Purdue-NCKU program

Lecture 8
Multiple Linear Regression

Dr. Qifan Song

The Data and Model

- Still have single response variable Y
- Now have multiple explanatory variables
- Examples:
  - Blood Pressure vs Age, Weight, Diet, Fitness Level
  - Traffic Count vs Time, Location, Population, Month
- Goal: There is a total amount of variation in Y (SSTO). We want to explain as much of this variation as possible using a linear model and our explanatory variables

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

- Have $p-1$ predictors $\longrightarrow p$ coefficients

General Linear Model

However, it can be much more flexible than just using the original response and explanatory variables in your data set

- Polynomial regression:

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i \\
&:= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i
\end{aligned}
$$

- cross product term:

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i2} * X_{i1} + \varepsilon_i \\
&:= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i
\end{aligned}
$$

- Transformed response:

$$
\log(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i
$$

- Factor analysis is also a multiple linear regression

Still linear models (of β's), while the meaning of β is different (will discussed later)

General Linear Regression In Matrix Terms

- After transformation and re-organization, a linear model ("linear" w.r.t. unknown coefficient, not to actual predictors) is obtained

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

- As an array

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1\ p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2\ p-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n\ p-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

- In matrix notation

$$\mathbf{Y} = \ \mathbf{X}\boldsymbol{\beta} \ + \boldsymbol{\varepsilon}$$

- Distributional assumptions:

$$\boldsymbol{\varepsilon} \sim N(0, \sigma^2 I) \longrightarrow \mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I)$$

Estimation, Fitted value and Residuals

- Least squares estimates $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$
- Fitted values: $\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{H}\mathbf{Y}$ define a (hyper)plane.
- Residuals: $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$
- Expected value $E(\mathbf{e}) = \mathbf{0}$
- Covariance Matrix

$$\begin{aligned} \boldsymbol{\sigma}^2(\mathbf{e}) &= \sigma^2(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})' \\ &= \sigma^2(\mathbf{I} - \mathbf{H}) \end{aligned}$$

  - Var$(e_i) = \sigma^2(1 - h_{ii})$ where $h_{ii} = \mathbf{X}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_i$
  - Residuals are usually correlated, i.e., $cov(e_i, e_j) = -\sigma^2 h_{ij}$, $i \neq j$

- Will use this information for diagnose

Estimation of σ^2

- Similar approach as before
- Estimate it from $\mathbf{e}$, since $\mathbf{e}$ has nothing to do with β_i's.
- Now p model parameters

$$
\begin{aligned}
s^2 &= \frac{\mathbf{e}'\mathbf{e}}{n-p} \\
&= \frac{(\mathbf{Y}-\mathbf{Xb})'(\mathbf{Y}-\mathbf{Xb})}{n-p} \\
&= \frac{\text{SSE}}{n-p} \\
&= \text{MSE}
\end{aligned}
$$

- Specifically, SSE$\sim \sigma^2\chi^2_{\text{rank of }(\mathbf{I}-\mathbf{H})}$

ANOVA TABLE

Source of Variation	df	SS	MS	F Value
Regression (Model)	$p-1$	SSR	MSR=SSR/$(p-1)$	MSR/MSE
Error	$n-p$	SSE	MSE=SSE/$(n-p)$	
Total	$n-1$	SSTO		

- F Test: Tests if the predictors *collectively* help explain the variation in Y
  - $H_0 : \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$
  - H_a : at least one $\beta_k \neq 0$, $1 \leq k \leq p-1$
  - $F^* = \frac{SSR/(p-1)}{SSE/(n-p)} \overset{H_0}{\sim} F(p-1, n-p)$
  - Reject H_0 if $F^* > F(1-\alpha, p-1, n-p)$
- No conclusions possible regarding individual predictors

Testing Individual Predictor

- Have already shown that $\mathbf{b} \sim N\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right)$
  - This implies $b_k \sim N(\beta_k, \sigma^2(b_k))$
- Perform t test
  - $H_0: \ \beta_k = \beta_k^0$ vs $H_a : \beta_k \neq \beta_k^0$
  - $t^* = \frac{b_k - \beta_k^0}{s(b_k)} \ \sim \ t_{n-p}$ under H_0
  - P-value $= Pr(|t_{n-p}| \geq t^*)$
  - Reject H_0 if $|t^*| \ > \ t(1-\alpha/2, n-p)$ or P-value $< \alpha$
- Confidence interval for β_k
  - $b_k \pm t(1-\alpha/2, n-p)s\{b_k\}$

Estimation of Mean Response $E(Y_h)$

- interested in making predictions for a new observation, represented by a p dimensional vector $\mathbf{X}_h$
  - Can show $\hat{Y}_h \sim N\left(\mathbf{X}_h'\boldsymbol{\beta}, \sigma^2\mathbf{X}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h\right)$
  - standard error $s\{\hat{Y}_h\} = \sqrt{\text{MSE}\mathbf{X}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h}$
- Individual CI for $\mathbf{X}_h$
  - $\hat{Y}_h \ \pm \ t(1-\alpha/2, n-p)s\{\hat{Y}_h\}$
- Bonferroni CI for g vectors $\mathbf{X}_h$
  - $\hat{Y}_h \ \pm \ t(1-\alpha/(2g), n-p)s\{\hat{Y}_h\}$
- Working-Hotelling confidence band for the whole regression line
  - $\hat{Y}_h \ \pm \ \sqrt{pF(1-\alpha, p, n-p)}\ s\{\hat{Y}_h\}$

Predict New Observation

- $Y_{h(new)} = E(Y_h) + \varepsilon$
  - $\widehat{Y}_h + \varepsilon \sim N\left(\mathbf{X}_h'\boldsymbol{\beta}, \sigma^2(1 + \mathbf{X}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h)\right)$
  - $s^2(pred) = s^2(\widehat{Y}_h) + \text{MSE}$
- Individual CI of $Y_{h(new)}$
  - $\widehat{Y}_h \ \pm \ t(1 - \alpha/2, n-p)s\{pred\}$
- Bonferroni CI for g vectors $\mathbf{X}_h$
  - $\widehat{Y}_h \ \pm \ t(1 - \alpha/(2g), n-p)s\{pred\}$

General Linear Test

- Comparison of a **full** model and **reduced** model that involves a subset of full model predictors (i.e., hierarchical structure)

- Involves a comparison of unexplained SS

- Consider a full model with k predictors (or k mean parameters) and reduced model with l predictors ($l < k$)

- One can prove that SSE(R) – SSE(F) $\geq$ 0.

- Can show that under null hypothesis

$$F^{\star} = \frac{(\text{SSE(R)} - \text{SSE(F)})/((n-1-l)-(n-k-1))}{\text{SSE(F)}/(n-k-1)} \sim F_{k-l,n-k-1} \text{ distribution}$$

- Degrees of freedom for F^* are the number of **extra** variables and the error degrees of freedom for the full model

Example

- Testing the null hypothesis that the regression coefficients for the **extra** variables are all zero.
- $H_0: \ \beta_k = 0$ vs $H_a: \beta_k \neq 0$
  - Full Model :
$$Y_i = \beta_0 + \sum_{j=1}^{p-1} \beta_j X_{ji} + \varepsilon_i$$
  - Reduced Model :
$$Y_i = \beta_0 + \sum_{j=1}^{k-1} \beta_j X_{ji} + \sum_{j=k+1}^{p-1} \beta_j X_{ji} + \varepsilon_i$$
  - $F^\star = \frac{(\text{SSE(R)}-\text{SSE(F)})/1}{\text{SSE(F)}/(n-p)}$
  - Reject H_0 if $F^* > F(1-\alpha, 1, n-p)$
- Can show that $F^* = (t^*)^2$, i.e., equivalent to the t test

Extra SS and Notation

- Consider $H_0 : X_1, X_3$ vs $H_a : X_1, X_2, X_3, X_4$
- Null can also be written $H_0 : \beta_2 = \beta_4 = 0$
- Write SSE(F) and SSE(R) as SSE(X_1, X_2, X_3, X_4) and SSE(X_1, X_3) respectively
- Difference in SSE's is the **extra SS**

$$\text{SSE}(X_2, X_4 | X_1, X_3) = \text{SSE}(X_1, X_3) - \text{SSE}(X_1, X_2, X_3, X_4)$$

- Recall SSM can also be used

$$\begin{aligned} \text{SSM}(X_2, X_4 | X_1, X_3) &= \text{SSM}(X_1, X_2, X_3, X_4) - \text{SSM}(X_1, X_3) \Longrightarrow \\ \text{SSM}(X_1, X_2, X_3, X_4) &= \text{SSM}(X_1, X_3) + \text{SSM}(X_2, X_4 | X_1, X_3) \end{aligned}$$

- Can rewrite F test as

$$F^{\star} = \frac{\text{SSE}(X_2, X_4 | X_1, X_3)/(4-2)}{\text{SSE}(X_1, X_2, X_3, X_4)/(n-5)}$$

- If it is possible that neither H_0 nor H_1 is correct, large p-value doesn't necessary provide evidence for H_0, but still serves as an evaluation tool for the usefulness of the additional predictors in H_1.

Type I SS and Type II SS

- Type I and Type II are very different
  - Type I is sequential, so it depends on `model` statement
  - Type II is conditional on all others, so it does not depend on model statement
- For example, `model y = x1 x2 x3` yields

Type I	Type II
SSM(X_1)	SSM($X_1 \| X_2, X_3$)
SSM($X_2 \| X_1$)	SSM($X_2 \| X_1, X_3$)
SSM($X_3 \| X_1, X_2$)	SSM($X_3 \| X_1, X_2$)

- Could variables be explaining same SS and "canceling" each other out, such that we need to cautions about testing results?

- A case study

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	396.98461	132.32820	21.52	<.0001
Error	16	98.40489	6.15031		
Corrected Total	19	495.38950			

Parameter Estimates

Variable	DF	Parameter Estimate	Pr > |t|	Type I SS	Type II SS
Intercept	1	117.08469	0.2578	8156.76050	8.46816
skinfold	1	4.33409	0.1699	352.26980	12.70489
thigh	1	-2.85685	0.2849	33.16891	7.52928
midarm	1	-2.18606	0.1896	11.54590	11.54590

- Set of three variables helpful in predicting body fat ($P < 0.0001$)
- None of the individual parameters is significant
  - Addition of each predictor to a model containing the other two is not helpful
  - More than 90% of Type I SS of `skinfold` can also be explained by `thigh` and `midarm`

Multicollinearity

- Numerical analysis problem is that the matrix $\mathbf{X}'\mathbf{X}$ is almost singular (linear dependent columns)
  - Makes it difficult to take the inverse
  - Generally handled with current algorithms
- Statistical problem: too much correlation among predictors
  - The coefficient estimation lacks interpretability.
  - Difficult to determine regression coefficients $\longrightarrow$ Increased standard error
  - May not affect prediction accuracy if the testing samples follow similar multicollinear correlation.
- Want to refine model to remove redundancy in the predictors

- Investigate the model via general linear tests: `fat=skinfold`

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	352.26980	352.26980	44.30	<.0001
Error	18	143.11970	7.95109		
Corrected Total	19	495.38950			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-1.49610	3.31923	-0.45	0.6576
skinfold	1	0.85719	0.12878	6.66	<.0001

- Skinfold now helpful. Note the change in coefficient estimate and standard error compared to the full model.

Residuals for Diagnostics

- $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$
  - $\mathbf{I} - \mathbf{H}$ symmetric and idempotent
- Expected value $E(\mathbf{e}) = \mathbf{0}$
- Covariance matrix

$$\begin{aligned} \boldsymbol{\sigma}^2(\mathbf{e}) &= \sigma^2(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})' \\ &= \sigma^2(\mathbf{I} - \mathbf{H}) \end{aligned}$$

  - $\text{Var}(e_i) = \sigma^2 \cdot (1 - h_{ii})$ where $h_{ii} = \mathbf{X}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_i$
  - $Cov(e_i, e_j) = \sigma^2 \cdot (0 - h_{ij}) = -\sigma^2 h_{ij}$
- Estimated variance and covariance
  - $\widehat{Var}(e_i) = MSE \cdot (1 - h_{ii})$
  - $\widehat{Cov}(e_i, e_j) = -MSE \cdot h_{ij}$

Residuals

- Ordinary residual

$$e_i = Y_i - \widehat{Y}_i \quad \rightarrow \quad \mathbf{e} \sim \text{MVN}(\mathbf{0}, (\mathbf{I} - \mathbf{H})\sigma^2)$$

  – residuals do not have the same variance, but depend on $\mathbf{X}_i$

- Semi-studentized residual

$$r_i = \frac{e_i}{\sqrt{\text{MSE}}}$$

  – denominator is not an estimate of SD of e_i

- (Internally) Studentized Residual

$$r_i = \frac{e_i}{\sqrt{\text{MSE}(1 - h_{ii})}}$$

  – denominator is the estimate of SD of e_i

- "Studentized" residual doesn't follow the student t distribution (but a τ distribution)

- Outlier may not have a outstanding studentized residual

Deleted Residual

- Deleted residual (a refinement of residual)

$$d_i = Y_i - \widehat{Y}_{i(i)} = \frac{e_i}{1 - h_{ii}}$$

  - $(\mathbf{X}_i, Y_i)$ was not used to fit the model
  - can calculate d_i in a single model fit

- Standard deviation of deleted residuals

$$\begin{aligned} s^2\{d_i\} &= MSE_{(i)} \cdot (1 + \mathbf{X}_i'(\mathbf{X}_{(i)}'\mathbf{X}_{(i)})^{-1}\mathbf{X}_i) \\ &= \frac{MSE_{(i)}}{1 - h_{ii}} \end{aligned}$$

- Studentized deleted residual (externally studentized residual)

$$\begin{aligned} t_i &= \frac{d_i}{s\{d_i\}} = \frac{e_i}{1 - h_{ii}} \cdot \sqrt{\frac{1 - h_{ii}}{MSE_{(i)}}} \\ &= \frac{e_i}{\sqrt{MSE_{(i)}(1 - h_{ii})}} \end{aligned}$$

Studentized Deleted Residuals

- If there is only one outlier, its studentized deleted residual will be outstanding
- Useful for identifying outlying Y observation
  - Test $H_{i0} : E[Y_i] = X_i\beta$ vs $H_{ia} : E[Y_i] \neq X_i\beta$
- If there are no outlying observations,

$$t_i \sim t_{n-1-p}$$

  - can compare t_i to this reference distribution
  - adjust for n tests using Bonferroni
  - an outlier has $|t_i| > t_{1-\alpha/(2n)}(n-1-p)$
  - t_i are not independent

Identifying Outlying X: Hat Matrix Diagonals

- Diagonals $0 \leq h_{ii} \leq 1$ and sum to p
- Also known as the leverage of ith case
- Is a measure of distance between the X value and the mean of the X values for all n cases $(\overline{X}_1, \overline{X}_2, ..., \overline{X}_{p-1})$
- Since $\hat{\mathbf{Y}} = \mathbf{HY}$

$$\widehat{Y}_i = h_{i1}Y_1 + h_{i2}Y_2 + \ldots + h_{in}Y_n$$

- Thus h_{ii} is a measure of how much Y_i is contributing to the prediction of $\widehat{Y}_i$

Hat Matrix Diagonals

- Residual

$$\begin{aligned} \mathbf{e} &= (\mathbf{I} - \mathbf{H})\mathbf{Y} \\ Var(\mathbf{e}) &= (\mathbf{I} - \mathbf{H})\sigma^2 \\ Var(e_i) &= (1 - h_{ii})\sigma^2 \end{aligned}$$

- Large h_{ii} means small residual variance
  - $\widehat{Y}_i$ will be close to Y_i (i.e., model is forced to fit this observation closely)

- Observations with large h_{ii} considered influential

  - large h_{ii} if it is more than double of the average value, i.e., $h_{ii} > 2p/n$

- Can compute $\mathbf{X}'_{new}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_{new}$ to check for hidden extrapolation

Identifying Influential Cases

Cook's Distance

- Measures influence of a case on the prediction of all $\widehat{Y}_i$'s
- Standardized version of sum of squared differences between fitted values with and without case i

$$D_i = \frac{\sum_{j=1}^{n} (\widehat{Y}_j - \widehat{Y}_{j(i)})^2}{p \cdot \text{MSE}} = \frac{(\mathbf{b}_{(i)} - \mathbf{b})'(\mathbf{X}'\mathbf{X})(\mathbf{b}_{(i)} - \mathbf{b})}{p \cdot \text{MSE}}$$

  – can be obtained in a single fit

$$D_i = \frac{e_i^2 h_{ii}}{p\text{MSE}(1 - h_{ii})^2}$$

- Compare with $F(p, n-p)$
- Concern if D_i is above the 50%-tile of $F(p, n-p)$

Multicollinearity Diagnostics: VIF

- Use **V**ariance **I**nflation **F**actor (VIF)
- VIF$_k$ is the the kth diagonal element of r_{XX}^{-1} (inverse of sample correlation matrix)

$$\text{VIF}_k = (r_{XX}^{-1})_{kk} = \frac{1}{1 - R_k^2}$$

  - where R_k^2 is the coefficient of multiple determination of X_k regressed versus all other $p-2$ variables.

- In standardized regression (all X columns are standardized)

$$\begin{aligned} Var(\mathbf{b}^*) &= (\sigma^*)^2 \, \mathbf{r}_{X'X}^{-1} \\ Var(\mathbf{b}_k^*) &= (\sigma^*)^2 (r_{XX}^{-1})_{kk} = (\sigma^*)^2 \, \text{VIF}_k \end{aligned}$$

- VIF of 10 or more suggests strong multicollinearity
- Also compare mean VIF to 1

Weighted Least Squares

Downweight influential observations

- The weighted least squares method minimizes

$$Q_w = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{W}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

  - where $\mathbf{W} = diag\{w_1, \cdots, w_n\}$,

- By taking a derivative of Q_w, obtain normal equations:

$$(\mathbf{X}'\mathbf{W}\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{W}\mathbf{Y}$$

- Solution of the normal equations:

$$b = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{Y}$$

- Can also be viewed as solution for unequal variance scenario

Unequal Error Variances

- Consider $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\sigma}^2(\boldsymbol{\varepsilon}) = \mathbf{W}^{-1}$
  - Potentially correlated errors and unequal variances
- Special case: $\mathbf{W} = diag\{w_1, w_2, \cdots, w_n\}$
  - Heterogeneous variance or *heteroscedasticity*
  - Homogeneous variance or *homoscedasticity* if $w_1 = w_2 = \cdots = w_n = 1/\sigma^2$
  - Least square estimation still yields unbiased estimation, but is no longer optimal, and gives wrong uncertainty quantification
- Consider a transformation based on a known $\mathbf{W}$

$$\begin{aligned} \mathbf{W}^{1/2}\mathbf{Y} &= \mathbf{W}^{1/2}\mathbf{X}\boldsymbol{\beta} + \mathbf{W}^{1/2}\boldsymbol{\varepsilon} \\ &\downarrow \\ \mathbf{Y}_w &= \mathbf{X}_w\boldsymbol{\beta} + \boldsymbol{\varepsilon}_w \end{aligned}$$

- Can show $\mathrm{E}(\boldsymbol{\varepsilon}_w) = 0$ and $\boldsymbol{\sigma}^2(\boldsymbol{\varepsilon}_w) = \mathbf{I}$

Connection

- Least square problem for $\mathbf{Y}_w, \mathbf{X}_w$

$$Q_w = (\mathbf{Y}_w - \mathbf{X}_w\boldsymbol{\beta})'(\mathbf{Y}_w - \mathbf{X}_w\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\beta)'\mathbf{W}(\mathbf{Y} - \mathbf{X}\beta)$$

- Must determine optimal weights
- Optimal weights $\propto$ 1/variance
- Methods to determine weights, if no prior information of variance
  - Find relationship between the absolute residual and another variable and use this as a model for the standard deviation
  - Instead of the absolute residual, use the squared residual and find function for the variance
  - Use grouped data or approximately grouped data to estimate the variance

Ridge Regression as Multicollinearity Remedy

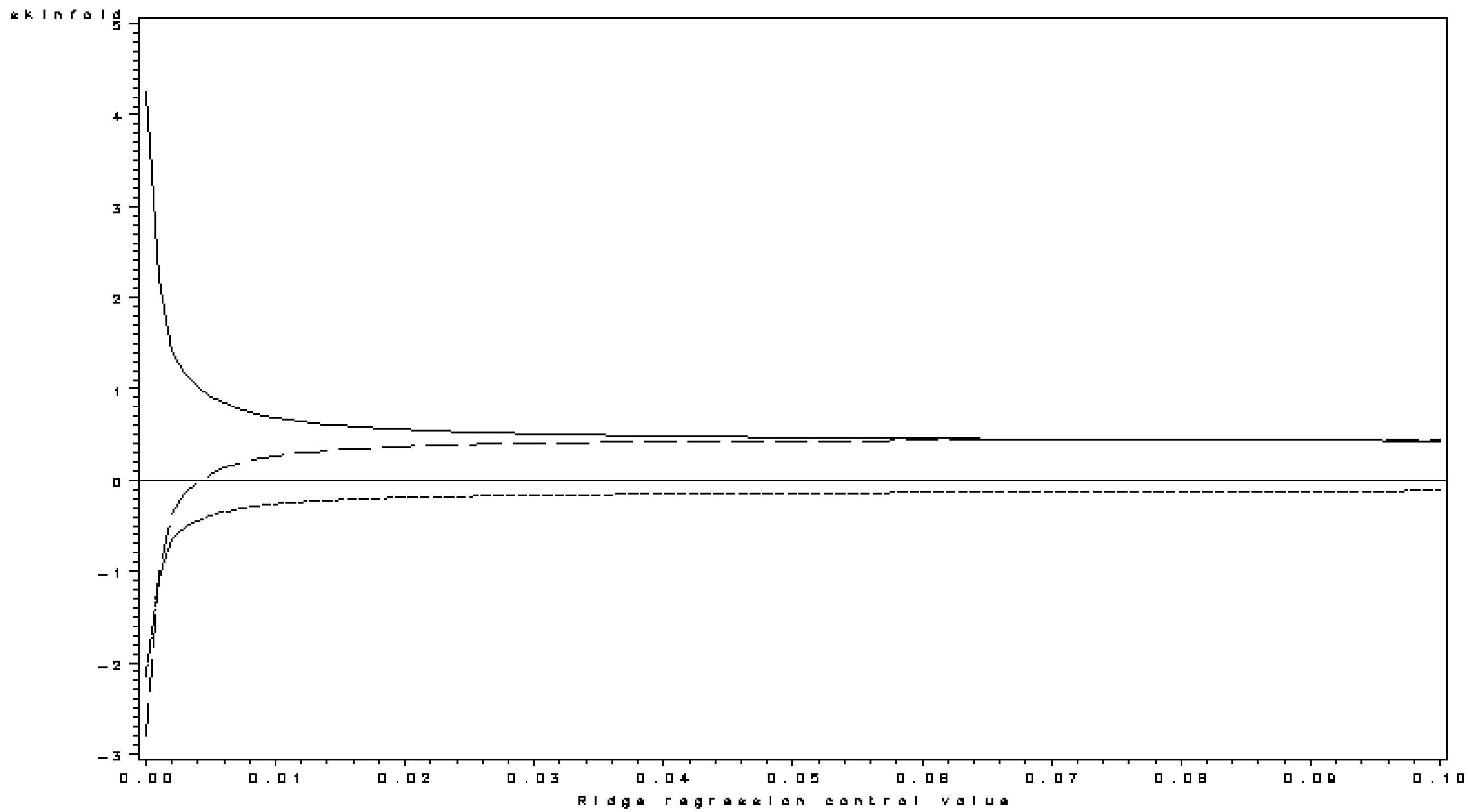
- Modification of least squares that overcomes multicollinearity problem
- Recall least squares suffers because $(\mathbf{X}'\mathbf{X})$ is almost singular thereby resulting in highly unstable parameter estimates
- Ridge regression results in biased but more stable estimates
- After standardizing data, we consider the correlation transformation so the normal equations are given by $\mathbf{r}_{XX}\mathbf{b} = \mathbf{r}_{YX}$. Since $\mathbf{r}_{XX}$ difficult to invert, we add a bias constant, c.

$$\mathbf{b}^R = (\mathbf{r}_{XX} + c\mathbf{I})^{-1}\mathbf{r}_{YX}$$

We then tranform it back to coefficient estimators for the orignal data.

Choice of c

- Key to approach is choice of c
- Common to use the *ridge trace* and VIF's
  - Ridge trace: simultaneous plot of $p-1$ parameter estimates for different values of $c \geq 0$. Curves may fluctuate widely when c close to zero but eventually stabilize and slowly converge to 0.
  - VIF's tend to fall quickly as c moves away from zero and then change only moderately after that
- Choose c where things tend to "stabilize"



skinfold
5
4
3
2
1
0
-1
-2
-3
0.00 0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09 0.10
Ridge regression control value

Chapter Review

- Multiple linear regression
- Estimation and inferences
- Diagnose and Remedy