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:

$$E(X) = \mu_X \quad \text{and} \quad E(Y) = \mu_Y$$

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

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

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

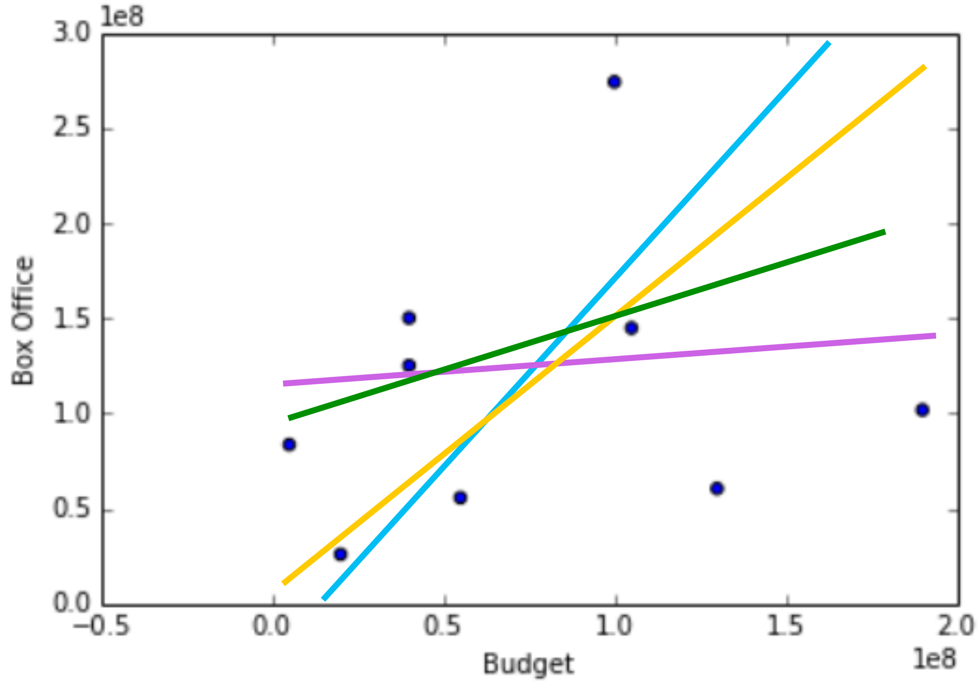
Covariance (more math facts)

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

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!



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

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

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

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): $Y = \beta_0 + e^{\beta_1} X^{\beta_2} + \varepsilon$

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

Examples:

- 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

Examples:

- Exact multicollinearity:

  X_1 = production budget

  X_2 = announced budget (= 2 × production budget)

- Near collinearity:

  X_1 = production budget

  X_2 = # opening theatres

Test:

- 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

Consequences:

- $\hat{\beta}$ 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

<u>Remedies:</u>

- 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

<u>Examples:</u>

- (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.

Linear Regression Assumptions: The Sequel

DATA SCIENCE BOOTCAMP

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.

6. Residuals (error terms, $\hat{\varepsilon}$) are normally distributed

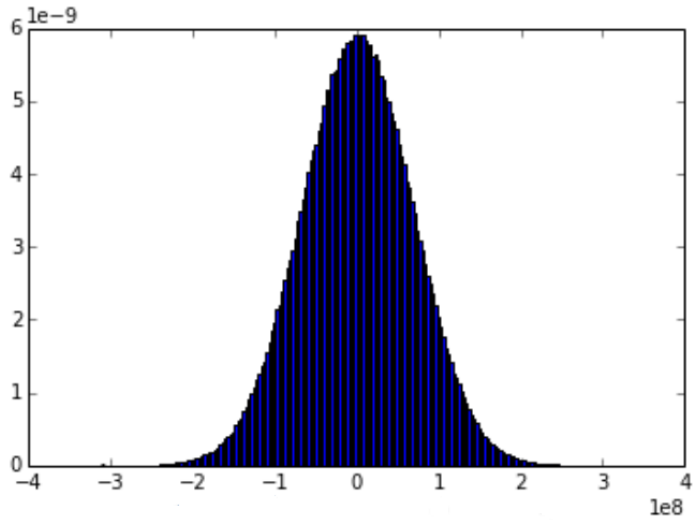
6. Residuals (error terms, $\hat{\varepsilon}$) are normally distributed

Examples:

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

- (good) residuals follow:

pdf (ε)



ε

(Note axes!)

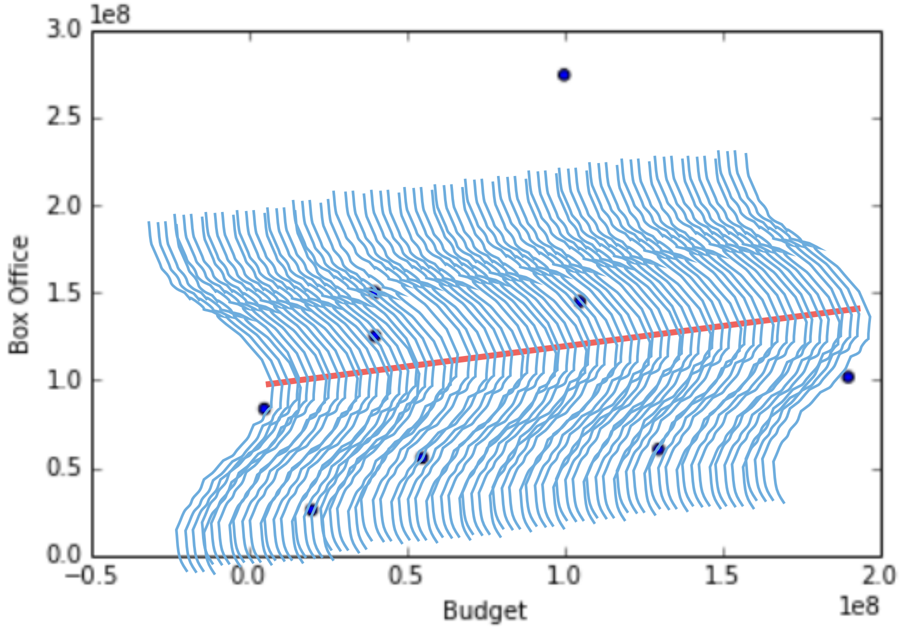
6. Residuals (error terms, $\hat{\varepsilon}$) are normally distributed

Examples:

- (good) residuals follow normal distribution:

$$\beta_0 = 94.68 million$$
$$\beta_1 = 0.1$$

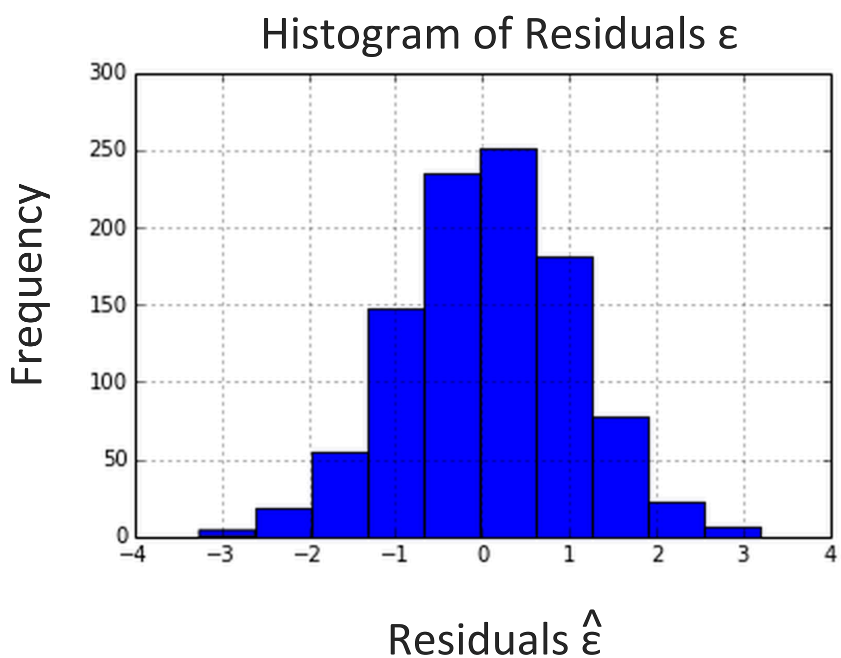
$$y_\beta(x) = \beta_0 + \beta_1 x + \varepsilon$$

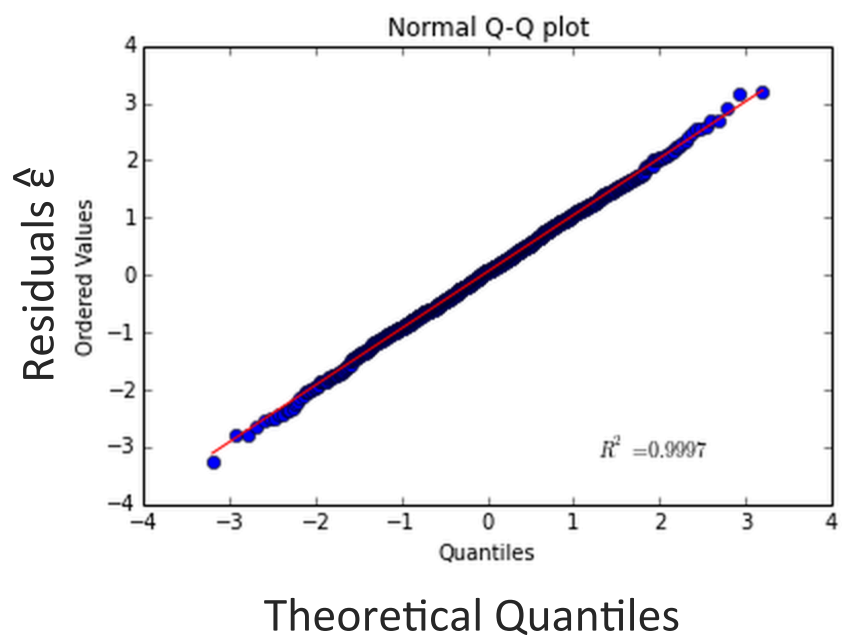


6. Residuals (error terms, $\hat{\varepsilon}$) are normally distributed

Tests:

- Normality tests (Jacques Bera / Omnibus)
- Q-Q plot (the residuals should follow the line)





6. Residuals (error terms, $\hat{\varepsilon}$) are normally distributed

Consequences

- $\hat{\beta}$ point estimates, $\hat{Y}$ point predictions still totally fine and BLUE (by the Central Limit Theorem)
- $\hat{\beta}$ would no longer follow t-distribution
- Standard errors and $\hat{Y}$ prediction intervals are slightly screwed up

6. Residuals (error terms, $\hat{\varepsilon}$) are normally distributed

<u>Consequences</u>

- $\hat{\beta}$ point estimates, $\hat{Y}$ point predictions still totally fine and BLUE (by the Central Limit Theorem)
- $\hat{\beta}$ would no longer follow t-distribution
- Standard errors and $\hat{Y}$ prediction intervals are slightly screwed up
  - Therefore t-tests, F-tests not valid, especially with small dataset
  - Confidence intervals for $\hat{\beta}$ not quite right
  - However, all t-tests, F-tests are *asymptotically correct*.

6. Residuals (error terms, $\hat{\varepsilon}$) are normally distributed

<u>Consequences</u>

- But with the Central Limit Theorem, error terms will tend toward the normal distribution anyways, as you get more data.

6. Residuals (error terms, $\hat{\varepsilon}$) are normally distributed

<u>Consequences</u>

- But with the Central Limit Theorem, error terms will tend toward the normal distribution anyways, as you get more data.

<u>Remedies</u>

- Mostly, as long as you have a lot of data, don't worry.
- Getting more data helps.
- Transformation of the response:
  - OLS on log(y) or f(y) generally
- Generalized Linear Models (later)

7. Mean of error terms $\hat{\varepsilon}$ is 0

i.e. $E(\hat{\varepsilon}) = 0$

7. Mean of error terms $\hat{\varepsilon}$ is 0

i.e. $E(\hat{\varepsilon}) = 0$

Examples

- (bad) Consistently underestimating movie gross

7. Mean of error terms $\hat{\varepsilon}$ is 0
i.e. $E(\hat{\varepsilon}) = 0$

Examples

- (bad) Consistently underestimating movie gross

Test:

- Can take the mean of the error terms, i.e. calculate $E(\hat{\varepsilon})$
- But really, don't worry

7. Mean of error terms $\hat{\varepsilon}$ is 0

i.e. $E(\hat{\varepsilon}) = 0$

Examples

- (bad) Consistently underestimating movie gross

Test:

- Can take the mean of the error terms, i.e. calculate $E(\varepsilon)$
- But really, don't worry

Consequences

- Intercept $\hat{\beta}_0$ estimate might lose some meaning.
- Everything else is totally fine.

7. Mean of error terms $\hat{\varepsilon}$ is 0
i.e. $E(\hat{\varepsilon}) = 0$

Examples

- (bad) Consistently underestimating movie gross

Test:

- Can take the mean of the error terms, i.e. calculate $E(\varepsilon)$
- But really, don't worry

Consequences

- Intercept $\hat{\beta}_0$ estimate might lose some meaning.
- Everything else is totally fine.

Remedies

- Don't worry. Perhaps be cautious not to trust $\hat{\beta}_0$ interpretation

8. Errors are homoskedastic

i.e. $Var(\varepsilon_i)$ is constant for all $i = 1, 2, \ldots, n$

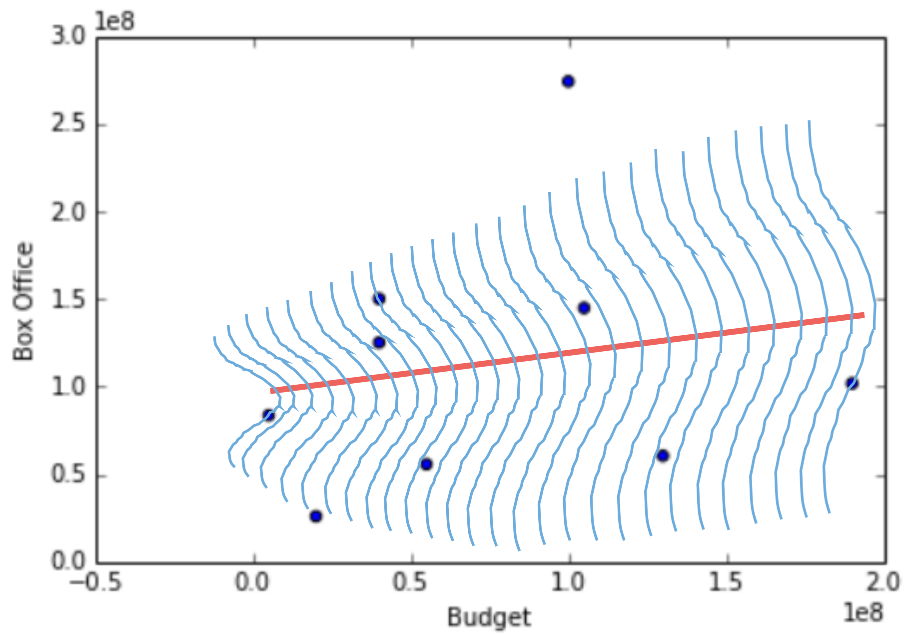
^

^

8. Errors are homoskedastic
i.e. Var(ε_i) is constant for all i = 1, 2, …, n

Examples

- (bad) Variance is different for different X values. As you invest more budget into a movie, there is more uncertainty in how it will perform

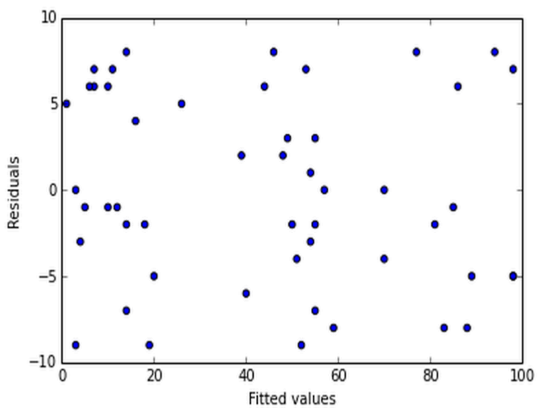


8. Errors are homoskedastic
i.e. Var(ε_i) is constant for all i = 1, 2, …, n

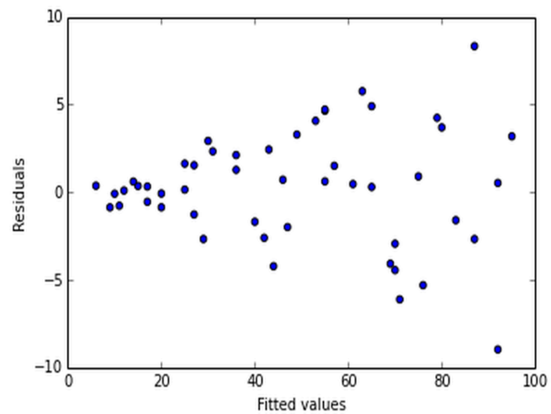
<u>Tests:</u>

- After fitting, plot the residuals $\hat{\varepsilon}$ versus fitted values $\hat{Y}$

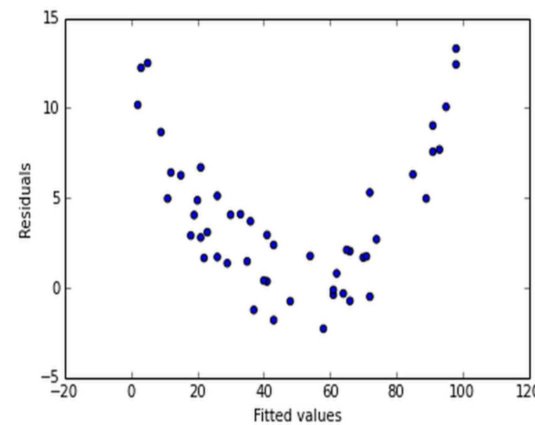
Good:



Heteroskedastic:



Nonlinear:



8. Errors are homoskedastic

i.e. $Var(\varepsilon_i)$ is constant for all $i = 1, 2, \ldots, n$

<u>Tests:</u>

- White-test:
  - Perform OLS ($Y \sim X$) as per usual; from this, get residuals $\hat{\varepsilon}_i$

8. Errors are homoskedastic

i.e. $Var(\varepsilon_i)$ is constant for all $i = 1, 2, \ldots, n$

<u>Tests:</u>

- White-test:
  - Perform OLS ($Y \sim X$) as per usual; from this, get residuals $\hat{\varepsilon}_i$
  - Perform a second Linear Regression:
    - $\hat{\varepsilon}_i \sim X$

8. Errors are homoskedastic

i.e. $Var(\varepsilon_i)$ is constant for all $i = 1, 2, \ldots, n$

<u>Tests:</u>

- White-test:
  - Perform OLS ($Y \sim X$) as per usual; from this, get residuals $\hat{\varepsilon}_i$
  - Perform a second Linear Regression:
    - $\hat{\varepsilon}_i \sim X$
    - Get those parameters: $\hat{\varepsilon}_i = \alpha_0 + \alpha_1 X_1 + \ldots + \alpha_p X_p + \varepsilon_i'$

8. Errors are homoskedastic
i.e. $Var(\varepsilon_i)$ is constant for all $i = 1, 2, \ldots, n$

Tests:

- White-test:
  - Perform OLS ($Y \sim X$) as per usual; from this, get residuals $\hat{\varepsilon}_i$
  - Perform a second Linear Regression:
    - $\hat{\varepsilon}_i \sim X$
    - Get those parameters: $\hat{\varepsilon}_i = \alpha_0 + \alpha_1 X_1 + \ldots + \alpha_p X_p + \varepsilon_i'$
    - The R^2 of this fit should be close to 0.

8. Errors are homoskedastic
i.e. $Var(\varepsilon_i)$ is constant for all $i = 1, 2, \ldots, n$

Tests:

- White-test:
  - Perform OLS ($Y \sim X$) as per usual; from this, get residuals $\hat{\varepsilon}_i$
  - Perform a second Linear Regression:
    - $\hat{\varepsilon}_i \sim X$
    - Get those parameters: $\hat{\varepsilon}_i = \alpha_0 + \alpha_1 X_1 + \ldots + \alpha_p X_p + \varepsilon_i'$
    - The R^2 of this fit should be close to 0.

- Statsmodels has an implementation of this test:

```
from statsmodels.stats.diagnostic import het_white
het_white(residuals, X)
```

8. Errors are homoskedastic

i.e. $Var(\varepsilon_i)$ is constant for all i = 1, 2, …, n

<u>Consequences:</u>

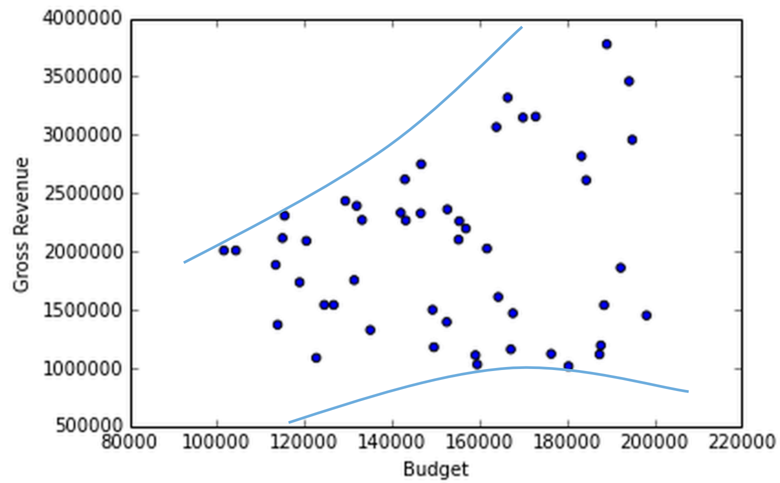
- $\hat{\beta}_0$ estimates are still unbiased
- But these estimates are no longer "best," i.e. "efficient"
  - Thus, standard errors are wrong.
  - Thus, t-tests, F-tests lose meaning for any inference

^

8. Errors are homoskedastic
i.e. $Var(\varepsilon_i)$ is constant for all i = 1, 2, ..., n

Remedies:

- Transformation of Y (i.e. this will also transform ε).
  - Example: $\log(Y) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon_i$
  - Then re-check for heteroskedasticity.
  - Remember: with log(Y), your prediction intervals no longer uniform:



8. Errors are homoskedastic
i.e. $Var(\varepsilon_i)$ is constant for all i = 1, 2, …, n

Remedies:

- (rare) if you know $Var(\varepsilon_i)$ exactly for each i

  → Weighted Least Squares (WLS)
- (if your dataset is large enough) use White's heteroskedasticity-consistent variances
  - a.k.a. "Sandwich estimators"
  - This will produce "robust standard errors"

8. Errors are homoskedastic
i.e. $Var(\varepsilon_i)$ is constant for all $i = 1, 2, \ldots, n$

Remedies:

- (rare) if you know $Var(\varepsilon_i)$ exactly for each i

  → Weighted Least Squares (WLS)
- (if your dataset is large enough) use White's heteroskedasticity-consistent variances
  - a.k.a. "Sandwich estimators"
  - This will produce "robust standard errors"
- If you only want the point estimates of β, then with large n (number of observations), don't worry too much. Just be aware of the caveats.

9. No Autocorrelation

i.e. $Cov(\varepsilon_{i,} \varepsilon_{j}) = 0$ for all I ≠ j

9. No Autocorrelation

i.e. $Cov(\varepsilon_{i,} \varepsilon_j) = 0$ for all $I \neq j$

<u>Example:</u>

- (bad) Movie gross time series
- (bad) Spatial correlation: snowfall over latitude/longitude

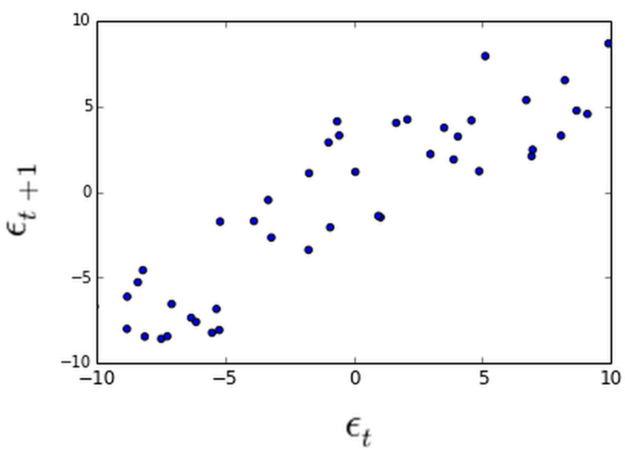
9. No Autocorrelation

i.e. $Cov(\varepsilon_{i,} \varepsilon_j) = 0$ for all I ≠ j

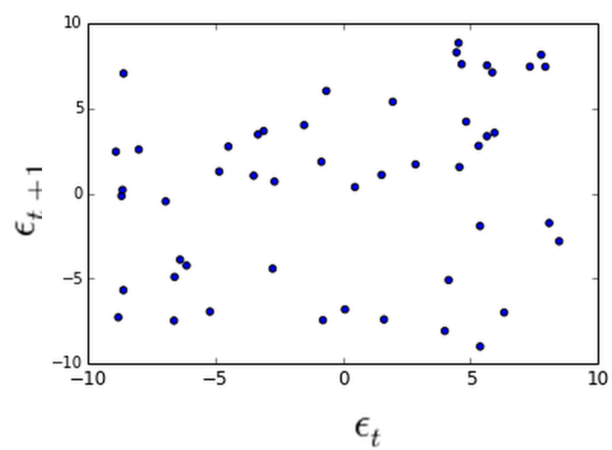
Tests:

- If possible, plot residuals over time:
  - Plot ε_t versus ε_{t+1}

Bad:



Good:



- Durbin-Watson test (in statsmodels summary already)

9. No Autocorrelation

i.e. $Cov(\varepsilon_{i,}\ \varepsilon_j) = 0$ for all I ≠ j

Consequences:

- $\hat{\beta}$, $\hat{Y}$ still unbiased, but:
- Standard errors of $\hat{\beta}$ will be wrong (generally they'll be big)
  - Thus confidence intervals will be wrong
- Variance of the residuals $\hat{\varepsilon}$, $\hat{\sigma}^2$, will be underestimated
  - Thus prediction intervals will be smaller than they should be, i.e: size of the prediction interval will be underestimated
  - Likely to overestimate R^2

9. No Autocorrelation

i.e. $Cov(\varepsilon_{i,}\ \varepsilon_{j}) = 0$ for all $I \neq j$

Remedies:

- Time series analysis examples:
  - Autoregressive model, AR(1):
    - $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$
  - Moving average model, MA(q):
    - Captures long term trends over time, with seasonality effects
- (rare) if you could identify how the errors change over time, could use Generalized Least Squares;
  - But this is rare, so don't worry about it.

10. Correctly Specified Model

10. Correctly Specified Model

<u>Example:</u>

- (not ideal) Using $Y = \beta_0 + \beta_1 X_1 + \varepsilon$, but the true underlying model is $Y = \beta_0 + \beta_1 X_1^2 + \beta_2 \log(X_2) + \beta_3 X_3 + \varepsilon$

10. Correctly Specified Model

Example:

- (not ideal) Using $Y = \beta_0 + \beta_1 X_1 + \varepsilon$, but the true underlying model is $Y = \beta_0 + \beta_1 X_1^2 + \beta_2 \log(X_2) + \beta_3 X_3 + \varepsilon$

Test:

- Practice the art of trying different model forms and measuring their performances (i.e. with R^2, cross-validation, AIC, etc.)

10. Correctly Specified Model

<u>Consequences:</u>

- Your model won't perform as well as you want.
- Might be under/over-fitting

10. Correctly Specified Model

Consequences:

- Your model won't perform as well as you want.
- Might be under/over-fitting

Remedies:

- This is the whole art of linear regression!
- Model selection via feature selection
- Try different models, compare them
- For different transformations of variables, make scatterplots