Regression Analysis I

Multiple Linear Regression

Hossein Moradi Rekabdarkolaee

South Dakota State University

email: *hossein.moradirekabdarkolaee@sdstate.edu*.

Office: Arch, Math, and Engineering Building, *Room: 254*.

Multiple Linear Regression

- **Model**:
  - Let y denotes the response variable and we have p independent variable $x_1, x_2, \ldots, x_k$.
  - There are n subjects in the experiment, and for each subject the dependent and independent variables are measured.
- The multiple linear regression model is given by

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \epsilon_i$$

  where $i = 1, 2, \ldots, n$.
- This model is linear in the parameters.
- This model has $p = k + 1$ unknown parameters.

Multiple Linear Regression

- The data of the experiment can be expressed as

Y	X_1	X_2	$\dots$	X_k
y_1	x_{11}	x_{12}	$\dots$	x_{1k}
y_2	x_{21}	x_{22}	$\dots$	x_{2k}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
y_n	x_{n1}	x_{n2}	$\dots$	x_{nk}

- The model can be written in matrix notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Multiple Linear Regression

- The model can be written in matrix notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = [\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k] = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix},$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

Multiple Linear Regression

- The **X** matrix is referred to as the model or data matrix.
- **X** is a $n \times p$ matrix.
- **y** is a $n \times 1$ vector of response.
- β is a $p \times 1$ vector of the regression coefficients.
- ϵ is a $n \times 1$ vector of the error.
- $\mathbf{y} = \mathbf{X}\beta + \epsilon$ is the linear regression model.

Multiple Linear Regression

- **Assumptions**:
  - The assumptions for a multiple linear regression model are very similar to those for a simple linear regression model.
  - All independent variable observations are nonrandom and measured with negligible error.
  - These variables also could be transformations of observed variables or interaction variables created by multiplying two or more variable together.
  - The key is that the model must remain linear in the parameter.

Multiple Linear Regression

- **Assumptions**:
  - The covariate matrix is a full ranked non-stochastic matrix, i.e. $rank(\mathbf{X}) = p + 1$
  - The ϵ_i are the random error components of the model and have the following assumptions made about them
    - $E(\epsilon_i) = 0$, implying that the model is appropriate.
    - $Var(\epsilon_i) = \sigma^2$, implying homogeneous variance.
    - $E(\epsilon_i \epsilon_j) = 0$ for $i \neq j$, implying uncorrelated errors.
  - In matrix notation, the assumption are: $E(\epsilon) = \mathbf{0}$ and $Var(\epsilon) = \sigma^2 \mathbf{I}_n$, where $\mathbf{I}_n$ is a $n \times n$ identity matrix.
  - $Var(\epsilon)$ is referred to as a variance-covariance matrix of the random errors.
  - A distribution assumption on ϵ is often made (although not necessary), in particular $\epsilon \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$.

Multiple Linear Regression

- These assumptions are used to study the statistical properties of estimator of regression coefficients.
- The following assumption is required to study particularly the large sample properties of the estimators:

$$lim_{n\to\infty}\left(\frac{\mathbf{X}^T\mathbf{X}}{n}\right) = \Delta$$

exists and is a non-stochastic and nonsingular matrix (with finite elements).

Multiple Linear Regression

- The multiple linear regression model is given by

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \epsilon_i$$

- In multiple regression, the estimated model is

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \ldots + \hat{\beta}_k x_{ki}$$

- Thus, we wish to minimize error given a particular measure.
- A general procedure for the estimation of regression coefficient vector is to minimize

$$\sum_{i=1}^{n} M(\epsilon_i) = \sum_{i=1}^{n} M(y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \ldots - \beta_k x_{ki})$$

for a suitably chosen function M.

Multiple Linear Regression

- Some examples of choice of M are

$$\begin{aligned} M(\epsilon_i) &= |\epsilon_i| \\ M(\epsilon_i) &= \epsilon_i^2 \\ M(\epsilon_i) &= |\epsilon_i|^p, \quad \text{in general} \end{aligned}$$

- If we consider the least squares approach then the objective function becomes

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \ldots + \hat{\beta}_k x_{pi})^2.$$

- We wish to find the estimates for coefficients that minimizes the above objective function.

Multiple Linear Regression (Least Squares Method)

- The model can be written in matrix notation as

$$\mathbf{y} = \mathbf{X}\beta + \epsilon,$$

- Thus the objective function is

$$f(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)$$

- To find the OLS estimates, we take the partial derivative of $(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)$ with respect to β, set it equal 0, and solve for β.

$$\frac{\partial f(\beta)}{\partial \beta} = \frac{\partial (\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\beta)}{\partial \beta} = -2\mathbf{X}^T\mathbf{y} + 2(\mathbf{X}^T\mathbf{X})\beta.$$

- Therefore, $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.

Multiple Linear Regression (Least Squares Method)

- In order to find that the estimate is actually minimize the objective function, we need to take the second derivative

$$\frac{\partial^2 f(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = 2(\mathbf{X}^T\mathbf{X})$$

  which is atleast non-negative definite.

- Thus, $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ minimizes the objective function $f(\boldsymbol{\beta})$

- The equation

$$\frac{\partial f(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0 \Rightarrow (\mathbf{X}^T\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y}$$

  is called normal equation.

Multiple Linear Regression (OLS properties)

- Since it is assumed that $rank(\mathbf{X}) = p$ (full rank), then $(\mathbf{X}^T\mathbf{X})$ is positive definite and unique solution of normal equation exists.
- Expected Value

With $\mathbf{y} = \mathbf{X}\beta + \epsilon$ and $E(\epsilon) = 0$, then

$$E(\mathbf{y}) = E(\mathbf{X}\beta + \epsilon) = \mathbf{X}\beta,$$

and

$$\begin{aligned} E(\hat{\beta}) &= E[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}] \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T E(\mathbf{y}) \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\beta \\ &= \mathbf{I}\beta \\ &= \beta. \end{aligned}$$

Multiple Linear Regression (OLS properties)

- With $var(\epsilon) = E(\epsilon\epsilon^T) = \sigma^2 \mathbf{I}_n$, we have

$$var(\mathbf{y}) = var(\mathbf{X}\beta + \epsilon) = var(\epsilon) = \sigma^2 \mathbf{I}_n.$$

- For the coefficients, we have

$$\begin{aligned} var(\hat{\beta}) &= var[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}] \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T var(\mathbf{y})[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]^T \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\sigma^2\mathbf{I}_n\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \end{aligned}$$

- Note that this contains two components: the model error and variability of the data.
- This is called the variance-covariance of the coefficients.

Multiple Linear Regression (OLS properties)

- **Gauss-Markov Theorem:** Without a distribution assumption on error, OLS estimators have minimum variance (best) among all linear unbiased estimators (also called BLUE).
- In addition, when error are normally distributed, then OLS estimators are uniformly minimum variance unbiased estimators (UMVUE).
- Usin LS criterion, we have $\mathbf{e} = \mathbf{y} - \mathbf{H}\mathbf{y}$.
- In addition, $SSE = \mathbf{y}^T(\mathbf{I} - \mathbf{H})\mathbf{y}$
- The unbiased estimation for σ^2 is

$$\hat{\sigma}^2 = \frac{(\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta})}{n-p} = \frac{\mathbf{y}^T[\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]\mathbf{y}}{n-p}.$$

Multiple Linear Regression (Coefficients of determination)

- Similar to simple linear regression, the total variability in response can be partitioned as :

$$SS_{TOTAL} = SS_{MODEL} + SS_{ERROR},$$

and degrees of freedom can be partitioned as

$$Df_{TOTAL} = Df_{MODEL} + Df_{ERROR},$$

which is

$$n - 1 = (p - 1) + (n - p).$$

Multiple Linear Regression (Coefficients of determination)

- Given the partitioning of the total variability, the coefficient of determination is
$$R^2 = \frac{SS_{MODEL}}{SS_{TOTAL}}.$$
- In multiple linear regression R^2 always increases as more variables are added to the model.
- This is due to the fact that by adding more variables to the model, more and more of the variability of response is explained.
- Due to this fact, it is often incorrect to use R^2 as a test of model adequacy without first testing for the appropriateness of the covariates in the model.

Multiple Linear Regression (adjusted R^2)

- We have stated that the coefficient of determination always increases as additional regressors are added to the model.
- Hence we can always improve R^2 just by adding "junk" variables to the model.
- However, the increase in SS_{MODEL} is offset by an increase in $Df_{MODEL} = p - 1$.
- Even worse, the decrease in SS_{ERROR} is offset by a decrease in $Df_{ERROR} = n - p$.
- Since the inferential procedures involves Df_{ERROR}, to improve the power, we want this value to be as large as possible.
- Hence, $MS_{ERROR} = \frac{SS_{ERROR}}{Df_{ERROR}}$ is not predictable.

Multiple Linear Regression (adjusted R^2)

- Recall
$$R^2 = \frac{SSR}{SST}.$$
- One way to check for model adequacy while also making sure that all the variables are important is to incorporate the Df_{ERROR} and calculate the adjusted R^2 as
$$R^2_{adj} = 1 - \frac{SS_{ERROR}/Df_{ERROR}}{SS_{TOTAL}/Df_{TOTAL}}.$$
- This adjusted R^2 is not predictable and if the model is made more complicated, adjusted R^2 does not necessarily increase.
- In fact, adjusted R^2 moves the same way as the error mean square moves.

Multiple Linear Regression (Confidence Interval Estimation)

- A $100(1-\alpha)\%$ confidence interval for β_j is

$$\hat{\beta}_j \pm t_{n-p,\alpha/2}\hat{\sigma}\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{jj}}$$

- The Bonferroni joint confidence intervals can be used to estimate several regression coefficients simultaneously.

- If g parameters are to be estimated jointly (where $g \leq p$), the confidence limits with family confidence coefficient $1-\alpha$ are:

$$\hat{\beta}_j \pm t_{n-p,\alpha/2g}\hat{\sigma}\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{jj}}$$

Multiple Linear Regression (Confidence Interval Estimation)

- Mean prediction for a new point in the variable space, $\mathbf{x}_0$, is

$$\mu(\mathbf{x}_0) = E[y(\mathbf{x}_0)] = \mathbf{x}_0^T \beta,$$

- The standard error of prediction is

$$SE[\hat{y}(\mathbf{x}_0)] = \hat{\sigma}\sqrt{\mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0}.$$

- A $100(1-\alpha)\%$ confidence interval for Mean prediction is

$$\mathbf{x}_0^T \beta \pm t_{n-p,\alpha/2}\hat{\sigma}\sqrt{\mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0}.$$

Multiple Linear Regression (Confidence Interval Estimation)

- Predicted value for a single observation is $\mathbf{x}_{new}^T\beta$ which has standard error given by

$$\hat{\sigma}\sqrt{1+\mathbf{x}_{new}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_{new}}.$$

- A $100(1-\alpha)\%$ confidence interval for a single observation is

$$\mathbf{x}_{new}^T\beta \pm t_{n-p,\alpha/2}\hat{\sigma}\sqrt{1+\mathbf{x}_{new}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_{new}}.$$

- Note that the midpoints of the confidence and prediction intervals are the same.

Multiple Linear Regression (Confidence Region for Regression Surface)

- The confidence region for the entire regression surface is an extension of the Working-Hotelling confidence band

$$\hat{\mathbf{y}} \pm WSE[\hat{\mathbf{y}}]$$

where

$$W^2 = pF(p, n-p)$$

- **Simultaneous Confidence Intervals for Several Mean Responses**:
  - Use the Working-Hotelling confidence region bounds.
  - Use Bonferroni simultaneous confidence intervals.

Multiple Linear Regression (ANOVA)

- The sums of squares for the analysis of variance in matrix terms for multiple linear regression are

$$
\begin{aligned}
SS_{TOTAL} &= SST = \mathbf{y}^T\mathbf{y} - \left(\frac{1}{n}\right)\mathbf{y}^T\mathbf{J}\mathbf{y} = \mathbf{y}^T\left[\mathbf{I} - \left(\frac{1}{n}\right)\mathbf{J}\right]\mathbf{y} \\
SS_{MODEL} &= SSR = \hat{\beta}^T\mathbf{X}^T\mathbf{y} - \left(\frac{1}{n}\right)\mathbf{y}^T\mathbf{J}\mathbf{y} = \mathbf{y}^T\left[\mathbf{H} - \left(\frac{1}{n}\right)\mathbf{J}\right]\mathbf{y} \\
SS_{ERROR} &= SSE = \mathbf{e}^T\mathbf{e} = (\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta}) = \mathbf{y}^T\mathbf{y} - \hat{\beta}^T\mathbf{X}^T\mathbf{y} \\
&= \mathbf{y}^T\left[\mathbf{I} - \mathbf{H}\right]\mathbf{y}
\end{aligned}
$$

Table: Analysis of variance table for multiple linear regression.

Source	df	Sum of Squares (SS)	Mean of SS	F-value
Model	p-1	SSR	$MSR = \frac{SSR}{p-1}$	$\frac{MSR}{MSE}$
Error	n-p	SSE	$MSE = \frac{SSE}{n-p}$	
Total	n-1	SST		

Multiple Linear Regression (Global F-test)

- The covariates in the regression model (hopefully) explain most of the variability in the response.
- If this is the case, then the estimated regression model will make a good prediction equation of the response.
- We can measure the simultaneous significant of all p regressors in a single test called an overall or Global F-test.
- We want to test the null hypothesis $H_0: \beta_1 = \beta_2 = \ldots = \beta_p = 0$ versus research hypothesis H_1 : at least one β_j is not equal 0, for $j = 1, 2, \ldots, p$.

Multiple Linear Regression (Global F-test)

- If we fail to reject the null hypothesis, then we conclude that none of the regressors, influence the response variable and hence a regression model using these covariates is worthless.
- In this case we either:
  1. Consider transformation,
  2. determine if there are other covariates that can be used to predict the response, or
  3. end the analysis.
- If we reject the null hypothesis, then at least one of the covariates is influencing the response.
- In this case we want to continue with further analysis.

Multiple Linear Regression (Global F-test)

- The test statistics is

$$F_{obs} = \frac{MS_{MODEL}}{MS_{ERROR}},$$

which follows a F distribution with p-1 and n-p degrees of freedom.

- Since we want the model to explain a large part of the variability in response, we want MS_{MODEL} and hence F_{obs} to be large.
- Thus, we reject the null hypothesis if $F_{obs} > F_{p-1,n-p}$.
- The $P - value = P(F_{obs} > F_{p-1,n-p})$.

Multiple Linear Regression (Global F-test)

- The Global F-test does not give any information on which regressors are having the influence or on which are more influential than others.
- F is also insensitive to the quality of the fit itself. for instance, if F is small (fail to reject) then R^2 will also be small, indicating a poor fit. However, if F is large (reject), R^2 may still be quite small, indicating poor fitting model.
- As mentioned earlier, additional covariates leads to an increase in SS_{MODEL}. However, this also involves an increase in p. Thus, for the F-test to be more powerful, the increase in SS_{MODEL} should offset the increase in $Df_{MODEL} = p - 1$.
- A significant F-test indicate that the full model is useful in predicting the response.

Multiple Linear Regression (F-test on Subset of Regressors)

- The full model might contain some regressors that do not contribute to the prediction of the response.
- For this reason, we may want to test if there is a subset model that in terms of prediction capability, is statistically equivalent to the full model.
- The full regression model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- suppose we partition $\mathbf{X}$ and $\boldsymbol{\beta}$ as $\mathbf{X} = [\mathbf{X}_1 \;\; \mathbf{X}_2]$ and $\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix}$, where $\mathbf{X}_1$ is $n \times p_1$, $\mathbf{X}_2$ is $n \times p_2$, $\boldsymbol{\beta}_1$ is $p_1 \times 1$, and $\boldsymbol{\beta}_2$ is $p_2 \times 1$ with $p_1 + p_2 = p$.

Multiple Linear Regression (F-test on Subset of Regressors)

- The general (full) linear model can be written as

$$\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \epsilon.$$

- Suppose that we put all the questionable regressors in $\mathbf{X}_1$.
- Then we have a subset model: $\mathbf{y} = \mathbf{X}_2\beta_2 + \epsilon$, and we may wish to test $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$.
- A F-test for this hypothesis requires the calculation of the reduction sum of squares, which is the difference between the SS_{MODEL} for the full model and the SS_{MODEL} for the subset. model.

Multiple Linear Regression (F-test on Subset of Regressors)

- The reduction degrees of freedom associated with the reduction sum of squares is p_1.
- Note that this is the number of parameters in the null hypothesis.
- It is also the difference between the number of columns in the full model matrix (p) and the number of columns in the subset model matrix (p_2).
- Since the full model is currently the "accepted" model, the denominator of the F-test is $\hat{\sigma}^2 = MS_{ERROR}$ for the full model.

Multiple Linear Regression (F-test on Subset of Regressors)

- This denominator is independent of the numerator, so the test statistic for $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$ is

$$F_{obs} = \frac{(SS_{MODEL,FULL} - SS_{MODEL,Reduced})/p_1}{MS_{ERROR,FULL}}.$$

- This statistic is distributed as an F with p_1 and n-p degrees of freedom.

- If we fail to reject the null hypothesis then the full and reduced models are statistically equivalent.

- If we reject the null hypothesis, then at least one of the regressors in the full model that is not in the reduced model is significant and the full model is preferred to the reduced model.

Multiple Linear Regression (Partial t-Test)

- Suppose we have unordered regressors and want to test the $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$. This could be done via a partial F-test.
- For the partial F-test, we have

$$F_{obs} = \frac{SS_{MODEL,FULL} - SS_{MODEL,REDUCED}}{MS_{ERROR,FULL}}$$

  which has a $F_{DF_{REDUCED},n-p}$.
- In above case, reduced model will be a model that contains all the covariates but the j-th covariate.
- Thus, $DF_{REDUCED} = 1$.

Multiple Linear Regression (Partial t-Test)

- Since $var(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$, then the standard deviation of an estimate $\hat{\beta}_j$ is $\hat{\sigma}\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{jj}}$.
- As we know, $\hat{\sigma} = \sqrt{MS_{ERROR,FULL}}$.
- Thus, the above test could be made using a partial t-test with test statistics

$$t_{obs} = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{jj}}} \sim t_{n-p}.$$

- Note that $t^2_{n-p} = F_{1,n-p}$.

Multiple Linear Regression (Residuals)

- Recall that a residual is the difference between an observed response value and a predicted response.
- The residuals are items that mirror the complexion of the ϵ's.
- Therefore, they can be used as diagnostics regarding assumptions.
- Consider the vector of residuals

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\beta} = \mathbf{y} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = [\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]\mathbf{y}$$

- The matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is referred to as the HAT matrix.
- The diagonal elements of the hat matrix $h_{ii} = \mathbf{x}_i(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i^T$, are called the HAT diagonals.

Multiple Linear Regression (Residuals)

- If the model is correct, such that $E(\mathbf{y}) = \mathbf{X}\beta$, then

$$\begin{aligned} E(\mathbf{e}) &= E\{[\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]\mathbf{y}\} \\ &= [\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]E[\mathbf{y}] \\ &= [\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]\mathbf{X}\beta \\ &= [\mathbf{X} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}]\beta \\ &= [\mathbf{X} - \mathbf{X}]\beta \\ &= \mathbf{0}\beta \\ &= \mathbf{0}. \end{aligned}$$

Multiple Linear Regression (Residuals)

- If the model is correct, such that $E(\mathbf{y}) = \mathbf{X}\beta$, then

$$\begin{aligned} var(\mathbf{e}) &= var\{[\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]\mathbf{y}\} \\ &= [\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]var[\mathbf{y}][\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]^T \\ &= [\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]\sigma^2\mathbf{I}[\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]^T \\ &= \sigma^2[\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]^T. \end{aligned}$$

- Therefore, $var(\mathbf{e}) \neq \sigma^2\mathbf{I}$, but instead $var(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H})$.
- Then $var(e_i) = \sigma^2(1 - h_{ii}) \neq \sigma^2$ (variance is not constant and $cov(e_i, e_j) = -\sigma^2 h_{ij} \neq 0$.
- Hence, even in an ideal setting, where all the assumptions are satisfied, the residuals do not behave like the error.

Multiple Linear Regression (Residuals)

- Recall that the prediction variance was

$$var[\hat{y}(\mathbf{x}_0)] = \sigma^2(\mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0).$$

- If $\mathbf{x}_i$ is a data point, then

$$var[\hat{y}(\mathbf{x}_i)] = \sigma^2(\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i) = \sigma^2 h_{ii}.$$

- In a sense, h_{ii} express a standardize squared distance that $\mathbf{x}_i$ is from the point $\bar{\mathbf{X}}$, a vector of averages.

Multiple Linear Regression (Residuals)

- As an example, consider simple linear regression, where

$$var[\hat{y}(\mathbf{x}_i)] = \sigma^2 \left(\frac{1}{n} + \frac{(\mathbf{x}_i - \bar{\mathbf{X}})^2}{\sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{X}})^2} \right)$$

  hence $h_{ii} = \frac{1}{n} + \frac{(\mathbf{x}_i - \bar{\mathbf{X}})^2}{S_{xx}}$.

- From this we can recognize that for regression with an intercept in the model, $\frac{1}{n} \leq h_{ii} \leq 1$.

- Finally, the HAT matrix is sometimes called a projection matrix that projects $\mathbf{y}$ to $\hat{\mathbf{y}}$:

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

Multiple Linear Regression (Indicator Variables)

- In simple linear regression with a single independent variable it is necessary for the independent variable to be quantitative variable.
- With multiple independent variable, it is possible to have qualitative (or categorical) variables in addition to qualitative variable.
- Qualitative variables with two categories

$$Z = \begin{cases} 0 & \text{If the qualitative variable takes one category} \\ 1 & \text{If the qualitative variable takes the other category} \end{cases}$$

- The indicator variable is then added to the regression model and analyzed in the same way as other variables.
- Thus the regression model becomes:

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \beta_{p+1} z_i + \epsilon_i$$

Multiple Linear Regression (Indicator Variables)

- Note that if Z = 0 then the β_{p+1} drops from the model while if Z = 1 then it remains.
- Hence, the qualitative variable results in a shift in the intercept when the observation falls in the second category.
- For example, with one independent variable and one indicator, the model becomes:

$$Model = \begin{cases} y_i = \beta_0 + \beta_1 x_i + \epsilon_i & \text{when Z = 0} \\ y_i = (\beta_0 + \beta_2) + \beta_1 x_i + \epsilon_i & \text{when Z = 1} \end{cases}$$

Multiple Linear Regression (Indicator Variables)

- Qualitative variables with multiple categories:
- Suppose a qualitative variable can take any one of the q possible categories with $q \geq 3$.
- In such a case we must define $q - 1$ indicator variable to specify the q different categories.
- Each indicator variable takes only two possible values (0 and 1).
- Each of these indicator variables is then added to the regression model and the analysis is as before.

Comparing different covariates

- Working with nonstandardized multiple regression model means that the regression coefficients cannot be compared because of differences in the units involved.

- Suppose the following model:

$$\hat{y} = 20 + 200x_1 + 0.3x_2.$$

Correlation Transformation

- Using correlation transformation leads to a standardized regression model.
- The regression model with the transformed variables is

$$y_i^* = \sum_{j=1}^{p-1} \beta_j^* x_{i,j} + \epsilon_i^*,$$

where

$$y_i^* = \frac{y_i - \bar{y}}{\sqrt{(n-1)s_y^2}}$$

and

$$x_i^* = \frac{x_i - \bar{x}}{\sqrt{(n-1)s_{x_k}^2}}$$

Correlation Transformation

- The reason why there is no intercept parameter in the standardized regression model is that the least squares or maximum likelihood calculations always would lead to an estimated intercept term of zero if an intercept parameter were present in the model.

- It can be shown that, for $j = 1, \ldots, p-1$,

$$\beta_j = \left(\frac{s_y}{s_{x_k}} \right) \beta_j^*$$

and for β_0,

$$\beta_0 = \bar{y} - \sum_{j=1}^{p-1} \beta_j \bar{x}_j$$

Multicollinearity

- Multicollinearity is a linear association among the regressor variables and exists when at least two regressors have high empirical correlation.
- Mathematically, multicollinearity exists when there exists a vector $\mathbf{c} \neq 0$ such that $\sum_{i=1}^{p} c_i \mathbf{x}_i = 0$.
- To diagnose the existence of the multicollinearity, we often center and scale the covariate matrix.
- Then we calculate $\mathbf{X^*}^T \mathbf{X^*}$, where $\mathbf{X^*}$ denotes the center and scale the covariate matrix.
- This gives the correlation matrix for the model matrix.

Multicollinearity (Source)

- Method of data collection.
- Model and population constraints.
- Existence of identities or definitional relationships.

  Example: Soil texture in the Homework problem.
- Imprecise formulation of model.

  Example: Interaction and polinomial terms.
- An over-determined model.

  Example: number of predictors are larger than sample size.

Multicollinearity (Impact)

- Poor and unstable estimation of the coefficients.

  If collinearity exists, then a small changes in response can cause a large change in the coefficients.

  Collinearity can also cause the coefficients to be too large or have wrong sign.

- Poor prediction.
- The partial t-test have very low power.
- The fit of the regression is not damaged.

Multicollinearity (Impact)

- To illustrate the consequences of presence of multicollinearity, consider a model

$$y = \beta_1 x_1 + \beta_2 x_2 + \epsilon; \quad E(\epsilon) = 0, \quad Var(\epsilon) = \sigma^2 \mathbf{I}$$

where x_1, x_2 and y are centered and scaled.

- The normal equation for this model is:

$$\begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} r_{x_1,y} \\ r_{x_2,y} \end{pmatrix}$$

- Thus

$$\begin{aligned} \hat{\beta}_1 &= \frac{r_{x_1,y} - r r_{x_2,y}}{1 - r^2} \\ \hat{\beta}_2 &= \frac{r_{x_2,y} - r r_{x_2,y}}{1 - r^2} \end{aligned}$$

Multicollinearity (Symptoms)

- Large empirical correlation among regressors.

  If the absolute value of the off-diagonal element of this matrix is large, then there is collinearity between two regressors.

  This is only for pairwise correlations. However, multicollinearity could involve more than two regressors.

- Coefficients have signs that are opposite of what the scientist expected.

- Models have a large R^2 but only a few of the independent variables (or none of them) are significant by partial t-test.

- The estimated regression coefficients have large standard errors.

- Forward, backward, and stepwise procedures gives very different models.

Multicollinearity (Diagnostic)

- Determinant of $\mathbf{X}^{*T}\mathbf{X}^{*}$.
- Inspection of correlation matrix.
- Variance Inflation Factor (VIF):

  VIF represents the inflation of the variance of $\hat{\beta}_j$ from σ^2.
- $(VIF)_j = c_{jj}$ where c_{jj} is the j-th diagonal element of $\mathbf{X}^{*T}\mathbf{X}^{*}$, with $\mathbf{X}^{*}$ denotes the center and scale the covariate matrix.
- It can be shown that $(VIF)_j = \frac{1}{1-R_j^2}$, where R_j^2 denotes the coefficient of determination when we regress the j-th variable against all other covariates.

Multicollinearity (Diagnostic)

- Unlike the correlation matrix, the VIF looks at the relationship among multiple covariates.
- Any $VIF \geq 4$ would indicate that multicollinearity may be a problem.
- Any $VIF \geq 10$ indicate severe multicollinearity may be a problem.
- VIF do not tell how many dependencies exists. It just indicates the existence of multicollinearity problem.

Multicollinearity (Diagnostic)

- Multicollinearity can be also detected using the eigenvalues of the $\mathbf{X^*}^T\mathbf{X^*}$.
- If $\mathbf{X^*}^T\mathbf{X^*}$ is full rank, then k real eigenvalues does exists.
- In ideal case, all these eigenvalues are equal to 1.
- If we standardize and center the model matrix, then the sum of all the eigenvalues is equal to k (the number of regressors).
- This is due to the fact that trace of a squared matrix is the sum of the eigenvalues of that matrix and we have $trace(\mathbf{X^*}^T\mathbf{X^*}) = k$.
- Since the sum of the eigenvalues is equal to k, then the small eigenvalues must be balanced by large one.
- The existence of one or more near 0 eigenvalues indicates the existence of multicollinearity problem.

Multicollinearity (Diagnostic)

- In order to determine which variables are involved in the dependency.
- To do this, we calculate the condition number which is

$$\frac{\lambda_{MAX}}{\lambda_{MIN}}.$$

- We also can calculate the following to decide which eigenvalue is small

$$\phi_j = \frac{\lambda_{MAX}}{\lambda_j}.$$

- where ϕ_j is called condition indices.
- Any condition index greater than 1000 indicates the associated eigenvalue is small.

Multicollinearity (Diagnostic)

- The VIF's, eigenvalues, and condition indices tell us if there exist the multicollinearity problem.
- Variance proportion determines what are those dependencies.
- A variance decomposition is the proportion of the variance of $\hat{\beta}_i$ that can be attributed to the collinearity characterized by the j-th eigenvalue and is calculated by

$$p_{ij} = \frac{v_{ij}^2 / \lambda_j}{(VIF)_i},$$

- Note that $\sum_{j=1}^{p} p_{ij} = 1$ which implies that for each variable the variance proportions sum to 1.

Multicollinearity (Unstable coefficients)

- If $\mathbf{X}$'s are held constant, then a small changes in the y's can cause a large change in the coefficients.
- This can be shown using eigenvalue decomposition.
- By definition, for an square matrix $\mathbf{A}$, there exists an orthogonal matrix $\mathbf{V}$ ($\mathbf{V}^T = \mathbf{V}^{-1}$) such that

$$\mathbf{V}^T\mathbf{A}\mathbf{V} = \mathbf{\Lambda},$$

  where $\mathbf{\Lambda}$ is diagonal matrix whose diagona elements are the eigen values of $\mathbf{A}$.
- Let $\mathbf{A} = \mathbf{X^*}^T\mathbf{X^*}$, then we have $\mathbf{X^*}^T\mathbf{X^*} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$ and $(\mathbf{X^*}^T\mathbf{X^*})^{-1} = \mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{V}^T$.

Multicollinearity (Unstable coefficients)

- Thus the vector of estimates become:

$$\hat{\beta} = (\mathbf{X}^{*T}\mathbf{X}^{*})^{-1}\mathbf{X}^{*T}\mathbf{y} = \mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{V}^{T}\mathbf{X}^{*T}\mathbf{y} = \mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{c} = \sum_{i=1}^{k}\mathbf{v}_i\left(\frac{1}{\lambda_i}\right)c_i,$$

  where $\mathbf{c} = \mathbf{V}^{T}\mathbf{X}^{*T}\mathbf{y}$ and $c_i = \mathbf{v}_i^{T}\mathbf{X}^{*T}\mathbf{y}$.

- Note that the response variable is only involved in c_i (not $\mathbf{v}_i$ or λ_i).
- One of the diagnostic of the collinearity is that at least one of the eigenvalues is very small ($\lambda_j \approx 0$ for some j).
- Hence an small change in $\mathbf{y}$ creates a large changes in $\hat{\beta}$.
- In addition, note that all the $\hat{\beta}$'s are affected, but they are not equally damaged.
- The coefficients whose regressors are involved in collinearity are the ones which are radically unstable.

Multicollinearity (Variance of coefficients)

- Recall that
$$var(\hat{\beta}) = \sigma^2(\mathbf{X^*}^T\mathbf{X^*})^{-1}.$$
- Using eigenvalue decomposition, we have
$$\frac{var(\hat{\beta})}{\sigma^2} = (\mathbf{X^*}^T\mathbf{X^*})^{-1} = \mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{V}^T = \sum_{i=1}^{k} \frac{\mathbf{v}_i\mathbf{v}_i^T}{\lambda_i}.$$
- Small eigenvalues have a negative impact because the variance will be large.

Multicollinearity (Prediction)

- We know that there exists an orthogonal matrix $\mathbf{V}$ such that $\mathbf{V}^T(\mathbf{X}^{*T}\mathbf{X}^*)\mathbf{V} = \mathbf{\Lambda}$.
- Let $\lambda_j \approx 0$ for some j, then

$$\lambda_j = \mathbf{v}_j^T(\mathbf{X}^{*T}\mathbf{X}^*)\mathbf{v}_j \approx 0,$$

- Since $(\mathbf{X}^*\mathbf{v}_j)^T(\mathbf{X}^*\mathbf{v}_j) \approx 0$, then $\mathbf{X}^*\mathbf{v}_j \approx 0$ which implies that

$$\sum_{i=1}^{k} \mathbf{x}_i^* v_{ij} \approx 0.$$

- So, the "weights" in the definition of multicollinearity are elements of the eigenvector associated with λ_j.

Multicollinearity (Prediction)

- To address prediction, we look at the rows of $\mathbf{X}^*$ matrix.
- Since $\mathbf{X}^*\mathbf{v}_j = 0$, then for $i = 1, 2, \ldots, n$, $\mathbf{x}_i^*\mathbf{v}_j = 0$, and hence the data points are orthogonal to $\mathbf{v}_j$.
- Thus, prediction at (or near) a data point will be ok.
- This means, prediction along the "mainstream" of the data will be ok.
- At a data point $\mathbf{x^*}_i^T$,

$$\frac{var[\hat{y}(\mathbf{x^*}_i^T)]}{\sigma^2} = \mathbf{x^*}_i^T(\mathbf{X^{*T}X^*})^{-1}\mathbf{x^*}_i = \mathbf{x^*}_i^T\mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{V}^T\mathbf{x^*}_i = \mathbf{z}_i^T\mathbf{\Lambda}^{-1}\mathbf{z}_i$$

where $\mathbf{z}_i^T = \mathbf{x^*}_i^T\mathbf{V}$.

Multicollinearity (Prediction)

- Suppose that collinearity exists i.e. $\lambda_j \approx 0$, implying that $1/\lambda_j$ will be large.
- However, since the data point are orthogonal to $\mathbf{v}_j$, then $z_{ij} = \mathbf{v}_j^T \mathbf{x}^*{}_i$ will be near 0.
- Thus, $z_{ij}(1/\lambda_j)z_{ij}$ will be bounded.
- Therefore, at a data point, the prediction variance will not "blow up".
- The bound for the variance apart from σ^2 is $1 - \frac{1}{n}$.

Multicollinearity (Prediction)

- Consider an arbitrary point that is not in the mainstream of the data.
- Then we have

$$\mathbf{z}_0^T = \mathbf{x^*}_0^T \mathbf{V}, \quad \text{and} \quad \frac{var[\hat{y}(\mathbf{x^*}_0^T)]}{\sigma^2} = \mathbf{z}_0^T \mathbf{\Lambda}^{-1} \mathbf{z}_0$$

- Since $\mathbf{x^*}_0$ is not orthogonal to the columns of $\mathbf{V}$, so $\mathbf{z}_0$ will not be near 0.
- Hence, the prediction variance is not bounded.
- The size of the prediction variance depends on the degree of orthogonality of the data point to the eigenvectors.

Multicollinearity (Power of tests)

- Suppose, we partition the $\mathbf{X}^* = [\mathbf{x}^*_j | \mathbf{X}^*_2]$ and consider a test of $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$.
- The test statistics is

$$t = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{c_{jj}}}$$

where

$$c_{jj} = [\mathbf{x}^{*T}_j \mathbf{x}^*_j - \mathbf{x}^{*T}_j \mathbf{X}^*_2 (\mathbf{X}^{*T}_2 \mathbf{X}^*_2)^{-1} \mathbf{X}^*_2 \mathbf{x}^*_j]^{-1}$$

is the j-th variance inflation factor (VIF).

- To find the power of the test, we have to compute the noncentrality parameter ψ.

Multicollinearity (Power of tests)

- To do this, we replace the estimates with the parameters they are estimating:

$$\psi = \frac{\beta_j}{\sigma\sqrt{c_{jj}}} = \frac{\beta_j}{\sigma\sqrt{VIF}}.$$

- To have a good power, we need the noncentrality parameter to be large.
- In this case β_j and σ are parameters and ψ depends on VIF.
- In ideal case VIF is equal to 1.
- However, when collinearity exists, VIF's are inflated and hence ψ is small.
- Thus, the power of the test will be low.

Multicollinearity (Fit)

- A fitted value is just a predicted value at a data point.
- We already showed that this is not adversely affected by collinearity.
- Under normality assumption, we have

$$\hat{\sigma}^2 = MSE \sim \frac{\sigma^2 \chi^2_{n-p}}{n-p}.$$

- Therefore, $SSE \sim \sigma^2 \chi^2_{n-p}$ which does not involve $\mathbf{X}$. Therefore, it is not affected by the collinearity.
- Thus, the sum of square surface is “flat” and hence the fitted values are fairly stable.

Multicollinearity

- When collinearity is diagnosed, we have to do something about it.
- The most obvious method is to drop one or more regressors.
- Here, we hope that the quality of the fit and prediction capability of model is not compromised.
- An alternative is to use a biased form of estimation.
- The common approaches are:

  Ridge regression,

  Principle component regression.

Ridge Regression

- Ridge regression is a biased estimation technique that attempts to reduce the effects of the collinearity.
- Ridge regression tries to reduce the variance and increase the stability of the regression coefficients.
- The negative aspect of Ridge regression is the fact that the estimators are not unbiased anymore.
- The problem with collinearity is that the off diagonal elements of the $\mathbf{X}^{*T}\mathbf{X}^{*}$ are large.
- This lead to the diagonal elements of $(\mathbf{X}^{*T}\mathbf{X}^{*})^{-1}$ (VIF) to be large.
- This problem can be reduced by adding a small positive constant k to the diagonal elements of $\mathbf{X}^{*T}\mathbf{X}^{*}$.

Ridge Regression

- Ridge regression is a generalized form of the least squares estimation.
- In other word, we want to minimize following objective function

$$(\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta})$$

subject to $\hat{\beta}^T\hat{\beta} \leq c$ where c is a nonegative constant.

- Using Lagrangian multipliers:

$$(\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta}) + k(\hat{\beta}^T\hat{\beta} - c) =$$
$$\mathbf{y}^T\mathbf{y} - 2\hat{\beta}^T\mathbf{X}^T\mathbf{y} + \hat{\beta}^T\mathbf{X}^T\mathbf{X}\hat{\beta} + k(\hat{\beta}^T\hat{\beta} - c).$$

- Then, one can take the derivative of the objective function with respect to $\hat{\beta}$ to get the Ridge estimate.

Ridge Regression

- The derivative of the objective function with respect to $\hat{\beta}$ is

$$-2\mathbf{X}^T\mathbf{y} + 2(\mathbf{X}^T\mathbf{X})\hat{\beta} + 2k\hat{\beta}$$

- Recall that the normal equation for OLS are:

$$(\mathbf{X}^T\mathbf{X})\hat{\beta} = \mathbf{X}^T\mathbf{y}.$$

- Thus, the normal equations for Ridge regression are modified to

$$(\mathbf{X}^T\mathbf{X} + k\mathbf{I})\hat{\beta}_R = \mathbf{X}^T\mathbf{y}.$$

- Therefore, the Ridge estimates for the coefficients are:

$$\hat{\beta}_R = (\mathbf{X}^T\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}.$$

- Ridge regression is also Bayesian estimator for regression coefficients with a normal prior.

Ridge Regression

- The Ridge estimators are smaller in magnitude ($||\hat{\beta}_R|| \leq ||\hat{\beta}_{OLS}||$), more stable, and thus produce better prediction.
- The variance of the coefficients are smaller:

$$\sum_{i=1}^{p} \frac{var(\hat{\beta}_{i,R})}{\sigma^2} \leq \sum_{i=1}^{p} \frac{var(\hat{\beta}_{i,OLS})}{\sigma^2}.$$

- $\sum_{i=1}^{p} \frac{var(\hat{\beta}_{i,R})}{\sigma^2}$ is monotonically decreasing function of k.
- $\hat{\beta}_R$ is biased:

$$\begin{aligned} E[\hat{\beta}_R] &= E[(\mathbf{X}^T\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}] \\ &= (\mathbf{X}^T\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^T E[\mathbf{y}] \\ &= (\mathbf{X}^T\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}\hat{\beta} \\ &= (\mathbf{X}^T\mathbf{X} + k\mathbf{I})^{-1}(\mathbf{X}^T\mathbf{X} + k\mathbf{I})\hat{\beta} - k(\mathbf{X}^T\mathbf{X} + k\mathbf{I})^{-1}\hat{\beta} \\ &= \hat{\beta} - k(\mathbf{X}^T\mathbf{X} + k\mathbf{I})^{-1}\hat{\beta}, \end{aligned}$$

Ridge Regression

- The sum of the squared biased is

$$\sum_{i=1}^{p}[BIAS(\hat{\beta}_{i,R})]^2 = k^2\hat{\beta}^T(\mathbf{X}^T\mathbf{X}+k\mathbf{I})^{-1}\hat{\beta}.$$

- In addition, the sum of the variance is

$$\sum_{i=1}^{n}\frac{var(\hat{\beta}_{i,R})}{\sigma^2} = \sum_{i=1}^{n}\frac{\lambda_i}{(\lambda_i+k)^2}.$$

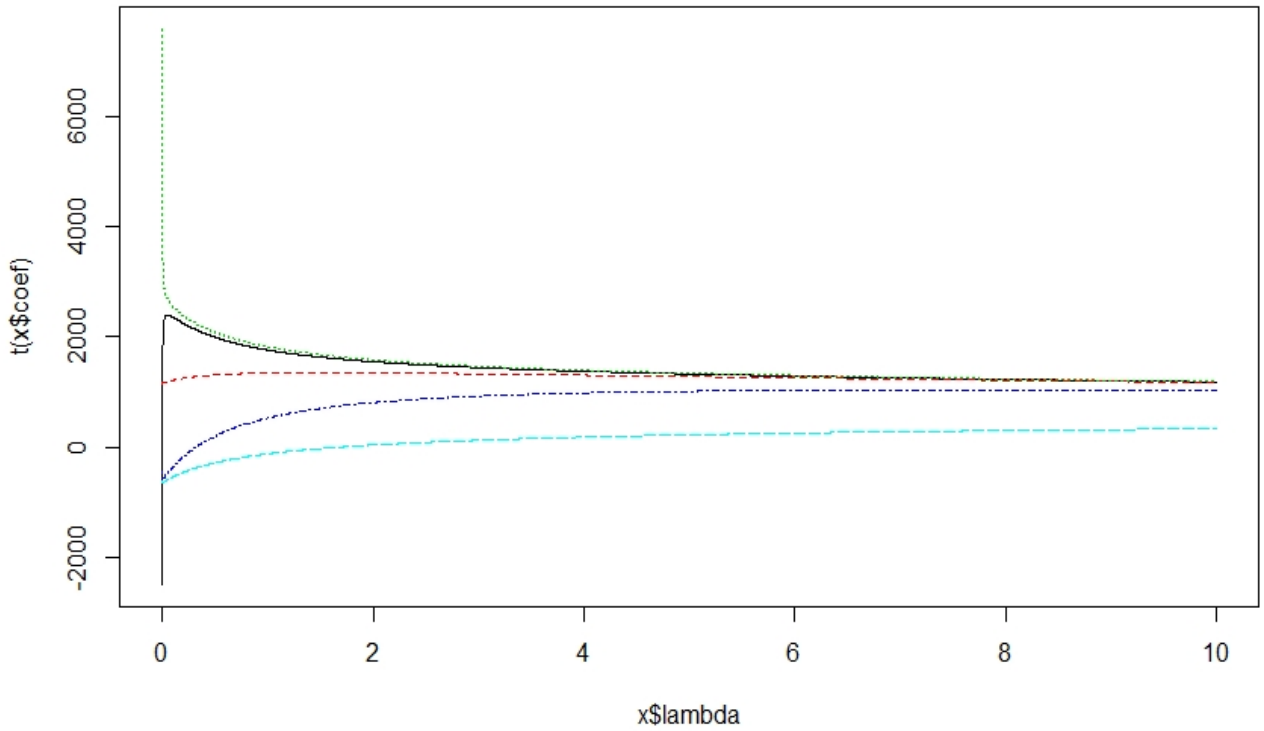
- k is called shrinkage parameter or shrinkage constant.
- It shrinks the length of the vector of $\hat{\beta}$ and shrinks the variance of the regressors.
- When choosing k, one should keep in mind that larger values increase the bias. So, we want k to be as small as possible.

Ridge Regression (Choosing Shrinkage Parameter)

- There are many different methods to choose k.
  - Ridge trace.
  - C_p-like statistic.
  - PRESS-like statistic.
  - VIF.
  - DF-trace.

Ridge Regression (Ridge trace)

- If the goal is the stability and interpretability of the regression coefficients, then the Ridge trace is a good procedure.
- The Ridge coefficients will be calculated for several different values of k.



Ridge Regression (C_p-like statistic)

- The C_p statistic is

$$\sum_{i=i}^{n} \frac{var[\hat{y}(\mathbf{x}_i)] + \{BIAS[\hat{y}(\mathbf{x}_i)]\}^2}{\sigma^2}$$

- Thus $C_p = p + \frac{(n-p)(\hat{\sigma}^2 - \sigma^2)}{\sigma^2}$, where $p = trace(\mathbf{H})$.
- Let

$$\mathbf{H}_k = \mathbf{X}^*(\mathbf{X}^{*^T}\mathbf{X}^* + k\mathbf{I})^{-1}\mathbf{X}^{*^T}$$

be the Ridge HAT matrix for center and scaled data.

- Let

$$\mathbf{A}_k = \mathbf{X}\begin{bmatrix} n & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & (\mathbf{X}^{*^T}\mathbf{X}^* + k\mathbf{I}) & \\ 0 & & & \end{bmatrix}^{-1}\mathbf{X}$$

with $\mathbf{X} = [\mathbf{1}|\mathbf{X}^*]$, be Ridge HAT matrix that involves the intercept.

Ridge Regression (C_p-like statistic)

- Unlike the HAT matrix, Ridge HAT matrix is not idempotent unless $k = 0$.
- This statistic is

$$\begin{aligned} C_k &= \sum_{i=i}^{n} \frac{var[\hat{y}(\mathbf{x}_i)] + \{BIAS[\hat{y}(\mathbf{x}_i)]\}^2}{\sigma^2} \\ &= [trace(\mathbf{A}_k)^2] + \left\{ \frac{SSE_k - \sigma^2 trace(\mathbf{I} - \mathbf{A}_k)^2}{\sigma^2} \right\} \\ &= \frac{SSE_k}{\sigma^2} - n + 2 + 2trace(\mathbf{H}_k) \end{aligned}$$

- This statistic addresses the same V-B trade off as the C_p-statistic.
- The goal is to choose the k value such that minimizes the C_k.
- It is worth mentioning that increases in k leads to rapid decreases in $trace(\mathbf{H}_k)$ and slow increases in SSE_k.
- The C_k is a criteria to use if the prediction capability of the model is more important.

Ridge Regression (PRESS-like statistic)

- The PRESS statistic is

$$PRESS = \sum_{i=1}^{n} e_{i,-i}^2.$$

- Another alternative for PRESS is

$$PRESS = \sum_{i=1}^{n} \left(\frac{e_i}{1 - h_{ii}} \right)^2.$$

- For the Ridge regression the exact formula for PRESS is

$$PRESS_k = \sum_{i=1}^{n} e_{i,-i,k}^2.$$

- However, there is not an exact short-cut to computing $PRESS_k$.

Ridge Regression (PRESS-like statistic)

- When we center and scale the data, the setting aside of the data point changes the centering and scaling constant.
- We can approximate the PRESS statistic with

$$PR(Ridge) = \sum_{i=1}^{n} \left(\frac{e_{i,k}}{1 - \frac{1}{n} - h_{ii,k}} \right)^2 .$$

- This will be a good approximate if the sample size is not small and if there are no high leverage points.
- We want to choose a k value such that minimizes the PR(Ridge).
- PRESS-like statistic is a criteria to use if the prediction capability of the model is more important.

Ridge Regression (VIF)

- The first three statistics uses the response variable in their calculation.
- Since collinearity only involves the independent variables, we desire to have a procedure that does not involves response.
- Such non-stochastic procedure allows k to increase until all the VIF's are "reduced significantly".
- It should be noted that VIF have a decreasing relation with k.
- One may choose k such that all VIF are either less than 4 or 10 depending on how much one want to be conservative.

Ridge Regression (DF-trace)

- Another non-stochastic procedure is DF-trace which is $trace(\mathbf{H}_k)$.
- Choose a k value that stabilizes DF-trace.
- Recall that in OLS $trace(\mathbf{H})$ is equal to the number of regressors.
- Thus,

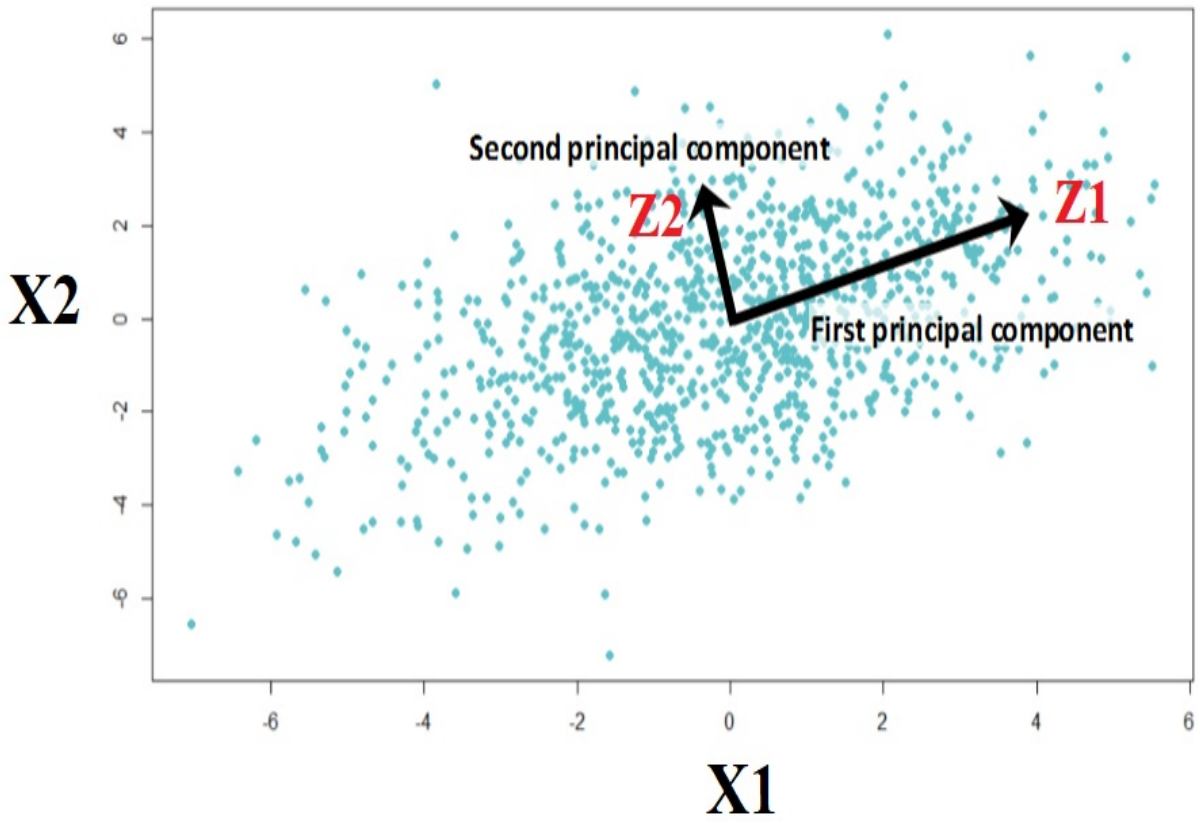
$$trace(\mathbf{H}_k) = \sum_{i=1}^{p} \frac{\lambda_i}{\lambda_i + k},$$

  which is called the effective regression degrees of freedom.
- This shows the number of regressors left in the model after Ridge regression has removed the collinearity.
- **What k do we choose**

  Completely depends on the goal.

Principla Component Regression

- Another biased estimation procedure that can be used to reduce the effect of collinearity is principal component regression.
- PCR involves a rotation of the axes to a set of axes that better cover the data.



Principla Component Regression

- Note that the PCs are orthogonal to each other. Thus, PCR hit the collinearity head on.
- The principal components are supplied with normalized version of original predictors.
- This is because, the original predictors may have different scales.
- **Example:**

  Imagine a data set with variables measuring units as gallons, kilometers, light years etc.

Principla Component Regression

- Recall that the centered and scaled model is

$$\mathbf{y} = \mathbf{1}\beta_0 + \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

- From eigenvalue decomposition method, we know that there exists an orthogonal matrix $\mathbf{V}$ such that

$$\mathbf{V}^T(\mathbf{X}^{*^T}\mathbf{X}^*)\mathbf{V} = \boldsymbol{\Lambda}.$$

  where $\boldsymbol{\Lambda}$ is a diagonal matrix of the eigenvalues of $\mathbf{X}^{*^T}\mathbf{X}^*$.

- Since $\mathbf{V}\mathbf{V}^T = \mathbf{I}$, then we can rewrite the model as

$$\mathbf{y} = \mathbf{1}\beta_0 + \mathbf{X}^*\mathbf{V}\mathbf{V}^T\boldsymbol{\beta} + \boldsymbol{\epsilon} = \mathbf{1}\beta_0 + \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\epsilon}$$

  where $\mathbf{Z} = \mathbf{X}^*\mathbf{V}$ and $\boldsymbol{\alpha} = \mathbf{V}^T\boldsymbol{\beta}$.

Principla Component Regression

- For this new model, we have

$$\mathbf{Z}^T\mathbf{Z} = (\mathbf{X}^*\mathbf{V})^T(\mathbf{X}^*\mathbf{V}) = \mathbf{V}^T{\mathbf{X}^*}^T\mathbf{X}^*\mathbf{V} = \mathbf{\Lambda}$$

- Thus, the columns of $\mathbf{Z}$ are orthogonal to each other.
- In term of $\mathbf{Z}$, there is **no collinearity**.
- It is worth mentioning that $\mathbf{z}_j = \mathbf{X}^*\mathbf{v}_j$ is the j-th principle component.
- Once the orthogonal transformation is made that produces the PCs, we can use the OLS to regress $\mathbf{y}$ versus $\mathbf{z}$'s.
- This procedure provide estimates for α that are not affected by collinearity problem.

Principla Component Regression

- The new parameter α may be meaningless to us, so we should rotate back to the original metric:

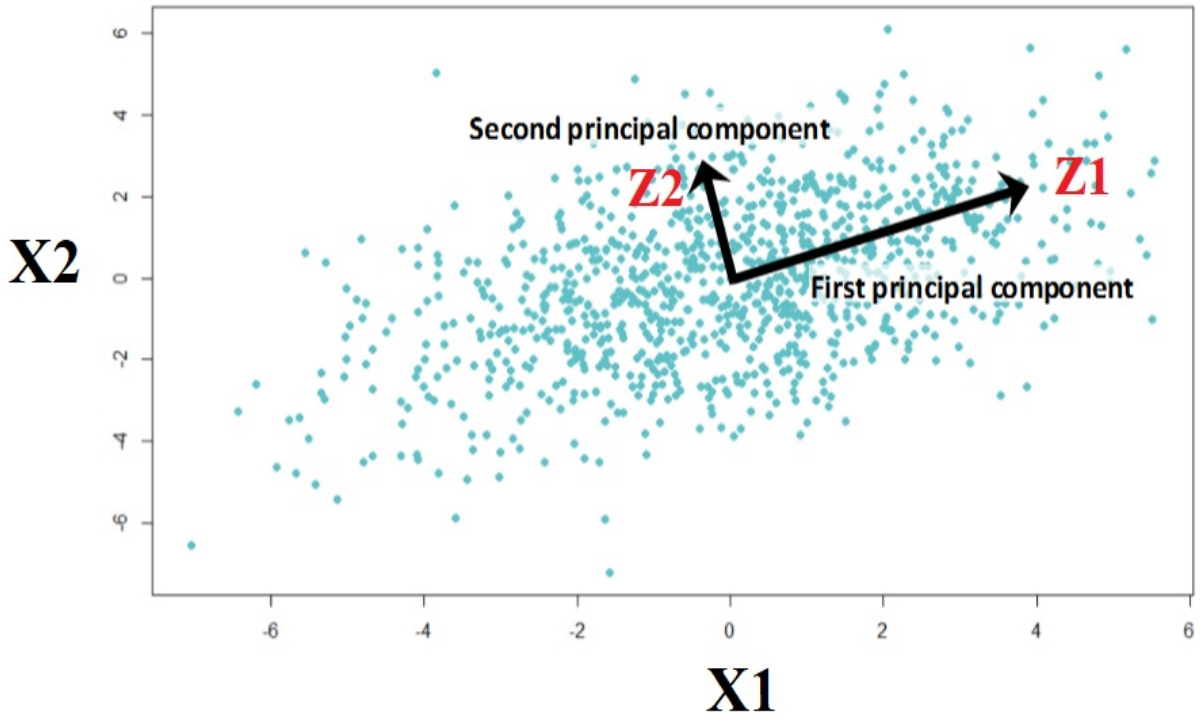
$$\beta = \mathbf{V}\alpha.$$

- In addition, while we have eliminated the collinearity, we created large variance; Thus the overall problem still exists.

- The idea behind the PCR is that since PCs are orthogonal to each other, we eliminate PCs to achieve a substantial reduction in variance.

Principla Component Regression

- **Example:**



- The second PC contains small information compared to the first PC.

- However, there is a large variability around the second PC.

Principla Component Regression

- Since

$$\mathbf{Z}^T\mathbf{Z} = \mathbf{\Lambda}$$

then $\lambda_j = \mathbf{z}_j^T\mathbf{z}_j$. This means sum of squares of the PCs are eigenvalues.

- Thus, if we have a near zero eigenvalue, then $\mathbf{z}_j^T\mathbf{z}_j \approx 0$, which can happen only if $\mathbf{z}_j \approx 0$.

- The PCs that we want to eliminate are those associated with near zero eigenvalues.

Principla Component Regression

- Considering the variance, we have

$$\frac{var(\hat{\alpha}_j)}{\sigma^2} = \frac{var(\mathbf{v}_j^T\hat{\beta})}{\sigma^2} = \frac{\mathbf{v}_j^T var(\hat{\beta})\mathbf{v}_j}{\sigma^2} = \mathbf{v}_j^T(\mathbf{X}^{*T}\mathbf{X}^*)^{-1}\mathbf{v}_j = \frac{1}{\lambda_j}$$

- Thus

$$\sum_{j=1}^{k}\frac{var(\hat{\alpha}_j)}{\sigma^2} = \sum_{j=1}^{k}\frac{1}{\lambda_j}$$

- This implies that eliminating the PC associated with the near zero eigenvalue will reduce the variance of the estimations.

Principla Component Regression

- Suppose that there are r near zero eigenvalues and hence we drop r PCs and keep the $s = k - r$.
- This leads to partitioning **Z, V**, α, and $\mathbf{\Lambda}$ as follow:

$$\mathbf{Z} = \left[\begin{array}{c|c} \mathbf{Z}_r & \mathbf{Z}_s \end{array} \right], \quad \mathbf{V} = \left[\begin{array}{c|c} \mathbf{V}_r & \mathbf{V}_s \end{array} \right], \quad \alpha = \begin{bmatrix} \alpha_r \\ \cdots \\ \alpha_s \end{bmatrix}, \quad \mathbf{\Lambda} = \left[\begin{array}{c|c} \mathbf{\Lambda}_r & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{\Lambda}_s \end{array} \right]$$

where $\mathbf{\Lambda}_r$ and $\mathbf{\Lambda}_s$ are diagonal matrices, where $\mathbf{\Lambda}_r$ contains the near zero eigenvalues.

Principla Component Regression

- If we drop $\mathbf{Z}_r$, it is equivalent to let $\alpha_r = 0$.
- The model become:

$$\begin{aligned} \mathbf{y} &= \mathbf{1}\beta_0 + \mathbf{X}^*\beta + \epsilon \\ &= \mathbf{1}\beta_0 + \mathbf{Z}\alpha + \epsilon \\ &= \mathbf{1}\beta_0 + \mathbf{Z}_r\alpha_r + \mathbf{Z}_s\alpha_s + \epsilon \\ &\approx \mathbf{1}\beta_0 + \mathbf{Z}_s\alpha_s + \epsilon \end{aligned}$$

- Using OLS:

$$\hat{\alpha} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{y} = \mathbf{\Lambda}^{-1}\mathbf{V}^T{\mathbf{X}^*}^T\mathbf{y}$$

- Thus

$$\begin{bmatrix} \hat{\alpha}_r \\ \cdots \\ \hat{\alpha}_s \end{bmatrix}, \quad \mathbf{\Lambda} = \left[\begin{array}{c|c} \mathbf{\Lambda}_r^{-1} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{\Lambda}_s^{-1} \end{array}\right] \begin{bmatrix} \mathbf{V}_r^T \\ \cdots \\ \mathbf{V}_s^T \end{bmatrix} {\mathbf{X}^*}^T\mathbf{y}$$

Principla Component Regression

- Therefore, $\hat{\alpha}_s = \mathbf{\Lambda}_s^{-1}\mathbf{V}_s^T\mathbf{X}^{*^T}\mathbf{y}$.
- This is the set of coefficients retained when we do principal components.
- Now we need to transform back to the original metric using $\hat{\boldsymbol{\beta}} = \mathbf{V}\hat{\alpha}$.
- Thus the PC estimates are

$$\hat{\boldsymbol{\beta}}_{PC} = \mathbf{V}_s\hat{\alpha}_s = \mathbf{V}_s\mathbf{\Lambda}_s^{-1}\mathbf{V}_s^T\mathbf{X}^{*^T}\mathbf{y}.$$

- $\hat{\boldsymbol{\beta}}_{PC}$ is a $k \times 1$ vector which means we eliminated the PCs not the regressors.

Principla Component Regression (Properties)

- **Bias of $\hat{\beta}_{PC}$:**
- Since $\alpha = \mathbf{V}^T\beta$, and $\alpha = \begin{bmatrix} \hat{\alpha}_r \\ \cdots \\ \hat{\alpha}_s \end{bmatrix} = \begin{bmatrix} \mathbf{V}_r^T \\ \cdots \\ \mathbf{V}_s^T \end{bmatrix} \beta$, then $\hat{\alpha}_s = \mathbf{V}_s^T\beta$.
- In addition, we have $\mathbf{V}\mathbf{V}^T = \mathbf{I}$. Thus

$$\left[\begin{array}{c|c} \mathbf{V}_r & \mathbf{V}_s \end{array} \right] \begin{bmatrix} \mathbf{V}_r^T \\ \cdots \\ \mathbf{V}_s^T \end{bmatrix} = \mathbf{V}_r\mathbf{V}_r^T + \mathbf{V}_s\mathbf{V}_s^T = \mathbf{I}.$$

- Thus:

$$\begin{aligned} E[\hat{\beta}_{PC}] &= \mathbf{V}_s E[\hat{\alpha}_s] = \mathbf{V}_s\alpha_s = \mathbf{V}_s\mathbf{V}_s^T\beta \\ &= [\mathbf{I} - \mathbf{V}_r\mathbf{V}_r^T]\beta = \beta - \mathbf{V}_r\mathbf{V}_r^T\beta = \beta - \mathbf{V}_r\alpha_r. \end{aligned}$$

- This shows that the bias of the coefficients is $\mathbf{V}_r\alpha_r$ and depends on the orientation of β with respect to $\mathbf{V}_r$.

Principla Component Regression (Properties)

- **Variance of $\hat{\beta}_{PC}$:**
- Since

$$\hat{\alpha} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{y}, \quad \text{and} \quad \frac{var(\hat{\alpha})}{\sigma^2} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{\Lambda}^{-1}.$$

  Then $\frac{var(\hat{\alpha}_s)}{\sigma^2} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{\Lambda}_s^{-1}$.

- Note that these are "good" eigenvalues which means the elements of $\mathbf{\Lambda}_s^{-1}$ are not extremely large.

Principla Component Regression (Properties)

- Since $\mathbf{Z} = \mathbf{X}^*\mathbf{V}$, thus

$$\begin{aligned}\frac{var(\hat{\beta}_{OLS})}{\sigma^2} &= \left[\begin{array}{c|c} \mathbf{V}_r & \mathbf{V}_s \end{array}\right] \left[\begin{array}{c|c} \mathbf{\Lambda}_r^{-1} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{\Lambda}_s^{-1} \end{array}\right] \begin{bmatrix} \mathbf{V}_r^T \\ \cdots \\ \mathbf{V}_s^T \end{bmatrix} \\ &= \mathbf{V}_r\mathbf{\Lambda}_r^{-1}\mathbf{V}_r^T + \mathbf{V}_s\mathbf{\Lambda}_s^{-1}\mathbf{V}_s^T\end{aligned}$$

- For PCR, we drop $\mathbf{Z}_r$. Thus

$$\frac{var(\hat{\beta}_{PC})}{\sigma^2} = \frac{var(\mathbf{V}_s\hat{\alpha})}{\sigma^2} = \frac{\mathbf{V}_s var(\hat{\alpha})\mathbf{V}_s^T}{\sigma^2} = \mathbf{V}_s\mathbf{\Lambda}_s^{-1}\mathbf{V}_s^T$$

- This means the difference is $\mathbf{V}_r\mathbf{\Lambda}_r^{-1}\mathbf{V}_r^T$.
- How many PC should we drop?

Principla Component Regression (Summary)

- **Dimensionality reduction**

  Using PCR, one can reduce the dimension of a high dimensional data set and fit a linear regression model to a smaller set of variables, while preserve most of the variability and information of the original data.

- **Avoiding multicollinearity**

  PCR is able to avoid the multicollinearity problem since performing PCA on the raw data produces linear combinations of the predictors that are uncorrelated.

Principla Component Regression (Summary)

- **Overfitting mitigation**

  Under PCR assumptions, fitting a least squares model to the principal components will lead to better results than fitting a least squares model to the original data.

  This is due to the fact that PCs contain most of the information related to the dependent variable. Thus, by estimating less coefficients one can reduce the risk of overfitting.

- **Drawback**

  PCR is not good for variable selection since each of the PCs is a linear combination of the original variables.

Impact of Model Misspecification

- Setting: We have several regressor variables, and thus several candidate models.
- The selection process to choose a ”best” model is a compromise between underspecifying (leaving out important variables) and overspecifying (including variables which do not belong).
- Underfitting creates bias,
- Overfitting inflate the variance.

Impact of Model Misspecification

- Underfitting creates bias:
- Suppose we fit the model $\mathbf{y} = \mathbf{X}_1\beta_1 + \epsilon^*$.
- Then the OLS estimate for β_1 is $(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{y}$.
- However, suppose the correct model is $\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \epsilon$, where β_1 contains p_1 and β_2 contains p_2 parameters.
- Therefore the fitted model is an underspecification of the true model.

Impact of Model Misspecification (Bias in the coefficients)

- This creates a bias in the coefficients of the fitted model.

$$\begin{aligned} E(\hat{\beta}_1) &= E[(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{y}] \\ &= (\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T E(\mathbf{y}) \\ &= (\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T(\mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2) \\ &= (\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{X}_1\beta_1 + (\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{X}_2\beta_2 \\ &= \beta_1 + (\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{X}_2\beta_2 \\ &= \beta_1 + \mathbf{A}\beta_2 \end{aligned}$$

  where $\mathbf{A} = (\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{X}_2$.

- The matrix **A** is called Alias Matrix.

Impact of Model Misspecification (Bias in the coefficients)

- The bias in the coefficients is $\mathbf{A}\beta_2$.
- This bias can be zero if either of the following in true.
  1. If $\beta_2 = 0$. This implies that the model that we fit is the correct model.
  2. If $\mathbf{X}_1^T\mathbf{X}_2 = 0$. This usually results from planned or designed experiments in which the column of the $\mathbf{X}_1$ and $\mathbf{X}_2$ are orthogonal.
- Underfitting creates a bias in prediction.

Impact of Model Misspecification (Bias in prediction)

- Consider a predicted value $\hat{y}(\mathbf{x}_0)$ at location $\mathbf{x}_0^T = [\mathbf{x}_{1,0}^T, \mathbf{x}_{2,0}^T]$.
- With fitted model $\mathbf{y} = \mathbf{X}_1\beta_1 + \epsilon^*$.
- A predicted value is $\hat{\mathbf{y}} = \mathbf{x}_{1,0}^T\hat{\beta}_1$ while the true mean response is $E[\mathbf{y}] = \mathbf{x}_{1,0}^T\beta_1 + \mathbf{x}_{2,0}^T\beta_2$.
- Then, we have:

$$\begin{aligned} E[\hat{\mathbf{y}}(\mathbf{x}_0)] &= E[\mathbf{x}_{1,0}^T\beta_1] \\ &= \mathbf{x}_{1,0}^T E[\beta_1] \\ &= \mathbf{x}_{1,0}^T[\beta_1 + \mathbf{A}\beta_2] \\ &= \mathbf{x}_{1,0}^T\beta_1 + \mathbf{x}_{1,0}^T\mathbf{A}\beta_2 \\ &= \mathbf{x}_{1,0}^T\beta_1 + \mathbf{x}_{1,0}^T(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{X}_2. \end{aligned}$$

Impact of Model Misspecification (Bias in prediction)

- Then the bias in prediction is:

$$\begin{aligned} E[\hat{\mathbf{y}}(\mathbf{x}_0) - \mathbf{y}(\mathbf{x}_0)] &= E[\hat{\mathbf{y}}(\mathbf{x}_0)] - E[\mathbf{y}(\mathbf{x}_0)] \\ &= [\mathbf{x}_{1,0}^T \boldsymbol{\beta}_1 + \mathbf{x}_{1,0}^T (\mathbf{X}_1^T \mathbf{X}_1)^{-1} \mathbf{X}_1^T \mathbf{X}_2] - [\mathbf{x}_{1,0}^T \boldsymbol{\beta}_1 + \mathbf{x}_{2,0}^T \boldsymbol{\beta}_2] \\ &= \mathbf{x}_{1,0}^T (\mathbf{X}_1^T \mathbf{X}_1)^{-1} \mathbf{X}_1^T \mathbf{X}_2 - \mathbf{x}_{2,0}^T \boldsymbol{\beta}_2 \\ &= \mathbf{x}_{1,0}^T \mathbf{A} \boldsymbol{\beta}_2 - \mathbf{x}_{2,0}^T \boldsymbol{\beta}_2 \\ &= (\mathbf{x}_{1,0}^T \mathbf{A} - \mathbf{x}_{2,0}^T) \boldsymbol{\beta}_2. \end{aligned}$$

- Note that the bias does not include the $\boldsymbol{\beta}_1$ parameters.

- This bias is a function of $\boldsymbol{\beta}_2$ and all the regressors variables.

Impact of Model Misspecification (Bias in Error Mean Square)

- Suppose we fit the model $\mathbf{y} = \mathbf{X}_1\beta_1 + \epsilon^*$.
- Therefore,

$$\hat{\sigma}^2 = \frac{\mathbf{y}^T[\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T]\mathbf{y}}{n - p_1} = \frac{\mathbf{y}^T[\mathbf{I} - \mathbf{H}_1]\mathbf{y}}{n - p_1}.$$

- This is a quadratic form of $\mathbf{y}^T\mathbf{B}\mathbf{y}$.
- By definition, $E(\mathbf{y}^T\mathbf{B}\mathbf{y}) = var(\mathbf{y})trace(\mathbf{B}) + [E(\mathbf{y})]^T\mathbf{B}[E(\mathbf{y})]$. (proof?)
- To determine the bias of the $\hat{\sigma}^2$, we need to find the $E(\hat{\sigma}^2)$.

Impact of Model Misspecification (Bias in Error Mean Square)

- The expected value of the error in mean square is

$$\begin{aligned} E(\hat{\sigma}^2) &= \frac{1}{n-p_1} E[\mathbf{y}^T[\mathbf{I}-\mathbf{H}_1]\mathbf{y}] \\ &= \frac{1}{n-p_1}[var(\mathbf{y})trace(\mathbf{I}-\mathbf{H}_1)+[E(\mathbf{y})]^T(\mathbf{I}-\mathbf{H}_1)[E(\mathbf{y})], \end{aligned}$$

- We have:

$$\begin{aligned} trace(\mathbf{I}-\mathbf{H}_1) &= trace(\mathbf{I}) - trace[\mathbf{H}_1] \\ &= n - trace[\mathbf{X}_1(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T] \\ &= n - trace[\mathbf{X}_1^T\mathbf{X}_1(\mathbf{X}_1^T\mathbf{X}_1)^{-1}] \\ &= n - trace[\mathbf{I}_{p_1\times p_1}] \\ &= n - p_1, \end{aligned}$$

Impact of Model Misspecification (Bias in Error Mean Square)

- Furthermore,

$$\begin{aligned}(\mathbf{X}_1\beta_1)^T[\mathbf{I}-\mathbf{H}_1] &= \beta_1^T\mathbf{X}_1^T[\mathbf{I})-\mathbf{X}_1(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T] \\ &= \beta_1^T[\mathbf{X}_1^T-\mathbf{X}_1^T\mathbf{X}_1(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T] \\ &= \beta_1^T[\mathbf{X}_1^T-\mathbf{X}_1^T]=\beta_1^T\mathbf{0}=\mathbf{0}.\end{aligned}$$

- Thus, we have

$$\begin{aligned}E(\hat{\sigma}^2) &= \frac{1}{n-p_1}[\sigma^2\mathbf{I}(n-p_1)+(\mathbf{X}_1\beta_1+\mathbf{X}_2\beta_2)^T[\mathbf{I}-\mathbf{H}_1](\mathbf{X}_1\beta_1+\mathbf{X}_2\beta_2)] \\ &= \sigma^2+\frac{1}{n-p_1}[(\mathbf{X}_1\beta_1+\mathbf{X}_2\beta_2)^T[\mathbf{I}-\mathbf{H}_1](\mathbf{X}_1\beta_1+\mathbf{X}_2\beta_2)] \\ &= \sigma^2+\frac{1}{n-p_1}[\beta_1^T\mathbf{X}_1^T[\mathbf{I}-\mathbf{H}_1]\mathbf{X}_1\beta_1+\beta_1^T\mathbf{X}_1^T[\mathbf{I}-\mathbf{H}_1]\mathbf{X}_2\beta_2 \\ &\quad +\beta_2^T\mathbf{X}_2^T[\mathbf{I}-\mathbf{H}_1]\mathbf{X}_1\beta_1+\beta_2^T\mathbf{X}_2^T[\mathbf{I}-\mathbf{H}_1]\mathbf{X}_2\beta_2] \\ &= \sigma^2+\frac{1}{n-p_1}\beta_2^T\mathbf{X}_2^T[\mathbf{I}-\mathbf{H}_1]\mathbf{X}_2\beta_2\end{aligned}$$

Impact of Model Misspecification (Bias in Error Mean Square)

- Thus, the bias is $\frac{1}{n-p_1}\beta_2^T\mathbf{X}_2^T[\mathbf{I}-\mathbf{H}_1]\mathbf{X}_2\beta_2$.
- This bias can be written as

$$\begin{aligned} & \frac{1}{n-p_1}\beta_2^T[\mathbf{X}_2^T\mathbf{X}_2-\mathbf{X}_2^T\mathbf{X}_1(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{X}_2]\beta_2 \\ = & \frac{1}{n-p_1}\beta_2^T\mathbf{A}_{22}^{-1}\beta_2 \\ = & \frac{1}{n-p_1}\beta_2^T[var(\hat{\beta}_2)]^{-1}\beta_2. \end{aligned}$$

Impact of Model Misspecification (Overfitting)

- Suppose, we fit the model $\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \epsilon$.
- However, $\mathbf{X}_2\beta_2$ is only of marginal importance.
- We would expect this to create an increase in variances.
- In general, suppose we have two competing models:
  1. A short model that contains $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$, and
  2. A long model that contains $\hat{\beta}_1^*, \hat{\beta}_2^*, \ldots, \hat{\beta}_k^*, \hat{\beta}_{k+1}^*$.
- If we partition $\mathbf{X} = [\mathbf{x}_j|\mathbf{X}_2]$,

$$\frac{var(\hat{\beta}_j)}{\sigma^2} = [\mathbf{x}_j^T\mathbf{x}_j - \mathbf{x}_j^T\mathbf{X}_2(\mathbf{X}_2^T\mathbf{X}_2)^{-1}\mathbf{X}_2^T\mathbf{x}_j]^{-1}.$$

- It turns out that $var(\hat{\beta}_j^*) \geq var(\hat{\beta}_j)$ for $j = 1, 2, \ldots, k$.

Impact of Model Misspecification (Overfitting)

- Suppose, we have a short and a long model and we want to predict at $\mathbf{x}_0^T = [\mathbf{x}_{1,0}^T, \mathbf{x}_{2,0}^T]$.
- The short model only uses $\mathbf{x}_{1,0}^T$, while the long model uses both.
- Let, $\hat{y}_S(\mathbf{x}_{1,0}^T)$ denotes the prediction via short model and $\hat{y}_L(\mathbf{x}_0^T)$ denotes the prediction using long model.
- It can be shown that $var[\hat{y}_L(\mathbf{x}_0^T)] \geq var[\hat{y}_S(\mathbf{x}_{1,0}^T)]$.
- Therefore, the prediction variance for the longer model is greater than short model.

Impact of Model Misspecification (Overfitting)

- Assuming the normality of the ϵ_i, the $\hat{\sigma}^2$ follows a chi-square distribution i.e.

$$\frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p},$$

which implies that

$$\hat{\sigma}^2 \sim \frac{\sigma^2 \chi^2_{n-p}}{n-p}.$$

- Thus we have:

$$\begin{aligned} var(\hat{\sigma}^2) &= var\left(\frac{\sigma^2 \chi^2_{n-p}}{n-p}\right) = \left(\frac{\sigma^4 var(\chi^2_{n-p})}{(n-p)^2}\right) \\ &= \frac{2\sigma^4(n-p)}{(n-p)^2} = \frac{2\sigma^4}{n-p}. \end{aligned}$$

- Thus, when the model is overfit, the variance of $\hat{\sigma}^2$ is inflated.

Model Selection

- Given a set of possible regressor variables, there are many different models that one may choose to use.
- The goal is to select the ”best” model.
- What is the ”best” model???
- We want a simple or parsimonious model that explains a large proportion of the variation in response.
- We want to avoid underfitting, so the results be unbiased.
- We want to avoid overfitting, so model do not suffer the variance inflation.

Model Selection

- In order to choose a model, one should decide the goal of the modeling before hand.
- Cross-Validation, Press statistics, C_p, etc. are among the criteria that one can use for model selection.
- Model selection can be based on an information criteria such as AIC, BIC, etc.
- Using some penalization to select the variables in the model.

Model Selection (Cross-Validation)

- Any variable selection routine is using data to help determine the model.
- Thus, it is possible to overfit the tailor the model too precisely to the data at hand.
- This event, generally, leads to a model that is great for our sample data but does not work very well in general.
- Cross-validation (CV) can be used to verify the prediction capability of the model.
- Ideally, CV should be a part of all model selection application.

Model Selection (Cross-Validation)

- The CV procedure is as follow:

1. Randomly split the data into two parts (They do not need to be the same size).
2. Use a part of data to fit the model and estimate the model parameters. This part is called training data set.
3. Use the estimated parameter for the model and predict the response variable for the second data set. This part of the data is called test data.
4. Calculate

$$CV = \frac{1}{n_2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

   where n_2 denotes the sample size for the test data.

5. Choose the model that minimizes the CV.

Model Selection (Cross-Validation)

- Another way to do cross-validation is to use Pearson correlation.
- Do the first three steps of the CV procedure.
- Calculate $r(y, \hat{y})$ to test $H_0 : \rho = 0$ versus $H_1 : \rho > 0$.
- The test statistic is

$$t_{obs} = r(y, \hat{y}) \sqrt{\frac{n_2 - 2}{1 - [r(y, \hat{y})]^2}} \sim t_{n_2 - 2}.$$

- If we fail to reject the null hypothesis, then CV is unsuccessful. This implies the model has been tailored too precisely to the training data and one should fit another model.
- If we reject the null hypothesis, then CV is successful. This implies the model is good. Thus, we rerun the model using all the data to obtain the the most stable estimate for the regression parameters.

Model Selection

- Another approach for model selection is to use different statistics from regression model.
- This approach is all-possible regression approach.
- Here we want to simultaneously:
  - Maximize the R^2 or the adjusted R^2.
  - Minimize the error mean square, standard errors of the coefficients' estimate, of standard error of the prediction.
- The advantage of this model is that every possible model is considered.
- The disadvantage of this model is that every possible model is considered.

Model Selection (Sequential Procedure)

- The all-possible regression approach can be computationally expensive when the number of covariates is large.
- One can use the sequential approach which involves adding or deleting one variable at a time from a current model in a sequential approach to choose a best model.
- The procedure make use of partial F-test, to determine if the j-th regressor should be added (or deleted) from the model given by:

$$F_{obs} = \frac{SS_{MODEL,FULL} - SS_{MODEL,SUB}}{MS_{ERROR,FULL}}.$$

  where the full model includes all the variables currently in the model plus j-th regressors.
- Since MS_{ERROR} is always used, the procedure occasionally have poor estimate of the σ^2. Thus, large critical values are typically used for this procedure.

Model Selection (Sequential Procedure)

- The first sequential approach is Forward Selection:
  1. Begin with no variables in the model.
  2. Add the variable that has the largest significant partial F (if any).
  3. Now add the variable that has the largest significant partial F given the previous variable is in the model.
  4. Continue adding variable in this manner until no other variables are significant or all the variables are included in the model.
- Advantage: This is an intuitive approach.
- Disadvantage: Inflexible: If one variable is added then it cannot be removed. Thus, the best model may never be considered.

Model Selection (Sequential Procedure)

- Backward Elimination:

  1. Begin with all the variables in the model.
  2. Remove the variable with the smallest significant partial F (if any).
  3. Remove the variable with the smallest significant partial F in presence of all the variables still in the model.
  4. Continue removing variables until all remaining variables are significant or all the variables have been removed.

- Advantage: Intuitively appealing.
- Disadvantage: Inflexible: Once a variable is removed then it cannot reenter. Thus, the best model may never be considered.

Model Selection (Sequential Procedure)

- Stepwise Procedure:
  1. We begin with no variable in the model.
  2. Add the variable that has the largest significant partial F (if any).
  3. Test to see if this variable should remain in the model.
  4. Now add the variable that has the largest significant partial F given the previous variable is in the model.
  5. Remove any one variable that is not significant in the presence of the other variable.
  6. Continue adding and removing variable in this manner until all significant variables have been added to the model.
- Advantage: Much more flexible compared to forward and backward procedures.

Model Selection (Sequential Procedure)

- The procedures leave many models uncomputed and therefore, we are not guaranteed to find the best model.
- The results depend on the significance level chosen.
- The three procedure often give different models.
- Partial F-test may not be valid F-test.
- Many statisticians prefer backward elimination over forward selection, feeling it safer to delete terms from an overly complex model than to add terms to an overly simple one.
- Forward selection can stop prematurely because a particular test in the sequence has low power.
- Neither strategy necessarily yields a meaningful model.
- Use variable selection procedures with caution!

Model Selection (PRESS Statistic)

- The PRESS statistic is not based on the fitting capability of the model.
- Instead, it is based on model validation.
- Since $\hat{\mathbf{y}} = \mathbf{Hy}$, then $\hat{y}_i = h_{ii} y_i$ and obviously y_i and $\hat{y}_i$ are not independent.
- The PRESS statistic uses residuals of the form $y_i - \hat{y}_{i,-i}$ where $\hat{y}_{i,-i}$ is the predicting the response without the use of i-th observation in the model fitting process.
- This makes the y_i and $\hat{y}_i$ independent from each other.
- The PRESS statistic can be seen as a special case of the cross validation where the training data contains $n-1$ observations and test data includes the other one.

Model Selection (PRESS Statistic)

- PRESS statistic is calculated by

$$PRESS = \sum_{i=1}^{n} (y_i - \hat{y}_{i,-i})^2.$$

- Although the calculation of the PRESS may sounds like a lot of work, it can be shown that

$$e_{i,-i} = y_i - \hat{y}_{i,-i} = \frac{y_i - \hat{y}_i}{1 - h_{ii}} = \frac{e_i}{1 - h_{ii}},$$

- Thus, large PRESS residuals are associated with large HAT diagonals.
- For model comparison, the smaller PRESS is the better.
- We can also turn PRESS into a R^2-type statistic by

$$R^2_{PRED} = 1 - \frac{PRESS}{SS_{TOTAL}},$$

Model Selection (PRESS Statistic)

- Comments on PRESS statistic:
- Advantage: a single number criteria that react to collinearity and that focuses on the prediction capabilities of the model.
- Disadvantage: It is not scale-free (standardized), and can ”blow up” due to a single high leverage point.

Model Selection (C_p Statistic)

- Recall that in an underfit situation there is the creation of bias, and in the overfit situation there are inflation in variances.
- The Mallow's C_p statistic strives for compromise by considering both the bias and variance simultaneously.
- To do this, the mean square error for prediction is used:

$$MSE[\hat{y}(\mathbf{x}_0)] = E\{\hat{y}(\mathbf{x}_0) - E[\hat{y}(\mathbf{x}_0)]\}^2 = var[\hat{y}(\mathbf{x}_0)] + \{BIAS[\hat{y}(\mathbf{x}_0)]\}^2.$$

- Consider an underfit situation:

$$\begin{aligned} \text{Proposed Model} &: \mathbf{y} = \mathbf{X}_1\beta_1 + \epsilon; \text{where } \mathbf{X}_1 \text{ is } n \times p \\ \text{True Model} &: \mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \epsilon, \end{aligned}$$

Model Selection (C_p Statistic)

- For this situation, we have
- $var[\hat{y}(\mathbf{x}_{1,0})] = \sigma^2 \mathbf{x}_{1,0}^T (\mathbf{X}_1^T \mathbf{X}_1)^{-1} \mathbf{x}_{1,0}$ and
- $BIAS[\hat{y}(\mathbf{x}_0)] = (\mathbf{x}_{1,0}^T \mathbf{A} - \mathbf{x}_{2,0}^T)\beta_2$, where $\mathbf{A} = (\mathbf{X}_1^T \mathbf{X}_1)^{-1} \mathbf{X}_1^T \mathbf{X}_2$.
- Therefore, the mean square error for prediction becomes:

$$MSE[\hat{y}(\mathbf{x}_0)] = \sigma^2 \mathbf{x}_{1,0}^T (\mathbf{X}_1^T \mathbf{X}_1)^{-1} \mathbf{x}_{1,0} + \beta_2^T (\mathbf{x}_{1,0}^T \mathbf{A} - \mathbf{x}_{2,0}^T)^T (\mathbf{x}_{1,0}^T \mathbf{A} - \mathbf{x}_{2,0}^T)\beta_2.$$

- C_p works with the predicted values $\hat{y}(\mathbf{x}_0)$, and is an estimate of the total mean square error.
- We creat a standardize (or scale free) statistics by dividing by σ^2, creating

$$C_p = \frac{\sum_{i=1}^n MSE[\hat{y}(\mathbf{x}_0)]}{\sigma^2} = \frac{\sum_{i=1}^n var[\hat{y}(\mathbf{x}_i)]}{\sigma^2} + \frac{\sum_{i=1}^n \{BIAS[\hat{y}(\mathbf{x}_i)]\}^2}{\sigma^2}.$$

Model Selection (C_p Statistic)

- The Variance part:

$$var[\hat{y}(\mathbf{x}_i)] = \sigma^2 \mathbf{x}_i^T (\mathbf{X}_1^T \mathbf{X}_1)^{-1} \mathbf{x}_i = \sigma^2 h_{ii}.$$

- Thus we have

$$\begin{aligned}
\frac{\sum_{i=1}^n var[\hat{y}(\mathbf{x}_i)]}{\sigma^2} &= \frac{\sum_{i=1}^n \sigma^2 h_{ii}}{\sigma^2} \\
&= \sum_{i=1}^n h_{ii} = trace(\mathbf{H}_1) = trace[\mathbf{X}_1 (\mathbf{X}_1^T \mathbf{X}_1)^{-1} \mathbf{X}_1^T] \\
&= trace[\mathbf{X}_1^T \mathbf{X}_1 (\mathbf{X}_1^T \mathbf{X}_1)^{-1}] = trace(\mathbf{I}_{p \times p}) \\
&= p = \text{number of parameters in the model}
\end{aligned}$$

Model Selection (C_p Statistic)

- The Bias part:

$$BIAS[\hat{y}(\mathbf{x}_0)] = (\mathbf{x}_{1,0}^T\mathbf{A} - \mathbf{x}_{2,0}^T)\beta_2,$$

  where $\mathbf{A} = (\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{X}_2$.
- Then a vector of Bias is $\mathbf{BIAS} = (\mathbf{X}_1\mathbf{A} - \mathbf{X}_2)\beta_2 = -(\mathbf{X}_2 - \mathbf{X}_1\mathbf{A})\beta_2$.
- Thus:

$$\begin{aligned}
\sum_{i=1}^{n}\{BIAS[\hat{y}(\mathbf{x}_i)]\}^2 &= \beta_2^T(\mathbf{X}_2 - \mathbf{X}_1\mathbf{A})^T(\mathbf{X}_2 - \mathbf{X}_1\mathbf{A})\beta_2 \\
&= \beta_2^T(\mathbf{X}_2^T\mathbf{X}_2 - \mathbf{A}^T\mathbf{X}_1^T\mathbf{X}_2 - \mathbf{X}_2^T\mathbf{X}_1\mathbf{A} + \mathbf{A}^T\mathbf{X}_1^T\mathbf{X}_1\mathbf{A})\beta_2 \\
&= \beta_2^T\{\mathbf{X}_2^T\mathbf{X}_2 - [(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{X}_2]^T\mathbf{X}_1^T\mathbf{X}_2 \\
&- \mathbf{X}_2^T\mathbf{X}_1[(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{X}_2] \\
&+ [(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{X}_2]^T\mathbf{X}_1^T\mathbf{X}_1[(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{X}_2]\}\beta_2 \\
&= \beta_2^T\mathbf{X}_2^T(\mathbf{I} - \mathbf{H}_1)\mathbf{X}_2\beta_2
\end{aligned}$$

Model Selection (C_p Statistic)

- Recall that for an underfit model,

$$E(\hat{\sigma}^2) = \sigma^2 + \frac{1}{n-p}[\beta_2^T \mathbf{X}_2^T (\mathbf{I} - \mathbf{H}_1)\mathbf{X}_2 \beta_2],$$

- Then, we have

$$\sum_{i=1}^{n} \{BIAS[\hat{y}(\mathbf{x}_i)]\}^2 = (n-p)[E(\hat{\sigma}^2) - \sigma^2].$$

- So an unbiased estimate of $\frac{\sum_{i=1}^{n} \{BIAS[\hat{y}(\mathbf{x}_i)]\}^2}{\sigma^2}$ is

$$\frac{\sum_{i=1}^{n} (n-p)[\hat{\sigma}^2 - \sigma^2]}{\sigma^2},$$

$\hat{\sigma}^2$ is the MS_{ERROR} for the model under consideration.
σ^2 is unknown and thus should be estimated.
An unbiased estimate for σ^2 is $MS_{ERROR,FULL}$.

Model Selection (C_p Statistic)

- Thus, C_p statistic is

$$C_p = p + \frac{(n-p)(\hat{\sigma}^2 - \hat{\sigma}^2_{FULL})}{\hat{\sigma}^2_{FULL}}.$$

- We want the C_p to be small.
- For the full model, $\hat{\sigma}^2 = \hat{\sigma}^2_{FULL}$. Therefore, $C_p = p$.
- Since $\hat{\sigma}^2_{FULL}$ is not necessarily the smallest MS_{ERROR}. Therefore, some C_p values can be less than p.
- $C_p >> p^*$ implies an underfit model (p^* is the number of parameters in the considered model).
- $C_p \approx p^* (or < p^*)$ implies good model.
- For a list of candidate models, we often plot C_p versus p^*. Point on or below the line $C_p = p^*$ indicate good models.
- C_p does not punish collinearity as much as it should. Thus, it favors more complicated model.

Model Selection (Information Criteria)

- **AIC:**
  - The Akaike information criterion (AIC) is an estimator of the relative quality of statistical models for a given set of data.
  - It can be calculate by:

$$AIC = 2p - 2ln(L).$$

- **BIC:**
  - Bayesian Information Criterion (BIC) is also an estimator of the relative quality of a statistical model.
  - BIC can be calculated by:

$$BIC = pln(n) - ln(L).$$

Variable Selection

- We already discussed the partial t-test and F-test approaches for selecting the variables.
- One can also use information criteria to select informative variables.
- In addition, one can use penalization to select the variables.
- Recall the penalization with Ridge regression.
- LASSO.

Variable Selection

- Objective function of Ridge:

$$\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2$$

subject to $\sum_{j=1}^{p} \beta_j^2 \leq t$.

- Drawback of the ridge regression:
- Unlike best subset selection or stepwise selection ridge regression does not select a model.
- That is at the beginning, we start with p predictors and fi nally we end up with all p predictors.

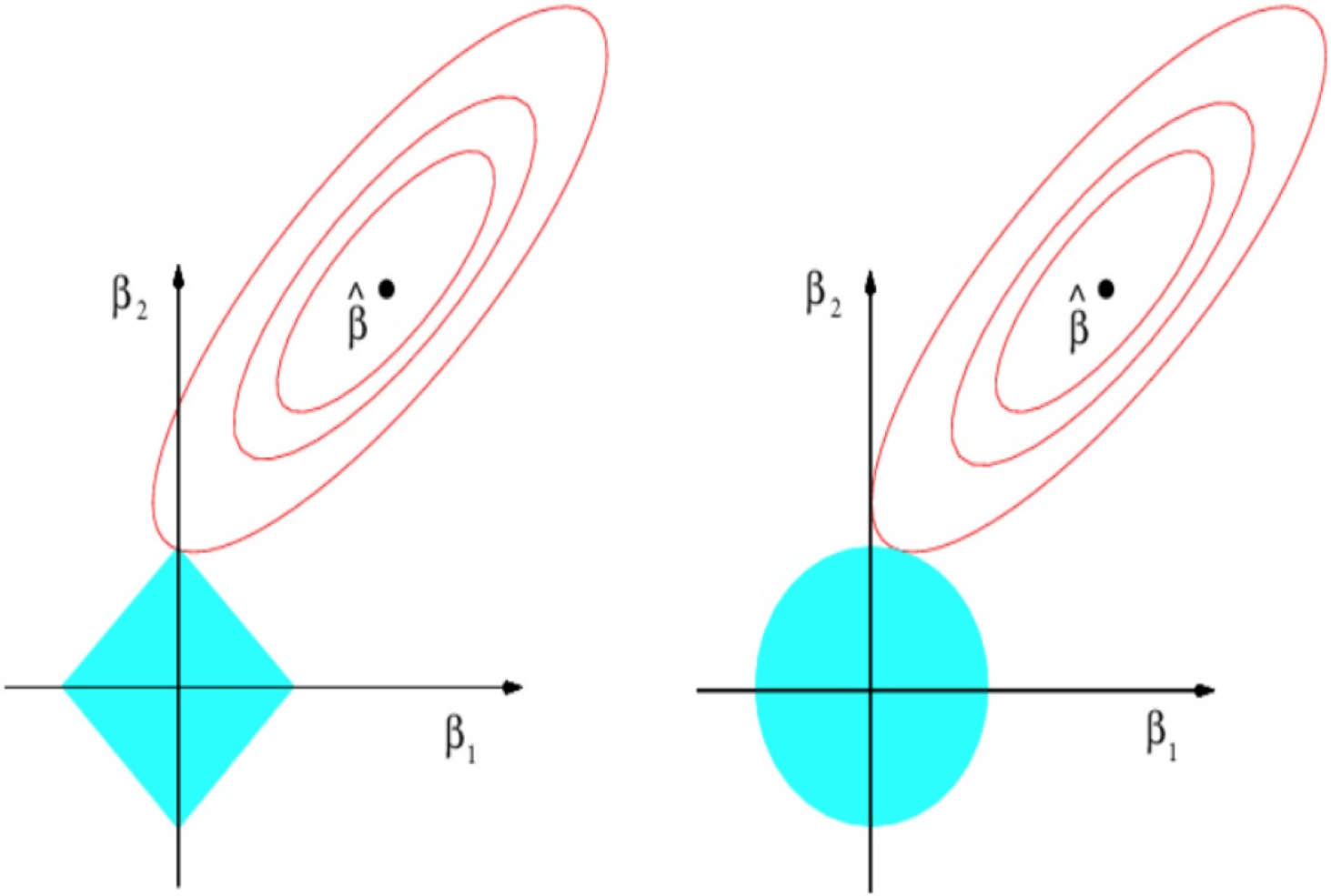
Variable Selection

- LASSO: Least Absolute Shrinkage And Selection Operator.
  - The Lasso is a shrinkage and selection method for linear regression.
  - It minimizes the usual sum of squared errors, with a bound on the sum of the absolute values of the coefficients.
  - It has connections to soft-thresholding of wavelet coefficients, forward stagewise regression, and boosting methods.
- LASSO estimate are obtain by solving following optimization problem:

$$\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2$$

subject to $\sum_{j=1}^{p} |\beta_j| \leq t$.

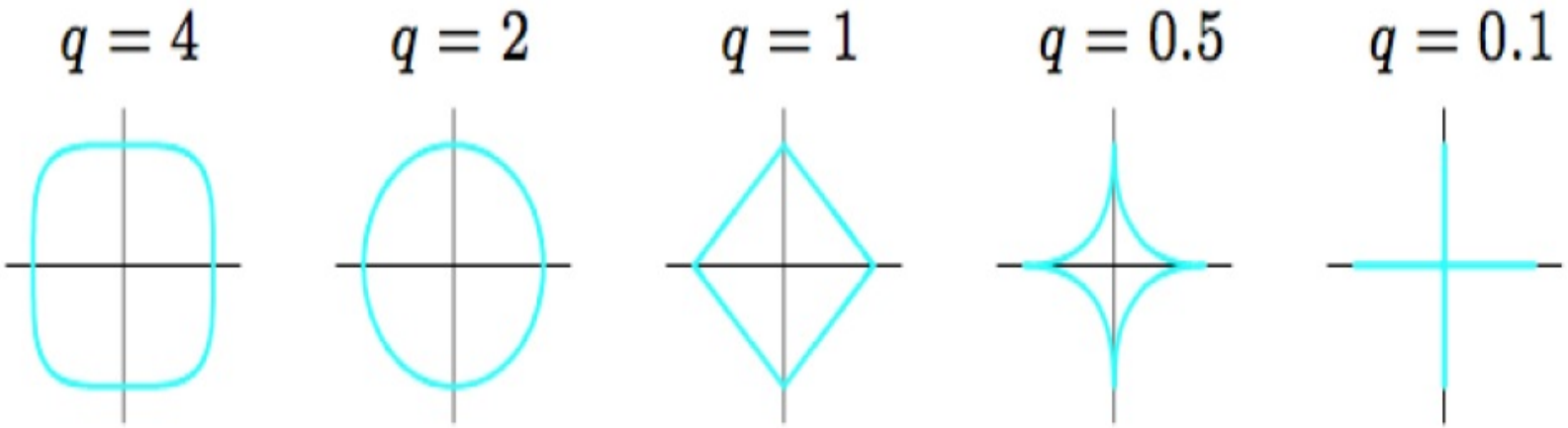
Variable Selection



Variable Selection

- To make it more general

$$\hat{\beta}^{PEN} = \text{argmin}_{\beta} \left\{ (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right\}$$



Transformation

- The regression techniques have certain assumptions.
- Violation of the assumptions can kead to major problems and make the regression model to be useless.
- Thus, we should always check the assumptions to see if they are at least approximately met.
- Data transformation is one thing that we can do if the assumptions are not met.
- The data transformation will hopefully do one (or more) of these things:
  1. Create a better fit and better prediction,
  2. Stabilize variance,
  3. Make errors more normal.

Transformation

- Partial residual plots are one tool that can indicate the possible need for transformations.
- This usually results if standard deviation of the y's (σ_y) is proportional to the mean of the y's (μ_y) raised to a power; i.e.

$$\sigma_y \propto \mu_y^r.$$

- The regression model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- When $\epsilon_i \sim N(0, \sigma_i^2$, then

$$var(\boldsymbol{\epsilon}) = diag(\sigma_1^2, \ldots, \sigma_n^2)$$

Weighted Least Square

- If we define a weight matrix $\mathbf{W}$ as

$$\mathbf{W} = diag(w_1, \ldots, w_n); \quad \forall i: \; w_i = \frac{1}{\sigma_i^2} \;\; or \;\; w_i = \frac{k}{\sigma_i^2}$$

  Then the objective function of OLS becomes

$$(\mathbf{y} - \mathbf{X}\beta)^T \mathbf{W} (\mathbf{y} - \mathbf{X}\beta)$$

- Thus we have

$$\hat{\beta} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{y}$$

- With

$$var(\hat{\beta}) = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1}$$

Transformation

- In the situation that the error variance is not independent of the mean, often stabilize variance and make errors more normal are tied together.
- Examples in which standard deviation is moving with the mean are:
  - Poisson: Mean and variance are both equal to λ ($\sigma_y \propto \mu_y^{\frac{1}{2}}$).
  - Exponential: Variance equals to the square of the mean ($\sigma_y \propto \mu_y$).
- It turns out that the only error distribution in which homogeneous variance is natural is the normal or Gaussian distribution.
- If the response is not normal (discrete, binary, etc.), the variance will not be stable.
- Hence a transformation to stabilize the variance is also will make the errors more normal.
- Often $\sigma_y = c\mu_y^r$, with c being a positive constant.

Transformation to Stabilize Variance

- We want to transform y, using $z = f(y)$, such that $\sigma_z = c$.
- Using a Taylor series expansion, it can be shown that if $z = f(y)$, then

$$\sigma_z \cong \left| \left\{ \frac{\partial f(y)}{\partial y} \right\}_{y=\mu_y} \right| \sigma_y$$

- We want to choose the z=f(y) such that σ_z is a constant.
- It turns out that if $\sigma_y = c\mu_y^r$, then the appropriate variance stabilizing transformation is:

$$z = \begin{cases} y^{-r+1} & ifr \neq 1 \\ ln(y) & ifr = 1 \end{cases}$$

Transformation to Stabilize Variance

- Once a transformation is made, we should check residual plots, influence diagnostics, etc. to be sure that the problems have been corrected.
- Usually a transformation used to clear up one problem will often help other things too.
- Exponential distribution.
- Poisson distribution.
- Proportional response.
- Proportion data assuming a binomial distribution.

Transformation to Improve Fit and Prediction

- The goal is no longer to stabilize the variance, but instead to improve the fit and prediction capabilities of the model.
- There are two procedures that are design to accomplish this.
  - Box-Tidwell Procedure.
  - Box-Cox procedure.

Transformation Box-Tidwell

- The Box-Tidwell procedure looks to transform one or more of the regressors.
- The Box-Tidwell model is

$$y_i = \beta_0 + \beta_1 x_{1i}^{\alpha_1} + \ldots + \beta_p x_{pi}^{\alpha_p} + \epsilon_i.$$

- Therefore, we can have just about everything except interactions.
- To leave a regressor untransformed, set $\alpha_j = 1$.
- If $\alpha_j = 0$, we use the log transformation.

Transformation Box-Tidwell

- In the Box-Tidwell transformation procedure, the parameter set is

$$\{\beta_0, \beta_1, \ldots, \beta_p, \alpha_1, \ldots, \alpha_p\}.$$

- With this parameter set the model is not linear in the parameters.
- This is an iterative procedure.
- Often, researcher use the one step procedure because after that, the nonlinear regression is more appropriate.

One Step Box-Tidwell

- For simplicity, let p = 2 where p denotes the number of regressors.
- The Box-Tidwell transformation procedure involves taylor series expansion around $\alpha_1 = \alpha_{1,0}$ and $\alpha_2 = \alpha_{2,0}$.
- We start with no transformation ($\alpha_{1,0} = \alpha_{2,0} = 1$). Then, we have

$$
\begin{aligned}
E(\mathbf{y}) &= f(\boldsymbol{\alpha}_0^T) + \sum_{j=1}^{k} \left|\frac{\partial f}{\partial \alpha_j}\right|_{\alpha_j=\alpha_{j,0}} (\alpha_j - \alpha_{j,0}) \\
&= \beta_0 + \beta_1 \mathbf{x}_1^{\alpha_{1,0}} + \beta_2 \mathbf{x}_2^{\alpha_{2,0}} \\
&+ \left|\frac{\partial f}{\partial \alpha_1}\right|_{\alpha_1=\alpha_{1,0}} (\alpha_1 - \alpha_{1,0}) + \left|\frac{\partial f}{\partial \alpha_2}\right|_{\alpha_2=\alpha_{2,0}} (\alpha_2 - \alpha_{2,0}) \\
&= \beta_0 + \beta_1 \mathbf{x}_1^{\alpha_{1,0}} + \beta_2 \mathbf{x}_2^{\alpha_{2,0}} \\
&+ \beta_1 \mathbf{x}_1 ln(\mathbf{x}_1)(\alpha_1 - \alpha_{1,0}) + \beta_2 \mathbf{x}_2 ln(\mathbf{x}_2)(\alpha_2 - \alpha_{2,0}) \\
&= \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 \\
&+ \beta_1 \mathbf{x}_1 ln(\mathbf{x}_1)(\alpha_1 - 1) + \beta_2 \mathbf{x}_2 ln(\mathbf{x}_2)(\alpha_2 - 1).
\end{aligned}
$$

One Step Box-Tidwell

- Let $\mathbf{z}_1 = \mathbf{x}_1 ln(\mathbf{x}_1)$, $\mathbf{z}_2 = \mathbf{x}_2 ln(\mathbf{x}_2)$, $\gamma_1 = \beta_1(\alpha_1 - 1)$, and $\gamma_2 = \beta_2(\alpha_2 - 1)$.

- Thus we have

$$E(\mathbf{y}) = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \gamma_1 \mathbf{z}_1 + \gamma_2 \mathbf{z}_2.$$

One Step Box-Tidwell

1. Using OLS, fit $\mathbf{y} = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \epsilon$. This gives us estimates for $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$.

2. Using OLS, fit $\mathbf{y} = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \gamma_1 \mathbf{z}_1 + \gamma_2 \mathbf{z}_2 + \epsilon$. This gives us estimate for $\hat{\gamma}_1$ and $\hat{\gamma}_2$. The estimates for $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ are not important here.

3. Solve for α's usig

$$\hat{\alpha}_j = \frac{\hat{\gamma}_j}{\hat{\beta}_j} + 1.$$

One Step Box-Tidwell

- If one of the $\hat{\alpha}_j$ is close to 0, then the appropriate transformation is $ln(\mathbf{x}_j)$, otherewise the transforation is $\mathbf{x}_j^{\hat{\alpha}_j}$.
- If $\hat{\gamma}_j \approx 0$, then it means that transformation is not necessary for that variable.
- The general Box-Tidwell procedure is a repeated procedure that updates $\hat{\alpha}_j$ in each iterations.
- For general p-regressors, the procedure will produces p $\hat{\alpha}_j$, where

$$\hat{\alpha}_j = \frac{\hat{\gamma}_j}{\hat{\beta}_j} + 1.$$

Transformation Box-Cox

- The Box-Cox transformation tries to improve the fit and prediction by transforming response variable instead of regressors.
- This method tryies to estimate the λ in th emodel

$$y_i^{\lambda} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + \epsilon_i.$$

- Suppose we have model $z_i = \mathbf{x}_i^T \beta + \epsilon_i$, where z_i are measured in the same units as y, and we let ϵ_i to be normally distributed.
- Using ML approach, we want to maximize the $L(z_i)$. With normality assumptions, this is equivalent to minimizing the SSE.

Transformation Box-Cox

- Let $\tilde{y}$ to be the geometric means of y's, then

$$z_i = \begin{cases} \frac{y_i^{\lambda}-1}{\lambda \tilde{y}^{\lambda-1}} & If \quad \lambda \neq 0 \\ \frac{ln(y_i)}{\tilde{y}^{\lambda-1}} & If \quad \lambda = 0 \end{cases}$$

- The MLE of $\hat{\sigma}^2$ is $\frac{SSE}{n}$, which is biased by $\frac{n}{n-p-1}$.

  Actually $\hat{\sigma}^2$ depends on the value of λ and can be written as $\hat{\sigma}^2_{\lambda}$.

Transformation Box-Cox

1. Select a grid of λ values.
2. For each fixed λ regress **z** vs. **x**.
3. Compute *SSE* and $-\frac{n}{2} ln(\hat{\sigma}^2_\lambda)$.
4. Plot $-\frac{n}{2} ln(\hat{\sigma}^2_\lambda)$ vs. λ.
5. Choose λ value that produces the largest value of $-\frac{n}{2} ln(\hat{\sigma}^2_\lambda)$.
6. Regress $\mathbf{y}^{\hat{\lambda}}$ vs. **x** and do regular regression analysis.

Transformation Box-Cox

- To decide what natural transformation to make, it would be useful to have an interval of λ values that are statistically equivalent to the $\hat{\lambda}$ that maximize the likelihood function.
- Such a confidence interval would consist of any λ_0 values that a test of $H_0 : \lambda = \lambda_0$ does not find significantly different from $\hat{\lambda}$.
- This procedure involves using the likelihood ratio statistic.

Transformation Box-Cox

- **The General likelihood ratio test:**
  - To test any null hypothesis, the proper likelihood ratio test is

$$LR = \frac{L_{MAX}(\text{restricted to the null hypothesis})}{L_{MAX}(\text{unrestricted})}.$$

  - Evidence favoring the null hypothesis results in a likelihood ratio close to 1.
  - It can be shows that asymptotically

$$-2ln(LR) \sim \chi^2_{\Delta}.$$

    where Δ denotes [the number of parameters under the under unrestricted scenario] - [the number of parameters under the under restricted scenario]

Transformation Box-Cox

- In our case, we are testing $H_0 : \lambda = \lambda_0$ and the range of λ_0 that are not rejected is an upper one-sided test from a confidence interval.
- We know that $ln[L_{MAX}(\mathbf{z})] = -\frac{n}{2} ln(\hat{\sigma}^2_\lambda)$.
- For the restricted case, the best estimate is the hypothesis value i.e. $\lambda = \lambda_0$.
- For the unrestricted case, the best estimate is the MLE i.e. $\lambda = hat\lambda$.
- Thus, we have:

$$\gamma = -2ln(LR) = -2\left\{[-\frac{n}{2} ln(\hat{\sigma}^2_{\lambda_0})] - [-\frac{n}{2} ln(\hat{\sigma}^2_\lambda)]\right\} = nln(\hat{\sigma}^2_{\lambda_0}) - nln(\hat{\sigma}^2_\lambda)$$

- with $\Delta = 1$.
- Therefore, any λ_0 such that

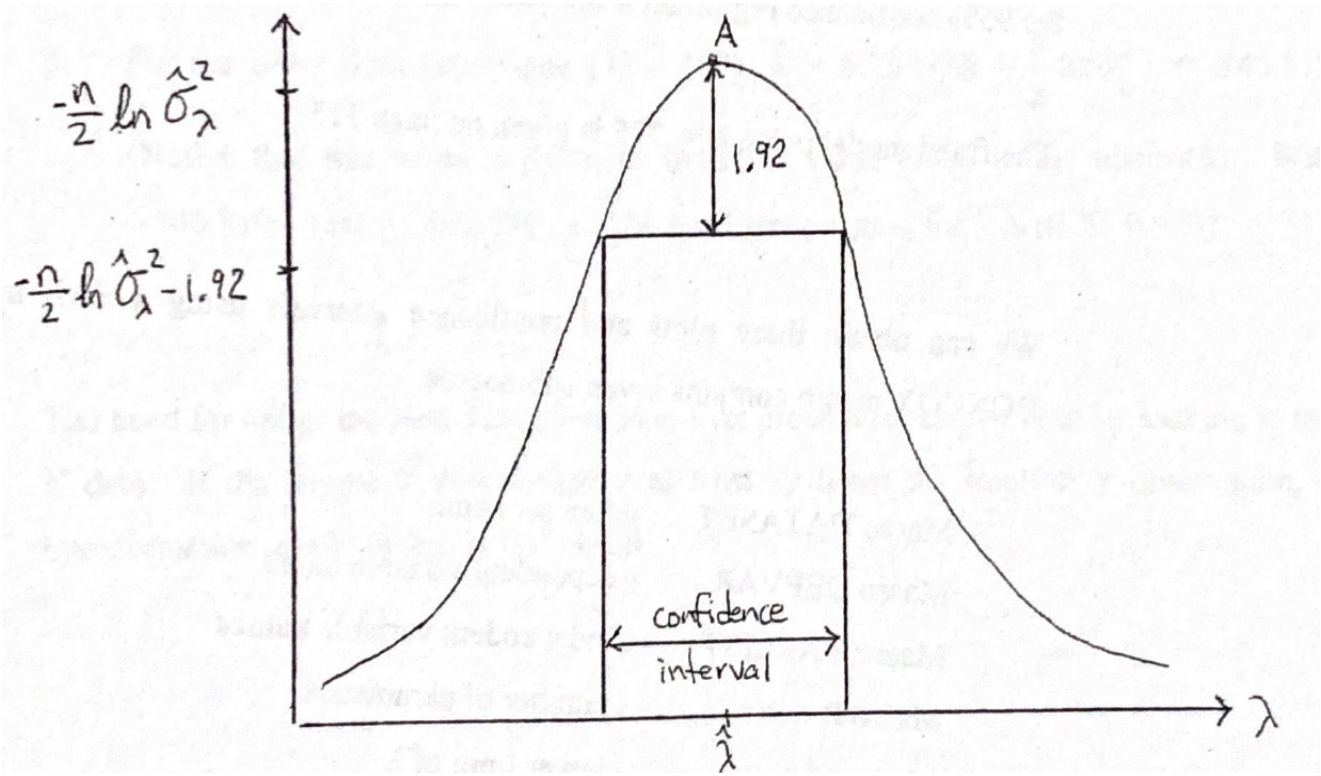
$$nln(\hat{\sigma}^2_{\lambda_0}) - nln(\hat{\sigma}^2_\lambda) \leq \chi^2_{0.05,1} = 3.84,$$

will fall inside the 95% CI of λ.

Transformation Box-Cox

- We can divide everything by 2 and have

$$\frac{n}{2} ln(\hat{\sigma}^2_{\lambda_0}) - \frac{n}{2} ln(\hat{\sigma}^2_{\lambda}) \leq 1.92.$$



Residual Analysis

- Suppose that we have identified several candidate models using model selection techniques.
- Now, we have to identify outliers and high leverage point.
- This is called residual analysis.
- Recall that the vector of residuals is $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$.
- We expect that $E(\mathbf{e}) = 0$ and $var(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H})$ where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.
- Since $var(\mathbf{e}) \neq \sigma^2\mathbf{I}$, thus the residuals do not have a constant variance and are correlated.

Studentized Residual

- The vector of residuals is

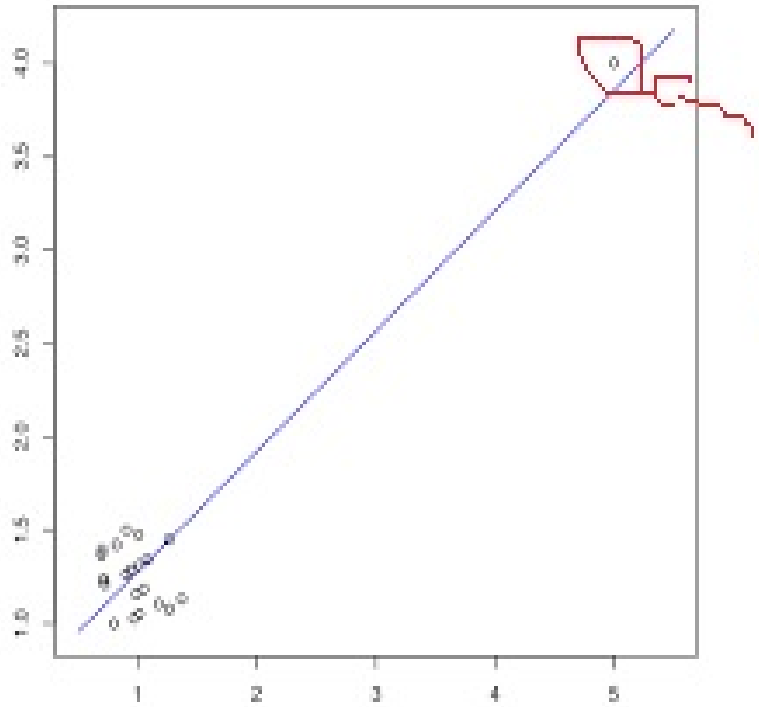
$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\beta} = \mathbf{y} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y}.$$

- Thus, the i-th residual is $e_i = (1 - h_{ii})y_i$.
- Since the residuals have heterogeneous variance, they should not be used (alone) as diagnostic tools.
- However, if we standardize a residual, we create a scale-free statistic that can be used as diagnostic tool.
- This will be

$$\frac{e_i - E(e_i)}{SE(e_i)} = \frac{y_i - \hat{y}_i}{\sigma\sqrt{1 - h_{ii}}}.$$

- A high leverage point is a point with a large HAT diagonal.
- We use $\sqrt{MS_{ERROR}}$ as the estimate for σ.

Residual Plot



Large h_{ii},

but small residual

Studentized Residual

- The $\frac{y_i - \hat{y}_i}{\sigma\sqrt{1-h_{ii}}}$ follow a t-like distribution.
- Also, we can use this to see if the residuals are truly have a mean of zero.
- As a rule of thumb, the standardized residual less than -2 or greater than +2 may indicate that the point is an outlier.
- If those points have large h_{ii}, then that point is a potential high influence point.
- A model that has several large standardized residuals together with several large HAT diagonals is usually in violation of one or more of the assumptions.
- Recall that large $\frac{y_i - \hat{y}_i}{\sigma\sqrt{1-h_{ii}}}$ denotes outlier, large h_{ii} denotes high leverage point, and when both of these are high, it is a high influence points.

Studentized Residual

- The U.S. Navy attempts to develop equations for estimation of manpower needs for manning installations such as Bachelor Officers Quarters. The data were collected at 25 BOQ sites.
  - X1: Average daily occupancy
  - X2: Monthly average number of check-ins
  - X3: Weekly hours of service desk operation
  - X4: Square feet of common use area
  - X5: Number of building wings
  - X6: Operational berthing capacity
  - X7: Number of rooms
  - y: Monthly man-hours

R-Student

- The studentized residuals do not have an exact t-distribution. (Why??)
- The dependence between the numerator and denominator can be eliminated if instead of estimating σ with $\sqrt{MS_{ERROR}}$, use $\hat{\sigma}_{-i}$ which is $\sqrt{MS_{ERROR}}$ calculated with the i-th observation removed.
- Thus, we have

$$t_i = \frac{y_i - \hat{y}_i}{\hat{\sigma}_{-i}\sqrt{1 - h_{ii}}} \sim t_{n-p-1}$$

- R-student is more sensitive to violation in the model than studentized residuals.

PRESS residuals

- Note that both studentized and R-student are standardized PRESS residuals.

- The variance of a PRESS residuals is

$$\begin{aligned} var(e_{i,-i}) &= var\left(\frac{e_i}{1-h_{ii}}\right) = \left(\frac{1}{1-h_{ii}}\right)^2 var(e_i) \\ &= \left(\frac{1}{1-h_{ii}}\right)^2 [\sigma^2(1-h_{ii})] = \frac{\sigma^2}{1-h_{ii}}. \end{aligned}$$

- Then a standardized PRESS residual is

$$\frac{e_{i,-i} - E(e_{i,-i})}{\sqrt{var(e_{i,-i})}} = \frac{\frac{e_i}{1-h_{ii}} - 0}{\sqrt{\frac{\sigma^2}{1-h_{ii}}}} = \frac{e_i}{\sigma\sqrt{1-h_{ii}}}.$$

where the studentized residuals estimates σ with $\hat{\sigma}$ and R-student estimates σ with $\hat{\sigma}_{-i}$

PRESS residuals

- We define a point to be an outlier if $E(\epsilon_i) = \Delta_i \neq 0$.

  This represents a location shift off the regression line in the response direction.

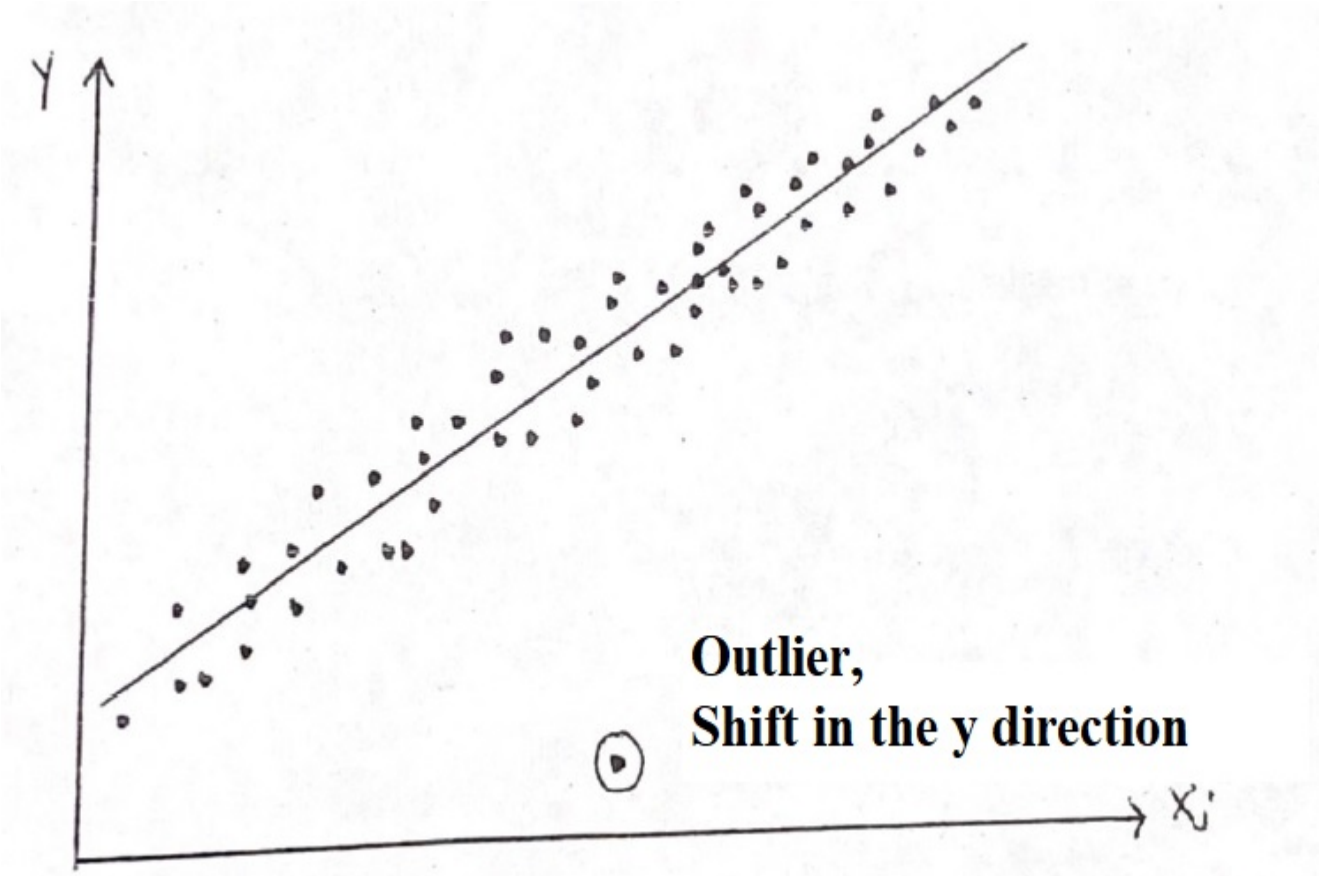
- A high leverage point is a point that potentially exerts undue influence on at least one regression coefficients.

- A high leverage point falls outside the mainstream of the regressors data along with the general trend.

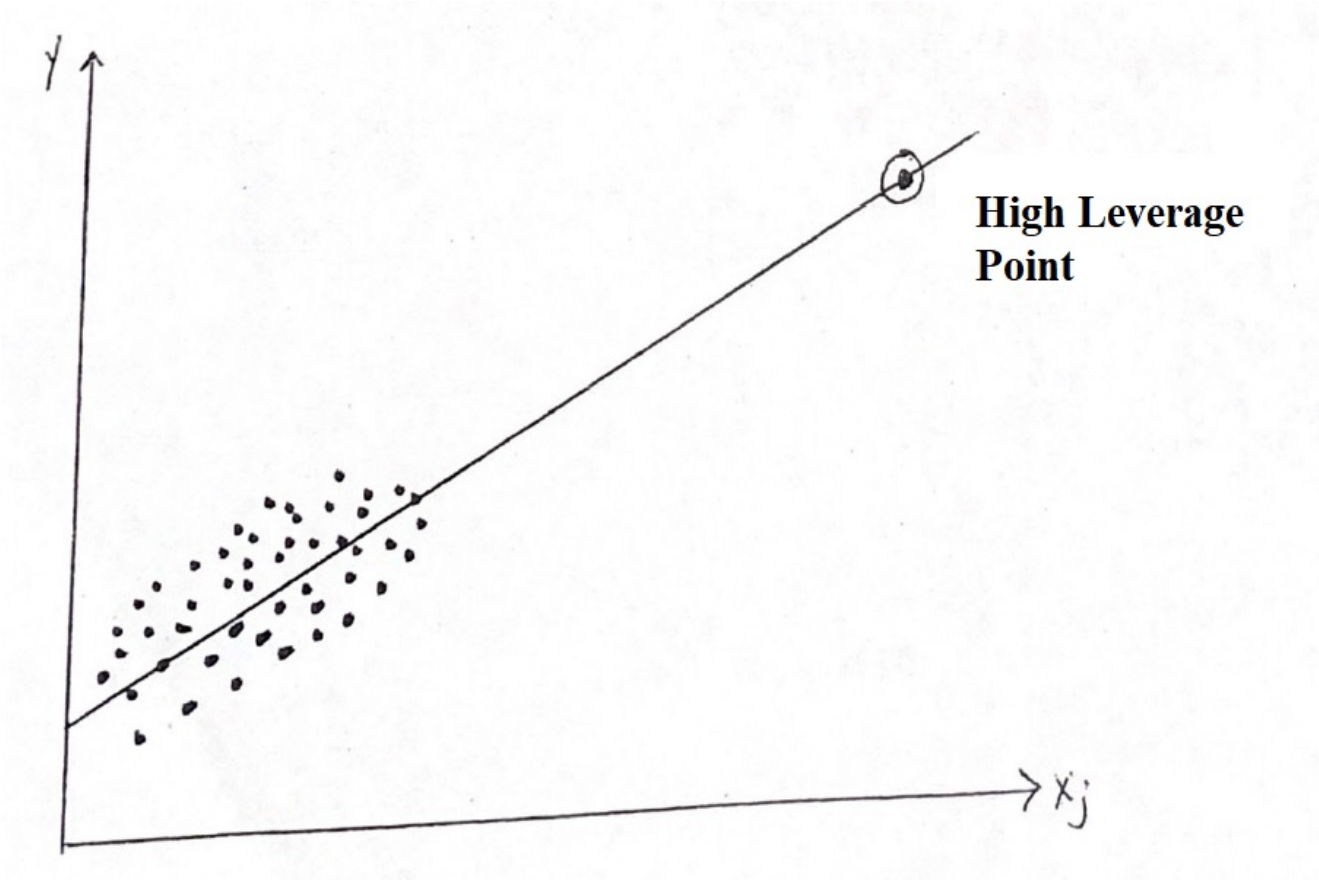
- A high influence point is one that exhibits high leverage and can be consider an outlier.

- Rule of thumb for R-student is to test if $|t_i| > t_{\alpha/2,n-p-1}$ for a single observation at the α level.

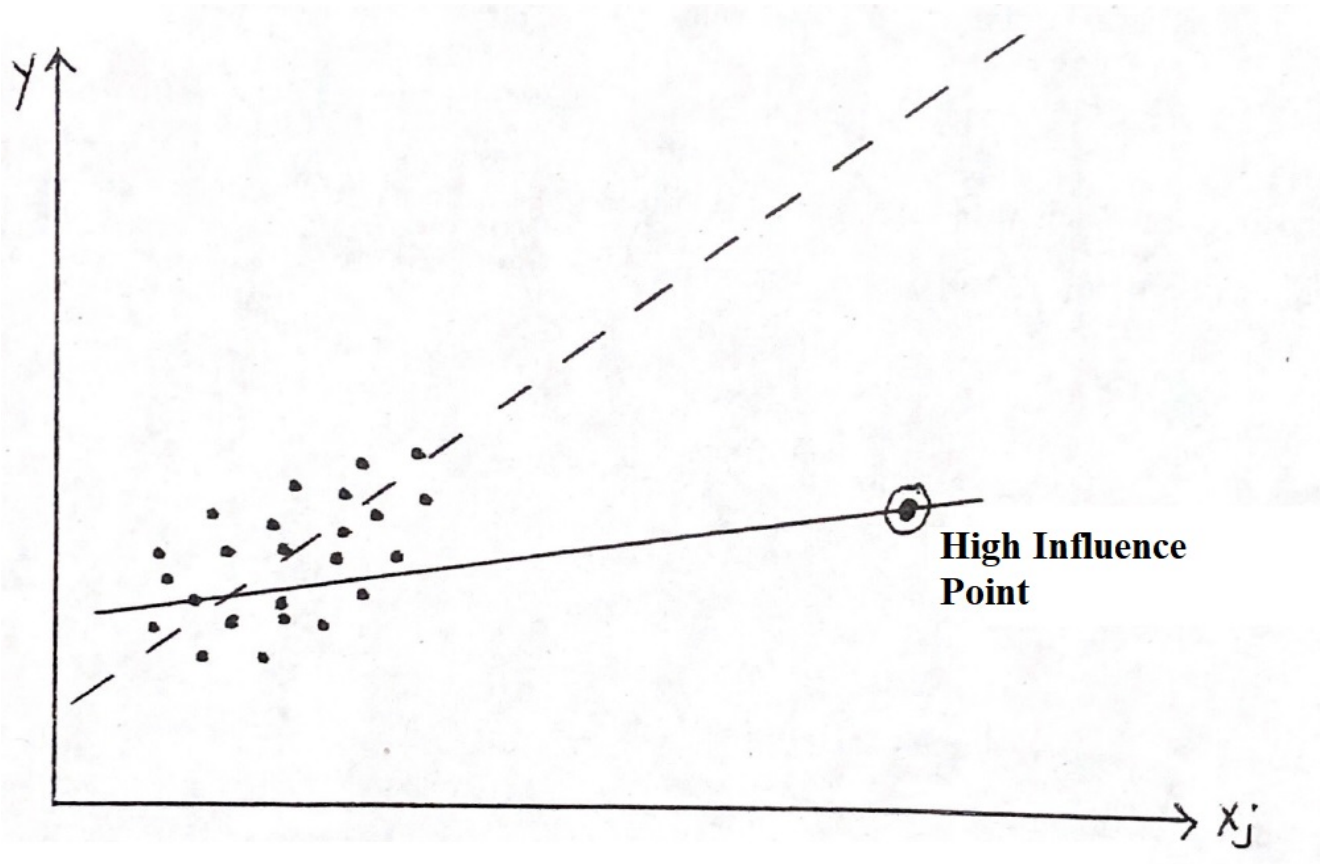
Residual Plot



Residual Plot



Residual Plot



Influential Diagnostics

- Recall that we talked about how to use the HAT diagonals and R-student to identify highly influential observations.
- Now we introduce some other diagnostics measures that are used to determine the extend of the influence, namely:

  DFFITS,

  DFBETAS,

  Cook's D,

  COVRATIO.

Influential Diagnostics (DFFITS)

- Difference-Fit stands for standardized distance between fitted values.
- DFFITS is calculated by

$$(DFFITS)_i = \frac{\hat{y}_i - \hat{y}_{i,-i}}{\hat{\sigma}_{-i}\sqrt{h_{ii}}}.$$

- It can be shown that:

$$(DFFITS)_i = \left(\frac{y_i - \hat{y}_i}{\hat{\sigma}_{-i}\sqrt{1-h_{ii}}}\right)\sqrt{\frac{h_{ii}}{1-h_{ii}}}.$$

- The i-th observation has high influence on the fitted value if $|(DFFITS)_i| \geq 2$.

Influential Diagnostics (DFBETAS)

- DFBETAS are the standardized difference between the regression coefficient estimates when the i-th observation is included and when it is excluded.
- DFBETAS is calculated by

$$(DFBETAS)_{j,i} = \frac{\hat{\beta}_j - \hat{\beta}_{j,-i}}{\hat{\sigma}_{-i}\sqrt{c_{jj}}},$$

  where c_{jj} is the j-th diagonal element of $(\mathbf{X}^T\mathbf{X})^{-1}$.
- Since $var(\hat{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$, thus $(DFBETAS)_{j,i}$ is the number of standard errors that the j-th coefficient changes when i-th observation is excluded.
- The i-th observation has high influence on the regression coefficient estimates if $|(DFBETAS)_{j,i}| \geq 2$.

Influential Diagnostics (DFBETAS)

- Let $\mathbf{R} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ denote a $p \times n$ matrix and $\mathbf{r}_j^T$ be the j-th row of $\mathbf{R}$, for $j = 1, 2, \ldots, p$.
- Let $r_{j,i}$ be the (j, i)-th element of $\mathbf{R}$.
- We have

$$(DFBETAS)_{j,i} = \frac{r_{j,i}}{\sqrt{\mathbf{r}_j^T\mathbf{r}_j}}\left(\frac{y_i - \hat{y}_i}{\hat{\sigma}_{-i}\sqrt{1-h_{ii}}}\right).$$

Influential Diagnostics (Cook's D)

- Cook's Distance is a composite of the DFBETAS and is more abstract since there is not a standardized scaling.
- Cook's D is calculated by

$$(Cook's D)_i = \frac{(\hat{\beta} - \hat{\beta}_{-i})^T (\mathbf{X}^T \mathbf{X})(\hat{\beta} - \hat{\beta}_{-i})}{p\sigma^2}.$$

- In application, one usually looks at Cook's D to determine which observations may be affecting the regression coefficient estimation. Then use the DFBETAS to determine exactly which coefficients those observations are affecting.
- A computational version of the Cook's D is

$$(Cook's D)_i = \left(\frac{r_i^2}{p}\right)\left(\frac{h_{ii}}{1 - h_{ii}}\right)$$

where r_i is the standardized residual.

Influential Diagnostics (COVRATIO)

- DFFITS, DFBETAS, and Cook's D are diagnostic tools for identifying the high influential observations.
- These statistics indicate what component of the regression is being influenced.
- They do not determine the extent of the influence.
- One single value that can give an insight about the regression performance is the generalized variance of the coefficients which is

$$det[var(\hat{\beta})] = det[\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}].$$

Influential Diagnostics (COVRATIO)

- Another statistic that can be used is COVRATIO:

$$(COVRATIO)_i = \frac{det[\sigma^2_{-i}(\mathbf{X}_{-i}^T\mathbf{X}_{-i})^{-1}]}{det[\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}]}.$$

- If we have $n > p$, then we have the following:
  1. If $(COVRATIO)_i > 1 + \frac{3p}{n}$, then the i-th observation have a positive impact on regression.
  2. If $(COVRATIO)_i > 1 - \frac{3p}{n}$, then the i-th observation have a negative impact on regression.
- COVRATIO is always positive.

Influential Diagnostics (Summary)

- Are there any observations having a large impact on the regression?

  R-Student and HAT diagonals.

- If yes, what part of regression (if any) does the observation affect?

  DFFITS, DFBETAS, Cook's D.

- Is the influence positive, negative, or neutral?

  COVRATIO.

Multiple Observation Influential Diagnostics

- DFFITS, DFBETAS, Cook's D, and COVRATIO are all single observation diagnostics.
- Suppose the outlier, leverage point and/or high influential points occure in more than one point.
- We need some multiple observation analogs to DFFITS, DFBETAS, Cook's D, and COVRATIO.
- Suppose we partition $\mathbf{y}$ and $\mathbf{X}$ as

$$\mathbf{y} = \begin{bmatrix} \mathbf{y_I} \\ \mathbf{y_{-I}} \end{bmatrix} \quad \text{and} \quad \mathbf{X} = \begin{bmatrix} \mathbf{X_I} \\ \mathbf{X_{-I}} \end{bmatrix}$$

- Then we can use extended Sherman-Morrison-Woodbury Theorem to come up with such diognostics.

Multiple Observation Influential Diagnostics

- Extended Sherman-Morrison-Woodbury Theorem:

$$(\mathbf{X}^T\mathbf{X}-\mathbf{X}_\mathbf{I}^T\mathbf{X}_\mathbf{I})^{-1} = (\mathbf{X}^T\mathbf{X})^{-1}+(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_\mathbf{I}^T[\mathbf{I}_m-\mathbf{X}_\mathbf{I}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_\mathbf{I}^T]^{-1}\mathbf{X}_\mathbf{I}(\mathbf{X}^T\mathbf{X})^{-1}$$

- Now the $\mathbf{H}_\mathbf{I} = \mathbf{X}_\mathbf{I}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}_\mathbf{I}^T$ is a leverage matrix.
- This can be used to develope multiple observations diagnostics.
- Disadvantage: It is hard to compe up with a rule on these diagnostics.

Multiple Observation Influential Diagnostics

- Consider following linear model

$$\mathbf{y} = \mathbf{X}\beta + \begin{bmatrix} \epsilon_{\mathbf{I}} \\ \epsilon_{-\mathbf{I}} \end{bmatrix}$$

  where $E(\epsilon) = \begin{bmatrix} \Delta \\ 0 \end{bmatrix}$.

- All other regression assumption hold.
- The vector of residuals of the questionable observation is

$$\mathbf{e}_{\mathbf{I}} = \mathbf{y}_{\mathbf{I}} - \mathbf{X}_{\mathbf{I}}\hat{\beta}$$

- Under the null hypothesis $\Delta = 0$ and $\mathbf{e}_{\mathbf{I}} \sim N[0, \sigma^2(\mathbf{I} - \mathbf{H}_{\mathbf{I}})]$.

Multiple Observation Influential Diagnostics

- Thus, under the null hypothesis

$$\frac{\mathbf{e}_{\mathbf{I}}^T(\mathbf{I}-\mathbf{H}_{\mathbf{I}})^{-1}\mathbf{e}_{\mathbf{I}}}{m\hat{\sigma}^2_{-\mathbf{I}}} \sim F_{m,n-p-m}$$

where m is the number of the questionable points and

$$\hat{\sigma}^2_{-\mathbf{I}} = \frac{SSE - \mathbf{e}_{\mathbf{I}}^T(\mathbf{I}-\mathbf{H}_{\mathbf{I}})^{-1}\mathbf{e}_{\mathbf{I}}}{n-p-m}.$$

Multiple Observation Influential Diagnostics

- Now we use an indicator variable to examine the model.

$$\mathbf{y} = \begin{bmatrix} \mathbf{y_I} \\ \mathbf{y_{-I}} \end{bmatrix} = \mathbf{X}\beta + \mathbf{Z}\Delta + \epsilon.$$

where

$$\Delta \begin{bmatrix} \delta_1 \\ \vdots \\ \delta_m \end{bmatrix}_{m\times 1}, \quad \mathbf{Z} = [Z_1, \ldots, Z_m] = \left[\begin{array}{cccc} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & & & \\ 0 & 0 & \ldots & 1 \\ \hline & & \mathbf{0} & \end{array}\right]$$

- Note that this is a regular full rank regression.

Multiple Observation Influential Diagnostics

- Now we can test $\Delta = 0$ using

$$\begin{aligned} F &= \frac{SSE_{Reduced} - SSE_{Full}/m}{\hat{\sigma}^2_{Full}} \\ &= \frac{SSE_{Reduced} - SSE_{Full}}{m\hat{\sigma}^2_{-\mathbf{I}}} \\ &= \frac{\mathbf{e}_{\mathbf{I}}^T(\mathbf{I} - \mathbf{H}_{\mathbf{I}})^{-1}\mathbf{e}_{\mathbf{I}}}{m\hat{\sigma}^2_{-\mathbf{I}}} \sim F_{m,n-p-m} \end{aligned}$$

Polynomial Regression Models

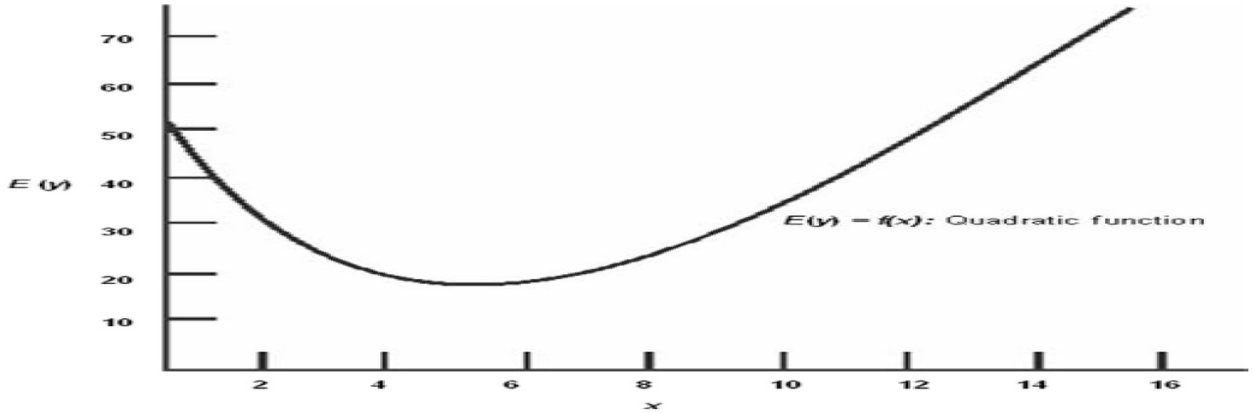
- The $k-th$ order polynomial model in one variable is given by

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x^k + \epsilon$$

- For example a polynomial regression model with one variable which is called as second order model or quadratic model, is:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

where β_1 denotes the linear effect parameter and β_2 is the quadratic effect parameter.

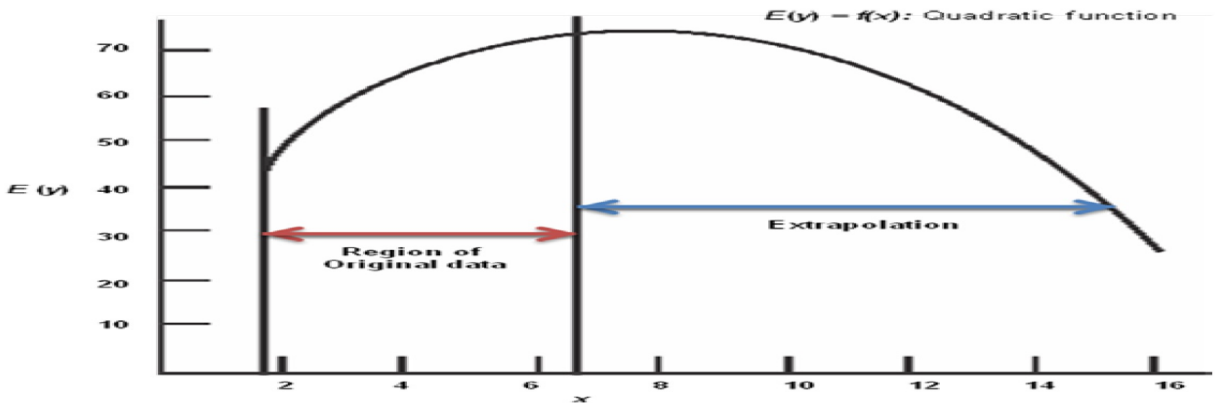


Polynomial Regression Models

- The order of the polynomial model is kept as low as possible.
- Some transformations can be used to keep the model to be of first order.
- Arbitrary fitting of higher order polynomials can be a serious abuse of regression analysis.
- It is always possible for a polynomial of order (n-1) to pass through n points so that a polynomial of sufficiently high degree can always be found that provides a “good” fit to the data.
- Such models neither enhance the understanding of the unknown function nor be a good predictor.
- One has to be very cautions in extrapolation with polynomial models.

Polynomial Regression Models

- The curvatures in the region of data and region of extrapolation can be different.



- In polynomial regression models, as the order increases, the $\mathbf{X}^T\mathbf{X}$ matrix becomes ill-conditioned.
- As a result, the $(\mathbf{X}^T\mathbf{X})^{-1}$ may not be accurate and parameters will be estimated with considerable error.
- If values of x lie in a narrow range then the degree of ill-conditioning increases and multicollinearity in the columns of $\mathbf{X}$ matrix enters.

Polynomial Regression Models (Analysis)

- Consider the polynomial model of order k is one variable as

$$y_i = \beta_0 + \beta_1 x_i + \ldots + \beta_k x_i^k + \epsilon_i$$

- Consider the fitting of following model:

$$y_i = \alpha_0 P_0(x_i) + \alpha_1 P_1(x_i) + \ldots + \alpha_k P_k(x_i) + \epsilon_i$$

where $P_u(x_i)$ denotes the $u - th$ order orthogonal polynomial defined as

$$\begin{aligned} \sum_{i=1}^{n} P_r(x_i) P_s(x_i) &= 0, \quad r \neq s = 0, \ldots, k \\ P_0(x_i) &= 1. \end{aligned}$$

Polynomial Regression Models

- The $\mathbf{X}$ in this case is

$$\mathbf{X} = \begin{bmatrix} P_0(x_1) & P_1(x_1) & \dots & P_k(x_1) \\ \vdots & \vdots & \dots & \vdots \\ P_0(x_n) & P_1(x_n) & \dots & P_k(x_n) \end{bmatrix}$$

- Thus, $\mathbf{X}^T\mathbf{X}$ matrix becomes

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} \sum_{i=1}^n P_0^2(x_i) & 0 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & \sum_{i=1}^n P_k^2(x_i) \end{bmatrix}$$

- The OLS is $\hat{\alpha} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ with for each α_j

$$\hat{\alpha}_j = \frac{\sum_{i=1}^n P_j(x_i)\mathbf{y}}{\sum_{i=1}^n P_j^2(x_i)}$$

and $var(\hat{\alpha}_j) = \frac{\sigma^2}{\sum_{i=1}^n P_j^2(x_i)}$

Polynomial Regression Models

- When x_i are equally spaced, the tables of orthogonal polynomials are available and the orthogonal polynomials can be easily constructed.

$$P_0(x_i) = 1$$

$$P_1(x_i) = \lambda_1 \left[\frac{x_i - \overline{x}}{d} \right]$$

$$P_2(x_i) = \lambda_2 \left[\left(\frac{(x_i - \overline{x})}{d} \right)^2 - \left(\frac{n^2 - 1}{12} \right) \right]$$

$$P_3(x_i) = \lambda_3 \left[\left(\frac{(x_i - \overline{x})}{d} \right)^3 - \left(\frac{x_i - \overline{x}}{d} \right) \left(\frac{3n^2 - 7}{20} \right) \right]$$

$$P_4(x_i) = \lambda_4 \left[\left(\frac{(x_i - \overline{x})}{d} \right)^4 - \left(\frac{x_i - \overline{x}}{d} \right)^2 \left(\frac{3n^2 - 13}{14} \right) + \frac{3(n^2 - 1)(n^2 - 9)}{560} \right]$$

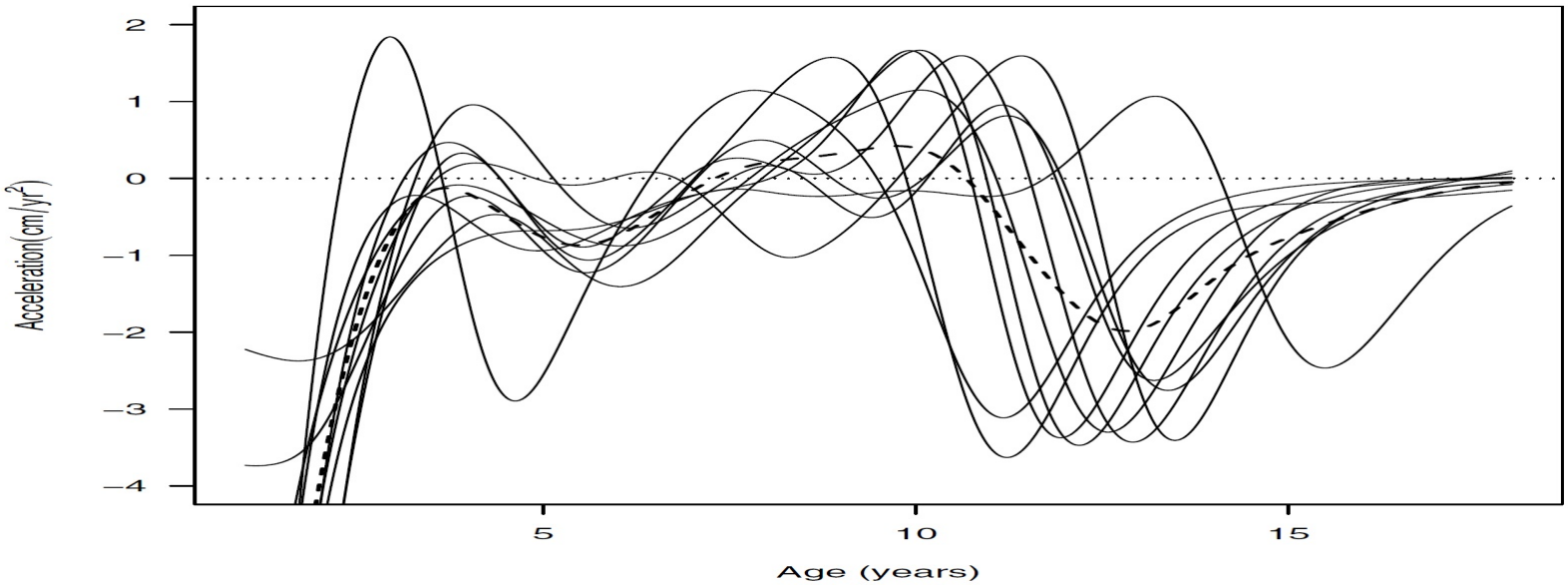
$$P_5(x_i) = \lambda_5 \left[\left(\frac{(x_i - \overline{x})}{d} \right)^5 - \frac{5}{18}\left(n^2 - 7\right) \left(\frac{x_i - \overline{x}}{d} \right)^3 + \frac{1}{1008}\left(15n^4 - 230n^2 + 407\right) \left(\frac{x_i - \overline{x}}{d} \right) \right]$$

$$P_6(x_i) = \lambda_6 \left[\left(\frac{(x_i - \overline{x})}{d} \right)^6 - \frac{5}{44}\left(3n^2 - 31\right) \left(\frac{x_i - \overline{x}}{d} \right)^4 + \frac{1}{176}\left(5n^4 - 110n^2 + 329\right) \left(\frac{x_i - \overline{x}}{d} \right)^2 - \frac{5}{14784}(n^2 - 1)(n^2 - 9)(n^2 - 25) \right]$$

Polynomial Regression Models

- The orthogonal polynomials can also be constructed when x's are not equally spaced.
- Piecewise polynomial is called Splines.

Image is from Functional Data Analysis with R and MATLAB (Ramsay et al.; 2009)



- This is also termed as response surface.
- The methodology of response surface methodology is used to fit such models and helps in designing an experiment.