Interpretation in Multiple Regression

Topics:

- 1. R-squared and Adjusted R-squared
- 2. Interpretation of parameter estimates
- 3. Linear combinations of parameter estimates
 - variance-covariance matrix
 - standard errors of combinations
 - standard error for the mean

We will use the final model from last time to illustrate these concepts. Summaries of the model - least squares estimates with standard errors given below in parentheses:

 $\widehat{logit}(proportion) = -2.71 + 0.89 * \log(duration) + 0.57 * I$ (0.38) (.14) (0.24)

 $\hat{\sigma} = 0.65$ with 44 degrees of freedom R-squared = 0.6068029

R-squared and Adjusted R-squared: The R-squared value means that 61% of the variation in the logit of proportion of pollen removed can be explained by the regression on log duration and the group indicator variable. As R-squared values increase as we ass more variables to the model, the adjusted R-squared is often used to summarize the fit as it takes into account the the number of variables in the model.

Adjusted R-squared = 1 - Mean Square Error /Total Mean Square

where Mean Square Error is $\hat{\sigma}^2$ from the regression model and the Total mean square is the sample variance of the response ($s_Y^2 = \hat{\sigma}^2$ is a good estimate if all the regression coefficients are 0). For this example,

Adjusted R-squared = $1 - 0.65^{2}/1.034 = 0.59$.

Intercept: the intercept in a multiple regression model is the mean for the response when all of the explanatory variables take on the value 0.

In this problem, this means that the dummy variable I = 0 (code = 1, which was the queen bumblebees) and log(duration) = 0, or duration is 1 second. For queenbumblebees, with visits of 1 second, we are 95% confident that the mean logit(proportion of pollen removed) is between $-2.71\pm2.02*0.38$ or between -3.49 to -1.93. The Student t quantile 2.02 is based on 44 degrees of freedom; qt(.975, 44).

To convert back to the original units, we can take the inverse of the logit transformation. I.e. if logit(p) = log(p/(1-p)), then p = exp(x)/(1 + exp(x)). To get the confidence interval

for the proportion just apply the inverse transformation. So for queen bumblebees with visits lasting 1 second, we are 95% confident that the mean proportion of pollen removed is between 0.03 and 0.13. [exp(-3.49)/(1 + exp(-3.49)) to exp(-1.93)/(1 + exp(-1.93))] Note: while this is the interpretation of the intercept, we are extrapolating.

<u>Regression Coefficients</u>: Typically the coefficient of a variable is interpreted as the change in the response based on a 1-unit change in the corresponding explanatory variable keeping all other variables held constant. In some problems, keeping all other variables held fixed is impossible (i.e. A quadratic model, or the model with different slopes for queen and worker bees).

For this example, we have the estimated coefficient of log(duration) is 0.89. Because we have taken the log transformation of duration, the interpretation of the coefficient is easier to understand by looking at a doubling of duration (review page 208 chapter 8).

A doubling of the duration of visit corresponds to a $\beta_1 \log(2)$ change in the mean logit(proportion of pollen removed) or $0.89*\log(2) = 0.62$. The 95% confidence interval for β_1 is $0.89\pm2.02*0.14$ or 0.61 to 1.17. The interval under the doubling of duration is obtained by multiplying this interval by log(2). So a 95% confidence interval for the change in the mean logit(proportion pollen removed) is 0.42 to 0.81. Further simplification is not possible.

Dummy variable coefficients: A 1 unit change for a dummy variable implies going from level 0 to level 1, so the the interpretation of the dummy variable coefficient is the amount by which the mean logit(proportion) for worker bees exceeds the mean logit(proportion) for queen bumble bees. i.e the logit of the proportion pollen removed for worker bees is 0.56 higher than the logit for queen bumble bees. A 95% confidence interval for the amount is 0.09 to 1.05. (this is the case for parallel regression lines; if we still had the interaction variable we could not make this statement, since the interaction of the dummy*log(duration) cannot be held constant).

In the model derivation, we said that the intercept plus the dummy variable coefficient corresponded to the intercept for the worker bees, which is estimated as - 2.71 + .57 or - 2.14. This can be translated ac to the original scale as we did the intercept for the queen bumble bees. As this has a more interesting meaning, let's find a confidence interval for $\beta_0 + \beta_2$. To do this we need to find the standard error for a linear combination.

Linear Combination of Parameters

To find the variance (and then standard deviation) of the estimator of $\beta_0 + \beta_2$ we need to take into account the individual variances plus how the estimates will vary together from sample to sample (their covariance). The variance of the sum is the sum of the variances plus 2 times the covariance. We can get the covariance from the correlation of the estimates (recall the correlation is the covariance divided by the product of the standard deviations, so the covariance is the correlation times the product of the standard deviations. Since the standard deviations are unknown, we use the estimated covariance matrix calculated using the standard errors. In the Results options for Regression, check

off the box to get the estimated correlation matrix of the coefficients.

Correlation of Coefficients: (Intercept) log.duration log.duration -0.9579857 ----I 0.2361514 -0.3860614

The correlation between something and itself is one, so this part has been omitted. Since the correlation of (b0, b1) is the same as the correlation of (b1, b0) the table only includes the elements below the diagonal.

The estimated covariance matrix is symmetric (just like the correlation matrix). The diagonal elements are the covariance between βi and βi which are the variances, or the square of the standard errors:

| Covariance Matrix of the Parameter Estimates | | | |
|---|-------------|--------------|---------|
| coefficient | (Intercept) | log.duration | Ι |
| (Intercept) | 0.1476 | -0.0516 | 0.0214 |
| log.duration | -0.0515 | 0.0196 | -0.0127 |
| Ι | 0.0215 | -0.0127 | 0.0558 |

The covariance between the intercept and the dummy variable I coefficient is estimated as the (correlation between the intercept and the coefficient for I) * (SE(intercept))(SE(of the coefficient for I)) or 0.2361514 * 0.38 * 0.24 = 0.0215.

The standard error for the estimate of $\beta_0 + \beta_2$ is the sqrt($0.38^2 + 0.24^2 + 2*0.0215$) = 0.50.

Thus a 95% Confidence interval for the intercept for worker bees is $-2.15\pm2.02*0.50=(-3.16,-1.14)$

Transforming back to the original units, we are 95% confident that the mean proportion of pollen removed by worker bees for visits of 1 second is between 0.04 to 0.24.

Standard Error of the Mean:

The same approach is used to find the standard error of the mean for a visit of a given duration. i.e the estimate of the mean of logit for a visit of 2 seconds for a queen bumble bee is $2.71 * 1 + 0.89 * \log(2) + 0.56 * 0$. The values C0 = 1, $C1 = \log(2)$, C2 = 0 are the multipliers Ci in the linear combination

 $\sum_{i=0}^{p} \hat{\beta}_{i} C_{i}$ which gives us the estimate of the linear combination (the mean in this case)

The variance of the mean at this point is found by $\sum_{i=0}^{p} \sum_{j=0}^{p} cov(\hat{\beta}_{i}, \hat{\beta}_{j}) C_{i}C_{j}$

which in this case simplifies to

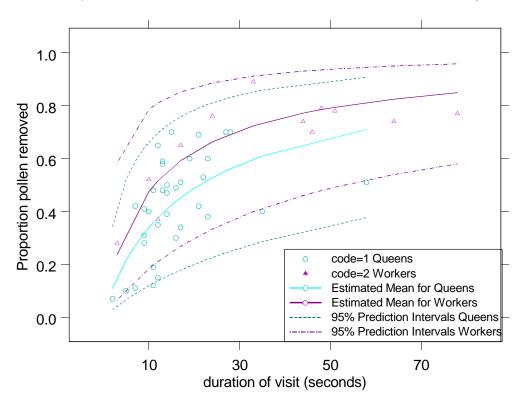
$$var(\hat{\beta}_0) * 1 + var(\hat{\beta}_1) \log(2)^2 + 2 * cov(\hat{\beta}_0, \hat{\beta}_1) * 1 * \log(2) = 0.0841$$

For more details see section 10.4.3 and exercises 21-23. This is how the se.fit variable is obtained. From the SE(mean) we can get the SE(prediction)

$$SE(prediction(Y|X=x)) = \sqrt{\{Xhat \sigma^2 + SE(\mu(Y|X=x)^2)\}}$$

To get a prediction interval first calculate the prediction interval in the logit scale, then transform the interval using the inverse transformation applied to each endpoint of the interval.

Putting this all together we can find the estimates and prediction intervals in the original units.



Proportion of Pollen Removed for Queen Bumblebees and Worker Honeybees