Lecture 22: Multiple Linear Regression

Announcements and reminders

- HW 5 due next Friday

Previously on CSCI 3022…

Given data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, fit a simple linear regression of the form

$$y_i = \alpha + \beta x_i + \epsilon_i \quad \text{where} \quad \epsilon_i \sim N(0, \sigma^2)$$

Estimates of the intercept and slope parameters are given by minimizing

$$SSE = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$$

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

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

The least-squares estimates of the parameters are:

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \quad \text{and} \quad \hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Previously on CSCI 3022…

We can perform inference on slope parameter β to determine if relationship is significant

$$\hat{\sigma}^2 = \frac{SSE}{n-2} \qquad SE(\hat{\beta}) = \frac{\hat{\sigma}}{\sqrt{\sum_i (x_i - \bar{x})^2}} \qquad \text{CI:} \quad \hat{\beta} \pm t_{\alpha/2,\ n-2} \times SE(\hat{\beta})$$

We can use the Coefficient of Determination to evaluate the goodness-of-fit of SLR model

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad SSE = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2 \qquad R^2 = 1 - \frac{SSE}{SST}$$

If R^2 is close to 1, then the model fits the data relatively well.

$R^2 \approx 0$, $SSE \approx SST$

Regression with multiple features

Turns out, in most practical applications, there are multiple features/predictors that potentially have an effect on the response

Example: S'pose that Y represents the sale price of a house. What are some reasonable features associated with the sale price?

Regression with multiple features

Turns out, in most practical applications, there are multiple features/predictors that potentially have an effect on the response

Example: S'pose that Y represents the sale price of a house. What are some reasonable features associated with the sale price?

- x_1 -- interior size of the house
- x_2 -- size of the lot
- x_3 -- number of bedrooms
- x_4 -- number of bathrooms
- x_5 -- age of the house … and others?

Regression with multiple features

SLR : $\begin{cases} H_0 : \beta = 0 \\ H_1 : \beta \neq 0 \end{cases}$

Questions we would like to answer**:**

- Is at least one of the features useful in predicting the response?
- Do all of the feature help to explain the response? Or can we reduce to just a few?
- How well does the model fit the data? How well does just a subset of features do?
- Given a set of predictor values, what response should we predict, and how accurate is our prediction?

We'll look at these questions today and into next week

But first, let's do a little exploration of a multiple feature data set and remind ourselves of SLR

Advertising budget example

Get in groups, get out your laptops, and let's open the **nb 22** in-class notebook

Example: Data is provided about the sales of a particular product in 200 different markets, along with the advertising budgets for each market for three different media types: TV, radio and newspaper.

The sales response is given in thousands of units, and each of the advertising budget features are given in thousands of dollars.

Let's begin by fitting three individual SLR models with the advertising budget as the feature and the sales as the response.

Multiple linear regression

We've seen from the Advertising example that SLR analysis has indicated that there is a significant relationship between each of the media types (TV/radio/newspaper) and the sales of the product.

But individual SLR models only show the effect of each media type in a vacuum.

To get a clearer picture of what's going on, we want to consider the effect of all three advertising types on sales simultaneously

→ This is where Multiple Linear Regression (MLR) comes in!

Multiple linear regression

sales = $\beta_0 + \beta_1 \times TV + \beta_2 \times Radio + \beta_3 \times News$

$p = 3$ features

Definition: In MLR, the data is assumed to come from a model of the form:

SLP : $Y = \alpha + \beta X$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$$

$Y = \beta_0 + \beta_1 X_1$

p features

So for each of the n data points $(x_{i1}, x_{i2}, \ldots, x_{ip}, y_i)$, for i = 1, 2, …, n, we assume:

data point

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i$$

β_j = coefficient for the j^{th} feature

x_{ij} = j^{th} feature from the i^{th}

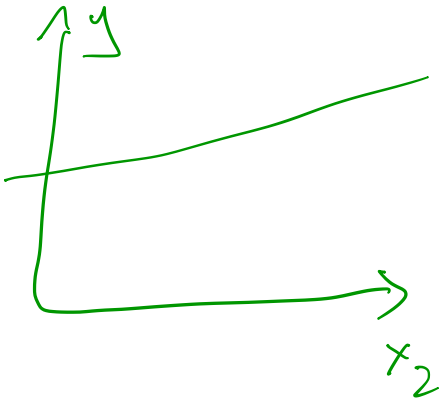
We make similar assumptions as in the case of SLR:

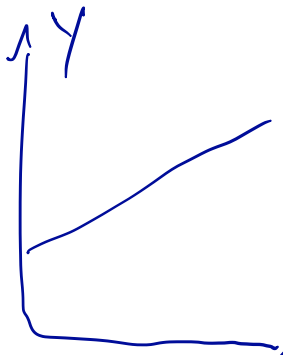
- Each ϵ_i is independent
- $\epsilon_i \sim N(0, \sigma^2)$

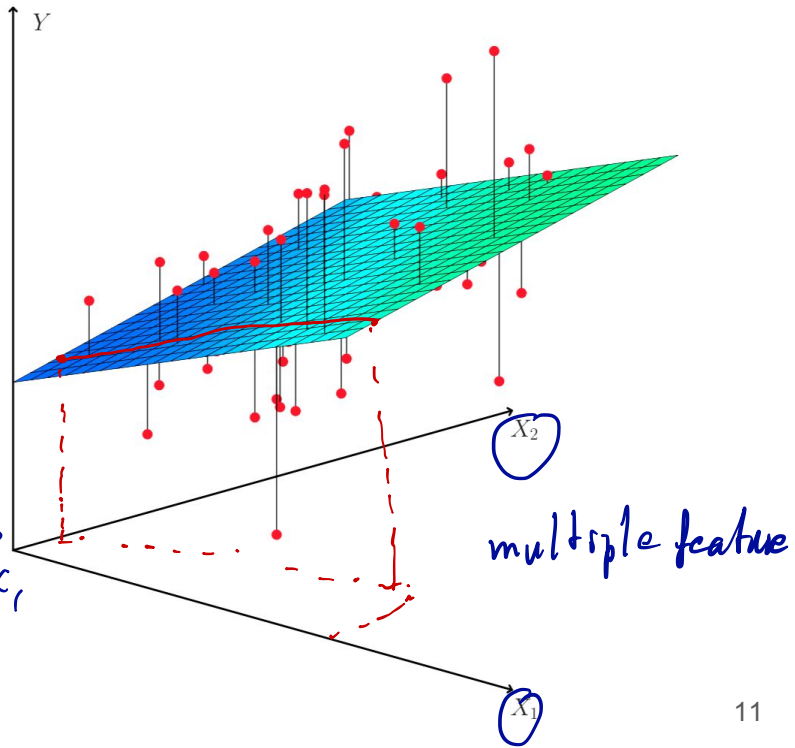
Multiple linear regression

Note that our model is no longer a simple line. Instead, it is a **linear surface**

- If you held all of the variables constant except for one of them, would look like a line as viewed from that variable's axis







Multiple linear regression

The interpretation of the model parameters is similar to that of SLR

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i$$

Parameter β_k is the expected change in the response associated with a unit change in the value of feature x_k **while all of the other features are held fixed**

$ 1000s k

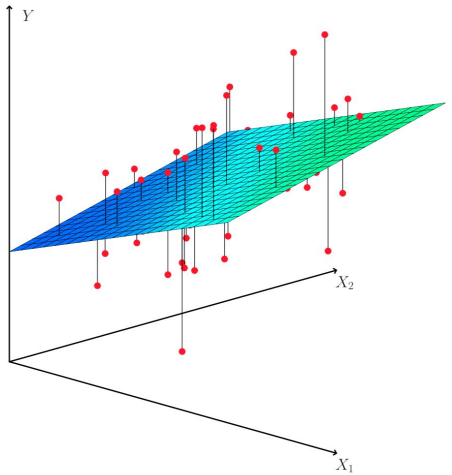
Example: House sale prices:

Y = $15 + 50 x_1 + 25 x_2 + 0.1 x_3$

x_1 = house s.q. ft. (1000 sq. ft.)

x_2 = # bedrooms

x_3 = # pizza that come w/ house



Estimating the MLR parameters

Just as in the case of SLR, we have no hope of discovering the true model parameters.

→ Need to **estimate** them from the data. Our estimated model will be:

As before, we will find the estimated parameters by minimizing the **sum of squared errors**:

The SSE is again interpreted as the measure of how much variation is left in the data that cannot be explained by the model.

Estimating the MLR parameters

Just as in the case of SLR, we have no hope of discovering the true model parameters.

→ Need to **estimate** them from the data. Our estimated model will be:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p$$

As before, we will find the estimated parameters by minimizing the **sum of squared errors**:

$$SSE = \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_1 + + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p))^2$$

$\hat{y}_i$, our model will predict

The SSE is again interpreted as the measure of how much variation is left in the data that cannot be explained by the model.

coeff of data: $R^2 = 1 - \frac{SSE}{SST}$

Estimating the MLR parameters

Just as in the case of SLR, we have no hope of discovering the true model parameters.

→ Need to **estimate** them from the data. Our estimated model will be:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p$$

As before, we will find the estimated parameters by minimizing the **sum of squared errors**:

$$SSE = \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_1 + + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p))^2$$

Note: Without linear algebra, it is difficult to write down an explicit expression for the parameter estimates. For now, we'll see how to obtain them in Python. Later, we can return to this problem using **stochastic gradient descent**!

Advertising budget example

Back to our groups! Let's see how we can find a MLR model for the advertising data.

Advertising budget example

Back to our groups! Let's see how we can find a MLR model for the advertising data.

… sale (TV=1, Radio=2, News=0) = 2.94 + 0.046 · 1 + 0.189 · 2 − 0.001 · 0

So. We see that the MLR model for the advertising data is:

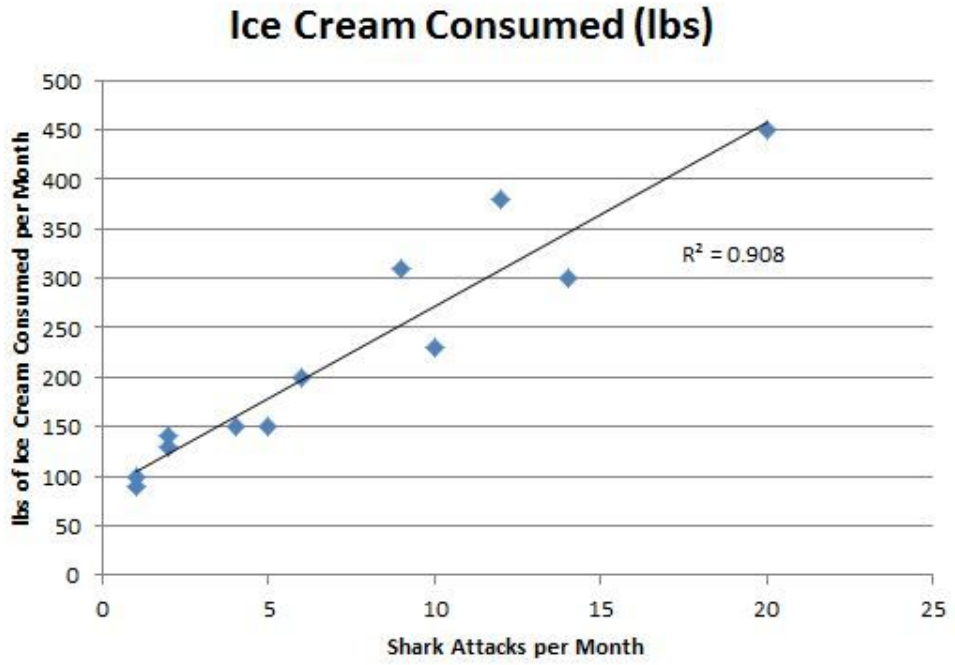
$$\text{sales} = 2.94 + 0.046 \times \text{TV} + 0.189 \times \text{radio} - 0.001 \times \text{news}$$

Baseline — Coeff how much we spend on TV

Question: Why did our SLR models indicate a positive relationship between **newspaper advertising** and **sales**, but our MLR model did not?

A correlation parable

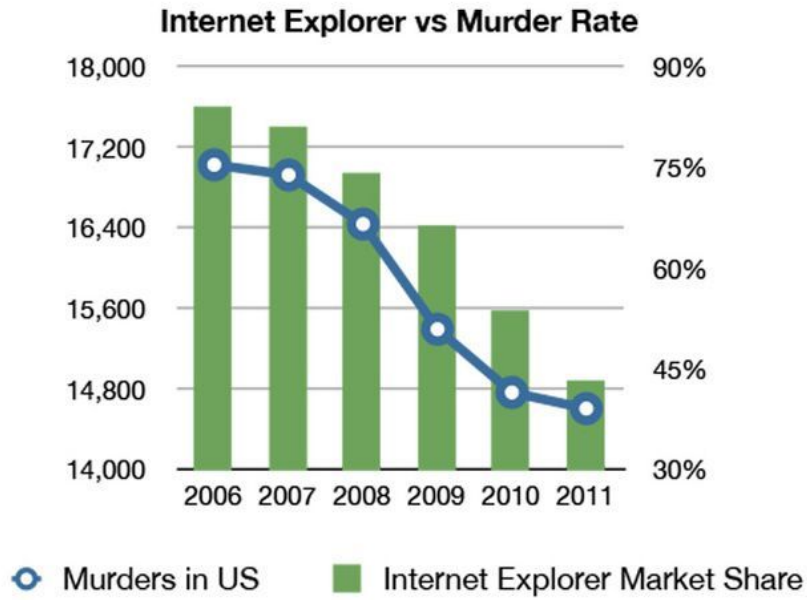
Example: A SLR analysis of shark attacks vs ice cream sales at a Southern California beach indicates that there is a strong relationship between the two.



A correlation parable

Example: 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.

Another example: Internet Explorer use vs Murders

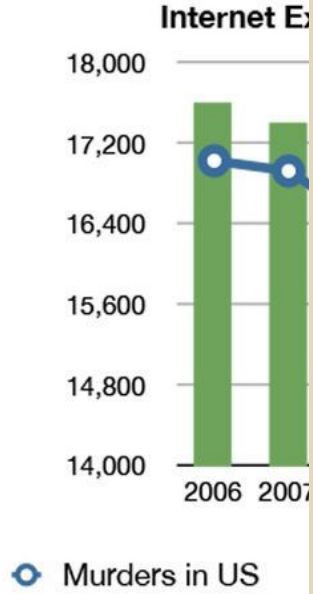


Question: Do you think that these relationships are real?

A correlation para

Example: A SLR anal rn California beach indicates that there is a

Another example: Int



hink that these hips are real?

A correlation parable

Example: 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.

Another example: Internet Explorer use vs Murders

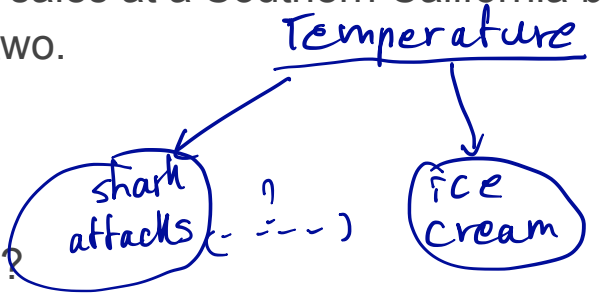
Question: Do you think that these relationships are real?

Answer: Probably not. **Temperature** explains both shark attacks and ice cream consumption.

- If we did an MLR analysis with shark attacks as the response and both temperature and ice cream sales as the features, our model would show the strong relationship between temperature and shark attacks, and an insignificant relationship between ice cream sales and shark attacks.
- If we **adjust** or **control** for temperature, then the relationship between ice cream sales and shark attacks disappears

A correlation parable

Example: 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.



Another example: Internet Explorer use vs Murders

Question: Do you think that these relationships are real?

Answer: Probably not. **Temperature** explains both shark attacks and ice cream consumption.

- If we did an MLR analysis with shark attacks as the response and both temperature and ice cream sales as the features, our model would show the strong relationship between temperature and shark attacks, and an insignificant relationship between ice cream sales and shark attacks.
- If we **adjust** or **control** for temperature, then the relationship between ice cream sales and shark attacks disappears

Advertising budget example

Question: Based on our absurd shark attack example, can you explain why newspaper spending became less significant in our MLR of product sales?

$$\text{sales} = 2.94 + 0.046 \times \text{TV} + 0.189 \times \text{radio} - 0.001 \times \text{news}$$

controlled for radio & TV, relation b/w news & sales disappears

-> radio & TV advertisng explain the response in sales

-> news doesn't give any

Covariance and correlation of features

$Var(x) = E[(x - E[x])^2]$

$= E[(x - E[x])(x - E[x])]$

One way to discover these relationships among features is to do a correlation analysis.

We want to know: if the value of one features changes, how will this affect the other features?

Definition: Let X and Y be random variables. The **covariance** between X and Y is given by

$$\mathrm{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$$

Definition: The **correlation coefficient** $\rho(X, Y)$ is a measure between -1 and 1, and given by

\rho (x, y)
↑ latex

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\ \mathrm{Var}(Y)}}$$

$\rho \simeq 1$ → strongly positively correla

$\rho \simeq -1$ → strongly negatively correla

$\rho > 0$

$\rho < 0$

$\rho \simeq 0$ → uncorrelad.

Covariance and correlation of features

We can estimate these relationships from the data using formulas analogous to the sample variance

Definition: The **sample covariance** is given by

Definition: The **sample correlation coefficient** is given by

Covariance and correlation of features

We can estimate these relationships from the data using formulas analogous to the sample variance

look exactly like variance s_x^2 in sample x

Definition: The **sample covariance** is given by

$$S_{XY}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Definition: The **sample correlation coefficient** is given by

$$\hat{\rho}(X,Y) = \frac{S_{XY}^2}{\sqrt{S_X^2 S_Y^2}}$$

$s_x^2 = \frac{1}{n_x - 1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Advertising budget example

Let's compute the pairwise correlation coefficients for the TV, radio and newspaper spending features in the advertising data

```
In [9]: dfAd[["tv", "radio", "news"]].corr()
```

Out[9]:

	tv	radio	news
tv	1.000000	0.054809	0.056648
radio	0.054809	1.000000	0.354104
news	0.056648	0.354104	1.000000

p(x, y)

p(radio, news)

Question: What do you notice?

radio & news are correlated!
TV & radio explained the variability in sales
∴ news had nothing else to add.

Polynomial regression

Now we will look at how we can use MLR to explain **nonlinear** relationships between single-feature data and the response

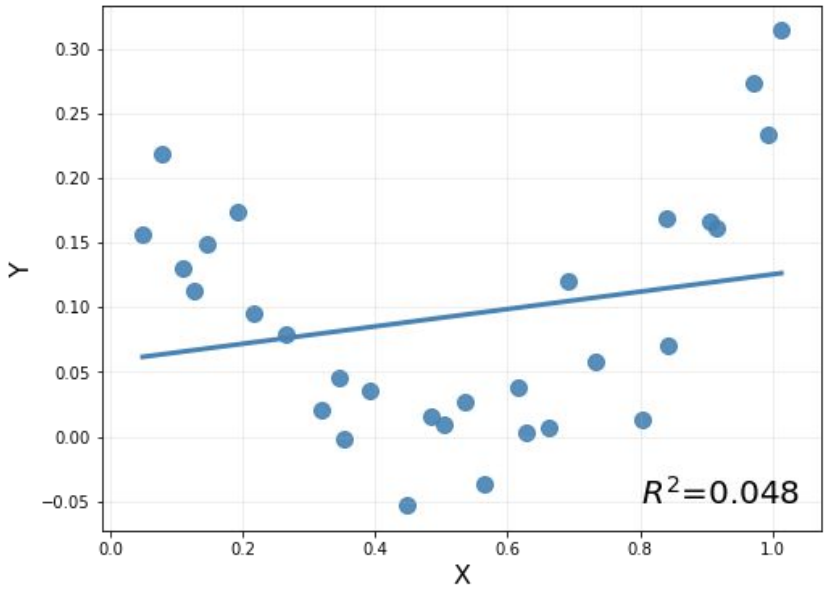
look alot more like a parabolu!

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

$Y = \beta_0 + \beta_1 x_1 + \beta_2 x$

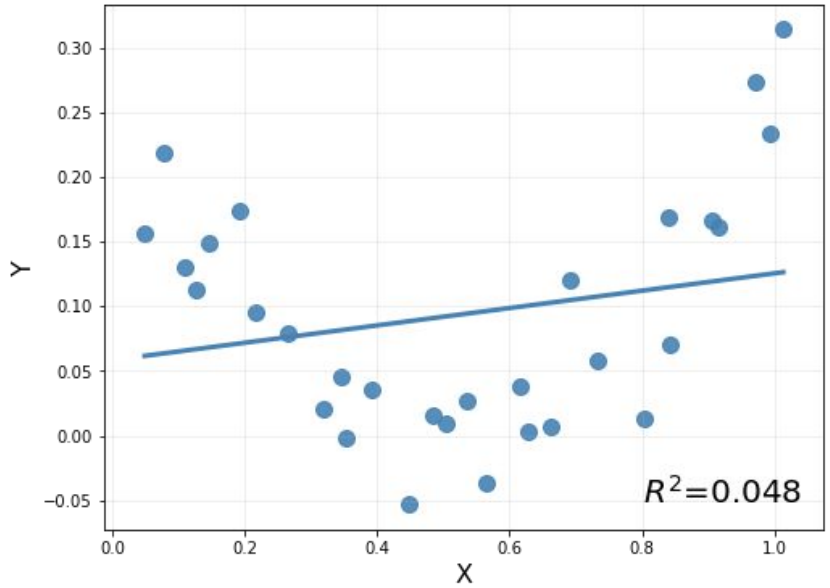
$x_2 \cong x^2$

x	x^2	y
1	1	⋮
2	4	
5	25	
⋮	⋮	



Polynomial regression

For single-feature data, we can fit a polynomial regression model by casting it as a multiple linear regression, where the additional features are **powers** of the original single feature, x



Residual plots, in polynomial regression

$\varepsilon \sim N(0, \sigma^2)$

Recall that the assumed nature of our true model is: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots \beta_p x_p + \varepsilon$

if true model ▷ : $Y = \beta_0 + \beta_1 x + \varepsilon$

& we assume : $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Residuals : $r = Y - \hat{Y} = \varepsilon \sim N(0, \sigma^2)$

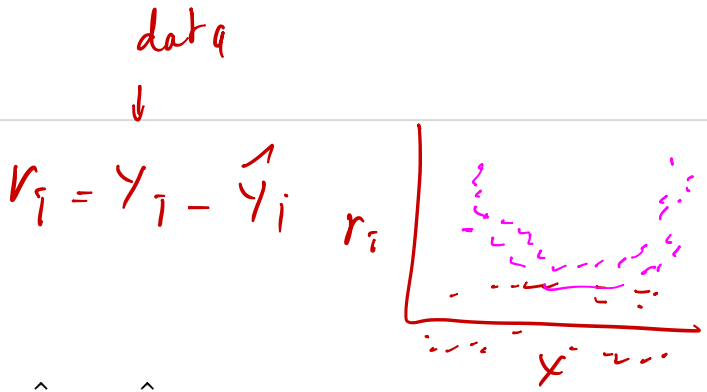
if true model : $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$

& we assume : $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$

residual : $r = y - \hat{y} = \beta_2 x^2 + \varepsilon \sim N(\beta_2 x^2, \sigma^2)$

residuals centred around the missing feature

Residual plots, in polynomial regression



Recall that the assumed nature of our true model is:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_p x^p + \epsilon$$

If true model is $y = \beta_0 + \beta_1 x + \epsilon$, and our model is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \epsilon$

then $r = y - \hat{y} \sim N(0, \sigma^2)$

If true model is $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$, and our model is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \epsilon$

then $r = y - \hat{y} \sim N(\beta_2 x^2, \sigma^2)$

Residual plots, in polynomial regression

Recall that the assumed nature of our true model is:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_p x^p + \epsilon$$

If true model is $y = \beta_0 + \beta_1 x + \epsilon$, and our model is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \epsilon$

then $r = y - \hat{y} \sim N(0, \sigma^2)$

If true model is $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$, and our model is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \epsilon$

then $r = y - \hat{y} \sim N(\beta_2 x^2, \sigma^2)$

In general: If you plot the residuals $r = y - \hat{y}$, they should be **normally distributed around the missing feature**. So add that to your model!

Okay! Let's get (back) to work!

Get in groups, get out laptops, and open **nb 22** notebook back up

Let's…

- Dig into comparisons of **multiple linear regression** versus **simple linear regression**!
- Get some practice working with **polynomial regression**!
- Get some practice guessing **polynomial features** using **residual plots**!
- Figure out why **Internet Explorer use causes murders**! (not really)