STATS 330: Lecture 7
Prediction
5.08.2014

Office hours

Lecturers

<table>
<thead>
<tr>
<th>Name</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>3035.265</td>
<td>aj.lee</td>
<td>Fri,</td>
<td>10:00–12:00</td>
</tr>
</tbody>
</table>

Tutors (Room 303.326)

<table>
<thead>
<tr>
<th>Name</th>
<th>aucklanduni.ac.nz</th>
<th>day</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Savannah Post</td>
<td>spo008</td>
<td>Mon,</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>lxu472</td>
<td>Tue,</td>
<td>10:30–12:00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>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>hgu0033</td>
<td>Tue,</td>
<td>11:00–12:00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Wed,</td>
<td>14:00–16:00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thu,</td>
<td>10:00–11:00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fri,</td>
<td>11:00–12:00</td>
</tr>
</tbody>
</table>

R-hint(s) of the day

Seeing variable names of R objects

> names(cherry.df)

[1] "diameter" "height" "volume"

Identifying interesting chunks of data

> 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?
Prediction: how to do it

- 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 error

- Prediction error is measured by
  \[ s_p^2 = \text{Var} (\text{observation} - \text{predictor})^2 = \text{Var} (\text{predictor}) + \sigma^2. \]
- \( \text{Var} (\text{predictor}) \) 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...

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: \( \mathbf{x} = (1, x_1, \ldots, x_k) \);
- Inner product: \( \mathbf{x}^T \hat{\beta} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k. \)
Variance of the predictor

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

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

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

$$\text{Var}(\text{predictor}) = \sigma^2 \hat{x}^T (X^T X)^{-1} \hat{x}.$$ 

Doing it in R

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.

```r
# 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
1 22.63846 29.04288
```

Hand calculation

$$\text{predictor} = 22.63846$$

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

$$= \sqrt{1.712901^2 + 3.881832^2} = 4.242953$$

$$t_{n-k-1}(0.975) = t_{28}(0.975) = 2.048407$$

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

$$= 22.63846 \pm 4.242953 \times 2.048407$$

$$= [13.94717, 31.32975]$$

Doing it in R

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

$fit

<table>
<thead>
<tr>
<th>fit</th>
<th>lwr</th>
<th>upr</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13.94717</td>
<td>31.32975</td>
</tr>
<tr>
<td>2</td>
<td>19.97235</td>
<td>38.11340</td>
</tr>
</tbody>
</table>

$se.fit

1.712901
2.130571

$df

[1] 28

$residual.scale

[1] 3.881832
```

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

$$\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k.$$
Standard error of the estimate

- 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!
  \[ SE(\text{predictor}) = \sqrt{\text{Var}(\text{predictor}) + \sigma^2}, \]
  \[ SE(\text{estimate}) = \sqrt{\text{Var}(\text{predictor})}. \]
- Confidence interval
  \[ \text{predictor} \pm SE(\text{estimate}) \times t_{n-1}(1 - \alpha/2) \]

Doing it in R

- 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.

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:

- tt.temp: initial tank temperature (degrees F)
- p.temp: temperature of dispensed petrol (degrees F)
- tvp: initial vapour pressure in tank (psi)
- pvp: vapour pressure of dispensed petrol (psi)
- hc: emitted dispensed hydrocarbons (g) (response)
Preliminary conclusions

- 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

```r
> vapour.reg <- lm(hc ~ t.temp + p.temp + t.vp + p.vp, data = vapour.df)
> summary(vapour.reg)
```

Call:  
`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   | 0.10855 |
| p.temp | 0.18317    | 0.04063 | 4.508  | 1.53e-05 *** |
| t.vp | -4.46230  | 1.56614 | -2.843   | 0.00626 ** |
| p.vp | 10.27271   | 1.60882 | 6.385   | 3.37e-09 *** |

Residual standard error: 2.723 on 120 degrees of freedom  
Multiple R-squared: 0.8989, Adjusted R-squared: 0.8925  
F-statistic: 258.2 on 4 and 120 DF,  p-value: < 2.2e-16

Conclusions

- Large R²
- Significant parameters except for t.temp
- Model seems satisfactory
- Move on to prediction
- Let us predict hydrocarbon emissions when  
  t.temp = 28, p.temp = 30, t.vp = 3, p.vp = 3.5

Prediction

```r
> vapour.pred <- predict(vapour.reg, vapour.pred, interval = "p")
> predict(vapour.reg, vapour.pred, interval = "p")
```

1 26.08471 20.20945 31.93898

Hydrocarbon emissions are predicted to be between 20 and 32 grams.
CREATURES TO SURPRISE AT FUNDRAISERS CONTINUE TO BE
BEST SHOT FOR DETERMINING WHICH OF TWO THINGS IS LARGER.

http://xkcd.com/1131/