Regression Exercise

Christopher Nowzohour

09.04.2014

Regression: Line Fitting

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$\mathbf{y}$	$(n \times 1)$-vector of observations of dependent variable
X	$(n \times p)$-matrix of observations of independent variables (one column per variable, first columnt constant)
$\boldsymbol{\beta}$	$(p \times 1)$-vector of parameters
$\boldsymbol{\epsilon}$	$(n \times 1)$-vector of errors

Regression: Line Fitting

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$\mathbf{y}$	$(n \times 1)$-vector of observations of dependent variable
X	$(n \times p)$-matrix of observations of independent variables (one column per variable, first columnt constant)
$\boldsymbol{\beta}$	$(p \times 1)$-vector of parameters
$\boldsymbol{\epsilon}$	$(n \times 1)$-vector of errors

Goals:

1. Prediction: Accurately predict $\mathbf{y}$ for new X

Regression: Line Fitting

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$\mathbf{y}$	$(n \times 1)$-vector of observations of dependent variable
X	$(n \times p)$-matrix of observations of independent variables (one column per variable, first columnt constant)
$\boldsymbol{\beta}$	$(p \times 1)$-vector of parameters
$\boldsymbol{\epsilon}$	$(n \times 1)$-vector of errors

Goals:

1. Prediction: Accurately predict $\mathbf{y}$ for new X
2. Statistical Inference: How confident are we about the parameter values $\boldsymbol{\beta}$?

Regression: Line Fitting

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$\mathbf{y}$	$(n \times 1)$-vector of observations of dependent variable
X	$(n \times p)$-matrix of observations of independent variables (one column per variable, first columnt constant)
$\boldsymbol{\beta}$	$(p \times 1)$-vector of parameters
$\boldsymbol{\epsilon}$	$(n \times 1)$-vector of errors

Goals:

1. Prediction: Accurately predict $\mathbf{y}$ for new X
2. Statistical Inference: How confident are we about the parameter values $\boldsymbol{\beta}$?
3. Causal Inference: Can we change $\mathbf{y}$ by changing X?

Regression: Line Fitting

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$\mathbf{y}$	$(n \times 1)$-vector of observations of dependent variable
X	$(n \times p)$-matrix of observations of independent variables (one column per variable, first columnt constant)
$\boldsymbol{\beta}$	$(p \times 1)$-vector of parameters
$\boldsymbol{\epsilon}$	$(n \times 1)$-vector of errors

Goals:

1. Prediction: Accurately predict $\mathbf{y}$ for new X
2. Statistical Inference: How confident are we about the parameter values $\boldsymbol{\beta}$?
3. Causal Inference: Can we change $\mathbf{y}$ by changing X?
   - Careful – need extra assumptions to make causal statements (e.g. no hidden variables, known causal direction)

Regression: Line Fitting

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$\mathbf{y}$	$(n \times 1)$-vector of observations of dependent variable
X	$(n \times p)$-matrix of observations of independent variables (one column per variable, first columnt constant)
$\boldsymbol{\beta}$	$(p \times 1)$-vector of parameters
$\boldsymbol{\epsilon}$	$(n \times 1)$-vector of errors

Goals:

1. Prediction: Accurately predict $\mathbf{y}$ for new X
2. Statistical Inference: How confident are we about the parameter values $\boldsymbol{\beta}$?
3. Causal Inference: Can we change $\mathbf{y}$ by changing X?
   - Careful – need extra assumptions to make causal statements (e.g. no hidden variables, known causal direction)
   - Otherwise: Confounding, Simpson's Paradox, ...

Fitting criteria: three examples

What are "good" parameter estimates $\widehat{\beta}$?

Fitting criteria: three examples

What are "good" parameter estimates $\widehat{\boldsymbol{\beta}}$?

1. Small squared residuals (L^2 regression / least squares):

$$\widehat{\boldsymbol{\beta}}_{L^2} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2 = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i \cdot \boldsymbol{\beta})^2$$

Fitting criteria: three examples

What are "good" parameter estimates $\widehat{\boldsymbol{\beta}}$?

1. Small squared residuals (L^2 regression / least squares):

$$\widehat{\boldsymbol{\beta}}_{L^2} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2 = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i \cdot \boldsymbol{\beta})^2$$

2. Small absolute residuals (L^1 regression / robust regression):

$$\widehat{\boldsymbol{\beta}}_{L^1} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - X\boldsymbol{\beta}\|_1 = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} |y_i - \mathbf{x}_i \cdot \boldsymbol{\beta}|$$

Fitting criteria: three examples

What are "good" parameter estimates $\widehat{\boldsymbol{\beta}}$?

1. Small squared residuals (L^2 regression / least squares):

$$\widehat{\boldsymbol{\beta}}_{L^2} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - X\boldsymbol{\beta}\|_2^2 = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i \cdot \boldsymbol{\beta})^2$$

2. Small absolute residuals (L^1 regression / robust regression):

$$\widehat{\boldsymbol{\beta}}_{L^1} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - X\boldsymbol{\beta}\|_1 = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} |y_i - \mathbf{x}_i \cdot \boldsymbol{\beta}|$$

3. Maximum likelihood:

$$\widehat{\boldsymbol{\beta}}_{ML} = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} \log f_\epsilon(y_i - \mathbf{x}_i \cdot \boldsymbol{\beta})$$

Finding optimal parameters $\widehat{\beta}$

1. Small squared residuals (L^2 regression / least squares):

Finding optimal parameters $\widehat{\beta}$

1. Small squared residuals (L^2 regression / least squares):

$$\nabla \|\mathbf{y} - X\widehat{\boldsymbol{\beta}}_{L^2}\|_2^2 = -2X^T(\mathbf{y} - X\widehat{\boldsymbol{\beta}}_{L^2}) \stackrel{!}{=} \mathbf{0}$$

Hence $\widehat{\boldsymbol{\beta}}_{L^2} = (X^T X)^{-1} X^T \mathbf{y}$

Finding optimal parameters $\widehat{\beta}$

1. Small squared residuals (L^2 regression / least squares):

$$\nabla \|\mathbf{y} - X\widehat{\beta}_{L^2}\|_2^2 = -2X^T(\mathbf{y} - X\widehat{\beta}_{L^2}) \stackrel{!}{=} \mathbf{0}$$

   Hence $\widehat{\beta}_{L^2} = (X^T X)^{-1} X^T \mathbf{y}$

2. Small absolute residuals (L^1 regression / robust regression):

Finding optimal parameters $\widehat{\beta}$

1. Small squared residuals (L^2 regression / least squares):

$$\nabla \|\mathbf{y} - X\widehat{\boldsymbol{\beta}}_{L^2}\|_2^2 = -2X^T(\mathbf{y} - X\widehat{\boldsymbol{\beta}}_{L^2}) \overset{!}{=} \mathbf{0}$$

Hence $\widehat{\boldsymbol{\beta}}_{L^2} = (X^TX)^{-1}X^T\mathbf{y}$

2. Small absolute residuals (L^1 regression / robust regression):
   - No analytic solution possible :-(
   - But numerical optimization works in practice (e.g. gradient descent)

Finding optimal parameters $\widehat{\beta}$

1. Small squared residuals (L^2 regression / least squares):

$$\nabla \|\mathbf{y} - X\widehat{\beta}_{L^2}\|_2^2 = -2X^T(\mathbf{y} - X\widehat{\beta}_{L^2}) \stackrel{!}{=} \mathbf{0}$$

   Hence $\widehat{\beta}_{L^2} = (X^T X)^{-1} X^T \mathbf{y}$

2. Small absolute residuals (L^1 regression / robust regression):
   - No analytic solution possible :-(
   - But numerical optimization works in practice (e.g. gradient descent)

3. Maximum likelihood:

Finding optimal parameters $\widehat{\boldsymbol{\beta}}$

1. Small squared residuals (L^2 regression / least squares):

$$\nabla \|\mathbf{y} - X\widehat{\boldsymbol{\beta}}_{L^2}\|_2^2 = -2X^T(\mathbf{y} - X\widehat{\boldsymbol{\beta}}_{L^2}) \stackrel{!}{=} \mathbf{0}$$

   Hence $\widehat{\boldsymbol{\beta}}_{L^2} = (X^T X)^{-1} X^T \mathbf{y}$

2. Small absolute residuals (L^1 regression / robust regression):
   - No analytic solution possible :-(
   - But numerical optimization works in practice (e.g. gradient descent)

3. Maximum likelihood:
   - If $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 I_{n \times n})$, for some $\sigma > 0$: $\widehat{\boldsymbol{\beta}}_{ML} = \widehat{\boldsymbol{\beta}}_{L^2}$!
   - In general: can be difficult ($\rightarrow$ numerical optimization)

Typical Assumptions

In descending order of importance:

Typical Assumptions

In descending order of importance:

1. Our sample $(X, \mathbf{y})$ is representative of the population

Typical Assumptions

In descending order of importance:

1. Our sample $(X, \mathbf{y})$ is representative of the population
2. X has full column rank ($n \geq p$ and no collinear predictors)

Typical Assumptions

In descending order of importance:

1. Our sample $(X, \mathbf{y})$ is representative of the population
2. X has full column rank ($n \geq p$ and no collinear predictors)
3. Unbiased errors: $E[\epsilon_i] = 0 \quad \forall i$

Typical Assumptions

In descending order of importance:

1. Our sample $(X, \mathbf{y})$ is representative of the population
2. X has full column rank ($n \geq p$ and no collinear predictors)
3. Unbiased errors: $E[\epsilon_i] = 0 \quad \forall i$
4. Uncorrelated errors: $E[\epsilon_i \epsilon_j] = 0 \quad \forall i, j \ (i \neq j)$

Typical Assumptions

In descending order of importance:

1. Our sample $(X, \mathbf{y})$ is representative of the population
2. X has full column rank ($n \geq p$ and no collinear predictors)
3. Unbiased errors: $E[\epsilon_i] = 0 \quad \forall i$
4. Uncorrelated errors: $E[\epsilon_i \epsilon_j] = 0 \quad \forall i, j \ (i \neq j)$
5. Exactly measured (but possibly still random) covariates X

Typical Assumptions

In descending order of importance:

1. Our sample $(X, \mathbf{y})$ is representative of the population
2. X has full column rank ($n \geq p$ and no collinear predictors)
3. Unbiased errors: $E[\epsilon_i] = 0 \quad \forall i$
4. Uncorrelated errors: $E[\epsilon_i \epsilon_j] = 0 \quad \forall i, j \ (i \neq j)$
5. Exactly measured (but possibly still random) covariates X
6. Constant error variance: $E[\epsilon_i^2] = \sigma^2 \quad \forall i$

Typical Assumptions

In descending order of importance:

1. Our sample $(X, \mathbf{y})$ is representative of the population
2. X has full column rank ($n \geq p$ and no collinear predictors)
3. Unbiased errors: $E[\epsilon_i] = 0 \quad \forall i$
4. Uncorrelated errors: $E[\epsilon_i \epsilon_j] = 0 \quad \forall i, j \ (i \neq j)$
5. Exactly measured (but possibly still random) covariates X
6. Constant error variance: $E[\epsilon_i^2] = \sigma^2 \quad \forall i$
7. Jointly Gaussian errors: $\boldsymbol{\epsilon} \sim \mathcal{N}$

Typical Assumptions

In descending order of importance:

1. Our sample $(X, \mathbf{y})$ is representative of the population
2. X has full column rank ($n \geq p$ and no collinear predictors)
3. Unbiased errors: $E[\epsilon_i] = 0 \quad \forall i$
4. Uncorrelated errors: $E[\epsilon_i \epsilon_j] = 0 \quad \forall i, j \ (i \neq j)$
5. Exactly measured (but possibly still random) covariates X
6. Constant error variance: $E[\epsilon_i^2] = \sigma^2 \quad \forall i$
7. Jointly Gaussian errors: $\boldsymbol{\epsilon} \sim \mathcal{N}$

Assumptions 3,4,6,7 are often summarized as $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 I_{n \times n})$

Properties of $\widehat{\boldsymbol{\beta}}_{L^2}$

If we have $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 I_{n \times n})$, then the following hold:

Properties of $\widehat{\boldsymbol{\beta}}_{L^2}$

If we have $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 I_{n\times n})$, then the following hold:

1. Unbiasedness: $E[\widehat{\boldsymbol{\beta}}_{L^2}] = \boldsymbol{\beta}$

Properties of $\widehat{\boldsymbol{\beta}}_{L^2}$

If we have $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 I_{n \times n})$, then the following hold:

1. Unbiasedness: $E[\widehat{\boldsymbol{\beta}}_{L^2}] = \boldsymbol{\beta}$

2. Minimal variance among all unbiased estimators (Gauss-Markov Theorem)

Properties of $\widehat{\boldsymbol{\beta}}_{L^2}$

If we have $\boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 I_{n \times n})$, then the following hold:

1. Unbiasedness: $E[\widehat{\boldsymbol{\beta}}_{L^2}] = \boldsymbol{\beta}$
2. Minimal variance among all unbiased estimators (Gauss-Markov Theorem)
3. $\widehat{\boldsymbol{\beta}}_{L^2} \sim \mathcal{N}_p(\boldsymbol{\beta}, \sigma^2 (X^T X)^{-1})$, and $\widehat{\boldsymbol{\beta}}_{L^2}$ is independent of $\widehat{\sigma}^2$
   - t-tests for components of $\widehat{\boldsymbol{\beta}}_{L^2}$ possible
   - F-test for the whole of $\widehat{\boldsymbol{\beta}}_{L^2}$ possible
   - Confidence interval for $E[y_0 | \mathbf{x}_0]$ and prediction interval for y_0 possible (where y_0 is a new observation at $\mathbf{x}_0$)

What happens if assumptions fail?

What happens if assumptions fail?

1. Non-representative sample: cannot infer about population

What happens if assumptions fail?

1. Non-representative sample: cannot infer about population
2. $X^T X$ non invertible: cannot compute $\widehat{\beta}_{L^2}$

What happens if assumptions fail?

1. Non-representative sample: cannot infer about population
2. $X^T X$ non invertible: cannot compute $\widehat{\boldsymbol{\beta}}_{L^2}$
3. Biased errors:
   - $\widehat{\boldsymbol{\beta}}_{L^2}$ will be biased
   - $\rightarrow$ Transformations? More predictors?

What happens if assumptions fail?

1. Non-representative sample: cannot infer about population
2. $X^T X$ non invertible: cannot compute $\widehat{\boldsymbol{\beta}}_{L^2}$
3. Biased errors:
   - $\widehat{\boldsymbol{\beta}}_{L^2}$ will be biased
   - $\rightarrow$ Transformations? More predictors?
4. Correlated errors:
   - Wrong p-values & confidence intervals
   - Estimator less precise (higher variance)
   - $\rightarrow$ Generalized Least Squares

What happens if assumptions fail?

1. Non-representative sample: cannot infer about population
2. $X^T X$ non invertible: cannot compute $\widehat{\boldsymbol{\beta}}_{L^2}$
3. Biased errors:
   - $\widehat{\boldsymbol{\beta}}_{L^2}$ will be biased
   - $\rightarrow$ Transformations? More predictors?
4. Correlated errors:
   - Wrong p-values & confidence intervals
   - Estimator less precise (higher variance)
   - $\rightarrow$ Generalized Least Squares
5. Noisy covariates: $\widehat{\boldsymbol{\beta}}_{L^2}$ will be biased

What happens if assumptions fail?

1. Non-representative sample: cannot infer about population
2. $X^T X$ non invertible: cannot compute $\widehat{\boldsymbol{\beta}}_{L^2}$
3. Biased errors:
   - $\widehat{\boldsymbol{\beta}}_{L^2}$ will be biased
   - $\rightarrow$ Transformations? More predictors?
4. Correlated errors:
   - Wrong p-values & confidence intervals
   - Estimator less precise (higher variance)
   - $\rightarrow$ Generalized Least Squares
5. Noisy covariates: $\widehat{\boldsymbol{\beta}}_{L^2}$ will be biased
6. Non-constant error variance:
   - Estimator less precise (higher variance)
   - $\rightarrow$ Generalized Least Squares, Transformations?

What happens if assumptions fail?

1. Non-representative sample: cannot infer about population
2. $X^T X$ non invertible: cannot compute $\widehat{\beta}_{L^2}$
3. Biased errors:
   - $\widehat{\beta}_{L^2}$ will be biased
   - $\rightarrow$ Transformations? More predictors?
4. Correlated errors:
   - Wrong p-values & confidence intervals
   - Estimator less precise (higher variance)
   - $\rightarrow$ Generalized Least Squares
5. Noisy covariates: $\widehat{\beta}_{L^2}$ will be biased
6. Non-constant error variance:
   - Estimator less precise (higher variance)
   - $\rightarrow$ Generalized Least Squares, Transformations?
7. Non-normal errors:
   - Only weak version of Gauss-Markov Theorem
   - $\widehat{\beta}_{L^2}$ is only approximately Gaussian (under weak assumptions on X), therefore slightly wrong p-values & confidence intervals
   - $\rightarrow$ Transformations?

Confidence and Prediction intervals / bands

Confidence and Prediction intervals / bands

95%-Confidence band: Area that includes true regression line $E[y|\mathbf{x}]$ with 95% probability.

Confidence and Prediction intervals / bands

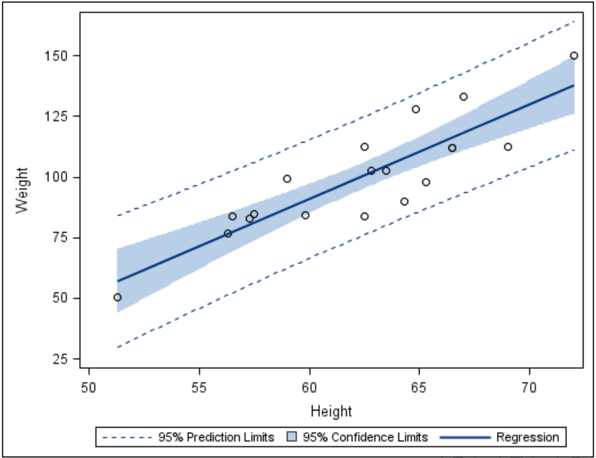
95%-Confidence band: Area that includes true regression line $E[y|\mathbf{x}]$ with 95% probability.

95%-Prediction band: Area that includes new observations $(X, \mathbf{y})$ with 95% probability.

Confidence and Prediction intervals / bands

95%-Confidence band: Area that includes true regression line $E[y|\mathbf{x}]$ with 95% probability.

95%-Prediction band: Area that includes new observations $(X, \mathbf{y})$ with 95% probability.



Diagnostic Plots

Diagnostic Plots

Tukey-Anscombe Plot: Residuals against fitted values

Diagnostic Plots

Tukey-Anscombe Plot: Residuals against fitted values

- Check for bias in errors

Diagnostic Plots

Tukey-Anscombe Plot: Residuals against fitted values

- Check for bias in errors
- Check for correlated errors

Diagnostic Plots

Tukey-Anscombe Plot: Residuals against fitted values

- Check for bias in errors
- Check for correlated errors
- Check for non-constant error variance

QQ-Plot: Theoretical Gaussian quantiles against empirical quantiles

Diagnostic Plots

Tukey-Anscombe Plot: Residuals against fitted values

- Check for bias in errors
- Check for correlated errors
- Check for non-constant error variance

QQ-Plot: Theoretical Gaussian quantiles against empirical quantiles

- Check for non-Gaussian errors