Inference for Regression

Ryan Miller

Grinnell College
Statistics

Introduction

Last week we learned about *one-way ANOVA*, which involves the *statistical model*:

$$y_i = \mu_i + \epsilon_i$$

This model is equivalent to linear regression with a single categorical predictor.

Introduction (cont.)

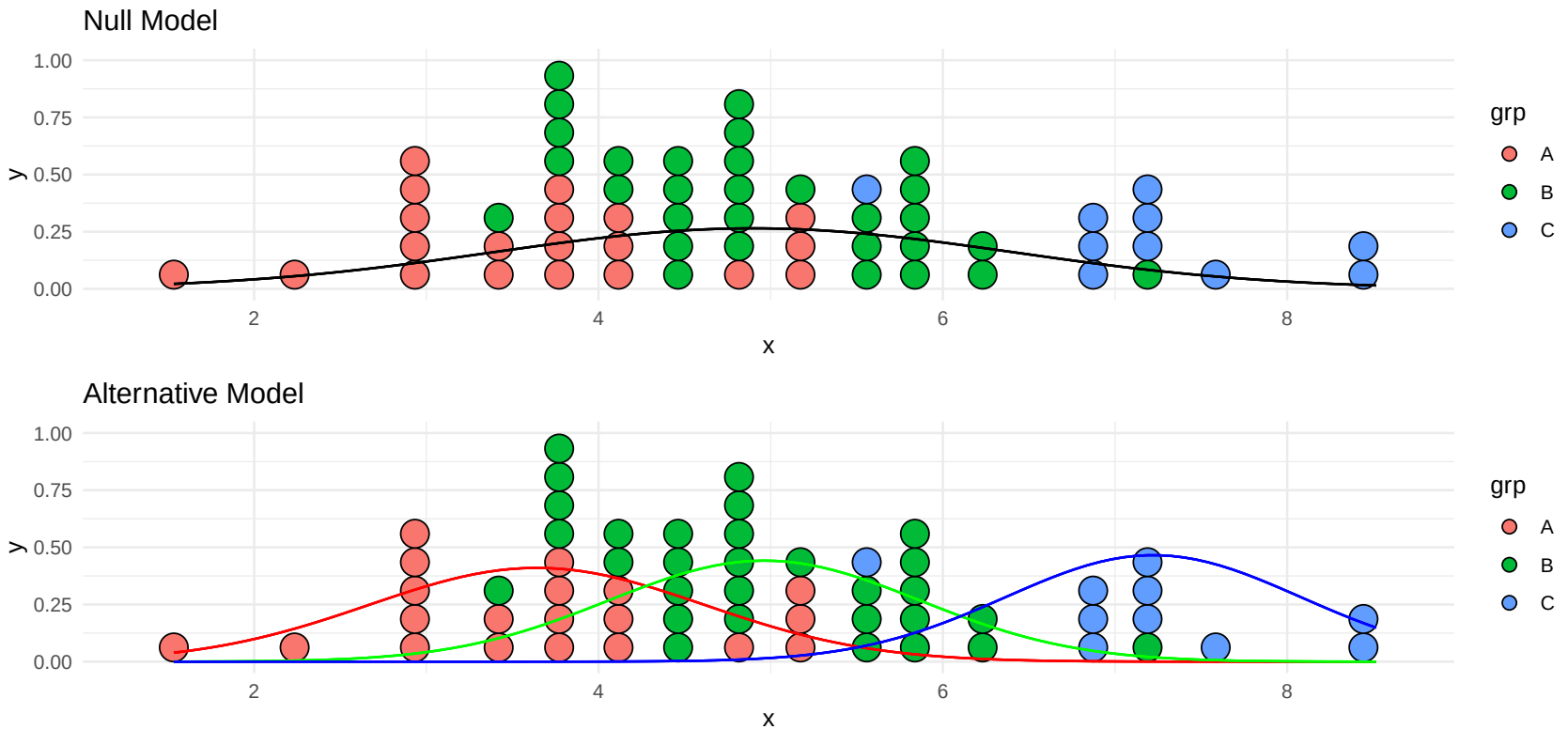
In the context of one-way ANOVA we also discussed two types of hypothesis tests:

1. Global tests - $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$
2. Pairwise tests - $H_0 : \mu_1 = \mu_2$, $H_0 : \mu_1 = \mu_3$, etc.

Our focus today will be extending these ideas to other linear regression models (ie: those with quantitative predictors, multiple predictors, etc.)

Introduction (cont.)

The global hypothesis test in one-way ANOVA compares two models for the data:



Introduction (cont.)

- If the alternative model was superior to the null model, it's sum of squared residuals (SSE) will be significantly smaller than the sum of squared residuals of the null model (SST)
  - In other words, the F-test in ANOVA is just a comparison of SSE and SST for two models
  - We can use this F-test for any two models that are *nested*, meaning the smaller model is a special case of the more complex model
- For example, we can consider the following two models:

$$\text{Model 1}: y_i = b_0 + b_2 * x_i + \epsilon_i$$
$$\text{Model 2}: y_i = b_0 + \epsilon_i$$

Model 2 is a special case of Model 1 where $b_2 = 0$, thus these models are nested.

F-tests for Linear Regression

In the examples that follow we'll consider data from a study of occupational prestige involving $n = 98$ job categories. We'll use the variables:

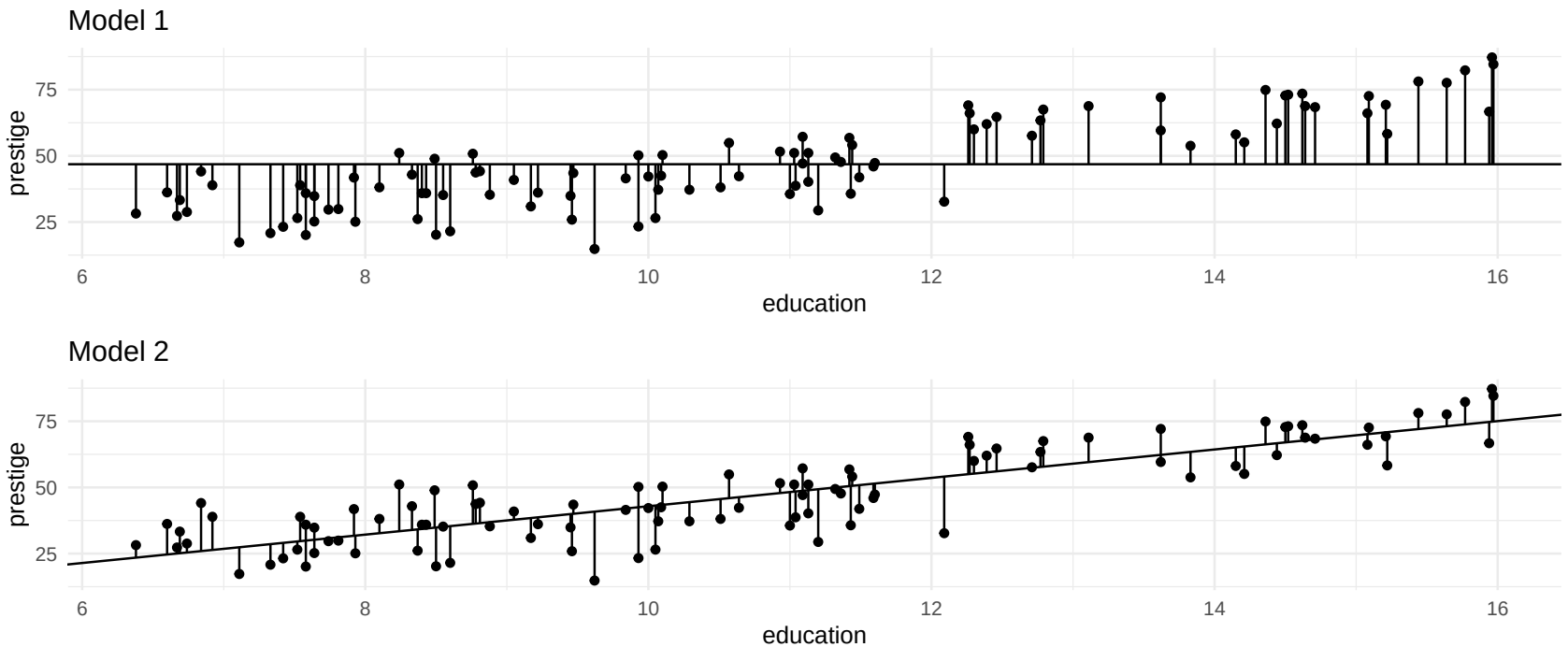
- **prestige**: the average prestige rating of the job (from 0 to 100)
- **education**: the average number of years of schooling for people holding the job
- **type**: the type of job, either skilled professional (prof), blue collar (bc), or white collar (wc)

We'll begin by comparing:

- Model 1: `prestige ~ 1` (null model of $y_i = \mu + \epsilon_i$)
- Model 2: `prestige ~ education` (alternative model of $y_i = b_0 + b_1 * \text{education} + \epsilon_i$)

F-tests for Linear Regression

Here's what the F-test comparing these models is based upon:

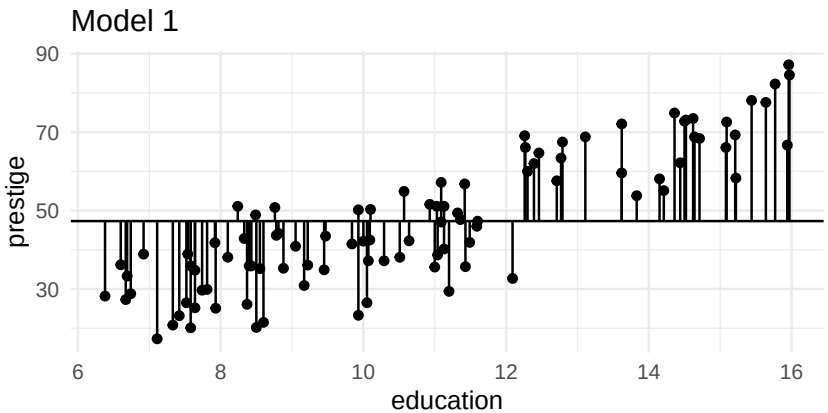


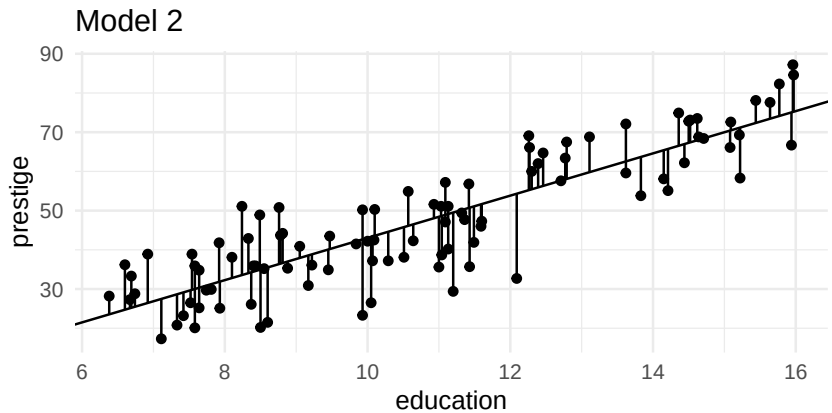
Remember that each sum of squares is defined: $SS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$,

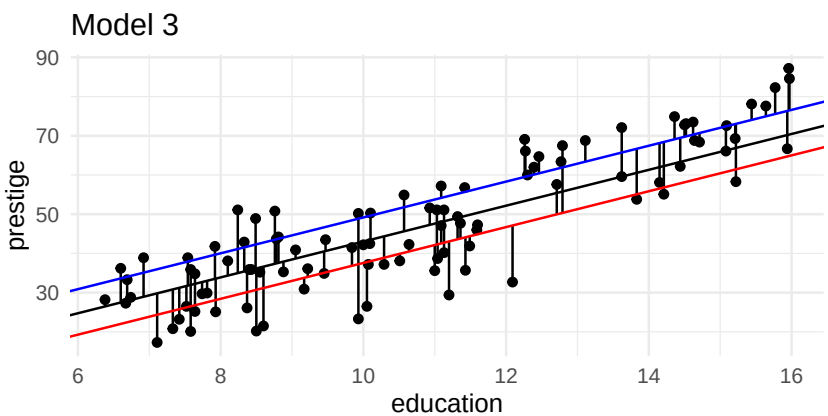
and $F = \dfrac{(SST - SSE)/(d_1 - d_0)}{SE}$

F-tests for Linear Regression

We could also compare an even larger model, Model 3, which includes one categorical and one quantitative predictor vs. Model 1 or vs. Model 2







F-tests for Linear Regression

We'll exclusively use R for F-tests involving regression models:

```
mod1 = lm(prestige ~ 1, data = df)
mod2 = lm(prestige ~ education, data = df)

anova(mod1, mod2)
```

```
## 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 ' ' 1
```

We reject the null hypothesis of "no difference between models" and conclude that there is overwhelming evidence ($p < 0.0001$) that Model 2 is a better fit for these data.

F-tests for Linear Regression

Here's the comparison between Model 3 and Model 2:

```
mod2 = lm(prestige ~ education, data = df)
mod3 = lm(prestige ~ education + type, data = df)

anova(mod2, mod3)
```

```
## 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 ' ' 1
```

We see that Model 3 is an even better fit for the data than Model 2.

F-test for Linear Regression

For illustrative purposes, let's add another explanatory variable that's just a bunch of random values (the output of `rnorm()`):

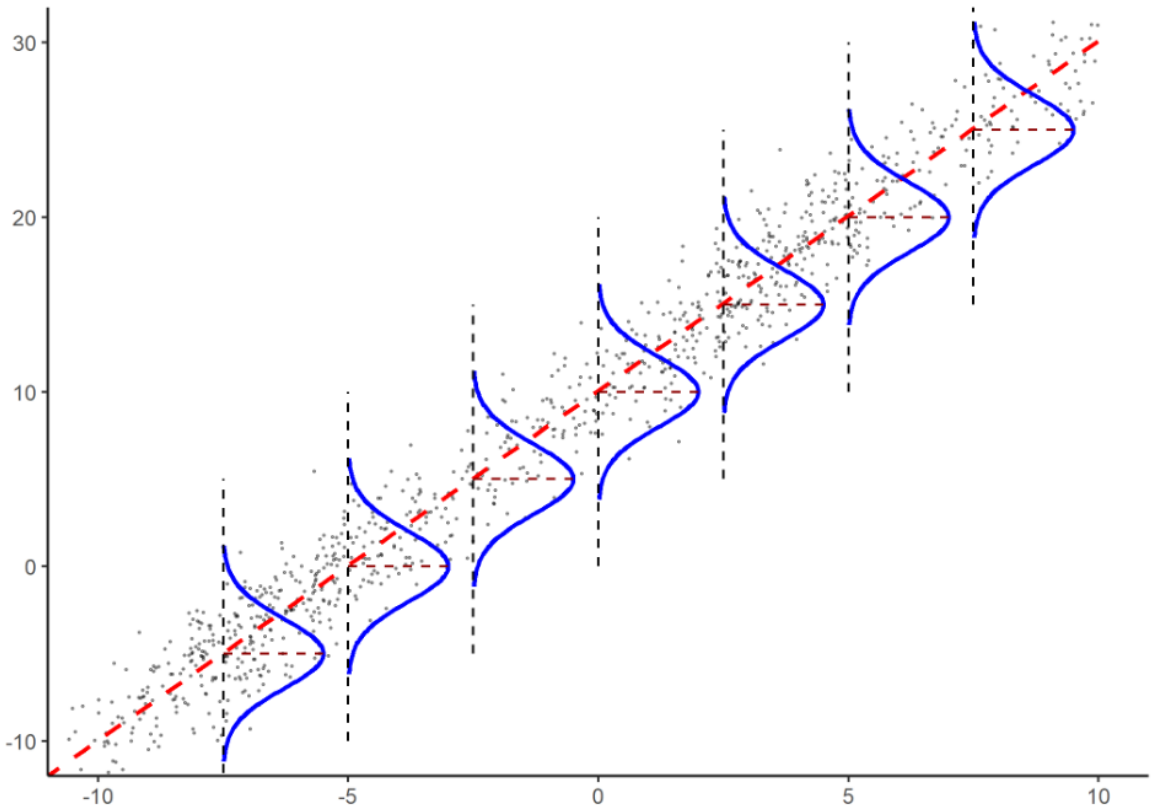
```
df <- df %>% mutate(RX = rnorm(nrow(df)))
mod4 = lm(prestige ~ education + type + RX, data = df)

anova(mod3, mod4)
```

```
## 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 5739.1  1   0.89813 0.0146 0.9042
```

As we'd expect, the sum of squares drops slightly, but the p-value is very high. So, we conclude there's no evidence that this larger model, `mod4`, is better.

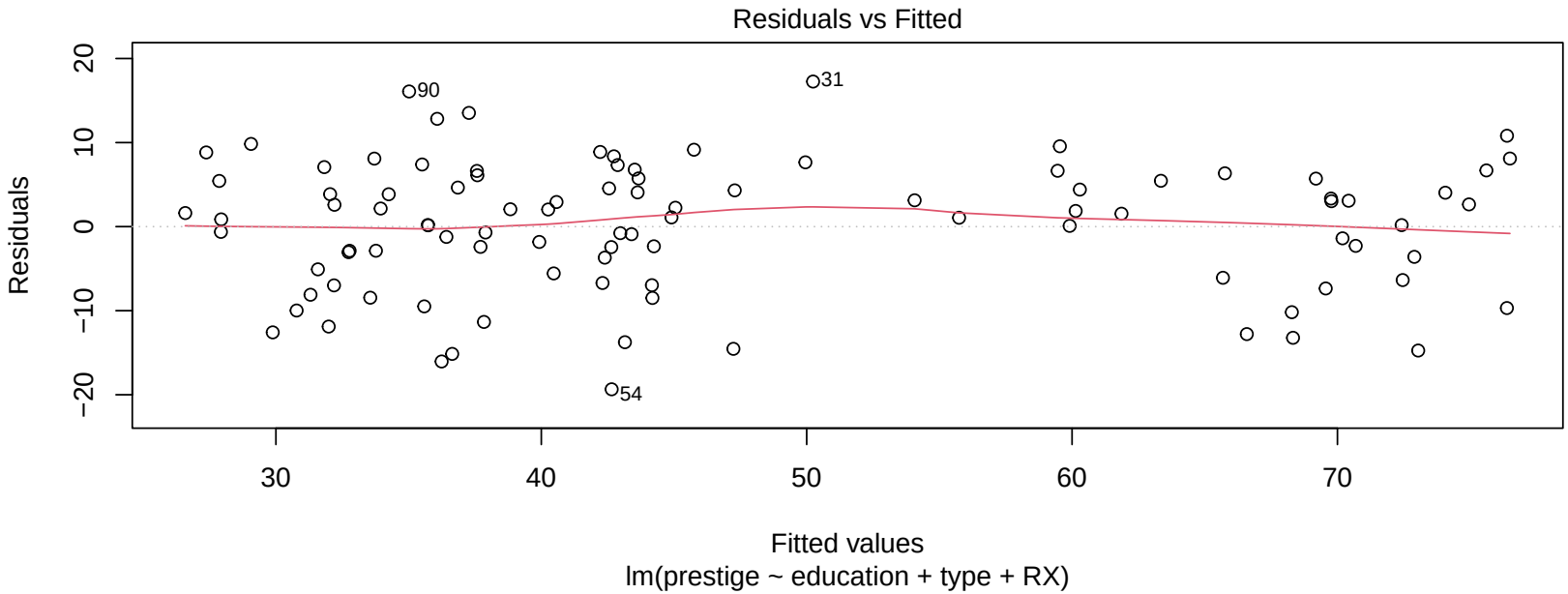
F-test Assumptions

The primary assumptions of the F-test for comparing nested regression models is that the errors of the larger model are *independent* and *Normally distributed*:



F-test Assumptions

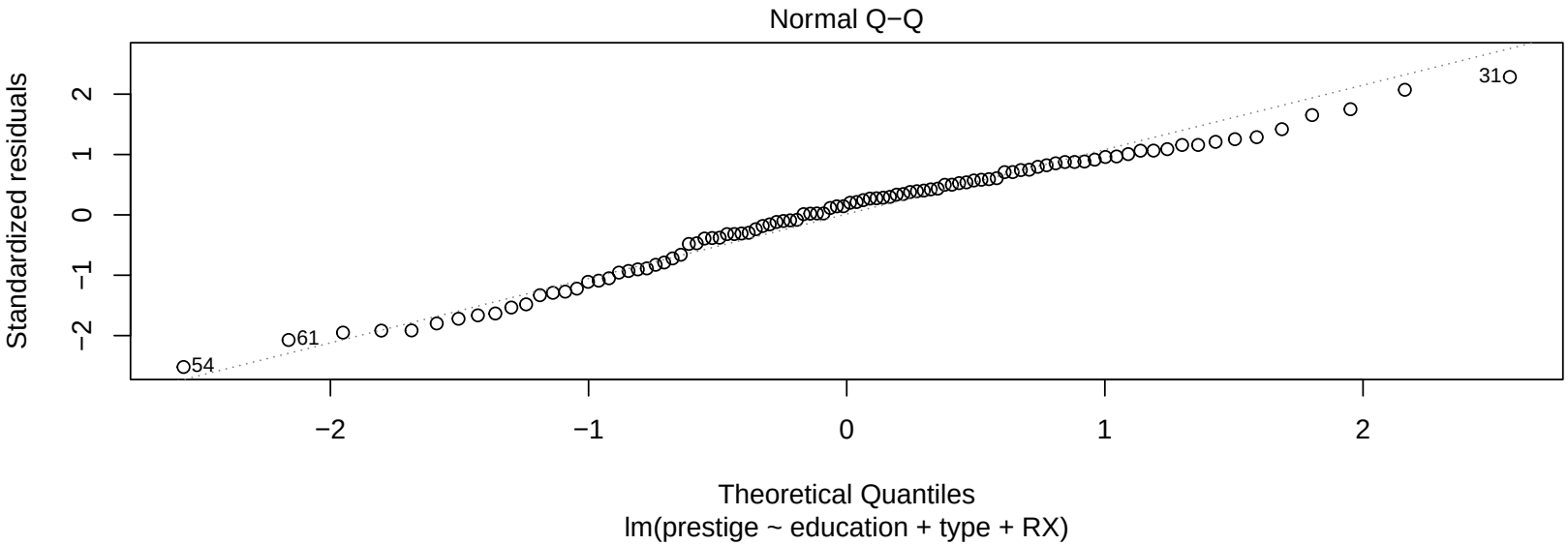
We can assess **independence** by graphing the residuals vs. the model's predictions:



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

F-test Assumptions

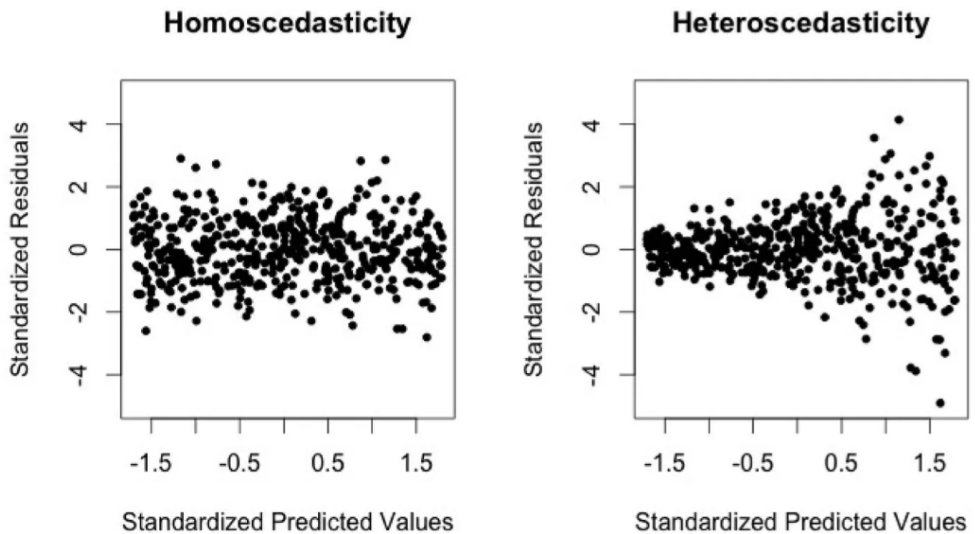
We can assess **Normality** using a quantile-quantitle, or QQ-plot:



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 QQ-plot

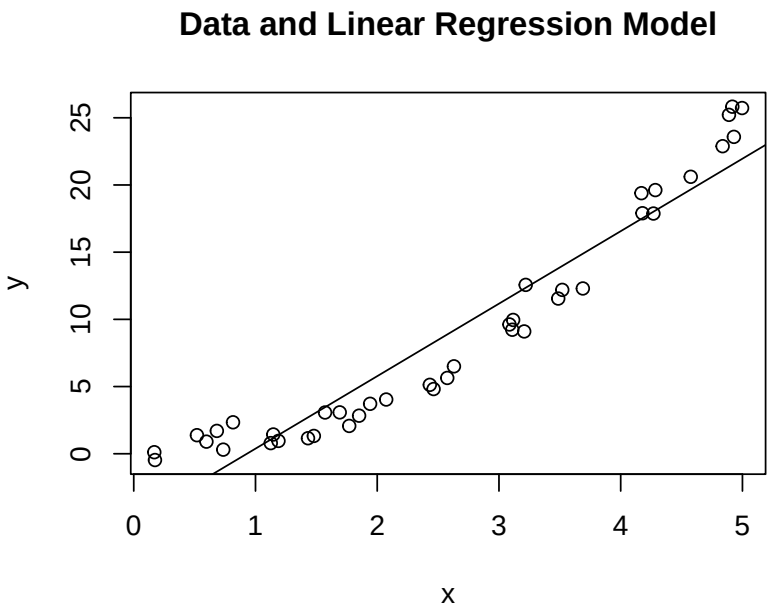
F-test Assumptions

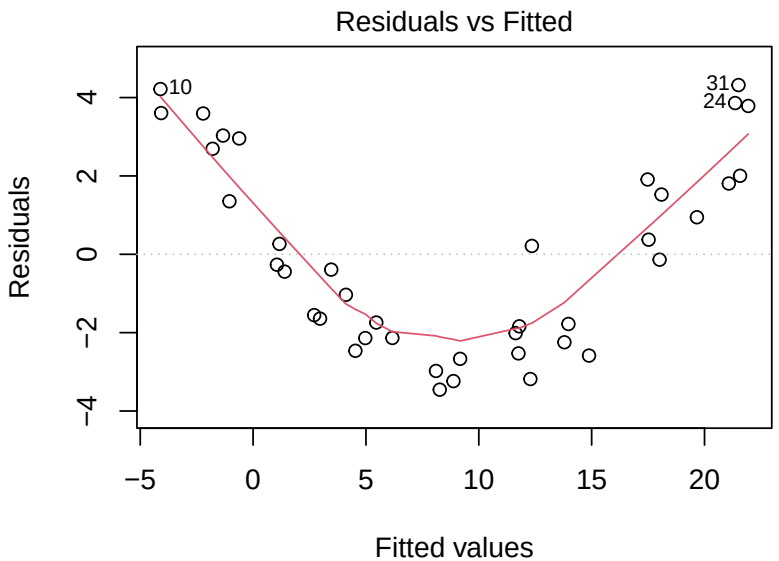
- The residuals not only need to follow a Normal distribution, but must follow the *same* Normal curve everywhere, meaning they need to exhibit the same amount of variability
  - The term **heteroscedasticity** describes scenarios when the variability of the residuals differs throughout the model



F-test Assumptions

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





We see a pattern in the residuals because this model tries to use a straight line to represent a quadratic relationship

F-test 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 (covered in today's lab)
2. Improving the fit of the model by including omitted variables or changing the functional form of the included variables using polynomials
3. Reporting our results with caution

t-tests for Regression Coefficients

- Regression is commonly used to isolate the effect of an explanatory variable on the outcome after adjusting for other factors
  - It's possible to use a t-test to evaluate $H_0 : b_j = 0$, or the null hypothesis that variable j has no effect on the outcome (after adjusting for everything else in the model)

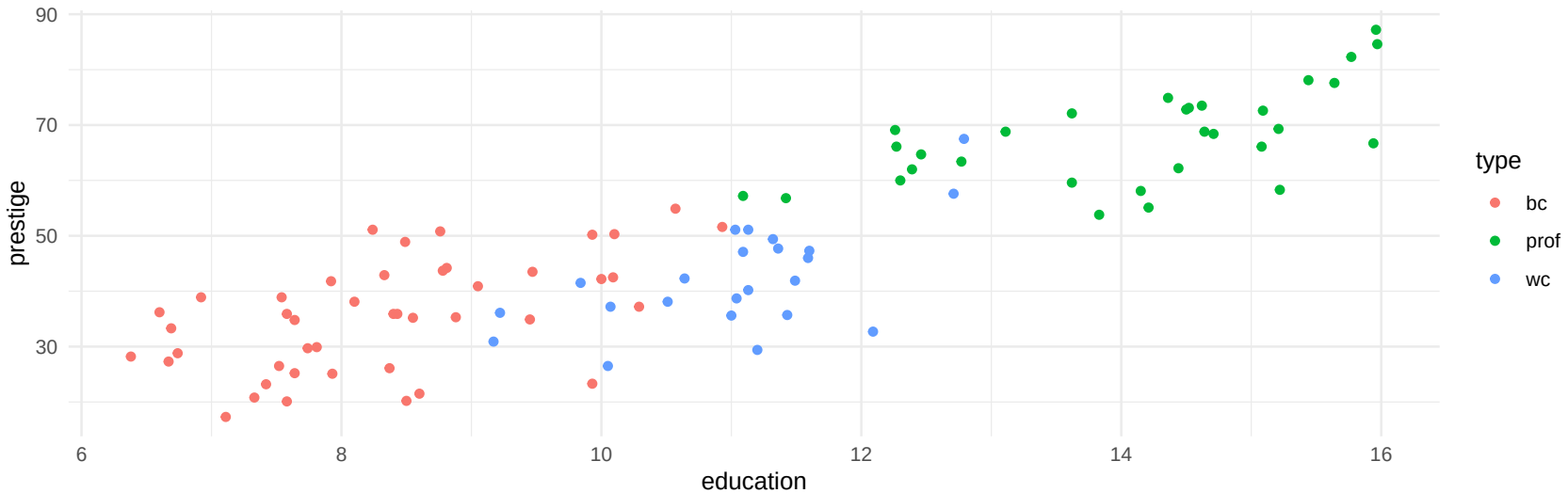
```
mod3 = lm(prestige ~ education + type, data = df)
summary(mod3)$coefficients
```

```
##              Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) -2.698159   5.736093 -0.4703828 6.391712e-01
## education    4.572793   0.671564  6.8091697 9.159877e-10
## typeprof     6.142444   4.258961  1.4422401 1.525583e-01
## typewc      -5.458495   2.690667 -2.0286769 4.532001e-02
```

- After adjusting for education, our model suggests that white collar jobs have significantly less prestige than blue collar jobs

Hidden Extrapolation

A final cause for concern when interpreting adjusted effects is *hidden extrapolation*, or making conclusions about segments of the population that don't exist.



We may want to avoid comparisons of `prof` and `bc` jobs adjusted to have the same level of education because no such jobs exist in our data.