Statistics 305/605: Introduction to Biostatistical Methods for Health Sciences

R Demo for Chapter 18, part 2: Inference in Simple Linear Regression

Jinko Graham

2018-10-19

Load the Data and Fit the Regression Model

- Load the data on low-birthweight babies:

```
uu <- url("http://people.stat.sfu.ca/~jgraham/Teaching/S305_17/Data/lbwt.csv")
lbwt <- read.csv(uu)
```

- Fit the regression model and print out the fitted regression coefficients:

```
lfit <- lm(headcirc ~ gestage,data=lbwt)
coefficients(lfit)
```

```
## (Intercept)     gestage
##   3.9142641   0.7800532
```

- The regression parameters are estimated by $\hat{\alpha} = 3.91$ (intercept) and $\hat{\beta} = 0.78$ (gestage).

Testing Example

- For the low-birthweight data, the model summary includes the p-value from the tests of $H_0 : \beta = 0$ *vs.* $H_a : \beta \neq 0$

```
summary(lfit)
```

```
##
## Call:
## lm(formula = headcirc ~ gestage, data = lbwt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5358 -0.8760 -0.1458  0.9041  6.9041
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.91426    1.82915    2.14   0.0348 *
## gestage      0.78005    0.06307   12.37   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.59 on 98 degrees of freedom
## Multiple R-squared:  0.6095, Adjusted R-squared:  0.6055
## F-statistic: 152.9 on 1 and 98 DF,  p-value: < 2.2e-16
```

- We're interested specifically in the `coefficients` component of the summary.
- Can extract the `coefficients` component with the `$` operator.

```
summary(lfit)$coefficients
```

```
##               Estimate Std. Error  t value    Pr(>|t|)
## (Intercept) 3.9142641 1.82914689  2.13994 3.48424e-02
## gestage     0.7800532 0.06307441 12.36719 1.00121e-21
```

- The test statistic value is about 12.37 and the p-value is tiny.
- Strong statistical evidence for an association between gestational age and head circumference.

Software Notes

- The output of the `summary()` function includes a lot of components that we are not yet ready for.
  - However `summary(lfit)$coefficients` extracts just the coefficients table.
- Statistics related to a particular coefficient are in the row of the table labelled by the name of the explanatory variable.
  - E.G., Below the summaries related to the slope of the regression line are in the row labelled `gestage`.

```
summary(lfit)$coefficients
```

```
##               Estimate Std. Error  t value    Pr(>|t|)
## (Intercept) 3.9142641 1.82914689  2.13994 3.48424e-02
## gestage     0.7800532 0.06307441 12.36719 1.00121e-21
```

CI Example

- We can use the `confint()` function in R to extract a confidence interval.

```
confint(lfit,conf.level=0.95)
```

```
##                 2.5 %    97.5 %
## (Intercept) 0.2843817 7.5441466
## gestage     0.6548841 0.9052223
```

- The 95% CI for β is about $(0.65, 0.91)$.
- Interpret: "With 95% confidence, we estimate that a one-week increase in gestational age is associated with an increase in head circumference of between 0.65 to 0.91 cm."

CIs at Observed Values of Explanatory Variable

- Use the `predict()` function to get predictions and confidence intervals for each observed value of the explanatory variable.
- Default coverage probability or level for the CI is $C = 0.95$

```
lpred <- predict(lfit,interval="confidence")
head(lpred)
```

```
##        fit      lwr      upr
## 1 26.53581 26.21989 26.85172
## 2 28.09591 27.68437 28.50745
## 3 29.65602 29.05247 30.25956
## 4 28.09591 27.68437 28.50745
## 5 27.31586 26.97102 27.66070
## 6 23.41559 22.83534 23.99584
```

- In a given row, 1st entry is fitted value $\hat{y}$, 2nd entry is lower bound of CI and last entry is upper bound of CI.
- Attach `lpred` to the corresponding values for y (`headcirc`) and x (`gestage`), to create a new R object called `lbwtFits`

...

```
lbwtFits <- data.frame(headcirc=lbwt$headcirc,gestage=lbwt$gestage,lpred)
head(lbwtFits)
```

```
##   headcirc gestage      fit      lwr      upr
## 1       27      29 26.53581 26.21989 26.85172
## 2       29      31 28.09591 27.68437 28.50745
## 3       30      33 29.65602 29.05247 30.25956
## 4       28      31 28.09591 27.68437 28.50745
## 5       29      30 27.31586 26.97102 27.66070
## 6       23      25 23.41559 22.83534 23.99584
```

- The values of the response, y, are in the column `headcirc`.
- The values of the explanatory variable, x, are in the column `gestage`.
- The fitted values $\hat{y}$ are in the column `fit`.
- The lower limits of the CIs are in the column `lwr` and the upper limits are in the column `upr`.

90% CIs at New Values of Explanatory Variable

- Suppose that we want 90% CI's at new values of the explanatory variable, such as `gestage` of 25.5 and 30.5 weeks.
- Create a dataset with the new values of the explanatory variables and pass this to `predict()`.
- Specify the level of the confidence interval with the `level` argument.

```
newdat <- data.frame(gestage = c(25.5,30.5))
predict(lfit,newdata = newdat, interval="confidence", level=.90)
```

```
##        fit      lwr      upr
## 1 23.80562 23.36311 24.24813
## 2 27.70589 27.39254 28.01923
```

- The fitted values $\hat{y}$ of `headcirc` for `gestages` of 25.5 and 30.5 are in the column `fit` and are about 23.8 and 27.7, respectively.
- The lower limits of the 90% CIs are in the column `lwr` and the upper limits are in the column `upr`.

Adding CIs to a Scatterplot

- Pointwise 95% CIs around the fitted regression line are added by `ggplot() + geom_smooth(method="lm")`
  - `geom_smooth(method="lm", level=.9)` will give 90% CIs.

```
library(ggplot2)
ggplot(lbwt, aes(x=gestage,y=headcirc)) +
   geom_point() +
   geom_smooth(method="lm", level=.9)
```