Correlation and Regression
ASSUMPTIONS

Assumptions in Linear Regression

- In order for the regression model to work as expected (minimum distance and maximum correlation), the data (and residuals must meet some specific assumptions:

1. The relationship between the DV and IV(s) is linear (or can be approximated by a straight line after some transformation)

Assumptions (cont)

2. The error term has a mean of zero
3. The error term has a constant variance
4. The errors are uncorrelated
5. The errors are normally distributed
6. If we add the assumption that the independent variables are fixed, then we also have as a consequence that the Least Squares method is the Best Linear Unbiased Estimator (BLUE)

Why is this important?

- If we use a model with faulty assumptions, the model is unstable:
  - A different sample might produce a very different model
- Problems in the assumptions cannot be detected just by looking at some statistics like F-tests, t-tests, R^2.
- We need to run some special tests to check that the assumptions are Ok

Exploratory Data Analysis (EDA; Hamilton, 1992)

- Checking normality on every variable on its own
  - Shapiro-Wilks and Lilliefors (SPSS): Shapiro-Wilks seems to be more powerful in detecting deviations from normality
  - M-estimators: Weight the values in a variable so they have less impact on the estimation procedure.
    - All common M-estimators assign weights so that they decrease as distance from the center of the distribution increases.
  - Four common estimators: Huber's, Tukey's biweight, Hampel's, and Andrew's. They are different in that they use different methods to assign weights

Exploratory Data Analysis (EDA)

- Check the correlation Matrix between all the variables involved in the model
- Use Graphical approaches to check correlations
- May also help detect some other potential problems

Residual analysis

Residual analysis

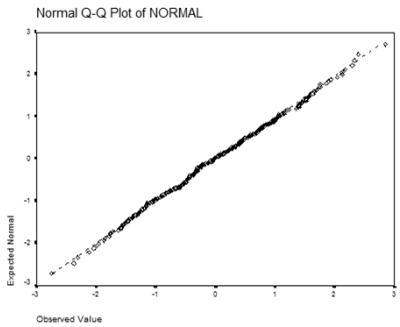
- We said that the residuals are basically composed of unknown variables plus error
  - By analyzing residuals, we might gain some understanding about what goes into the error term
- Residual analysis is more of an art, because the best indices are visual (you check graphs for problems)

Normal plots

- Some departure from normality is always expected
  - Mendenhall: Regression is robust to some non-normality
- However, serious departure from normality can be a concern, because all the statistical tests depend on the assumption of normality (F-tests and t-tests; confidence and prediction intervals)

Normal plots

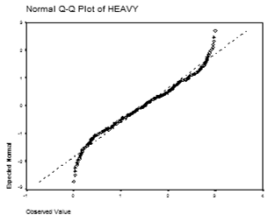
- The easiest way: plot the errors.
  - Stem-and-leaf graphs, Histograms, or Normal probability plots



Y axis: Cumulative value of a theoretical normal curve. X axis: ranked (from smaller to larger) residuals.

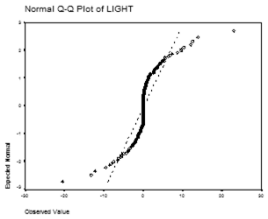
If the distribution is approximately normal, the plot will look pretty much like a straight line

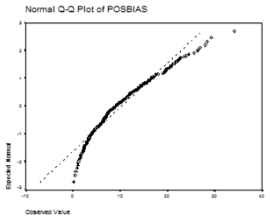
Example of non-normal distributions



Heavy tails: lots of outliers

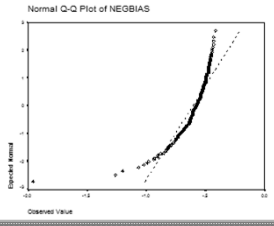
Light tails: few outliers



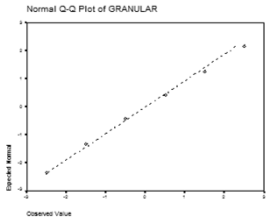


Positive bias: positive skew

Negative bias: negative skew

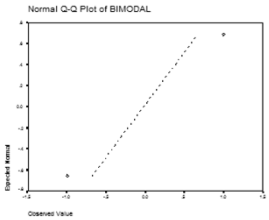


Example of non-normal distributions (cont)



Granular: non-continuous distribution

Bimodal: two observations

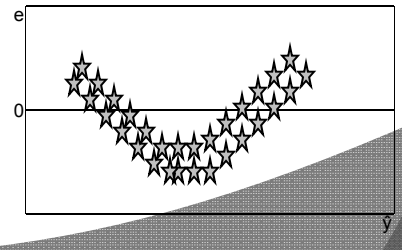


How non-normal is not normal?

- Not even data coming from a normal sample will always look normal
  - Daniel & Wood (1980): examples of normal probability plots from samples taken from normal distributions(different sizes, easier to compare our sample)
- 68-95-99 rule: If residuals normal, then about 68% ~ 1 sd; 95% ~ 2 sd, and 99% ~ 3 sd
  - Best to use standardized residuals

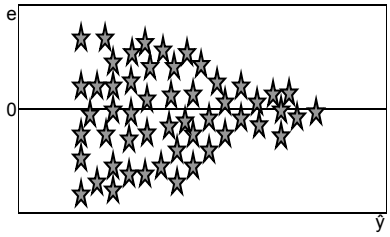
Residuals vs. ŷ

- Ideally, residuals should be in a band centered around 0, with no tendencies in the plot
- If they don't, it may mean problems:
  - Some sort of pattern? Maybe a non-linear relationship



Residuals vs. ŷ

- If you see a pattern where the width of the band thins-out or gets fatter, then it is pretty likely that you have a problem of heterogeneity of variance

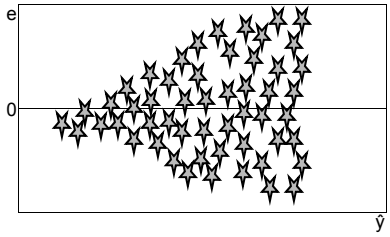


Residuals are positively skewed

This type of pattern can also be observed if the dependent variable is Poisson

Residuals vs. ŷ

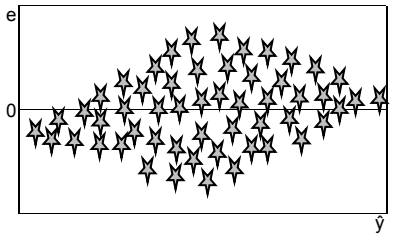
- If you see a pattern where the width of the band thins-out or gets fatter, then it is pretty likely that you have a problem of heterogeneity of variance



Residuals are negatively skewed

Residuals vs. ŷ

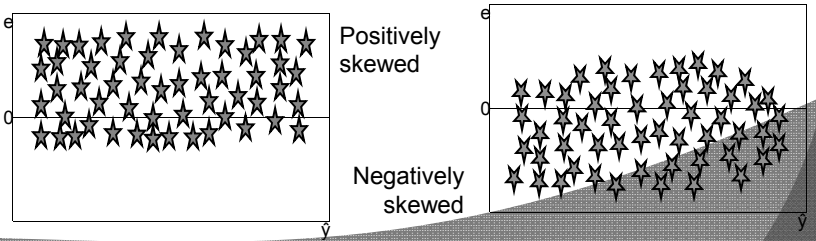
- If you see a pattern where the width of the band thins-out or gets fatter, then it is pretty likely that you have a problem of heterogeneity of variance



A fat section in the middle suggests that the DV is binomial

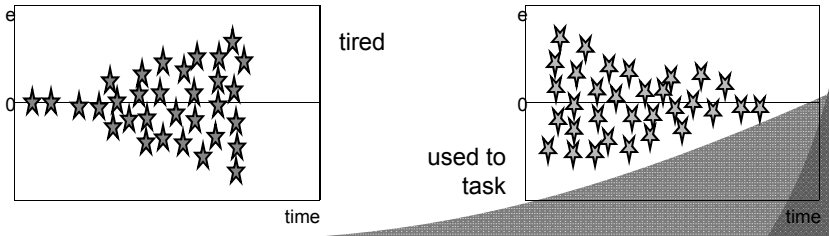
Residuals vs. ŷ

- Distribution of the residuals around 0 also tells you something about the shape of the distribution:
  - Positively skewed: more points above the mean (zero point)
  - Negatively skewed: more points above the mean (zero point)



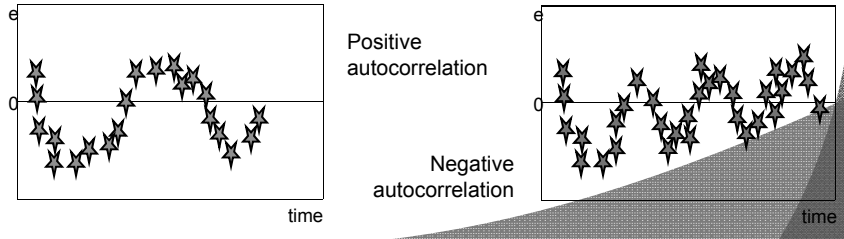
Residuals vs. time

- When order is important, plot of residuals versus order (time)
  - If we see some pattern (fans opening to right or left), the variance is changing with time (e.g., the subject might be getting tired, or getting used to the task)



Residuals vs. time

- If pattern that goes back and forth between positive and negative values, errors are correlated with those from the period before/after it
  - Slow-moving: positive autocorrelation
  - Fast-moving: negative autocorrelation



Lack of fit

- We can propose a miss-specified model, which may yield good statistics
  - A model where a curve will produce the best fit, can produce an adequate fit to a line
  - Significant F, significant t-test
- Residual analysis may help us detect this lack of fit
- We can also use partial regression plots
  - Partial residuals

Partial regression plots

- Estimate the best fit for the DV while holding the other variables constant
  - Remove the influence of all the variables (not including the target) from the DV (residuals)
  - Remove the influence of all the other variables (not including the DV) from the target (residuals)
  - The residuals from both models are plotted against each other

Statistical tests to check assumptions

Normality

- Chi-square data versus normal distribution
  - Overly sensitive when the sample size is large
- Values for mean, median, mode
  - The 68-95-99% rule
- If problems, robust models (weighted regression)

Heteroscedasticity

- The variance of the distribution is not equal (slides 15-17)
  - If you have only one IV, you can split the data in half (around the area where you see the problem), and fit two regression models
  - Find the Mean of Squares of the Error (MSE) for each model
  - Compute an F-test with the large MSE at the top:

Correlation with time

- When order is important and we suspect that the errors are correlated over time (slide 20)
  - Se(b) are under-estimated (we need to use a different tool: Time series analysis). But when will correlation become a problem?
- Durbin-Watson (DW): tests for residual correlation

$$DW = \frac{\sum(e_t - e_{t-1})^2}{\sum(e_t)^2}$$

  - Critical values between 0 and 4

Durbin-Watson test

- DW created tables to determine if the degree of correlation may affect se(b)
- Upper and Lower bounds for k (number of IV's) and N (number of subjects). Rule:
  - if DW < DL
    - Reject null hypothesis of no positive correlation among residuals
  - If (4 – DW) < DL
    - Reject the null hypothesis of no negative correlation among residuals

Potential solutions to some of these problems

Skewed distributions

- Try a transformation of the data (Hamilton's handout)
- q > 1: Reduce negative skewness. The higher the power, the higher the effect.
- q = 1: raw data: no transformation
- q < 1: Reduce positive skew. The lower the power, the stronger the pull.
  - lower-than-one powers are common, as well as logarithms.

Nonlinearity

- Plots of Residuals vs. ŷ show a pattern
- Partial regression plots will show some potential alternatives
  - Use the residual trick to try to find what fit may be best

Heteroscedasticity

- Try some variance-stabilization transformations
  - Fan opening to the right:

$$y_{transf} = \sqrt{y}$$
$$y_{transf} = \log(y)$$
$$y_{transf} = \frac{1}{\sqrt{y}}$$
$$y_{transf} = \frac{1}{y}$$

  - Distributions wide in the center: $y_{transf} = \arcsin(y)$
  - Fan opening to the left: can be associated with bias (skewness). Try some of the transformations described in Hamilton

Outliers

Outliers

- An extreme observation
  - Observations which are three or more standard deviations away from the rest of the data
- There are two important reasons to spend some time checking for their presence:
  1. These points are not typical of our data set. Therefore it is important to detect them, and then, do something about them

Outliers (cont)

2. It will be nice to have an explanation for them (typos, weird conditions in measurement)
   - Maybe very important measures to determine whether our model is doing a good job in explaining the data
   - An outlier may point out to some unusual circumstances where the data deviates from what the model says

- Influential observations are not outliers

Standardized and Studentized residuals

- Recommended for identification of outliers. Standardized residual:

$$zresid_i = \frac{e_i}{sd_{resid}\sqrt{1 - h_{i,i}}}$$

- Studentized residual:

$$tresid_i = \frac{e_i}{sd_{resid-i}\sqrt{1 - h_{i,i}}}$$

$sd_{resid\text{-}i}$: standard deviation of the residuals when we remove the i-th observation

  - We can interpret this statistic as the result of a t-test with (n-k-1) degrees of freedom to test the hypothesis that the observation is an outlier

Mahalanobis distance

- This statistic is associated with values in the IV.
  - Tries to determine if the values are *unusual*, compared with the rest of the observations
    - In the case of one IV, this distance is defined as the square of the standardized value:

$$Mahalanobis_i = \left(\frac{x_i - \bar{x}}{sd(x)}\right)^2$$

    - Values above 3 absolute standard deviations (i.e., 9), are very suspicious

Influential observations

Influential observations

- Have a strong effect on the values of the statistics (F-test, t-test, R^2 sd(e), se(b))
- One of the strategies to determine how influential the observations are
  - If the values of R^2, b's, sd(e), se(b) change when we remove some observation, then the observation must be influential

Cook distance

- Establishes the effect of the influential observation in all the independent variables at once
- Establishes whether the observation is influential, but for all the variables at the same time
  - Evaluates the changes in all the residuals when case "i" is omitted

Use when the suspicion that the influence of an observation may be spread over several IV's rather than one IV

$$CookD = \left(\frac{(y_i - \hat{y}_i)^2}{(k+1)MSE}\right)\left(\frac{h_{i,i}}{(1-h_{i,i})^2}\right)$$

"$h_{i,i}$" elements of the hat matrix:

$$(X)(X^TX)^{-1}(X^T)_{ii}$$

Check values > 1

Leverage

- How much a value on the IV is pulling the slope in its direction

$$(X)(X^TX)^{-1}(X^T)_{ii}$$ "$h_{i,i}$" elements of the hat matrix:

  - The farther an observation is from its mean (X from $\dot{X}$), the stronger its leverage
  - Check values > 0.5. More detailed:
    - $\max(h_{i,i}) \leq 0.2$ considered safe
    - $0.2 < \max(h_{i,i}) \leq 0.5$ considered risky
    - $\max(h_{i,i}) > 0.5$ avoid if possible

Leverage (cont)

- SPSS has what they call "centered leverage values", and they suggest that:
  - OK observations that are close to:

$$\frac{number\ of\ IV's}{n}$$

  - Check observations that are closer to:

$$\frac{2\ x\ number\ of\ IV's}{n}$$

DFBETA's

- DFBETAS provide a comparison of what the beta for every coefficient will be like if we remove each observation in turn:

$$DFbeta_{i,k} = \frac{b_k - b_k^{(i)}}{\frac{sd_{b(k)}}{sd_{resid}}}$$

  - We calculate as many slopes as observations
  - We are going to get as many of these DFbetas as IV's we have in the equation

DFBETA's (cont)

- Criteria:
  - If DFBETA > 0, case is pulling the slope up
  - If DFBETA < 0, case is pulling the slope down
- SPSS has both this version, and a Standardized version. Given that we don't have a good criterion, use standardized DFBETA's
  - Criteria: suspicious observations if:

$$Standardized\ DFbeta > \left|\frac{2}{\sqrt{n}}\right|$$

Deleted Residuals

- Calculate the residuals with and without every observation
  - If there is a big change in the residual between when the observation is in, and when the observation is out, we know that the observation must be influential

Deleted Residuals (cont)

- SPSS gives you both the DELETED RESIDUAL (DRESID)

$$deleted\ residual_i = \left(y_i - \hat{y}_{(i)}\right)$$

- And the Studentized deleted residual (DSRESID)

$$Studentized\ deleted\ residual_i = \frac{\left(y_i - \hat{y}_{(i)}\right)}{se(-i)}$$

Change in ŷ

- Calculate the predicted value (i.e., ŷ) for each value in the IV with and without each observation
  - If there is a big change in the fit between when the observation is in, and when the observation is out, we know that the observation must be influential

Change in ŷ (cont)

- SPSS gives you both the change in ŷ (DFFIT):

$$change\ in\ fit_j = \left(\hat{y}_i - \hat{y}_{(i)}\right)$$

- And the standardized fit (SDFFIT)

$$Standardized\ change\ in\ fit_j = \frac{\left(\hat{y}_i - \hat{y}_{(i)}\right)}{se(-i)}$$

Criteria: $std\ change\ fit > \left|\frac{2}{\sqrt{\frac{p}{n}}}\right|$

COVRATIO (SPSS manual)

- This measure is dependent on the fact that the model will produce a variance-covariance matrix, and the knowledge that the determinant of that matrix will be unique.
  - If an observation has a strong influence, removing that observation from the model will affect in a very specific way the determinant of the variance-covariance matrix

COVRATIO (cont)

- Creates a ratio between the two determinants (with and without observation included in the model), and compares it against the following criterion:

$$\left|\frac{VC_{without\ obs}}{VC_{with\ obs}}\right| - 1 > \frac{3 \times number\ of\ IV's}{n}$$

  - If the ratio is larger than the criterion, then the observation may be influential

What to do with influential observations?

- Report the study both with and without the influential observations.
  - Let the readers draw their own conclusion about how important the points might be
- Find an explanation for them
  - They might just be the most important points in the model you are trying to fit, because they represent "the odd man out“
- If located at the extremes, consider running some transformations on the data set
  - You might get more normal results if the data are transformed