Stats 330: Lecture 7

Prediction

5.08.2014
Office hours

▶ Lecturers

<table>
<thead>
<tr>
<th></th>
<th>Office</th>
<th>auckland.ac.nz</th>
<th>day</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steffen Klaere</td>
<td>303.219</td>
<td>s.klaere</td>
<td>Thu</td>
<td>10:00–12:00</td>
</tr>
<tr>
<td>Alan Lee</td>
<td>303S.265</td>
<td>aj.lee</td>
<td>Tue, Thu</td>
<td>10:30–12:00</td>
</tr>
</tbody>
</table>

▶ Tutors (Room 303.326)

<table>
<thead>
<tr>
<th></th>
<th>aucklanduni.ac.nz</th>
<th>day</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Savannah Post</td>
<td>spos008</td>
<td>Mon, Thu</td>
<td>10:00–12:00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thu</td>
<td>14:30–15:30</td>
</tr>
<tr>
<td>Leshun Xu</td>
<td>l xu472</td>
<td>Tue, Wed</td>
<td>13:00–14:00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thu</td>
<td>13:00–14:00</td>
</tr>
<tr>
<td>Hongbin Guo</td>
<td>hguo033</td>
<td>Tue, Wed</td>
<td>11:00–12:00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thu, Fri</td>
<td>10:00–11:00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>14:00–16:00</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>11:00–12:00</td>
</tr>
</tbody>
</table>

Deepayan Sarkar SpringerLink (Online service)

Available. The library has access.

**Title:** Lattice [electronic resource] : multivariate data visualization with R / Deepayan Sarkar.

**Author(s):** Deepayan Sarkar

**Published:** New York ; London : Springer Science+Business Media, c2008.

**Description:** xvii, 285 p. : ill. (some col.) ; 24 cm.

**ISBN:** 9780387759892; 0387759897


**Subjects:** Lattice theory ; R (Computer program language) ; Computer graphics ; Electronic books

**Related Titles:** Series: Use R!

**Notes:** Includes bibliographical references (p. [255]-268) and index.

**Data Source:** Voyager: BID1983312

---

Interlibrary Loan (VDX) request | Voyager | The University of Auckland Library | Contact: digital.services@auckland.ac.nz
Seeing variable names of R objects

```r
> names(cherry.df)
[1] "diameter" "height" "volume"
```

Identifying interesting chunks of data

```r
> which(cherry.df$height>85)
[1] 18 31
> cherry.df[cherry.df$height>85,]
   diameter height volume
 18     13.3  86    27.4
 31     20.6  87    77.0
```
Aims of today’s lecture

- Describe how to use the regression model to predict future values of the response, given the values of the covariates.

- Describe how to estimate the mean response, for particular values of the covariates.
Typical questions

▶ Given the height and diameter, can we predict the volume of a cherry tree? E.g. if a particular tree has diameter 11 inches and height 85 feet, can we predict its volume?

▶ What is the average volume of all cherry trees having a given height and diameter? E.g. if we consider every cherry tree with diameter 11 inches and height 85 feet, can we estimate their mean volume?
\[ Y = \mu + \varepsilon \]
Suppose we want to predict the response of an individual whose covariates are 

\[ x_1, \ldots, x_k \]

E.g., in the cherry tree example, we have \( k = 2 \), and we want to predict the average volume for height \( x_1 = 85 \), and diameter \( x_2 = 11 \).

Since the response we want to predict is \( Y = \mu + \varepsilon \), we obtain \( Y \) by making separate predictions of \( \mu \) and \( \varepsilon \), and adding the results together.
Prediction: how to do it

- Since $\mu$ is the value of the true regression plane at $x_1, \ldots, x_k$, we predict $\mu$ by the value of the fitted plane at $x_1, \ldots, x_k$. This is
  \[
  \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k
  \]

- We have no information about $\varepsilon$ other than the fact that it is sampled from a normal distribution with zero mean. Thus, we predict $\varepsilon = 0$.

- Therefore, to predict the response $y$ for the covariates $x_1, \ldots, x_k$ we use the fitted plane at $x_1, \ldots, x_k$ as the predictor.
The inner product form of a predictor

- Vector of regression coefficients: $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)$;
- Vector of predictor variables: $x = (1, x_1, \ldots, x_k)$;
- Inner product: $x^T \hat{\beta} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$. 
Prediction error

- Prediction error is measured by

\[ s_P^2 = \mathbb{E} (\text{observation} - \text{predictor})^2 = \text{Var}(\text{predictor}) + \sigma^2. \]

- \( \text{Var}(\text{predictor}) \) comes from predicting the mean, \( \sigma^2 \) comes from predicting the error.
- \( \text{Var}(\text{predictor}) \) depends on the covariates and on \( \sigma^2 \).
- **Standard error of prediction** is \( s_P \).
- Prediction interval approximately

  \[ \text{predictor} \pm 2 \times \text{standard error}, \]

  or more precisely...
A prediction interval is an interval which contains the actual value of the response with a given probability, most commonly 0.95.

The $1 - \alpha$ interval for the inner product $\mathbf{x}^T \widehat{\mathbf{\beta}}$ is

$$\mathbf{x}^T \widehat{\mathbf{\beta}} \pm s_p \times t_{n-k-1}(1 - \alpha/2)$$
Variance of the predictor

Arrange covariate data (used for fitting the model) into a matrix $X$ such that

$$X = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{nk} \end{pmatrix}.$$

Then the variance of the prediction for point $\hat{x}$ is

$$\text{Var}(\text{predictor}) = \sigma^2 \hat{x}^T \left(X^T X\right)^{-1} \hat{x}.$$
Suppose we want to predict the volume of 2 cherry trees. The first has diameter 11 inches and height 85 feet, the second has diameter 12 inches and height 90 feet.

First step is to generate a data frame containing the new data. Names must be the same as in the original data.

Then we need to combine the results of the fit (regression coefficients, estimate of error variance) with the new data (the predictor variables) to calculate the predictor.
# Calculate the fit from the original data
> cherry.lm <- lm(volume~diameter+height,data=cherry.df)

# Make a new data frame
> new.df <- data.frame(diameter=c(11,12),height=c(85,90))

# Do the prediction
> predict(cherry.lm,new.df)

# Output
<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>22.63846</td>
<td>29.04288</td>
</tr>
</tbody>
</table>
Doing it in R

```r
> predict(cherry.lm,new.df,se.fit=T,interval="prediction")

$fit
   fit  lwr  upr
1 22.63846 13.94717 31.32976
2 29.04288 19.97235 38.11340

$se.fit
     1     2
1 1.712901 2.130571

$df
[1] 28

$residual.scale
[1] 3.881832
Hand calculation

predictor = 22.63846

\[
\text{SE(predictor)} = \sqrt{\text{se.fit}^2 + \text{residual.scale}^2}
\]

\[
= \sqrt{1.712901^2 + 3.881832^2} = 4.242953
\]

\[
t_{n-k-1}(0.975) = t_{28}(0.975) = 2.048407
\]

Prediction interval = predictor ± SE(predictor) \times t_{n-k-1}(0.975)

\[
= 22.63846 ± 4.242953 \times 2.048407
\]

\[
= [13.94717, 31.32975]
\]
Estimating the mean response

- Suppose we want to estimate the mean response of all individuals whose covariates are $x_1, \ldots, x_k$.

- Since the mean we want to predict is the height of the true regression plane at $x_1, \ldots, x_k$, we use as our estimate the height of the fitted plane at $x_1, \ldots, x_k$. This is

$$\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$$
The standard error of the estimate is the square root of the variance of the predictor.

Note that this is less than the standard error of the prediction!

\[
\begin{align*}
SE(\text{predictor}) &= \sqrt{\text{Var}(\text{predictor}) + \sigma^2}, \\
SE(\text{estimate}) &= \sqrt{\text{Var}(\text{predictor})}.
\end{align*}
\]

Confidence interval

\[
\text{predictor} \pm SE(\text{estimate}) \times t_{n-k-1}(1 - \alpha/2)
\]
Suppose we want to estimate the mean volume of all cherry trees having diameter 11 inches and height 85 feet, or diameter 12 inches and height 90 feet.

As before, the first step is to make a data frame containing the new data. Names must be the same as in the original data.
Doing it in R

```r
> predict(cherry.lm,new.df,se.fit=T,interval="confidence")
$fit
     fit  lwr  upr
 1 22.64 19.13 26.15
 2 29.04 24.68 33.41

$se.fit
     1  2
 1 1.71 2.13
 2

$df
[1] 28

$residual.scale
[1] 3.881832
```
Example: Hydrocarbon data

When petrol is pumped into a tank, hydrocarbon vapours are forced into the atmosphere. To reduce this significant source of air pollution, devices are installed to capture the vapour. A laboratory experiment was conducted in which the amount of vapour given off was measured under carefully controlled conditions. In addition to the response, there were four variables which were thought relevant for prediction:

- **t.temp**: initial tank temperature (degrees F)
- **p.temp**: temperature of dispensed petrol (degrees F)
- **t vp**: initial vapour pressure in tank (psi)
- **p vp**: vapour pressure of dispensed petrol (psi)
- **hc**: emitted dispensed hydrocarbons (g) (response)
```r
$se.fit
1   2
1.71290 2.13057

$df
[1] 28

$residual.scale
[1] 3.881832

> data(vapour.df)
> View(vapour.df)
> pairs(vapour.df,col="steelblue",pch=16)
```
Pairs plot
Pairs plot
Pairs plot

hc

t.temp

0.81

p.temp

0.88

0.81

0.85

0.94

0.77

t.vp

0.85

0.94

0.77

t.vp

0.91

0.93

0.83

0.98

p.vp
All variables seem related to the response

p.vp and t.vp seem highly correlated

Quite strong correlations between some of the other variables

No obvious outliers
Fitting the “full” model

\[
\text{vapour.reg <- lm(hc~t.temp+p.temp+t.vp+p.vp, data=vapour.df)}
\]

\[
> \text{summary(vapour.reg)}
\]

Call:
\[
\text{lm(formula = hc ~ t.temp + p.temp + t.vp + p.vp, data = vapour.df)}
\]

Coefficients:

|            | Estimate | Std. Error | t value | Pr(>|t|) |
|------------|----------|------------|---------|----------|
| (Intercept)| 0.16609  | 1.02198    | 0.163   | 0.87117  |
| t.temp     | -0.07764 | 0.04801    | -1.617  | 0.10850  |
| p.temp     | 0.18317  | 0.04063    | 4.508   | 1.53e-05 *** |
| t.vp       | -4.45230 | 1.56614    | -2.843  | 0.00526  ** |
| p.vp       | 10.27271 | 1.60882    | 6.385   | 3.37e-09 *** |

Residual standard error: 2.723 on 120 degrees of freedom
Multiple R-squared: 0.8959, Adjusted R-squared: 0.8925
F-statistic: 258.2 on 4 and 120 DF, p-value: < 2.2e-16
Conclusions

- Large $R^2$
- Significant parameters except for $t$.$\text{temp}$
- Model seems satisfactory
- Move on to prediction
- Let us predict hydrocarbon emissions when

$$t$.$\text{temp} = 28, \ p$.$\text{temp} = 30, \ t$.$\text{vp} = 3, \ p$.$\text{vp} = 3.5$$
Hydrocarbon emissions are predicted to be between 20 and 32 grams.
BREAKING: TO SURPRISE OF PUNDITS, NUMBERS CONTINUE TO BE BEST SYSTEM FOR DETERMINING WHICH OF TWO THINGS IS LARGER.