More Linear Modeling

Fundamental Techniques in Data Science with R

Utrecht University

Kyle M. Lang

Department of Methodology & Statistics
Utrecht University

Outline

Categorical Predictors
- Dummy Coding
- Significance Testing for Dummy Codes

Moderation
- Categorical Moderators

Assumptions & Diagnostics
- Regression Diagnostics

Influential Observations
- Treating Influential Observations

Categorical Predictors

Categorical Predictors

Most of the predictors we've considered thus far have been *quantitative*.

- Continuous variables that can take any real value in their range
- Interval or Ratio scaling

We often want to include grouping factors as predictors.

- These variables are *qualitative*.
  - Their values are simply labels.
  - There is no ordering of the categories.
  - Nominal scaling

How to Model Categorical Predictors

We need to be careful when we include categorical predictors into a regression model.

- The variables need to be coded before entering the model

Consider the following indicator of major:
$X_{maj} = \{1 = Law, 2 = Economics, 3 = Data\ Science\}$

- What would happen if we naïvely used this variable to predict program satisfaction?

How to Model Categorical Predictors

```
mDat <- readRDS("../data/major_data.rds")
mDat[seq(25, 150, 25), ]

    sat majF majN
25  1.9  law    1
50  1.4  law    1
75  4.3 econ    2
100 4.1 econ    2
125 5.7   ds    3
150 5.1   ds    3

out1 <- lm(sat ~ majN, data = mDat)
```

How to Model Categorical Predictors

```
partSummary(out1, -1)

Residuals:
   Min     1Q Median     3Q    Max
-1.303 -0.313 -0.113  0.342  1.342

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.33200    0.12060  -2.753  0.00664
majN         2.04500    0.05582  36.632  < 2e-16

Residual standard error: 0.5582 on 148 degrees of freedom
Multiple R-squared:  0.9007,Adjusted R-squared:    0.9
F-statistic:  1342 on 1 and 148 DF,  p-value: < 2.2e-16
```

Dummy Coding

The most common way to code categorical predictors is *dummy coding*.

- A G-level factor must be converted into a set of $G - 1$ dummy codes.
- Each code is a variable on the dataset that equals 1 for observations corresponding to the code's group and equals 0, otherwise.
- The group without a code is called the *reference group*.

Example Dummy Code

Let's look at the simple example of coding biological sex:

	sex	male
1	female	0
2	male	1
3	male	1
4	female	0
5	male	1
6	female	0
7	female	0
8	male	1
9	female	0
10	female	0

Example Dummy Codes

Now, a slightly more complex example:

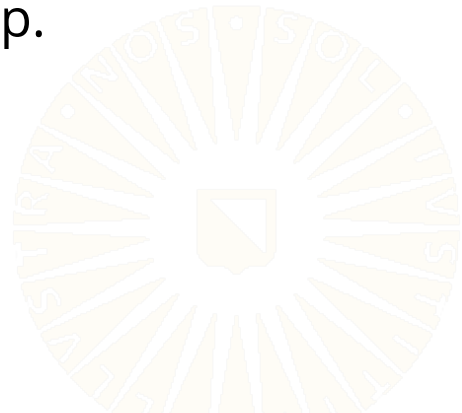
	drink	juice	tea
1	juice	1	0
2	coffee	0	0
3	tea	0	1
4	tea	0	1
5	tea	0	1
6	tea	0	1
7	juice	1	0
8	tea	0	1
9	coffee	0	0
10	juice	1	0

Using Dummy Codes

To use the dummy codes, we simply include the $G-1$ codes as $G-1$ predictor variables in our regression model.

$$Y = \beta_0 + \beta_1 X_{male} + \varepsilon$$
$$Y = \beta_0 + \beta_1 X_{juice} + \beta_2 X_{tea} + \varepsilon$$

- The intercept corresponds to the mean of Y for the reference group.
- Each slope represents the difference between the mean of Y in the coded group and the mean of Y in the reference group.



Example

First, an example with a single, binary dummy code:

```
## Read in some data:
cDat <- readRDS("../data/cars_data.rds")

## Fit and summarize the model:
out2 <- lm(price ~ mtOpt, data = cDat)
```

Example

```
partSummary(out2, -1)

Residuals:
    Min      1Q  Median      3Q     Max
-10.341  -6.338  -3.141   2.662  38.059

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   23.841      1.623  14.691   <2e-16
mtOpt         -6.603      2.004  -3.295   0.0014

Residual standard error: 9.18 on 91 degrees of freedom
Multiple R-squared:  0.1066,Adjusted R-squared:  0.09679
F-statistic: 10.86 on 1 and 91 DF,  p-value: 0.001403
```

Interpretations

- The average price of a car without the option for a manual transmission is $\hat{\beta}_0 = 23.84$ thousand dollars.
- The average difference in price between cars that have manual transmissions as an option and those that do not is $\hat{\beta}_1 = -6.6$ thousand dollars.

Example

Fit a more complex model:

```
out3 <- lm(price ~ front + rear, data = cDat)
partSummary(out3, -1)

Residuals:
    Min      1Q  Median      3Q     Max
-14.050  -6.250  -1.236   3.264  32.950

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.63000    2.76119   6.385 7.33e-09
front       -0.09418    2.96008  -0.032  0.97469
rear        11.32000    3.51984   3.216  0.00181

Residual standard error: 8.732 on 90 degrees of freedom
Multiple R-squared:  0.2006,Adjusted R-squared:  0.1829
F-statistic: 11.29 on 2 and 90 DF,  p-value: 4.202e-05
```

Interpretations

- The average price of a four-wheel-drive car is $\hat{\beta}_0 = 17.63$ thousand dollars.
- The average difference in price between front-wheel-drive cars and four-wheel-drive cars is $\hat{\beta}_1 = -0.09$ thousand dollars.
- The average difference in price between rear-wheel-drive cars and four-wheel-drive cars is $\hat{\beta}_2 = 11.32$ thousand dollars.

Example

Include two sets of dummy codes:

```
out4 <- lm(price ~ mtOpt + front + rear, data = cDat)
partSummary(out4, -c(1, 2))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  21.7187     2.9222   7.432 6.25e-11
mtOpt        -5.8410     1.8223  -3.205  0.00187
front        -0.2598     2.8189  -0.092  0.92677
rear         10.5169     3.3608   3.129  0.00237

Residual standard error: 8.314 on 89 degrees of freedom
Multiple R-squared:  0.2834,Adjusted R-squared:  0.2592
F-statistic: 11.73 on 3 and 89 DF,  p-value: 1.51e-06
```

Interpretations

- The average price of a four-wheel-drive car that does not have a manual transmission option is $\hat{\beta}_0 = 21.72$ thousand dollars.
- After controlling for drive type, the average difference in price between cars that have manual transmissions as an option and those that do not is $\hat{\beta}_1 = -5.84$ thousand dollars.
- After controlling for transmission options, the average difference in price between front-wheel-drive cars and four-wheel-drive cars is $\hat{\beta}_2 = -0.26$ thousand dollars.
- After controlling for transmission options, the average difference in price between rear-wheel-drive cars and four-wheel-drive cars is $\hat{\beta}_3 = 10.52$ thousand dollars.

Significance Testing

For variables with only two levels, we can test the overall factor's significance by evaluating the significance of a single dummy code.

```
partSummary(out2, -c(1, 2))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   23.841      1.623  14.691   <2e-16
mtOpt         -6.603      2.004  -3.295   0.0014

Residual standard error: 9.18 on 91 degrees of freedom
Multiple R-squared:  0.1066,Adjusted R-squared:  0.09679
F-statistic: 10.86 on 1 and 91 DF,  p-value: 0.001403
```

Significance Testing

For variables with more than two levels, we need to simultaneously evaluate the significance of each of the variable's dummy codes.

```
partSummary(out4, -c(1, 2))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  21.7187     2.9222   7.432 6.25e-11
mtOpt        -5.8410     1.8223  -3.205  0.00187
front        -0.2598     2.8189  -0.092  0.92677
rear         10.5169     3.3608   3.129  0.00237

Residual standard error: 8.314 on 89 degrees of freedom
Multiple R-squared:  0.2834,Adjusted R-squared:  0.2592
F-statistic: 11.73 on 3 and 89 DF,  p-value: 1.51e-06
```

Significance Testing

```
summary(out4)$r.squared - summary(out2)$r.squared

[1] 0.1767569

anova(out2, out4)

Analysis of Variance Table

Model 1: price ~ mtOpt
Model 2: price ~ mtOpt + front + rear
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     91 7668.9
2     89 6151.6  2    1517.3 10.976 5.488e-05 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Significance Testing

For models with a single nominal factor is the only predictor, we use the omnibus F-test.

```
partSummary(out3, -c(1, 2))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.63000    2.76119   6.385 7.33e-09
front       -0.09418    2.96008  -0.032  0.97469
rear        11.32000    3.51984   3.216  0.00181

Residual standard error: 8.732 on 90 degrees of freedom
Multiple R-squared:  0.2006,Adjusted R-squared:  0.1829
F-statistic: 11.29 on 2 and 90 DF,  p-value: 4.202e-05
```

Moderation

Moderation

So far we've been discussing *additive models*.

- Additive models allow us to examine the partial effects of several predictors on some outcome.
  - The effect of one predictor does not change based on the values of other predictors.

Now, we'll discuss *moderation*.

- Moderation allows us to ask *when* one variable, X, affects another variable, Y.
  - We're considering the conditional effects of X on Y given certain levels of a third variable Z.

Equations

In additive MLR, we might have the following equation:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon$$

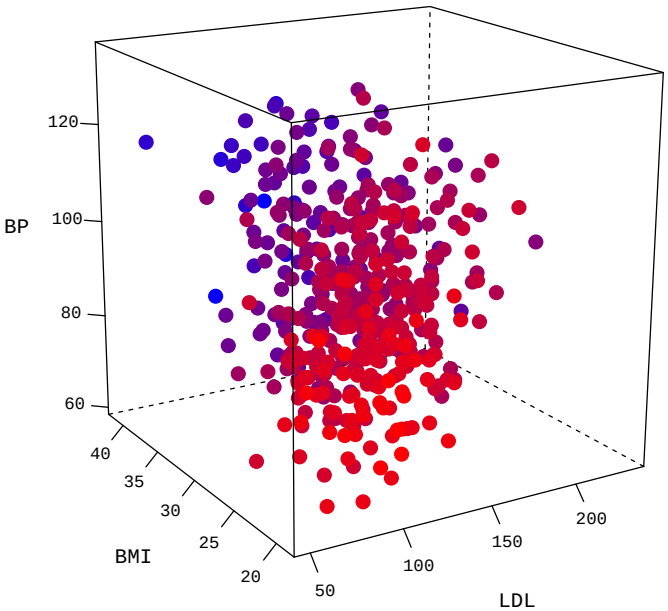
This additive equation assumes that X and Z are independent predictors of Y.

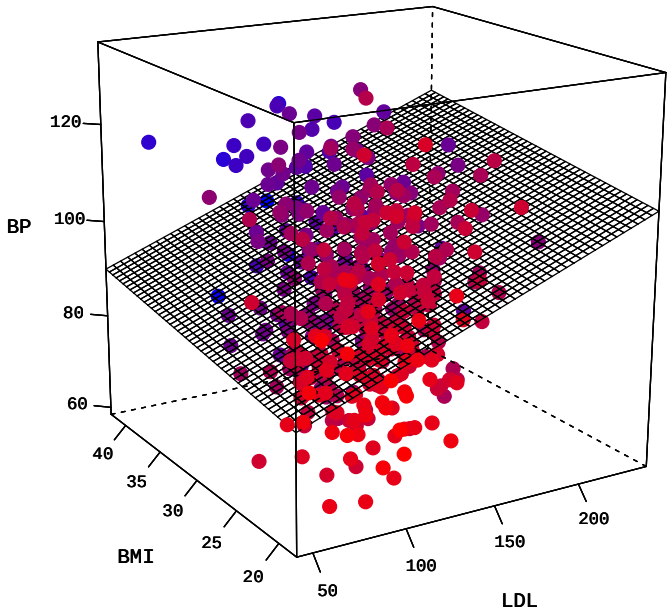
When X and Z are independent predictors, the following are true:

- X and Z *can* be correlated.
- β_1 and β_2 are *partial* regression coefficients.
- The effect of X on Y is the same at **all levels** of Z, and the effect of Z on Y is the same at **all levels** of X.

Additive Regression

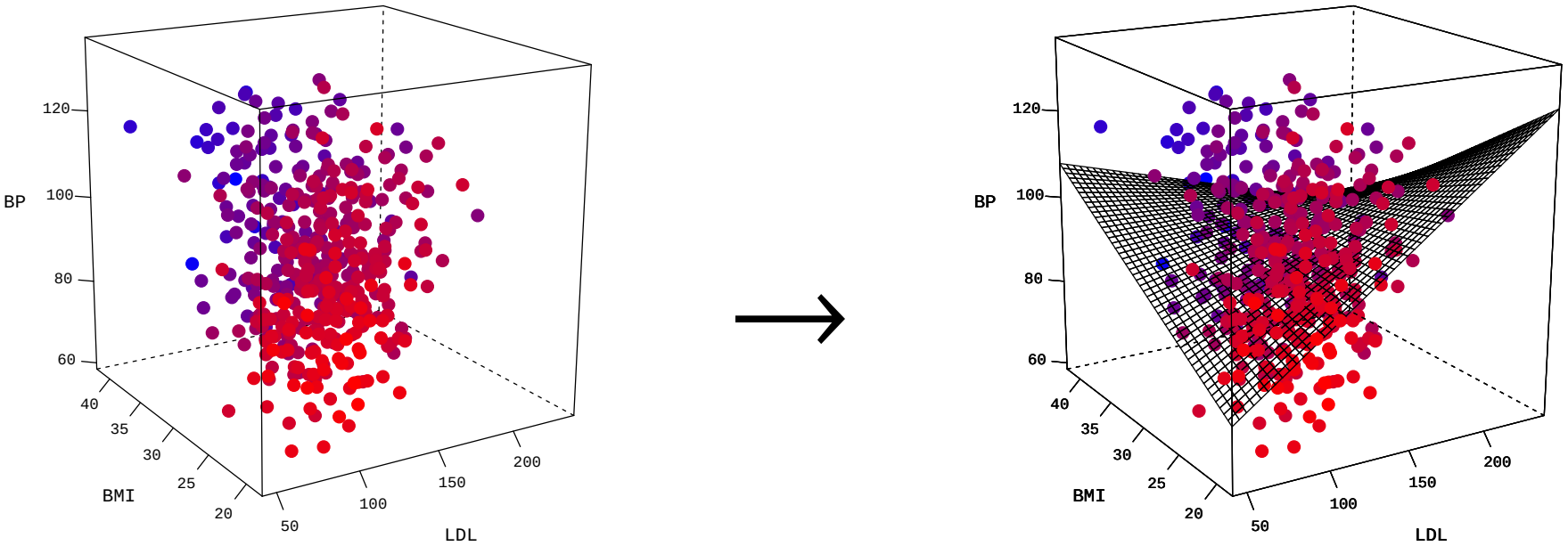
The effect of X on Y is the same at **all levels** of Z.





Moderated Regression

The effect of X on Y varies **as a function** of Z.



Equations

The following derivation is adapted from Hayes (2017).

- When testing moderation, we hypothesize that the effect of X on Y varies as a function of Z.
- We can represent this concept with the following equation:

$$Y = \beta_0 + f(Z)X + \beta_2 Z + \varepsilon \tag{1}$$

Equations

The following derivation is adapted from Hayes (2017).

- When testing moderation, we hypothesize that the effect of X on Y varies as a function of Z.
- We can represent this concept with the following equation:

$$Y = \beta_0 + f(Z)X + \beta_2 Z + \varepsilon \tag{1}$$

- If we assume that Z linearly (and deterministically) affects the relationship between X and Y, then we can take:

$$f(Z) = \beta_1 + \beta_3 Z \tag{2}$$

Equations

- Substituting Equation 2 into Equation 1 leads to:

$$Y = \beta_0 + (\beta_1 + \beta_3 Z)X + \beta_2 Z + \varepsilon$$

Equations

- Substituting Equation 2 into Equation 1 leads to:

$$Y = \beta_0 + (\beta_1 + \beta_3 Z)X + \beta_2 Z + \varepsilon$$

- Which, after distributing X and reordering terms, becomes:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + \varepsilon$$

Testing Moderation

Now, we have an estimable regression model that quantifies the linear moderation we hypothesized.

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + \varepsilon$$

- To test for significant moderation, we simply need to test the significance of the interaction term, XZ.
  - Check if $\hat{\beta}_3$ is significantly different from zero.

Interpretation

Given the following equation:

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 Z + \hat{\beta}_3 XZ + \hat{\varepsilon}$$

- $\hat{\beta}_3$ quantifies the effect of Z on the focal effect (the $X \rightarrow Y$ effect).
  - For a unit change in Z, $\hat{\beta}_3$ is the expected change in the effect of X on Y.
- $\hat{\beta}_1$ and $\hat{\beta}_2$ are *conditional effects*.
  - Interpreted where the other predictor is zero.
  - For a unit change in X, $\hat{\beta}_1$ is the expected change in Y, when $Z = 0$.
  - For a unit change in Z, $\hat{\beta}_2$ is the expected change in Y, when $X = 0$.

Example

Still looking at the *diabetes* dataset.

- We suspect that patients' BMIs are predictive of their average blood pressure.
- We further suspect that this effect may be differentially expressed depending on the patients' LDL levels.

Example

```
## Focal Effect:
out0 <- lm(bp ~ bmi, data = dDat)
partSummary(out0, -c(1, 2))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  61.9973     3.6659   16.91   <2e-16
bmi           1.2379     0.1371    9.03   <2e-16

Residual standard error: 12.72 on 440 degrees of freedom
Multiple R-squared:  0.1563,Adjusted R-squared:  0.1544
F-statistic: 81.54 on 1 and 440 DF,  p-value: < 2.2e-16
```

Example

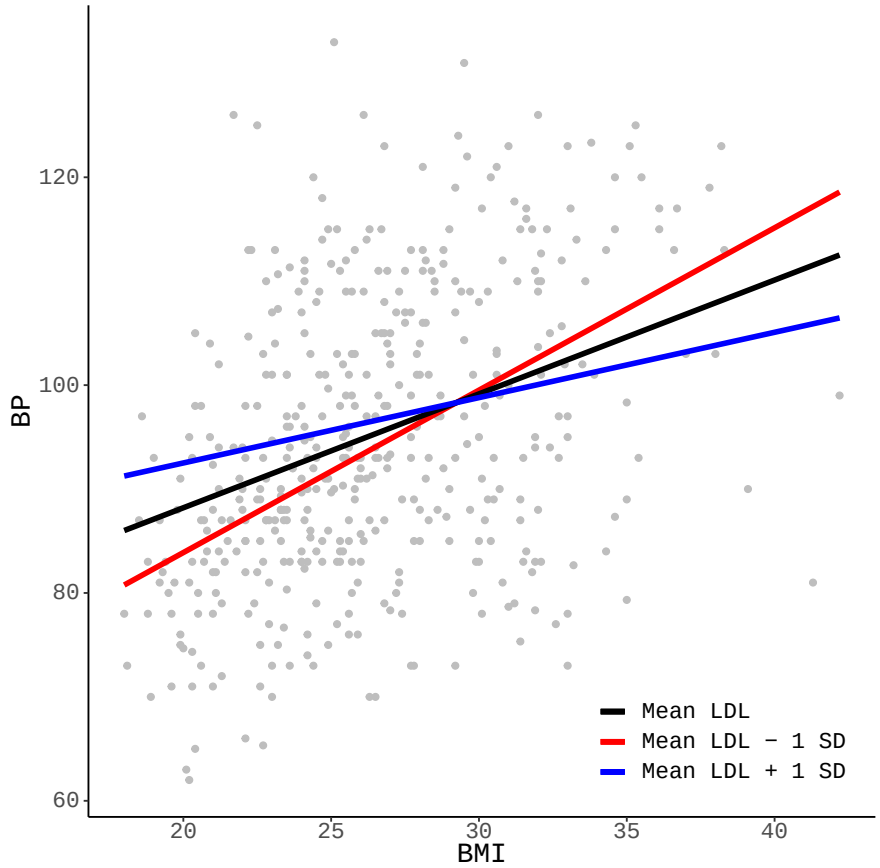
```
## Additive Model:
out1 <- lm(bp ~ bmi + ldl, data = dDat)
partSummary(out1, -c(1, 2))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 59.26577    3.91281  15.147  < 2e-16
bmi          1.16567    0.14156   8.235 2.08e-15
ldl          0.04016    0.02056   1.953   0.0515

Residual standard error: 12.68 on 439 degrees of freedom
Multiple R-squared:  0.1636,Adjusted R-squared:  0.1598
F-statistic: 42.94 on 2 and 439 DF,  p-value: < 2.2e-16
```

Example

```
## Moderated Model:
out2 <- lm(bp ~ bmi * ldl, data = dDat)
partSummary(out2, -c(1, 2))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.480616  14.291677   1.013 0.311514
bmi          2.867825   0.541312   5.298 1.86e-07
ldl          0.448771   0.127160   3.529 0.000461
bmi:ldl     -0.015352   0.004716  -3.255 0.001221

Residual standard error: 12.54 on 438 degrees of freedom
Multiple R-squared:  0.1834,Adjusted R-squared:  0.1778
F-statistic: 32.78 on 3 and 438 DF,  p-value: < 2.2e-16
```

Visualizing the Interaction

We can get a better idea of the patterns of moderation by plotting the focal effect at conditional values of the moderator.



Categorical Moderators

Categorical moderators encode *group-specific* effects.

- E.g., if we include *sex* as a moderator, we are modeling separate focal effects for males and females.

Given a set of codes representing our moderator, we specify the interactions as before:

$$Y_{total} = \beta_0 + \beta_1 X_{inten} + \beta_2 Z_{male} + \beta_3 X_{inten} Z_{male} + \varepsilon$$

$$Y_{total} = \beta_0 + \beta_1 X_{inten} + \beta_2 Z_{lo} + \beta_3 Z_{mid} + \beta_4 Z_{hi} + \beta_5 X_{inten} Z_{lo} + \beta_6 X_{inten} Z_{mid} + \beta_7 X_{inten} Z_{hi} + \varepsilon$$

Example

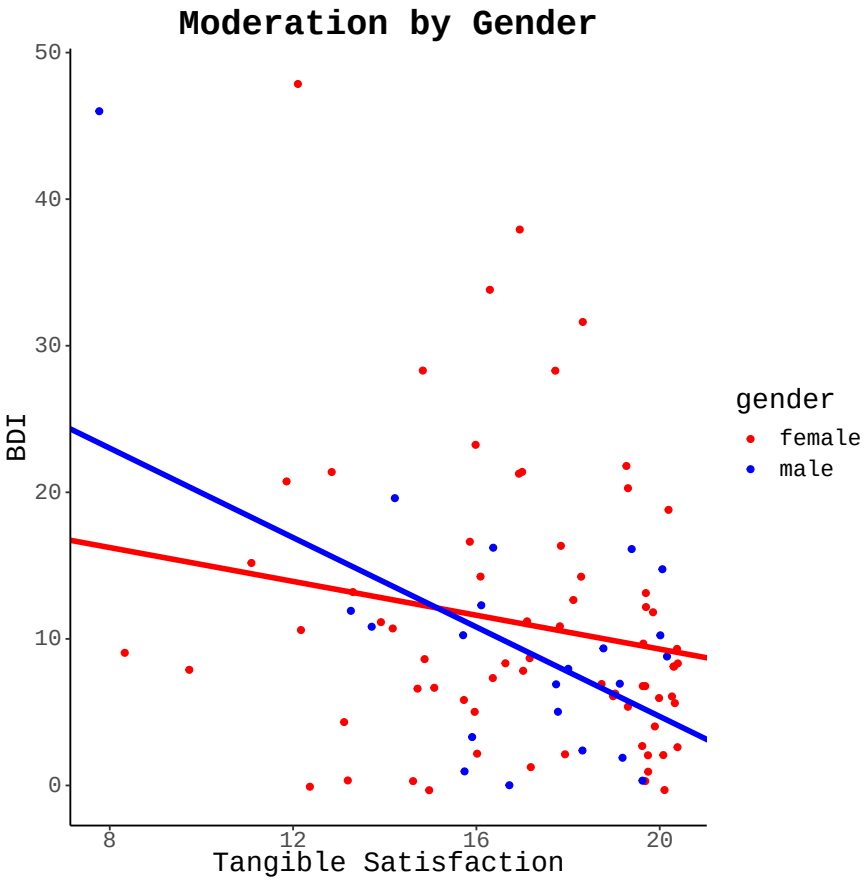
```
## Load data:
socSup <- readRDS(paste0(dataDir, "social_support.rds"))

## Focal effect:
out3 <- lm(bdi ~ tanSat, data = socSup)
partSummary(out3, -c(1, 2))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  24.4089     5.3502   4.562 1.54e-05
tanSat       -0.8100     0.3124  -2.593   0.0111

Residual standard error: 9.278 on 93 degrees of freedom
Multiple R-squared:  0.06742,Adjusted R-squared:  0.05739
F-statistic: 6.723 on 1 and 93 DF,  p-value: 0.01105
```

Example

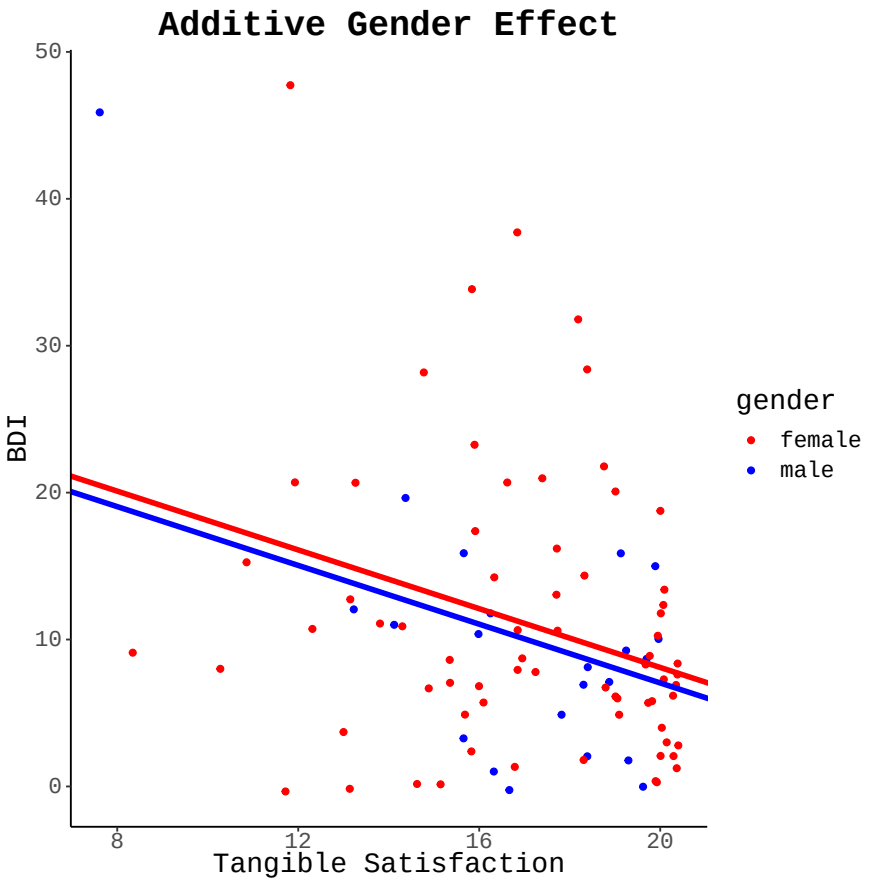
```
## Estimate the interaction:
out4 <- lm(bdi ~ tanSat * sex, data = socSup)
partSummary(out4, -c(1, 2))

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     20.8478     6.2114   3.356  0.00115
tanSat          -0.5772     0.3614  -1.597  0.11372
sexmale         14.3667    12.2054   1.177  0.24223
tanSat:sexmale  -0.9482     0.7177  -1.321  0.18978

Residual standard error: 9.267 on 91 degrees of freedom
Multiple R-squared:  0.08955,Adjusted R-squared:  0.05954
F-statistic: 2.984 on 3 and 91 DF,  p-value: 0.03537
```

Visualizing Categorical Moderation

$$\hat{Y}_{BDI} = 20.85 - 0.58X_{tsat} + 14.37Z_{male} - 0.95X_{tsat}Z_{male}$$



$$\hat{Y}_{BDI} = 28.10 - 1.00X_{tsat} - 1.05Z_{male}$$



Assumptions & Diagnostics

Assumptions of MLR

The assumptions of the linear model can be stated as follows:

1. The model is linear in the parameters.
   - This is OK: $Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + \beta_4 X^2 + \beta_5 X^3 + \varepsilon$
   - This is not: $Y = \beta_0 X^{\beta_1} + \varepsilon$
2. The predictor matrix is *full rank*.
   - $N > P$
   - No X_p can be a linear combination of other predictors.

Assumptions of MLR

3. The predictors are strictly exogenous.
   - The predictors do not correlated with the errors.
   - $\text{Cov}(\hat{Y}, \varepsilon) = 0$
   - $\text{E}[\varepsilon_n] = 0$
4. The errors have constant, finite variance.
   - $\text{Var}(\varepsilon_n) = \sigma^2 < \infty$
5. The errors are uncorrelated.
   - $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0, \ i \neq j$
6. The errors are normally distributed.
   - $\varepsilon \sim \text{N}(0, \sigma^2)$

Assumptions of MLR

The assumption of *spherical errors* combines Assumptions 4 and 5.

$$\text{Var}(\varepsilon) = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & & 0 \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{I}_N$$

We can combine Assumptions 3, 4, 5, and 6 by assuming independent and identically distributed normal errors:

- $\varepsilon \overset{iid}{\sim} \text{N}(0, \sigma^2)$

Consequences of Violating Assumptions

1. If the model is not linear in the parameters, then we're not even working with linear regression.
   - We need to move to entirely different modeling paradigm.
2. If the predictor matrix is not full rank, the model is not estimable.
   - The parameter estimates cannot be uniquely determined from the data.
3. If the predictors are not exogenous, the estimated regression coefficients will be biased.
4. If the errors are not spherical, the standard errors will be biased.
   - The estimated regression coefficients will be unbiased, though.
5. If errors are non-normal, small-sample inferences may be biased.
   - The justification for some tests and procedures used in regression analysis may not hold.

Regression Diagnostics

If some of the assumptions are (grossly) violated, the inferences we make using the model may be wrong.

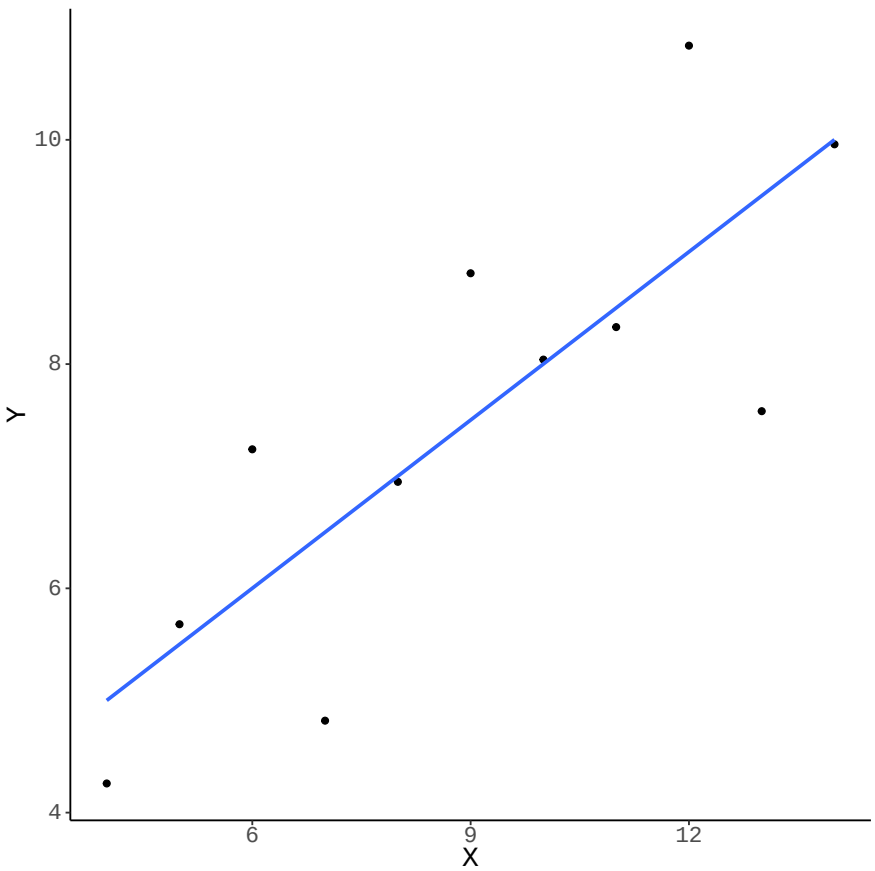
- We need to check the tenability of our assumptions before leaning too heavily on the model estimates.

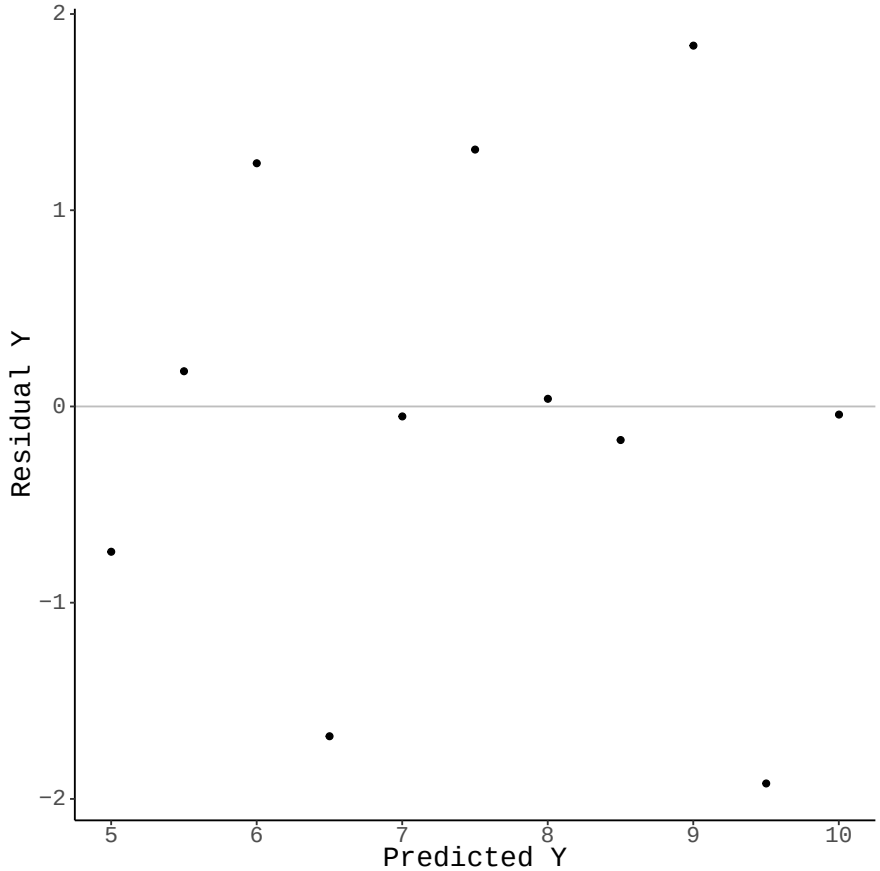
These checks are called *regression diagnostics*.

- Graphical visualizations
- Quantitative indices/measures
- Formal statistical tests

Residual Plots

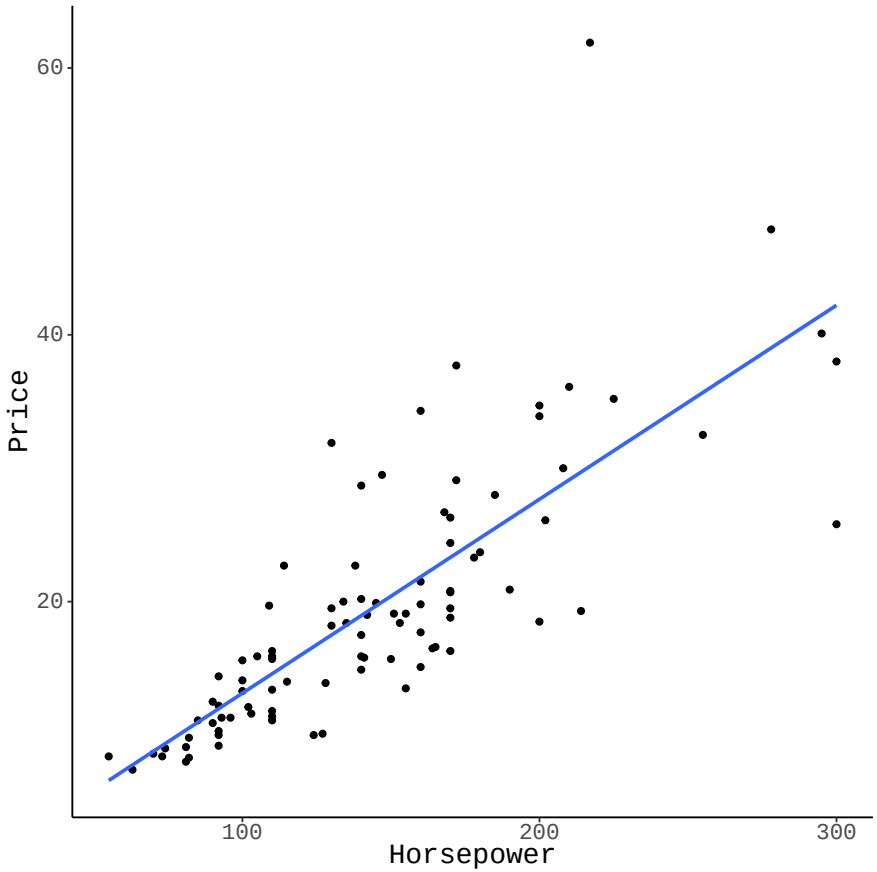
One of the most useful diagnostic graphics is the plot of residuals vs. predicted values.

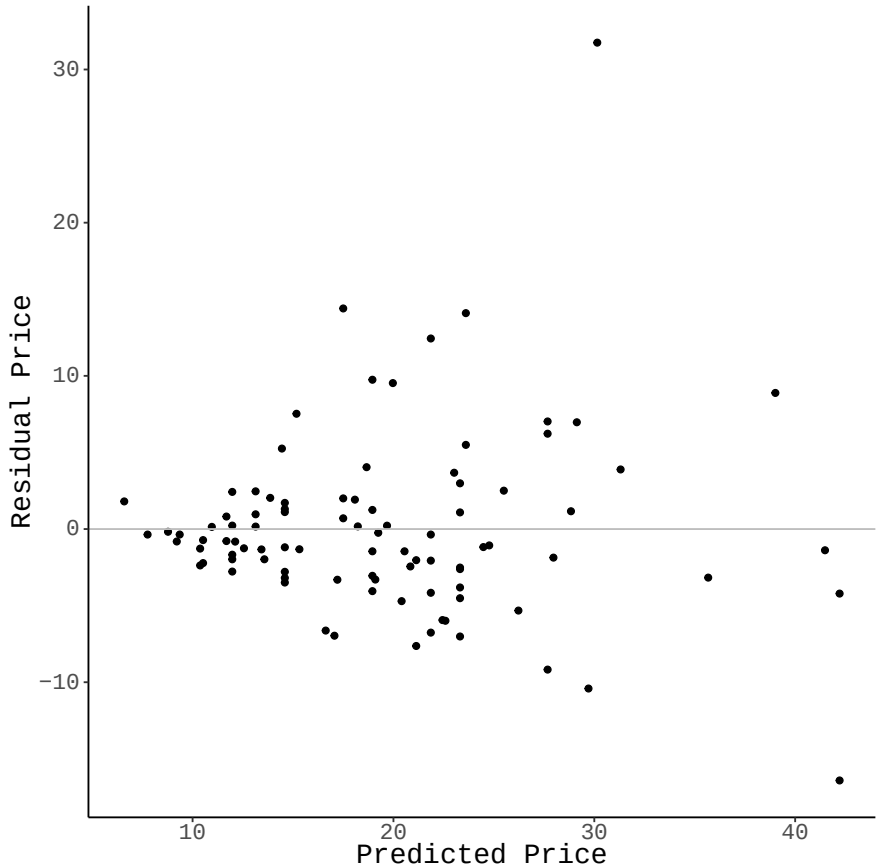




Heteroscedasticity

One commonly encountered problem is non-constant error variance (i.e., *heteroscedasticity*) which violates Assumption 4.

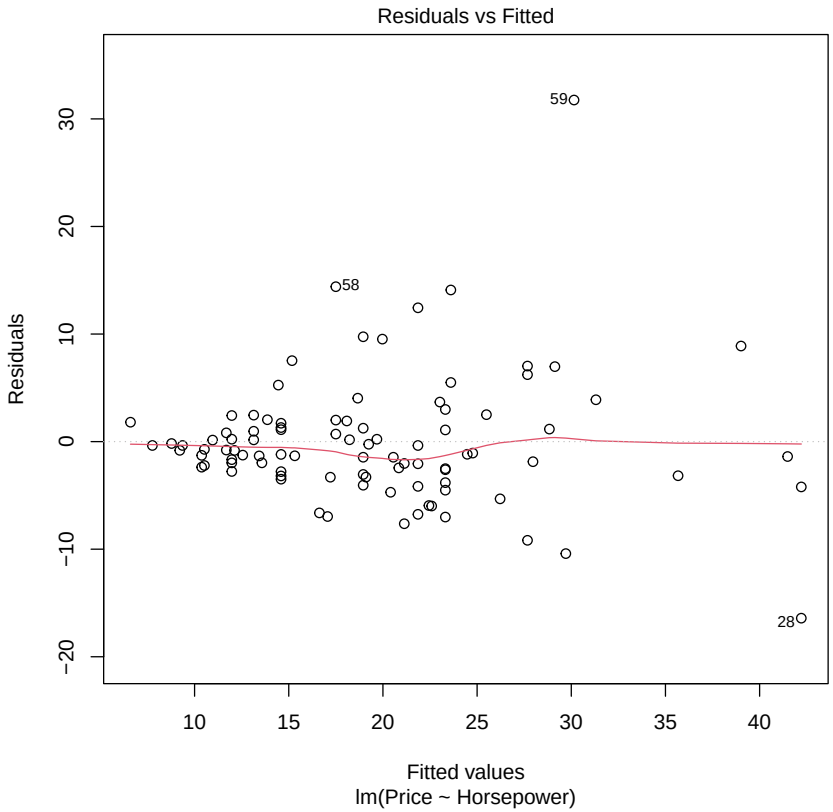




Heteroscedasticity

We can easily generate a simple plot of residuals vs. fitted values by plotting the fitted `lm()` object in R.

```
out1 <- lm(Price ~ Horsepower,
           data = Cars93)

plot(out1, 1)
```



Consequences of Heteroscedasticity

Non-constant error variance will not bias the parameter estimates.

- The best fit line is still correct.
- Our measure of uncertainty around that best fit line is wrong.

Heteroscedasticity will bias standard errors (usually downward).

- Test statistics will be too large.
- CIs will be too narrow.
- We will have inflated Type I error rates.

To get valid inference, we need to address (severe) heteroscedasticity.

Treating Heteroscedasticity

1. Transform your outcome using a concave function (e.g., $\ln(Y)$, $\sqrt{Y}$).
   - These transformations will shrink extreme values more than small/moderate ones.
2. Refit the model using *weighted least squares*.
   - Create inverse weights using functions of the residual variances or quantities highly correlated therewith.
3. Use a *Heteroscedasticity Consistent* (HC) estimate of the asymptotic covariance matrix.
   - Robust standard errors, Huber-White standard errors, Sandwich estimators
   - HC estimators correct the standard errors for non-constant error variance.

Correlated Errors

Errors can become correlated in two basic ways:

1. Serial dependence
   - When modeling longitudinal data, the errors for a given observational unit are correlated over time.
   - We can detect temporal dependence by examining the *autocorrelation* of the residuals.

2. Clustering
   - Your data have some important, unmodeled, grouping structure.
     - Children nested within classrooms
     - Romantic couples
     - Departments within a company
   - We can detect problematic levels of clustering with the *intraclass correlation coefficient* (ICC).
     - We need to know the clustering variable to apply the ICC.

Treating Correlated Errors

Serially dependent errors in a longitudinal model usually indicate an inadequate model.

- Your model is ignoring some important aspect of the temporal variation that is being absorbed by the error terms.
- Hopefully, you can add the missing component to your model.

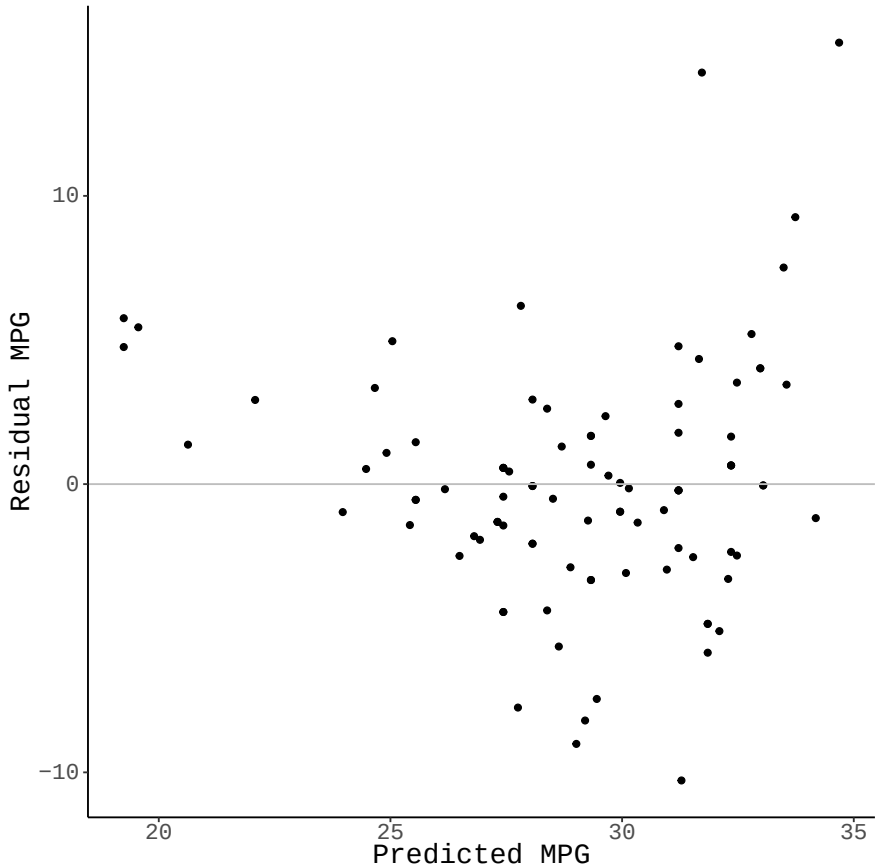
Treating Correlated Errors

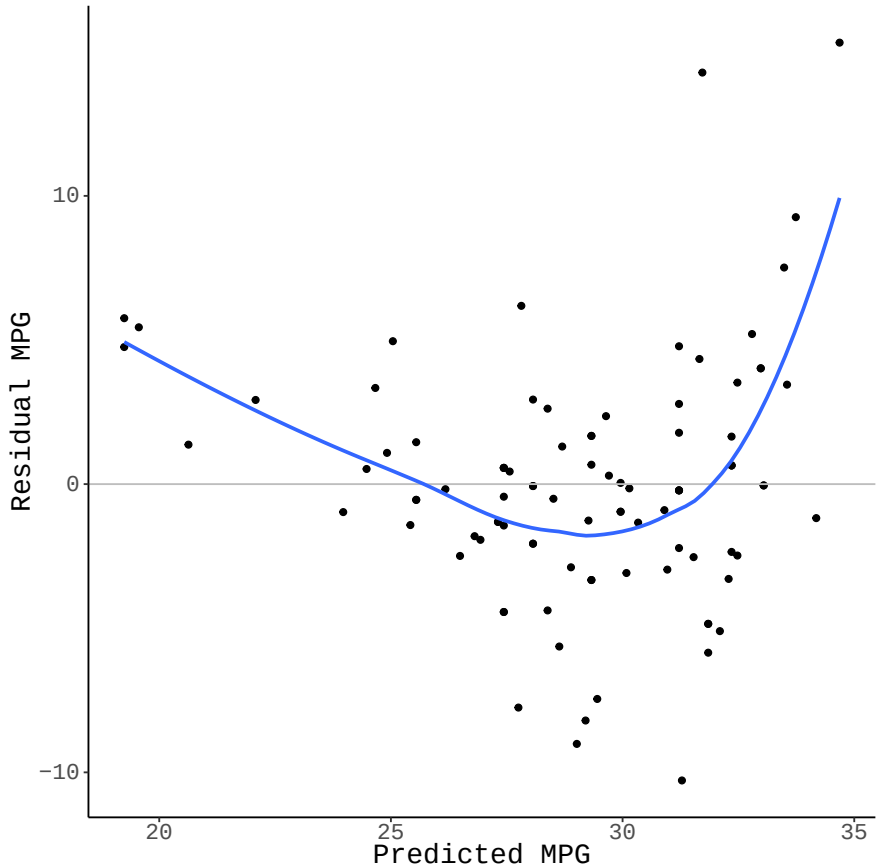
Clustering can be viewed as theoretically meaningful or as a nuisance factor that just needs to be controlled.

- If the clustering is meaningful, you should model the data using *multilevel modeling*.
  - Hierarchical linear regression
  - Mixed models
  - Random effects models
- If the clustering is an uninteresting nuisance, you can use specialized HC variance estimators that deal with clustering.

Model Specification

Our assumptions mostly focus on the errors, so incorrect model specification can lead to violations of many assumptions.

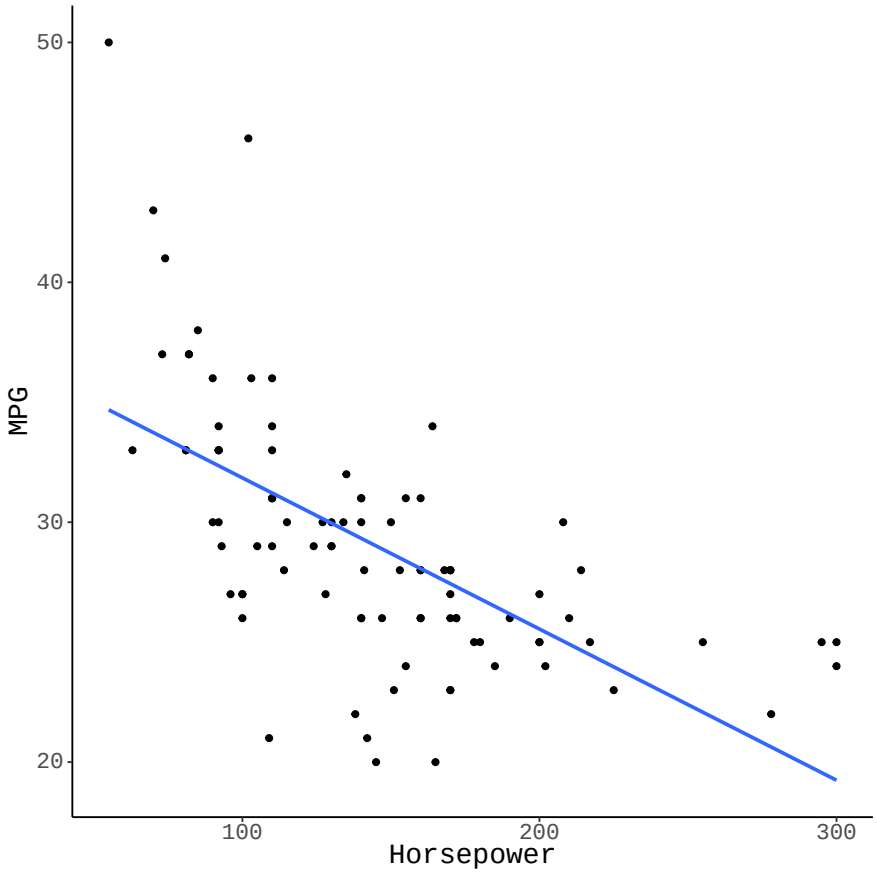


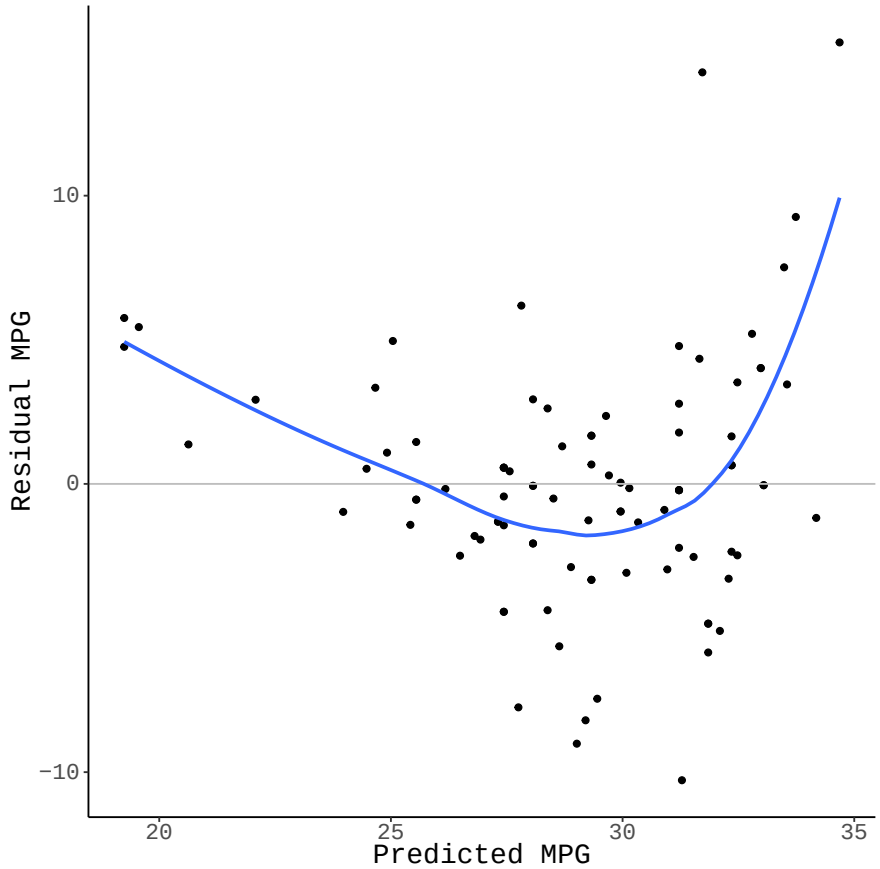


Nonlinear Trends in Residual Plots

Clearly, the linear trend fits these data poorly.

- We should probably add some polynomial terms





Treating Residual Nonlinearity

Nonlinearity in the residual plots is usually a sign of either:

1. Model misspecification
2. Influential observations

This type of model misspecification usually implies omitted functions of modeled variables.

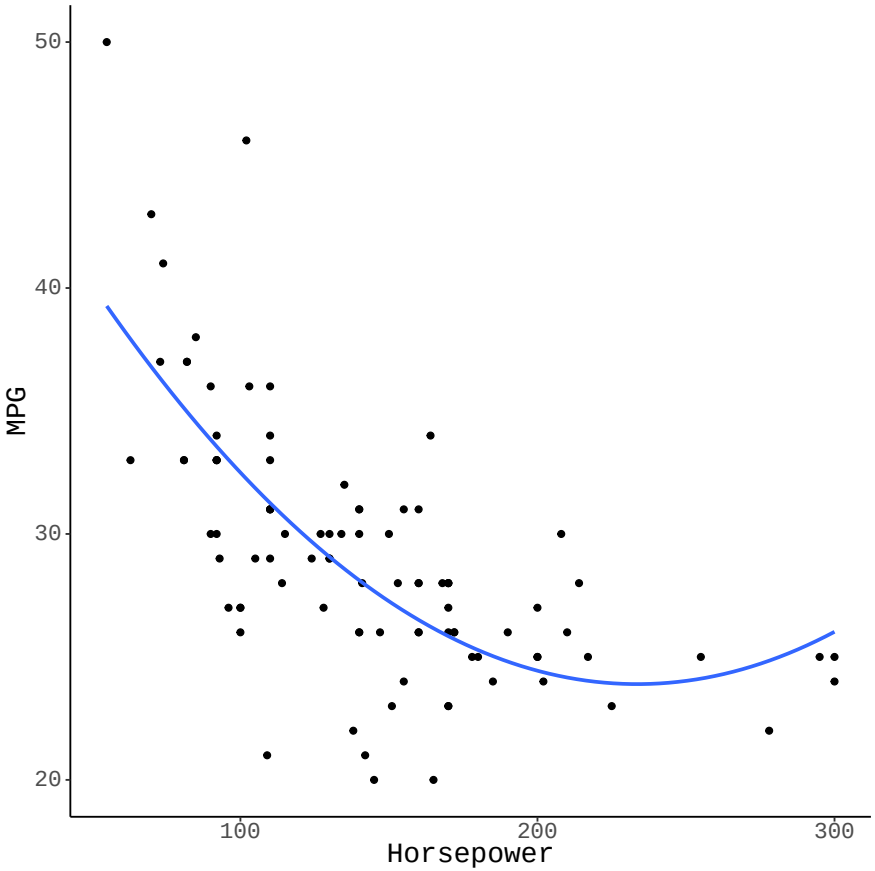
- Polynomial terms
- Interactions

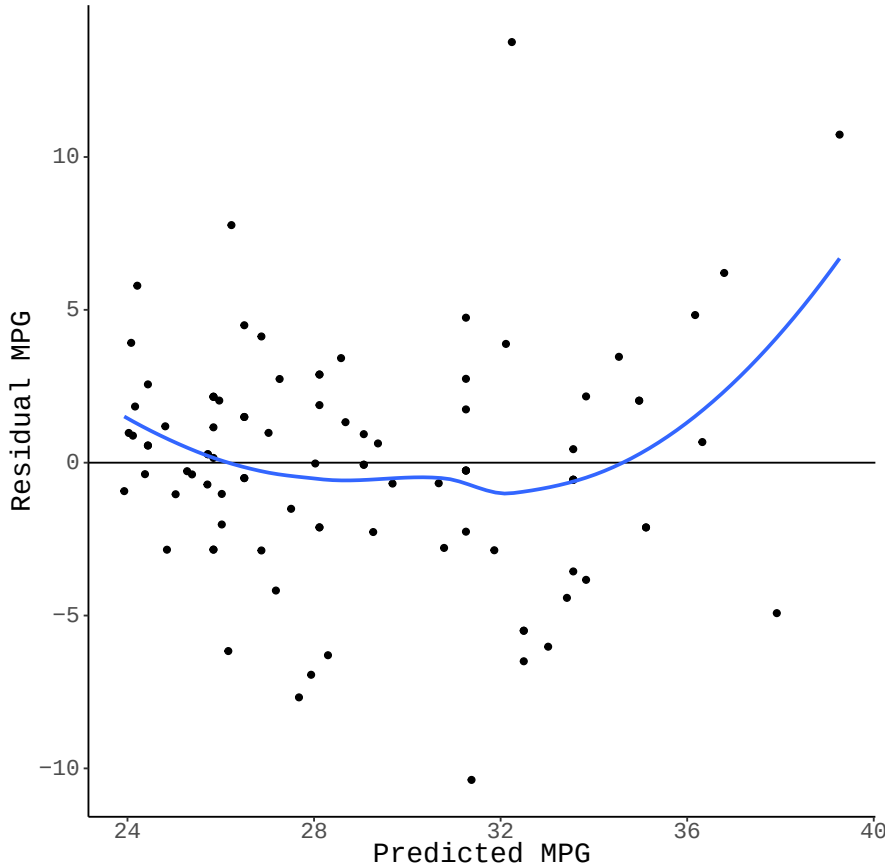
The solution is to include the omitted term into the model and refit.

- This is very much easier said than done.

Residual Plots

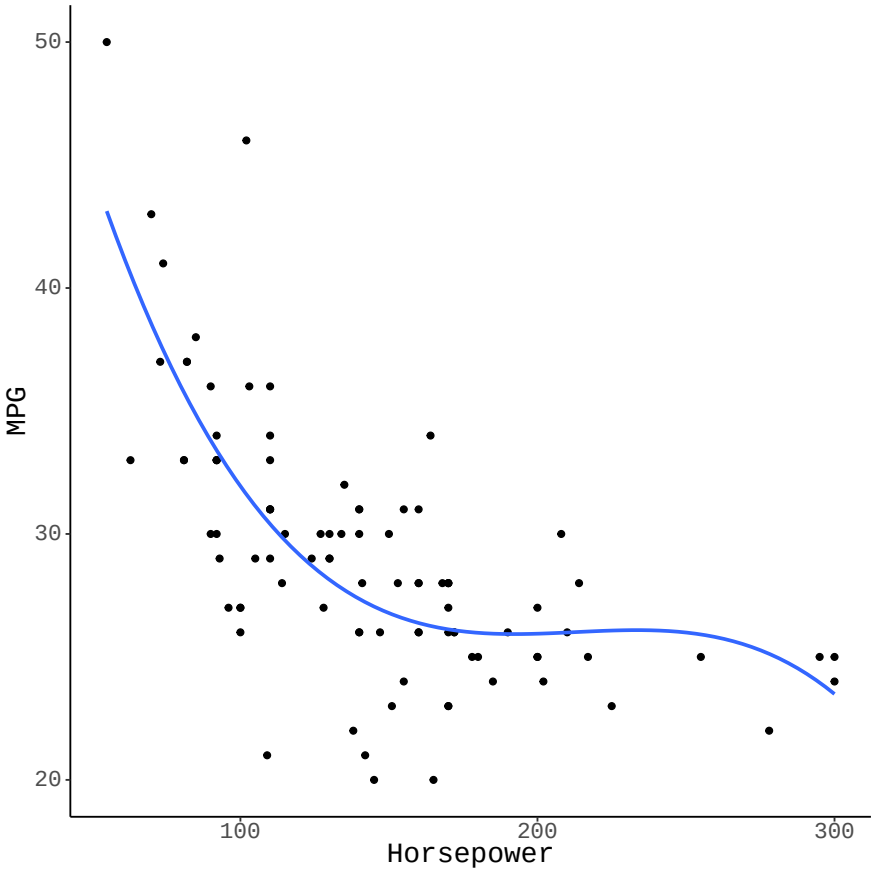
Certainly looks better, but not ideal.

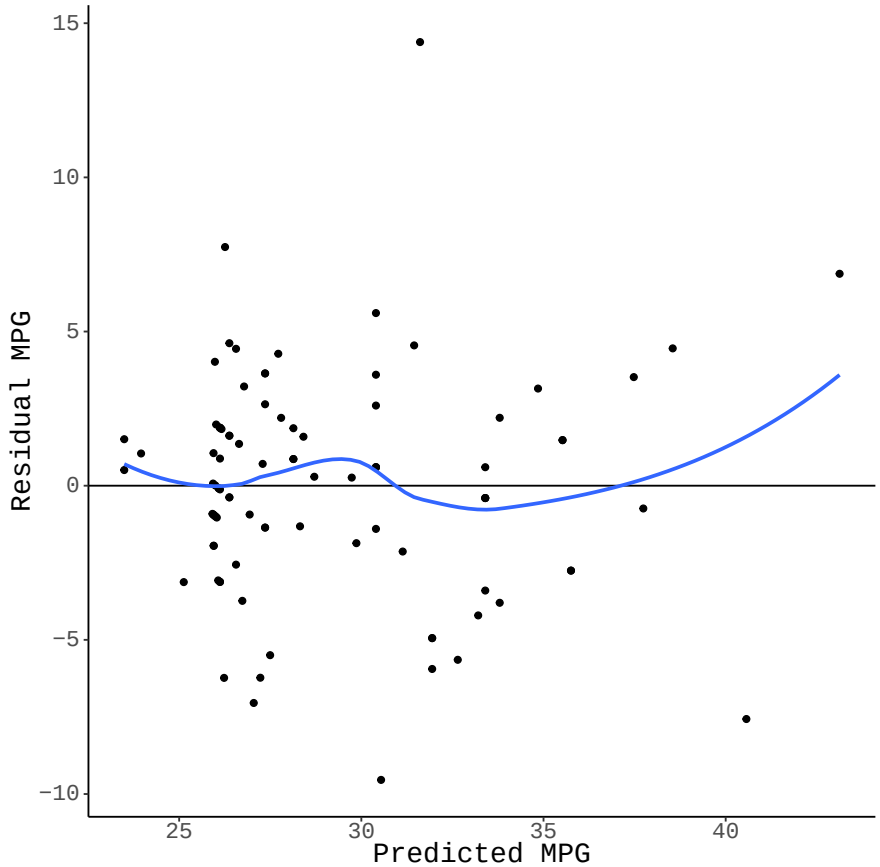




Residual Plots

Further improvement (perhaps).





Omitted Variables

The most common cause of endogeneity (i.e., violating Assumption 3) is *omitted variable bias*.

- If we leave an important predictor variable out of our equation, some modeled predictors will become endogenous and their estimated regression slopes will be biased.
- The omitted variable must be correlated with Y and at least one of the modeled X_p, to be a problem.

Omitted Variables

Assume the following is the true regression model.

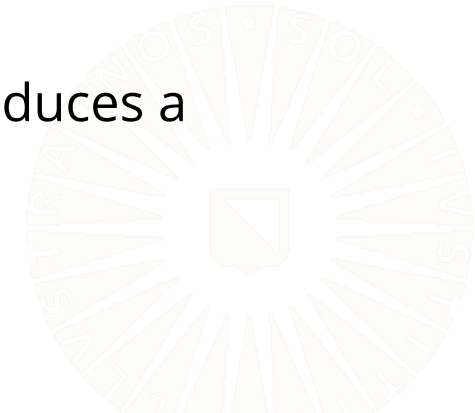
$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon$$

Now, suppose we omit Z from the model:

$$Y = \beta_0 + \beta_1 X + \omega$$
$$\omega = \varepsilon + \beta_2 Z$$

Our new error, ω, is a combination of the true error, ε, and the omitted term, $\beta_2 Z$.

- Consequently, if X and Z are correlated, omitting Z induces a correlation between X and ω (i.e., endogeneity).



Treating Omitted Variable Bias

Omitted variable bias can have severe consequences, but you can't really test for it.

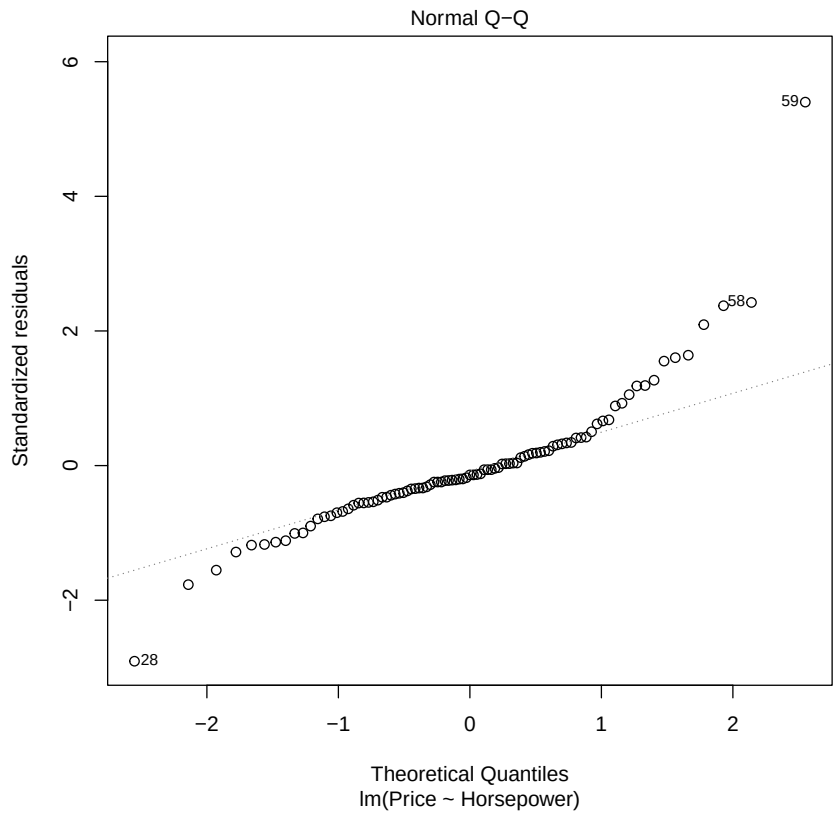
- The *errors* are correlated with the predictors, but our model is estimated under the assumption of exogeneity, so the *residuals* from our model will generally be uncorrelated with the predictors.
- We mostly have to pro-actively work to include all relevant variables in our model.

Normality Assumption

```
plot(out1, 2)
```

One of the best ways to evaluate the normality of the error distribution with a Q-Q Plot.

- Plot the quantiles of the residual distribution against the theoretically ideal quantiles.
- We can actually use a Q-Q Plot to compare any two distributions.



Consequences of Violating Normality

In small samples, with *fixed* predictors, normally distributed errors imply normal sampling distributions for the regression coefficients.

- In large samples, the central limit theorem implies normal sampling distributions for the coefficients, regardless of the error distribution.

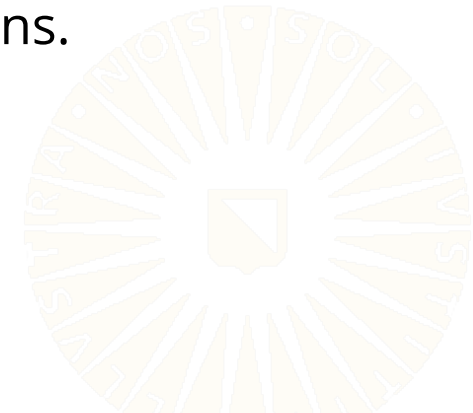
Consequences of Violating Normality

In small samples, with *fixed* predictors, normally distributed errors imply normal sampling distributions for the regression coefficients.

- In large samples, the central limit theorem implies normal sampling distributions for the coefficients, regardless of the error distribution.

Prediction intervals require normally distributed errors.

- Confidence intervals for predictions share the same normality requirements as the coefficients' sampling distributions.



Treating Violations of Normality

We usually don’t need to do anything about non-normal errors.

- The CLT will protect our inferences.

Treating Violations of Normality

We usually don't need to do anything about non-normal errors.

- The CLT will protect our inferences.

We can use *bootstrapping* to get around the need for normality.

1. Treat your sample as a synthetic population from which you draw many new samples (with replacement).
2. Estimate your model in each new sample.
3. The replicates of your estimated parameters generate an empirical sampling distribution that you can use for inference.

Treating Violations of Normality

We usually don't need to do anything about non-normal errors.

- The CLT will protect our inferences.

We can use *bootstrapping* to get around the need for normality.

1. Treat your sample as a synthetic population from which you draw many new samples (with replacement).
2. Estimate your model in each new sample.
3. The replicates of your estimated parameters generate an empirical sampling distribution that you can use for inference.

Bootstrapping can be used for inference on pretty much any estimable parameter, but it won't work with small samples.

- Need to assume that your sample is representative of the population

Influential Observations

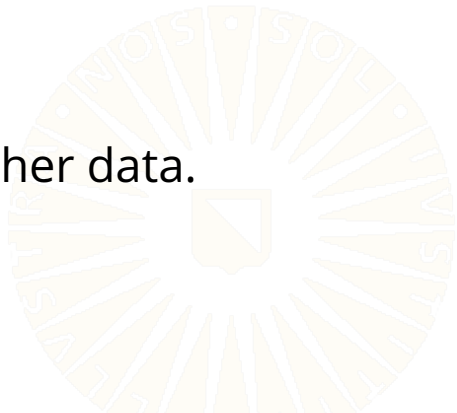
Influential Observations

Influential observations contaminate analyses in two ways:

1. Exert too much influence on the fitted regression model
2. Invalidate estimates/inferences by violating assumptions

There are two distinct types of influential observations:

1. Outliers
    - Observations with extreme outcome values, relative to the other data.
    - Observations with outcome values that fit the model very badly.
2. High-leverage observations
    - Observation with extreme predictor values, relative to other data.



Outliers

Outliers can be identified by scrutinizing the residuals.

- Observations with residuals of large magnitude may be outliers.
- The difficulty arises in quantifying what constitutes a “large” residual.

If the residuals do not have constant variance, then we cannot directly compare them.

- We need to standardize the residuals in some way.

Detecting Outliers

We are specifically interested in *externally studentized residuals*.

- We can't simply standardize the ordinary residuals.
  - *Internally studentized residuals*
  - Outliers can pull the regression line towards themselves.
  - The internally studentized residuals for outliers will be too small.

Begin by defining the concept of a *deleted residual*:

$$\hat{\varepsilon}_{(n)} = Y_n - \hat{Y}_{(n)}$$

- $\hat{\varepsilon}_{(n)}$ quantifies the distance of Y_n from the regression line estimated after excluding the nth observation.

Studentized Residuals

If we standardize the deleted residual, $\hat{\varepsilon}_{(n)}$, we get the externally studentized residual:

$$t_{(n)} = \frac{\hat{\varepsilon}_{(n)}}{SE_{\hat{\varepsilon}_{(n)}}}$$

The externally studentized residuals have two very useful properties:

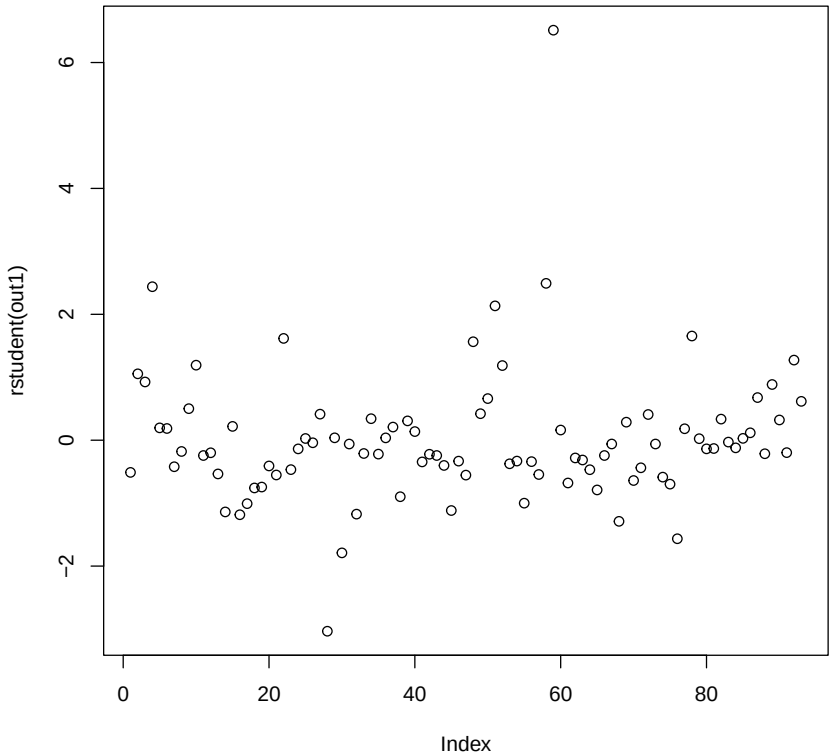
1. Each $t_{(n)}$ is scaled equivalently.
   - We can directly compare different $t_{(n)}$.
2. The $t_{(n)}$ are *Student's t* distributed.
   - We can quantify outliers in terms of quantiles of the t distribution.
   - $|t_{(n)}| > 3.0$ is a common rule of thumb for flagging outliers.

Studentized Residual Plots

```
plot(rstudent(out1))
```

Index plots of the externally studentized residuals can help spotlight potential outliers.

- Look for observations that clearly "stand out from the crowd."



High-Leverage Points

We identify high-leverage observations through their *leverage* values.

- An observation's leverage, h_n, quantifies the extent to which its predictors affect the fitted regression model.
- Observations with X values very far from the mean, $\bar{X}$, affect the fitted model disproportionately.

High-Leverage Points

We identify high-leverage observations through their *leverage* values.

- An observation's leverage, h_n, quantifies the extent to which its predictors affect the fitted regression model.
- Observations with X values very far from the mean, $\bar{X}$, affect the fitted model disproportionately.

In simple linear regression, the nth leverage is given by:

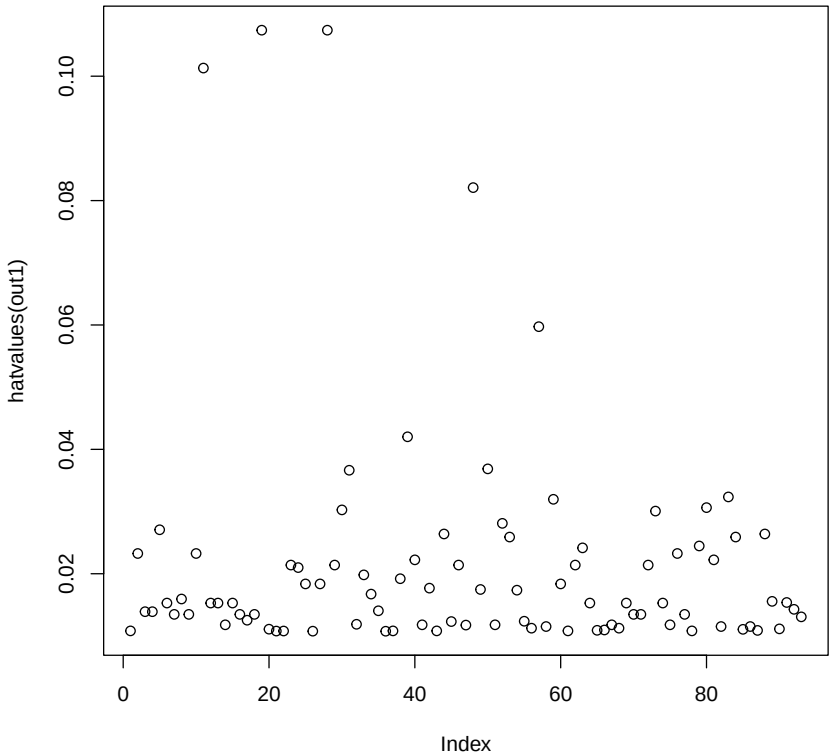
$$h_n = \frac{1}{N} + \frac{\left(X_n - \bar{X}\right)^2}{\sum_{m=1}^{N}\left(X_m - \bar{X}\right)^2}$$

Leverage Plots

```
plot(hatvalues(out1))
```

Index plots of the leverage values can help spotlight high-leverage points.

- Again, look for observations that clearly "stand out from the crowd."



Outliers & Leverages → Influential Points

Observations with high leverage or large (externally) studentized residuals are not necessarily influential.

- High-leverage observations tend to be more influential than outliers.
- The worst problems arise from observations that are both outliers and have high leverage.

Measures of influence simultaneously consider extremity in both X and Y dimensions.

- Observations with high measures of influence are very likely to cause problems.

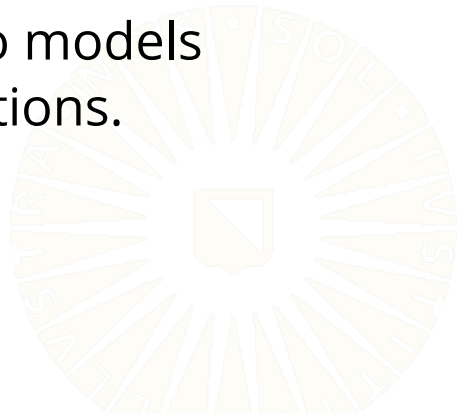
Measures of Influence

Measures of influence come in two flavors.

1. Global measures of influence
    - Cook's Distance
2. Coefficient-specific measures of influence
    - DFBETAS

All measures of influence use the same logic as the deleted residual.

- Compare models estimated from the whole sample to models estimated from samples excluding individual observations.



Global Measures of Influence

Each observation gets a Cook's Distance value.

$$\text{Cook's } D_n = \frac{\sum_{n=1}^{N} \left(\hat{Y}_n - \hat{Y}_{(n)} \right)^2}{(P+1)\,\hat{\sigma}^2}$$

$$= (P+1)^{-1} t_n^2 \frac{h_n}{1-h_n}$$

Each regression coefficient (including the intercept) gets a DFBETAS value for each observation.

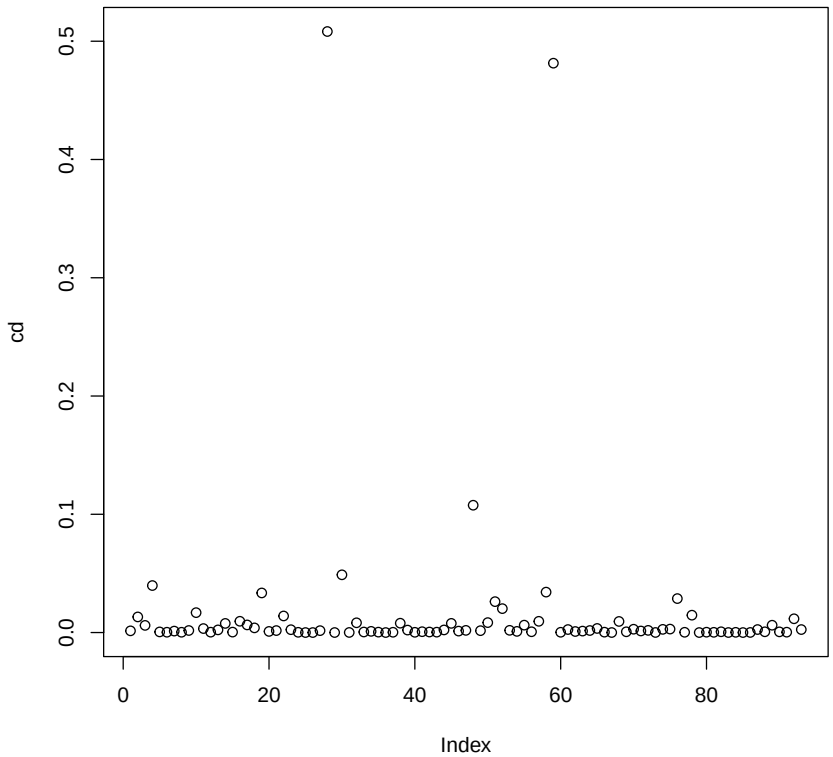
$$\text{DFBETAS}_{np} = \frac{\hat{\beta}_p - \hat{\beta}_{p(n)}}{\text{SE}_{\hat{\beta}_{p(n)}}}$$

Plots of Cook's Distance

```
cd <- cooks.distance(out1)
plot(cd)
```

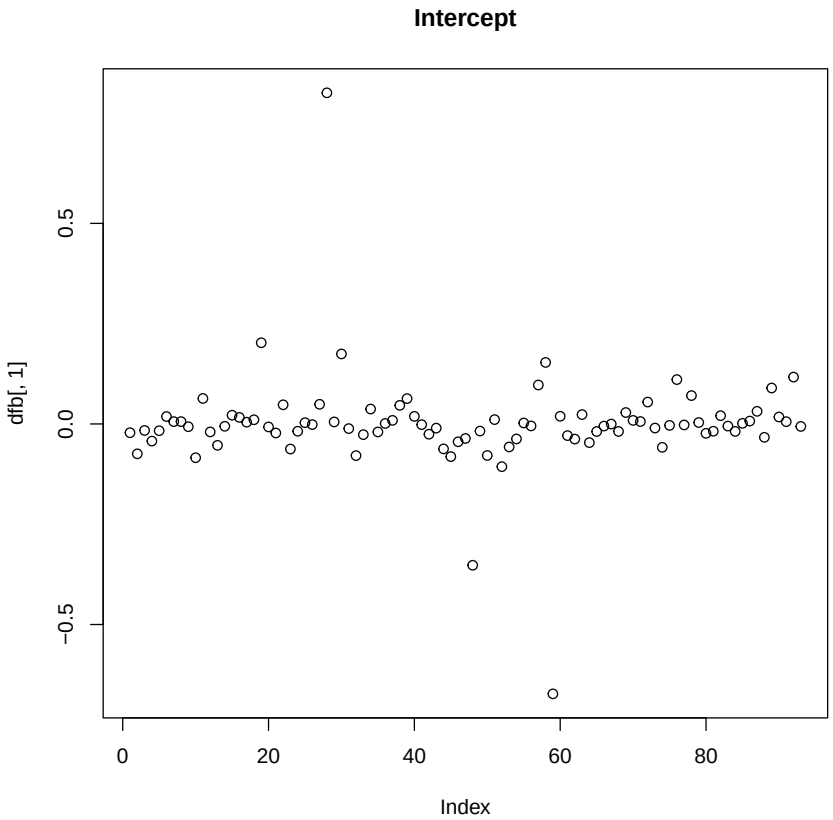
Index plots of Cook's distances can help spotlight the influential points.

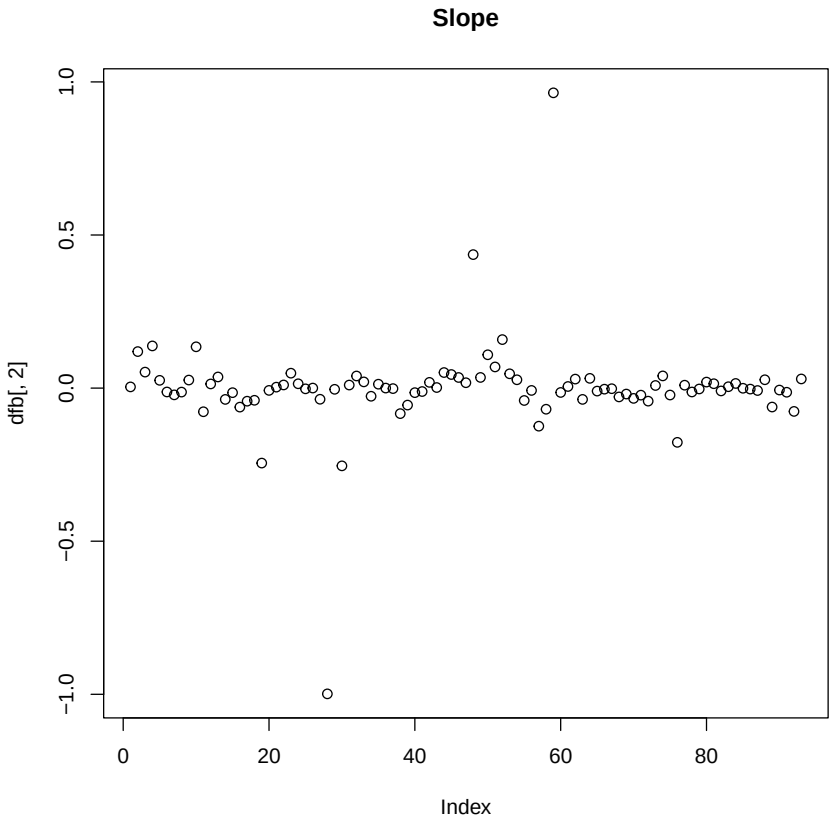
- Look for observations that clearly "stand out from the crowd."



Plots of DFBETAS

```
dfb <- dfbetas(out1)
plot(dfb[ , 1], main = "Intercept")
plot(dfb[ , 2], main = "Slope")
```



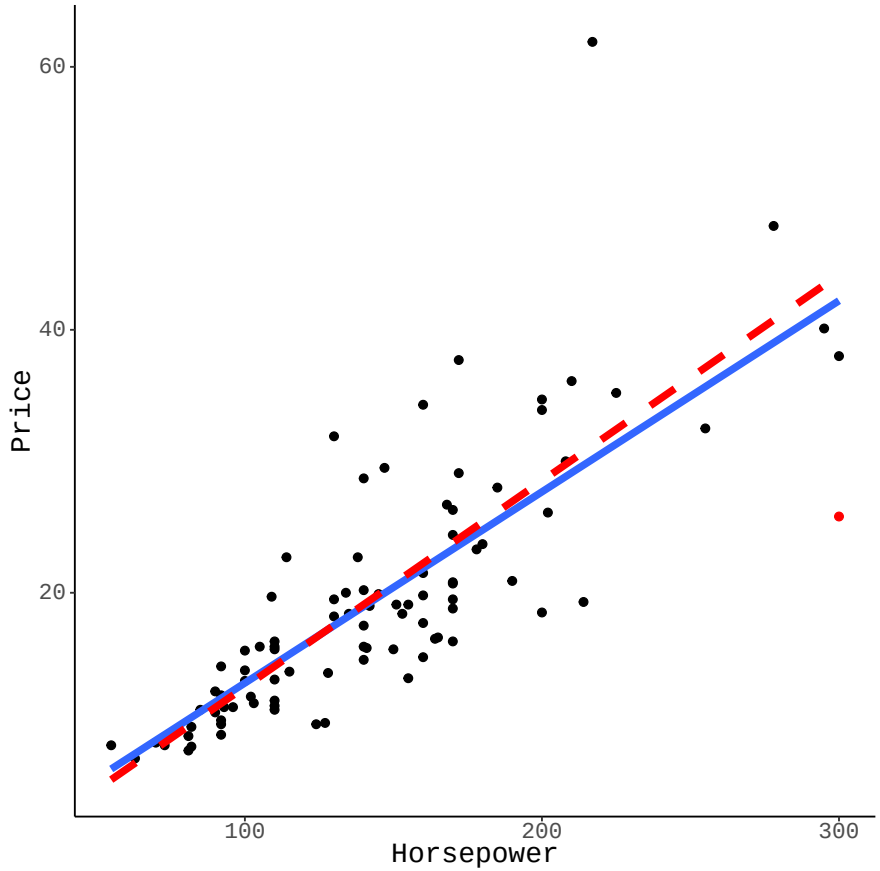


Removing Influential Observations

```
(maxD <- which.max(cd))

28
28
```

Observation number 28 was the most influential according to Cook's Distance.

- Removing that observation has a small impact on the fitted regression line.
- Influential observations don't only affect the regression line, though.



Removing Influential Observations

```
## Exclude the influential case:
Cars93.2 <- Cars93[-maxD, ]

## Fit model with reduced sample:
out2 <- lm(Price ~ Horsepower, data = Cars93.2)

round(summary(out1)$coefficients, 6)

             Estimate Std. Error   t value Pr(>|t|)
(Intercept) -1.398769   1.820016 -0.768548 0.444152
Horsepower   0.145371   0.011898 12.218325 0.000000

round(summary(out2)$coefficients, 6)

             Estimate Std. Error   t value Pr(>|t|)
(Intercept) -2.837646   1.806418 -1.570868 0.119722
Horsepower   0.156750   0.011996 13.066942 0.000000
```

Removing Influential Observations

```
partSummary(out1, 2)

Residuals:
    Min      1Q  Median      3Q     Max
-16.413  -2.792  -0.821   1.803  31.753

partSummary(out2, 2)

Residuals:
     Min       1Q   Median       3Q      Max
-11.4069  -3.0349  -0.5912   1.8530  30.7229
```

Removing Influential Observations

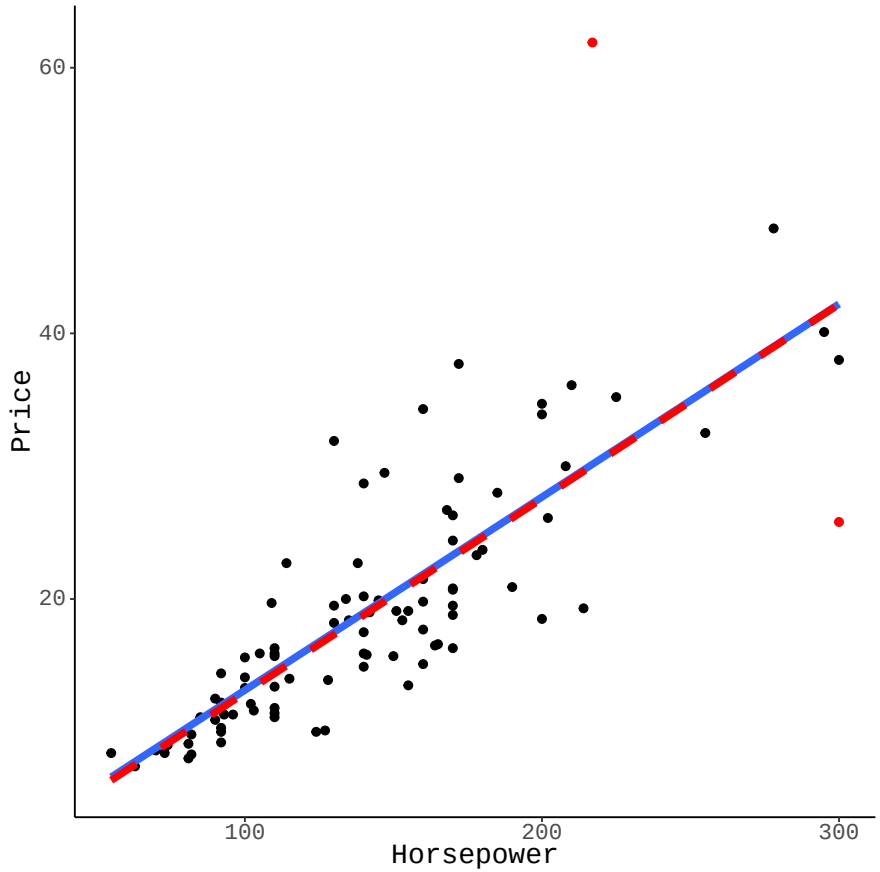
```
summary(out1)[c("sigma", "r.squared", "fstatistic")] %>%
    unlist() %>%
    head(3)

           sigma        r.squared fstatistic.value
        5.976953         0.621287       149.287468

summary(out2)[c("sigma", "r.squared", "fstatistic")] %>%
    unlist() %>%
    head(3)

           sigma        r.squared fstatistic.value
       5.7243112        0.6548351      170.7449721
```

Removing Influential Observations

```
(maxDs <- sort(cd) %>% names() %>% tail(2) %>% as.numeric())

[1] 59 28
```

If we remove the two most influential observations, 59 and 28, the fitted regression line barely changes at all.

- The influences of these two observations were counteracting one another.
- We're probably still better off, though.



Removing Influential Observations

```
## Exclude influential cases:
Cars93.2 <- Cars93[-maxDs, ]

## Fit model with reduced sample:
out2.2 <- lm(Price ~ Horsepower, data = Cars93.2)

round(summary(out1)$coefficients, 6)

             Estimate Std. Error   t value Pr(>|t|)
(Intercept) -1.398769   1.820016 -0.768548 0.444152
Horsepower   0.145371   0.011898 12.218325 0.000000

round(summary(out2.2)$coefficients, 6)

             Estimate Std. Error   t value Pr(>|t|)
(Intercept) -1.695315   1.494767 -1.134166  0.25977
Horsepower   0.146277   0.009986 14.648807  0.00000
```

Removing Influential Observations

```
partSummary(out1, 2)

Residuals:
    Min      1Q  Median      3Q     Max
-16.413  -2.792  -0.821   1.803  31.753

partSummary(out2.2, 2)

Residuals:
     Min       1Q   Median       3Q      Max
-10.3079  -2.5786  -0.6084   1.9775  14.5793
```

Removing Influential Observations

```
summary(out1)[c("sigma", "r.squared", "fstatistic")] %>%
    unlist() %>%
    head(3)

           sigma        r.squared fstatistic.value
        5.976953         0.621287       149.287468

summary(out2.2)[c("sigma", "r.squared", "fstatistic")] %>%
    unlist() %>%
    head(3)

           sigma        r.squared fstatistic.value
       4.7053314        0.7068391      214.5875491
```

Treating Influential Points

The most common way to address influential observations is simply to delete them and refit the model.

- This approach is often effective—and always simple—but it is not fool-proof.
- Although an observation is influential, we may not be able to justify excluding it from the analysis.

Robust regression procedures can estimate the model directly in the presence of influential observations.

- Observations in the tails of the distribution are weighted less in the estimation process, so outliers and high-leverage points cannot exert substantial influence on the fit.

References

Hayes, A. F. (2017). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. New York: Guilford Press.