Lecture 18

October 22, 2018

Checking Assumptions on $e_1, \ldots, e_n$

Checking Assumptions on $e_1, \ldots, e_n$

- The confidence intervals and hypothesis tests for $\hat{\beta}_j, j = 0, \ldots, p$ in the linear model depend crucially on the assumption that the errors $e_1, \ldots, e_n$ are independent, have equal variance σ^2 and are normally distributed.

Checking Assumptions on $e_1, \ldots, e_n$

- The confidence intervals and hypothesis tests for $\hat{\beta}_j, j = 0, \ldots, p$ in the linear model depend crucially on the assumption that the errors $e_1, \ldots, e_n$ are independent, have equal variance σ^2 and are normally distributed.
- The errors $e_1, \ldots, e_n$ are of course unobservable. How does one then check the assumptions of independence, constant variance and normality of the errors?

Checking Assumptions on $e_1, \ldots, e_n$

- The confidence intervals and hypothesis tests for $\hat{\beta}_j, j = 0, \ldots, p$ in the linear model depend crucially on the assumption that the errors $e_1, \ldots, e_n$ are independent, have equal variance σ^2 and are normally distributed.
- The errors $e_1, \ldots, e_n$ are of course unobservable. How does one then check the assumptions of independence, constant variance and normality of the errors?
- The idea is to use the residuals $\hat{e}_1, \ldots, \hat{e}_n$ which act as proxies for the errors. It is important to note that the residuals are not exactly interchangeable with the errors however.

Checking Assumptions on $e_1, \ldots, e_n$

- The confidence intervals and hypothesis tests for $\hat{\beta}_j, j = 0, \ldots, p$ in the linear model depend crucially on the assumption that the errors $e_1, \ldots, e_n$ are independent, have equal variance σ^2 and are normally distributed.
- The errors $e_1, \ldots, e_n$ are of course unobservable. How does one then check the assumptions of independence, constant variance and normality of the errors?
- The idea is to use the residuals $\hat{e}_1, \ldots, \hat{e}_n$ which act as proxies for the errors. It is important to note that the residuals are not exactly interchangeable with the errors however.
- For example, $var(\hat{e}_i) = \sigma^2(1 - h_{ii})$ where h_{ii} is the ith leverage and $cov(\hat{e}_i, \hat{e}_j) = -\sigma^2 h_{ij}$ where h_{ij} is the (i, j)th entry of the hat matrix.

- Because each h_{ii} lies between 0 and 1 and their average is $(1+p)/n$ which is usually small, in most of the cases, it turns out that each h_{ii} is small.

- Because each h_{ii} lies between 0 and 1 and their average is $(1+p)/n$ which is usually small, in most of the cases, it turns out that each h_{ii} is small.
- Hence each $\hat{e}_i$ has roughly variance equal to σ^2.

- Because each h_{ii} lies between 0 and 1 and their average is $(1+p)/n$ which is usually small, in most of the cases, it turns out that each h_{ii} is small.
- Hence each $\hat{e}_i$ has roughly variance equal to σ^2.
- Similarly, because $\sum_{j=1}^{n} h_{ij}^2 = h_{ii}$ for each i, it also turns out that h_{ij} is typically close to zero for most i and j.

- Because each h_{ii} lies between 0 and 1 and their average is $(1 + p)/n$ which is usually small, in most of the cases, it turns out that each h_{ii} is small.
- Hence each $\hat{e}_i$ has roughly variance equal to σ^2.
- Similarly, because $\sum_{j=1}^{n} h_{ij}^2 = h_{ii}$ for each i, it also turns out that h_{ij} is typically close to zero for most i and j.
- Thus the residuals $\hat{e}_1, \dots, \hat{e}_n$ have variance roughly equal to σ^2 and correlation roughly equal to zero.

- Because each h_{ii} lies between 0 and 1 and their average is $(1+p)/n$ which is usually small, in most of the cases, it turns out that each h_{ii} is small.
- Hence each $\hat{e}_i$ has roughly variance equal to σ^2.
- Similarly, because $\sum_{j=1}^{n} h_{ij}^2 = h_{ii}$ for each i, it also turns out that h_{ij} is typically close to zero for most i and j.
- Thus the residuals $\hat{e}_1, \dots, \hat{e}_n$ have variance roughly equal to σ^2 and correlation roughly equal to zero.
- This is true under the assumption that $e_1, \dots, e_n$ are independent with variance σ^2. The residuals can therefore be used to test assumptions on $e_1, \dots, e_n$.

- Because each h_{ii} lies between 0 and 1 and their average is $(1+p)/n$ which is usually small, in most of the cases, it turns out that each h_{ii} is small.
- Hence each $\hat{e}_i$ has roughly variance equal to σ^2.
- Similarly, because $\sum_{j=1}^{n} h_{ij}^2 = h_{ii}$ for each i, it also turns out that h_{ij} is typically close to zero for most i and j.
- Thus the residuals $\hat{e}_1, \dots, \hat{e}_n$ have variance roughly equal to σ^2 and correlation roughly equal to zero.
- This is true under the assumption that $e_1, \dots, e_n$ are independent with variance σ^2. The residuals can therefore be used to test assumptions on $e_1, \dots, e_n$.
- Alternately, one might use standardized residuals.

Constant Variance Assumption

Constant Variance Assumption

- Plot residuals (y-axis) against the fitted values (x-axis). If all is well, you should see constant variance in the vertical direction and the scatter should be symmetric vertically about zero.

Constant Variance Assumption

- Plot residuals (y-axis) against the fitted values (x-axis). If all is well, you should see constant variance in the vertical direction and the scatter should be symmetric vertically about zero.
- Things to look for are heteroscedasticity (nonconstant variance) and nonlinearity (which indicates that some change in the model is necessary).

Constant Variance Assumption

- Plot residuals (y-axis) against the fitted values (x-axis). If all is well, you should see constant variance in the vertical direction and the scatter should be symmetric vertically about zero.
- Things to look for are heteroscedasticity (nonconstant variance) and nonlinearity (which indicates that some change in the model is necessary).
- Also plot the residuals (y-axis) against each explanatory variable values (for explanatory variables that are both in and out of the model; we will be looking at variable selection methods later).

Constant Variance Assumption

- Plot residuals (y-axis) against the fitted values (x-axis). If all is well, you should see constant variance in the vertical direction and the scatter should be symmetric vertically about zero.
- Things to look for are heteroscedasticity (nonconstant variance) and nonlinearity (which indicates that some change in the model is necessary).
- Also plot the residuals (y-axis) against each explanatory variable values (for explanatory variables that are both in and out of the model; we will be looking at variable selection methods later).
- Look for the same things as the residuals against fitted values plot; except that in the case of plots against explanatory variables that are not in the model, look for any relationship that might indicate that this explanatory variable should be included.

- If indeed there is some evidence of nonconstant variance, two common ways of dealing with it are (a) using weighted least squares and (b) using variable transformations.

- If indeed there is some evidence of nonconstant variance, two common ways of dealing with it are (a) using weighted least squares and (b) using variable transformations.
- We will look at weighted least squares later. The most common variable transformations are taking powers (most common power is square root) and logarithms. Some heuristic for these transformations is given below.

- If indeed there is some evidence of nonconstant variance, two common ways of dealing with it are (a) using weighted least squares and (b) using variable transformations.
- We will look at weighted least squares later. The most common variable transformations are taking powers (most common power is square root) and logarithms. Some heuristic for these transformations is given below.
- Suppose y is a random variable with mean μ. For a function $h(y)$, using a Taylor expansion of order one of y around μ, we get $h(y) \approx h(\mu) + h'(\mu)(y - \mu)$.

- If indeed there is some evidence of nonconstant variance, two common ways of dealing with it are (a) using weighted least squares and (b) using variable transformations.
- We will look at weighted least squares later. The most common variable transformations are taking powers (most common power is square root) and logarithms. Some heuristic for these transformations is given below.
- Suppose y is a random variable with mean μ. For a function $h(y)$, using a Taylor expansion of order one of y around μ, we get $h(y) \approx h(\mu) + h'(\mu)(y - \mu)$.
- From here, we obtain that $var(h(y)) \approx (h'(\mu))^2 var(y)$. Thus if $var(y) \propto \mu^2$, then use $h(y) = \log y$. If $var(y) \propto \mu$, use $h(y) = \sqrt{y}$.

- If indeed there is some evidence of nonconstant variance, two common ways of dealing with it are (a) using weighted least squares and (b) using variable transformations.
- We will look at weighted least squares later. The most common variable transformations are taking powers (most common power is square root) and logarithms. Some heuristic for these transformations is given below.
- Suppose y is a random variable with mean μ. For a function $h(y)$, using a Taylor expansion of order one of y around μ, we get $h(y) \approx h(\mu) + h'(\mu)(y - \mu)$.
- From here, we obtain that $var(h(y)) \approx (h'(\mu))^2 var(y)$. Thus if $var(y) \propto \mu^2$, then use $h(y) = \log y$. If $var(y) \propto \mu$, use $h(y) = \sqrt{y}$.
- Note however that a square root or logarithm can only be taken for nonnegative data.

Normality

Normality

- Normality of the errors is checked by checking normality of the residuals. This is done via a qq plot.

Normality

- Normality of the errors is checked by checking normality of the residuals. This is done via a qq plot.
- In R, The normal qq-plot (*qqnorm*) of the residuals plots the sorted residuals against $z_1, \ldots, z_n$ where

$$z_i = \Phi^{-1}\left(\frac{i-a}{n+1-2a}\right) \quad \text{for } \; i = 1, \ldots, n$$

where $a = 3/8$ if $n \leq 10$ and 0.5 if $n > 10$.

Normality

- Normality of the errors is checked by checking normality of the residuals. This is done via a qq plot.
- In R, The normal qq-plot (*qqnorm*) of the residuals plots the sorted residuals against $z_1, \ldots, z_n$ where

$$z_i = \Phi^{-1}\left(\frac{i-a}{n+1-2a}\right) \quad \text{for } \; i = 1, \ldots, n$$

  where $a = 3/8$ if $n \leq 10$ and 0.5 if $n > 10$.
- The general idea behind the qq plot is the following. The *i*th sorted data point $x_{(i)}$ satisfies the property that the fraction of data points less than or equal to $x_{(i)}$ is i/n (assume that there all observed values are distinct).

Normality

- Normality of the errors is checked by checking normality of the residuals. This is done via a qq plot.
- In R, The normal qq-plot (*qqnorm*) of the residuals plots the sorted residuals against $z_1, \dots, z_n$ where

$$z_i = \Phi^{-1}\left(\frac{i-a}{n+1-2a}\right) \quad \text{for } \ i = 1, \dots, n$$

  where $a = 3/8$ if $n \leq 10$ and 0.5 if $n > 10$.
- The general idea behind the qq plot is the following. The ith sorted data point $x_{(i)}$ satisfies the property that the fraction of data points less than or equal to $x_{(i)}$ is i/n (assume that there all observed values are distinct).
- If the data are normal with mean μ and variance σ^2, then $x_{(i)}$ should be comparable to the point t such that

$$\frac{i}{n} = \mathbb{P}\{N(\mu, \sigma^2) \leq t\} = \mathbb{P}\left\{N(0,1) \leq \frac{t-\mu}{\sigma}\right\}$$

  which means that $t = \mu + \sigma\Phi^{-1}(i/n)$.

- Therefore if the data comes from a normal distribution, we expect a plot of the sorted values $x_{(i)}$ against $\Phi^{-1}(i/n)$ to be linear.

- Therefore if the data comes from a normal distribution, we expect a plot of the sorted values $x_{(i)}$ against $\Phi^{-1}(i/n)$ to be linear.
- *R* essentially does but uses a slightly more involved theoretical quantile z_i which is similar to but not exactly equal to $\Phi^{-1}(i/n)$.

- Therefore if the data comes from a normal distribution, we expect a plot of the sorted values $x_{(i)}$ against $\Phi^{-1}(i/n)$ to be linear.
- *R* essentially does but uses a slightly more involved theoretical quantile z_i which is similar to but not exactly equal to $\Phi^{-1}(i/n)$.
- Given a value of n, the points $z_1, \ldots, z_n$ can be gotten in R by the command *qnorm*(*ppoints*(n)).

- Therefore if the data comes from a normal distribution, we expect a plot of the sorted values $x_{(i)}$ against $\Phi^{-1}(i/n)$ to be linear.
- *R* essentially does but uses a slightly more involved theoretical quantile z_i which is similar to but not exactly equal to $\Phi^{-1}(i/n)$.
- Given a value of n, the points $z_1, \ldots, z_n$ can be gotten in R by the command *qnorm*(*ppoints*(n)).
- When the errors are not normal, least squares estimators may not be optimal (although they are still best linear unbiased estimators). Other robust estimators may be more effective.

- Therefore if the data comes from a normal distribution, we expect a plot of the sorted values $x_{(i)}$ against $\Phi^{-1}(i/n)$ to be linear.
- *R* essentially does but uses a slightly more involved theoretical quantile z_i which is similar to but not exactly equal to $\Phi^{-1}(i/n)$.
- Given a value of n, the points $z_1, \ldots, z_n$ can be gotten in R by the command $qnorm(ppoints(n))$.
- When the errors are not normal, least squares estimators may not be optimal (although they are still best linear unbiased estimators). Other robust estimators may be more effective.
- More importantly, tests and confidence intervals are not exact. However, only long-tailed distributions cause large inaccuracies.

- Therefore if the data comes from a normal distribution, we expect a plot of the sorted values $x_{(i)}$ against $\Phi^{-1}(i/n)$ to be linear.
- *R* essentially does but uses a slightly more involved theoretical quantile z_i which is similar to but not exactly equal to $\Phi^{-1}(i/n)$.
- Given a value of n, the points $z_1, \ldots, z_n$ can be gotten in R by the command *qnorm*(*ppoints*(n)).
- When the errors are not normal, least squares estimators may not be optimal (although they are still best linear unbiased estimators). Other robust estimators may be more effective.
- More importantly, tests and confidence intervals are not exact. However, only long-tailed distributions cause large inaccuracies.
- Mild nonnormality can safely be ignored.

- The resolution of nonnormality depends on the type of problem found. For short-tailed distributions, the consequences of nonnormality are not serious and can reasonably be ignored.

- The resolution of nonnormality depends on the type of problem found. For short-tailed distributions, the consequences of nonnormality are not serious and can reasonably be ignored.
- For skewed errors, a transformation of the response might solve the problem. For long-tailed errors, we might just accept the nonnormality and base the inference on the assumption of another distribution or use resamplign based methods such as permutation tests or bootstrap.

- The resolution of nonnormality depends on the type of problem found. For short-tailed distributions, the consequences of nonnormality are not serious and can reasonably be ignored.
- For skewed errors, a transformation of the response might solve the problem. For long-tailed errors, we might just accept the nonnormality and base the inference on the assumption of another distribution or use resamplign based methods such as permutation tests or bootstrap.
- Alternatively, one may use robust methods which give less weight to outlying observations.

- The resolution of nonnormality depends on the type of problem found. For short-tailed distributions, the consequences of nonnormality are not serious and can reasonably be ignored.
- For skewed errors, a transformation of the response might solve the problem. For long-tailed errors, we might just accept the nonnormality and base the inference on the assumption of another distribution or use resamplign based methods such as permutation tests or bootstrap.
- Alternatively, one may use robust methods which give less weight to outlying observations.
- Also you may find that other diagnostics suggest changes to the model.

- The resolution of nonnormality depends on the type of problem found. For short-tailed distributions, the consequences of nonnormality are not serious and can reasonably be ignored.
- For skewed errors, a transformation of the response might solve the problem. For long-tailed errors, we might just accept the nonnormality and base the inference on the assumption of another distribution or use resamplign based methods such as permutation tests or bootstrap.
- Alternatively, one may use robust methods which give less weight to outlying observations.
- Also you may find that other diagnostics suggest changes to the model.
- In this changed model, the problem of nonnormal errors might not occur.

- The resolution of nonnormality depends on the type of problem found. For short-tailed distributions, the consequences of nonnormality are not serious and can reasonably be ignored.
- For skewed errors, a transformation of the response might solve the problem. For long-tailed errors, we might just accept the nonnormality and base the inference on the assumption of another distribution or use resamplign based methods such as permutation tests or bootstrap.
- Alternatively, one may use robust methods which give less weight to outlying observations.
- Also you may find that other diagnostics suggest changes to the model.
- In this changed model, the problem of nonnormal errors might not occur.
- The Shapiro-Wilk test is a formal test for normality where a small p-value indicates non-normality. We will not go into the details of this test.

Correlated Errors

Correlated Errors

- This is only problematic if there is a natural time (or spatial) structure to the way in which data on the subjects are collected. The simplest way of assessing correlation is to plot the sample autocorrelation function of the residuals (or standardized residuals).

Correlated Errors

- This is only problematic if there is a natural time (or spatial) structure to the way in which data on the subjects are collected. The simplest way of assessing correlation is to plot the sample autocorrelation function of the residuals (or standardized residuals).
- The sample autocorrelation of the residuals $\hat{e}_1, \ldots, \hat{e}_n$ at lag k is defined by

$$\rho_k := \frac{\frac{1}{n-k}\sum_{t=1}^{n-k} \hat{e}_t \hat{e}_{t+k}}{\frac{1}{n}\sum_{t=1}^{n} \hat{e}_t^2}.$$

Correlated Errors

- This is only problematic if there is a natural time (or spatial) structure to the way in which data on the subjects are collected. The simplest way of assessing correlation is to plot the sample autocorrelation function of the residuals (or standardized residuals).
- The sample autocorrelation of the residuals $\hat{e}_1, \ldots, \hat{e}_n$ at lag k is defined by

$$\rho_k := \frac{\frac{1}{n-k}\sum_{t=1}^{n-k} \hat{e}_t \hat{e}_{t+k}}{\frac{1}{n}\sum_{t=1}^{n} \hat{e}_t^2}.$$

- If there is no correlation structure in the residuals, we expect $\rho_1, \rho_2, \ldots$ to behave like independent normal random variables with mean zero and standard deviation $n^{-1/2}$.

Correlated Errors

- This is only problematic if there is a natural time (or spatial) structure to the way in which data on the subjects are collected. The simplest way of assessing correlation is to plot the sample autocorrelation function of the residuals (or standardized residuals).
- The sample autocorrelation of the residuals $\hat{e}_1, \ldots, \hat{e}_n$ at lag k is defined by

$$\rho_k := \frac{\frac{1}{n-k}\sum_{t=1}^{n-k} \hat{e}_t \hat{e}_{t+k}}{\frac{1}{n}\sum_{t=1}^{n} \hat{e}_t^2}.$$

- If there is no correlation structure in the residuals, we expect $\rho_1, \rho_2, \ldots$ to behave like independent normal random variables with mean zero and standard deviation $n^{-1/2}$.
- In R, one can plot the sample autocorrelations by the function *acf*().

- This plot also gives two horizontal blue bars at the levels $\pm 1.96 n^{-1/2}$.

- This plot also gives two horizontal blue bars at the levels $\pm 1.96n^{-1/2}$.
- If about 95% of the sample autocorrelations $\rho_1, \rho_2, \ldots$ (note that ρ_0 is always equal to 1) lie between the horizontal blue bars, then one need not worry about the errors being correlated.

- This plot also gives two horizontal blue bars at the levels $\pm 1.96 n^{-1/2}$.
- If about 95% of the sample autocorrelations $\rho_1, \rho_2, \ldots$ (note that ρ_0 is always equal to 1) lie between the horizontal blue bars, then one need not worry about the errors being correlated.
- There is also a formal test for checking correlation between errors. This is the Durbin-Watson test. We won't go into the details of this test.

Variable Selection

- Consider a regression problem with a response variable y and p explanatory variables $x_1, \ldots, x_p$.

Variable Selection

- Consider a regression problem with a response variable y and p explanatory variables $x_1, \dots, x_p$.
- Should we just go ahead and fit a linear model to y with all the p explanatory variables or should we throw out some unnecessary explanatory variables and then fit a linear model for y based on the remaining variables?

Variable Selection

- Consider a regression problem with a response variable y and p explanatory variables $x_1, \dots, x_p$.
- Should we just go ahead and fit a linear model to y with all the p explanatory variables or should we throw out some unnecessary explanatory variables and then fit a linear model for y based on the remaining variables?
- One often does the latter in practice. The process of selecting important explanatory variables to include in a regression model is called variable selection.

The following are reasons for performing variable selection:

The following are reasons for performing variable selection:

1. Removing unnecessary variables results in a simpler model. Simpler models are always preferred to complicated models.

The following are reasons for performing variable selection:

1. Removing unnecessary variables results in a simpler model. Simpler models are always preferred to complicated models.
2. Unnecessary explanatory variables will add noise to the estimation of quantities that we are interested in.

The following are reasons for performing variable selection:

1. Removing unnecessary variables results in a simpler model. Simpler models are always preferred to complicated models.
2. Unnecessary explanatory variables will add noise to the estimation of quantities that we are interested in. For example, the variance of $\hat{\beta}_0$ in the model $y_i = \beta_0 + e_i$ is σ^2/n while the variance of $\hat{\beta}_0$ in the model $y_i = \beta_0 + \beta_1 x_i + e_i$ is $\sigma^2 \sum_{i=1}^{n} x_i^2/(n \sum_i x_i^2 - n^2 \bar{x}^2)$ where $\bar{x} := \sum_i x_i/n$.

The following are reasons for performing variable selection:

1. Removing unnecessary variables results in a simpler model. Simpler models are always preferred to complicated models.
2. Unnecessary explanatory variables will add noise to the estimation of quantities that we are interested in. For example, the variance of $\hat{\beta}_0$ in the model $y_i = \beta_0 + e_i$ is σ^2/n while the variance of $\hat{\beta}_0$ in the model $y_i = \beta_0 + \beta_1 x_i + e_i$ is $\sigma^2 \sum_{i=1}^n x_i^2/(n\sum_i x_i^2 - n^2\bar{x}^2)$ where $\bar{x} := \sum_i x_i/n$.
3. Collinearity is a problem with having too many variables trying to do the same job.

The following are reasons for performing variable selection:

1. Removing unnecessary variables results in a simpler model. Simpler models are always preferred to complicated models.
2. Unnecessary explanatory variables will add noise to the estimation of quantities that we are interested in. For example, the variance of $\hat{\beta}_0$ in the model $y_i = \beta_0 + e_i$ is σ^2/n while the variance of $\hat{\beta}_0$ in the model $y_i = \beta_0 + \beta_1 x_i + e_i$ is $\sigma^2 \sum_{i=1}^n x_i^2/(n\sum_i x_i^2 - n^2\bar{x}^2)$ where $\bar{x} := \sum_i x_i/n$.
3. Collinearity is a problem with having too many variables trying to do the same job.
4. We can save time and/or money by not measuring redundant explanatory variables.

There are two broad ways of performing variable selection in linear models:

There are two broad ways of performing variable selection in linear models:

1. Stepwise Regression

There are two broad ways of performing variable selection in linear models:

1. Stepwise Regression
2. Criteria-based variable selection

There are two broad ways of performing variable selection in linear models:

1. Stepwise Regression
2. Criteria-based variable selection

Stepwise Regression Methods for Variable Selection

The two main stepwise regression methods are backward elimination and forward selection.

Backward Elimination

Backward Elimination

1. Start with all the explanatory variables in the model.

Backward Elimination

1. Start with all the explanatory variables in the model.
2. Remove the explanatory variable with highest *p*-value larger than a critical value.

Backward Elimination

1. Start with all the explanatory variables in the model.
2. Remove the explanatory variable with highest *p*-value larger than a critical value.
3. Refit the model and go to the previous step.

Backward Elimination

1. Start with all the explanatory variables in the model.
2. Remove the explanatory variable with highest *p*-value larger than a critical value.
3. Refit the model and go to the previous step.
4. Stop when all the *p*-values are less than the critical value.

Backward Elimination

1. Start with all the explanatory variables in the model.
2. Remove the explanatory variable with highest p-value larger than a critical value.
3. Refit the model and go to the previous step.
4. Stop when all the p-values are less than the critical value.

The critical value is sometimes called the p-to-remove and does not have to be 0.05. If prediction performance is the goal, then a 0.15-0.20 cut-off may work best, although methods designed more directly for optimal prediction should be preferred.

Forward Selection

Forward Selection

This just reverses the backward method:

Forward Selection

This just reverses the backward method:

1. Start with no variables in the model.

Forward Selection

This just reverses the backward method:

1. Start with no variables in the model.
2. For all predictors not in the model, check their p-value if they are added to the model. Choose the one with lowest p-value less than the critical value.

Forward Selection

This just reverses the backward method:

1. Start with no variables in the model.
2. For all predictors not in the model, check their p-value if they are added to the model. Choose the one with lowest p-value less than the critical value.
3. Continue until no new predictors can be added.

Other Stepwise Regression Methods

Other Stepwise Regression Methods

- There are several other stepwise regression methods. These are all combinations of backward elimination and forward selection.

Other Stepwise Regression Methods

- There are several other stepwise regression methods. These are all combinations of backward elimination and forward selection.
- These might be better than backward elimination or forward selection by addressing the situation where variables are added or removed early in the process and we want to change our mind about them later.

Other Stepwise Regression Methods

- There are several other stepwise regression methods. These are all combinations of backward elimination and forward selection.
- These might be better than backward elimination or forward selection by addressing the situation where variables are added or removed early in the process and we want to change our mind about them later.
- At each stage a variable may be added or removed and there are several variations on exactly how this is done.

Drawbacks of Stepwise Regression

Drawbacks of Stepwise Regression

Stepwise procedures are relatively cheap computationally but they do have the following drawbacks:

Drawbacks of Stepwise Regression

Stepwise procedures are relatively cheap computationally but they do have the following drawbacks:

1. Because of the one-at-a-time nature of adding/dropping variables, it is possible to miss the optimal model.

Drawbacks of Stepwise Regression

Stepwise procedures are relatively cheap computationally but they do have the following drawbacks:

1. Because of the one-at-a-time nature of adding/dropping variables, it is possible to miss the optimal model.
2. The procedures are not directly linked to final objectives of prediction or explanation and so may not really help solve the problem of interest.

Drawbacks of Stepwise Regression

Stepwise procedures are relatively cheap computationally but they do have the following drawbacks:

1. Because of the one-at-a-time nature of adding/dropping variables, it is possible to miss the optimal model.
2. The procedures are not directly linked to final objectives of prediction or explanation and so may not really help solve the problem of interest.
3. Stepwise variable selection tends to pick models that are smaller than desirable for prediction purposes.