CSCI 3022 Intro to Data Science
Multiple Linear Regression

Workflow for SLR so far: → Picture #1

1. Plot the data as a scatter plot
    1. 1.1 Does **linearity** seem appropriate?
    2. 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:
    1. 2.1 Plot a histogram of the residuals: are they **normal**?
    2. 2.2 Plot the residuals against x: are they changing?
3. Perform inference (on βs or on values of $Y|X$)

Picture #2

plt.scatter(x, errors_of_x)

→ plt.hist(errors)

1: CI on slope/intercept

2: values of R^2, $\hat{\sigma}^2$

3: questions about Y given X

Announcements and Reminders

office hr, today: 3p-5p.

- Short HW for next week posted. : reminder: 2 hw are dropped
- First half of final practicum posted!

Common HW questions

1 "maximum" means the maximum of the two numbers. like the np.max(A,B) function.

2 "$X \sim P(\lambda)$" in problem 2 is X is distributed *Poisson*.

Pois

3 For the last problem of problem 3, consider the same p_{alt} values as you did in the prior part. I.e.: "if a coin has 51% *true* rate of heads, what sample size is necessary to achieve both 0.05 or less type one error **and** 0.05 or less type two error?"

Where we at?

Definition: *Simple Linear Regression* (SLR)

The *Simple Linear Regression* model is a model of the form → 599.5 !

1. $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ (**linearity**) with 3 assumptions on ε:
2. $\text{Cov}[\varepsilon_i, \varepsilon_j] = 0 \quad \forall i, j$ or **Independence** of errors Sign
3. $Var(\varepsilon_i) = \sigma^2 \quad \forall i$ or **Homoskedasticity** of errors Size

} of errors

4. $\varepsilon_i \sim N(0, \overset{\sigma^2}{1})$ or **distribution** of errors Shape of errors

The estimators in the model are: Summary table coefficients

1. $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$
2. $\hat{\beta}_1 = \frac{Cov[X,Y]}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}$ → total variance summed all points

→ variance per point

3. $\hat{\sigma}^2 = \frac{SSE}{n-2}$ where $\boxed{\text{SSE}} = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right)^2$

CIs on Betas

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

Spread of X matters

more spread = less error.

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

depends on $\sigma^2 \approx \hat{\sigma}^2 \Rightarrow$ use t's.

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 $Y = \bar{Y}$.

3: those #s live in summary table

Tests for the slope of the linear model are t tests. $t = \frac{\hat{\beta}_1 - \beta_{1,H_0}}{\sqrt{\frac{\sigma^2}{\sum_i (X_i - \bar{X})^2}}}$

Coef $\hat{\beta}_0$ std err $s.e.(\hat{\beta}_0)$

sm.OLS(y, x).fit()

1) is $\beta_1 = 0$? $\Rightarrow$ CI

CIs on Y

-> summary_table

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

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

2)

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.

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

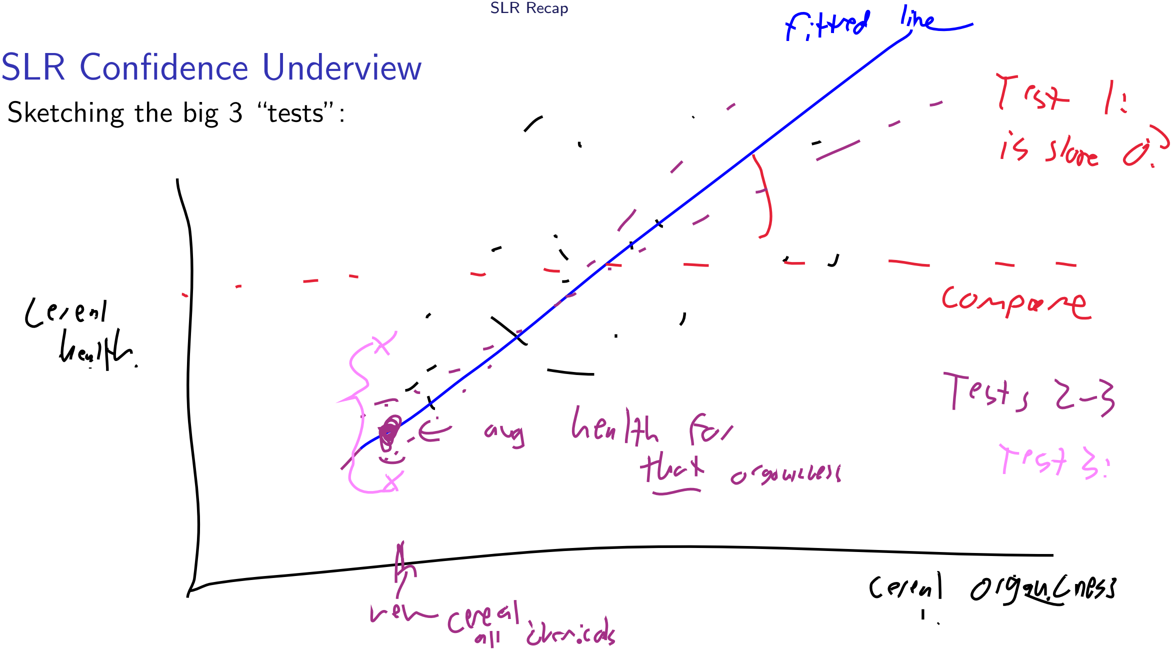
3)

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* ε.

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

SLR Confidence Underview

Sketching the big 3 “tests”:



Covariance

(x,y) values

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:

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

- + if X is "above average"
- − if that X is "below average"

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.

(Compare to **variance**: $E[(X-\mu_X)(X-\mu_X)]$.)

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

combines

→ "spread X times spread of Y" TOGETHER

Definition: *Correlation*
The *correlation* coefficient of X and Y, denoted by $Cov[X, Y]$ or just ρ, is the *unitless* measure of covariance defined by:

$$\rho = \frac{Cov[X, Y]}{\boxed{\sigma_X \sigma_Y}} \rightarrow \text{becomes unitless}$$

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

Covariance Pictured

$Cov(X,Y)$: units X · units Y

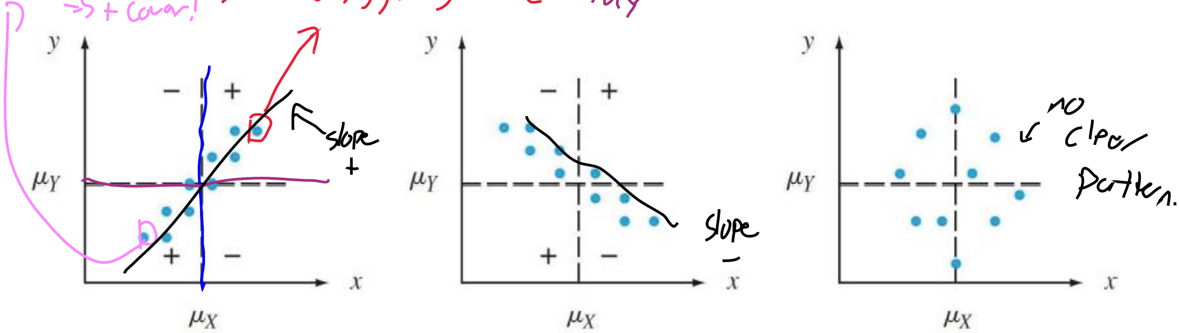
σ_X: units X σ_Y: units Y

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":

x,y: x value $< \mu_X$
y value $< \mu_y$
=> + covar!

(x,y): x value $> \mu_X$
y value $> \mu_Y$ => + covar!



slope +

slope −

no clear pattern.

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.

$\rho \in [-1, 1]$.

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

$[0, 1]$.

(correlation)^2

Interpreting Correlation

SLR

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

multiple X's !

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.

   ↳ $x_{new} = \frac{x - \bar{x}}{s_x}$

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.

   ↑ intercept

   slope

   concavity

3. Remove 'bad' points ?!

MLR

Multiple (x-values).

Linear

Regression

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, +, probably

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

x_3 = the number of bedrooms, + ... maybe? but what if house size is the same?

x_4 = the number of bathrooms, and +

x_5 = the house's age.

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

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.:

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

Slope for "size"

Slope for "size of lot"

slope for "age".

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 $x_1, x_2, \ldots, x_p$. (ex. Blood sugar level vs. weight and age).

Multiple Linear Regression

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:

$$y = \beta_0 + \beta_1 \overset{\text{feature \#1}}{x} + \beta_2 \underset{\text{feature \#2}}{x^2} + \varepsilon$$

3 estimated values

intercept, slope, concavity.

Multiple Linear Regression

slope #1

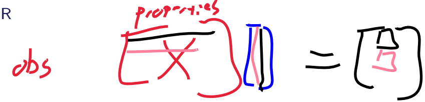
The regression parameter β_1 is interpreted as the expected change in Y associated with a 1-unit increase in x_1 while $x_2, \ldots, x_p$ are held fixed.

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

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

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

$$X := \left.\underbrace{\begin{bmatrix} 1 & X_{1,1} & \dots & X_{1,p} \\ 1 & X_{2,1} & \dots & X_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & \dots & X_{n,p} \end{bmatrix}}_{\text{Feature}}\right\}\text{obs}$$

Multiple Linear Regression

The multiple regression model can be written in matrix form:

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

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

$$X := \begin{bmatrix} 1 & X_{1,1} & \ldots & X_{1,p} \\ 1 & X_{2,1} & \ldots & X_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & \ldots & X_{n,p} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}$$

$$Y_3 = \beta_0 \cdot 1 + \beta_1 \underset{\text{feature 1, obs 3}}{X_{3,1}} + \beta_2 \cdot \underset{\text{feature 2, obs 3}}{X_{3,2}} + \cdots + \beta_p \underset{\text{row 3 feature p}}{X_{3,p}}$$

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

$$\text{each } Y = \beta_0 + \beta_1 \cdot (\text{that } \underset{\text{size of house}}{X_1}) + \beta_2 (\text{that } \underset{\text{size of lot}}{X_2}) + \ldots$$

Multiple Linear Regression

The multiple regression model can be written in matrix form:
We use vectors $\underline{Y} := [Y_1, Y_2, \ldots Y_n]^T$
$\underline{\beta} := [\beta_0, \beta_1, \ldots \beta_p]^T$

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

So our model is $\underline{Y} = \boldsymbol{X}\underline{\beta} + \underline{\varepsilon}$. $\boldsymbol{X}$ 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, $X\underline{\beta}$ 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 $\underline{\varepsilon}$:

Multiple Linear Regression

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

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

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

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.

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

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

   Independence of errors

3. $$Var(\varepsilon_i) = \sigma^2 \qquad \forall i$$

   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 $\underline{\varepsilon}$:

vector

2. $$\text{Cov}[\varepsilon_i, \varepsilon_j] = 0 \qquad \forall i, j$$ indep.

   Independence of errors

3. $$Var(\varepsilon_i) = \sigma^2 \qquad \forall i$$ identical

   Homoskedasticity of errors

4. Distribution

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

   Distribution of errors

Multiple Linear Regression Estimators

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

Multiple Linear Regression Estimators

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

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

Multiple Linear Regression Estimators

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

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

The $\left(X^T X\right)^{-1}$ bit corresponds to the $1/\sum\left(X_i - \bar{X}\right)^2$ 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:

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

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:

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

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

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

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:

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

And the regression sum of squares is:

$$SSR = \sum_{i=1}^{n} \left(\hat{Y_i} - \bar{Y_i}\right)^2$$

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:

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

And the regression sum of squares is:

$$SSR = \sum_{i=1}^{n} \left(\hat{Y_i} - \bar{Y_i}\right)^2$$

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.

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.

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 R2 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 R2 to take account of the size of the model:

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

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 R2 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, σ^2:

MLR Errors

SSR is still the basis for estimating the remaining model parameter, σ^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:

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

We can also conduct hypothesis tests:

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

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

Collinearity

The $\left(X^T X\right)^{-1}$ 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. $y = ax + bx^{1.000001}; \qquad \forall a + b = 1.$

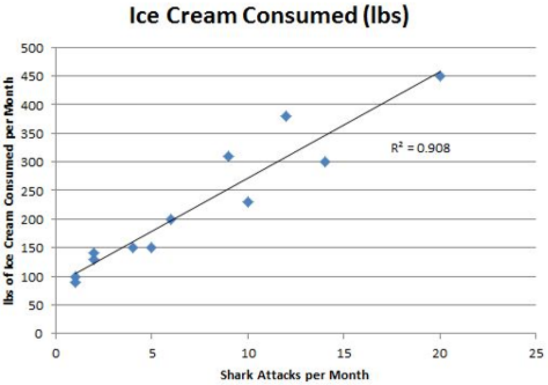
Collinearity

This is scary! The distribution of $\underline{\beta}$ has its own covariance, because the best choices for β_1 and β_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:

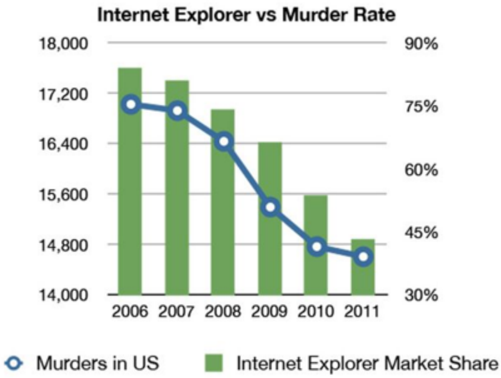
A SLR analysis of shark attacks vs ice cream sales at a Southern California beach indicates that there is a strong relationship between the two.



MLR:

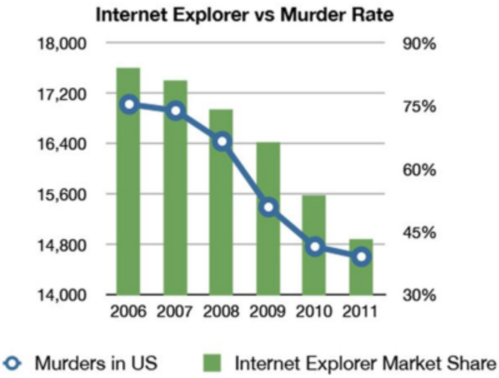
Suppose we included both **temperature** and shark attacks as features in our model of ice cream sales. What would happen? Which one should we probably exclude, and why?

Correlations:



Why is this correlated?

Correlations:



Why is this correlated?

What are we?
BROWSERS!
BROWSERS!
BROWSERS!
What do we want?
MORE SPEED!
MORE SPEED!
MORE SPEED!
And when do we want it?
RIGHT NOW!!!
RIGHT NOW!!!
RIGHT NOW!!!
BROWSERS!

Daily Recap

Today we learned

1. Multiple Linear Regression

Moving forward:

- nb day Friday

Next time in lecture:

- More Regression! More predictor!