Class 7: Linear Regression 1

MGSC 310

Prof. Jonathan Hersh

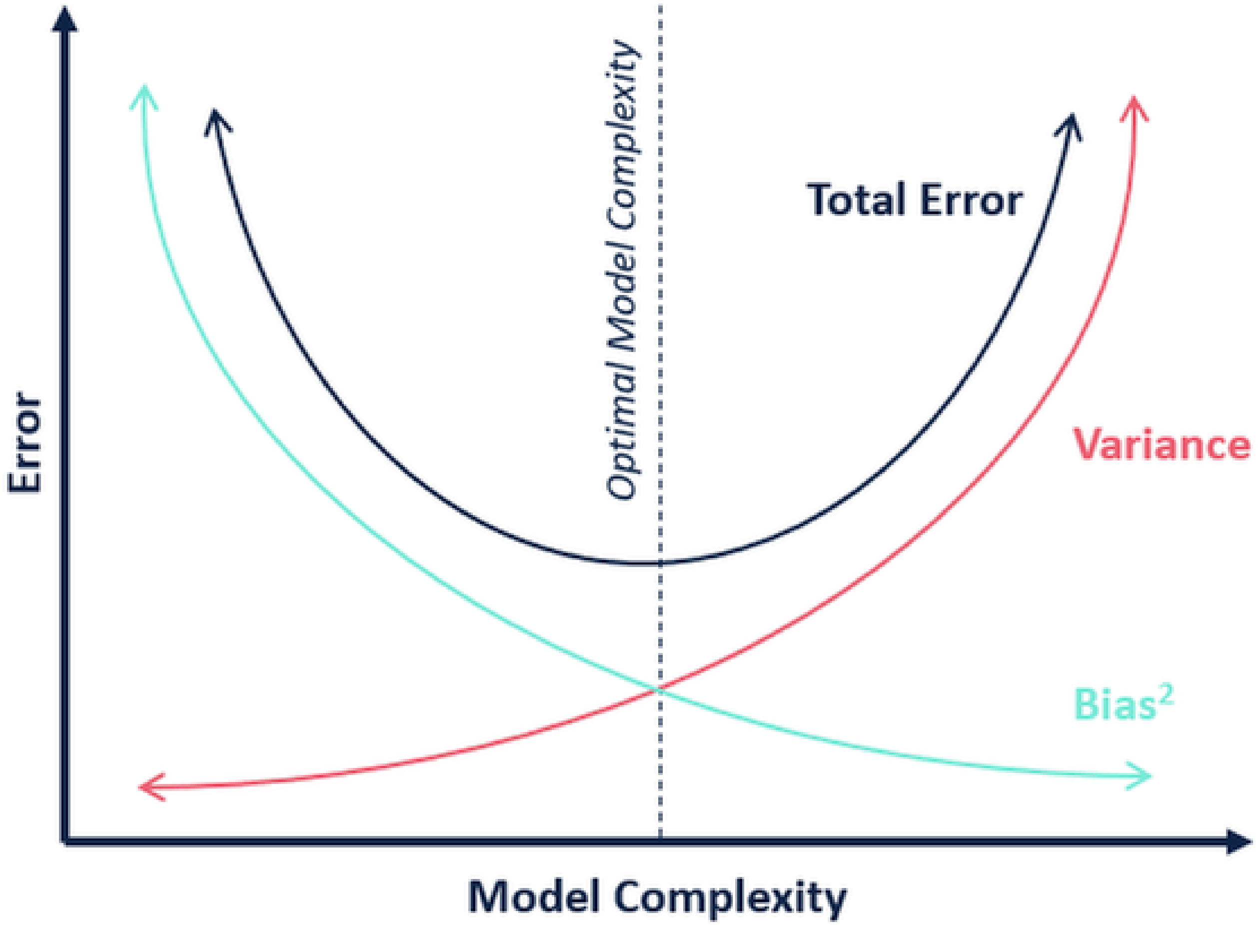
Class 7: Announcements

1. TA Office Hours:
   - Tuesdays: 5:30 – 7
   - Thursdays: 12:30-2
   - Mondays: 5-6:30
2. Quiz 3 posted, due Thursday @ midnight
3. Be sure you are following along with the course reading (ISLR pp 15-36 covered today)
4. Problem Set 2 Posted, Due Sept 29
5. **Problem Set Solutions:**
   - Typically submit these via hard copy b/c of cheating
   - Cannot do this this year ☹
   - For now: TA/Instructor Office Hours will share solutions for specific Qs. Hard copies will be available on campus if you care to pick these up.

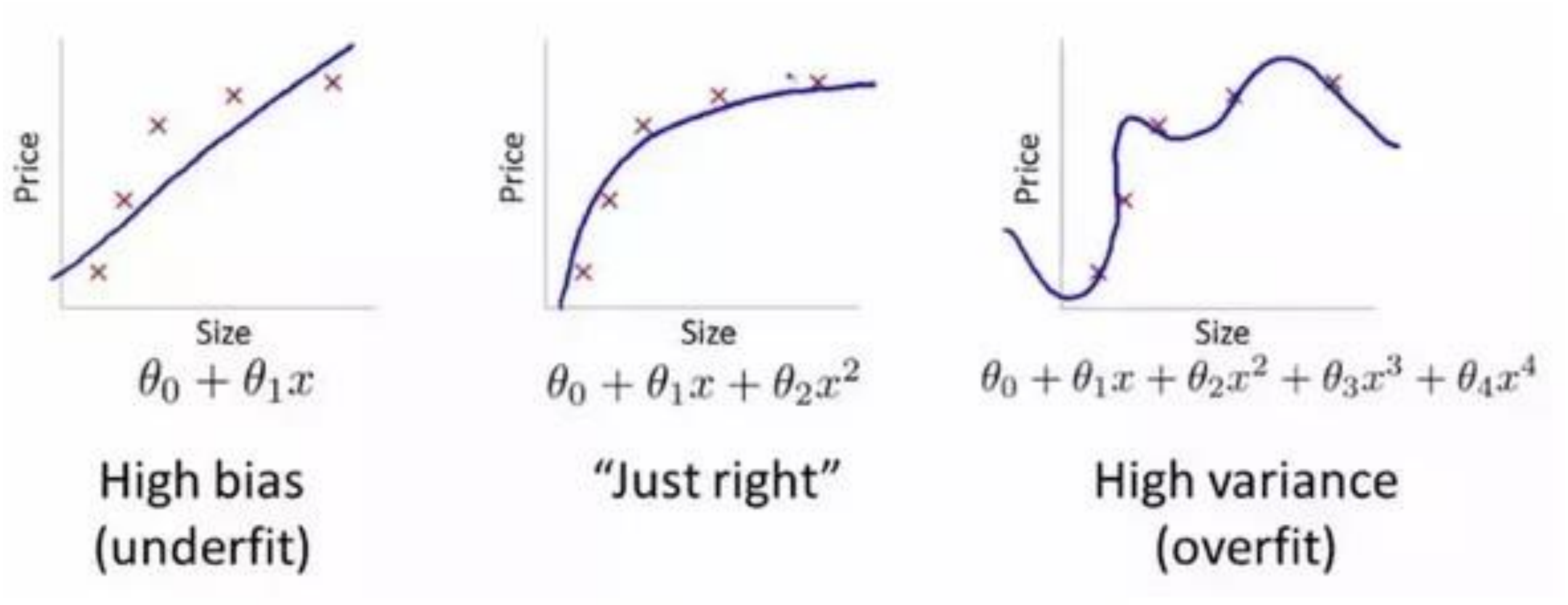
Class 7: Outline

1. Last Class Review:
    - Bias, Variance, Overfit, Underfit, Mean Squared Error
2. Linear Regression Review
3. Estimating Linear Models in R
4. Interpreting Model Coefficients
5. Regression Lab

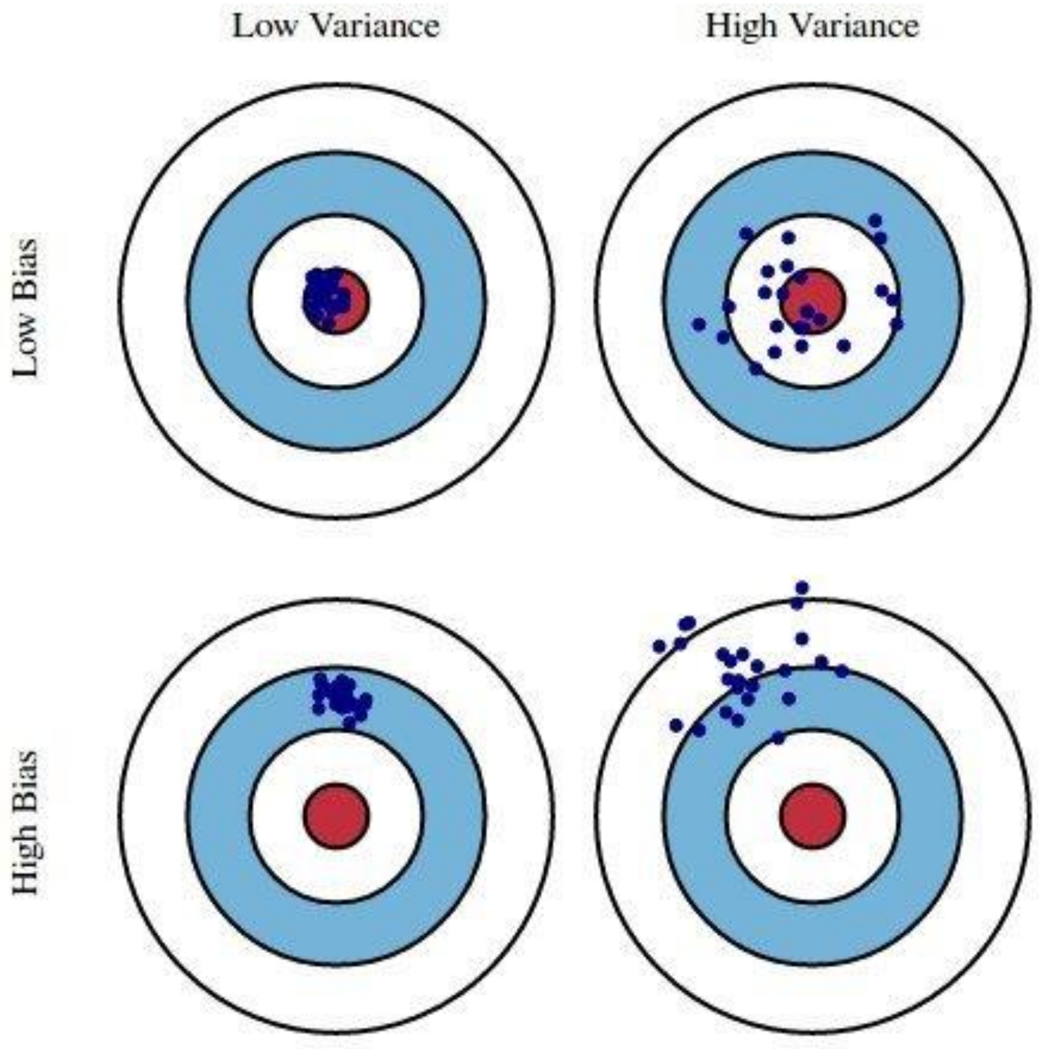
Key: Finding Optimal Model Complexity



Optimal Model Complexity: Neither Underfit Nor Overfit



Bias-Variance Tradeoff



Mean Squared Error in Practice

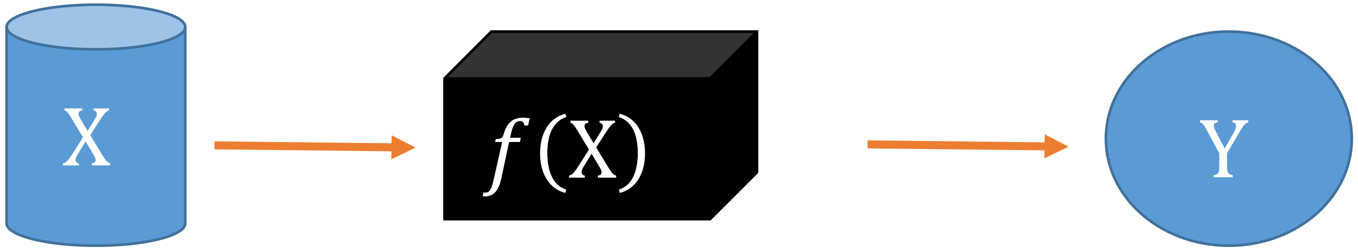
$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2$$

$\sum$ means we add up anything with i, starting at $i = 1$ to $i = n$

y_i	$\hat{y}_i$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
5	5	0	0
5	7	2	2^2=4
9	8	1	1^2=1
10	1	1	9^2=81
13	13	0	0

Recipes for learning $f(X)$: Ordinary Linear Models

$$Y = f(X) + \epsilon$$



Ordinary Linear Models

$$f(X) = \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_3 \cdot x_3$$

OLS: Only allows linear combinations of Xs

Class 7: Outline

1. Review Bias, Variance, Overfit, Underfit
2. **Linear Regression Review**
3. Estimating Linear Models in R
4. Interpreting Model Coefficients
5. Regression Lab

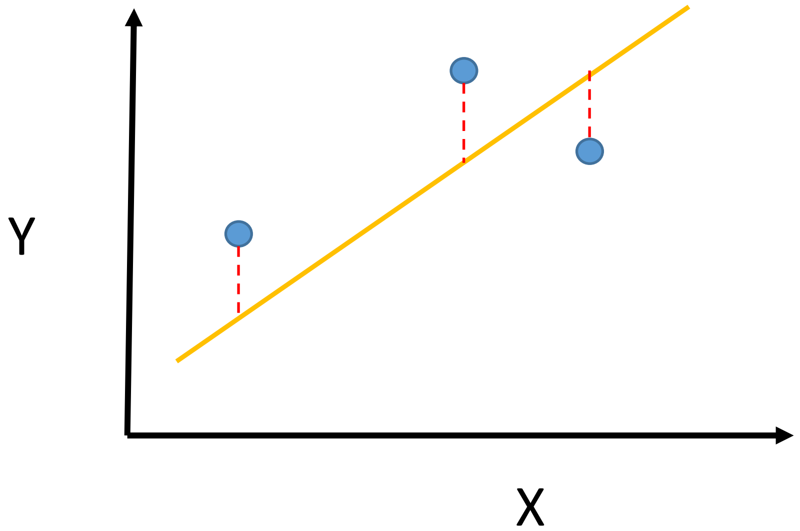
How Are Linear Regression Coefficients Chosen?

$$\hat{\beta}\ minimizes: \sum_{i=1}^{N} (y_i - \beta_0 - x_{i1}\beta_1 - x_{i2}\beta_2 - \cdots - x_{ik}\beta_k)^2$$

Least Squares Minimizes the **sum of squared residuals**

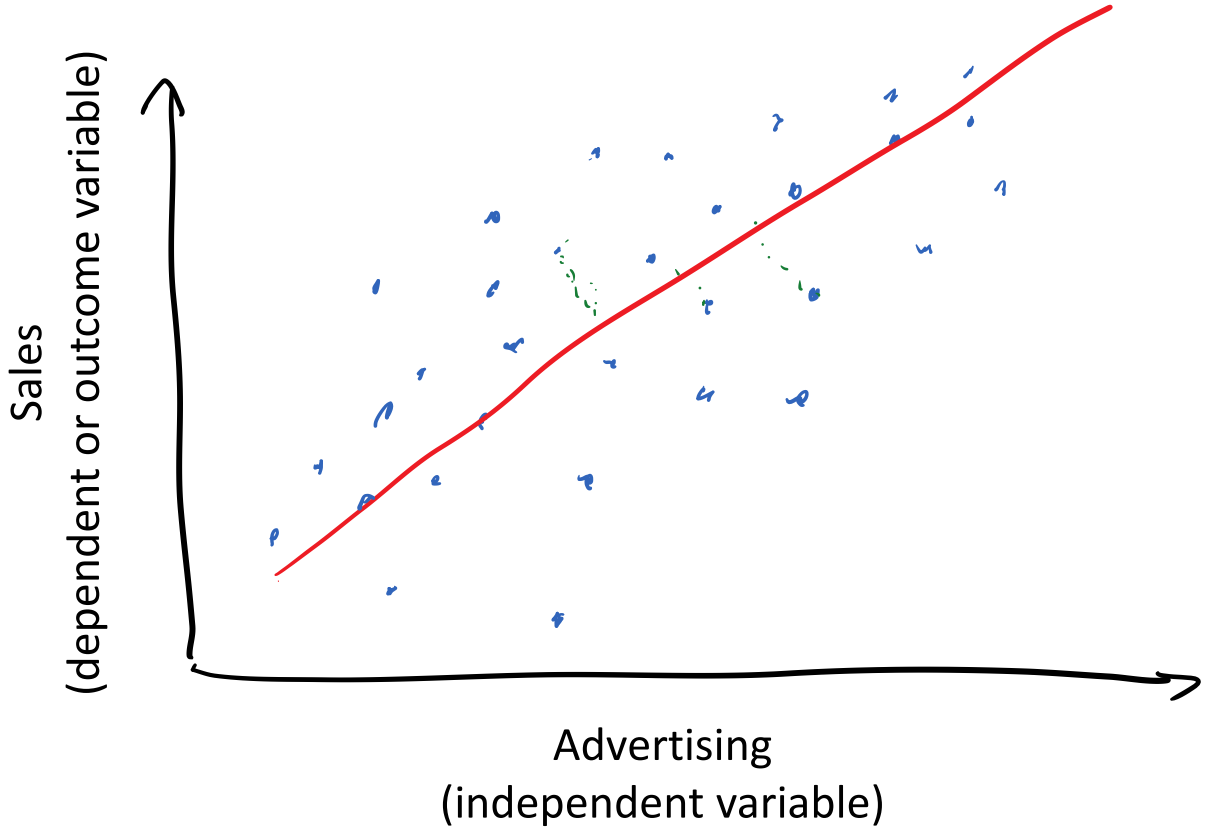
Sum through all observations

$$\epsilon_i = y_i - \hat{y}$$
$$= \mathrm{y_i} - \beta_0 - x_1\beta_1 - \cdots - x_2\beta_2$$



Visually, the slope (β_1) minimizes the difference between the points and the yellow line (red lines)

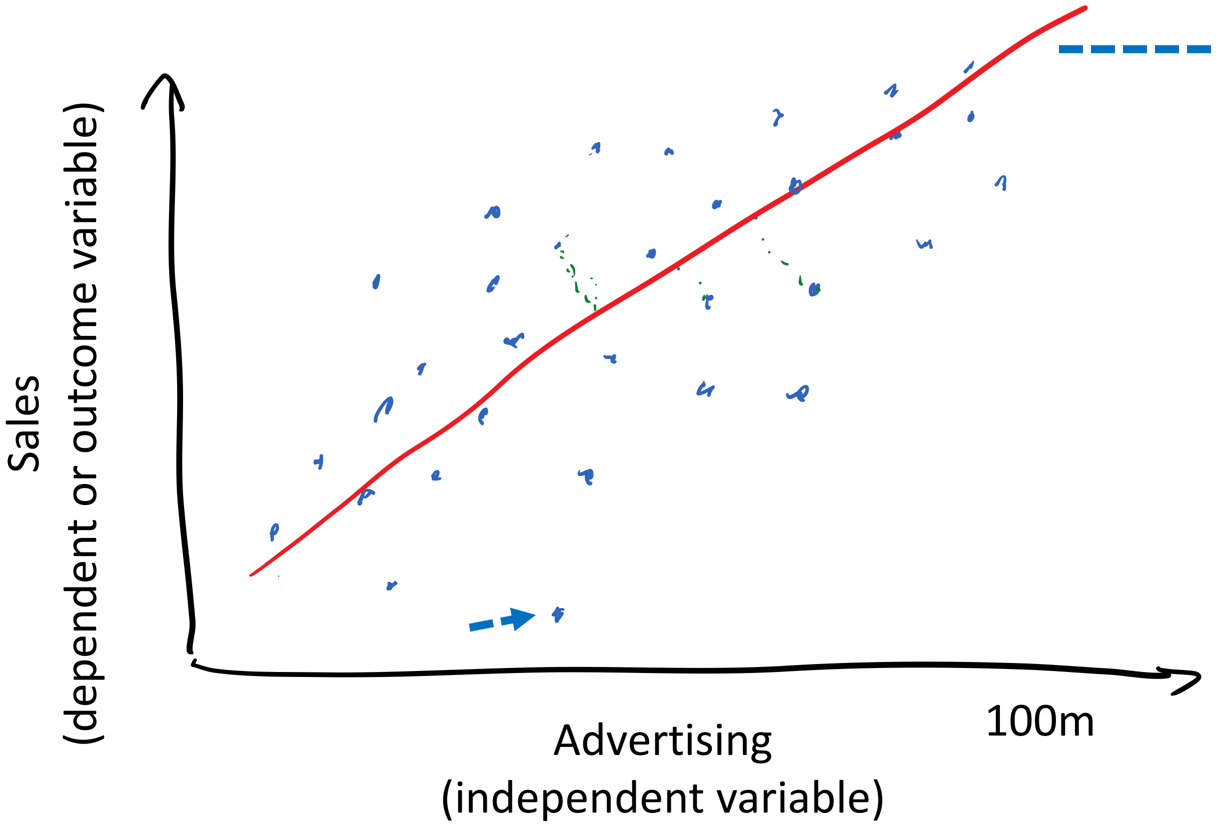
What is Linear Regression?



Regression: statistical process of estimating relationship between an <u>outcome</u> and and one or more <u>predictors</u> or <u>independent variables</u>

Linear Regression: restricting relationship between predictors and outcome to be linear

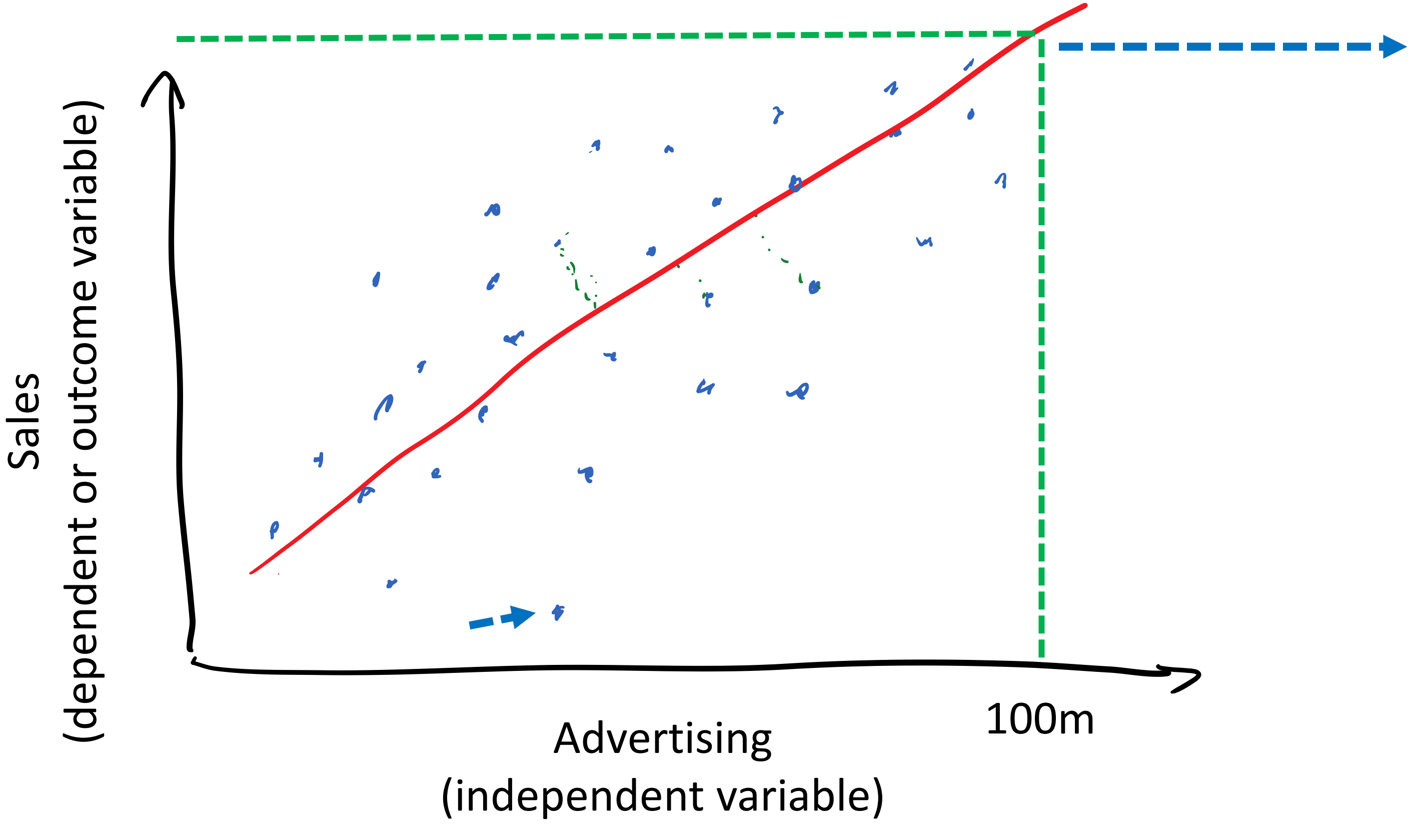
Linear Regression Equation



$$y_i = \beta_0 + \beta_1 x_{Advertising} + \epsilon_i$$

Red line “explains” the data the best.

Predictions from Linear Regression



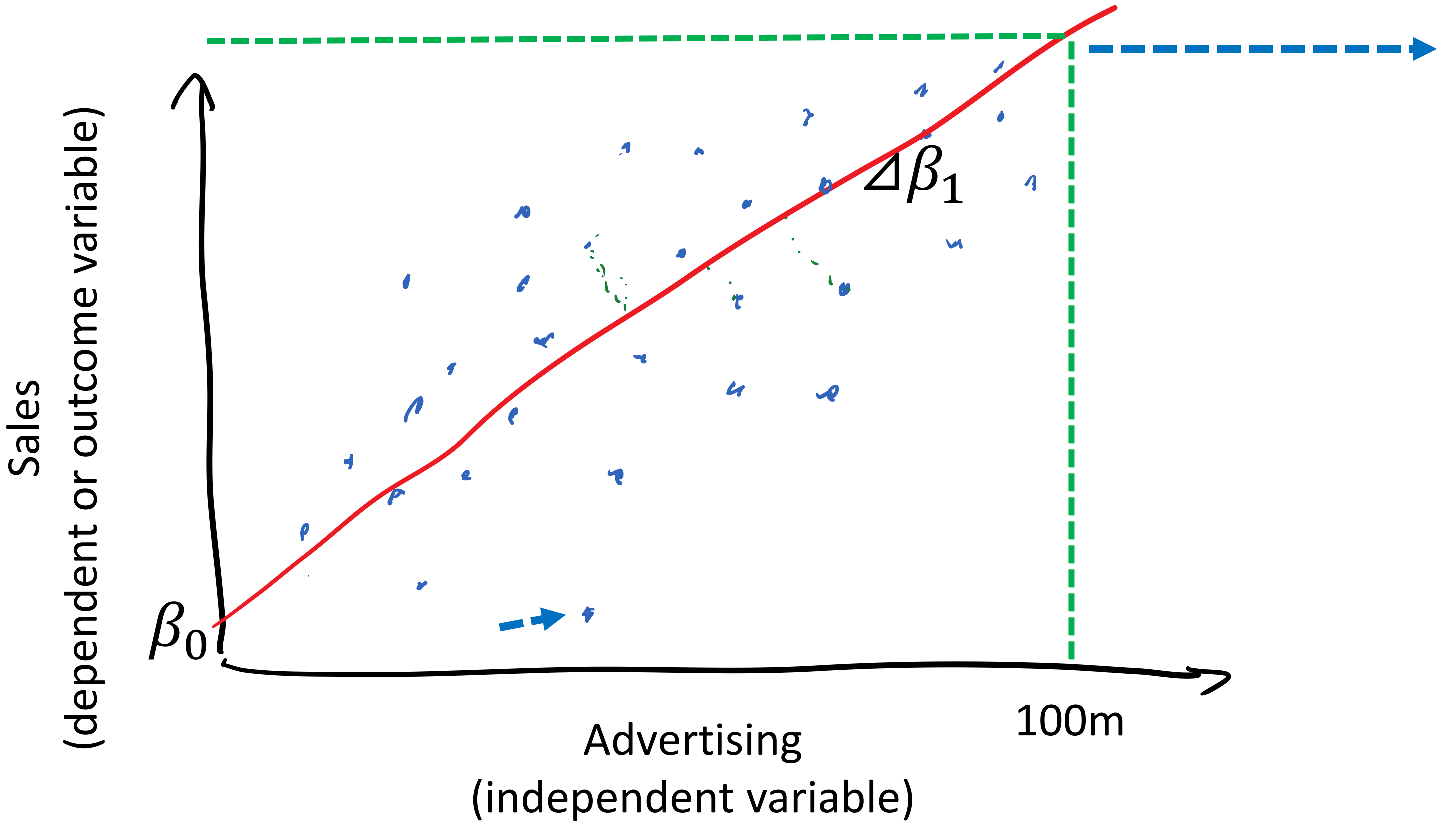
$$y_i = \beta_0 + \beta_1 x_{Advertising} + \epsilon_i$$

Suppose we spend 100m on advertising?
What's our expected sales?

$$? = \widehat{\beta_0} + \widehat{\beta_1} 100m$$

"Hat", e.g. $\widehat{\beta_0}$, means we've estimated this relationship from data.

Predictions from Linear Regression



$$y_i = \beta_0 + \beta_1 x_{Advertising} + \epsilon_i$$

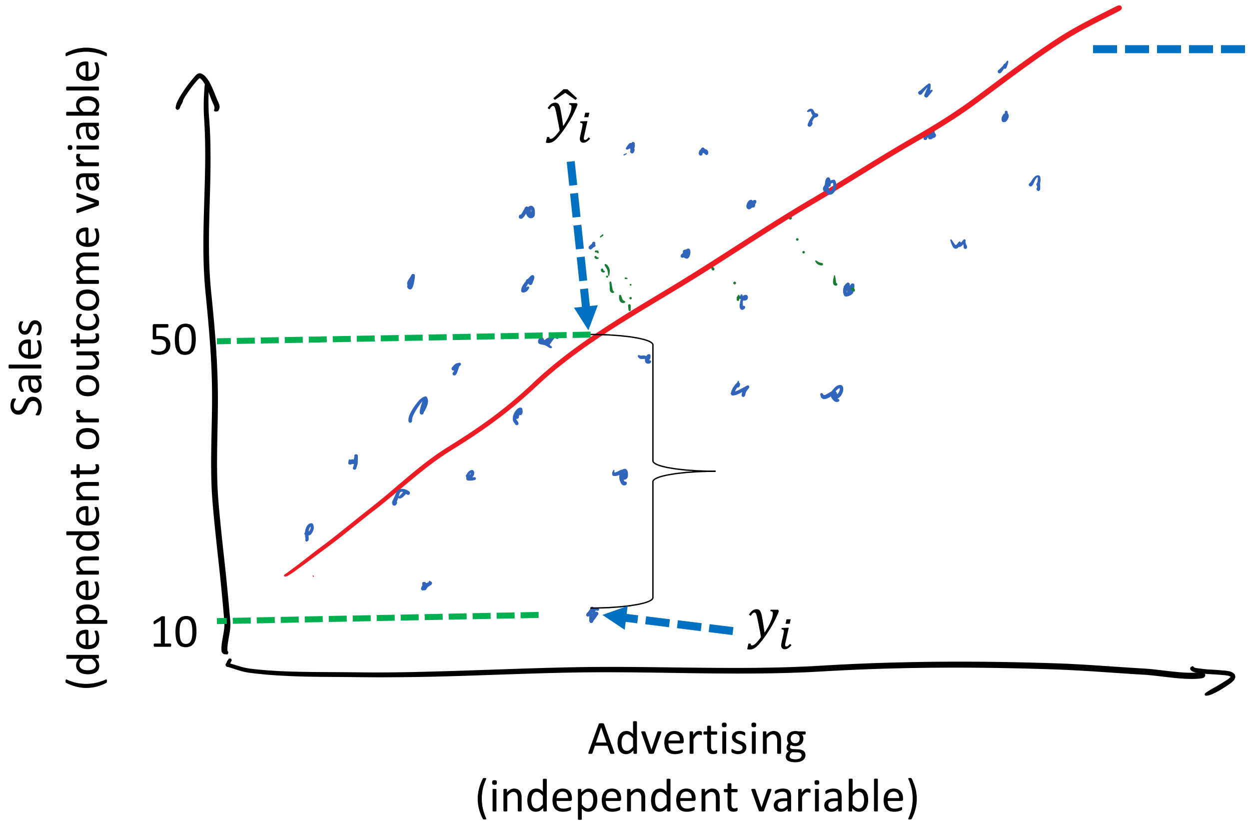
Suppose we spend 100m on advertising?
What's our expected sales?

$$? = \widehat{\beta_0} \;\; + \widehat{\beta_1}100m$$

$$? = 10 \;\; + 1 * 100$$
$$110 = 10 \;\; + 1 * 100$$

"Hat", e.g. $\widehat{\beta_0}$, means we've estimated this relationship from data.

Measuring Errors



$$y_i = \beta_0 + \beta_1 x_{Advertising} + \epsilon_i$$

Errors: $\epsilon_i = y_i - \hat{y}_i$

Error: $\hat{\epsilon}_i = 10 - 50 = -40$

Errors are the difference between what we predict ($\hat{y}_i$) and the actual values (y_i).

Class 7: Outline

1. Review Bias, Variance, Overfit, Underfit
2. Linear Regression Review
3. **Estimating Linear Models in R**
4. Interpreting Model Coefficients
5. Regression Lab

Model Formulas in R

- Formulas in R start with the dependent variable on the left hand side (LHS)
- Followed by ”~” tilde
- Then all dependent variables separated by plus signs

```
>
>
>
> data(mpg)
> hwy ~ year + displ + cyl
hwy ~ year + displ + cyl
>
```

- **The above translates to a regression equation of:**
- $\boldsymbol{hwy = \beta_0 + \beta_1 \cdot year + \beta_2 \cdot displ + \beta_3 \cdot cyl}$

Estimating Linear Models Using lm()

- Estimate a linear model using the ‘lm()’ function in R
- We must pass the dataset on which to estimate our model
- Then we store the regression model as ‘mod1’ (or whatever name you like
- Summary() outputs a summary of the estimated model

```
# estimate a linear model with displacement, and
# cycl on the RHS, and hwy as the
# development variable (LHS)
# Use the 'mpg' dataframe to estimate the model
# and store the regression equation as 'mod1'
mod1 <- lm(hwy ~ displ + cyl,
           data = mpg)

# print out a summary of the linear model
summary(mod1)

# or just view the whole "list" object of
# the model results
str(mod1)
```

Viewing Regression Output Using “Summary”

```
> summary(mod1)

Call:
lm(formula = hwy ~ displ + cyl, data = mpg)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.5098 -2.1953 -0.2049  1.9023 14.9223 

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)    
(Intercept)  38.2162     1.0481  36.461 < 0.0000000000000002 ***
displ        -1.9599     0.5194  -3.773             0.000205 ***
cyl          -1.3537     0.4164  -3.251             0.001323 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.759 on 231 degrees of freedom
Multiple R-squared:  0.6049,	Adjusted R-squared:  0.6014 
F-statistic: 176.8 on 2 and 231 DF,  p-value: < 0.00000000000000022
```

Coefficient standard errors

Estimated Coefficients or “betas”

Independent variables

Coefficient T-Statistic

P-values for coefficients

R^2, or “coefficient of determination” (model fit)

Making “Pretty” Version of Regression Output Table

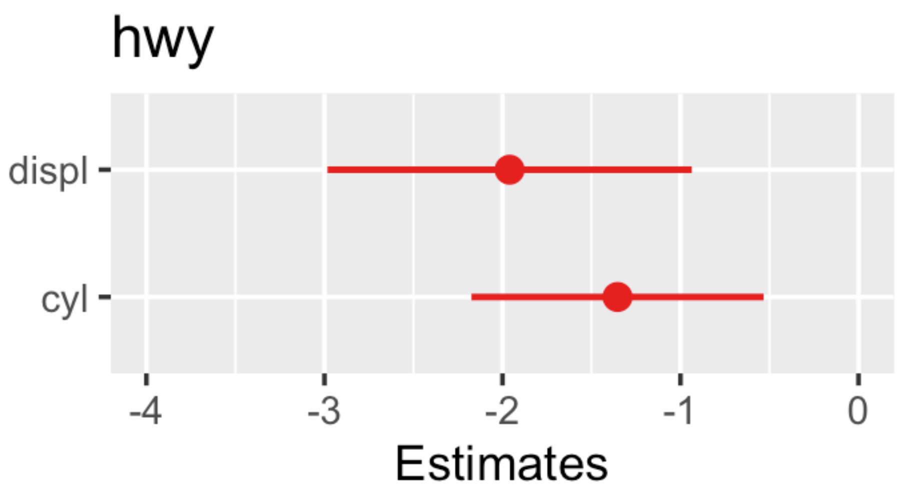
```r
# install.packages('sjPlot')
library('sjPlot')
# output a prettier table of results
# looks very nice in RMarkdown!
tab_model(mod1)

# output a plot of regression coefficients
plot_model(mod1)

# output a table of nice coefficients
tidy(mod1)
```

```
# A tibble: 3 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)    38.2      1.05      36.5  8.57e-98
2 displ          -1.96     0.519     -3.77 2.05e- 4
3 cyl            -1.35     0.416     -3.25 1.32e- 3
>
```

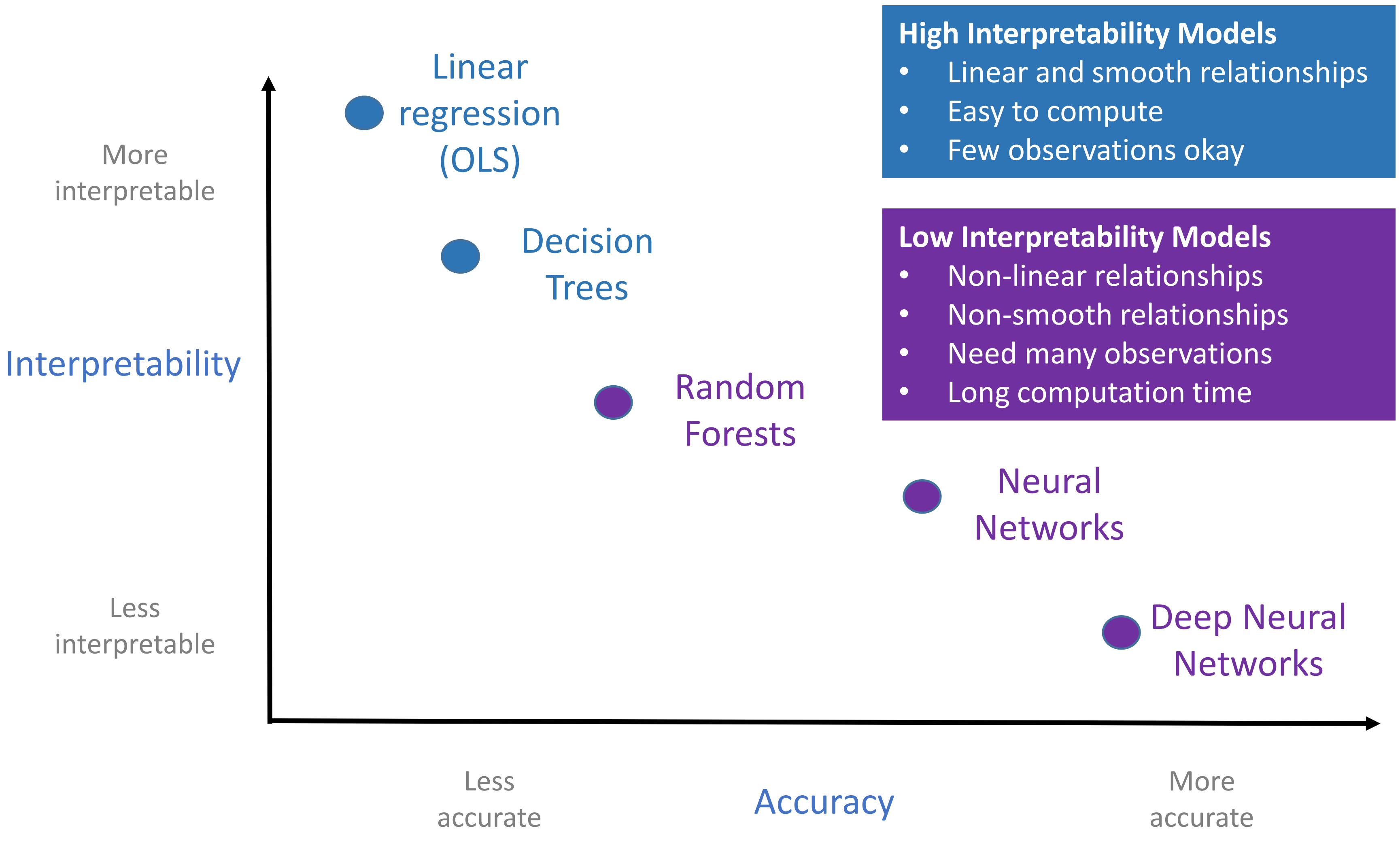
	hwy		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	38.22	36.15 – 40.28	**<0.001**
displ	-1.96	-2.98 – -0.94	**<0.001**
cyl	-1.35	-2.17 – -0.53	**0.001**
Observations	234		
R^2 / R^2 adjusted	0.605 / 0.601		



Class 7: Outline

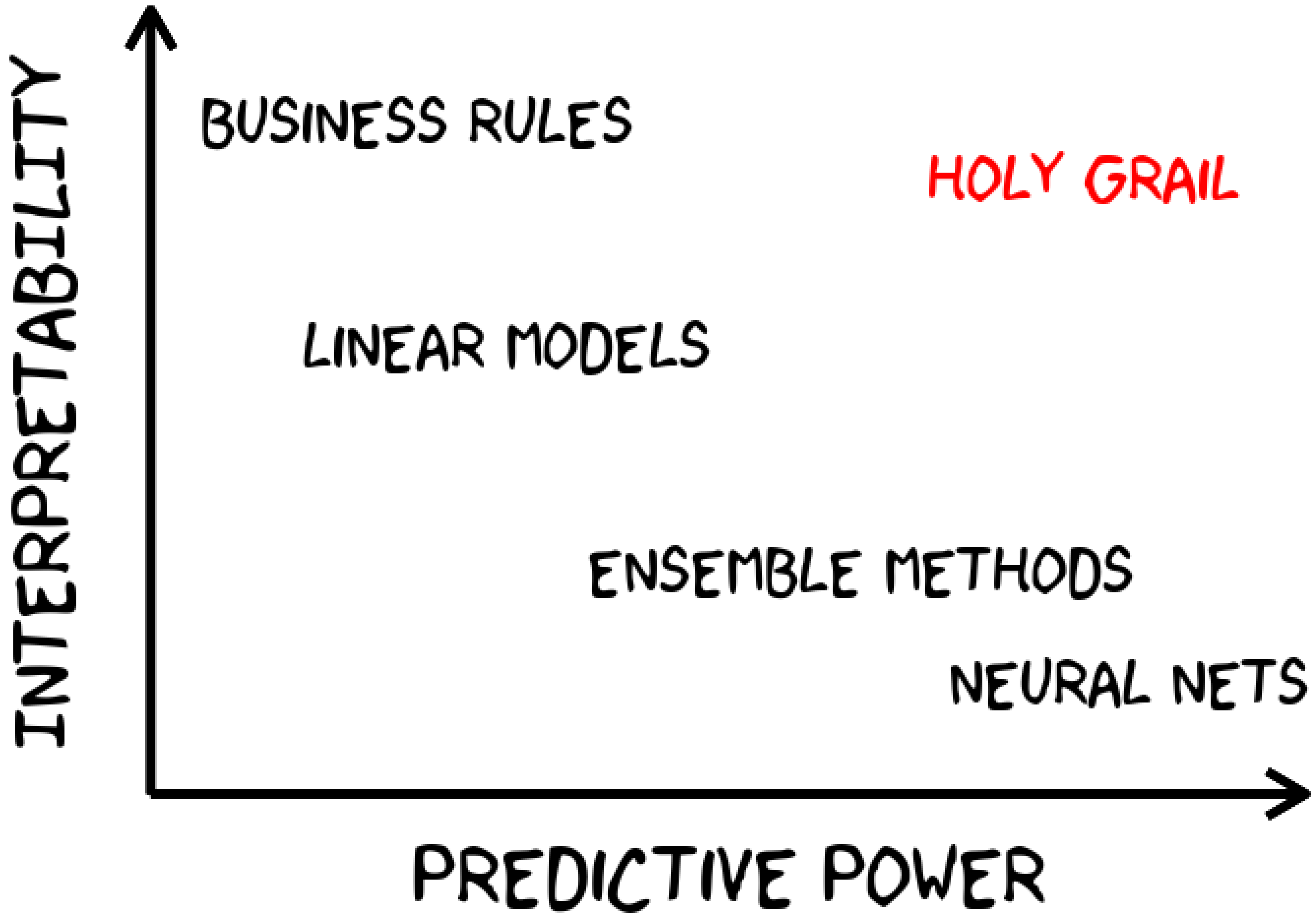
1. Review Bias, Variance, Overfit, Underfit
2. Linear Regression Review
3. Estimating Linear Models in R
4. **Interpreting Model Coefficients**
5. Regression Lab

What Is Model Interpretability?



- **Model interpretability:**
  - “the degree to which a human can understand the cause of a decision” (Miller, 2017)
- The higher the interpretability, the easier it is for someone to comprehend why a decision has been made

Of Course We Care About Both!



Why Do We Care About Model Interpretability?

1. **Strengthen Trust and Transparency**
   - People trust things they can understand, and don’t trust things they don’t (5G)

2. **Explain decisions**
   - An interpretable model allows humans to understand the proposed decision, and diagnose and analyzed the solution

3. **Regulatory Requirements**
   - Certain regulatory schemes (GDPR, Anti-Discrimination) require transparency.

4. **Improve the models**
   - Interpretability ensures the model is right or wrong for the right reasons. Interpretability offers new feature engineering and helps debugging.

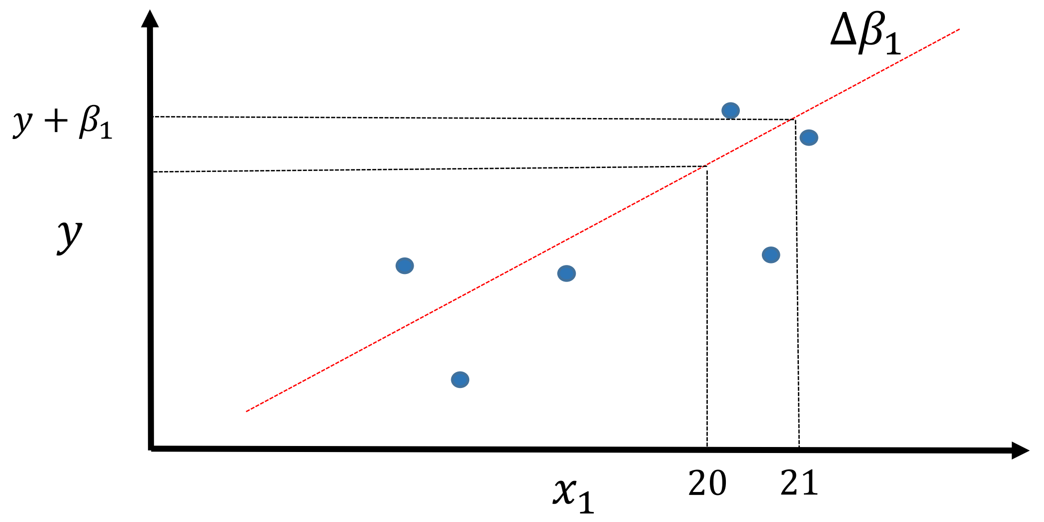
Interpreting Linear Model Coefficients

- $\boldsymbol{\beta_1}$ mathematically explains how y changes when we increase x_1 by one unit
- Suppose we change x_1 by one unit of x_1. By how much does y change?
- Well, it changes by exactly $\boldsymbol{\beta_1}$

$$y = \beta_0 + \boldsymbol{\beta_1} \cdot x_1 \; + \cdots + \beta_3 \cdot x_3$$

$$? = \beta_0 + \boldsymbol{\beta_1} \cdot (x_1 + 1) \; + \cdots + \beta_3 \cdot x_3$$

$$y + \beta_1 = \beta_0 + \boldsymbol{\beta_1} \cdot (x_1 + 1) \; + \cdots + \beta_3 \cdot x_3$$



Interpreting Linear Coefficients In Words

- **Communicating effect of coefficient**

  Increasing displacement by **one liter** (communicate units!) **decreases** **highway mile per gallon** (y variable) by -**1.96** **miles per gallon**

  - **X-variable**
  - **X-variable units**
  - **Direction (pos/neg)**
  - **Y-variable (outcome)**
  - **Estimated coefficient (magnitude)**
  - **Y-units**

```
> summary(mod1)

Call:
lm(formula = hwy ~ displ + cyl, data = mpg)

Residuals:
    Min      1Q  Median      3Q     Max
-7.5098 -2.1953 -0.2049  1.9023 14.9223

Coefficients:
            Estimate Std. Error t value              Pr(>|t|)
(Intercept)  38.2162     1.0481  36.461 < 0.0000000000000002 ***
displ        -1.9599     0.5194  -3.773             0.000205 ***
cyl          -1.3537     0.4164  -3.251             0.001323 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.759 on 231 degrees of freedom
Multiple R-squared:  0.6049,    Adjusted R-squared:  0.6014
F-statistic: 176.8 on 2 and 231 DF,  p-value: < 0.00000000000000022
```

DO NOT JUST SAY WHEN X GOES UP Y GOES UP
THIS IS OBVIOUS AND YOU WILL GET FIRED

Class 7: Outline

1. Review Bias, Variance, Overfit, Underfit
2. Linear Regression Review
3. Estimating Linear Models in R
4. Interpreting Model Coefficients
5. **Regression Lab**

Class 7 Lab

lab_class_7_linear_regression.R

Source on Save

```
#------------------------------------------------------------
# Basic Linear Regression
#------------------------------------------------------------
# remove all existing objects in memory
rm(list = ls())

# load these libraries
library('tidyverse')


#------------------------------------------------------------
# Formulas in R
#------------------------------------------------------------
# formulas in R start with the dependent variable on the
# left hand side (LHS), then by a "~" tilde, following by
# all the dependent variables you wish to estimate on the
# right hand side (RHS)

# e.g. y ~ x1 + x2

data(mpg)
hwy ~ year + displ + cyl
```

1. Estimate a regression model of city mpg on year, displacement, and engine cylinders and store this as 'mod3’

2. Interpret in words the coefficient for year

3. Interpret in words the coefficient for engine cylinders

4. Upload answers to 1-3 to Canvas (.R code with answers in comments is fine.)

5. If you finish and still have time, try using 'plot_model()’ 'tab_model' and 'tidy' on 'mod3' (may need to load/install the packages tidymodels and sjPlot)

Class 7 Summary

- Regression estimates a relationship between an outcome (y) and one or more predictor variables (Xs)
- Linear regression or OLS restricts these relationships to be linear
- Estimate linear models in R using the lm() package
- Model interpretability means we can easily communicate why a model makes certain choices
- We should strive to build the most interpretable models whose accuracy is acceptable
- To interpret a coefficient in words we must state the X variable, its units, the direction of the effect, and the magnitude of the effect in appropriate units