Introduction to Linear Modeling

Fundamental Techniques in Data Science with R

Utrecht University

Kyle M. Lang

Department of Methodology & Statistics
Utrecht University

Outline

Regression Problem

Some of the most ubiquitous and useful statistical models are *regression models*.

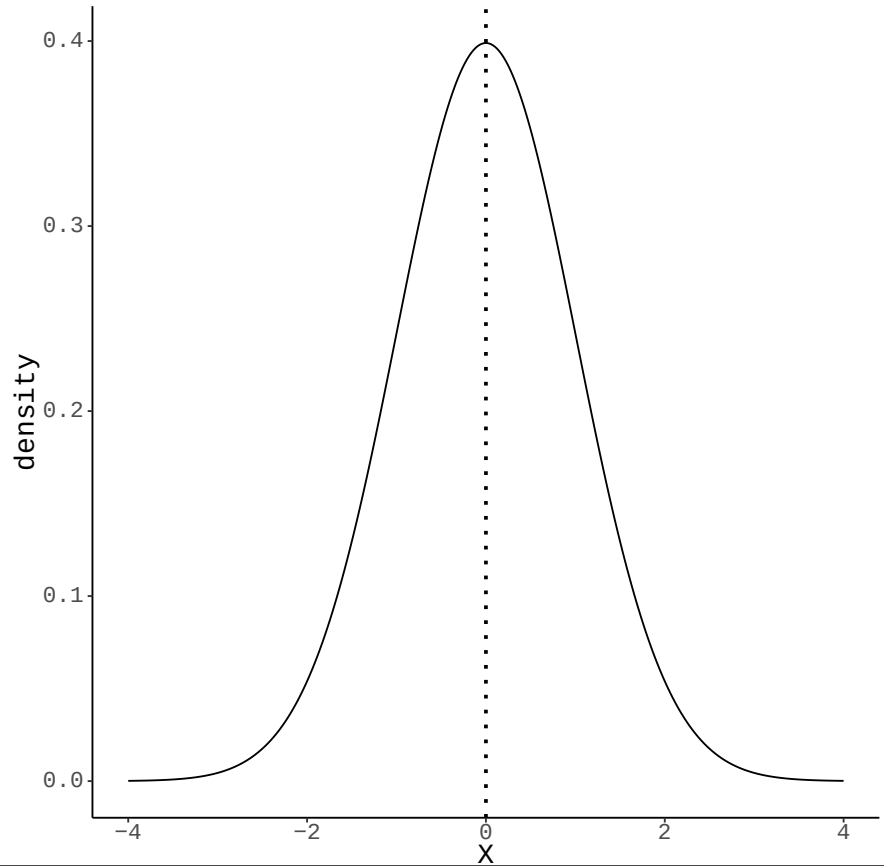
- *Regression* problems (as opposed to *classification* problems) involve modeling a quantitative response.
- The regression problem begins with a random outcome variable, Y.
- We hypothesize that the mean of Y is dependent on some set of fixed covariates, $\mathbf{X}$.

Flavors of Probability Distribution

```
Warning:  Using ‘size‘ aesthetic for
lines was deprecated in ggplot2
3.4.0.
i Please use ‘linewidth‘ instead.
```

The distributions with which you're probably most familiar imply a constant mean.

- Each observation is expected to have the same value of Y, regardless of their individual characteristics.
- This type of distribution is called "marginal" or "unconditional."



Flavors of Probability Distribution

The distributions we consider in regression problems have *conditional means*.

- The value of Y that we expect for each observation is defined by the observations' individual characteristics.
- This type of distribution is called "conditional."

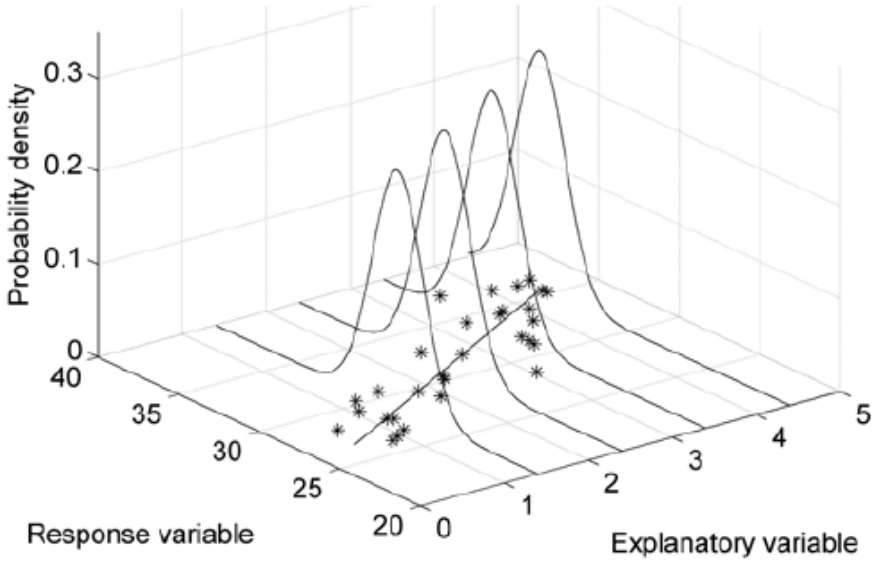
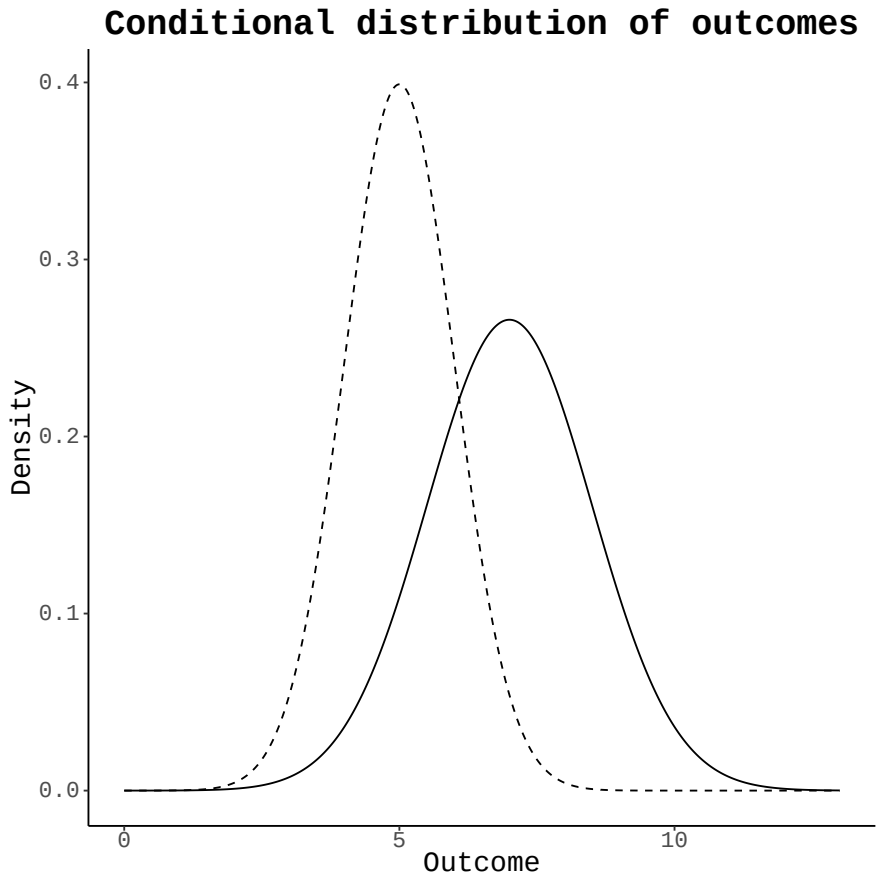


Image retrieved from:
`http://www.seaturtle.org/mtn/archives/mtn122/mtn122p1.shtml`

Flavors of Probability Distribution

Even a simple comparison of means implies a conditional distribution.

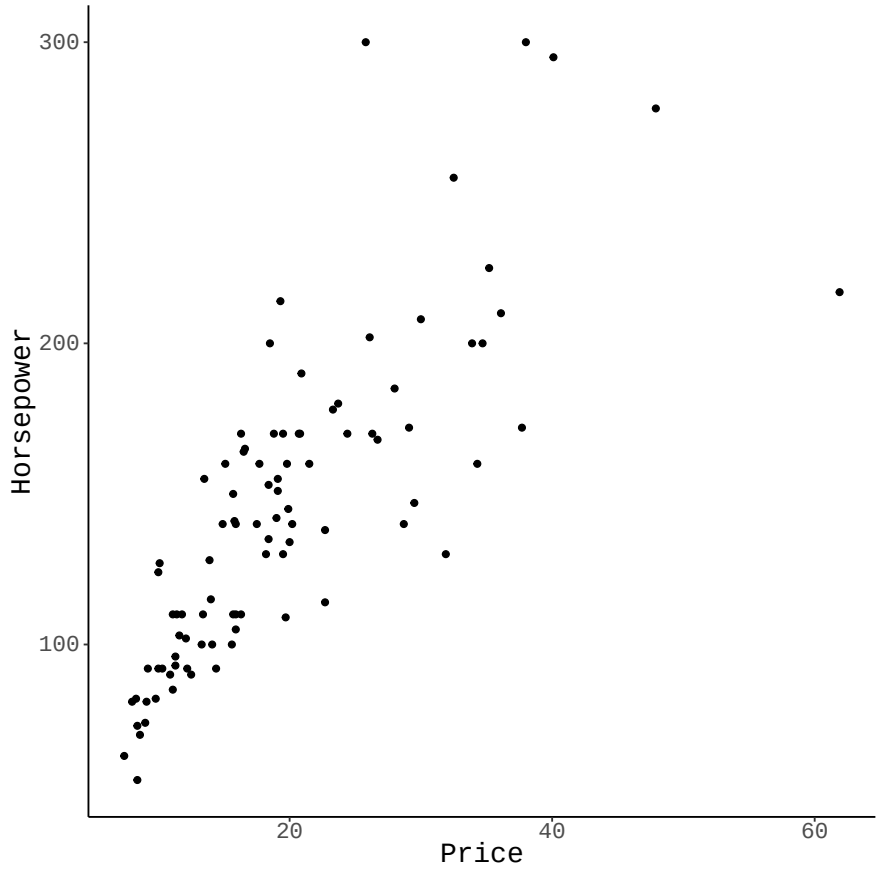
- The solid curve corresponds to outcome values for one group.
- The dashed curve represents outcomes from the other group.



Projecting a Distribution onto the Plane

In practice, we only interact with the X-Y plane of the previous 3D figure.

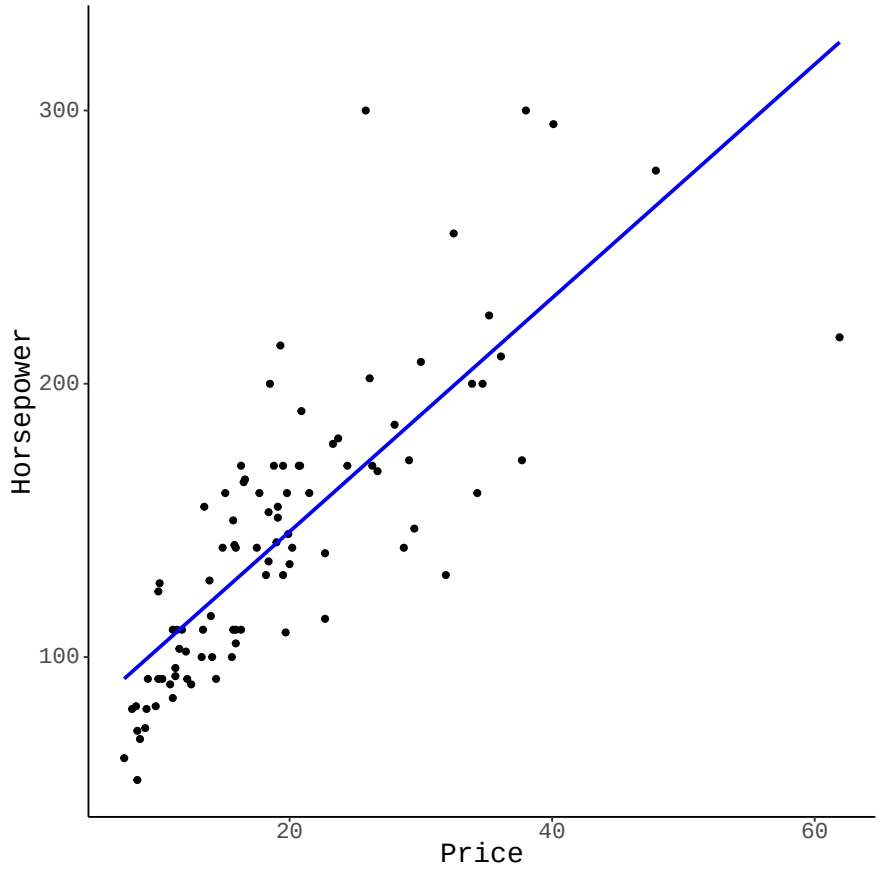
- On the Y-axis, we plot our outcome variable
- The X-axis represents the predictor variable upon which we condition the mean of Y.



Modeling the X–Y Relationship in the Plane

We want to explain the relationship between Y and X by finding the line that traverses the scatterplot as "closely" as possible to each point.

- This is the "best fit line".
- For any given value of X the corresponding point on the best fit line is our best guess for the value of Y, given the model.



Simple Linear Regression

Simple Linear Regression

The best fit line is defined by a simple equation:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

The above should look very familiar:

$$\begin{aligned} Y &= mX + b \\ &= \hat{\beta}_1 X + \hat{\beta}_0 \end{aligned}$$

$\hat{\beta}_0$ is the *intercept*.

- The $\hat{Y}$ value when $X = 0$.
- The expected value of Y when $X = 0$.

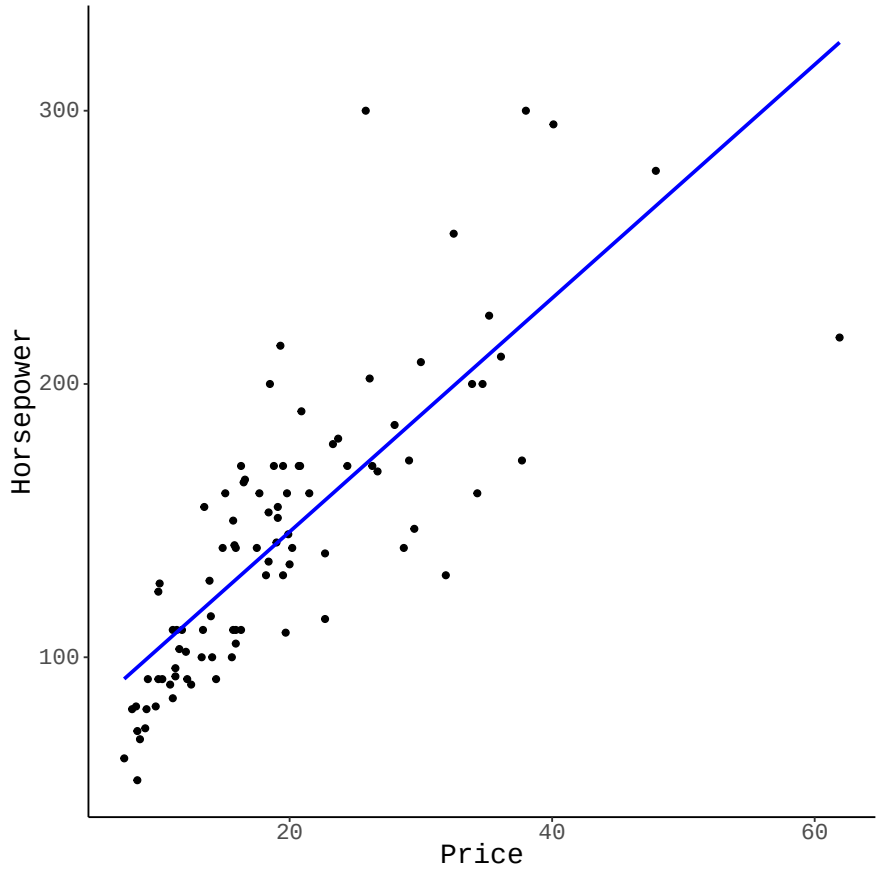
$\hat{\beta}_1$ is the *slope*.

- The change in $\hat{Y}$ for a unit change in X.
- The expected change in Y for a unit change in X.

Thinking about Error

The equation $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ only describes the best fit line.

- It does not fully quantify the relationship between Y and X.



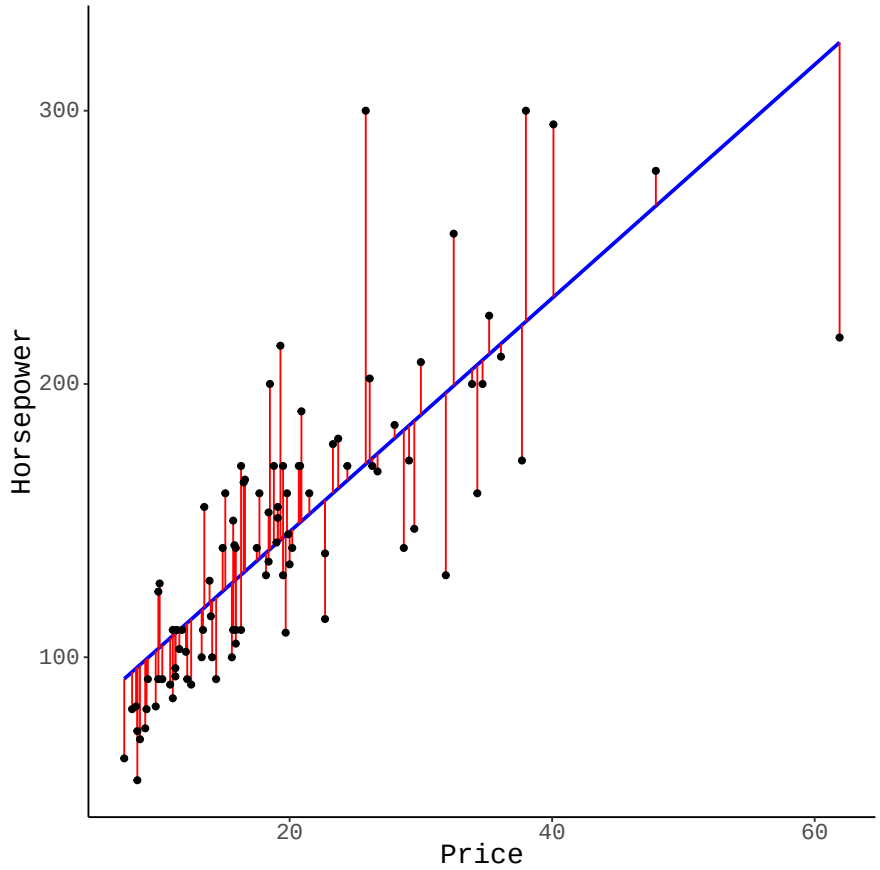
Thinking about Error

The equation $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ only describes the best fit line.

- It does not fully quantify the relationship between Y and X.

We still need to account for the estimation error.

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\varepsilon}$$



Estimating the Regression Coefficients

The purpose of regression analysis is to use a sample of N observed $\{Y_n, X_n\}$ pairs to find the best fit line defined by $\hat{\beta}_0$ and $\hat{\beta}_1$.

- The most popular method of finding the best fit line involves minimizing the sum of the squared residuals.
- $RSS = \sum_{n=1}^{N} \hat{\varepsilon}_n^2$

Residuals as the Basis of Estimation

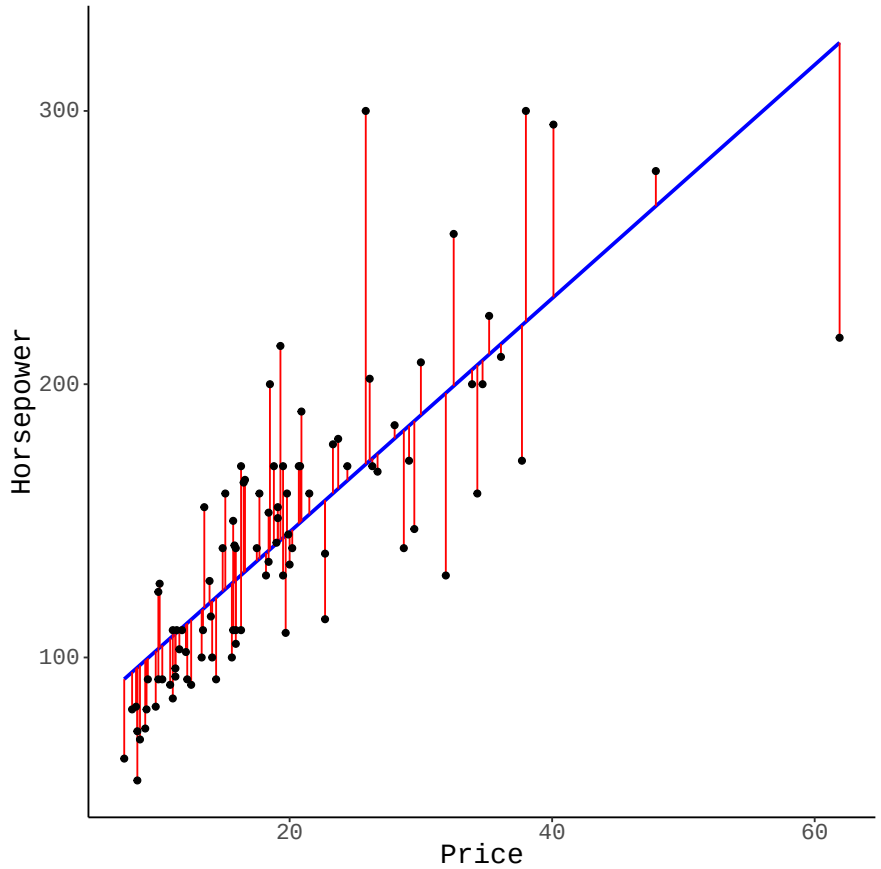
The $\hat{\varepsilon}_n$ are defined in terms of deviations between each observed Y_n value and the corresponding $\hat{Y}_n$.

$$\hat{\varepsilon}_n = Y_n - \hat{Y}_n = Y_n - \left(\hat{\beta}_0 + \hat{\beta}_1 X_n\right)$$

Each $\hat{\varepsilon}_n$ is squared before summing to remove negative values.

$$RSS = \sum_{n=1}^{N} \hat{\varepsilon}_n^2 = \sum_{n=1}^{N} \left(Y_n - \hat{Y}_n\right)^2$$

$$= \sum_{n=1}^{N} \left(Y_n - \hat{\beta}_0 - \hat{\beta}_1 X_n\right)^2$$



Least Squares Example

Estimate the least squares coefficients for our example data:

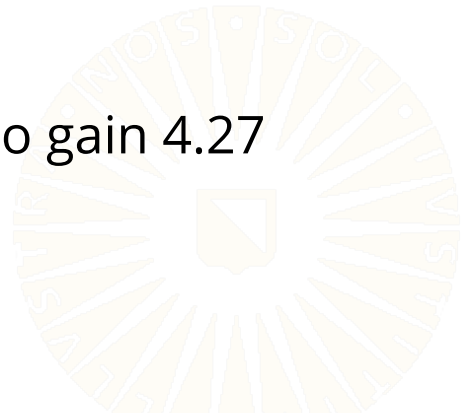
```
#data(Cars93)
out1 <- lm(Horsepower ~ Price, data = Cars93)
coef(out1)

(Intercept)       Price
  60.447578    4.273796
```

The estimated intercept is $\hat{\beta}_0 = 60.45$.

- A free car is expected to have 60.45 horsepower.

The estimated slope is: $\hat{\beta}_1 = 4.27$.

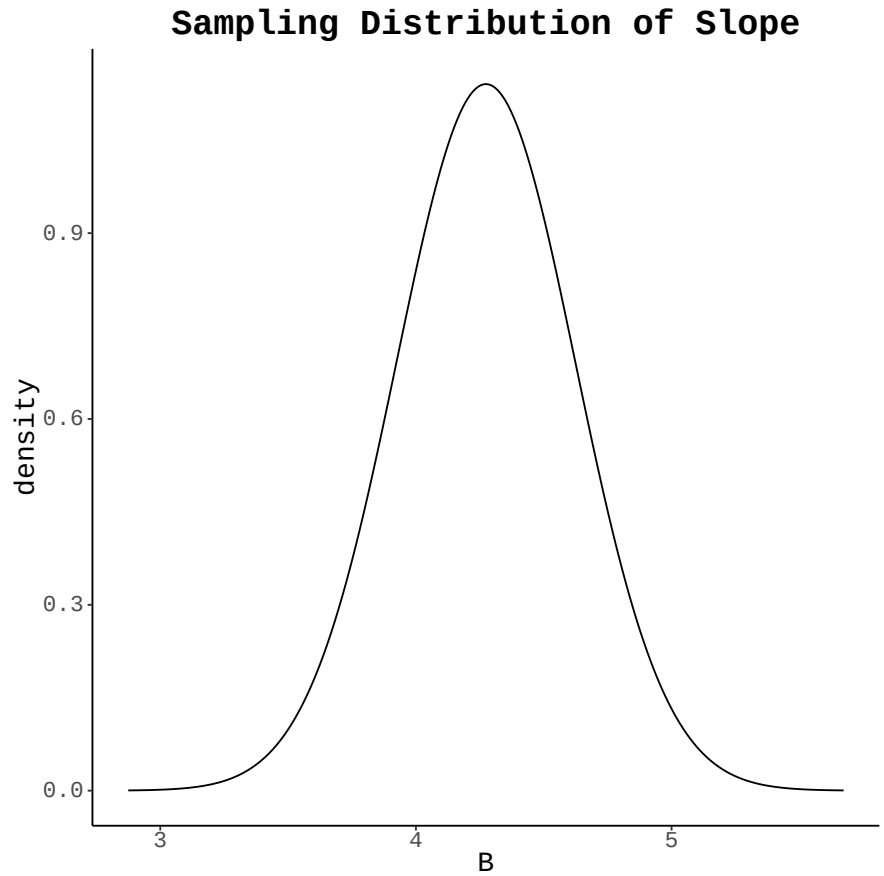
- For every additional $1000 in price, a car is expected to gain 4.27 horsepower.



Sampling Distribution

Sampling distribution = Probability distribution of a parameter.

- The *population* is defined by an infinite sequence of repeated estimations.
  - The sampling distribution quantifies the possible values of the statistic over infinite repeated sampling.
- The area of a region under the curve represents the probability of observing a *statistic* within the corresponding interval.



Intuition: `http://onlinestatbook.com/stat_sim/sampling_dist/`

Test Statistics

To "test" a slope coefficient, $\hat{\beta}$, we need a point of comparison.

- The *null-hypothesized* value of the slope, $H_0 : \beta = \tilde{\beta}$.

Our hypothesis test is actually a test for the size of the difference: $\hat{\beta} - \tilde{\beta}$

- We define a *test statistic*, t, to quantify the size of this difference accounting for the precision with which we've estimated $\hat{\beta}$.

We can construct the test statistic for $\hat{\beta}$ as follows:

$$t = \frac{\hat{\beta} - \tilde{\beta}}{SE\left(\hat{\beta}\right)} \overset{\tilde{\beta}=0}{\Rightarrow} t = \frac{\hat{\beta} - 0}{SE\left(\hat{\beta}\right)} = \frac{\hat{\beta}}{SE\left(\hat{\beta}\right)}$$

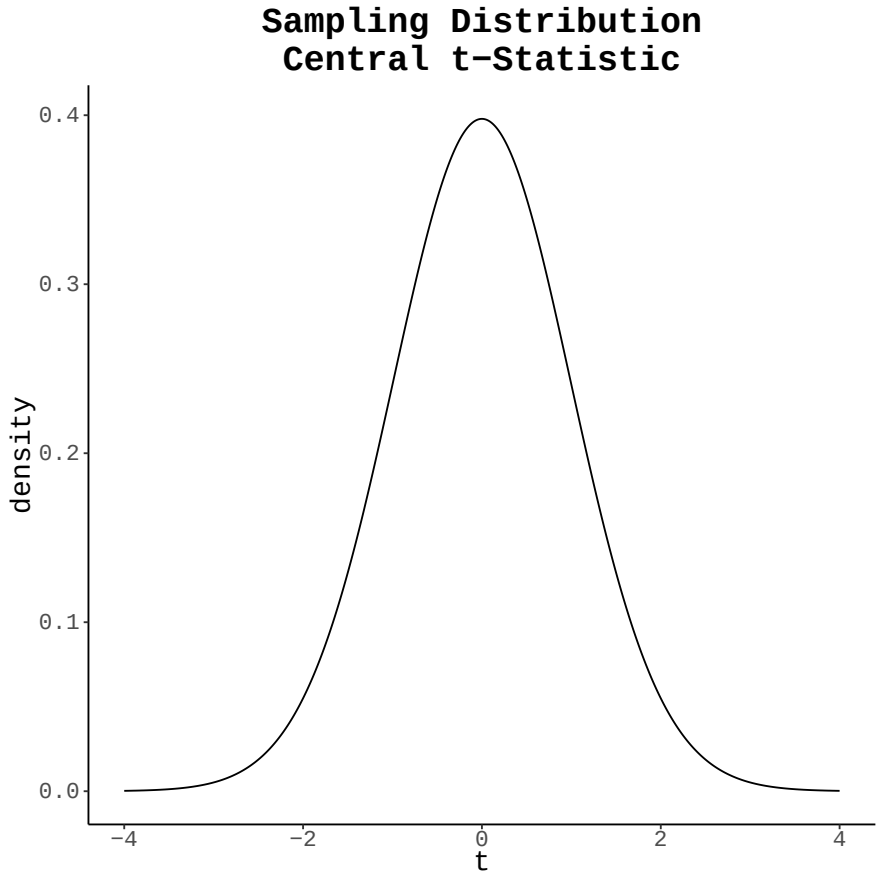
For the slope in our example, we get a test statistic of:

$$t = \frac{\hat{\beta}_1}{SE\left(\hat{\beta}_1\right)} = \frac{4.27}{0.35} = 12.2$$

Sampling Distribution of Test Statistic

The t-statistic also has a sampling distribution.

- Quantifies the possible values we could get if we repeatedly drew samples, of the same size, from the same population and re-computed a t-statistic each time.
- The distribution under the null hypothesis assumes a population wherein $\hat{\beta} = \tilde{\beta}$, and, consequently, $t = 0$.

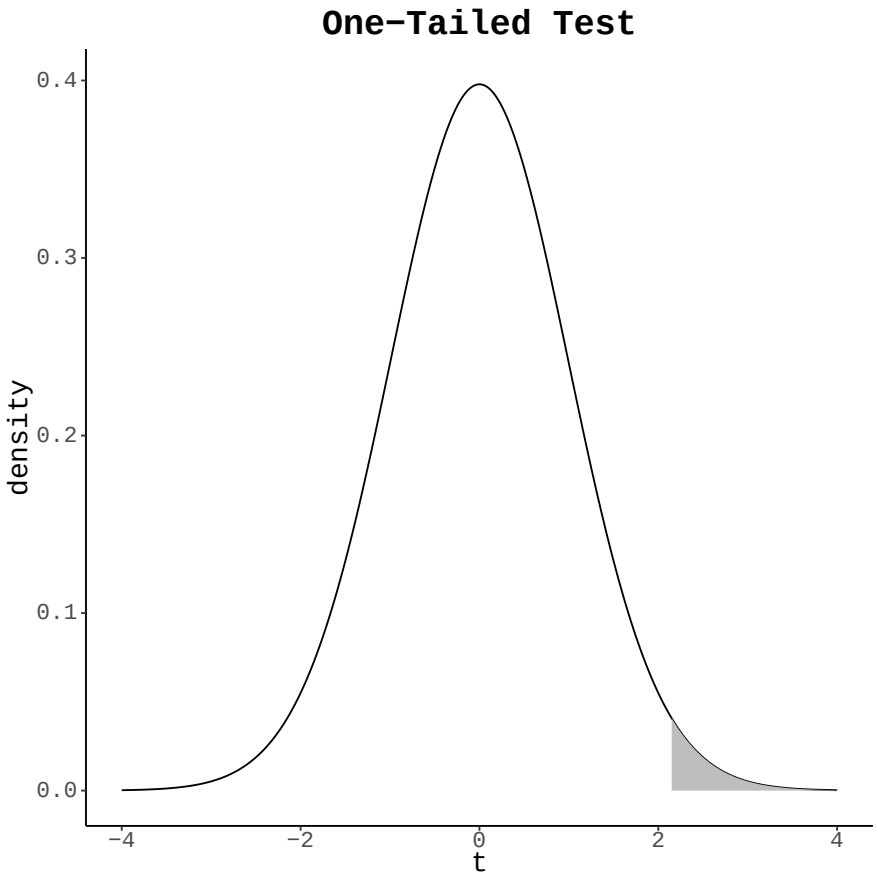


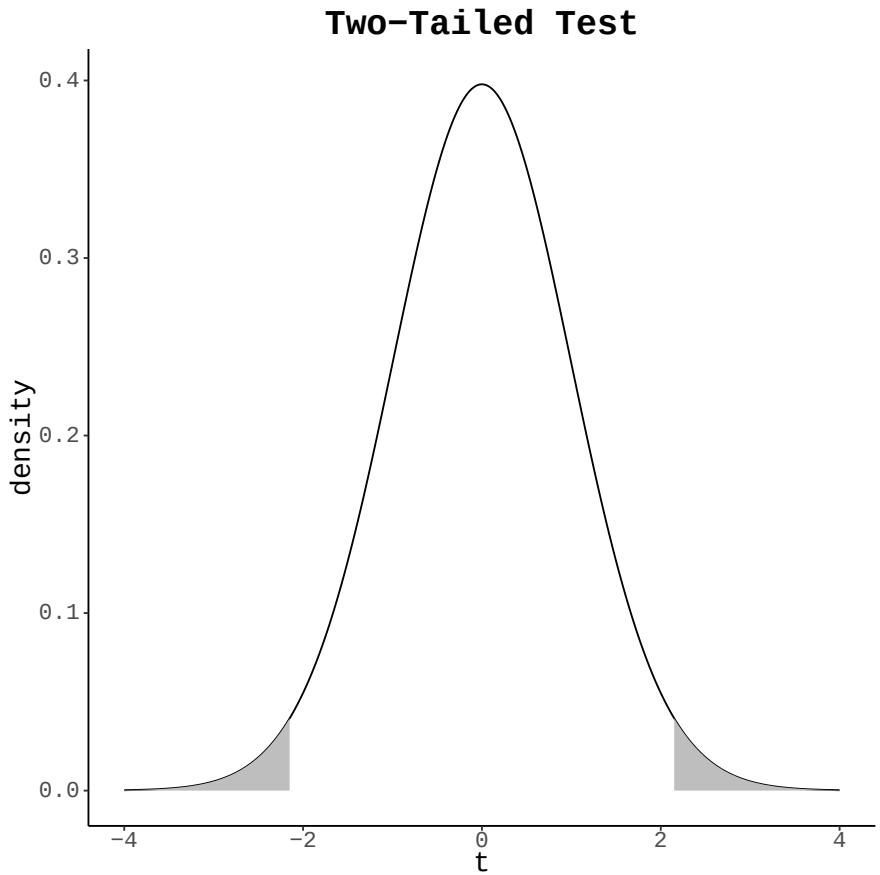
P-Values

Once we compute our estimated test statistic, $\hat{t}$, we compare it to the appropriate null-hypothesized sampling distribution.

- By calculating the area in the null distribution that exceeds our estimated test statistic, we can compute the probability of observing the given test statistic, or one more extreme, if the null hypothesis were true.
    - In other words, we can compute the probability of having sampled the data we observed, or more unusual data, from a population wherein there is no true difference between $\hat{\beta}$ and $\tilde{\beta}$.
- This value is the infamous *p-value*.

P-Values


One−Tailed Test
density
0.4
0.3
0.2
0.1
0.0
−4
−2
0
2
4
t


Two−Tailed Test
density
0.4
0.3
0.2
0.1
0.0
−4
−2
0
2
4
t

Interpreting P-Values

Consider the one-tailed test for our estimated test-statistic of $\hat{t} = 2.15$ that produces a p-value of $p = 0.017$.

- We *cannot* say that there is a 0.017 probability that the true mean difference is greater than zero.
- We *cannot* say that there is a 0.017 probability that the alternative hypothesis is true.
- We *cannot* say that there is a 0.017 probability that the null hypothesis is false.
- We *cannot* say that there is a 0.017 probability that the observed result is due to chance alone.
- We *cannot* say that there is a 0.017 probability of replicating the observed effect in future studies.

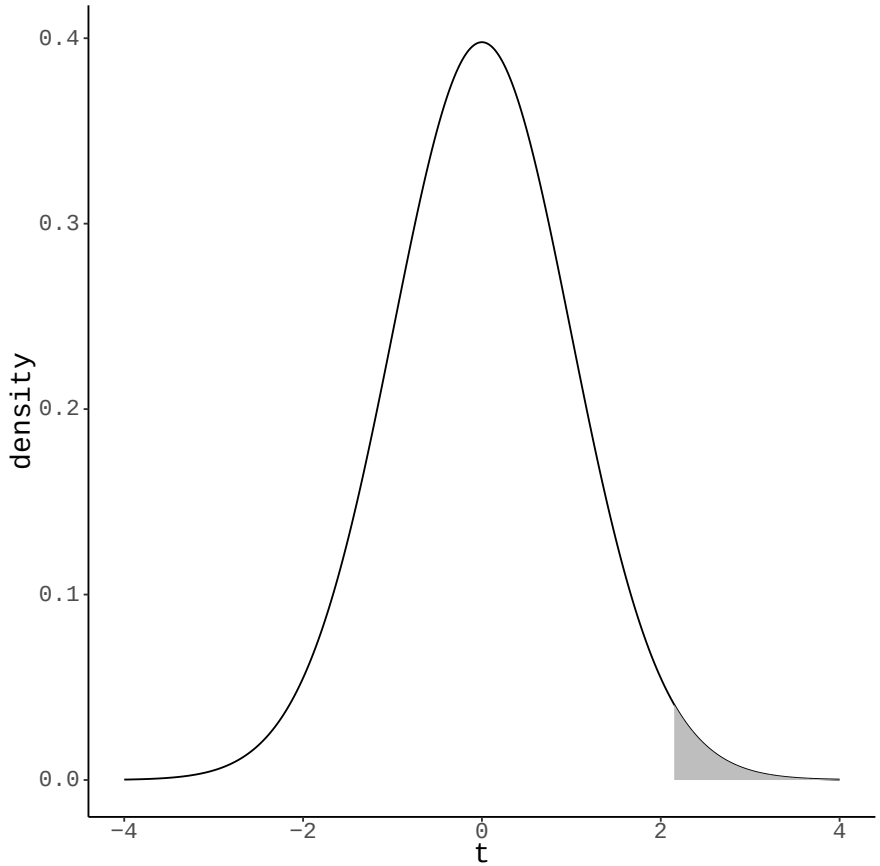
Interpreting P-Values

The p-value tells us $P(t \geq \hat{t}|H_0)$

- What we really want to know is $P(H_0|t \geq \hat{t})$.

All that we *can* say is that there is a 0.017 probability of observing a test statistic at least as large as $\hat{t}$, if the null hypothesis is true.

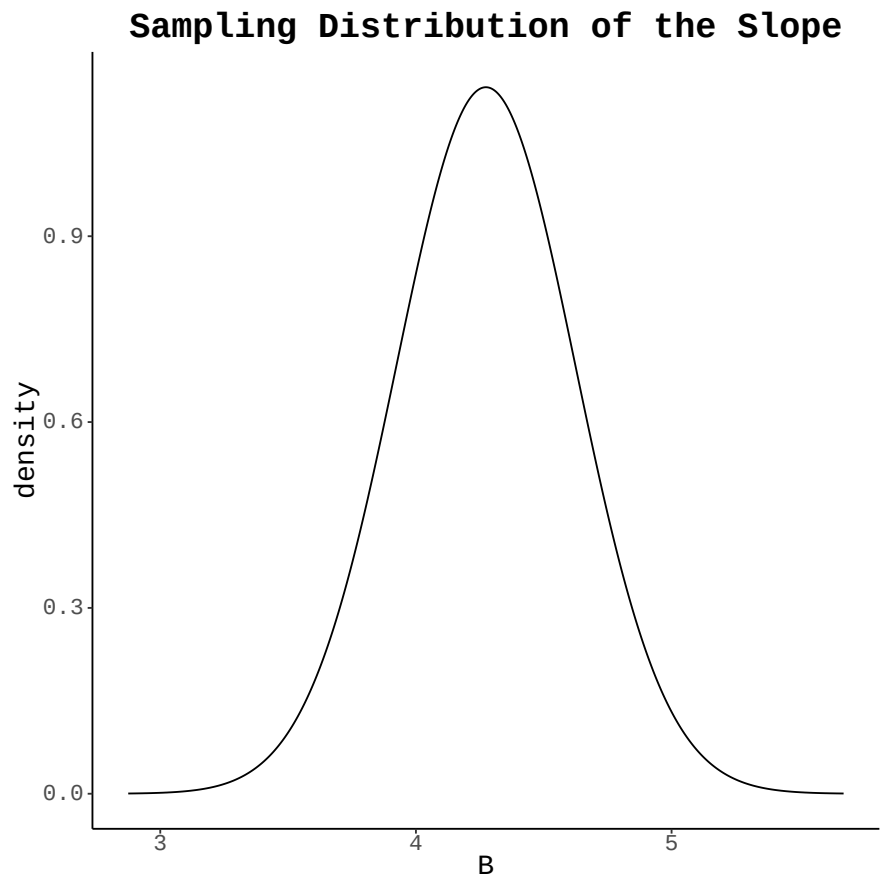
- Our test uses the same logic as *proof by contradiction*.



Confidence Intervals

A sampling distribution quantifies the possible values of the statistic.

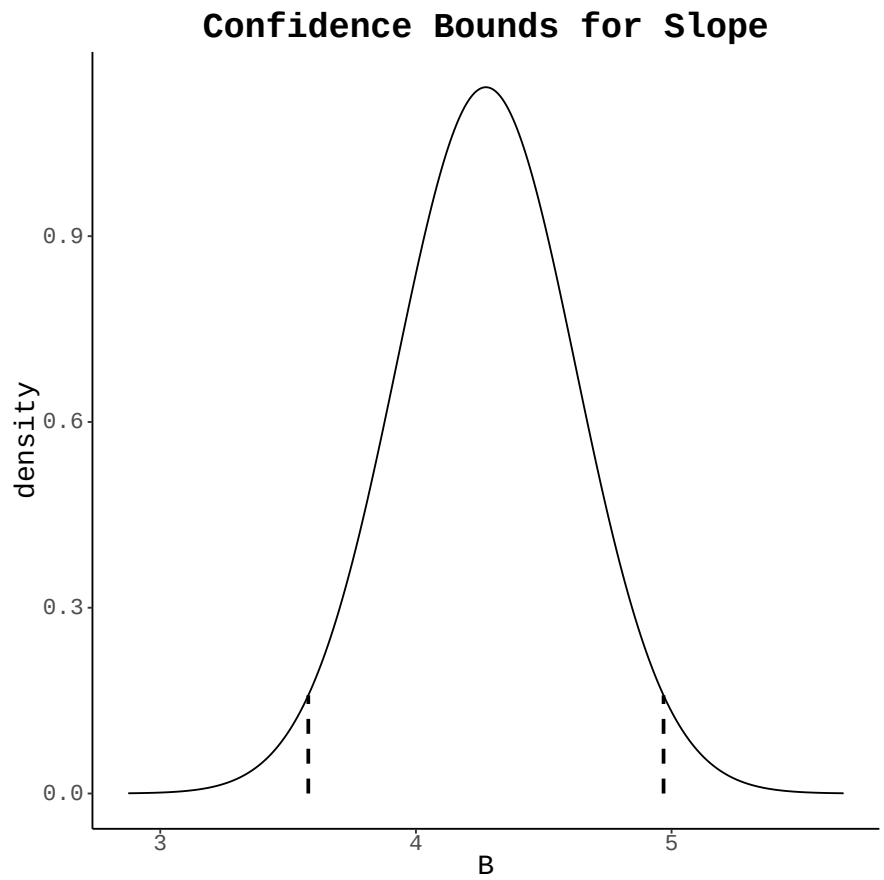
- We can use this distribution to estimate a *plausible range* for the population parameter.



Confidence Intervals

A sampling distribution quantifies the possible values of the statistic.

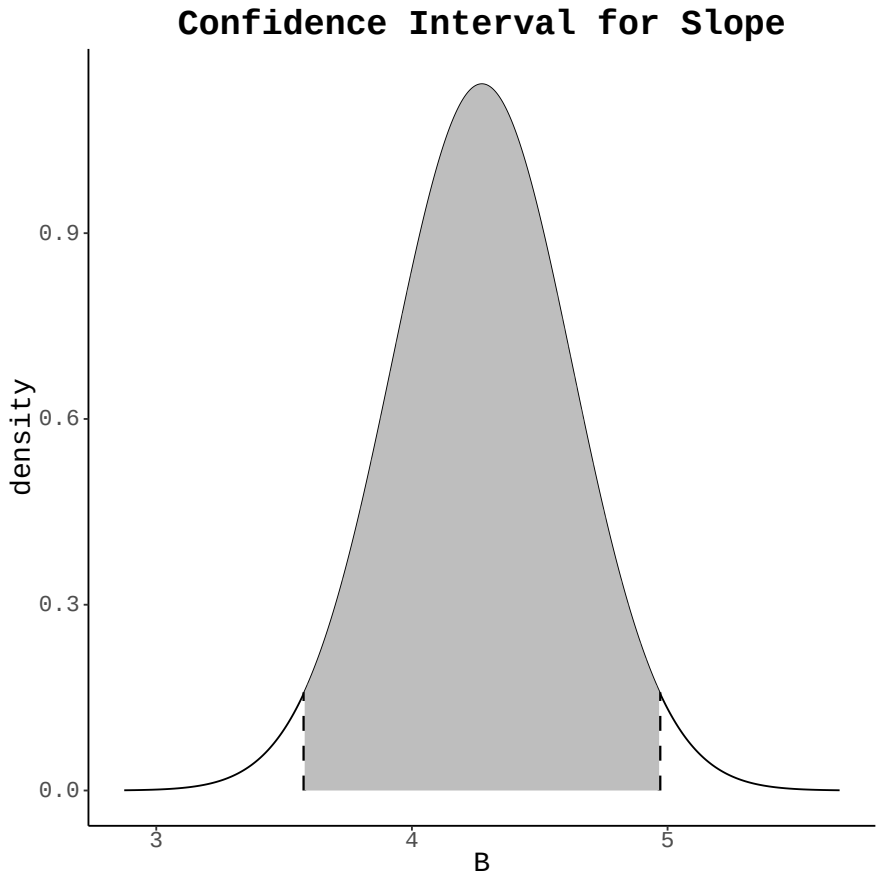
- We can use this distribution to estimate a *plausible range* for the population parameter.
  1. Exclude the tails of the distribution.



Confidence Intervals

A sampling distribution quantifies the possible values of the statistic.

- We can use this distribution to estimate a *plausible range* for the population parameter.
  1. Exclude the tails of the distribution.
  2. The remaining values represent a good guess for plausible population values of the parameter.
- This range is known as the *confidence interval*.



Confidence Intervals

We can construct confidence intervals by:

$$CI = \hat{\beta} \pm t_{crit} \times SE\left(\hat{\beta}\right)$$

For our example slope, we get a 95% CI of:

$$CI_{95} = 4.27 \pm 1.99 \times 0.35 = [3.57; 4.97]$$

Which suggests that we can be 95% certain that the true value of β_1 is somewhere between 3.57 and 4.97.

- We are *95% certain* in the sense that if we repeat this analysis an infinite number of times, 95% of the CIs that we calculate will surround the true value of β_1.

Interpreting Confidence Intervals

Say we estimate a regression slope of $\hat{\beta}_1 = 0.5$ with an associated 95% confidence interval of $CI = [0.25; 0.75]$.

Interpreting Confidence Intervals

Say we estimate a regression slope of $\hat{\beta}_1 = 0.5$ with an associated 95% confidence interval of $CI = [0.25; 0.75]$.

- We *cannot* say that there is 95% chance that the true value of β_1 is between 0.25 and 0.75.
- We *cannot* say that the true value of β_1 is between 0.25 and 0.75, with probability 0.95.

Interpreting Confidence Intervals

Say we estimate a regression slope of $\hat{\beta}_1 = 0.5$ with an associated 95% confidence interval of $CI = [0.25; 0.75]$.

- We *cannot* say that there is 95% chance that the true value of β_1 is between 0.25 and 0.75.
- We *cannot* say that the true value of β_1 is between 0.25 and 0.75, with probability 0.95.

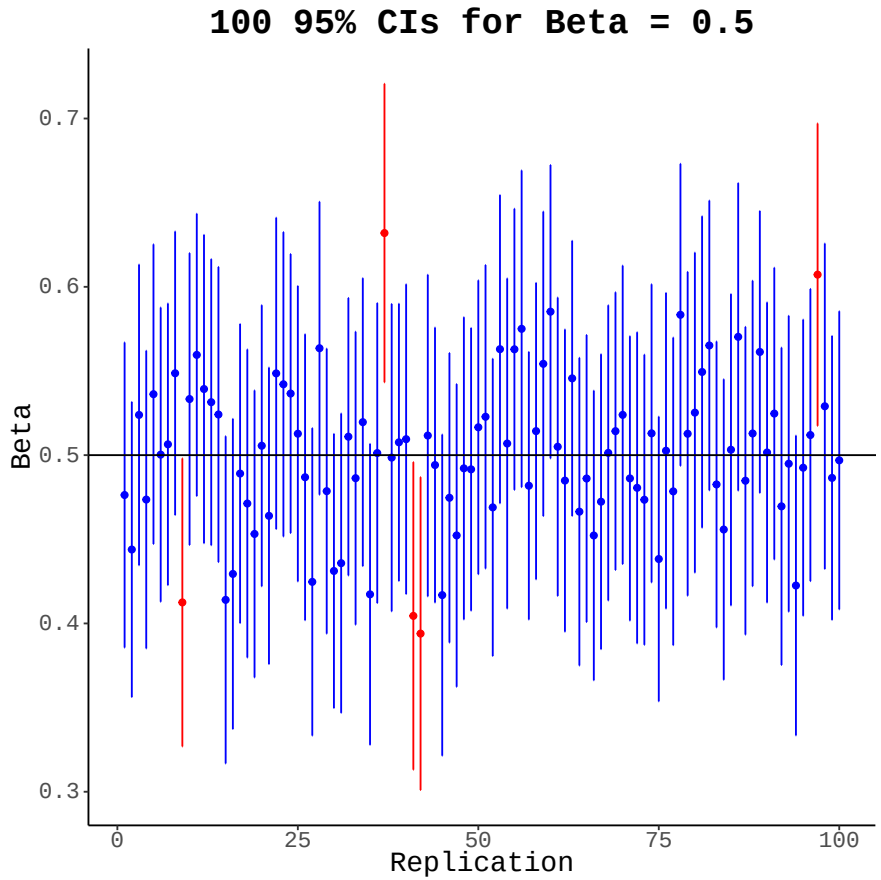
The true value of β_1 is fixed; it's a single quantity.

- β_1 is either in our estimated interval or it is not; there is no uncertainty.
- The probability that β_1 is within our estimated interval is either exactly 1 or exactly 0.

Interpreting Confidence Intervals

We don't talk about 95% probabilities when interpreting CIs; instead, we talk about 95% confidence.

- If we collected a new sample—of the same size—re-estimated our model, and re-computed the 95% CI for $\hat{\beta}_1$, we would get a different interval.
- Repeating this process an infinite number of times would give us a distribution of CIs.
- 95% of those CIs would surround the true value of β_1.



Model-Based Prediction

In the social and behavioral sciences, regression modeling is often focused on inference about estimated model parameters.

- The association between the price of a car and its power.
- We model the system and scrutinize $\hat{\beta}_1$ to make inferences about the association between price and power.

Model-Based Prediction

In the social and behavioral sciences, regression modeling is often focused on inference about estimated model parameters.

- The association between the price of a car and its power.
- We model the system and scrutinize $\hat{\beta}_1$ to make inferences about the association between price and power.

In data science applications, we're often more interested in predicting the outcome for new observations.

- After we estimate $\hat{\beta}_0$ and $\hat{\beta}_1$, we can plug in new predictor data and get a predicted outcome value for any new case.
- In our example, these predictions represent the projected horsepower ratings of cars with prices given by the new X_{price} values.

Inference vs. Prediction

When doing statistical inference, we focus on how certain variables relate to the outcome.

- Do men have higher job-satisfaction than women?
- Does increased spending on advertising correlate with more sales?
- Is there a relationship between the number of liquor stores in a neighborhood and the amount of crime?

Inference vs. Prediction

When doing statistical inference, we focus on how certain variables relate to the outcome.

- Do men have higher job-satisfaction than women?
- Does increased spending on advertising correlate with more sales?
- Is there a relationship between the number of liquor stores in a neighborhood and the amount of crime?

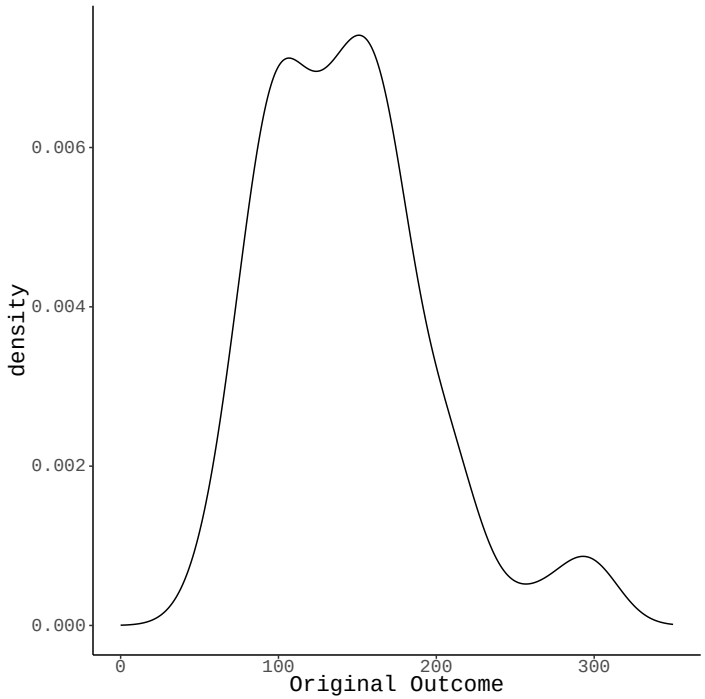
When doing prediction (or classification), we want to build a tool that can accurately guess future values.

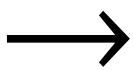
- Will it rain tomorrow?
- How much will a company earn from investing in a certain research profile?
- What is a patient's risk of heart disease based on their medical history and test results?

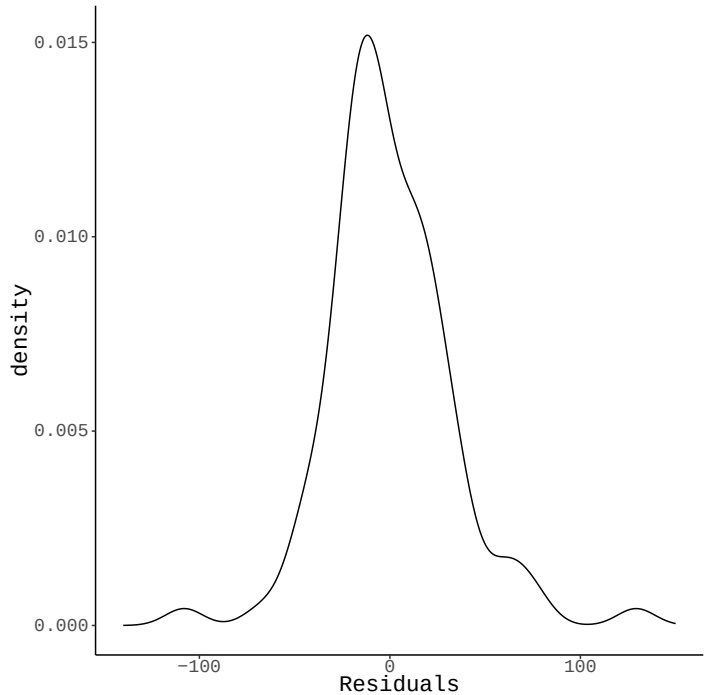
Model Fit

We may also want to know how well our model explains the outcome.

- Our model explains some proportion of the outcome's variability.
- The residual variance $\hat{\sigma}^2 = \mathbf{Var}(\hat{\varepsilon})$ will be less than $\mathbf{Var}(Y)$.



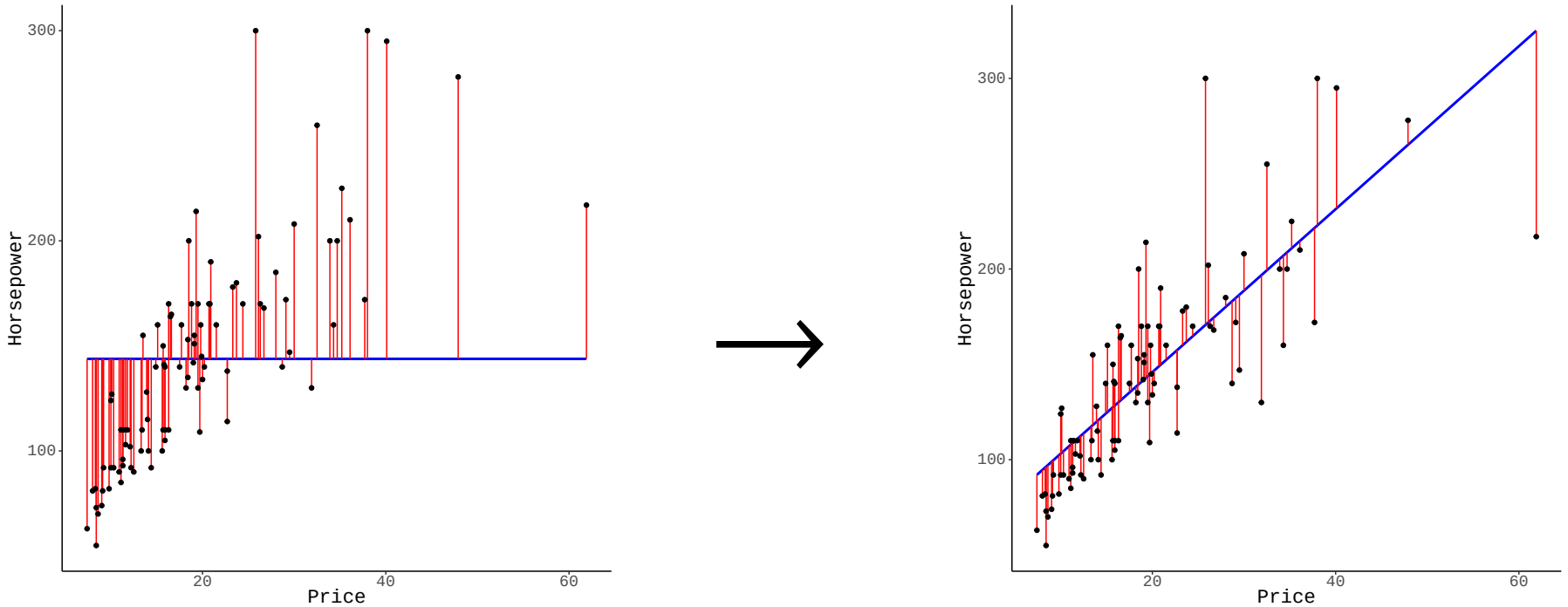




Model Fit

We may also want to know how well our model explains the outcome.

- Our model explains some proportion of the outcome's variability.
- The residual variance $\hat{\sigma}^2 = \mathbf{Var}(\hat{\varepsilon})$ will be less than $\mathbf{Var}(Y)$.



Model Fit

We quantify the proportion of the outcome's variance that is explained by our model using the R^2 statistic:

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

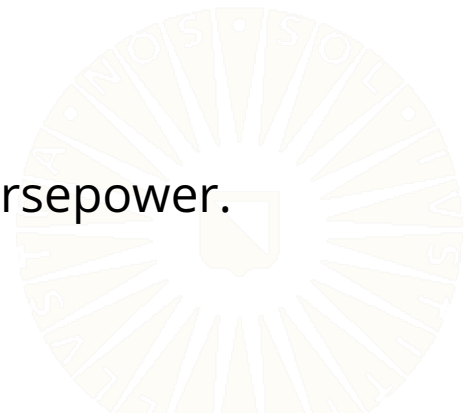
where

$$TSS = \sum_{n=1}^{N} \left(Y_n - \bar{Y}\right)^2 = \text{Var}(Y) \times (N - 1)$$

For our example problem, we get:

$$R^2 = 1 - \frac{95573}{252363} \approx 0.62$$

Indicating that car price explains 62% of the variability in horsepower.



Model Fit for Prediction

When assessing predictive performance, we will most often use the *mean squared error* (MSE) as our criterion.

$$
\begin{aligned}
MSE &= \frac{1}{N}\sum_{n=1}^{N}\left(Y_n - \hat{Y}_n\right)^2 \\
&= \frac{1}{N}\sum_{n=1}^{N}\left(Y_n - \hat{\beta}_0 - \sum_{p=1}^{P}\hat{\beta}_p X_{np}\right)^2 \\
&= \frac{RSS}{N}
\end{aligned}
$$

For our example problem, we get:

$$
MSE = \frac{95573}{93} \approx 1027.67
$$

Interpreting MSE

The MSE quantifies the average squared prediction error.

- Taking the square root improves interpretation.

$$RMSE = \sqrt{MSE}$$

The RMSE estimates the magnitude of the expected prediction error.

- For our example problem, we get:

$$RMSE = \sqrt{\frac{95573}{93}} \approx 32.06$$

- When using price as the only predictor of horsepower, we expect prediction errors with magnitudes of 32.06 horsepower.

Information Criteria

We can use *information criteria* to quickly compare *non-nested* models while accounting for model complexity.

- Akaike's Information Criterion (AIC)

$$AIC = 2K - 2\hat{\ell}(\theta|X)$$

- Bayesian Information Criterion (BIC)

$$BIC = K\ln(N) - 2\hat{\ell}(\theta|X)$$

Information Criteria

We can use *information criteria* to quickly compare *non-nested* models while accounting for model complexity.

- Akaike's Information Criterion (AIC)

$$AIC = 2K - 2\hat{\ell}(\theta|X)$$

- Bayesian Information Criterion (BIC)

$$BIC = K\ln(N) - 2\hat{\ell}(\theta|X)$$

Information criteria balance two competing forces.

- The optimized loglikelihood quantifies fit to the data.
- The penalty term corrects for model complexity.

Information Criteria

For our example, we get the following estimates of AIC and BIC:

$$\begin{aligned} AIC &= 2(3) - 2(-454.44) \\ &= 914.88 \end{aligned}$$

$$\begin{aligned} BIC &= 3\ln(93) - 2(-454.44) \\ &= 922.48 \end{aligned}$$

To compute the AIC/BIC from a fitted `lm()` object in R:

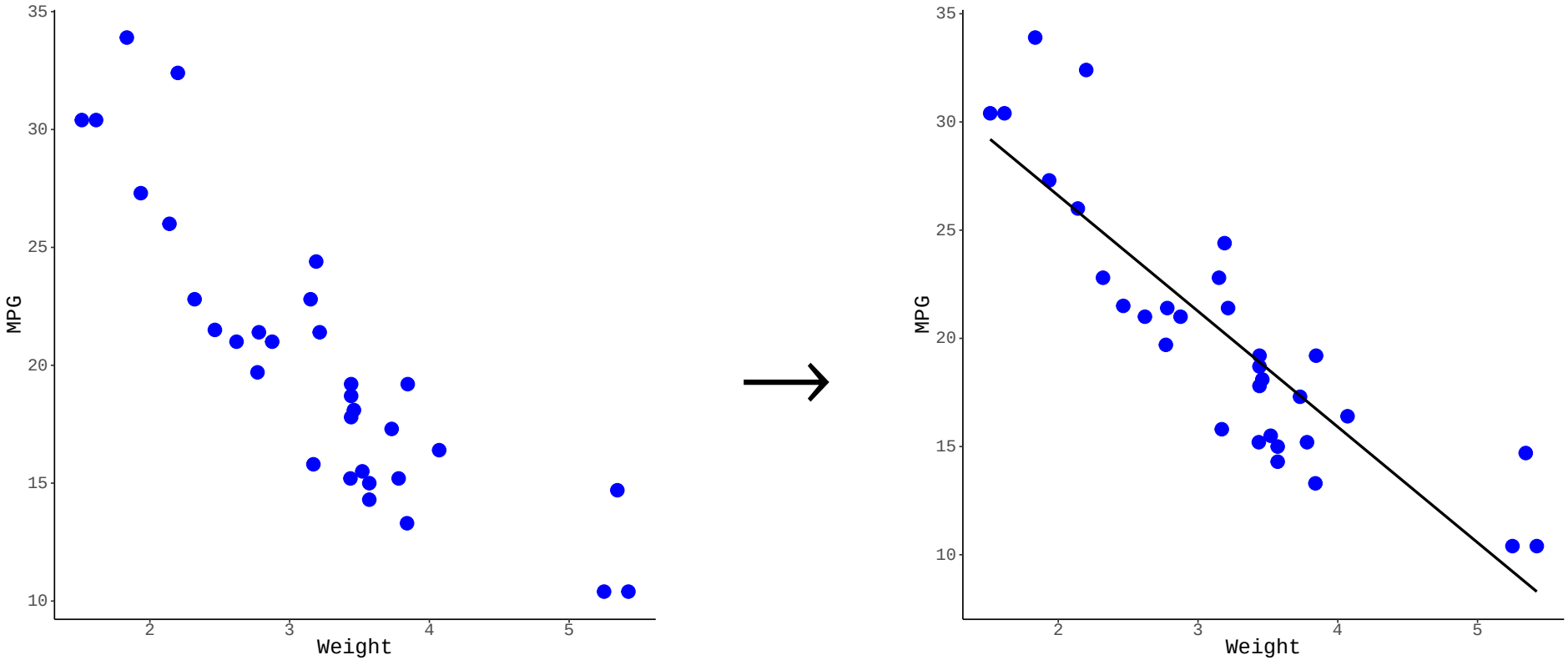
```
AIC(out1)

[1] 914.8821

BIC(out1)

[1] 922.4799
```

Multiple Linear Regression

Graphical Representations

A regression of two variables can be represented on a 2D scatterplot.

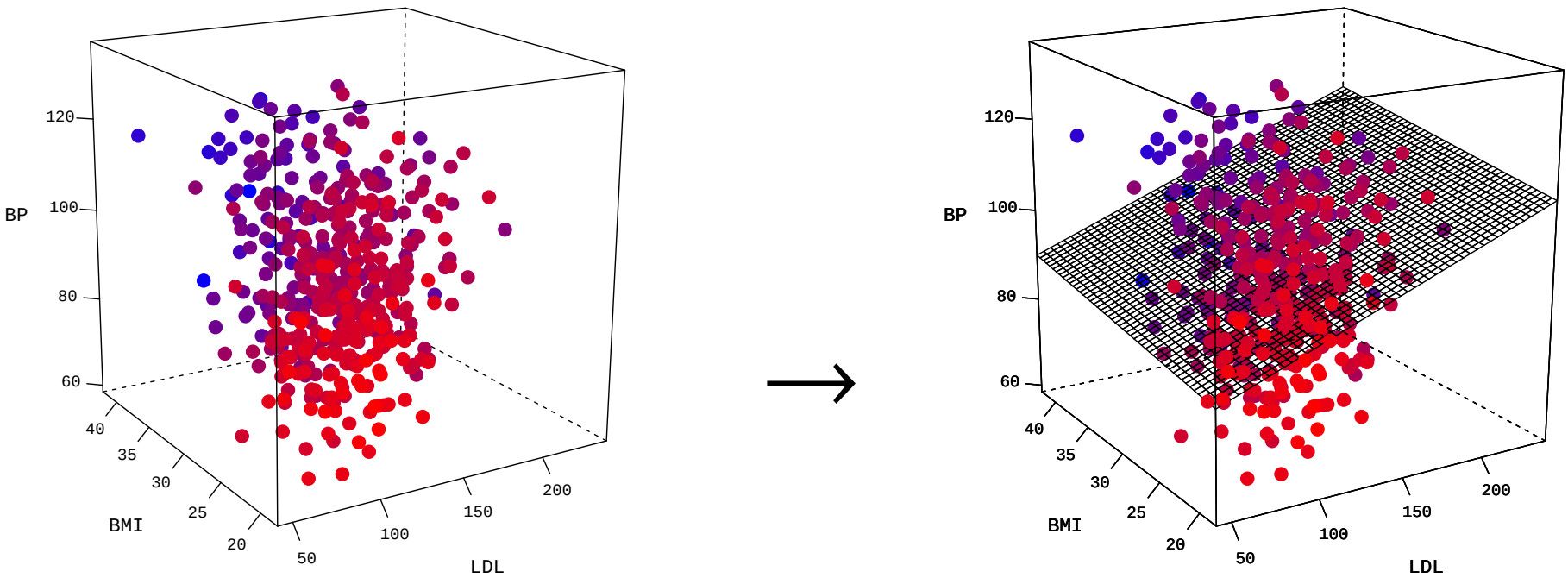
- Simple linear regression implies a 1D line in 2D space.



Graphical Representations

Adding an additional predictor leads to a 3D point cloud.

- A regression model with two IVs implies a 2D plane in 3D space.



Partial Effects

In MLR, we want to examine the *partial effects* of the predictors.

- What is the effect of a predictor after controlling for some other set of variables?

This approach is crucial to controlling confounds and adequately modeling real-world phenomena.

Example

```
## Read in the 'diabetes' dataset:
dDat <- readRDS("../data/diabetes.rds")

## Simple regression with which we're familiar:
out1 <- lm(bp ~ age, data = dDat)
```

Asking: What is the effect of age on average blood pressure?

Example

```
partSummary(out1, -1)

Residuals:
    Min      1Q  Median      3Q     Max
-31.188  -8.897  -1.209   8.612  39.952

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 77.47605    2.38132  32.535  < 2e-16
age          0.35391    0.04739   7.469 4.39e-13

Residual standard error: 13.04 on 440 degrees of freedom
Multiple R-squared:  0.1125,Adjusted R-squared:  0.1105
F-statistic: 55.78 on 1 and 440 DF,  p-value: 4.393e-13
```

Example

```
## Add in another predictor:
out2 <- lm(bp ~ age + bmi, data = dDat)
```

Asking: What is the effect of BMI on average blood pressure, *after controlling for age?*

- We're partialing age out of the effect of BMI on blood pressure.

Example

```
partSummary(out2, -1)

Residuals:
    Min      1Q  Median      3Q     Max
-29.287  -8.198  -0.178   8.413  41.026

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 52.24654    3.83168  13.635  < 2e-16
age          0.28651    0.04504   6.362 5.02e-10
bmi          1.08053    0.13363   8.086 6.06e-15

Residual standard error: 12.18 on 439 degrees of freedom
Multiple R-squared:  0.2276,Adjusted R-squared:  0.224
F-statistic: 64.66 on 2 and 439 DF,  p-value: < 2.2e-16
```

Interpretation

- The expected average blood pressure for an unborn patient with a negligible extent is 52.25.
- For each year older, average blood pressure is expected to increase by 0.29 points, after controlling for BMI.
- For each additional point of BMI, average blood pressure is expected to increase by 1.08 points, after controlling for age.

Multiple R^2

How much variation in blood pressure is explained by the two models?

- Check the R^2 values.

```
## Extract R^2 values:
r2.1 <- summary(out1)$r.squared
r2.2 <- summary(out2)$r.squared

r2.1

[1] 0.1125117

r2.2

[1] 0.2275606
```

F-Statistic

How do we know if the R^2 values are significantly greater than zero?

- We use the F-statistic to test $H_0 : R^2 = 0$ vs. $H_1 : R^2 > 0$.

```
f1 <- summary(out1)$fstatistic
f1

    value     numdf     dendf
 55.78116   1.00000 440.00000

pf(q = f1[1], df1 = f1[2], df2 = f1[3], lower.tail = FALSE)

       value
4.392569e-13
```

F-Statistic

```
f2 <- summary(out2)$fstatistic
f2

   value    numdf    dendf
 64.6647   2.0000 439.0000

pf(f2[1], f2[2], f2[3], lower.tail = FALSE)

       value
2.433518e-25
```

Comparing Models

How do we quantify the additional variation explained by BMI, above and beyond age?

- Compute the ΔR^2

```
## Compute change in R^2:
r2.2 - r2.1

[1] 0.115049
```

Significance Testing

How do we know if ΔR^2 represents a significantly greater degree of explained variation?

- Use an F-test for $H_0 : \Delta R^2 = 0$ vs. $H_1 : \Delta R^2 > 0$

```
## Is that increase significantly greater than zero?
anova(out1, out2)

Analysis of Variance Table

Model 1: bp ~ age
Model 2: bp ~ age + bmi
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1    440 74873
2    439 65167  1    9706.1 65.386 6.057e-15 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comparing Models

We can also compare models based on their prediction errors.

- For OLS regression, we usually compare MSE values.

```
mse1 <- MSE(y_pred = predict(out1), y_true = dDat$bp)
mse2 <- MSE(y_pred = predict(out2), y_true = dDat$bp)

mse1

[1] 169.3963

mse2

[1] 147.4367
```

In this case, the MSE for the model with BMI included is smaller.

- We should prefer the the larger model.

Comparing Models

Finally, we can compare models based on information criteria.

```
AIC(out1, out2)

     df       AIC
out1  3 3528.792
out2  4 3469.424

BIC(out1, out2)

     df       BIC
out1  3 3541.066
out2  4 3485.789
```

In this case, both the AIC and the BIC for the model with BMI included are smaller.

- We should prefer the the larger model.

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

Example Dummy Codes

Now, a slightly more complex example:

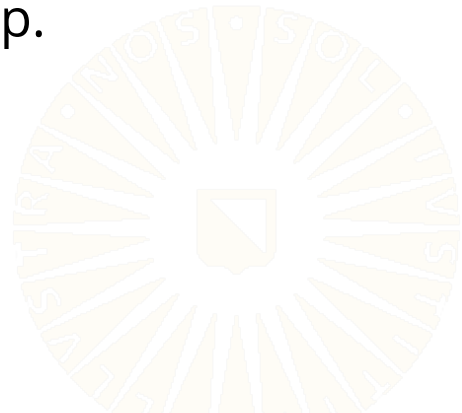
	drink	juice	tea
1	coffee	0	0
2	tea	0	1
3	coffee	0	0
4	coffee	0	0
5	coffee	0	0
6	coffee	0	0
7	juice	1	0
8	coffee	0	0
9	coffee	0	0
10	coffee	0	0

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

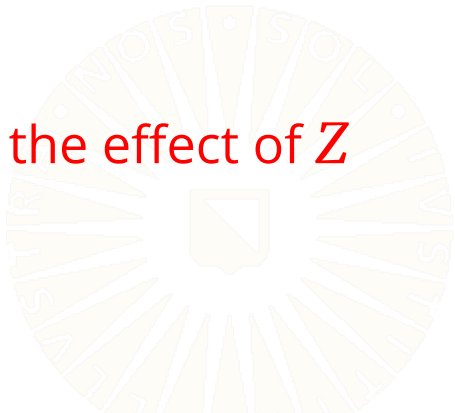
In additive MLR, we might have the following equation:

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

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

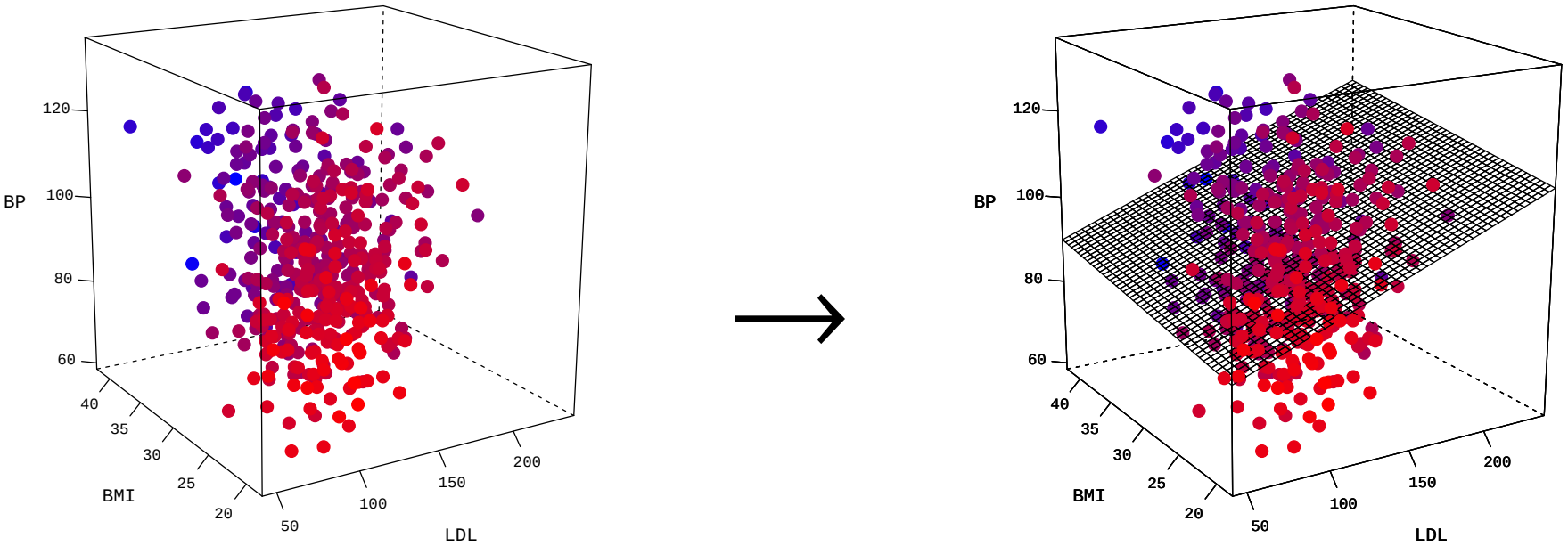
When X and Z are independent predictors, the following are true:

- 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.



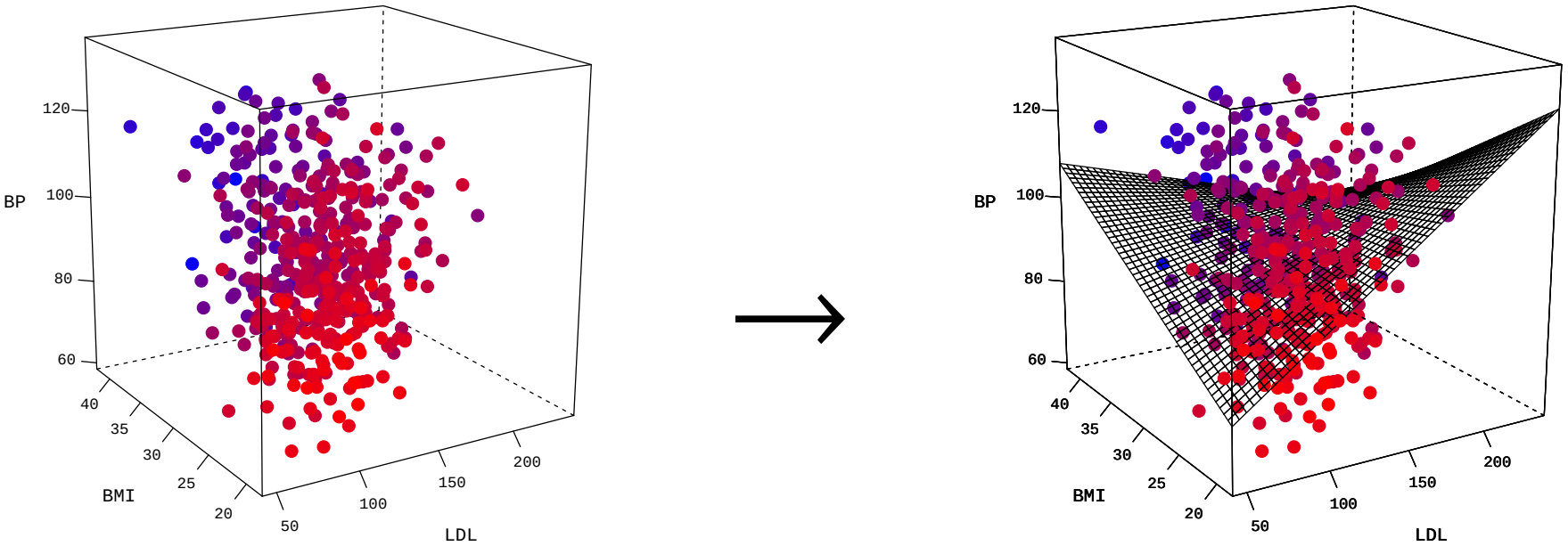
Additive Regression

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

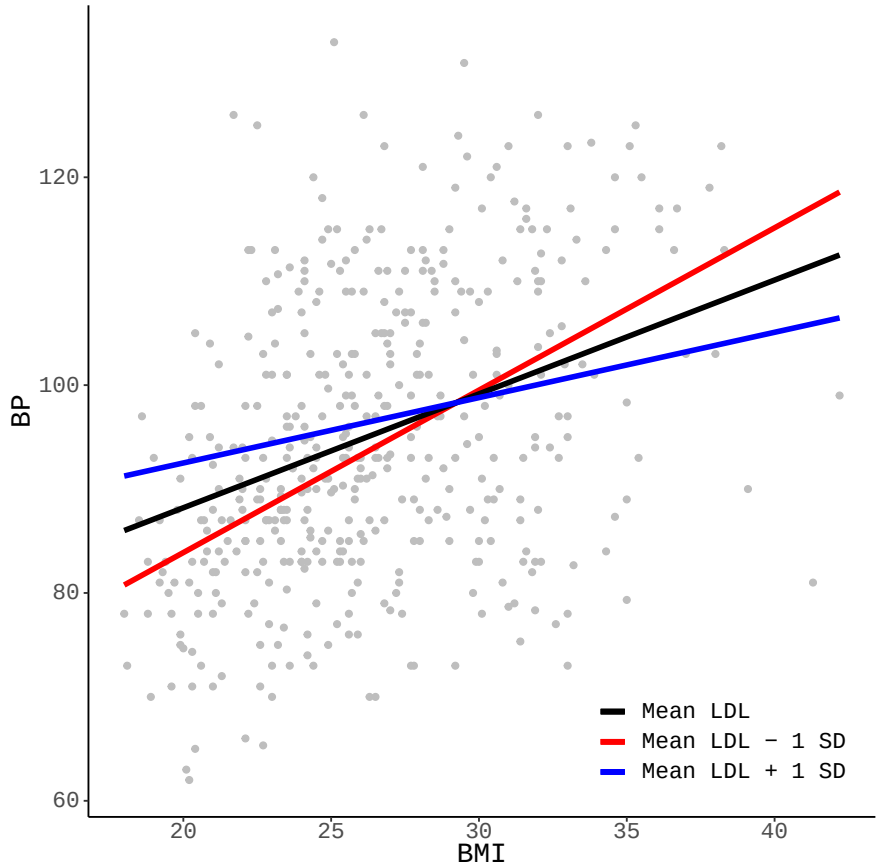
```
## Additive Model:
out1 <- lm(bp ~ bmi + ldl, data = dDat)
partSummary(out1, -c(1, 2))

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"))

Error in paste0(dataDir, "social_support.rds"):  object 'dataDir' not found

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

Error in is.data.frame(data):  object 'socSup' not found

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
```

Example

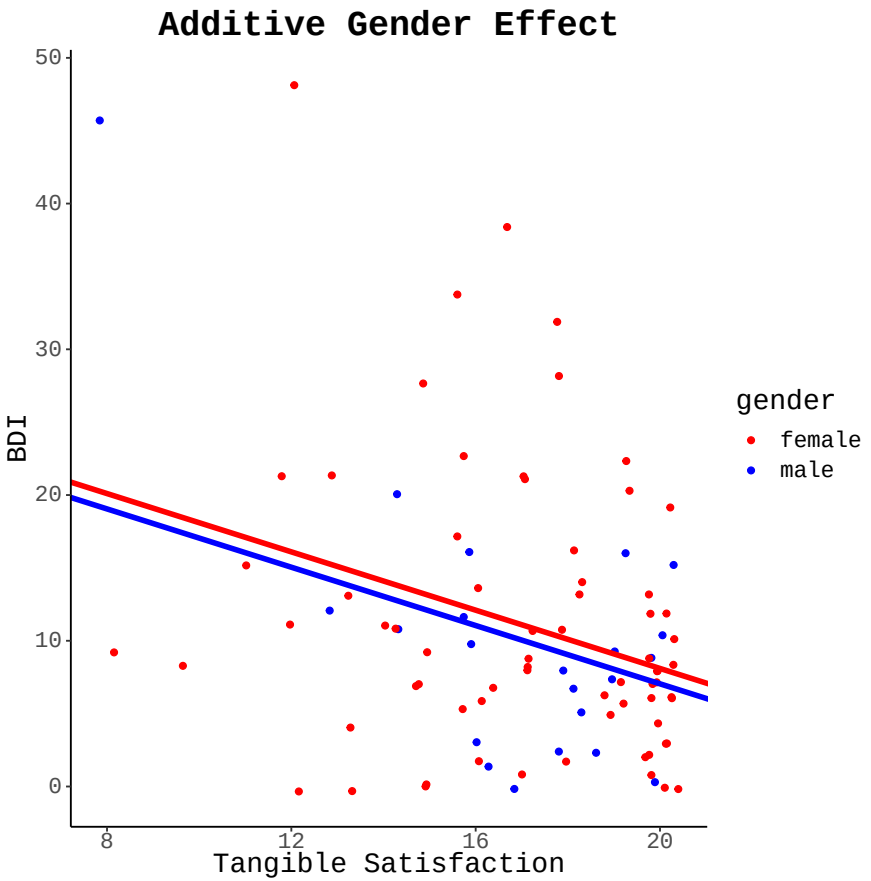
```
## Estimate the interaction:
out4 <- lm(bdi ~ tanSat * sex, data = socSup)

Error in is.data.frame(data):  object 'socSup' not found

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
```

Visualizing Categorical Moderation

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

$$\hat{Y}_{BDI} = 21.72 - 5.84X_{tsat} + -0.26Z_{male}$$
$$10.52X_{tsat}Z_{male}$$

```
Error in relevel(socSup$sex, ref =
"male"):  object 'socSup' not found
Error in is.data.frame(data):  object
'socSup' not found
Error in coef(out5):  object 'out5'
not found
```



Model-Building Example

Let's walk through an example of the model-building process.

- We'll take $Y_{bp} = \beta_0 + \beta_1 X_{age.30} + \varepsilon$ as our baseline model.
- Next, we simultaneously add predictors of LDL and HDL cholesterol.

```
diabetes <- readRDS("../data/diabetes.rds")

## Center predictor variables:
diabetes <- mutate(diabetes,
                   ldl100 = ldl - 100,
                   hdl60 = hdl - 60,
                   age30 = age - 30)
```

```
Error in mutate(diabetes, ldl100 = ldl - 100, hdl60 = hdl - 60, age30 = age
- :  could not find function "mutate"
```

```
## Baseline model:
out1 <- lm(bp ~ age30, data = diabetes)
```

```
Error in eval(predvars, data, env):  object 'age30' not found
```

```
## Simultaneously add two predictors:
out2 <- lm(bp ~ age30 + ldl100 + hdl60, data = diabetes)
```

Model-Building Example

```
partSummary(out1, -1)

Residuals:
    Min      1Q  Median      3Q     Max
-30.536  -8.508  -0.863   9.099  39.938

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
```

Model-Building Example

```
partSummary(out2, -1)

Residuals:
    Min      1Q  Median      3Q     Max
-32.877  -8.427  -0.966   8.931  39.368

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
```

Interpretations

- The expected average blood pressure for a 30 year old patient with LDL = 100 and HDL = 60 is 14.48.
- For each additional year older, average blood pressure is expected to increase by NA, after controlling for LDL and HDL levels.
- For each additional unit of LDL level, average blood pressure is expected to increase by NA, after controlling for age and HDL.
- For each additional unit of HDL level, average blood pressure is expected to decrease by NA, after controlling for age and LDL.

Model Comparison

```
## Compute change in R^2:
summary(out2)$r.squared - summary(out1)$r.squared

[1] 0.01975773

## Significance test for change in R^2:
anova(out1, out2)

Analysis of Variance Table

Model 1: bp ~ bmi + ldl
Model 2: bp ~ bmi * ldl
  Res.Df   RSS Df Sum of Sq      F   Pr(>F)
1    439 70562
2    438 68895  1    1666.9 10.597 0.001221 **
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model Comparison

```
(mse1 <- MSE(y_pred = predict(out1), y_true = diabetes$bp))

[1] 159.6421

(mse2 <- MSE(y_pred = predict(out2), y_true = diabetes$bp))

[1] 155.8709

AIC(out1, out2)

     df      AIC
out1  4 3504.579
out2  5 3496.012

BIC(out1, out2)

     df      BIC
out1  4 3520.944
out2  5 3516.469
```

Interpretations

- Age, LDL, and HDL explain a combined 18.3% of the variation in blood pressure.
  - This proportion of variation explained is significantly greater than zero.
- Adding LDL and HDL produces a model that explains 2% more variation in blood pressure than a model with age as the only predictor.
  - This increase in variation explained is significantly greater than zero.
- Adding LDL and HDL produces a model with lower prediction error (i.e., MSE = 155.87 vs. MSE = 159.64).
- Both the AIC and the BIC also suggest that adding LDL and HDL produces a better model.

Continue Building the Model

So far we've established that age, LDL, and HDL are all significant predictors of average blood pressure.

- We've also established that adding LDL and HDL, together, explain significantly more variation than age alone.

Next, we'll add BMI to see what additional explanatory role it can play above and beyond age and cholesterol.

```
## Center BMI:
diabetes <- mutate(diabetes, bmi25 = bmi - 25)

Error in mutate(diabetes, bmi25 = bmi - 25):  could not find function
"mutate"

## Now, add bmi:
out3 <- lm(bp ~ age30 + ldl100 + hdl60 + bmi25, data = diabetes)

Error in eval(predvars, data, env):  object 'age30' not found
```

Model-Building Example

```
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

BMI seems to have a pretty strong effect on average blood pressure, after controlling for age and cholesterol levels.

- After controlling for BMI, cholesterol levels no longer seem to be important predictors.
- Let's take a look at what happens to the cholesterol effects when we add BMI:

	LDL	HDL
Without BMI	0.449	-0.015
With BMI	11.320	11.320

Model Comparison

How much additional variability in blood pressure is explained by BMI above and beyond age and cholesterol levels?

```
r2.3 <- summary(out3)$r.squared
r2.3 - r2.2

[1] 0.01726571
```

Model Comparison

Is the additional 1.73% variation explained a significant increase?

```
anova(out2, out3)

Warning in anova.lmlist(object, ...):  models with response '"price"'
removed because response differs from model 1

Analysis of Variance Table

Response: bp
           Df Sum Sq Mean Sq F value    Pr(>F)
bmi         1  13190 13190.5 83.8586 < 2.2e-16 ***
ldl         1    613   612.9  3.8968  0.049005 *
bmi:ldl     1   1667  1666.9 10.5971  0.001221 **
Residuals 438  68895   157.3
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model Comparison

```
mse3 <- MSE(y_pred = predict(out3), y_true = diabetes$bp)
```

```
Warning in y_true - y_pred:  longer object length is not a multiple of
shorter object length
```

```
mse2
```

```
[1] 155.8709
```

```
mse3
```

```
[1] 5854.724
```

```
AIC(out2, out3)
```

```
Warning in AIC.default(out2, out3):  models are not all fitted to the same
number of observations
```

```
     df       AIC
out2  5 3496.0121
out3  4  671.9264
```

```
BIC(out2, out3)
```

Model Modification

Maybe cholesterol levels are not important features once we've accounted for BMI.

- Let's try a model including BMI but excluding cholesterol levels.

```
## Take out the cholesterol variables:
out4 <- lm(bp ~ age30 + bmi25, data = diabetes)

Error in eval(predvars, data, env):  object 'age30' not found
```

Model-Building Example

```
partSummary(out4, -1)

Residuals:
    Min      1Q  Median      3Q     Max
-17.336  -5.559  -2.218   4.082  29.664

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
```

Model Comparison

How much explained variation did we loose by removing the LDL and HDL variables?

```
r2.4 <- summary(out4)$r.squared
r2.3 - r2.4

[1] -0.08272339
```

Model Comparison

Is this -8.27% loss in explained variance significant?

```
anova(out4, out3)

Analysis of Variance Table

Model 1: price ~ mtOpt + front + rear
Model 2: price ~ front + rear
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     89 6151.6
2     90 6861.7 -1    -710.1 10.274 0.001874 **
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model Comparison

```
mse4 <- MSE(y_pred = predict(out4), y_true = diabetes$bp)
```

```
Warning in y_true - y_pred:  longer object length is not a multiple of
shorter object length
```

```
mse3
```

```
[1] 5854.724
```

```
mse4
```

```
[1] 5851.933
```

```
AIC(out3, out4)
```

```
     df      AIC
out3  4 671.9264
out4  5 663.7668
```

```
BIC(out3, out4)
```

```
     df      BIC
out3  4 682.0568
out4  5 676.4298
```

Model-Based Prediction

Prediction

So far, we've focused mostly on inferences about the estimated regression coefficients.

- Asking questions about how X is related to Y.

We can also use linear regression for *prediction*.

- Given a new observation, X_m, what outcome value, $\hat{Y}_m$, does our model attribute to the *m*th observation?

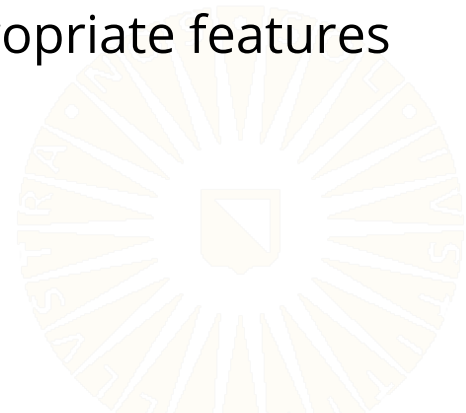
Prediction

Train a model to predict psychological well-being from diet-related and exercise-related features.

- Plug-in new feature values corresponding to an experimental wellness program to see the expected well-being for a hypothetical patient treated with the new program.

Predict future gasoline prices based on geo-political events in oil-producing countries.

- If conflict escalates in the Middle East, adjust the appropriate features and project likely changes in gasoline prices.



Prediction Example

To fix ideas, let's reconsider the *diabetes* data and the following model:

$$Y_{LDL} = \beta_0 + \beta_1 X_{BP} + \beta_2 X_{gluc} + \beta_3 X_{BMI} + \varepsilon$$

Training this model on the first $N = 400$ patients' data produces the following fitted model:

$$\hat{Y}_{LDL} = 22.135 + 0.089 X_{BP} + 0.498 X_{gluc} + 1.48 X_{BMI}$$

Prediction Example

To fix ideas, let's reconsider the *diabetes* data and the following model:

$$Y_{LDL} = \beta_0 + \beta_1 X_{BP} + \beta_2 X_{gluc} + \beta_3 X_{BMI} + \varepsilon$$

Training this model on the first $N = 400$ patients' data produces the following fitted model:

$$\hat{Y}_{LDL} = 22.135 + 0.089 X_{BP} + 0.498 X_{gluc} + 1.48 X_{BMI}$$

Suppose a new patient presents with $BP = 121$, $gluc = 89$, and $BMI = 30.6$. We can predict their LDL score by:

$$\begin{aligned} \hat{Y}_{LDL} &= 22.135 + 0.089(121) + 0.498(89) + 1.48(30.6) \\ &= 122.463 \end{aligned}$$

Interval Estimates for Prediction

To quantify uncertainty in our predictions, we want to use an appropriate interval estimate.

- Two flavors of interval are applicable to predictions:
  1. Confidence intervals for $\hat{Y}_m$
  2. Prediction intervals for a specific observation, Y_m

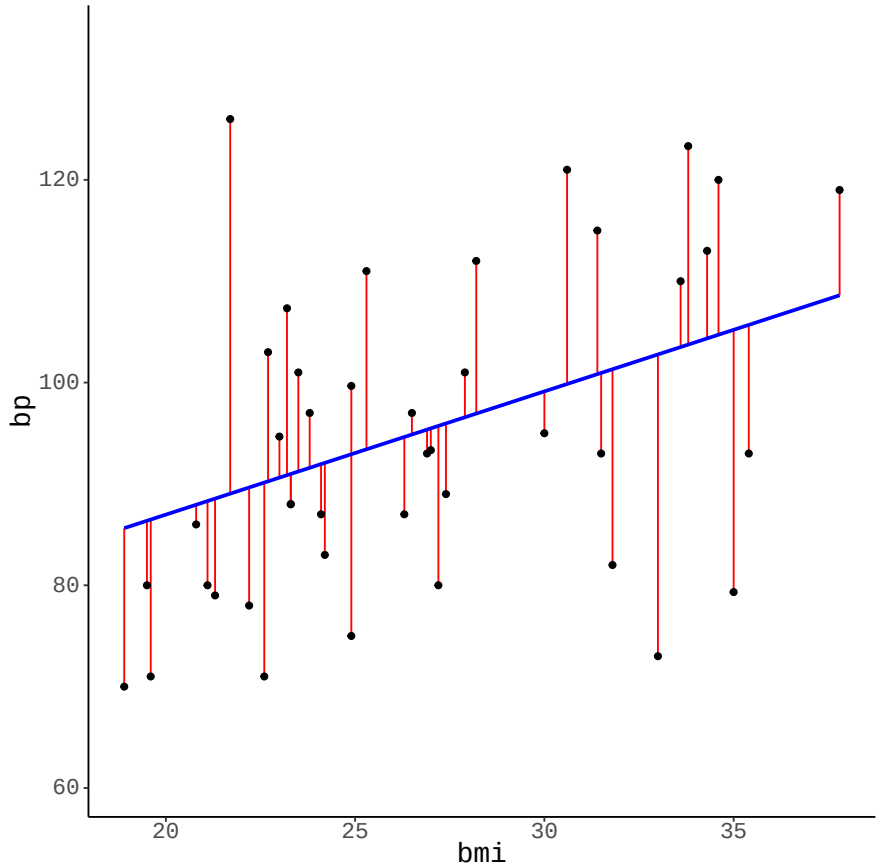
- The CI for $\hat{Y}_m$ gives a likely range (in the sense of coverage probability and "confidence") for the *m*th value of the true conditional mean.
  - CIs only account for uncertainty in the estimated regression coefficients, $\{\hat{\beta}_0, \hat{\beta}_p\}$.

- The prediction interval for Y_m gives a likely range (in the same sense as CIs) for the *m*th outcome value.
  - Prediction intervals also account for the regression errors, ε.

Confidence vs. Prediction Intervals

Let's visualize the predictions from a simple model:

$$Y_{BP} = \hat{\beta}_0 + \hat{\beta}_1 X_{BMI} + \hat{\varepsilon}$$

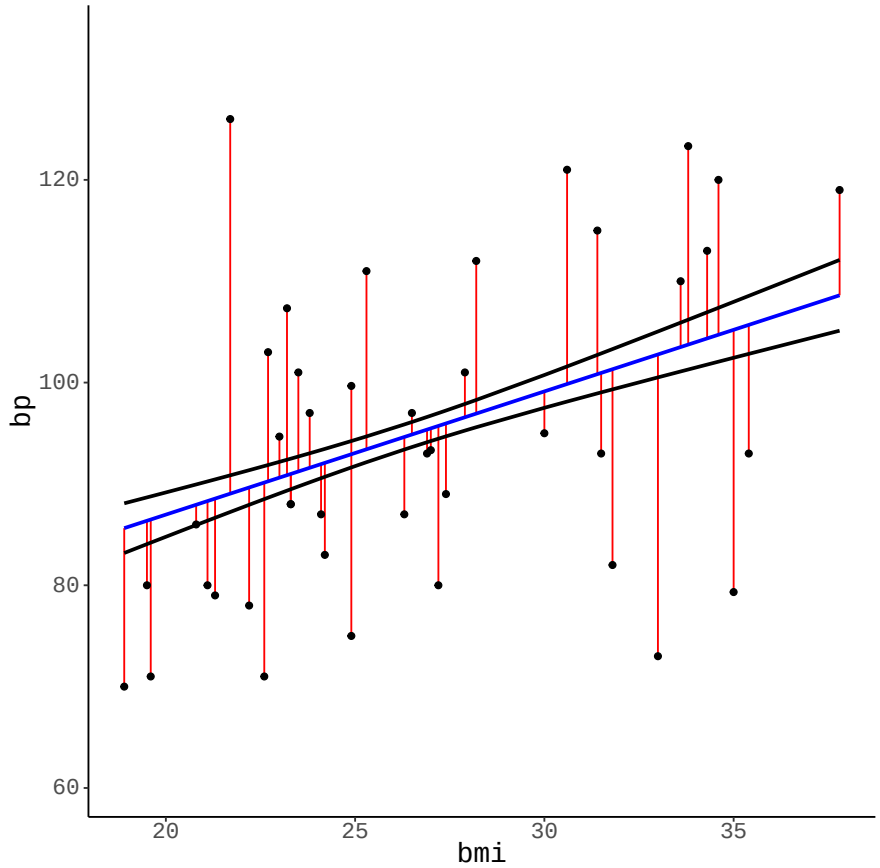


Confidence vs. Prediction Intervals

Let's visualize the predictions from a simple model:

$$Y_{BP} = \hat{\beta}_0 + \hat{\beta}_1 X_{BMI} + \hat{\varepsilon}$$

- CIs for $\hat{Y}$ ignore the errors, ε.
  - They only care about the best-fit line, $\beta_0 + \beta_1 X_{BMI}$.

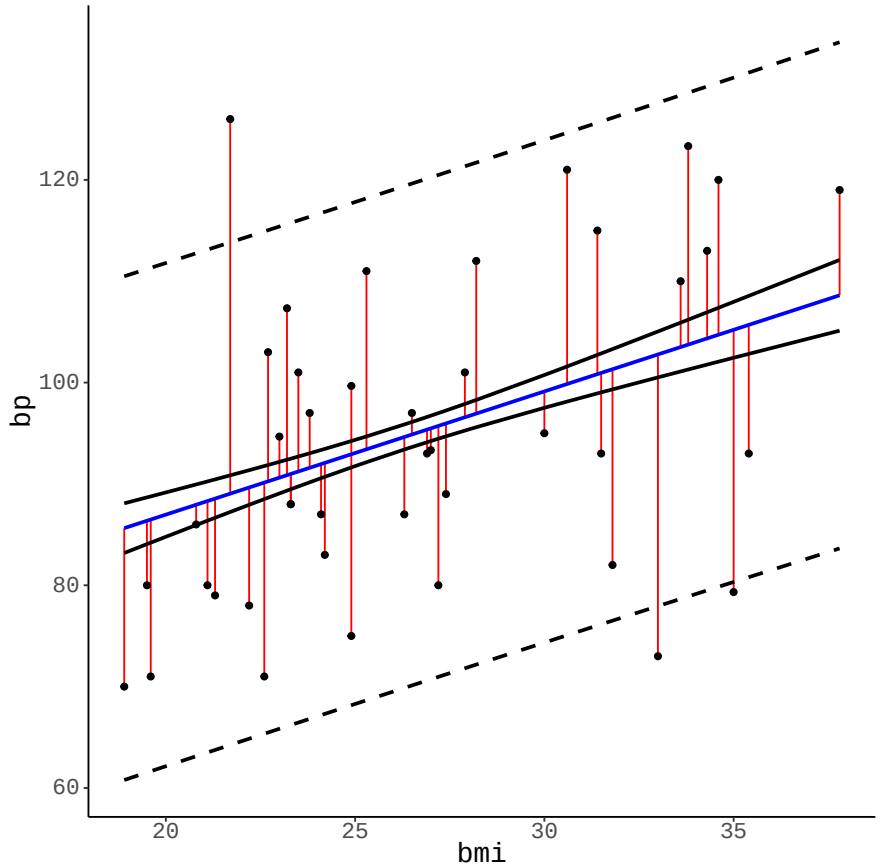


Confidence vs. Prediction Intervals

Let's visualize the predictions from a simple model:

$$Y_{BP} = \hat{\beta}_0 + \hat{\beta}_1 X_{BMI} + \hat{\varepsilon}$$

- CIs for $\hat{Y}$ ignore the errors, ε.
  - They only care about the best-fit line, $\beta_0 + \beta_1 X_{BMI}$.
- Prediction intervals are wider than CIs.
  - They account for the additional uncertainty contributed by ε.



Interval Estimates Example

Going back to our hypothetical "new" patient, we get the following 95% interval estimates:

$$95\%\ CI_{\hat{Y}} = [115.6; 129.33]$$

$$95\%\ PI = [66.56; 178.37]$$

- We can be 95% confident that the average *LDL* of patients with *Glucose* = 89, *BP* = 121, and *BMI* = 30.6 will be somewhere between 115.6 and 129.33.
- We can be 95% confident that the *LDL* of a specific patient with *Glucose* = 89, *BP* = 121, and *BMI* = 30.6 will be somewhere between 66.56 and 178.37.

References