Regression Assumptions

Preparation

We will use the `homework-achievement.csv` data. We read this into a data frame called `math`.

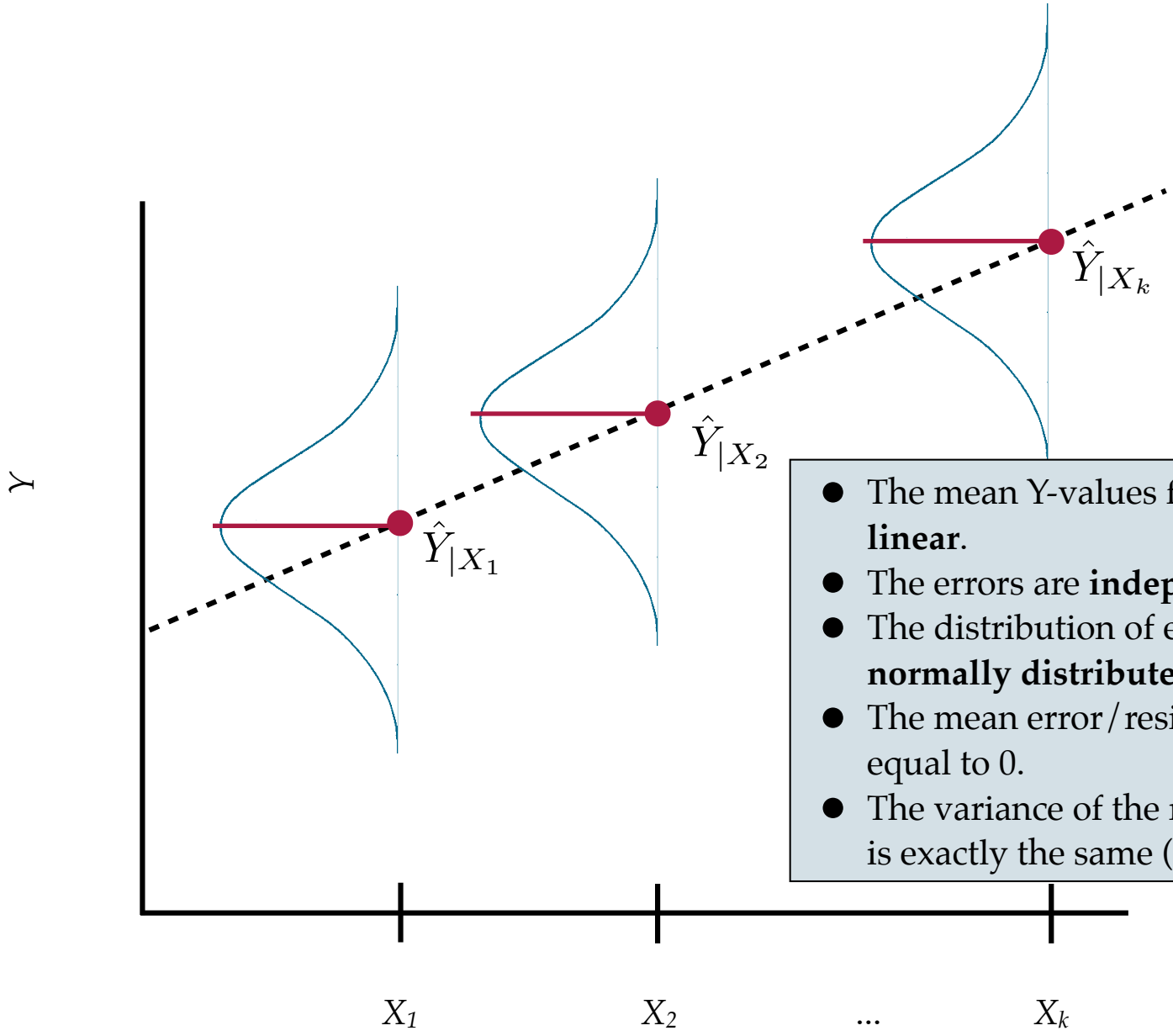
```
# fit the model
> lm.a = lm(achievement ~ homework, data = math)
> summary(lm.a)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  47.0316     1.6940  27.763  < 2e-16 ***
homework      1.9902     0.5952   3.344  0.00117 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.75 on 98 degrees of freedom
Multiple R-squared:  0.1024,   Adjusted R-squared:  0.09324
F-statistic: 11.18 on 1 and 98 DF,  p-value: 0.001173
```

- 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)

- 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)

- 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)

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**.

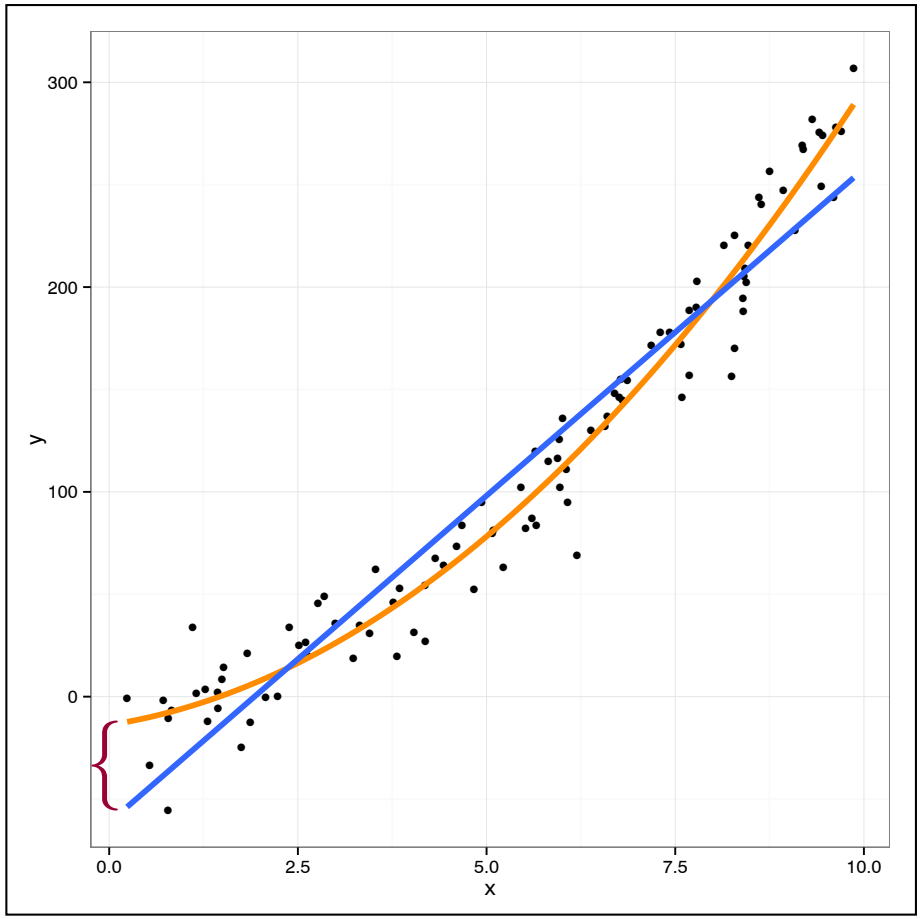
Recall that the sample e_i are approximations of the ε_i

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.

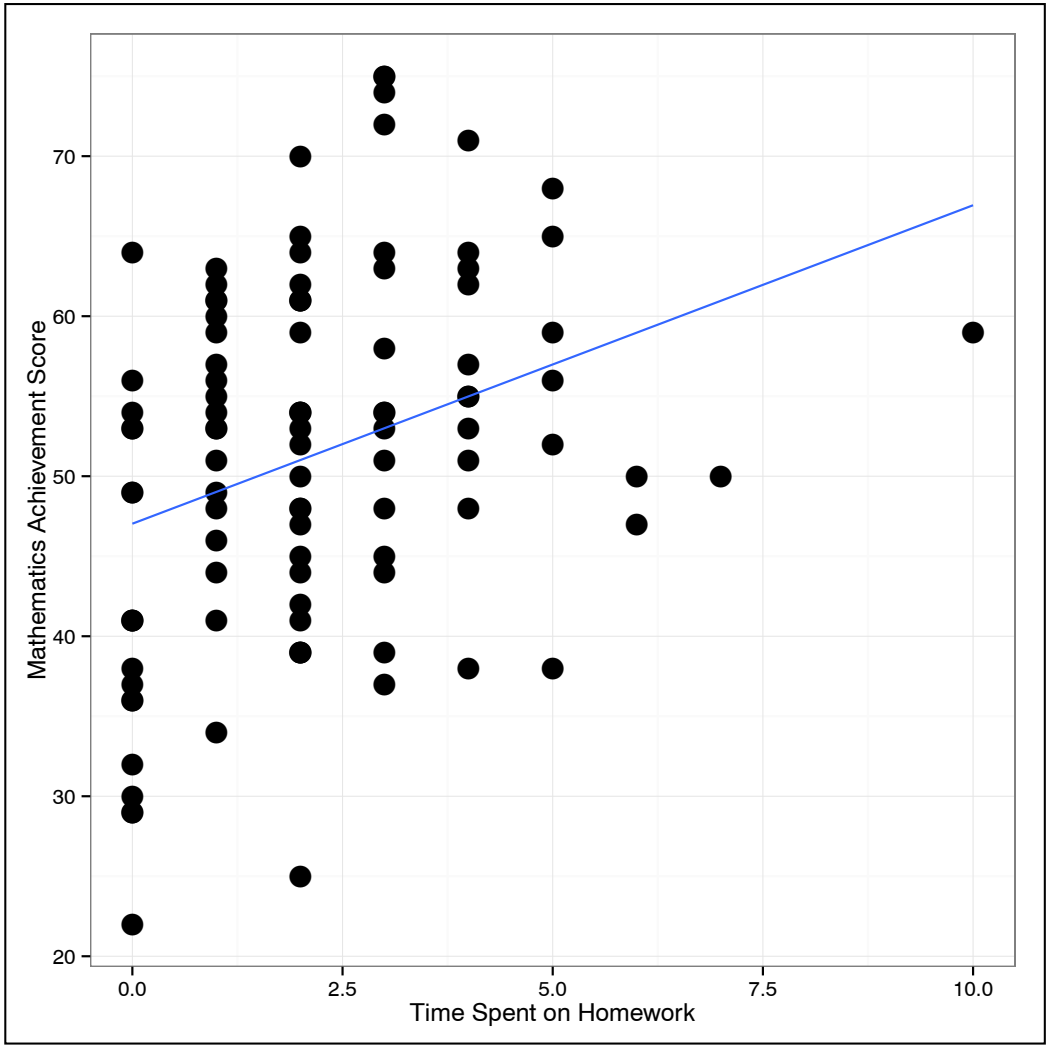
Fitting a linear model when the TRUE relationship is non-linear may, or may not be problematic.



- 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.

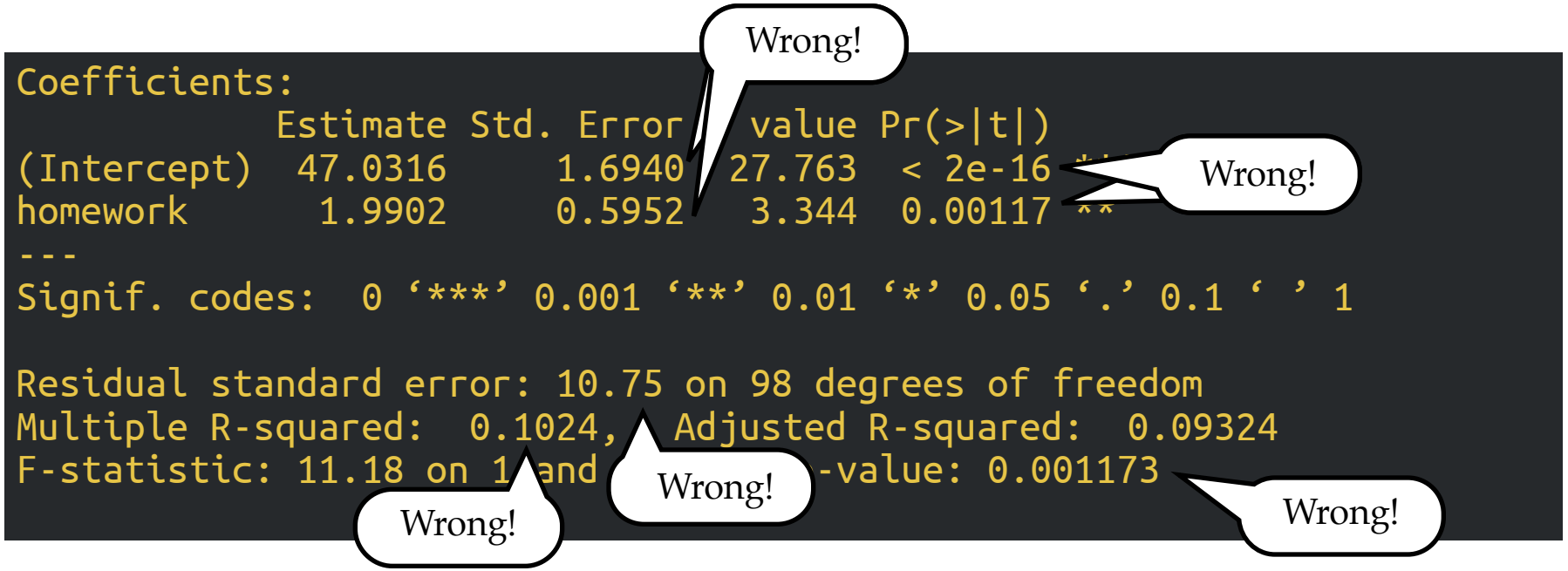
Assessing the independence assumption is primarily a logical argument.

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.



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:

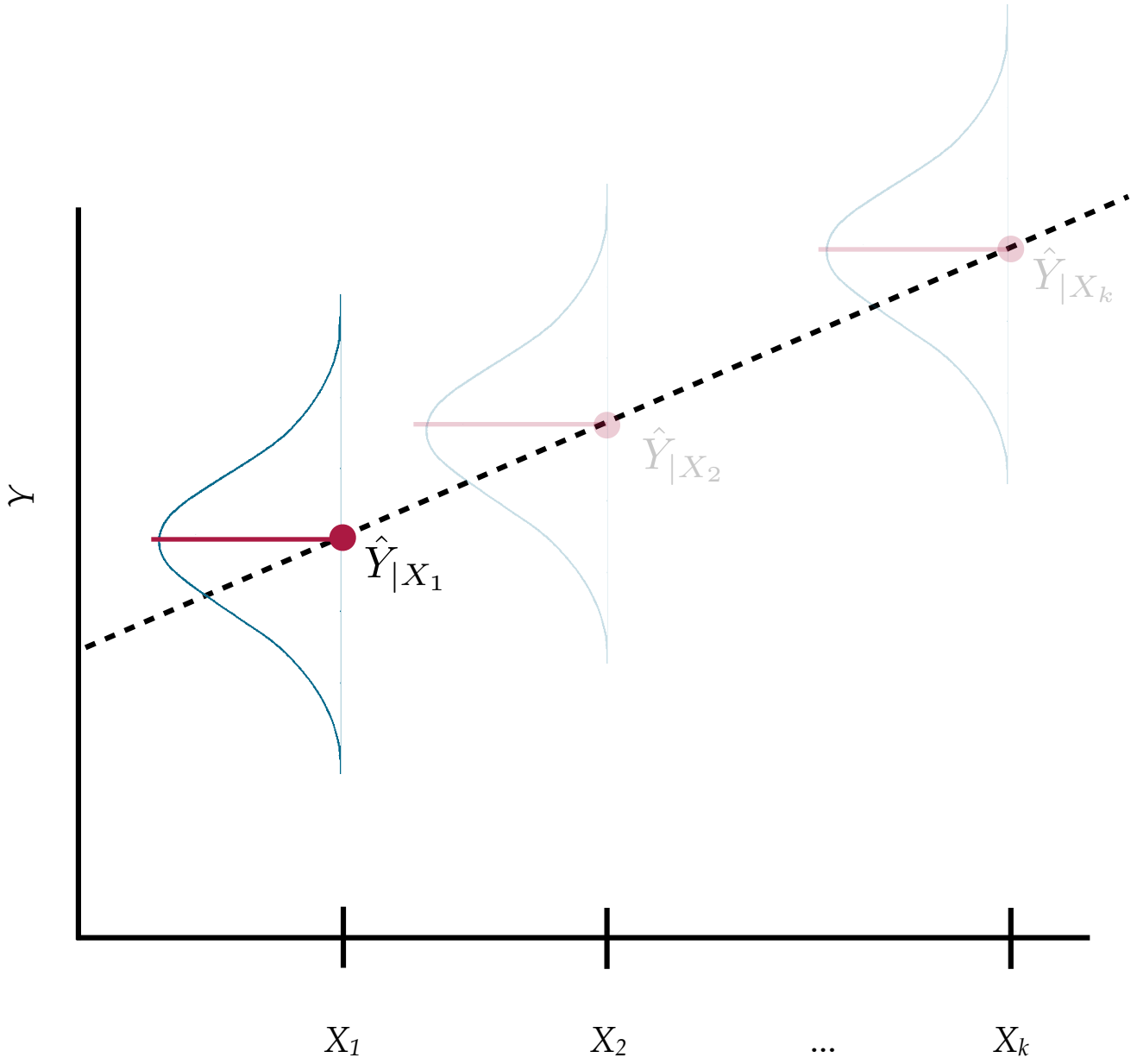
- The distribution of errors at each value of *X* is normal.
- The distribution of errors at each value of *X* has a mean of 0
- The distribution of errors at each value of *X* has the same variance

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

```
# fortify the model
> out_a = fortify(lm.a)
> head(out_a)

  achievement homework       .hat   .sigma      .cooksd  .fitted      .resid  .stdresid
1          54        2 0.01012270 10.79802 0.0003992961 51.01196   2.988037  0.2794506
2          53        0 0.02484663 10.78488 0.0040292826 47.03160   5.968405  0.5623823
3          53        4 0.01993865 10.80040 0.0003566990 54.99233  -1.992331 -0.1872599
4          56        0 0.02484663 10.76290 0.0090979090 47.03160   8.968405  0.8450619
5          59        2 0.01012270 10.77152 0.0028536673 51.01196   7.988037  0.7470664
6          30        0 0.02484663 10.65944 0.0328111777 47.03160 -17.031595 -1.6048286
```

These are the predicted values.

These are the errors.



Consider the distribution of *Y* values at X_1

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

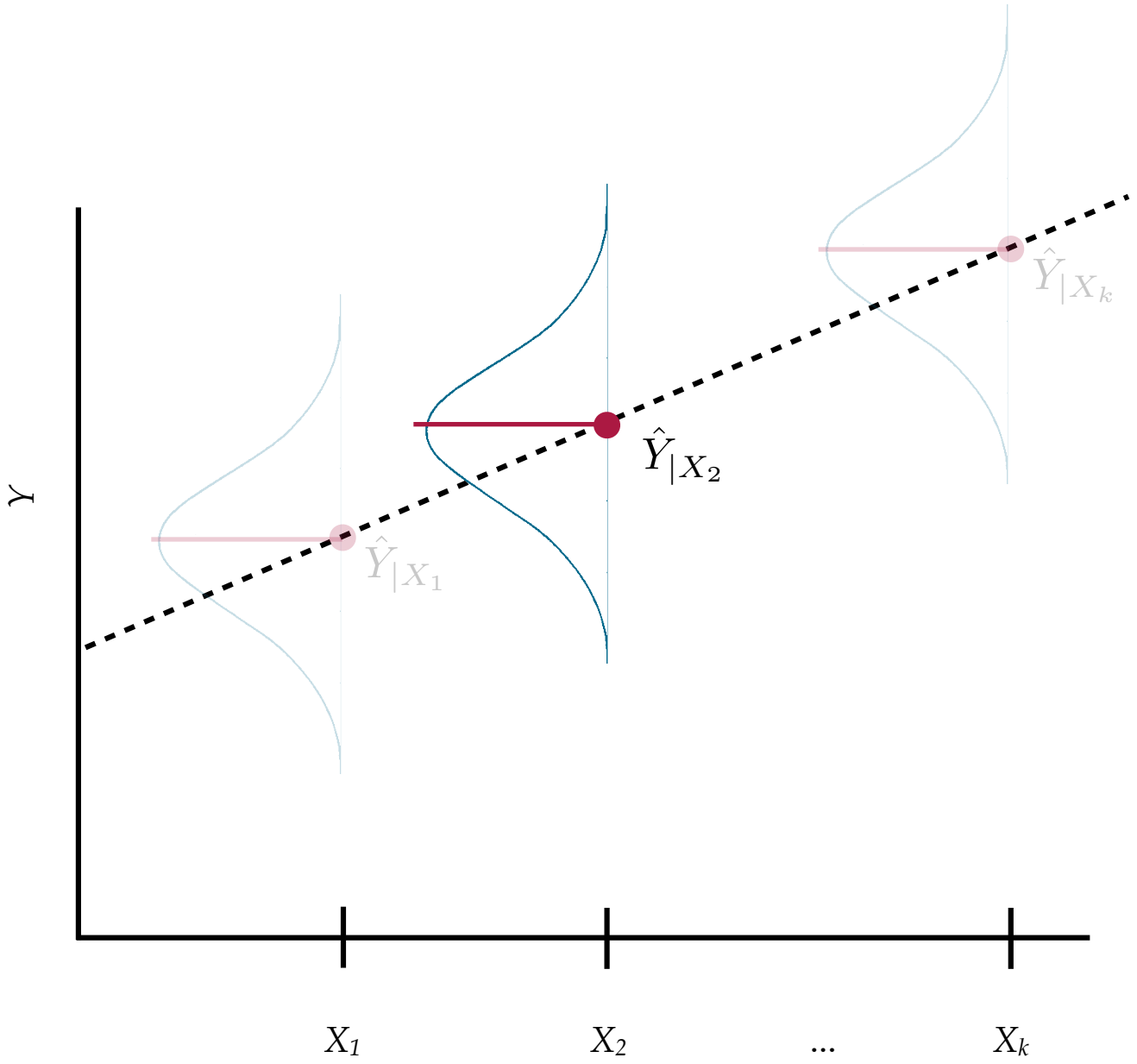
where

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

At X_1, the predicted value is constant.

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

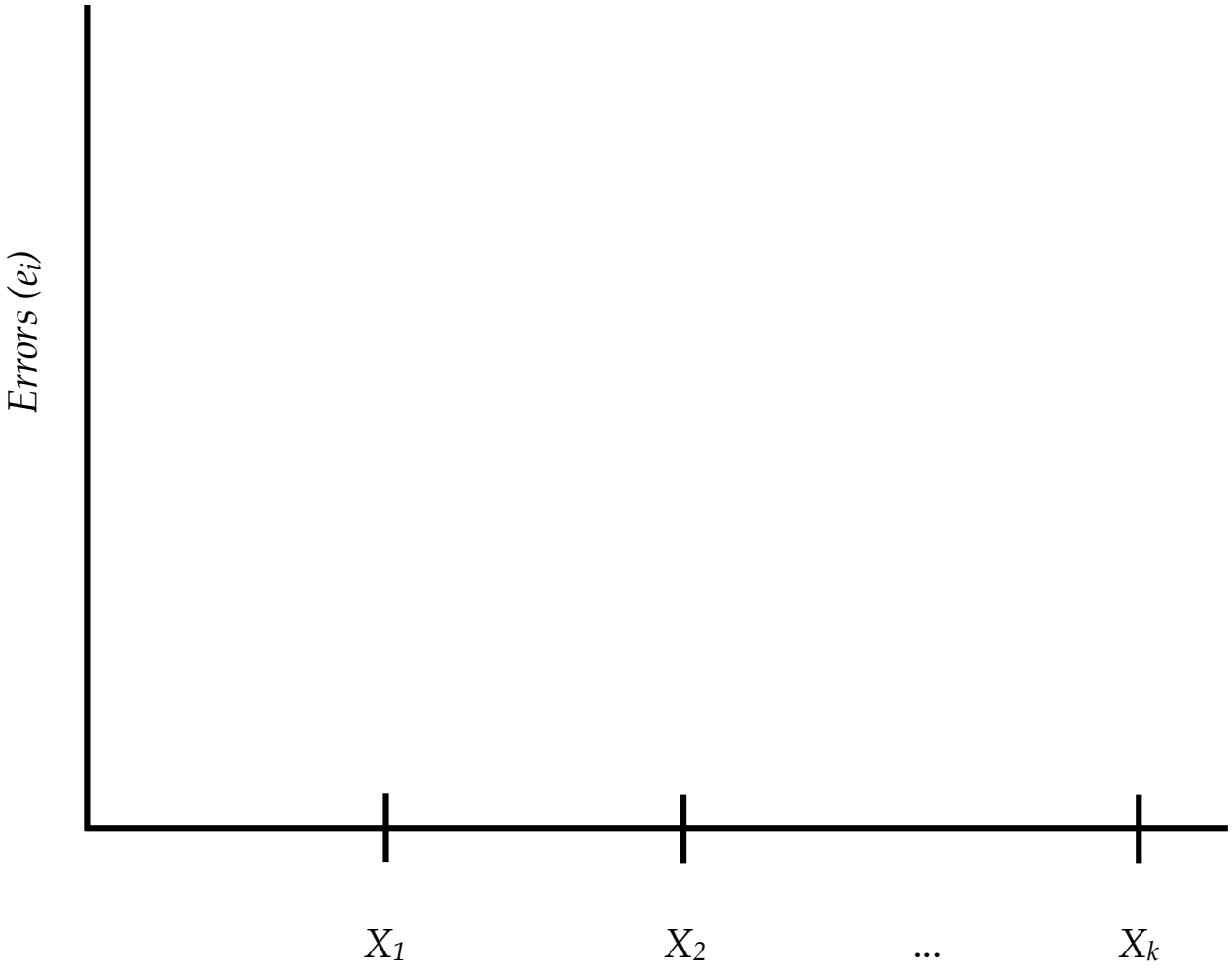
What would the distribution of errors look like at $X = X_1$?



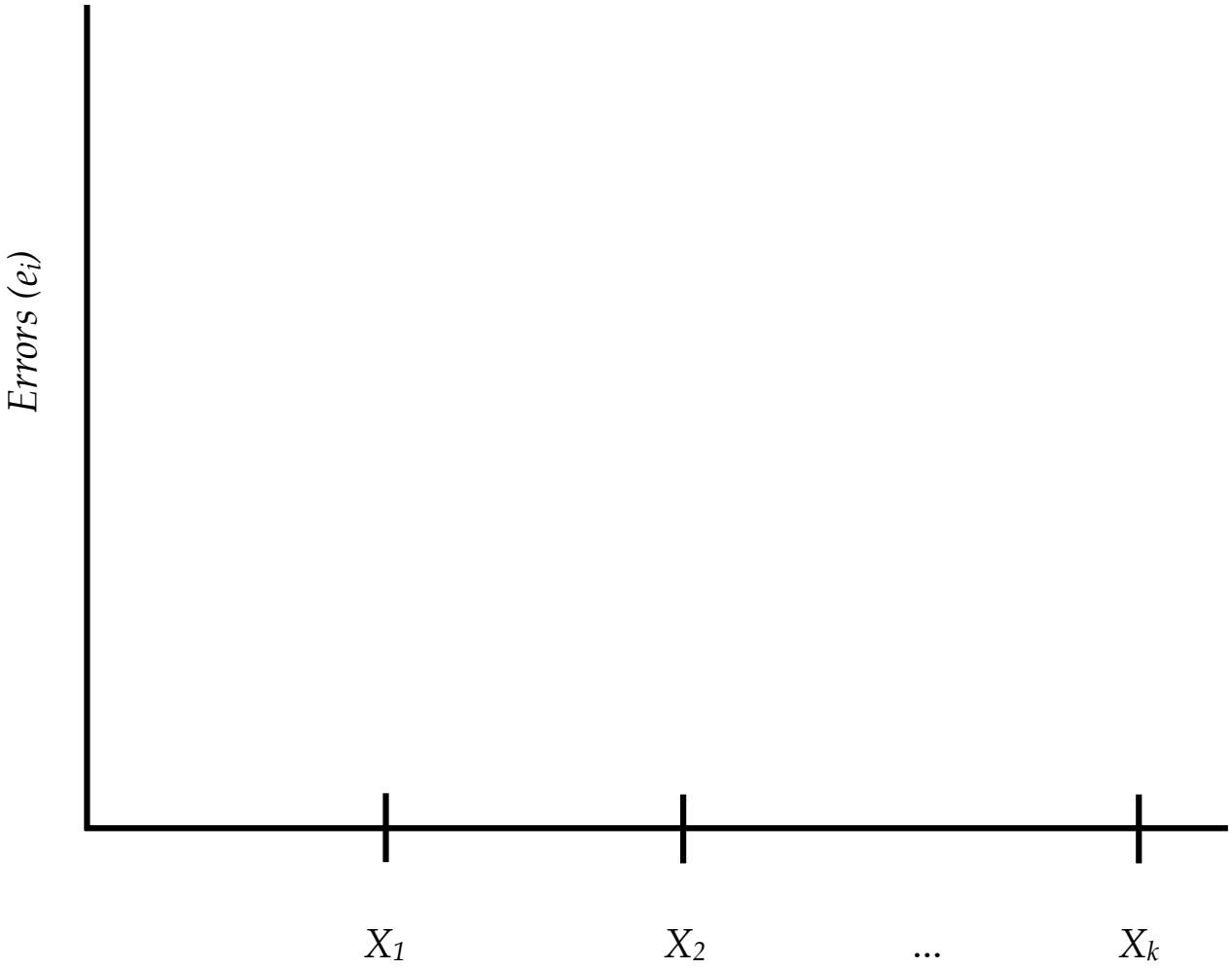
Consider the distribution of *Y* values at X_2

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.

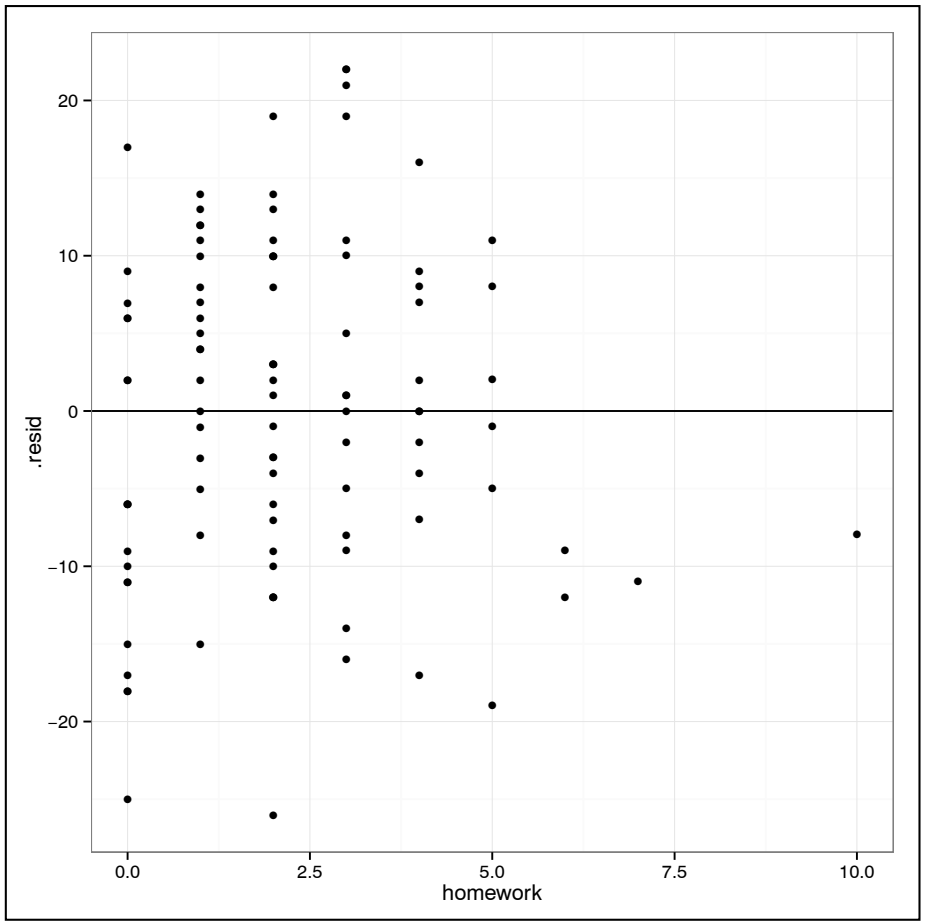


What would a scatterplot of the errors vs. X look like? Sketch it.



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_a, aes(x = homework, y = .resid)) +
    geom_point() +
    theme_bw() +
    geom_hline(yintercept = 0)
```



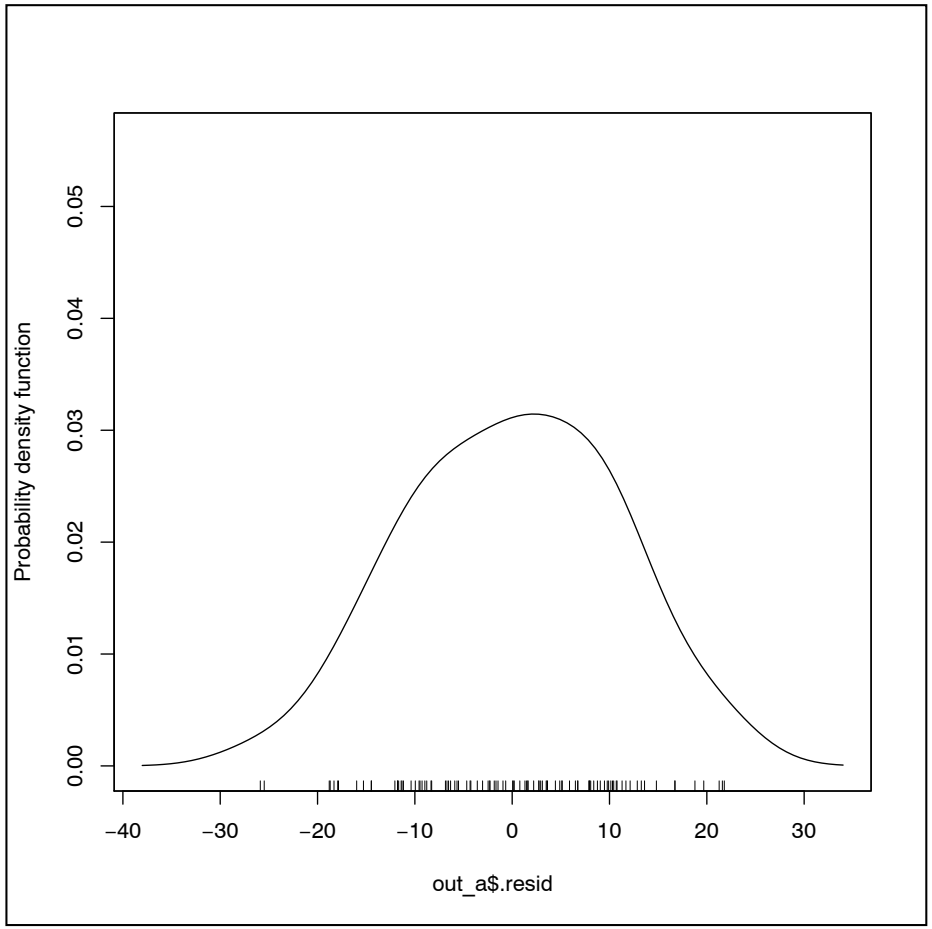
Do the conditional distributions of the errors suggest that the conditional distributions in the population might be normal?

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.

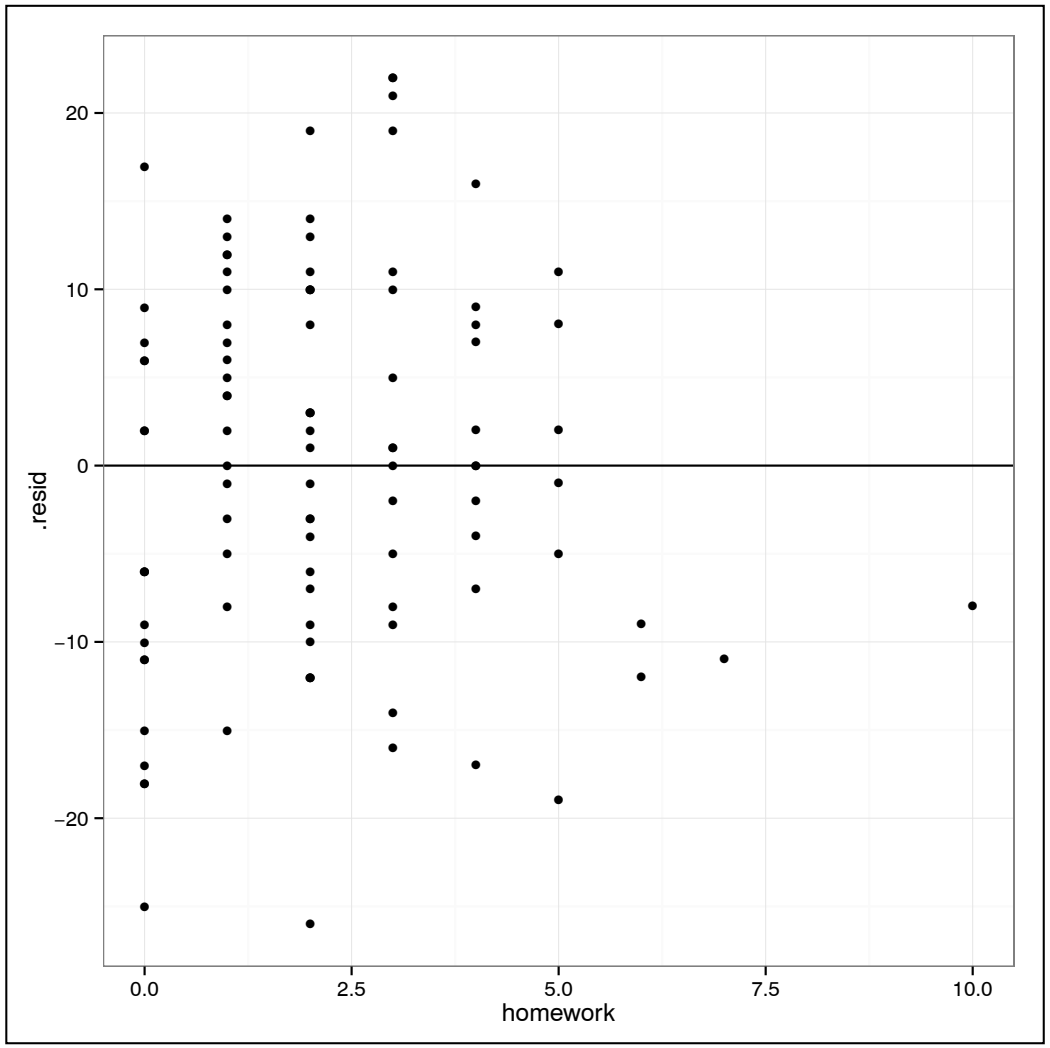
```
# Plot the marginal distribution of the errors
> library(sm)
> sm.density(out_a$.resid)
```



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)

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 homework = 0 is slightly larger than the ranges at homework = 5 or homework = 10.

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

Have the Model Assumptions been Met?

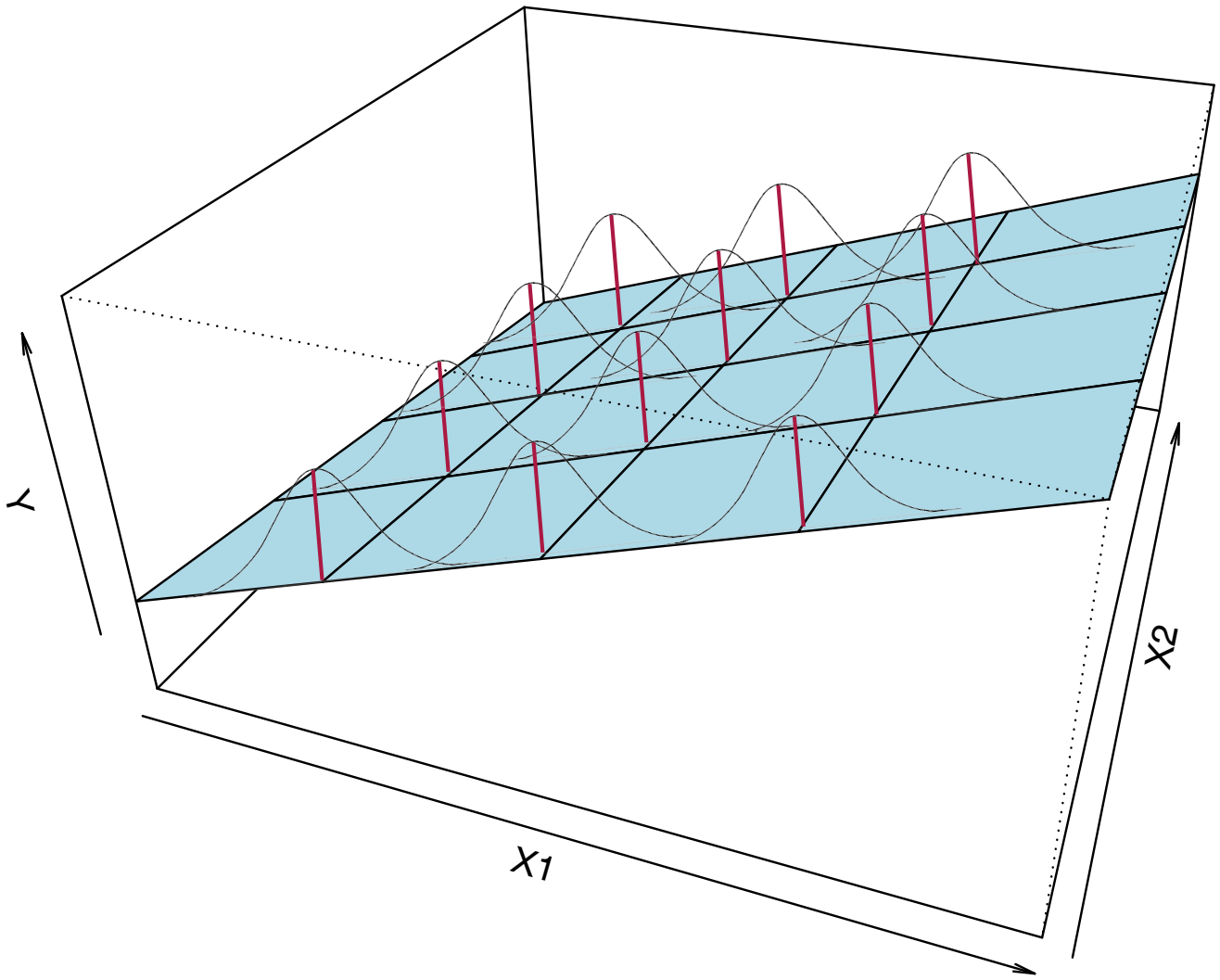
- ☑ Independence
- ☑ The mean of each conditional distribution of the residuals is zero.
- ☑ The conditional distributions of the residuals are normally distributed.
- ☑ Each conditional distributions of the residuals has the same variance.

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.

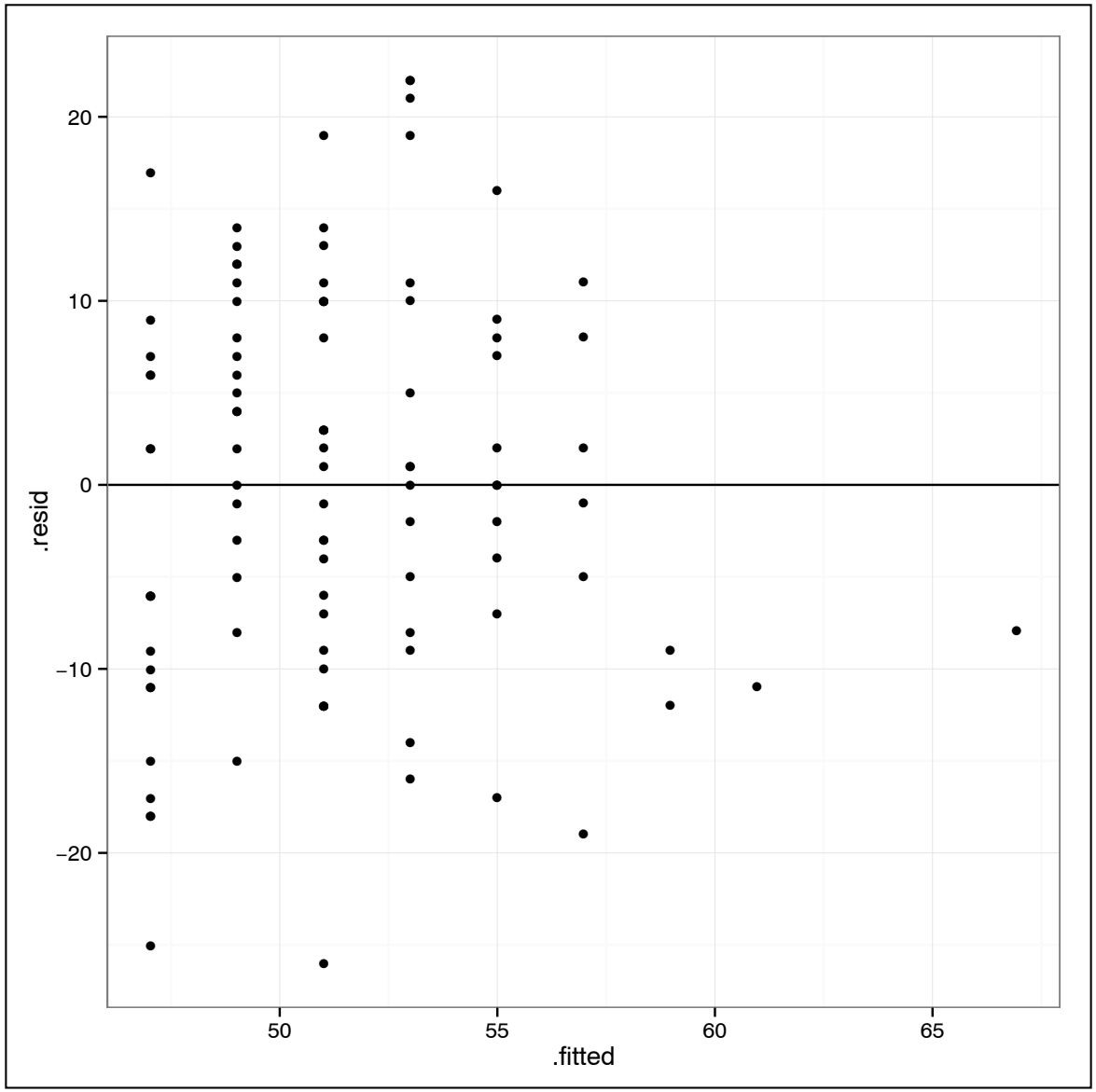
We will look at transformations as one possible solution in a later set of notes.

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)

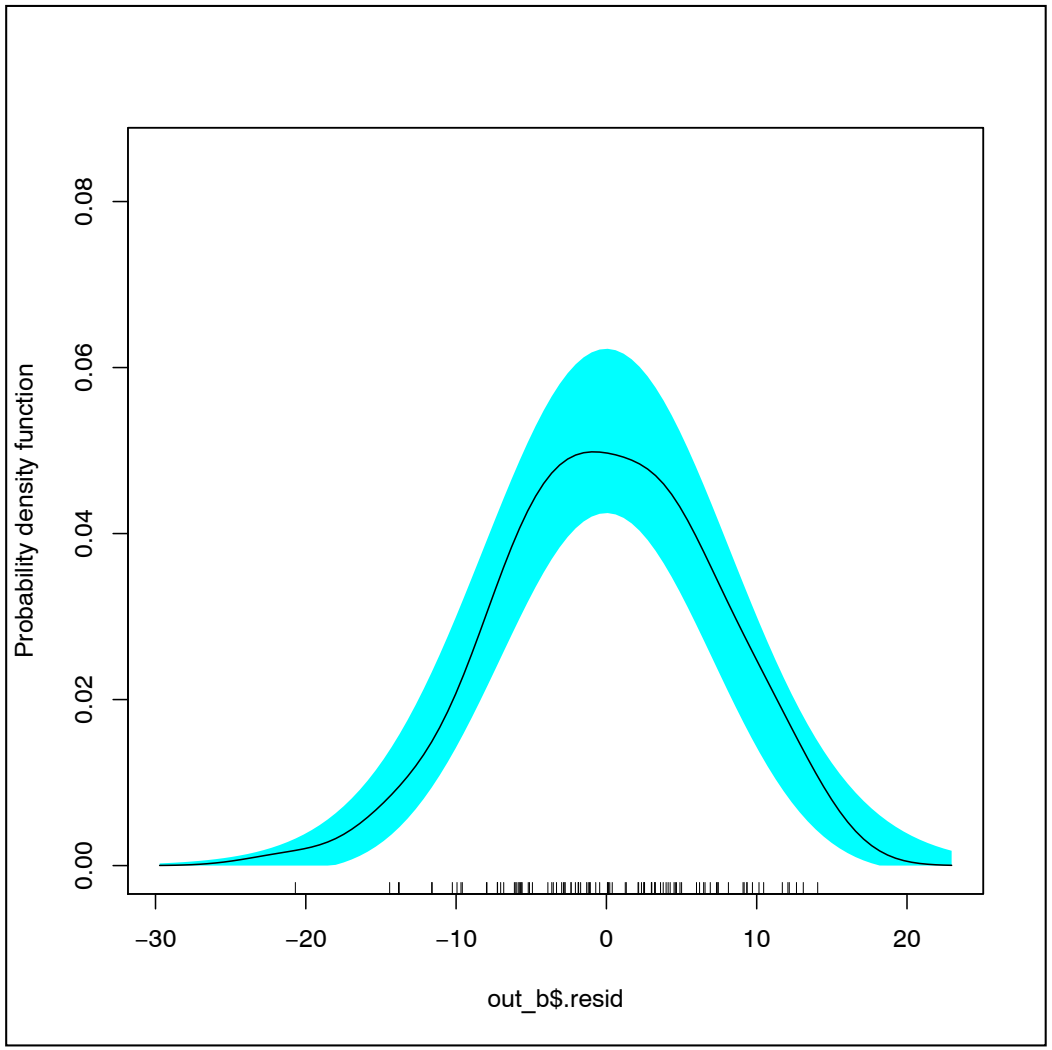
When there is more than one predictor in the model we create the **residual plot** by plotting the errors vs. the predicted values.



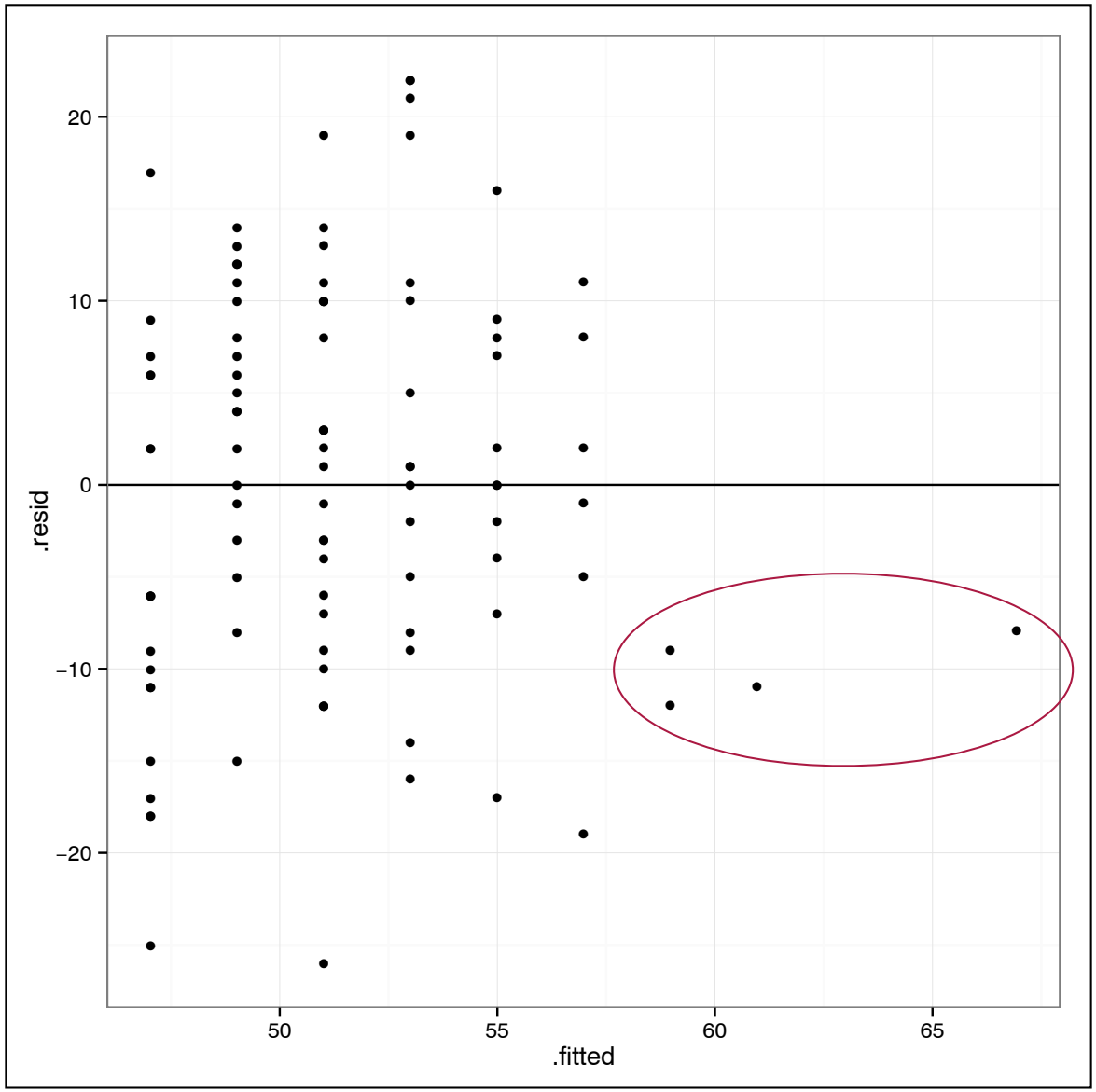
The fitted values represent the combinations of X_1 and X_2

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.

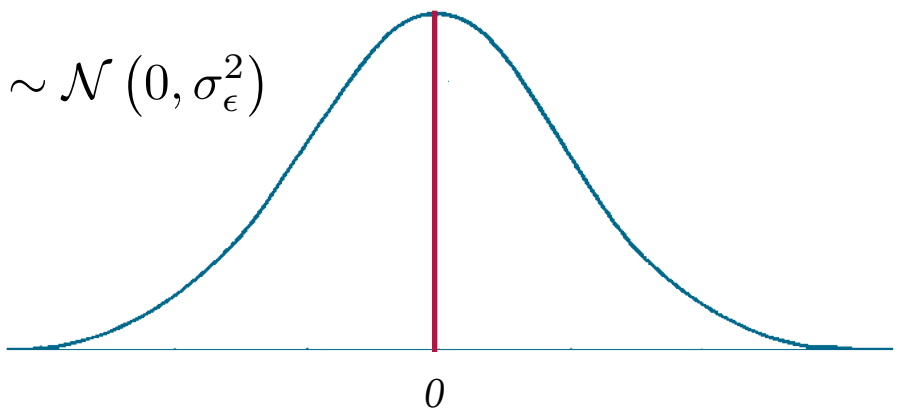
If we pick any fitted value, we should see this principle in the plot.

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).

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.

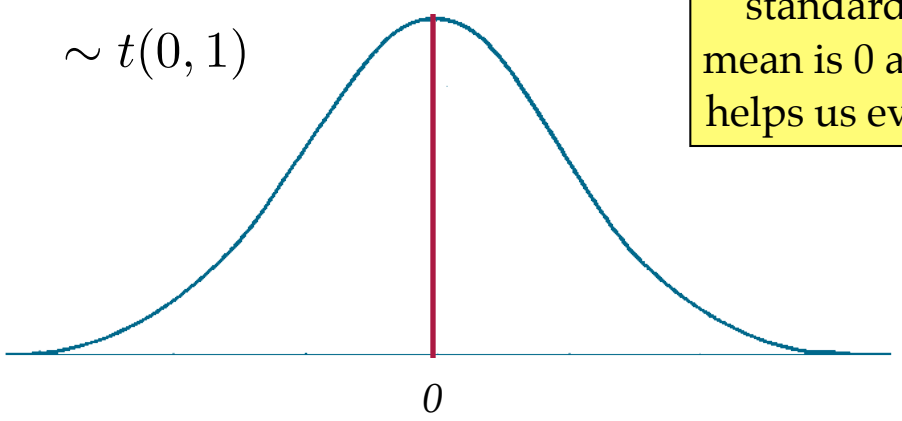
$$Z_{\epsilon_i} = \frac{\epsilon_i - 0}{\text{SD}}$$

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.

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

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



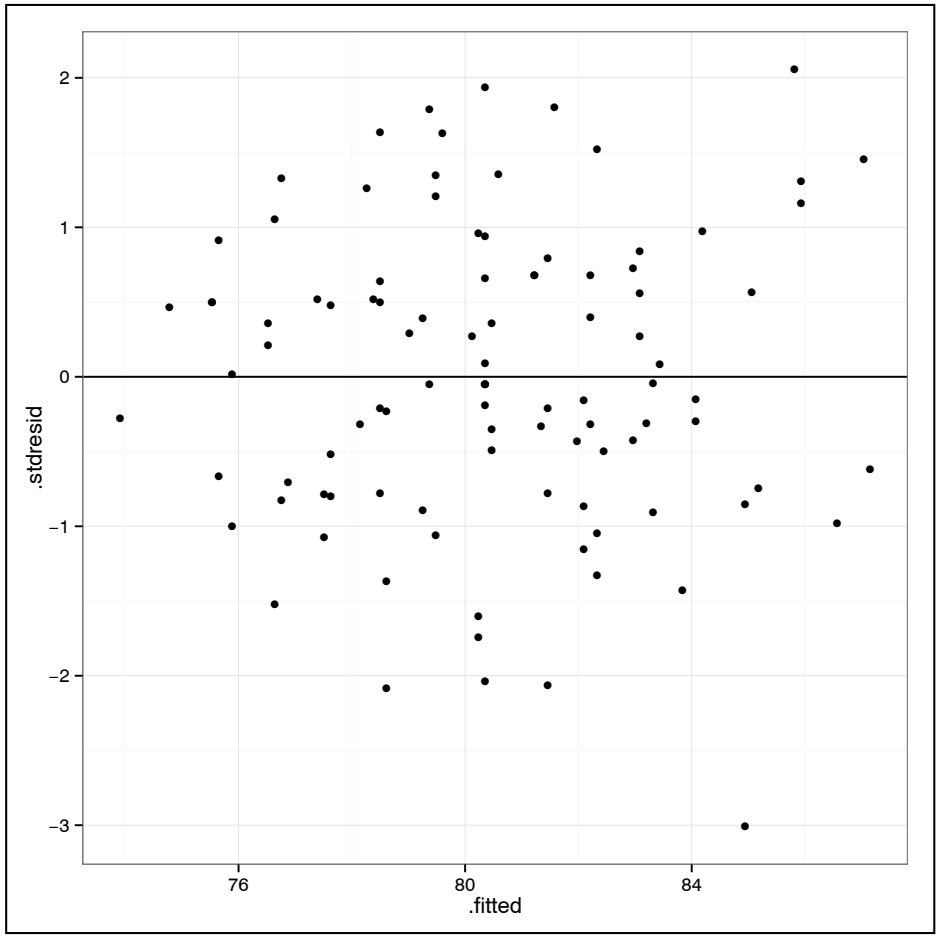
Given the assumption of normality, what percentage of the standardized residuals within 2 standard errors of the mean?

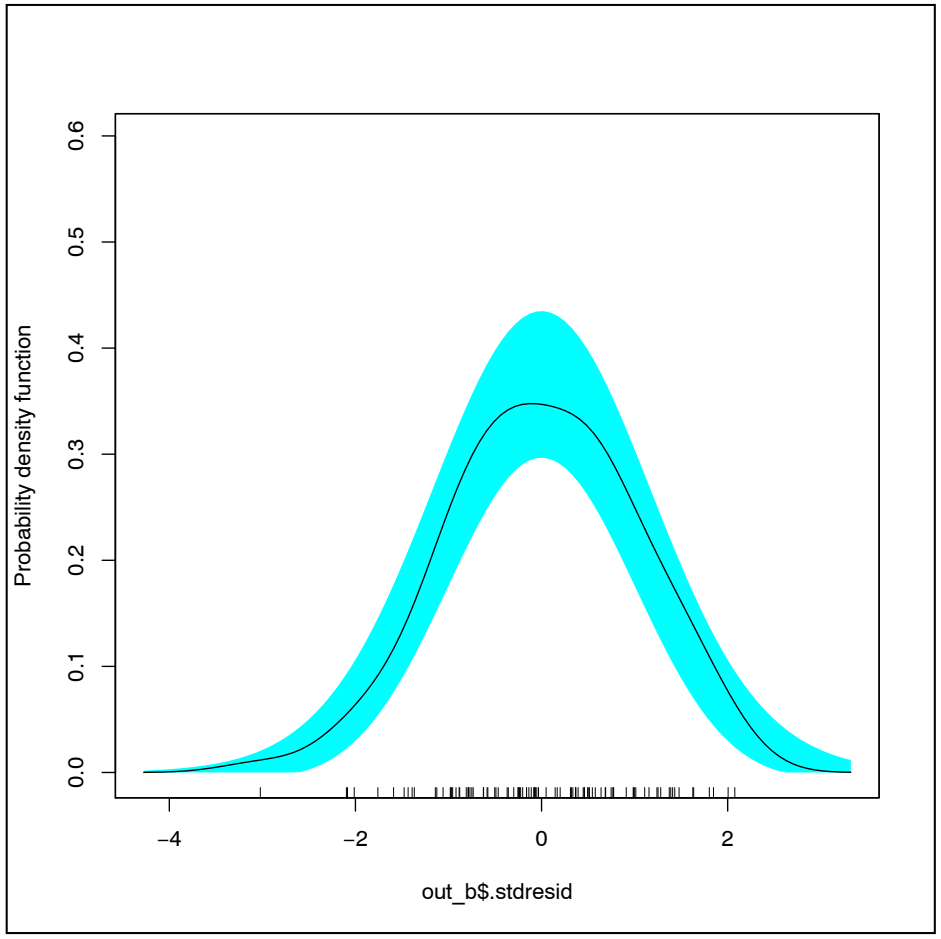
These are the standardized (studentized) errors.

```
# fortified data
> head(out_a)

  achievement homework       .hat   .sigma      .cooksd  .fitted      .resid  .stdresid
1          54        2 0.01012270 10.79802 0.0003992961 51.01196   2.988037  0.2794506
2          53        0 0.02484663 10.78488 0.0040292826 47.03160   5.968405  0.5623823
3          53        4 0.01993865 10.80040 0.0003566990 54.99233  -1.992331 -0.1872599
4          56        0 0.02484663 10.76290 0.0090979090 47.03160   8.968405  0.8450619
5          59        2 0.01012270 10.77152 0.0028536673 51.01196   7.988037  0.7470664
6          30        0 0.02484663 10.65944 0.0328111777 47.03160 -17.031595 -1.6048286
```

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)





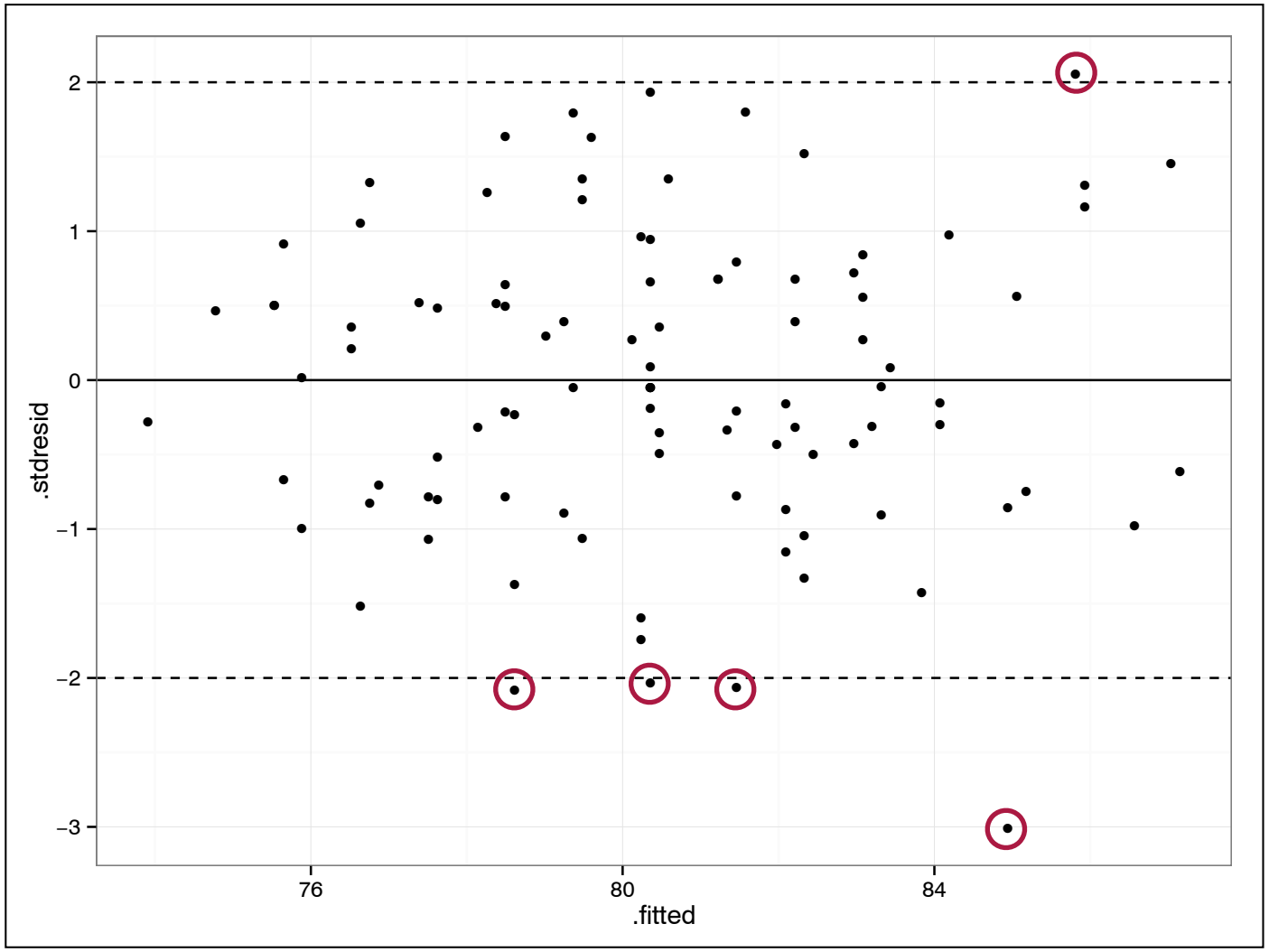
Since standardizing is a linear transformation, the shapes and relationships we examine will be identical whether you use the raw or standardized residuals.

Examining Observations that have an Extreme Residual

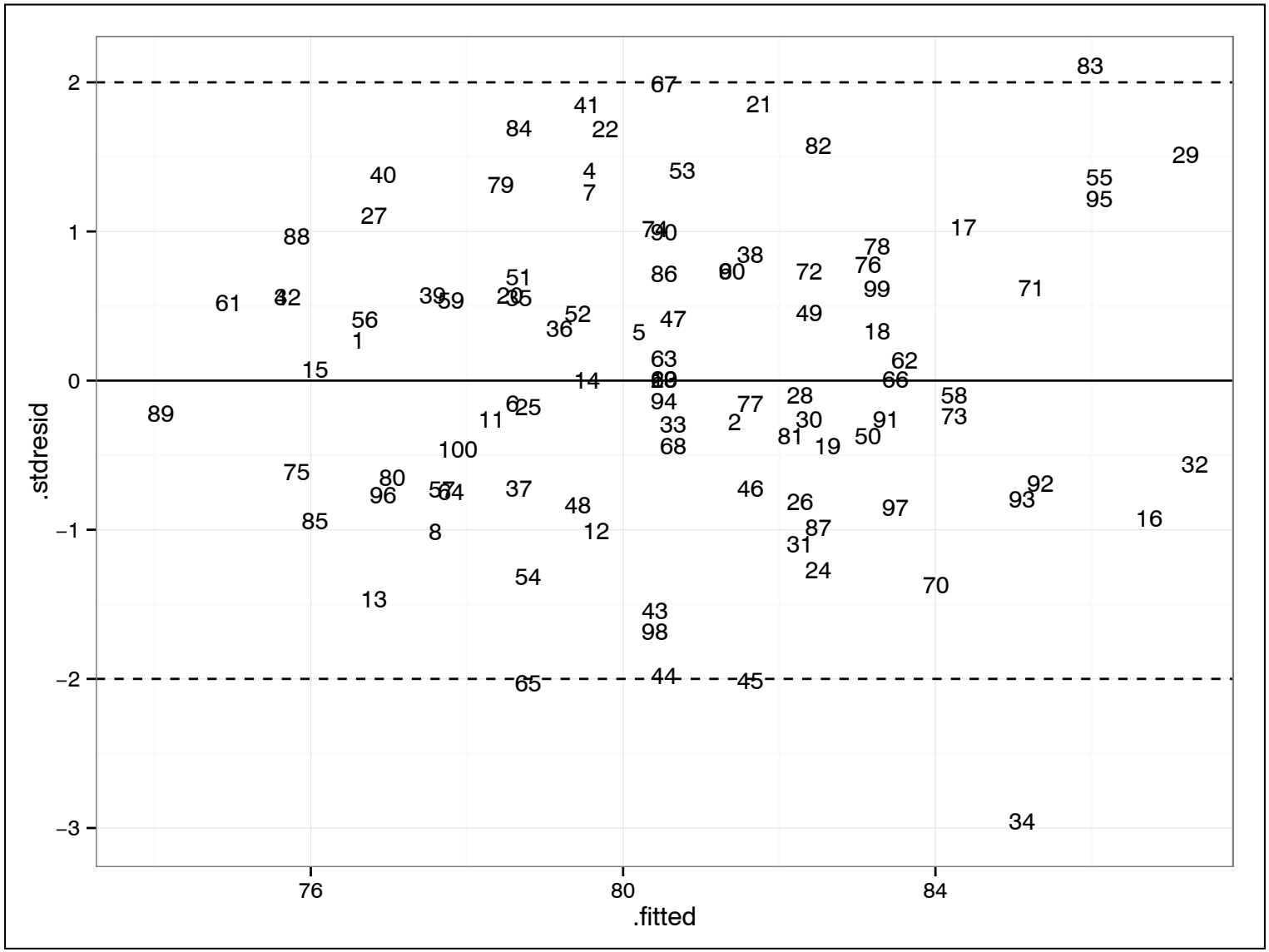
When examining the residual plot, it can be helpful to add horizontal lines at ±2. This helps us easily see observations that have an outcome value that is more than 2 standard errors from where we would predict the outcome to be given their X values.

These 5 observations have "large" residuals relative to the rest of the data. One has a "really large" residual.

Given these students' homework and parent education values, we would predict them to have a much higher (4 obs) or lower (1 obs) GPA than they actually do.



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



If you need the actual value of the standardized (or raw) residual, you need to to go back to the data.

```
# Load dplyr library
> library(dplyr)

# Get observations with standardized residual >= 2
> out_b %>% filter(.stdresid >= 2)

  gpa homework parentEd       .hat   .sigma    .cooksd  .fitted   .resid .stdresid
1 100        7       18 0.05414066 6.971209 0.08073007 85.81316 14.18684  2.056979

# Get observations with standardized residual <= -2
> out_b %>% filter(.stdresid <= -2)

  gpa homework parentEd       .hat   .sigma    .cooksd  .fitted    .resid .stdresid
1  64        7       17 0.03668296 6.787611 0.11491528 84.94254 -20.94254 -3.008860
2  66        5       14 0.01001950 6.974672 0.01396339 80.35498 -14.35498 -2.034447
3  67        7       13 0.02556513 6.969870 0.03731427 81.46004 -14.46004 -2.065620
4  64        5       12 0.02183960 6.967067 0.03231023 78.61373 -14.61373 -2.083595
```

It would really help if there were a variable that told us the row number or ID each of these observations were in the original data!!!!

```
# Add an ID number to the fortified data
> out_b$id = 1:100

# Examine data
> head(out_b)

  gpa homework parentEd       .hat   .sigma      .cooksd  .fitted    .resid  .stdresid id
1  78        2       13 0.03298270 7.126761 0.0005115121 76.52082  1.479185  0.2121107  1
2  79        6       14 0.01218886 7.124353 0.0004544506 81.34282 -2.342821 -0.3323988  2
3  79        1       13 0.05000447 7.119163 0.0044144279 75.53297  3.467030  0.5015962  3
4  89        5       13 0.01299390 7.061067 0.0080051578 79.48435  9.515648  1.3506281  4
5  82        3       16 0.03899033 7.125728 0.0009884747 80.12053  1.879470  0.2703515  5
6  77        4       13 0.01447741 7.126753 0.0002212628 78.49651 -1.496507 -0.2125703  6

# Get observations with standardized residual >= 2
> out_b %>% filter(.stdresid >= 2)

  gpa homework parentEd       .hat   .sigma    .cooksd  .fitted   .resid .stdresid id
1 100        7       18 0.05414066 6.971209 0.08073007 85.81316 14.18684  2.056979 83

# Get observations with standardized residual <= -2
> out_b %>% filter(.stdresid <= -2)

  gpa homework parentEd       .hat   .sigma    .cooksd  .fitted    .resid .stdresid id
1  64        7       17 0.03668296 6.787611 0.11491528 84.94254 -20.94254 -3.008860 34
2  66        5       14 0.01001950 6.974672 0.01396339 80.35498 -14.35498 -2.034447 44
3  67        7       13 0.02556513 6.969870 0.03731427 81.46004 -14.46004 -2.065620 45
4  64        5       12 0.02183960 6.967067 0.03231023 78.61373 -14.61373 -2.083595 65
```

Student #34 in the data has an extremely large, negative residual. Given his/her homework value of 7 and parent education value of 17 we would predict that students GPA to be much higher (hat(GPA) = 84.9) than it is (GPA = 64).

Should we remove this student from the data or not?

Removing Case from Data and Refitting Regression

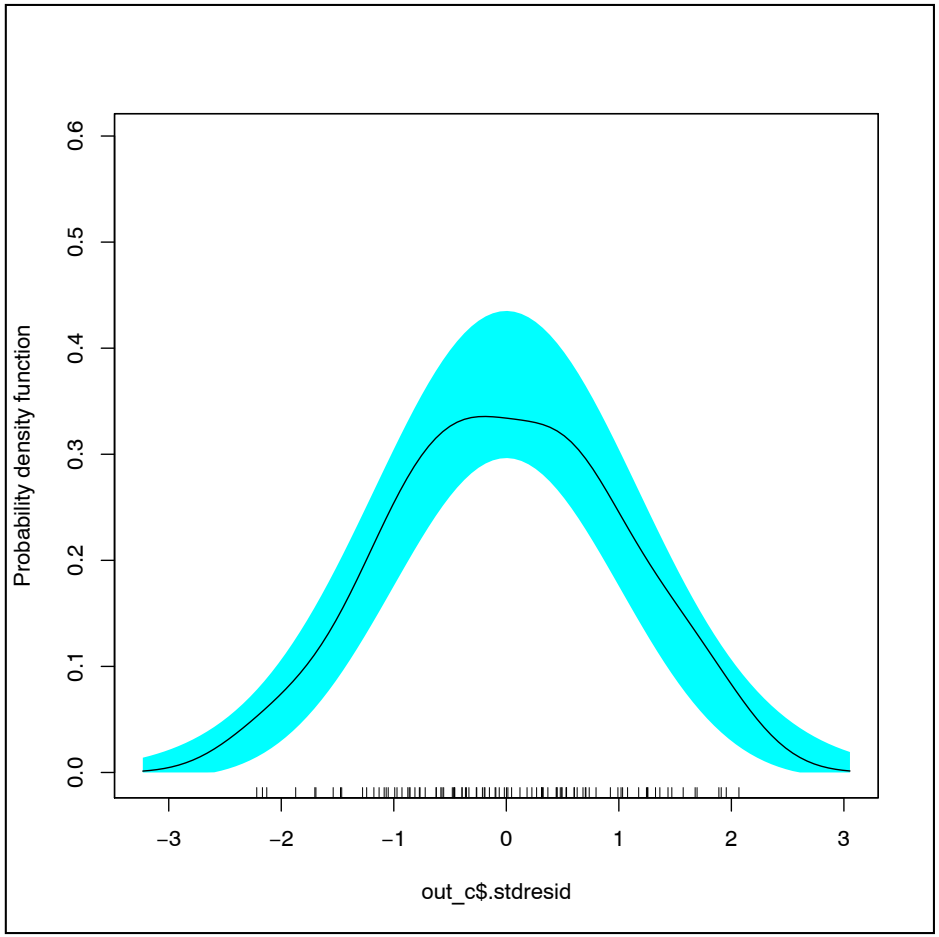
If you decide to remove a case from the data, the regression models need to be re-fitted, and *all* of the assumptions need to be re-checked.

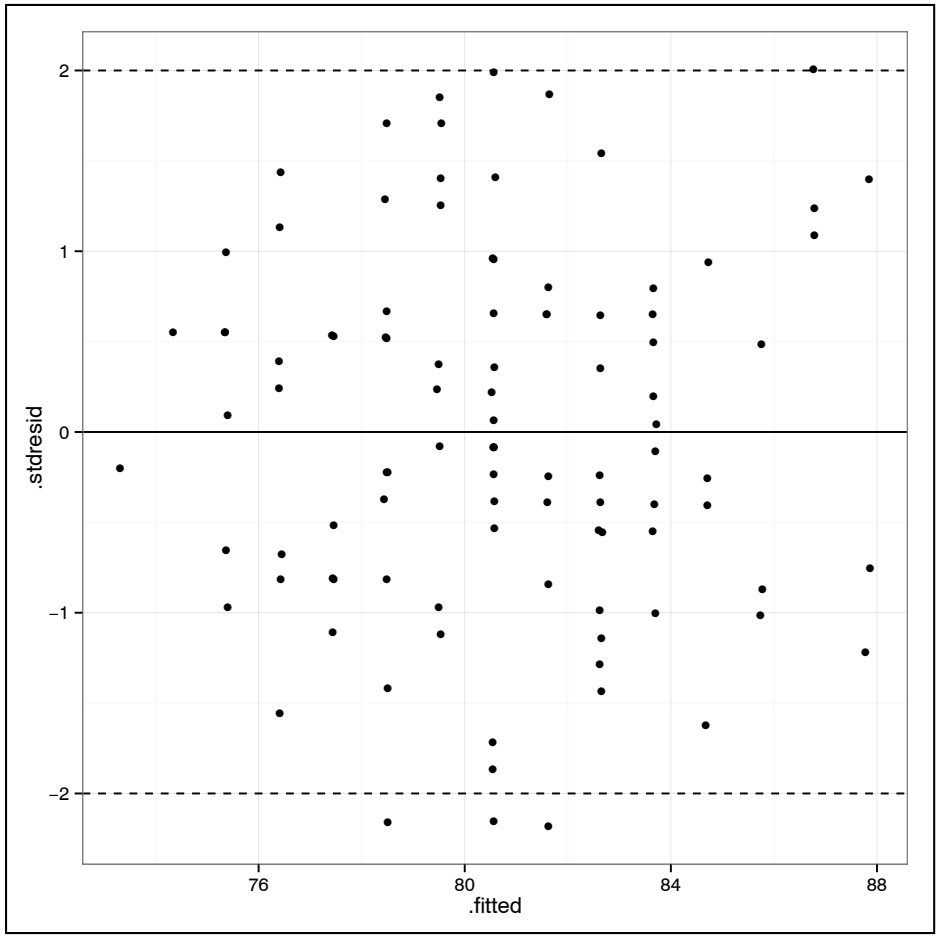
```
# Add an ID number to the original data
> multReg$id = 1:100

# Examine data
> head(multReg)

  gpa parentEd homework id
1  78       13        2  1
2  79       14        6  2
3  79       13        1  3
4  89       13        5  4
5  82       16        3  5
6  77       13        4  6

# Double-check that Case 34 is the problematic observation
> multReg[34, ]

   gpa parentEd homework id
34  64       17        7 34
```

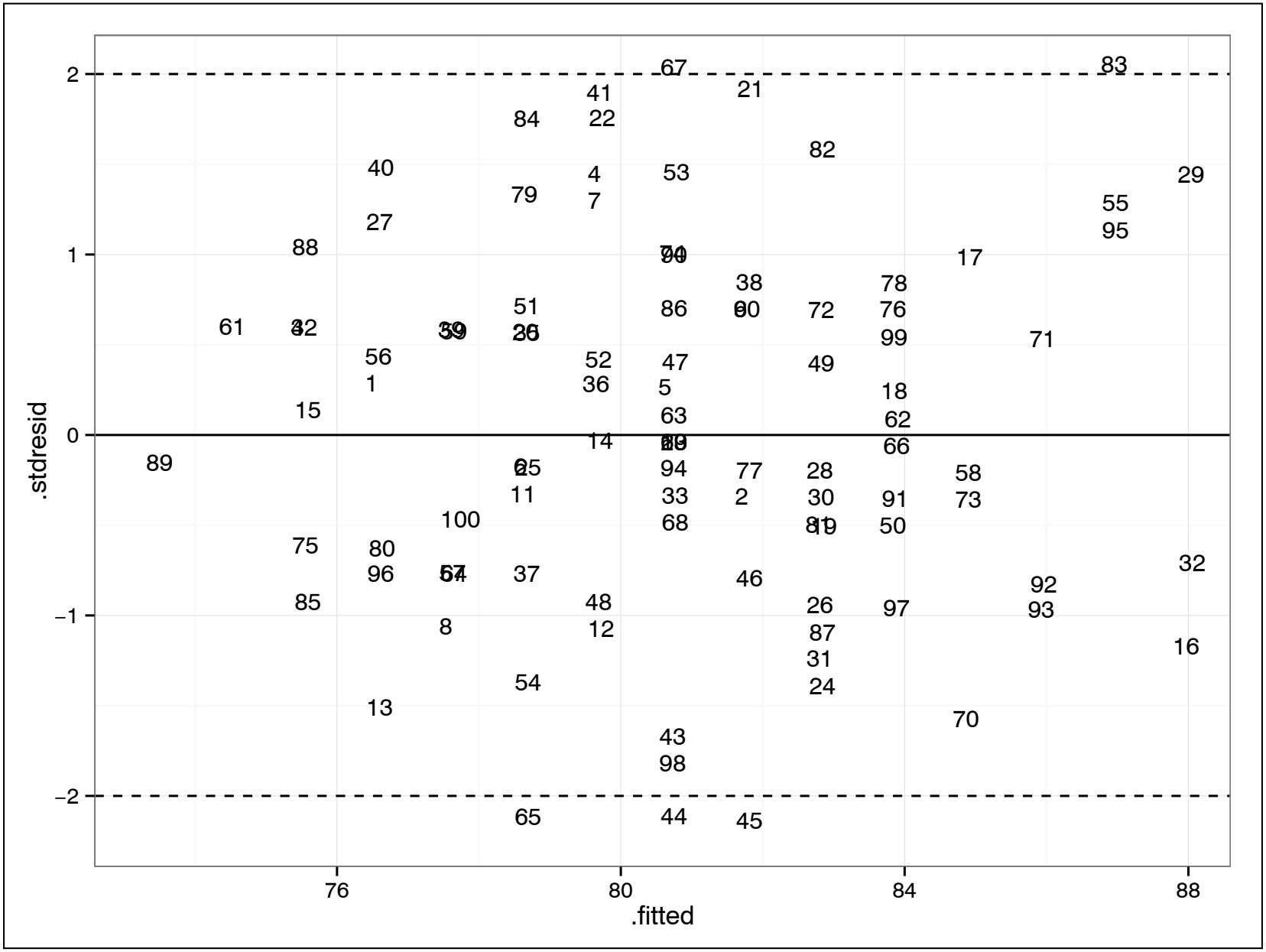
```
# Create a new data frame that removes row 34
> multReg2 = multReg[-c(34), ]

# Examine data
> multReg

    gpa parentEd homework id
      ⋮        ⋮        ⋮  ⋮
31   74       16        5 31
32   83       15       11 32
33   78       13        6 33
35   82       13        4 35
36   81       17        1 36
      ⋮        ⋮        ⋮  ⋮

# Re-fit regression model
> lm.c = lm(gpa ~ homework + parentEd, data = multReg2)
> summary(lm.c)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  60.9332     5.0680  12.023  < 2e-16 ***
homework      1.0461     0.3459   3.024  0.00320 **
parentEd      1.0285     0.3712   2.771  0.00671 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.788 on 96 degrees of freedom
Multiple R-squared:  0.1927,   Adjusted R-squared:  0.1759
F-statistic: 11.46 on 2 and 96 DF,  p-value: 3.44e-05
```

	Model B		Model C	
Coefficient	B	SE	B	SE
Homework	0.99**	0.36	1.05**	0.35
Parent education	0.87*	0.38	1.03**	0.37
(Intercept)	63.24***	5.24	60.93***	5.07
R^2	0.135		0.193	

Which values changed?


Probability density function
0.0
0.1
0.2
0.3
0.4
0.5
0.6
−3
−2
−1
0
1
2
3
out_c$.stdresid


.stdresid
2
1
0
−1
−2
76
80
84
88
.fitted



Now observations have "large" residuals that did not have large residuals before…what should we do?

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).

Correlation Between the Fitted Values and the Residuals

```
# For Model A (simple regression)
> cor(out_a$.fitted, out_a$.resid)

[1] -2.914861e-16

# For Model B (multiple regression)
> cor(out_a$.fitted, out_a$.resid)

[1] -6.630209e-17
```

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

Correlation Between the Fitted Values and the Outcome

```
# For Model A (simple regression)
> cor(out_a$.fitted, out_a$achievement)

[1] 0.3199936

> cor(math$homework, out_a$achievement)

[1] 0.3199936


# For Model B (multiple regression)
> cor(out_a$.fitted, out_a$gpa)

[1] 0.3899378
```

In a simple regression, the correlation between the fitted values and the outcome is exactly the same as the correlation coefficient between X and Y. In a multiple regression the correlation between the fitted values and the outcome is called the **multiple correlation coefficient**. If we square this value (in any regression) we obtain the R^2 value.