



MONASH University

ETC3250

Business Analytics

Week 3

Linear regression

13 March 2018

Outline

Week	Topic	Chapter	Lecturer
1	Introduction	1	Souhaib
2	Statistical learning	2	Souhaib
3	Regression	3	Souhaib
4	Classification	4	Souhaib
5	Clustering	10	Souhaib
Semester break			
6	Model selection and resampling methods	5	Souhaib
7	Dimension reduction	6,10	Souhaib
8	Advanced regression	6	Souhaib
9	Advanced regression	6	Souhaib
10	Advanced classification	9	Souhaib
11	Tree-based methods	8	Souhaib
12	Project presentation		Souhaib

Outline

1 Revision: multiple regression

2 Matrix formulation

Revision: Multiple regression

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

- Each $X_{j,i}$ is numerical and is called a “predictor”.
- The coefficients β_1, \dots, β_p measure the effect of each predictor after taking account of the effect of all other predictors in the model.
- Predictors may be transforms of other predictors. e.g., $X_2 = X_1^2$.
- The model describes a line, plane or hyperplane in the predictor space.

Revision: Multiple regression

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

- Each $X_{j,i}$ is numerical and is called a “predictor”.
- The coefficients β_1, \dots, β_p measure the effect of each predictor after taking account of the effect of all other predictors in the model.
- Predictors may be transforms of other predictors. e.g., $X_2 = X_1^2$.
- The model describes a line, plane or hyperplane in the predictor space.

Revision: Multiple regression

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

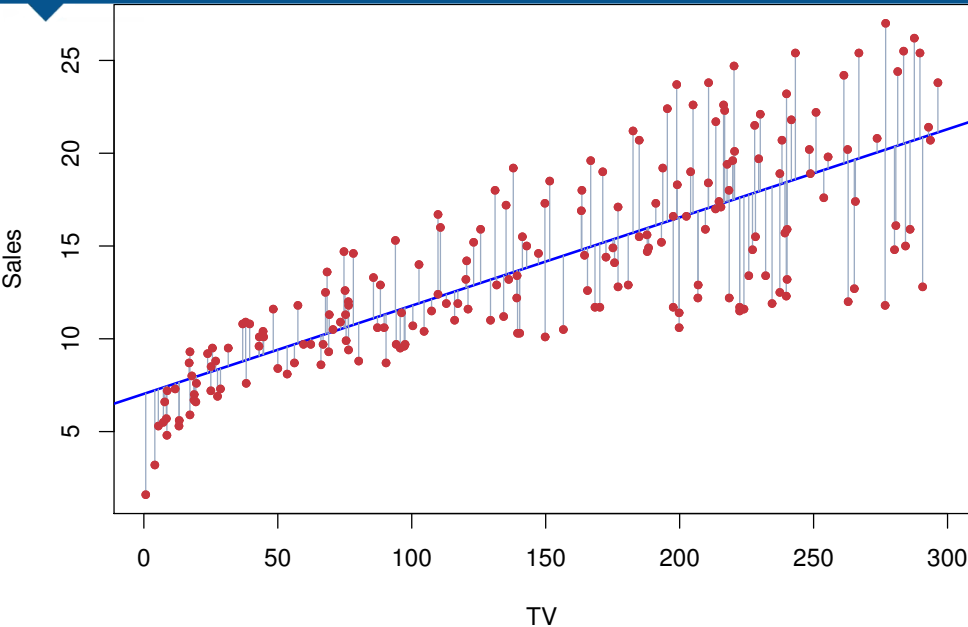
- Each $X_{j,i}$ is numerical and is called a “predictor”.
- The coefficients β_1, \dots, β_p measure the effect of each predictor after taking account of the effect of all other predictors in the model.
- Predictors may be transforms of other predictors. e.g., $X_2 = X_1^2$.
- The model describes a line, plane or hyperplane in the predictor space.

Revision: Multiple regression

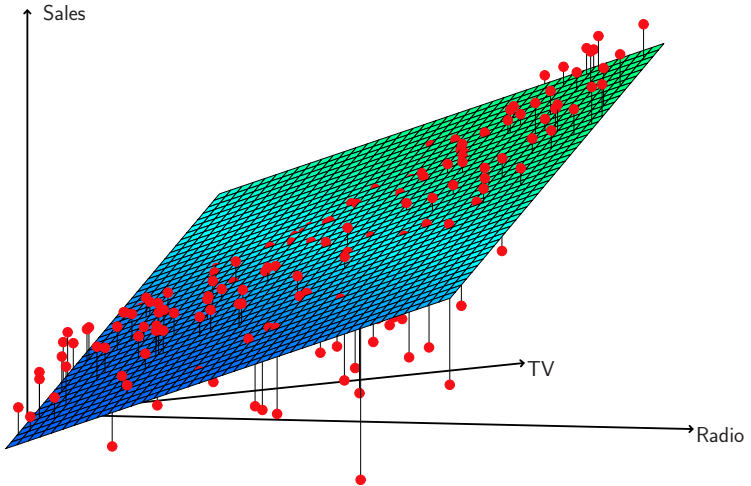
$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \cdots + \beta_p X_{p,i} + e_i.$$

- Each $X_{j,i}$ is numerical and is called a “predictor”.
- The coefficients β_1, \dots, β_p measure the effect of each predictor after taking account of the effect of all other predictors in the model.
- Predictors may be transforms of other predictors. e.g., $X_2 = X_1^2$.
- The model describes a line, plane or hyperplane in the predictor space.

Revision: Multiple regression



Revision: Multiple regression



Revision: Multiple regression

- Dummy variables
- OLS
- R^2
- se of coefficients
- residual standard error
- F statistic

Revision: Multiple regression

- Dummy variables
- OLS
- R^2
- se of coefficients
- residual standard error
- F statistic

Revision: Multiple regression

- Dummy variables
- OLS
- R^2
- se of coefficients
- residual standard error
- F statistic

Revision: Multiple regression

- Dummy variables
- OLS
- R^2
- se of coefficients
- residual standard error
- F statistic

Revision: Multiple regression

- Dummy variables
- OLS
- R^2
- se of coefficients
- residual standard error
- F statistic

Revision: Multiple regression

- Dummy variables
- OLS
- R^2
- se of coefficients
- residual standard error
- F statistic

Important questions

- 1 Is at least one of the predictors useful in predicting the response?
- 2 Do all the predictors help to explain Y , or is only a subset of the predictors useful?
- 3 How well does the model fit the data?
- 4 Given a set of predictor values, what response value should we predict and how accurate is our prediction?

Credit scores

Banks score loan customers based on a lot of personal information. A sample of 500 customers from an Australian bank provided the following information.

Score	Savings \$'000	Income \$'000	Time current address Months	Time current job Months
39.40	0.01	111.17	27	8
51.79	0.65	56.40	29	33
32.82	0.75	36.74	2	16
57.31	0.62	55.99	14	7
37.17	4.13	62.04	2	14
33.69	0.00	43.75	7	7
25.56	0.94	79.01	4	11
32.04	0.00	45.41	3	3
41.34	4.26	55.22	16	18
⋮	⋮	⋮	⋮	⋮

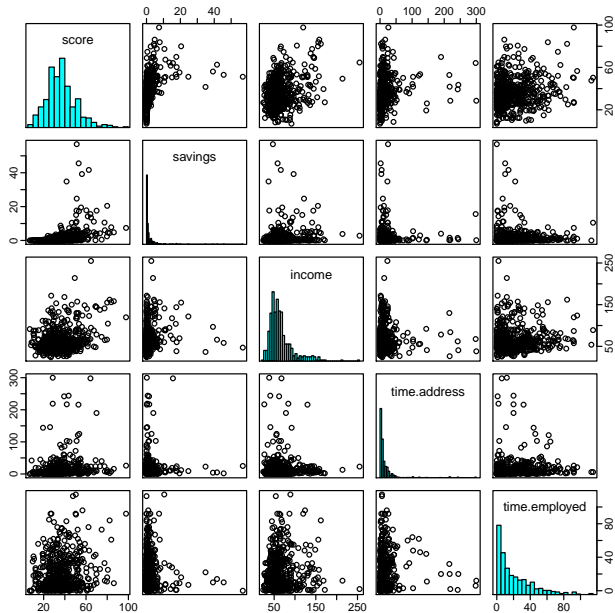
Credit scores

Banks score loan customers based on a lot of personal information. A sample of 500 customers from an Australian bank provided the following information.

Score	Savings \$'000	Income \$'000	Time current address Months	Time current job Months
39.40	0.01	111.17	27	8
51.79	0.65	56.40	20	33
32.82	0.7			6
57.31	0.6			7
37.17	4.1			4
33.69	0.0			7
25.56	0.94	79.01	4	11
32.04	0.00	45.41	3	3
41.34	4.26	55.22	16	18
:	:	:	:	:

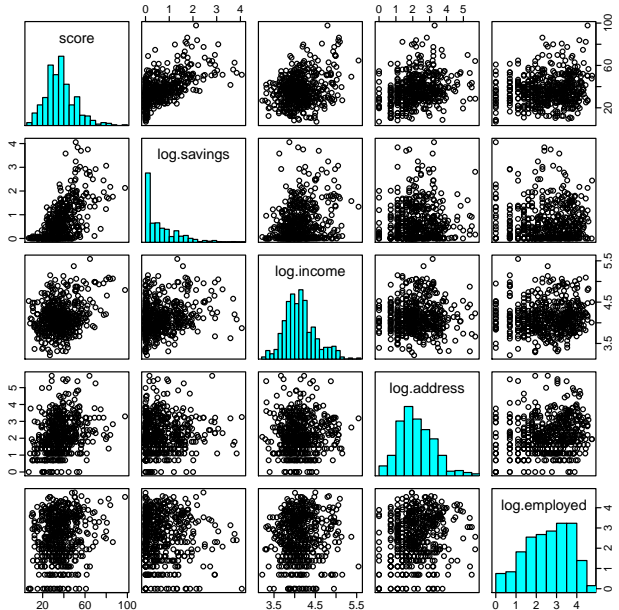
Can we use only savings, income, time @ address and time employed to predict the score of a new customer?

Credit scores



Credit scores

- Taking logarithms reduces the skewness in the predictor variables.
- Because of zeros, I used $\log(x + 1)$.



Credit scores

Proposed model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + e$$

where Y = Credit score,

X_1 = log savings,

X_2 = log income,

X_3 = log time at current address,

X_4 = log time in current job,

e = error.

Credit scores

```
lm(formula = score ~ log.savings + log.income + log.address +  
    log.employed, data = creditlog)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.219	5.231	-0.04	0.9667
log.savings	10.353	0.612	16.90	< 2e-16
log.income	5.052	1.258	4.02	6.8e-05
log.address	2.667	0.434	6.14	1.7e-09
log.employed	1.314	0.409	3.21	0.0014

Residual standard error: 10.2 on 495 degrees of freedom

Multiple R-squared: 0.47, Adjusted R-squared: 0.466

F-statistic: 110 on 4 and 495 DF, p-value: <2e-16

Interpreting regression coefficients

- The ideal scenario is when the predictors are uncorrelated.
 - Each coefficient can be interpreted and tested separately.
- Correlations amongst predictors cause problems.
 - The variance of all coefficients tends to increase, sometimes dramatically.
 - Interpretations become hazardous – when X_j changes, everything else changes.
 - Predictions still work provided new X values are within the range of training X values.
- Claims of causality should be avoided for observational data.

Interactions

- An interaction occurs when the one variable changes the effect of a second variable. (e.g., spending on radio advertising increases the effectiveness of TV advertising).
- To model an interaction, include the product X_1X_2 in the model in addition to X_1 and X_2 .
- **Hierarchy principle:** If we include an interaction in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant. (This is because the interactions are almost impossible to interpret without the main effects.)

Residual patterns

- If a plot of the residuals vs any predictor in the model shows a pattern, then the relationship is nonlinear.
- If a plot of the residuals vs any predictor **not** in the model shows a pattern, then the predictor should be added to the model.
- If a plot of the residuals vs fitted values shows a pattern, then there is heteroscedasticity in the errors. (Could try a transformation.)

Residual patterns

- If a plot of the residuals vs any predictor in the model shows a pattern, then the relationship is nonlinear.
- If a plot of the residuals vs any predictor **not** in the model shows a pattern, then the predictor should be added to the model.
- If a plot of the residuals vs fitted values shows a pattern, then there is heteroscedasticity in the errors. (Could try a transformation.)

Residual patterns

- If a plot of the residuals vs any predictor in the model shows a pattern, then the relationship is nonlinear.
- If a plot of the residuals vs any predictor **not** in the model shows a pattern, then the predictor should be added to the model.
- If a plot of the residuals vs fitted values shows a pattern, then there is heteroscedasticity in the errors. (Could try a transformation.)

Outline

1 Revision: multiple regression

2 Matrix formulation

Matrix formulation

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

Let $\mathbf{Y} = (Y_1, \dots, Y_n)'$, $\mathbf{e} = (e_1, \dots, e_n)'$,
 $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)'$ and

$$\mathbf{X} = \begin{bmatrix} 1 & X_{1,1} & X_{2,1} & \cdots & X_{p,1} \\ 1 & X_{1,2} & X_{2,2} & \cdots & X_{p,2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{1,n} & X_{2,n} & \cdots & X_{p,n} \end{bmatrix}.$$

Then

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}.$$

Matrix formulation

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

Let $\mathbf{Y} = (Y_1, \dots, Y_n)'$, $\mathbf{e} = (e_1, \dots, e_n)'$,
 $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)'$ and

$$\mathbf{X} = \begin{bmatrix} 1 & X_{1,1} & X_{2,1} & \cdots & X_{p,1} \\ 1 & X_{1,2} & X_{2,2} & \cdots & X_{p,2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{1,n} & X_{2,n} & \cdots & X_{p,n} \end{bmatrix}.$$

Then

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}.$$

Matrix formulation

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

Let $\mathbf{Y} = (Y_1, \dots, Y_n)'$, $\mathbf{e} = (e_1, \dots, e_n)'$,
 $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)'$ and

$$\mathbf{X} = \begin{bmatrix} 1 & X_{1,1} & X_{2,1} & \cdots & X_{p,1} \\ 1 & X_{1,2} & X_{2,2} & \cdots & X_{p,2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{1,n} & X_{2,n} & \cdots & X_{p,n} \end{bmatrix}.$$

Then

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}.$$

Matrix formulation

Least squares estimation

Minimize: $(\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta)$

Differentiate wrt β and equal to zero gives

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

(The “normal equation”.)

$$\hat{\sigma}^2 = \frac{1}{n - p - 1}(\mathbf{Y} - \mathbf{X}\hat{\beta})'(\mathbf{Y} - \mathbf{X}\hat{\beta})$$

Note: If you fall for the dummy variable trap, $(\mathbf{X}'\mathbf{X})$ is a singular matrix.

Matrix formulation

Least squares estimation

Minimize: $(\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta)$

Differentiate wrt β and equal to zero gives

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

(The “normal equation”.)

$$\hat{\sigma}^2 = \frac{1}{n - p - 1}(\mathbf{Y} - \mathbf{X}\hat{\beta})'(\mathbf{Y} - \mathbf{X}\hat{\beta})$$

Note: If you fall for the dummy variable trap, $(\mathbf{X}'\mathbf{X})$ is a singular matrix.

Matrix formulation

Least squares estimation

Minimize: $(\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta)$

Differentiate wrt β and equal to zero gives

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

(The “normal equation”.)

$$\hat{\sigma}^2 = \frac{1}{n - p - 1}(\mathbf{Y} - \mathbf{X}\hat{\beta})'(\mathbf{Y} - \mathbf{X}\hat{\beta})$$

Note: If you fall for the dummy variable trap, $(\mathbf{X}'\mathbf{X})$ is a singular matrix.

Matrix formulation

Least squares estimation

Minimize: $(\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta)$

Differentiate wrt β and equal to zero gives

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

(The “normal equation”.)

$$\hat{\sigma}^2 = \frac{1}{n - p - 1}(\mathbf{Y} - \mathbf{X}\hat{\beta})'(\mathbf{Y} - \mathbf{X}\hat{\beta})$$

Note: If you fall for the dummy variable trap, $(\mathbf{X}'\mathbf{X})$ is a singular matrix.

Likelihood

If the errors are iid and normally distributed, then

$$\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}).$$

So the likelihood is

$$L = \frac{1}{\sigma^n (2\pi)^{n/2}} \exp \left(-\frac{1}{2\sigma^2} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \right)$$

which is maximized when $(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$ is minimized.

So **MLE** \equiv **OLS**.

Likelihood

If the errors are iid and normally distributed, then

$$\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}).$$

So the likelihood is

$$L = \frac{1}{\sigma^n (2\pi)^{n/2}} \exp \left(-\frac{1}{2\sigma^2} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \right)$$

which is maximized when $(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$ is minimized.

So **MLE** \equiv **OLS**.

Likelihood

If the errors are iid and normally distributed, then

$$\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}).$$

So the likelihood is

$$L = \frac{1}{\sigma^n (2\pi)^{n/2}} \exp \left(-\frac{1}{2\sigma^2} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \right)$$

which is maximized when $(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$ is minimized.

So **MLE** \equiv **OLS**.

Likelihood

If the errors are iid and normally distributed, then

$$\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}).$$

So the likelihood is

$$L = \frac{1}{\sigma^n (2\pi)^{n/2}} \exp \left(-\frac{1}{2\sigma^2} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \right)$$

which is maximized when $(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$ is minimized.

So MLE \equiv OLS.

Multiple regression predictions

Optimal predictions

$$\hat{Y}^* = E(Y^* | \mathbf{X}^*, \mathbf{Y}, \mathbf{X}) = \mathbf{X}^* \hat{\beta} = \mathbf{X}^* (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y}$$

where \mathbf{X}^* is a row vector containing the values of the regressors for the predictions (in the same format as \mathbf{X}).

Prediction variance

$$\text{Var}(Y^* | \mathbf{X}^*, \mathbf{Y}, \mathbf{X}) = \sigma^2 [1 + \mathbf{X}^* (\mathbf{X}'\mathbf{X})^{-1} (\mathbf{X}^*)']$$

- This ignores any errors in \mathbf{X}^* .

Multiple regression predictions

Optimal predictions

$$\hat{Y}^* = E(Y^* | \mathbf{X}^*, \mathbf{Y}, \mathbf{X}) = \mathbf{X}^* \hat{\beta} = \mathbf{X}^* (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y}$$

where \mathbf{X}^* is a row vector containing the values of the regressors for the predictions (in the same format as \mathbf{X}).

Prediction variance

$$\text{Var}(Y^* | \mathbf{X}^*, \mathbf{Y}, \mathbf{X}) = \sigma^2 [1 + \mathbf{X}^* (\mathbf{X}'\mathbf{X})^{-1} (\mathbf{X}^*)']$$

- This ignores any errors in \mathbf{X}^* .
- 95% prediction intervals assuming normal errors: $\hat{Y}^* \pm 1.96 \sqrt{\text{Var}(Y^* | \mathbf{Y}, \mathbf{X}, \mathbf{X}^*)}$.

Multiple regression predictions

Optimal predictions

$$\hat{Y}^* = E(Y^* | \mathbf{X}^*, \mathbf{Y}, \mathbf{X}) = \mathbf{X}^* \hat{\beta} = \mathbf{X}^* (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y}$$

where \mathbf{X}^* is a row vector containing the values of the regressors for the predictions (in the same format as \mathbf{X}).

Prediction variance

$$\text{Var}(Y^* | \mathbf{X}^*, \mathbf{Y}, \mathbf{X}) = \sigma^2 [1 + \mathbf{X}^* (\mathbf{X}'\mathbf{X})^{-1} (\mathbf{X}^*)']$$

- This ignores any errors in \mathbf{X}^* .
- 95% prediction intervals assuming normal errors: $\hat{Y}^* \pm 1.96 \sqrt{\text{Var}(Y^* | \mathbf{Y}, \mathbf{X}, \mathbf{X}^*)}$.

Multiple regression predictions

Optimal predictions

$$\hat{Y}^* = E(Y^* | \mathbf{X}^*, \mathbf{Y}, \mathbf{X}) = \mathbf{X}^* \hat{\beta} = \mathbf{X}^* (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y}$$

where \mathbf{X}^* is a row vector containing the values of the regressors for the predictions (in the same format as \mathbf{X}).

Prediction variance

$$\text{Var}(Y^* | \mathbf{X}^*, \mathbf{Y}, \mathbf{X}) = \sigma^2 [1 + \mathbf{X}^* (\mathbf{X}'\mathbf{X})^{-1} (\mathbf{X}^*)']$$

- This ignores any errors in \mathbf{X}^* .
- 95% prediction intervals assuming normal errors: $\hat{Y}^* \pm 1.96 \sqrt{\text{Var}(Y^* | \mathbf{Y}, \mathbf{X}, \mathbf{X}^*)}$.