MPV: 9.0-9.6

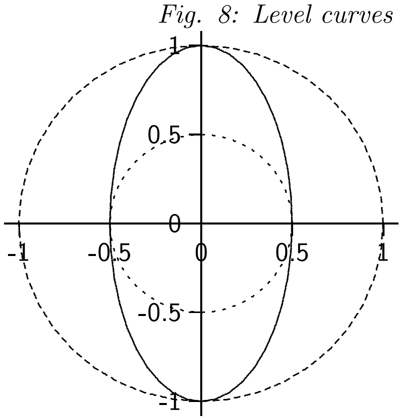
D. Collinearity
1. Collinearity Definition
   a. Recall model $\mathrm{E}[\boldsymbol{Y}] = \boldsymbol{X}\boldsymbol{\beta}$.
   b. Opposite: orthogonality:
      i. Inner product of columns of $\boldsymbol{X}$ is zero.
      ii. $\sum_{i=1}^{n} x_{ij} x_{ik} = 0$ if $j \neq k$.
      iii. Careful with notation: does $\boldsymbol{x}_j$ represent row j or column j?
   c. Extreme collinearity:
      i. Exist constants ω_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 $x_{iJ} = \sum_{j \neq J} x_{ij}\omega_j/(-\omega_J)$
      iv. Makes $\boldsymbol{X}^\top \boldsymbol{X}$ singular.
      v. More transparently, this makes $\boldsymbol{\beta}$ and $\boldsymbol{\beta} + \boldsymbol{\lambda}$ give the same fitted values, and so models with these parameters cannot be distinguished from $\boldsymbol{Y}$.
   d. More commonly, approximate collinearity:
      i. Exist constants ω_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.
      iv. Furthermore, $\mathrm{E}\left[\hat{\beta}_j^2\right] = \mathrm{Var}\left[\hat{\beta}_j\right] + \mathrm{E}\left[\hat{\beta}_j\right]^2 = \mathrm{Var}\left[\hat{\beta}_j\right] + \beta_j^2$

- So if $\mathrm{Var}\left[\hat{\beta}_j\right]$ is inflated, so is the typical value of $\hat{\beta}_j^2$.

2. Detection of Multicollinearity:
   a. Examine correlations between covariates.
      i. Will not necessarily catch effects of three or more variables.
   b. Or Variance Inflation Factor.
      i. See Trevor A. Craney & James G. Surles (2002) Model-Dependent Variance Inflation Factor Cutoff Values, Quality Engineering, 14:3, 391-403, DOI: 10.1081/QEN-120001878
   c. Can also examine eigenvalues.
      i. We want $\boldsymbol{\omega}$ so that $\boldsymbol{X}\boldsymbol{\omega} = 0$, for exact collinearity
      ii. For approximate collinearity, find $\boldsymbol{\omega}$ minimizing $\|\boldsymbol{X}\boldsymbol{\omega}\|$.
         - subject to $\|\boldsymbol{\omega}\| = 1$.
         - $\|\boldsymbol{\omega}\|$ is defined to be the vector norm $\sqrt{\sum_j \omega_j^2}$.
      iii. Easier to picture finding $\boldsymbol{\omega}$ minimizing $\|\boldsymbol{X}\boldsymbol{\omega}\|^2$
         - Lagrangian is $\boldsymbol{\omega}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\omega} - \lambda(\boldsymbol{\omega}^\top\boldsymbol{\omega} - 1)$.
      iv. Stationary Points
         - Stationary points satisfy $2\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\omega} - 2\lambda\boldsymbol{\omega} = \mathbf{o}$ and $\boldsymbol{\omega}^\top\boldsymbol{\omega} = 1$
      v. Vectors $\boldsymbol{\omega}$ satisfying $\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{\omega} = \lambda\boldsymbol{\omega}$ are called eigenvectors of $\boldsymbol{X}^\top\boldsymbol{X}$.
         - λ 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. 8and 9.
      vi. Eigenvalues shown in picture.
         - The picture is here:

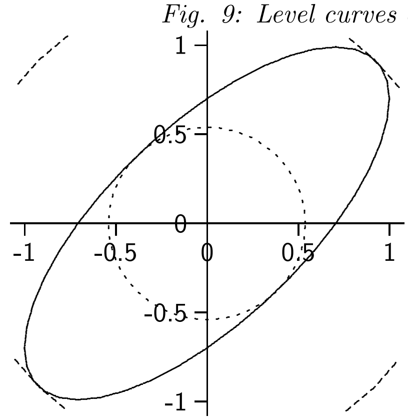
Fig. 8: Level curves of $\omega^\top \boldsymbol{X}^\top \boldsymbol{X}\omega$



Dotted circles represent level curves of $\|\boldsymbol{\omega}\|^2$

      vii. Eigenvalues shown in picture.
         - The picture is here:
      viii. Closeness to singularity measured by ratio of largest to smallest eigenvalue.
         - Called the condition number.
3. Origins of collinearity

Fig. 9: Level curves of $\boldsymbol{\omega}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\omega}$



Dotted circles represent level curves of $\|\boldsymbol{\omega}\|^2$

   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

applicant 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 that 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 $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon}$ independent and homoscedastic.
   b. Least squares estimates $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y}$
   c. Problematic if $\boldsymbol{X}^\top \boldsymbol{X}$ is close to singular
   d. Ridge regression solution: $\tilde{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X} + k\boldsymbol{I})^{-1}\boldsymbol{X}^\top \boldsymbol{Y}$ 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. $\mathrm{E}\left[\tilde{\boldsymbol{\beta}}\right]$ generally $\neq \boldsymbol{\beta}$ if $k > 0$
      i. Estimates are biased.
      ii. k is called biasing constant.
      iii. Generally $\mathrm{Var}\left[\tilde{\beta}_j\right] \leq \mathrm{Var}\left[\hat{\beta}_j\right]$
      iv. k can be thought of as reflecting prior belief about the size of $\boldsymbol{\beta}$.
         - with distribution centered at zero.
      v. Estimates go to zero as $k \to \infty$.
         - Text suggests trying values $k \in [0, 1]$.
      vi. HKB estimator: $\hat{k} = p\hat{\sigma}^2/(\hat{\boldsymbol{\beta}}^\top \hat{\boldsymbol{\beta}})$ for $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}$ from least-squares estimate.

MPV: 10.1-10.1.2

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. $\check{\boldsymbol{\beta}}_p = (\boldsymbol{X}_p^\top \boldsymbol{X}_p)^{-1}\boldsymbol{X}_p^\top \boldsymbol{Y}$
      iii. $\mathrm{E}\left[\check{\boldsymbol{\beta}}_p\right] = (\boldsymbol{X}_p^\top \boldsymbol{X}_p)^{-1}\boldsymbol{X}_p^\top \mathrm{E}\left[\boldsymbol{Y}\right] = (\boldsymbol{X}_p^\top \boldsymbol{X}_p)^{-1}\boldsymbol{X}_p^\top(\boldsymbol{X}_p\boldsymbol{\beta}_p + \boldsymbol{X}_r\boldsymbol{\beta}_r) = \boldsymbol{\beta}_p + (\boldsymbol{X}_p^\top \boldsymbol{X}_p)^{-1}\boldsymbol{X}_p^\top \boldsymbol{X}_r\boldsymbol{\beta}_r$.
   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 $(\boldsymbol{X}_p, \boldsymbol{X}_r)$
         - Choose $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$ so that
            ▷ $\boldsymbol{A}$ is square with as many columns as $\boldsymbol{X}_p$ has,
            ▷ and so that $(\boldsymbol{X}_p, \boldsymbol{X}_r) = (\boldsymbol{Z}_p, \boldsymbol{Z}_r)\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \mathbf{o} & \boldsymbol{C} \end{pmatrix}^{-1}$ for $(\boldsymbol{Z}_p, \boldsymbol{Z}_r)$ orthogonal.
         - Let $\boldsymbol{e}_j$ the vector with 1 in component j and 0 everywhere else.
         - $\boldsymbol{X}_p^\top \boldsymbol{X}_p = \boldsymbol{A}^{-1\top}\boldsymbol{A}^{-1}$
         - Variance of incorrect model estimator for β_j is $\boldsymbol{e}_j^\top \boldsymbol{A}\boldsymbol{A}^\top \boldsymbol{e}_j\sigma^2$
         - $\boldsymbol{X}^\top \boldsymbol{X} = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \mathbf{o} & \boldsymbol{C} \end{pmatrix}^{-1\top}\begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \mathbf{o} & \boldsymbol{C} \end{pmatrix}^{-1}$
         - $(\boldsymbol{X}^\top \boldsymbol{X})^{-1} = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \mathbf{o} & \boldsymbol{C} \end{pmatrix}\begin{pmatrix} \boldsymbol{A}^\top & \mathbf{o} \\ \boldsymbol{B}^\top & \boldsymbol{C}^\top \end{pmatrix} = \begin{pmatrix} \boldsymbol{A}\boldsymbol{A}^\top + \boldsymbol{B}\boldsymbol{B}^\top & \boldsymbol{B}\boldsymbol{C}^\top \\ \boldsymbol{C}\boldsymbol{B}^\top & \boldsymbol{C}\boldsymbol{C}^\top \end{pmatrix}$

         - Variance of correct model estimator for β_j is $\boldsymbol{e}_j^\top(\boldsymbol{A}\boldsymbol{A}^\top + \boldsymbol{B}\boldsymbol{B}^\top)\boldsymbol{e}_j\sigma^2$

:

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.
      i. Removing parameter associated with one level of a factor collapses the associated level into baseline.
      ii. Model selection becomes dependent baseline choice, which is usually arbitrary.
   c. Removing one sine-cosine pair members fixes start of cycle.
      i. Arises as before from the sine-of-difference and cosine-of-difference formulae.
      ii. 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.