Review of Linear Regression

Utrecht University

Kyle M. Lang

Department of Methodology & Statistics
Utrecht University

Outline

1. Introduction to the “regression” problem
2. Simple linear regression
3. Multiple linear regression
4. Categorical predictors

Regression Problem

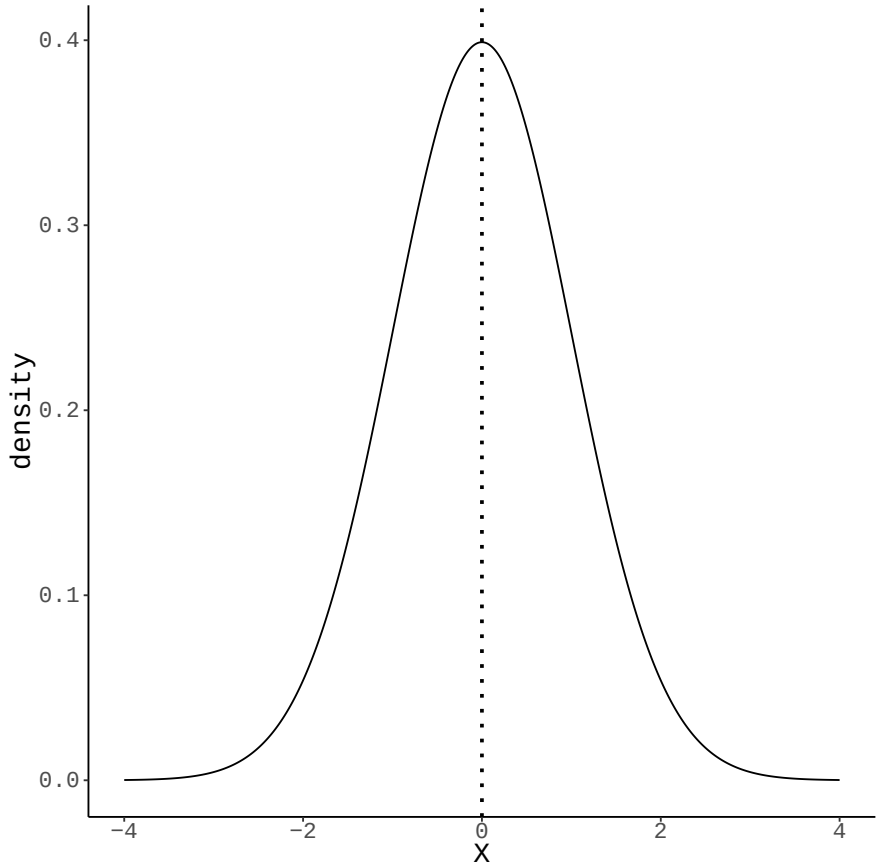
Some of the most ubiquitous and useful statistical models are *regression models*.

- *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

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."

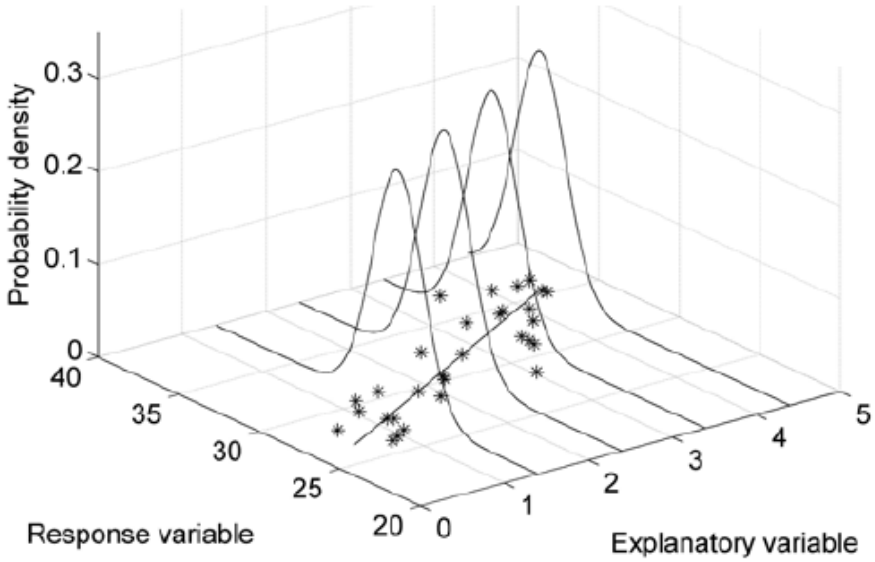
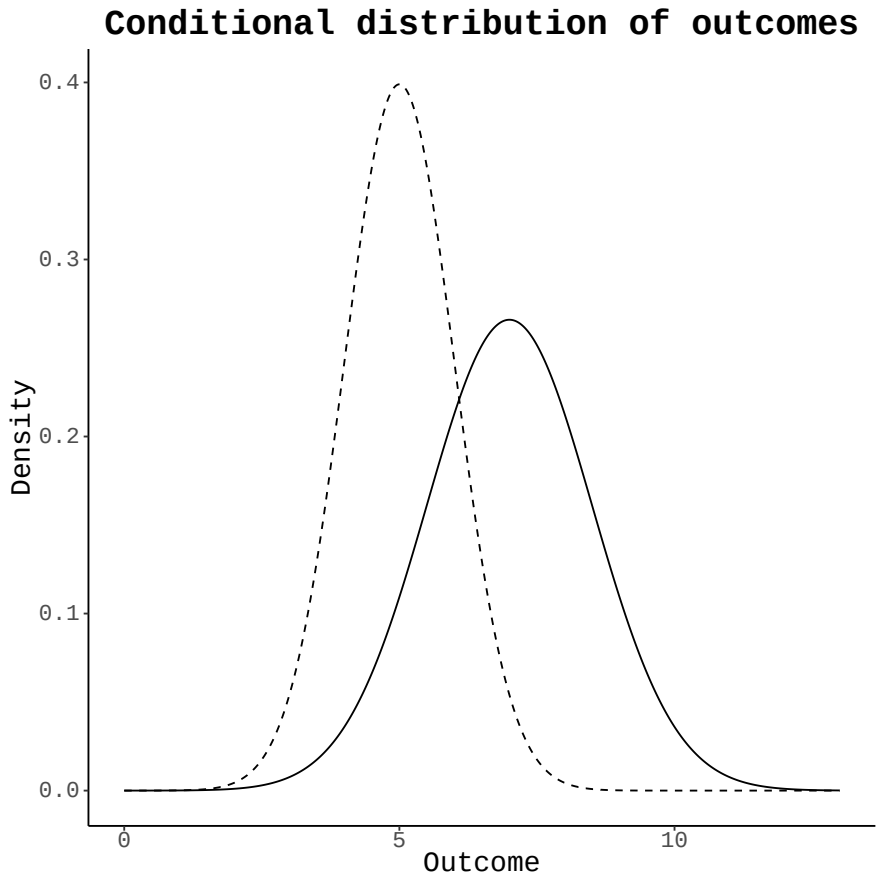


Image retrieved from:
`http://www.seaturtle.org/mtn/archives/mtn122/mtn122p1.shtml`

Flavors of Probability Distribution

Even a simple comparison of means implies a conditional distribution.

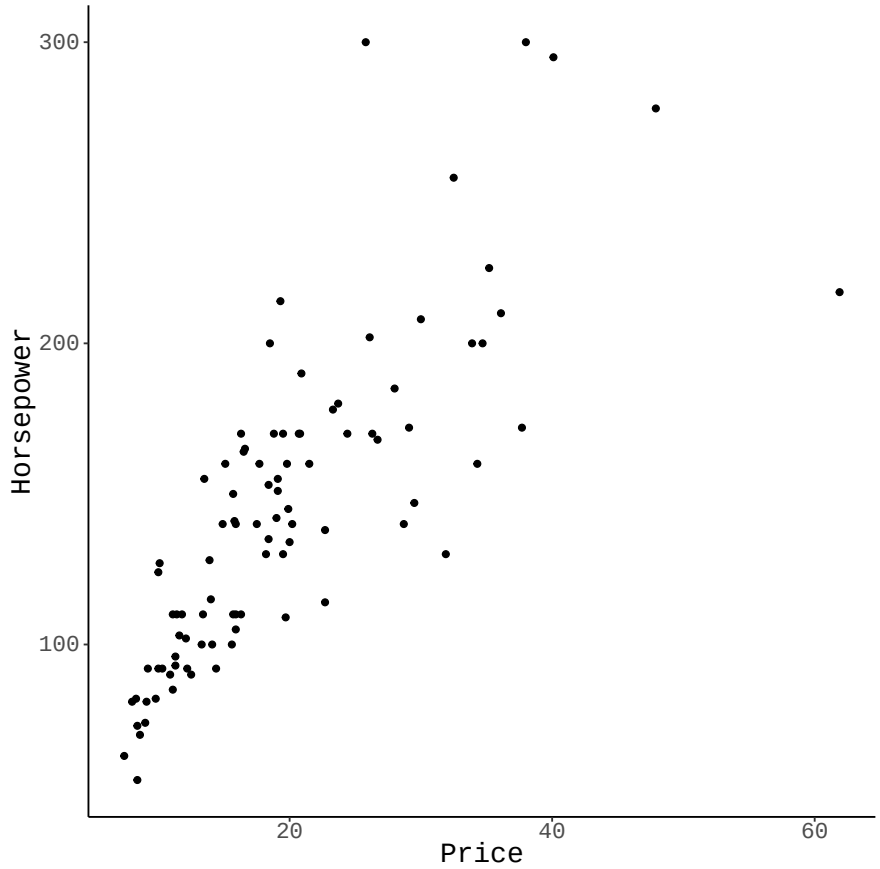
- The solid curve corresponds to outcome values for one group.
- The dashed curve represents outcomes from the other group.



Projecting a Distribution onto the Plane

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

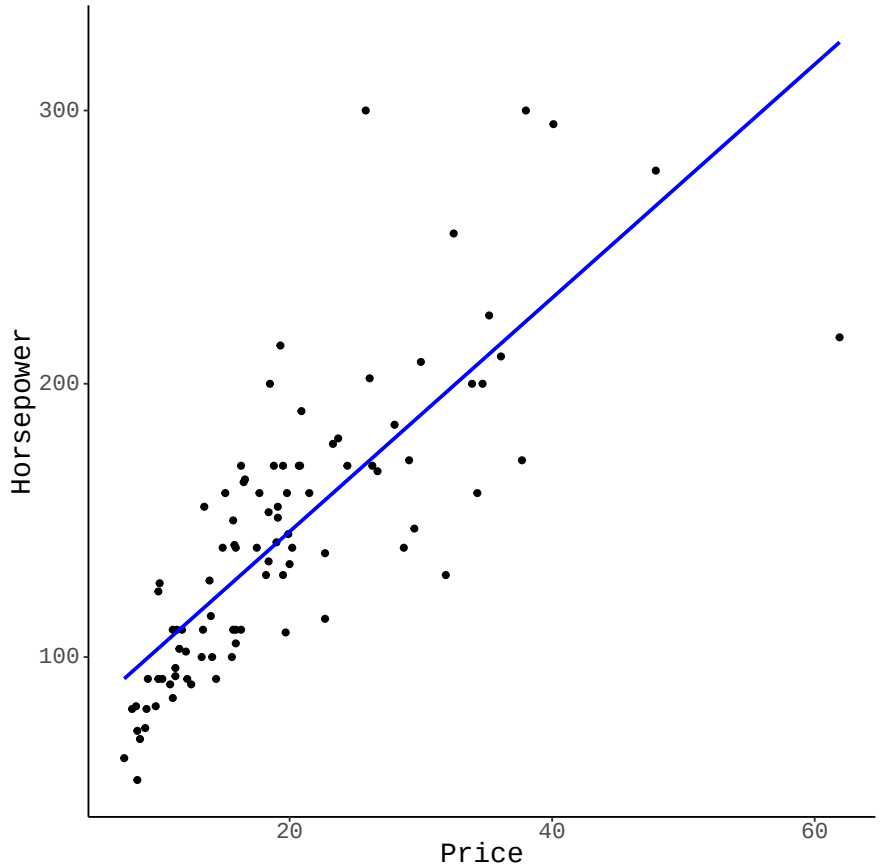
- On the Y-axis, we plot our outcome variable
- The X-axis represents the predictor variable upon which we condition the mean of Y.



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

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 $X = 0$.

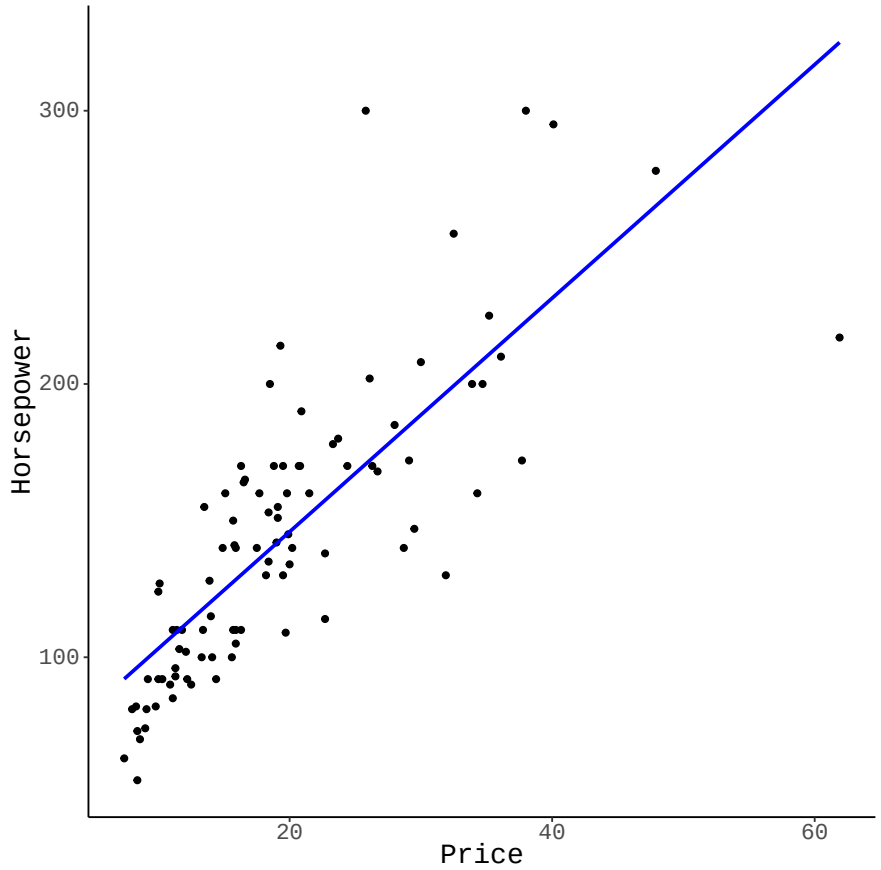
$\hat{\beta}_1$ is the *slope*.

- 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.



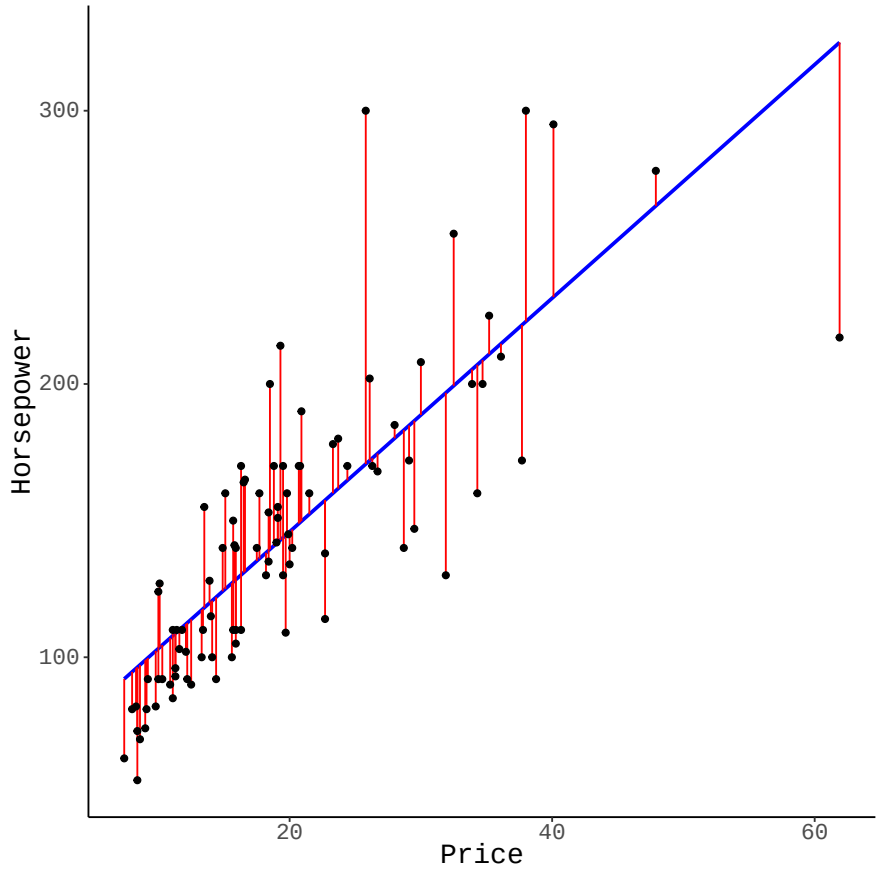
Thinking about Error

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

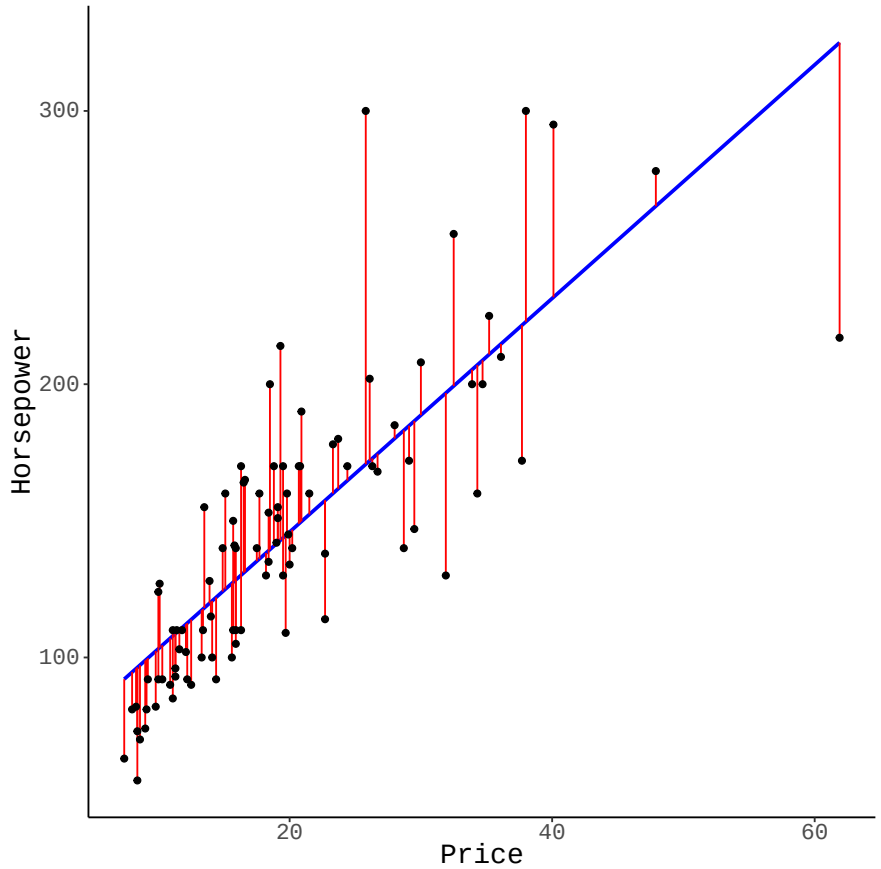
The $\hat{\varepsilon}_n$ are defined in terms of deviations between each observed Y_n value and the corresponding $\hat{Y}_n$.

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

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:

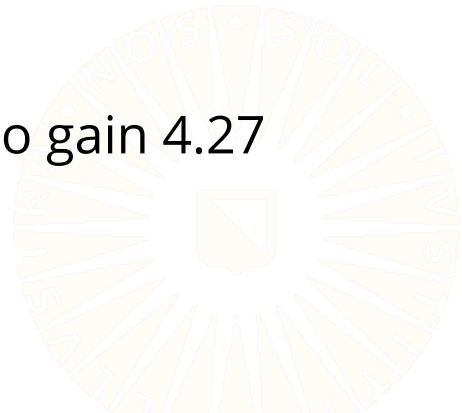
```
#data(Cars93)
out1 <- lm(Horsepower ~ Price, data = Cars93)
coef(out1)

## (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.



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?

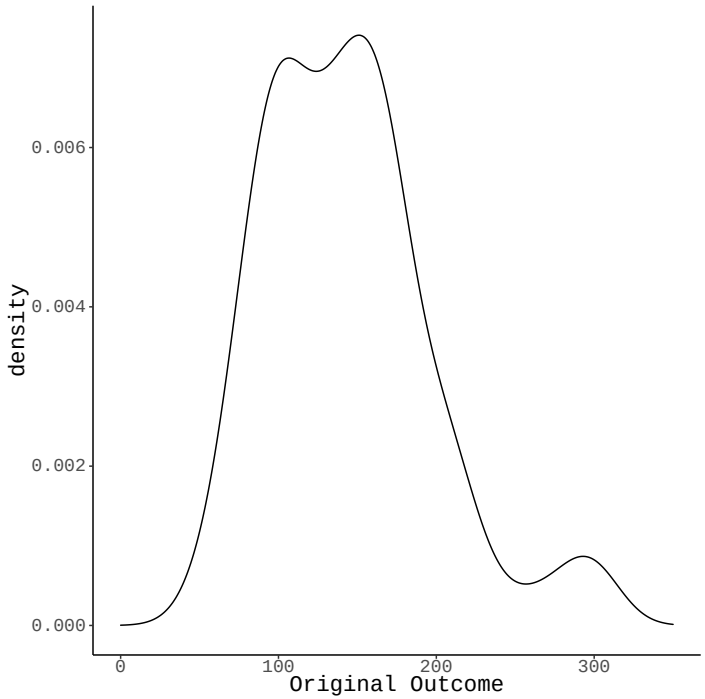
When doing prediction, we want to build a tool that can accurately guess future values.

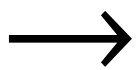
- Will it rain tomorrow?
- Will this investment turn a profit within one year?
- Will increasing the number of contact hours improve grades?

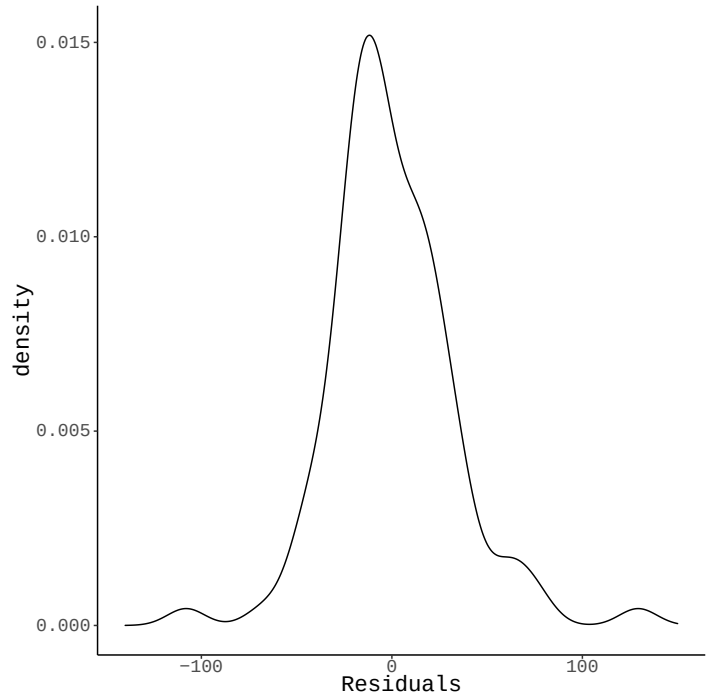
Model Fit

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 $\hat{\sigma}^2 = \text{Var}(\hat{\varepsilon})$ will be less than $\text{Var}(Y)$.



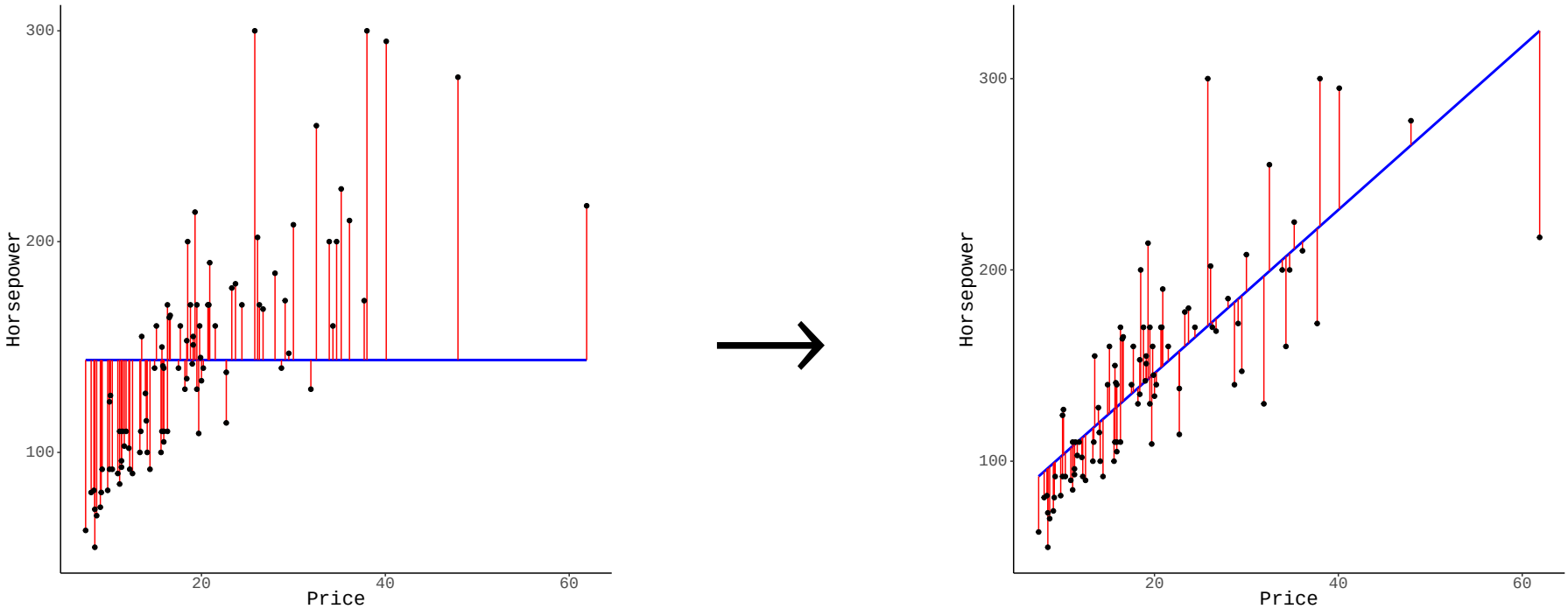




Model Fit

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 $\hat{\sigma}^2 = \text{Var}(\hat{\varepsilon})$ will be less than $\text{Var}(Y)$.



Model Fit

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

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

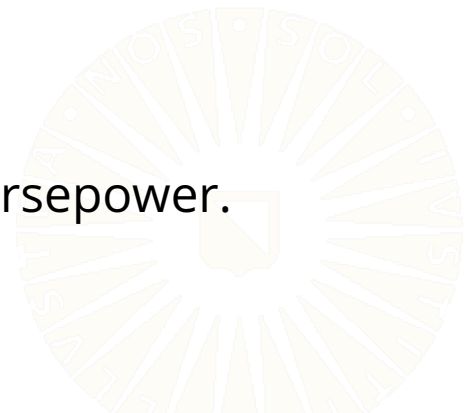
where

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

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:

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

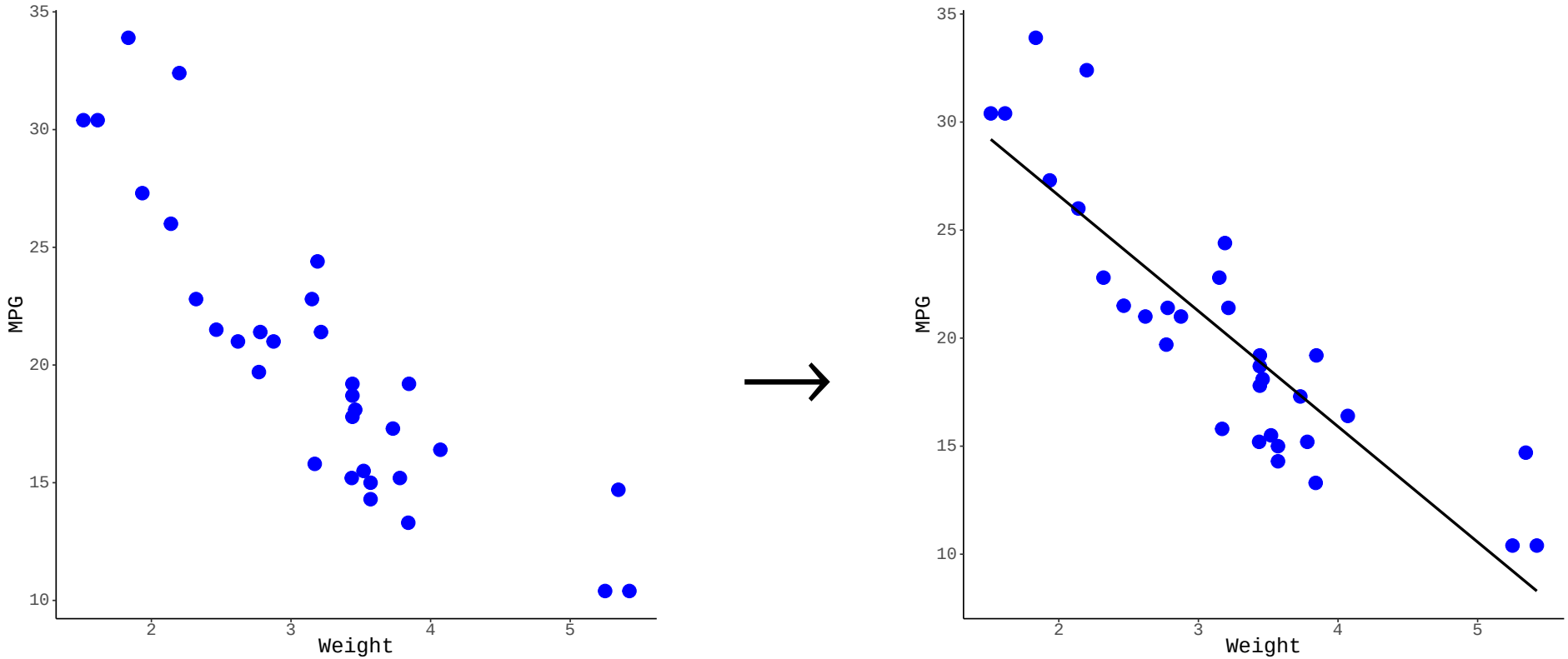
- When using price as the only predictor of horsepower, we expect prediction errors with magnitudes of 32.06 horsepower.

Multiple Linear Regression

Graphical Representations of Regression Models

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

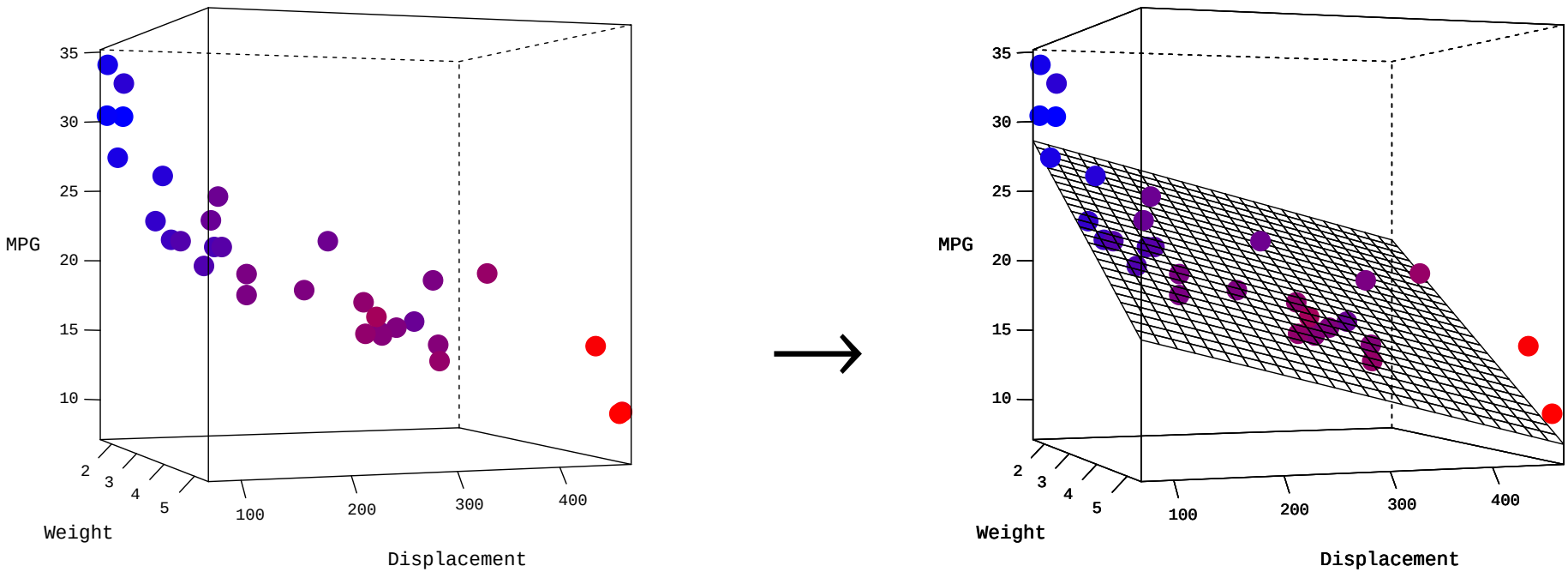
- Simple linear regression implies a 1D line in 2D space.



Graphical Representations of Regression Models

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?*

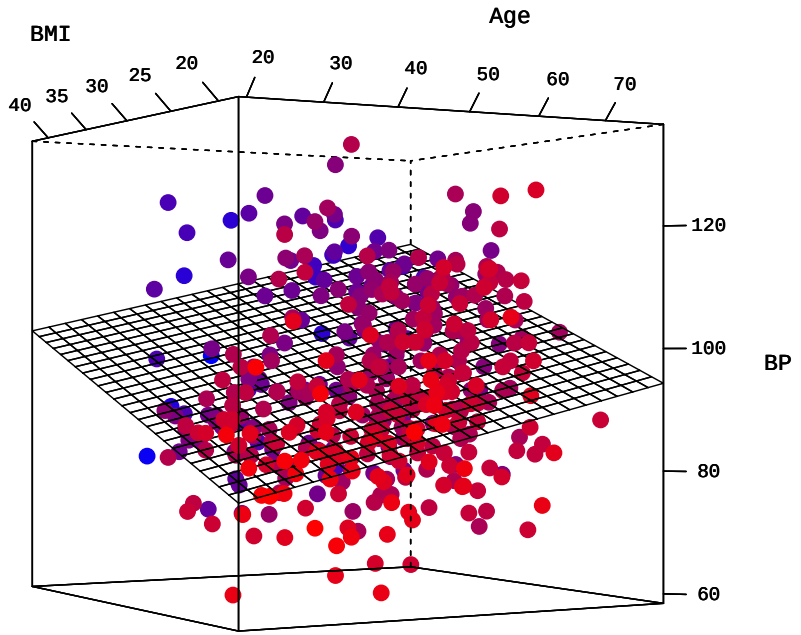
- We're partialing age out of the effect of BMI on blood pressure.

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
```

Model Comparison

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.

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.

```
## Center predictor variables:
dDat$ldl100 <- dDat$ldl - 100
dDat$hdl60  <- dDat$hdl - 60
dDat$age30  <- dDat$age - 30

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

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

Model-Building 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) 88.09330    1.07470  81.970  < 2e-16
## age30        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
```

Model-Building Example

```
partSummary(out2, -1)

## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.297  -8.106  -0.979   8.141  40.677
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 86.53984    1.13885  75.989  < 2e-16
## age30        0.32178    0.04784   6.727 5.43e-11
## ldl100       0.04166    0.02097   1.987  0.04757
## hdl60       -0.14740    0.04824  -3.055  0.00239
##
## Residual standard error: 12.84 on 438 degrees of freedom
## Multiple R-squared:  0.1439,Adjusted R-squared:  0.1381
## F-statistic: 24.55 on 3 and 438 DF,  p-value: 1.064e-14
```

Interpretations

- The expected average blood pressure for a 30 year old patient with LDL = 100 and HDL = 60 is 86.54.
- For each additional year older, average blood pressure is expected to increase by 0.32, after controlling for LDL and HDL levels.
- For each additional unit of LDL level, average blood pressure is expected to increase by 0.04, after controlling for age and HDL.
- For each additional unit of HDL level, average blood pressure is expected to decrease by -0.15, after controlling for age and LDL.

Model Comparison

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

## [1] 0.03142445

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

## Analysis of Variance Table
##
## Model 1: bp ~ age30
## Model 2: bp ~ age30 + ldl100 + hdl60
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    440 74873
## 2    438 72222  2    2651.1 8.0391 0.0003726 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model Comparison

```
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] 163.3983
```

Interpretations

- Age, LDL, and HDL explain a combined 14.4% 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 3.1% 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 = 163.4 vs. MSE = 169.4).

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:
dDat$bmi25 <- dDat$bmi - 25

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

Model-Building Example

```
partSummary(out3, -1)

## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.970  -8.145  -0.300   8.456  41.135
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 87.46233    1.08944  80.282  < 2e-16
## age30        0.27949    0.04582   6.099 2.35e-09
## ldl100       0.01646    0.02024   0.814    0.416
## hdl60       -0.03478    0.04856  -0.716    0.474
## bmi25        1.01743    0.14568   6.984 1.07e-11
##
## Residual standard error: 12.19 on 437 degrees of freedom
## Multiple R-squared:  0.2299,Adjusted R-squared:  0.2228
## F-statistic: 32.61 on 4 and 437 DF,  p-value: < 2.2e-16
```

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	0.042	-0.147
With BMI	0.016	-0.035

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] 0.08595543
```

Model Comparison

Is the additional 8.6% variation explained a significant increase?

```
anova(out2, out3)

## Analysis of Variance Table
##
## Model 1: bp ~ age30 + ldl100 + hdl60
## Model 2: bp ~ age30 + ldl100 + hdl60 + bmi25
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    438 72222
## 2    437 64970  1    7251.7 48.776 1.074e-11 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model Comparison

What about the relative predictive performance?

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

mse2

## [1] 163.3983

mse3

## [1] 146.9918
```

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 = dDat)
```

Model-Building Example

```
partSummary(out4, -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) 87.85488    1.00420  87.487  < 2e-16
## age30        0.28651    0.04504   6.362 5.02e-10
## bmi25        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
```

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] 0.002330906
```

Model Comparison

Is this 0.23% loss in explained variance significant?

```
anova(out4, out3)

## Analysis of Variance Table
##
## Model 1: bp ~ age30 + bmi25
## Model 2: bp ~ age30 + ldl100 + hdl60 + bmi25
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1    439 65167
## 2    437 64970  2    196.65 0.6613 0.5167
```

Model Comparison

How do the prediction errors compare?

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

mse3

## [1] 146.9918

mse4

## [1] 147.4367
```

Categorical Predictors

Categorical Predictors

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

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

We often want to include grouping factors as predictors.

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

How to Model Categorical Predictors

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

- The variables need to be coded before entering the model

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

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

How to Model Categorical Predictors

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

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

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

How to Model Categorical Predictors

```
partSummary(out1, -1)

## Residuals:
##    Min     1Q Median     3Q    Max
## -1.303 -0.313 -0.113  0.342  1.342
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.33200    0.12060  -2.753  0.00664
## majN         2.04500    0.05582  36.632  < 2e-16
##
## Residual standard error: 0.5582 on 148 degrees of freedom
## Multiple R-squared:  0.9007,Adjusted R-squared:    0.9
## F-statistic:  1342 on 1 and 148 DF,  p-value: < 2.2e-16
```

Dummy Coding

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

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

Example Dummy Code

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

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

Example Dummy Codes

Now, a slightly more complex example:

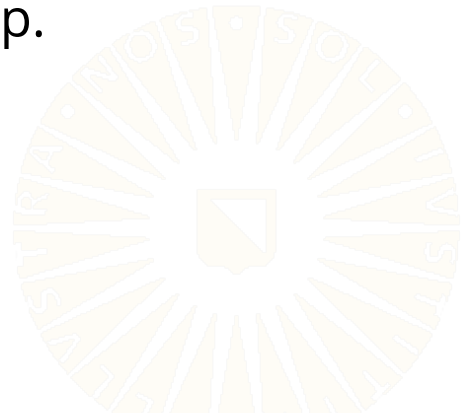
	drink	juice	tea
1	juice	1	0
2	coffee	0	0
3	tea	0	1
4	tea	0	1
5	tea	0	1
6	tea	0	1
7	juice	1	0
8	tea	0	1
9	coffee	0	0
10	juice	1	0

Using Dummy Codes

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

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

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



Example

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

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

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

Example

```
partSummary(out2, -1)

## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.341  -6.338  -3.141   2.662  38.059
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   23.841      1.623  14.691   <2e-16
## mtOpt         -6.603      2.004  -3.295   0.0014
##
## Residual standard error: 9.18 on 91 degrees of freedom
## Multiple R-squared:  0.1066,Adjusted R-squared:  0.09679
## F-statistic: 10.86 on 1 and 91 DF,  p-value: 0.001403
```

Interpretations

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

Example

Fit a more complex model:

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

## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.050  -6.250  -1.236   3.264  32.950
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.63000    2.76119   6.385 7.33e-09
## front       -0.09418    2.96008  -0.032  0.97469
## rear        11.32000    3.51984   3.216  0.00181
##
## Residual standard error: 8.732 on 90 degrees of freedom
## Multiple R-squared:  0.2006,Adjusted R-squared:  0.1829
## F-statistic: 11.29 on 2 and 90 DF,  p-value: 4.202e-05
```

Interpretations

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

Example

Include two sets of dummy codes:

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

## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  21.7187     2.9222   7.432 6.25e-11
## mtOpt        -5.8410     1.8223  -3.205  0.00187
## front        -0.2598     2.8189  -0.092  0.92677
## rear         10.5169     3.3608   3.129  0.00237
##
## Residual standard error: 8.314 on 89 degrees of freedom
## Multiple R-squared:  0.2834,Adjusted R-squared:  0.2592
## F-statistic: 11.73 on 3 and 89 DF,  p-value: 1.51e-06
```

Interpretations

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

Significance Testing

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

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

## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   23.841      1.623  14.691   <2e-16
## mtOpt         -6.603      2.004  -3.295   0.0014
##
## Residual standard error: 9.18 on 91 degrees of freedom
## Multiple R-squared:  0.1066,Adjusted R-squared:  0.09679
## F-statistic: 10.86 on 1 and 91 DF,  p-value: 0.001403
```

Significance Testing

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

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

## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  21.7187     2.9222   7.432 6.25e-11
## mtOpt        -5.8410     1.8223  -3.205  0.00187
## front        -0.2598     2.8189  -0.092  0.92677
## rear         10.5169     3.3608   3.129  0.00237
##
## Residual standard error: 8.314 on 89 degrees of freedom
## Multiple R-squared:  0.2834,Adjusted R-squared:  0.2592
## F-statistic: 11.73 on 3 and 89 DF,  p-value: 1.51e-06
```

Significance Testing

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

## [1] 0.1767569

anova(out2, out4)

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

Significance Testing

What about models where a single nominal factor is the only predictor?

```
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
```

Significance Testing

We can compare back to an “intercept-only” model.

```
out0 <- lm(price ~ 1, data = cDat)
partSummary(out0, -1)

## Residuals:
##    Min     1Q Median     3Q    Max
## -12.11  -7.31  -1.81   3.79  42.39
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   19.510      1.002   19.48   <2e-16
##
## Residual standard error: 9.659 on 92 degrees of freedom
```

Significance Testing

```
r2Diff <- summary(out3)$r.squared - summary(out0)$r.squared
r2Diff

## [1] 0.2006386

anova(out0, out3)

## Analysis of Variance Table
##
## Model 1: price ~ 1
## Model 2: price ~ front + rear
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     92 8584.0
## 2     90 6861.7  2    1722.3 11.295 4.202e-05 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Significance Testing

We don't actually need to do the explicit model comparison, though.

```
r2Diff

## [1] 0.2006386

summary(out3)$r.squared

## [1] 0.2006386

anova(out0, out3)[2, "F"]

## [1] 11.29494

summary(out3)$fstatistic[1]

##    value
## 11.29494
```

Conclusion

- Each variable in a regression model corresponds to a dimension in the data-space.
  - A regression model with P predictors implies a P-dimensional (hyper)-plane in (P + 1)-dimensional space.
- The coefficients in MLR are partial coefficients.
  - Each effect is interpreted as holding other predictors constant.
- Categorical predictors must be coded before they can be used in our models.
  - The regression coefficients represent group mean differences.