Simple Linear Regression

ST552 Lecture 2

Charlotte Wickham

January 6, 2015

Review

High level review

Since simple linear regression is a special case of multiple linear regression, we'll leave the "whys?" to when we cover multiple linear regression.

- the simple linear regression model
- interpretation
- assumptions
- how the estimates are found
- properties of the estimates
- F-tests

The simple linear regression model

n observations are collected in pairs, $(x_i, y_i), i = 1, \ldots, n$ where the y_i are generated according to the model,

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

What is random?

where ϵ_i are independent and identically distributed with expected value zero, and variance σ^2.

For inference, we often also assume the ϵ_i are Normally distributed,

$$\epsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$$

The assumptions in words

- Linearity: the mean response is a straight line function of the explanatory variable
- Constant spread: the standard deviation around the mean response is the same at all values of the explanatory variable
- Normality: the deviations from the mean response, the errors, are Normally distributed.
- Independence: the deviations from the mean response are independent.

Interpretation of the parameters

Intercept, β_0,
When the explanatory variable is zero, the mean response is β_0.

Slope, β_1,
An increase in the explanatory variable of one unit is associated with a change in mean response of β_1.

(Careful with causal language. . . is it justified?)

But we don't know β_0 and β_1. . .

The least squares estimates

Find $\hat{\beta}_0$ and $\hat{\beta}_1$ such that

$$\sum_{i=1}^{n} \Big(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\Big)^2$$

is minimized.

Fitted values: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
Residuals: $e_i = y_i - \hat{y}_i$

We don't require any properties of random variables to derive these estimates.

There are formulae for $\hat{\beta}_0$ and $\hat{\beta}_1$, how do you derive them?

Properties of the least squares estimates

Using the moment assumptions of ϵ_i, the least squares estimates can be shown to be unbiased. You can derive their variances, Var (β_0), Var (β_1), Cov (β_0, β_1), but they depend on the unknown σ.

An unbiased estimate of σ is

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2$$

Intuition: $\frac{1}{n} \sum_{i=1}^{n} e_i^2$ seems a reasonable place to start to estimate the variance of the errors, but this tends to underestimate the variance because we picked our estimates to make the sum of squared errors as small as possible.

Inference on the coefficients

With the addition of the Normality assumption,

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\widehat{\mathsf{Var}}\left(\beta_0\right)}} \sim t_{n-2} \quad \text{and} \quad \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\widehat{\mathsf{Var}}\left(\beta_1\right)}} \sim t_{n-2}$$

where $\widehat{\mathsf{Var}}\left(.\right)$ is the variance of the estimate with $\hat{\sigma}$ plugged in for σ.

Leads to confidence intervals and hypothesis tests of the individual coefficients.

Also under Normality the least squares estimates of slope and intercept **are** the maximum likelihood estimates.

Prediction

New?

Consider some new observation with explanatory value x_0. The true response is,

$$y_0 = \beta_0 + \beta_1 x_0 + \epsilon$$

with expected value

$$\mathsf{E}\left(y_0\right) = \beta_0 + \beta_1 x_0$$

There are two things we might be interested in:

- estimating the mean response at this value, $\hat{\mathsf{E}}(y_0)$
- predicting the response at this value, $\mathsf{Pred}(y_0)$

For both cases the point prediction is,

$$\mathsf{Pred}(y_0) = \hat{y_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

Prediction Intervals

When estimating the mean response, uncertainty only comes from the uncertainty in our estimates of the slope and intercept.

$$\mathsf{Var}\left(\hat{y_0}\right) = \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

Leads to confidence intervals of the form

$$\hat{y_0} \pm t_{n-2,1-\alpha/2}\sqrt{\widehat{\mathsf{Var}}\left(\hat{y_0}\right)}$$

When predicting a new response, uncertainty also comes from the variation about the mean.

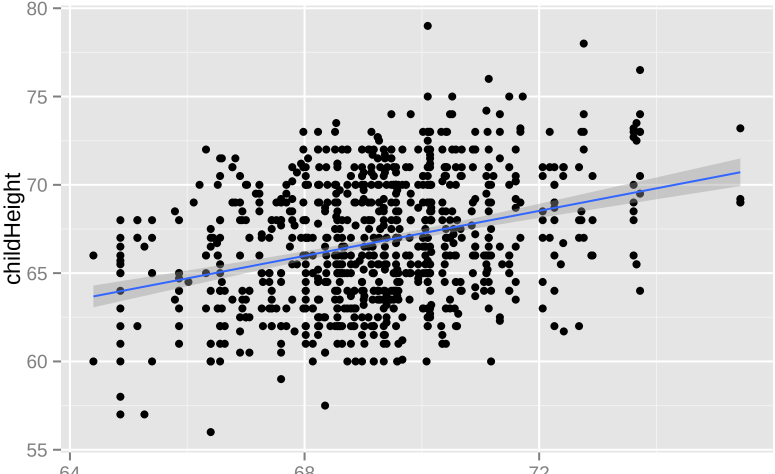
$$\mathsf{Var}\left(\mathsf{Pred}(y_0)\right) = \mathsf{Var}\left(\hat{y_0}\right) + \sigma^2$$

Leads to **prediction** intervals of the form

$$\hat{y_0} \pm t_{n-2,1-\alpha/2}\sqrt{\widehat{\mathsf{Var}}\left(\mathsf{Pred}(y_0)\right)}$$

Example

```r
library(ggplot2)
data(GaltonFamilies, package = "HistData")
qplot(midparentHeight, childHeight, data = GaltonFamilies) +
  geom_smooth(method = "lm")
```



Simple linear regression in R

```r
slr <- lm(childHeight ~ midparentHeight,
  data = GaltonFamilies)
slr
```

```
##
## Call:
## lm(formula = childHeight ~ midparentHeight, data = GaltonFamilies)
##
## Coefficients:
##     (Intercept)  midparentHeight
##         22.6362           0.6374
```

```r
summary(slr)
```

```
##
## Call:
## lm(formula = childHeight ~ midparentHeight, data = GaltonFamilies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.9570 -2.6989 -0.2155  2.7961 11.6848
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     22.63624    4.26511   5.307 1.39e-07 ***
## midparentHeight  0.63736    0.06161  10.345  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.392 on 932 degrees of freedom
## Multiple R-squared:  0.103,  Adjusted R-squared:  0.102
## F-statistic:   107 on 1 and 932 DF,  p-value: < 2.2e-16
```

Next

Lab:

- good coding practices

Homework:

- R data structures
- Simple linear regression review
- Matrix warm up

Next lecture

- Partioning the variability in simple linear regression
- F-tests