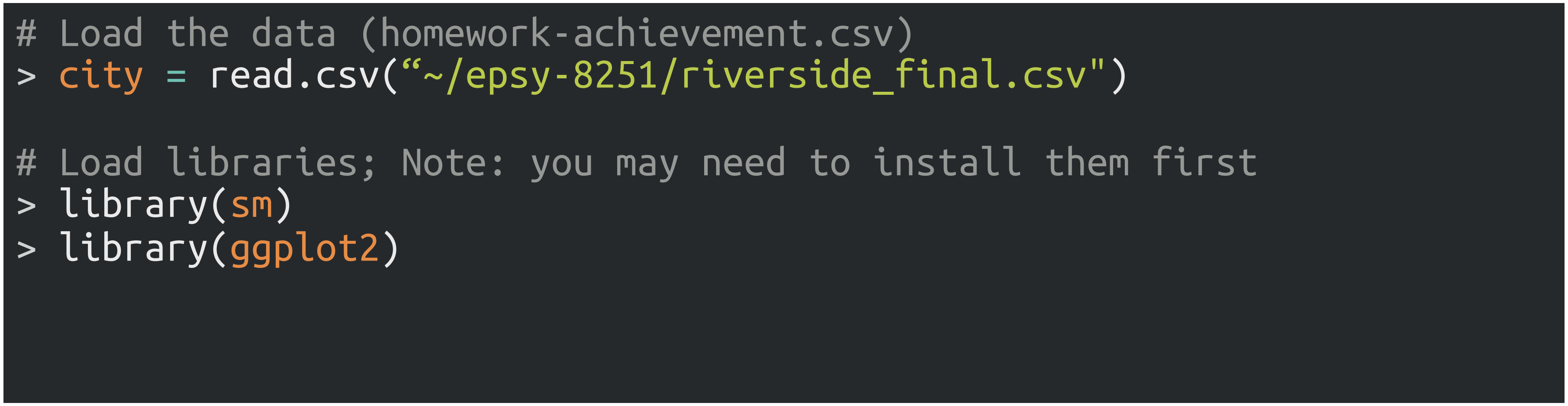
Regression Assumptions

Andrew Zieffler

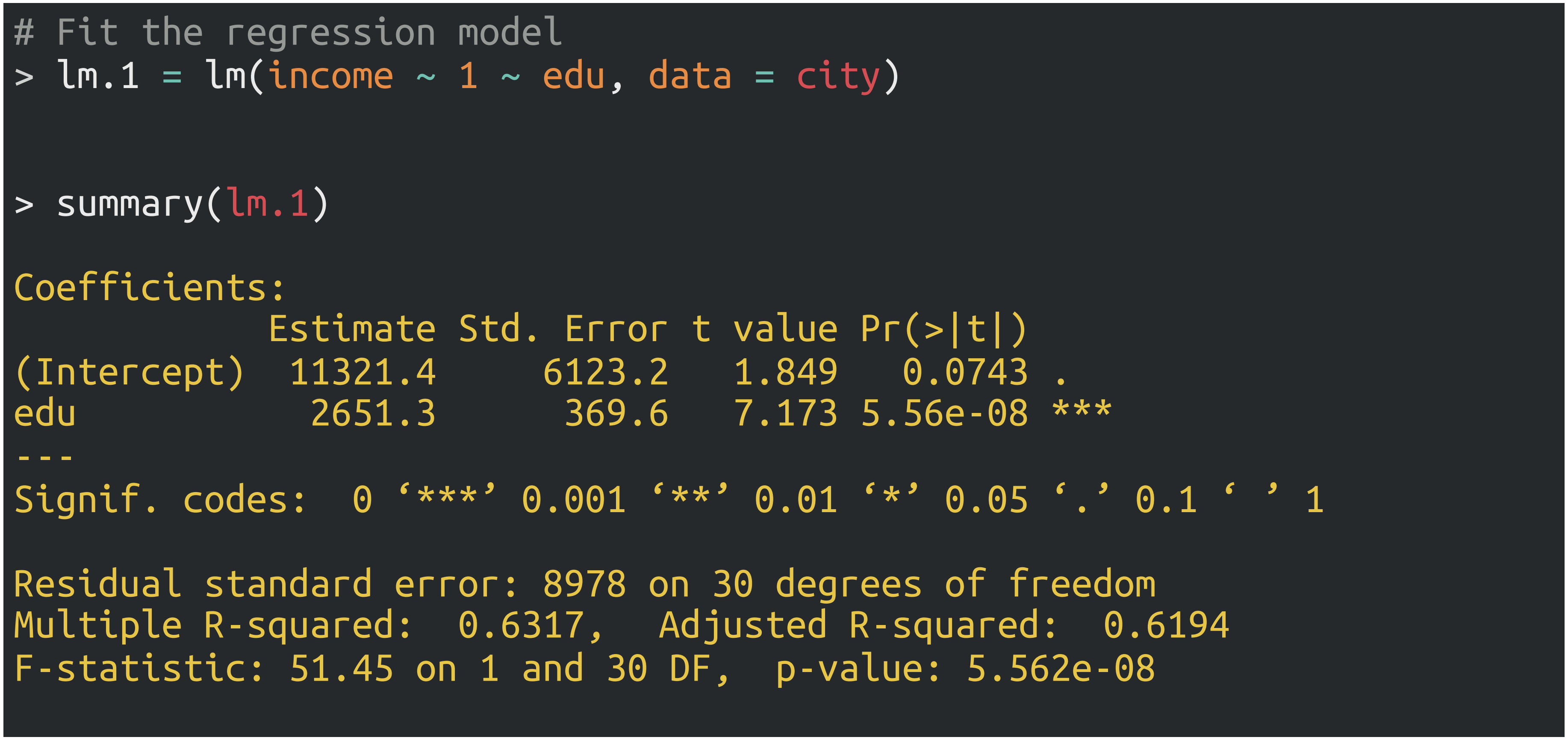
Prepare



```
# Load the data (homework-achievement.csv)
> city = read.csv(“~/epsy-8251/riverside_final.csv")

# Load libraries; Note: you may need to install them first
> library(sm)
> library(ggplot2)
```

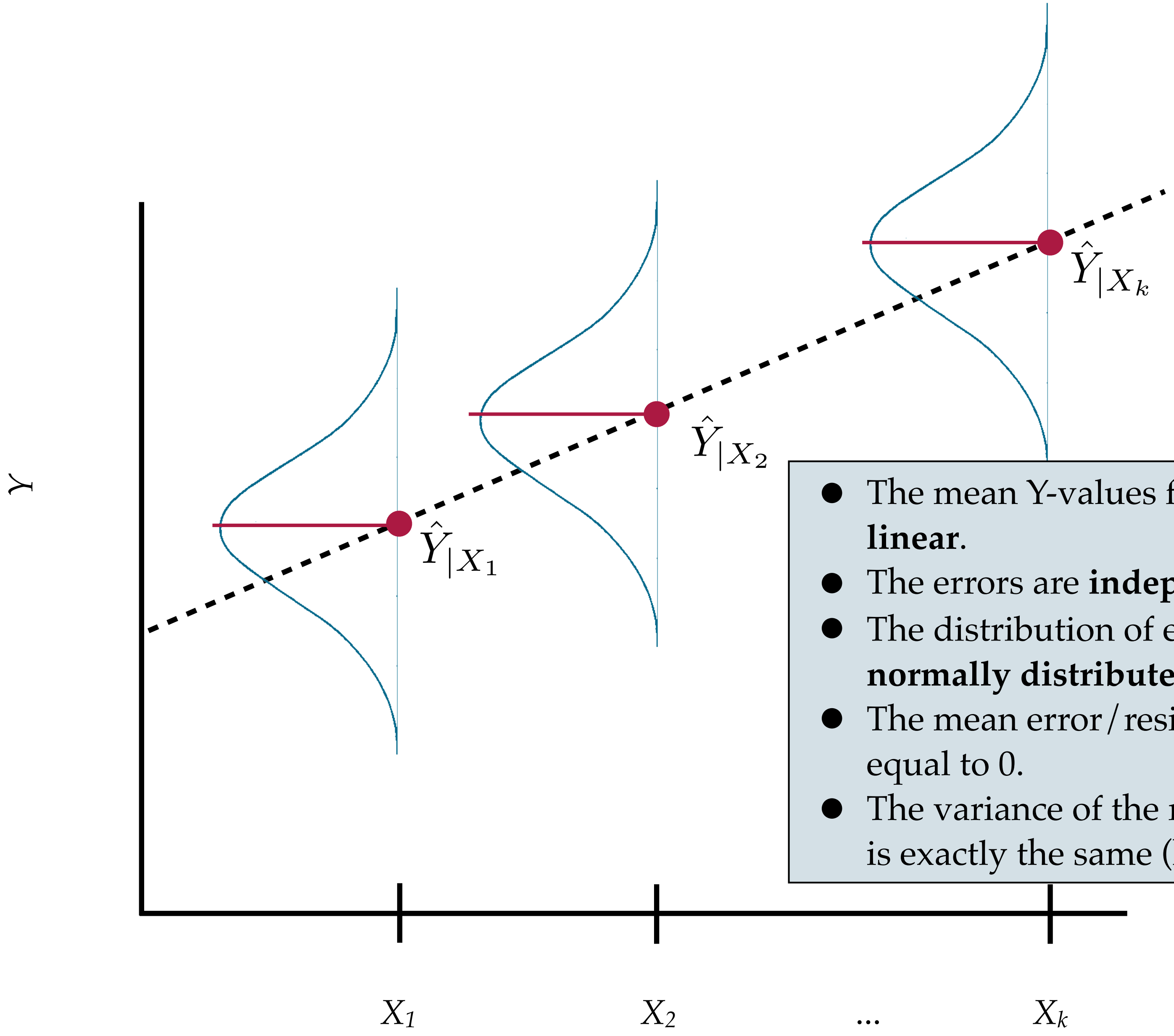
Fit the Regression Model and Examine the Output

Use the summary() function to display the fitted regression coefficients and their standard errors.



```
# Fit the regression model
> lm.1 = lm(income ~ 1 ~ edu, data = city)


> summary(lm.1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11321.4     6123.2   1.849   0.0743 .
edu           2651.3      369.6   7.173 5.56e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8978 on 30 degrees of freedom
Multiple R-squared:  0.6317,   Adjusted R-squared:  0.6194
F-statistic: 51.45 on 1 and 30 DF,  p-value: 5.562e-08
```

Regression Assumptions



- The mean Y-values from each X are **linear**.
- The errors are **independent.**
- The distribution of errors at each X is **normally distributed**.
- The mean error/residual at each X is equal to 0.
- The variance of the residuals at each X is exactly the same (homoskedasticity)

Two important caveats:

1. The assumptions are about the **distribution of errors at each level of *X***.
2. The assumptions refer to the the distribution of errors in the **population**.

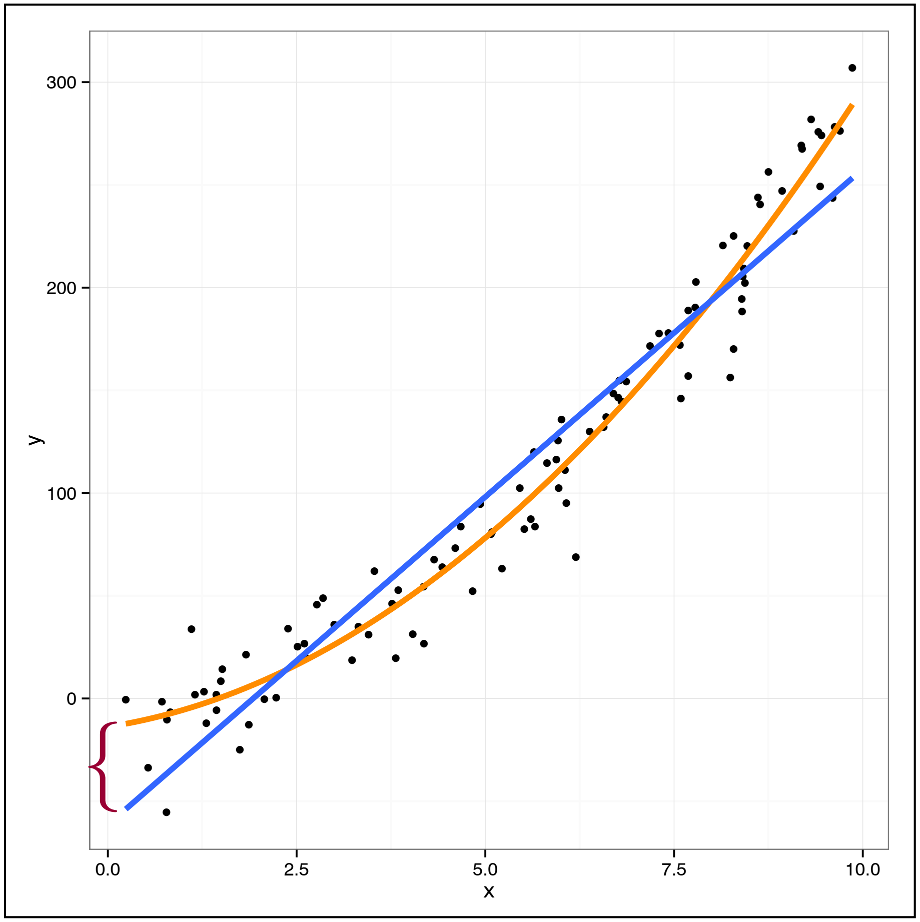
Recall that the sample e_i are approximations of the ε_i

Examining the e_i gives a good indication of how the ε_i behave...but remember that sample data can deviate from what would be expected because they are a sample.

Assumption: The mean Y-values at each X are linear.

This is an assumption that allows us to specify the structural part of the model. This assumption can be evaluated **theoretically** (literature supporting a linear relationship between *X* and *Y*) or **empirically**, by examining scatterplots of the outcome vs. predictor.

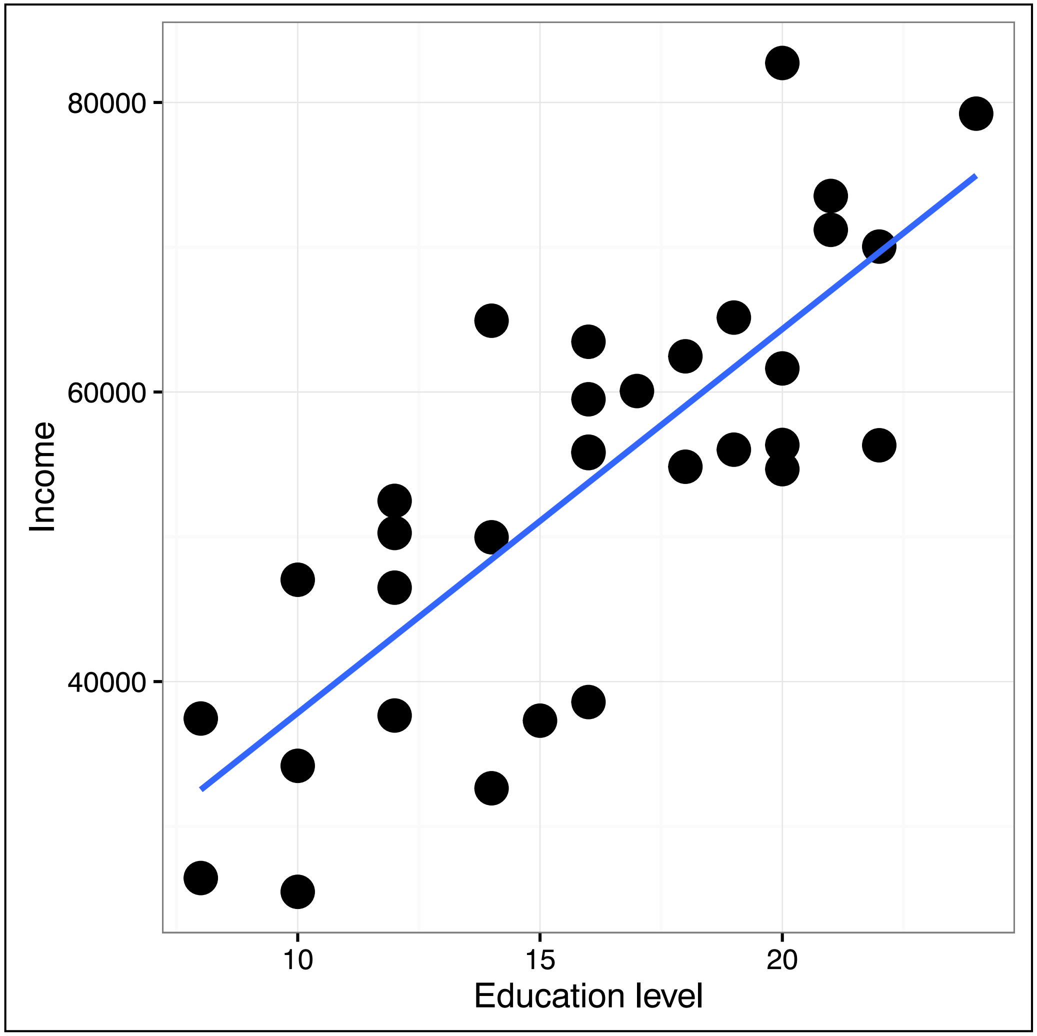
Fitting a linear model when the TRUE relationship is non-linear may, or may not be problematic.



- Coefficients may be wrong
- Predictions may be wrong, especially at the extreme values for *X*
- Mis-specified models lead to misinformed understandings of the world.

Notice that when we fit a linear model to non-linear data that the line is consistently above, or below, the data at different *X*-values. This would be evidence that we did not meet the linearity assumption.

Evaluation: The mean Y-values at each X are linear.



Plotting Y vs. X using a scatterplot can give us an initial look at the linearity assumption.

The line seems to fit the Goldilocks principle that at most values of X, roughly half of the points are above the line, and half are below.

The linearity assumption seems satisfied for these data. However, later, we will double-check this.

Assumption: The Errors (in the Population) are Independent

The definition of independence relies on formal mathematics. Loosely speaking **a set of observations is independent if knowing that one observation is above or below its mean value conveys no information about whether any other observation is above or below its mean value**. If observations are not independent, we say they are dependent or correlated.

Using a **random chance** in the study (to either select observations or assign them to levels of the predictor) will guarantee independence of the observations.

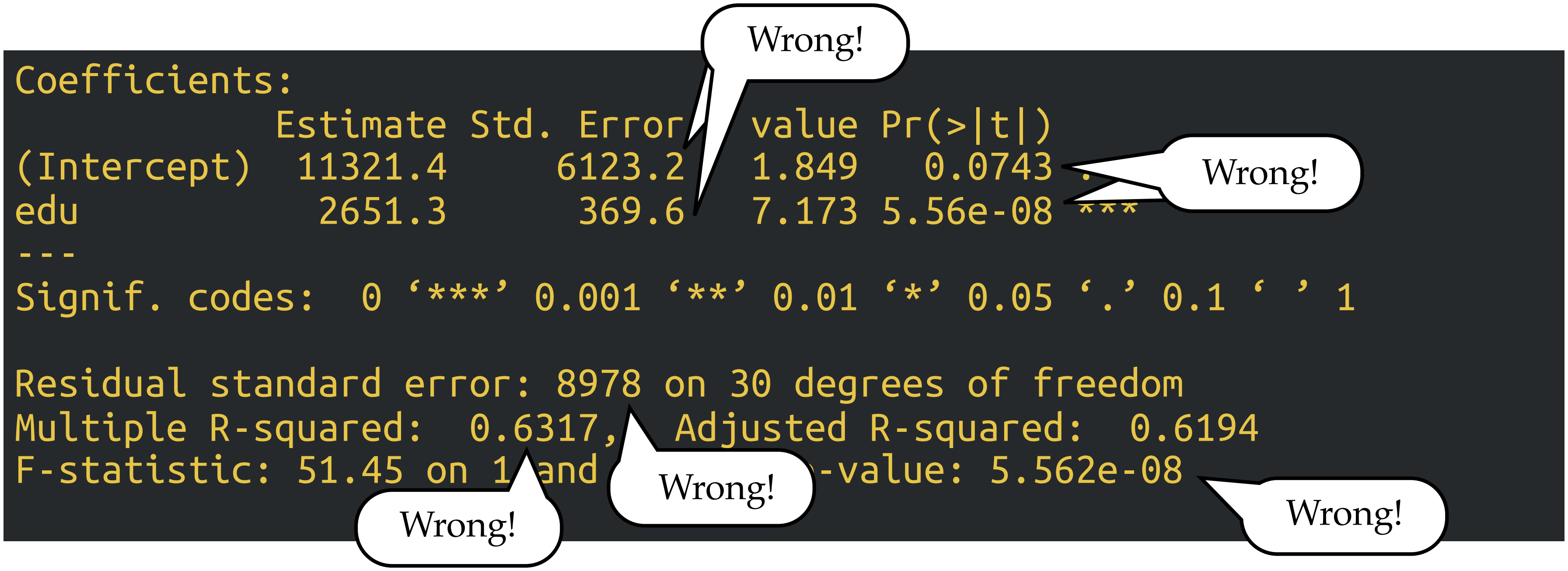
Assessing the independence assumption is primarily a logical argument.

Aspects of data collection and analysis that **violate** independence:

- Physical (spatial) proximity in the collection of observations (e.g., convenience sampling based on location)
- Observations collected longitudinally (especially when they are the same subjects' data collected repeatedly)
- Analysis: When the level of assignment does not correspond to the level of analysis.

What Happens if the Independence Assumption is Violated?

Violation of the independence assumption is a BIG problem.



```
Coefficients:
            Estimate Std. Error   value Pr(>|t|)
(Intercept)  11321.4     6123.2   1.849   0.0743 .
edu           2651.3      369.6   7.173 5.56e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8978 on 30 degrees of freedom
Multiple R-squared:  0.6317,  Adjusted R-squared:  0.6194
F-statistic: 51.45 on 1 and             -value: 5.562e-08
```

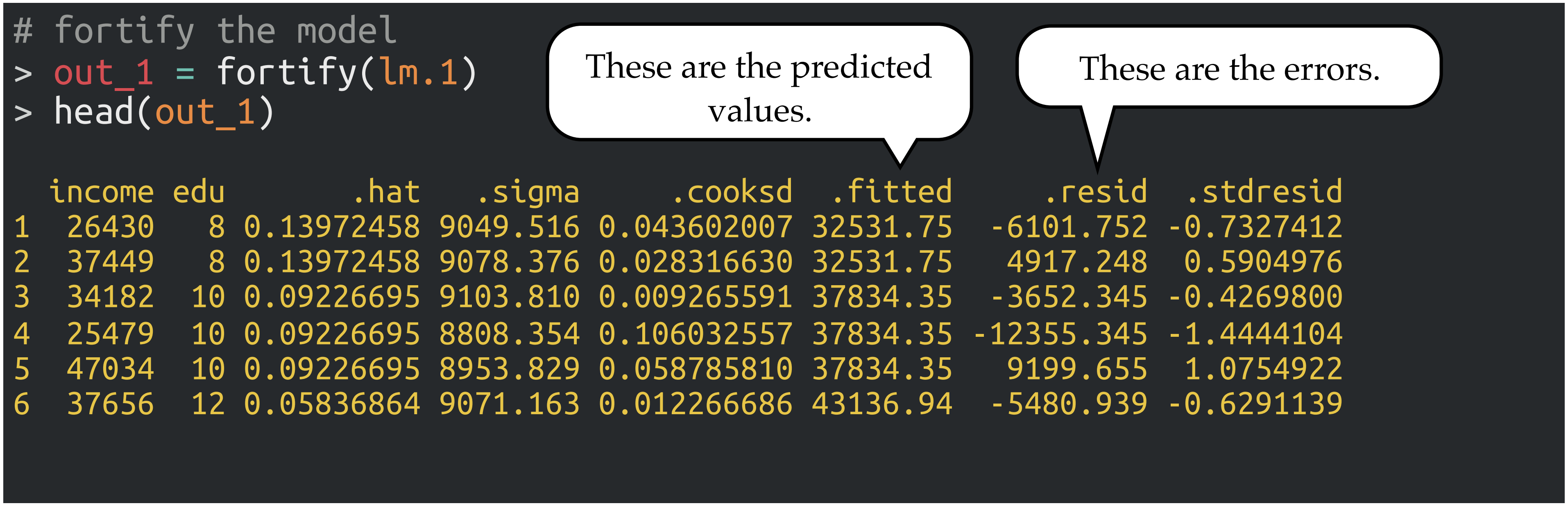
What to do: Use a method for correlated (non-independent) data
(Take EPsy 8252 to find out more!)

Assumption: $\varepsilon_{ij} \sim N(0, \sigma^2)$

These assumptions are about the distribution of errors at each level of X. This assumption has three parts to examine:

- The distribution of errors at each value of X is normal.
- The distribution of errors at each value of X has a mean of 0
- The distribution of errors at each value of X has the same variance

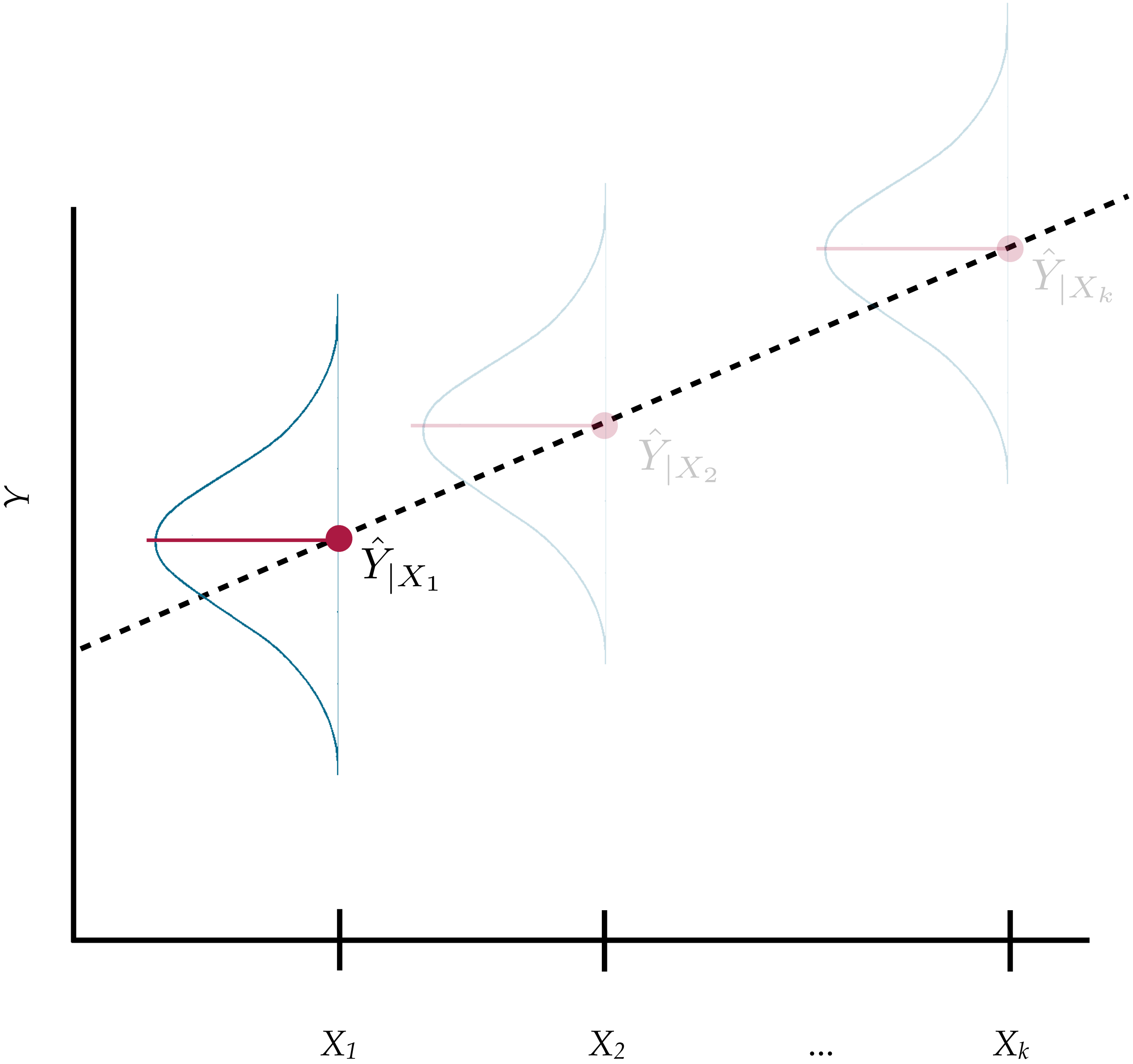
We will use the `fortify()` function from the **ggplot2** library to obtain the sample errors and evaluate the assumption.



```
# fortify the model
> out_1 = fortify(lm.1)
> head(out_1)

  income edu        .hat    .sigma      .cooksd  .fitted     .resid  .stdresid
1  26430   8 0.13972458 9049.516 0.043602007 32531.75  -6101.752 -0.7327412
2  37449   8 0.13972458 9078.376 0.028316630 32531.75   4917.248  0.5904976
3  34182  10 0.09226695 9103.810 0.009265591 37834.35  -3652.345 -0.4269800
4  25479  10 0.09226695 8808.354 0.106032557 37834.35 -12355.345 -1.4444104
5  47034  10 0.09226695 8953.829 0.058785810 37834.35   9199.655  1.0754922
6  37656  12 0.05836864 9071.163 0.012266686 43136.94  -5480.939 -0.6291139
```

These are the predicted values.

These are the errors.



Consider the distribution of Y values at X_1

$$Y_i = \beta_0 + \beta_1(X_1) + \epsilon_i$$

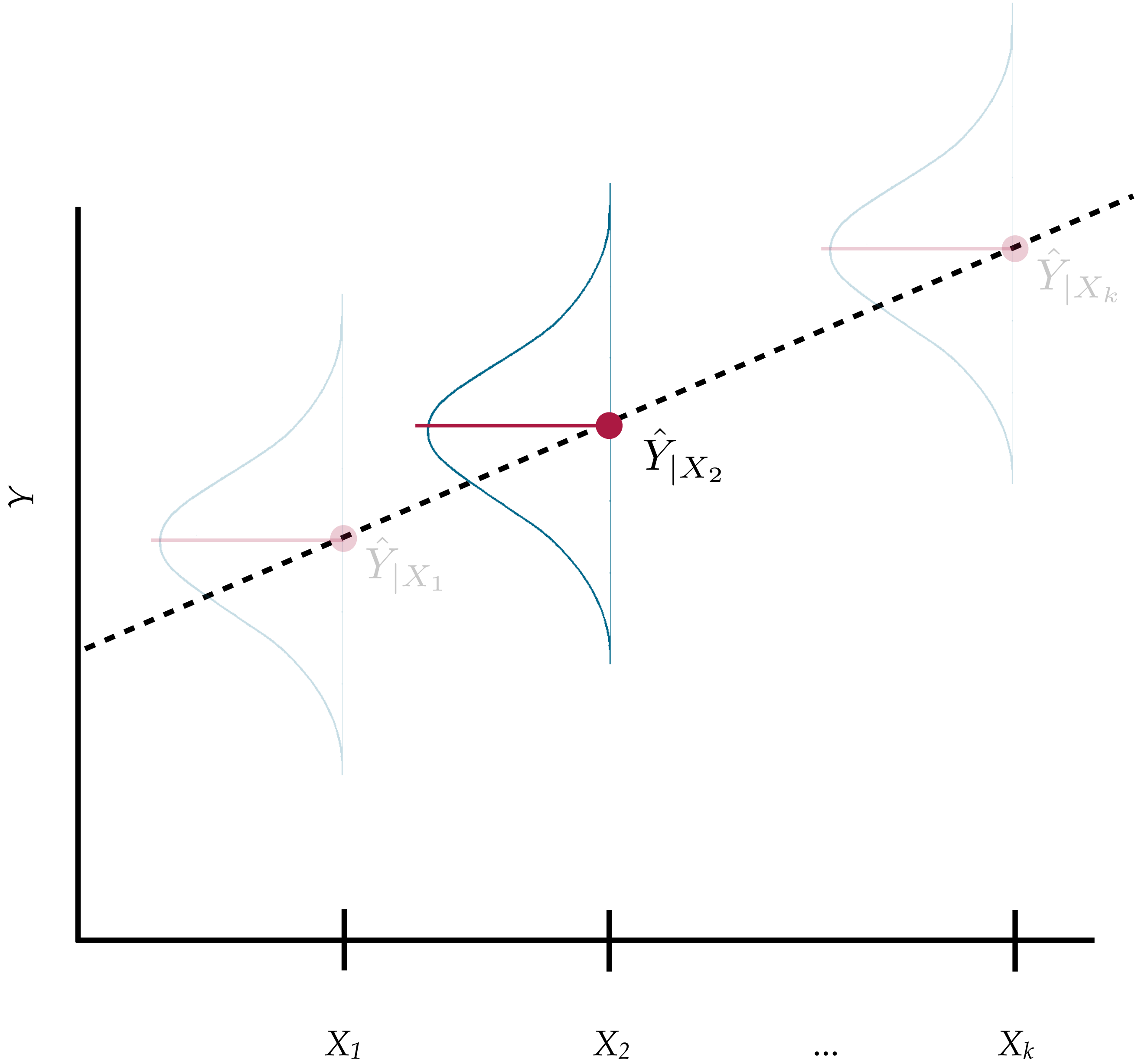
where

$$\hat{Y} = \beta_0 + \beta_1(X_1)$$

At X_1, the predicted value is constant.

$$\epsilon_i = Y_i - \hat{Y}$$

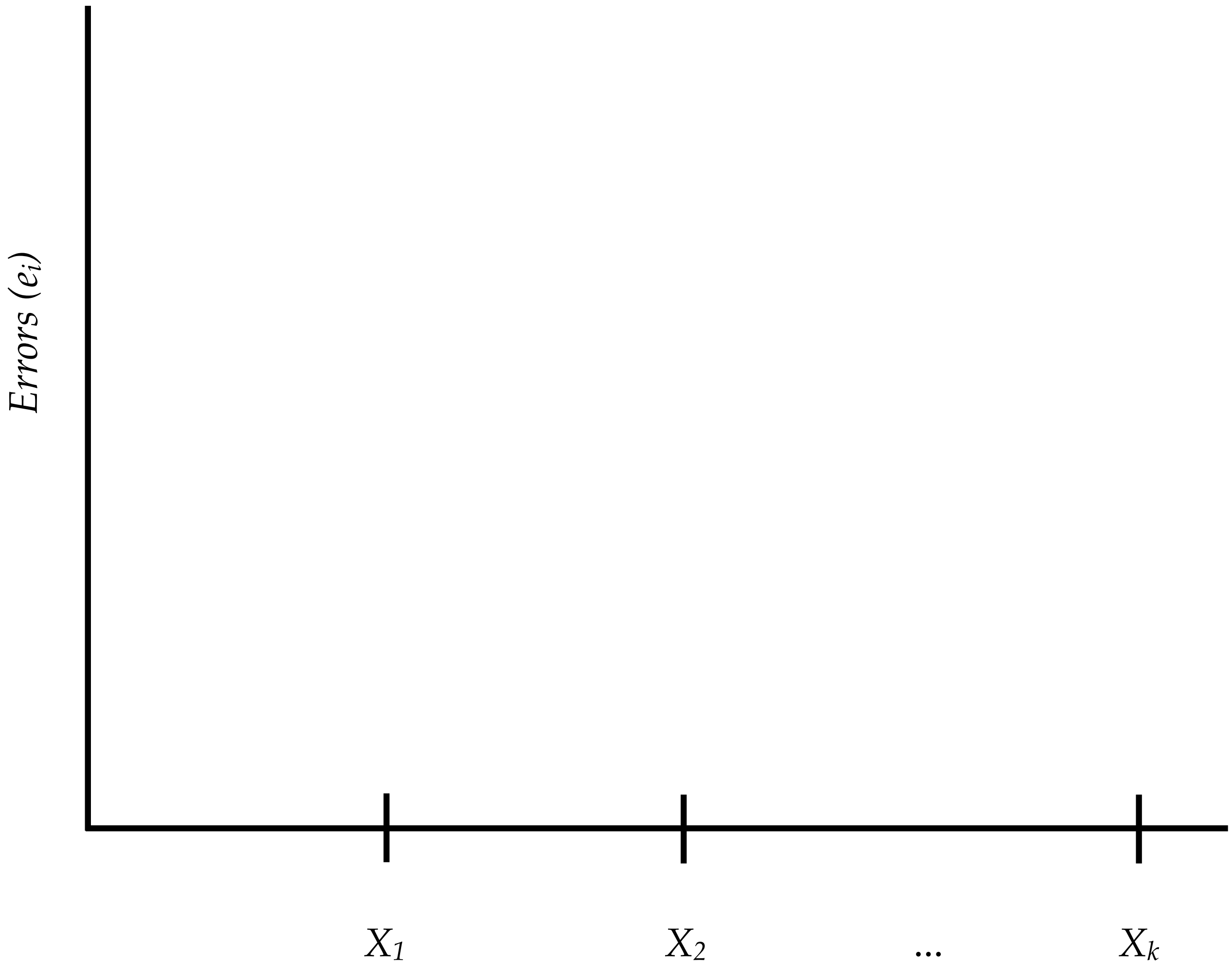
What would the distribution of errors look like at $X = X_1$?



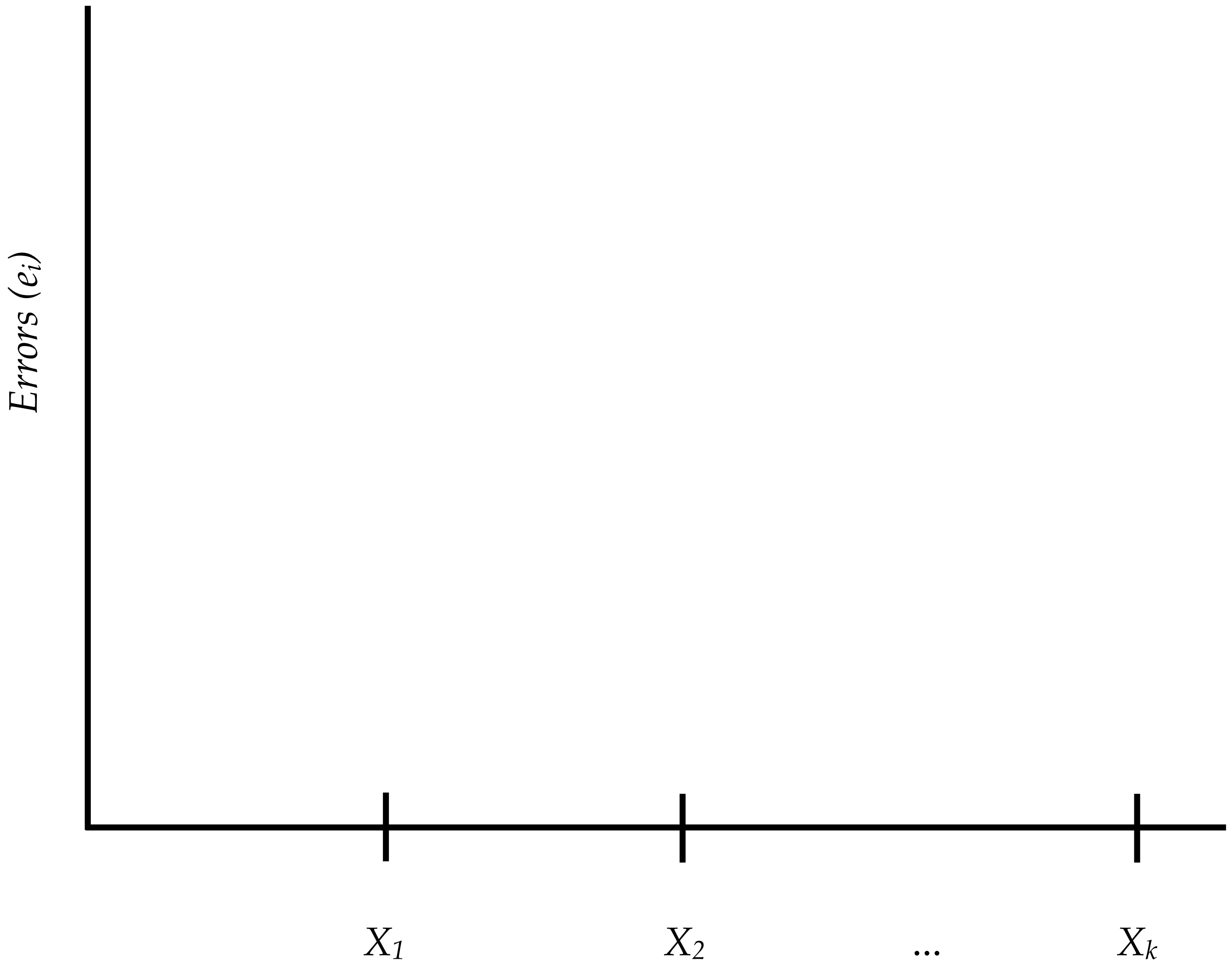
Consider the distribution of Y values at X_2

What would the distribution of errors look like at $X = X_2$?

Consider the distribution of error values at each level of X.
Sketch them.

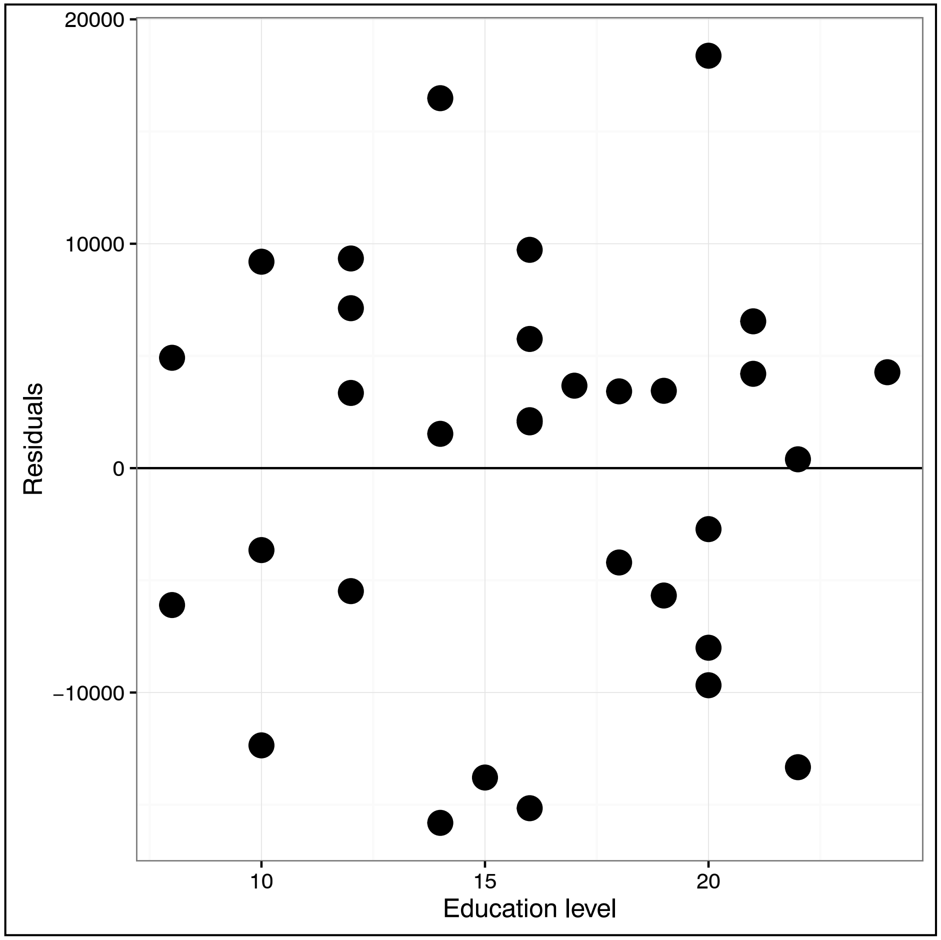


What would a scatterplot of the errors vs. *X* look like? Sketch it.



We can examine a scatterplot of the residuals (on Y) vs. the predictor (on X). This is a **residual plot**.

```
# Plot the marginal distribution of the errors
> library(ggplot)
> ggplot(data = out_1, aes(x = edu, y = .resid)) +
    geom_point() +
    theme_bw() +
    geom_hline(yintercept = 0)
```



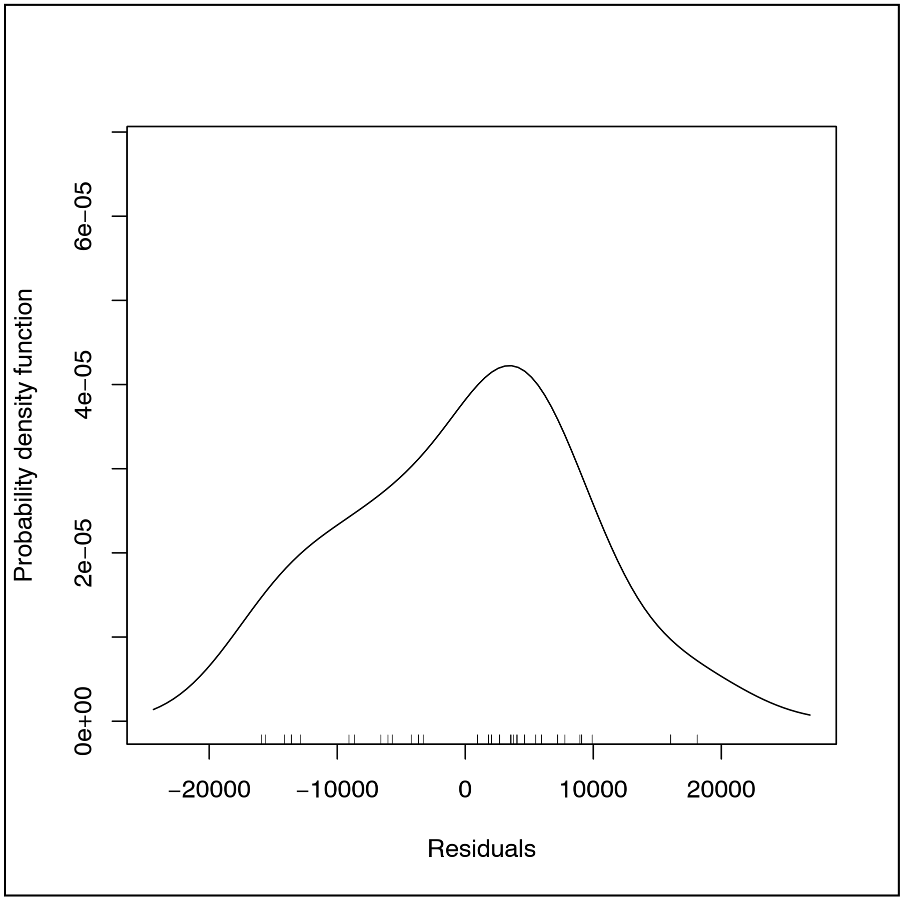
Do the conditional distributions of the errors suggest that the conditional distributions in the population might be normal?

What makes this hard to determine?

Now consider the distribution of *marginal* distribution of error values across all levels of X. Sketch what that distribution would look like.

To evaluate the normality assumption, we typically examine a plot of the **marginal distribution of the errors**. A histograms or density plot is a useful plot to examine the shape of a distribution.

```
# Plot the marginal distribution of the errors
> library(sm)
> sm.density(out_1$.resid)
```

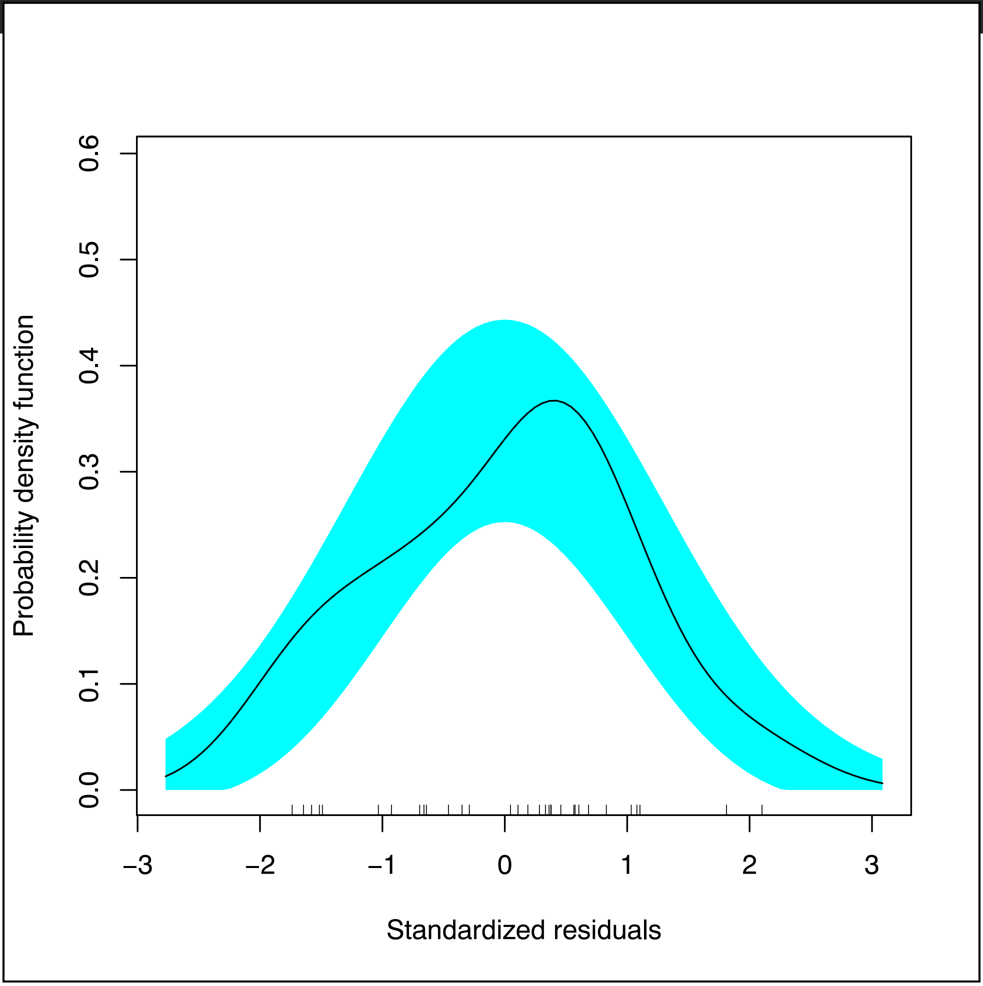


Does the distribution of the sample errors suggest that the population distribution might be normal?

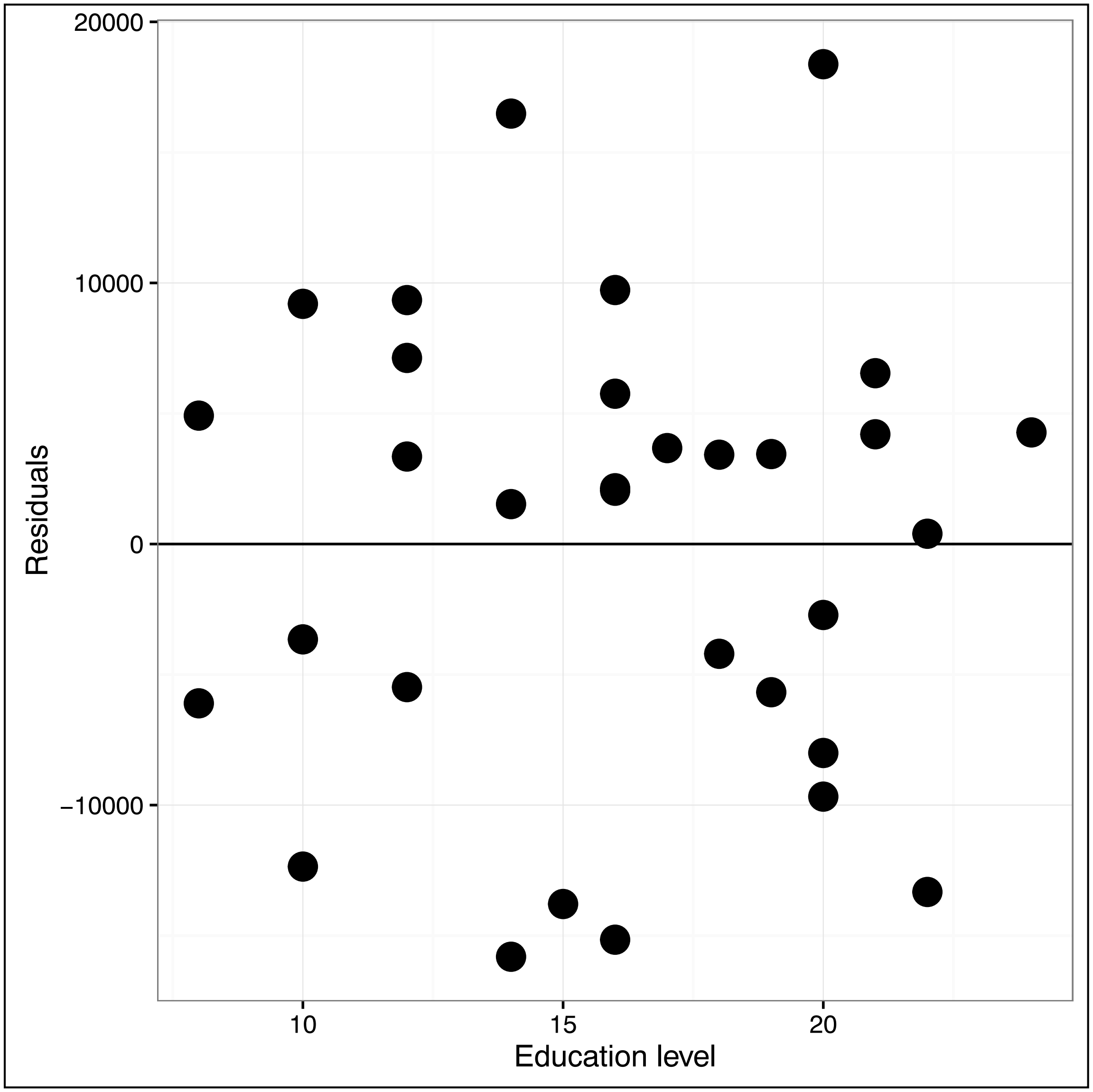
What does this suggest about the conditional distributions (i.e., the distribution of errors at each value of *X*)

One thing to help us evaluate the normality assumption, is to think about the uncertainty we expect in the plot when drawing from a normal distribution.

```
# Plot the marginal distribution of the errors with a
# confidence envelope for the normal model
> sm.density(out_1$.resid, model = "normal")
```



To evaluate the homogeneity of variance assumption, called homoskedasticity in regression analyses, we typically look at the residual plot for roughly constant ranges.



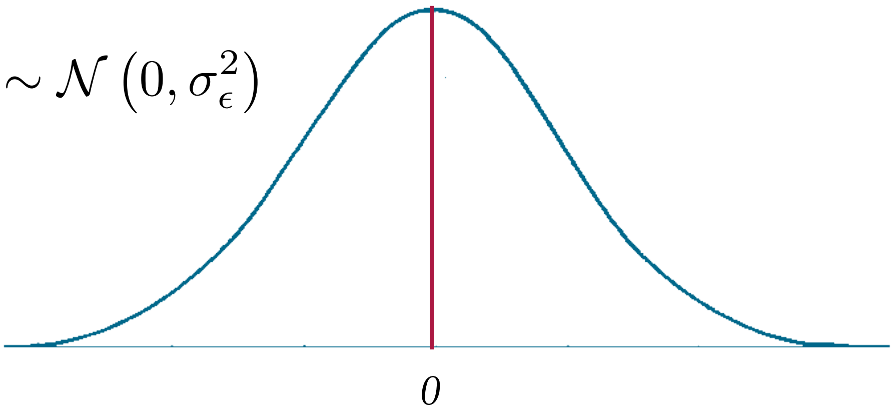
Here we see some differences in the ranges of the errors. For example, the range of errors at education = 16 is slightly larger than the ranges at education = 10 or education = 20.

This is probably related to differences in sample sizes at these *X* values, and does not reflect actual differences in the population variances.

Standardizing the Residual

The residuals are in the same metric as the outcome. The magnitude of the residuals thus needs to be judged in that metric.

To make it easier to judge whether an observation has an extreme residual, they are typically standardized.



The distribution of residuals at each *X* is normally distributed with a mean of 0 and constant variance.

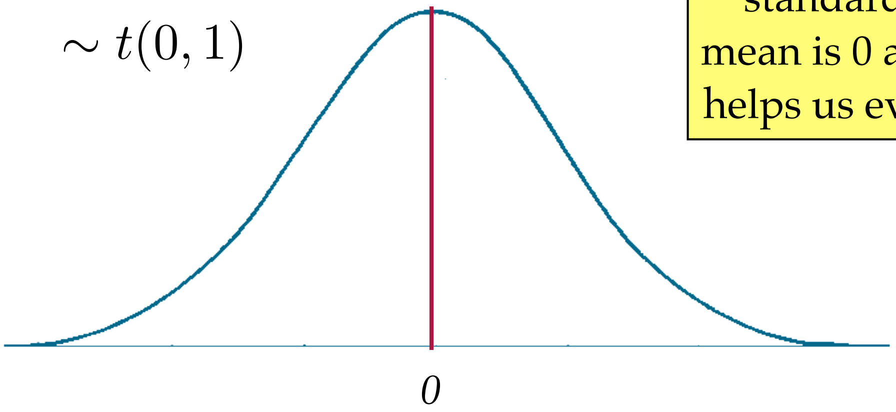
$$Z_{\epsilon_i} = \frac{\epsilon_i - 0}{\text{SD}}$$

Since residuals are statistics, the SD is called a standard error (SE)

When the SE is estimated from the data we call it *Studentizing* rather than standardizing, since the resulting score is a *t*-value rather than a *z*-value.

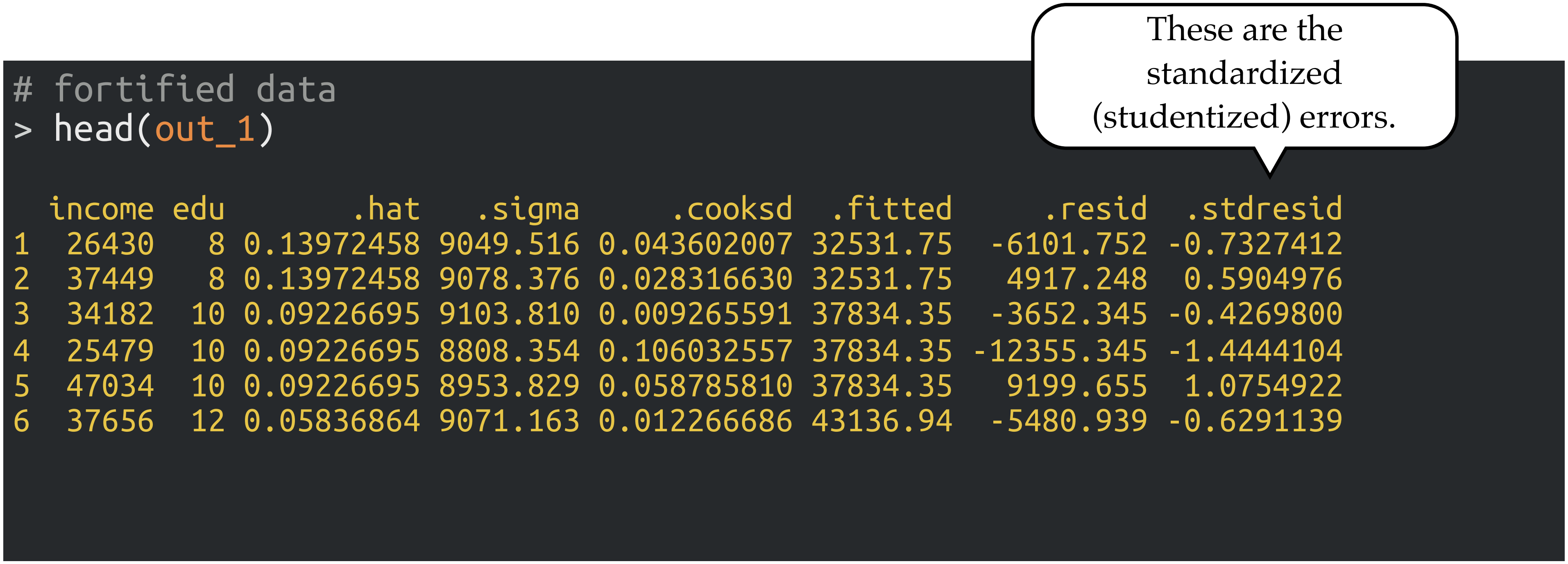
$$t_{\epsilon_i} = \frac{\epsilon_i - 0}{\text{SE}}$$

$\sim t(0, 1)$



The studentized residuals are standardized in that their new mean is 0 and the SD (SE) is 1. This helps us evaluate their magnitude.

Given the assumption of normality, what percentage of the standardized residuals within 2 standard errors of the mean?

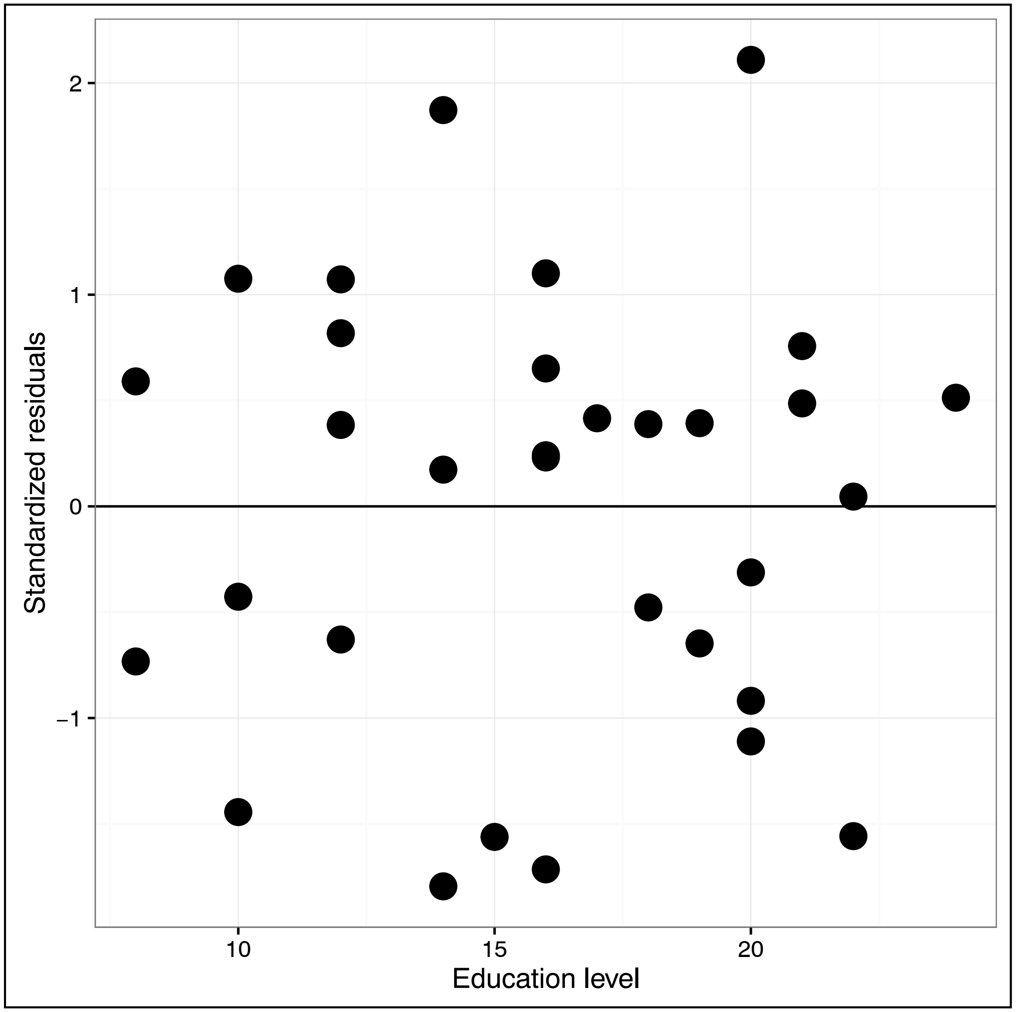


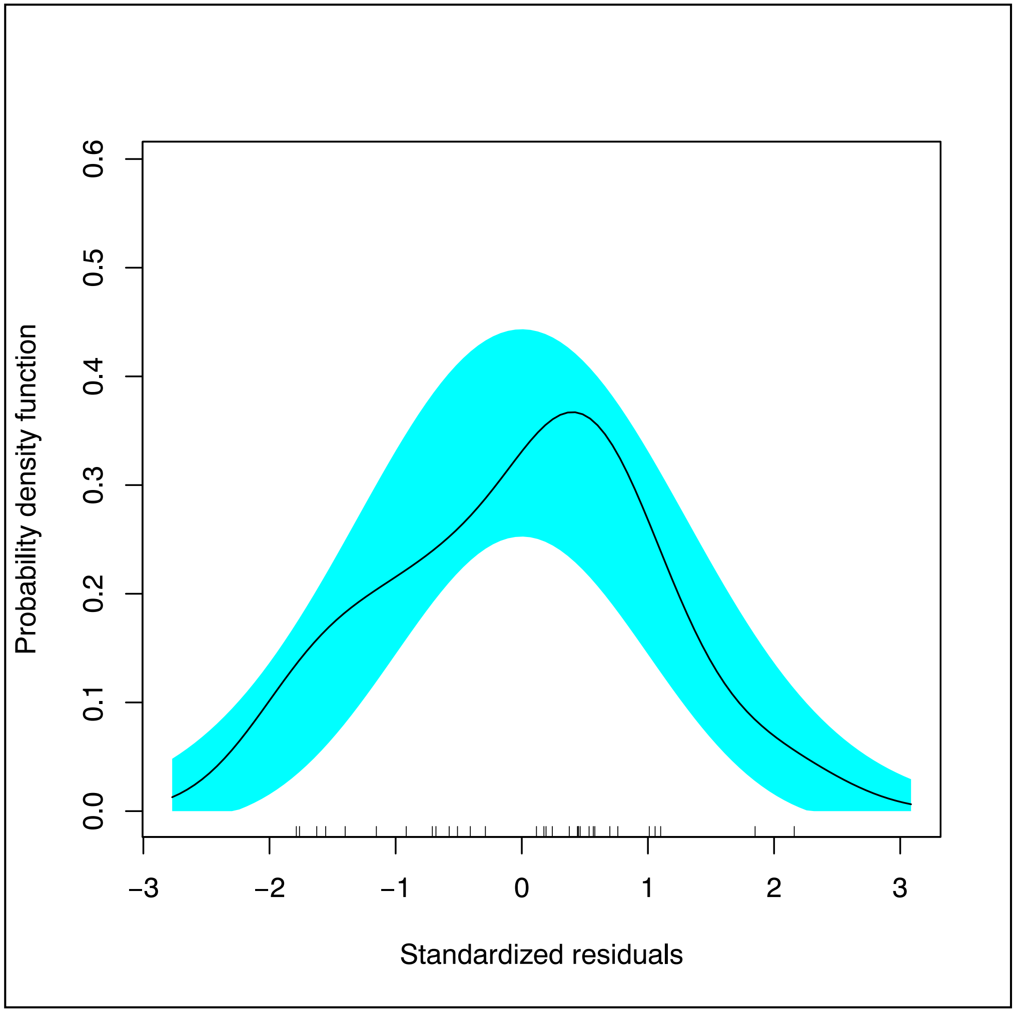
```
# fortified data
> head(out_1)

  income edu       .hat   .sigma     .cooksd  .fitted     .resid  .stdresid
1  26430   8 0.13972458 9049.516 0.043602007 32531.75  -6101.752 -0.7327412
2  37449   8 0.13972458 9078.376 0.028316630 32531.75   4917.248  0.5904976
3  34182  10 0.09226695 9103.810 0.009265591 37834.35  -3652.345 -0.4269800
4  25479  10 0.09226695 8808.354 0.106032557 37834.35 -12355.345 -1.4444104
5  47034  10 0.09226695 8953.829 0.058785810 37834.35   9199.655  1.0754922
6  37656  12 0.05836864 9071.163 0.012266686 43136.94  -5480.939 -0.6291139
```

These are the standardized (studentized) errors.

All of the plots we use to evaluate regression assumptions should be created using the standardized residuals rather than the raw residuals.

Plots Using the Standardized Residuals (Model B)





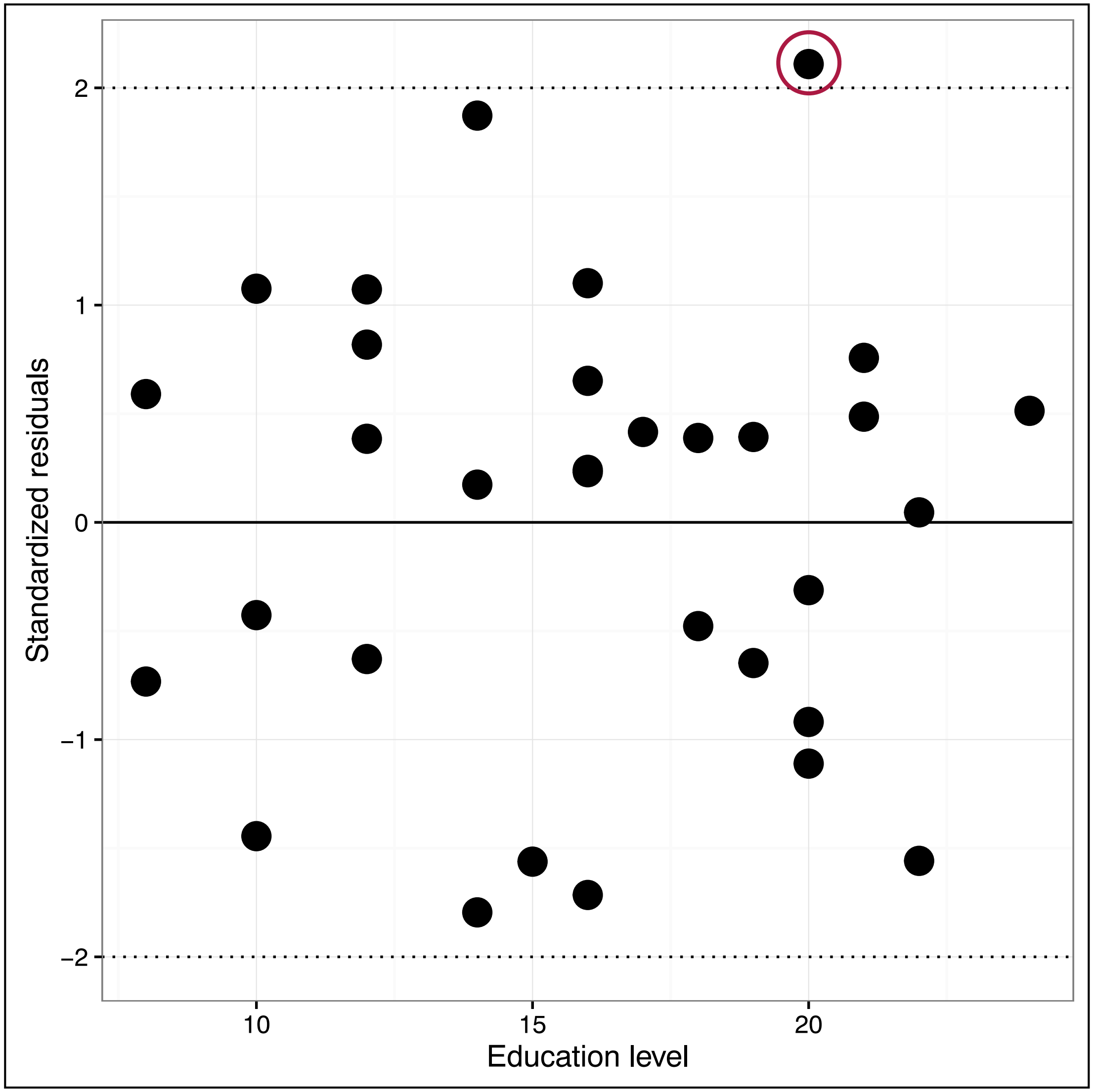
Since standardizing is a linear transformation, the shapes and relationships we examine will be identical whether you use the raw or standardized residuals.

Examining Observations that have an Extreme Residual

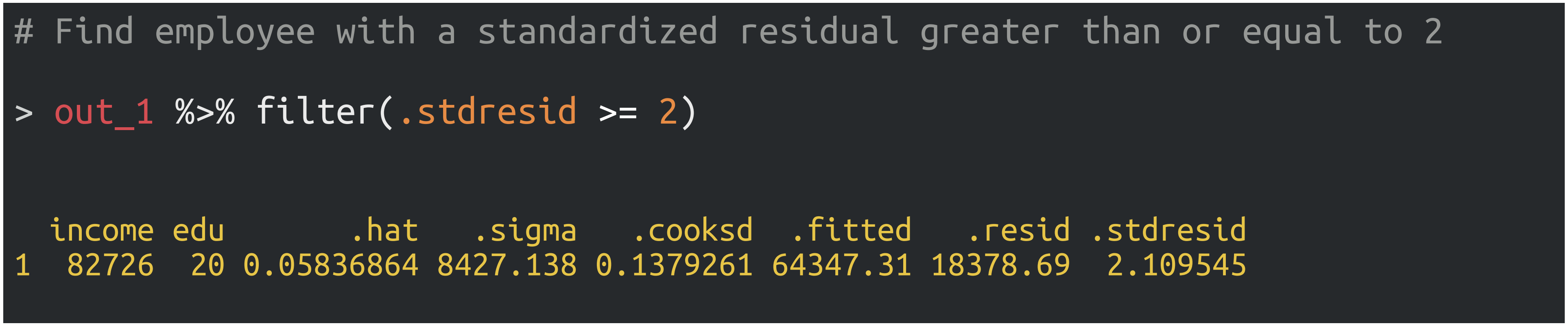
When examining the residual plot, it can be helpful to add horizontal lines at ±2. This helps us easily see observations that have an outcome value that is more than 2 standard errors from where we would predict the outcome to be given their *X* values.

This employee has a "large" residual relative to the rest of the data.

Given this employee's education level, we would predict her/him to have a much lower income than they actually do.

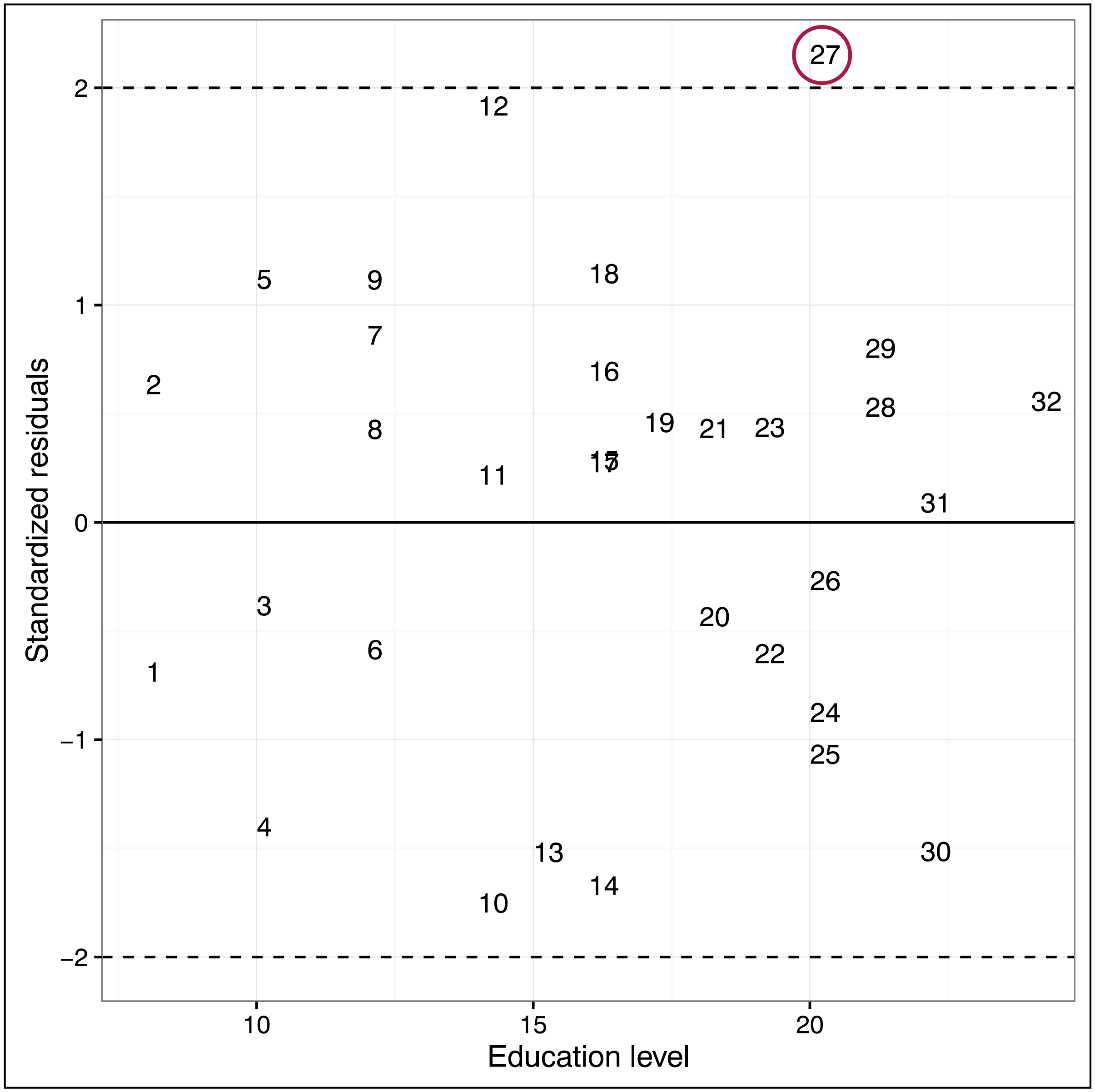


We can use `filter()` from **dplyr** to find out which employee this is.

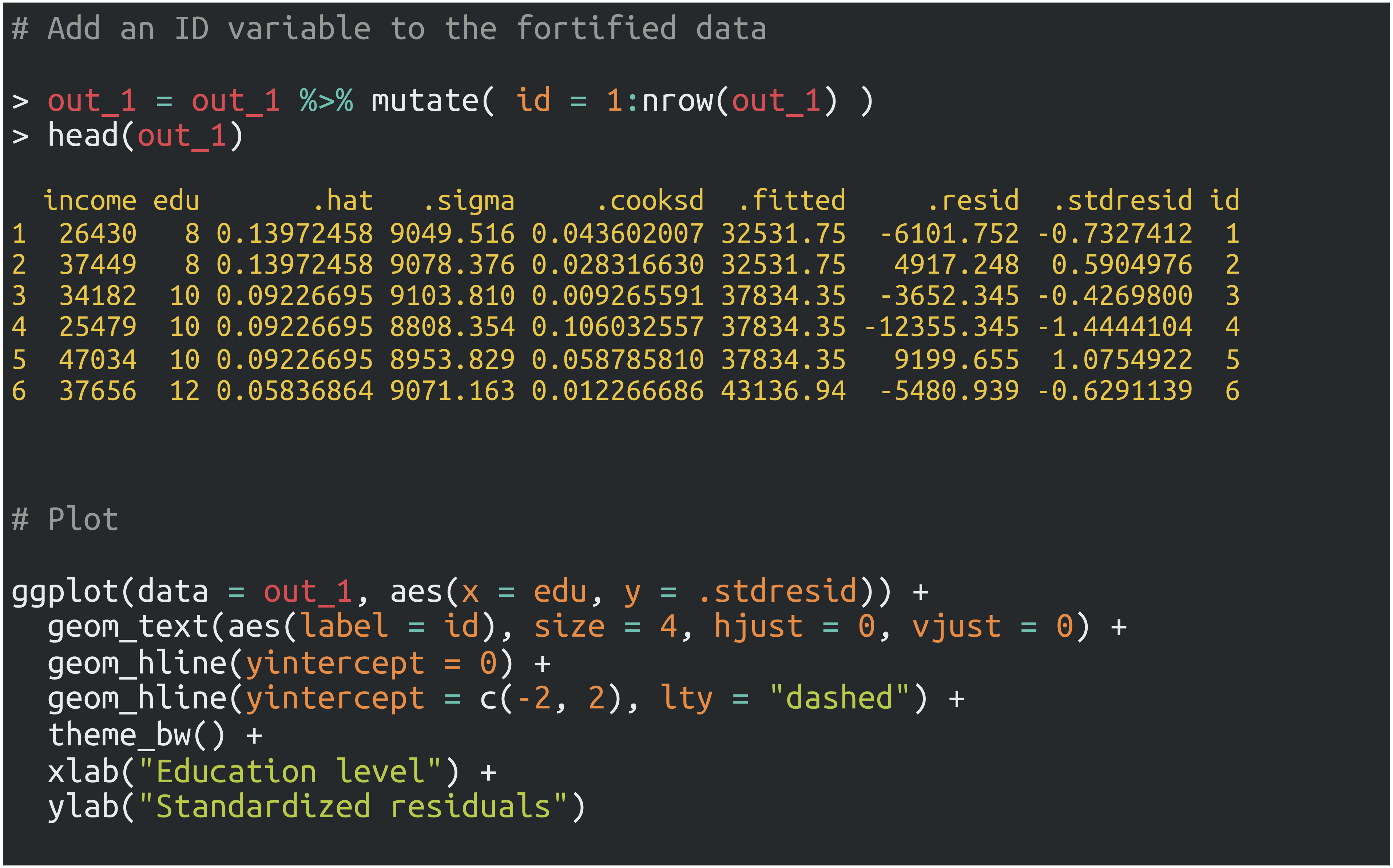


```
# Find employee with a standardized residual greater than or equal to 2

> out_1 %>% filter(.stdresid >= 2)


  income edu        .hat   .sigma   .cooksd  .fitted   .resid .stdresid
1  82726  20 0.05836864 8427.138 0.1379261 64347.31 18378.69  2.109545
```

Rather than plotting points for each observation, we can plot text that provides each observation's row number.



To do this, we first need to create an ID variable in the fortified data, then use `geom_text()` rather than `geom_point()` in the ggplot syntax.



```
# Add an ID variable to the fortified data

> out_1 = out_1 %>% mutate( id = 1:nrow(out_1) )
> head(out_1)

  income edu       .hat   .sigma     .cooksd  .fitted     .resid  .stdresid id
1  26430   8 0.13972458 9049.516 0.043602007 32531.75  -6101.752 -0.7327412  1
2  37449   8 0.13972458 9078.376 0.028316630 32531.75   4917.248  0.5904976  2
3  34182  10 0.09226695 9103.810 0.009265591 37834.35  -3652.345 -0.4269800  3
4  25479  10 0.09226695 8808.354 0.106032557 37834.35 -12355.345 -1.4444104  4
5  47034  10 0.09226695 8953.829 0.058785810 37834.35   9199.655  1.0754922  5
6  37656  12 0.05836864 9071.163 0.012266686 43136.94  -5480.939 -0.6291139  6



# Plot

ggplot(data = out_1, aes(x = edu, y = .stdresid)) +
  geom_text(aes(label = id), size = 4, hjust = 0, vjust = 0) +
  geom_hline(yintercept = 0) +
  geom_hline(yintercept = c(-2, 2), lty = "dashed") +
  theme_bw() +
  xlab("Education level") +
  ylab("Standardized residuals")
```

Outliers and Problematic Observations

We have identified observations with large residuals. In regression these are typically referred to as **outliers**. Outliers are problematic because their observed Y does not "fit" with what we expect (i.e., they have a very different outcome from their predicted value).

There are several other types of problematic observations in regression. Two of these are **leverage observations** and **influential observations**. There are several ways to measure these types of observations.

These methods are beyond the scope of this class. The textbook has information about and methodology used to identify these types of problematic observations for the interested student. This would be discussed further in an advanced regression course (e.g., EPsy 8264).

Have the Model Assumptions been Met?

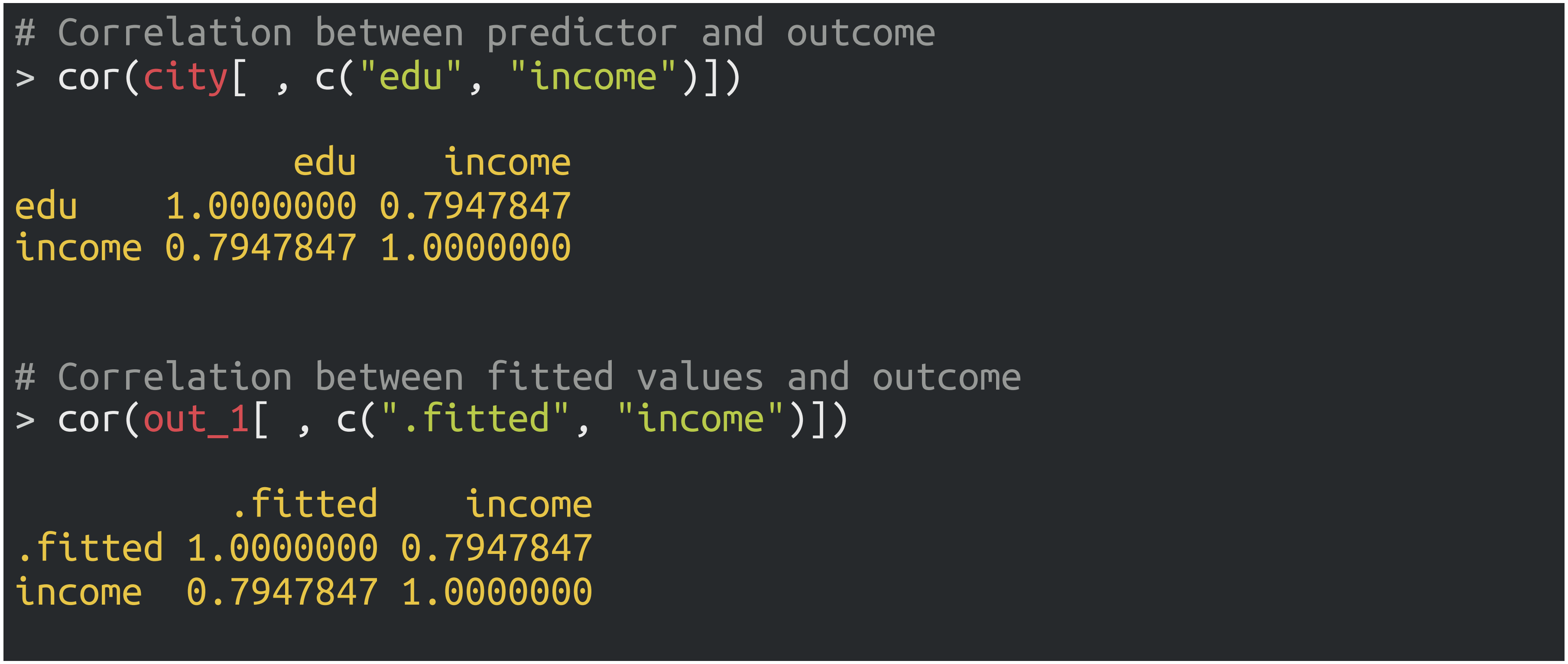
- ☑ Independence
- ☑ The mean of each conditional distribution of the residuals is zero.
- ☑ The conditional distributions of the residuals are normally distributed.
- ☑ Each conditional distributions of the residuals has the same variance.

In practice, we would say the model's assumptions have been reasonably satisfied.

If the assumptions are not satisfied, we should not believe the estimates of the coefficients and standard errors, nor subsequently, the *p*-values and CIs.

We will look at transformations as one possible solution later in the course.

Correlations



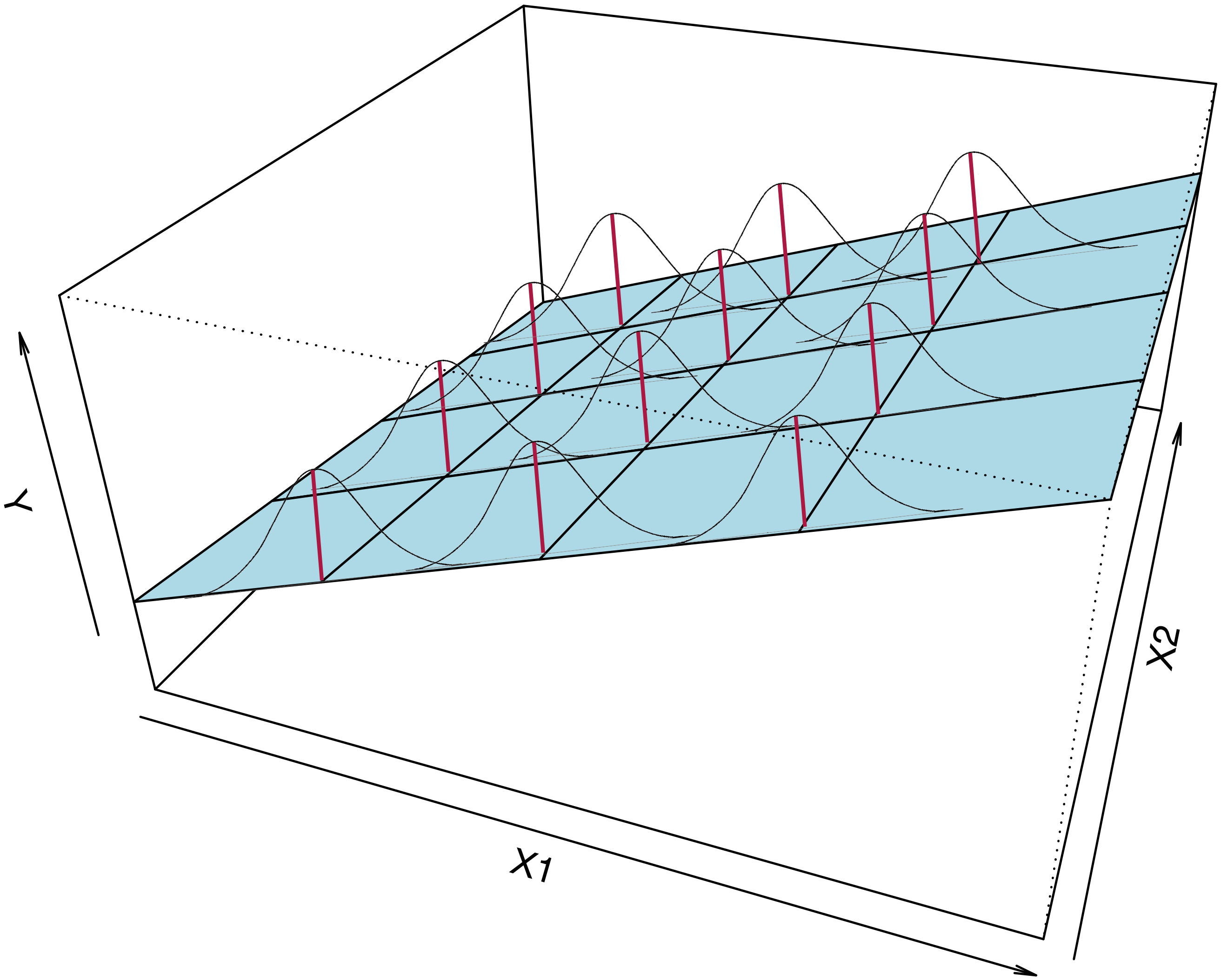
```
# Correlation between predictor and outcome
> cor(city[ , c("edu", "income")])

             edu    income
edu    1.0000000 0.7947847
income 0.7947847 1.0000000


# Correlation between fitted values and outcome
> cor(out_1[ , c(".fitted", "income")])

          .fitted    income
.fitted 1.0000000 0.7947847
income  0.7947847 1.0000000
```

In simple regression, the correlation between X and Y is the same as the correlation between X and the fitted values. Why?

```
# Correlation between fitted values and residuals
> cor(out_1[ , c(".fitted", ".resid")])

               .fitted        .resid
.fitted   1.000000e+00 -3.551972e-18
.resid   -3.551972e-18  1.000000e+00
```

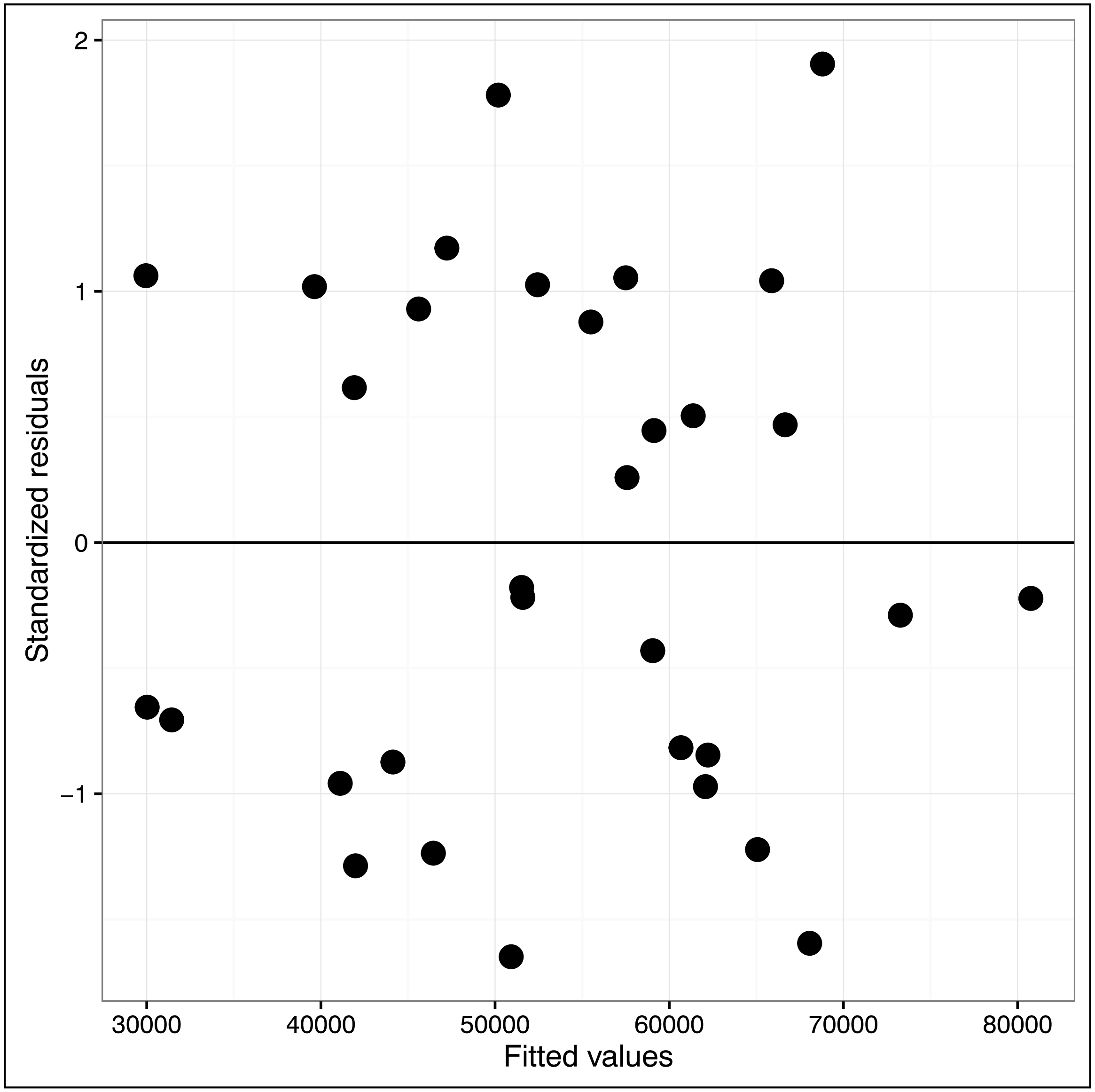
The correlation between the fitted values and the residuals will always be zero. Why?

Multiple Regression



- The mean Y-values from each combination of X_1 *and* X_2 are **linear**.
- The errors are **independent.**
- The distribution of errors at each combination of X_1 *and* X_2 is **normally distributed**.
- The mean error / residual at each combination of X_1 *and* X_2 is equal to 0.
- The variance of the residuals at each combination of X_1 *and* X_2 is exactly the same (homoskedasticity)

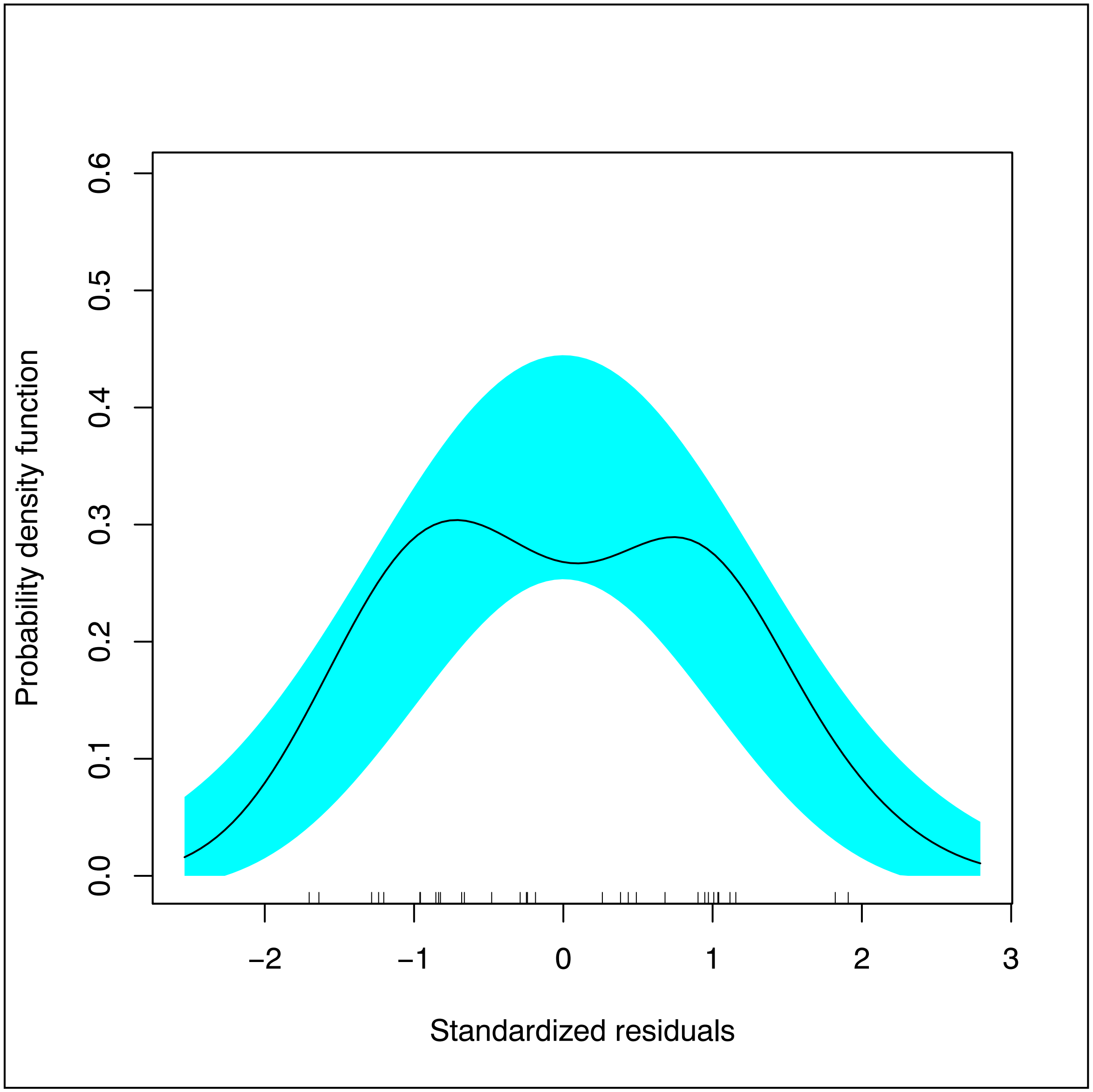
When there is more than one predictor in the model we create the **residual plot** by plotting the errors vs. the predicted values.



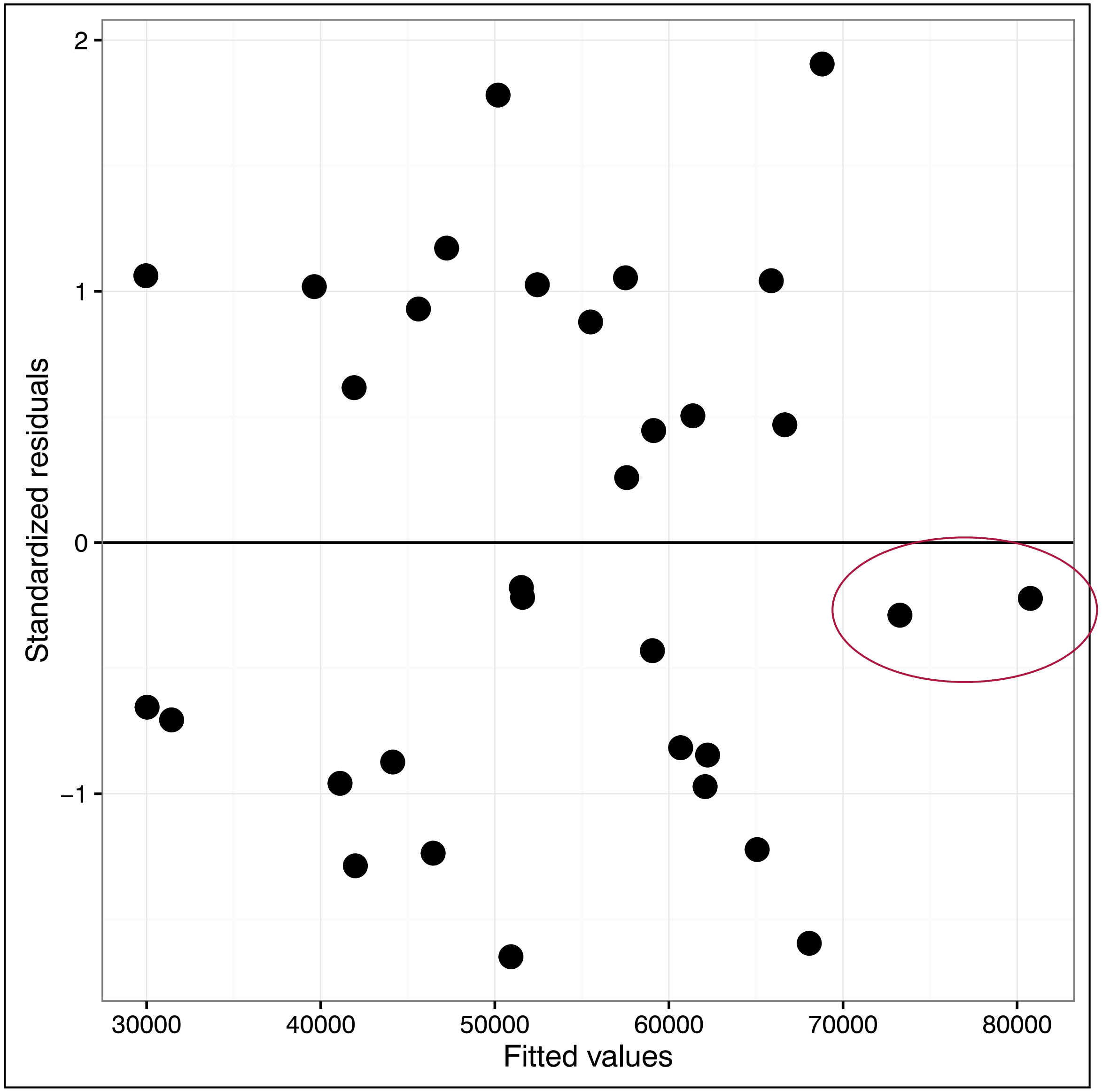
The fitted values represent the combinations of X_1 and X_2

We evaluate this plot the same way we evaluated the residual plot for the simple regression.

When there is more than one predictor in the model we examine normality by again looking at the **marginal distribution of the residuals**.



Residual Plot to Evaluate Linearity



We can also evaluate this plot to assess the linearity assumption.

The "Goldilocks" principle says some residuals will be positive, some will be negative, but, on average, they are 0.

If we pick any fitted value, we should see this principle in the plot.

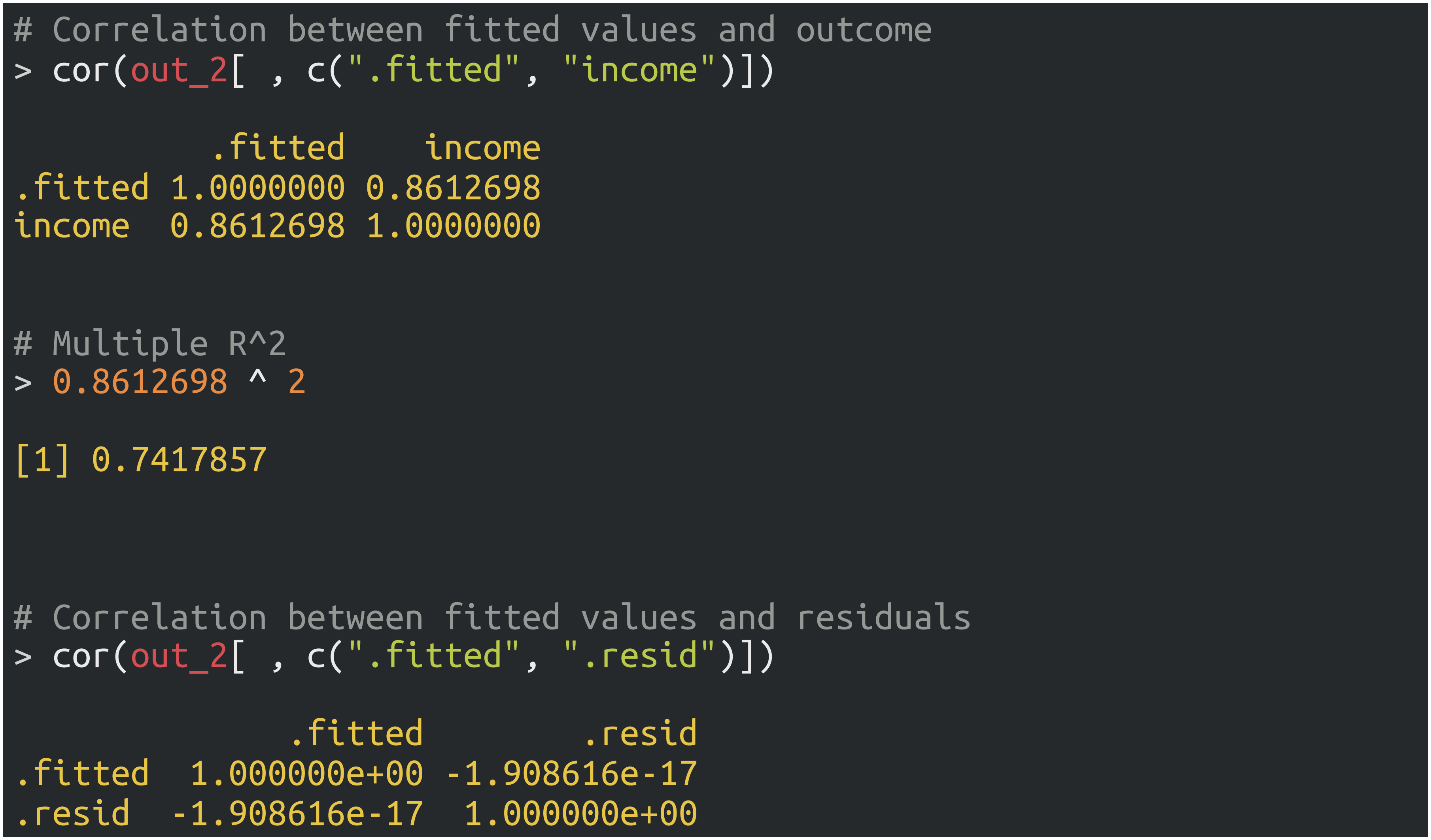
For higher fitted values the residuals tend to be negative. This is probably because of the small number of these fitted values in the sample data (remember the assumption is about the population residuals).

Table 1

Regression Results for Fitting a Taxonomy of Models to Predict Employee Income ($n = 32$)

	Model 1		Model 2	
Coefficient	B	SE	B	SE
Education level	2651***	370	2252**	335
Seniority			739**	210
(Intercept)	11321	6123	6769	5373
R^2	0.631		0.742	
RMSE	8978		7646	

Correlations: Multiple Regression



```
# Correlation between fitted values and outcome
> cor(out_2[ , c(".fitted", "income")])

          .fitted    income
.fitted 1.0000000 0.8612698
income  0.8612698 1.0000000


# Multiple R^2
> 0.8612698 ^ 2

[1] 0.7417857



# Correlation between fitted values and residuals
> cor(out_2[ , c(".fitted", ".resid")])

              .fitted        .resid
.fitted  1.000000e+00 -1.908616e-17
.resid  -1.908616e-17  1.000000e+00
```