PM 592
Regression Analysis for Public Health Data Science

Week 4

Regression II

Regression II

Checking Assumptions

ANOVA

Transformations

Categorical Binary Predictors

Categorical Nominal Predictors

Lecture Objectives

- Assess conformity to the assumptions of linear regression.
- Distinguish between regression/model variance and error variance.
- Determine situations in which a variable transformation is necessary.
- Interpret beta coefficients for binary and nominal predictors.

1. Review

- ✓ The form of a linear regression equation
- ✓ Interpretation of coefficients and p-values
- ✓ Centering and multiplicative transformations
- ✓ Correlation and its relation to regression

2. Checking Linear Regression Assumptions

Last class we discussed the assumptions of linear regression.

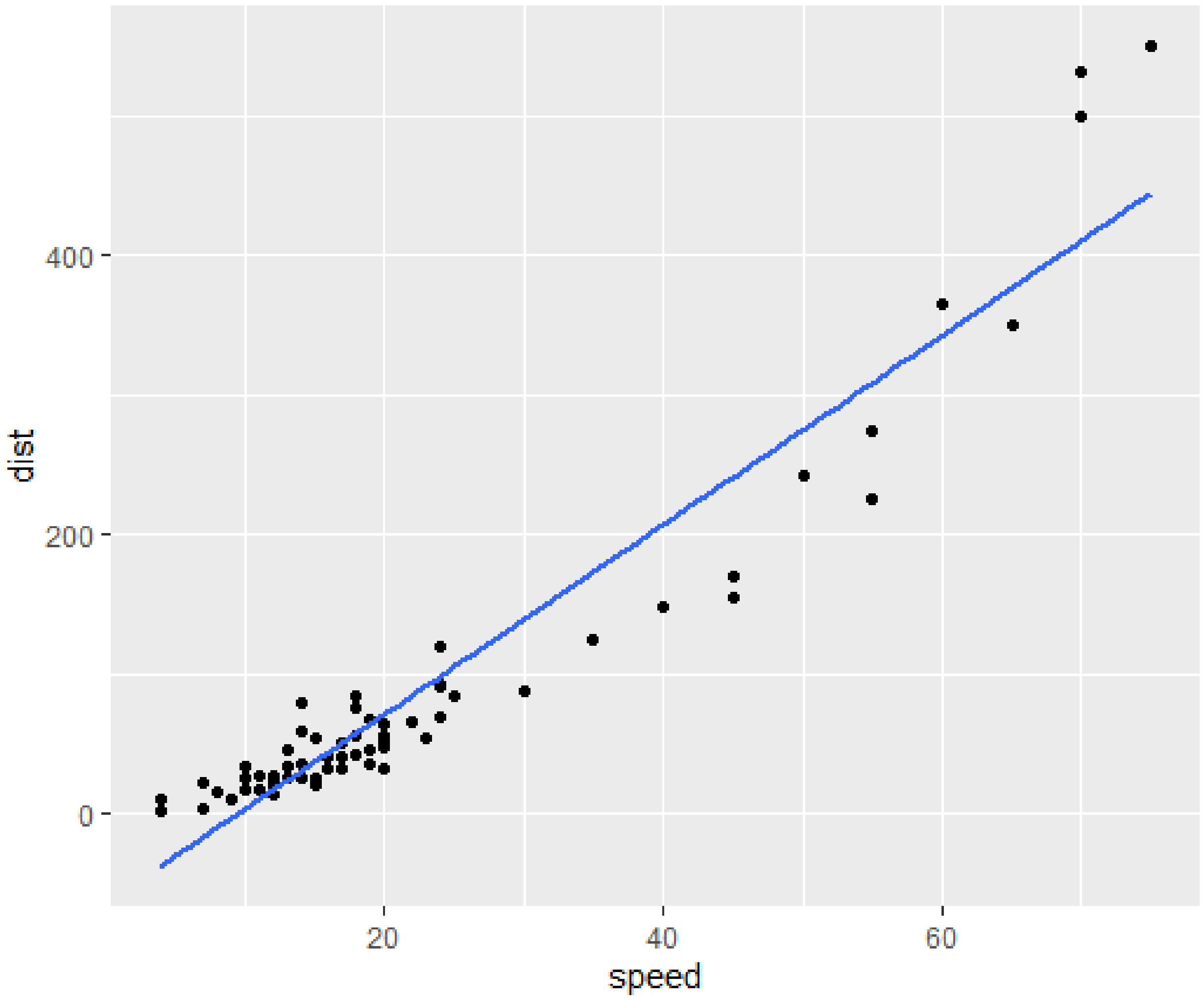
Here, we will go through how to assess these assumptions.

Remember, the assumptions are:

- **Linearity**. Scatterplots should indicate some degree of linearity. If there is nonlinearity, you may be able to transform variables.
- **Independence.** You must assume this based on the study design.
- **Normality.** The residuals should be normally distributed.
- **Equal Variance (Homoscedasticity).** Do the residuals have a common variance across the x values?

Recall our model from last time: is speed (mph) related to stopping distance (feet)?

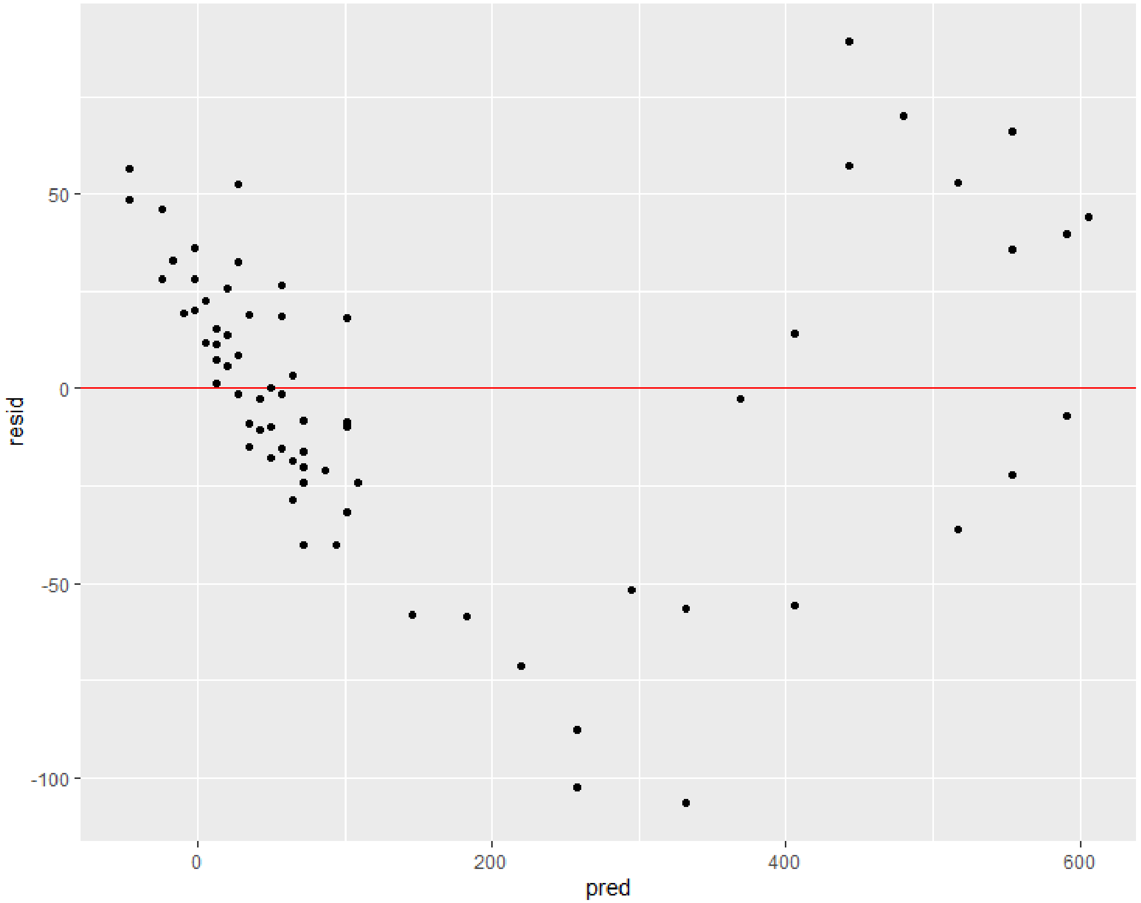
```
> model1 %>% summary()

Call:
lm(formula = dist ~ speed, data = carstot)

Residuals:
     Min       1Q   Median       3Q      Max 
-106.666  -20.336    0.656   26.010   89.104 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -76.1783     7.2667  -10.48 5.42e-16 ***
speed         7.4154     0.1809   40.99  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.

Residual standard error: 39.59 on 70 degrees of freedom
Multiple R-squared:   0.96, Adjusted R-squared:  0.9594
F-statistic:  1680 on 1 and 70 DF,  p-value: < 2.2e-16
```



Linearity

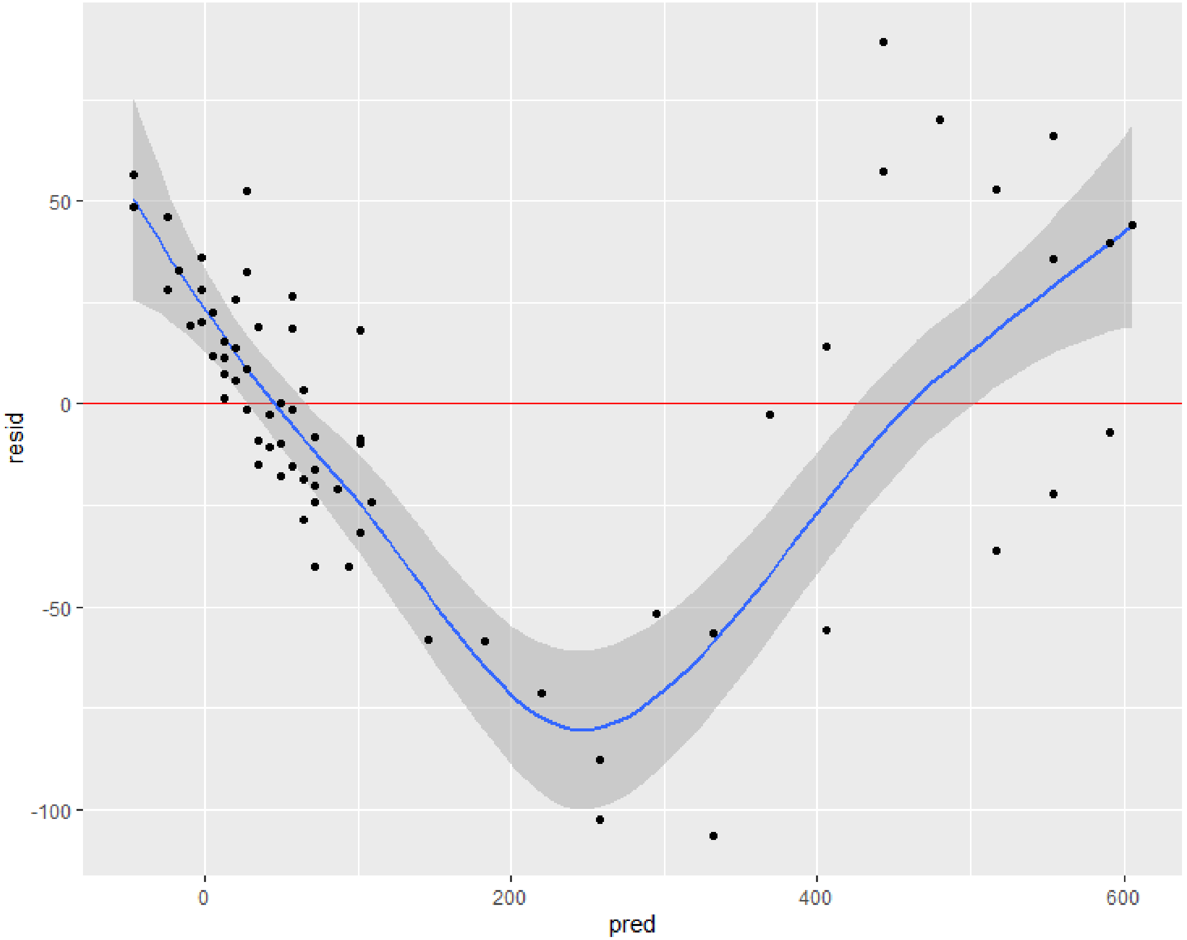
If the relationship is linear, then the residuals will show a flat scatter around 0 when plotted by the predicted value of Y.

Here the residuals droop down, and back up.



Linearity

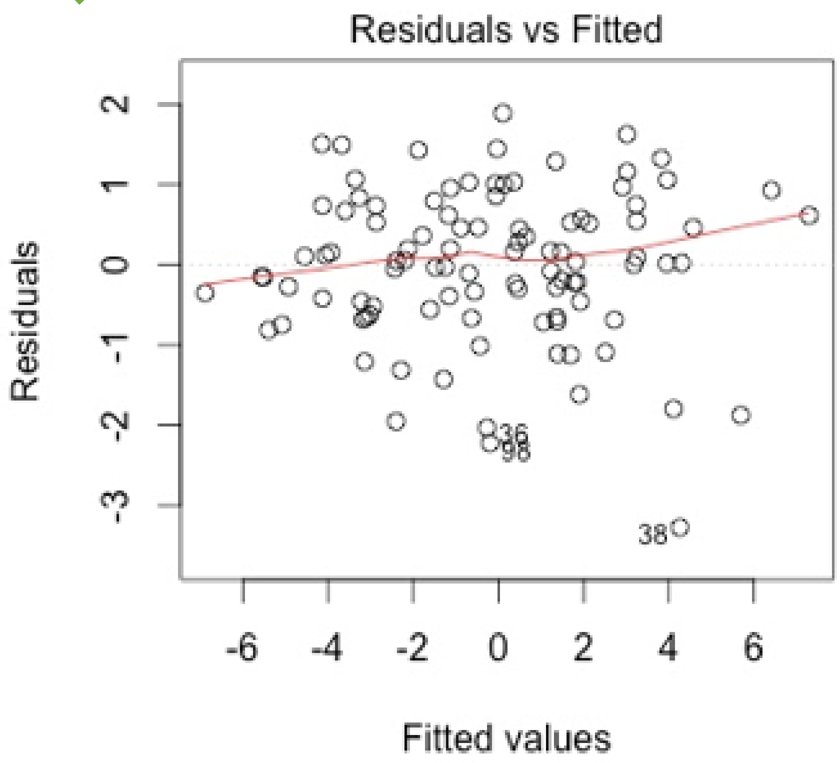
To help us examine the relationship, we can add a LOWESS (locally-weighted scatterplot smoother) line of the relationship.



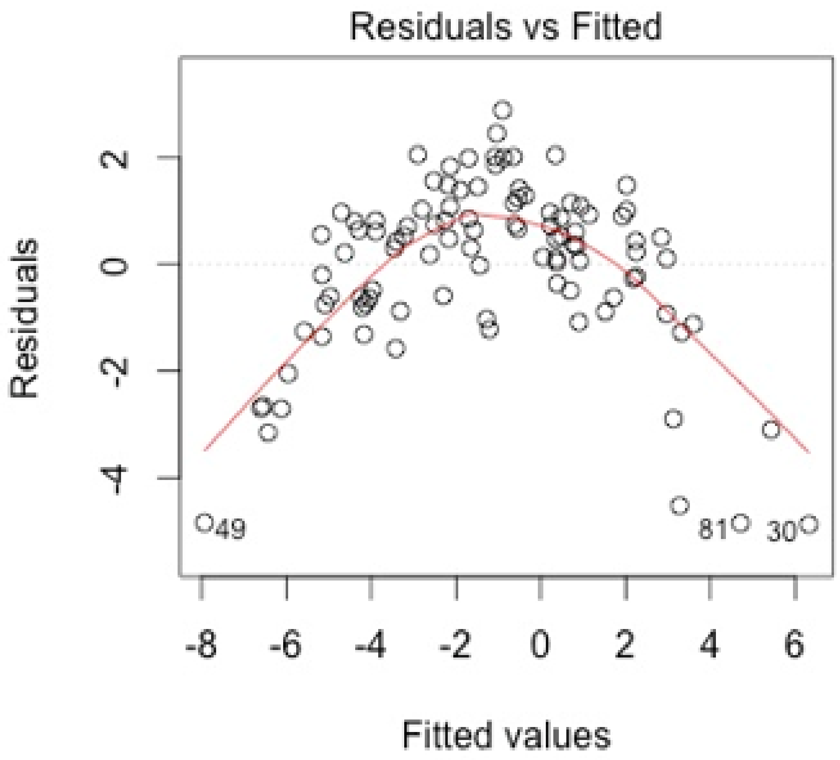
2. Checking Linear Regression Assumptions

Linearity

✓ No Pattern Evident



Non-Linearity Evident ✗



Normality

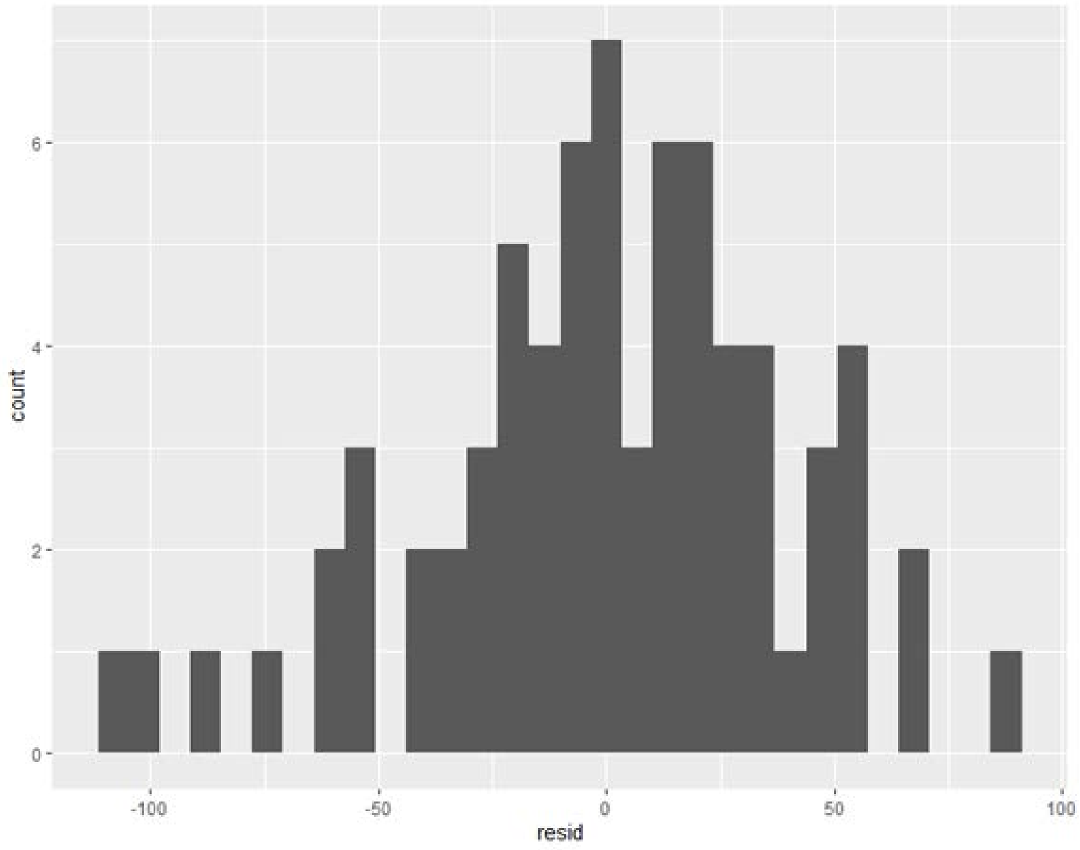
We can evaluate the normality of residuals in the same way we would typically examine normality.

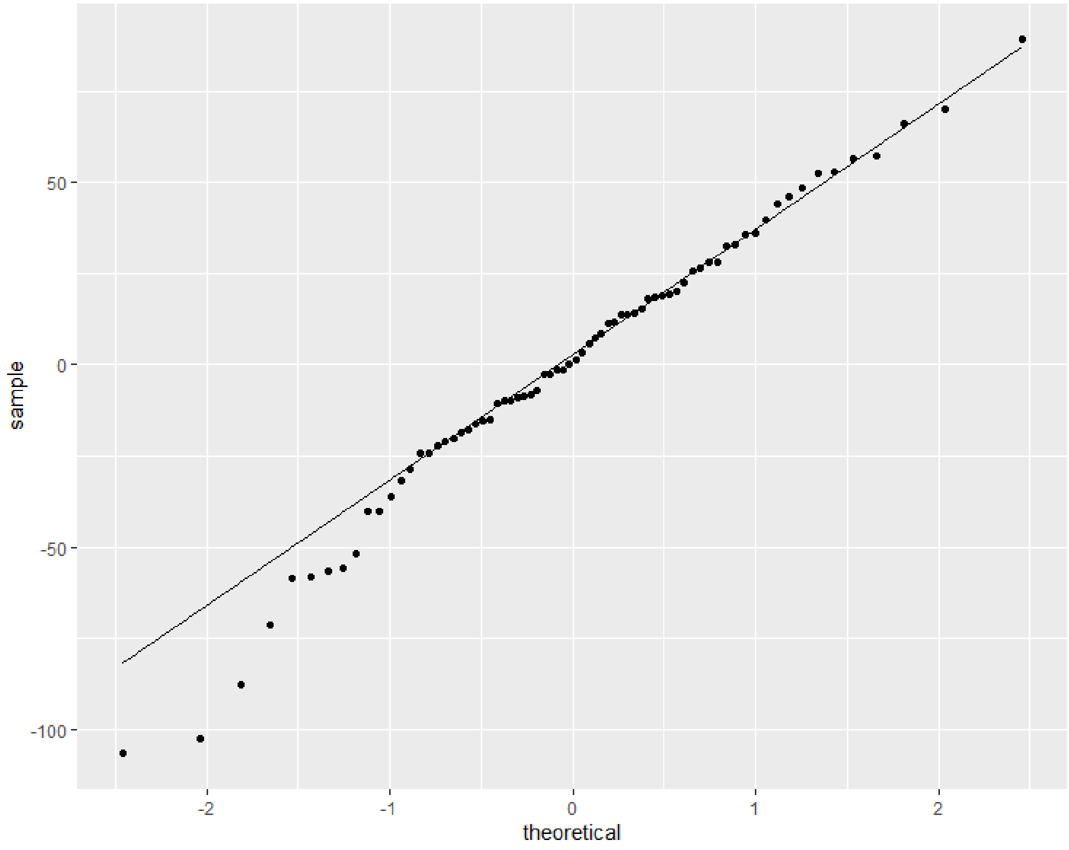
First, examining residual statistics:

```
> carstot_model1 %>%
+   select(resid) %>%
+   psych::describe()
   vars  n mean    sd median trimmed   mad     min  max  range  skew kurtosis   se
X1    1 72    0 39.32   0.66    1.68 33.16 -106.67 89.1 195.77 -0.43     0.21 4.63

> shapiro.test(carstot_model1$resid)

       Shapiro-Wilk normality test

data:  carstot_model1$resid
W = 0.98256, p-value = 0.4191
```

Normality

Then, a histogram and QQ plot for the residuals:





Normality

The Central Limit Theorem makes the inference robust to non-normality of residuals when the sample size is large enough (a few hundred or greater).

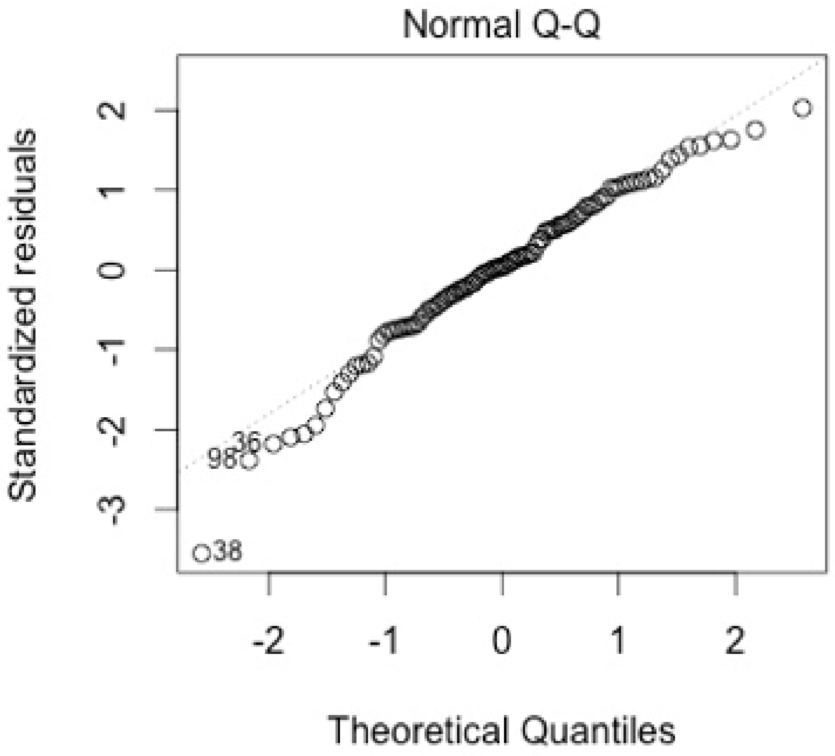
Normality

General Guidelines:

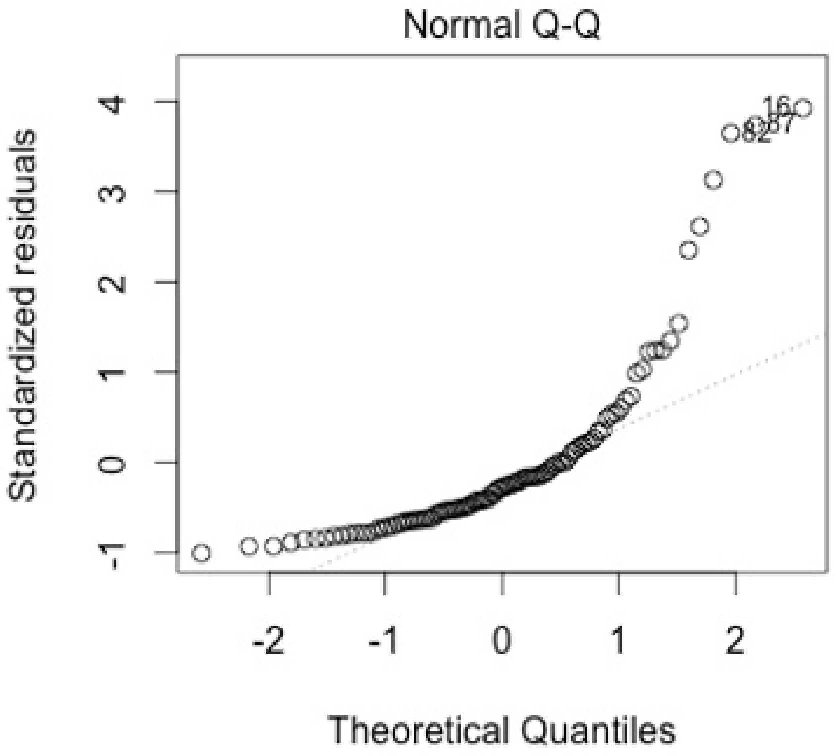
- Are the median and mean within 20% of 1 SD?
- Are skewness and kurtosis < |1|?
- Does a histogram of the residuals look normal?
- Does the Q-Q plot follow a straight line?
- Is the Shapiro-Wilk test not rejected?

Normality

✓ Normal Distribution Evident

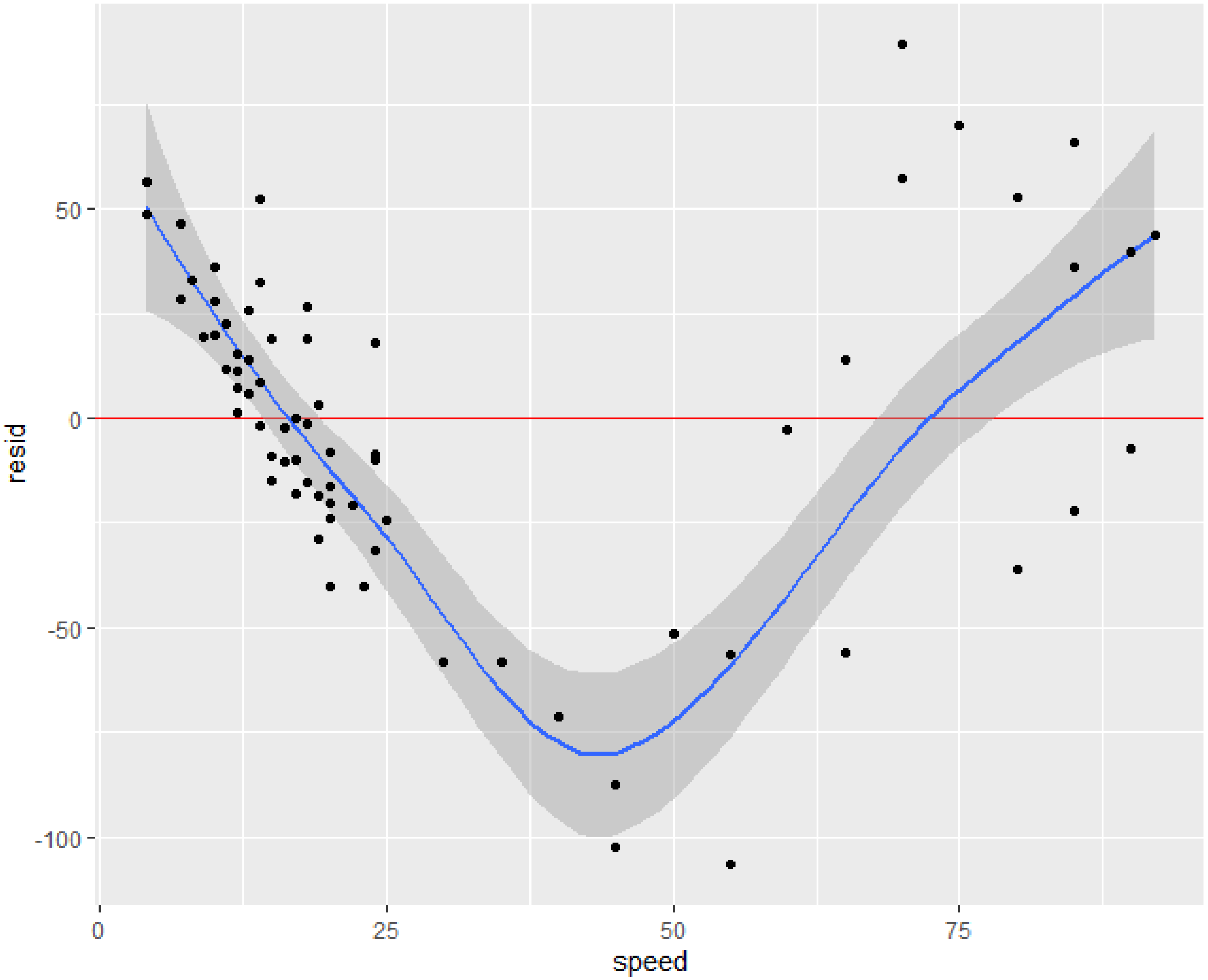


Not Normally Distributed ✗



Homoscedasticity

Is the variance of the residuals consistent across all X values?



Homoscedasticity

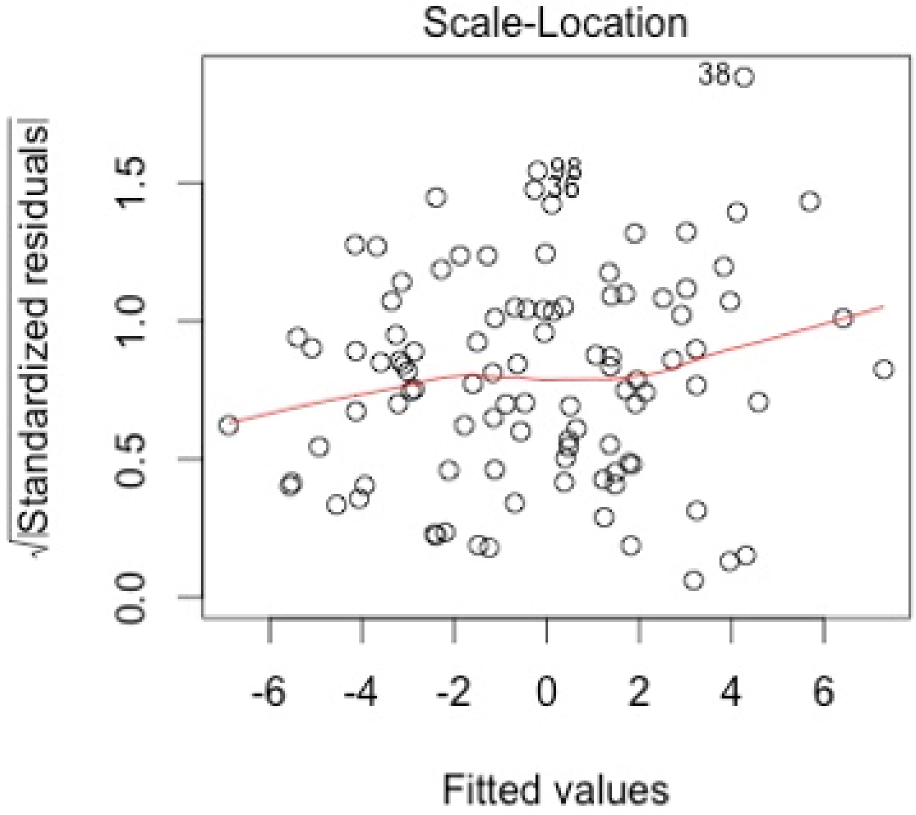
Some ways to assess homoscedasticity visually:

- A plot of the residual vs. X
- A plot of the residual vs. the predicted value (this will become more relevant in multiple regression)
- A plot of the square root of the standardized residual vs. the predicted value

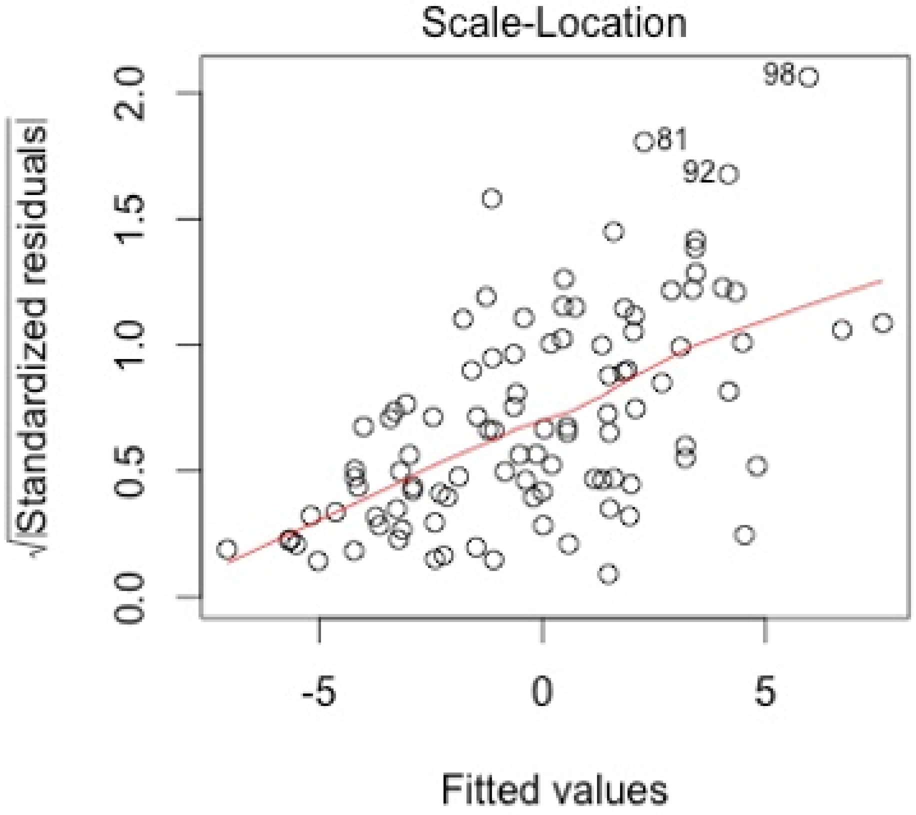
These will produce similar results, but in each you want to see that the spread of the points is consistent across the x-axis.

Homoscedasticity

✓ Homoskedasticity is Evident



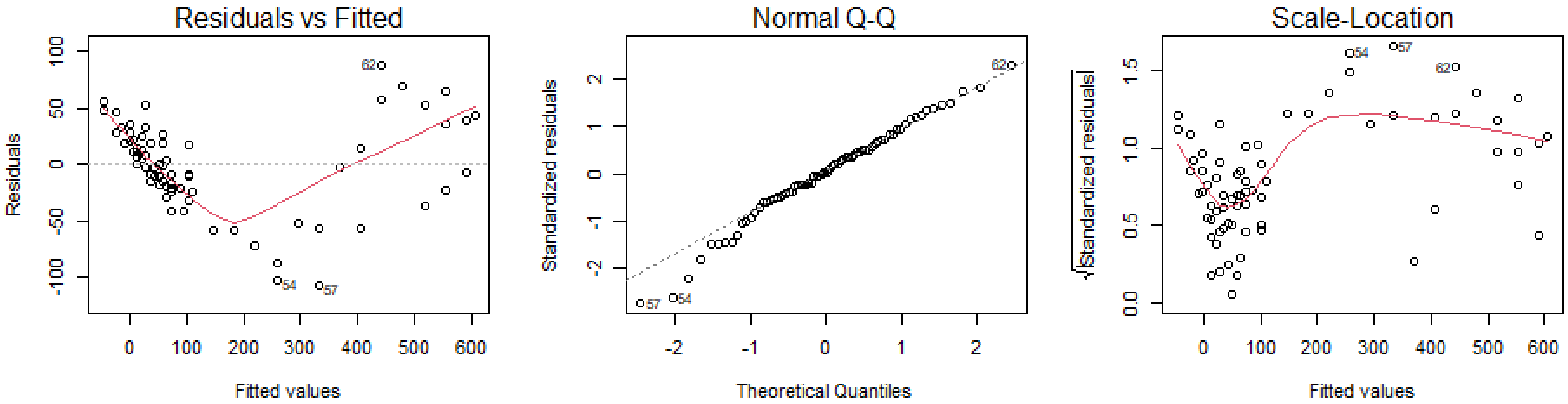
Heteroskedasticity is Evident ✗



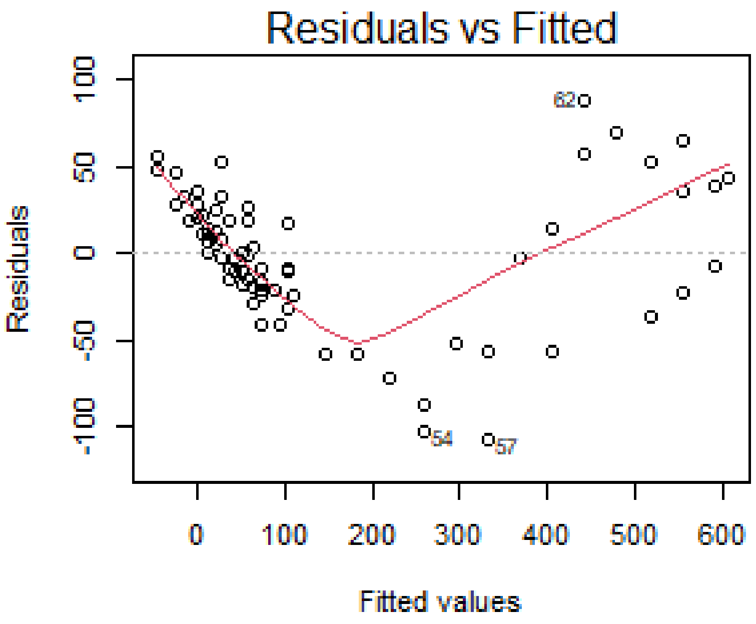
Plotting Graphs for Assumptions

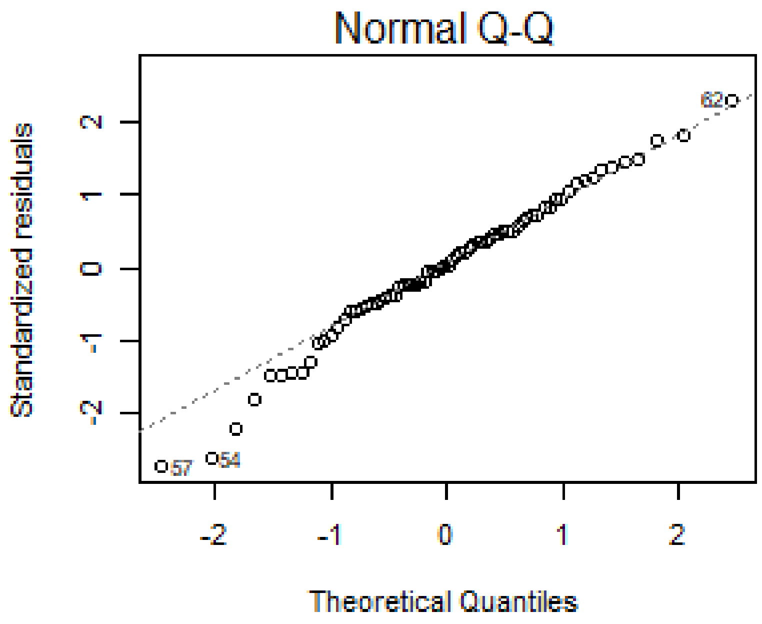
The “plot” command is quite versatile, as the output depends on the type of object that is fed into it.

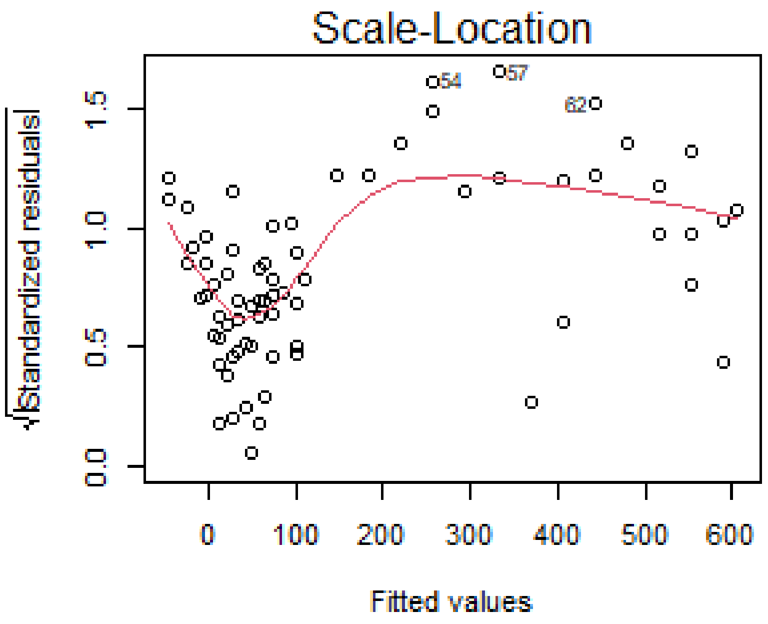
When plot() sees a lm object, it knows to plot model diagnostics.



Are any of the assumptions violated?







Linearity: Violated. The points dip down and then back up.

Normality: No violation. The residuals seem normal.

Homoscedasticity: Violated. The variance of the residuals is smaller on the left.

What can we do when our assumptions are violated?

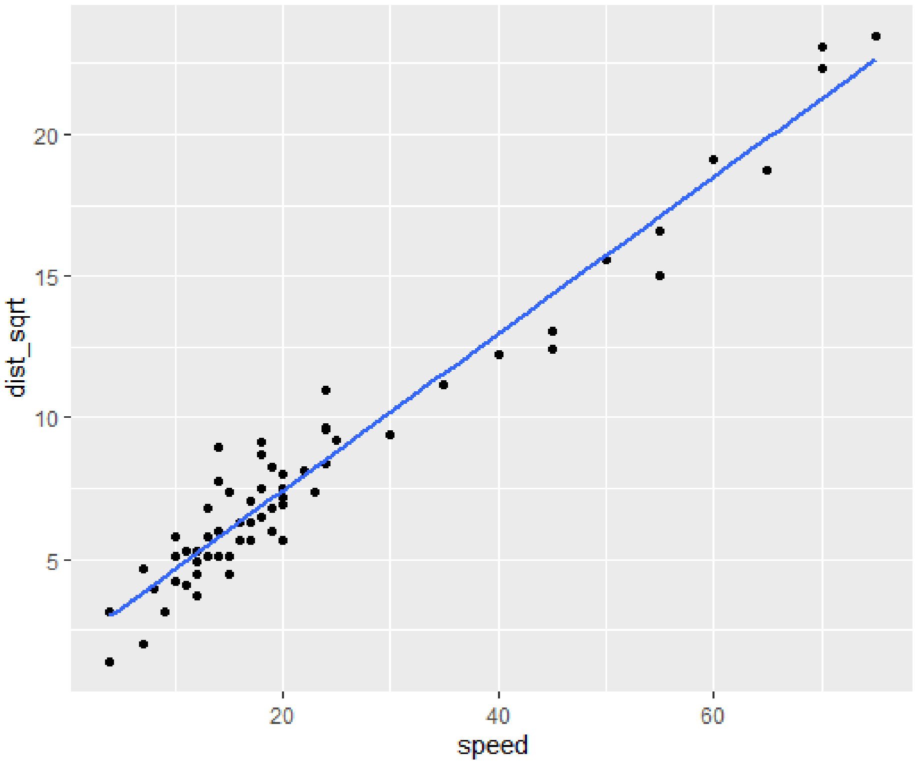
- Change our variables
  - Convert variables to categorical
  - Transform the outcome variable
  - Transform the predictor variable
- Examine your predictors
  - You may be omitting important predictors – we will discuss in multiple regression
- Change your modeling approach
  - Use another model such as logistic, Poisson, etc.

2. Checking Linear Regression Assumptions

As we saw last time, transforming the outcome (square root) provided a better fit.

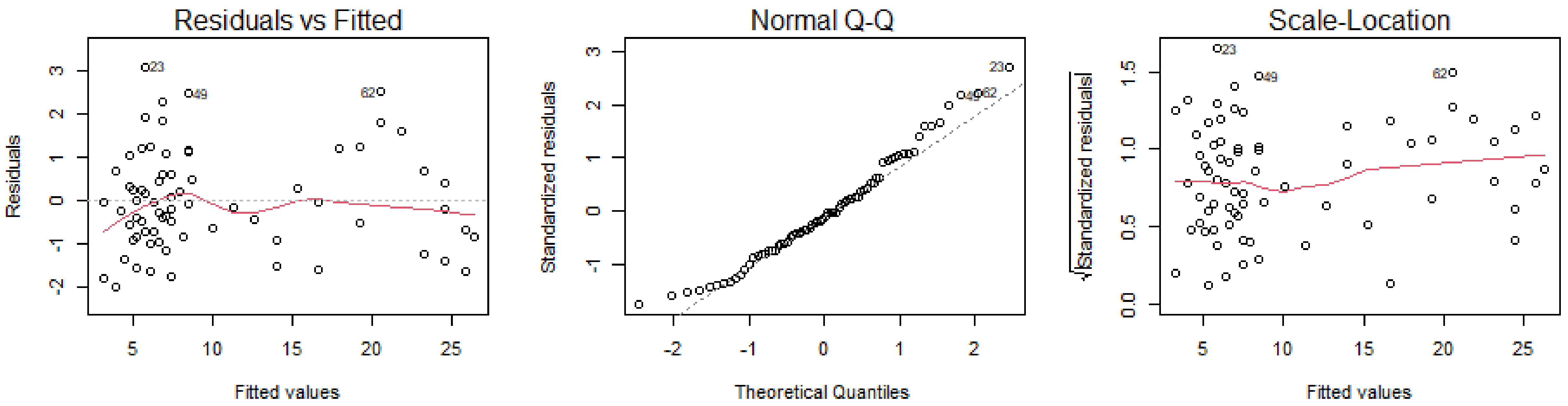
Let’s see if this model satisfies the assumptions..

```
> lm(dist_sqrt ~ speed, data = carstot) %>%
+   summary()

Call:
lm(formula = dist_sqrt ~ speed, data = carstot)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.1057 -0.7780 -0.1337  0.6287  3.1834 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.887082   0.242072   7.796 9.89e-11 ***
speed       0.276702   0.008362  33.092  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.145 on 61 degrees of freedom
Multiple R-squared:  0.9472,	Adjusted R-squared:  0.9464 
F-statistic:  1095 on 1 and 61 DF,  p-value: < 2.2e-16
```



2. Checking Linear Regression Assumptions

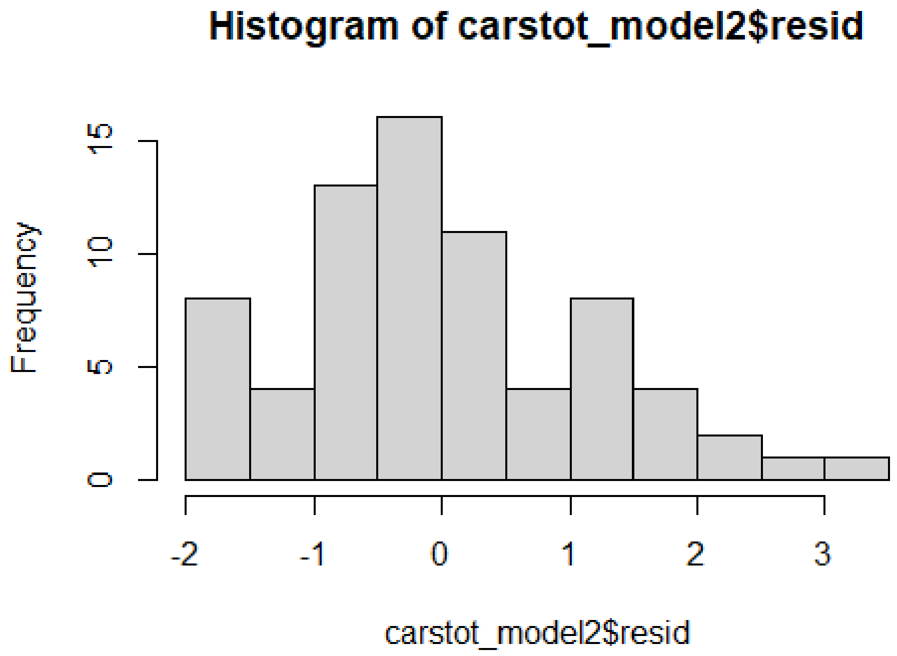
Much better! The linearity, normality, and homoscedasticity assumptions appear to hold.



If you had some questions about normality, you can further examine the residuals:

```
> psych::describe(carstot_model2$resid)
   vars  n mean   sd median trimmed  mad   min  max range skew kurtosis   se
X1    1 72    0 1.15  -0.13   -0.07 1.07 -1.99 3.11  5.11 0.51    -0.24 0.14
```

```
> shapiro.test(carstot_model2$resid)

	Shapiro-Wilk normality test

data:  carstot_model2$resid
W = 0.97115, p-value = 0.09664
```



2. Checking Linear Regression Assumptions

This is a great site for some examples of how assumptions can be violated. I’d highly recommend examining it in your free(?) time.

https://www.qualtrics.com/support/stats-iq/analyses/regression-guides/interpreting-residual-plots-improve-regression/

Recap

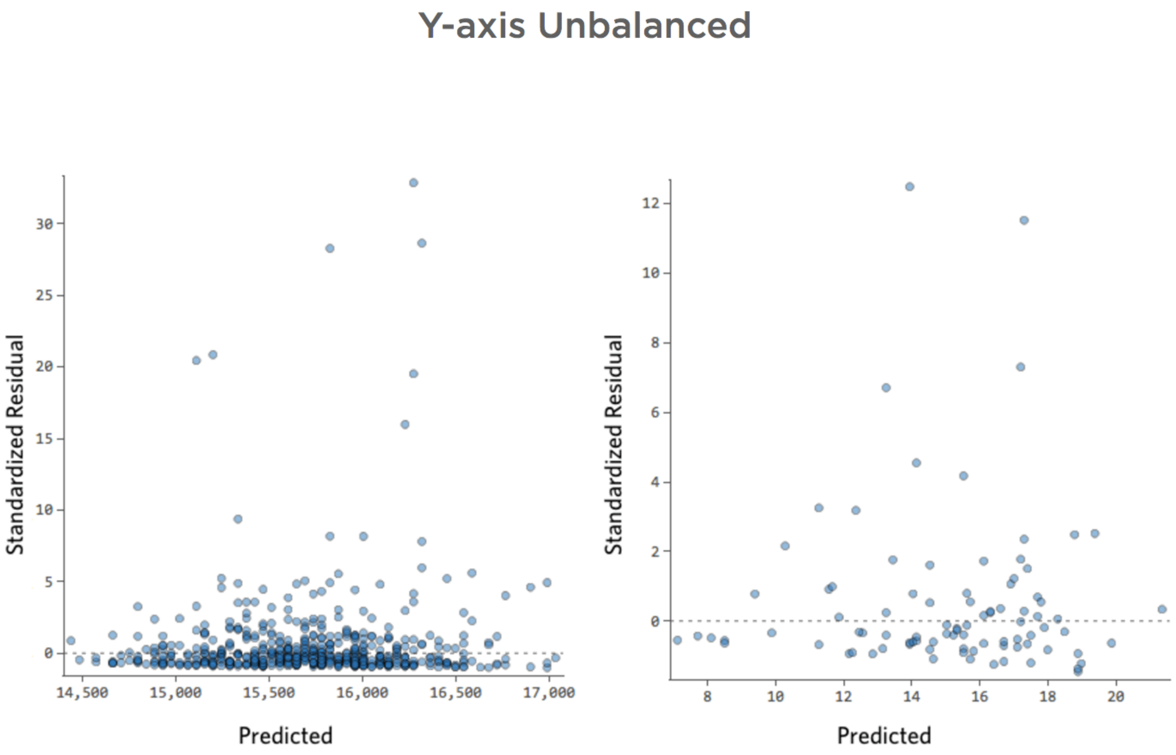
- Linear regression models are only valid if the LINE assumptions hold; it is therefore important to check these assumptions.

Recap

- Assess the 4 LINE assumptions, given a regression model
- Suggest alternative strategies if the assumptions do not hold

Test Yourself

The following reflect plots of the standardized residual vs. predicted value for two separate models. Which assumption is violated?



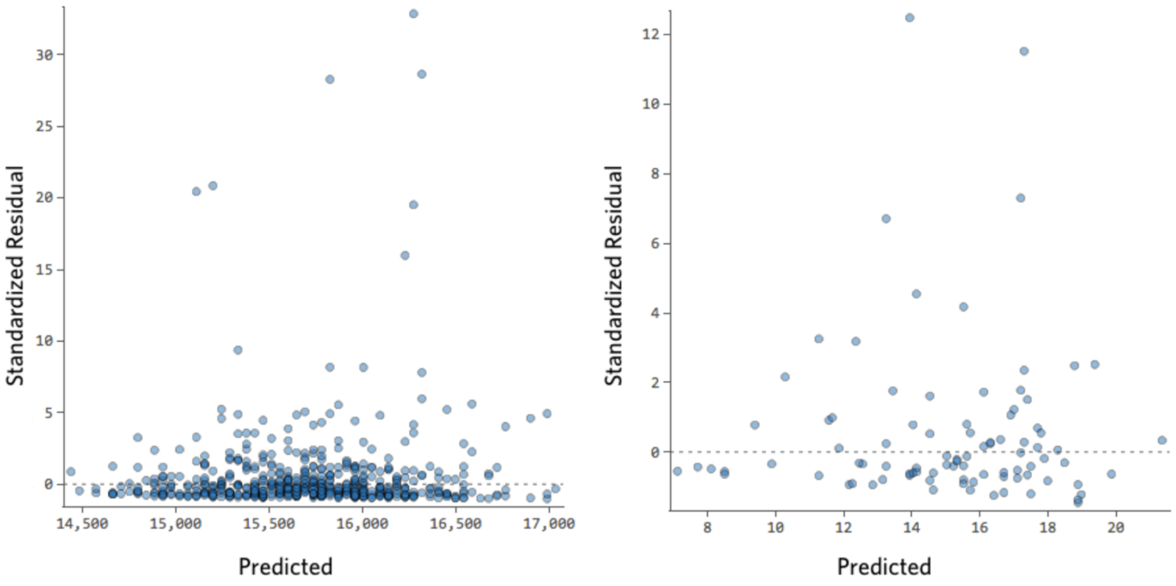
Y-axis Unbalanced

Test Yourself

The following reflect plots of the standardized residual vs. predicted value for two separate models. Which assumption is violated?

Normality

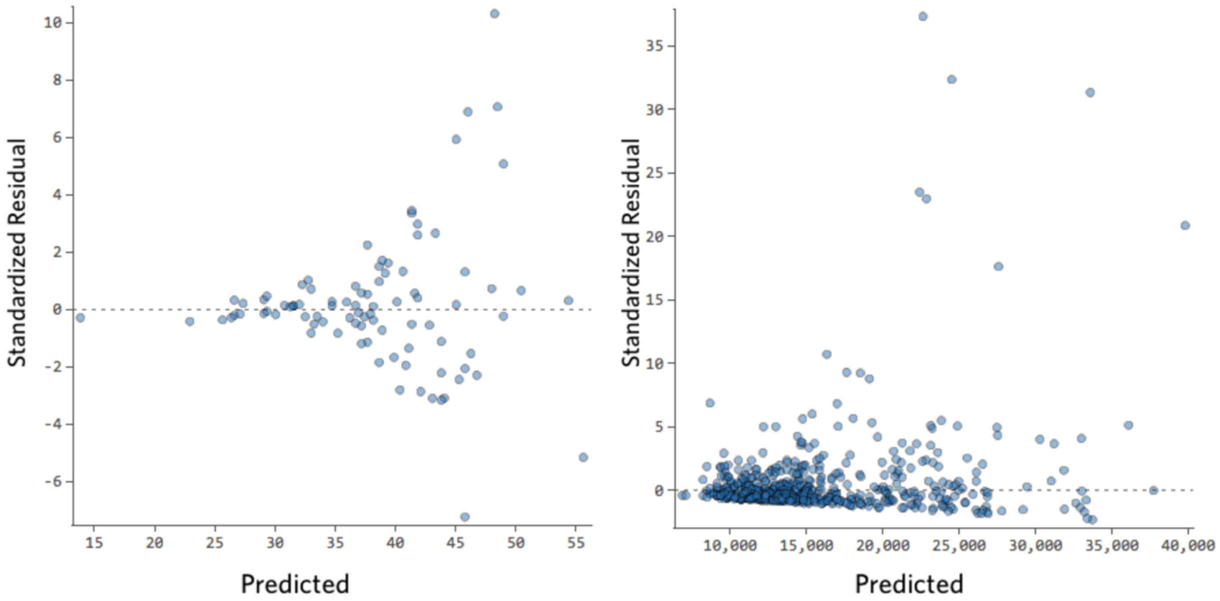
Y-axis Unbalanced



Test Yourself

The following reflect plots of the standardized residual vs. predicted value for two separate models. Which assumption is violated?

Heteroscedasticity

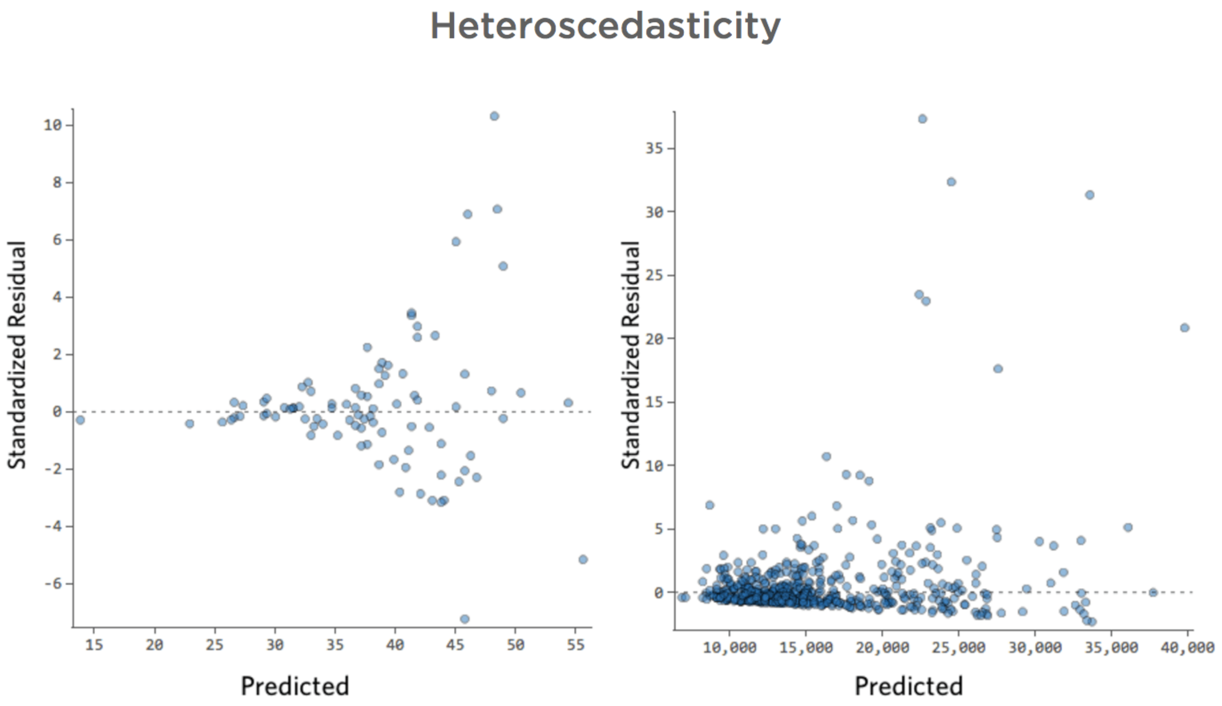


2. Checking Linear Regression Assumptions

Test Yourself

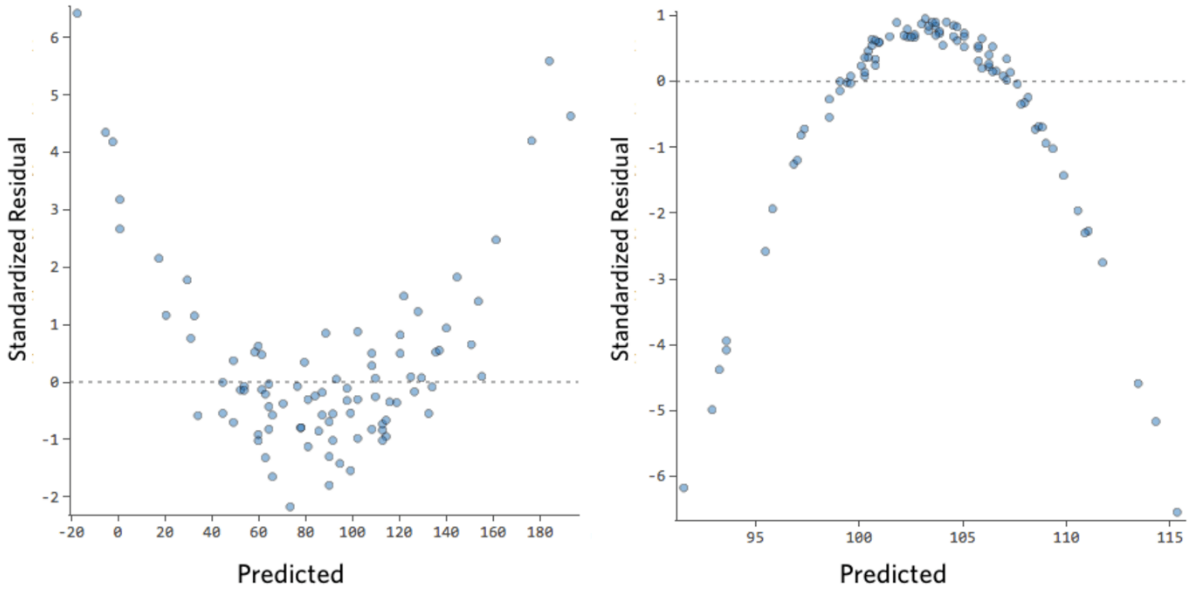
The following reflect plots of the standardized residual vs. predicted value for two separate models. Which assumption is violated?

Equal Variance



Test Yourself

The following reflect plots of the standardized residual vs. predicted value for two separate models. Which assumption is violated?



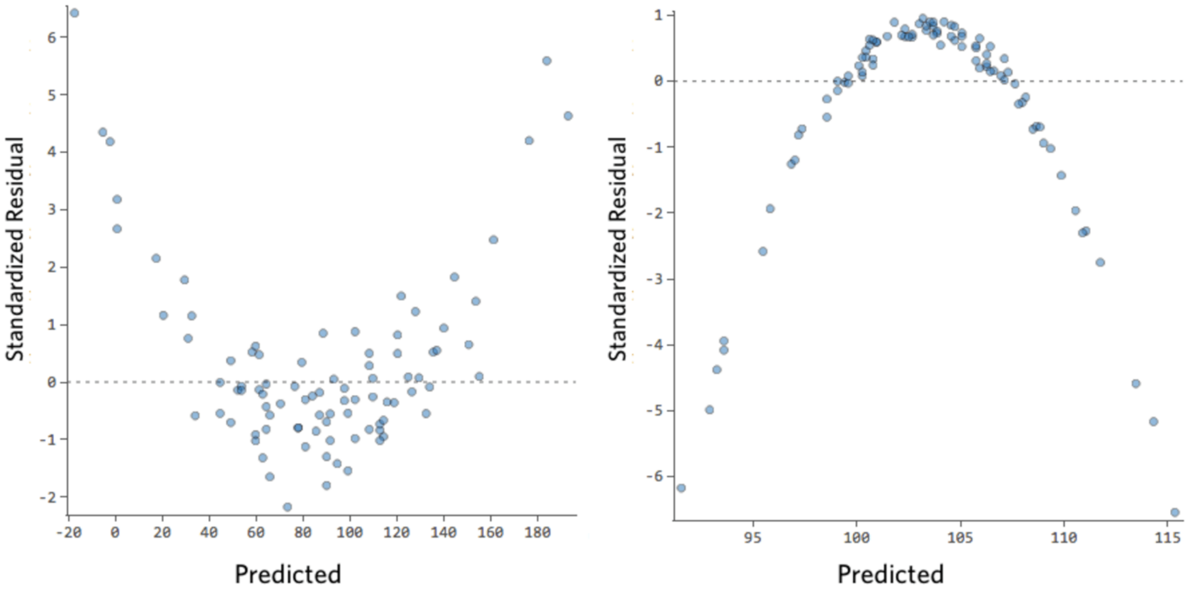
Nonlinear

Test Yourself

The following reflect plots of the standardized residual vs. predicted value for two separate models. Which assumption is violated?

Linearity

Nonlinear



3. The ANOVA Table

The ANOVA table is a way to tell us how “good” a regression model is.

The basic idea of the ANOVA table is to decompose each Y value into:

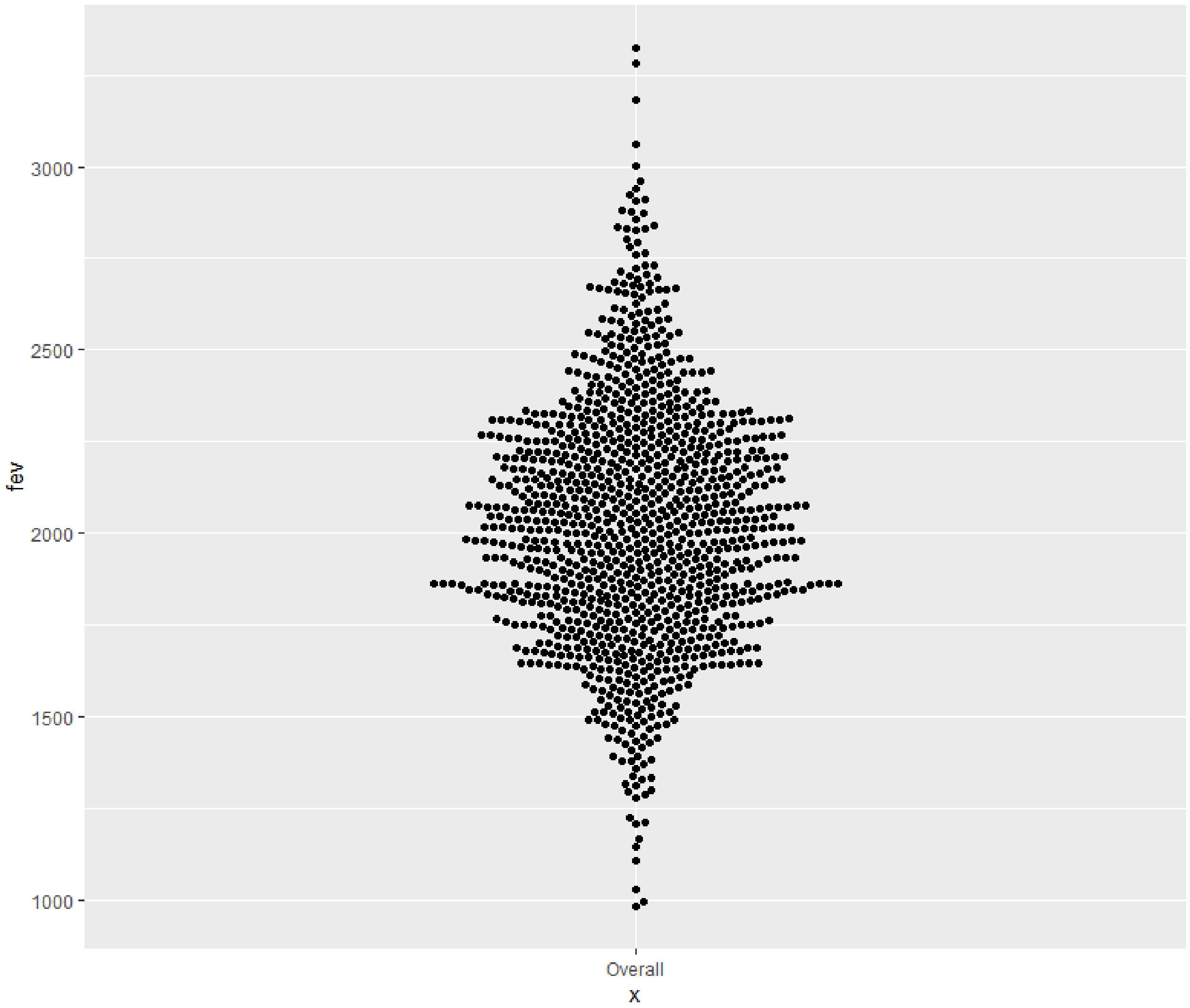
- The part that is explained by the regression model (the predicted value)
- The part that is not explained by the regression model (residuals)

For example, there is a lot of variation in children's FEV values in the Children's Health Study – values range from approximately 1000 to 3000 with a mean value of 2031.

The question is: **why do children's FEV values vary** around the mean?

Is the variation random?

Does some X variable contribute to the variation?



Created with ggbeeswarm()

3. The ANOVA Table

In a naïve (i.e., null, unconditional) model, we would use the overall mean to predict FEV. In this case, the sample mean is 2,031, so our best prediction for each individual would be 2,031.

```
> lm(fev ~ 1, data = chs) %>% summary()

Call:
lm(formula = fev ~ 1, data = chs)

Residuals:
     Min       1Q   Median       3Q      Max 
-1046.42  -222.30    -8.52   218.45  1292.42 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2031.265      9.947   204.2   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 330.7 on 1104 degrees of freedom
  (95 observations deleted due to missingness)
```

3. The ANOVA Table

The error (or residual) would be each person's Y value vs. the mean.

The ANOVA table will tell us the sum of squares of the residuals.

```
> lm(fev ~ 1, data = chs) %>% anova()
Analysis of Variance Table

Response: fev
            Df    Sum Sq Mean Sq F value Pr(>F)
Residuals 1104 120713111  109342
```

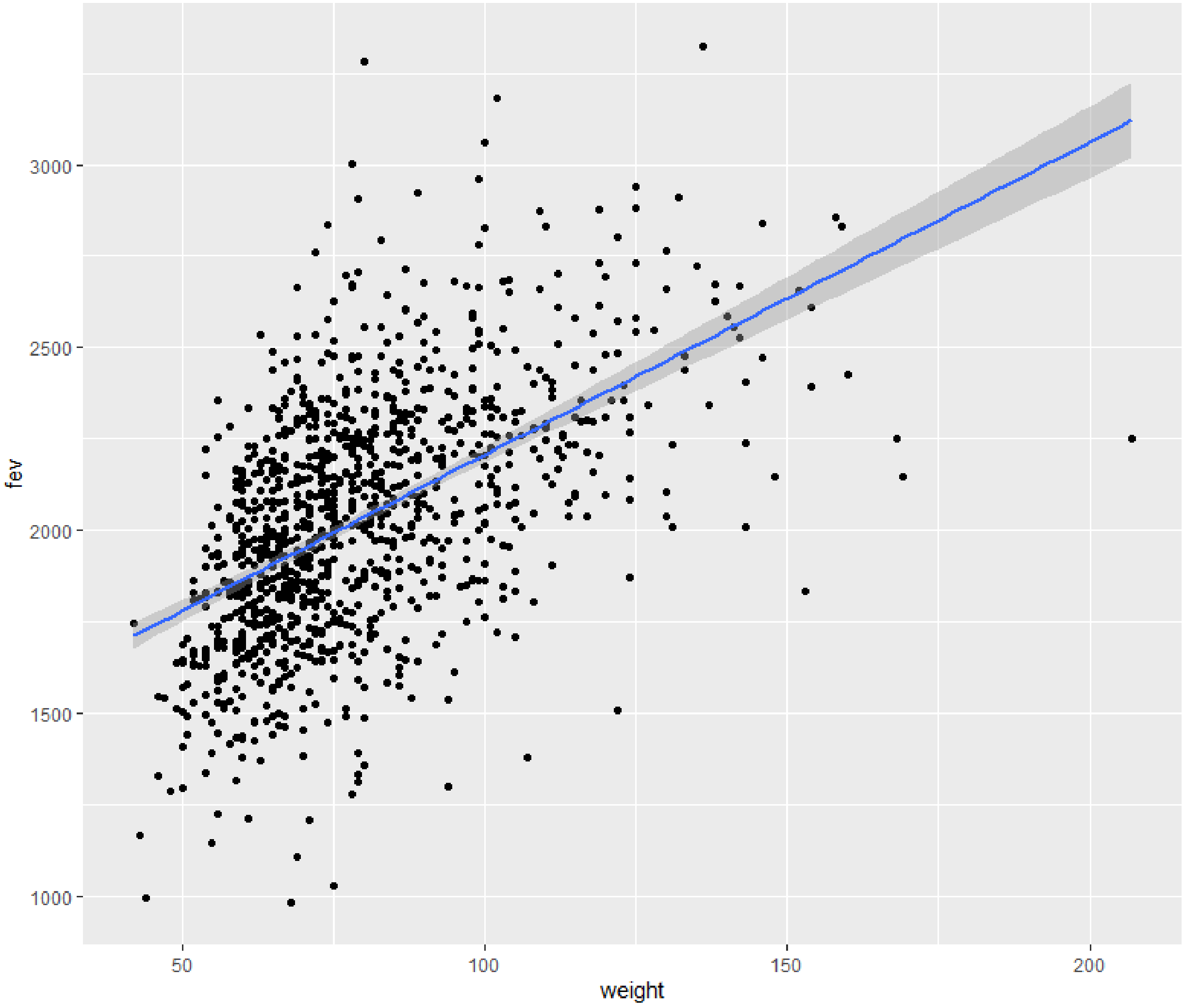
This is the amount of variation present in our Y values that is unexplained.

Can we do better at explaining our Y (FEV) values?

Let’s try looking at an independent variable: weight.

It does appear that weight can explain some of the variation in FEV.

And this comes in the form of the regression line.



Now we see that weight is significantly related to FEV.

```
> lm(fev ~ weight, data = chs) %>% summary()

Call:
lm(formula = fev ~ weight, data = chs)

Residuals:
    Min      1Q  Median      3Q     Max 
-962.08 -175.37   -6.39  181.60 1246.60 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1353.6432    33.4330   40.49   <2e-16 ***
weight         8.5392     0.4077   20.94   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 279.8 on 1103 degrees of freedom
  (95 observations deleted due to missingness)
Multiple R-squared:  0.2845,	Adjusted R-squared:  0.2839 
F-statistic: 438.6 on 1 and 1103 DF,  p-value: < 2.2e-16
```

3. The ANOVA Table

The ANOVA table is now broken into two components:

1. The sum of squares that is explained from the regression line

2. The sum of squares that is unexplained

```
> lm(fev ~ weight, data = chs) %>% anova()
Analysis of Variance Table

Response: fev
            Df   Sum Sq  Mean Sq F value    Pr(>F)    
weight       1 34344022 34344022   438.6 < 2.2e-16 ***
Residuals 1103 86369089    78304                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The 120,713,111 sum of squares that was previously unexplained is now split into: 34,344,022 SS explained by the regression and 86,369,089 SS unexplained.

If a large enough proportion of the variance in FEV is explained by the regression, then the F-test will be significant.

3. The ANOVA Table

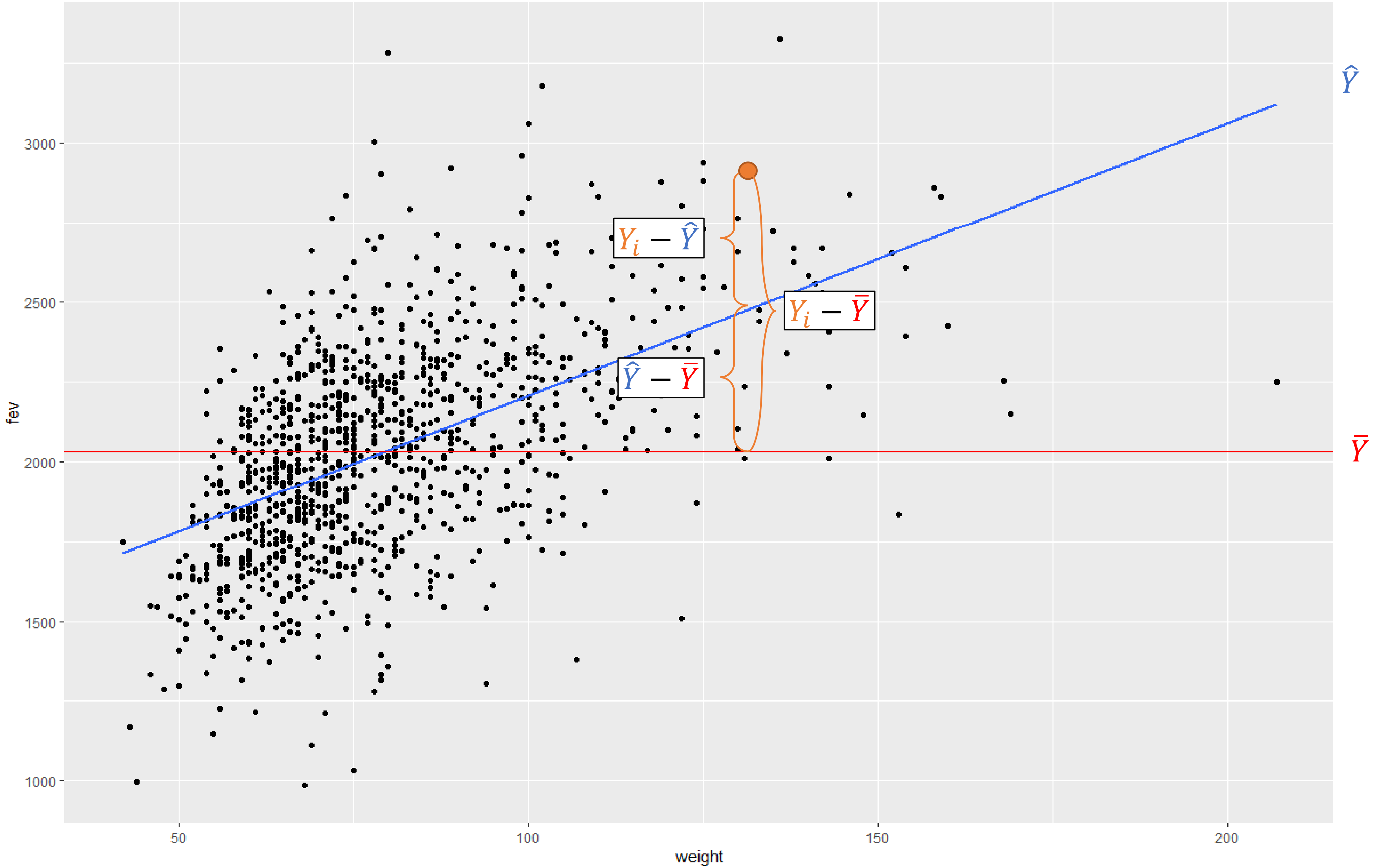
The ANOVA table is now broken into two components:

1. The sum of squares that is explained from the regression line ($SS_{Regression}$)
2. The sum of squares that is unexplained (SSE)

$$SS_{Total} = SS_{Reg} + SS_{Error}$$

$$SS_{Total} = \sum_{i=1}^{N}(Y_i - \bar{Y})^2,\ SS_{Reg} = \sum_{i=1}^{N}\left(\hat{Y} - \bar{Y}\right)^2,\ SS_{Error} = \sum_{i=1}^{N}\left(Y_i - \hat{Y}\right)^2$$

3. The ANOVA Table



3. The ANOVA Table

A typical ANOVA table will include a row for SS_{Total}, but the anova() function doesn’t do this for you.

If you want you can compute the Total SS manually, or write a function to compute it.

```
> lm(fev ~ weight, data = chs) %>% anova.full()
# A tibble: 3 x 6
  rowname      Df     Sum.Sq   Mean.Sq F.value    Pr..F.
  <chr>     <int>      <dbl>     <dbl>   <dbl>     <dbl>
1 weight        1  34344022. 34344022.    439.  2.93e-82
2 Residuals  1103  86369089.    78304.      NA  NA
3 Total      1104 120713111.   109342.      NA  NA
```

Some things to note about the ANOVA table.

```
> lm(fev ~ weight, data = chs) %>% anova.full()
# A tibble: 3 x 6
  rowname      Df      Sum.Sq    Mean.Sq F.value     Pr..F.
  <chr>     <int>       <dbl>      <dbl>   <dbl>      <dbl>
1 weight        1   34344022.  34344022.    439.   2.93e-82
2 Residuals  1103   86369089.     78304.      NA   NA
3 Total      1104  120713111.    109342.      NA   NA
```

```
> var(chs$fev, na.rm=T)
[1] 109341.6
```

The MS is the SS/df.
The total model df is N-1.
So the $\text{MS}_{\text{Total}} = \frac{\sum(Y_i-\bar{Y})^2}{N-1} = \sigma_Y^2 = \text{Var(Y)}$

i.e., The Mean Squares Total is the variance of Y.

The MS error provides an unbiased estimate of the variance of the errors (technical note: this is not the same as the variance of the residuals).
$\text{MS}_{\text{Error}} = \sigma^2$

Recall R^2 is "the proportion of variation in Y that is explained by our X variables." We can compute R^2 as $\text{SS}_{\text{Reg}}/\text{SS}_{\text{Total}}$ =
34344022/120713111 = 0.2845
(Which is equivalent to the output provided by lm())

Recap

- Each observation's Y score can be broken down into:
  - A component that is explained by the regression model
  - A component that is left unexplained (residual)
- The ANOVA table breaks down how much of the overall variation in Y is due to X (the "model") and how much is unexplained

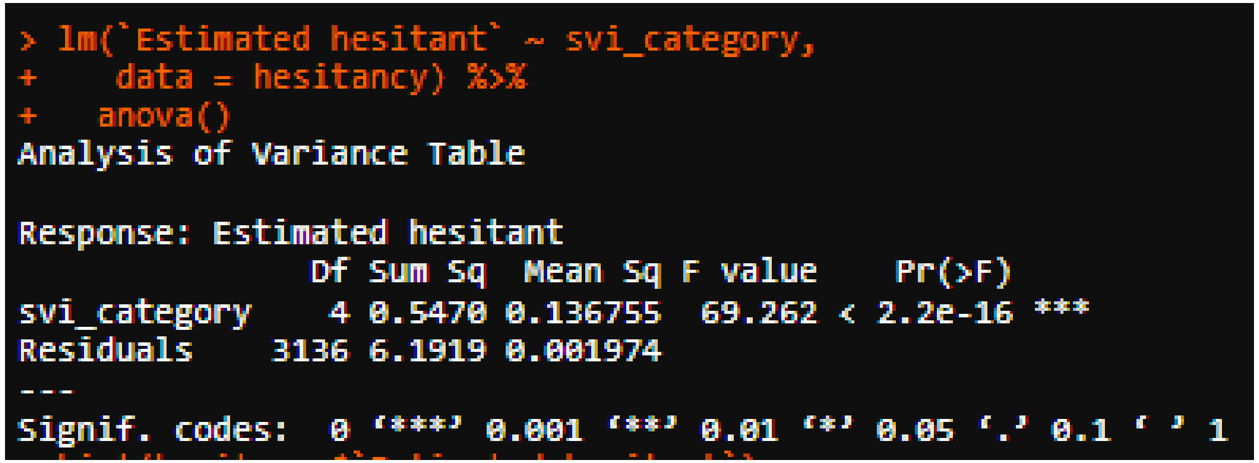
Recap

- Recreate the components of the ANOVA table given partial output
- Explain how the ANOVA table relates to parts of the regression output (e.g., R^2, standard error of the residuals)
- Use the ANOVA table to assess the fit of a linear model

Test Yourself

A regression was performed of percent of population hesitant of vaccines on “social vulnerability index”, using census tract as the unit of observation.

- Is the relationship between these variables statistically significant?
- What percent of the variation in vaccine hesitancy is explained by social vulnerability index?

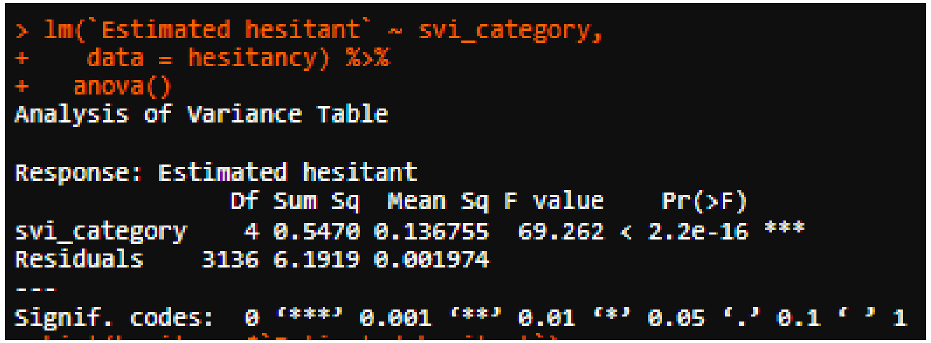


```
> lm(`Estimated hesitant` ~ svi_category,
+     data = hesitancy) %>%
+    anova()
Analysis of Variance Table

Response: Estimated hesitant
               Df Sum Sq  Mean Sq F value    Pr(>F)    
svi_category    4 0.5470 0.136755  69.262 < 2.2e-16 ***
Residuals    3136 6.1919 0.001974                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Test Yourself

A regression was performed of percent of population hesitant of vaccines on “social vulnerability index”, using census tract as the unit of observation.

- Is the relationship between these variables statistically significant? Yes (p<0.001). The p-value from the ANOVA table tests whether the overall model is associated with outcome, and the only independent variable is SVI.
- What percent of the variation in vaccine hesitancy is explained by social vulnerability index? $R^2 = (0.547)/(0.547 + 6.1919) = 0.08$, or 8%



```
> lm(`Estimated hesitant` ~ svi_category,
+     data = hesitancy) %>%
+   anova()
Analysis of Variance Table

Response: Estimated hesitant
               Df Sum Sq  Mean Sq F value    Pr(>F)
svi_category    4 0.5470 0.136755  69.262 < 2.2e-16 ***
Residuals    3136 6.1919 0.001974
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

4. Log Transformations

Occasionally when the data does not conform to our linear regression assumptions, we may want to transform either the X or Y variable.

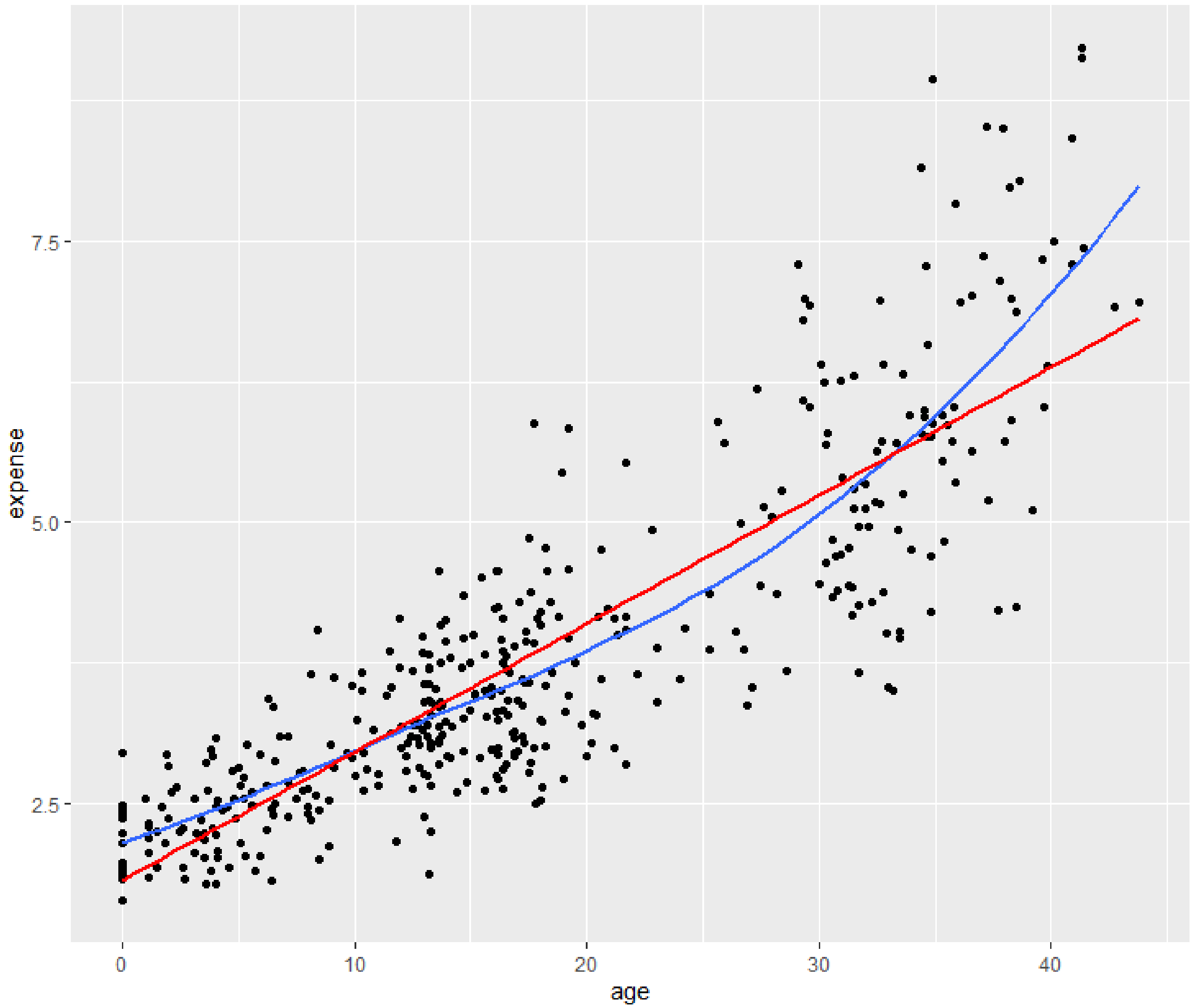
How do we know whether to transform the X or Y variable?

- Transformations on Y will help shrink the errors at higher values, and can help a model conform to homoscedasticity.
- Transformations on X or Y can help a model conform to linearity, although transformations on X are more desirable.

Let’s use the real estate data on “expense” ($/month expended due to inefficiencies) vs. house age.

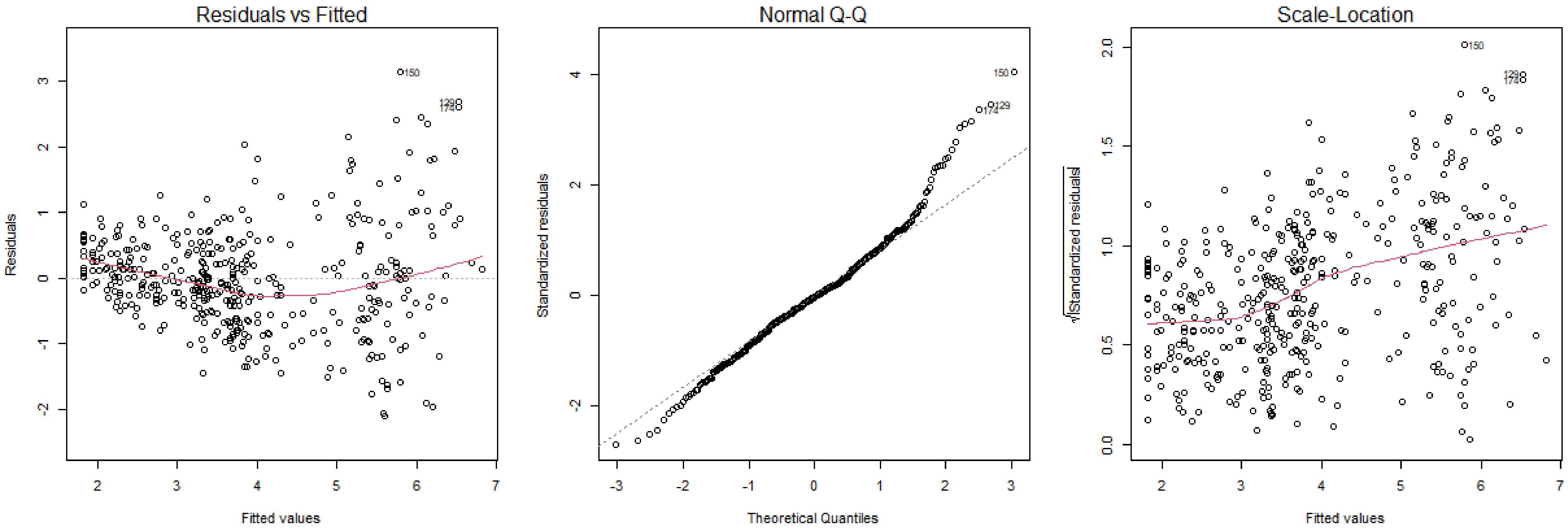
The red line is a linear fit, while the blue line is a smoothed fit.

What problems do you think we will encounter here?



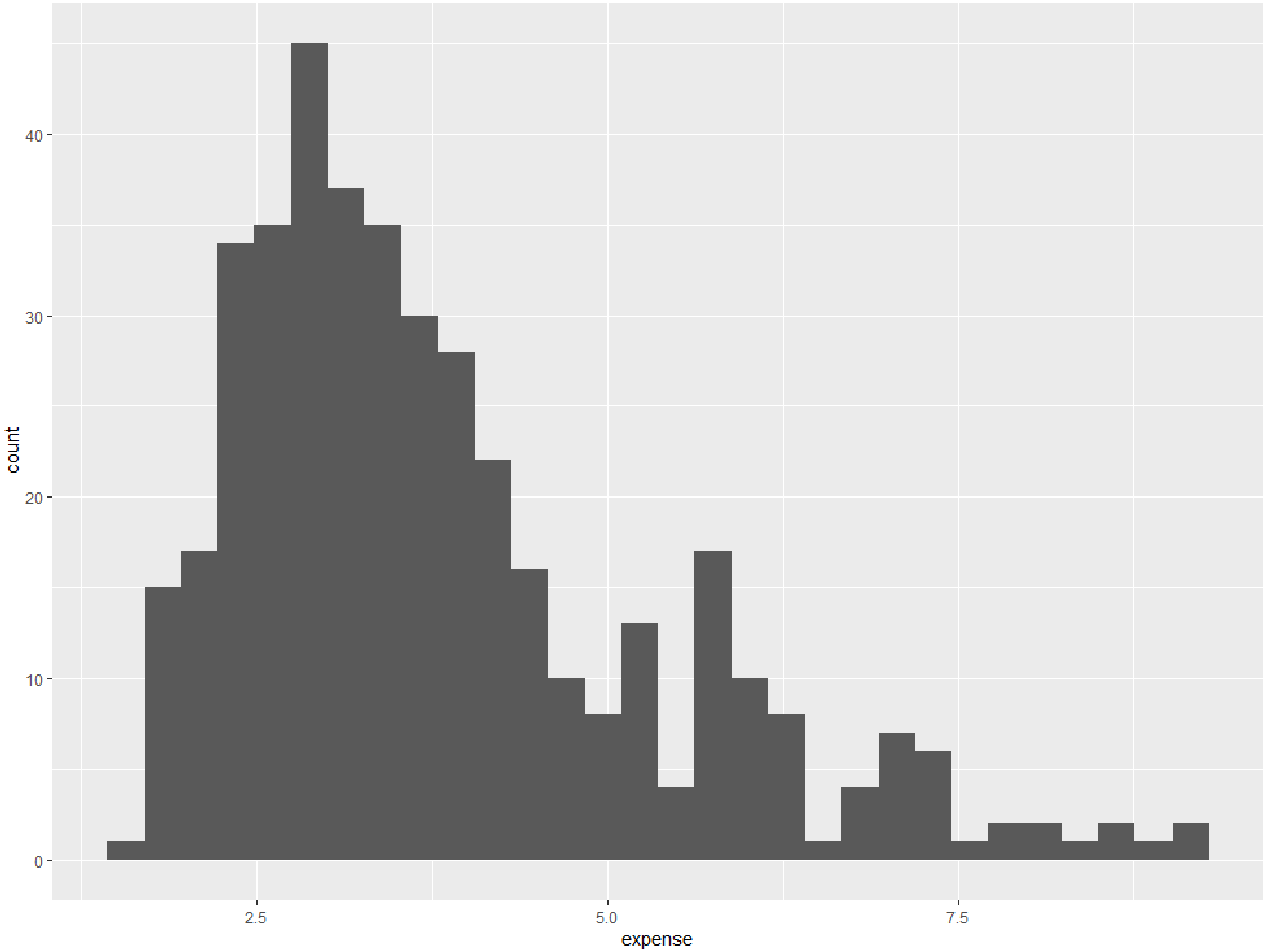
4. Log Transformations

After performing the regression, we have a problem with the residuals. Not only is linearity slightly violated, but there appears to be a violation of homoscedasticity.



4. Log Transformations

Part of the reason that this is a problem is because of the skewed distribution of Y. As X increases, Y increases nonlinearly. And since Y is increasing faster than X, the residuals will be higher at these values.



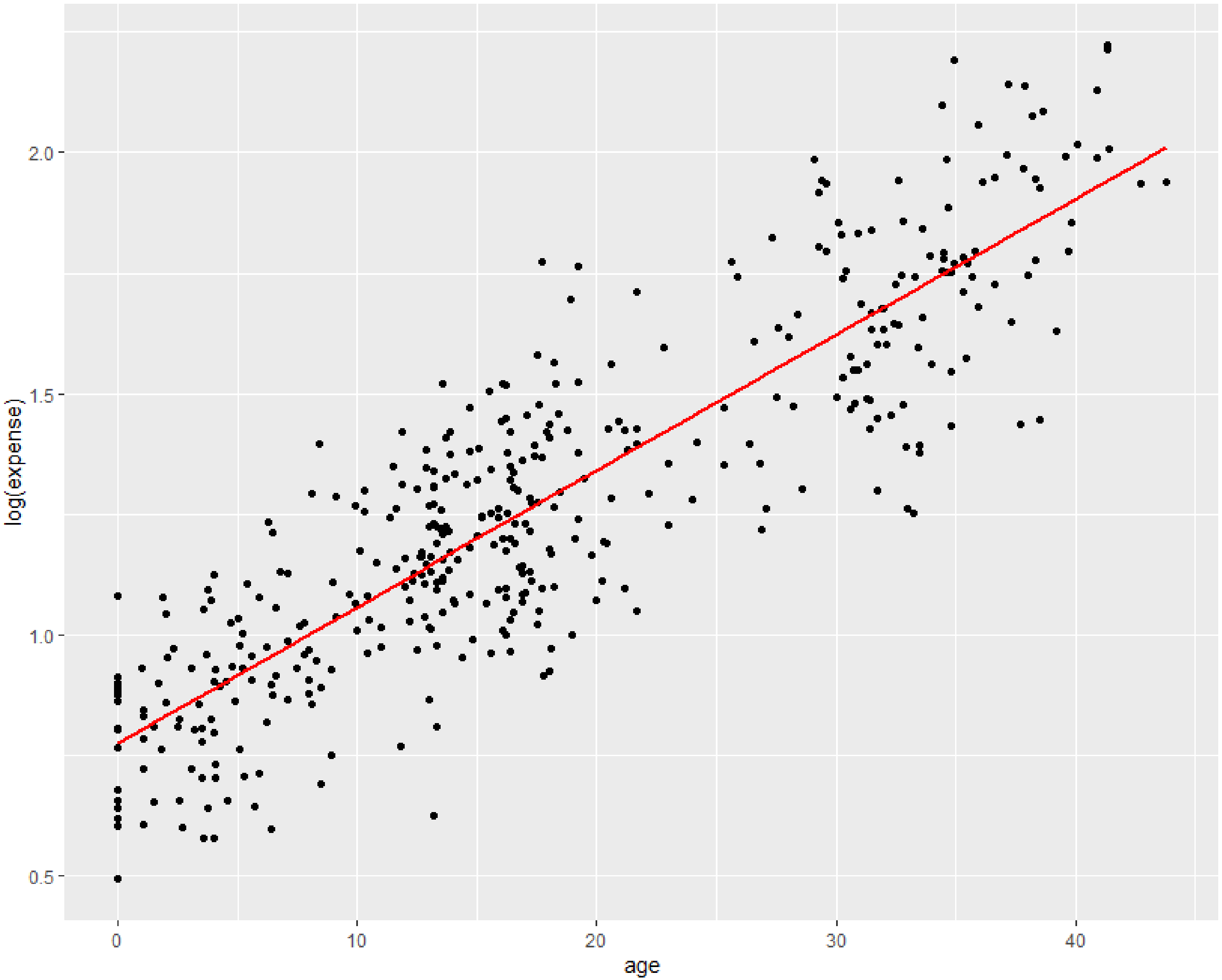
4. Log Transformations

A **variance stabilizing transformation** is one that reduces the variance of the residuals at higher values, providing a consistent value of σ^2 across all X values.

The log (or ln) transformation is perhaps the most common variance-stabilizing transformation.

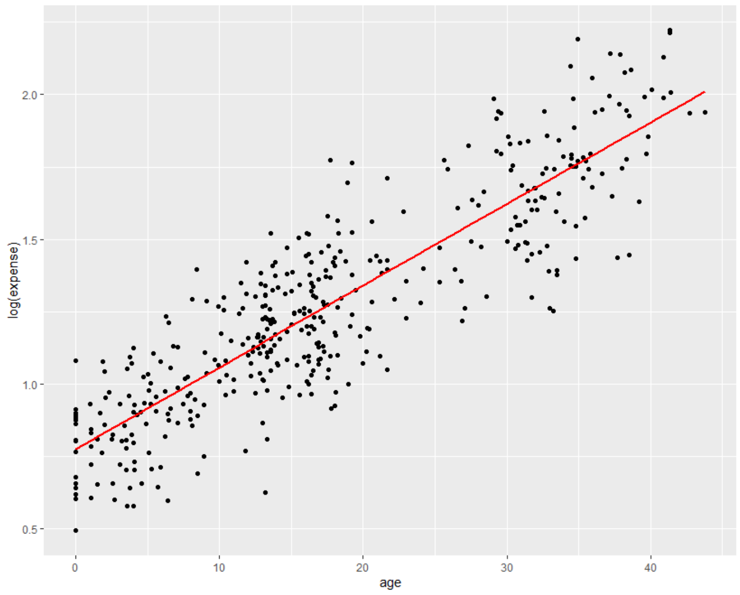
As we will see, the log transformation has some nice properties when it comes to interpretation.

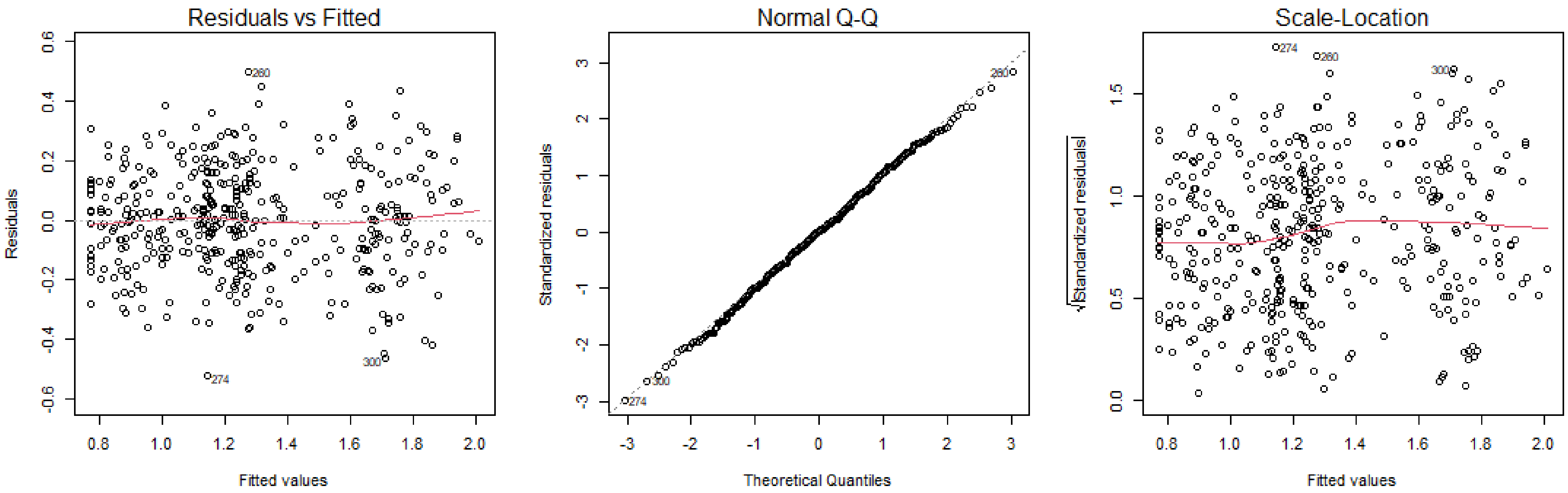
It appears that when we take the natural log of “expense,” the regression assumptions are much better satisfied!



4. Log Transformations

It appears that when we take the natural log of "expense," the regression assumptions are much better satisfied!





But what is the interpretation?

```
> lm(log(expense) ~ age, data = re) %>% summary()

Call:
lm(formula = log(expense) ~ age, data = re)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.52278 -0.11180  0.00334  0.12202  0.49750 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.7747629  0.0159542   48.56   <2e-16 ***
age         0.0282453  0.0007578   37.27   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1755 on 412 degrees of freedom
Multiple R-squared:  0.7713,	Adjusted R-squared:  0.7707 
F-statistic:  1389 on 1 and 412 DF,  p-value: < 2.2e-16
```

A one-unit (year) increase in age is associated with a 0.028-unit increase in log expense.

To make this more interpretable, we can take an “anti-log” transformation using the exponentiation. Recall, $e^{\log(x)}$=x.

If we look at the log(Y) value associated with a a 1-unit increase in X:

$$\log(Y|X=x+1)=\beta_0+\beta_1(x+1)+e$$

$$\log(Y|X=x)=\beta_0+\beta_1(x)+e$$

$$\log(Y|X=x+1)-\log(Y|X=x)=\beta_1$$

$$\log\left(\frac{(Y|X=x+1)}{(Y|X=x)}\right)=\beta_1$$

4. Log Transformations

$$\log\left(\frac{(Y|X = x + 1)}{(Y|X = x)}\right) = \beta_1$$

$$\frac{(Y|X = x + 1)}{(Y|X = x)} = e^{\beta_1}$$

In a non-transformed regression, a 1-unit change in X is associated with a β_1-unit change in Y.

In a log-transformed regression, a 1-unit change in X is associated with a e^{β_1} multiplicative change in Y.

An interpretation that makes more sense:

```
> lm(log(expense) ~ age, data = re) %>% summary()

Call:
lm(formula = log(expense) ~ age, data = re)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.52278 -0.11180  0.00334  0.12202  0.49750 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.7747629  0.0159542   48.56   <2e-16 ***
age         0.0282453  0.0007578   37.27   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1755 on 412 degrees of freedom
Multiple R-squared:  0.7713,	Adjusted R-squared:  0.7707 
F-statistic:  1389 on 1 and 412 DF,  p-value: < 2.2e-16
```

A one-unit (year) increase in age is associated with a 0.028-unit increase in log Y.

OR

A one-unit (year) increase in age is associated with a exp(0.028) = 1.0286 *times* increase in Y.

OR

A one-unit (year) increase in age is associated with a 100(e^{β}-1) = 2.86% increase in Y.

4. Log Transformations

To get the confidence interval, we must exponentiate the lower and upper boundaries individually.

$$95\%\ \text{CI} = (e^{\beta - 1.96SE(\beta)}, e^{\beta + 1.96SE(\beta)})$$

Do <u>not</u> exponentiate the parameter estimate and the standard error individually.

```
> lm(log(expense) ~ age, data = re) %>% confint()
                 2.5 %     97.5 %
(Intercept) 0.74340104 0.80612474
age         0.02675565 0.02973503

> lm(log(expense) ~ age, data = re) %>% confint() %>% exp(.)
               2.5 %   97.5 %
(Intercept) 2.103076 2.239214
age         1.027117 1.030182
```

A one-year increase in age is associated with a 2.86% increase in expense (95% CI = 2.71%, 3.02%).

How do we interpret the intercept?

```
> lm(log(expense) ~ age, data = re) %>% summary()

Call:
lm(formula = log(expense) ~ age, data = re)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.52278 -0.11180  0.00334  0.12202  0.49750 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.7747629  0.0159542   48.56   <2e-16 ***
age         0.0282453  0.0007578   37.27   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1755 on 412 degrees of freedom
Multiple R-squared:  0.7713,	Adjusted R-squared:  0.7707 
F-statistic:  1389 on 1 and 412 DF,  p-value: < 2.2e-16
```

The mean log expense is 0.77 when age = 0.

OR

The geometric mean expense is exp(0.77) = 2.17 when age = 0.

Conclusion Statement

To satisfy the assumptions of linear regression, the natural log of household monthly expense was regressed on house age. Predicted values and 95% confidence limits were computed on the log scale, and the values were un-transformed to obtain corresponding values on the original expense scale. We found that each year increase in house age was associated with a 2.86% (95% CI = 2.71%, 3.02%) increase in expense (p<.001).

4. Log Transformations

Sometimes we may wish to transform just the X variable, or both the X and Y variables.

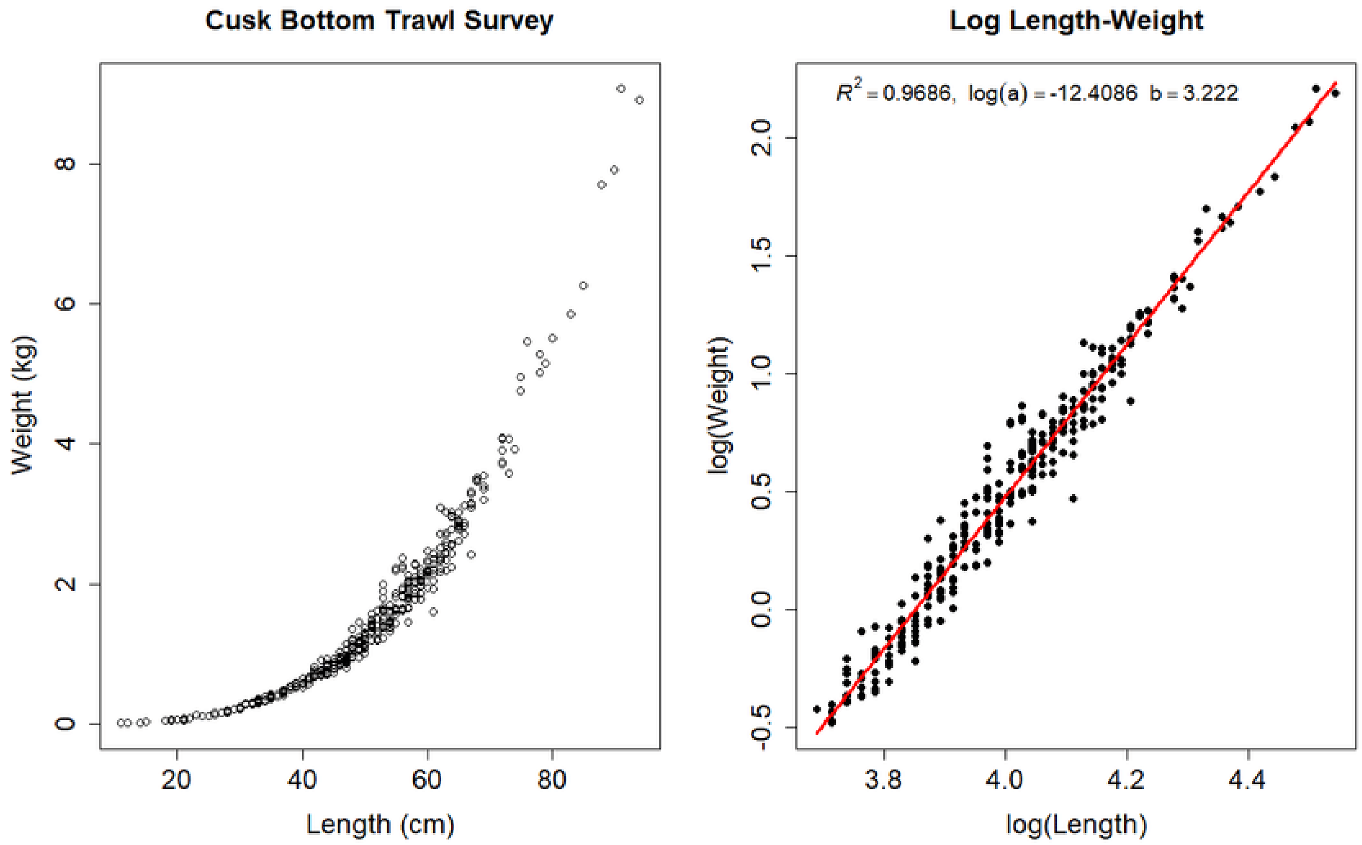
When we use the log transformation, the interpretation is easy:

Transformation	Equation	Interpretation
None	$\hat{Y} = \beta_0 + \beta_1 X$	A one-unit increase in X is associated with a β_1 unit increase in Y.
Ln(Y)	$\ln(\hat{Y}) = \beta_0 + \beta_1 X$ $\hat{Y} = e^{\beta_0} e^{\beta_1 X}$	A one-unit increase in X is associated with a $100(e^{\beta_1} - 1)$% increase in Y.
Ln(X)	$\hat{Y} = \beta_0 + \beta_1 \ln(X)$	A 1% increase in X is associated with a $(\beta_1/100)$ unit increase in Y.
Ln(Y) & Ln(X)	$\ln(\hat{Y}) = \beta_0 + \beta_1 \ln(X)$	A 1% increase in X is associated with a β_1% increase in Y.

Example

The Cusk is a species of fish. Biologists measured how length and height were related in a sample of Cusk in the Gulf of Maine.

These results show that a 1% increase in Cusk length is associated with a 3.22% increase in Cusk weight.



Recap

- The natural log transformation is a common way to transform either X or Y to better satisfy the assumptions of linear regression
- Variables that undergo log transformations are simple to interpret, compared to other transformations

Recap

- Decide when to implement a log transformation of X or Y
- Perform and interpret an analysis on log-transformed variables

5. Categorical Predictors: Binary

We’ve seen how we can relate continuous independent variables to a continuous outcome through linear regression.

How would we examine the effect of a categorical independent variable on an outcome?

Example: how does sex relate to FEV?

5. Categorical Predictors: Binary

Option 1. One way to examine the relationship between a dichotomous IV and a continuous DV is through a t-test.

Here, we see that males on average have a higher FEV than females (t_{1103}=-7.45, p<.001).

```
> t.test(fev ~ male,
+        var.equal = T,
+        data = chs)

	Two Sample t-test

data:  fev by male
t = -7.4514, df = 1103, p-value = 1.861e-13
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -182.8212 -106.6080
sample estimates:
mean in group 0 mean in group 1 
       1959.105        2103.819 
```

Notice this formula notation looks a lot like our lm() regression equation…

5. Categorical Predictors: Binary

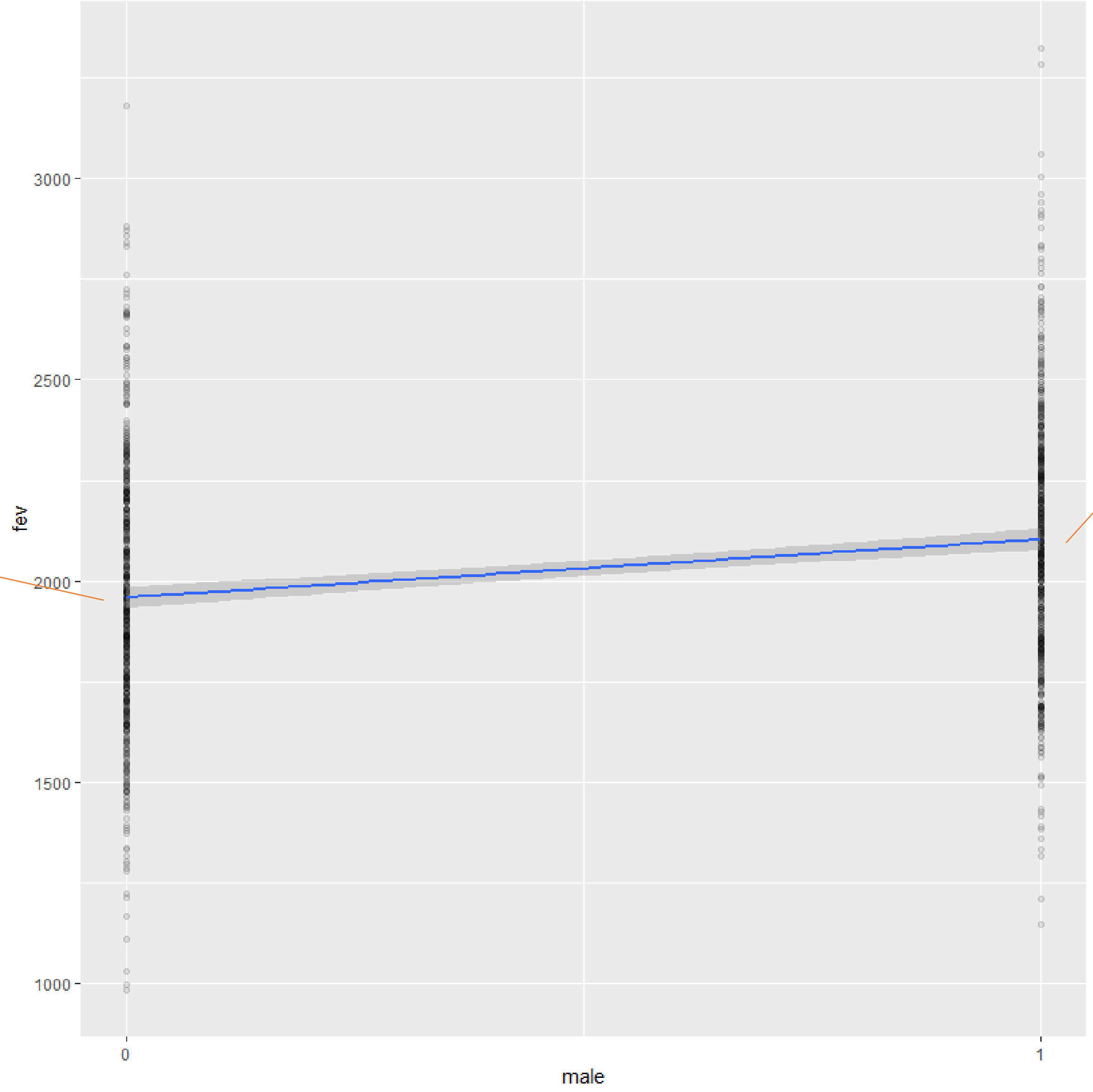
Option 2. Another way we can examine this is through the regression framework.

We need to make sure our X value is coded the correct way: as a “dummy” variable.

Some coding guidelines to make interpretation easier:

- The “baseline” category should be coded 0.
- The “other” category should be coded 1, which makes a 1-unit difference in X between the two groups.

5. Categorical Predictors: Binary



The Y-intercept (β_0) is the mean value of FEV when X=0 (i.e., females).

The mean value for males will be $\beta_0 + \beta_1$. This is the value of our regression equation when X=1 (i.e., for males).

□ What would β_1 be if there was no difference in FEV between males and females?

Let’s compare the t-test to the linear regression results.

```
> t.test(fev ~ male,
+        var.equal = T,
+        data = chs)

         Two Sample t-test

data:  fev by male
t = -7.4514, df = 1103, p-value = 1.861e-13
alternative hypothesis: true difference in means is
not equal to 0
95 percent confidence interval:
 -182.8212 -106.6080
sample estimates:
mean in group 0 mean in group 1
       1959.105        2103.819
```

```
> lm(fev ~ male,
+    data = chs) %>%
+   summary()

Call:
lm(formula = fev ~ male, data = chs)

Residuals:
    Min      1Q  Median      3Q     Max
-974.26 -235.28  -17.68  206.88 1220.69

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1959.10      13.71 142.852  < 2e-16 ***
male          144.71      19.42   7.451 1.86e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
1

Residual standard error: 322.8 on 1103 degrees of freedom
  (95 observations deleted due to missingness)
Multiple R-squared:  0.04793,      Adjusted R-squared:
0.04706
F-statistic: 55.52 on 1 and 1103 DF,  p-value: 1.861e-13
```

5. Categorical Predictors: Binary

```
> lm(fev ~ male,
+    data = chs) %>%
+   summary()

Call:
lm(formula = fev ~ male, data = chs)

Residuals:
    Min      1Q  Median      3Q     Max 
-974.26 -235.28  -17.68  206.88 1220.69 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1959.10      13.71 142.852  < 2e-16 ***
male          144.71      19.42   7.451 1.86e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 
1

Residual standard error: 322.8 on 1103 degrees of freedom
  (95 observations deleted due to missingness)
Multiple R-squared:  0.04793,	Adjusted R-squared:  
0.04706 
F-statistic: 55.52 on 1 and 1103 DF,  p-value: 1.861e-13
```

For females:

$$\hat{Y}_{(X=0)} = 1959.10 + 144.71(0) = 1959.10$$

For males:

$$\hat{Y}_{(X=1)} = 1959.10 + 144.71(1) = 2103.82$$

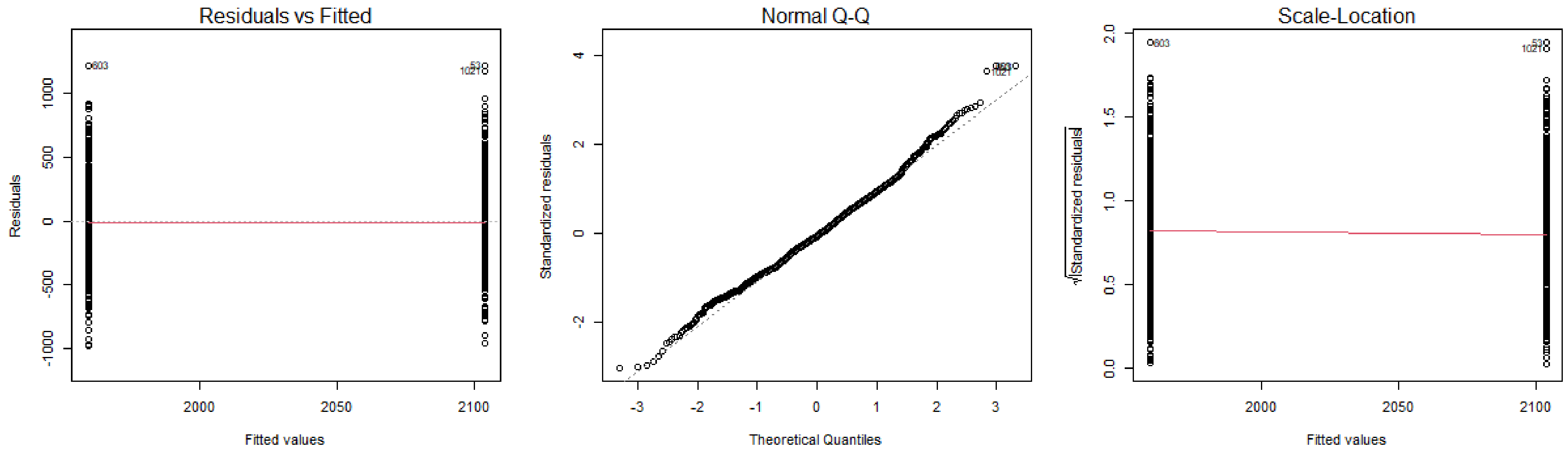
Are the assumptions of linear regression met?

✓ Linearity: The regression line fits perfectly through the sex-specific mean FEV.

✓ Independence: FEV is measured once per child (assume satisfied).

✓ Normality: The residuals look normally distributed.

✓ Homoscedasticity: The residuals appear to have equal variance for males and females.



Conclusion Statement

We examined the relationship between sex and FEV using linear regression. The estimated regression model was $1959.10 + 144.71X_{MALE}$, where X_{MALE} was an indicator variable for male sex. We rejected the null hypothesis that mean FEV was identical for both males and females; mean FEV in males was 144.71ml (95% CI = 106.61, 182.82) higher than females. The regression model assumptions of linearity, normality, and homoscedasticity were evaluated using analysis of residuals and appeared to be satisfied.

What if we had coded sex 1=female, 0=male?

```
> lm(fev ~ female,
+     data = chs %>% mutate(female = 1-male)) %>%
+   summary()

Call:
lm(formula = fev ~ female, data = chs %>% mutate(female = 1 -
    male))

Residuals:
    Min      1Q  Median      3Q     Max
-974.26 -235.28  -17.68  206.88 1220.69

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2103.82      13.75 152.989  < 2e-16 ***
female       -144.71      19.42  -7.451 1.86e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 322.8 on 1103 degrees of freedom
  (95 observations deleted due to missingness)
Multiple R-squared:  0.04793,	Adjusted R-squared:  0.04706
F-statistic: 55.52 on 1 and 1103 DF,  p-value: 1.861e-13
```

For females:

$$\hat{Y}(X = 1) = 2103.82 - 144.71(1) = 1959.10$$

For males:

$$\hat{Y}(X = 0) = 2103.82 - 144.71(0) = 2103.82$$

The intercept is the value of $\hat{Y}$ for the “baseline” group (the group where X=0).

Recap

- Binary X variables have the same interpretation approach in linear regression: a 1-unit increase in X is associated with a β-unit increase in Y

- Because of this, it is important to know which category is coded as X=1 and which is coded as X=0

Recap

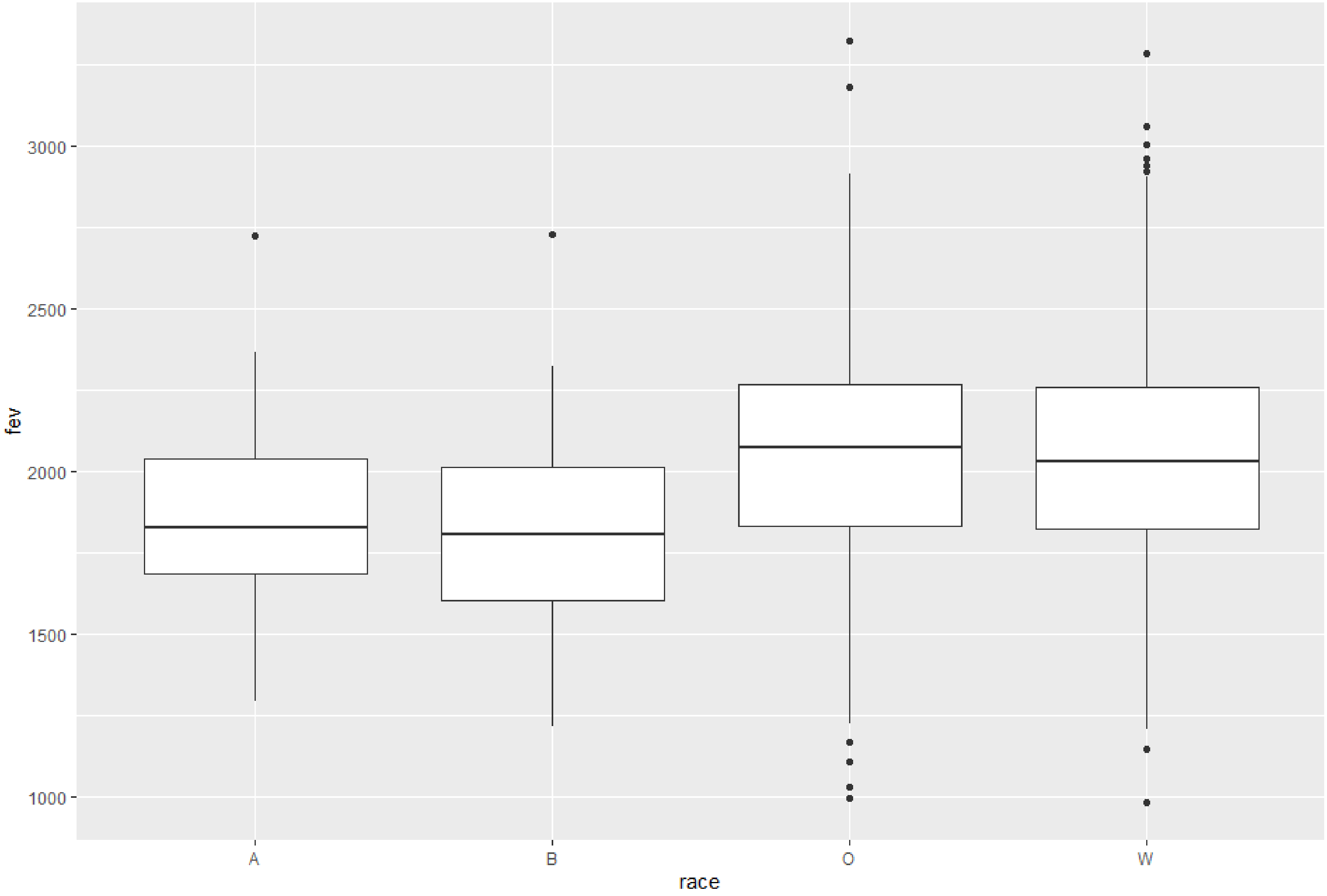
- Implement and interpret an analysis with a binary predictor
- Compare and contrast a t-test vs. a linear regression approach for binary X variables

How could we use regression with a multi-category predictor?

Note: I recoded some race values to make this example easier. See class R code.

```
> chs %>%
+   group_by(race) %>%
+   skim(fev)
-- Data Summary ------------------------
                           Values
Name                       Piped data
Number of rows             1200
Number of columns          23
_______________________
Column type frequency:
  numeric                  1
________________________
Group variables            race

-- Variable type: numeric ------------------------------------------------------------------------
# A tibble: 4 x 12
  skim_variable race  n_missing complete_rate  mean    sd    p0   p25   p50   p75  p100 hist
* <chr>         <chr>     <int>         <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <chr>
1 fev           A             1         0.980 1866.  276. 1296. 1686. 1829. 2040. 2724. ▂▇▅▂▁
2 fev           B             8         0.860 1810.  282. 1215. 1605. 1806. 2012. 2730. ▃▇▇▁▁
3 fev           O            36         0.917 2056.  336.  996. 1833. 2072  2268  3324. ▁▅▇▂▁
4 fev           W            50         0.924 2046.  325.  985. 1825. 2031. 2258. 3283. ▁▅▇▂▁
```

It appears that FEV varies by race.



We will code race with a series of **dummy variables** (an extension to what we did with sex).

To do this, we must pick a **reference group**. The reference group is somewhat arbitrary, but Hardy (1993) suggests the following considerations that should guide the choice of reference group:

- The reference group should serve as a useful “baseline” comparison (e.g., a control group).
- For clarity of interpretation, the baseline group should be well-defined and not a “catch-all” group (e.g., “other”).
- The reference group should not have small sample size relative to other groups.

6. Categorical Predictors

Let’s create a set of dummy variables using “white” as the reference group.

For any variable with *k* categories, you will need to create *k-1* dummy variables.

X_{RACE}	X_A	X_B	X_{Oth}
“A” (Asian)	1	0	0
“B” (Black)	0	1	0
“O” (Other)	0	0	1
“W” (White)	0	0	0

In other words:

$$X_A = \begin{cases} 1, X_{RACE} = Asian \\ 0, Otherwise \end{cases}$$

$$X_B = \begin{cases} 1, X_{RACE} = Black \\ 0, Otherwise \end{cases}$$

$$X_{Oth} = \begin{cases} 1, X_{RACE} = Other \\ 0, Otherwise \end{cases}$$

We will know a participant is white if they have “0” for all three of these.

Results from the dummy variable regression.

```
> lm(fev ~ race_a + race_b + race_o, data = chs) %>%
+   summary()

Call:
lm(formula = fev ~ race_a + race_b + race_o, data = chs)

Residuals:
     Min       1Q   Median       3Q      Max 
-1061.37  -217.26    -6.61   208.25  1267.24 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2046.22      13.18 155.195  < 2e-16 ***
race_a       -180.30      47.87  -3.767 0.000174 ***
race_b       -236.12      48.32  -4.887 1.18e-06 ***
race_o         10.23      20.99   0.487 0.626040    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 325.4 on 1101 degrees of freedom
  (95 observations deleted due to missingness)
Multiple R-squared:  0.03439,	Adjusted R-squared:  0.03176 
F-statistic: 13.07 on 3 and 1101 DF,  p-value: 2.161e-08
```

For white:

$$\hat{Y} = 2046.22 - 180.30(0) - 236.12(0) + 10.23(0) = 2046.22$$

For Asian:

$$\hat{Y} = 2046.22 - 180.30(1) - 236.12(0) + 10.23(0) = 1865.92$$

For Black:

$$\hat{Y} = 2046.22 - 180.30(0) - 236.12(1) + 10.23(0) = 1810.10$$

For other:

$$\hat{Y} = 2046.22 - 180.30(0) - 236.12(0) + 10.23(1) = 2056.45$$

For our equation:

$$\hat{Y} = \beta_0 + \beta_A X_A + \beta_B X_B + \beta_{Oth} X_{Oth}$$

β_A tests whether the mean FEV for Asian is different than for white
β_B tests whether the mean FEV for Black is different than for white
β_{Oth} tests whether the mean FEV for other race is different than for white

6. Categorical Predictors

We can test the overall effect of race. Namely,

H_0: FEV is not associated with race (or, equivalently)

H_0: $\beta_A = 0 \;\&\; \beta_B = 0 \;\&\; \beta_{Oth} = 0$

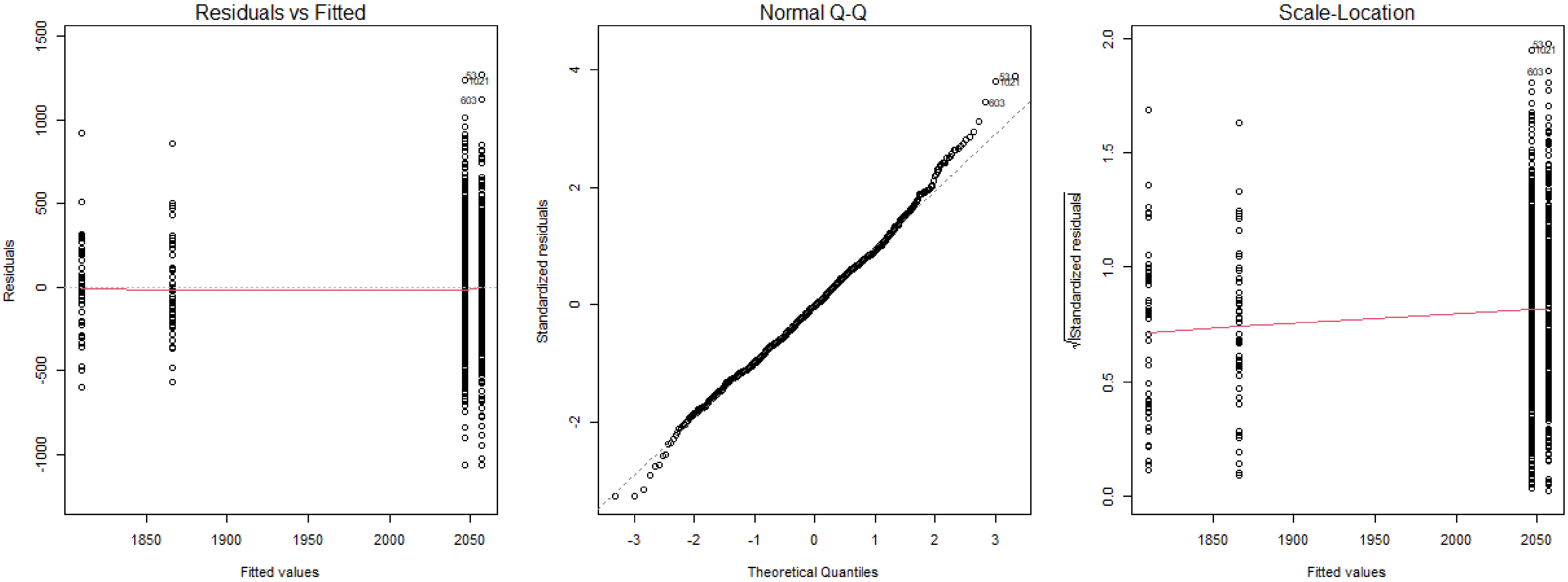
H_A: At least one $\beta \neq 0$

This is tested with the F-statistic:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2046.22      13.18 155.195  < 2e-16 ***
race_a       -180.30      47.87  -3.767 0.000174 ***
race_b       -236.12      48.32  -4.887 1.18e-06 ***
race_o         10.23      20.99   0.487 0.626040    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 325.4 on 1101 degrees of freedom
  (95 observations deleted due to missingness)
Multiple R-squared:  0.03439,	Adjusted R-squared:  0.03176 
F-statistic: 13.07 on 3 and 1101 DF,  p-value: 2.161e-08
```

These data indicate that there is a statistically significant association between race and FEV (p<.001). 3.44% of the variation in FEV is explained by race.

What do we think about the residuals?



Note the following about dummy variables:

- Each value of the original race variable will be translated into a unique combination of dummy code values.
- All dummy variable coefficients will be interpreted relative to the reference group (here, white).
- All codes must be included in the regression equation as a complete set.
- This method is analogous to performing a one-way ANOVA (the F-value for the model is identical).
- Many other types of coding systems are available for categorical variables (https://stats.idre.ucla.edu/spss/faq/coding-systems-for-categorical-variables-in-regression-analysis/)

We can use **factor variables** to automate our coding of dummy variables:

```
> lm(fev ~ factor(race), data = chs) %>% summary()

Call:
lm(formula = fev ~ factor(race), data = chs)

Residuals:
     Min       1Q   Median       3Q      Max 
-1061.37  -217.26    -6.61   208.25  1267.24 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    1865.92      46.01  40.550  < 2e-16 ***
factor(race)B   -55.82      65.41  -0.853 0.393597    
factor(race)O   190.53      48.83   3.902 0.000101 ***
factor(race)W   180.30      47.87   3.767 0.000174 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 325.4 on 1101 degrees of freedom
  (95 observations deleted due to missingness)
Multiple R-squared:  0.03439,	Adjusted R-squared:  0.03176 
F-statistic: 13.07 on 3 and 1101 DF,  p-value: 2.161e-08
```

Now, the interpretation of the β coefficients is the FEV of each race group in comparison to Asian (we see that the “A” dummy variable was omitted).

By default, factor() will use the lowest value of the reference group.

If we specify “W” as the reference group, then we get the same output as before:

```
> lm(fev ~ relevel(factor(race), ref = "W"), data = chs) %>% summary()

Call:
lm(formula = fev ~ relevel(factor(race), ref = "W"), data = chs)

Residuals:
     Min       1Q   Median       3Q      Max 
-1061.37  -217.26    -6.61   208.25  1267.24 

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)                        2046.22      13.18 155.195  < 2e-16 ***
relevel(factor(race), ref = "W")A  -180.30      47.87  -3.767 0.000174 ***
relevel(factor(race), ref = "W")B  -236.12      48.32  -4.887 1.18e-06 ***
relevel(factor(race), ref = "W")O    10.23      20.99   0.487 0.626040    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 325.4 on 1101 degrees of freedom
  (95 observations deleted due to missingness)
Multiple R-squared:  0.03439,	Adjusted R-squared:  0.03176 
F-statistic: 13.07 on 3 and 1101 DF,  p-value: 2.161e-08
```

Here we did a lot of variable manipulation in the lm() function. In practice, you should do this variable manipulation first, and then perform the lm().

Recap

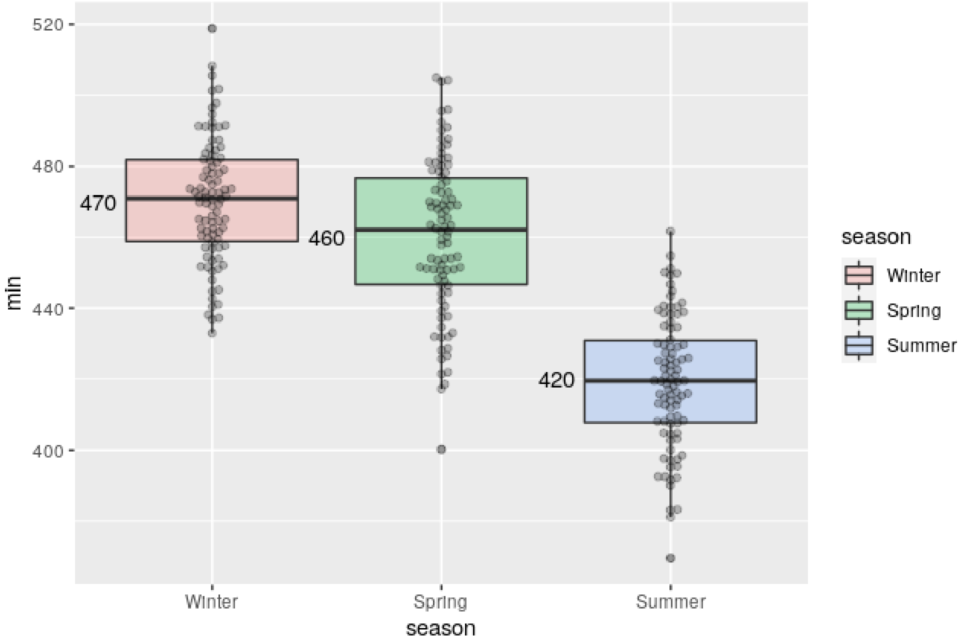
- Categorical predictors must first be dummy-coded relative to a reference group for use in regression
- The β coefficients for each group reflect the estimated difference in $\hat{Y}$ for each group relative to the reference group
- Dummy variables must be considered as a complete set in analysis
- A “factor” variable in R will automatically be treated as a dummy variable set in analysis

Recap

- Implement and interpret an analysis with a categorical predictor
- Explain the meaning of coefficients in a dummy-coded variable set
- Explain how to statistically test the collective effect of a dummy-coded variable set
- Explain how changing the reference group will impact the output of a regression model

Test Yourself

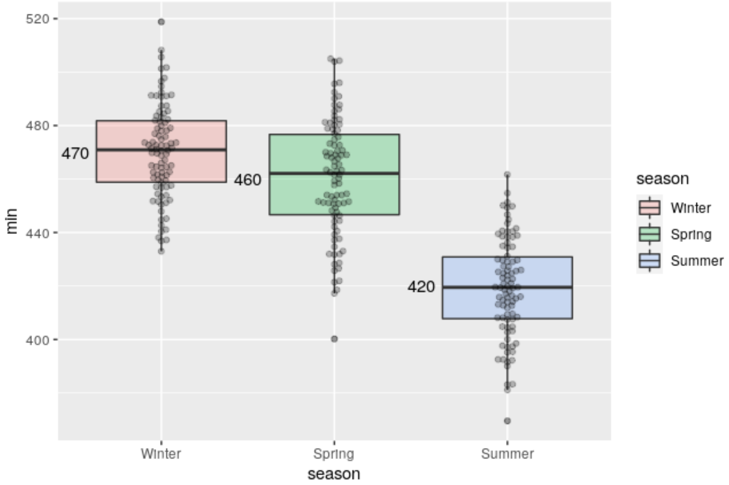
Suppose you regressed minutes of exercise (Y) on season (X). Winter is the baseline group. The estimated equation is $\hat{Y} = \beta_0 + \beta_1 X_{spring} + \beta_2 X\_summer$. What is the value of β_0? (The mean minutes within each season are displayed below.)



Test Yourself

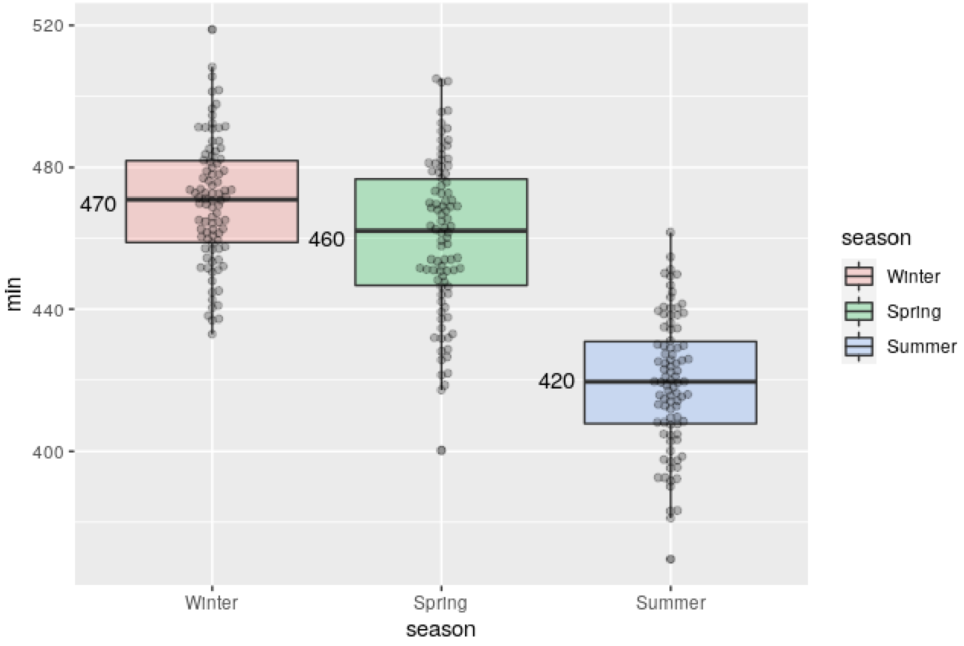
Suppose you regressed minutes of exercise (Y) on season (X). Winter is the baseline group. The estimated equation is $\hat{Y} = \beta_0 + \beta_1 X_{spring} + \beta_2 X\_summer$. What is the value of β_0? (The mean minutes within each season are displayed below.)

Since winter is the reference group, it would equal 470, the mean number of minutes during winter.



Test Yourself

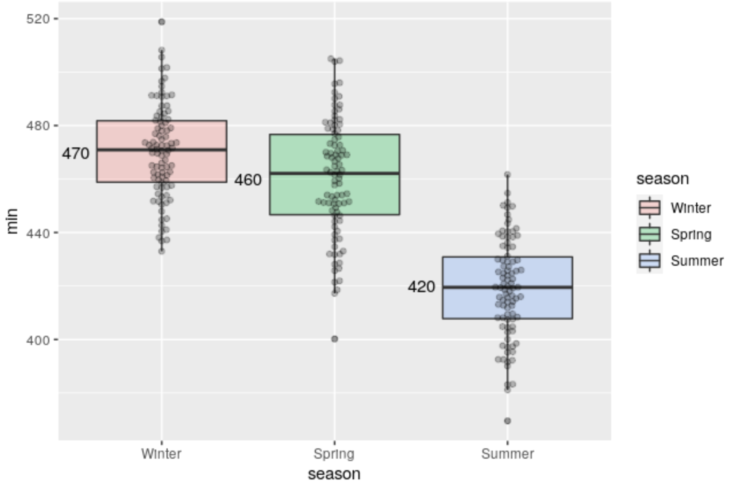
Suppose you regressed minutes of exercise (Y) on season (X). Winter is the baseline group. The estimated equation is $\hat{Y} = \beta_0 + \beta_1 X_{spring} + \beta_2 X\_summer$. What is the value of β_1? (The mean minutes within each season are displayed below.)



Test Yourself

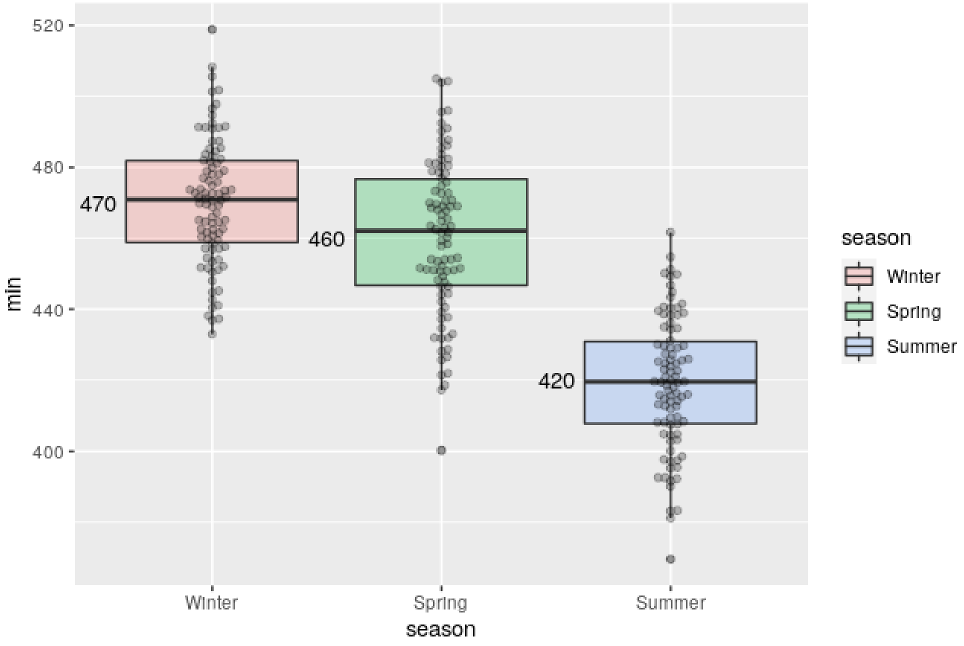
Suppose you regressed minutes of exercise (Y) on season (X). Winter is the baseline group. The estimated equation is $\hat{Y} = \beta_0 + \beta_1 X_{spring} + \beta_2 X\_summer$. What is the value of β_1? (The mean minutes within each season are displayed below.)

This is the difference in mean minutes for a 1-unit increase in Xspring (which reflects comparing spring vs. winter). Therefore the value would be (460 – 470 = -10).



Test Yourself

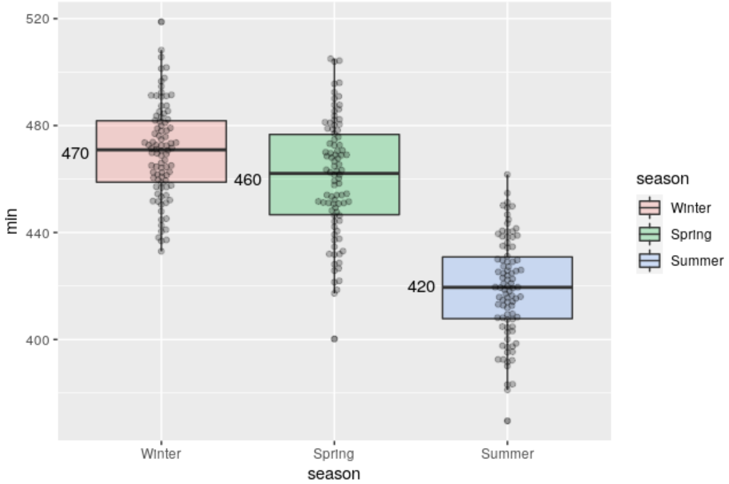
Suppose you regressed minutes of exercise (Y) on season (X). Winter is the baseline group. The estimated equation is $\hat{Y} = \beta_0 + \beta_1 X_{spring} + \beta_2 X\_summer$. What is the value of β_2? (The mean minutes within each season are displayed below.)



Test Yourself

Suppose you regressed minutes of exercise (Y) on season (X). Winter is the baseline group. The estimated equation is $\hat{Y} = \beta_0 + \beta_1 X_{spring} + \beta_2 X\_summer$. What is the value of β_1? (The mean minutes within each season are displayed below.)

This is the difference in mean minutes for a 1-unit increase in Xsummer (which reflects comparing summer vs. winter). Therefore the value would be (420 – 470 = -50).



7. Recap

- Check your assumptions are met after performing a regression model.
- Log transformations of the Y variable (and sometimes of the X variable) can better satisfy some assumptions.
- Whenever you transform a variable, there will be more difficulty in interpreting your results; use only when necessary.
- Regression with a binary predictor is equivalent to a t-test.
- Regression with a multi-category predictor is equivalent to an ANOVA.

Packages and Functions

- `plot(lm_object)`