ST5202: Applied Regression Analysis

Department of Statistics and Applied Probability
National University of Singapore

16-April-2018
Lecture 11

Announcement

- Assignment #5 due today by 9:00 pm
- Today is the LAST lecture
- Final exam scheduled on 5 May between 13:00-15:00.
  - Venue: S16-04-30/41
  - A non-programmable calculator is allowed.
  - One sheet of A4-sized help sheet is allows. You can write or print anything on BOTH sides.
  - Statistics table will be provided (e.g., *t table*, *F table*)

Lecture 11

Review

Estimating the mean response versus prediction

- Mean response at $\mathbf{X}_i$:
  - $E\{y_i\} = \mathbf{X}_i^t\beta$
  - Estimate by $\mathbf{X}_i^t\mathbf{b}$
  - Idea: mean of ALL the y values at $\mathbf{X}_i$
- Prediction
  - Want to predict **a single future value** of y given $\mathbf{X}_i$
  - A future value has a distribution of $\mathbf{X}_i^t\beta + \epsilon \sim N(\mathbf{X}_i^t\beta, \sigma^2)$

Multiple Linear Regression

- $\mathbf{y} \in \mathbb{R}^n$ is a response vector
- $y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i$
  where $\epsilon_i \sim N(0, \sigma^2)$ i.i.d.
- $E\{\mathbf{y}\} = N\left(\mathbf{X}\beta, \sigma^2 I_n\right)$ where

$$\mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,p-1} \\ 1 & x_{2,1} & \cdots & x_{2,p-1} \\ \vdots & & \ddots & \\ 1 & x_{n,1} & \cdots & x_{n,p-1} \end{bmatrix}$$

Multiple Linear Regression

- LS estimator $\mathbf{b}$
$$\mathbf{b} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}$$
- fitted values $\hat{\mathbf{y}}$:
$\mathbf{y} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y} = \mathbf{H}\mathbf{y}$
- residuals $\mathbf{e}$:
$\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{y}$

Simultaneous testing

- Bonferroni procedure
  - suppose we have g tests to test in simultaneous manner.
  - given the family confidence level α, set the quantiles of decision theresholds to be $1-\alpha/g$ (e.g., $F(1-\alpha/g)$, $t(1-\alpha/(2g))$, and so on)
  - provides conservative results:
    the family confidence level is AT LEAST $1-\alpha$
- Many other approaches depending on situations
  - Working-Hotelling procedure, Scheffe procedure, etc
  - Given many options, we can choose one with tighter bounds

General linear test approach–extra sum of squares and partial F-tests

- Given a full model
  $y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_{q-1} x_{q-1} + \cdots + \beta_{p-1} x_{p-1} + \epsilon_i$
- We want to test
  $H_0 : \beta_q = \cdots = \beta_{p-1} = 0$ vs. $H_a : \exists \beta_k \neq 0$ $(k = q, \cdots, p-1)$
- We have reduced model which conforms to the null hypothesis:
  $y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_{q-1} x_{q-1} + \epsilon_i$

General linear test approach–extra sum of squares and partial F-tests

- SSE(F): SSE of full model
- SSE(R): SSE of reduced model
- SSE of reduced model $\geq$ SSE of full model

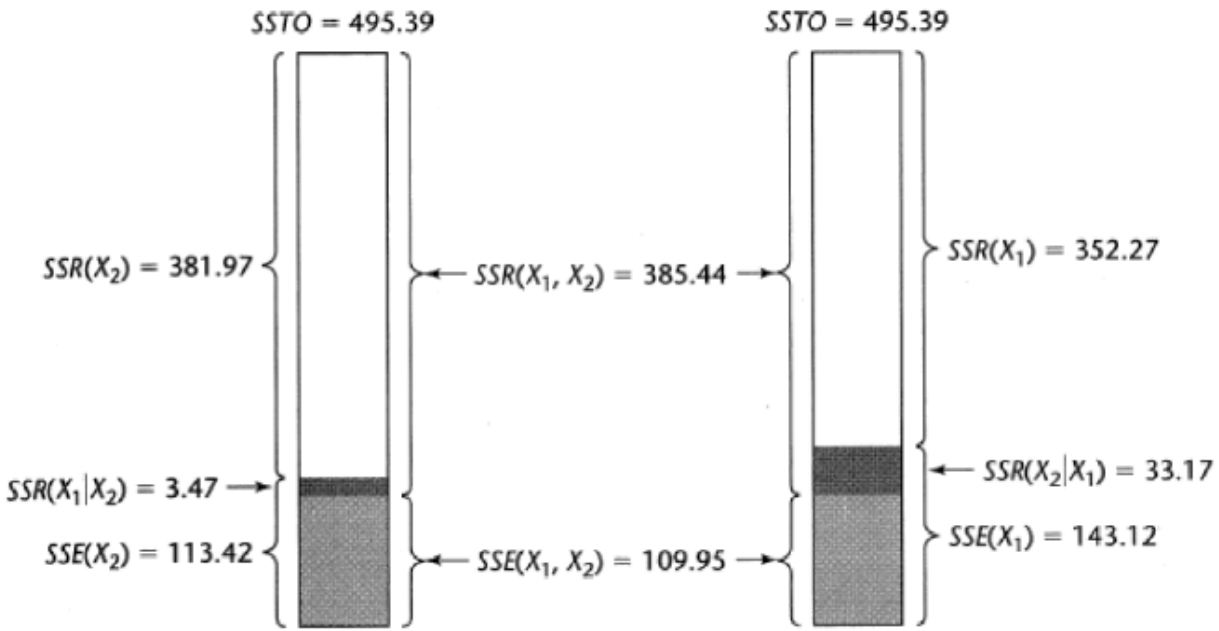
General linear test approach–extra sum of squares and partial F-tests

- We have

$$\begin{aligned} SSTO &= SSR(X_1, \cdots, X_{p-1}) + SSE(X_1, \cdots, X_{p-1}) \\ &= SSR(X_1, \cdots, X_{q-1}) + SSE(X_1, \cdots, X_{q-1}) \end{aligned}$$

  $(p > q)$
- $SSE(X_1, \cdots, X_{p-1}) \leq SSE(X_1, \cdots, X_{q-1})$
  $SSR(X_1, \cdots, X_{p-1}) \geq SSR(X_1, \cdots, X_{q-1})$
- Decompose SSR to measure marginal reduction in error sum of squares when extra variables are added to the model
  e.g., $SSR(X_q, \cdots, X_{p-1} | X_1, \cdots, X_{q-1}) =$
  $SSR(X_1, \cdots, X_{p-1}) - SSR(X_1, \cdots, X_{q-1})$

General linear test approach–extra sum of squares and partial F-tests



General linear test approach–extra sum of squares and partial F-tests

- Idea: if the full model is far better than the reduced model, then $\frac{SSE(R)-SSE(F)}{SSE(F)}$ tends to be large
- Under H_0,
  $F^* = \frac{(SSE(R)-SSE(F))/(df_R-df_F)}{SSE(F)/df_F} = \frac{SSR(X_q,\cdots,X_{p-1}|X_1,\cdots,X_{q-1})/(df_R-df_F)}{SSE(F)/df_F} \sim F(df_R - df_F, df_F)$
  where df_R and df_F are df's of the corresponding SSE's.
- Given level α, we reject H_0 when
  $F^* > F(1-\alpha, df_R - df_F, df_F)$

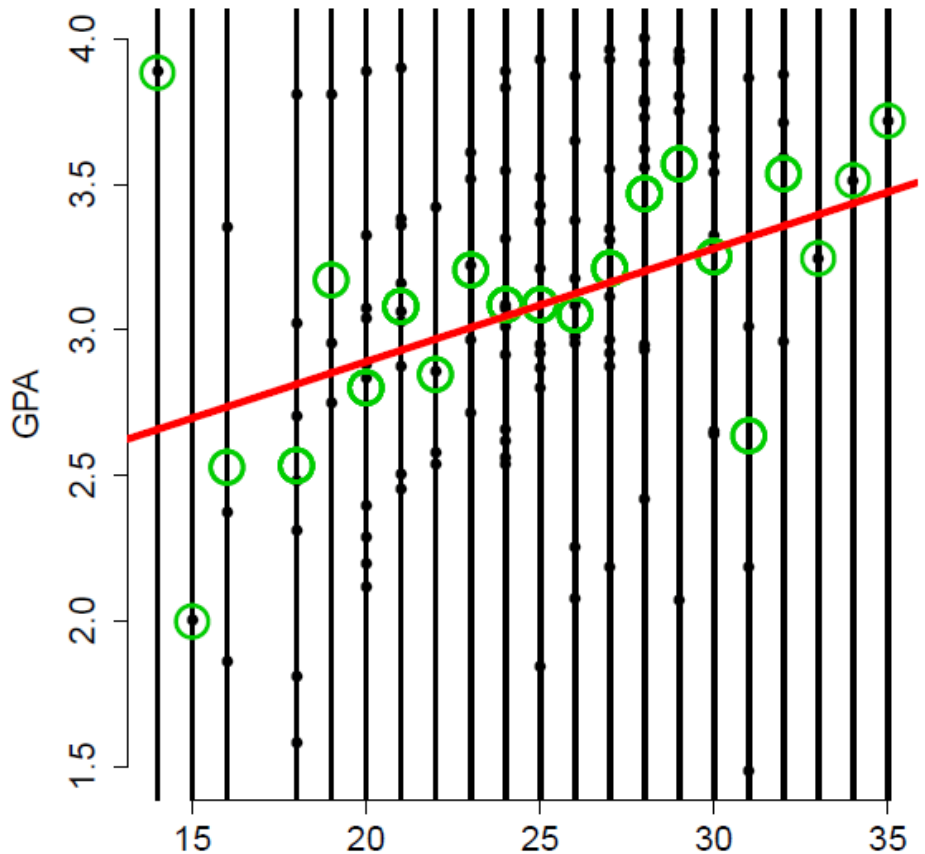
General linear test approach–lack-of-fit test

- Want to test whether the given linear model $E\{y_i\} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1}$ fits the data well
- $H_0 : E\{y_i\} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1}$ versus
  $H_a : E\{y_i\} \neq \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1}$

General linear test approach–lack-of-fit test

- Full model:
  - $y_{ij} = \mu_j + \epsilon_{ij}$ where j denotes the j^{th} level of X.
    μ_j denotes the expectation of y's at j^{th} level of X
  - Estimator of $E\{y_{ij}\} = \mu_j$ is $\bar{Y}_j$
  - $SSE(F) = \sum_j \sum_i \left(Y_{ij} - \bar{Y}_j\right)^2$
- Reduced model:
  - $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i$
  - Estimator of $E\{y_{ij}\} = \mathbf{x}_i^t \mathbf{b}$
  - $SSE(R) = \sum_i \left(y_i - \left(b_0 + b_1 x_{i,1} + \cdots + b_{p-1} x_{i,p-1}\right)\right)^2$
- Test statistic: $F^* = \frac{(SSE(R) - SSE(F))/(df_R - df_F)}{SSE(F)/df_F}$
  $df_R = n - p$, $df_F = n - c$
  where c is the number of distinct levels of X.

General linear test approach–lack-of-fit test (example)

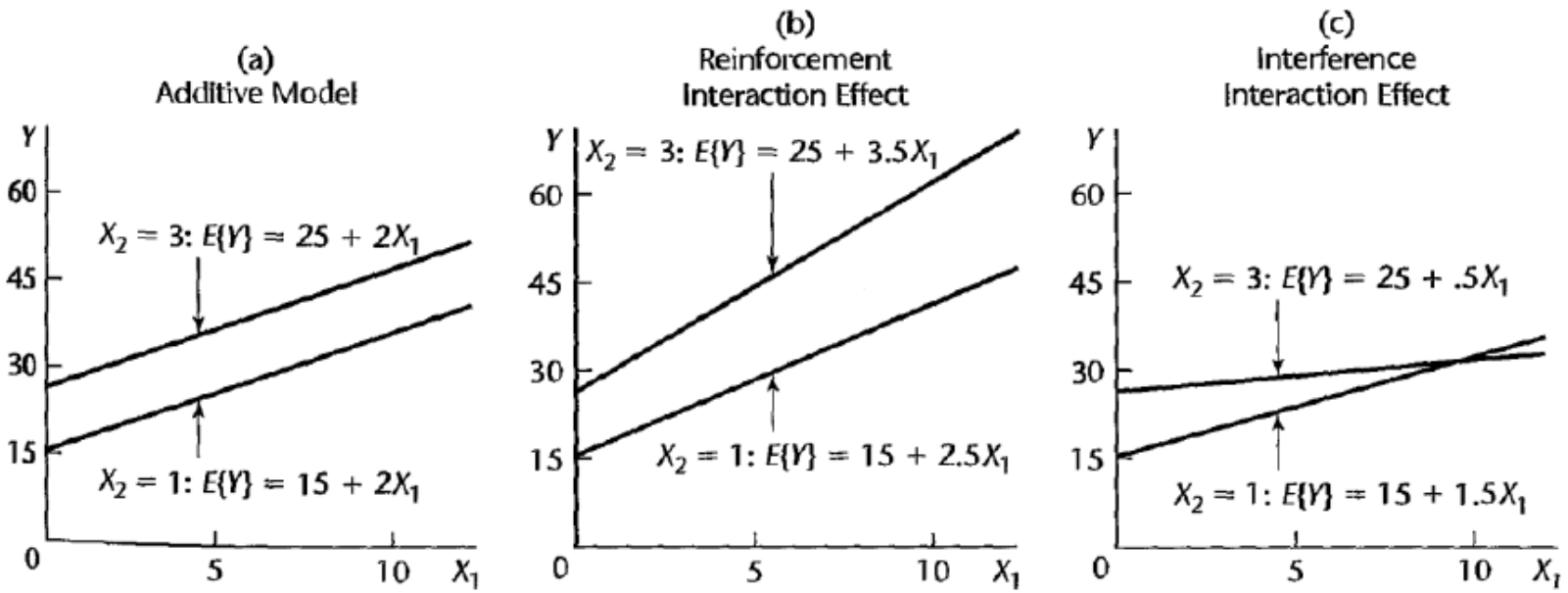


Interaction terms

- Model with an interaction term:
  $E\{y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$
- The association between Y and X_1 depends on the level of X_2 (and vice versa)

$$\begin{aligned} E\{Y|X_2 = x_2\} &= (\beta_0 + \beta_2 x_2) + (\beta_1 + \beta_3 x_2) X_1 \\ E\{Y|X_1 = x_1\} &= (\beta_0 + \beta_1 x_1) + (\beta_2 + \beta_3 x_1) X_2 \end{aligned}$$

Interaction terms



Qualitative variables

- Qualitative variables=categorical variables (e.g., male/female, university A, B, or C)
- How to code?
  - Utilize dummy variables
  - e.g., For university A, B, or C

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

$$X_1 = \begin{cases} 1, & \text{university } A \\ 0, & \text{otherwise} \end{cases}, \quad X_2 = \begin{cases} 1, & \text{university } B \\ 0, & \text{otherwise} \end{cases}$$

  - For university A, $E\{Y\} = \beta_0 + \beta_1$
    For university B, $E\{Y\} = \beta_0 + \beta_2$
    For university C, $E\{Y\} = \beta_0$

Diagnostic and remedial measures OVERVIEW

- Non-linear regression function:
  - Diagnose with residual plots (e_i versus X_i) and $F-$test for lack of fit
  - Solve it with:
    - Transformation of X (not Y, why?)
    - Polynomial regression
    - Non-linear regression
    - etc...
- Non-constancy of error variance
  - Diagnose with residual plots ($|e_i|$ versus X_i) and Brown-Forsythe/Breusch-Pagan test
  - Solve it with:
    - Transformation of Y
    - Weighted least-squares
    - etc

Diagnostic measures

- Diagnostic plots
  - Residuals versus fitted Y's, squared residuals versus fitted Y's, qq-plot of the residuals, time series plot of the residuals
  - Plot residuals versus each predictor variable to examine of linear relation was appropriate.
  - Plot residuals versus product of predictor variables to examine whether exists interaction.

Diagnostic measures

- Diagnostic tests
  - Constant variance:
    - Brown-Forsythe test
      idea: divide residuals into two groups, and examine whether the size of residuals differ between two groups
    - Breush-Pagan test
      idea: examine whether σ_i^2 is related to the level of X in the following way:
      $\log_e \sigma_i^2 = \gamma_0 + \gamma_1 X_i$
  - F-test for lack of fit

Diagnostic measures

- Unusual points
  - Regression outliers
    - Semi-studentized residuals: $\frac{e_i}{\sqrt{MSE}}$
    - Internally studentized residuals: $\frac{e_i}{\sqrt{MSE(1-h_{ii})}}$
    - Externally studentized residuals: $\frac{y_i - \hat{y}_{i(i)}}{s\{y_i - \hat{y}_{i(i)}\}}$
  - Leverage points (outlying X observations)
    - h_{ii} is a measure of the distance between X values for the i^{th} case and the mean of the X values for all n cases.
    - large h_{ii} indicates that i^{th} case is far away from center of all X observations
    - For a new observation, measure for distance to observed cases is $h_{new,new} = \mathbf{X}_{new}^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}_{new}$

Diagnostic measures

- Unusual points
  - Influential data points
    - A case is influential:
      · if it has "a large" influence on the fitted line, on the estimated regression coefficients
      · if excluding it causes "major" changes in the fitted regression function
    - Influence = Leverage × "Outlyingness"
    - DFFITS measures influence on single fitted value $\hat{y}_i$: $\frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)} h_{ii}}}$
    - Cook's distance measures influence on all fitted values: $\frac{\sum_{j=1}^{n} (\hat{Y}_j - \hat{Y}_{j(i)})^2}{pMSE}$
    - DFBETAS measures influence on regression coefficient: $\frac{b_k - b_{k(i)}}{\sqrt{MSE_{(i)} (\mathbf{X^t X})^{-1}_{[kk]}}}$

What to do with unusual data?

- Check for data entry errors
- Think of reasons why the observations are unusual
- Change the model
- Fit the model with and without the observations to see the effect
- Robust regression
- And so on...

Multicollinearity

- What happens?
  - Estimation becomes unstable (small change makes make big difference in results)
  - SE's of the b_k gets large
- Diagnostics
  - VIF: $\frac{1}{1-R_k^2}$ where R_k^2 is the R-squared when regressing X_k on the other X_{-k} predictor variables.
- Remedial measures
  - Correlation transform
  - Ridge regression
  - etc...

Multicollinearity

- Ridge regression
  -

$$\begin{aligned}(\mathbf{r}_{XX} + c\mathbf{I})\mathbf{b}^R &= \mathbf{r}_{XY} \\ \mathbf{b}^R &= (\mathbf{r}_{XX} + cI)^{-1}\mathbf{r}_{XY} \; c > 0\end{aligned}$$

  - Bias/Variance trade off: when $c = 0$, variance is large but unbiased, while when c is large, variance is small but bias is large
  - Equivalent to finding $\hat{\beta}$ that minimizes

$$\sum_{i=1}^{n}(y_i - \mathbf{X}_i^t\beta)^2 + \lambda||\beta||_2^2 \; (\lambda > 0)$$

  - Bias/Variance trade off: when $\lambda = 0$, variance is large but unbiased, while when λ is large, variance is small but bias is large

Variable selection–bias/variance trade off

- $E\{e^2\} = (E\{e\})^2 + Var(e) = Bias(prediction)^2 + Var(prediction)$
- Variable selection is a trade off between the bias and variance:
  - Adding some important predictor variables to a model $\hat{Y} = 0$ may increase a variance by a little bit, but the bias will drop significantly.
  - Deleting some unimportant variables from the full model may increase the bias a bit, but may decrease the variance significantly.
  - Want to find a balance.

Variable selection

- Methods for variable selection
  - Stepwise method: include or exclude a predictor variable based on its p-value at each step
    - Backward elimination:
      At given step, remove a predictor variable with the highest p-value, greater than given level α_{drop}. Continue until all p-values are smaller than α_{drop}
    - Forward selection:
      For all predictors not in the current model, compute their p-values for adding them to the model. Choose the one with the lowest p-value, lower than α_{add}. Continue until no new predictors can be added.
    - Both way
  - All subset methods considering all possible models
  - Lasso regression
  - etc

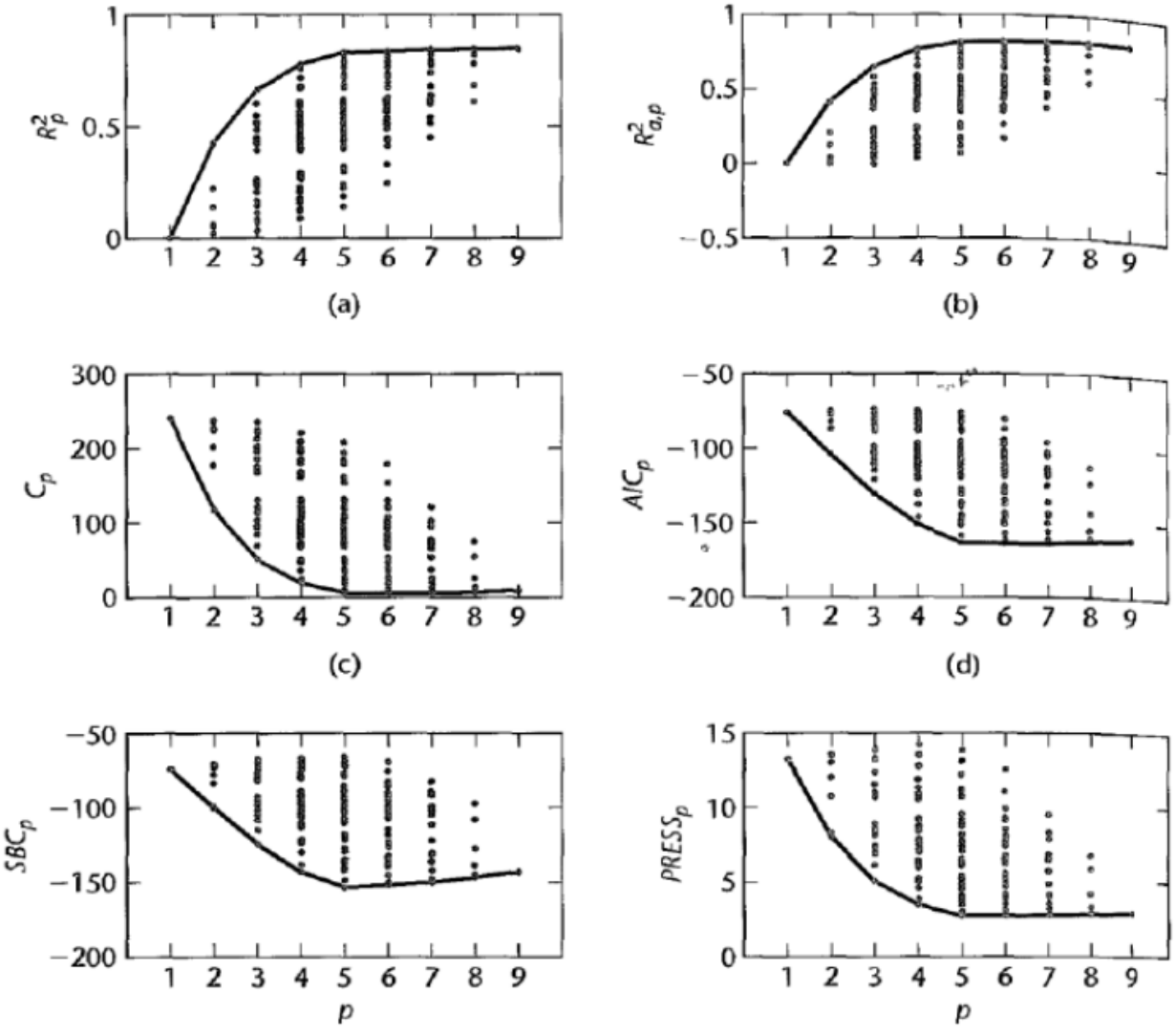
Sample Question

- Methods for variable selection
  - Stepwise method: include or exclude a predictor variable based on its p-value at each step
    - Backward elimination:
      At given step, remove a predictor variable with the highest p-value, greater than given level α_{drop}. Continue until all p-values are smaller than α_{drop}
    - Forward selection:
      For all predictors not in the current model, compute their p-values for adding them to the model. Choose the one with the lowest p-value, lower than α_{add}. Continue until no new predictors can be added.
    - Both way
  - All subset methods considering all possible models
  - Lasso regression
  - etc

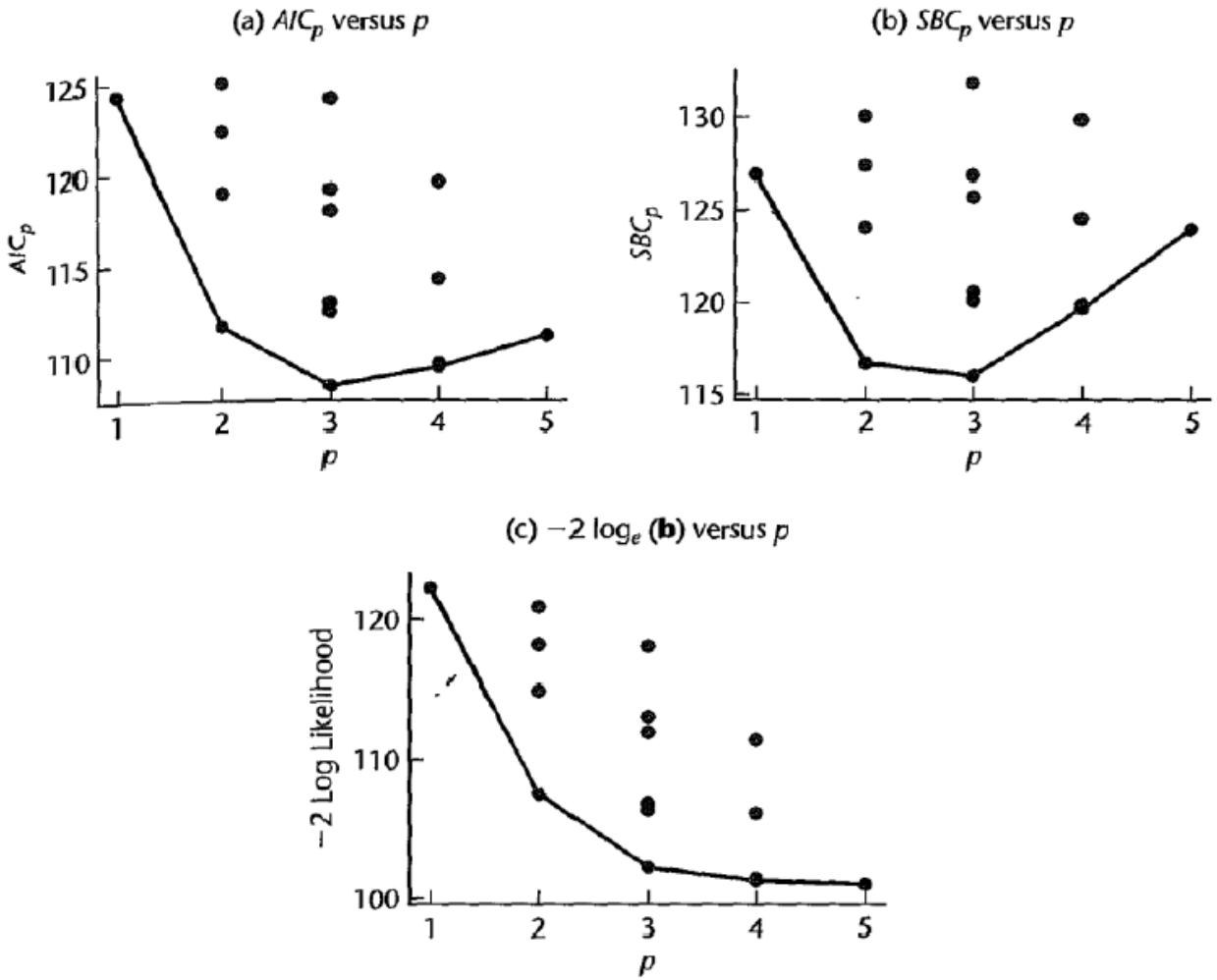
Variable selection criteria

- Adjusted coefficient of multiple determination: $R^2_{a,p} = 1 - \frac{MSE_p}{SSTO/(n-1)}$
- Mallow's C_p: $\frac{SSE_p}{MSE(X_1,\cdots,X_{P-1})} - (n-2p)$
  Estimates $\Gamma_p = \frac{1}{\sigma^2}\sum_i E\{(\hat{Y}_i - \mu_u)^2\}$
- $AIC_p = -2\log l(\mathbf{b}) + 2p$
- $BIC_p = -2\log l(\mathbf{b}) + p\log(n)$
- $PRESS_p = \sum_{i=1}\left(Y_i - \hat{Y}_{i(i)}\right)^2 = \sum_i \left(\frac{e_i}{1-h_{ii}}\right)^2$

Variable selection criteria



Variable selection criteria



Variable selection criteria

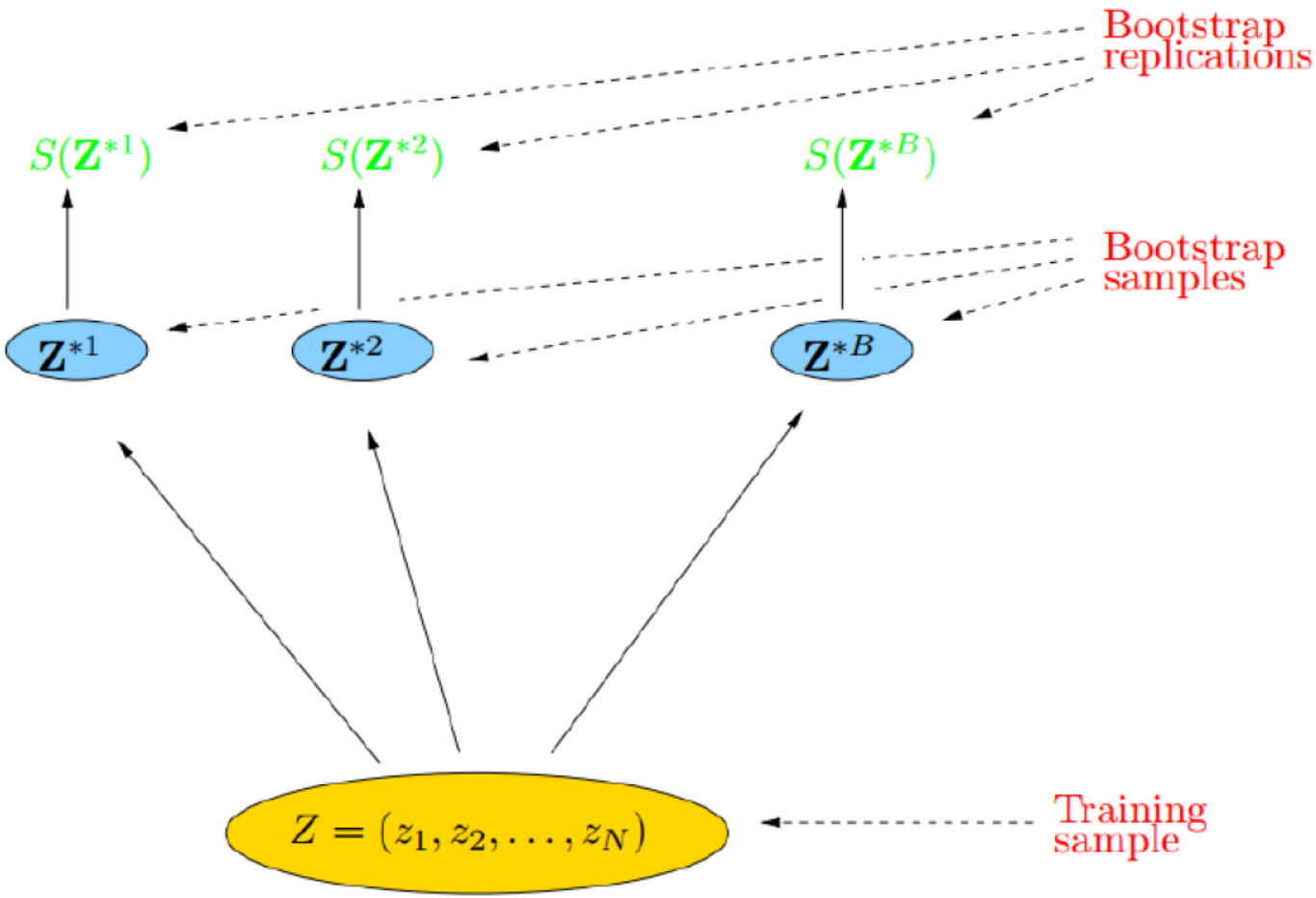
- Cross validation
  - Estimate prediction error Δ
  - Leave-one-out cross validation
    $n\hat{\Delta}_{CV} = CV = \sum_{j=1}^{n}(y_j - \mathbf{X}_j^t\hat{\beta}_{-j})^2$
    where $\hat{\beta}_j$ is the estimator produced when dropping the j^{th} observation.

Non-constant variance: weighted linear regression

- $Y_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i$
  where $\epsilon_i \sim N(0, k \cdot \sigma_i^2)$ with $k > 0$
- $Y_i \sim N(E\{Y_i\}, k/w_i)$
- $Q_w = \sum w_i (Y_i - E\{Y_i\})^2 = (\mathbf{Y} - \mathbf{X}\beta)^t \mathbf{W} (\mathbf{Y} - \mathbf{X}\beta)$
  where $\mathbf{W}$ is a diagonal matrix with $w_i = 1/\sigma_i^2$
- Minimizing Q_w gives:

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

Bootstrapping



Accuracy or prediction?

- The leave-one out bootstrap relaxes the over-fitting problem

$$\widehat{Err}^{(1)} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|C^{-i}|}\sum_{b\in C^{-i}} L\left(y_i, \hat{f}^{*b}(x_i)\right)$$

(C^{-i} is the set of indices of the bootstrap samples b that do not contain observation i, and $|C^{-i}|$ is the number of such samples)

Robust regression

- Robust regression gives automated procedure to reduce the impact of influential cases (but remember that examining outliers is important!)
- Fitted regression line not found by minimizing the sum of the squared errors, but something more robust to outliers, e.g., minimize:
  1. Sum of the absolute errors:
     least absolute residuals (LAR) regression
     $$\sum |Y_i - (\beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1})|$$
  2. The median of squared errors (instead of mean):
     least median of squares (LMS) regression
     $$\mathrm{median_i}\{(\mathrm{Y_i} - (\beta_0 + \beta_1 \mathrm{X_{i1}} + \cdots + \beta_{\mathrm{p-1}} \mathrm{X_{i,p-1}}))^2\}$$
- Or use iteratively reweighted least squares estimation (IRLS)
- No analytical expression for the b_k's

Iteratively reweighted least squares (IRLS)

- Weighted linear regression, with the weights w_i for each case are based on "how far an outlying case is", measured by the scaled residuals $u_i = \frac{e_i}{\hat{\sigma}}$
- An example of a weight function: Huber

$$w(u_i) = \begin{cases} 1 & |u_i| \leq 1.345 \\ 1.345/|u_i| & |u_i| > 1.345 \end{cases}$$

- Continue fitting iteratively until the fit doesn't change much anymore.

Generalized linear model

- $g(E\{\mathbf{Y}\}) = \mathbf{X}\beta$
  - For logistic regression, $E\{Y_i\} = P(Y_i = 1)$ and $g(\cdot)$ is a logit function so that $\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_{p-1} x_{i,p-1}$
- Estimate β by maximizing $l(\beta)$ (usually numerical optimization)
  - For logistic regression, $l(\beta) = \sum_{i=1}^{n} Y_i \cdot \mathbf{X}_i^t \beta + \sum_{i=1}^{n} \log\{1 + \exp(\mathbf{X_i^t}\beta)\}$
- For large sample, $s^2(\mathbf{b}) = -G^{-1}(\mathbf{b})$
- For testing $H_0 : \beta_q = \cdots = \beta_{p-1} = 0$ versus $H_a : H_0$ is not true $G^2 = -2\{l(R) - l(F)\} \approx \chi^2_{p-q}$ under the null
  - $l(R)$ is the maximized log-likelihood under H_0.
  - $l(F)$ is the maximized log-likelihood under the full model.

Sample Question (from Assignment #5)

Cosmetics sales. An assistant in the district sales office of a national cosmetics firm obtained data, shown below, on advertising expenditures and sales last year in the district's 44 territories. X_1 denotes expenditures for point-of-sale displays in beauty salons and department stores (in thousand dollars), and X_2 and X_3 represent the corresponding expenditures for local media advertising and prorated share of national media advertising, respectively. Y denotes sales (in thousand cases). The assistant was instructed to estimate the increase in expected sales when X_1 is increased by 1 thousand dollars and X_2 and X_3 are held constant, and was told to use an ordinary multiple regression model with linear terms for the predictor variables and with independent normal error terms.

i:	1	2	3	...	42	43	44
X_{i1}:	5.6	4.1	3.7	...	3.6	3.9	5.5
X_{i2}:	5.6	4.8	3.5	...	3.7	3.6	5.0
X_{i3}:	3.8	4.8	3.6	...	4.4	2.9	5.5
Y_i:	12.85	11.55	12.78	...	10.47	11.03	12.31

Sample Question

Q. Test whether there is a regression relation between sales and the three predictor variables, using significance level $\alpha = 0.05$. State (1) the null and the alternative hypotheses, (2) the value of test statistic, (3) decision rule, (4) whether the null is rejected. Also, state (5) your conclusion on whether there exists a regression relation between sales and the three predictor variables.

Sample Question

Q. Test whether there is a regression relation between sales and the three predictor variables, using significance level $\alpha = 0.05$. State (1) the null and the alternative hypotheses, (2) the value of test statistic, (3) decision rule, (4) whether the null is rejected. Also, state (5) your conclusion on whether there exists a regression relation between sales and the three predictor variables.

A. (1) $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ versus H_a : at least one of β_1, β_2, and β_3 is non-zero.

Incorrect answer examples

- $H_0 : \beta_0 = \beta_1 = \beta_2 = \beta_3 = 0$ versus H_a : at least one of $\beta_0, \beta_1, \beta_2$, and β_3 is non-zero.
- $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ versus $H_a : \beta_1, \beta_2$, and β_3 are non-zero.
- $H_0 : \beta_0 = \beta_1 = \beta_2 = \beta_3 = 0$ versus H_a : otherwise .

Sample Question

Q. Test whether there is a regression relation between sales and the three predictor variables, using significance level $\alpha = 0.05$. State (1) the null and the alternative hypotheses, (2) the value of test statistic, (3) decision rule, (4) whether the null is rejected. Also, state (5) your conclusion on whether there exists a regression relation between sales and the three predictor variables.

A. (2) $F^* = 38.279$.

Sample Question

Q. Test whether there is a regression relation between sales and the three predictor variables, using significance level $\alpha = 0.05$. State (1) the null and the alternative hypotheses, (2) the value of test statistic, (3) decision rule, (4) whether the null is rejected. Also, state (5) your conclusion on whether there exists a regression relation between sales and the three predictor variables.

A. (3) We reject the null when $F^* > F(1-\alpha, 3, n-4) = 2.839$

Incorrect answer examples

- $F^* > F(1-\alpha, 3, n-4) = 2.839$
  (not stating whether this rejects the null or accepts the null)
- $F(1-\alpha, 3, n-4)$ (not informative enough)
- $F^* > F(1-\alpha, 3, n-4)$
  (not evaluating $F(1-\alpha, 3, n-4)$)
- Since $F^* = 38.279 > F(1-\alpha, 3, n-4) = 2.839$, we reject H_0
  (not stating the decision rule)

Sample Question

Q. Test whether there is a regression relation between sales and the three predictor variables, using significance level $\alpha = 0.05$. State (1) the null and the alternative hypotheses, (2) the value of test statistic, (3) decision rule, (4) whether the null is rejected. Also, state (5) your conclusion on whether there exists a regression relation between sales and the three predictor variables.

A. (4) Since $F^* = 38.279 > F(1-\alpha, 3, n-4) = 2.839$, we reject H_0.

Incorrect answer examples

- Since $F^* = 38.279 > F(1-\alpha, 3, n-4) = 2.839$, we accept H_0.

Sample Question

Q. Test whether there is a regression relation between sales and the three predictor variables, using significance level $\alpha = 0.05$. State (1) the null and the alternative hypotheses, (2) the value of test statistic, (3) decision rule, (4) whether the null is rejected. Also, state (5) your conclusion on whether there exists a regression relation between sales and the three predictor variables.

A. (5) We can conclude that there exists a regression relation between sales and the three predictor variables.

Incorrect answer examples

- We can conclude that a regression relation does not exist between sales and the three predictor variables.
- We can conclude that the model is lack of fit.