Inference for Linear Regression Models

Part 1 - F-tests

Ryan Miller

Grinnell College
Statistics

Introduction

We recently learned about *one-way ANOVA*, a statistical test for scenarios involving a nominal categorical explanatory variable and a quantitative outcome. The alternative model in one-way ANOVA can be expressed as a regression model:

Group	one-way ANOVA	Regression
1	$y_i = \mu_1 + \epsilon_i$	
2	$y_i = \mu_2 + \epsilon_i$	$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$
3	$y_i = \mu_3 + \epsilon_i$	

X_1 and X_2 are dummy variables (0's and 1's), such that:

- $\mu_1 = \beta_0$ (the reference group when both X_1 and X_2 are zero)
- $\mu_2 = \beta_0 + \beta_1$ (when $X_1 = 1$)
- $\mu_3 = \beta_0 + \beta_2$ (when $X_2 = 1$)

The F-test for Regression

Similarly, the null model in one-way ANOVA can also be expressed as an *intercept-only* regression model:

Group	one-way ANOVA	Regression
All	$y_i = \mu + \epsilon_i$	$y_i = \beta_0 + \epsilon_i$

- Thus, the F-test performed in one-way ANOVA is actually comparison of *nested regression models*
  - Two models are nested when one model, known as the *reduced model*, is a special case of a larger model
  - $y_i = \beta_0 + \epsilon_i$ is a special case of the larger model: $y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$ that occurs when $\beta_1 = 0$ and $\beta_2 = 0$

Grinnell College
Statistics

The F-test for Regression

In regression, the F-test's null hypothesis is that additional predictors in the larger model provide no improvement over the predictors already present in the reduced model, or in statistical symbols:

$$H_0 : \beta_j = \beta_{j+1} = \beta_{j+2} = \ldots = \beta_p = 0$$

- $\{\beta_j, \beta_{j+1}, \ldots, \beta_p\}$ are the respective coefficients of all variables included in the larger, alternative model that are *not present* in reduced model.
  - Thus, *both models* contain $\{\beta_1, \beta_2, \ldots, \beta_{j-1}\}$ and we don't learn anything about the corresponding variables from the test

The F-test for Regression

The F-test assesses whether the *sum of squared* residuals decreases by more than would be expected by chance when additional predictors are included in a regression model:

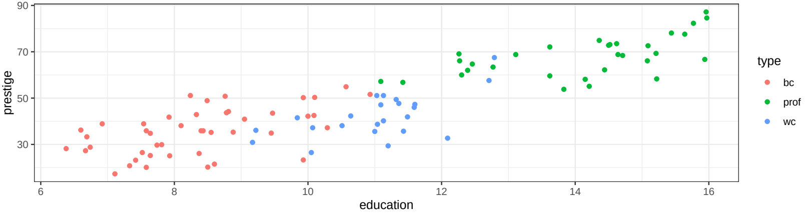
$$F = \frac{(SS_0 - SS_1)/(d_1 - d_0)}{SS_1/(n - d_1)}$$

- Here SS_0 is the sum of squares of the *reduced model*, which the smaller one that represents the null hypothesis
- SS_1 is the sum of squares of the *larger model*, which represents the alternative hypothesis
- d_0 is the number of parameters in the reduced model, while d_1 is the number of parameters in the larger model

Example - Modeling Occupational Prestige

A study collected data on $n = 98$ occupations. We'll consider the variables:

- **prestige**: the average prestige rating of the job (from 0 to 100)
- **education**: the average number of years of schooling for people holding the job
- **type**: the type of job, either skilled professional (prof), blue collar (bc), or white collar (wc)



Grinnell College
Statistics

Example (cont.)

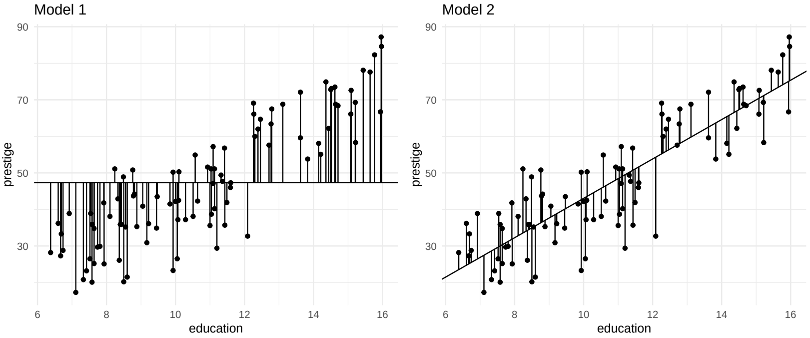
In this application we'll consider the following models:

- Model 1: `prestige ~ 1`
  - The intercept-only model $y_i = \beta_0 + \epsilon_i$
- Model 2: `prestige ~ education`
  - An alternative model $y_i = \beta_0 + \beta_1 \text{Educ}_i + \epsilon_i$
- Model 3: `prestige ~ education + type`
  - An even larger model $y_i = \beta_0 + \beta_1 \text{Educ}_i + \beta_2(\text{type} = \text{``prof"}) + \beta_3(\text{type} = \text{``wc"}) + \epsilon_i$

Questions: Is model 1 nested within model 3? Is model 2 nested within model 3? Why?

Example (cont.)

The F-test compares models using their *sums of squares* (squared residuals):



Recall: $SS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ and $F = \frac{(SS_0 - SS_1)/(d_1 - d_0)}{SS_1/(n - d_1)}$

Grinnell College
Statistics

Example (cont.)

Below we use `anova()` to evaluate whether Model 2 is superior to Model 1:

```
## Analysis of Variance Table
##
## Model 1: prestige ~ 1
## Model 2: prestige ~ education
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     97 28346.9
## 2     96  7064.4  1     21283 289.21 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

The p-value of the F-test is very small, thereby providing strong evidence that Model 2 better fits the data

Example (cont.)

We can repeat the same process to evaluate whether Model 3 is superior to Model 2:

```
## Analysis of Variance Table
##
## Model 1: prestige ~ education
## Model 2: prestige ~ education + type
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     96 7064.4
## 2     94 5740.0  2    1324.4 10.844 5.787e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

Here we conclude that Model 3 is indeed superior to Model 2

Example (cont.)

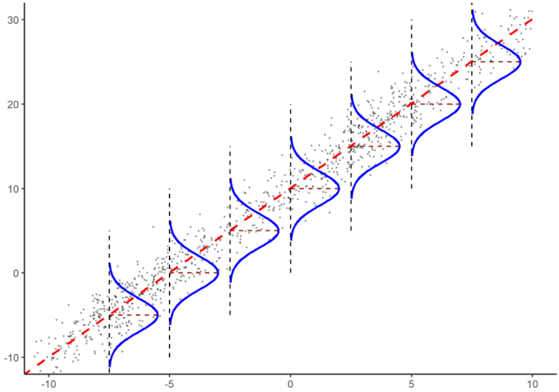
What if we add another variable, `RX`, consisting of randomly generated values with no relation to the response?

```
## Analysis of Variance Table
##
## Model 1: prestige ~ education + type
## Model 2: prestige ~ education + type + RX
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     94 5740.0
## 2     93 5671.5  1    68.574 1.1245 0.2917
```

The F-test indicates a lack of evidence that this variable improves the model, so we should omit it as the complexity it adds to the model isn't warranted.

F-test Assumptions

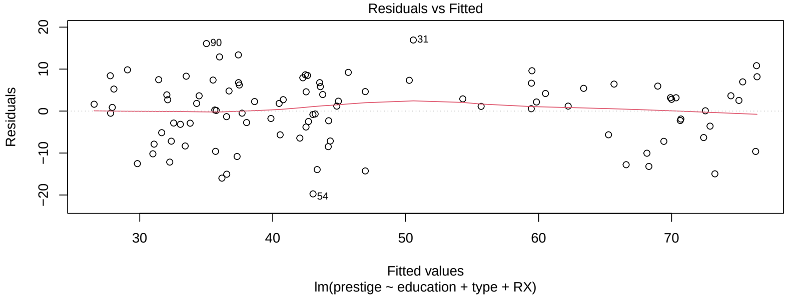
The primary assumption of the F-test for regression is that alternative model's errors are independent and follow identical Normal distributions



Grinnell College
Statistics

F-test Assumptions

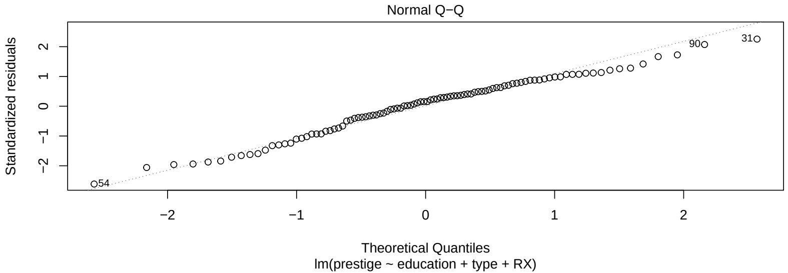
We can assess **independence** by graphing the residuals vs. the model's predictions:



If errors are independent, we expect there to be *no pattern*, since an error of a given magnitude is equally likely to occur anywhere

F-test Assumptions

We can assess **Normality** using a Q-Q plot:

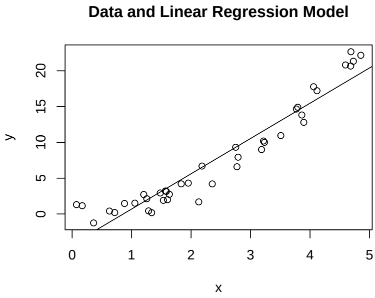


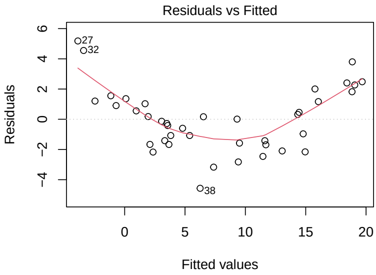
If errors are Normally distributed, we expect the standardized residuals (Z-scores of the observed residuals) to match the Z-scores for those observation's percentiles in a Normal distribution, leading to a 45-degree line in the Q-Q plot

Grinnell College
Statistics

Assumptions and Lack of Fit

Checking these assumptions can also help us determine if we're using an inappropriate model for our data:





A U-shaped pattern in the residuals happens when the model uses a straight line to represent a quadratic relationship

Handling Violated Assumptions

If one or more of the assumptions of our linear regression model are not met, any p-values we calculate might not be accurate. There are a variety of proposed solutions, we'll focus on the following:

1. Transforming the outcome variable using logarithms
2. Improving the fit of the model by including omitted variables or changing the functional form of the included variables using polynomials
3. Reporting the results with caution