ST5202: Applied Regression Analysis

Department of Statistics and Applied Probability
National University of Singapore

26-March-2018
Lecture 9

Announcement

Assignment $\#4$ due today.

Lecture 9

Model Validation and Diagnostics (Ch. 10)

Outline

- Model validation
- Added-variable plots
- Outlying observations
- Influential observations
- Multicollinearity and Ridge regression

Model Validation

Many possible choices, no universally accepted paradigm. Some (classical) possibilities:

- Prediction error based criteria (CV)
- Information criteria (AIC, BIC, etc.)
- Mallow's C_p statistic

Before looking at these, let's introduce terminology:
suppose that the true model is $\mathbf{y} = X\beta + \epsilon$ but with $\beta_r = 0$ for some subset β_r of β

- The *true* model contains only the columns for which $\beta_r \neq 0$
- A *correct* model is the true model plus extra columns
- A wrong model is a model that does not contain all the columns of the true model.

Expected Prediction Error

We may wish to choose a model by minimizing the error we make on average, when predicting a future observation given our model.
Our "experiment" is:

- Design matrix $\mathbf{X}$
- response $\mathbf{y}$ at $\mathbf{X}$

Every model f, will yield fitted valued $\hat{\mathbf{y}}(f) = \mathbf{H_f}\mathbf{y}$. And suppose we now obtain new independent response $\mathbf{y}_+$ for the same "experimental setup" $\mathbf{X}$.
Then, one approach is to select the model

$$f^* = \arg \min_{f \in 2^{|\mathbf{X}|}} \underbrace{\frac{1}{n} E\left\{ ||\mathbf{y}_+ - \hat{\mathbf{y}}(f)||^2 \right\}}_{\Delta(f)}$$

The bias/variance trade-off

Let $\mathbf{X}$ be a design matrix, and let $\mathbf{X}_\diamondsuit(n \times p)$ and $\mathbf{X}_\heartsuit(n \times q)$ be matrices built using columns of $\mathbf{X}$. Suppose that the true relationship between $\mathbf{y}$ and $\mathbf{X}$ is given by

$$\mathbf{y} = \underbrace{X_\heartsuit \beta}_{\mu} + \epsilon$$

but we use the matrix $\mathbf{X}_\diamondsuit$ instead of $\mathbf{X}_\heartsuit$ (i.e., we fit a different model). Therefore our fitted values are

$$\hat{\mathbf{y}} = (\mathbf{X}_\diamondsuit^\mathbf{T}\mathbf{X}_\diamondsuit)^{-1}\mathbf{X}_\diamondsuit^\mathbf{T}\mathbf{y} = \mathbf{H}_\diamondsuit \mathbf{y}$$

Now suppose that we obtain new observation $\mathbf{y}_+$ corresponding to the same design $\mathbf{X}$

$$\mathbf{y}_+ = \mathbf{X}_\heartsuit \beta + \epsilon_+ = \mu + \epsilon_+$$

Then observe that

$$\mathbf{y}_+ - \hat{\mathbf{y}} = \mu + \epsilon_+ - \mathbf{H}_\diamondsuit(\mu + \epsilon) = (\mathbf{I} - \mathbf{H}_\diamondsuit)\mu + \epsilon_+ - \mathbf{H}_\diamondsuit \epsilon$$

The bias/variance trade-off

It follows that

$$\begin{aligned} ||\mathbf{y}_+ - \hat{\mathbf{y}}(f)||^2 &= (\mathbf{y}_+ - \hat{\mathbf{y}}(f))^T(\mathbf{y}_+ - \hat{\mathbf{y}}(f)) \\ &= \mu^{\mathbf{T}}(\mathbf{I} - \mathbf{H}_\diamondsuit)\mu + \epsilon^{\mathbf{T}}\mathbf{H}_\diamondsuit\epsilon + \epsilon_+^{\mathbf{T}}\epsilon_+ + [\text{cross term}] \end{aligned}$$

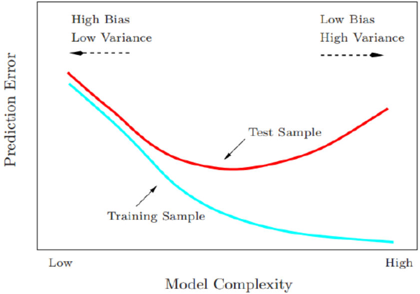
Since $E\{\text{cross term}\} = 0$, we observe the following

$$\Delta = \begin{cases} n^{-1}\mu^{\mathbf{T}}(\mathbf{I} - \mathbf{H}_\diamondsuit)\mu + (1 + p/n)\sigma^2 & \text{if model is wrong} \\ (1 + p/n)\sigma^2 & \text{if model is correct} \\ (1 + q/n)\sigma^2 & \text{if model is true} \end{cases}$$

Selecting a correct model instead of the true model brings in additional variance.
Selecting a wrong model instead of the true model results in bias.
Must find a balance!

Test and training error as a function of model complexity

- The training error tends to decrease whenever we increase the model complexity, that is , whenever we fit the data harder.
- However, with too much fitting, the model adapts itself too closely to the training data, and will not generalize well (i.e., have large test error)



Cross validation

Impossible to calculate Δ (depends on unknowns...). Must find a proxy (estimator) $\hat{\Delta}$. Suppose that n is large so that we can split the data in two pieces:

- $\mathbf{X}^*, \mathbf{y}^*$ used to estimate the model
- $\mathbf{X}', \mathbf{y}'$ used to estimate the prediction error for the model

The estimator of the prediction error will be

$$\hat{\Delta} = (n')^{-1}||\mathbf{y}' - \mathbf{X}'\hat{\beta}^*||^2$$

In practice n is small and we cannot afford to split the data. Instead we use the *leave-one-out* cross validation sum of squares:

$$n\hat{\Delta}_{CV} = CV = \sum_{j=1}^{n}(\mathbf{y_j} - \mathbf{X_j^T}\hat{\beta}_{-\mathbf{j}})^2$$

where $\hat{\beta}_{-j}$ is the estimate produced when dropping the jth case

Cross validation

No need to perform n regressions since

$$CV = \sum_{j=1}^{n} \frac{(\mathbf{y_j} - \mathbf{X_j^T}\hat{\beta})^2}{(1 - h_{jj})^2}$$

so the full regression may be used. Alternatively, one may use a more stable version:

$$GCV = \sum_{j=1}^{n} \frac{(\mathbf{y_j} - \mathbf{X_j^T}\hat{\beta})^2}{(1 - \mathrm{trace}(\mathbf{H})/n)^2}$$

where "G" stands for "generalized". It holds that:

$$E\{GCV\} = \frac{\mu^{\mathbf{T}}(\mathbf{I} - \mathbf{H})\mu}{(1 - p/n)^2} + \frac{n\sigma^2}{1 - p/n} \approx n\Delta$$

Suggests strategy: pick variables that minimize (G)CV

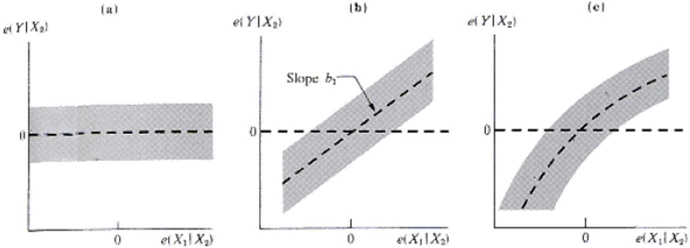
Added-variable plots

- Goal is to examine the marginal relationship between X_k and Y, given that other predictor variables are already in the model for Y
- Fit 3 models (also discussed in Chapter 3):
  - Model(a): $Y \sim X_{-k}$, denote residuals with $e(Y|X_{-k})$
  - Model (b): $X_k \sim X_{-k}$, denote residuals with $e(X_k|X_{-k})$
  - Model (c): $e(Y|X_{-k}) \sim e(X_k|X_{-k})$
- In (c) we are modeling the part of Y that is not explained by the other predictors X_{-k}, with the part of X_k that is not explained by X_{-k} (doesn't work if relation(s) between Y and X_{-k} have been misspecified)

Added-variable plots for a simple case

- The slope of the partial regression of $e_i(y|X_2)$ on $e_i(X_1|X_2)$ is *equal* to the estimated regression coefficient b_1 of X_1 in the multiple regression model $y = b_0 + b_1 X_1 + b_2 X_2 + \epsilon$.
- Thus the added-variable plot allows one to isolate the role of the specific independent variable in the multiple regression model.
- In practice one scrutinizes the plot patterns such as the ones shown in the next slide.

Prototype added variable plots

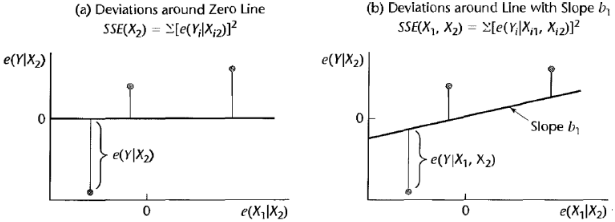


Prototype added variable plots–continued

- Plot of $e_i(Y|X_{-k})$ against $e_i(X_k|X_{-k})$ is called the added variable plot (for association between X_k and Y, after controlling for X_{-k})
- If linear relation is appropriate, then what's SSR and the regression coefficient in model (c)?
    - $\text{SSR} = \text{SSR}(X_k|X_{-k})$
    - $R^2 = R^2_{Yk|-k}$
    - The regression coefficient for $e_i(X_k|X_{-k})$ is the regression coefficient of X_k in model $Y \sim X_{-k} + X_k$

Illustration of deviation in an added-variable plot

$$\hat{Y}_i(X_2) = b_0 + b_2 X_{i2}$$
$$e_i(Y|X_2) = Y_i - \hat{Y}_i(X_2)$$
$$\hat{X}_{i1}(X_2) = b_0^* + b_2^* X_{i2}$$
$$e_i(X_1|X_2) = X_{i1} - \hat{X}_{i1}(X_2)$$



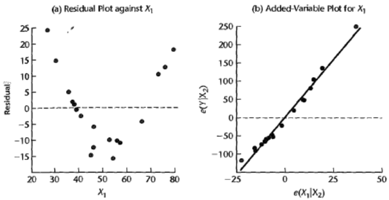
Example–life insurance

A few facts: $R^2_{Y1|2} = 0.984$ and $r_{12} = 0.254$

$$\hat{Y} = -205.72 + 6.2880X_1 + 4.738X_2$$

$$\hat{Y}(X_2) = 50.70 + 15.54X_2$$

$$\hat{X}_1(X_2) = 40.779 + 1.718X_2$$



Unusual data points

- Univariate outlier:
  unusual value for one of the X's or for Y
- In regression analysis:
  - Y is an outlier if the value of Y conditional on X's is unusual
  - a combination of predictor variables is an outlier if it has one or more unusual X values, and/or an unusual combination of X's
- Y outliers are called regression outliers
- X outliers are called leverage points

How to find regression outliers?

- Approach: examine the residuals $e_i = Y_i - \hat{Y}_i$
- Semi-studentized residuals from Ch. 3:

$$e_i^* = \frac{e_i}{\sqrt{MSE}}$$

- Refine:
  Internally studentized residuals

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

do these residuals have constant variance?

How to find regression outliers?

$$
\begin{aligned}
\mathbf{e} &= (\mathbf{1}-\mathbf{H})\mathbf{y} \\
\sigma^2(\mathbf{e}) &= \sigma^2(\mathbf{I}-\mathbf{H}) \\
Var(e_j) &= \sigma^2(1-h_{ii}) \\
Cov(e_i, e_j) &= \sigma^2(-h_{ij}), \text{ for } i \neq j
\end{aligned}
$$

- However, an outlying Y value might draw the fitted response function more towards itself, thus may note be detectable using residuals or studentized residuals

How to find regression outliers?

- Deleted residuals
$$d_i = Y_i - \hat{Y}_{i(i)} = \frac{e_i}{1 - h_{ii}} \text{(no need for re-computation)}$$
with $\hat{Y}_{i(i)}$ fitted mean response without using observation i
- Variance $Var\{d_i\} = \frac{\sigma^2}{1-h_{ii}}$, estimate σ^2 by $MSE_{(i)}$ (MSE based on model without using observation i):
$$s\{d_i\} = \sqrt{\frac{MSE_{(i)}}{1 - h_{ii}}} \text{ (recall: } s^2\{d_i\} = MSE_{(i)}(1 + \mathbf{X_i^T}(\mathbf{X_{(i)}^T}\mathbf{X_{(i)}})^{-1}\mathbf{X_i})$$
- t_I's are called the externally studentized residuals:
$$t_i = \frac{d_i}{s\{d_i\}} \sim t(n - p - 1)$$
(why $(n - p - 1)$ degrees of freedom?)
- Compare t_i to $t(1 - \alpha/2n, n - p - 1)$ which adjusts for the n comparisons for n observations by Bonferroni

Externally studentized residuals

- Non-independence:

$$t_i = \frac{e_i}{\sqrt{MSE_{(i)}(1 - h_{ii})}}$$

- To avoid having to fit the model without case i to get $MSE_{(i)}$:

$$(n - p)MSE = (n - p - 1)MSE_{(i)} + e_i^2/(1 - h_{ii})$$

- The externally studentized residuals are then given by:

$$t_i = e_i \left(\frac{n - p - 1}{SSE(1 - h_{ii}) - e_i^2} \right)^{1/2}$$

Body fat example

- Externally studentized residuals

i	(1) e_i	(2) h_{ii}	(3) t_i
1	−1.683	.201	−.730
2	3.643	.059	1.534
3	−3.176	.372	−1.656
4	−3.158	.111	−1.348
5	.000	.248	.000
6	−.361	.129	−.148
7	.716	.156	.298
8	4.015	.096	1.760
9	2.655	.115	1.117
10	−2.475	.110	−1.034
11	.336	.120	.137
12	2.226	.109	.923
13	−3.947	.178	−1.825
14	3.447	.148	1.524
15	.571	.333	.267
16	.642	.095	.258
17	−.851	.106	.344
18	−.783	.197	.335
19	−2.857	.067	−1.176
20	1.040	.050	.409

Body fat example–continued

- The estimated function (see lecture 8)

$$\hat{Y} = -19.174 + .2224X_1 + .6594X_2$$

- For $X_{11} = 19.5$ and $X_{12} = 43.1$,we have

$$\hat{Y}_1 = -19.174 + .2224(19.5) + .6594(43.1) = 13.583$$

- The residual $e_1 = 11.9 - 13.583 = -1.683$
- We find, given $SSE = 109.95$ from lecture 8

$$t_i = -1.683\left(\frac{20-3-1}{109.95(1-.201)-(-1.683)^{1/2}}\right)^{1/2} = -.730$$

Body fat example–continued

- Test case 13 using Bonferroni at $\alpha = 0.10$
- $|t_{13}| = 1.825 \leq 3.252$, conclude that the case 13 is not an outlier

$$t(1 - \alpha/2n; n - p - 1) = t(0.9975; 16) = 3.252$$

- Still would like to see if case 13 is influential (why and how)

Outlying X observations (leverage points)

- Use **H** to identify outlying X observations:
  - h_{ii} is a measure of the distance between X values for the i^{th} case and the means of the X values for all n cases
  - large h_{ii} indicates that i^{th} case is far away from center of all X observations
  - h_{ii} is called the leverage of the i^{th} case
- A point with high leverage will draw the fitted response function more towards itself, as $\hat{Y}_i = \sum_j h_{ij} Y_j$
- The following holds true:

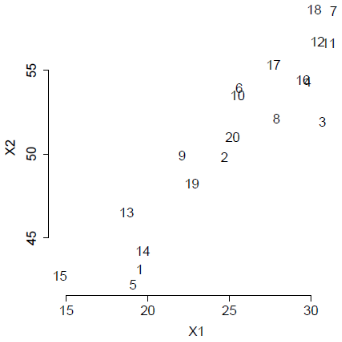
$$1/n \leq h_{ii} \leq 1, \;\; \bar{h} = \frac{\sum_i h_{ii}}{n} = \frac{p}{n}$$

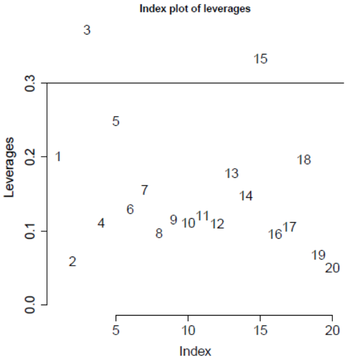
- Leverage $> 2p/n$ indicates outlying case with regard to X values
- For a new observation, measure for distance to observed cases:

$$h_{new,new} = \mathbf{X}_{\mathbf{new}}^{\mathbf{T}}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}_{\mathbf{new}}$$

$h_{new,new}$ larger than observed h_{ii}'s indicates extrapolation

Body fat example–continued





Influential data points

- A case is influential
  - if it as "a large" influence on the fitted regression line, on the estimated regression coefficients
  - if excluding it causes "major" changes in the fitted regression function
- Influence = Leverage $\times$ "Outlyingness"
- Different measures for identifying influential cases, each based on the omission of a single case to measure its influence
- Note:
  Diagnostics based on leaving i^{th} case don't work if there are more outliers in same area (consider leaving out several cases simultaneously)

Influence on single fitted values

- Influence on single fitted value $\hat{Y}_i$:

$$(DFFITS)_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)} h_{ii}}} = t_i \left(\frac{h_{ii}}{1 - h_{ii}} \right)^{1/2}$$

- ("Outlyingness $\times$ leverage")
- Rule of thumb:
  Case is influential if DFFITS exceeds 1 for small to medium data sets, and $2\sqrt{p/n}$ for large data sets

Body fat example–DFFITS

- $(DFFITS)_3$ for case 3 with $t_3 = -1.656$ and $h_{33} = 0.372$

$$(DFFITS)_3 = -1.656\left(\frac{.372}{1-.372}\right)^{1/2} = -1.27$$

- Case 3 is influential as $|(DFFITS)_3| > 1$, but might not be influential enough to require remedial action

Influence on all fitted values

- Cook's distance:

$$D_i = \frac{\sum_{j=1}^{n} \left(\hat{Y}_j - \hat{Y}_{j(i)} \right)^2}{pMSE} = \frac{e_i^2}{pMSE(1 - h_{ii})} \cdot \frac{h_{ii}}{(1 - h_{ii})}$$

  - large e_i and only moderate leverage h_{ii}
  - large leverage h_{ii} and only a moderately sized e_i
  - both a large e_i and a large leverage h_{ii}
- "Outlyingness" $\times$ leverage
- Compare D_i to $F(p, n-p)$ distribution:
  - i^{th} case is influential if percentile $F(D_i; p, n-p) > 0.5$, if n is moderately large (can be sensitive)

Body fat example–Cook's distance

- D_3 for case 3 with $e_3 = -3.176$ and $h_{33} = 0.372$ and $MSE = 6.47$ (lecture 8)

$$D_3 = \frac{(-3.176)^2}{3(6.47)} \left(\frac{0.372}{(1-0.372)^2} \right) = 0.490$$

- Substantially larger than the second largest one $D_{13} = 0.212$
- .490 is the 30.6th percentile of $F(p, n-p) = F(3, 20-3)$, given the fact from R that
  "qf(0.306,3,17) = 0.4897561"
- Case 3 appears as an influential point but not quite substantial that needs remedial action

Influence on regression coefficients

- Influence on b_k:

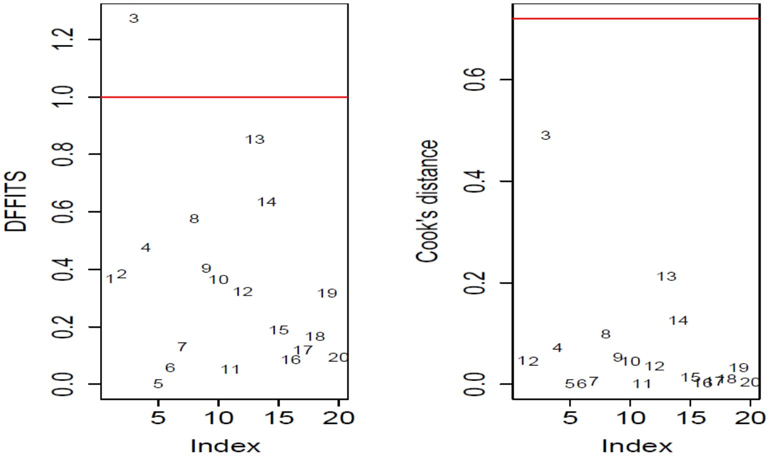
$$(DFBETAS)_{k(i)} = \frac{b_k - b_{k(i)}}{\sqrt{MSE_{(i)}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}_{[kk]}}}$$

- Rule of thumb:
  case is influential if absolute $(DFBETAS) > 1$ for small/medium data sets, and $> 2/\sqrt{n}$ for large data sets.

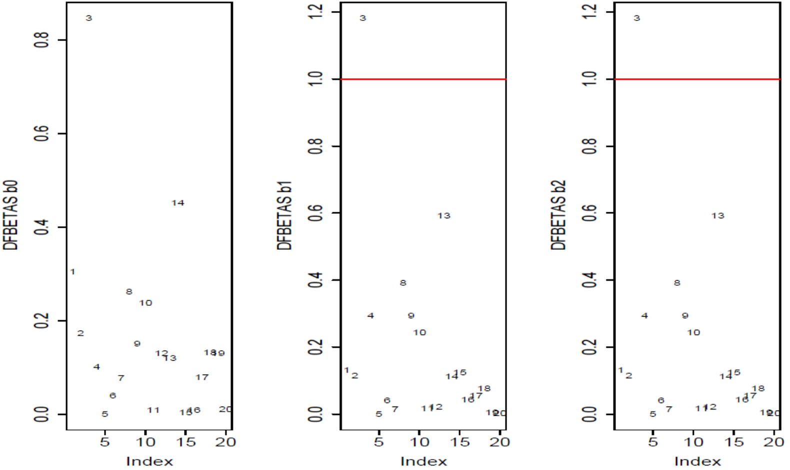
Body fat example–continued

	(1)	(2)	(3)	(4) DFBETAS	(5)
i	$(DFFITS)_i$	D_i	b_0	b_1	b_2
1	−.366	.046	−.305	−.132	.232
2	.384	.046	.173	.115	−.143
3	−1.273	.490	−.847	−1.183	1.067
4	−.476	.072	−.102	−.294	.196
5	.000	.000	.000	.000	.000
6	−.057	.001	.040	.040	−.044
7	.128	.006	−.078	−.016	.054
8	.575	.098	.261	.391	−.333
9	.402	.053	−.151	−.295	.247
10	−.364	.044	.238	.245	−.269
11	.051	.001	−.009	.017	−.003
12	.323	.035	−.131	.023	.070
13	−.851	.212	.119	.592	−.390
14	.636	.125	.452	.113	−.298
15	.189	.013	−.003	−.125	.069
16	.084	.002	.009	.043	−.025
17	−.118	.005	.080	.055	−.076
18	−.166	.010	.132	.075	−.116
19	−.315	.032	−.130	−.004	.064
20	.094	.003	.010	.002	−.003

Body fat example–continued



Body fat example–continued



What to do with unusual data?

- Check for data entry errors
- Think of reasons why the observations might be different:
  - Are the influential cases part of your population interest
- Change the model
- Fit the model with and without the observations to see the effect
- Robust regression(Ch.11)
- Don't throw them out without thinking, example...
  - In 1985, British Antarctic service observed a large decrease in atmospheric ozone over the Antarctic
  - In 1985, NASA Numbus 7 satellite had been recording atmospheric information for several years. However, they didn't discover the hole: low values had been excluded automatically, assuming that they were mistakes, thus delaying the discovery of the Antarctic ozone hole for several years

Multicollinearity and Ridge regression(Ch. 7, 10, and 11)

- Ch. 7: when using a correlation transformation of Y and X's, the normal equations are given by:

$$\mathbf{r_{XX}} \mathbf{b}^* = \mathbf{r_{XY}}$$

- Inverse of correlation matrix instable if the X's are highly correlated
- Problems:
  - Variance estimation of the coefficients may become very large
  - Regression coefficients may change signs
  - Marginal significance highly depends on which predictor variables are included in the model
  - Significance may be masked by correlated variables in the model
- Use "variance inflation factor" to detect multicollineariry

Variance inflation factor (Ch. 10)

- VIF = how much the variances of the b_k's are inflated as compared to when the predictor variables are not linearly related
- VIF for predictor variable k:

$$(VIF)_k = \frac{1}{1 - R_k^2}$$

with R_k^2 the R-squared when regressing X_k on the other X_{-k} predictor variables. This is derived from $\sigma^2\{b_k^*\}$:

$$\sigma^2\{b_k^*\} = (\sigma^*)^2 \, [\mathbf{r_{XX}}^{-1}]_{[kk]} = (\sigma^*) \, (VIF)_k$$

- Rules of thumb for diagnosing serious multicolinearity:
  - $\max\left((VIF)_k\right) > 10$
  - mean VIF considerably larger than 1

Ridge regression

- Main idea:
  Make the $b_k's$ slightly biased, to reduce their variance (plot!)
- Ridge regression with constant $c > 0$:

$$
\begin{aligned}
(\mathbf{r_{XX}} + \mathbf{cI})\mathbf{b}^R &= \mathbf{r}_{XY} \\
\mathbf{b}^R &= (\mathbf{r}_{XX} + c\mathbf{I})^{-1}\mathbf{r}_{XY}
\end{aligned}
$$

- How to choose c?
  - Empirical evaluation of the ridge trace (changes in b_k^R's) and the variance inflation factors for different c values
  - (Use bootstrap method to evaluate the variance, Ch 11)

Example of Ridge regression

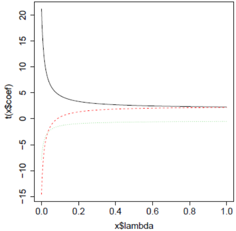
- Body fat (Y) against skinfold thickness, thigh and midarm circumference:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   117.085     99.782   1.173    0.258
Triceps         4.334      3.016   1.437    0.170
Thigh          -2.857      2.582  -1.106    0.285
Midarm         -2.186      1.595  -1.370    0.190
```

- VIF's:

```
Triceps       Thigh       Midarm
708.8 564.3 104.6
```

Example of Ridge regression for body fat example



- Ridge trace using "lm.ridge" in R":
  - λ is ridge constant on original scale
  - Coefficients are the standardized coefficients: they stabilize around $\lambda = 0.3$
- To get the coefficients on the original scale:

```
lm.ridge(y ~ ., lambda = 0.3, data = bf[,-4])
Triceps      Thigh      Midarm
0.6   0.3 -0.2
```

Ridge regression–continued

Find $\hat{\gamma}$ that minimizes

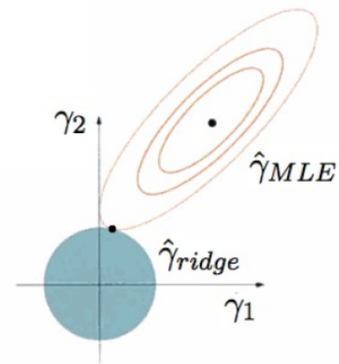
$$\sum_{i=1}^{n}(y_i - \mathbf{X_i^T}\gamma)^2 + \lambda \sum_{j=1}^{p} \gamma_j^2$$

The solution to the Ridge regression problem is given by

$$\hat{r}^{ridge} = (\mathbf{X^T X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{Y}$$

- $\lambda \to 0$, $\hat{\gamma}^{ridge} \to \hat{\gamma}^{OLS}$
- $\lambda \to \infty$, $\hat{\gamma}^{ridge} \to 0$

L_2 Shrinkage (Ridge regression)



Lasso

Motivated from Ridge regression we can consider a formulation: find $\hat{\gamma}$ that minimizes

$$\sum_{i=1}^{n}(y_i - \mathbf{X}_i^T \gamma)^2 + \lambda \sum_{j=1}^{p} |\gamma_j|$$

L_1 penalty (almost) produces a "continuous" model selection!
Shinks coefficient size by different version of magnitude:

- Resulting estimator non-linear in $\mathbf{y}$
- Why choose a different type of norm?

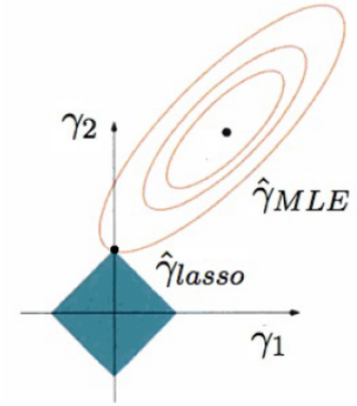
L_1 norm induces "sharp" balls

Extreme case: L^0 "Norm", gives best subsets selection!

$$\|\gamma\|_0 = \sum_{j=1}^{p-1} |\gamma_j|^0 = \sum_{j=1}^{p-1} \mathbf{1}_{\{\gamma_j \neq 0\}} = \#\{j : \gamma_j \neq 0\}$$

Generally: $\|\gamma\|_p^p = \sum_{j=1}^{p-1} |\gamma_j|^p$, sharp balls for $0 < p \leq 1$

Lasso profile for body fat data



L_1 Shrinkage (Lasso regression)

LASSO and CV for different values of $r(\lambda)/\|\hat{\gamma}\|_1$