Week 3

BUS 696: Machine Learning for Managers

Prof. Jonathan Hersh

BUS 696: Week 3 Outline

1. **AI/ML in the News**
2. Bias Variance Tradeoff
3. Testing/Training split of Data
4. Creating testing/training splits in R
5. Simple Linear Regression
6. Simple Linear Regression in R

TECHNOLOGY

How a Feel-Good AI Story Went Wrong in Flint

A machine-learning model showed promising results, but city officials and their engineering contractor abandoned it.

ALEXIS C. MADRIGAL JAN 3, 2019

Workers in Flint, Michigan, replace a lead water-service pipe. (BILL PUGLIANO / GETTY)

BUS 696: Week 3 Outline

1. AI/ML in the News
2. **Bias Variance Tradeoff**
3. Testing/Training split of Data
4. Creating testing/training splits in R
5. Simple Linear Regression
6. Simple Linear Regression in R

Recall: $\mathbf{X}$, x_i, $\mathbf{x}_j$ and $\mathbf{y}$ for Data

$\mathbf{X}$	x_i	$\mathbf{x}_j$	$\mathbf{y}$
$n \times p$	$1 \times p$	$n \times 1$	$n \times 1$

$i = 1, \ldots, n$

$j = 1, \ldots, p$

Admit	GRE	GPA
0	380	3.61
1	660	3.67
1	800	4.00
0	520	2.93
1	760	3.00

Notation and Matrix Algebra: data matrix $\mathbf{X}$

$$\mathbf{X}$$
$$n \times p$$

$$i = 1, \ldots, n$$

$$j = 1, \ldots, p$$

Admit	GRE	GPA
0	380	3.61
1	660	3.67
1	800	4.00
0	520	2.93
1	760	3.00

$n = 5 \times p = 3$

i_{th} Row of X

$$x_i$$
$$1 \times p$$

$$x_2 = [1 \quad 660 \quad 3.67]$$

Admit	GRE	GPA
0	380	3.61
1	660	3.67
1	800	4.00
0	520	2.93
1	760	3.00

j_{th} column of X

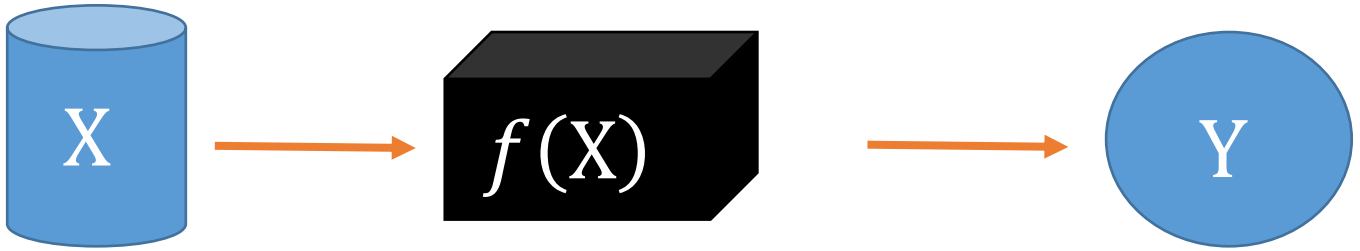
$$\underset{n\times 1}{\mathrm{x}_j}$$

$$\mathrm{x}_2 = \begin{bmatrix} 3.61 \\ 3.67 \\ 4.00 \\ 2.93 \\ 3.00 \end{bmatrix}$$

Admit	GRE	GPA
0	380	3.61
1	660	3.67
1	800	4.00
0	520	2.93
1	760	3.00

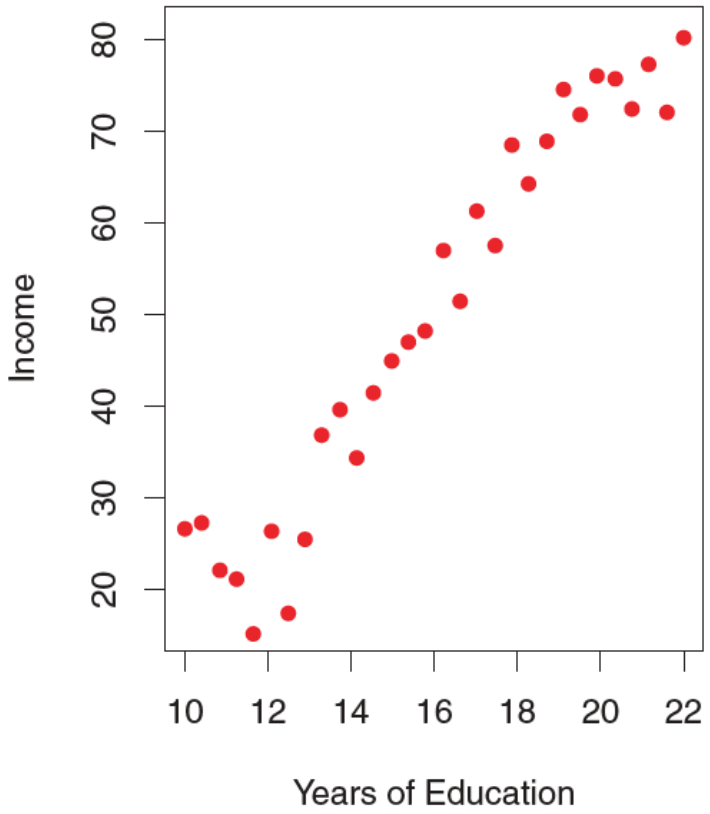
$f(\mathrm{X})$: our predicted out given inputs

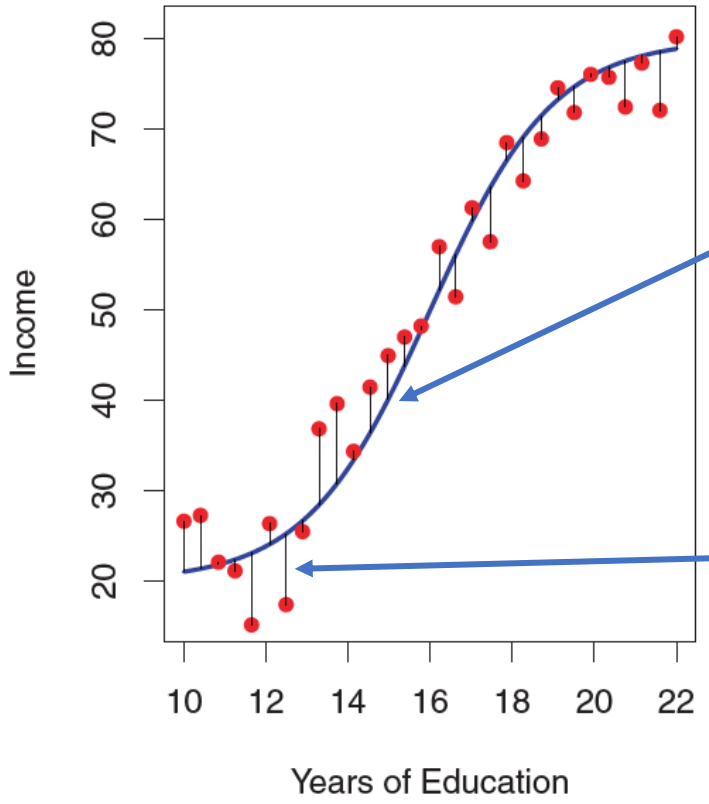
$$\mathrm{Y} = f(\mathrm{X}) + \epsilon$$



ϵ = “epsilon” (unexplained portion)

Example: education and income





Blue line is our Prediction

$\hat{y} = f(x)$

Distance to blue Line are errors

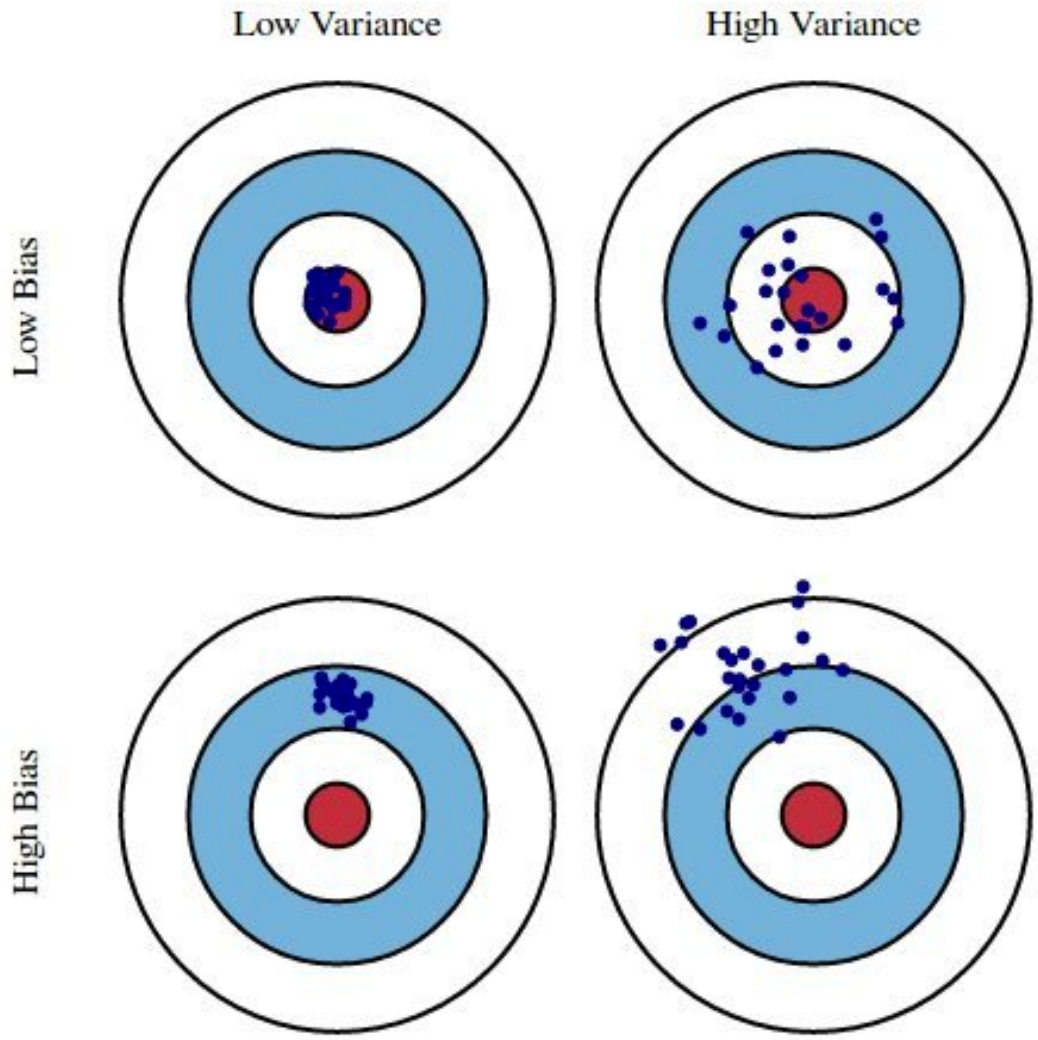
$\epsilon_i = y_i - \hat{y}_i$

"Estimating" $\hat{f}(X)$

- $\mathrm{Y} = f(\mathrm{X}) + \epsilon$ is the true value
- We can only use data to "guess" at $f(\mathrm{X})$
- We call this guess $\hat{f}(X)$
- And our guess of y is $\hat{y} = \hat{f}(X)$

How do we know when we've selected a "good" $\hat{f}(X)$?

Bias-Variance Tradeoff



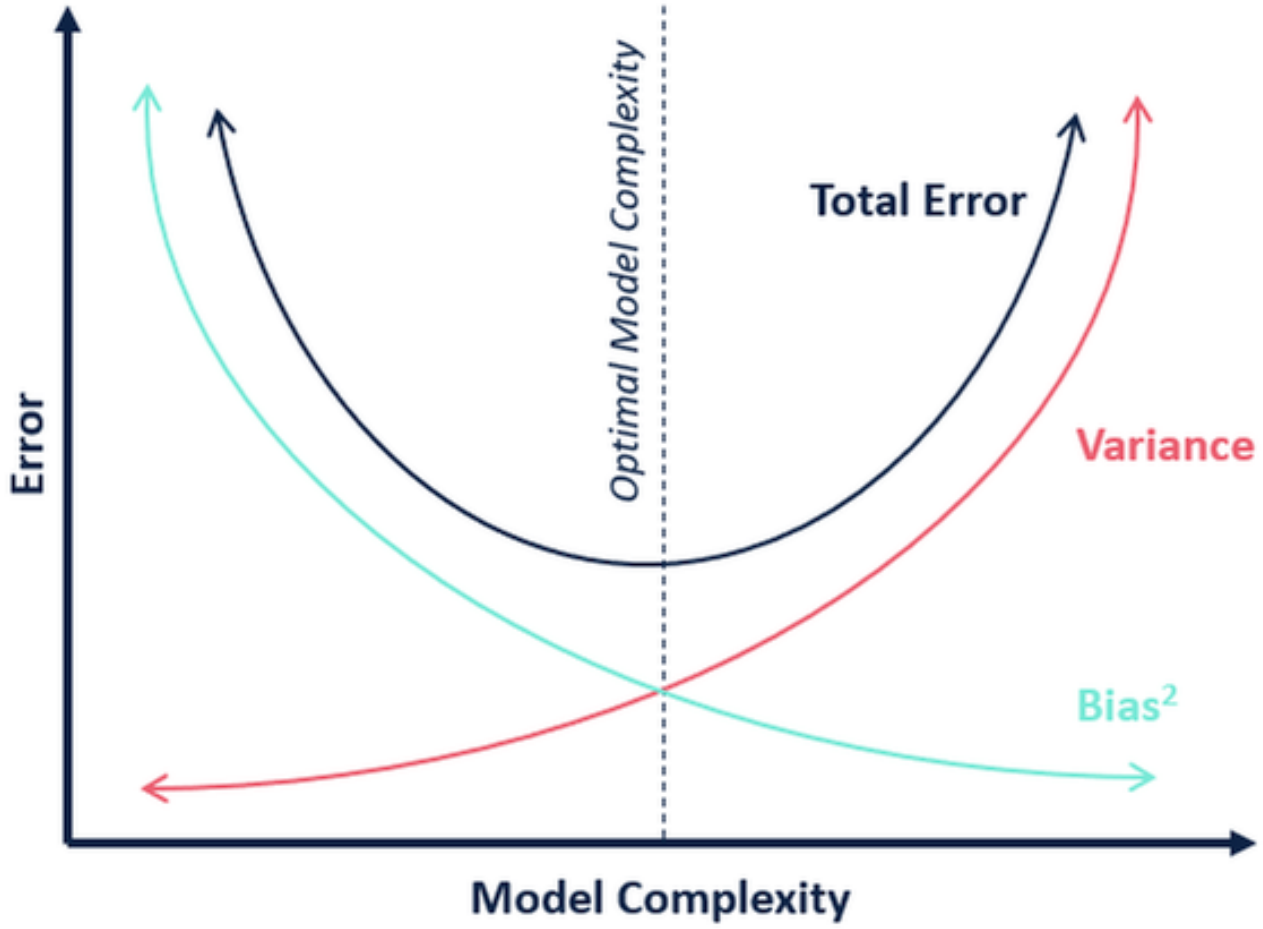
Bias and Variance (conceptually)

Bias: **Tendency of an in-sample statistic to over or under estimate the statistic in the *population***

Variance: **Tendency to noisily estimate a statistic**.

E.g., sensitivity to small fluctuations in the training dataset.

Bias-Variance Tradeoff



How to Estimate Bias and Variance? Testing/Training Split

Training set: (observation-wise) subset of data used to develop models

Testing/Training Split

Training set: (observation-wise) subset of data used to develop models

Test set: subset of data used during intermediate stages to “tune” model parameters

Rule of thumb 75% training 25% test -ish

Assessing Model Accuracy: Mean Squared Error

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

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

Mean Squared Error in Practice

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

y_i	$\hat{y}_i$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
5	5	0	$0^2 = 0$
6	7	-1	$-1^2 = 1$
9	8	1	$1^2 = 1$
10	1	9	$9^2 = 81$
14	13		

Mean Squared Error in Practice

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

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 84$$

$$\frac{1}{5}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{84}{5} = 16.8$$

y_i	$\hat{y}_i$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
5	5	0	$0^2 = 0$
6	7	-1	$-1^2 = 1$
9	8	1	$1^2 = 1$
10	1	9	$9^2 = 81$
14	13	1	$1^2 = 1$

Root Mean Squared Error

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

$$\sqrt{16.8} = 4.0987$$

Why is this useful?
Transformed
Back to units of
original y_i

y_i	$\hat{y}_i$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
5	5	0	$0^2 = 0$
6	7	-1	$-1^2 = 1$
9	8	1	$1^2 = 1$
10	1	9	$9^2 = 81$
14	13	1	$1^2 = 1$

$$\bar{y} = \frac{1}{n}\sum y_i = 8.8$$

Example: Overfitting (True Relationship Linear)

3 Models

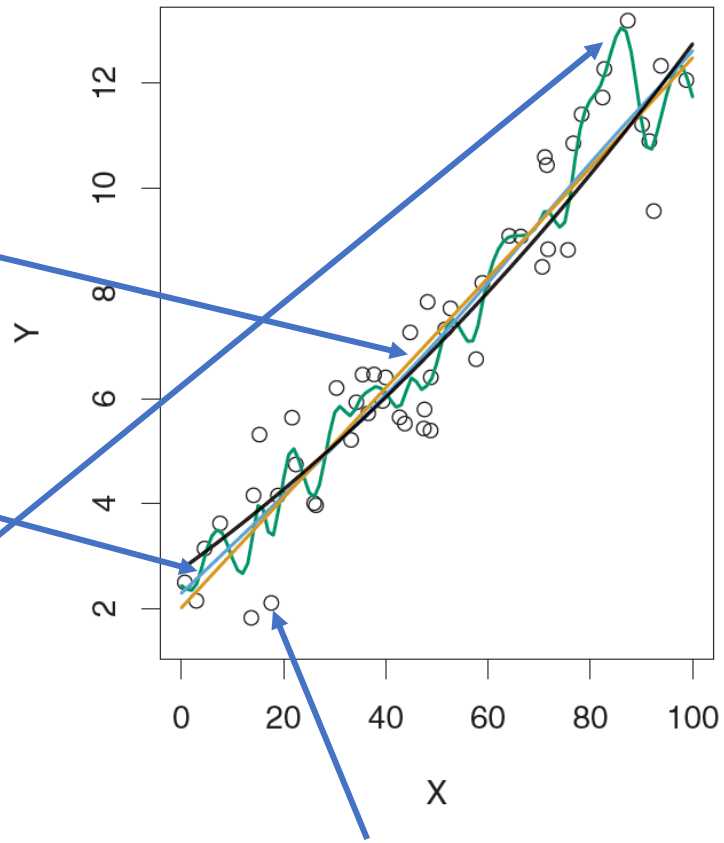
Simple (linear): gold line

Moderate complexity (linear): blue line

Very complicated: green line

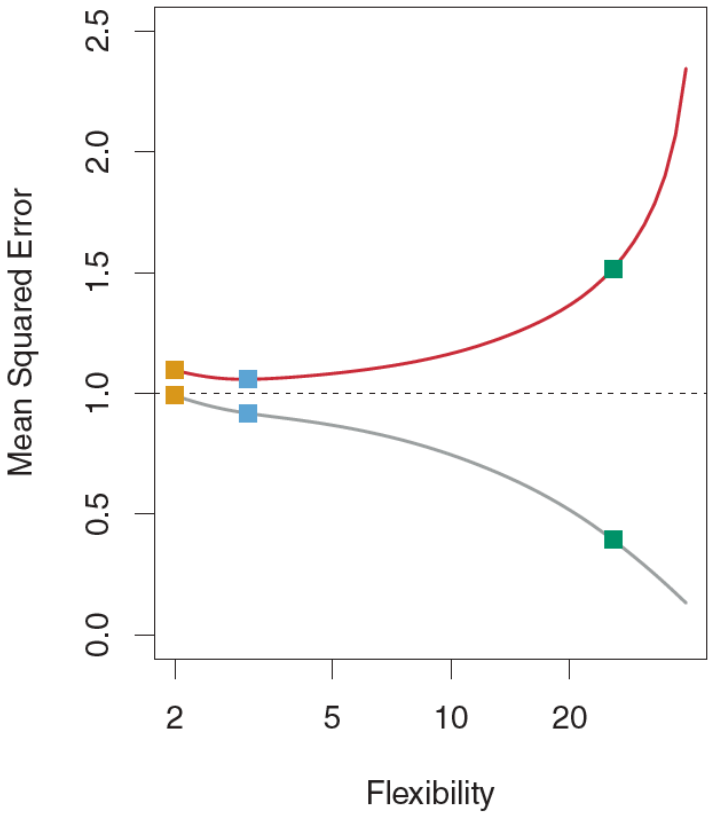
Black line: true relationship

Training Data



Data points (true relationship observed with noise)

Results on Test Data



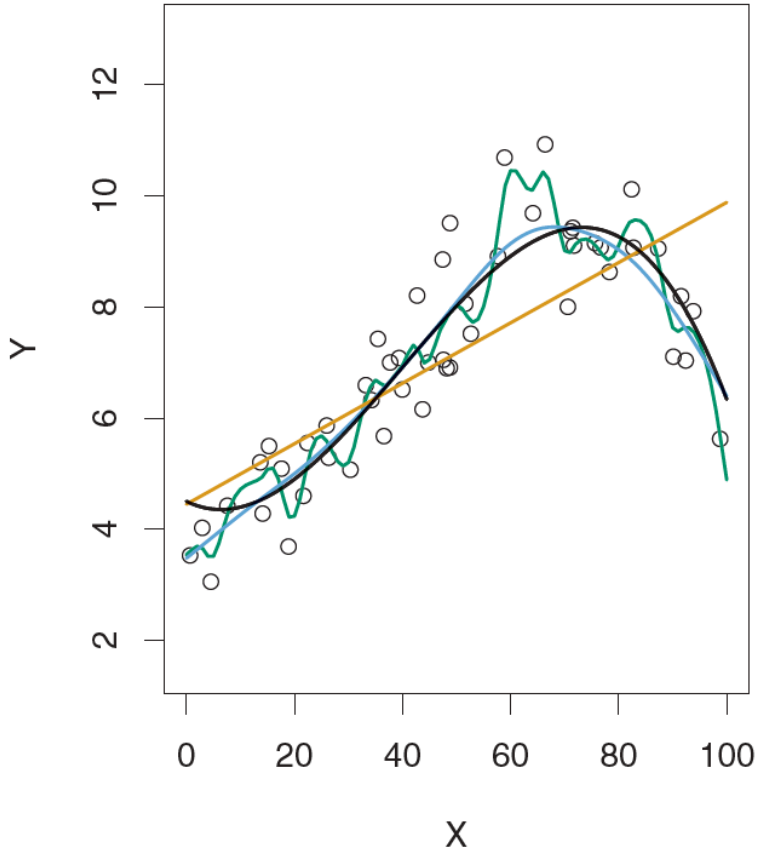
Example: Overfitting (True Relationship Slightly Complicated)

Simple (linear): gold line

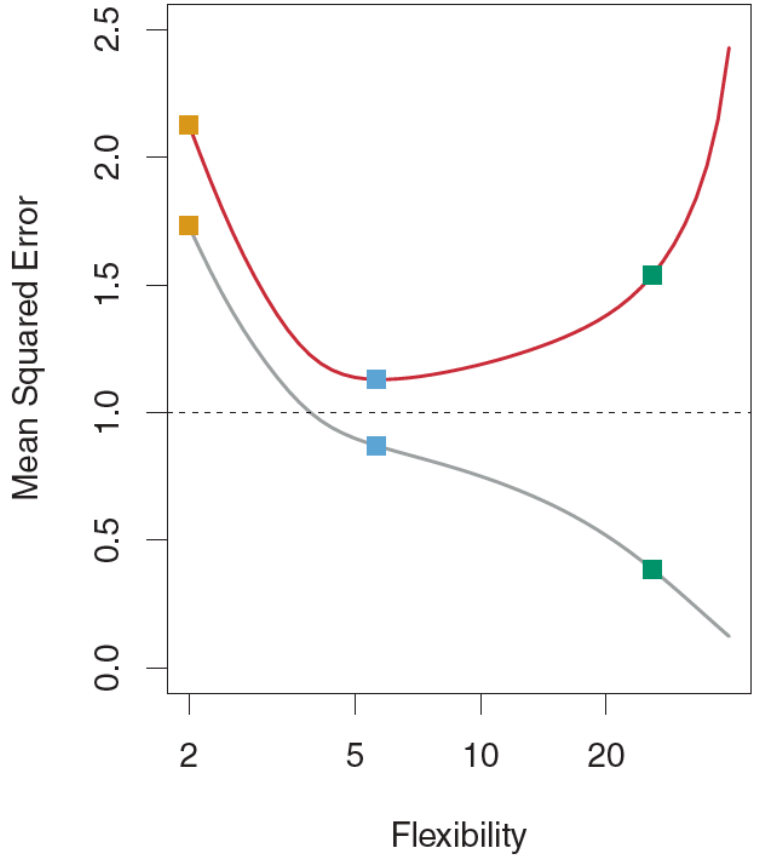
Moderate complexity (linear): blue line

Very complicated: green line

Training Data



Results on Test Data

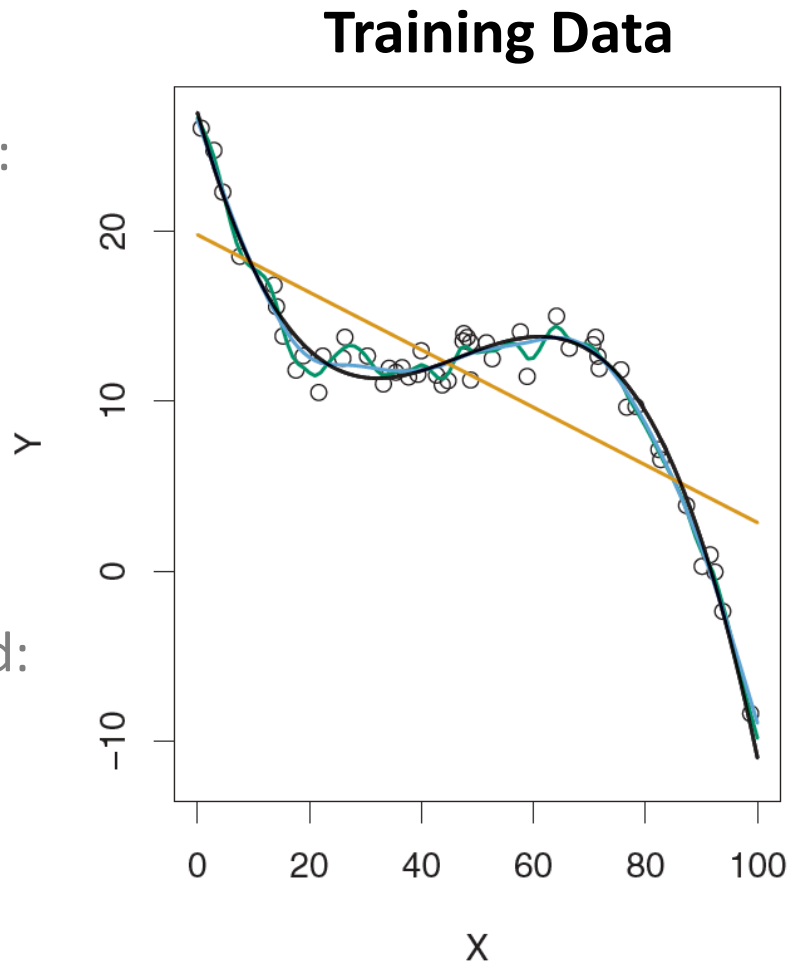


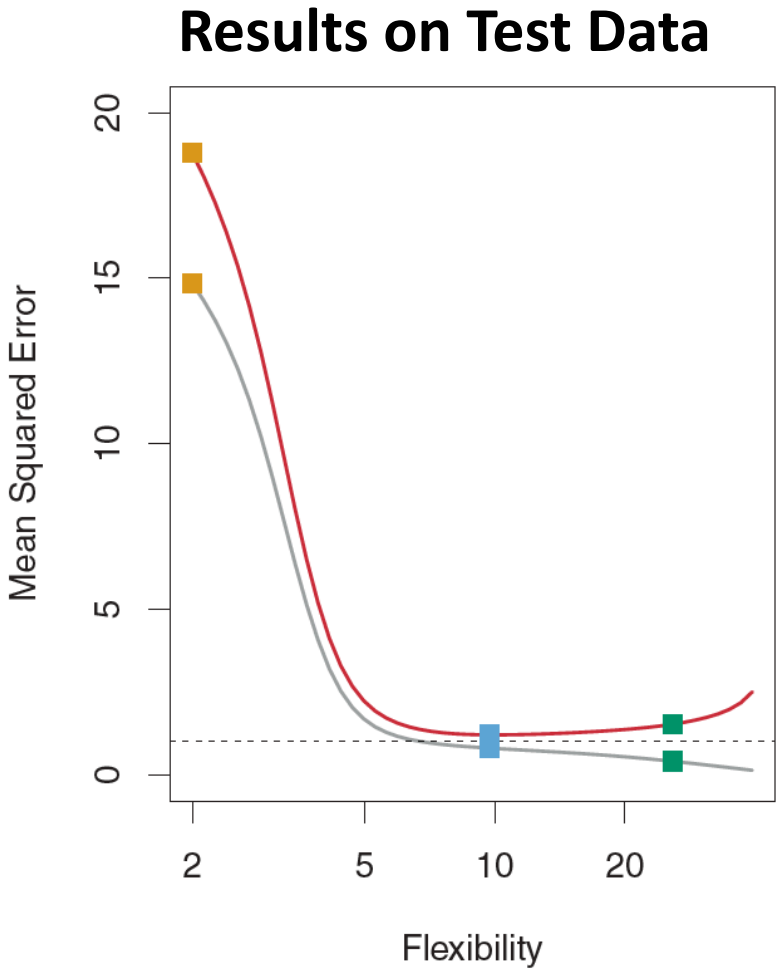
Example: Overfitting (True Relationship Very Complicated)

Simple (linear): gold line

Moderate complexity (linear): blue line

Very complicated: green line





Linear Regression

Training set: (observation-wise) subset of data used to develop models

Test set: subset of data used during intermediate stages to “tune” model parameters

Rule of thumb 75% training 25% test -ish

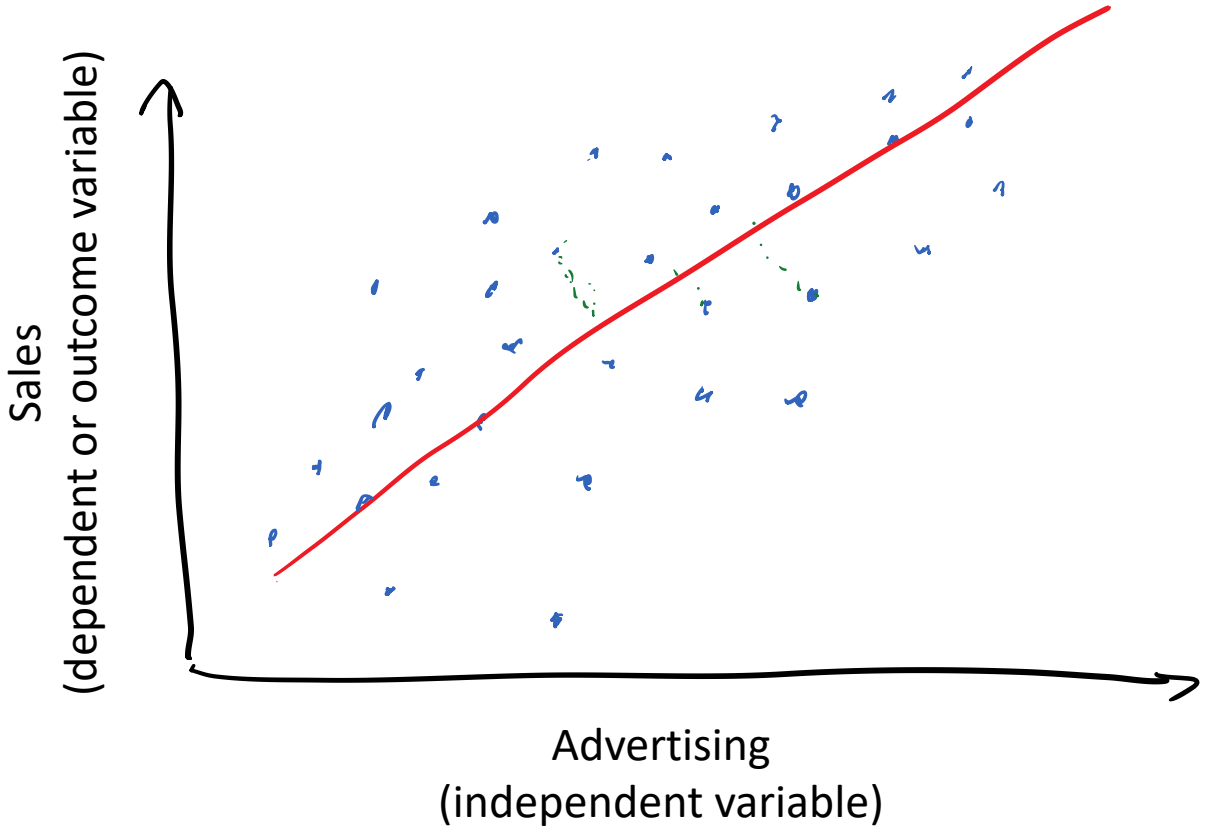
BUS 696: Week 3 Outline

1. AI/ML in the News
2. Bias Variance Tradeoff
3. Testing/Training split of Data
4. **Creating testing/training splits in R**
5. Simple Linear Regression
6. Simple Linear Regression in R

BUS 696: Week 3 Outline

1. AI/ML in the News
2. Bias Variance Tradeoff
3. Testing/Training split of Data
4. Creating testing/training splits in R
5. **Simple Linear Regression**
6. Simple Linear Regression in R

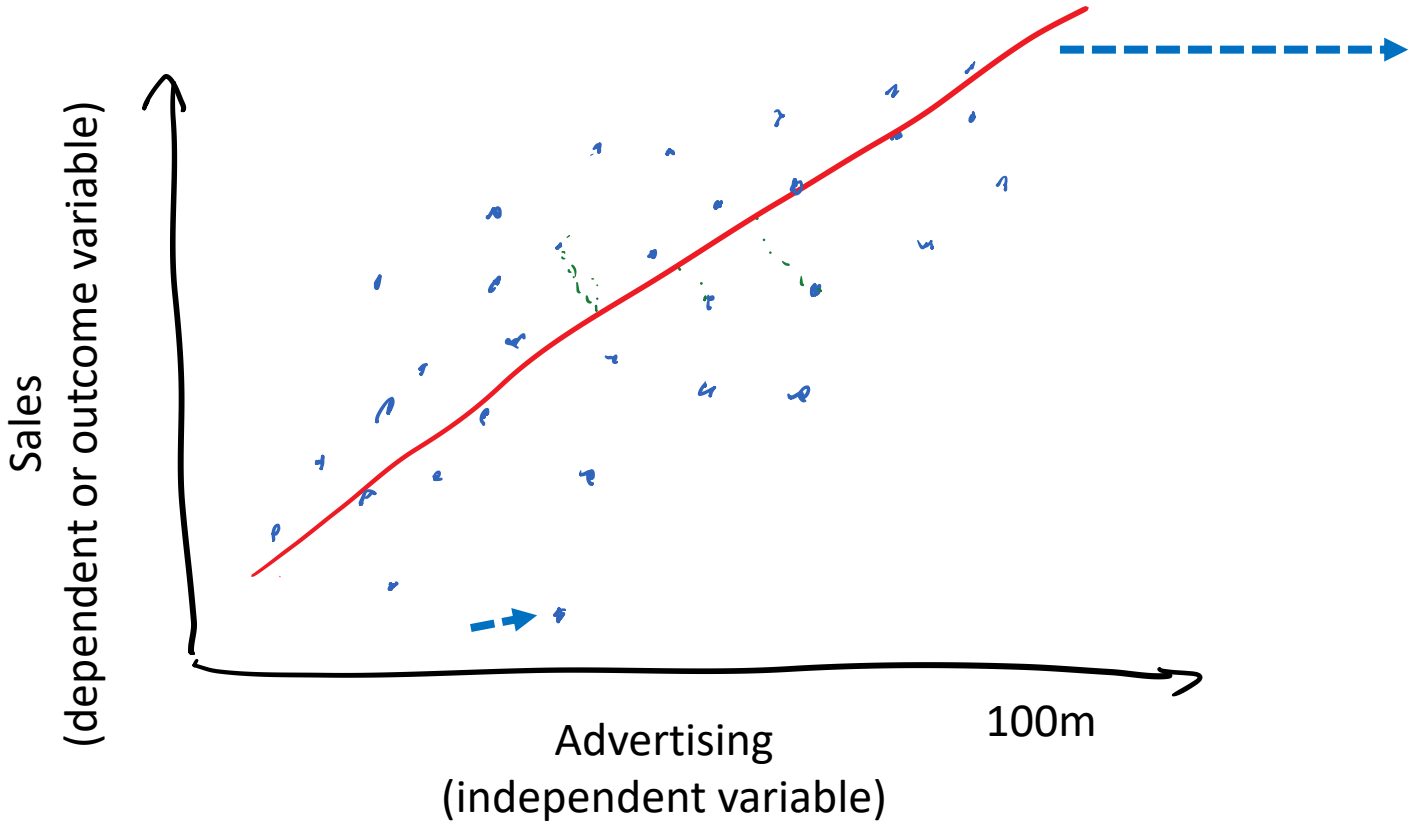
What is Linear Regression?



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

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

Linear Regression Equation

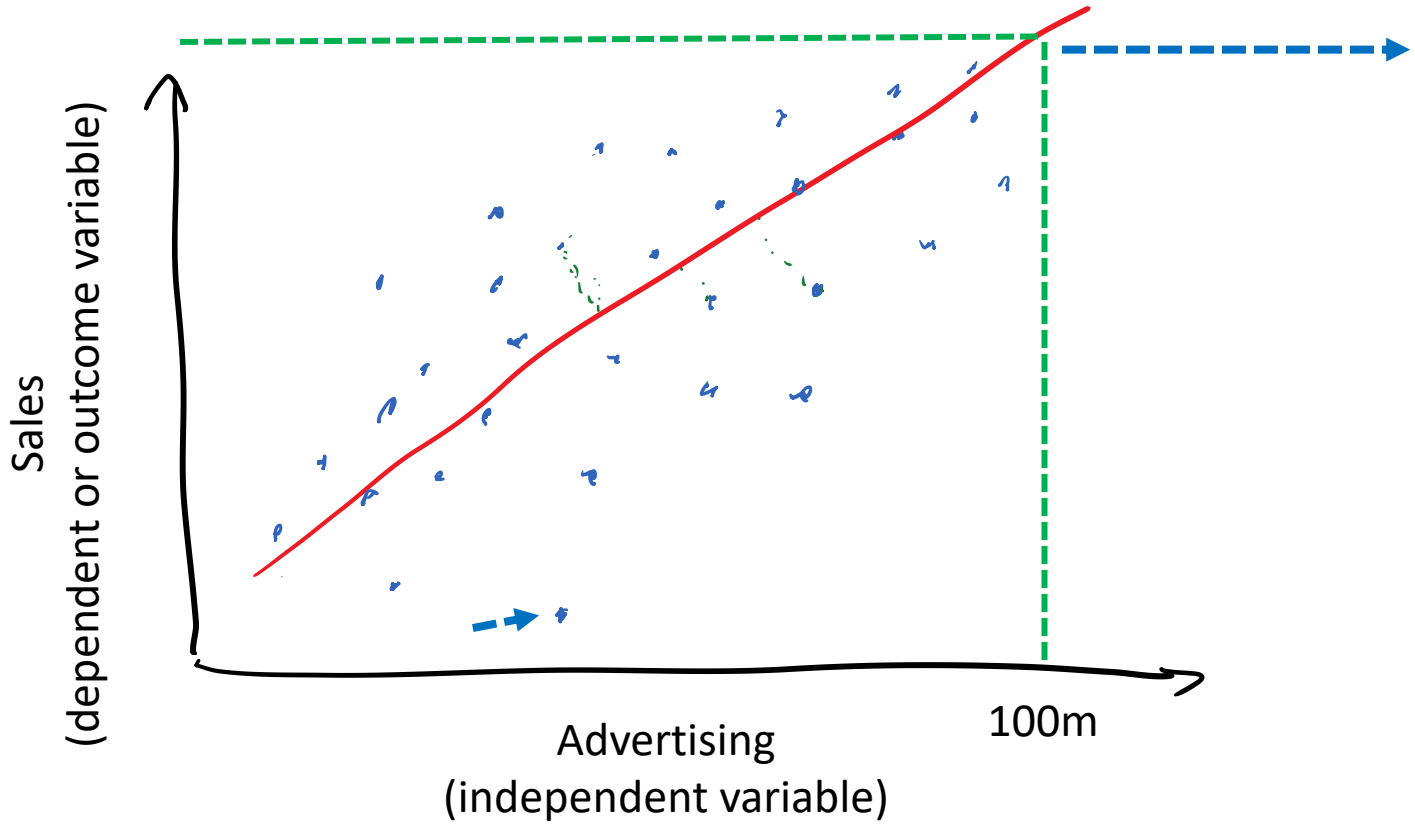


$$y_i = \beta_0 + \beta_1 x_{Advertising} + \epsilon_i$$

Red line chosen in a special way…we'll get to that

For now, it "explains" the data the best

Linear Regression as Prediction

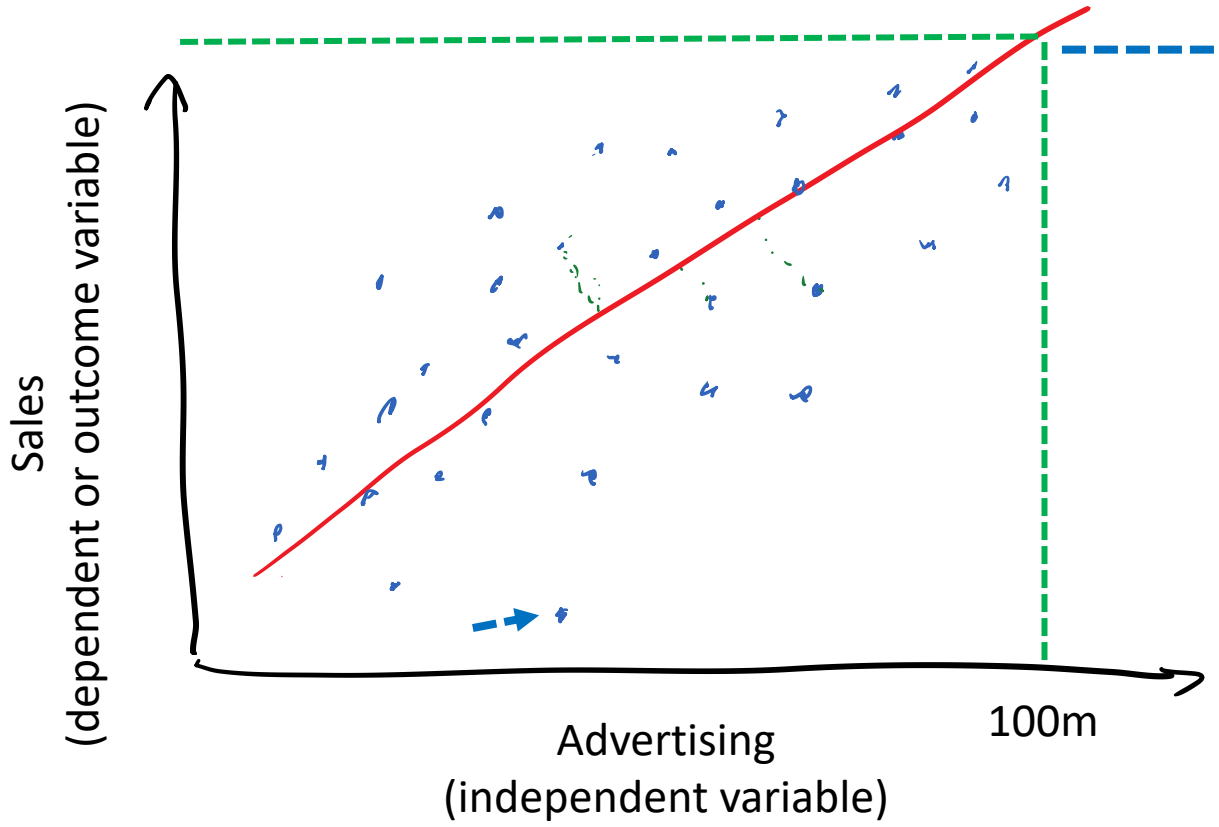


$$y_i = \beta_0 + \beta_1 x_{Advertising} + \epsilon_i$$

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

$$? = \hat{\beta}_0 + \widehat{\beta_1} 100m$$

Estimated Coefficients: $\widehat{\beta_0}$ and $\widehat{\beta_1}$



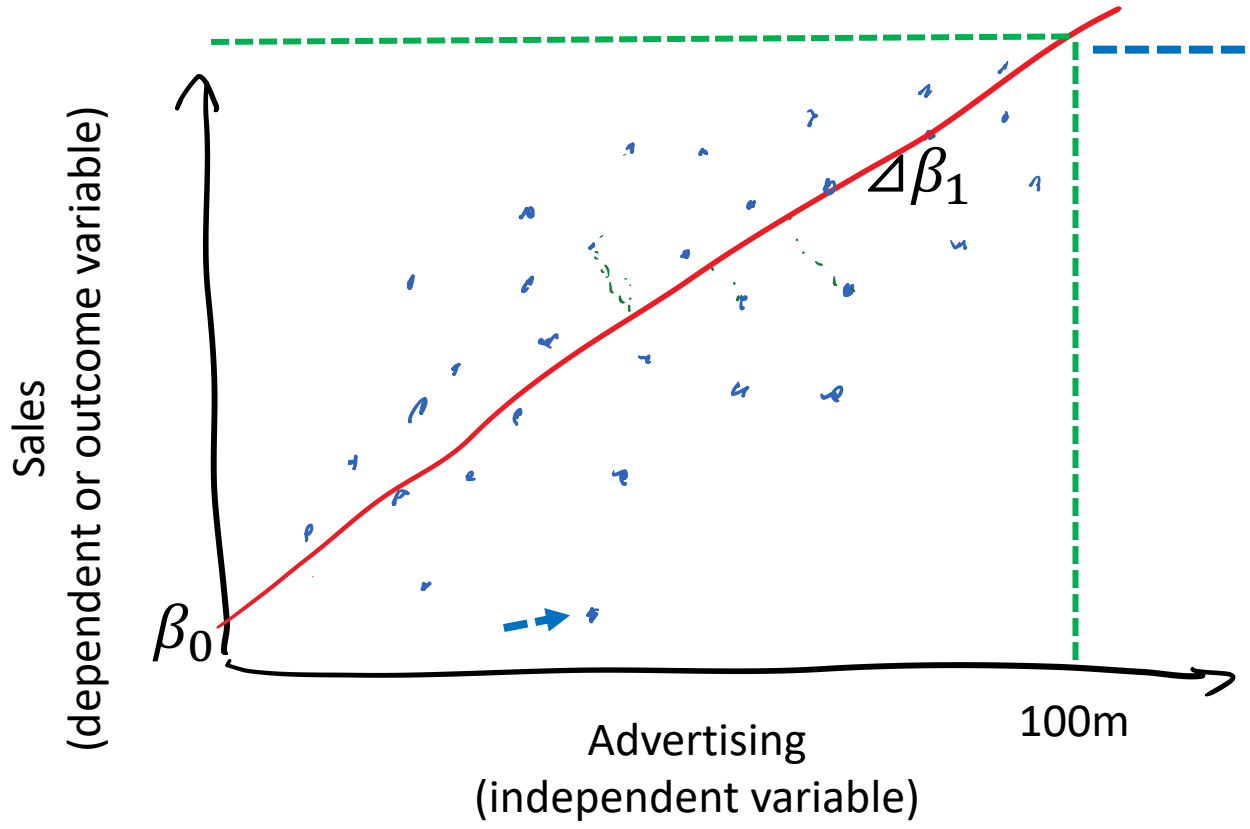
$$y_i = \beta_0 + \beta_1 x_{Advertising} + \epsilon_i$$

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.

Estimated Coefficients: $\widehat{\beta_0}$ and $\widehat{\beta_1}$



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

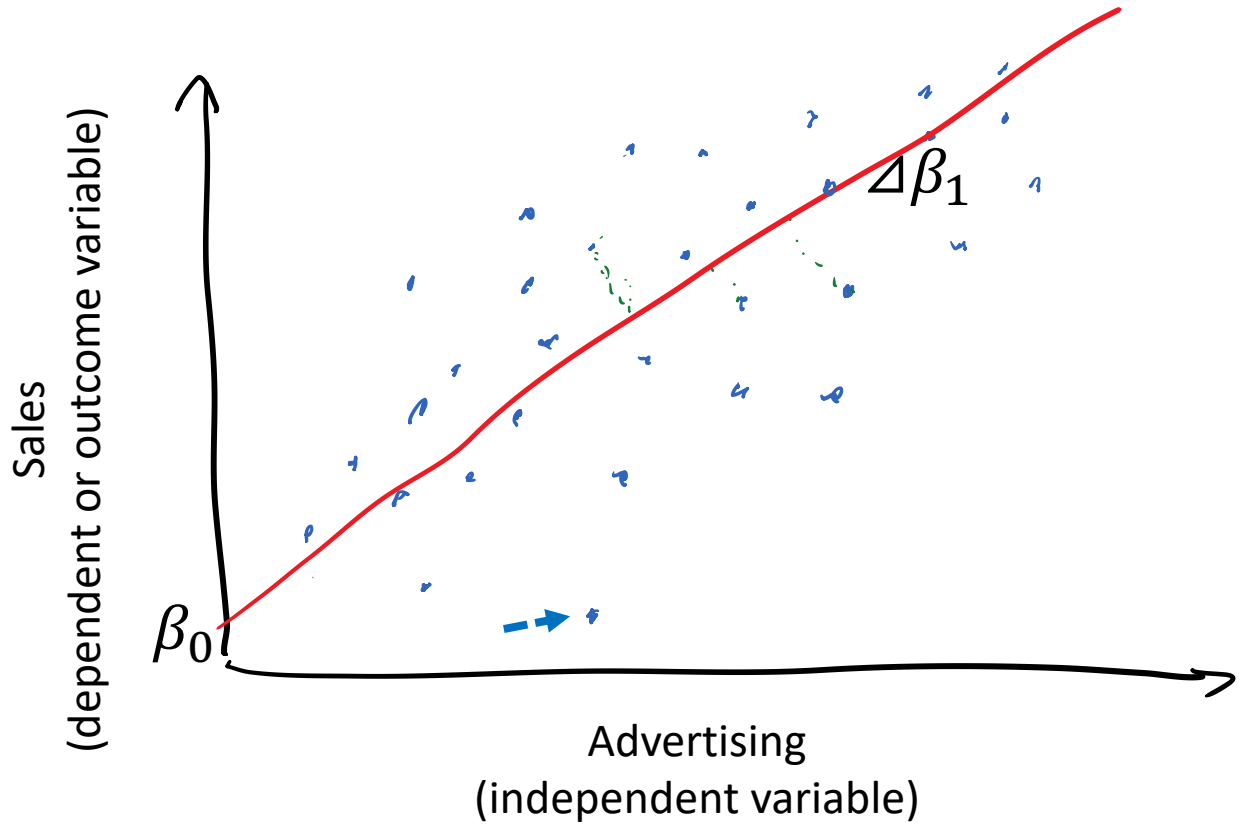
Suppose we spend 100m on advertising?
What's our expected sales?

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

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

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

A Note on Units and Interpreting β_1



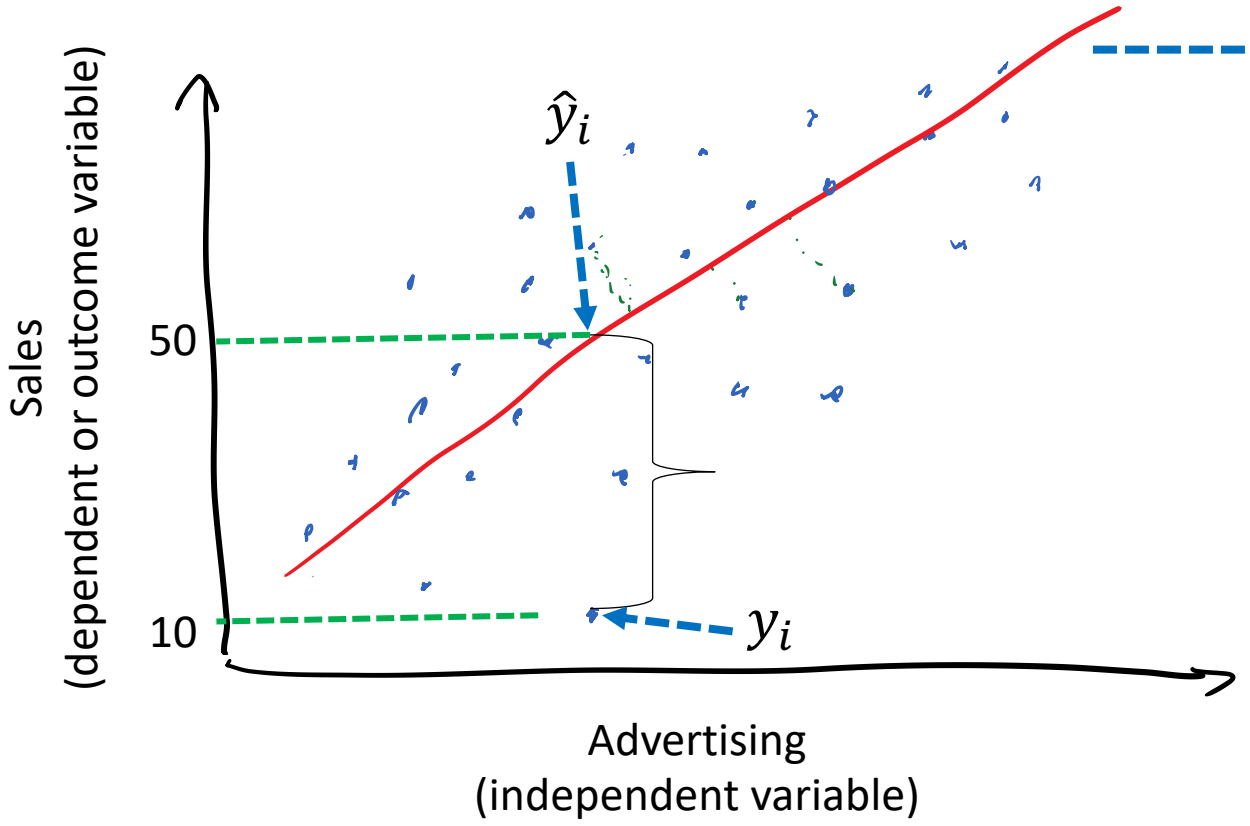
We interpret β_1 as the effect of one unit increase in x_1 leads to a β_1 unit increase in y

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

A one unit change in advertising (millions) leads to a 1 (units of y) increase in sales

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

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).

BUS 696: Week 3 Outline

1. AI/ML in the News
2. Bias Variance Tradeoff
3. Testing/Training split of Data
4. Creating testing/training splits in R
5. Simple Linear Regression
6. **Simple Linear Regression in R**

Interpreting Regression Output

```
> summary(mod1)

Call:
lm(formula = grossM ~ budgetM, data = movies_train)

Residuals:
    Min      1Q  Median      3Q     Max
-420.97  -44.31  -22.97   15.22  696.74

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 56.650257   1.293979  43.780  < 2e-16 ***
budgetM      0.030005   0.005506   5.449 5.43e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 72.32 on 3274 degrees of freedom
  (506 observations deleted due to missingness)
Multiple R-squared:  0.008988,  Adjusted R-squared:  0.008685
F-statistic: 29.69 on 1 and 3274 DF,  p-value: 5.434e-08
```

1. Residual stats
2. Coefficient Magnitude
3. Std. Error
4. T-stat (t-value)
5. P-Value
6. R-squared
7. F-Stat

Hypothesis Testing Coefficients

H_0: There is no linear relationship between X and Y

Null hypothesis: $\beta = 0$

H_1: There is some linear relationship between X and Y

P-value tells us the likelihood that, if the null hypothesis were true, we would receive a result as extreme as the one seen

What do p-values measure?

P-value tells us the likelihood that, if the null hypothesis were true, we would receive a result as extreme as the one seen

Small p-values: unlikely that we would receive a result as extreme if the null is true

P-value does not measure

- Size of effect
- Importance of a result
- Probability the alternative hypothesis is true

- It just tells us how likely we would see the coefficient we see if beta is really equal to zero

T-Stat and F-Statistic

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

- Large t-stat in abs value -> big effect size

$$F = \frac{(TSS - RSS)/p}{RSS/(n - p - 1)}$$

- F –stat tells us joint significance of model

Total Sum of Squares, Residual Sum of Squares and R-Squared

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}$$

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

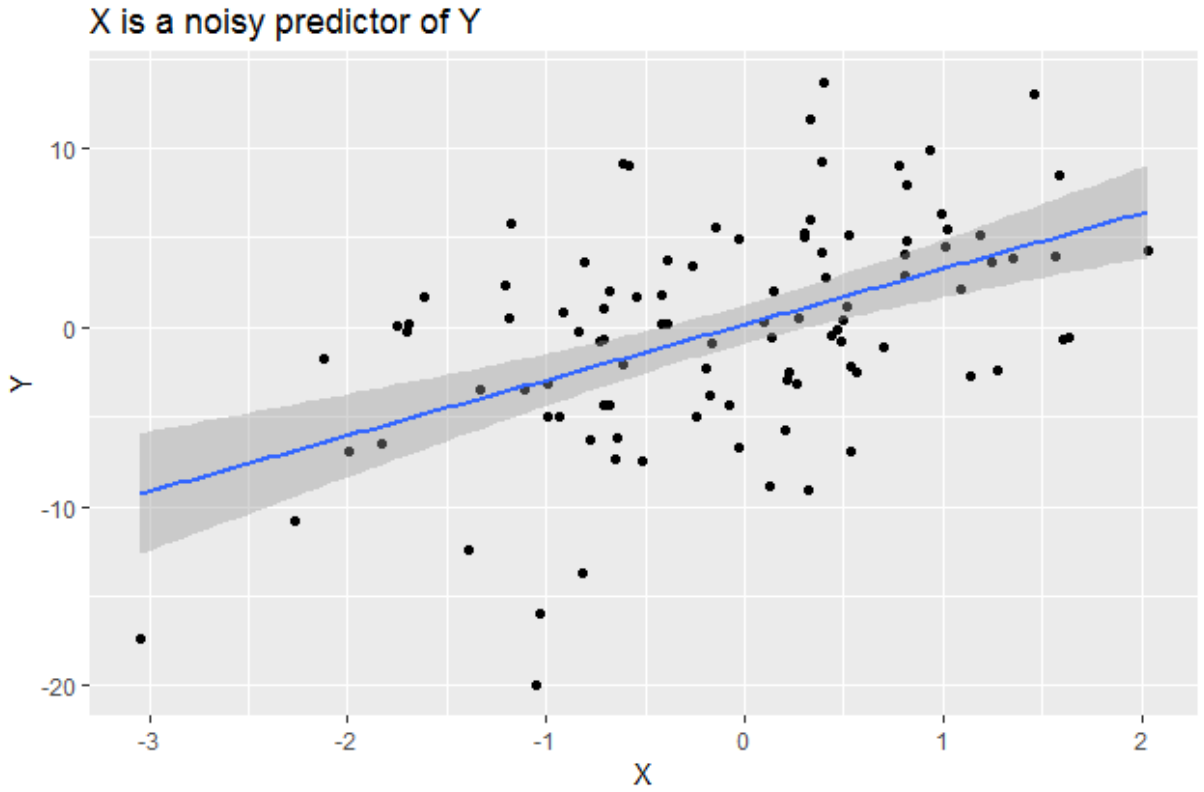
$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

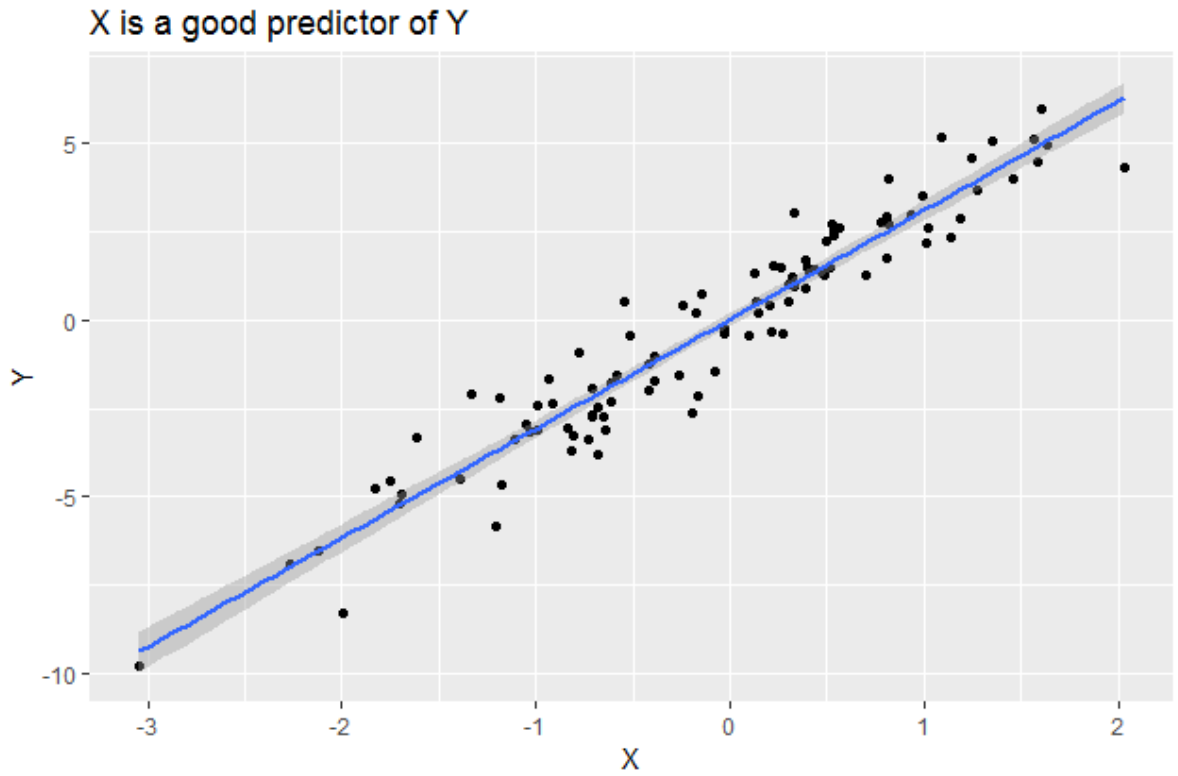
- TSS: Total Sum of Squares
- RSS: Residual Sum of Squares

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

$\Rightarrow R^2 \in [0,1]$

R^2: Measuring goodness of fit





Adding More Variables

$$gross_i = \beta_0 + \beta_1 x_{budget} + \beta_2 \cdot x_{imdb\_score} + \epsilon_i$$

$$\widehat{gross}_i = -54.5 + 0.0273 \cdot x_{budget} + 17.29 \cdot x_{imdb\_score} + \epsilon_i$$

```
Call:
lm(formula = grossM ~ budgetM + imdb_score, data = movies_train)

Residuals:
    Min      1Q  Median      3Q     Max
-398.62  -42.15  -17.19   18.57  671.86

Coefficients:
             Estimate Std. Error t value          Pr(>|t|)
(Intercept) -54.506824   7.552397  -7.217 0.000000000000657 ***
budgetM       0.027361   0.005332   5.132 0.000000303879390 ***
imdb_score   17.299975   1.159150  14.925           < 2e-16 ***
---
```

Including Qualitative Predictors (Dummy Variables)

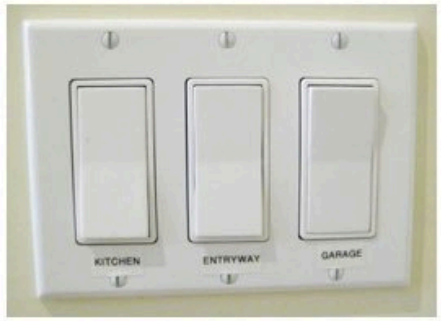
```
> DF <- data.frame(y = rnorm(5),
+                  x1 = 1:5,
+                  x2 = c("A","B","B","A","C"))
> DF
           y x1 x2
1 -0.3475685  1  A
2 -0.5332828  2  B
3  1.4417620  3  B
4 -2.0573151  4  A
5 -0.7681746  5  C
```

```
> model.matrix(y ~ x1 + x2, DF)
  (Intercept) x1 x2B x2C
1           1  1   0   0
2           1  2   1   0
3           1  3   1   0
4           1  4   0   0
5           1  5   0   1
```

- Every qualitative predictor (factor) has been transformed into a column of binary information corresponding to that factor level

Factor: Switches, Continuous: Sliders

$$gross_i = \beta_0 + \beta_1 x_{budget} + \beta_2 \cdot x_{imdb\_score} + \epsilon_i$$



$$gross_i$$
$$= \beta_0 + \beta_1 x_{StevenSpielberg} + \beta_2 \cdot x_{MichaelBay} + \epsilon_i$$

Source: https://twitter.com/andrewheiss/status/1171084259660107777?s=20

Factor: Switches, Continuous: Sliders

$$gross_i = \beta_0 + \beta_1 x_{budget} + \beta_2 \cdot x_{imdb\_score} + \epsilon_i$$

$$gross_i = \beta_0 + \beta_1 x_{budget} + \beta_2 \cdot x_{imdb_{score}} + \beta_3 \cdot MichaelBay + \beta_4 \cdot StevenSpielberg + \epsilon_i$$

Source: https://twitter.com/andrewheiss/status/1171084259660107777?s=20

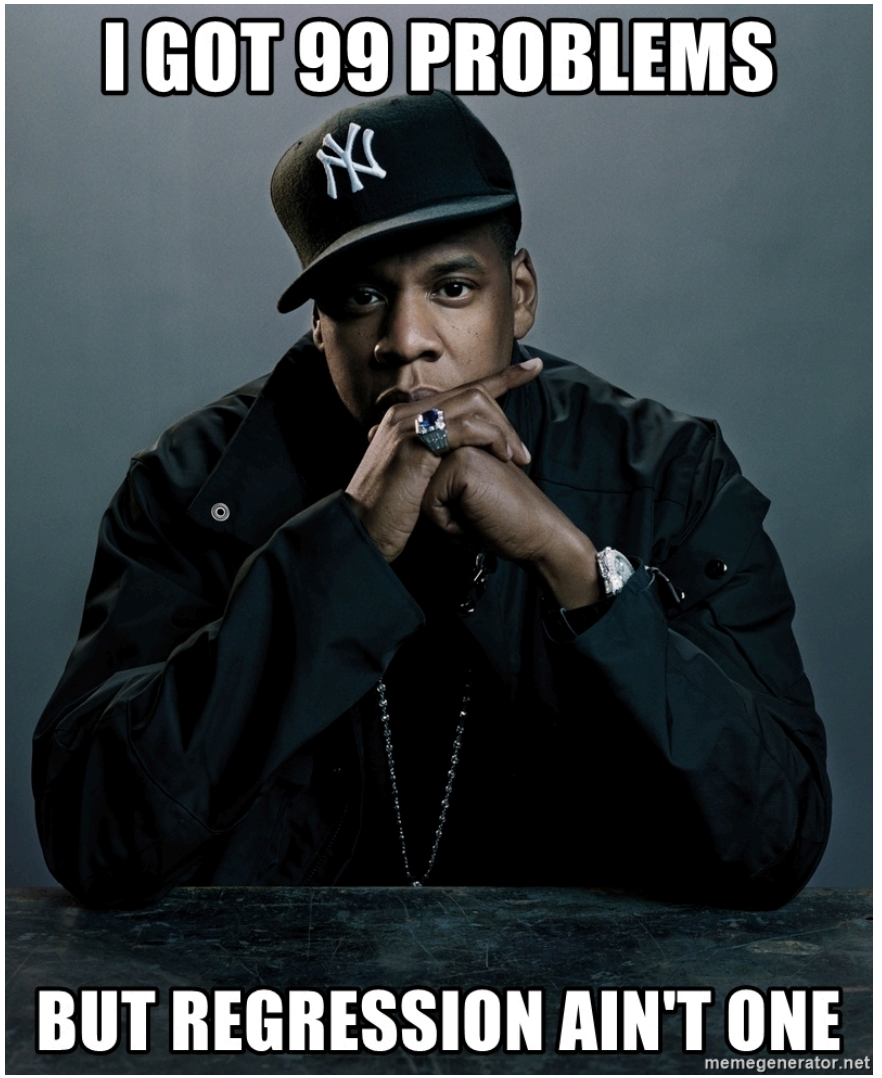
“Base” Levels and Interpreting Factors

budget	Director	Director_Michael_Bay	Director_StevenSpielBerg
14M	Werner Herzog	0	0
250M	Michael Bay	1	0
160M	Michael Bay	1	0
17M	Wes Anderson	0	0
100M	David Fincher	0	1

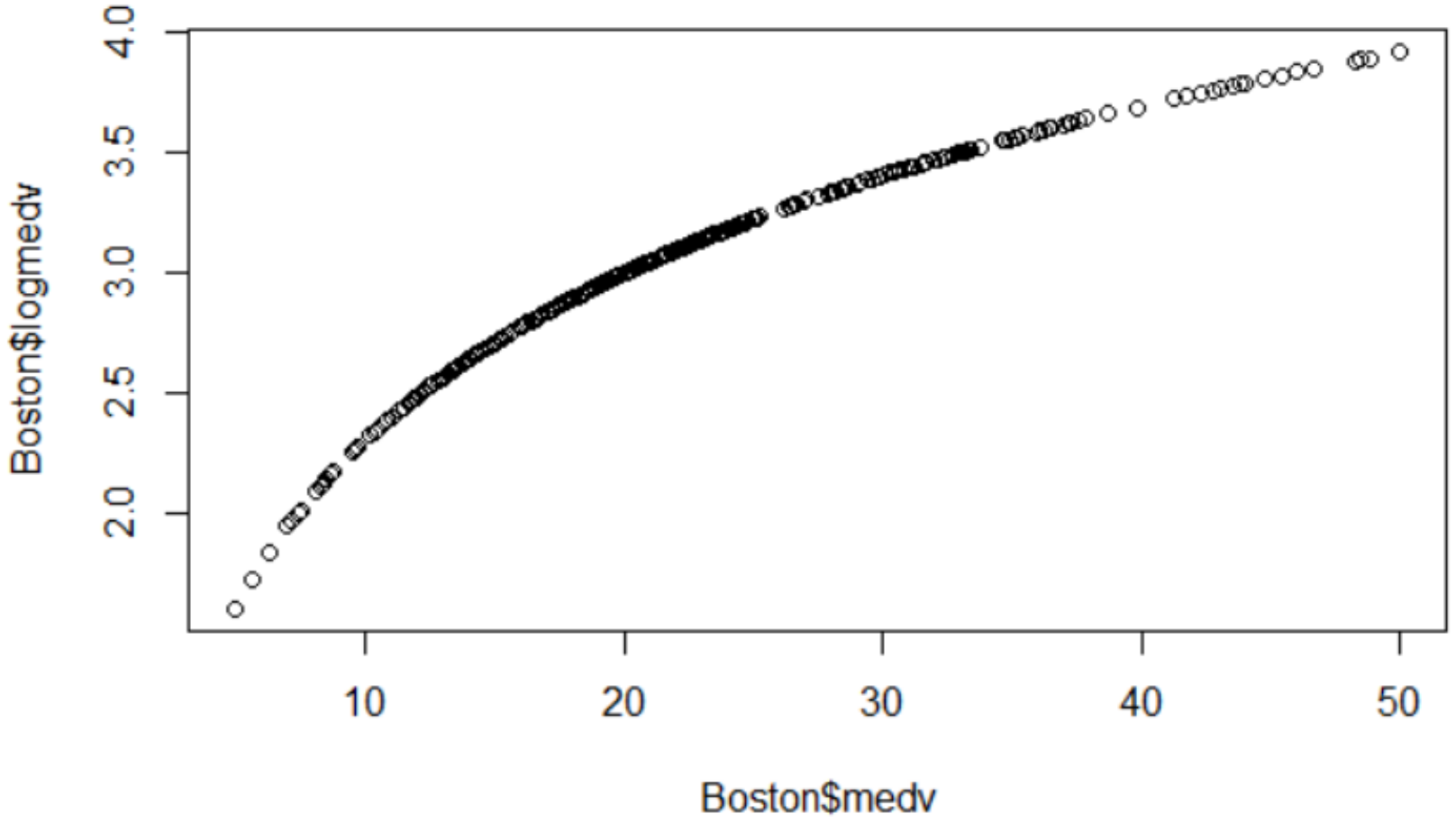
Source: https://twitter.com/andrewheiss/status/1171084259660107777?s=20

Potential Problems with Regressions

1. Non-linearity between Y and X
2. Correlation of error terms
3. Non-constant variance of error term
4. Outliers
5. High-leverage points
6. Collinearity
7. Correlation vs Causation



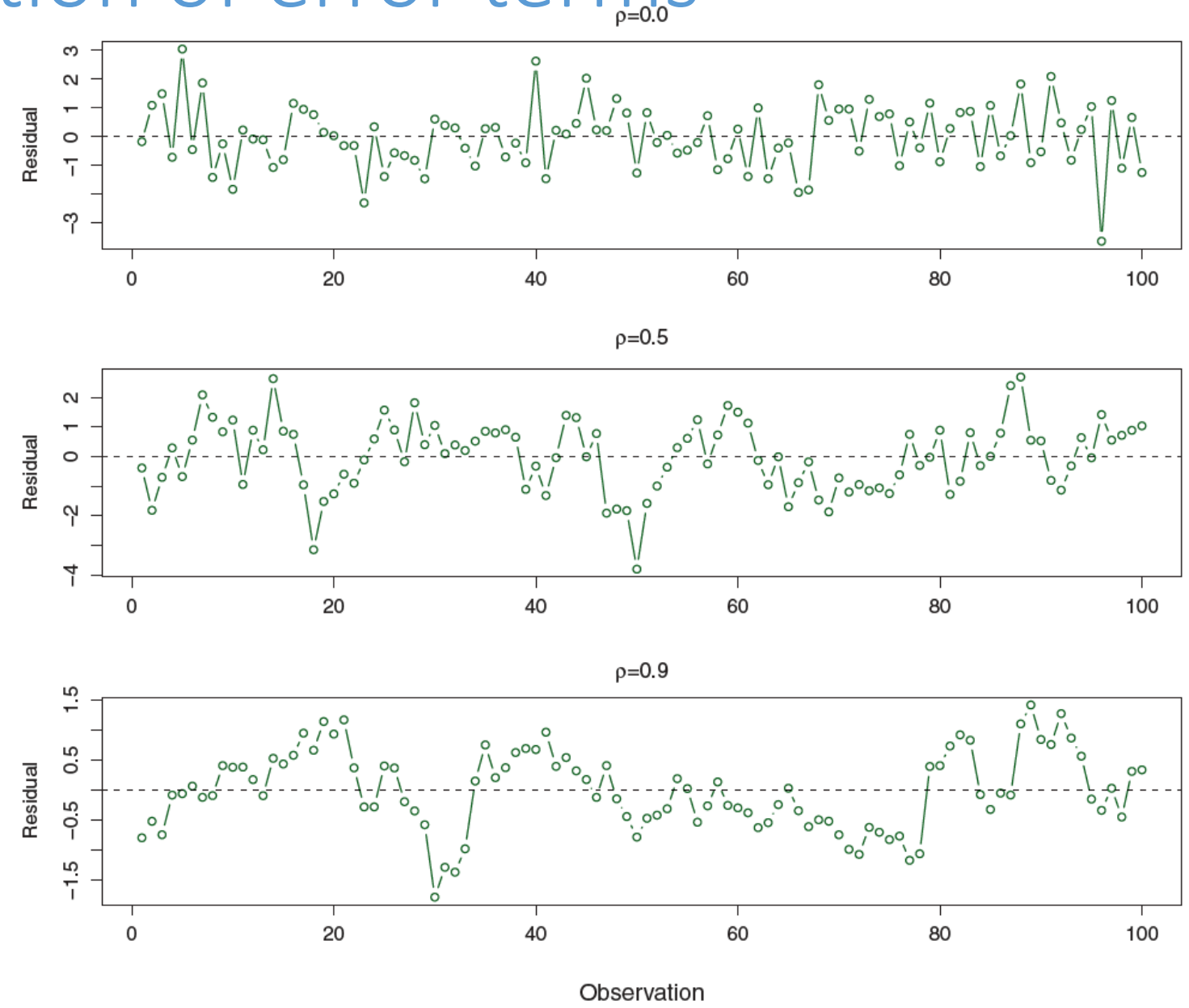
Log transformations



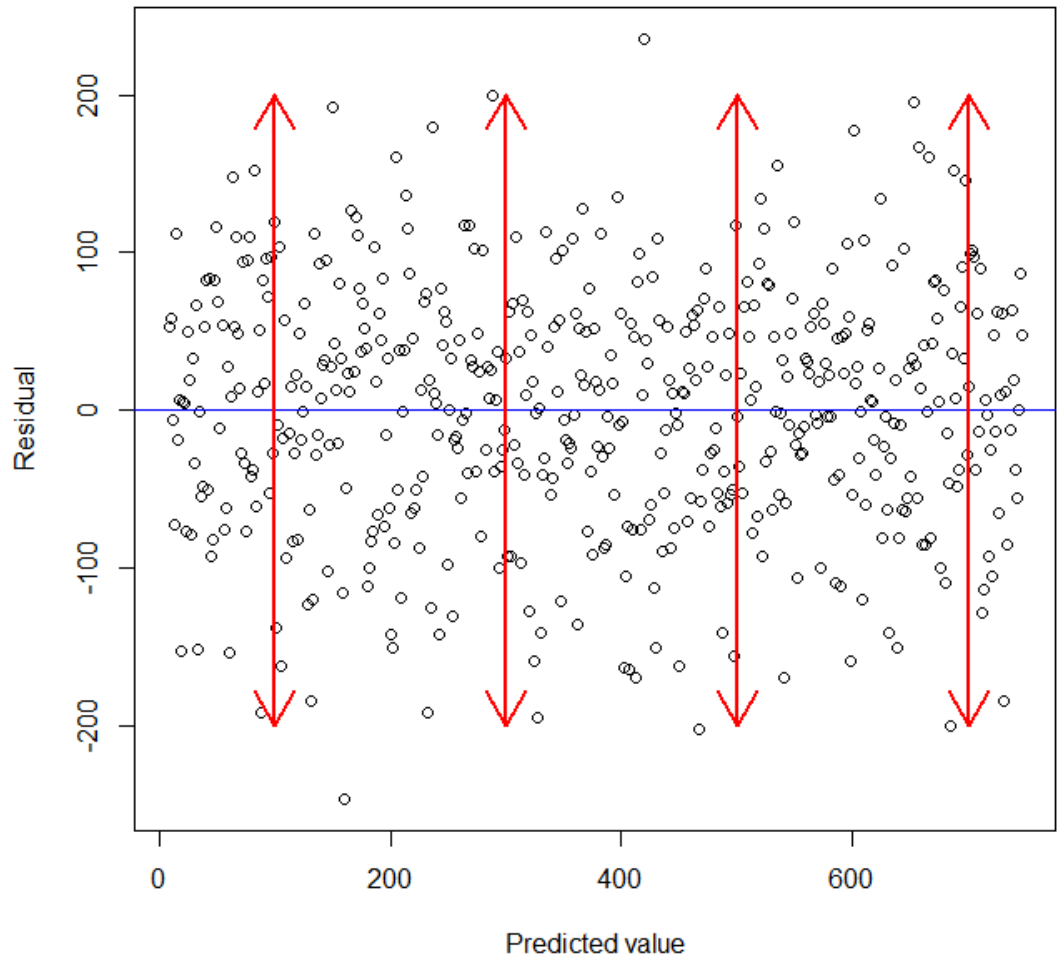
Interactions and log transformations in R

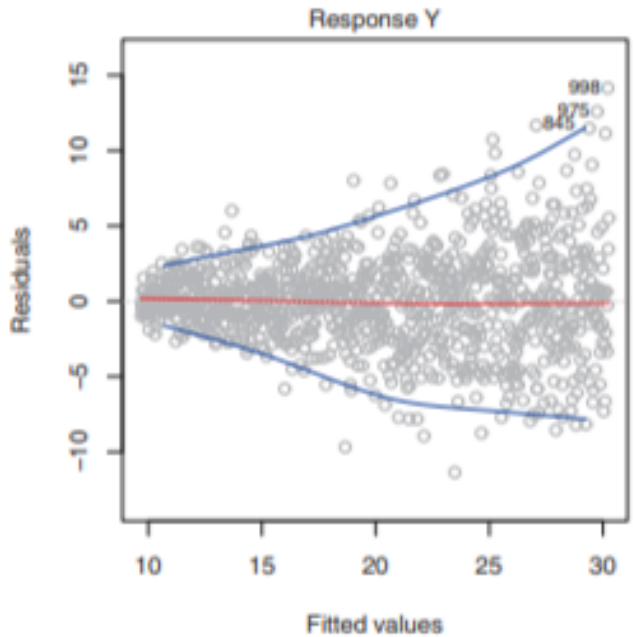
```
> Boston$ageSq <- Boston$age * Boston$age
> Boston$logmedv <- log(Boston$medv)
> summary(data.frame(Boston$ageSq, Boston$age, Boston$logmedv, Boston$medv))
  Boston.ageSq        Boston.age      Boston.logmedv   Boston.medv
 Min.   :    8.41   Min.   :  2.90   Min.   :1.609   Min.   : 5.00
 1st Qu.: 2027.25   1st Qu.: 45.02   1st Qu.:2.835   1st Qu.:17.02
 Median : 6006.29   Median : 77.50   Median :3.054   Median :21.20
 Mean   : 5493.31   Mean   : 68.57   Mean   :3.035   Mean   :22.53
 3rd Qu.: 8850.11   3rd Qu.: 94.08   3rd Qu.:3.219   3rd Qu.:25.00
 Max.   :10000.00   Max.   :100.00   Max.   :3.912   Max.   :50.00
>
```

Correlation of error terms



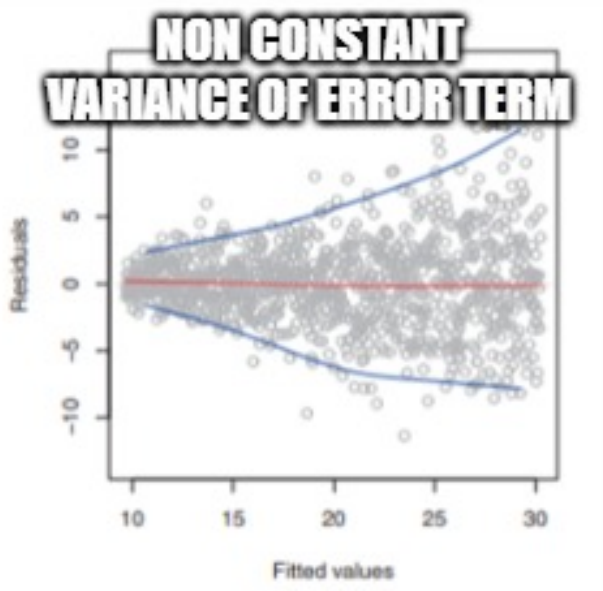
Diagnostics: non-constant variance of error

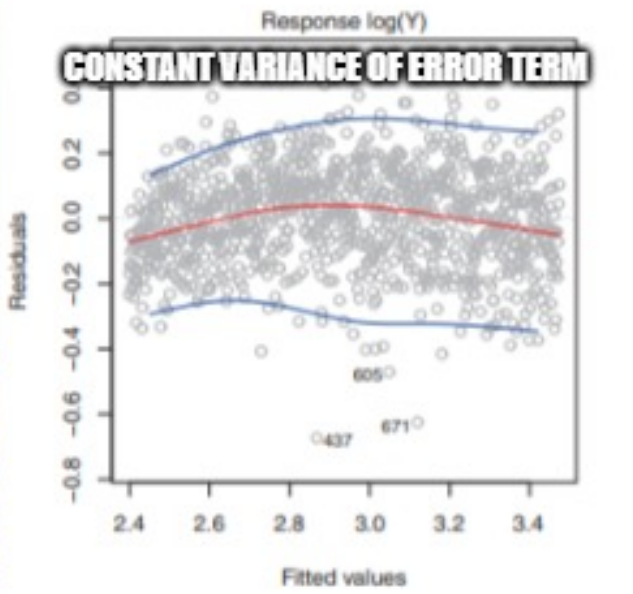




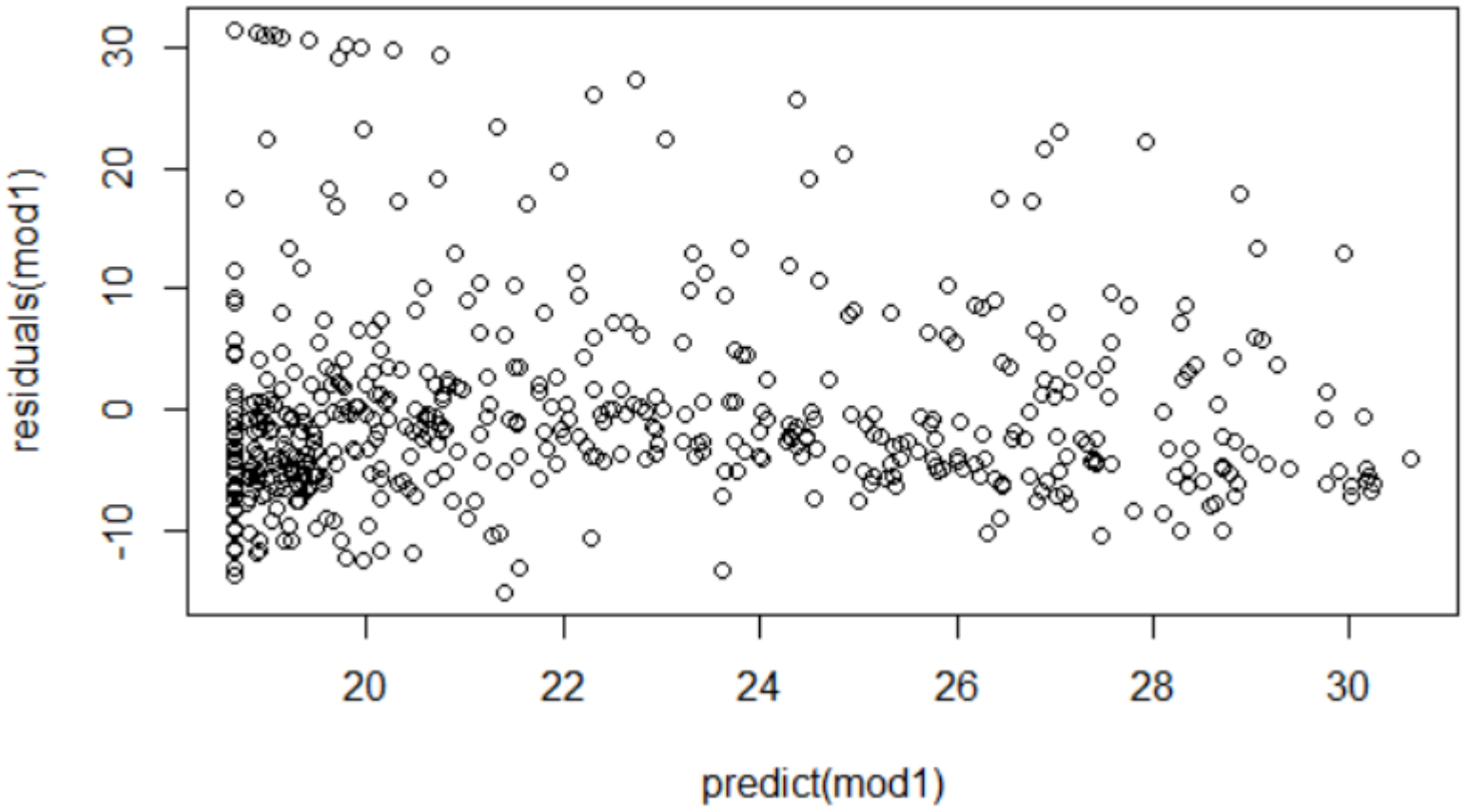

FUTUBANDERA.CL


imgflip.com

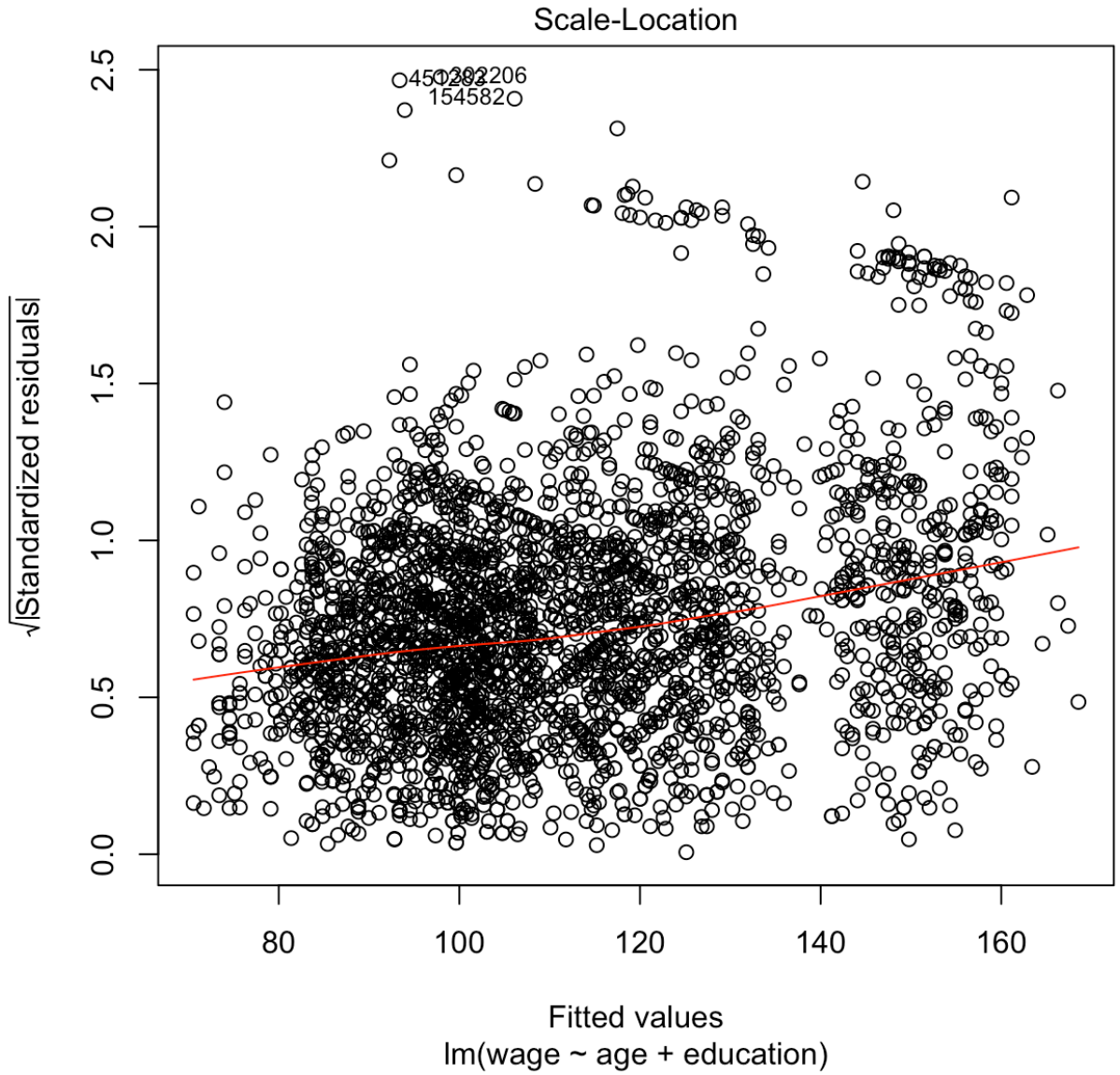

NON CONSTANT
VARIANCE OF ERROR TERM
Residuals
10
5
0
−5
−10
10
15
20
25
30
Fitted values


Response log(Y)
CONSTANT VARIANCE OF ERROR TERM
Residuals
0.2
0.0
−0.2
−0.4
−0.6
−0.8
605
671
437
2.4
2.6
2.8
3.0
3.2
3.4
Fitted values

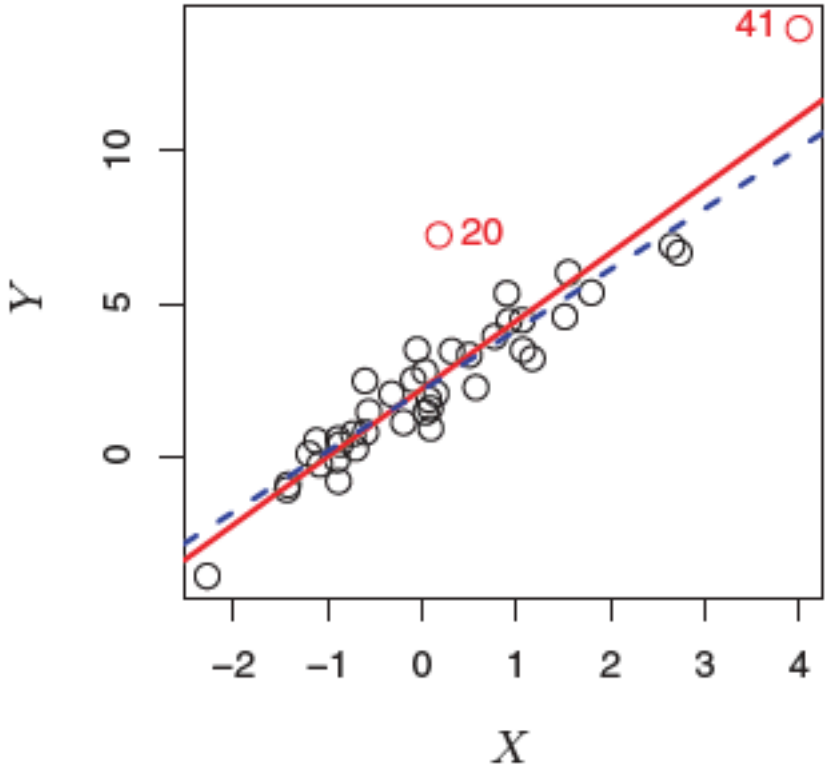
Plot fitted values against residuals

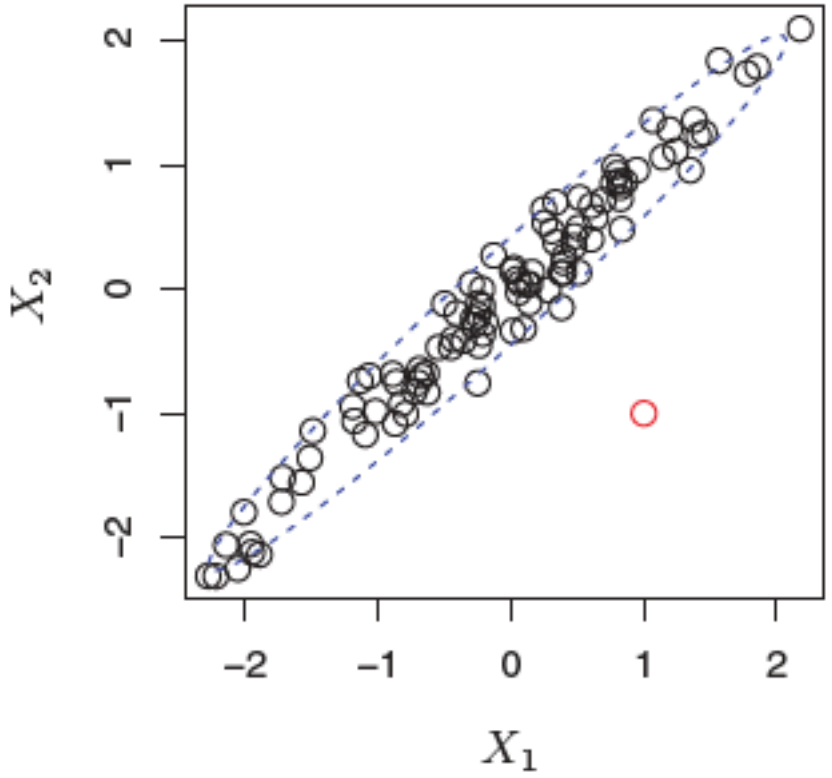


Scale-Location plots (note standardized residuals)



Outliers





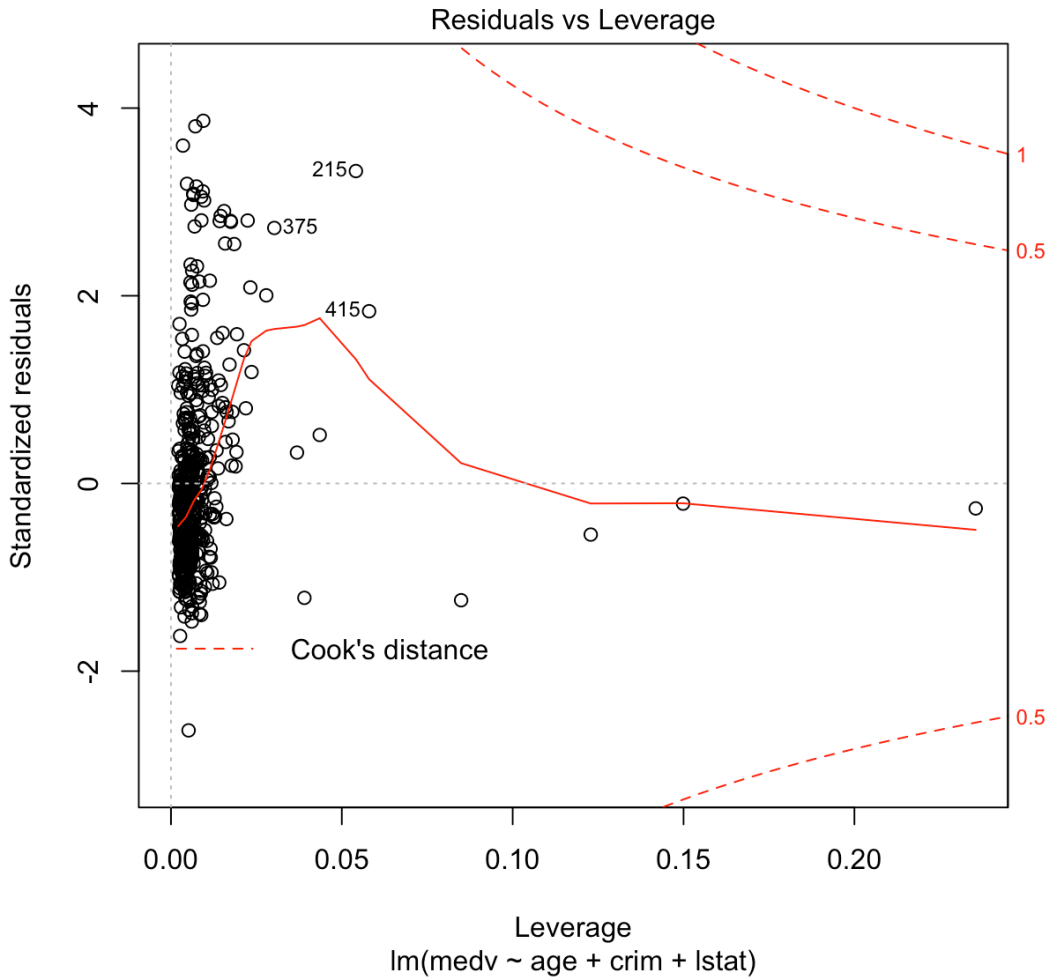
Leverage statistic (hat statistic)

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2}$$

$$h_i \in \left[\frac{1}{n}, 1\right]$$

$Average = (p+1)/n$

```
> mod1 <- lm(medv ~ age + crim + lstat, data = Boston)
> hats <- hatvalues(mod1)
> top5hats <- sort(hats, decreasing = TRUE)[1:5]
> top5hats
       381        419        406        411
0.23550080 0.14986108 0.12283704 0.08491197 0.
```



Collinearity

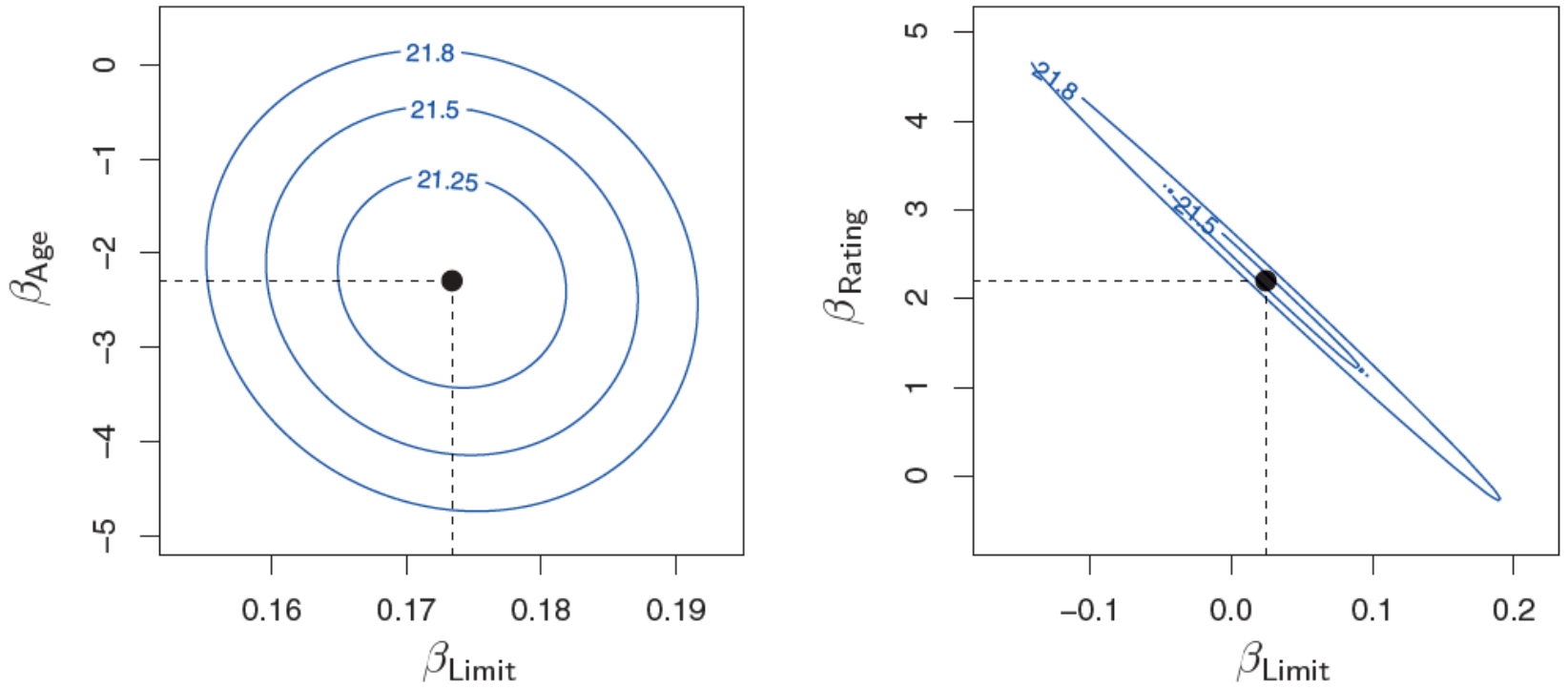


FIGURE 3.15. *Contour plots for the RSS values as a function of the parameters* β *for various regressions involving the* `Credit` *data set. In each plot, the black dots represent the coefficient values corresponding to the minimum RSS.* Left: *A contour plot of RSS for the regression of* `balance` *onto* `age` *and* `limit`*. The minimum value is well defined.* Right: *A contour plot of RSS for the regression of* `balance` *onto* `rating` *and* `limit`*. Because of the collinearity, there are many pairs* $(\beta_{\text{Limit}}, \beta_{\text{Rating}})$ *with a similar value for RSS.*

Impact of collinearity

		Coefficient	Std. error	t-statistic	p-value
Model 1	Intercept	−173.411	43.828	−3.957	< 0.0001
	age	−2.292	0.672	−3.407	0.0007
	limit	0.173	0.005	34.496	< 0.0001
Model 2	Intercept	−377.537	45.254	−8.343	< 0.0001
	rating	2.202	0.952	2.312	0.0213
	limit	0.025	0.064	0.384	0.7012

TABLE 3.11. *The results for two multiple regression models involving the* `Credit` *data set are shown. Model 1 is a regression of* `balance` *on* `age` *and* `limit`*, and Model 2 a regression of* `balance` *on* `rating` *and* `limit`*. The standard error of* $\hat{\beta}_{\text{limit}}$ *increases 12-fold in the second regression, due to collinearity.*

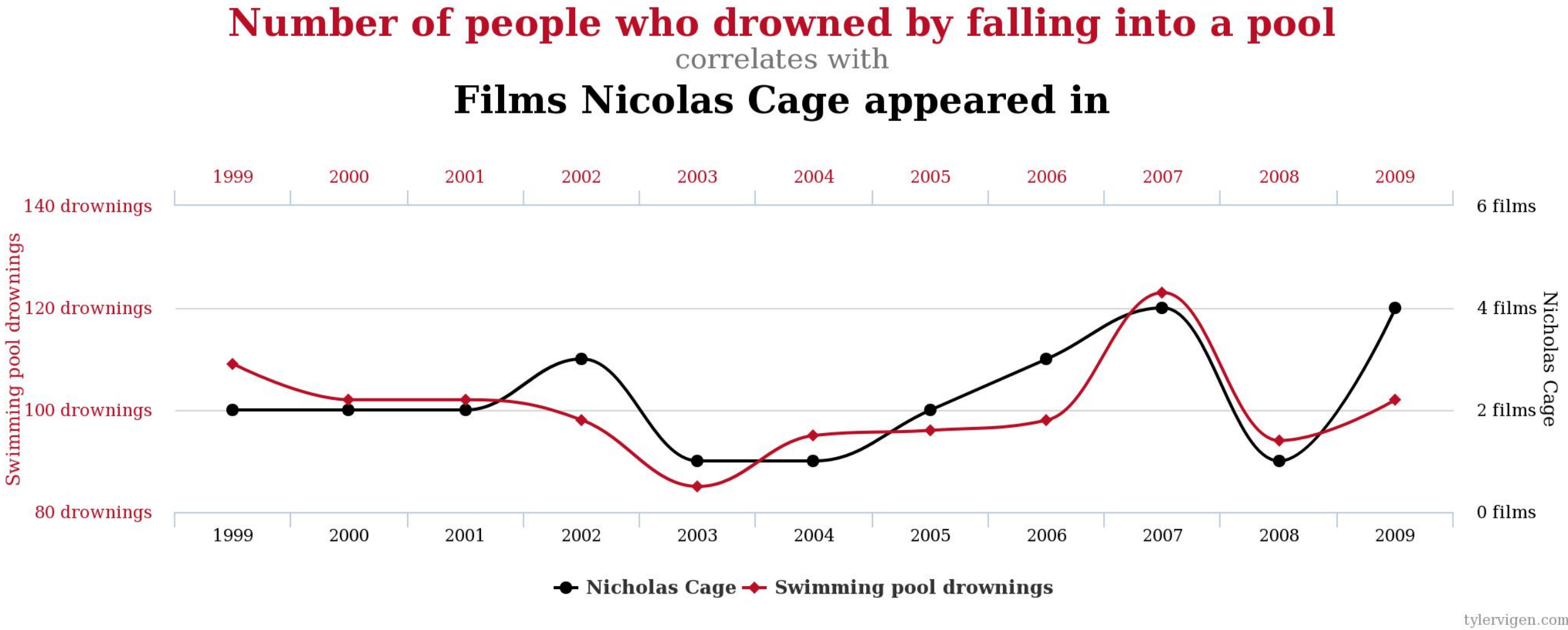
Testing for collinearity: VIF (olsrr package)

$$\mathrm{VIF}(\hat{\beta}_j) = \frac{1}{1-R^2_{X_j|X_{-j}}}$$

```
> ols_vif_tol(mod3)
# A tibble: 3 x 3
  Variables   Tolerance        VIF
      <chr>       <dbl>      <dbl>
1       Age 0.988743466   1.011385
2     Limit 0.006226926 160.592880
3    Rating 0.006224003 160.668301
```

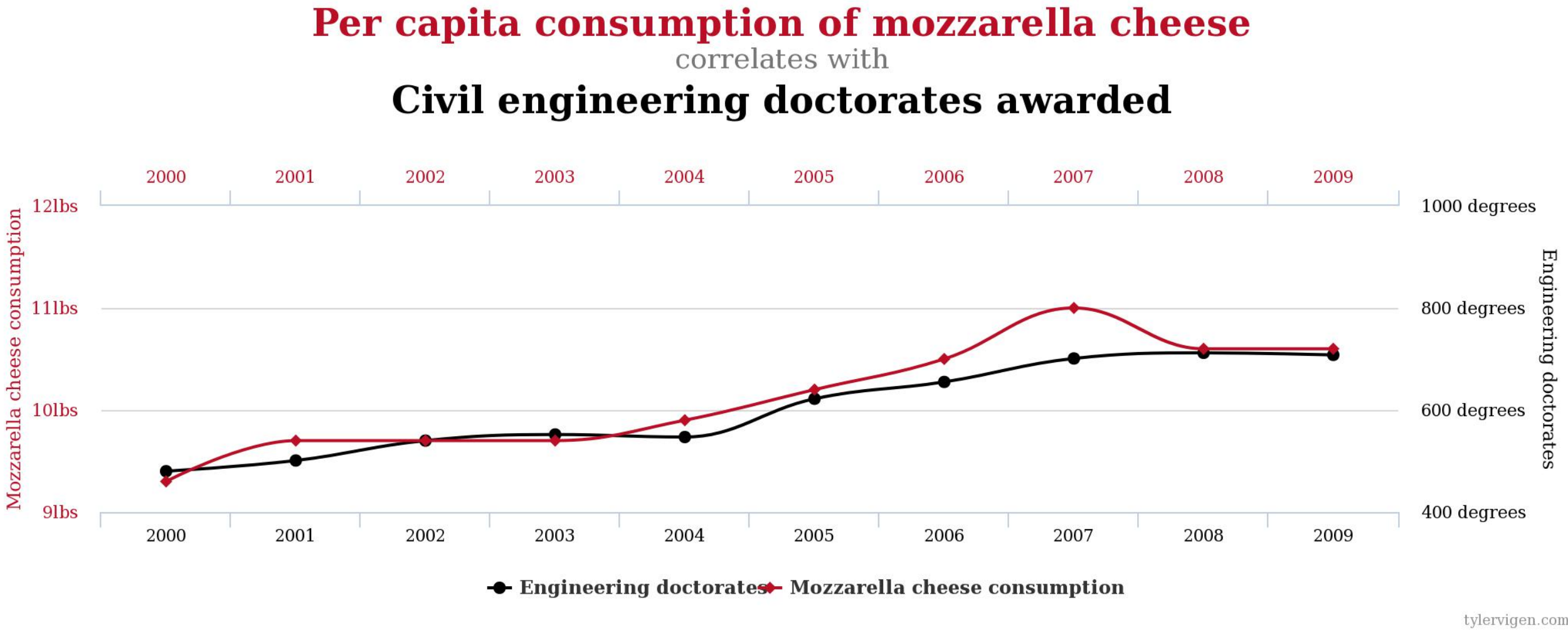
Type equation here.

Spurious Relationship?



Source: http://www.tylervigen.com/spurious-correlations

Spurious Relationship?



Source: http://www.tylervigen.com/spurious-correlations