

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

Chapter 19, part 2: Inference in Multiple Regression

Jinko Graham

Load the Low-Birthweight Data and Fit Regression Models

```
uu <- url("http://people.stat.sfu.ca/~jgraham/Teaching/S305_17/Data/lbwt.csv")
lbwt <- read.csv(uu)
fit1 <- lm(headcirc ~ gestage, data=lbwt) #SLR
summary(fit1)$coefficients #SLR Coeffs
```

```
##              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
```

```
fit2 <- lm(headcirc ~ gestage + birthwt, data=lbwt) #MLR
summary(fit2)$coefficients #MLR Coeffs
```

```
##              Estimate Std. Error  t value    Pr(>|t|)
## (Intercept)  8.308015388  1.578942936  5.261758 8.535816e-07
## gestage      0.448732848  0.067245982  6.673006 1.555501e-09
## birthwt      0.004712283  0.000631179  7.465843 3.596527e-11
```

- ▶ In the SLR, the slope estimate for gestage is $\hat{\beta} = 0.78$.
- ▶ By contrast, in the MLR, the slope estimate for gestage is $\hat{\beta}_1 = 0.45$.

Inference in Multiple Linear Regression

- ▶ Estimate $q + 1$ population parameters $\alpha, \beta_1, \dots, \beta_q$ by $\hat{\alpha}, \hat{\beta}_1, \dots, \hat{\beta}_q$.
- ▶ If we could observe the errors, $\epsilon = Y - \mu_{Y|X_1, \dots, X_q}$, we could estimate σ_Y , the SD of Y , by

$$\sqrt{\frac{1}{n} \sum_{i=1}^n \epsilon_i^2}$$

- ▶ Instead, substitute *residuals* $e_i = y - \hat{\mu}_{Y|X_1, \dots, X_q} = y - \hat{y}$, and estimate σ_Y by:

$$s_y = \sqrt{\frac{1}{n - q - 1} \sum_{i=1}^n e_i^2} = \sqrt{\frac{1}{n - q - 1} \sum_{i=1}^n (y_i - \hat{y}_i)^2},$$

called the **residual standard error** in R model summaries.

- ▶ The degrees of freedom (df) are $n - (q + 1) = n - q - 1$, the number of observations, n , less the number of parameters, $q + 1$, used to estimate the population mean $\mu_{Y|X_1, \dots, X_q}$.

Hypothesis Tests for β_j 's

- ▶ We can test the null hypothesis that X_j is not associated with Y vs. the alternative that it is; i.e.,

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

- ▶ The test statistic is derived from the sampling distribution of $\hat{\beta}_j$ which is normal with mean β_j under our modelling assumptions.
- ▶ From this we can get that the pivotal quantity

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

has a t -distribution with $n - q - 1$ df.

- ▶ The test statistic is

$$T_j = \frac{\hat{\beta}_j - 0}{SE(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}.$$

Testing Example

- ▶ For the low-birthweight data, the model summary from the `lm()` function includes the p-values from the tests of $H_0 : \beta_j = 0$ versus $H_a : \beta_j \neq 0$.

```
summary(fit2)$coefficients
```

##	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	8.308015388	1.578942936	5.261758	8.535816e-07
## gestage	0.448732848	0.067245982	6.673006	1.555501e-09
## birthwt	0.004712283	0.000631179	7.465843	3.596527e-11

- ▶ For inference of β_1 (gestage) the test statistic value is about 6.67 and the p-value is tiny.
- ▶ For inference of β_2 (birthwt) the test statistic value is about 7.47 and the p-value is tiny.
- ▶ The tiny pvalues for both gestage and birthwt indicate strong statistical evidence for both being associated with head circumference.

Confidence intervals

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

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

- ▶ The level- C CI is of the usual form

estimate \pm margin of error,

where

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

CI Example

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

```
confint(fit2,conf.level=0.95)
```

```
##                2.5 %        97.5 %  
## (Intercept) 5.174250734 11.441780042  
## gestage      0.315268189  0.582197507  
## birthwt      0.003459568  0.005964999
```

- ▶ The 95% CI for β_1 , the slope term for gestage, is about (0.32, 0.58):
 - ▶ “With 95% confidence, we estimate that, for a given birth weight, a one week increase in gestational age is associated with an increase in head circumference of between 0.32 to 0.58cm.”
- ▶ The interpretation of the 95% CI for β_2 is analogous.

Inference about the Regression Line

- ▶ The fitted value $\hat{y} = \hat{\alpha} + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_q x_q$ is an estimate of the population conditional-mean response

$$\mu_{y|x_1, \dots, x_q} = \alpha + \beta_1 x_1 + \dots + \beta_q x_q$$

and is also our prediction of a future y at x_1, x_2, \dots, x_q .

- ▶ As in SLR, level- C CIs for $\mu_{y|x_1, \dots, x_q}$ and level- C PIs for y at x_1, x_2, \dots, x_q are of the usual form

estimate \pm margin of error.

- ▶ R's `predict()` function can be used to obtain these CIs and PIs.

90% CIs and PIs at New Values of Explanatory Variables

- ▶ Suppose that we want 90% CIs or PIs at new values of the explanatory variables; E.G.,
 - ▶ gestage of 25.5 weeks and birthwt of 1050 grams for one individual, and
 - ▶ gestage of 30.5 weeks and birthwt of 1450 grams for another individual.
- ▶ Create a dataframe newdat containing these new values:

```
newdat <- data.frame(gestage = c(25.5,30.5),birthwt=c(1050,1450))  
newdat
```

```
##   gestage birthwt  
## 1    25.5    1050  
## 2    30.5    1450
```

- ▶ Pass these values to `predict()`.

```
predict(fit2,newdata = newdat, interval="confidence",level=.90)
```

```
##           fit      lwr      upr
## 1 24.69860 24.29225 25.10495
## 2 28.82718 28.47331 29.18104
```

- ▶ Let's focus on the 1st set of new values, gestage=25.5 weeks and birthwt=1050 g.
 - ▶ The fitted value of headcirc is in the column fit and is about 24.7 cm.
 - ▶ The lower limit of the 90% CI is in the column lwr and is about 24.3 cm.
 - ▶ The upper limit of the 90% CI is in the column upr and is about 25.1 cm.
- ▶ Replace the argument **interval = "confidence"** with **interval = "prediction"** for the Pls.
- ▶ Omit the argument **newdata=newdat** for CIs/Pls at the observed values of the explanatory variables.

Coefficient of Determination (R^2)

- ▶ In MLR, the coefficient of determination, R^2 , is the fraction of the variation in the values of Y that is explained by the least-squares regression of Y on X_1, \dots, X_q .
 - ▶ Large R^2 means observed responses fall close to fitted values.

- ▶ E.G. adding birthwt to gestage as an explanatory variable increases R^2 from about 0.61 to 0.75:

```
summary(fit1)$call #SLR model without birthwt
```

```
## lm(formula = headcirc ~ gestage, data = lbwt)
```

```
summary(fit1)$r.squared # its  $R^2$ 
```

```
## [1] 0.6094799
```

```
summary(fit2)$call # MLR model with birthwt
```

```
## lm(formula = headcirc ~ gestage + birthwt, data = lbwt)
```

```
summary(fit2)$r.squared # its  $R^2$ 
```

```
## [1] 0.751992
```

Comparing Models

- ▶ For comparing 2 regression models, R^2 is not useful because it **always increases** as we add explanatory variables, even if they have no actual effect on Y .
- ▶ By contrast, the *adjusted R^2* doesn't always increase.
 - ▶ It is instead designed to increase whenever adding an explanatory variable improves the model's ability to **predict** new values.
- ▶ Adjusted R^2 is one of a number of tools for *model selection* based on the predictive ability of a model that we won't have time to cover.
 - ▶ If you are interested, consider taking STAT 452: **Statistical Learning and Prediction** after this course. You will have the STAT 305 pre-requisite.