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

Chapter 18, part 2: Inference in Simple Linear Regression

Jinko Graham

2018-10-19

Inference in Regression

- Estimate the population conditional means $\mu_{y|x} = \alpha + \beta x$ by
$$\hat{\mu}_{y|x} = \hat{y} = \hat{\alpha} + \hat{\beta} x.$$
- If we could observe the errors, $\epsilon = Y - \mu_{y|x}$, we could estimate $\sigma_{y|x}$ by
$$\sqrt{\frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2}$$
- But we can't observe the errors because we don't know the population conditional means $\mu_{y|x}$.
- Instead, substitute the *residuals*, $e = y - \hat{\mu}_{y|x} = y - \hat{y}$, and estimate $\sigma_{y|x}$ by:
$$s_{y|x} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n} e_i^2} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}.$$
Divide by $n-2$, the sample size less the number of parameters used to estimate the conditional mean.

Hypothesis Test for β

- We can test the null hypothesis of no association between X and Y *vs.* the alternative of association; i.e.,

$$H_0 : \beta = 0 \text{ vs. } H_a : \beta \neq 0.$$

- The test statistic is derived from the sampling distribution of $\hat{\beta}$.
- Assuming that the error terms, ϵ, in the regression model are normally distributed, the sampling distribution of $\hat{\beta}$ is normal with mean β and SD

$$SD(\hat{\beta}) = \frac{\sigma_{y|x}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

- Replace $\sigma_{y|x}$ by $s_{y|x}$ to get standard error of $\hat{\beta}$, $SE(\hat{\beta})$.

- Replace $\sigma_{y|x}$ by $s_{y|x}$ to get standard error of $\hat{\beta}$:

$$SE(\hat{\beta}) = \frac{s_{y|x}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

.

- We will always use computer software to get the SE.
- The pivotal quantity
$$\frac{\hat{\beta} - \beta}{SE(\hat{\beta})}$$
has a t-distribution with $n - 2$ df.
- To test $H_0 : \beta = 0$ *vs.* $H_a : \beta \neq 0$, the test statistic is

$$T = \frac{\hat{\beta} - 0}{SE(\hat{\beta})} = \frac{\hat{\beta}}{SE(\hat{\beta})}$$

Testing Example

- For the low-birthweight babies, let X be the gestational age (in weeks) and Y be the head circumference (in cm).
- The regression coefficient β summarizes the association between X and Y. Test for association using hypotheses $H_0 : \beta = 0$ *vs.* $H_a : \beta \neq 0$

```
##             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.
- There is strong statistical evidence of an association between gestational age and head circumference in the data.

Confidence Intervals

- Following the typical development, a CI for β can be derived from the pivotal quantity

$$\frac{\hat{\beta} - \beta}{SE(\hat{\beta})}$$

.

- The level-C CI is of the usual form

$$\text{estimate} \pm \text{margin of error},$$

where

  - the estimate is $\hat{\beta}$,
  - the margin of error is $SE(\hat{\beta})$ times a critical value t^*, the upper $(1 - C)/2$ critical value from the t-distribution with $n - 2$ df.

Confidence Interval Example

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

- From the above R output, the 95% CI for β is about $(0.65, 0.91)$; i.e., in 95 out of 100 samples, we expect the CI to cover the true β
- One way to interpret (from text):
  - "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."

Inference about the Regression Line

- The conditional mean, $\mu_{y|x} = \alpha + \beta x$, is a population parameter.
- The fitted value at x, $\hat{y} = \hat{\alpha} + \hat{\beta}x$, is an estimate of $\mu_{y|x}$
- The statistic $\hat{y}$ has a sampling distribution whose SD can be estimated by $SE(\hat{y})$, the standard error given on page 429 of the text (text's notation is $\widehat{se}(\hat{y})$).
- We can construct a level-C CI for $\mu_{y|x}$ of the usual form estimate $\pm$ margin of error, where
  - the estimate is $\hat{y}$, and
  - the margin of error is $SE(\hat{y})$ times t^*, the upper $(1 - C)/2$-critical value of the t-distribution with $n - 2$ df.
- We will always use a computer to calculate CIs for the regression line.

CIs at Observed Values of Explanatory Variable

```
##   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
```

In the R output above:

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

CIs at New Values of Explanatory Variable

- The output below gives 90% CIs at **new** values of the explanatory variable; i.e., `gestage` of 25.5 and 30.5 weeks.

```
##   gestage      fit      lwr      upr
## 1    25.5 23.80562 23.36311 24.24813
## 2    30.5 27.70589 27.39254 28.01923
```

- The fitted values $\hat{y}$ of `headcirc` for `gestage`s 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`.