Categorical Predictors with Two Levels and Inference in Regression

Chapter 6, Lab 3: Solutions

OpenIntro Biostatistics

Topics

- Categorical predictors with two levels
- Inference in regression

This lab introduces the idea of using a categorical predictor variable (specifically, a categorical predictor with two levels) in regression and also discusses the extension of statistical inference to the regression context.

The material in this lab corresponds to Sections 6.3.3 and 6.4 of *OpenIntro Biostatistics*.

Introduction

Categorical predictors with two levels

Although the response variable in linear regression is necessarily numerical, the predictor variable may be either numerical or categorical. Simple linear regression only allows for categorical predictors with two levels; examining categorical predictors with more than two levels requires multiple linear regression.

Fitting a simple linear regression model with a categorical predictor that has two levels is analogous to comparing the means of two groups, where the groups are defined by the categorical variable. The equation of the regression line has intercept b_0, which equals the mean of one of the groups, and slope b_1, which equals the difference in means between the two groups.[1]

Inference in regression

When conducting inference in a regression context, observed data (x_i, y_i) used for fitting a regression line are assumed to have been randomly sampled from a population where the explanatory variable X and response variable Y follow a population model

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

where $\epsilon \sim N(0, \sigma)$. Under this assumption, the intercept and slope of the regression line, b_0 and b_1, are estimates of the population parameters β_0 and β_1.

Hypothesis tests and confidence intervals for regression population parameters have the same basic structure as tests and intervals about population means. Inference is usually done about the slope, β_1. Under the null hypothesis, the variables X and Y are not associated; $H_0 : \beta_1 = 0$. Under the alternative hypothesis, the variables X and Y are associated; $H_1 : \beta_1 \neq 0$.

[1] The group for which b_0 is the mean is usually referred as the *baseline* group or *reference* group.

Categorical predictors with two levels

Obesity is known to be a leading risk factor for diabetes. The following questions step through exploring the association between BMI (`BMI`) and presence of diabetes (`DM`) in a random sample of $n = 500$ individuals from the PREVEND data.

1. Run the code chunk shown in the template to load `prevend.samp`, the random sample of 500 individuals from the PREVEND data used in the previous labs in this chapter.

   The code chunk has been run, but the code and output have been suppressed in this PDF using `echo = FALSE`.

2. Examine the variable `DM` and identify how many individuals in `prevend.samp` have diabetes. Presence of diabetes is coded as `1` and absence is coded as `0`.

   33 individuals in `prevend.samp` have diabetes.

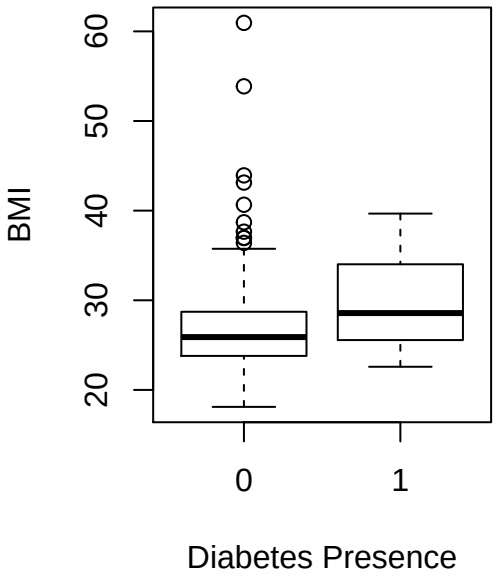
```
#examine DM
table(prevend.samp$DM)
```

```
##
##   0   1
## 467  33
```

3. Create plots that show the association between BMI and presence of diabetes.

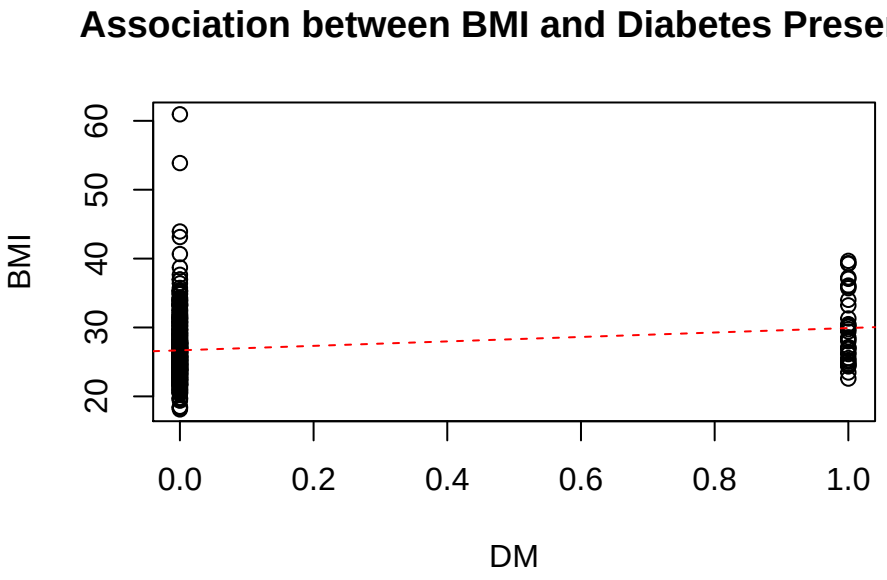
   a) Create a boxplot that shows the association between BMI and presence of diabetes.

```
boxplot(prevend.samp$BMI ~ prevend.samp$DM,
    ylab = "BMI", xlab = "Diabetes Presence")
```



b) Create a scatterplot of BMI versus presence of diabetes and plot the least-squares line.

```
plot(BMI ~ DM, data = prevend.samp,
 main = "Association between BMI and Diabetes Presence")
abline(lm(BMI ~ DM, data = prevend.samp), col = "red", lty = 2)
```



c) Based on the plots, comment briefly on the nature of the association.

Overall, individuals with diabetes tend to have a higher BMI than individuals without diabetes. As seen in the boxplot, median BMI for individuals with diabetes is higher than median BMI for individuals without diabetes.

4. In the Chapter 1 Lab Notes, the **factor** data structure was introduced. Factors are ideal for storing categorical data. In a factor variable, the levels of the variable are displayed as characters (such as "Female" and "Male") while the data remain stored as integer values (such as `0` and `1`); each level of the variable is associated with a specific integer value.

a) Run the following code chunk to create the factor variable `DM.factor` from the integer vector `DM`. Note that while the `DM` variable is part of the `prevend.samp` dataframe, the variable `DM.factor` is not.

```
#create DM.factor
DM.factor = factor(prevend.samp$DM, levels = c(0, 1),
                                labels = c("Absent", "Present"))
```

b) Run the `summary()` function on both `DM.factor` and `DM`. Compare the output and comment on which one has interpretive meaning.

Running `summary()` on the factor variable returns the number of individuals in each level of the variable; this is informative. Running `summary()` on the integer vector gives a numerical summary of the 0 and 1 values, which has no interpretive meaning since the 0 and 1 values represent levels of a categorical variable rather than numeric data.

```
#use summary( ) on DM.factor
summary(DM.factor)
```

```
##  Absent Present
##     467      33
```

```
#use summary( ) on DM
summary(prevend.samp$DM)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   0.066   0.000   1.000
```

c) Run the following code chunk to overwrite the `DM` variable (in `prevend.samp`) with the factor variable `DM.factor`. Confirm that the overwrite was successful by running `summary()` on `DM`.

```
#overwrite DM with DM.factor
prevend.samp$DM <- DM.factor

#confirm the overwrite is successful
summary(prevend.samp$DM)
```

```
##  Absent Present
##     467      33
```

5. Calculate mean BMI for diabetic individuals and individuals without diabetes.

   Mean BMI for diabetic individuals is 29.92 and mean BMI for individuals without diabetes is 26.69.

```
tapply(prevend.samp$BMI, prevend.samp$DM, mean)
```

```
##   Absent  Present
## 26.68580 29.91872
```

6. Use a linear regression model to relate BMI and diabetes presence.

   a) Using a residual plot and Q-Q plot, check the assumptions for linear regression. It is reasonable to assume that these observations are independent.

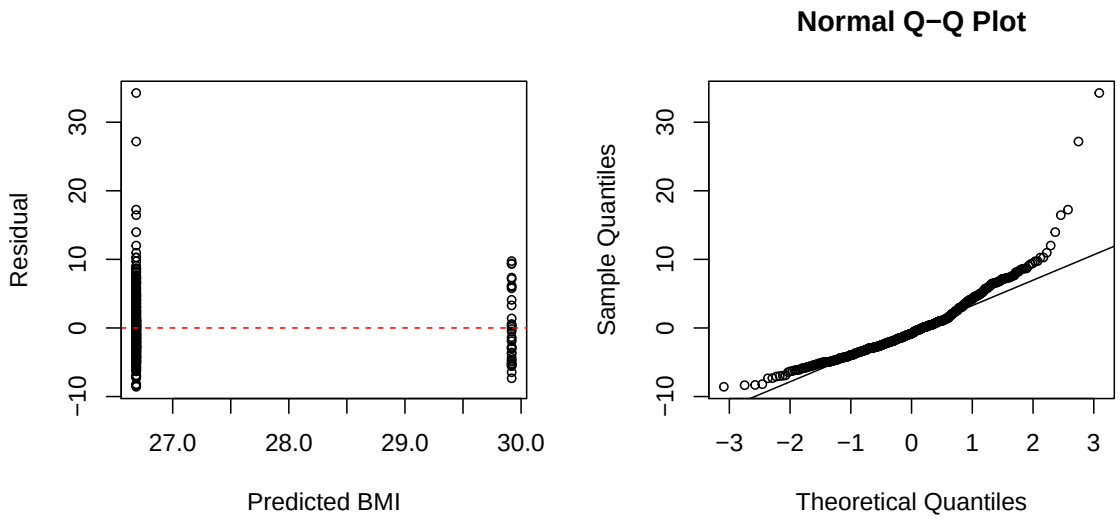
      Note how linearity is automatically satisfied for a categorical predictor variable; since the data are lined up at either of two points on the x-axis, the best fit line simply passes through the center of both groups. The variability in BMI is roughly constant between the groups; calculating the variance of BMI in each group can confirm this. The residuals are close to a normal distribution in the center, but there is deviation from normality in the upper tail.

```
par(mfrow = c(1, 2))

#residual plot
residuals = resid(lm(BMI ~ DM, data = prevend.samp))
predicted = predict(lm(BMI ~ DM, data = prevend.samp))
plot(residuals ~ predicted,
```

```
 ylab = "Residual", xlab = "Predicted BMI",
 cex = 0.75)
abline(h = 0, lty = 2, col = "red")

#calculate variances
tapply(prevend.samp$BMI, prevend.samp$DM, var)
```

```
##   Absent  Present
## 20.09413 26.54779
```

```
#Q-Q plot
qqnorm(residuals, cex = 0.75)
qqline(residuals)
```



b) Write the equation of the least-squares line in terms of the variable names (e.g., *BMI*).

$$\widehat{BMI} = 26.69 + 3.23(DMPresent)$$

```
#print the values of the coefficients
coef(lm(BMI ~ DM, data = prevend.samp))
```

```
## (Intercept)   DMPresent
##   26.685800    3.232923
```

c) Based on part b), solve for the two possible values of $\widehat{BMI}$ and interpret the values.

The predictor variable can take on either 0 or 1. When *DMPresent* is 0, $\widehat{BMI}$ is 26.69; this is the estimated average BMI of non-diabetic individuals. When *DMPresent* is 1, $\widehat{BMI}$ is 29.92; this is the estimated average BMI of diabetic individuals.

```
#use r as a calculator
b0 = coef(lm(BMI ~ DM, data = prevend.samp))[1]
b1 = coef(lm(BMI ~ DM, data = prevend.samp))[2]
```

```r
x = c(0, 1)

b0 + b1*x
```

```
## [1] 26.68580 29.91872
```

Alternatively, use the `predict( )` function.

```r
#store the model fit
model.BMIvsDM = lm(BMI ~ DM, data = prevend.samp)

#use predict( )
predict(model.BMIvsDM, newdata = data.frame(DM = "Absent"))
```

```
##       1
## 26.6858
```

```r
predict(model.BMIvsDM, newdata = data.frame(DM = "Present"))
```

```
##        1
## 29.91872
```

d) Confirm that the numbers obtained in part c) match those from Question 5.

The numbers from part c) are identical to the answers from Question 5.

Inference in regression

Inference in a regression context is usually for the slope parameter, β_1.

The null hypothesis in regression is most commonly a hypothesis of 'no association', similar to how the null hypothesis when testing for a difference of means is often one of 'no difference'. When two variables are not associated, plotting them against each other results in a cloud of points with no apparent trend; in this setting, the slope of a least-squares line equals 0.

Thus, the hypotheses in regression can be written as:

- $H_0 : \beta_1 = 0$, the X and Y variables are not associated
- $H_A : \beta_1 \neq 0$, the X and Y variables are associated

The t-statistic for a null hypothesis $H_0 : \beta_1 = \beta_1^0$ has degrees of freedom $df = n - 2$, where n is the number of ordered pairs in the dataset. The value β_1^0 equals 0 when the null hypothesis is one of no association.

$$t = \frac{b_1 - \beta_1^0}{\text{s.e.}(b_1)} = \frac{b_1}{\text{s.e.}(b_1)}$$

A 95% confidence interval for β_1 has the following formula, where $t^\star$ is the point on a t-distribution with $n-2$ degrees of freedom and $\alpha/2$ area to the right.

$$b_1 \pm \left(t^\star \times \text{s.e.}(b_1)\right)$$

The formulas for calculating the standard error of b_1 (s.e.(b_1)) are in Section 6.4 of *OpenIntro Biostatistics*. In practice, statistical software like R is used to obtain t-statistics and p-values for

linear models.

7. Carry out inference based on the linear model from Question 6.

```r
#view model summary
summary(lm(BMI ~ DM, data = prevend.samp))
```

```
##
## Call:
## lm(formula = BMI ~ DM, data = prevend.samp)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.578 -2.938 -0.800  2.058 34.262
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  26.6858     0.2096 127.341  < 2e-16 ***
## DMPresent     3.2329     0.8157   3.963 8.47e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.529 on 498 degrees of freedom
## Multiple R-squared:  0.03058,    Adjusted R-squared:  0.02863
## F-statistic: 15.71 on 1 and 498 DF,  p-value: 8.474e-05
```

   a) Conduct a formal hypothesis test of no association between BMI and diabetes presence using `prevend.samp`, at the $\alpha = 0.05$ significance level.

      i. State the hypotheses.

         The null hypothesis is that BMI and diabetes presence are not associated, $H_0 : \beta_1 = 0$. The alternative hypothesis is that BMI and diabetes presence are associated, $H_A : \beta_1 \neq 0$.

      ii. Identify the relevant t-statistic and p-value from the output of the `summary(lm( ))` function.

         The t-statistic of the slope coefficient is 3.96, with associated p-value of 8.47×10^{-5}.

      iii. State a conclusion in the context of the data.

         Since $p < \alpha$, there is sufficient evidence to reject the null hypothesis in favor of the alternative that BMI and diabetes presence are associated. From the observed slope of 3.23, there is evidence of a positive association between BMI and diabetes; presence of diabetes is associated with a higher BMI of 3.23 units, on average.

b) Calculate and interpret the 95% confidence interval for the slope coefficient of the model.

With 95% confidence, the interval (1.63, 4.84) captures the difference in average BMI between diabetic and non-diabetic individuals in the population. Note that since individuals selected for the PREVEND study were chosen on the basis of their urinary albumin excretion (which is associated with abnormalities in renal function), they are most likely not representative of the general population of adults in the Netherlands.

```
#use r as a calculator

#define parameters
b1 = coef(summary(lm(BMI ~ DM, data = prevend.samp)))[2, 1]
se.b1 = coef(summary(lm(BMI ~ DM, data = prevend.samp)))[2, 2]
n = length(prevend.samp$BMI) - sum(is.na(prevend.samp$BMI))

#calculate interval
t.star = qt(0.975, df = n - 2)
m = t.star*se.b1
b1 - m; b1 + m
```

```
## [1] 1.63025
```

```
## [1] 4.835596
```

8. Use `t.test( )` to conduct a t-test for the difference in mean BMI between diabetic and non-diabetic individuals. Compare the results of inference based on the linear model to those based on a two-group test.

The final conclusions are the same, although the p-value from the linear model is considerably smaller than the one from the two-group test. Since the estimated difference in means is the same, the different p-values arise from a difference in the associated standard error of the estimate; the standard error of the estimate is larger in the two-group test procedure. As expected, then, the confidence interval associated with the two-group test is wider than the one for the model slope coefficient.

The sign of the t-statistic from the two-group test is negative, unlike the one from the linear model. Note that this is only reversed because for the two-group test, the difference is calculated as (non-diabetic BMI - diabetic BMI); the model slope specifically quantifies the difference *from* the non-diabetic BMI and so is positive. A difference in sign does not affect the p-value.

The difference in t-statistics and p-values is due to a modeling assumption. Linear regression assumes constant variance between groups, while the two-sample t-test does not. When constant variance is assumed...

- The standard error of the difference in means is no longer $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$, as in formulas shown previously for calculating the t-statistic in an independent two-group test.
- The p-value is calculated using a distribution with degrees of freedom $n-2$, unlike for a two-group test, which uses the Satterthwaite approximation for degrees of freedom.

It is possible to conduct a two-group t-test under the assumption that variance between groups is constant, as discussed in Section 5.3.5 of *Openintro Biostatistics*. The relevant R output is shown below, and illustrates that once this assumption is in place, the t-statistic and p-value are identical to those from the linear model output.

```
#t-test without constant variance assumption (default)
t.test(BMI ~ DM, data = prevend.samp)
```

```
##
##  Welch Two Sample t-test
##
## data:  BMI by DM
## t = -3.5118, df = 35.508, p-value = 0.001232
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.100888 -1.364958
## sample estimates:
##  mean in group Absent mean in group Present
##              26.68580              29.91872
```

```
#t-test with constant variance assumption
t.test(BMI ~ DM, data = prevend.samp, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  BMI by DM
## t = -3.9633, df = 498, p-value = 8.474e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.835596 -1.630250
## sample estimates:
##  mean in group Absent mean in group Present
##              26.68580              29.91872
```

The following questions return to the investigation of RFFT score as a main response variable of interest in the PREVEND data.

9. The previous labs in this chapter have focused on exploring the association between RFFT score and age. The linear model with RFFT score as a response variable and age as a predictor was shown to reasonably satisfy the assumptions of linear regression.

   a) Briefly discuss whether there is significant evidence of an association between RFFT score and age; be sure to report the relevant numerical evidence.

   There is highly significant evidence at the $\alpha = 0.05$ level that RFFT score and age are associated; $p < 0.0001$. An increase in age of one year is associated with a decrease in average RFFT score of 1.26 points.

```
#assess significance
coef(summary(lm(RFFT ~ Age, data = prevend.samp)))
```

```
##                Estimate Std. Error    t value      Pr(>|t|)
## (Intercept) 137.549716 5.01614122  27.42142 1.406325e-101
## Age          -1.261359 0.08952532 -14.08941  3.503269e-38
```

b) Compute and interpret the 99% confidence interval for the model slope.

With 99% confidence, the interval (-1.49, -1.03) captures the average change in RFFT score associated with an increase in one year of age.

```
#compute 99% confidence interval
confint(lm(RFFT ~ Age, data = prevend.samp), level = 0.99)
```

```
##                   0.5 %    99.5 %
## (Intercept) 124.579291 150.52014
## Age          -1.492848  -1.02987
```

10. Use a linear regression model to relate BMI and gender (`Gender`).

a) Convert `Gender` to a factor variable. In the original variable, males are coded as `0` and females are coded as 1.

```
#convert Gender to a factor variable
prevend.samp$Gender <- factor(prevend.samp$Gender, levels = c(0, 1),
                              labels = c("Male", "Female"))
```

b) Fit the model and interpret the model intercept and slope.

The model intercept is 26.84; this is the estimated average BMI for males. The model slope is 0.116; thus, the estimated average BMI for females is $26.84 + 0.116 = 26.96$.

```
#fit the BMI ~ Gender model
coef(summary(lm(BMI ~ Gender, data = prevend.samp)))
```

```
##                Estimate Std. Error    t value      Pr(>|t|)
## (Intercept)  26.8444875  0.2825244 95.0165111 2.658073e-321
## GenderFemale  0.1163529  0.4121043  0.2823385  7.778013e-01
```

c) Evaluate whether gender is a significant predictor of BMI.

Since $p = 0.77$, gender is not a significant predictor of BMI at the $\alpha = 0.05$. The observed difference in average BMI between males and females may well be due to random variation.