Linear Regression Assumptions

DATA SCIENCE BOOTCAMP

First, some notes about Covariance

Covariance

- It’s a measure of how two random variables change together.

Covariance

- It's a measure of how two random variables change together.

- The sign (+/-) shows the tendency of the linear relationship between X and Y

- The magnitude is harder to interpret.

Covariance (math)

Let's say X and Y are two random variables, where:

$E(X) = \mu_X$ and $E(Y) = \mu_Y$

The <u>covariance</u> between X and Y is:

$$
\begin{aligned}
Cov(X, Y) &= E[(X - \mu_X)(Y - \mu_Y)] \\
&= E(XY) - \mu_X \mu_Y \\
&= \sigma_{XY}
\end{aligned}
$$

Covariance, normalized

- Pearson's correlation coefficient

$$\rho_{X,Y} = \frac{\mathrm{cov}(X, Y)}{\sigma_X \sigma_Y}$$

- Here, magnitude shows strength of the linear relationship.
- $-1 \leq \rho_{X,Y} \leq 1$

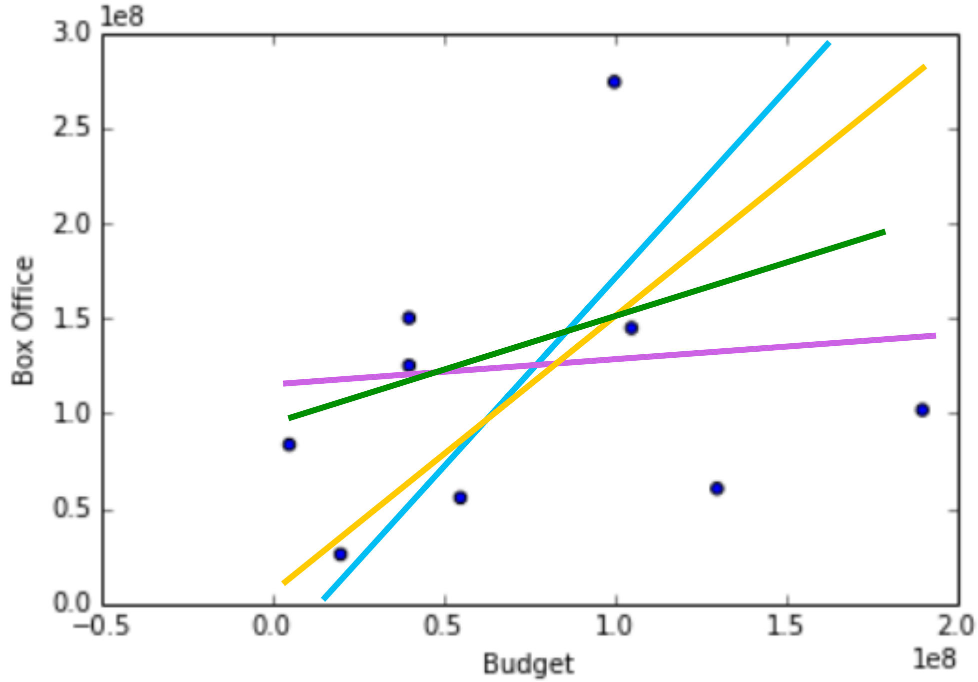
Covariance (more math facts)

- $\text{Cov}(X, X) = \text{Var}(X)$
- $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
- $\text{Cov}(X, aY) = a\text{Cov}(X, Y)$; a is any constant number
    - $\text{Var}(aX) = \text{Cov}(aX, aX)$
      $= a^2\text{Var}(X)$

Covariance

- If random variables X and Y are independent,
  Then $Cov(X, Y) = 0$

- BUT if $Cov(X, Y) = 0$, it *does not necessarily* mean that X and Y are independent!



$$y_\beta(x) = \beta_0 + \beta_1 x + \varepsilon$$

$\beta_0 = 80million$	$\beta_0 = 0$	$\beta_0 = 120million$	$\beta_0 = 30million$
$\beta_1 = 0.5$	$\beta_1 = 1.5$	$\beta_1 = 0.1$	$\beta_1 = 2$

Classical Assumptions of Ordinary Least Squares

1. Linear in parameters
2. Identifiability / No exact pairwise collinearity
   / No exact multicollinearity
3. Either: the covariates (X_i's) are fixed, OR,
   if X_i's are random variables, then X_i's are independent of ε
   i.e.: $Cov(X_1, \varepsilon) = Cov(X_2, \varepsilon) = \ldots = Cov(X_p, \varepsilon) = 0$
4. Number of observations > number of β parameters
5. Sufficient variation in the values of the X variables
6. Errors ε are normally distributed
7. Mean of the errors ε is 0
   i.e.: $E(\varepsilon) = 0$
8. Homoskedasticity. $Var(\varepsilon_i) = x^2$ for all i observations
9. No autocorrelation / no serial correlation
   i.e.: $Cov(\varepsilon_i, \varepsilon_j) = 0$ for any $i \neq j$
10. The model is correctly specified.

Classical Assumptions of Ordinary Least Squares

1. Linear in parameters
2. Identifiability / No exact pairwise collinearity
   / No exact multicollinearity
3. Either: the covariates (X_i's) are fixed, OR,
   if X_i's are random variables, then X_i's are independent of ε
   i.e.: $Cov(X_1, \varepsilon) = Cov(X_2, \varepsilon) = \ldots = Cov(X_p, \varepsilon) = 0$
4. Number of observations > number of β parameters
5. Sufficient variation in the values of the X variables
6. Errors ε are normally distributed
7. Mean of the errors ε is 0
   i.e.: $E(\varepsilon) = 0$
8. Homoskedasticity. $Var(\varepsilon_i) = x^2$ for all i observations
9. No autocorrelation / no serial correlation
   i.e.: $Cov(\varepsilon_i, \varepsilon_j) = 0$ for any $i \neq j$
10. The model is correctly specified.

Classical Assumptions of Ordinary Least Squares

If those assumptions highlighted in blue are true, then the OLS estimators $\hat{\beta}_0, \hat{\beta}_0, \ldots, \hat{\beta}_p$ are the

Best:

Linear

Unbiased:

Estimators

Classical Assumptions of Ordinary Least Squares

If those assumptions highlighted in blue are true, then the OLS estimators $\hat{\beta}_0$, $\hat{\beta}_0$, ..., $\hat{\beta}_p$ are the

Best: smallest variances among all linear unbiased estimators (*efficient*)

Linear

Unbiased:

Estimators

Classical Assumptions of Ordinary Least Squares

If those assumptions highlighted in blue are true, then the OLS estimators $\hat{\beta}_0, \hat{\beta}_0, \ldots, \hat{\beta}_p$ are the

Best: smallest variances among all linear unbiased estimators (*efficient*)

Linear

Unbiased: $E(\hat{\beta}_i) = \beta_i =$ for i from 1, 2, …, p

Estimators

Classical Assumptions of Ordinary Least Squares

Note: an unbiased estimator with the least variance is known as an *efficient* estimator.

Classical Assumptions of Ordinary Least Squares

Note: an unbiased estimator with the least variance is known as an *efficient* estimator.

i.e.: if you have an *efficient* estimator, you require the least amount of data to get an estimate $\hat{\beta}_i$ with reasonable variance.

Classical Assumptions of Ordinary Least Squares

Note: an unbiased estimator with the least variance is known as an *efficient* estimator.

i.e.: if you have an *efficient* estimator, you require the least amount of data to get an estimate $\hat{\beta}_i$ with reasonable variance.

If an estimate is *not efficient* (but still unbiased), you're still generally OK if you use enough data, i.e.: your estimate will be asymptotically correct.

1. Linear in Parameters

1. Linear in Parameters

Examples:

- (good): $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2^2 + \varepsilon$

1. Linear in Parameters

<u>Examples:</u>

- (good): $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2{}^2 + \varepsilon$
- (bad) Y: ranks of films, or categorical variable
    - Here, the underlying models are nonlinear
- (bad): $Y = \beta_0 + e^{\beta_1} X^{\beta_2} + \varepsilon$

1. Linear in Parameters

Examples:

- (good): $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2^{\,2} + \varepsilon$
- (bad) Y: ranks of films, or categorical variable
  - Here, the underlying models are nonlinear
- (bad): $Y = \beta_0 + e^{\beta_1} X^{\beta_2} + \varepsilon$

Test:

- Ask yourself: is Y numerical? Are you sure Y is not a rank?
- Try partial regressions and plots: $Y \sim X_i$, see if there's a linear relationship
- If all your standard errors are really big, might suspect nonlinearity
- (Assumption #8) residuals vs fitted plot: nonlinear

1. Linear in Parameters

Consequences:

- Estimates for $\hat{\beta}_0$ and their standard errors will be wrong, so predictions Y will be wrong.
- Whole model will be wrong.

1. Linear in Parameters

Consequences:

- Estimates for $\hat{\beta}_0$ and their standard errors will be wrong, so predictions Y will be wrong.
- Whole model will be wrong.

Remedies:

- Give up (try a nonlinear model)

2. Identifiability

a.k.a. No exact pairwise collinearity,

No exact multicollinearity

2. Identifiability

a.k.a. No exact pairwise collinearity, No exact multicollinearity

<u>Examples:</u>

- Exact multicollinearity:

  X_1 = production budget

  X_2 = announced budget (= 2 × production budget)

2. Identifiability

a.k.a. No exact pairwise collinearity,
No exact multicollinearity

Examples:

- Exact multicollinearity:
  X_1 = production budget
  X_2 = announced budget (= 2 × production budget)
- Near collinearity:
  X_1 = production budget
  X_2 = # opening theatres

2. Identifiability

a.k.a. No exact pairwise collinearity, No exact multicollinearity

<u>Examples:</u>

- Exact multicollinearity:
  - X_1 = production budget
  - X_2 = announced budget (= 2 × production budget)
- Near collinearity:
  - X_1 = production budget
  - X_2 = # opening theatres

<u>Test:</u>

- Check 1 versus 1 scatterplots of suspect pairs of X_i's
- High R^2 and significant F-statistic but mostly insignificant t-statistics

2. Identifiability

a.k.a. No exact pairwise collinearity, No exact multicollinearity

<u>Consequences:</u>

- $\hat{\beta}$ and $\hat{Y}$ estimates are still BLUE:
  - Still unbiased (best point estimates)
  - Still minimum possible variances

2. Identifiability

a.k.a. No exact pairwise collinearity, No exact multicollinearity

<u>Consequences:</u>

- $\hat{\beta}$ and $\hat{Y}$ estimates are still BLUE:
    - Still unbiased (best point estimates)
    - Still minimum possible variances
- BUT:
    - Large variances (large standard errors)

2. Identifiability

a.k.a. No exact pairwise collinearity, No exact multicollinearity

<u>Consequences:</u>

- $\hat{\beta}$ and $\hat{Y}$ estimates are still BLUE:
  - Still unbiased (best point estimates)
  - Still minimum possible variances
- BUT:
  - Large variances (large standard errors)
    - In perfect collinearity, standard errors would be infinite
  - Wide confidence intervals

2. Identifiability

a.k.a. No exact pairwise collinearity, No exact multicollinearity

Consequences (continued):

- t-tests tend to fail to reject the null (statistically insignificant covariates)
    - Thus you would be incorrectly concluding that covariates aren't related to Y, when in actuality, they are.
- Tiny changes in data → large differences in $\hat{\beta}$ and $\hat{Y}$

2. Identifiability

a.k.a. No exact pairwise collinearity,

No exact multicollinearity

Remedies:

- Feature selection; then see if standard errors get smaller
  - Regularize (Ridge/Lasso) – this gets rid of some of the overlapping
- Be careful: sometimes it's better to have near-collinearity than a loss in signal.

3. Either: all X's are fixed, OR
some X's are random, but independent of ε

3. Either: all X's are fixed, OR some X's are random, but independent of ε

Examples:

- (good) Store offering % discounts. Experimenting with sales revenue.
    - % discount levels (10%, 20%, 25%, etc.) are fixed.

3. Either: all X's are fixed, OR
some X's are random, but independent of ε

Examples:

- (good) Store offering % discounts. Experimenting with sales revenue.
  - % discount levels (10%, 20%, 25%, etc.) are fixed.
- (non-experimental) treat movie budget as a random variable; then movie budget must not be correlated with ε

3. Either: all X's are fixed, OR some X's are random, but independent of ε

Examples:

- (good) Store offering % discounts. Experimenting with sales revenue.
  - % discount levels (10%, 20%, 25%, etc.) are fixed.
- (non-experimental) treat movie budget as a random variable; then movie budget must not be correlated with ε

Test:

- Mostly you can assume the former. Else: don't worry about it.

3. Either: all X's are fixed, OR
some X's are random, but independent of ε

Consequences:
- Model may be mis-specified (see Assumption #10)

Remedies:
- Mostly you can assume X is fixed.
- Specify the model as best you can.
- Most importantly: be aware, but don't worry too much.

4. Number of Observations > Number of β Parameters

4. Number of Observations > Number of β Parameters

Examples:

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$
  - Number of data points > (p + 1)
- (bad) Fitting all possible X_i's, their interactions (1000 covariates), but only having 100 movies in your dataset

4. Number of Observations > Number of β Parameters

Examples:

- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$
  - Number of data points > (p + 1)
- (bad) Fitting all possible X_i's, their interactions (1000 covariates), but only having 100 movies in your dataset

Test:

- Count.

Consequences:

- Overfit

Remedies:

- Feature selection / Regularization
- Get more data

5. Sufficient variation in your covariates (X_i's)

5. Sufficient variation in your covariates (X_i's)

Examples:

- (bad) revenue = $\beta_0 + \beta_1$budget;

  but all the budget values in your dataset are $100 million

  - Then you try to predict revenues of films with $50,000 budget

5. Sufficient variation in your covariates (X_i's)

Examples:

- (bad) revenue = $\beta_0 + \beta_1$budget;

  but all the budget values in your dataset are $100 million

  - Then you try to predict revenues of films with $50,000 budget

Test:

- Look. (Be smart).

Consequences:

- Wrong about anything outside of your covariate (X_i) range.

Remedies:

- Don't.

Classical Assumptions of Ordinary Least Squares

1. Linear in parameters
2. Identifiability / No exact pairwise collinearity
   / No exact multicollinearity
3. Either: the covariates (X_i's) are fixed, OR,
   if X_i's are random variables, then X_i's are independent of ε
   i.e.: $Cov(X_1, \varepsilon) = Cov(X_2, \varepsilon) = \ldots = Cov(X_p, \varepsilon) = 0$
4. Number of observations > number of β parameters
5. Sufficient variation in the values of the X variables
6. Errors ε are normally distributed
7. Mean of the errors ε is 0
   i.e.: $E(\varepsilon) = 0$
8. Homoskedasticity. $Var(\varepsilon_i) = x^2$ for all i observations
9. No autocorrelation / no serial correlation
   i.e.: $Cov(\varepsilon_i, \varepsilon_j) = 0$ for any $i \neq j$
10. The model is correctly specified.

Classical Assumptions of Ordinary Least Squares

1. Linear in parameters
2. Identifiability / No exact pairwise collinearity
   / No exact multicollinearity
3. Either: the covariates (X_i's) are fixed, OR,
   if X_i's are random variables, then X_i's are independent of ε
   i.e.: $Cov(X_1, \varepsilon) = Cov(X_2, \varepsilon) = \ldots = Cov(X_p, \varepsilon) = 0$
4. Number of observations > number of β parameters
5. Sufficient variation in the values of the X variables
6. Errors ε are normally distributed
7. Mean of the errors ε is 0
   i.e.: $E(\varepsilon) = 0$
8. Homoskedasticity. $Var(\varepsilon_i) = x^2$ for all i observations
9. No autocorrelation / no serial correlation
   i.e.: $Cov(\varepsilon_i, \varepsilon_j) = 0$ for any $i \neq j$
10. The model is correctly specified.