Assumptions & Diagnostics

Statistics & Methodology Lecture 9

Kyle M. Lang

Department of Methodology & Statistics
Tilburg University

Outline

1. Assumptions of MLR
2. Regression diagnostics
3. Influence measures

Intuition of Assumptions

Do you trust your senses?

- How do you know that you are really observing this lecture?
- How do you know that anything you see, hear, touch, taste, or smell actually exists?
- How do you know that you actually exist in the form that you perceive for yourself?

Intuition of Assumptions

You cannot *know* any of the things I just asked.

- You cannot *prove* any of those conditions to be veridical without *assuming*, at least, one key pre-condition.
- You must assume that your senses are trustworthy to meaningfully interact with the world.

Algebraic Example

Consider the following equation:

$$5 = x + y$$

What are the values of x and y?

Algebraic Example

Consider the following equation:

$$5 = x + y$$

What are the values of x and y?

$$y = 5 - x$$

Algebraic Example

Consider the following equation:

$$5 = x + y$$

What are the values of x and y?

$$y = 5 - x$$

What if we assume that $y = x$?

Algebraic Example

Consider the following equation:

$$5 = x + y$$

What are the values of x and y?

$$y = 5 - x$$

What if we assume that $y = x$?

$$5 = x + y$$
$$0 = x - y$$

Algebraic Example

Consider the following equation:

$$5 = x + y$$

What are the values of x and y?

$$y = 5 - x$$

What if we assume that $y = x$?

$$5 = x + y$$
$$0 = x - y$$

Now we have enough information:

$$5 = x + x = 2x \quad \Rightarrow \quad x = y = 2.5$$

Assumptions in Statistics

Why do we need *assumptions* in statistics?

- Data, by themselves, do not offer enough information to support statistical analysis.
- We need to assume some properties of the population model that generated the data.
- We also use assumptions to simplify problems.

Gauss-Markov Theorem: Setup

Consider linear regression models of the form:

$$Y = \beta_0 + \sum_{p=1}^{P} \beta_p X_p + \varepsilon$$

- Assume **X** is fixed over repeated sampling.
- Assume the following properties of the errors:
  1. The errors have a mean of zero:
     - $E[\varepsilon_n] = 0$
  2. The errors have constant, finite variance.
     - $\text{Var}(\varepsilon_n) = \sigma^2 < \infty$
  3. The errors are uncorrelated.
     - $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0, \ i \neq j$

Gauss-Markov Theorem

Given the preceding conditions, the Gauss-Markov Theorem states that the OLS estimates—when they exist—are the *Best Linear Unbiased Estimators* of β.

- OLS estimates are "best" in the sense that they will have the lowest variance (i.e., smallest SEs) of any unbiased estimator of β.

Gauss-Markov Theorem

Given the preceding conditions, the Gauss-Markov Theorem states that the OLS estimates—when they exist—are the *Best Linear Unbiased Estimators* of β.

- OLS estimates are "best" in the sense that they will have the lowest variance (i.e., smallest SEs) of any unbiased estimator of β.

Note that the Gauss-Markov Theorem does not require normally distributed errors.

- We assume normally distributed errors to facilitate inference in finite samples.
- If we assume normally distributed errors, the OLS estimate of β is also the *maximum likelihood* estimate.

Assumptions of MLR

The typical assumptions for linear regression extend the Gauss-Markov assumptions by:

- Removing the requirement for fixed $\mathbf{X}$
- Adding the assumption of normally distributed errors

So, our final assumptions 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, $\mathbf{X}$, 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.
   - $\text{Cov}(\mu_{Y|X}, \varepsilon) = 0$
   - $\text{E}(\varepsilon_n | X_n) = 0$
4. The errors have constant, finite variance.
   - $\text{Var}(\varepsilon_n | \mathbf{X}) = \sigma^2 < \infty$
5. The errors are uncorrelated.
   - $\text{Cov}(\varepsilon_i, \varepsilon_j | \mathbf{X}) = 0, \;\; i \neq j$
6. The errors are normally distributed.
   - $\varepsilon | \mathbf{X} \sim \text{N}(0, \sigma^2)$

Assumptions of MLR

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

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

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

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

Keeping the assumptions stated in finer levels of detail, however, is helpful for diagnosing violations.

- We will work with the fine-grained definition of six assumptions.

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.

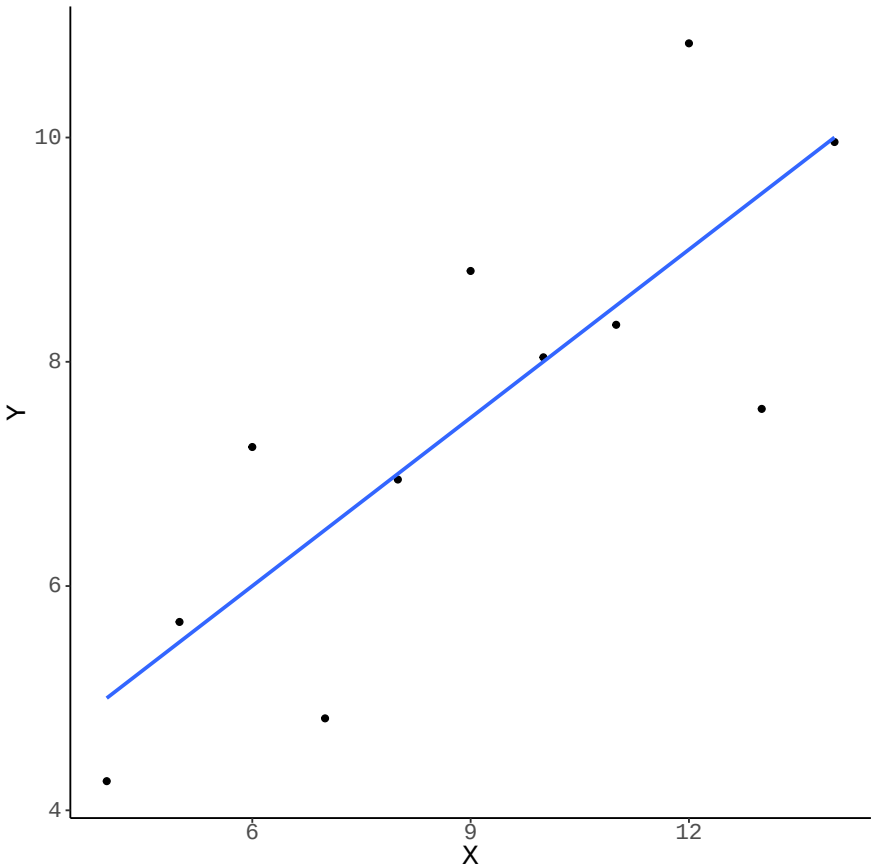
- We need to check the tenability of our assumptions before leaning too heavily on the model estimates.

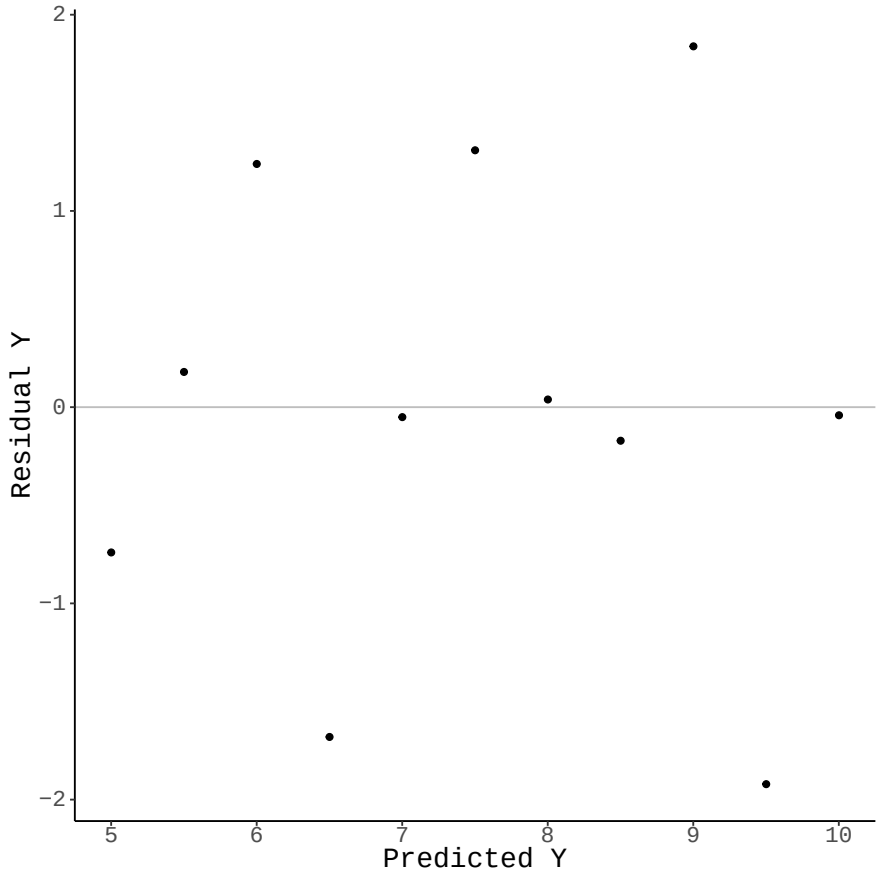
These checks are called *regression diagnostics*.

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

Residual Plots

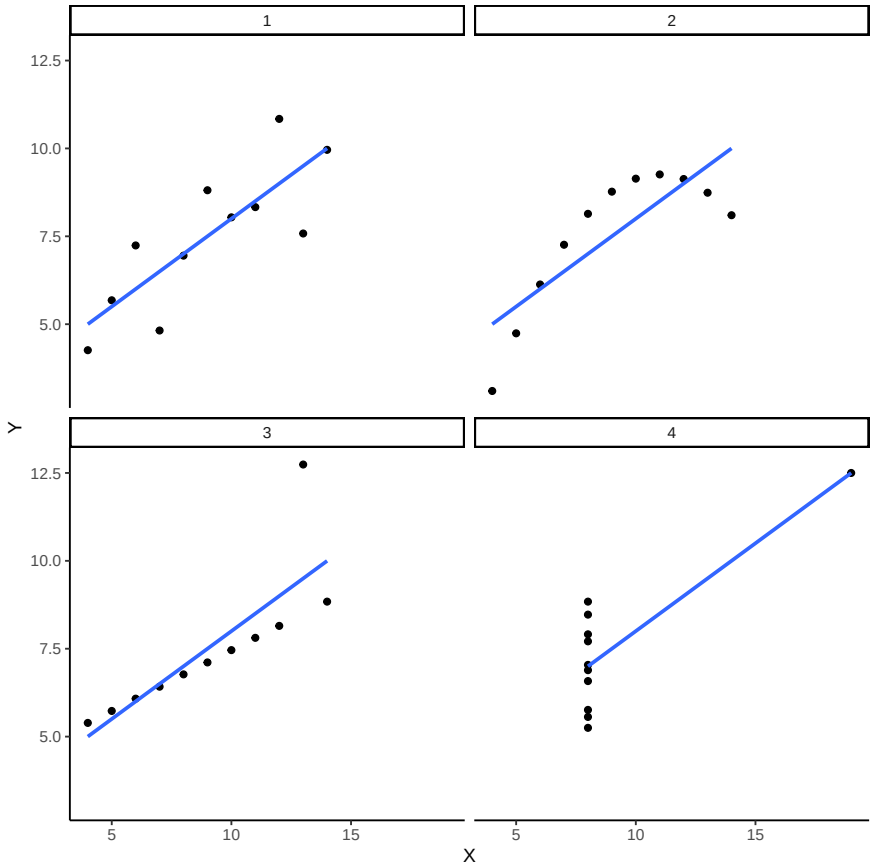
One of the most useful diagnostic graphics is the plot of residuals vs. predicted values.





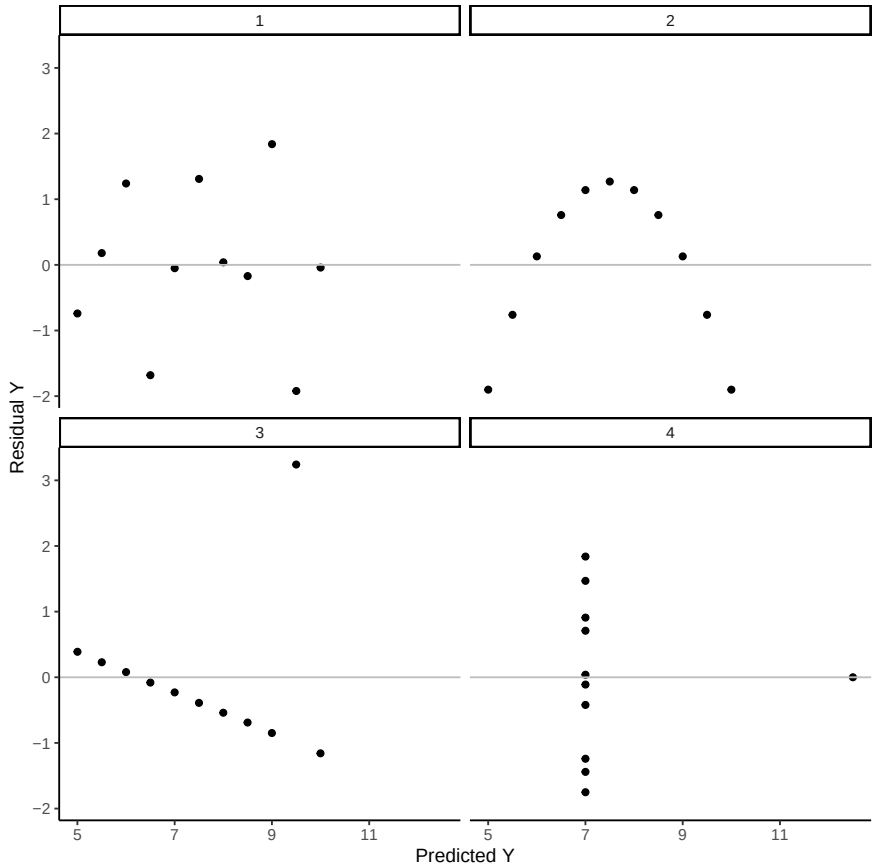
Residual Plots

Recall the Anscombe data that we saw when discussing EDA:



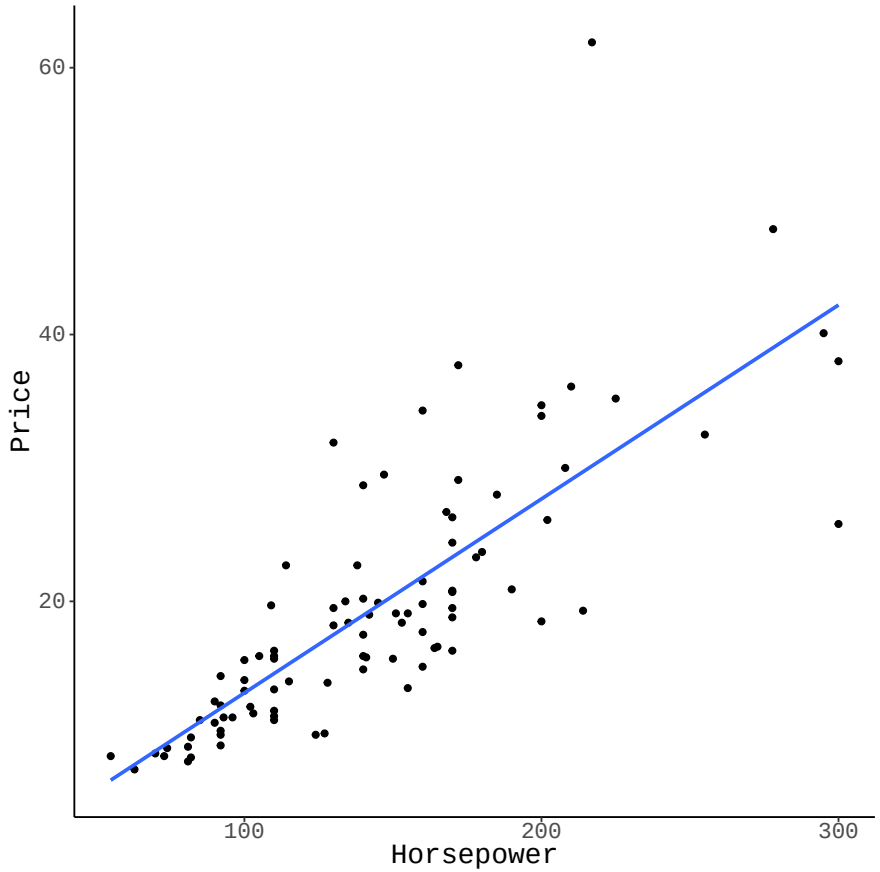
Residual Plots

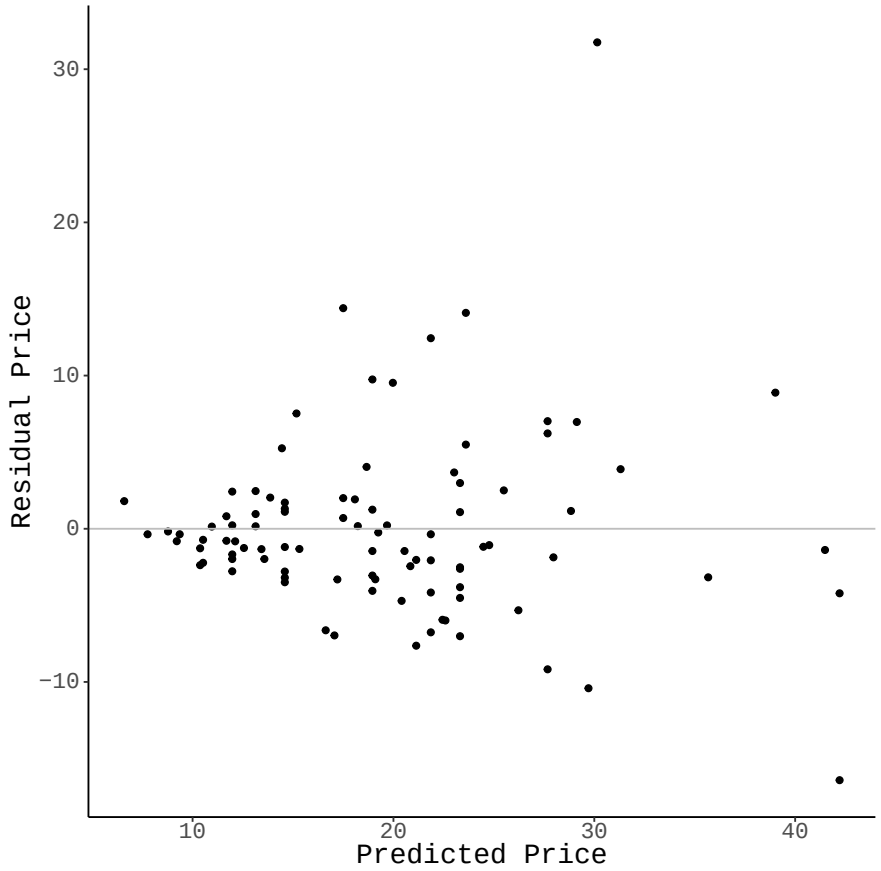
The residual plots clearly highlight the problems:



Heteroscedasticity

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





Testing for Heteroscedasticity

Residual plots—like the one shown on the last slide—are probably the best way to detect potential heteroscedasticity, but we can also do statistical tests.

- When heteroscedasticity is present, the variance of the residuals will vary along the regression line.
- If the pattern of this change is approximately monotonic, we can detect it by regression an estimate of the pointwise residual variance onto the predictors.
  - This is the intuition for the *Breusch-Pagan Test*.

$$\hat{\varepsilon}^2 = \gamma_0 + \sum_{p=1}^{P} \gamma_p X_p + \omega$$

Example

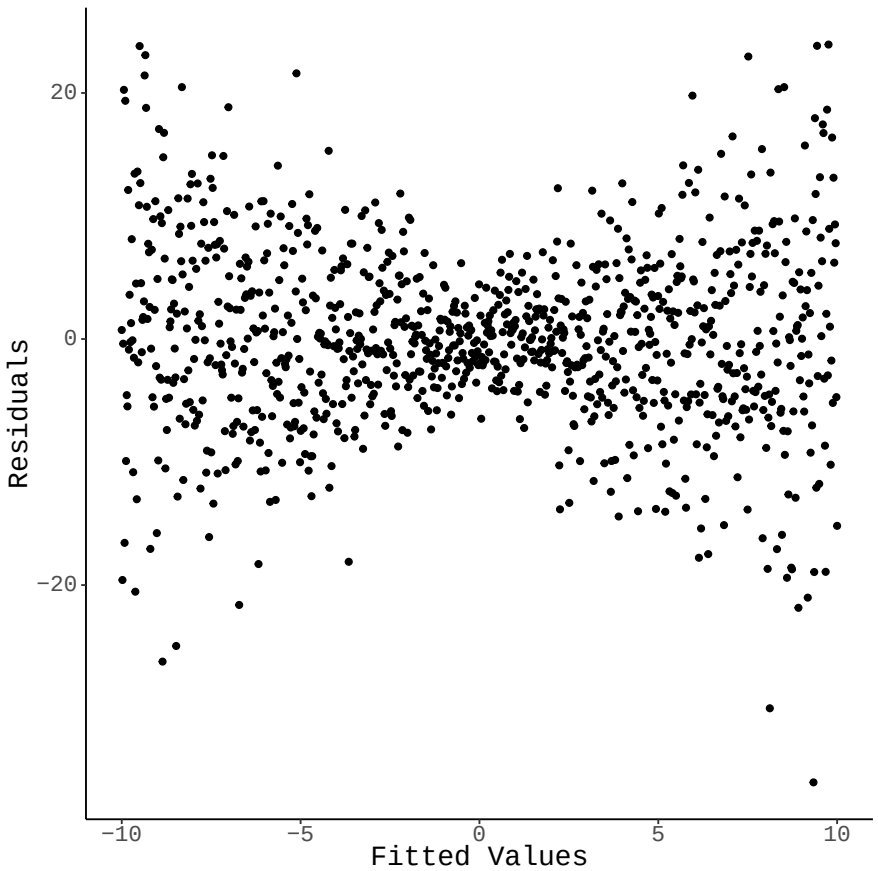
Let's apply the Breusch-Pagan Test to the regression we plotted above.

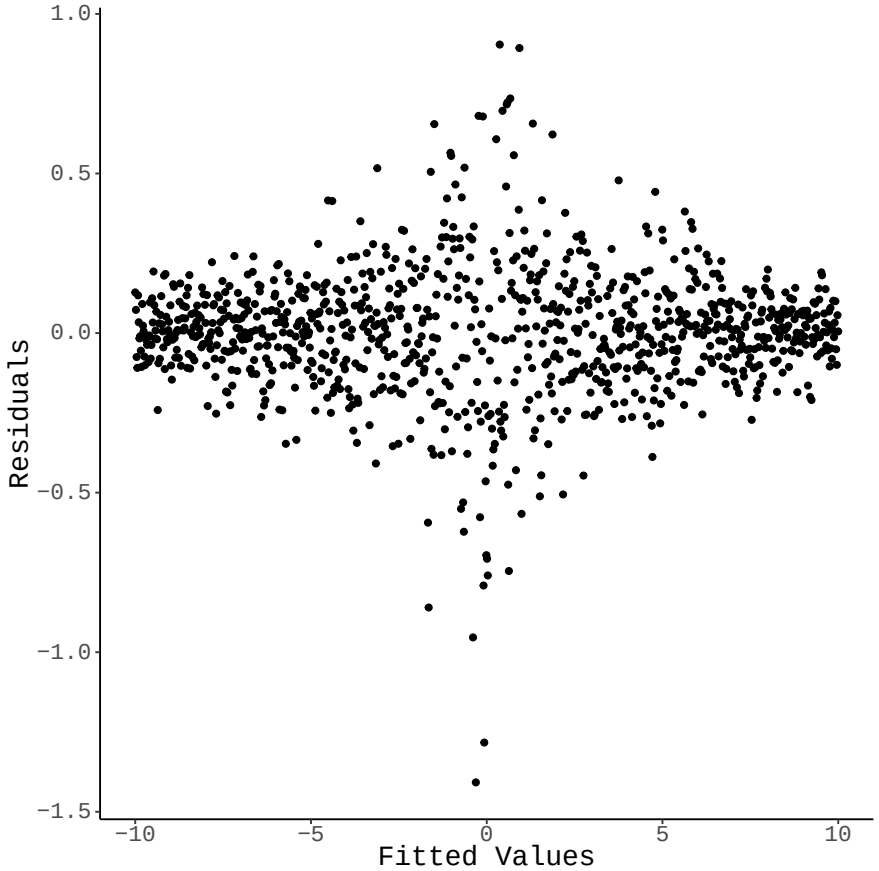
```
suppressMessages(library(lmtest))
data(Cars93)

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

##
##  studentized Breusch-Pagan test
##
## data:  out1
## BP = 7.9292, df = 1, p-value = 0.004864
```

We have a significant test statistic, so we reject the null hypothesis of homoscedasticity.

Limitations of the Breusch-Pagan Test

The Breusch-Pagan test can have trouble detecting non-monotonic heteroscedasticity.





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.

Example

```
## The 'sandwich' package provides several HC estimators:
library(sandwich)

## Use sandwich estimator to compute ACOV matrix:
hcCov <- vcovHC(out1)

## Test coefficients with robust SEs:
robTest <- coeftest(out1, vcov = hcCov)

## Test coefficients with default SEs:
defTest <- summary(out1)$coefficients
```

Example

```
## Compare robust and default approaches:
robTest

##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -1.398769   2.078200 -0.6731    0.5026
## Horsepower   0.145371   0.017164  8.4696 4.051e-13

defTest

##               Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) -1.3987691  1.8200164 -0.7685475 4.441519e-01
## Horsepower   0.1453712  0.0118978 12.2183251 6.837464e-21
```

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.

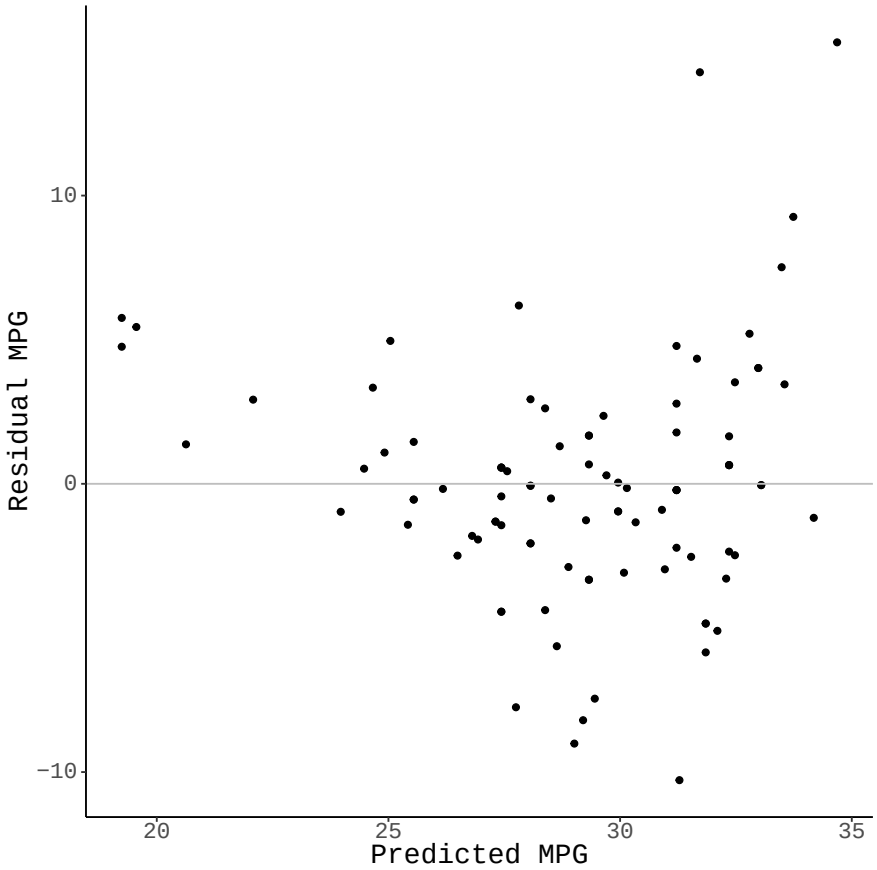
Treating Correlated Errors

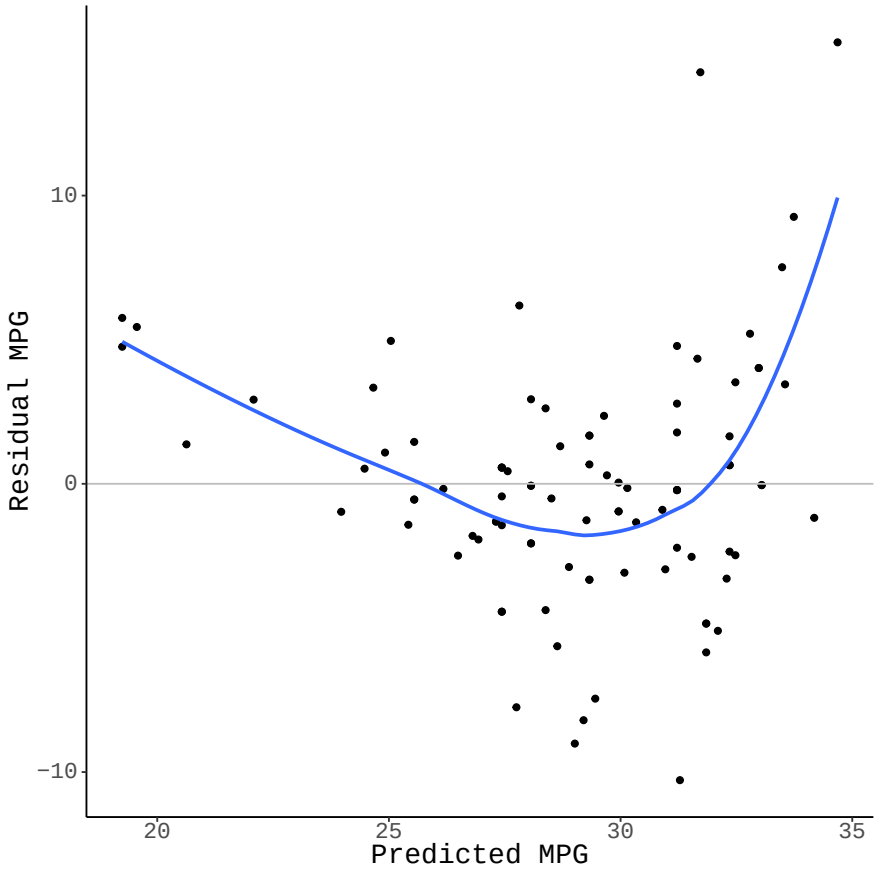
Clustering can be viewed as theoretically meaningful or as a nuisance factor that just needs to be controlled.

- 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.

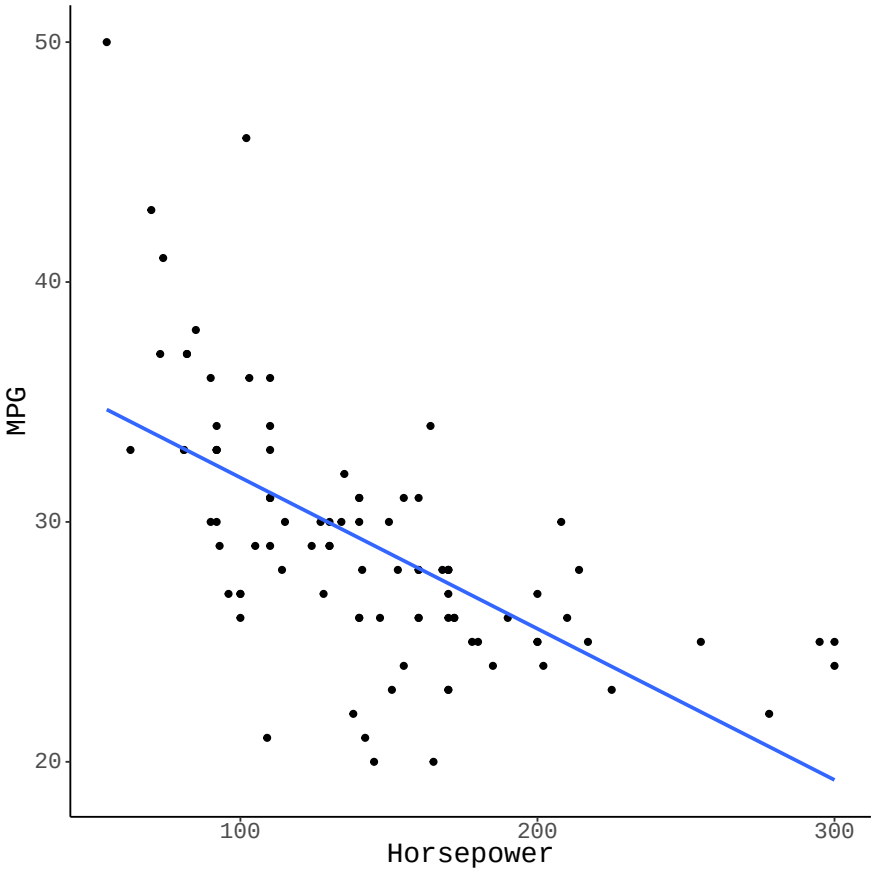


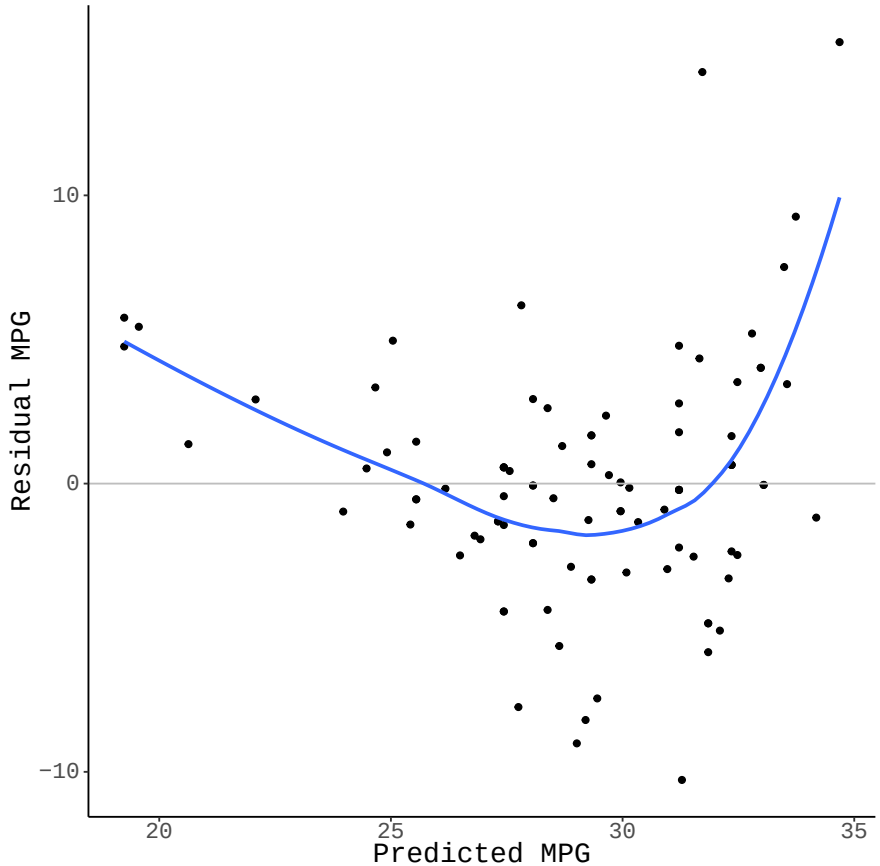


Nonlinear Trends in Residual Plots

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.

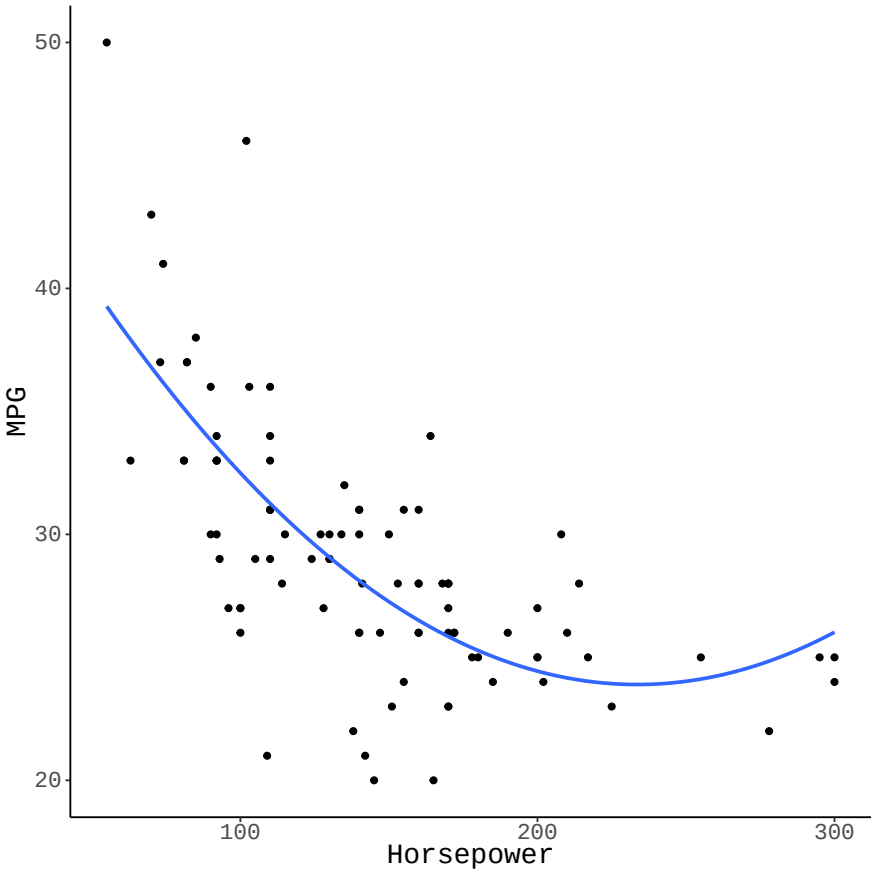
- Polynomial terms
- Interactions

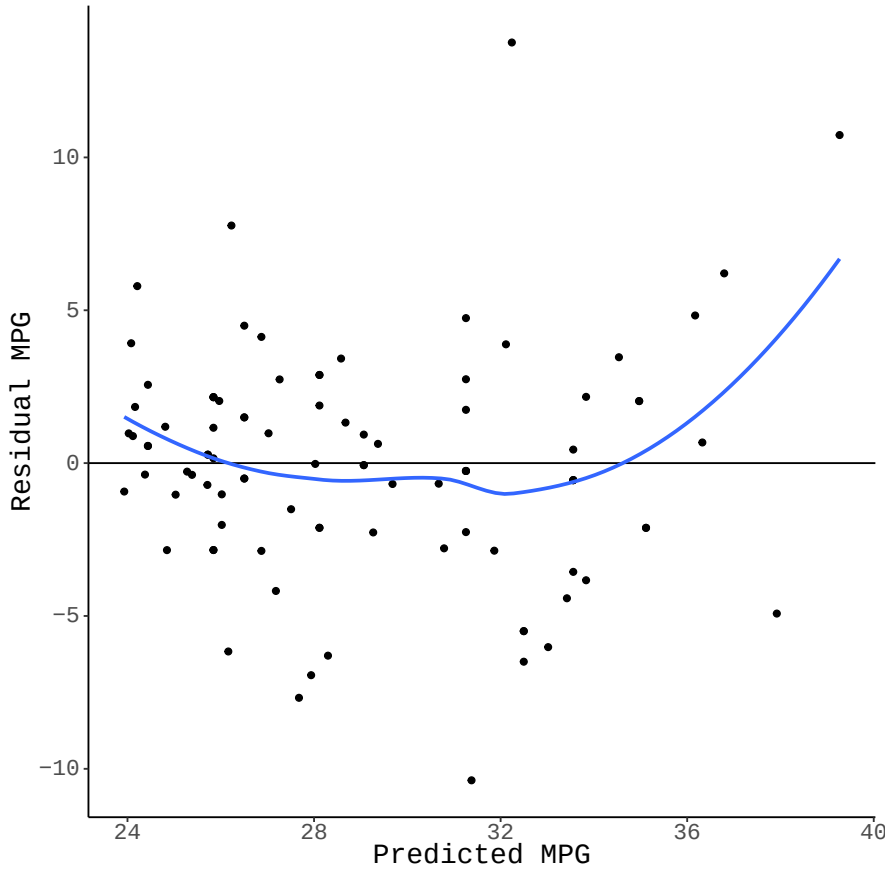
The solution is to include the omitted term into the model and refit.

- This is very much easier said than done.

Residual Plots

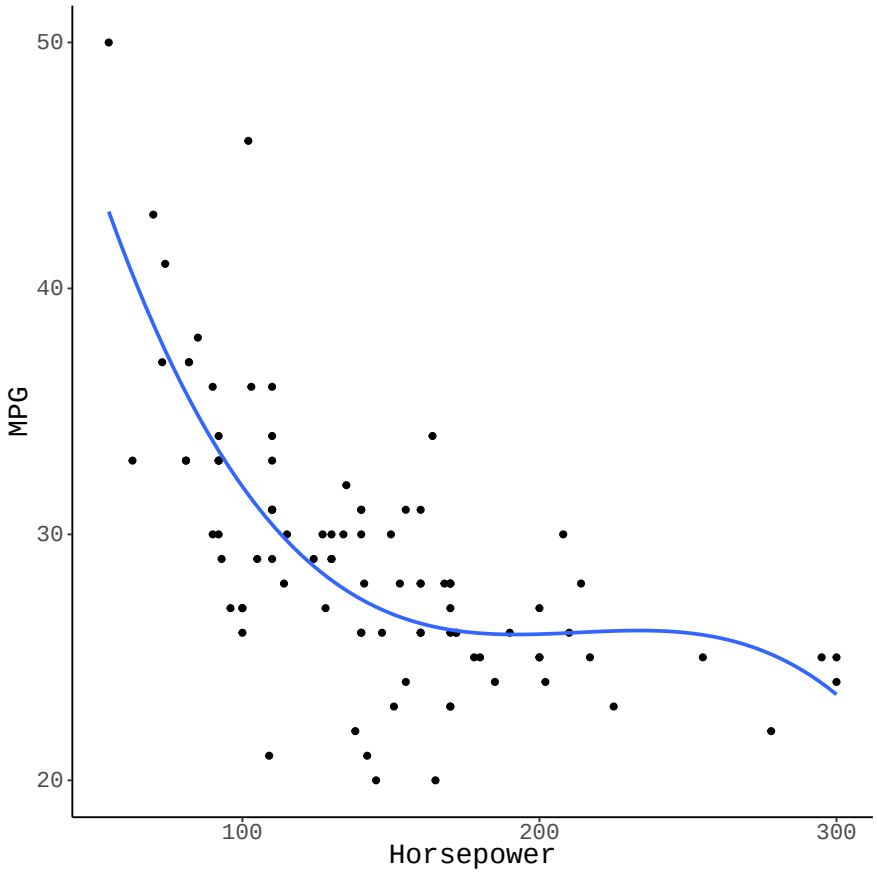
Certainly looks better, but not ideal.

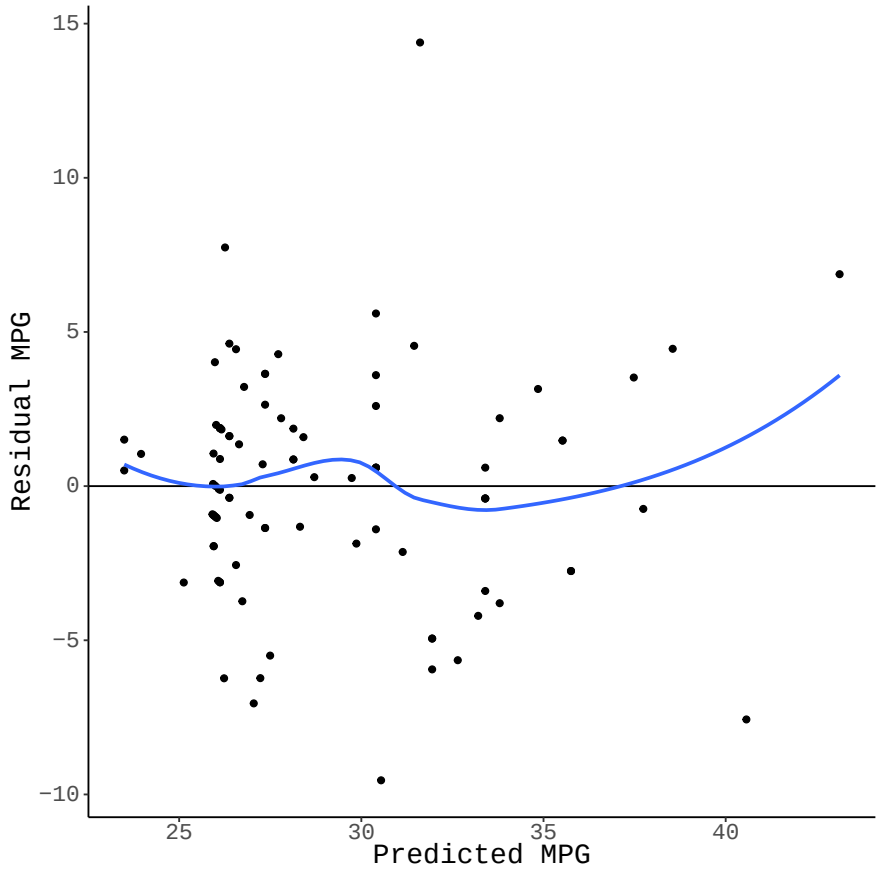




Residual Plots

Further improvement (perhaps).





Model Specification Tests

As with heteroscedasticity, residual plots are probably the best way to assess model misspecification, but we can do statistical tests here, too.

- One of the most popular specification tests is the Ramsey *Regression Equation Specification Error Test* (RESET).
- If the functional form of the model is misspecified, then including additional nonlinear predictors should improve fit.
  - In particular, we usually including polynomial transformations of the predictors or the fitted values.

Ramsey RESET

Given the estimated model:

$$Y = \hat{\beta}_0 + \sum_{p=1}^{P} \hat{\beta}_p X_p + \hat{\varepsilon}$$

We'll estimated the augmented model:

$$Y = \hat{\beta}_0 + \sum_{p=1}^{P} \hat{\beta}_p X_p + \sum_{q=1}^{Q} \hat{\gamma}_q Z_q + \hat{\varepsilon}$$

Then we check if the augmented model fits better than the original.

- The Z_q are usually taken to be powers of X_p or $\hat{Y}$.
- If the augmented model fits better, we reject the null hypothesis of correct specification.

Example

Let's apply the RESET to the model plotted above:

```
out2 <- lm(MPG.highway ~ Horsepower, data = Cars93)

resettest(out2)

##
##  RESET test
##
## data:  out2
## RESET = 16.718, df1 = 2, df2 = 89, p-value =
## 6.852e-07
```

The test definitely suggests misspecification.

Example

What happens when we add the square of horsepower?

```
out3 <- update(out2, ". ~ . + I(Horsepower^2)")
resettest(out3)

##
##  RESET test
##
## data:  out3
## RESET = 5.1653, df1 = 2, df2 = 88, p-value = 0.007567
```

We still reject the null.

- The test is telling us that our model is still incorrect.

Example

What about the cubic model?

```
out4 <- update(out3, ". ~ . + I(Horsepower^3)")
resettest(out4)

##
##  RESET test
##
## data:  out4
## RESET = 1.2906, df1 = 2, df2 = 87, p-value = 0.2803
```

Now we've finally failed to reject the null of correct specification.

- The test cannot tell us that our model is incorrect.
- We still *cannot* conclude that our model is correctly specified.

Limitations of the Ramsey RESET

The RESET is only meant to detect nonlinear misspecifications.

- It will not detect omitted variables that linearly predict Y.

The RESET is also sensitive to heteroscedasticity.

- The model comparison used to test the $\hat{\gamma}_q$ is based on significance tests that are sensitive to non-constant errors.
- With severe heteroscedasticity, the test can reject the null hypothesis when the model is correctly specified.

We can run the RESET with robust standard errors.

- This robust version tends to spuriously indicate misspecification, if the errors are not heteroscedastic (Long & Trivedi, 1993).

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.

If we suspect omitted variables, but we don't have access to key variables, we can try:

- Proxy variables
- Instrumental variables
- Fixed effects regression

Other Causes of Endogeneity

Simultaneity (i.e., the third variable problem)

- X and Y are both caused by an unmodeled third variable.
- The absence of the true cause induces a spurious correlation between X and Y.
- This is actually special type of omitted variable problem.

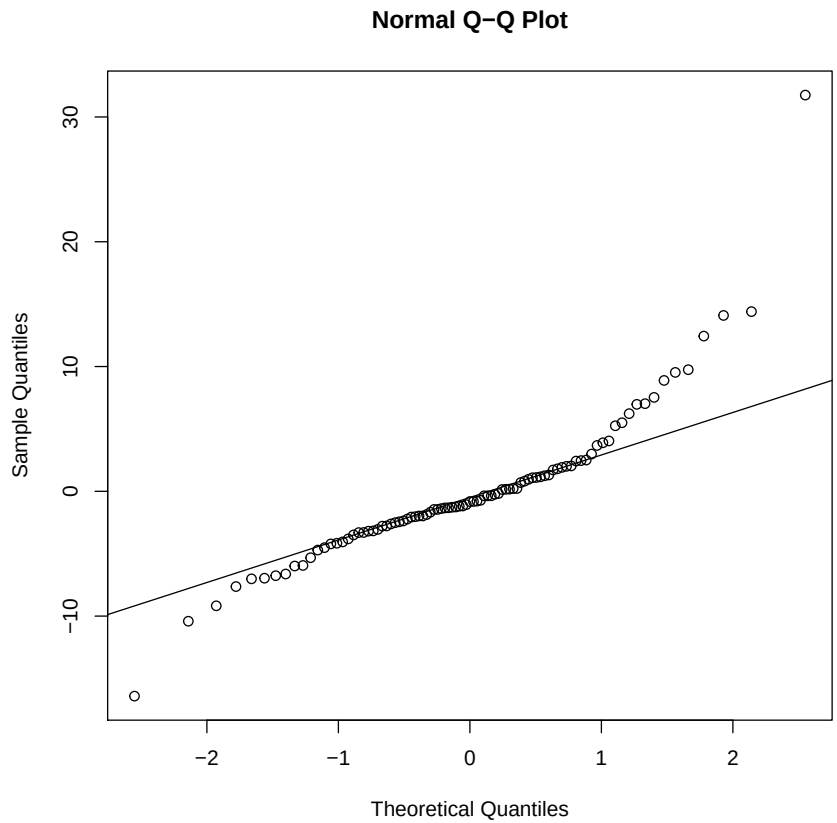
Measurement error in the X_p variables

- If some predictors are measured with error, their estimated effects will be biased.
- We can use fancy-pants corrections for this bias.
- We can also use *latent variable* models.
- Measurement error in Y increases residual variance, but does not bias parameter estimates.

Normality Assumption

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.



Evaluating Normality

Next to Q-Q Plots, one of the simplest ways to evaluate the normality of the error distribution is through the *skewness* and *kurtosis*.

- Skewness and kurtosis are functions of the third and fourth moments, respectively.

The *moments* of a distribution are single number summaries of the distribution's characteristics.

- The zeroth moment is the total probability mass.
- The first moment is the mean.
- The second central moment is the variance.
- The third standardized moment is the skewness.
- The fourth standardized moment is the kurtosis.

Moments

The kth raw moment of X (around zero):

$$M_k = \mathrm{E}\left[X^k\right]$$

The kth central moment of X:

$$\widetilde{M}_k = \mathrm{E}\left[(X - \mathrm{E}[X])^k\right] = \mathrm{E}\left[(X - \mu)^k\right]$$

The kth standardized moment of X:

$$M_k^* = \frac{\mathrm{E}\left[(X - \mathrm{E}[X])^k\right]}{\left(\sqrt{\mathrm{E}\left[(X - \mathrm{E}[X])^2\right]}\right)^k} = \frac{\mathrm{E}\left[(X - \mu_1)^k\right]}{\sigma^k}$$

Skewness

The skewness, γ, is the third standardized moment of X.

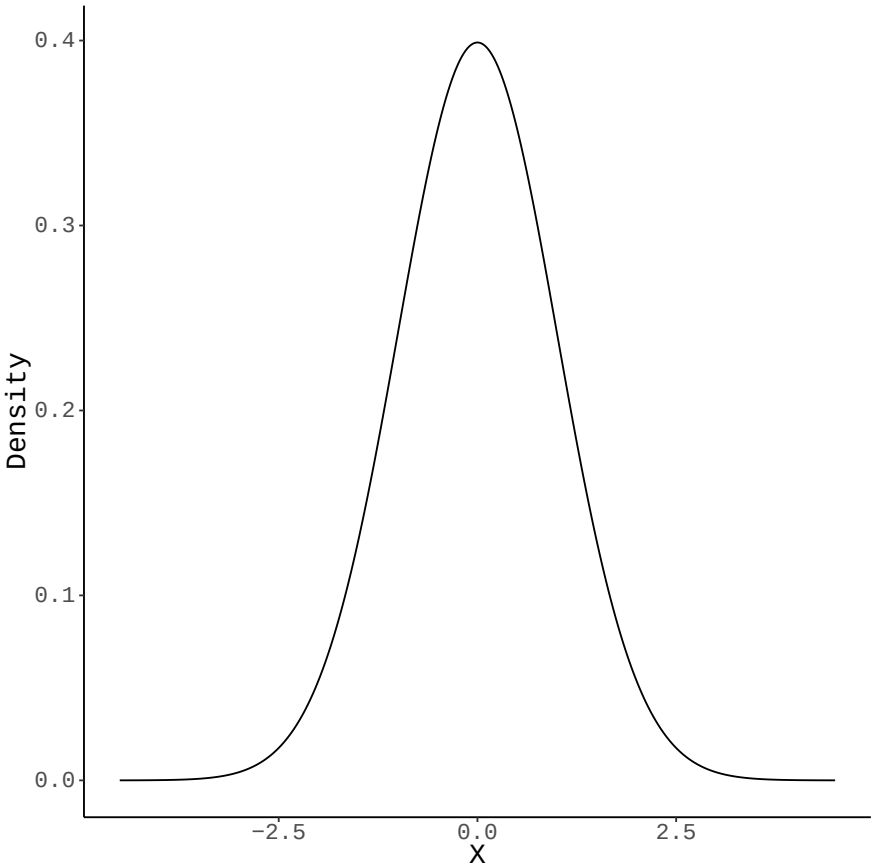
$$\gamma = M_3^* = \frac{E\left[(X-\mu)^3\right]}{\sigma^3}$$

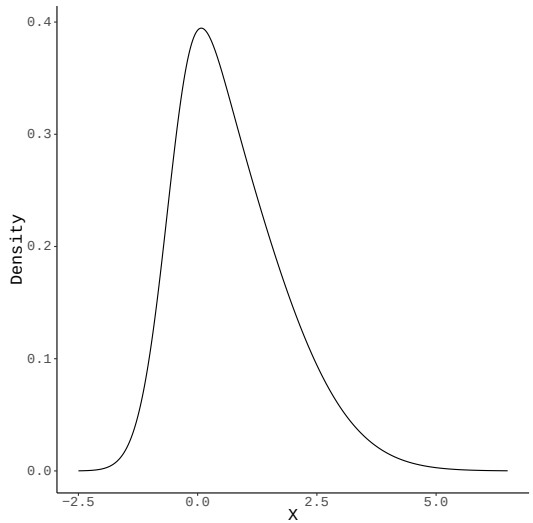
Skewness tells us about asymmetry in the distribution.

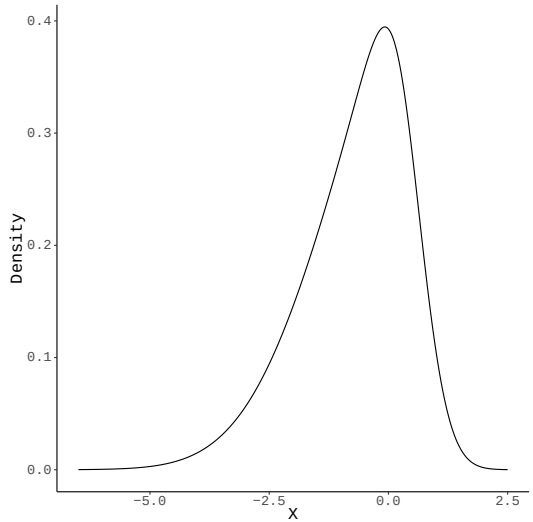
- The normal distribution has skewness $\gamma = 0$.
- Positively skewed distributions (i.e., $\gamma > 0$) will have long right tails.
- Negatively skewed distributions (i.e., $\gamma < 0$) will have long left tails.

A common cutoff says that skewness with a magnitude greater than 1 (i.e., $|\gamma| > 1$) indicates substantial non-normality.

Skewed Distributions







Kurtosis

The kurtosis, κ_0, is the fourth standardized moment of X.

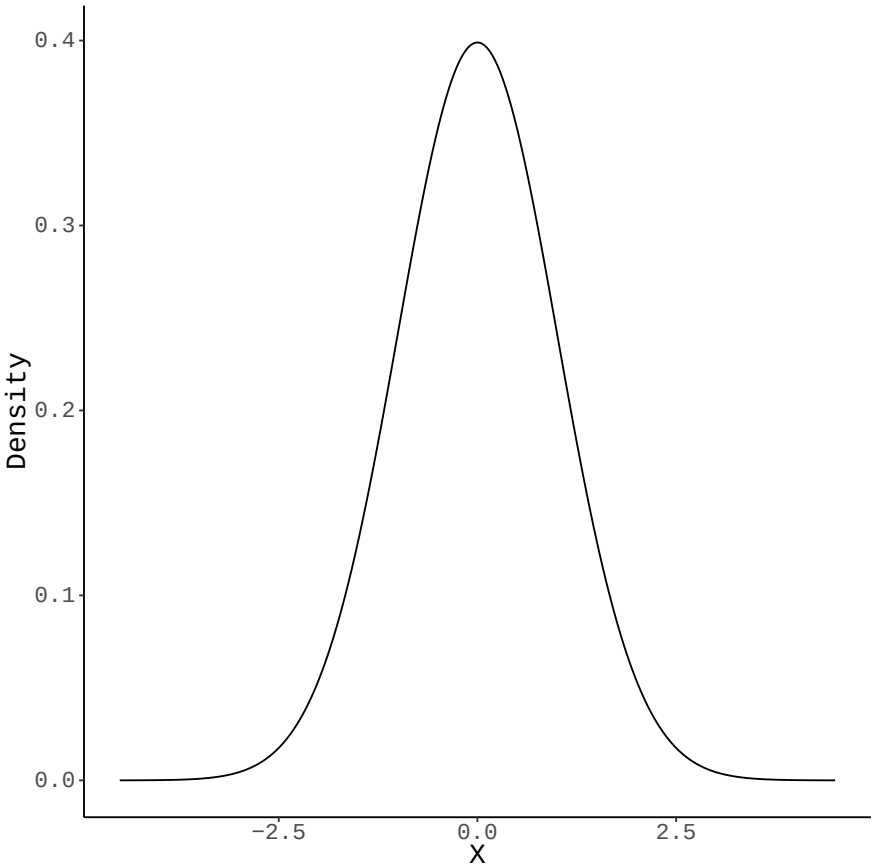
$$\kappa_0 = M_4^* = \frac{E\left[(X-\mu)^4\right]}{\sigma^4}$$

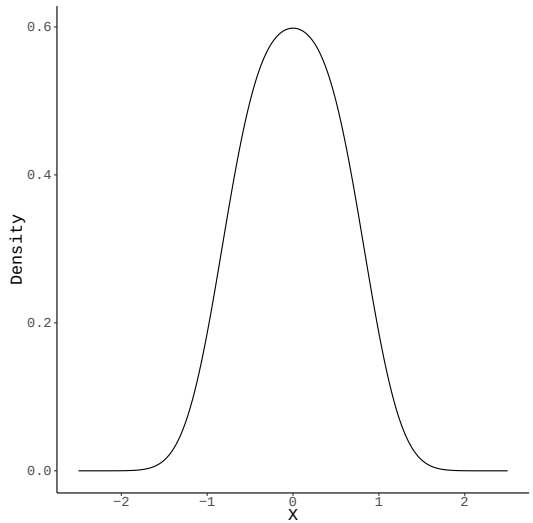
Kurtosis tells us about the relative heaviness of the distribution's tails.

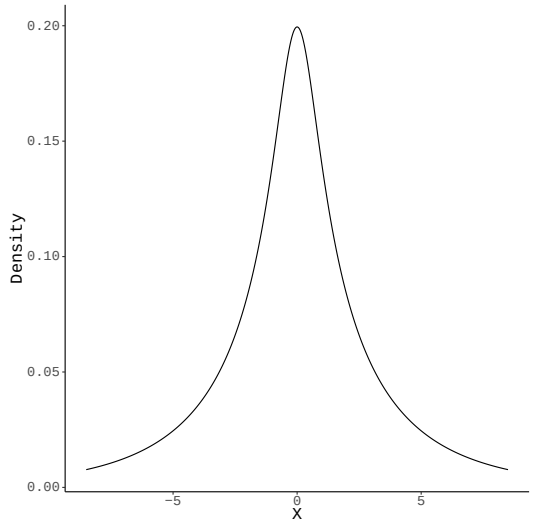
- The normal distribution has a raw kurtosis of $\kappa_0 = 3$.
- The *excess kurtosis*, $\kappa = \kappa_0 - 3$, is a more common measure.
- Positive excess kurtosis indicates heavy tails.
- Negative excess kurtosis indicates light tails.

A common cutoff says that excess kurtosis greater than 7 (i.e., $\kappa_0 > 10$) indicates substantial non-normality.

Kurtotic Distributions







Tests of Normality

We can also conduct formal statistical tests of the normality assumption.

- Shapiro-Wilks test
- Kolmogorov-Smirnov test / Lilliefors test
- Cramer-von Mises test
- Anderson-Darling test

These tests tend to be sensitive to sample size.

- The Shapiro-Wilks test will have the best power in small samples.
- The Kolmogorov-Smirnov test is probably best for large samples.

Consequences of Violating Normality

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

- In small samples, with *random* predictors, normal errors *do not* imply normally distributed 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 small samples, with *random* predictors, normal errors *do not* imply normally distributed 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.

Central Limit Theorem

Let $\mathcal{Z} = \{z_1, z_2, \ldots, z_N\}$ be a set of $i.i.d.$ random variables with mean μ_Z and (finite) standard deviation σ_Z.

- Define the mean of $\mathcal{Z}$ as:

$$\bar{Z} \equiv N^{-1} \sum_{n=1}^{N} z_n$$

- The distribution of $\bar{Z}$ tends towards normality as N increases:

$$P(\bar{Z}) \rightarrow \mathrm{N}\left(\mu_Z, \frac{\sigma_Z}{\sqrt{N}}\right)$$

Lindeberg-Feller CLT

The Lindeberg-Feller CLT relaxes the requirement for $i.i.d.$ variables.

- Let $\mathcal{Z} = \{z_1, z_2, \ldots, z_N\}$ be a set of *independent* random variables.
  - Every z_n has finite variance
  - No z_n has an overwhelmingly large variance
- The distribution of $\bar{Z}$ approaches normality as N increases.

Lindeberg-Feller CLT

The Lindeberg-Feller CLT relaxes the requirement for $i.i.d.$ variables.

- Let $\mathcal{Z} = \{z_1, z_2, \ldots, z_N\}$ be a set of *independent* random variables.
  - Every z_n has finite variance
  - No z_n has an overwhelmingly large variance
- The distribution of $\bar{Z}$ approaches normality as N increases.

The Lindeberg-Feller CLT implies normal sampling distributions for the regression coefficients and predicted values in linear regression models.

Central Limit Theorem for Regression

Recall the definition of our regression coefficients:

$$\hat{\beta}_p = \left[\sum_{n=1}^{N} \left(X_n - \bar{X}\right)\left(Y_n - \bar{Y}\right)\right] \Big/ \sum_{n=1}^{N} \left(X_n - \bar{X}\right)^2$$

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y = \mathbf{A}Y$$

- $\mathbf{A} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is a $P \times N$ matrix of weights.

Central Limit Theorem for Regression

Recall the definition of our regression coefficients:

$$\hat{\beta}_p = \left[\sum_{n=1}^{N} \left(X_n - \bar{X}\right)\left(Y_n - \bar{Y}\right)\right] \Big/ \sum_{n=1}^{N} \left(X_n - \bar{X}\right)^2$$

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y = \mathbf{A}Y$$

- $\mathbf{A} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is a $P \times N$ matrix of weights.

Each of the P regression coefficient can be estimated as:

$$\hat{\beta}_p = \sum_{n=1}^{N} a_{pn} y_n$$

- a_{pn} represents the element in position (p, n) of $\mathbf{A}$

Central Limit Theorem for Regression

Recall the definition of our regression coefficients:

$$\hat{\beta}_p = \left[\sum_{n=1}^{N} (X_n - \bar{X})(Y_n - \bar{Y})\right] \Big/ \sum_{n=1}^{N} (X_n - \bar{X})^2$$

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y = \mathbf{A}Y$$

- $\mathbf{A} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is a $P \times N$ matrix of weights.

Each of the P regression coefficient can be estimated as:

$$\hat{\beta}_p = \sum_{n=1}^{N} a_{pn} y_n$$

- a_{pn} represents the element in position (p, n) of $\mathbf{A}$

If we take $a_{pn} \equiv z_n \in \mathcal{Z}$, the Lindeberg-Feller CLT implies normal sampling distributions for the $\hat{\beta}_p$, if N is large enough.

Central Limit Theorem for Regression

The same logic implies normal distributions for predicted values.

$$\begin{aligned}\hat{Y} &= \mathbf{X}\hat{\beta} \\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y \\ &= \mathbf{H}Y\end{aligned}$$

- $\mathbf{H}$ is an $N \times N$ projection matrix (the *hat* matrix).

Central Limit Theorem for Regression

The same logic implies normal distributions for predicted values.

$$\begin{aligned}\hat{Y} &= \mathbf{X}\hat{\beta} \\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y \\ &= \mathbf{H}Y\end{aligned}$$

- $\mathbf{H}$ is an $N \times N$ projection matrix (the *hat* matrix).

Each of the N predicted values can be estimated as:

$$\hat{Y}_n = \sum_{i=1}^{N} h_{ni} y_i$$

- h_{ni} represents the element in position (n, i) of $\mathbf{H}$.

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 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:

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

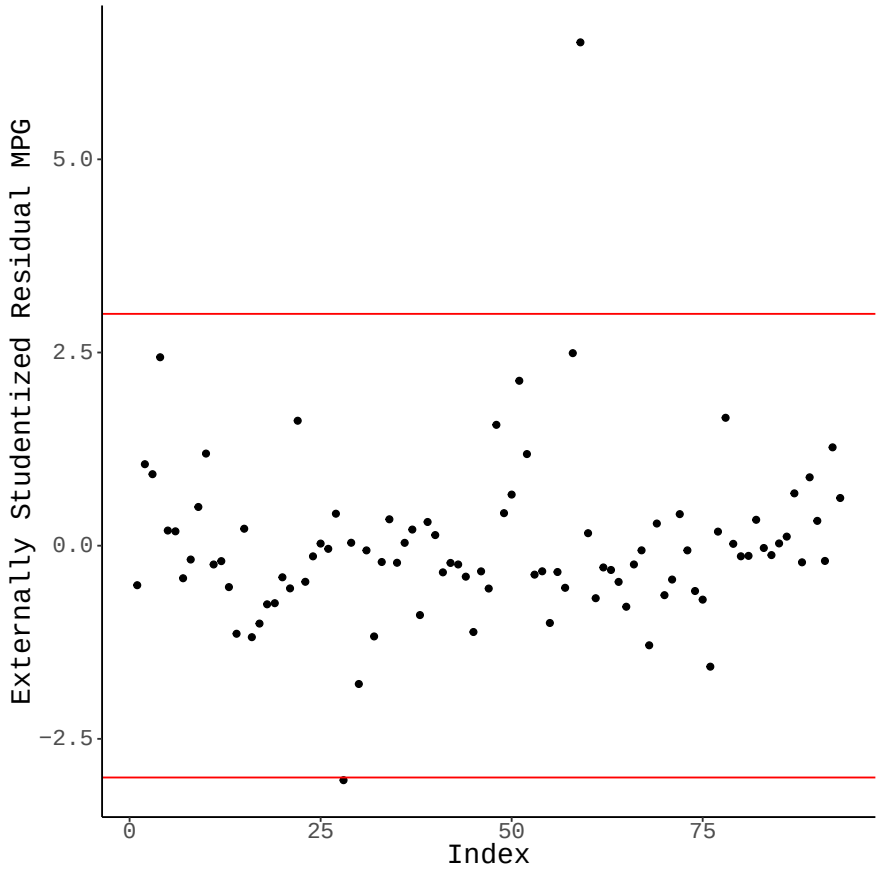
The externally studentized residuals have two very useful properties:

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

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:

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

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:

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

In multiple linear regression, the leverages are given by the diagonal of the hat matrix, $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$:

$$h_n = \mathbf{H}[n, n]$$

Properties of Leverages

Leverages have the following useful properties:

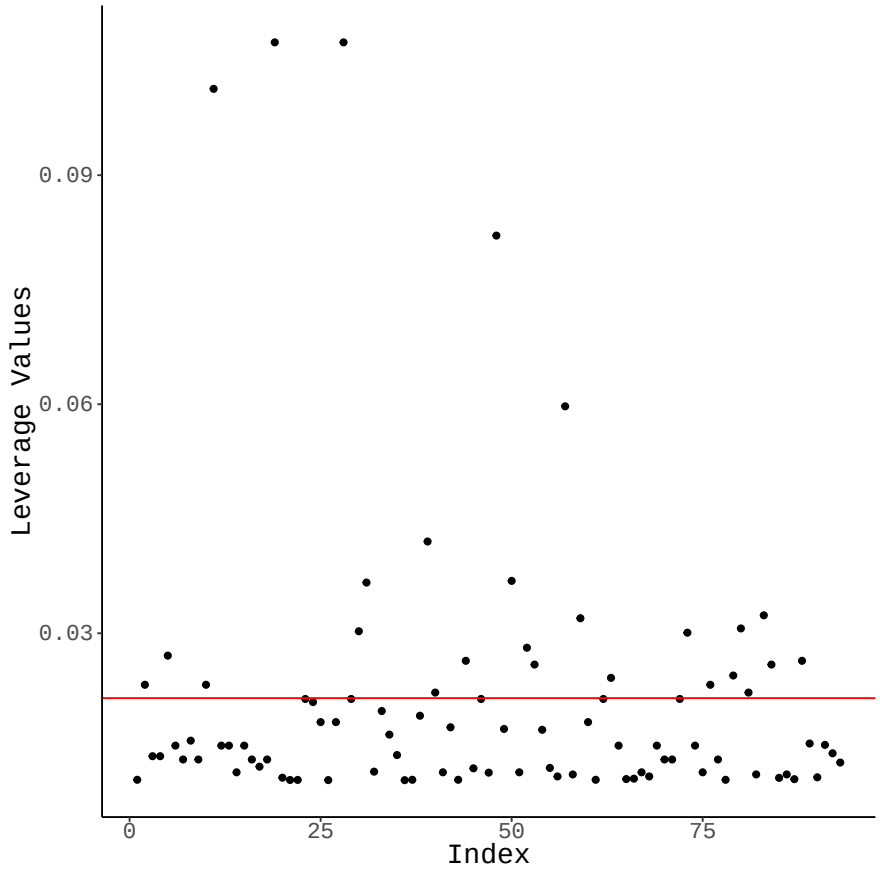
- h_n grows in direct proportion to the distance between X_n and $\bar{X}$.
- $h_n \in \left[\frac{1}{N}, 1.0\right]$
- $\bar{h} = \frac{P+1}{N}$

Observations with $h_n \gg \bar{h}$ are potentially influential.

Leverage Plots

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

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



From Outliers and Leverages to 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 *D*
   - *DFFITS*
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

Cook's *D* and *DFFITS* are more-or-less interchangeable measures.

- They each provide very similar information.
- They will tend to lead to the same conclusions.

$$DFFITS_n = \frac{\hat{Y}_n - \hat{Y}_{(n)}}{\sqrt{\hat{\sigma}^2_{(n)} h_n}} = t_{(n)} \sqrt{\frac{h_n}{1 - h_n}}$$

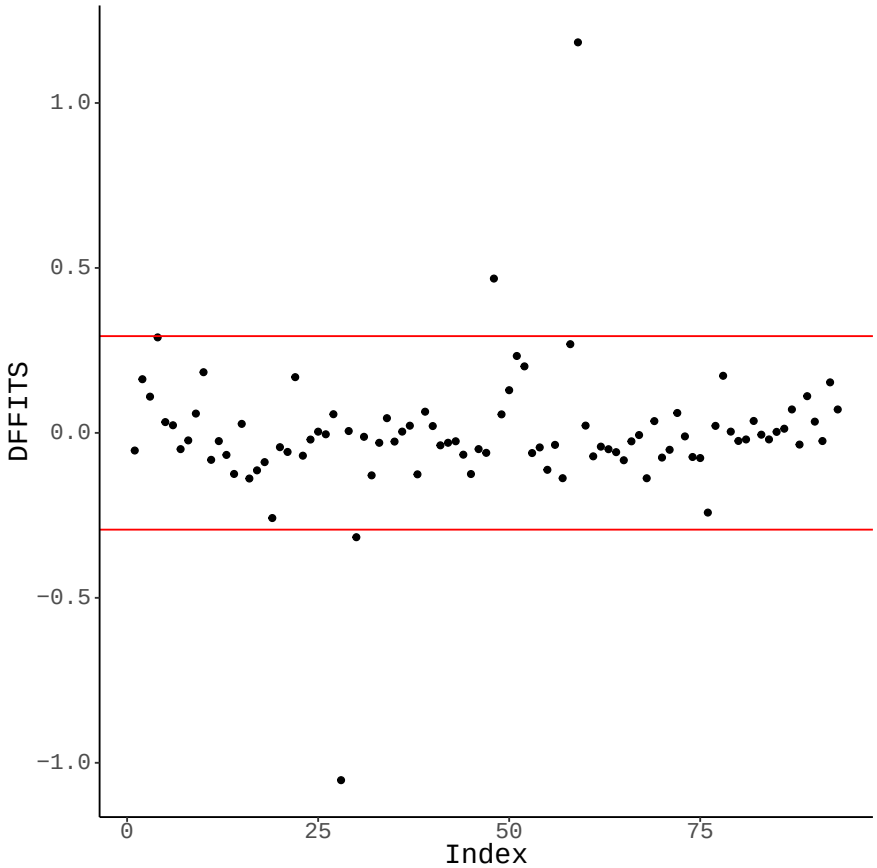
$$\text{Cook's } D_n = \frac{\sum_{n=1}^{N} \left(\hat{Y}_n - \hat{Y}_{(n)} \right)^2}{(P+1)\,\hat{\sigma}^2} = (P+1)^{-1} t_n^2 \frac{h_n}{1 - h_n}$$

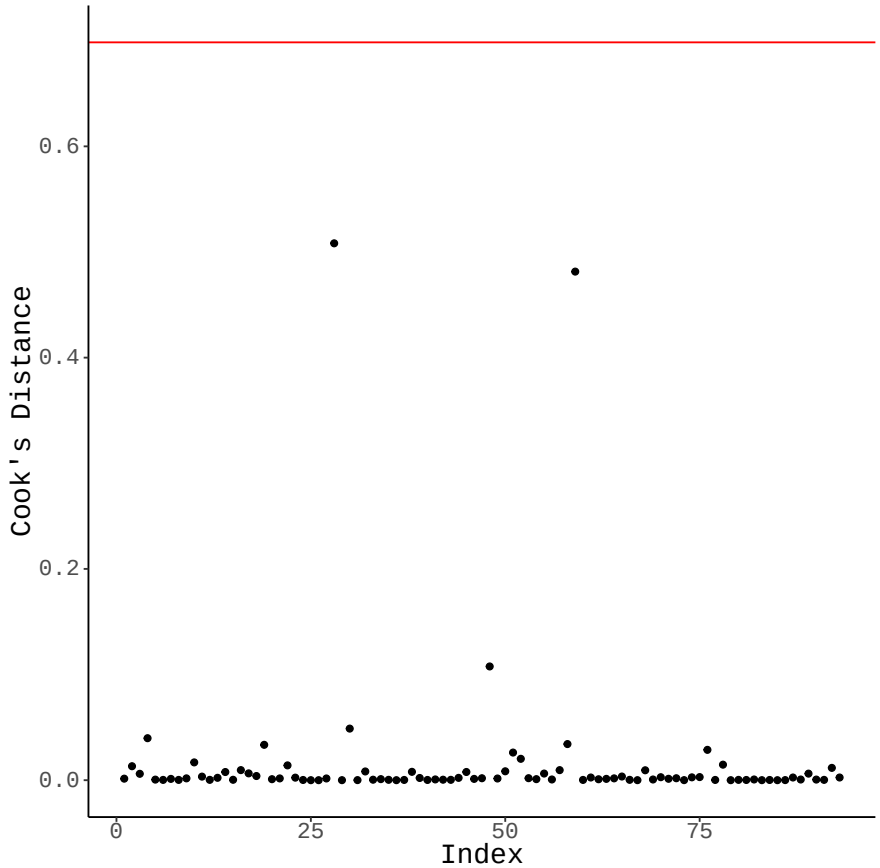
Rules-of-Thumb

The recommend thresholds for problematic degrees of global influence:

- $|DFFITS_n| > 2\sqrt{\frac{P+1}{N}}$
- Cook's $D_n >$ the critical F value for $\alpha = 0.5$, $df_{num} = P + 1$, and $df_{den} = N - P - 1$

Plots of Global Measures of Influence





Coefficient-Specific Measures of Influence

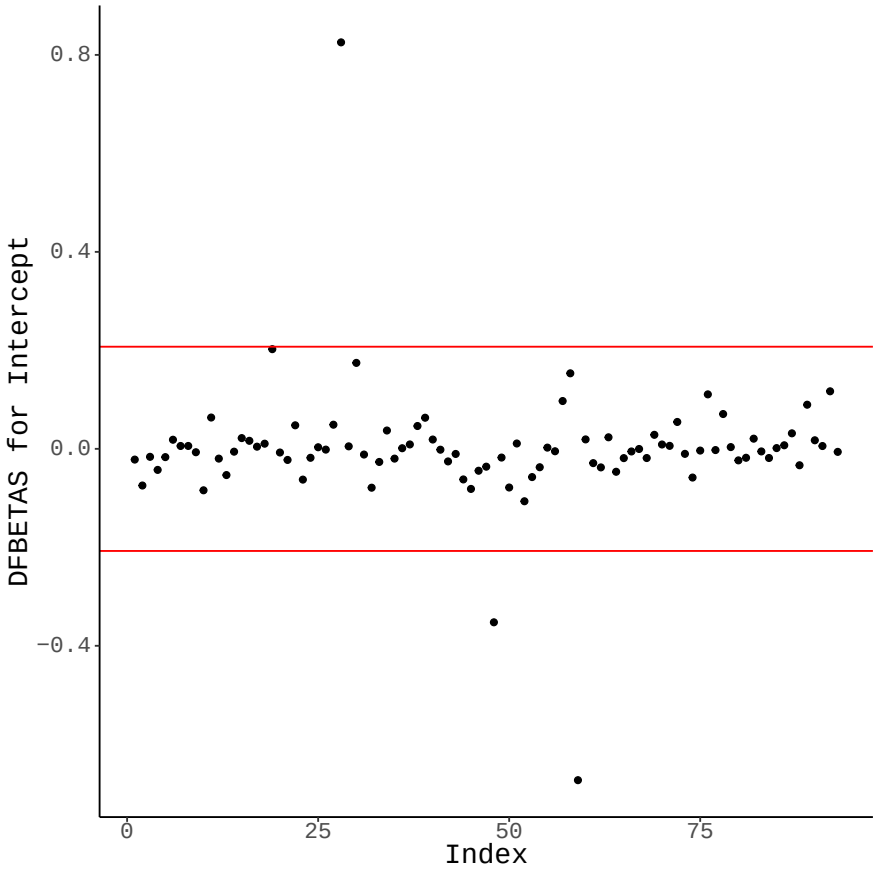
Each regression coefficient (including the intercept) gets a *DFBETAS* value for each observation.

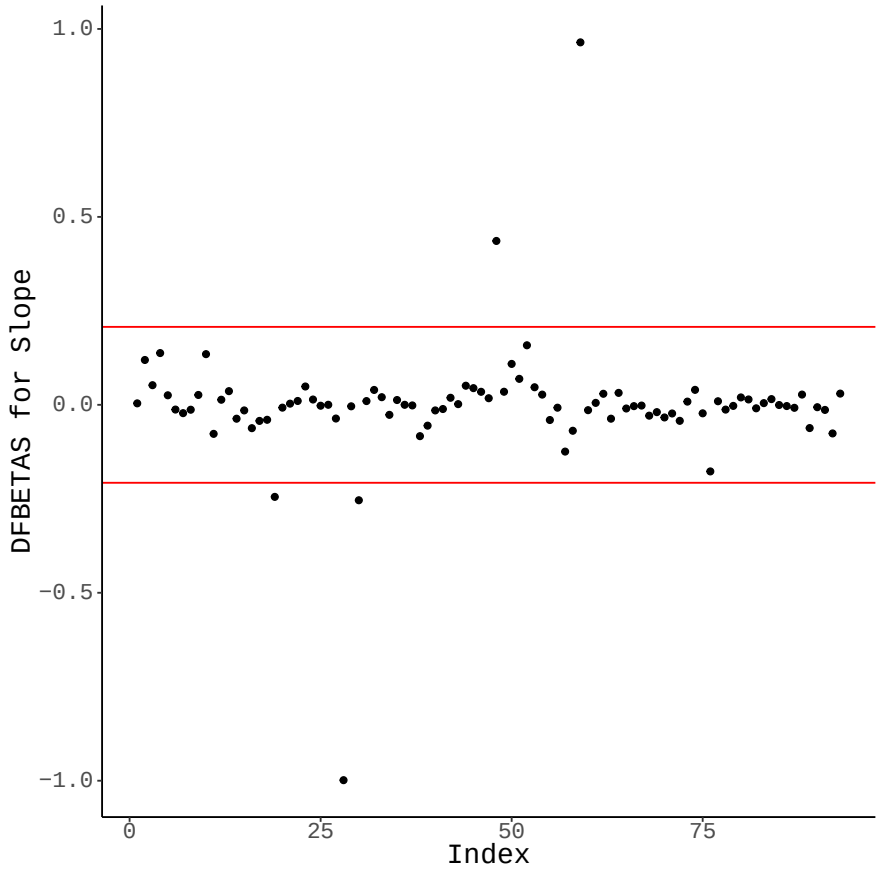
$$DFBETAS_{np} = \frac{\hat{\beta}_p - \hat{\beta}_{p(n)}}{SE_{\hat{\beta}_{p(n)}}}$$

The rule-of-thumb for flagging influential observations:

- $|DFBETAS_{np}| > \frac{2}{\sqrt{N}}$

Plots of Coefficient-Specific Measures

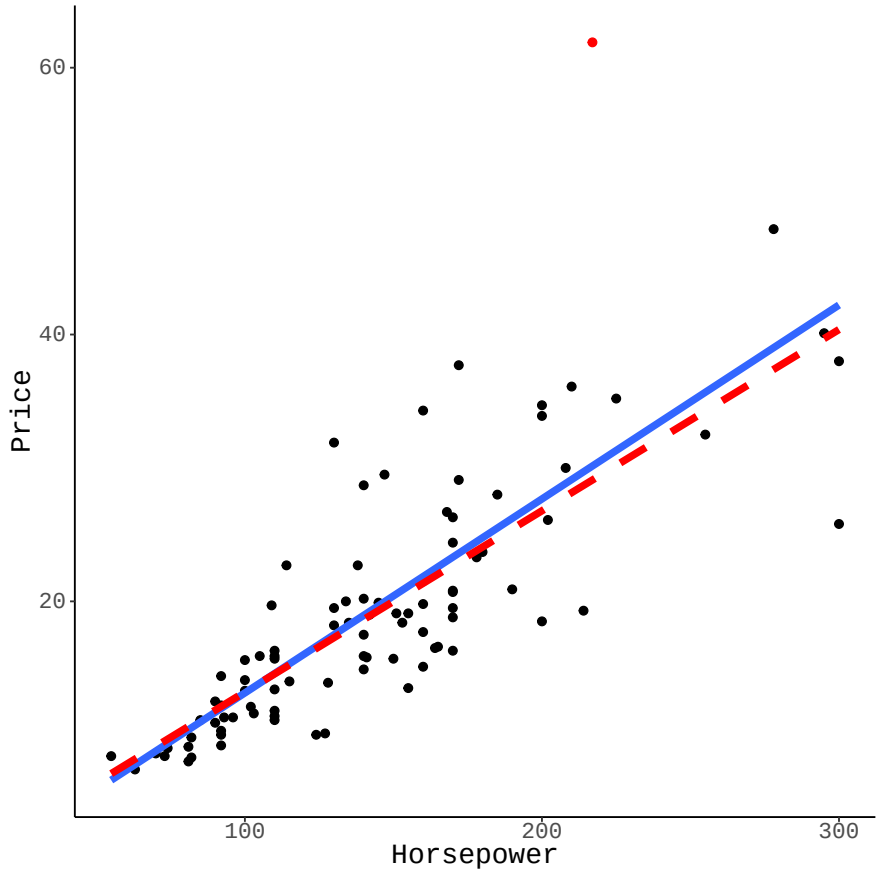




Consequences of Removing Influential Observations

Observation number 59 was the most influential according to *DFFITS*.

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



Consequences of Removing Influential Observations

```
## Fit model with full sample:
out1 <- lm(Price ~ Horsepower, data = Cars93)

## Exclude outliers:
Cars93.2 <- Cars93[-59, ]

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

Consequences of Removing Influential Observations

```
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) -0.383557    1.51673 -0.252884 0.800934
## Horsepower   0.135860    0.00997 13.626694 0.000000
```

Consequences of 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
## -14.5746  -2.6501  -0.9477   1.8087  14.7156
```

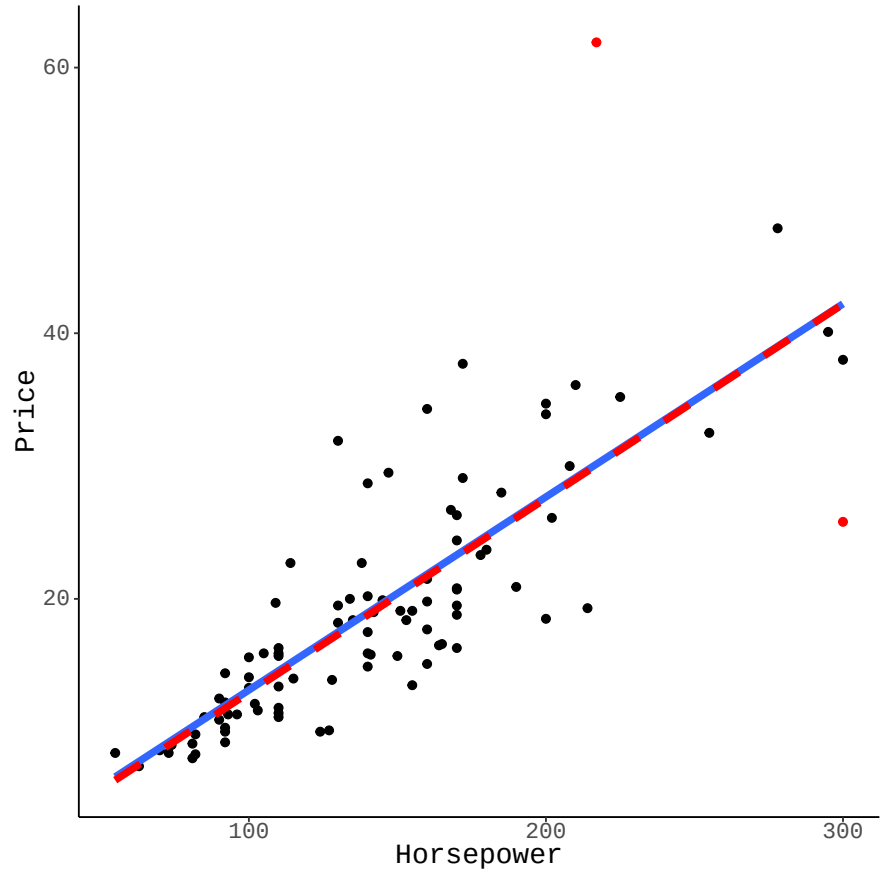
Consequences of Removing Influential Observations

```
unlist(
    summary(out1)[c("sigma", "r.squared", "fstatistic")]
)[1 : 3]

##             sigma         r.squared fstatistic.value
##          5.976953          0.621287        149.287468

unlist(
    summary(out2)[c("sigma", "r.squared", "fstatistic")]
)[1 : 3]

##             sigma         r.squared fstatistic.value
##         4.9545903         0.6735426       185.6867952
```

Consequences of Removing Influential Observations

If we remove the two most influential observations, $n = \{28, 59\}$, 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.



Consequences of Removing Influential Observations

```
## Fit model with full sample:
out1.2 <- lm(Price ~ Horsepower, data = Cars93)

## Exclude outliers:
Cars93.2 <- Cars93[-c(28, 59), ]

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

Consequences of Removing Influential Observations

```
round(summary(out1.2)$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
```

Consequences of Removing Influential Observations

```
partSummary(out1.2, 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
```

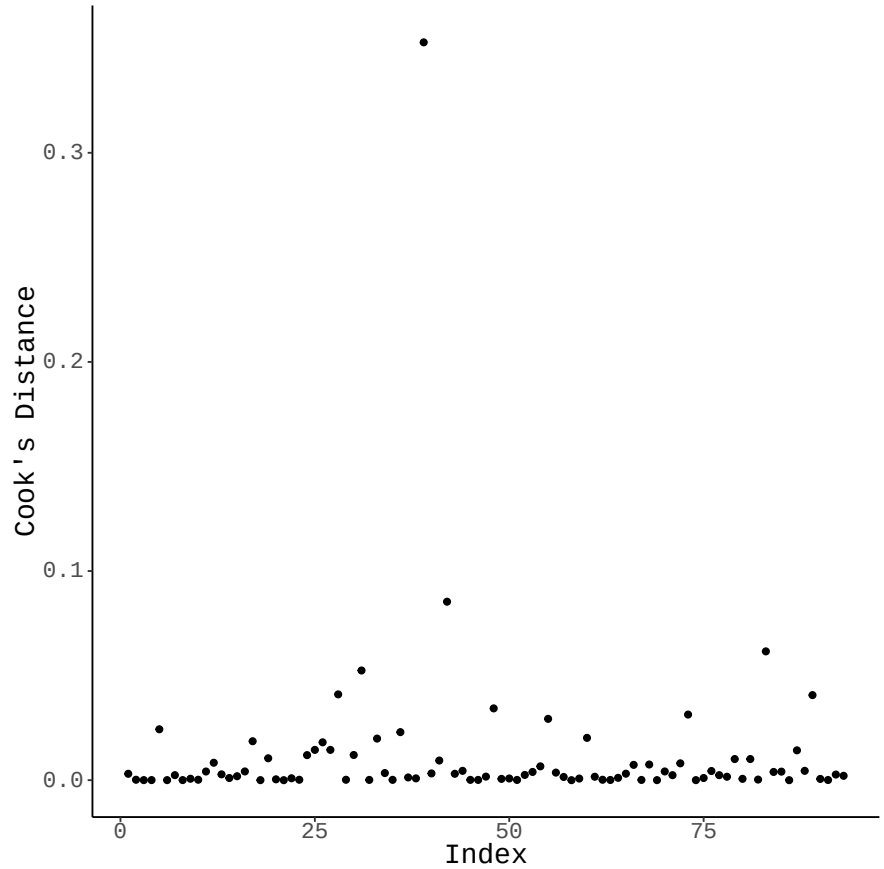
Consequences of Removing Influential Observations

```
unlist(
    summary(out1.2)[c("sigma", "r.squared", "fstatistic")]
)[1 : 3]

##             sigma         r.squared fstatistic.value
##          5.976953          0.621287        149.287468

unlist(
    summary(out2.2)[c("sigma", "r.squared", "fstatistic")]
)[1 : 3]

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

Influential Points Violating Assumptions

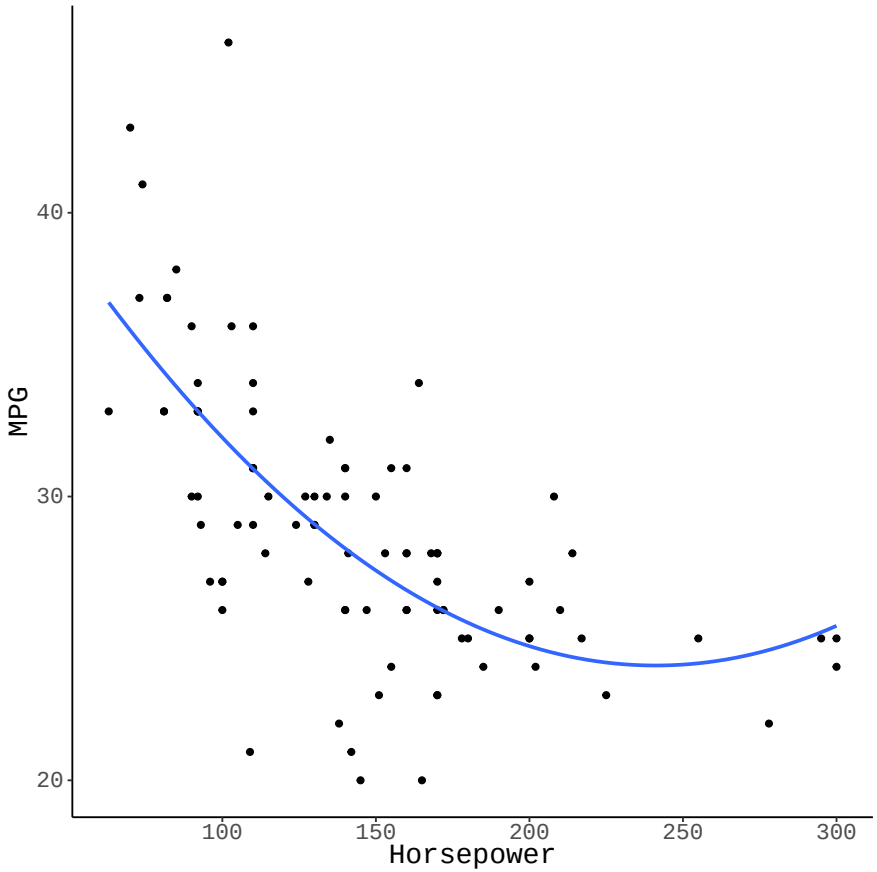
Let's revisit the nonlinear relationship between *Horsepower* and *MPG*.

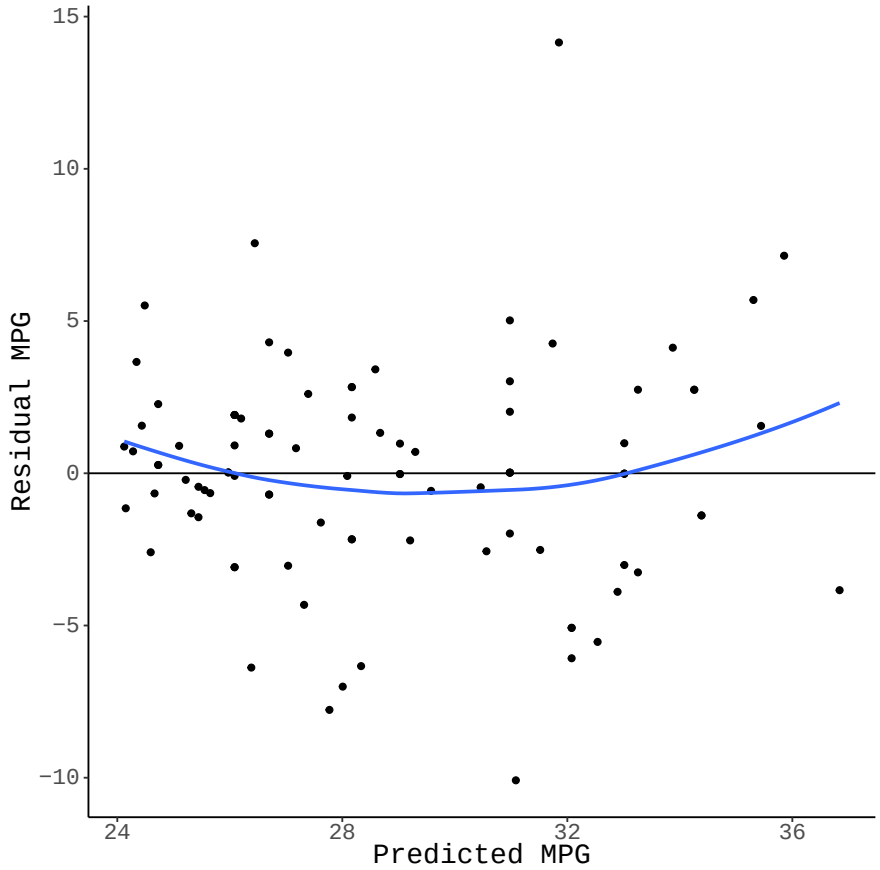
- The residual plots never really behaved.
- Maybe some influential observations are causing trouble?
  - $n = 39$ certainly looks like a jerk.



Quadratic Model

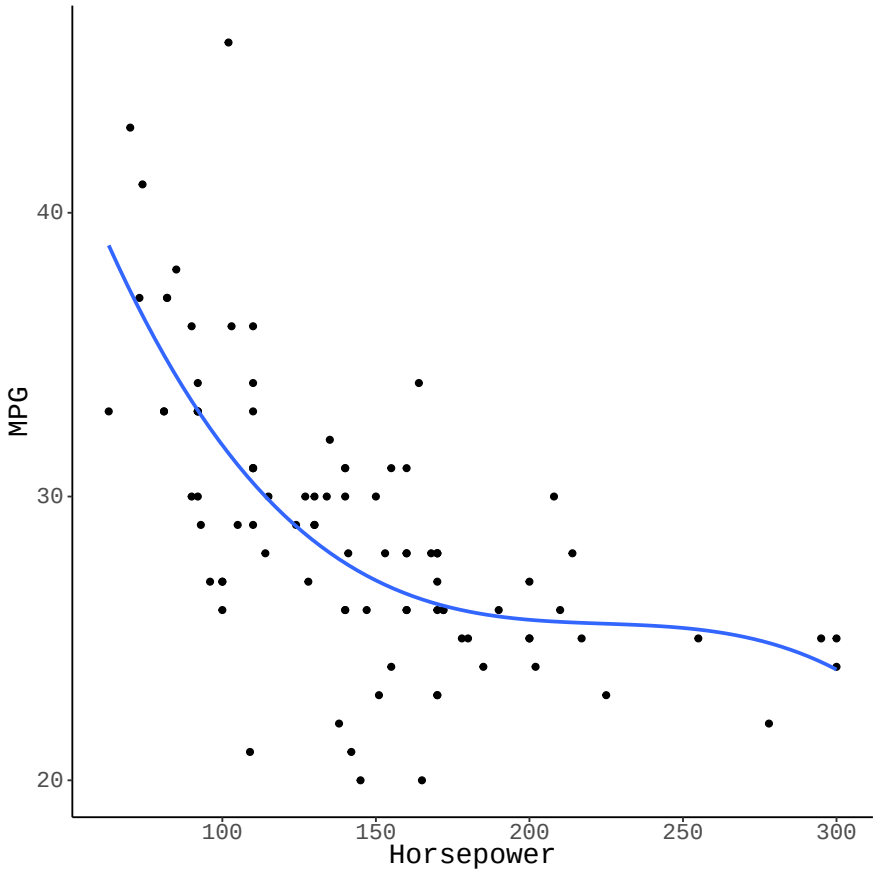
Residuals look much smoother after deleting the influential point.

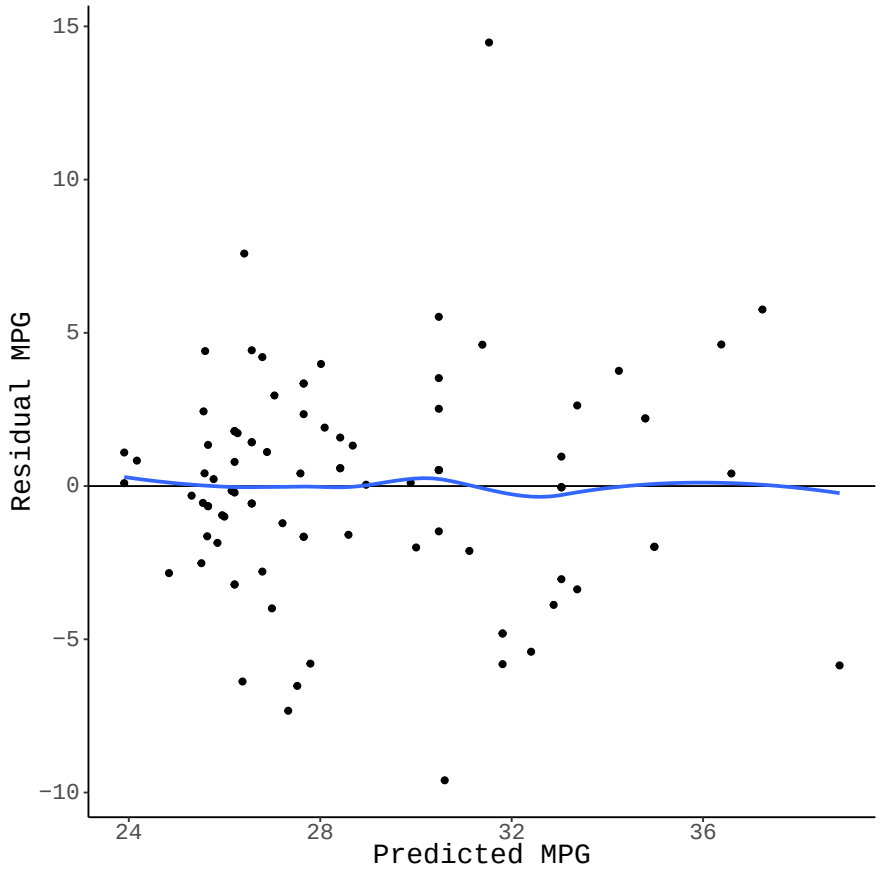




Cubic Model

Basically no trend, after fitting the cubic model to the reduced dataset.





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.

Conclusion

- The Gauss-Markov Theorem ensures that, when the assumptions are met, OLS is the BLUE of a linear regression model.
  - No other unbiased estimator will have lower variance than OLS.

- We stated the usual assumptions of OLS regression as:
  1. The model is linear in the parameters.
  2. The predictor matrix is full rank.
  3. The predictors are strictly exogenous.
  4. The errors have constant, finite variance.
  5. The errors are uncorrelated.
  6. The errors are normally distributed.

- Endogenous predictors will bias parameter estimates.
  - Omitted variables
  - Measurement error

Conclusion

- Non-spherical errors will bias standard errors.
  - Parameter estimates will remain unbiased.
- Non-normally distributed errors limit our inferential abilities.
- We use regression diagnostics to check the model assumptions and find influential observations.
- Influential observations come in two flavors:
  1. Outliers
  2. High-leverage points
- Measures of influence combine outlier checks and leverage checks to find highly influential observations.
  - Measures of influence can be global or parameter-specific.

References

Long, J. S., & Trivedi, P. K. (1993). Some specification tests for the linear regression model. *Sociological Methods & Research*, *21*(2), 161–204. doi: 10.1177/0049124192021002003