals the number of linear dependences. Remember that the rank of $\mathbf{X}$ is the same as the rank of $\mathbf{X}'\mathbf{X}$.

The eigenvalues of $\mathbf{X}'\mathbf{X}$, $\boldsymbol{C}$, and $\boldsymbol{R}$ are variances of linear

combinations of X_j's, and eigenvectors provide regression coefficients to create new variables (variates) having the eigenvalues as variances. The relative size of eigenvalues, especially largest to smallest, indicates the amount of collinearity, while the eigenvectors allow discovery of which variables overlap.

Troubles often reduce to a scaling problem or a variable with s^2 near zero.

Example: Collinearity

Consider records from one night at a hospital emergency room.

If 39 of 40 patients are male, then the indicator of gender (0 for female, 1 for male) is highly collinear with the intercept.

Consider the role of eigenvalues in determining the stability of $(\mathbf{X}'\mathbf{X})^{-1}$. Using the spectral decomposition,

$$\mathbf{X}'\mathbf{X} = \boldsymbol{V}\text{Diag}(\boldsymbol{\lambda})\boldsymbol{V}'$$

If $\mathbf{X}$ has full rank, then $(\mathbf{X}'\mathbf{X})^{-1} = \boldsymbol{V}\text{Diag}(\boldsymbol{\lambda}^{-1})\boldsymbol{V}'$. If any eigenvalues are 0, we cannot compute the usual inverse. A zero eigenvalue implies some linear combination of the predictors has a zero variance and provides no additional information.

Principal Component Analysis

Principal component analysis describes variables using eigenanalysis.

Let $\boldsymbol{V} = \begin{bmatrix} \boldsymbol{v}_1 & \dots & \boldsymbol{v}_p \end{bmatrix}$ indicate the $p \times p$ matrix of eigenvectors of $\mathbf{X}'\mathbf{X}/n$, with each column of $\boldsymbol{V}$ a $p \times 1$ eigenvector.

Define $\mathbf{X}_* = \mathbf{X}\boldsymbol{V}$ as the $n \times p$ matrix of principal component scores. Each column represents a new variable (a variate).

The jth element of kth eigenvector ($\boldsymbol{v}_k$) provides the coefficient for the jth column (variable) in $\mathbf{X}$ to compute the kth principal component score.

Using eigenvectors of $\boldsymbol{C} = \mathbf{X}_c'\mathbf{X}_c/n$ allows computing principal component scores for centered data, while using eigenvectors of $\boldsymbol{R} = \boldsymbol{D}_c^{-0.5}\mathbf{X}_c'\mathbf{X}_c\boldsymbol{D}_c^{-0.5}/n$ allows computing principal component scores for centered and scaled data.

Except in special cases, the eigenvalues, eigenvectors, and principal component scores of $\boldsymbol{C}$, $\boldsymbol{R}$, and the SSCP matrices have no simple correspondences between them.

Example: Eigenanalysis of Correlation Matrix for Ozone Data

```
proc princomp data=ozone;
var outdoor home time_out;
run;
```

```
********************************************************************************
                        The PRINCOMP Procedure

                      Observations          64
                      Variables              3


                          Simple Statistics

                    outdoor              home          time_out

      Mean       44.94541180       19.83328125      0.2817187500
      StD        21.90644120       11.98360958      0.2135754184


                        Correlation Matrix

         outdoor       Outdoor Ozone Concentration (ppb)
         home          Home Indoor Ozone Concentration (ppb)
         time_out      Proportion of Time Spent Outdoors
```

Correlation Matrix

	outdoor	home	time_out
outdoor	1.0000	0.5558	0.0782
home	0.5558	1.0000	-.0071
time_out	0.0782	-.0071	1.0000

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	1.56035176	0.55838231	0.5201	0.5201
2	1.00196944	0.56429064	0.3340	0.8541
3	0.43767880		0.1459	1.0000

Eigenvectors

outdoor	Outdoor Ozone Concentration (ppb)
home	Home Indoor Ozone Concentration (ppb)
time_out	Proportion of Time Spent Outdoors

Eigenvectors

	Prin1	Prin2	Prin3
outdoor	0.707674	0.012227	-.706434
home	0.700809	-.139231	0.699629
time_out	0.089803	0.990184	0.107099

None of the eigenvalues are too close to zero, indicating that there is not much sign of collinearity among the predictors.

Now, we also examine collinarity with the intercept by conducting an eigenanalysis of the scaled SSCP matrix.

```
data ozone; set ozone; int=1; run;

proc princomp data=ozone noint;
var int outdoor home time_out;
run;
```

```
************************************************************************
                        The PRINCOMP Procedure

                        Observations          64
                        Variables              4


                          Simple Statistics

                  int         outdoor            home        time_out

 Mean     1.000000000     44.94541180     19.83328125    0.2817187500
 UStD     1.000000000     49.92478235     23.12405860    0.3525155138


                   Uncorrected Correlation Matrix

                      int      outdoor        home      time_out
     int           1.0000       0.9003      0.8577        0.7992
     outdoor       0.9003       1.0000      0.8966        0.7399
     home          0.8577       0.8966      1.0000        0.6832
```

time_out	0.7992	0.7399	0.6832	1.0000

Eigenvalues of the Uncorrected Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	3.44397079	3.09272276	0.8610	0.8610
2	0.35124803	0.23327663	0.0878	0.9488
3	0.11797140	0.03116161	0.0295	0.9783
4	0.08680979		0.0217	1.0000

Eigenvectors

	Prin1	Prin2	Prin3	Prin4
int	0.517493	-.013105	-.672666	-.528724
outdoor	0.515072	-.282728	-.212037	0.780901
home	0.500584	-.471532	0.649407	-.324566
time_out	0.465099	0.835195	0.284308	0.072810

Condition Number and Condition Index

The *condition index* for the kth eigenvalue equals $\sqrt{\lambda_1/\lambda_k}$. The maximum condition index, called the *condition number*, $\max_k \sqrt{\lambda_1/\lambda_k}$, involves the first and last eigenvalue.

(Some authors exchange the names of condition number and condition index, so be careful!)

Note that $1 \leq \mathrm{CI}_k < \infty$, where large values of the condition index indicate greater collinearity.

Example: Eigenanalysis of Average SSCP, Scaled SSCP, $\mathbf{C}$, and $\mathbf{R}$

Using PROC PRINCOMP, we conduct an eigenanalysis for the ozone data. Note the VARDEF option, which is used to specify that n and not $n-1$ is used in calculations for average SSCP and covariance matrices.

```
/* SSCP */
proc princomp data=ozone noint cov vardef=n;
var int outdoor home time_out;
run;
/* Scaled SSCP */
proc princomp data=ozone noint;
var int outdoor home time_out;
```

```
run;
/* Covariance Matrix */
proc princomp data=ozone noint cov vardef=n;
var outdoor home time_out;
run;
/* Correlation Matrix */
proc princomp data=ozone;
var outdoor home time_out;
run;
```

The condition indices (calculated by hand) are provided in a table.

Matrix Type	Eigenvalue	Condition Index
Average SSCP	2939.11	1.00
	88.99	5.75
	0.19	23.90
	0.04	267.81
Scaled SSCP	3.44	1.00
	0.35	3.13
	0.12	5.40
	0.09	6.30

Matrix Type	Eigenvalue	Condition Index
Covariance	2938.29	1.00
	88.99	5.75
	0.06	229.10
Correlation	1.56	1.00
	1.00	1.25
	0.44	1.89

We next consider interpretation of these values.

Interpretation of Condition Number

One rule of thumb is that CN> 30 implies moderate to severe collinearity. This corresponds to a ratio of variances of roughly 1000, so single computations may involve numbers varying by 3 decimal places. Loss of precision occurs as such errors accumulate.

Note that λ_k very near zero indicates a redundancy in the predictors, and the eigenvector can identify the culprits. The relative sizes of $\boldsymbol{v}_k = \{v_{jk}\}_{p\times 1}$ give the relative importance of the columns of $\mathbf{X}$ in determining the undesirable kth variate. The kth variate has zero variance, and hence no predictive value. We interpret the $\{v_{jk}\}$ as regression coefficients.

Consider the SSCP and scaled SSCP matrices. A λ_k near zero indicates either collinearity with the intercept or among other predictors. Consider carefully the value of the eigenvector (with a bad condition number) weighting the intercept. A relatively large value

indicates a collinearity with the intercept.

Many simple mistakes may create collinearity with the intercept:

- using birth year rather than age,
- analyzing FVC in mL, rather than FVC in L,
- including variables with near zero variance (39/40 males),
- including redundant codings (including male and female codes), and
- some combination of the above.

$\boldsymbol{C}$ and $\boldsymbol{R}$ diagnose collinearity in predictors other than the intercept. $\boldsymbol{C}$ involves the relative scales of the variables, and $\boldsymbol{R}$ does not. Since few tests vary due to a change in scale or location, we prefer $\boldsymbol{R}$, assuming no loss of precision.

R_j^2, Tolerance, and VIF

Correlation and Collinearity

The simplest type of collinearity for variables other than the intercept is a correlation of 1.0 between two predictors.

Example: Exact Collinearity

Let X_1 represent temperature in F, and let X_2 represent temperature in C. Then $x_{i1} = 32 + 1.8x_{i2}$.

Fitting the model $x_{i1} = \beta_0 + \beta_1 x_{i2} + \varepsilon_i$ yields $\widehat{\sigma}^2 = 0$ and $r^2(x_1, x_2) = 1$.

Suppose we use X_j as the response in a regression with the $p - 2$ predictors (excluding the intercept),

$$\{X_1, \ldots X_{j-1}, X_{j+1}, \ldots X_{p-1}\}.$$

Then we compute the *squared multiple correlation*

$$R_j^2 = R^2(X_j, \{X_1, \ldots X_{j-1}, X_{j+1}, \ldots X_{p-1}\}) .$$

Clearly $R_1^2 = R_2^2 = 1$.

Large values (close to 1) of R_j^2 imply worse collinearity by indicating the extent of redundancy of a predictor with the remaining ones, and values close to zero indicate little or no collinearity.

Define *tolerance* as $1 - R_j^2$. Tolerance close to 1 is good, while tolerance close to 0 is bad.

Define the *variance inflation factor* as

$$\text{VIF}_j = \frac{1}{1 - R_j^2} = \frac{1}{\text{tolerance}}.$$

The name comes from the fact that $\text{var}(\widehat{\beta}_j)$ is proportional to VIF_j.

The variance inflation factor shows how multicollinearity has increased

the instability of the coefficient estimates. The variance of the parameter estimate is larger (by the VIF) than it would be if there were no multicollinearity.

A VIF close to 1 is good, and VIF $\longrightarrow \infty$ is bad.

Summary:

Good		Diagnostic		Bad
0	$\leq$	R_j^2	$\leq$	1
1	$\geq$	$1 - R_j^2$	$\geq$	0
1	$\leq$	$\frac{1}{1-R_j^2}$	$\leq$	∞

Eliminating a redundant variable does not reduce prediction!

As a rule of thumb,

$R_j^2 > .90$ merits some attention, and

$R_j^2 > .98$ suggests the need for removal of a useless variable.

Example: Diagnosing Collinearity in Ozone Data

Suppose that we fit the model $y_i = \beta_0 + \beta_1 OUTDOOR_i + \beta_2 HOME_i + \beta_3 {HOME_i}^2 + \beta_4 {HOME_i}^3 + \beta_5 TIMEOUT_i + \varepsilon_i$ to the ozone data so that we can investigate whether a polynomial in home provides a better fit to the data. If so, we obtain the following regression model output.

```
data ozone; set ozone;
home2=home*home;
home3=home*home2;
run;

proc reg;
model personal=outdoor home home2 home3 time_out/tol vif;
run;
```

```
*************************************************************************

                              The REG Procedure
                                Model: MODEL1
                          Dependent Variable: personal
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	5109.91477	1021.98295	5.88	0.0002
Error	58	10073	173.67557		
Corrected Total	63	15183			

Root MSE	13.17860	R-Square	0.3366
Dependent Mean	23.54578	Adj R-Sq	0.2794
Coeff Var	55.97012		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2.34851	8.45710	0.28	0.7822
outdoor	1	0.10435	0.09388	1.11	0.2709
home	1	0.56187	1.46423	0.38	0.7026
home2	1	0.01058	0.07103	0.15	0.8821
home3	1	-0.00024100	0.00098237	-0.25	0.8071
time_out	1	13.70723	7.95522	1.72	0.0902

Parameter Estimates

Variable	DF	Tolerance	Variance Inflation
Intercept	1	.	0
outdoor	1	0.65176	1.53431
home	1	0.00895	111.68575
home2	1	0.00160	624.92088
home3	1	0.00415	241.25289
time_out	1	0.95497	1.04715

Is there evidence of collinearity?

To investigate further, we conduct eigenanalyses of the scaled SSCP and correlation matrices.

```
/* Scaled SSCP matrix */
proc princomp data=ozone noint;
var int outdoor home home2 home3 time_out;
run;

/* correlation matrix */
proc princomp data=ozone;
```

```
var outdoor home home2 home3 time_out;
run;
```

****

The PRINCOMP Procedure

Observations	64
Variables	6

Simple Statistics

	int	outdoor	home
Mean	1.000000000	44.94541180	19.83328125
UStD	1.000000000	49.92478235	23.12405860

Simple Statistics

	home2	home3	time_out
Mean	534.7220859	17251.08446	0.2817187500
UStD	788.6788104	31240.89600	0.3525155138

Uncorrected Correlation Matrix

	int	outdoor	home	home2	home3	time_out
int	1.0000	0.9003	0.8577	0.6780	0.5522	0.7992
outdoor	0.9003	1.0000	0.8966	0.7937	0.7039	0.7399
home	0.8577	0.8966	1.0000	0.9459	0.8610	0.6832
home2	0.6780	0.7937	0.9459	1.0000	0.9771	0.5338
home3	0.5522	0.7039	0.8610	0.9771	1.0000	0.4355
time_out	0.7992	0.7399	0.6832	0.5338	0.4355	1.0000

Eigenvalues of the Uncorrected Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	4.81211368	3.97703127	0.8020	0.8020
2	0.83508241	0.60164583	0.1392	0.9412
3	0.23343658	0.14484901	0.0389	0.9801
4	0.08858757	0.05833043	0.0148	0.9949
5	0.03025714	0.02973452	0.0050	0.9999
6	0.00052262		0.0001	1.0000

The PRINCOMP Procedure

Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6
int	0.406570	0.394870	-.446481	-.481845	0.489442	-.087778
outdoor	0.428643	0.164398	-.405351	0.785773	-.086223	-.007426
home	0.447257	-.107067	-.121952	-.352657	-.660918	0.460919
home2	0.421044	-.411815	0.117787	-.107096	-.154855	-.777050
home3	0.387170	-.549914	0.264047	0.103256	0.539792	0.419447
time_out	0.351780	0.577573	0.733402	0.062342	-.029063	-.007115

The PRINCOMP Procedure

Observations	64
Variables	5

Simple Statistics

	outdoor	home	home2
Mean	44.94541180	19.83328125	534.7220859
StD	21.90644120	11.98360958	584.3126423

Simple Statistics

	home3	time_out

Mean	17251.08446	0.2817187500
StD	26251.89177	0.2135754184

Correlation Matrix

	outdoor	home	home2	home3	time_out
outdoor	1.0000	0.5558	0.5730	0.5697	0.0782
home	0.5558	1.0000	0.9642	0.9037	-.0071
home2	0.5730	0.9642	1.0000	0.9836	-.0181
home3	0.5697	0.9037	0.9836	1.0000	-.0116
time_out	0.0782	-.0071	-.0181	-.0116	1.0000

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	3.31662552	2.30161843	0.6633	0.6633
2	1.01500708	0.44509651	0.2030	0.8663
3	0.56991057	0.47248772	0.1140	0.9803
4	0.09742285	0.09638887	0.0195	0.9998

```
       5    0.00103398                        0.0002        1.0000
```

Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5
outdoor	0.389979	0.167238	0.905409	0.012712	-.004641
home	0.524979	-.038764	-.227787	0.749446	0.330663
home2	0.539900	-.050271	-.226048	-.094914	-.803663
home3	0.529906	-.042101	-.208732	-.655080	0.494700
time_out	0.004661	0.982970	-.183526	-.005519	-.006083

Detecting Numerical Inaccuracy

Which Matrix Should Be Examined?

Consider the ordered list of matrices: $\boldsymbol{R}$, $\boldsymbol{C}$, SSCP, and $\mathbf{X}$. The order corresponds to the following rank order of:

- most to least processed,
- least to most information about scaling problems,
- least to most information about location problems, and
- easiest to hardest in which to see collinearity (beyond intercept).

A general strategy for investigating collinearity is provided below.

1. Look at descriptive statistics (mean, variance) for all predictors. If necessary, change location or scale and remove variables with zero variance.

2. Examine eigenvalues and eigenvectors of scaled SSCP matrix to check for collinearity with the intercept and eliminate problems (by changing location or scale or deleting variables) if necessary.

3. Examine eigenvalues and eigenvectors of $\mathbf{R}$ to discover collinearity among other predictors.

Location Problems

Consider a full rank GLM that includes an intercept.

In theory, slopes are location invariant. In addition, a full rank GLM which spans an intercept (and a GLH test which does not) has an F statistic and p-value invariant to location. In practice, finite precision arithmetic may create violations.

Any variation in F and p-value due to location implies a problem (for a model spanning an intercept and a GLH test which does not).

Only the scaled SSCP, SSCP, and $\mathbf{X}$ directly detect location problems. Eigenanalysis of the scaled SSCP matrix, means, histograms, and extreme values for $\boldsymbol{y}$ and $\mathbf{X}$ help detect location problems.

Extreme ratios of data values and SSCP or scaled SSCP eigenvalues (or elements) warn of location-induced inaccuracy.

Scaling Problems

Slope estimates may vary due to changes in scale.

In theory, a full rank GLM which spans an intercept (and a GLH test which does not involve the intercept) has F statistic and p-value invariant to change in scale of $\boldsymbol{y}$ and/or $\mathbf{X}$.

Any variation in F statistic and p-value due to change of scale indicates a problem (if model spans an intercept and test does not).

$\boldsymbol{C}$, scaled SSCP, SSCP and $\mathbf{X}$ yield information about scale problems.

Extreme ratios of values of $\boldsymbol{C}$ (especially the variances) and histograms and extreme values of $\boldsymbol{y}$ and $\mathbf{X}$ can detect scale problems.

Collinearity Problems

In theory a GLH test that does not span an intercept in a full rank GLM which spans an intercept has an F statistic and p-value invariant to full rank linear transformation of $\boldsymbol{y}$ and/or the columns of $\mathbf{X}$

Any variation in F statistic and p-value due to a full rank linear transformation diagnoses a problem (if the model spans an intercept and the test does not).

$\boldsymbol{R}$ demonstrates invariance to location and scale changes.

$\boldsymbol{R}$, its eigenvalues, and R_j^2's detect collinearity problems.

The eigenvector for the smallest eigenvalue of $\mathbf{R}$ identifies the culprits.

Recommendations

1. Conduct data validation as part of data entry and file creation.
2. Treat non-independence of observations, if necessary. (For example, use multivariate methods, such as repeated measures ANOVA or random effects models.)
3. Minimize location and scale problems by roughly aligning variable ranges of variables. Centering and scaling may be necessary.
4. Minimize collinearity.
   - Eliminate predictors with near zero variance.
   - Eliminate redundant variables defined by unimportance or R_j^2.
   - Avoid unnecessary collinearity: center polynomials or use orthogonal polynomials (Chapter 9), and use cell-mean or effect coding (Chapter 12).
5. Treat "linearity" (model specification error). Find useful predictors and transformations (Chapter 10).

6. Treat non-normality and heterogeneity. Consider transformations. See Chapter 11 for a general treatment of exploratory regression.

We will address all of these issues in the coming weeks.

To summarize, first, clean the data. Then scale, center, and code it sensibly. If necessary, transform and delete variables. Do not be satisfied with the model unless all diagnostics create no substantial cause for concern. Chapter 11 centers on using this approach while creating statistically defensible estimates and hypothesis tests.

Next: Polynomial Regression

Reading Assignment:

- Muller and Fetterman, Chapter 9: "Polynomial Regression" (Required)