STAT115: Introduction to Biostatistics

University of Otago

Ōtākou Whakaihu Waka

Lecture 23: Prediction with Linear Regression

Outline

- R^2: the proportion of variance explained
- Another look at estimating the mean response
- Predicting a new observation
- Extrapolation

Recall: possum data

- The size of brushtail possums
  - Exploring relationship between total length (mm) and head length (mm)
- If we have a total length measurement
  - Can we predict the head length?
- Import the data into R

```
possum = read.csv('possum.csv')
```

- Fit a simple linear regression

```
m_possum = lm(head_l ~ total_l, data = possum)
```

Output

```
summary(m_possum)

##
## Call:
## lm(formula = head_l ~ total_l, data = possum)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.188 -1.534 -0.334  1.279  7.397
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 42.70979    5.17281    8.26  5.7e-13 ***
## total_l      0.05729    0.00593    9.66  4.7e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.6 on 102 degrees of freedom
## Multiple R-squared:  0.478, Adjusted R-squared:  0.472
## F-statistic: 93.3 on 1 and 102 DF,  p-value: 4.68e-16
```

R^2: Coefficient of determination

- R^2 is a commonly used measure of how well a regression model describes the data
  - In R summary: `Multiple R-squared` $=$ `0.4776`
- Look at two descriptions of R^2
  - Give us different perspectives on what it represents

R^2: squared correlation

- R^2 is the squared correlation between y and $\hat{y}$

```
y = possum$head_l # y values
yhat = fitted(m_possum) # y-hat values
R = cor(y, yhat)
R^2 # correlation^2

## [1] 0.4776
```

- Since $-1 \leq r \leq 1$ we have $0 < R^2 < 1$
  - The larger the value of R^2, the better the regression model describes the data
    - The fitted values are ‘close’ to the observations

R^2: percentage of variance explained

- The total sum of squares is $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$
  - Measures the variability of the outcome variable
- (Recall) the residual sum of squares $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
  - Measures the variability of the outcome variable after fitting regression model
- The explained sum of squares $ESS = TSS - RSS$
  - Amount of variation in the outcome variable that is explained by the regression model
- R^2 can be expressed as

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

- The proportion of variance explained by the model
  - R^2 is often reported as a percentage: $R^2 = 47.8\%$

Interpreting R^2

- R^2 is often reported when fitting a linear regression
- No absolute rule for what a good (or bad) R^2 value is
  - In one particular area of application: an R^2 of 0.3 might be good
  - In another area of application: an R^2 of 0.8 might be poor

Mean response

- Recall: linear regression model

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- Mean response at a given x value: $\mu_y = \beta_0 + \beta_1 x$
- The fitted model is an estimate of the mean response

$$\hat{\mu}_y = \hat{\beta}_0 + \hat{\beta}_1 x$$

- How precise is this estimate?
- Can we find a confidence interval for μ_y?
  - e.g. what is the confidence interval for mean head length of the subpopulation of possums with total length 850 mm

Confidence interval for mean response

- Goal: find a confidence interval for μ_{y_0}, the mean response when $x = x_0$
- Confidence interval will have the form

$$\text{estimate} \pm \text{multiplier} \times \text{std. error}$$

- Estimate: $\hat{\mu}_{y_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$.
- The (estimated) standard error for $\hat{\mu}_{y_0}$ is

$$s_{\hat{\mu}_{y_0}} = s_\varepsilon \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

- Multiplier: t-distribution with $\nu = n - 2$ degrees of freedom

Confidence interval for mean response

- A $100(1-\alpha)\%$ confidence interval for μ_{y_0} is given by

$$\hat{\mu}_{y_0} \pm t_{(1-\frac{\alpha}{2}, n-2)} \times s_{\hat{\mu}_{y_0}}$$

- This is an interval estimate for the mean response μ_{y_0}
- Finding this confidence interval by hand is tedious
  - Use R to help us
  - `predict` function
- The `predict` function requires a data frame
  - Contains x_0: the predictor variable values where we want to find the mean response

Excursion: data frames in R

- You have been using data frames all semester
- When we import data into R: it is in a data frame
  - Rows: Each row is an observation or data record
  - Columns: Each column is a variable (typically with a name)
- We can construct a data frame using function `data.frame`

```
first_df = data.frame(name = c("Bob","Mary","Lucy"), age = c(19,17,23),
                      height = c(173, 168, 176))
first_df
```

```
##   name age height
## 1  Bob  19    173
## 2 Mary  17    168
## 3 Lucy  23    176
```

Data from for `predict`: possum data

- We need to construct a data frame in R
  - Contain the x (predictor variable) values where we want to find the mean response
  - Same variable name as was used to fit the model in `lm`
- Recall:

```
m_possum = lm(head_l ~ total_l, data = possum)
```

- Predictor variable name: `total_l`
- Let's say we want to estimate the mean response at 850 mm

```
predictor1 = data.frame(total_l = 850)
```

- If we wanted to find the mean response at 850 mm and 900 mm

```
predictor2 = data.frame(total_l = c(850,900))
```

Mean response in R

- Use the predict function, with option interval = "confidence"

```
mean_resp = predict(m_possum, newdata = predictor1, interval = "confidence")
mean_resp
##     fit   lwr   upr
## 1 91.41 90.84 91.97
```

- First argument: model we are using (m_possum)
- Second argument (newdata): data frame of predictor values
- Third argument (interval): the kind of interval
  - Confidence interval for mean response: interval = "confidence"

Mean response: possum

- The estimated mean response is

$$\hat{\mu}_{y_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0 = 42.7098 + 0.0573 \times 850 = 91.4$$

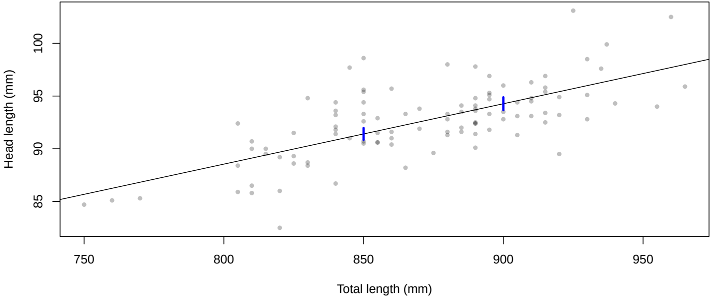
- Estimated mean head length for possums with total length 850 mm is 91.4 mm
  - Given by `fit` from `predict` output

```
mean_resp

##      fit   lwr   upr
## 1 91.41 90.84 91.97
```

- We are 95% confident that the mean head length for possums with total length 850 mm is between 90.8 mm and 92 mm
  - Given by `lwr` and `upr` in `predict` output

Mean response: visual

```r
mean_resp2 = predict(m_possum, newdata = predictor2, interval = "confidence")
mean_resp2

##      fit   lwr   upr
## 1 91.41 90.84 91.97
## 2 94.27 93.66 94.88
```



Prediction

- We can also use the model to predict a new observation y_0
- At a given value of $x = x_0$ (say $x_0 = 850$ mm)
  - The prediction ($\hat{y}_0$) is the same as the estimated mean response ($\hat{\mu}_{y_0}$)
    - Recall: fitted line was $\hat{y} = \hat{\mu}_y = \hat{\beta}_0 + \hat{\beta}_1 x$
- That means that at $x_0 = 850$ mm we have

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = 42.7098 + 0.0573 \times 850 = 91.4$$

- We predict that a (new) possum of 850 mm would have a head length of 91.4 mm
  - What about the possible error in the prediction?
  - We want to find a prediction interval?

Prediction error

- The prediction uncertainty is larger than the uncertainty about mean response
  - It needs to combine uncertainty about the mean response and individual variability
- Eg. if we are predicting the head length of a possum with total length 850 mm
  - The mean head length among the subpopulation of possums with total length 850 mm is uncertain
    - Standard error for mean response
  - There is possum to possum variability in head length among the subpopulation of possums with total length 850 mm
    - Not all possums with total length 850 mm will have the same head length
    - Given by the error ε in the linear regression model

Prediction error

- The prediction error takes account of both sources of uncertainty
- For prediction at $x = x_0$, the prediction error is

$$PE(\hat{y}_0) = s_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

  - Looks like standard error for mean response
    - Has an extra term in the square root: $1+$
    - Accounts for individual variation about the mean
- A $100(1-\alpha)\%$ prediction interval for y_0 is $\hat{y}_0 \pm t_{(1-\frac{\alpha}{2}, n-2)} \times PE(\hat{y}_0)$
- The prediction interval is a probability interval
  - There is a probability of $(1-\alpha)$ that y_0 will lie in this interval

Prediction in R

- Use the `predict` function, with option `interval = "prediction"`

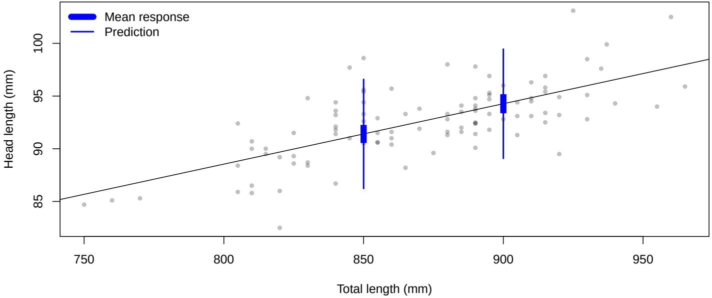
```
pred = predict(m_possum, newdata = predictor1, interval = "prediction")
pred
##     fit   lwr   upr
## 1 91.41 86.23 96.58
```

- There is a probability of 0.95 that a possum with total length 850 mm will have head length between 86.2 mm and 96.6 mm
- Note: we can find a 90% or 99% interval by including the argument `level`
    - Also applies when finding confidence interval for mean response

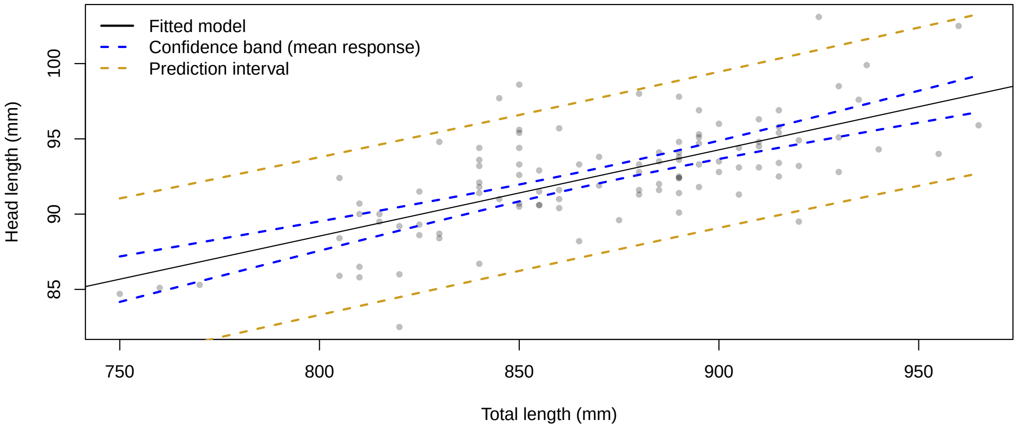
```
predict(m_possum, newdata = predictor1, interval = "prediction", level = 0.99)
##     fit   lwr   upr
## 1 91.41 84.55 98.26
```

Prediction: visual

```
pred2 = predict(m_possum, newdata = predictor2, interval = "prediction")
pred2

##      fit   lwr   upr
## 1 91.41 86.23 96.58
## 2 94.27 89.09 99.45
```



Mean response and prediction: visual



Mean response and prediction

- The mean response is most precise in middle of plot
  - Confidence interval is narrower
- Same is true of prediction interval (harder to see on plot)
- The standard error and prediction error both include the term

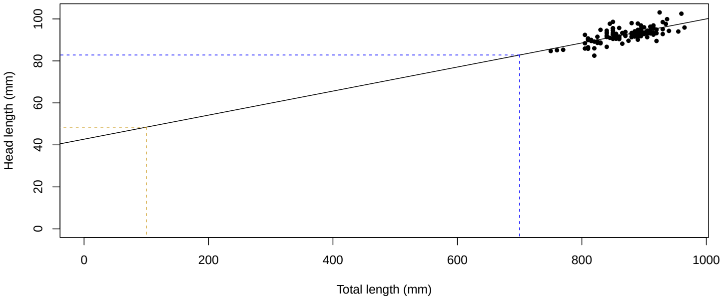
$$\frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- This is smallest when $x_0 = \bar{x}$
  - Estimation of mean response and prediction is most precise at $x_0 = \bar{x}$
  - Errors increase the further x_0 is from sample mean $\bar{x}$

Extrapolation

- When using linear regression models
  - Care is needed if extrapolating!
- Extrapolation: predicting values outside the range of the observed data
- Why is this a problem?
  - The linear regression model has limitations
    - It approximates the relationship between x and y across the range of data we observe
    - We don't necessarily believe it describes the true relationship between x and y
    - We don't know how data will behave outside the range we have observed
- If we decide to extrapolate
  - Important to know the risks and limitations

Extrapolation: possum



- The linear regression model provides a description of the relationship between total length and head length across the range of observed data
  - Total length between 750 mm and 950 mm
- We don't believe it describes the true relationship
  - We wouldn't use it to predict head length when total length is 100 mm
  - What about predicting head length when total length is 700 mm?

Summary

- Model summary: R^2
  - Squared correlation between fitted values and observations
  - Gives the percentage of variance explained by regression
- Looked again at mean response
  - Found confidence interval for mean response at $x = x_0$
- Looked at predicting a new observation
  - $\hat{y} = \hat{\mu}_y$
  - Prediction interval wider that confidence interval for mean response
- Looked at dangers of extrapolating