

Linear Regression (part 1)

Ryan Miller

Linear Regression

Using ANOVA, we modeled a quantitative outcome variable Y using a single categorical variable:

$$y_i = \mu_k + \epsilon_i$$

Generally speaking, **linear regression** models a quantitative outcome using a *linear combination* of variables:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \epsilon_i$$

ANOVA as Linear Regression

To fit our ANOVA model into the linear regression framework, we need to use *dummy variables* and *reference coding*:

Y	group	Y	X1	X2
8.5	C	8.5	0	1
11.6	B	11.6	1	0
9.0	C	9.0	0	1
9.1	B	9.1	1	0
8.0	A	8.0	0	0
9.7	A	9.7	0	0

- ▶ In this example, $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$
 - ▶ For those in group A, $X_1 = 0$ and $X_2 = 0$, suggesting $\beta_0 = \mu_A$
 - ▶ For those in group B, $X_1 = 1$ and $X_2 = 0$, suggesting $\beta_1 = \mu_B - \mu_A$
 - ▶ For those in group C, $X_1 = 0$ and $X_2 = 1$, suggesting $\beta_2 = \mu_C - \mu_A$

Model Estimation

- ▶ In regression, we assume the specified model *is true* at the population level
 - ▶ The specified model is something we decide, the data are then used to *estimate* the model's *parameters*
 - ▶ We understand the model through these estimates
- ▶ The using population model: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$ results in the estimated model:

$$\hat{y}_1 = b_0 + b_1 x_{i1} + b_2 x_{i2}$$

- ▶ The *parameter estimates* are found using least squares, they are: $b_0 = 9.77$, $b_1 = 0.25$, $b_2 = 0.03$
 - ▶ In our example, $\bar{y}_A = 9.77$, $\bar{y}_B = 10.02$, $\bar{y}_C = 9.80$; this is not a coincidence

Interpreting Model Parameter Estimates

For the drug use and tailgating example, the *estimated* ANOVA model is:

$$\hat{Y} = b_0 + b_1X_{MDMA} + b_2X_{NODRUG} + b_3X_{THC}$$

The *parameter estimates* are $b_0 = 36.8$, $b_1 = -9.2$, $b_2 = 10.5$, $b_3 = 5.8$, from this we know:

1. ALC is the reference group
2. The mean following distance in the ALC group is 36.8 ft (interpretation of b_0)
3. Predicted following distances in the MDMA group are 9.2 ft *less* than the reference group of ALC users (interpretation of b_1)
4. Predicted following distances in the THC group are 5.8 ft *more* than the reference group of ALC users (interpretation of b_3)

What About Uncertainty?

- ▶ Like any estimate, the regression estimates, b_0, b_1, \dots, b_p , won't *exactly* match the population parameters, $\beta_0, \beta_1, \dots, \beta_p$
- ▶ We won't go too far into the details, but most standard software will provide confidence interval estimates for the population parameters using the t -distribution

What About Inference?

- ▶ We can also use the regression estimates, b_0, b_1, \dots, b_p , to perform hypothesis testing
- ▶ How might you describe the hypothesis $H_0 : \beta_1 = 0$? (Hint: think about reference coding and what β_1 is in terms of group means)
- ▶ $\beta_1 = 0$ implies that the mean of group 1 is no different from the mean of the reference group, essentially a two-sample t-test!

Example #1

With your group, load the College Data into Minitab and fit a linear regression using REGION to predict AVGFACSAL (Hint: you'll need to change the variable REGION to text by right clicking the column and selecting "format column"). Answer the following questions:

1. By default, which region is chosen as the reference category?
2. Which region has the highest average faculty salaries?
3. Are the population mean salaries in Region 3 (Great Lakes) significantly different from those in Region 4 (Midwest) (Hint: change the reference category using "Coding" in the regression menu)

Example #1 - Solution

1. Region 1 (Northeast) is the default reference category
2. Region 1 (Northeast) has the highest average salaries
3. No, the estimated difference of -2572 has a p -value of 0.66, there is not sufficient evidence for us to believe that colleges in these regions have different average faculty salaries

Simple Linear Regression

- ▶ These inferential methods apply to quantitative predictor variables too
 - ▶ However, our interpretations must change
- ▶ *Simple linear regression* uses a single quantitative predictor and the population level model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$$

- ▶ The Tips Data documents the tips received by a server in a suburban national chain restaurant:

Example #2

With your group, load the Tips Data into Minitab and fit a regression model that uses TotBill to predict Tip, then answer the following:

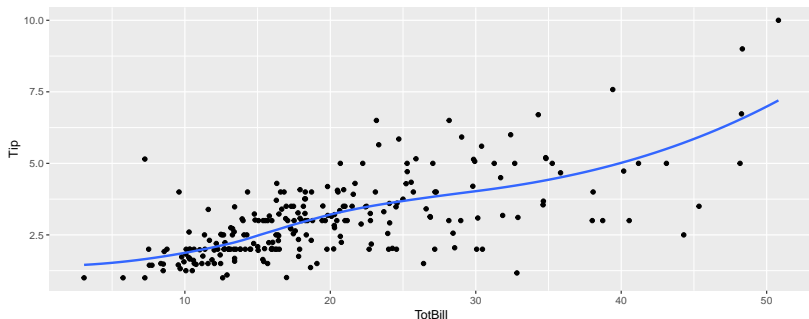
1. Interpret the estimate and 95% confidence interval for the slope coefficient of TotBill
2. From the 95% CI, it plausible that tips are unrelated to the total bill amount? Reconcile your conclusion with the hypothesis test on b_1

Example #2 - Solution

1. The estimate is $b_1 = 0.105$, indicating that each \$1 increase in the total bill on average leads to a 10.5 cent increase in the tip. The 95% confidence interval suggests we are confident the actual population level effect is between 9 cents and 12 cents (the population would be all possible tables this individual might serve)
2. No, the 95% confidence interval doesn't contain zero (similarly, the hypothesis test has a very small p -value)

Non-linear Effects

- ▶ Sometimes, the relationship between a predictor and the outcome won't be linear
- ▶ The plot below adds a *loess* smoothing line (essentially the average calculated within a moving window) applied the Tips Data



Non-linear Effects

- ▶ The loess line shows some *curvature*, but is it enough to suggest that a population model with a **quadratic effect** should be used?
 - ▶ We can compare these two models using ANOVA!
- ▶ Null model: $y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$
- ▶ Alternative model: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2 + \epsilon_i$
 - ▶ All we need is the *sum of squares* for each of these models to perform the test

Non-linear Effects - Example

For the Tips Data, use Minitab fit the following models:

1. A model which predicts Tip using a linear effect for TotBill
2. A model which predicts Tip using a quadratic effect for TotBill (Use the “Model” button, and add “Terms through order 2”)

Record the sum of squares for each model (This is record as “Error” in Minitab’s ANOVA table) and test for superiority of the quadratic model using your own ANOVA test (Hint: think about what d_0 and d_1 are here)

Non-linear Effects - Solution

1. Null model (linear) SS is 252.79 and $d_0 = 2$ (intercept and slope)
 2. Alternative model (quadratic) SS is also 252.78 and $d_1 = 3$ (intercept, b_1 , and b_2)
- ▶ The F -statistic is $\frac{(252.79 - 252.78)/(3 - 2)}{252.78/(244 - 3)} = 0.01$ and the p -value is 0.93
- ▶ Notice that this F -statistic and p -value are provided in Minitab's ANOVA table in the row "TotBill*TotBill"

Loose-ends - R-Squared

- ▶ In Chapter 2, we learned about the coefficient of variation or R^2
- ▶ We can express R^2 using sums of squares:

$$R^2 = SSE/SST$$

- ▶ Note: In calculating R^2 , SST refers to the null model that predicts each observation as the mean \bar{y} . This is in contrast to some situations (like the last example) where we consider another model to be the “null model”

Loose-ends - Residuals and Nesting

- ▶ Linear regression models assume ϵ 's (population level deviations from the model) are normally distributed with a mean of zero
 - ▶ We can check this assumption using the residuals
- ▶ ANOVA can only be used to compare *nested models*
 - ▶ $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2 + \epsilon_i$ and $y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$ are nested because forcing $\beta_2 = 0$ makes them identical

Conclusion

These notes cover Ch 9 of the textbook. Right now, you should. . .

1. Know the relationship between one-way ANOVA and linear regression
2. Understand how to perform on statistical inference on the parameters of linear regression model
3. Know how to compare two nested models using ANOVA

I encourage you to read Ch 9 of the book and its examples.