СТАТИСТИКА БОЛЬШИХ ДАННЫХ

Лекция 20. Задача множественной линейной регрессии. Часть 2

Особенности решения и интерпретации

Как отобрать значимые переменные?

Первый шаг в анализе множественной линейной регрессии состоит в вычислении F-статистики и проверке соответствующего p-значения. Если мы заключаем на основе p-значения, что хотя бы один из факторов связан с откликом, естественно поинтересоваться какие это факторы.

Возможно, что все факторы связаны с откликом, но более вероятно, что отклик связан с некоторым подмножеством факторов. Задача определения того, какие именно факторы связаны с откликом в целях подгонки к единственной модели, включающей эти переменные, называется *отбором переменных (variable selection)*.

Ниже мы кратко рассмотрим основные классические подходы к решению задачах отбора переменных.

Особенности решения и интерпретации

Прямой отбор (forward selection). Процедура начинается с **нулевой модели** (*null model*) — модели, которая содержит нулевой член и в которой нет других факторов. Затем перебирается p простых линейных регрессий и к нулевой модели добавляется переменная для модели с минимальным значением *RSS*. Далее к этой модели добавляется новая переменная с минимальным значением *RSS* в двухфакторной модели. Этот подход продолжается пока не будет выполнено некоторое правило останова.

Обратный отбор (backward selection). Процедура начинается с модели со всеми p факторами, и удаляется фактор с наибольшим p-значением (наименее статистически значимая переменная). Проверяется новая модель с p-1 факторами, и снова удаляется фактор с наибольшим p-значением. Эта процедура продолжается пока не сработает правило останова, например, когда все оставшиеся переменные будут иметь p-значения ниже некоторого порога.

Особенности решения и интерпретации

Смешанный отбор. Комбинация прямого и обратного отборов. Процедура начинается с нулевой модели и в модель добавляются новые переменные, обеспечивающие наилучшую подгонку. p-значения переменных могут увеличиваться по мере того, как в модель добавляются новые факторы. Если на некотором шаге p-значение некоторой переменной стало больше некоторого порога, то эта переменная удаляется из модели. Далее продолжают чередоваться прямые и обратные шаги пока все переменные в модели будут иметь достаточно малые p-значения, а все удаленные переменные будут иметь большие p-значения, если их добавить в модель.

Процедура обратного отбора не может быть использована если $p>n$, в то время как прямой отбор всегда возможен. Прямой отбор — "жадный" (greedy) алгоритм и может на ранних этапах включать переменные, которые потом сделаются избыточными. Смешанный отбор может справиться с этой ситуацией.

Особенности решения и интерпретации

Качество подгонки модели

Наиболее общепринятыми мерами качества подгонки модели являются стандартная ошибка остатков *RSE* и корреляционное отношение η^2 (доля объясненной дисперсии). Эти меры вычисляются и интерпретируются также, как для простой линейной регрессии.

В простой линейной регрессии η^2 равно квадрату коэффициента корреляции Пирсона между откликом и фактором. В множественной линейной регрессии оно равно квадрату коэффициента корреляции между откликом и его *МНК*-оценкой $r^2_{Y\hat{Y}}$ — действительно, одно из свойств *МНК*-оценки линейной модели состоит в том, что она максимизирует эту меру корреляции среди всех возможных линейных моделей.

Особенности решения и интерпретации

Значение корреляционного отношения η^2 близкое к 1 означает, что модель объясняет большую часть дисперсии отклика.

Стандартная ошибка остатков имеет вид

$$RSE = \sqrt{\frac{1}{n-p-1}RSS}\,.$$

Прогноз

Модель множественной линейной регрессии можно использовать для прогноза отклика Y на основе множества значений факторов $X_1, X_2, \dots, X_p$. Имеется **три источника неопределенности**, связанных с этим прогнозом

1. Оценки коэффициентов $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$. Плоскость регрессии

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$$

является только оценкой *истинной плоскости регрессии* $f(X)$. Неточность оценок коэффициентов является одним из источников неопределенности прогноза. Можно вычислить *доверительный интервал*, чтобы определить насколько близки $\hat{Y}$ и $f(X)$.

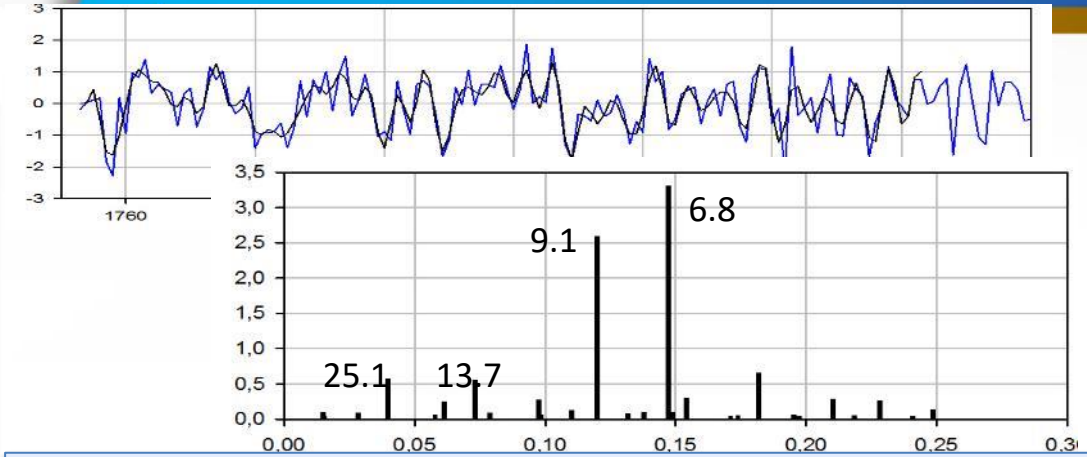
Прогноз

2. На практике линейная модель для $f(X)$ почти всегда некоторая аппроксимация реальности, так что этот дополнительный источник неопределенности прогноза называется *ошибкой модели*

3. Если бы мы знали точно $f(X)$ — т.е., если бы мы знали точно значения коэффициентов $\beta_0, \beta_1, \ldots, \beta_p$ — значение отклика невозможно было бы точно предсказать из-за случайной ошибки e результатов измерений отклика в модели регрессии.

Пример

```
1.  x<-scan("spbost.txt")
2.  n<-130
3.  x<-x[1:n]
4.  h<- 1
5.  t<-1:130
6.  m<-60
7.  C<-matrix(nrow=n-m,ncol=m)
8.  for (j in 1:m)
9.  {for (i in 1:(n-m)) {C[i,j]<-x[m-j+i]}}
10. b<-vector(length=n-m)
11. for (i in 1:(n-m)){b[i]<- -x[m+i]}
12. d1<-data.frame(b,C)
13. M1<-lm(b~C[,1:m]-1,data=d1)
14. summary(M1)
```



$$\sum_{k=1}^{m} a_k x_{i-k} = -x_i, \qquad i = m+1, m+2, \dots, n. \ (n > 2m)$$

$$RSS = \sum_{i=m}^{n} (-x_i - \hat{a}_1 c_{i1} - \hat{a}_2 c_{i2} - \cdots - \hat{a}_m c_{im})^2 .$$

Полиномиальная линейная регрессия

Уравнения полиномиальной регрессии для одного фактора X или модель полиномиальной линейной регрессии:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_p X^p + e$$

Сведение к модели множественной линейной регрессии введением новых факторов:

$$X_j = X^j, \qquad j = 1, \ldots, p.$$

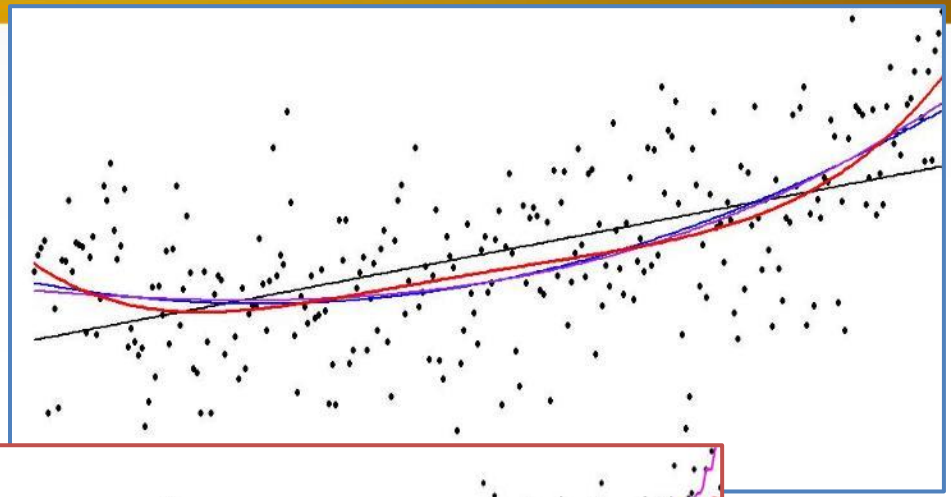
Параметры регрессии оцениваются по методу наименьших квадратов (МНК):

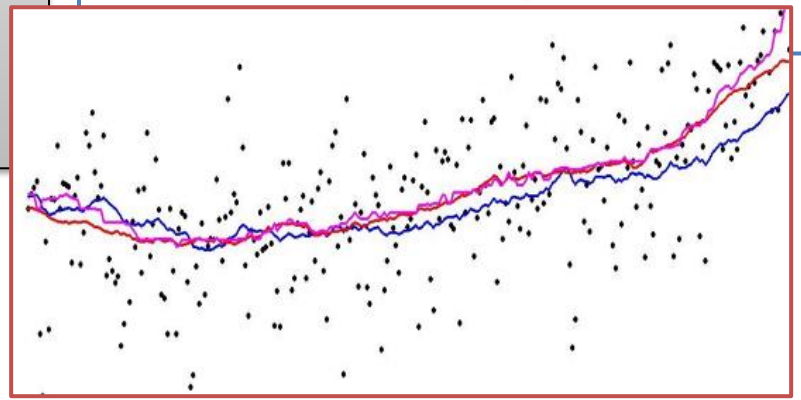
$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i - \hat{\beta}_2 x_i^2 - \cdots - \hat{\beta}_p x_i^p\right)^2 .$$

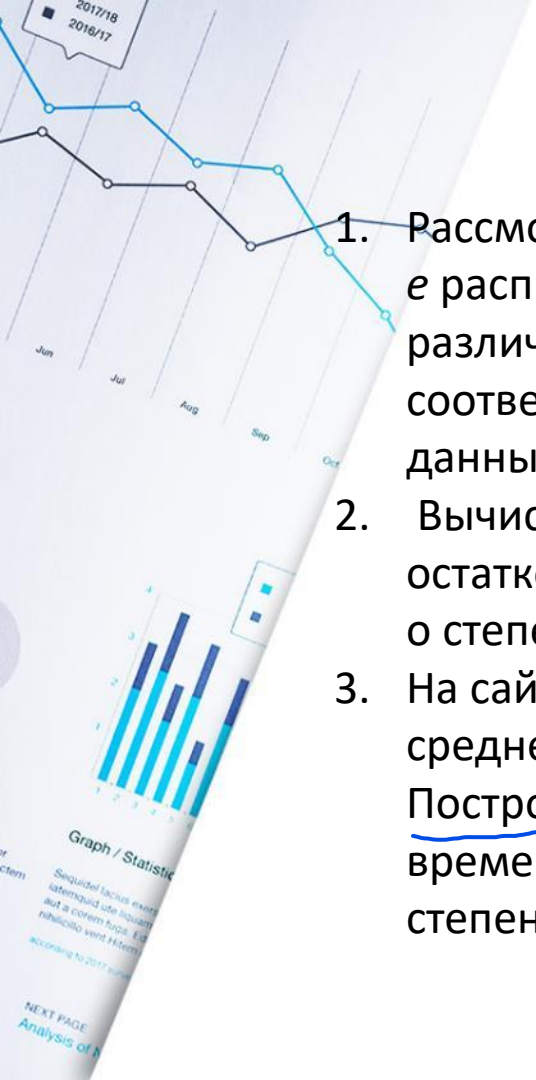
Пример полиномиальной регрессии

```
1.  y<-scan("SPB.txt")
2.  n<-length(y)
3.  x<-c(1753:(1752+n))
4.  data1<-data.frame(x,y)
5.  M1<-lm(y~x,data=data1)
6.  M2<-lm(y~x+I(x^2),data=data1)
7.  M3<-lm(y~x+I(x^2)+I(x^3),data=data1)
8.  M4<-lm(y~x+I(x^2)+I(x^3)+I(x^4),data=data1)
9.  summary(M1)
10. summary(M2)
11. summary(M3)
12. summary(M4)
```

Степень, p	R-squared
1	0.311
2	0.393
3	0.394
4	0.405







Контрольные вопросы и задания

1. Рассмотреть зависимость $y=1+3x_1-2x_2+x_3+e$, где случайная величина e распределена по нормальному закону N(0,1). Сгенерировать 20 различных замеров вектора факторов (x_{1k},x_{2k},x_{3k}) и вычислить соответствующие им значения отклика y_k. По сгенерированным данным построить множественную линейную регрессию.
2. Вычислить сумму квадратов остатков (RSS), стандартную ошибку остатков (RSE) и корреляционное отношение η^2 (R^2). Сделать выводы о степени адекватности линейной модели.
3. На сайте http://www.pogodaiklimat.ru/file.htm найти данные о среднегодовых температурах Москвы за период наблюдений. Построить полиномиальные линейные регрессии температуры от времени. Реализовать процедуру прямого отбора при выборе степени полинома.