Linear Regression

University of the Witwatersrand

2025

Review Question

- Invert the matrix $\begin{pmatrix} 1 & 2 \\ 2 & 3 \end{pmatrix}$

Lesson Plan

- Review Question

Lesson Plan

- Review Question
- Idea of multivariate regression

Lesson Plan

- Review Question
- Idea of multivariate regression
- Potential Issues

Lesson Plan

- Review Question
- Idea of multivariate regression
- Potential Issues
- Ridge and Lasso

Review Question

- Invert the matrix $\begin{pmatrix} 1 & 2 \\ 2 & 3 \end{pmatrix}$

Review Question

- Invert the matrix $\begin{pmatrix} 1 & 2 \\ 2 & 3 \end{pmatrix}$
- $\begin{pmatrix} 1 & 2 & : & 1 & 0 \\ 2 & 3 & : & 0 & 1 \end{pmatrix}$

Review Question

- Invert the matrix $\begin{pmatrix} 1 & 2 \\ 2 & 3 \end{pmatrix}$
- $\begin{pmatrix} 1 & 2 & : & 1 & 0 \\ 2 & 3 & : & 0 & 1 \end{pmatrix}$
- $\begin{pmatrix} 1 & 2 & : & 1 & 0 \\ 0 & -1 & : & -2 & 1 \end{pmatrix}$

Review Question

- Invert the matrix $\begin{pmatrix} 1 & 2 \\ 2 & 3 \end{pmatrix}$
- $\begin{pmatrix} 1 & 2 & : & 1 & 0 \\ 2 & 3 & : & 0 & 1 \end{pmatrix}$
- $\begin{pmatrix} 1 & 2 & : & 1 & 0 \\ 0 & -1 & : & -2 & 1 \end{pmatrix}$
- $\begin{pmatrix} 1 & 2 & : & 1 & 0 \\ 0 & 1 & : & 2 & -1 \end{pmatrix}$

Review Question

- Invert the matrix $\begin{pmatrix} 1 & 2 \\ 2 & 3 \end{pmatrix}$
- $\begin{pmatrix} 1 & 2 & : & 1 & 0 \\ 2 & 3 & : & 0 & 1 \end{pmatrix}$
- $\begin{pmatrix} 1 & 2 & : & 1 & 0 \\ 0 & -1 & : & -2 & 1 \end{pmatrix}$
- $\begin{pmatrix} 1 & 2 & : & 1 & 0 \\ 0 & 1 & : & 2 & -1 \end{pmatrix}$
- $\begin{pmatrix} 1 & 0 & : & -3 & 2 \\ 0 & 1 & : & 2 & -1 \end{pmatrix}$

Idea of Multivariate Regression

- In a single variable we fit the line. $y = \beta_0 + \beta_1 x + \epsilon$

Idea of Multivariate Regression

- In a single variable we fit the line. $y = \beta_0 + \beta_1 x + \epsilon$
- In multivariate regression we have multiple x-values. For example we might be interested in the gold price next week and have, dollar-pound exhange rate, current gold price, current silver price and so on.

Idea of Multivariate Regression

- In a single variable we fit the line. $y = \beta_0 + \beta_1 x + \epsilon$
- In multivariate regression we have multiple x-values. For example we might be interested in the gold price next week and have, dollar-pound exhange rate, current gold price, current silver price and so on.
- We fit a hyper-plane $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon$

Idea of Multivariate Regression

- In a single variable we fit the line. $y = \beta_0 + \beta_1 x + \epsilon$
- In multivariate regression we have multiple x-values. For example we might be interested in the gold price next week and have, dollar-pound exhange rate, current gold price, current silver price and so on.
- We fit a hyper-plane $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \epsilon$
- We'll use p as the number of variables because n will be reserved for the number of observations.

Idea of Multivariate Regression

- We get some observations n data points each of which has p variables

Idea of Multivariate Regression

- We get some observations n data points each of which has p variables
- For convenience we'll make one of these variables a constant.

Idea of Multivariate Regression

- We get some observations n data points each of which has p variables
- For convenience we'll make one of these variables a constant.
- Once again we'll try to minimize $\mathbb{E}[(Y - \hat{Y})^2]$

Idea of Multivariate Regression

- We get some observations n data points each of which has p variables
- For convenience we'll make one of these variables a constant.
- Once again we'll try to minimize $\mathbb{E}[(Y - \hat{Y})^2]$
- Here that means $\sum(Y_i - (\beta_0 + \beta_1 x_1^i + \beta_2 x_2^i + ... + \beta_p x_p^i)$.

Idea of Multivariate Regression

- We get some observations n data points each of which has p variables
- For convenience we'll make one of these variables a constant.
- Once again we'll try to minimize $\mathbb{E}[(Y - \hat{Y})^2]$
- Here that means $\sum(Y_i - (\beta_0 + \beta_1 x_1^i + \beta_2 x_2^i + \ldots + \beta_p x_p^i)$.
- Superscript refers to a particular observation/data point.

Idea of Multivariate Regression

- We get some observations n data points each of which has p variables
- For convenience we'll make one of these variables a constant.
- Once again we'll try to minimize $\mathbb{E}[(Y - \hat{Y})^2]$
- Here that means $\sum(Y_i - (\beta_0 + \beta_1 x_1^i + \beta_2 x_2^i + ... + \beta_p x_p^i)$.
- Superscript refers to a particular observation/data point.
- To minimize this error we can take p partial derivatives

Idea of Multivariate Regression

- We can use some nice notation to make this easier.

Idea of Multivariate Regression

- We can use some nice notation to make this easier.
- Put our $n \times p$ data points in a p by n matrix X. Each row is an observation, each column a variable.

Idea of Multivariate Regression

- We can use some nice notation to make this easier.
- Put our $n \times p$ data points in a p by n matrix X. Each row is an observation, each column a variable.
- Then We get $\hat{\beta} = (X^T X)^{-1} X^T Y$

Idea of Multivariate Regression

- We can use some nice notation to make this easier.
- Put our $n \times p$ data points in a p by n matrix X. Each row is an observation, each column a variable.
- Then We get $\hat{\beta} = (X^T X)^{-1} X^T Y$
- We also get that $\hat{Y} = X\hat{\beta} = X(X^T X)^{-1} X^T Y$

Idea of Multivariate Regression

- We can use some nice notation to make this easier.
- Put our $n \times p$ data points in a p by n matrix X. Each row is an observation, each column a variable.
- Then We get $\hat{\beta} = (X^T X)^{-1} X^T Y$
- We also get that $\hat{Y} = X\hat{\beta} = X(X^T X)^{-1} X^T Y$
- As with the single variable case this works just fine as a calculus problem but the assumption that the ϵ are distributed normally, independently and with common variance allows us to to inference!

Idea of Multivariate Regression

- What inference do we want to do? Well we're working in the world where $Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \epsilon$ and we have best estimates on the β_i.

Idea of Multivariate Regression

- What inference do we want to do? Well we're working in the world where $Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \epsilon$ and we have best estimates on the β_i.
- As the ϵ s are normal, the Xs are known our estimated βs can be shown to be normal.

Idea of Multivariate Regression

- What inference do we want to do? Well we're working in the world where $Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \epsilon$ and we have best estimates on the β_i.
- As the ϵ s are normal, the Xs are known our estimated βs can be shown to be normal.
- This is very much like our usual hypothesis-testing situation! There is a real parameter, we have some estimate for it and some idea of how variable it is.

Idea of Multivariate Regression

- What inference do we want to do? Well we're working in the world where $Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \epsilon$ and we have best estimates on the β_i.
- As the ϵ s are normal, the Xs are known our estimated βs can be shown to be normal.
- This is very much like our usual hypothesis-testing situation! There is a real parameter, we have some estimate for it and some idea of how variable it is.
- I'm hiding some linear algebra here but it turns out that we can test for individual $\beta_i = 0$.

Idea of Multivariate Regression

- What inference do we want to do? Well we're working in the world where $Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \epsilon$ and we have best estimates on the β_i.
- As the ϵ s are normal, the Xs are known our estimated βs can be shown to be normal.
- This is very much like our usual hypothesis-testing situation! There is a real parameter, we have some estimate for it and some idea of how variable it is.
- I'm hiding some linear algebra here but it turns out that we can test for individual $\beta_i = 0$.
- This matters because if some β are zero it means that we can ignore those variables!

Idea of inference

- A notable point about regression testing is that when the null-hypothesis is actually true we'll reject it five-percent of the time!

Idea of inference

- A notable point about regression testing is that when the null-hypothesis is actually true we'll reject it five-percent of the time!
- If you have a hundred useless variables then by chance alone you should expect to think five of them are significant at five percent.

Idea of inference

- A notable point about regression testing is that when the null-hypothesis is actually true we'll reject it five-percent of the time!
- If you have a hundred useless variables then by chance alone you should expect to think five of them are significant at five percent.
- In some situations with a lot of random variables it's useful to keep this kind of thing in mind.

Idea of inference

- A notable point about regression testing is that when the null-hypothesis is actually true we'll reject it five-percent of the time!
- If you have a hundred useless variables then by chance alone you should expect to think five of them are significant at five percent.
- In some situations with a lot of random variables it's useful to keep this kind of thing in mind.
- Options for dealing with this:

Idea of inference

- A notable point about regression testing is that when the null-hypothesis is actually true we'll reject it five-percent of the time!
- If you have a hundred useless variables then by chance alone you should expect to think five of them are significant at five percent.
- In some situations with a lot of random variables it's useful to keep this kind of thing in mind.
- Options for dealing with this:
  - Pre-set your α to something lower.

Idea of inference

- A notable point about regression testing is that when the null-hypothesis is actually true we'll reject it five-percent of the time!
- If you have a hundred useless variables then by chance alone you should expect to think five of them are significant at five percent.
- In some situations with a lot of random variables it's useful to keep this kind of thing in mind.
- Options for dealing with this:
  - Pre-set your α to something lower.
  - Have training set and a testing set. Or even several hold-out sets.

Idea of inference

- A notable point about regression testing is that when the null-hypothesis is actually true we'll reject it five-percent of the time!
- If you have a hundred useless variables then by chance alone you should expect to think five of them are significant at five percent.
- In some situations with a lot of random variables it's useful to keep this kind of thing in mind.
- Options for dealing with this:
  - Pre-set your α to something lower.
  - Have training set and a testing set. Or even several hold-out sets.
  - Ridge/Lasso regression (more later).

Other issues

- Your X variables might be well correlated. This is not too bad for estimation, but it makes it difficult to find accurate β (wide confidence intervals).

Other issues

- Your X variables might be well correlated. This is not too bad for estimation, but it makes it difficult to find accurate β (wide confidence intervals).
- Imagine we're trying to tell a child's age from his height and weight. This is pretty likely to be a predictive model because teenagers are taller and heavier than toddlers.

Other issues

- Your X variables might be well correlated. This is not too bad for estimation, but it makes it difficult to find accurate β (wide confidence intervals).
- Imagine we're trying to tell a child's age from his height and weight. This is pretty likely to be a predictive model because teenagers are taller and heavier than toddlers.
- On the other hand height and weight are really well correlated with each other.

Other issues

- Your X variables might be well correlated. This is not too bad for estimation, but it makes it difficult to find accurate β (wide confidence intervals).
- Imagine we're trying to tell a child's age from his height and weight. This is pretty likely to be a predictive model because teenagers are taller and heavier than toddlers.
- On the other hand height and weight are really well correlated with each other.
- Maybe the true model is $Y = X_1 + X_2 + \epsilon$ but with X_1 and X_2 really similar this would make $\hat{Y} = 2X_1$ and $\hat{Y} = 2X_2$ nearly as good fits as $\hat{Y} = X_1 + X_2$, worse so is $hatY = 200X_1 - 199X_2$. Your estimates on β_i become larger

Feature engineering

- We can usually add new variables by taking functions of our current ones.

Feature engineering

- We can usually add new variables by taking functions of our current ones.
- Maybe we add in an X^2 term because we think it that X might have a sweet spot and not be linear.

Feature engineering

- We can usually add new variables by taking functions of our current ones.
- Maybe we add in an X^2 term because we think it that X might have a sweet spot and not be linear.
- Perhaps we add an $X_3 = X_1 X_2$ (interaction term)

Feature engineering

- We can usually add new variables by taking functions of our current ones.
- Maybe we add in an X^2 term because we think it that X might have a sweet spot and not be linear.
- Perhaps we add an $X_3 = X_1 X_2$ (interaction term)
- You might log a variable if it has a natural interpretation

Feature engineering

- We can usually add new variables by taking functions of our current ones.
- Maybe we add in an X^2 term because we think it that X might have a sweet spot and not be linear.
- Perhaps we add an $X_3 = X_1 X_2$ (interaction term)
- You might log a variable if it has a natural interpretation
- This can get out of control quickly.

Overfitting

- With enough variables available or engineered we can fit every point exactly

Overfitting

- With enough variables available or engineered we can fit every point exactly
- Think of fitting n poitns with an $n - 1$ degree polynomial.

Overfitting

- With enough variables available or engineered we can fit every point exactly
- Think of fitting n poitns with an $n-1$ degree polynomial.
- The problem is that these don't generalize at all.

Overfitting

- With enough variables available or engineered we can fit every point exactly
- Think of fitting n poitns with an $n-1$ degree polynomial.
- The problem is that these don't generalize at all.
- You might log a variable if it has a natural interpretation

Overfitting

- With enough variables available or engineered we can fit every point exactly
- Think of fitting n poitns with an $n-1$ degree polynomial.
- The problem is that these don't generalize at all.
- You might log a variable if it has a natural interpretation
- This can get out of control quickly.

Overfitting

- With enough variables available or engineered we can fit every point exactly
- Think of fitting n poitns with an $n - 1$ degree polynomial.
- The problem is that these don't generalize at all.
- You might log a variable if it has a natural interpretation
- This can get out of control quickly.
- Overfitted_Data.png

Subset Selection

- Sometimes you have a lot of variables and suspect that some don't matter. Maybe you want to perform an intervention.

Subset Selection

- Sometimes you have a lot of variables and suspect that some don't matter. Maybe you want to perform an intervention.
- Problem. If you look at the best fitting model it's unlikely that any of the $\hat{\beta}_i$ are actually zero. And if you have high correlations coefficients your model coefficients can blow up and destroy model interpret ability!

Subset Selection

- Sometimes you have a lot of variables and suspect that some don't matter. Maybe you want to perform an intervention.
- Problem. If you look at the best fitting model it's unlikely that any of the $\hat{\beta}_i$ are actually zero. And if you have high correlations coefficients your model coefficients can blow up and destroy model interpret ability!
- We have criteria to deal with with AIC and BIC to favour lower dimensional models.

Forward and Backwards Selection

- Can't really check ALL models (2^p such models)

Forward and Backwards Selection

- Can't really check ALL models (2^p such models)
- Forward selection, try to add variables in

Forward and Backwards Selection

- Can't really check ALL models (2^p such models)
- Forward selection, try to add variables in
- Backwards selection, try to take them away

Ridge and Lasso Regression

- OK we're trying to avoid using all the variables and maybe don't want to worry about weather we're overfitting by rererefitting.

Ridge and Lasso Regression

- OK we're trying to avoid using all the variables and maybe don't want to worry about weather we're overfitting by rererefitting.
- We have another tool called Lasso regression

Ridge and Lasso Regression

- OK we're trying to avoid using all the variables and maybe don't want to worry about weather we're overfitting by rererefitting.
- We have another tool called Lasso regression
- Lasso Regression:
  Minimize $\mathbb{E}[(y - \hat{y})^2] + \lambda|\hat{\beta}|$

Ridge and Lasso Regression

- OK we're trying to avoid using all the variables and maybe don't want to worry about weather we're overfitting by rererefitting.
- We have another tool called Lasso regression
- Lasso Regression:
  Minimize $\mathbb{E}[(y - \hat{y})^2] + \lambda|\hat{\beta}|$
- Often sets some coefficents to zero.

Ridge and Lasso Regression

- OK we're trying to avoid using all the variables and maybe don't want to worry about weather we're overfitting by rererefitting.
- We have another tool called Lasso regression
- Lasso Regression:
  Minimize $\mathbb{E}[(y - \hat{y})^2] + \lambda|\hat{\beta}|$
- Often sets some coefficents to zero.
- Contrasts with Ridge Regression:
  Minimize $\mathbb{E}[(y - \hat{y})^2] + \lambda\hat{\beta}^2$, which keeps coefficents small but rarely actually gets them to zero

Idea of Logistic Regression

- Sometimes we want to classify things.

Idea of Logistic Regression

- Sometimes we want to classify things.
- There are a lot of good classification algorithms. K-NN, Support vector machines, Decision Tree, Random Forest and so on.

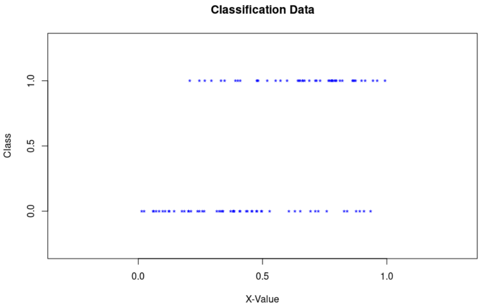
Idea of Logistic Regression

- Sometimes we want to classify things.
- There are a lot of good classification algorithms. K-NN, Support vector machines, Decision Tree, Random Forest and so on.
- The idea of logistic regression is that we'll set up an indicator random variable. 1 if in the group 0 if not (two class classification problem).

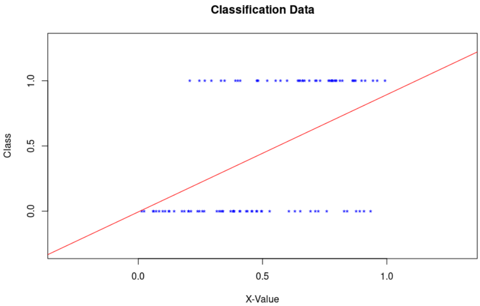
Idea of Logistic Regression

- Sometimes we want to classify things.
- There are a lot of good classification algorithms. K-NN, Support vector machines, Decision Tree, Random Forest and so on.
- The idea of logistic regression is that we'll set up an indicator random variable. 1 if in the group 0 if not (two class classification problem).
- Then we try to do regression on it.

A problem



A problem



Fitting a curve

- OK so we'd like to fit a curve that's bounded between zero and one.

Fitting a curve

- OK so we'd like to fit a curve that's bounded between zero and one.
- This can be thought of as representing probability of being in the class.

Fitting a curve

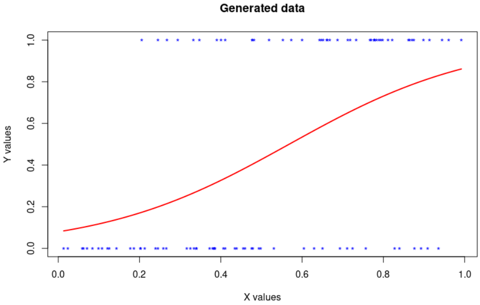
- OK so we'd like to fit a curve that's bounded between zero and one.
- This can be thought of as representing probability of being in the class.
- Could use a lot of functions. In machine learning we sometimes do. For logistic regression we use a sigmoid function

A problem

- Looks like:

A problem

- Looks like:



-

Sigmoid Function

- We us the model as $p(X) = \frac{e^{\beta^T X}}{1+e^{\beta^T X}}$

Sigmoid Function

- We us the model as $p(X) = \frac{e^{\beta^T X}}{1+e^{\beta^T X}}$
- This is really a function of $Z = e^{\beta^T X}$. It's monotonic and ranges from 0 to 1.

Sigmoid Function

- We us the model as $p(X) = \frac{e^{\beta^T X}}{1+e^{\beta^T X}}$
- This is really a function of $Z = e^{\beta^T X}$. It's monotonic and ranges from 0 to 1.
-

$$\begin{aligned}
p(X) &= \frac{e^{\beta^T X}}{1 + e^{\beta^T X}} \\
p(X) &= \frac{Z}{1 + Z} \\
p(X)(1 + Z) &= Z \\
p(X) + Zp(X) &= Z
\end{aligned}$$

Sigmoid Function

- $$\begin{aligned} p(X) + Zp(X) &= Z \\ p(X) &= Z(1 - p(X)) \\ Z &= \frac{p(X)}{1 - p(X)} \\ \frac{p(X)}{1 - p(X)} &= e^{\beta^T X} \\ \ln \frac{p(X)}{1 - p(X)} &= \beta^T X \end{aligned}$$

Sigmoid Function

- $$\begin{aligned} p(X) + Zp(X) &= Z \\ p(X) &= Z(1 - p(X)) \\ Z &= \frac{p(X)}{1 - p(X)} \\ \frac{p(X)}{1 - p(X)} &= e^{\beta^T X} \\ \ln \frac{p(X)}{1 - p(X)} &= \beta^T X \end{aligned}$$

- That is to say that the "log odds" is modeled as a linear function of the X variables.

Interpretation

- Writing this as $\frac{p(X)}{1-p(X)} = e^{\beta^T X}$ leads to the interpretation that increasing X_i by one unit multiplies the odds by e^{β_i}

Interpretation

- Writing this as $\frac{p(X)}{1-p(X)} = e^{\beta^T X}$ leads to the interpretation that increasing X_i by one unit multiplies the odds by e^{β_i}
- Parameters are found via maximum likelihood estimation.

Interpretation

- Writing this as $\frac{p(X)}{1-p(X)} = e^{\beta^T X}$ leads to the interpretation that increasing X_i by one unit multiplies the odds by e^{β_i}
- Parameters are found via maximum likelihood estimation.
- In the case of linear regression likelihood maximization and least squares are equivalent.

Interpretation

- Writing this as $\frac{p(X)}{1-p(X)} = e^{\beta^T X}$ leads to the interpretation that increasing X_i by one unit multiplies the odds by e^{β_i}
- Parameters are found via maximum likelihood estimation.
- In the case of linear regression likelihood maximization and least squares are equivalent.
- We will cover the derivation of this but the upshot is that we don't have a nice closed form solution for β and in practice rely of software to compute things (mostly these use Newton's method because gradient decent turns out to be slower for this problem).

Interpretation

- Writing this as $\frac{p(X)}{1-p(X)} = e^{\beta^T X}$ leads to the interpretation that increasing X_i by one unit multiplies the odds by e^{β_i}
- Parameters are found via maximum likelihood estimation.
- In the case of linear regression likelihood maximization and least squares are equivalent.
- We will cover the derivation of this but the upshot is that we don't have a nice closed form solution for β and in practice rely of software to compute things (mostly these use Newton's method because gradient decent turns out to be slower for this problem).
- Due to similarities with Linear Regression we can do inference on β. We won't here but your favourite software package will have tests.

Proof

- $$L(\theta) = \prod_{y_i=1} p(X_i) \prod_{y_i=0} 1 - p(X_i)$$

Proof

- $$L(\theta) = \prod_{y_i=1} p(X_i) \prod_{y_i=0} 1 - p(X_i)$$
- $$L(\theta) = \prod_{i} p(X_i)^{y_i}(1 - p(X_i))^{1-y_i}$$

Proof

- $$L(\theta) = \prod_{y_i=1} p(X_i) \prod_{y_i=0} 1 - p(X_i)$$
- $$L(\theta) = \prod_{i} p(X_i)^{y_i}(1 - p(X_i))^{1-y_i}$$
- $$\begin{aligned} l(\beta) &= \sum_i y_i \ln p(x_i) + (1 - y_i) \ln[1 - p(x_i)] \\ &= \sum_i y_i \ln \frac{p(x_i)}{1 - p(x_i)} + \ln[1 - p(x_i)] \\ &= \sum_i y_i(X^T\beta) - \ln[1 + e^{X^T\beta}] \end{aligned}$$

Proof

- $$L(\theta) = \prod_{y_i=1} p(X_i) \prod_{y_i=0} 1 - p(X_i)$$
- $$L(\theta) = \prod_i p(X_i)^{y_i} (1 - p(X_i))^{1-y_i}$$
- $$\begin{aligned} l(\beta) &= \sum_i y_i \ln p(x_i) + (1 - y_i) \ln[1 - p(x_i)] \\ &= \sum_i y_i \ln \frac{p(x_i)}{1 - p(x_i)} + \ln[1 - p(x_i)] \\ &= \sum_i y_i (X^T \beta) - \ln[1 + e^{X^T \beta}] \end{aligned}$$
- Numerics