D. Collinearity

1. Collinearity Definition
   a. Recall model $E[Y] = X\beta$.
   b. Opposite: orthogonality:
      i. Inner product of columns of $X$ is zero.
      ii. $\sum_{i=1}^{n} x_{ij}x_{ik} = 0$ if $j \neq k$.
      iii. Careful with notation: does $x_j$ represent row $j$ or column $j$?
   c. Extreme collinearity:
      i. Exist constants $\omega_j$ not all zero such that
         $\sum_{j=1}^{p} \omega_j x_{ij} = 0$.
      ii. Choose $J$ such that $\omega_j \neq 0$.
      iii. Then $x_{ij} = \sum_{j\neq i} \omega_j x_{ij} / (\omega_j)$.
      iv. Makes $X^\top X$ singular.
      v. More transparently, this makes $\beta$ and $\beta + \lambda$ give the same fitted values, and so models with these parameters cannot be distinguished from $Y$.
   d. More commonly, approximate collinearity:
      i. Exist constants $\omega_j$ not all zero such that
         $\sum_{j=1}^{p} \omega_j x_{ij} \approx 0$.
      ii. There are no two parameter vectors with exactly the same fitted values, but there are many that are close.
      iii. Consequence is that parameter estimates have inflated standard errors.
      v. So if $\text{Var}[\hat{\beta}_j]$ is inflated, so is the typical value of $\beta_j^2$.

2. Detection of Multicollinearity:
   a. Examine correlations between covariates.
   b. Or Variance Inflation Factor.
   c. Can also examine eigenvalues.
      i. We want $\omega$ so that $X\omega = 0$, for exact collinearity
      ii. For approximate collinearity, find $\omega$ minimizing $\|X\omega\|$. 
         • subject to $\|\omega\| = 1$.
         • $\|\omega\|$ is defined to be the vector norm $\sqrt{\sum_j \omega_j^2}$.
      iii. Easier to picture finding $\omega$ minimizing $\|X\omega\|^2$.
         • Lagrangian is $\omega^\top X^\top X\omega - \lambda(\omega^\top \omega - 1)$.
      iv. Stationary Points
         • Stationary points satisfy $2X^\top X\omega - 2\lambda \omega = 0$ and $\omega^\top \omega = 1$.
      v. Vectors $\omega$ satisfying $X^\top X\omega = \lambda \omega$ are called eigenvectors of $X^\top X$.
         • $\lambda$ is called an eigenvalue.
         • Symmetric real matrices as above have all eigenvalues real.

- There are no more eigenvalues than there are rows of the matrix.
- The smallest of these is the one giving the closest to collinear. See Fig. 8 and 9.

- Eigenvalues shown in picture.
  - The picture is here:

  ![Fig. 8: Level curves of $\omega^\top X^\top X\omega$](image)
  
  Dotted circles represent level curves of $\|\omega\|^2$

- Eigenvalues shown in picture.
  - The picture is here:

  ![Fig. 9: Level curves of $\omega^\top X^\top X\omega$](image)
  
  Dotted circles represent level curves of $\|\omega\|^2$

- Closeness to singularity measured by ratio of largest to smallest eigenvalue.
  - Called the condition number.

3. Origins of collinearity
   a. Data collection method?
      i. Investigators may choose to collect data in a way that makes variables collinear.
      ii. I don’t see this as particularly plausible.
   b. Constraints on the model or population
      i. If the population that is sampled from is a sub-manifold of the overall population, then resulting variables will be highly correlated.
      ii. Ex. Rutgers studies relationship between graduate GPA (the response) vs. undergraduate GPA and GRE (explanatory variables).
      iii. Those admitted and who accept lie in a narrower range of overall desirability than the general...
applicants pool.

c. model specification
   i. Ex., polynomial terms when data are constrained to a
      arrow range.

d. over-defined model.
   i. More regressors than variables.
   ii. Quite common, for ex. in genetic studies
      • Often times one wants to determine which genes
         among tens of thousands are associated with
         disease in a few hundred subjects.

4. Solutions to collinearity:
   a. Extend the range of the data set
      i. Text notes this is often infeasible because of
         cost or because new observations will no longer be
         typical.
   b. Re-specify variables:
      i. Ex., make orthogonal.
   c. Omit variables.

5. Ridge Regression:
   a. Model is still \( Y = X\beta + \epsilon, \epsilon \) independent and
      homoscedastic.
   b. Least squares estimates \( \hat{\beta} = (X^TX)^{-1}X^TY \)
   c. Problematic if \( X^TX \) is close to singular
   d. Ridge regression solution: \( \hat{\beta} = (X^TX + kI)^{-1}X^TY \)
      for some \( k \geq 0 \)
      i. \( k = 0 \) reduces to same least-squares approach.
   ii. \( k > 0 \) results in a matrix easier to invert.
   iii. Sometimes intercept term is not impacted.
   iv. Note that this does penalizes all parameters equally.

   • Might want to scale regressors first.
   e. \( E[\hat{\beta}] \) generally \( \neq \beta \) if \( k > 0 \)
      i. Estimates are biased.
      ii. \( k \) is called biasing constant.
      iii. Generally \( Var[\hat{\beta}_j] \leq Var[\hat{\beta}_j] \)
   iv. \( k \) can be thought of as reflecting prior belief about
      the size of \( \beta \).
   v. Estimates go to zero as \( k \to \infty \).
   vi. Text suggests trying values \( k \in [0,1] \).

E. Variable Selecton and Model Building

1. Build a model:
   a. Blindly-built regression model: add all seven covariates
      as linear predictors
   b. Smarter model will use mathematical and subject
      matter knowledge to build a better model.
      i. If response is always positive, and so taking log puts
         it on a scale that makes linear fits meaningful.
   ii. Log scale allows for multiplicative effects on original
      scale.
   iii. Enter cyclic effects: Season, hour in day, wind
      direction.
      • Treat these using sines and cosines.

2. Consequences of an incorrect model
   a. Leaving out a variable that should be in the model:
      i. Slope estimates (including intercept) are biased,
         unless omitted variable is orthogonal to variables of
         interest.
      ii. \( \hat{\beta}_p = (X_p^TX_p)^{-1}X_p^TY \)
      iii. \( E[\hat{\beta}_p] = (X_p^TX_p)^{-1}X_p^TE[Y] = \)
         \( (X_p^TX_p)^{-1}X_p^T(X_p\beta_p + X_r\beta_r) = \)
         \( \beta_p + (X_p^TX_p)^{-1}X_p^TX_r\beta_r \).
   b. Variability estimates are biased, and inflated.
      i. Variance of estimates of correct model are higher
         than in too-small model.
      • Represent the true regression matrix as \( (X_p, X_r) \)
      • Choose \( A, B, C \) so that
         \( A \) is square with as many columns as \( X_p \) has,
         and so that \( (X_p, X_r) = (Z_p, Z_r) \begin{pmatrix} A & B \\ o & C \end{pmatrix}^{-1} \)
         for \( (Z_p, Z_r) \) orthogonal.
      • Let \( e_j \) the vector with 1 in component \( j \) and 0
         everywhere else.
      • \( X_j^TX_j = A^{-1}A^{-1} \)
      • Variance of incorrect model estimator for \( \beta_j \) is
         \( e_j^T(AA^T + BB^T)e_j\sigma^2 \)

3. Which are Reasonable Submodels?
   a. Statistical intuition tells us which models are coherent.
      i. If powers of a term appear in the model, shifts in the
         origin of the measurement scale can arbitrarily knock
         out lower terms.
      ii. Hence do not consider removing lower order terms in
         the presence of higher-order terms.
      iii. Similar issues apply to interaction terms.
   b. Removing parameters associated with some factors
      collapses that category with the baseline category.
   c. Removing parameter associated with one level of a
      factor collapses the associated level into baseline.
   d. Model selection becomes dependent baseline choice,
      which is usually arbitrary.
   e. Removing one sine-cosine pair members fixes start of
      cycle.
      i. Arises as before from the sine-of-difference and
         cosine-of-difference formulae.
   f. Unless the model is parameterized to explicitly have
      a meaningful null-hypothesis start of the cycle, these
      coefficients should only be evaluated as a pair.