

# CSCI 3022-002 Intro to Data Science

## Multiple Linear Regression

↪ predict  $Y$  values given  $X$ .

Workflow for SLR so far:

1. Plot the data as a scatter plot
  - 1.1 Does linearity seem appropriate?
  - 1.2 Calculate  $\hat{\beta}_0, \hat{\beta}_1$  and overlay the best-fit line  $y = \beta_0 + \beta_1 X$ .
2. Consider assumptions of SLR:
  - 2.1 Plot a histogram of the residuals: are they **normal**?
  - 2.2 Plot the residuals against  $x$ : are they changing?
3. Perform inference (on  $\beta$ s or on values of  $Y|X$ )

# Announcements and Reminders

- ▶ Short HW for next week.
- ▶ NB day Friday.

## Where we at?

**Definition:** *Simple Linear Regression* (SLR)

The *Simple Linear Regression* model is a model of the form

1.  $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$  (**linearity**) with 3 assumptions on  $\varepsilon$ : *line*  
 - *int*, *slope*, *Adise*
2.  $\text{Cov}[\varepsilon_i, \varepsilon_j] = 0 \quad \forall i, j$  or **Independence** of errors
3.  $\text{Var}(\varepsilon_i) = \sigma^2 \quad \forall i$  or **Homoskedasticity** of errors
4.  $\varepsilon_i \sim N(0, \sigma^2)$  or **distribution** of errors

The estimators in the model are:

1.  $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$
2.  $\hat{\beta}_1 = \frac{\text{Cov}[X, Y]}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$
3.  $\hat{\sigma}^2 = \frac{\text{SSE}}{n-2}$  where  $\text{SSE} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$

## CI's on Betas

$$y = \beta_0 + 0 \cdot x$$

From those estimators, there are **two** types of confidence intervals and hypothesis tests we can make. If our assumptions about the model are correct, our “hat” estimates for  $\hat{\beta}$  will be

normals!  $\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2}{n} + \frac{\sigma^2 \bar{X}^2}{\sum_i (X_i - \bar{X})^2}\right)$   $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_i (X_i - \bar{X})^2}\right)$

... but of course, we don't know  $\sigma^2$ , so we estimate with  $\underline{SSE/(n-2)}$ .

For each of these we can draw hypothesis tests and make confidence intervals. The most common test is:  $H_0 : \beta_1 = 0$ . This tests “is the slope zero or not?” Put another way: it tests *whether the feature  $X$  is a statistically significant predictor of the response  $Y$* . We use it as a measure for whether the existence of our linear model  $Y = \beta_0 + \beta_1 X$  is any better than the model that **ignores**  $X$ , or  $\underline{Y = \bar{Y}}$ .

horizontal line

Tests for the slope of the linear model are  $t$  tests.  $t =$

sloped line

$$t = \frac{\hat{\beta}_1 - \boxed{\beta_{1,H_0}}}{\sqrt{\frac{\sigma^2}{\sum_i (X_i - \bar{X})^2}}}$$

estimates null hyp. value (typically 0)  
 $\rightarrow$  error of  $\hat{\beta}_1$

## CIs on Y

$\hat{\beta}_0$  has standard error (s.e.):  $\hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum_i (X_i - \bar{X})^2}}$  and  $\hat{\beta}_1$  has s.e.:  $\hat{\sigma} \sqrt{\frac{1}{\sum_i (X_i - \bar{X})^2}}$

line:  $\hat{\beta}_0 + \hat{\beta}_1 X_i$   
 error:  $\varepsilon_i$

$(+\varepsilon)$   
 $\uparrow$   
 $N(0, \sigma^2)$

The **second** type of confidence interval are confidence intervals on the  $Y$ -values themselves, of which there are *two*. Each follows from that  $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \varepsilon$ .

1. We can estimate the *mean*  $E[y_{new}]$  value for a given  $x_{new}$  value, often called the *confidence region*. Since  $E[\varepsilon] = 0$ , this depends on the standard errors of the betas.

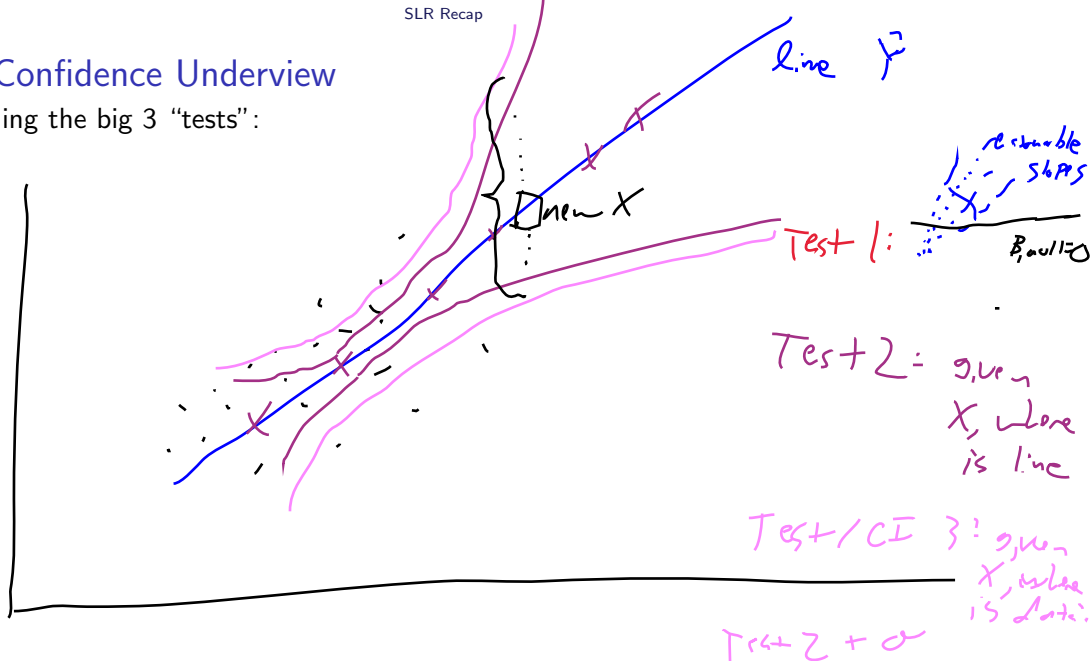
$$\text{s.e}(\text{mean}) = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_{new} - \bar{X})^2}{\sum_i (X_i - \bar{X})^2}}$$

2. We can also estimate the distribution of *individual* observations of  $y_{new}$  values for a *given*  $X_i$ . This is a *prediction band*, and includes an extra addition of the variance  $\hat{\sigma}^2$  since it *includes*  $\varepsilon$ .

$$\text{s.e}(\text{observation}) = \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_{new} - \bar{X})^2}{\sum_i (X_i - \bar{X})^2}}$$

# SLR Confidence Underview

Sketching the big 3 "tests":



# Covariance

When two random variables  $X$  and  $Y$  are not independent, it is frequently of interest to assess how strongly they are related to one another.

**Definition:** *Covariance:*

The covariance between two rv's  $X$  and  $Y$  is defined as:

$$E[ \underbrace{(X - \mu_X)}_{\text{X versus its mean}} \underbrace{(Y - \mu_Y)}_{\text{Y versus its mean}} ]$$

If both variables tend to deviate in the same direction (both go above their means or below their means at the same time), then the covariance will be positive.

If the opposite is true, the covariance will be negative.

If  $X$  and  $Y$  are not strongly related, the covariance will be near 0.

# Correlation

**Definition:** *Correlation*

The *correlation* coefficient of X and Y, denoted by \_\_\_\_\_ or just \_\_, is the *unitless* measure of covariance defined by:

It represents a "scaled" covariance: correlation ranges between -1 and 1.



# Correlation

**Definition:** *Correlation*

The *correlation* coefficient of X and Y, denoted by  $\underline{Cov[X, Y]}$  or just  $\underline{\rho}$ , is the *unitless* measure of covariance defined by:

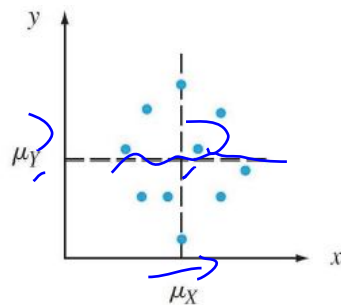
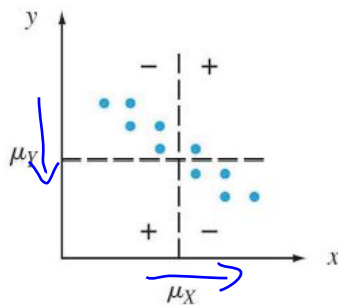
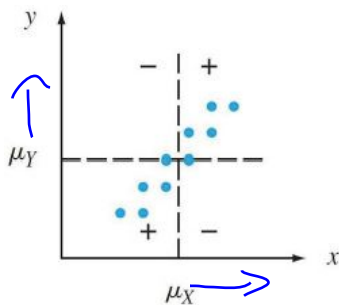
$$\rho = \frac{Cov[X, Y]}{\sigma_X \sigma_Y}$$

It represents a "scaled" covariance: correlation ranges between -1 and 1.

## Covariance Pictured

The covariance depends on both the set of possible pairs and the probabilities of those pairs.

Below are examples of 3 types of “co-varying”:



## Interpreting Correlation

If  $X$  and  $Y$  are independent, then  $\rho = 0$ , but  $\rho = 0$  does not imply independence.

The correlation coefficient is a measure of the *linear relationship* between  $X$  and  $Y$ , and only when the two variables are perfectly related in a *linear* manner will be as positive or negative as it can be.

Two variables could be uncorrelated yet highly dependent because there is a strong nonlinear relationship, so be careful not to conclude too much from low correlation.

NB: the *coefficient of determination*  $R^2$  is **exactly** equal to  $\rho_{X,Y}^2$  for simple linear regression.

# Interpreting Correlation

Today, we will care about both correlations between  $X$  and  $Y$  and also between different features.

## Where we at?

So we have the broad strokes of the assumptions of the linear regression model, but we have no tools in our toolbox to fix violations of the assumptions! That's unfortunate. This week we develop two tools to make our model more responsive to real world concerns:

1. Transformations of variables: does it make more sense to hit  $X$  or  $Y$  with the math stick before we compute a function? Common plans: replace  $Y$  with  $\log Y$  or  $e^Y$ , normalize  $X$  or  $Y$  - e.g. use  $Y_{new,i} = \frac{Y_i - \bar{Y}}{s_Y}$  or just that numerator - and so forth.
2. Add more predictive variables. These can be new columns: e.g. predict weight with *both* height and sex (and age? and wealth?) or more complicated functions of the original predictors, e.g.  $Y = ax^2 + bx + c$  where we estimate  $a, b, c$ .

## MLR

**Example:** Suppose  $y$  is the sale price of a house that we wish to predict. Then sensible predictors include

$x_1$  = the interior size of the house,

$x_2$  = the size of the lot on which the house sits,

$x_3$  = the number of bedrooms,

$x_4$  = the number of bathrooms, and

$x_5$  = the house's age.

expect!



# Multiple Linear Regression

**Definition:** *Multiple Linear Regression*

The multiple regression model is one where we allow each data point to have multiple characteristics (features/predictors) that we use to predict  $y$ . So for each data point we have  $p$  different  $X$ 's to predict  $y$ :

# Multiple Linear Regression

$X_{\#,i}$  :  $i$ th observation for feature " $\#$ "

## Definition: Multiple Linear Regression

The multiple regression model is one where we allow each data point to have multiple characteristics (features/predictors) that we use to predict  $y$ . So for each data point we have  $p$  different  $X$ 's to predict  $y$ :

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \cdots + \beta_p X_{p,i}$$

$\beta_{\#}$  : "slope" of feature/column/ $X$ -value " $\#$ "



# Multiple Linear Regression

**Definition:** *Multiple Linear Regression*

The multiple regression model is one where we allow each data point to have multiple characteristics (features/predictors) that we use to predict  $y$ . So for each data point we have  $p$  different  $X$ 's to predict  $y$ :

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \cdots + \beta_p X_{p,i}$$

This is not a regression line any longer, but a *regression surface*; we relate  $Y$  to more than one predictor variable  $x_1, x_2, \dots, x_p$ . (ex. Blood sugar level vs. weight and age).

## Multiple Linear Regression

$$(x, y) \rightarrow (x, x^2, y)$$

An important note:  $X_j$  could be related to  $X_i$  for any  $i$  or  $J$  of the regression model. For example, to fit a parabola  $y = ax^2 + bx + c$ , we're actually fitting a Multiple Linear Regression model where  $x_1 = x$ ,  $x_2 = x^2$ , and the model is:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

## Multiple Linear Regression

The regression parameter  $\beta_1$  is interpreted as the expected change in  $Y$  associated with a 1-unit increase in  $x_1$  while  $x_2, \dots, x_p$  are held fixed.

Analogous interpretations hold for  $\beta_2, \dots, \beta_p$ .

Thus, these parameters are called partial or adjusted regression parameters/coefficients.

In contrast, the simple regression slope is called the marginal (or unadjusted) coefficient.

# Multiple Linear Regression

The multiple regression model can be written in matrix form:

## Multiple Linear Regression

The multiple regression model can be written in matrix form:

We use vectors  $\underline{Y} := [Y_1, Y_2, \dots, Y_n]^T$

$\underline{\beta} := [\beta_0, \beta_1, \dots, \beta_p]^T$

$$X := \begin{bmatrix} 1 & X_{1,1} & \dots & X_{1,p} \\ 1 & X_{2,1} & \dots & X_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & \dots & X_{n,p} \end{bmatrix}$$

$\uparrow$   
# rows

$\uparrow$   
lot Sq. footage

columns: features  
rows: observation

## Multiple Linear Regression

The multiple regression model can be written in matrix form:

We use vectors  $\underline{Y} := [Y_1, Y_2, \dots, Y_n]^T$

$\underline{\beta} := [\beta_0, \beta_1, \dots, \beta_p]^T$

$$\mathbf{X} := \begin{bmatrix} 1 & X_{1,1} & \dots & X_{1,p} \\ 1 & X_{2,1} & \dots & X_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & \dots & X_{n,p} \end{bmatrix}$$

So our model is  $\underline{Y} = \mathbf{X}\underline{\beta} + \underline{\varepsilon}$ .  $\mathbf{X}$  is called the **design** matrix

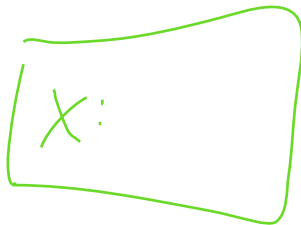
## Multiple Linear Regression

The multiple regression model can be written in matrix form:

We use vectors  $\underline{Y} := [Y_1, Y_2, \dots, Y_n]^T$

$\underline{\beta} := [\beta_0, \beta_1, \dots, \beta_p]^T$

$$\underline{X} := \begin{bmatrix} 1 & X_{1,1} & \dots & X_{1,p} \\ 1 & X_{2,1} & \dots & X_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & \dots & X_{n,p} \end{bmatrix} \quad \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \quad \underline{X} \cdot \underline{\beta}$$



So our model is  $\underline{Y} = \underline{X}\underline{\beta} + \underline{\varepsilon}$ .  $\underline{X}$  is called the **design** matrix

Consider e.g.  $n = 2; p = 2$ . The matrix system is

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{bmatrix} 1 & X_{1,1} & X_{1,2} \\ 1 & X_{2,1} & X_{2,2} \end{bmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} \beta_0 + \beta_1 X_{1,1} + \beta_2 X_{1,2} \\ \beta_0 + \beta_1 X_{2,1} + \beta_2 X_{2,2} \end{pmatrix}$$

In this case,  $\underline{X}\underline{\beta}$  is the mean model and  $\underline{\varepsilon}$  is the errors.

## Multiple Linear Regression

This model can be thought of as one with 4 (familiar) assumptions:

With 3 assumptions on  $\underline{\epsilon}$ :



## Multiple Linear Regression

This model can be thought of as one with 4 (familiar) assumptions:

1. A **linear** fit is appropriate:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \cdots + \beta_p X_{p,i}$$

With 3 assumptions on  $\underline{\varepsilon}$ :

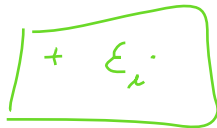
## Multiple Linear Regression

This model can be thought of as one with 4 (familiar) assumptions:

1. A **linear** fit is appropriate:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \cdots + \beta_p X_{p,i}$$

With 3 assumptions on  $\underline{\varepsilon}$ :



- 2.

$$\text{Cov}[\varepsilon_i, \varepsilon_j] = 0 \quad \forall i, j$$

**Independence** of errors

## Multiple Linear Regression

This model can be thought of as one with 4 (familiar) assumptions:

1. A **linear** fit is appropriate:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \cdots + \beta_p X_{p,i}$$

With 3 assumptions on  $\underline{\varepsilon}$ :

- 2.

$$\text{Cov}[\varepsilon_i, \varepsilon_j] = 0 \quad \forall i, j$$

**Independence** of errors

- 3.

$$\text{Var}(\varepsilon_i) = \sigma^2 \quad \forall i$$

**Homoskedasticity** of errors

# Multiple Linear Regression

This model can be thought of as one with 4 (familiar) assumptions:

1. A **linear** fit is appropriate:

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \cdots + \beta_p X_{p,i}$$

With 3 assumptions on  $\underline{\varepsilon}$ :

- 2.

$$\text{Cov}[\varepsilon_i, \varepsilon_j] = 0 \quad \forall i, j$$

**Independence** of errors

- 3.

$$\text{Var}(\varepsilon_i) = \sigma^2 \quad \forall i$$

**Homoskedasticity** of errors

- 4.

$$\varepsilon_i \sim N(0, 1)$$

**Distribution** of errors

# Multiple Linear Regression Estimators

Our estimators for the regression coefficients  $\beta$  rely on these assumptions, and look similar to those in SLR.

Recall:

$$\hat{\beta}_1 =$$

$$\frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

$X$  to itself?

$X$  &  $Y$  relations

## Multiple Linear Regression Estimators

Our estimators for the regression coefficients  $\beta$  rely on these assumptions, and look similar to those in SLR.

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix} = \boxed{(X^T X)^{-1}} X^T \underline{Y}$$

matrix inverse  $\approx$  denominator

## Multiple Linear Regression Estimators

Our estimators for the regression coefficients  $\beta$  rely on these assumptions, and look similar to those in SLR.

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix} = (X^T X)^{-1} X^T \underline{Y}$$

The  $(X^T X)^{-1}$  bit corresponds to the  $1 / \sum (X_i - \bar{X})^2$  part from before, where the  $X^T \underline{Y}$  part corresponds roughly to a covariance between  $X$  and  $Y$ .

## MLR Sums of Squares

Just as with simple regression, the residual sum of squares is:

It is again interpreted as a measure of how much variation in the observed  $y$  values is not explained by (not attributed to) the model relationship.

The number of df associated with SSE is  $n - (p + 1)$  because  $p + 1$  df are "lost" in estimating the  $p + 1$  coefficients. This leads to an estimate for the standard errors of



## MLR Sums of Squares

Just as with simple regression, the residual sum of squares is:

$$SSE = \sum_{i=1}^n \left( Y_i - \hat{Y}_i \right)^2$$

*Handwritten notes:*  
 -  $Y_i$  is labeled "data" with a pink arrow.  
 -  $\hat{Y}_i$  is labeled "estimated mean/surface" with a pink arrow.

It is again interpreted as a measure of how much variation in the observed y values is not explained by (not attributed to) the model relationship.

The number of df associated with SSE is  $n - (p + 1)$  because  $p + 1$  df are "lost" in estimating the  $p + 1$  coefficients. This leads to an estimate for the standard errors of

*Handwritten notes:*  
 - "p" indicates slope (with a pink arrow pointing down from the text to the p in the formula).  
 - "slope" (with a pink arrow pointing down from the text to the p in the formula).

## MLR Sums of Squares

Just as with simple regression, the residual sum of squares is:

$$SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

It is again interpreted as a measure of how much variation in the observed  $y$  values is not explained by (not attributed to) the model relationship.

The number of df associated with SSE is  $n - (p + 1)$  because  $p + 1$  df are "lost" in estimating the  $p + 1$  coefficients. This leads to an estimate for the standard errors of

$$\hat{\sigma}^2 = \frac{SSE}{n - (p + 1)}$$

(versus  $\frac{SSE}{n-2}$  for  $\beta_0, \beta_1$ )

## MLR Coefficient of Determination

Just as before, the total sum of squares is:

And the regression sum of squares is:

Then the coefficient of multiple determination  $R^2$  is:

It is interpreted in the same way as in SLR: it is the proportion of variability in  $Y$  captured by our linear model.

## MLR Coefficient of Determination

Just as before, the total sum of squares is:

$$SST = \sum_{i=1}^n (Y_i - \bar{Y}_i)^2$$

And the regression sum of squares is:

$$SSR = \sum_{i=1}^n (\hat{Y}_i - \bar{Y}_i)^2$$

Then the coefficient of multiple determination  $R^2$  is:

It is interpreted in the same way as in SLR: it is the proportion of variability in  $Y$  captured by our linear model.

## MLR Coefficient of Determination

Just as before, the total sum of squares is:

$$SST = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

And the regression sum of squares is:

$$SSR = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$$

Then the coefficient of multiple determination  $R^2$  is:

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST}$$

It is interpreted in the same way as in SLR: it is the proportion of variability in  $Y$  captured by our linear model.

Compare

$$y = ax + b$$

$$y = ax^2 + \underbrace{bx + c}_{\text{contains small model}}$$

## MLR Coefficient of Determination

Unfortunately, there is a problem with  $R^2$ : Its value can be inflated by adding lots of predictors into the model even if most of these predictors are frivolous.

From our example of predicting house pricing  $y$  before, suppose we also add these predictors to the model:

$x_6$  = the diameter of the doorknob on the coat closet,

$x_7$  = the thickness of the cutting board in the kitchen,

$x_8$  = the thickness of the patio slab.

} overfit

## Adjusted MLR Coefficient of Determination

The objective in multiple regression is not simply to explain the most of the observed  $y$  variation. Some variation is random (i.e., not associated with a predictor). So, too many predictors would be bad: we might start assigning that randomness to features that don't make any sense. We should build a model with relatively few predictors (that are easily interpreted: Occam's razor).

It is thus desirable to adjust  $R^2$  to take account of the size of the model:

## Adjusted MLR Coefficient of Determination

The objective in multiple regression is not simply to explain the most of the observed  $y$  variation. Some variation is random (i.e., not associated with a predictor). So, too many predictors would be bad: we might start assigning that randomness to features that don't make any sense. We should build a model with relatively few predictors (that are easily interpreted: Occam's razor).

It is thus desirable to adjust  $R^2$  to take account of the size of the model:

$$R^2: \frac{SSR}{SST}$$

$$R_a^2 = \frac{SSR / (\underbrace{p}_{\text{\# of predictors}} + 1)}{SST / (n - 1)}$$



## Adjusted MLR Coefficient of Determination

The objective in multiple regression is not simply to explain the most of the observed  $y$  variation. Some variation is random (i.e., not associated with a predictor). So, too many predictors would be bad: we might start assigning that randomness to features that don't make any sense. We should build a model with relatively few predictors (that are easily interpreted: Occam's razor).

It is thus desirable to adjust  $R^2$  to take account of the size of the model:

$$R_a^2 = \frac{SSR/(p+1)}{SST/(n-1)}$$

Idea: as  $p$  grows, we have to improve  $SSR$  proportionately just as fast, or it wasn't worth the new parameter.

## MLR Errors

SSR is still the basis for estimating the remaining model parameter,  $\sigma^2$ :

## MLR Errors

SSR is still the basis for estimating the remaining model parameter,  $\sigma^2$ :

$$\hat{\sigma}^2 = \frac{SSE}{n - (p + 1)}$$

## MLR Errors

We can use Python to compute the standard errors of the regression coefficients. By hand, one would use the distribution of the least squares estimator to calculate the standard errors:

With the standard error, we can compute confidence intervals:

We can also conduct hypothesis tests:

## MLR Errors

We can use Python to compute the standard errors of the regression coefficients. By hand, one would use the distribution of the least squares estimator to calculate the standard errors:

$$\underline{\hat{\beta}} \sim N(\underline{\beta}, Var[\underline{\hat{\beta}}])$$

or  $\hat{\beta}_j \sim N(\beta_j, (s.e.(\hat{\beta}_j))^2)$

With the standard error, we can compute confidence intervals:

$$CI \beta_j : \underline{\hat{\beta}}_j \pm t_{\alpha/2, n-(p+1)} \cdot s.e.(\hat{\beta}_j)$$

We can also conduct hypothesis tests:

$$H_0 : \beta_j = 0 ; \quad H_a : \beta_j \neq 0$$

for  $j = 1, 2, \dots, p$ , usually.

test in  
the presence  
of each  
other/  
variable

## Collinearity

The  $(X^T X)^{-1}$  term in our regression coefficient errors is very similar to the "spread of  $X$ " term in the SLR coefficients. This time, however, it's a little nastier: it's the spread of  $X$  across *all*  $p$  dimensions of the predictors. Example:

Suppose we have roughly linear data, and we decide to fit the data with the model

$$y = \beta_0 + \beta_1 x + \beta_2 x^{1.000001} + \varepsilon$$

## Collinearity

The predictors  $x_1 = x$  and  $x_2 = x^{1.000001}$  are very highly related. In particular, if the *true* value of the regression line is  $y = x$ , the following functions are visually nearly identical:

## Collinearity

The predictors  $x_1 = x$  and  $x_2 = x^{1.000001}$  are very highly related. In particular, if the *true* value of the regression line is  $y = x$ , the following functions are visually nearly identical:

1.  $y = x$
2.  $y = x^{1.000001}$



## Collinearity

The predictors  $x_1 = x$  and  $x_2 = x^{1.000001}$  are very highly related. In particular, if the *true* value of the regression line is  $y = x$ , the following functions are visually nearly identical:

1.  $y = x$
2.  $y = x^{1.000001}$
3.  $y = .5x + .5x^{1.000001}$
4.  $y = 2x^{1.000001} - x$

## Collinearity

The predictors  $x_1 = x$  and  $x_2 = x^{1.000001}$  are very highly related. In particular, if the *true* value of the regression line is  $y = x$ , the following functions are visually nearly identical:

1.  $y = x$
2.  $y = x^{1.000001}$
3.  $y = .5x + .5x^{1.000001}$
4.  $y = 2x^{1.000001} - x$
5.  $y = 10^6 x - 999999x^{1.000001}$
6.  $y = ax + bx^{1.000001}; \quad \forall a + b = 1.$

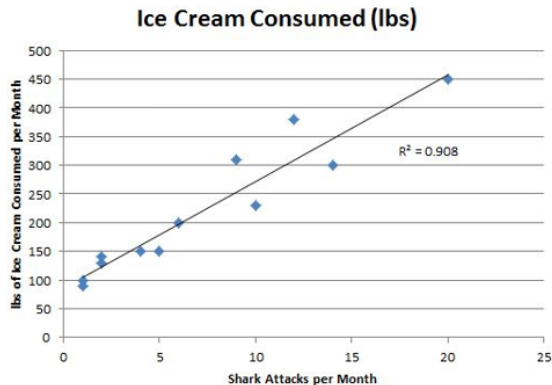
## Collinearity

This is scary! The distribution of  $\underline{\beta}$  has its own covariance, because the best choices for  $\beta_1$  and  $\beta_2$  may depend on each other. In the prior example, they would have a negative correlation of  $-1$ !

In general, the interactions between coefficients is a function of the *linear independence* of the columns of the  $X$  matrix. In other words, we get a lot of negative effects if one predictor is describing one of the same things that we already have!

## Correlations:

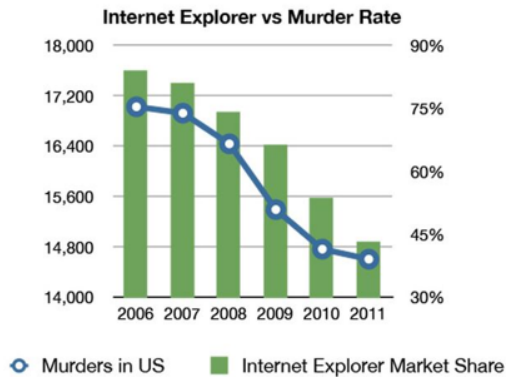
A SLR analysis of shark attacks vs ice cream sales at a Southern California beach indicates that there is a strong relationship between the two.



# MLR:

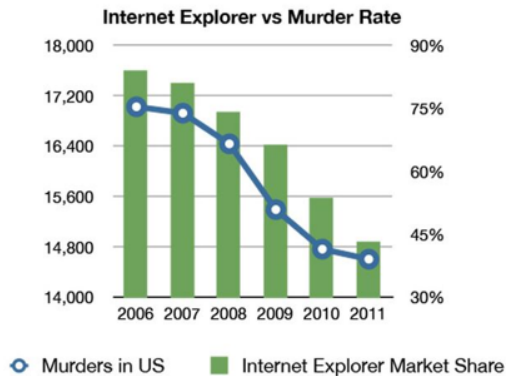
Suppose we included both **temperature** and shark attacks as features in our model of ice cream sales. What would happen? Which one should we probably exclude, and why?

## Correlations:



Why is this correlated?

# Correlations:



Why is this correlated?

