INFO 2950: Intro to Data Science

Lecture 11
2023-09-27

Agenda

1. **Admin**
2. **Logistic Regression Review**
   a. Logit interpretations
3. **Linear Regression Review**
   a. Log-log interpretations
   b. Multivariable Dummies

Admin

- HW2 solutions posted on Canvas
- HW3 **must** be submitted by Friday
  - Questions tagged on Gradescope correctly
  - PDFs not cut off
  - Problem 0 filled out fully

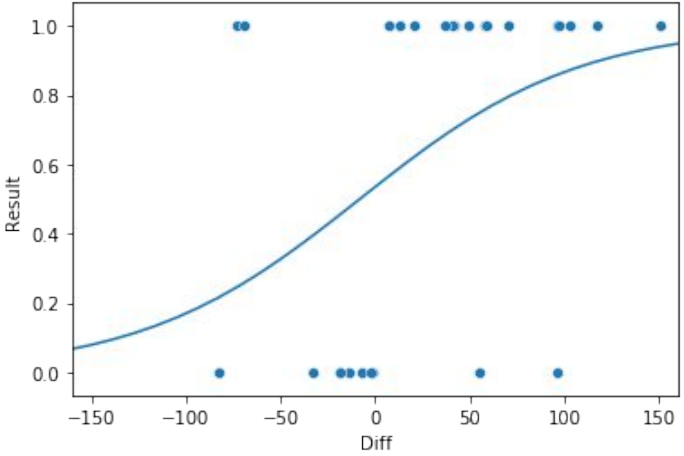
Prelim locations: Oct 2nd during class

- Last name A-T in Ives 305 (this room)
- Last name **U-Z** in Sage Hall B01
- SDS accommodations: emailed to you via the Alternative Testing Program (ATP); please let us know if you did not receive an email already

How do we derive interpretations for logistic regression?

Magnus Carlsen Hans Niemann

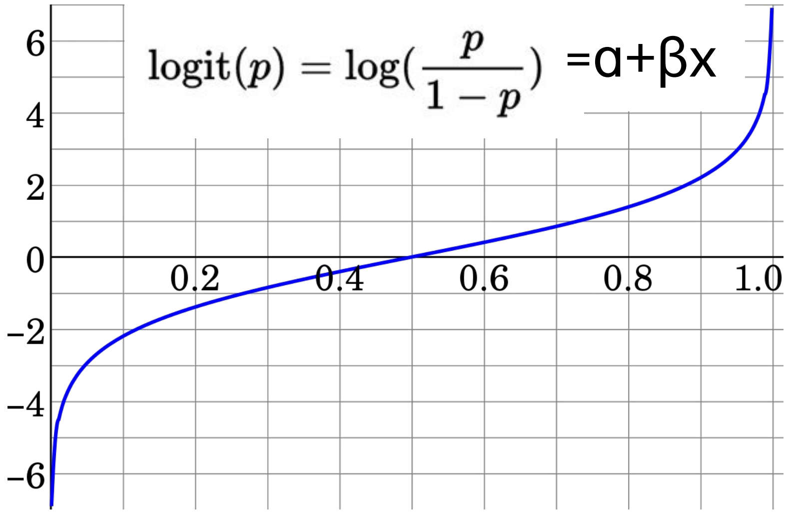
The two players involved in the controversy



https://en.wikipedia.org/wiki/Carlsen%E2%80%93Niemann_controversy

Deriving logit interpretations

1. Summarize
2. Prediction (intercept)
3. Oddities

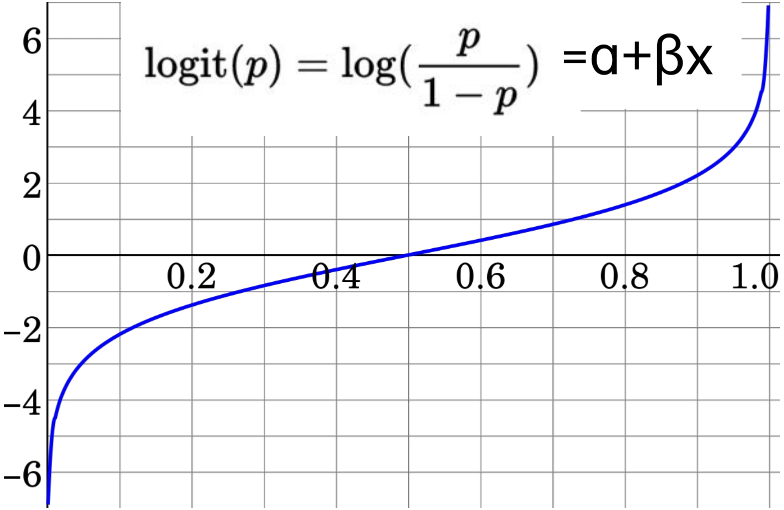


The probability that x=0 yields output y=1 is $e^{\alpha}/(e^{\alpha}+1)$

If x=0, then probability p = $e^{\alpha}/(e^{\alpha}+1)$

https://en.wikipedia.org/wiki/Logit

Deriving logit interpretations

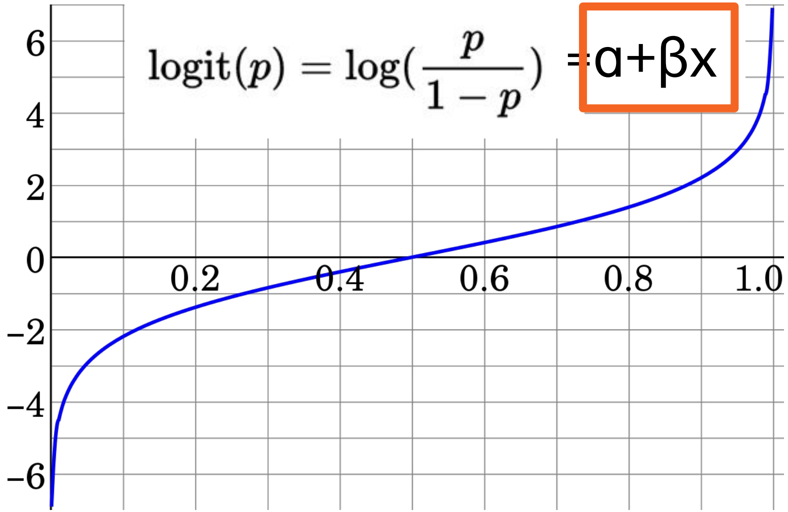


To deal with the log odds ratio, we just have to "solve" for p (the probability that yields output y=1)

The probability that x=0 yields output y=1 is $e^{\alpha}/(e^{\alpha}+1)$

https://en.wikipedia.org/wiki/Logit

Deriving logit interpretations



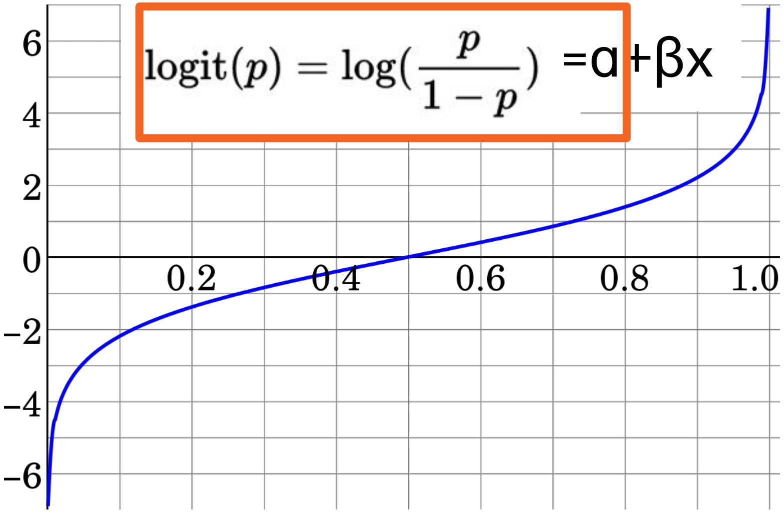
To deal with the log odds ratio, we just have to “solve” for p (the probability that yields output y=1)

The probability that x=0 yields output y=1 is $e^{\alpha}/(e^{\alpha}+1)$

- If x = 0, then RHS expression $\alpha+\beta x = \alpha$

https://en.wikipedia.org/wiki/Logit

Deriving logit interpretations



$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \alpha + \beta x$$

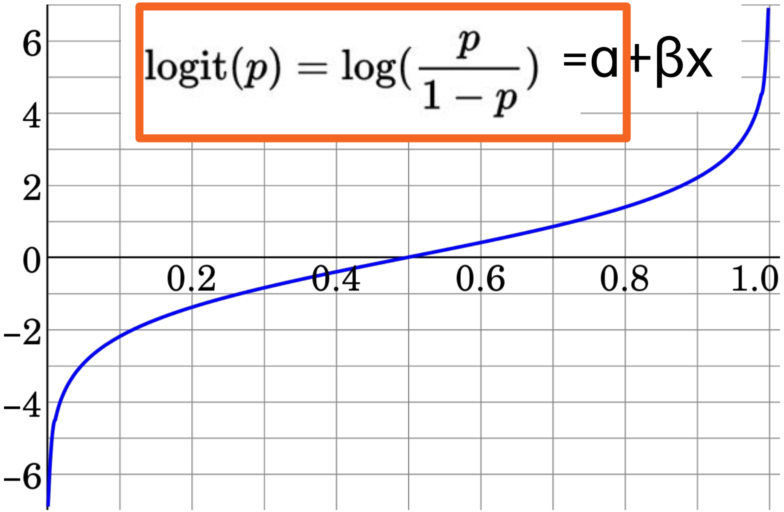
To deal with the log odds ratio, we just have to "solve" for p (the probability that yields output y=1)

The probability that x=0 yields output y=1 is $e^{\alpha}/(e^{\alpha}+1)$

- If x = 0, then RHS expression α+βx = α
- We set logit(p) = RHS = α

https://en.wikipedia.org/wiki/Logit

Deriving logit interpretations



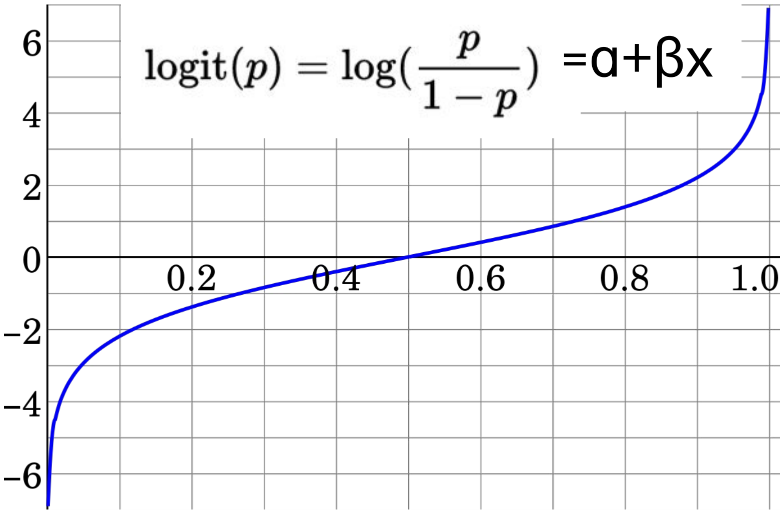
To deal with the log odds ratio, we just have to "solve" for p (the probability that yields output y=1)

The probability that x=0 yields output y=1 is $e^{\alpha}/(e^{\alpha}+1)$

- If x = 0, then RHS expression $\alpha+\beta x = \alpha$
- We set logit(p) = RHS = α
- logit(p) = log(p/[1-p]) = α

https://en.wikipedia.org/wiki/Logit

Deriving logit interpretations



$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \alpha+\beta x$$

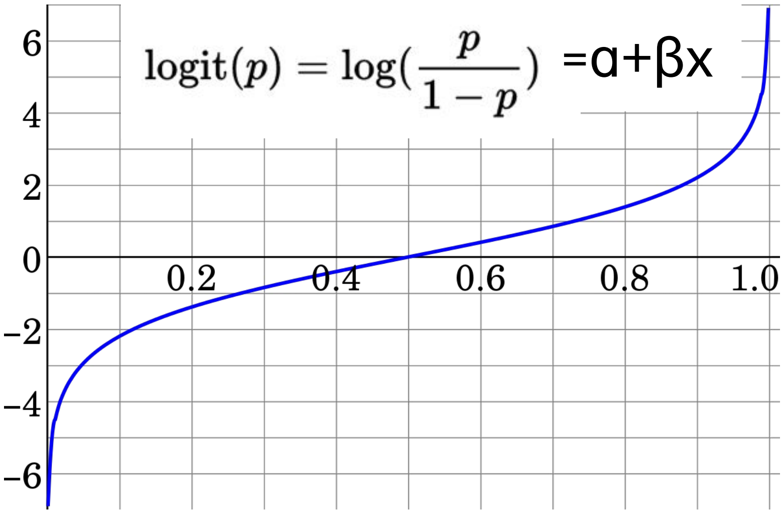
To deal with the log odds ratio, we just have to "solve" for p (the probability that yields output y=1)

> The probability that x=0 yields output y=1 is $e^{\alpha}/(e^{\alpha}+1)$

- If x = 0, then RHS expression $\alpha+\beta x = \alpha$
- We set logit(p) = RHS = α
- logit(p) = $\boxed{\log(p/[1-p]) = \alpha}$ exponentiate!
- $p/[1-p] = e^{\alpha}$

https://en.wikipedia.org/wiki/Logit

Deriving logit interpretations



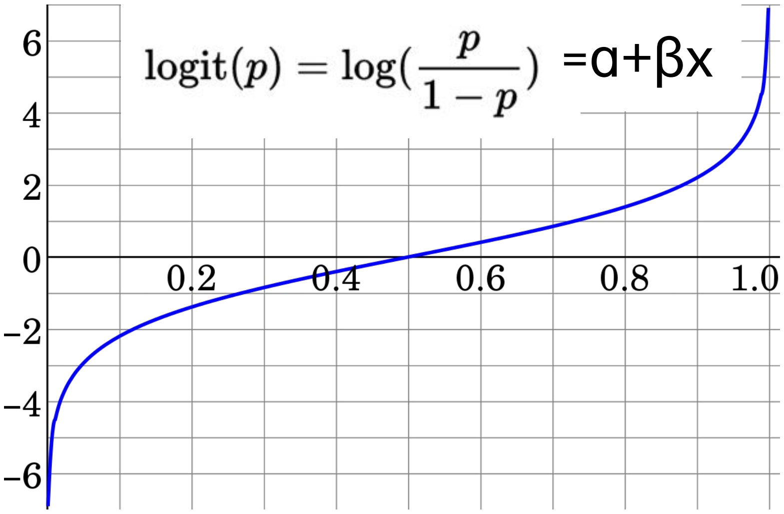
To deal with the log odds ratio, we just have to "solve" for p (the probability that yields output y=1)

The probability that x=0 yields output y=1 is $e^{\alpha}/(e^{\alpha}+1)$

- If x = 0, then RHS expression $\alpha+\beta x = \alpha$
- We set logit(p) = RHS = α
- logit(p) = log(p/[1-p]) = α
- $\boxed{p/[1-p] = e^{\alpha}}$ arithmetic: solve for p
- p = __________

https://en.wikipedia.org/wiki/Logit

Deriving logit interpretations



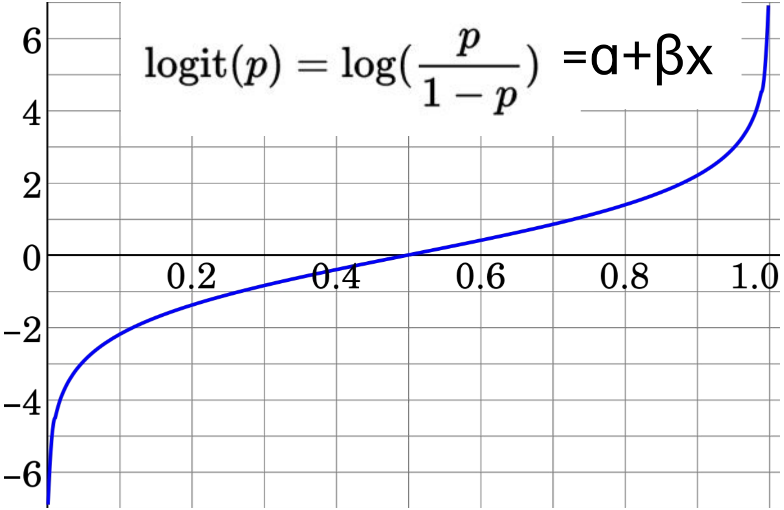
To deal with the log odds ratio, we just have to "solve" for p (the probability that yields output y=1)

The probability that x=0 yields output y=1 is $e^{\alpha}/(e^{\alpha}+1)$

- If x = 0, then RHS expression $\alpha+\beta x = \alpha$
- We set logit(p) = RHS = α
- $\text{logit}(p) = \log(p/[1-p]) = \alpha$
- $p/[1-p] = e^{\alpha}$
- $\boxed{p = e^{\alpha}/ (e^{\alpha}+1)}$

https://en.wikipedia.org/wiki/Logit

Deriving logit interpretations



$$\operatorname{logit}(p) = \log\left(\frac{p}{1-p}\right) = \alpha + \beta x$$

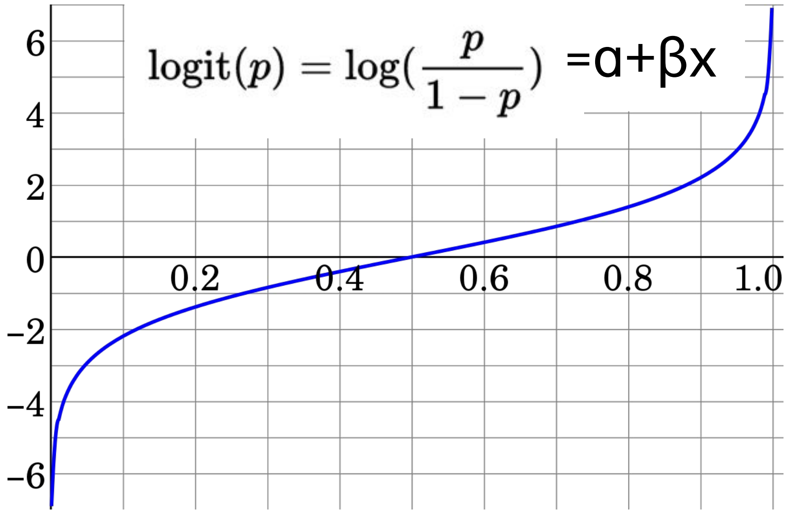
The probability that x=0 yields output y=1 is $e^{\alpha}/(e^{\alpha}+1)$

Note: you can use a similar process to do predictions with other values of x (instead of 0)!

https://en.wikipedia.org/wiki/Logit

Deriving logit interpretations

For a 1 unit change in x, we expect the odds of y to be multiplied by e^{β}

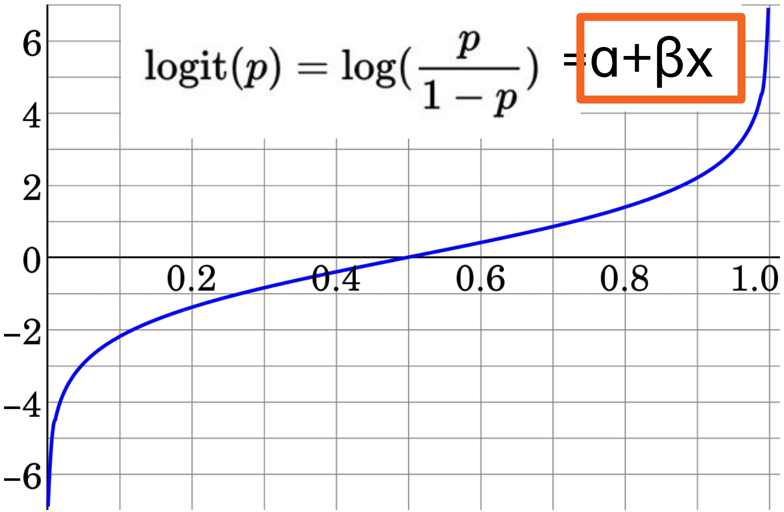


1. **Summarize**
2. **Prediction (intercept)**
3. **Oddities**

https://en.wikipedia.org/wiki/Logit

Deriving logit interpretations

For a 1 unit change in x, we expect the odds of y to be multiplied by e^{β}

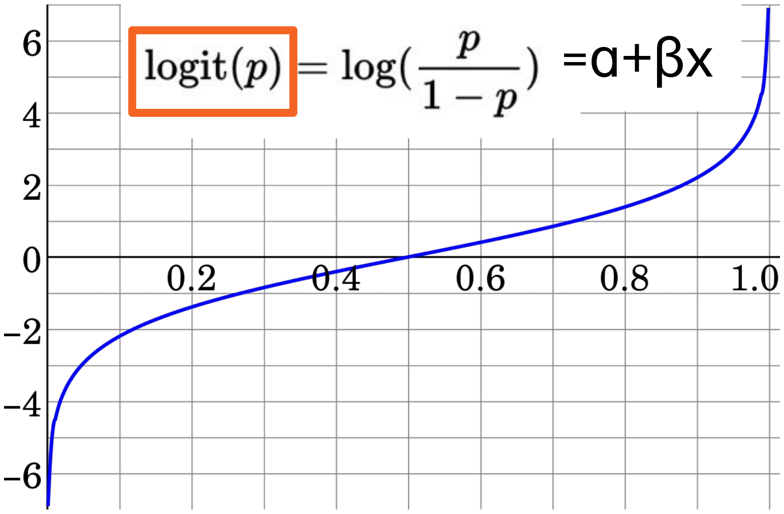


- If x increases by 1 unit,
  → RHS (α+βx) total increases by β

Deriving logit interpretations

For a 1 unit change in x, we expect the odds of y to be multiplied by e^{β}

$$\boxed{\mathrm{logit}(p)} = \log\left(\frac{p}{1-p}\right) = \alpha+\beta x$$

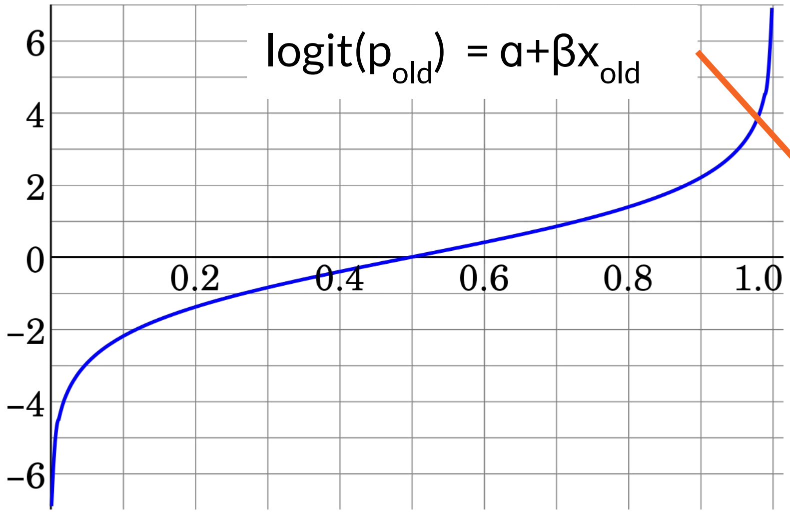


- If x increases by 1 unit,
  → RHS ($\alpha+\beta x$) total increases by β
  → LHS logit(p) increases by β

https://en.wikipedia.org/wiki/Logit

Deriving logit interpretations

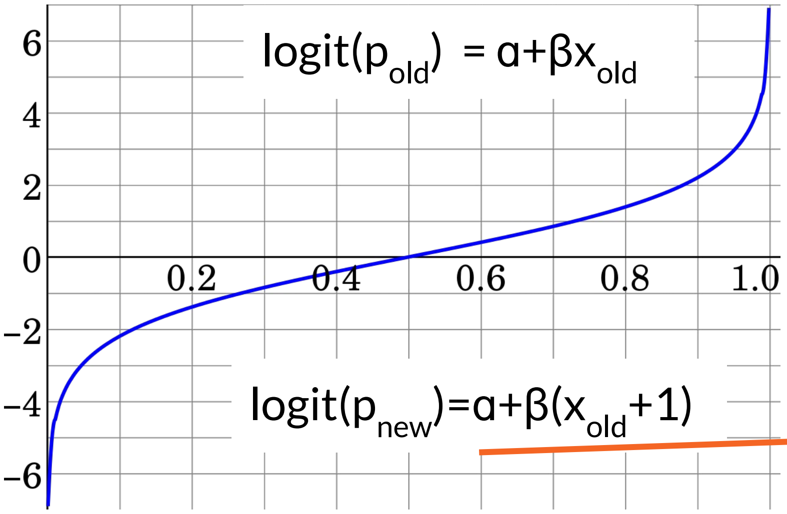
For a 1 unit change in x, we expect the odds of y to be multiplied by e^{β}



- If x increases by 1 unit,
  → RHS ($\alpha+\beta x$) total increases by β
  → LHS logit(p) increases by β
- Let our original value of x give us: $\text{logit}(p_{old})$

https://en.wikipedia.org/wiki/Logit

Deriving logit interpretations

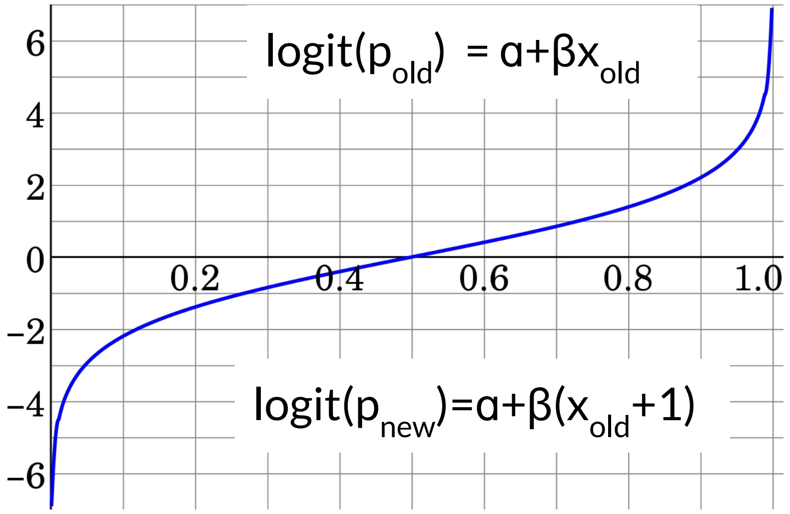


For a 1 unit change in x, we expect the odds of y to be multiplied by e^{β}

- If x increases by 1 unit,
  → RHS ($\alpha+\beta x$) total increases by β
  → LHS logit(p) increases by β
- Let our original value of x give us: $\text{logit}(p_{old})$
- Our new value of x+1 gives us $\text{logit}(p_{new})$

https://en.wikipedia.org/wiki/Logit

Deriving logit interpretations



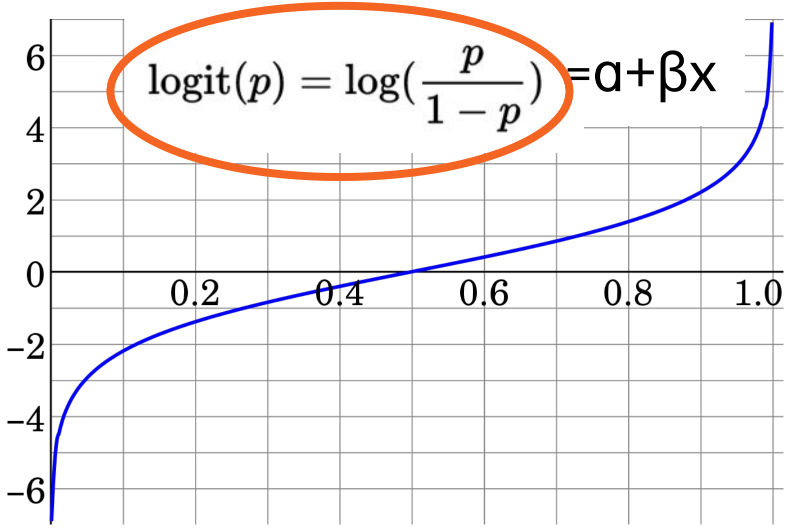
For a 1 unit change in x, we expect the odds of y to be multiplied by e^{β}

- $\text{logit}(p_{new}) = \text{logit}(p_{old}) + \beta$

https://en.wikipedia.org/wiki/Logit

Deriving logit interpretations

For a 1 unit change in x, we expect the odds of y to be multiplied by e^{β}

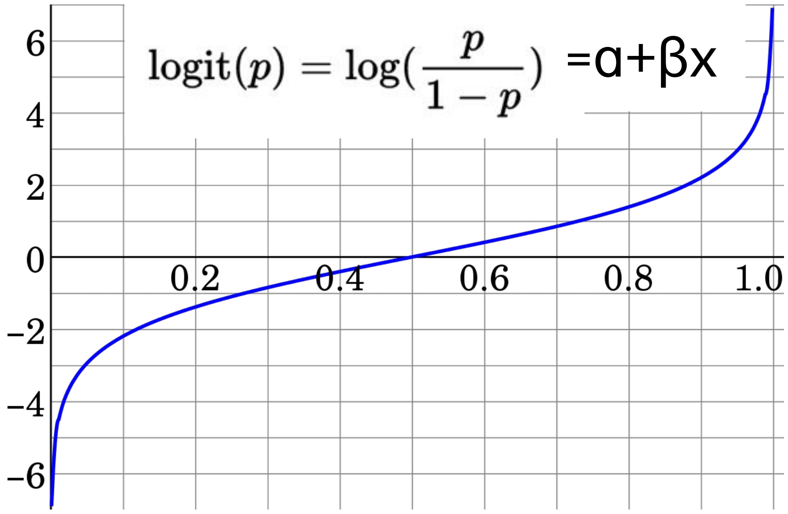


- $\mathrm{logit}(p_{new}) = \mathrm{logit}(p_{old}) + \beta$
- $\log(p_{new} / [1-p_{new}]) = \log(p_{old} / [1-p_{old}]) + \beta$

https://en.wikipedia.org/wiki/Logit

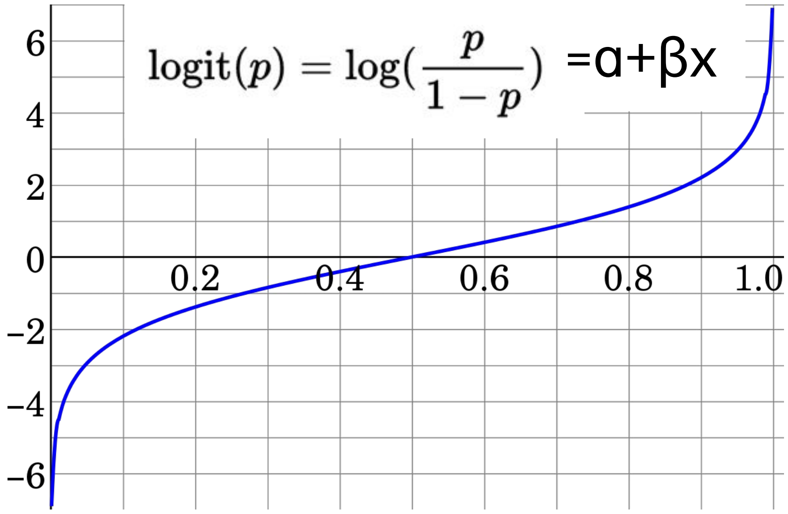
Deriving logit interpretations

For a 1 unit change in x, we expect the odds of y to be multiplied by e^{β}



- $\text{logit}(p_{new}) = \text{logit}(p_{old}) + \beta$
- $\log(p_{new} / [1-p_{new}]) = \log(p_{old} / [1-p_{old}]) + \beta$
- Solve for odds ratio difference:
  - $p_{new} / [1-p_{new}]$ in terms of $p_{old} / [1-p_{old}]$

Deriving logit interpretations



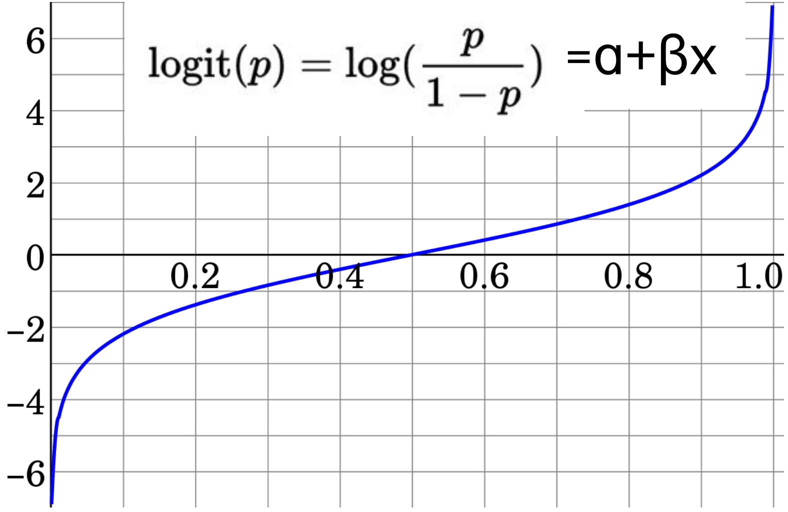
For a 1 unit change in x, we expect the odds of y to be multiplied by e^{β}

- $\text{logit}(p_{new}) = \text{logit}(p_{old}) + \beta$ exponentiate!
- $\boxed{\log(p_{new} / [1-p_{new}]) = \log(p_{old} / [1-p_{old}]) + \beta}$
- Solve for odds ratio difference:
  - $p_{new} / [1-p_{new}]$ in terms of $p_{old} / [1-p_{old}]$
- $\mathbf{p_{new} / [1-p_{new}] = e^{\beta} * p_{old} / [1-p_{old}]}$

https://en.wikipedia.org/wiki/Logit

Deriving logit interpretations

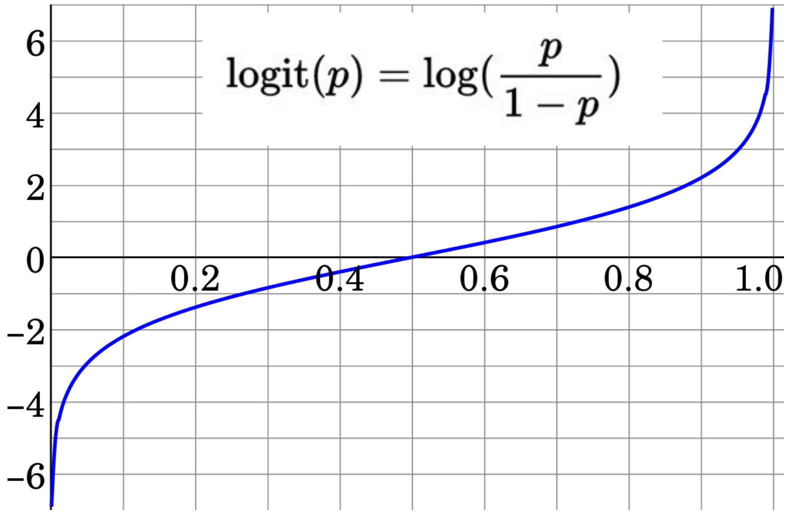
For a 1 unit change in x, we expect the odds of y to be multiplied by e^{β}



- $\text{logit}(p_{new}) = \text{logit}(p_{old}) + \beta$
- $\log(p_{new} / [1-p_{new}]) = \log(p_{old} / [1-p_{old}]) + \beta$
- Solve for odds ratio difference:
  - $p_{new} / [1-p_{new}]$ in terms of $p_{old} / [1-p_{old}]$
- $\boxed{p_{new} / [1-p_{new}] = e^{\beta} * p_{old} / [1-p_{old}]}$

https://en.wikipedia.org/wiki/Logit

Deriving logit interpretations



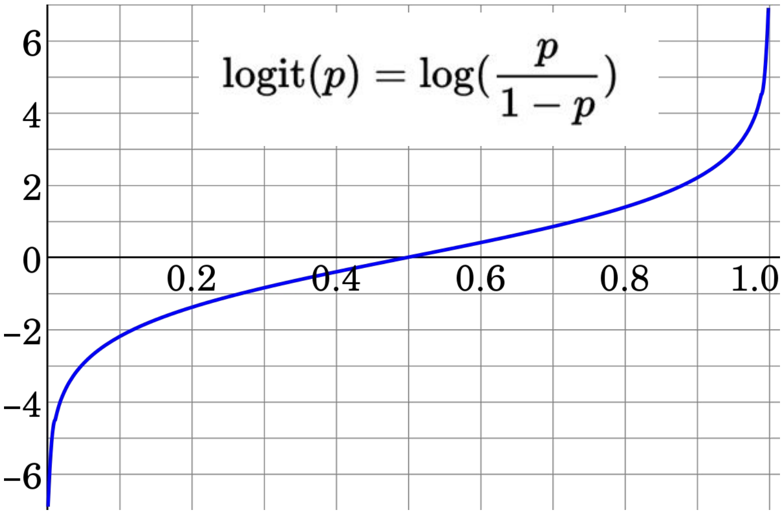
Summarizing

For a 1 unit change in x, we expect the odds of y to be multiplied by e^{β}

1 unit change in x is associated with a $100*(e^{\beta} - 1)\%$ change in y

https://en.wikipedia.org/wiki/Logit

Deriving logit interpretations



For a 1 unit change in x, we expect the odds of y to be multiplied by e^{β}

1 unit change in x is associated with a $100*(e^{\beta} - 1)$% change in y

- Simply convert the previous “summarizing” interpretation from a multiplicative value to a percentage!
    - E.g. if something is doubled (2x), there’s a 100*(2-1)=100% increase

https://en.wikipedia.org/wiki/Logit

https://www.femina.in/beauty/10-step-korean-skin-care-routine-203955.html

Formalizing multivariable regression

i	x	y
1	78	18
2	83	14
…	…	…

i	x_1	x_2	x_3	y
1	78	0	30.5	18
2	83	1	28.0	14
…	…	…	…	…

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

$$\Downarrow$$

$$y_i = \alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \ldots + \varepsilon_i$$

Formalizing multivariable regression

i	x	y
1	78	18
2	83	14
…	…	…

i	x_1	x_2	x_3	y
1	78	0	30.5	18
2	83	1	28.0	14
…	…	…	…	…

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

$$\Downarrow$$

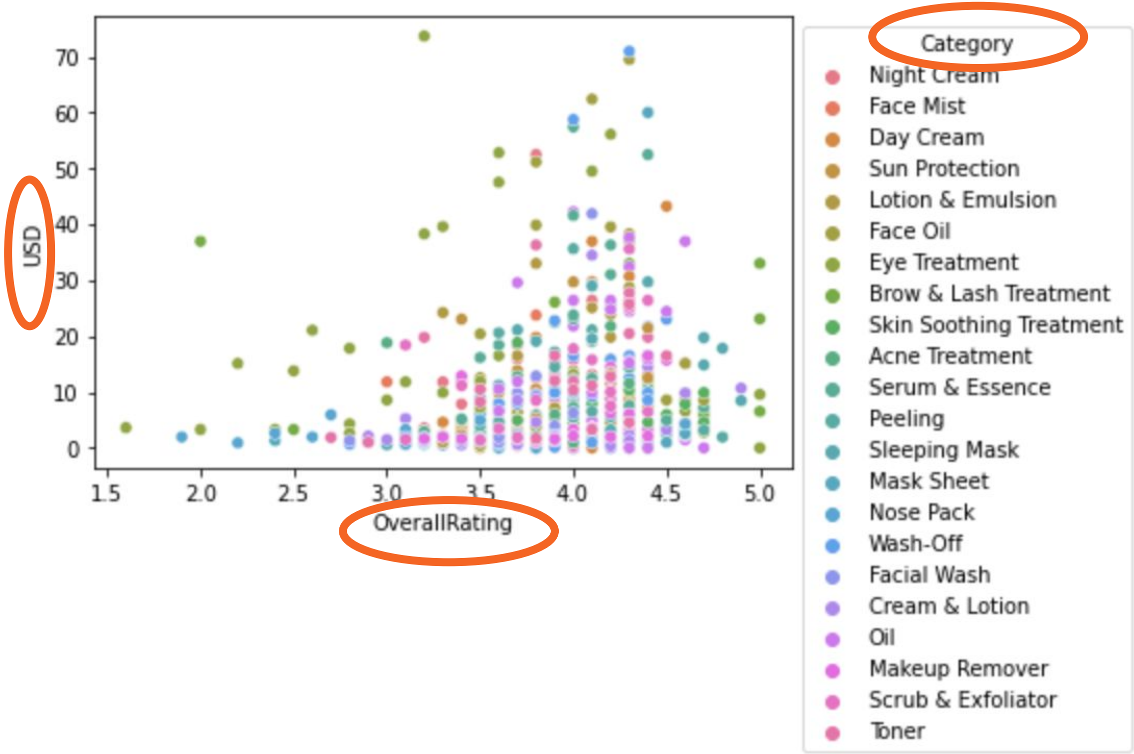
$$y_i = \alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \ldots + \varepsilon_i$$

skincare_df (data from Indonesia)

Product	USD	Category	Brand	OverallRating
Perfect 3D Gel	6.01	Night Cream	Hada Labo	3.8
Aqua Beauty Protecting Mist	1.78	Face Mist	PIXY	4.2
Thermal Spring Water	13.13	Face Mist	Avene	4.4
White Secret Night Cream	6.47	Night Cream	Wardah	3.6
Mineral Water Spray	10.56	Face Mist	Evian	3.8
...	...	...	...	...
Vitamin E Hydrating Toner	11.15	Toner	The Body Shop	4.1
Skin Perfecting 2% BHA Liquid Exfoliant	25.74	Toner	Paula's Choice	4.3
Facial Lotion	0.99	Toner	Ovale	2.9
Centella Water Alcohol-Free Toner	10.36	Toner	Cosrx	4.0
Rose Water Toner	12.76	Toner	Mamonde	4.2

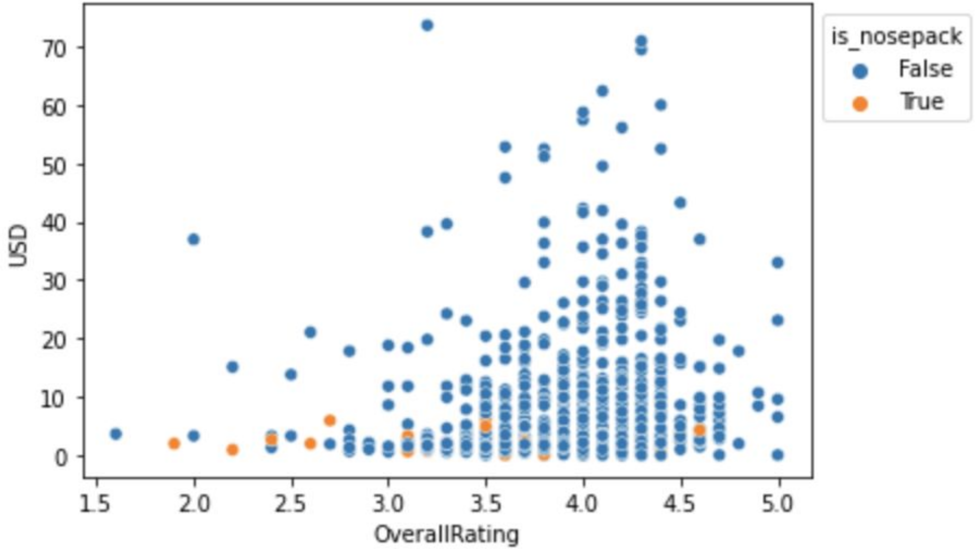
https://www.kaggle.com/datasets/hafidahmusthaanah/skincare-review

What if our outcome is a binary outcome?



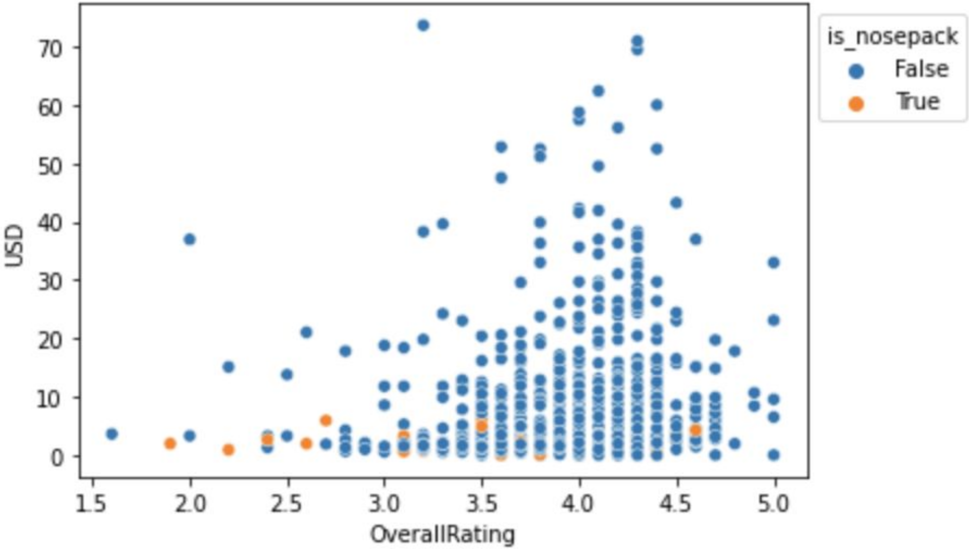
https://www.kaggle.com/datasets/hafidahmusthaanah/skincare-review

```python
ax = sns.scatterplot(data=skincare_df,
    x="OverallRating", y="USD",
    hue="is_nosepack")
```



```python
ax = sns.scatterplot(data=skincare_df,
    x="OverallRating", y="USD",
    hue="is_nosepack")
```

Hypothesis: we can use product *ratings* and *prices* to predict whether a product is a nose pack or not



Multivar Logistic Regression (sklearn)

Define X

```
X = skincare_df[["OverallRating", "USD"]]
y = skincare_df[["is_nosepack"]].values.ravel()
m2 = LogisticRegression().fit(X,y)
yhat = m2.predict(X)
m2.intercept_
m2.coef_
```

Multivar Logistic Regression (sklearn)

```
X = skincare_df[["OverallRating", "USD"]]
y = skincare_df[["is_nosepack"]].values.ravel()
m2 = LogisticRegression().fit(X,y)
yhat = m2.predict(X)
m2.intercept_
m2.coef_
```

Define y

ravel “flattens” 2d array
so [[1,2,3]] → [1,2,3]

Multivar Logistic Regression (sklearn)

Define model m2 using a Logistic Regression

```
X = skincare_df[["OverallRating", "USD"]]
y = skincare_df[["is_nosepack"]].values.ravel()
m2 = LogisticRegression().fit(X,y)
yhat = m2.predict(X)
m2.intercept_
m2.coef_
```

Multivar Logistic Regression (sklearn)

```
X = skincare_df[["OverallRating", "USD"]]
y = skincare_df[["is_nosepack"]].values.ravel()
m2 = LogisticRegression().fit(X,y)
yhat = m2.predict(X)
m2.intercept_
m2.coef_
```

Predict and get α, β’s; same as before

Multivar Logit: Formulation

- y = is product a nose pack?
- x_1 = avg customer rating; x_2 = price in \$
- Hypothesis: we expect to see a [positive? negative?] effect of x_i on y
- $y \sim \sigma\,(x_1 + x_2)$
- $y \sim \sigma(\alpha + \beta_1\,x_1 + \beta_2\,x_2)$

Multivar Logit: Formulation

- y = is product a nose pack?
- x_1 = avg customer rating; x_2 = price in $
- Hypothesis: I expect to see negative effects of both rating and price on y
- $y \sim \sigma\,(x_1 + x_2)$
- $y \sim \sigma(\alpha + \beta_1 x_1 + \beta_2 x_2)$

Multivar Logistic Regression

```
X = skincare_df[["OverallRating", "USD"]]
y = skincare_df[["is_nosepack"]].values.ravel()
m2 = LogisticRegression().fit(X,y)
yhat = m2.predict(X)
m2.intercept_    →    array([1.47402304])
m2.coef_   →   array([[-0.90925508, -0.30896428]])
```

For a 1 unit change in x, we expect the odds of y to be multiplied by e^{β}

1 unit change in x is associated with a $100*(e^{\beta} - 1)\%$ change in y

Interpret: summarize x_1

- y = is product a nose pack?
- x_1 = avg customer rating; x_2 = price in $
- $y \sim \text{sigmoid}(1.5 - 0.9x_1 - 0.3x_2)$
- $e^{-0.9} \approx 0.4$, $e^{-0.3} \approx 0.7$
- ______________________________

For a 1 unit change in x, we expect the odds of y to be multiplied by e^{β}

1 unit change in x is associated with a $100*(e^{\beta} – 1)\%$ change in y

Interpret: summarize x_1

- y = is product a nose pack?
- x_1 = avg customer rating; x_2 = price in $
- y ~ sigmoid($1.5 - 0.9x_1 - 0.3x_2$)
- $e^{-0.9} \approx 0.4$, $e^{-0.3} \approx 0.7$
- Our model estimates that, all else equal, for each additional star rating given to the product, the odds of the product being a nose pack are multiplied by 0.4 (a.k.a. 100*(0.4-1)% = decrease of 60% in odds)

Interpret: predict ŷ

- y = is product a nose pack?
- x_1 = avg customer rating; x_2 = price in $
- $y \sim \text{sigmoid}(1.5 - 0.9x_1 - 0.3x_2)$
- **Predict for $x_1 = 0$ and $x_2=0$:**
- **Hint:** $e^R/(e^R+1)\approx 0.82$ where **R refers to the RHS of solving:** $\text{logit}(p) = \log(p/[1-p]) = \alpha+\beta_1x_1 + \beta_2x_2$

Interpret: predict ŷ

- y = is product a nose pack?
- x_1 = avg customer rating; x_2 = price in $
- y ~ sigmoid($1.5 - 0.9x_1 - 0.3x_2$)
- $e^R/(e^R+1) = e^{1.5}/(e^{1.5}+1) \approx 0.82$

Remember, we want to solve prediction:
$\text{logit}(p) = \log(p/[1-p]) = \alpha+\beta_1 x_1 + \beta_2 x_2$

If $x_1=0$ and $x_2=0$, RHS is just $\alpha = 1.5$

$p = e^R/(e^R+1) = e^{1.5}/(e^{1.5}+1) = 0.82$

Interpret: predict ŷ

- y = is product a nose pack?
- x_1 = avg customer rating; x_2 = price in $
- $y \sim \text{sigmoid}(1.5 - 0.9x_1 - 0.3x_2)$
- $e^{1.5}/(e^{1.5}+1) \approx 0.82$
- Our model predicts that, for a product with an average rating of 0 stars and price of 0, there is a $e^R/(e^R+1) = 0.82$ probability that the product is a nose pack.

Remember, we want to solve prediction:
$\text{logit}(p) = \log(p/[1-p]) = \alpha+\beta_1 x_1 + \beta_2 x_2$

If $x_1=0$ and $x_2=0$, RHS is just $\alpha = 1.5$

$p = e^R/(e^R+1) = e^{1.5}/(e^{1.5}+1) = 0.82$

Interpret: predict ŷ

- y = is product a nose pack?
- x_1 = avg customer rating; x_2 = price in $
- $y \sim \text{sigmoid}(1.5 - 0.9x_1 - 0.3x_2)$
- Predict for $x_1 = 3$ and $x_2=3$:
- Hint: $e^R/(e^R+1) \approx 0.12$ where R refers to the RHS of solving: $\text{logit}(p) = \log(p/[1-p]) = \alpha+\beta_1 x_1 + \beta_2 x_2$

Interpret: predict $\hat{y}$

- y = is product a nose pack?
- x_1 = avg customer rating; x_2 = price in $
- $y \sim \text{sigmoid}(1.5 - 0.9x_1 - 0.3x_2)$
- Our model predicts that, for a product with an average rating of 3 stars and price of $3, there is a $e^R/(e^R+1) = 0.12$ probability that the product is a nose pack.

Remember, we want to solve prediction: $\text{logit}(p) = \log(p/[1-p]) = \alpha+\beta_1 x_1 + \beta_2 x_2$

If $x_1=3$ and $x_2=3$, RHS is $1.5 - 0.9*3 - 0.3*3 = -2.1$

$p = e^R/(e^R+1) = e^{-2.1}/(e^{-2.1}+1) = 0.12$

Interpret: Oddities

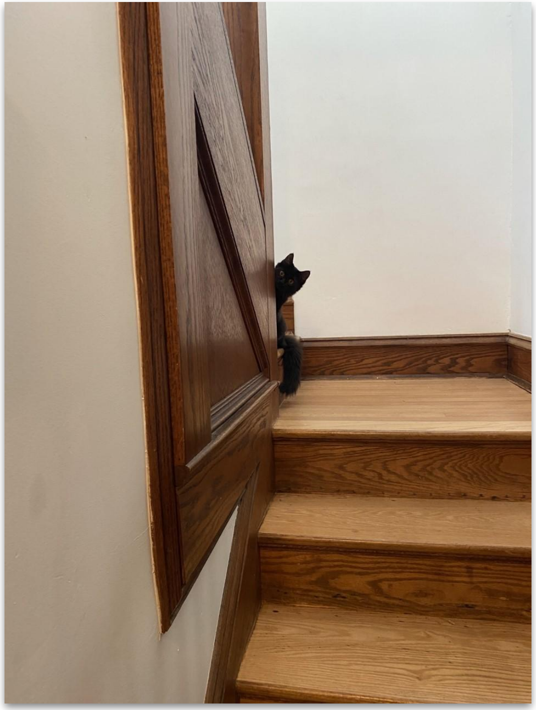
- y = is product a nose pack?
- x_1 = avg customer rating; x_2 = price in \$
- $y \sim \text{sigmoid}(1.5 - 0.9x_1 - 0.3x_2)$
-

Interpret: Oddities

- y = is product a nose pack?
- x_1 = avg customer rating; x_2 = price in $
- $y \sim \text{sigmoid}(1.5 - 0.9x_1 - 0.3x_2)$
- It doesn't make sense to have an x_1 value that's not between 1-5, and it doesn't make sense to have a negative x_2

1 minute break

The midterm is right around the corner

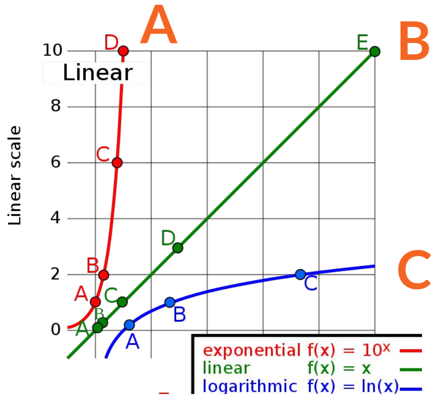


Regression type review

Model	Interpretation
Linear $y = \alpha + \beta x$	1 unit change in x is associated with a β unit change in y
Linear-log $y = \alpha + \beta \ln(x)$	If x is multiplied by *e*, we expect a β unit change in y 1% change in x is associated with a 0.01*β unit change in y
Log-linear $\ln(y) = \alpha + \beta x$	For a 1 unit change in x, we expect y to be multiplied by e^{β} 1 unit change in x is associated with a 100*(exp(β)-1)% change in y
Log-log $\ln(y) = \alpha + \beta \ln(x)$	If x is multiplied by *e*, we expect y to be multiplied by e^{β} 1% change in x is associated with a β% change in y (*elasticity*)

Regression type review

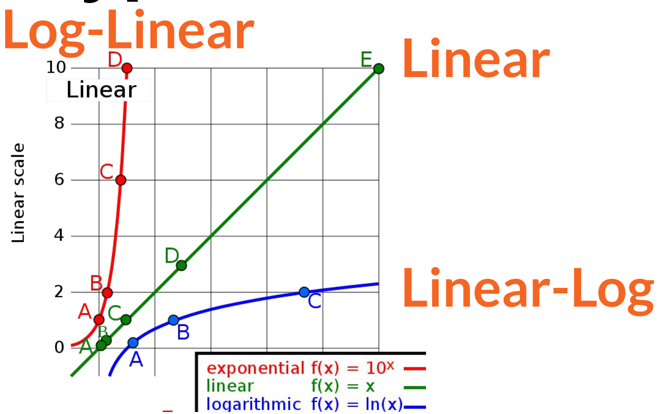
Model
Linear $y = \alpha + \beta x$
Linear-log $y = \alpha + \beta \ln(x)$
Log-linear $\ln(y) = \alpha + \beta x$
Log-log $\ln(y) = \alpha + \beta \ln(x)$



https://en.wikipedia.org/wiki/Logarithmic_scale#/media/File:Logarithmic_Scales-mkII.svg

Regression type review

Model
Linear $y = \alpha + \beta x$
Linear-log $y = \alpha + \beta \ln(x)$
Log-linear $\ln(y) = \alpha + \beta x$
Log-log $\ln(y) = \alpha + \beta \ln(x)$

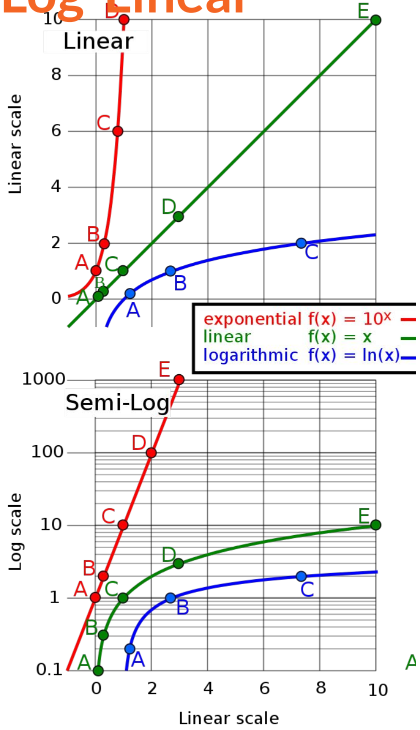


https://en.wikipedia.org/wiki/Logarithmic_scale#/media/File:Logarithmic_Scales-mkII.svg

Regression type review

Model
Linear $y = \alpha + \beta x$
Linear-log $y = \alpha + \beta \ln(x)$
Log-linear $\ln(y) = \alpha + \beta x$
Log-log $\ln(y) = \alpha + \beta \ln(x)$

Log-Linear

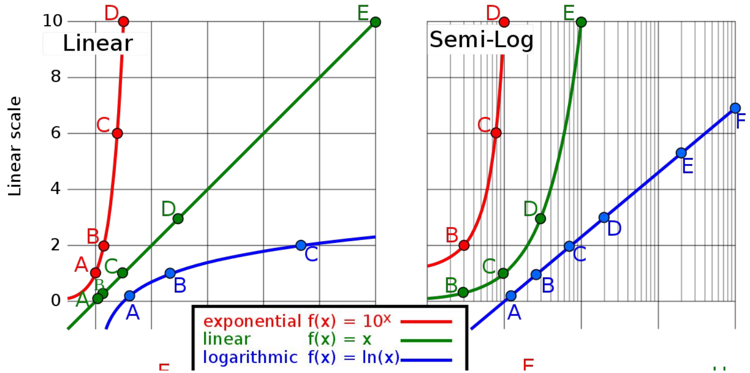


Log-Linear line looks straight if we plot the y-axis on a log scale!

https://en.wikipedia.org/wiki/Logarithmic_scale#/media/File:Logarithmic_Scales-mkII.svg

Regression type review

Model
Linear $y = \alpha + \beta x$
Linear-log $y = \alpha + \beta \ln(x)$
Log-linear $\ln(y) = \alpha + \beta x$
Log-log $\ln(y) = \alpha + \beta \ln(x)$



Linear-Log

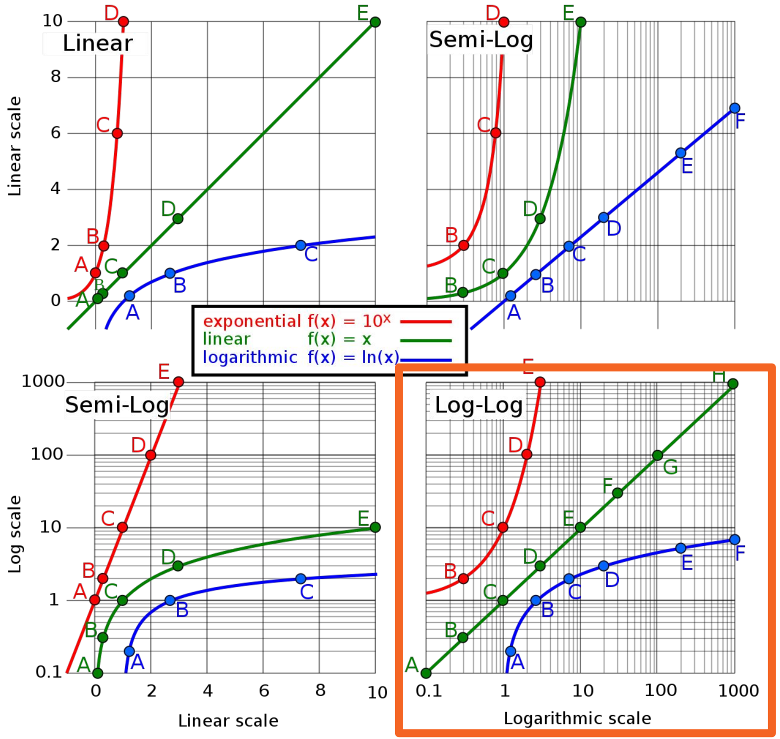
Linear-Log line looks straight if we plot the x-axis on a log scale!

https://en.wikipedia.org/wiki/Logarithmic_scale#/media/File:Logarithmic_Scales-mkII.svg

Regression type review

Model
Linear $y = \alpha + \beta x$
Linear-log $y = \alpha + \beta \ln(x)$
Log-linear $\ln(y) = \alpha + \beta x$
Log-log $\ln(y) = \alpha + \beta \ln(x)$

When to use log-log?

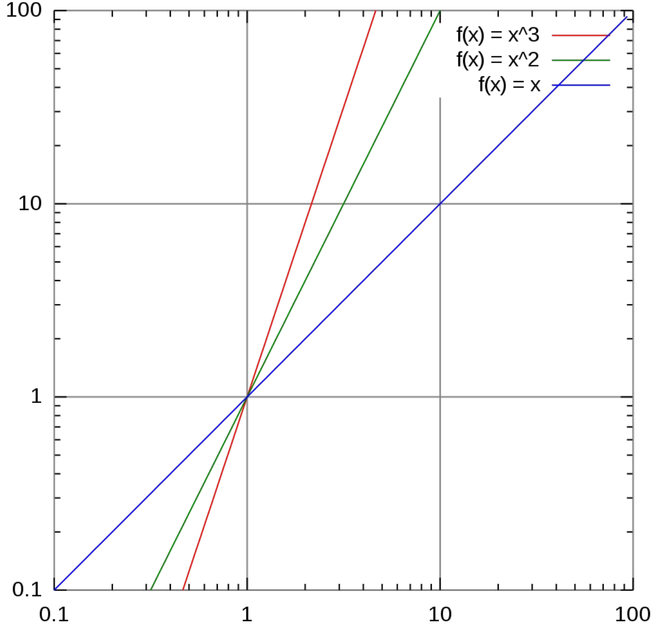


https://en.wikipedia.org/wiki/Logarithmic_scale#/media/File:Logarithmic_Scales-mkII.svg

Regression type review

Model
Linear y = α + βx
Linear-log y = α + β ln(x)
Log-linear ln(y) = α + βx
Log-log ln(y) = α + β ln(x)

When you want to 'smoosh' both axes, e.g. often with powers



https://en.wikipedia.org/wiki/Log%E2%80%93log_plot#/media/File:LogLog_exponentials.svg

Regression type review

Model	Interpretation
Linear $y = \alpha + \beta x$	1 unit change in x is associated with a β unit change in y
Linear-log $y = \alpha + \beta \ln(x)$	If x is multiplied by *e*, we expect a β unit change in y 1% change in x is associated with a 0.01*β unit change in y
Log-linear $\ln(y) = \alpha + \beta x$	For a 1 unit change in x, we expect y to be multiplied by e^{β} 1 unit change in x is associated with a 100*(exp(β)-1)% change in y
Log-log $\ln(y) = \alpha + \beta \ln(x)$	If x is multiplied by *e*, we expect y to be multiplied by e^{β} 1% change in x is associated with a β% change in y (*elasticity*)

Deriving interpretations

Let's do log-log

$\boxed{\textbf{\underline{Step 1, Write the model}}\text{: } \log(y) = a+b*\log(x)}$

Step 2, Define new variable for x: $x_{new} = x*e$

Step 3, Define new variable for y: $\log(y_{new}) = a+b*\log(x_{new})$

Step 4, Plug in Step 2 to Step 3: $\log(y_{new}) = a+b*\log(x*e)$

Step 5, Use Step 1 to Rewrite Step 4's RHS in terms of y:

$\log(y_{new}) = a+b*[\log(x)+\log(e)]$ because $\log(a*b)=\log(a)+\log(b)$

$\log(y_{new}) = a+b*\log(x)+b*\log(e)= a+b*\log(x)+b*1= a+b*\log(x)+b$

$\log(y_{new}) = \log(y)+b$

Step 6, Calculate the difference between y_{new} and y:

$y_{new} = e^{\log(y)+b}= e^{\log(y)} *e^b = y *e^b$

$y_{new} / y = e^b$

Deriving interpretations

For linear-log and log-log (any time we want to smoosh the x), the interpretation will involve multiplying by e (≈2.72)

Step 1, Write the model: $\log(y) = a+b*\log(x)$

Step 2, Define new variable for x: $x_{new} = x*e$

Step 3, Define new variable for y: $\log(y_{new}) = a+b*\log(x_{new})$

Step 4, Plug in Step 2 to Step 3: $\log(y_{new}) = a+b*\log(x*e)$

Step 5, Use Step 1 to Rewrite Step 4's RHS in terms of y:

$\log(y_{new}) = a+b*[\log(x)+\log(e)]$ because $\log(a*b)=\log(a)+\log(b)$

$\log(y_{new}) = a+b*\log(x)+b*\log(e)= a+b*\log(x)+b*1= a+b*\log(x)+b$

$\log(y_{new}) = \log(y)+b$

Step 6, Calculate the difference between y_{new} and y:

$y_{new} = e^{\log(y)+b}= e^{\log(y)} *e^b = y *e^b$

$y_{new} / y = e^b$

Deriving interpretations

If we didn't want to smoosh x, this would instead be x+1 (a 1 unit change)

Step 1, Write the model: $\log(y) = a+b*\log(x)$

Step 2, Define new variable for x: $x_{new} = x*e$

Step 3, Define new variable for y: $\log(y_{new}) = a+b*\log(x_{new})$

Step 4, Plug in Step 2 to Step 3: $\log(y_{new}) = a+b*\log(x*e)$

Step 5, Use Step 1 to Rewrite Step 4's RHS in terms of y:

$\log(y_{new}) = a+b*[\log(x)+\log(e)]$ because $\log(a*b)=\log(a)+\log(b)$

$\log(y_{new}) = a+b*\log(x)+b*\log(e)= a+b*\log(x)+b*1= a+b*\log(x)+b$

$\log(y_{new}) = \log(y)+b$

Step 6, Calculate the difference between y_{new} and y:

$y_{new} = e^{\log(y)+b}= e^{\log(y)} *e^b = y *e^b$

$y_{new} / y = e^b$

Deriving interpretations

Rewrite Step 1 using both new x and new y

<u>Step 1, Write the model</u>: $\log(y) = a+b*\log(x)$

<u>Step 2, Define new variable for x</u>: $x_{new} = x*e$

<u>Step 3, Define new variable for y</u>: $\log(y_{new}) = a+b*\log(x_{new})$

<u>Step 4, Plug in Step 2 to Step 3</u>: $\log(y_{new}) = a+b*\log(x*e)$

<u>Step 5, Use Step 1 to Rewrite Step 4's RHS in terms of y</u>:

$\log(y_{new}) = a+b*[\log(x)+\log(e)]$ because $\log(a*b)=\log(a)+\log(b)$

$\log(y_{new}) = a+b*\log(x)+b*\log(e)= a+b*\log(x)+b*1= a+b*\log(x)+b$

$\log(y_{new}) = \log(y)+b$

<u>Step 6, Calculate the difference between y_{new} and y</u>:

$y_{new} = e^{\log(y)+b}= e^{\log(y)} *e^b = y *e^b$

$y_{new} / y = e^b$

Deriving interpretations

Get rid of x_{new}

<u>**Step 1, Write the model**</u>: $\log(y) = a+b*\log(x)$

<u>**Step 2, Define new variable for x**</u>: $x_{new} = x*e$

<u>**Step 3, Define new variable for y**</u>: $\log(y_{new}) = a+b*\log(x_{new})$

<u>**Step 4, Plug in Step 2 to Step 3**</u>: $\boxed{\log(y_{new}) = a+b*\log(x*e)}$

<u>**Step 5, Use Step 1 to Rewrite Step 4's RHS in terms of y**</u>:

$\log(y_{new}) = a+b*[\log(x)+\log(e)]$ because $\log(a*b)=\log(a)+\log(b)$

$\log(y_{new}) = a+b*\log(x)+b*\log(e)= a+b*\log(x)+b*1= a+b*\log(x)+b$

$\log(y_{new}) = \log(y)+b$

<u>**Step 6, Calculate the difference between y_{new} and y**</u>:

$y_{new} = e^{\log(y)+b}= e^{\log(y)} *e^b = y *e^b$

$y_{new} / y = e^b$

Deriving interpretations

Get rid of x (often requires log / exp rules)

<u>Step 1, Write the model</u>: log(y) = a+b*log(x)
<u>Step 2, Define new variable for x</u>: x_{new} = x*e
<u>Step 3, Define new variable for y</u>: log(y_{new}) = a+b*log(x_{new})
<u>Step 4, Plug in Step 2 to Step 3</u>: log(y_{new}) = a+b*log(x*e)
<u>Step 5, Use Step 1 to Rewrite Step 4's RHS in terms of y</u>:

log(y_{new}) = a+b*[log(x)+log(e)] because log(a*b)=log(a)+log(b)
log(y_{new}) = a+b*log(x)+b*log(e)= a+b*log(x)+b*1= a+b*log(x)+b
log(y_{new}) = log(y)+b

<u>Step 6, Calculate the difference between y_{new} and y</u>:

$y_{new} = e^{\log(y)+b} = e^{\log(y)} * e^b = y * e^b$
$y_{new} / y = e^b$

Deriving interpretations

<u>Step 1, Write the model</u>: $\log(y) = a+b*\log(x)$

<u>Step 2, Define new variable for x</u>: $x_{new} = x*e$

<u>Step 3, Define new variable for y</u>: $\log(y_{new}) = a+b*\log(x_{new})$

<u>Step 4, Plug in Step 2 to Step 3</u>: $\log(y_{new}) = a+b*\log(x*e)$

<u>Step 5, Use Step 1 to Rewrite Step 4's RHS in terms of y</u>:

$\log(y_{new}) = a+b*[\log(x)+\log(e)]$ because $\log(a*b)=\log(a)+\log(b)$

$\log(y_{new}) = a+b*\log(x)+b*\log(e)= a+b*\log(x)+b*1= a+b*\log(x)+b$

$\log(y_{new}) = \log(y)+b$

<u>Step 6, Calculate the difference between y_{new} and y</u>:

$y_{new} = e^{\log(y)+b}= e^{\log(y)} *e^b = y *e^b$

$y_{new} / y = e^b$

Express y_{new} in terms of y. (If we're squishing y, it's multiplicative. If not, it's a unit change)

Regression type review

Squishing x with a log → interpret by multiplying x by e (or think about % change in x)

Model	Interpretation
Linear $y = \alpha + \beta x$	1 unit change in x is associated with a β unit change in y
Linear-log $y = \alpha + \beta \boxed{\ln(x)}$	If x is multiplied by *e* we expect a β unit change in y 1% change in x is associated with a 0.01*β unit change in y
Log-linear $\ln(y) = \alpha + \beta x$	For a 1 unit change in x, we expect y to be multiplied by e^{β} 1 unit change in x is associated with a 100*(exp(β)-1)% change in y
Log-log $\ln(y) = \alpha + \beta \boxed{\ln(x)}$	If x is multiplied by *e* we expect y to be multiplied by e^{β} 1% change in x is associated with a β% change in y (*elasticity*)

Regression type review

Squishing y with a log → interpret by multiplying y by e^{β} (or think about % change in y)

Model	Interpretation
Linear $y = \alpha + \beta x$	1 unit change in x is associated with a β unit change in y
Linear-log $y = \alpha + \beta \ln(x)$	If x is multiplied by *e*, we expect a β unit change in y 1% change in x is associated with a 0.01*β unit change in y
Log-linear $\boxed{\ln(y)} = \alpha + \beta x$	For a 1 unit change in x, we expect $\boxed{\text{y to be multiplied by } e^{\beta}}$ 1 unit change in x is associated with a 100*(exp(β)-1)% change in y
Log-log $\boxed{\ln(y)} = \alpha + \beta \ln(x)$	If x is multiplied by *e*, we expect $\boxed{\text{y to be multiplied by } e^{\beta}}$ 1% change in x is associated with a β% change in y (*elasticity*)

Back to multivariable regression

i	x	y
1	78	18
2	83	14
…	…	…

i	x_1	x_2	x_3	y
1	78	0	30.5	18
2	83	1	28.0	14
…	…	…	…	…

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

$$\Downarrow$$

$$y_i = \alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \ldots + \varepsilon_i$$

Interpreting *more* dummies

- y = temperature **(numerical)**
- x_1 = air pressure **(numerical)**
- x_2 = season [spring, summer, fall, winter] **(categorical)**

Interpreting dummies

Temp (F)	Pressure	Season
80	81	Summer
50	63	Fall
70	75	Spring
…	…	…

Regression needs numbers, not words!

Interpreting dummies

Temp (F)	Pressure	Season	Season_num
80	81	Summer	2
50	63	Fall	3
70	75	Spring	1
…	…	…	

If we map categories to numeric values, [Spring→Summer] = [Summer→Fall]... Problematic?

Interpreting dummies

Temp (F)	Pressure	Season	Season_num
80	81	Summer	2
50	63	Fall	3
70	75	Spring	1
…	…	…	

Problem: Spring and Winter far away (1 vs 4) but they are likely more similar attribute-wise than Spring and Fall (1 vs 3)

Interpreting dummies

Temp (F)	Pressure	Season	Spring	Summer	Fall	Winter
80	81	Summer	0	1	0	0
50	63	Fall	0	0	1	0
70	75	Spring	1	0	0	0
…	…	…	…	…	…	…

get_dummies (without drop_first=True) gives us a new column for each unique Season value

Why do we exclude Spring from regression?

y	x_1			x_2	x_3	x_4
Temp (F)	**Pressure**	**Season**	**Spring**	**Summer**	**Fall**	**Winter**
80	81	Summer	0	1	0	0
50	63	Fall	0	0	1	0
70	75	Spring	1	0	0	0
…	…	…	…	…	…	…

Why do we exclude “Spring”?

Why do we exclude Spring from regression?

y	x_1			x_2	x_3	x_4
Temp (F)	**Pressure**	**Season**	**Spring**	**Summer**	**Fall**	**Winter**
80	81			1	0	0
50	63			0	1	0
70	75			0	0	0
…	…			…	…	…

Why do we exclude "Spring"? We can derive that info from x_2, x_3, and x_4!

Why do we exclude Spring from regression?

y	x_1			x_2	x_3	x_4
Temp (F)	**Pressure**	**Season**	**Spring**	**Summer**	**Fall**	**Winter**
80	81			1	0	0
50	63			0	1	0
70	75			0	0	0
…	…			…	…	…

Why do we exclude "Spring"? We can derive that info from x_2, x_3, and x_4!

Why do we exclude Spring from regression?

y	x_1			x_2	x_3	x_4
Temp (F)	**Pressure**	**Season**	**Spring**	**Summer**	**Fall**	**Winter**
80	81			1	0	0
50	63			0	1	0
70	75			0	0	0
…	…			…	…	…

Why do we exclude "Spring"? If all the other dummies x_2, x_3, and x_4 are 0, then the season must be Spring.

Why do we exclude Spring from regression?

y	x_1			x_2	x_3	x_4
Temp (F)	**Pressure**	**Season**	**Spring**	**Summer**	**Fall**	**Winter**
80	81	Summer	0	1	0	0
50	63	Fall	0	0	1	0
70	75	Spring	1	0	0	0
…	…	…	…	…	…	…

Spring = 1 - (Summer + Fall + Winter)

Why do we exclude Spring from regression?

y	x_1			x_2	x_3	x_4
Temp (F)	**Pressure**	**Season**	**Spring**	**Summer**	**Fall**	**Winter**
80	81	Summer	0	1	0	0
50	63	Fall	0	0	1	0
70	75	Spring	1	0	0	0
…	…	…	…	…	…	…

Spring (a.k.a. x_2, x_3, and x_4 are 0) is our reference level.

Interpreting dummy variables

- Regression: $y \sim x_1 + x_2 + x_3 + x_4$
- y = temperature, x_1 = air pressure
- x_2, x_3, x_4 = Summer, Fall, Winter

- **How to understand difference between coefficients? Plug numbers in!**

Dummy variables

- Regression: $y \sim x_1 + x_2 + x_3 + x_4$
- y = temperature, x_1 = air pressure
- x_2, x_3, x_4 = Summer, Fall, Winter

Season	x_2	x_3	x_4	Equation	Simplified
Spring	0	0	0	$y^{hat} = \alpha + \beta_1 x_1 + \beta_2 {*0} + \beta_3 {*0} + \beta_4 {*0}$	
Summer	1	0	0	$y^{hat} = \alpha + \beta_1 x_1 + \beta_2 {*1} + \beta_3 {*0} + \beta_4 {*0}$	
Fall	0	1	0	$y^{hat} = \alpha + \beta_1 x_1 + \beta_2 {*0} + \beta_3 {*1} + \beta_4 {*0}$	
Winter	0	0	1	$y^{hat} = \alpha + \beta_1 x_1 + \beta_2 {*0} + \beta_3 {*0} + \beta_4 {*1}$	

Interpreting dummy variables

- Regression: $y \sim x_1 + x_2 + x_3 + x_4$
- y = temperature, x_1 = air pressure
- x_2, x_3, x_4 = Summer, Fall, Winter

Season	x_2	x_3	x_4	Equation	Simplified
Spring	0	0	0	$y^{hat} = \alpha + \beta_1 x_1 + \beta_2{*}0 + \beta_3{*}0 + \beta_4{*}0$	$y^{hat} = \alpha + \beta_1 x_1$
Summer	1	0	0	$y^{hat} = \alpha + \beta_1 x_1 + \beta_2{*}1 + \beta_3{*}0 + \beta_4{*}0$	$y^{hat} = \alpha + \beta_1 x_1 + \beta_2$
Fall	0	1	0	$y^{hat} = \alpha + \beta_1 x_1 + \beta_2{*}0 + \beta_3{*}1 + \beta_4{*}0$	$y^{hat} = \alpha + \beta_1 x_1 + \beta_3$
Winter	0	0	1	$y^{hat} = \alpha + \beta_1 x_1 + \beta_2{*}0 + \beta_3{*}0 + \beta_4{*}1$	$y^{hat} = \alpha + \beta_1 x_1 + \beta_4$

Interpreting dummy variables

- What is the difference between the two highlighted simplified equations?

Season	x_2	x_3	x_4	Equation	Simplified
Spring	0	0	0	$y^{hat} = \alpha + \beta_1 x_1 + \beta_2 {*0} + \beta_3 {*0} + \beta_4 {*0}$	$y^{hat} = \alpha + \beta_1 x_1$
Summer	1	0	0	$y^{hat} = \alpha + \beta_1 x_1 + \beta_2 {*1} + \beta_3 {*0} + \beta_4 {*0}$	$y^{hat} = \alpha + \beta_1 x_1 + \beta_2$
Fall	0	1	0	$y^{hat} = \alpha + \beta_1 x_1 + \beta_2 {*0} + \beta_3 {*1} + \beta_4 {*0}$	$y^{hat} = \alpha + \beta_1 x_1 + \beta_3$
Winter	0	0	1	$y^{hat} = \alpha + \beta_1 x_1 + \beta_2 {*0} + \beta_3 {*0} + \beta_4 {*1}$	$y^{hat} = \alpha + \beta_1 x_1 + \beta_4$

Interpreting dummy variables

β_2 represents the **difference** in output y^{hat} between summer and spring

Season	x_2	x_3	x_4	Equation	Simplified
Spring	0	0	0	$y^{hat} = \alpha + \beta_1 x_1 + \beta_2{*}0 + \beta_3{*}0 + \beta_4{*}0$	$y^{hat} = \alpha + \beta_1 x_1$
Summer	1	0	0	$y^{hat} = \alpha + \beta_1 x_1 + \beta_2{*}1 + \beta_3{*}0 + \beta_4{*}0$	$y^{hat} = \alpha + \beta_1 x_1 + \beta_2$
Fall	0	1	0	$y^{hat} = \alpha + \beta_1 x_1 + \beta_2{*}0 + \beta_3{*}1 + \beta_4{*}0$	$y^{hat} = \alpha + \beta_1 x_1 + \beta_3$
Winter	0	0	1	$y^{hat} = \alpha + \beta_1 x_1 + \beta_2{*}0 + \beta_3{*}0 + \beta_4{*}1$	$y^{hat} = \alpha + \beta_1 x_1 + \beta_4$

Interpreting dummy variables

- What is the difference between the two highlighted simplified equations?

Season	x_2	x_3	x_4	Equation	Simplified
Spring	0	0	0	$y^{hat} = \alpha + \beta_1 x_1 + \beta_2{*}0 + \beta_3{*}0 + \beta_4{*}0$	$\mathbf{y^{hat} = \alpha + \beta_1 x_1}$
Summer	1	0	0	$y^{hat} = \alpha + \beta_1 x_1 + \beta_2{*}1 + \beta_3{*}0 + \beta_4{*}0$	$y^{hat} = \alpha + \beta_1 x_1 + \beta_2$
Fall	0	1	0	$y^{hat} = \alpha + \beta_1 x_1 + \beta_2{*}0 + \beta_3{*}1 + \beta_4{*}0$	$\mathbf{y^{hat} = \alpha + \beta_1 x_1 + \beta_3}$
Winter	0	0	1	$y^{hat} = \alpha + \beta_1 x_1 + \beta_2{*}0 + \beta_3{*}0 + \beta_4{*}1$	$y^{hat} = \alpha + \beta_1 x_1 + \beta_4$

Interpreting dummy variables

β_3 represents the **difference** in output y^{hat} between fall and spring

Season	x_2	x_3	x_4	Equation	Simplified
Spring	0	0	0	$y^{hat} = \alpha + \beta_1 x_1 + \beta_2 {*0} + \beta_3 {*0} + \beta_4 {*0}$	$y^{hat} = \alpha + \beta_1 x_1$
Summer	1	0	0	$y^{hat} = \alpha + \beta_1 x_1 + \beta_2 {*1} + \beta_3 {*0} + \beta_4 {*0}$	$y^{hat} = \alpha + \beta_1 x_1 + \beta_2$
Fall	0	1	0	$y^{hat} = \alpha + \beta_1 x_1 + \beta_2 {*0} + \beta_3 {*1} + \beta_4 {*0}$	$y^{hat} = \alpha + \beta_1 x_1 + \beta_3$
Winter	0	0	1	$y^{hat} = \alpha + \beta_1 x_1 + \beta_2 {*0} + \beta_3 {*0} + \beta_4 {*1}$	$y^{hat} = \alpha + \beta_1 x_1 + \beta_4$

Interpreting dummy variables

- What is the difference between the two highlighted simplified equations?

Season	x_2	x_3	x_4	Equation	Simplified
Spring	0	0	0	$y^{hat} = \alpha + \beta_1 x_1 + \beta_2{*}0 + \beta_3{*}0 + \beta_4{*}0$	$y^{hat} = \alpha + \beta_1 x_1$
Summer	1	0	0	$y^{hat} = \alpha + \beta_1 x_1 + \beta_2{*}1 + \beta_3{*}0 + \beta_4{*}0$	$y^{hat} = \alpha + \beta_1 x_1 + \beta_2$
Fall	0	1	0	$y^{hat} = \alpha + \beta_1 x_1 + \beta_2{*}0 + \beta_3{*}1 + \beta_4{*}0$	$y^{hat} = \alpha + \beta_1 x_1 + \beta_3$
Winter	0	0	1	$y^{hat} = \alpha + \beta_1 x_1 + \beta_2{*}0 + \beta_3{*}0 + \beta_4{*}1$	$y^{hat} = \alpha + \beta_1 x_1 + \beta_4$

Interpreting dummy variables

- Regression: $y \sim x_1 + x_2 + x_3 + x_4$
- y = temperature, x_1 = air pressure
- x_2, x_3, x_4 = Summer, Fall, Winter

- β_2 represents the **difference** in output y^{hat} between summer and spring
- β_3 represents the **difference** in output y^{hat} between fall and spring
- β_4 represents the **difference** in output y^{hat} between winter and spring

Interpreting dummy variables

- Regression: $y \sim x_1 + x_2 + x_3 + x_4$
- $\mathbf{y}$ = temperature, $\mathbf{x_1}$ = air pressure
- x_2, x_3, x_4 = Summer, Fall, Winter

All interpretations are relative to spring!

By omitting spring from the regression, we've made it our "reference"

- $\boldsymbol{\beta_2}$ represents the **difference** in output y^{hat} between summer and spring
- $\boldsymbol{\beta_3}$ represents the **difference** in output y^{hat} between fall and spring
- $\boldsymbol{\beta_4}$ represents the **difference** in output y^{hat} between winter and spring

What is a reference variable?

The thing you interpret a β coefficient relative to.

INFO2950_Lec7_20230913

File Edit View Insert Format

Regression interpretations: summarize relationship

x = millimeters of rainfall

y = umbrellas sold

y = -19 + 0.45x

Summarize relationship between variables:

Our model shows a positive correlation between rain and sales of umbrellas; specifically, each additional mm of rain corresponds to an extra 0.45 umbrellas we expect to be sold.

x = {0 if no rain, 1 if any rain}

y = umbrellas sold

y = 0.0 + 8x

Summarize relationship between variables:

What is the reference for a single binary variable?

1 - (variable)

INFO2950_Lec7_20230913

File Edit View Insert Format

Regression interpretations: summarize relationship

x = millimeters of rainfall
y = umbrellas sold
y = -19 + 0.45x

Summarize relationship between variables:

Our model shows a positive correlation between rain and sales of umbrellas; specifically, each additional mm of rain corresponds to an extra 0.45 umbrellas we expect to be sold.

x = {0 if no rain, 1 if any rain}
y = umbrellas sold
y = 0.0 + 8x

Summarize relationship between variables:

____

____

____

Multivariable interpretation

- Regression: $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$
- y = temperature, x_1 = air pressure
- x_2, x_3, x_4 = Summer, Fall, Winter

- How do we summarize the interpretation of β_2 in English?

Multivariable interpretation

- Regression: $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$
- y = temperature, x_1 = air pressure
- x_2, x_3, x_4 = Summer, Fall, Winter

- How do we summarize the interpretation of β_2 in English?

When it is summer, all else equal, we expect the temperature to increase by β_2 degrees relative to when it is spring.

Multivariable interpretation

- Regression: $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$
- y = temperature, x_1 = air pressure
- x_2, x_3, x_4 = Summer, Fall, Winter

- How do we summarize the interpretation of β_1 in English?

Multivariable interpretation

- Regression: $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$
- y = temperature, x_1 = air pressure
- x_2, x_3, x_4 = Summer, Fall, Winter

- How do we summarize the interpretation of β_1 in English?

For a one unit increase in air pressure, all else equal, we expect the temperature to increase by β_1 degrees.

Multivariable interpretation

- Regression: $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$
- y = temperature, x_1 = air pressure
- x_2, x_3, x_4 = Summer, Fall, Winter

- How do we summarize the interpretation of β_4 in English?

Regression interpretation

- Regression: $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$
- y = temperature, x_1 = air pressure
- x_2, x_3, x_4 = Summer, Fall, Winter

- How do we summarize the interpretation of β_4 in English?

When it is winter, all else equal, we expect the temperature to increase by β_4 degrees relative to when it is spring.

How to decide on multivariable regression inputs

- Start with a hypothesis of what “covariates” (inputs) are relevant

How to decide on multivariable regression inputs

- Start with a hypothesis of what “covariates” (inputs) are relevant
- Exclude relevant covariates that are very collinear with other covariates in your model already (can check correlations)

What happens if we include “collinear” inputs?

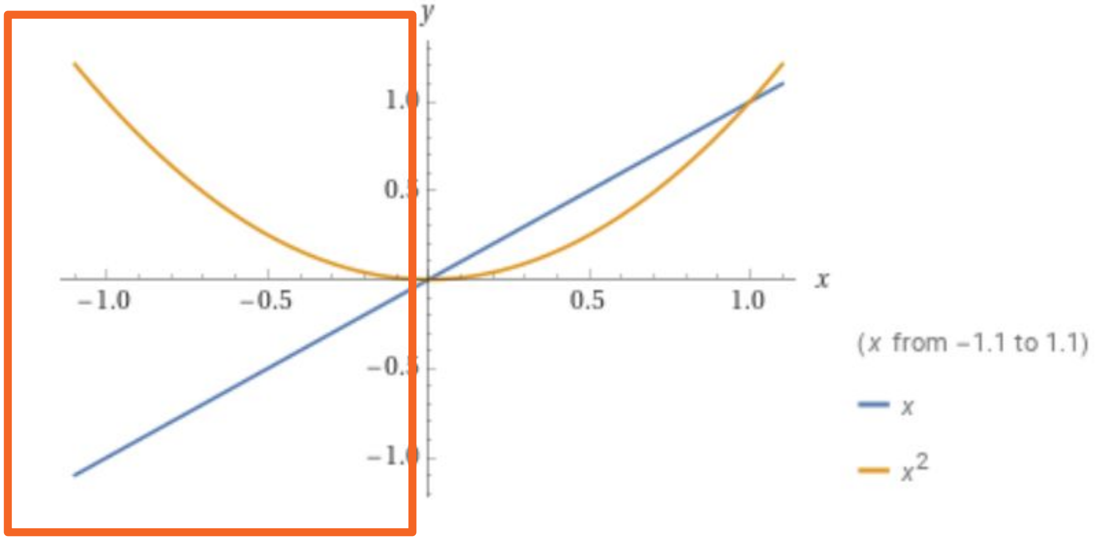
- **Collinearity =** correlation between inputs
- **Are x_1 and x_2 correlated? Yes (corr = -1)**
  - **x_1 =** binary: use oil cleanser daily
  - **x_2 =** binary: does not use oil cleanser daily

Multicollinearity

- $x_2 = x_1^2$
- Are x_1 and x_2 collinear?

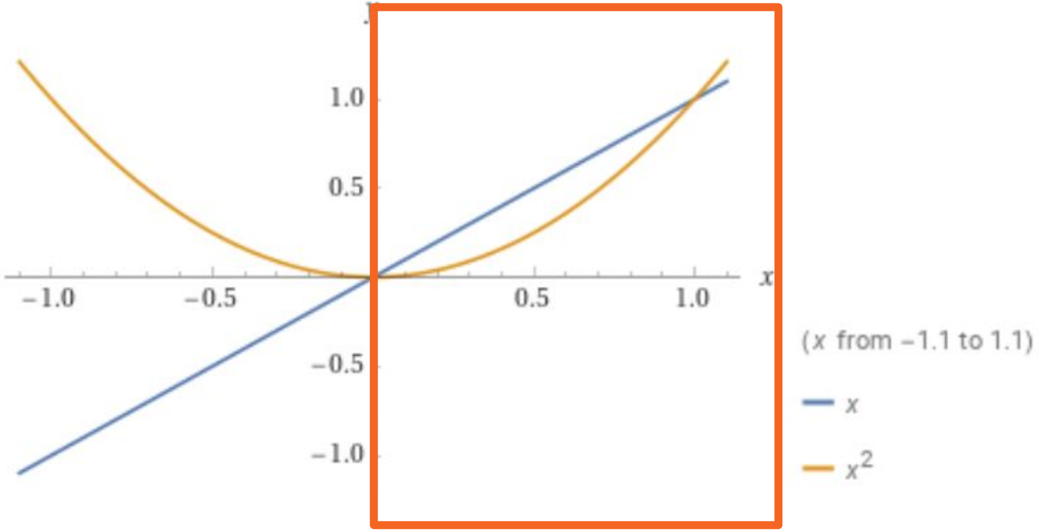
No, x and x² are not collinear!

Negative correlation



No, x and x² are not collinear!

Positive correlation



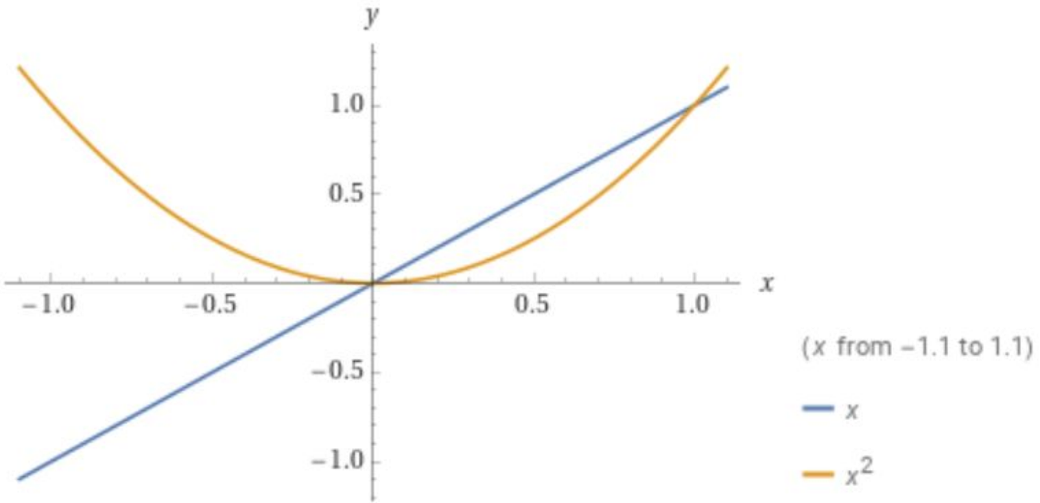
No, x and x^2 are not collinear!

```
x1  = np.random.normal(0,1,size=(1,1000))
x2 = x1*x1
np.corrcoef(x1,x2)
```

✓ 0.3s

```
array([[1.        , 0.02637914],
       [0.02637914, 1.        ]])
```

Simulation on 1000 random x's drawn from normal distribution shows us a tiny correlation of 0.03



How to decide on multivariable regression inputs

- Start with a hypothesis of what “covariates” (inputs) are relevant
- Exclude relevant covariates that are very collinear with other covariates in your model already (can check correlations)

https://twitter.com/ChristophMolnar/status/1423528092178518023

add noise and not new info to your model?

https://twitter.com/ChristophMolnar/status/1423528092178518023

How to decide on multivariable regression inputs

- Start with a hypothesis of what “covariates” (inputs) are relevant
- Exclude relevant covariates that are very collinear with other covariates in your model already (can check correlations)
- **Check your residual plots, just like before! (Look for randomness, use transformations as needed)**

What variables to include/exclude?

- Want to include any input (x) that might have some effect on output (y)
- Don't want to include input (x) that's noise

What variables to include/exclude?

- Want to include any input (x) that might have some effect on output (y)
- Don’t want to include input (x) that’s noise
- You can get both “**included** variable bias” and “**excluded** variable bias” – there’s no one-size-fits-all rule, but some guidelines…

Takeaways on multivariable regression inputs

- Choose covariates that…
  - **make sense**
  - **aren’t redundant** (i.e., aren’t collinear and don’t overfit the data)
  - allow you (with transformation) to get **random-looking residual plots**

INFO2950_Lec8_20230918

File Edit View Insert Format

Why might a residual plot not be random?

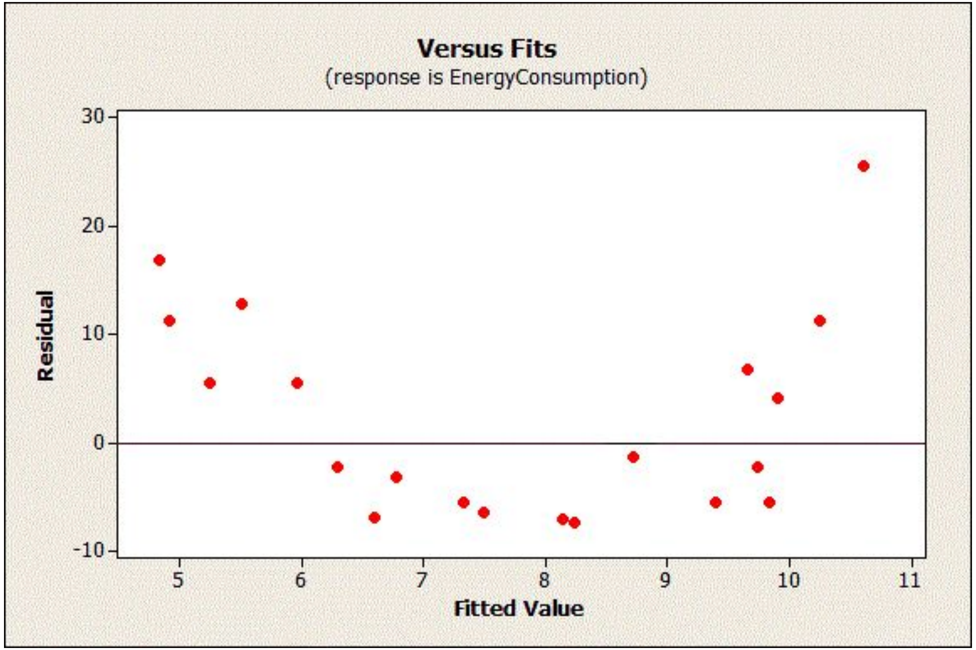
- The deterministic part of the model ($\alpha + \beta x$) is not capturing all of the information hidden in your data, which is leaking into your residual
  - Weird curvature – missing transformations
  - (for multivariable regressions) missing variables

Residual plots for multivar reg

Same method as for single variable regressions!

X-axis: prediction $\hat{y}_i$

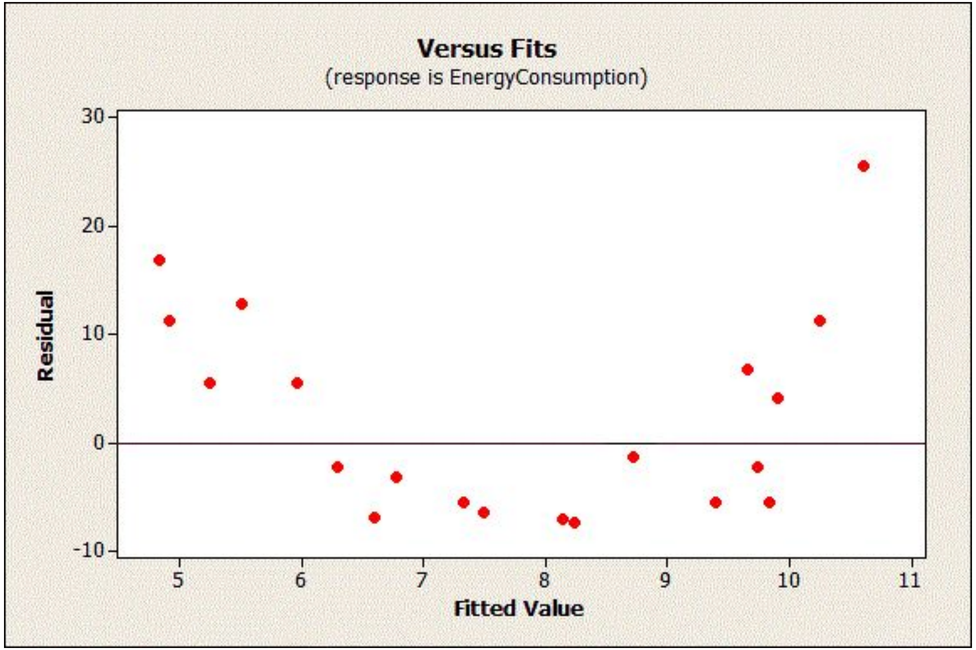
Y-axis: residual ε_i



https://statisticsbyjim.com/regression/check-residual-plots-regression-analysis/

Residual plots for multivar reg

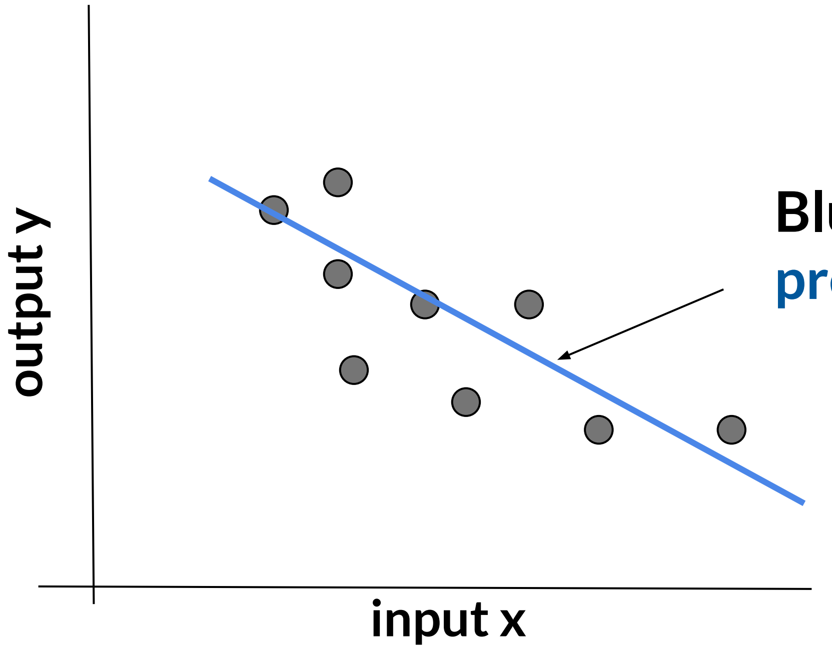
Does this residual plot imply issues with our model?



https://statisticsbyjim.com/regression/check-residual-plots-regression-analysis/

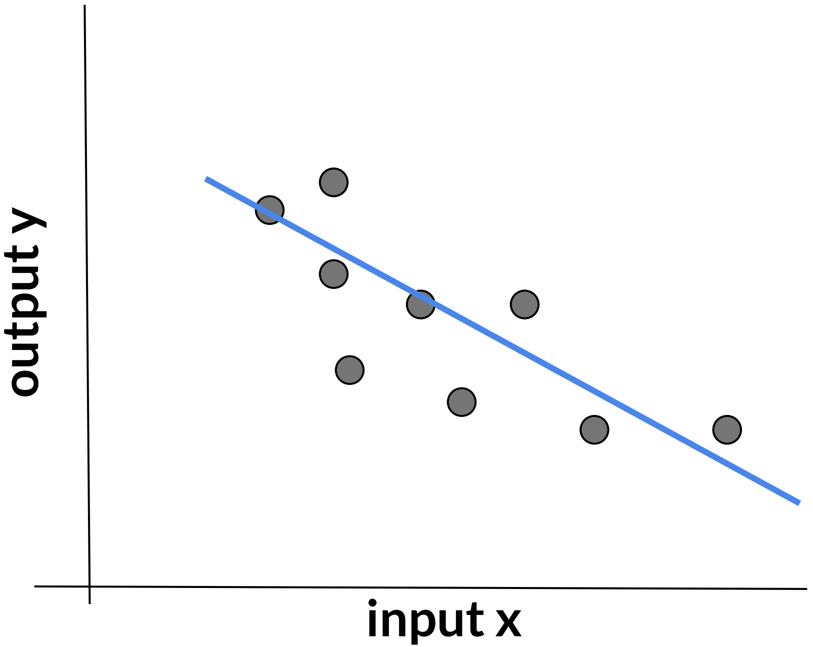
Predictions vs. Reality

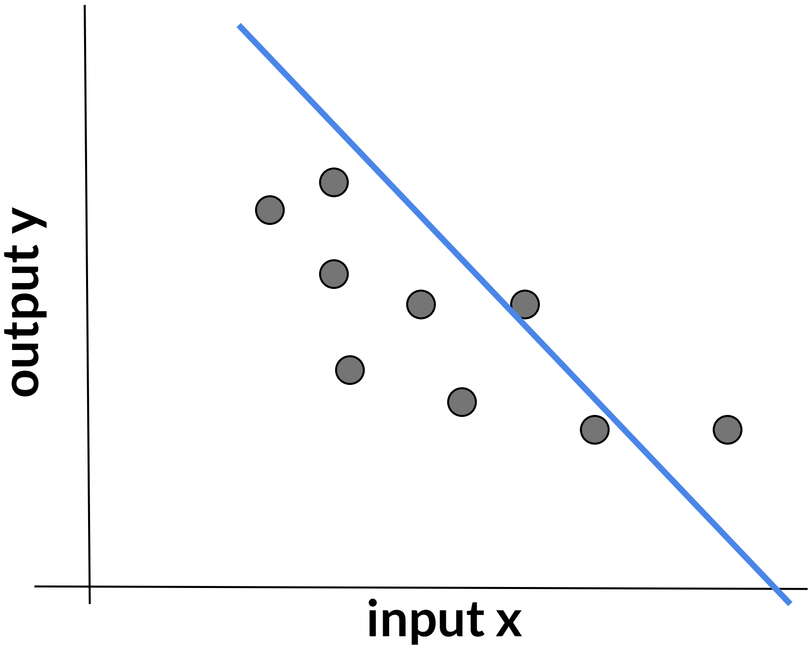
This is what you are used to seeing:



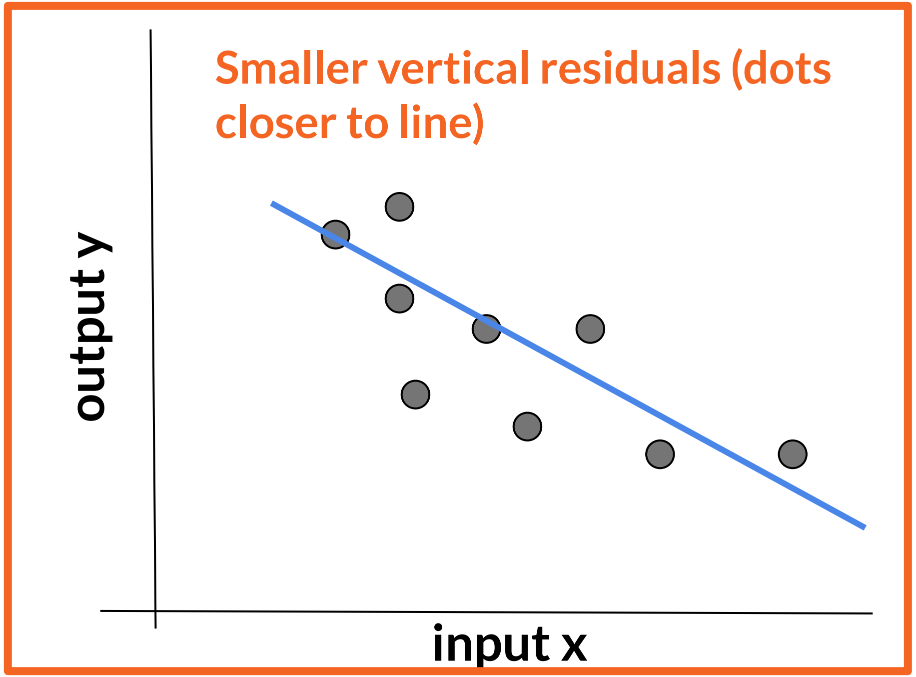
Blue line shows predictions $\alpha + \beta x$

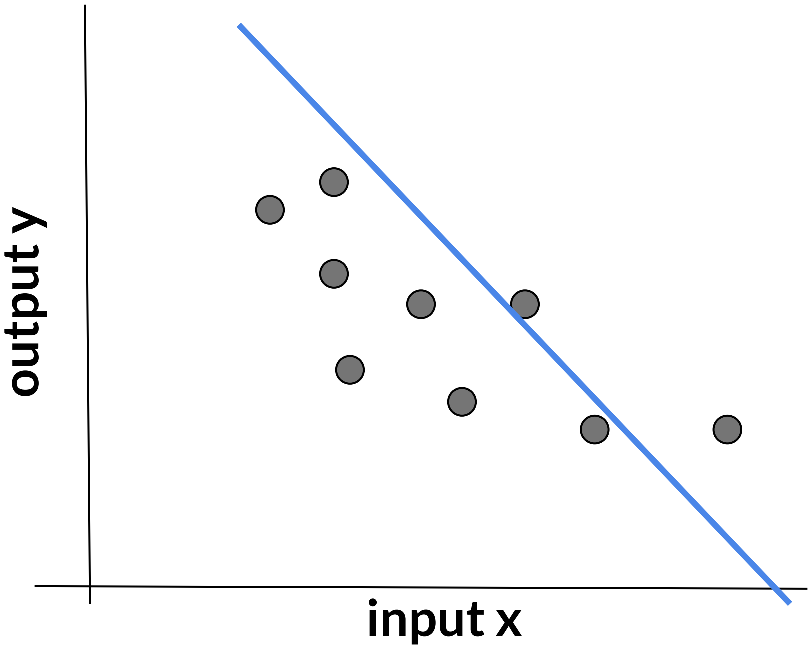
Which model has better predictions?



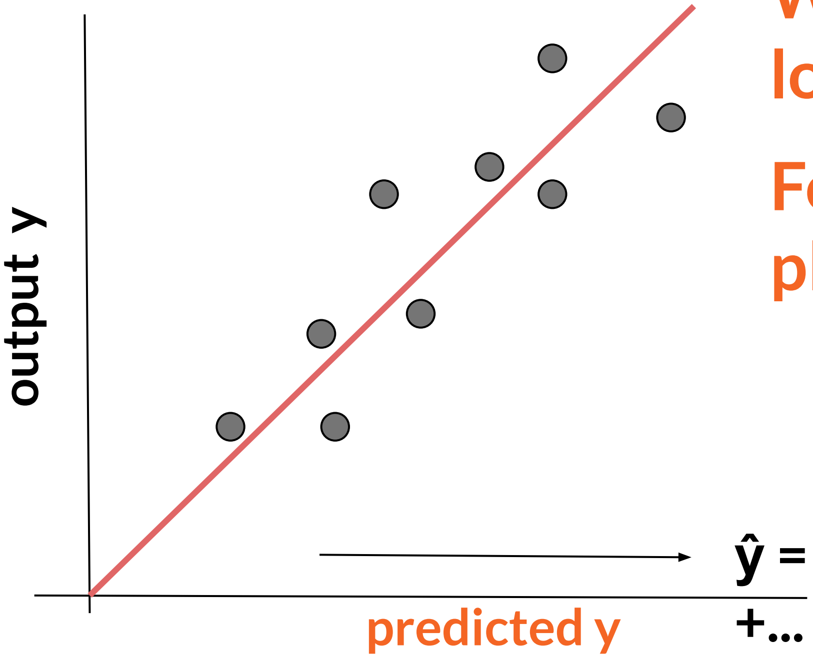


Which model has better predictions?





Predictions vs. Reality

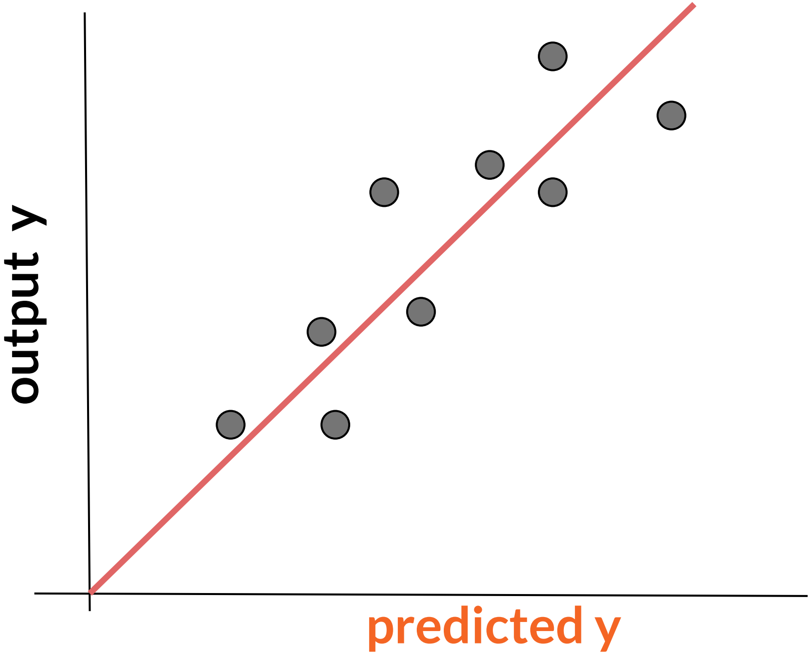


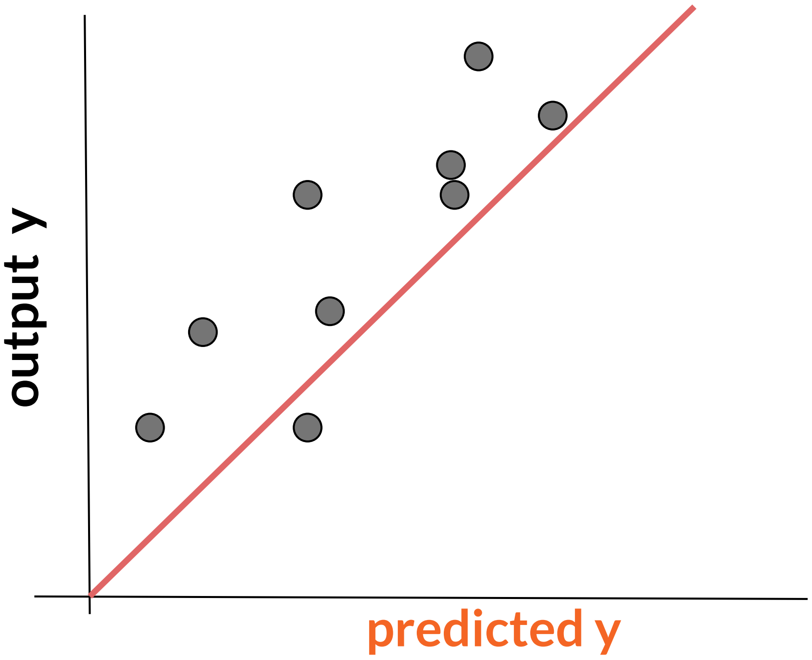
$$\hat{y} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots$$

What if we have lots of inputs?

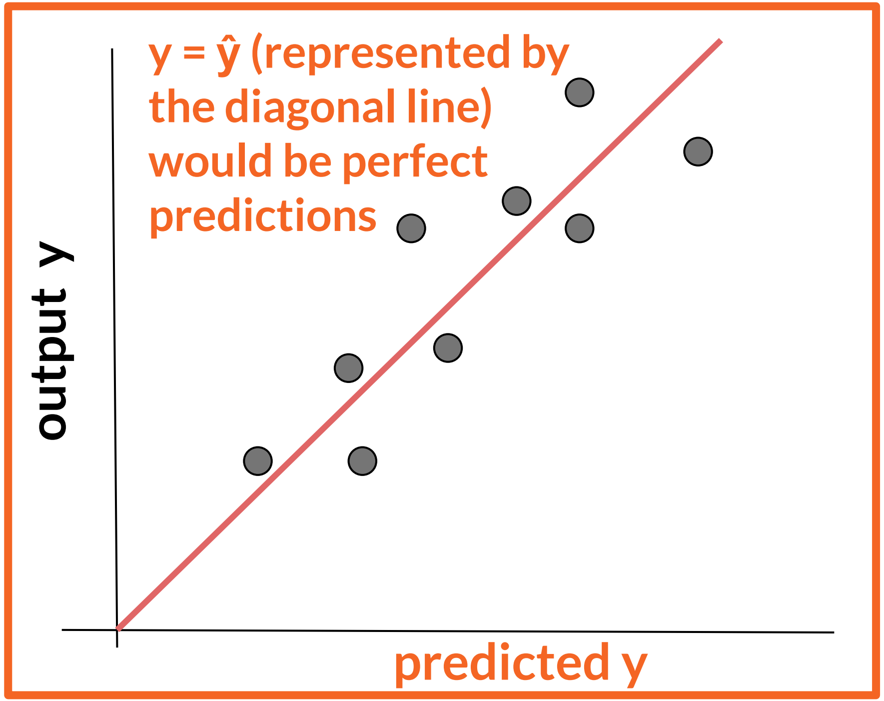
For each row i, plot ŷ vs. y

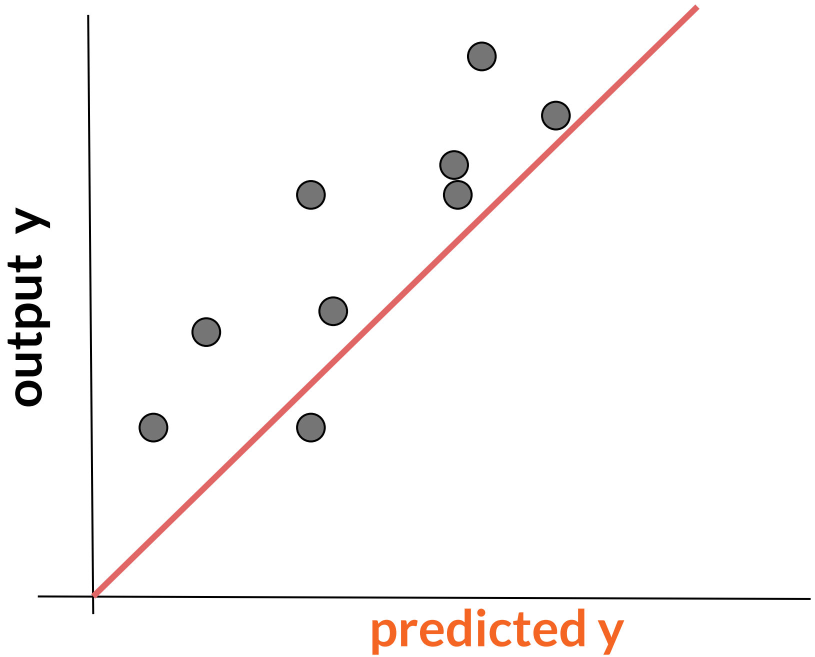
Which model has better predictions?





Which model has better predictions?

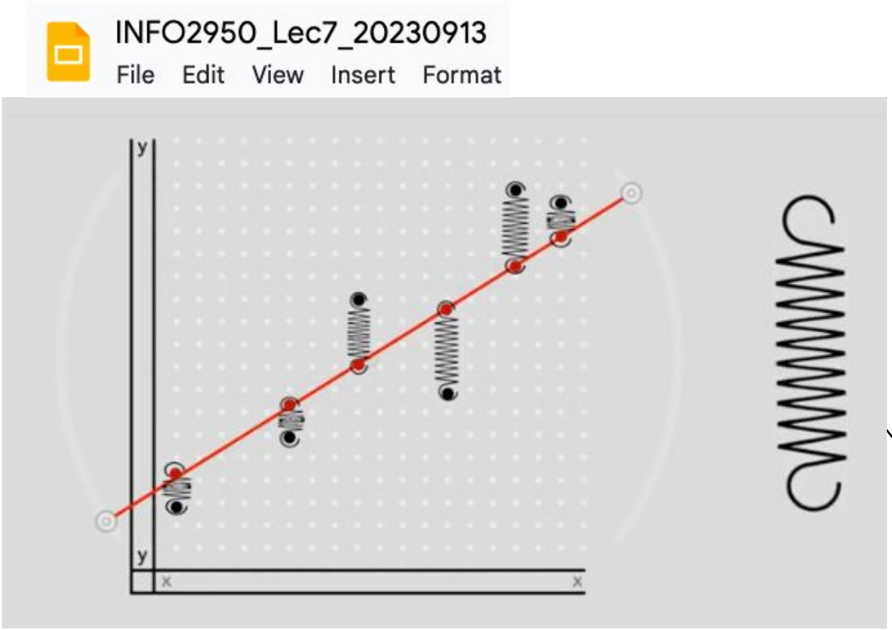




Multivariable Regression in Python

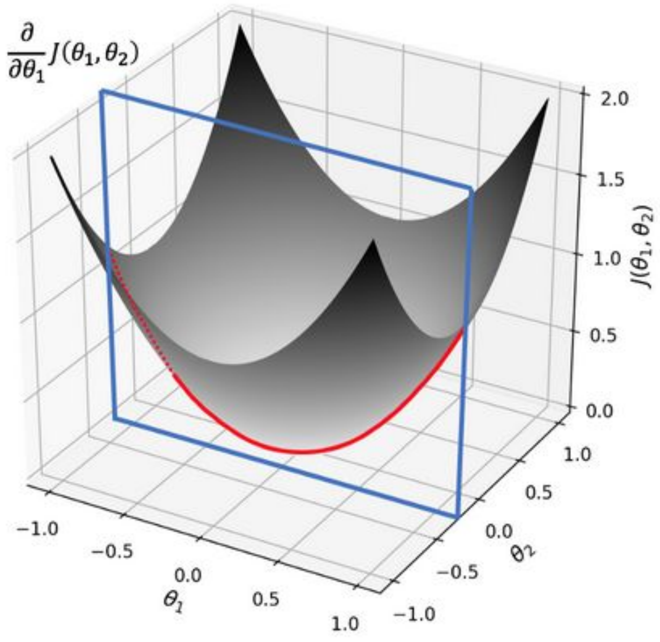
- We can use Python to get values for α, β_1, β_2, β_3, …
- For single variable regression, we learned about minimizing the squared error by hand (setting the derivative of Q to 0)
- What about with multiple variables?

Least Squares Intuition for Single Variable Regression



https://www.jmp.com/en_us/statistics-knowledge-portal/what-is-regression/the-method-of-least-squares.html

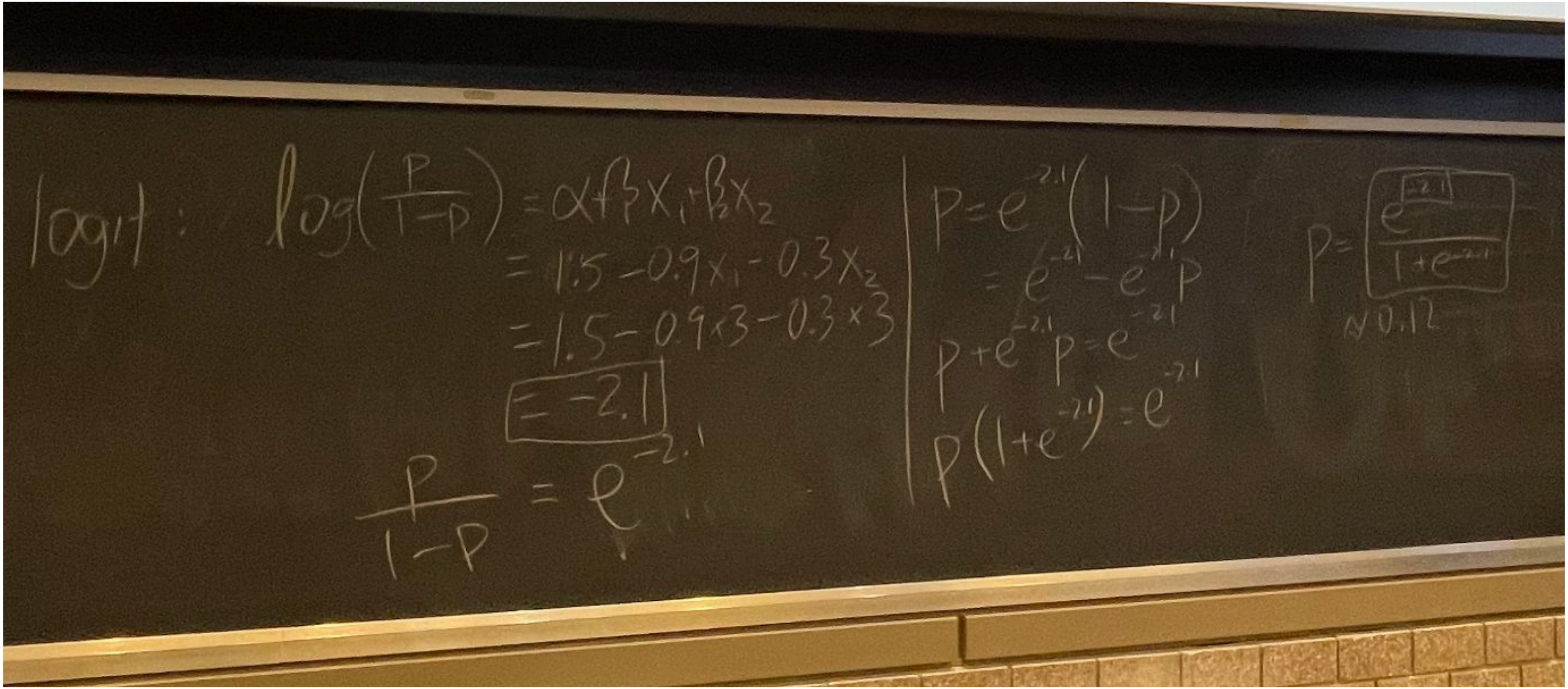
Multivariable Minimizing Intuition



https://towardsdatascience.com/machine-learning-bit-by-bit-multivariate-gradient-descent-e198fdd0df85

Attendance

tinyurl.com/yx3xvta5



Explanation for problem in slide 46