CS555B1 Data Analysis and Visualization

Lecture 8

Regression Diagnostics

Kia Teymourian

Regression Diagnostics

- After fitting a regression model it is important to determine **whether all the necessary model assumptions are valid** before performing inference.

- If there are any violations, subsequent inferential procedures may be invalid resulting in faulty conclusions.

- **Regression Diagnostics** is an activity that we perform after fitting a regression model to **check if the model fits the data well and if the assumptions underlying the theory were met** in order to have confidence in the inferences we made from the regression model.

- These techniques often involve **visualizing** the fit of the regression model via **examination of the residuals** in order to identify data points that don’t seem to fit the trend and issues with violations of the assumptions of the regression model.

Residual Plots

- One of the most **powerful tools in regression diagnostics are residual plots** which help visualize how well a regression equation fits the sample data.

- **Residual plots** are scatterplots of the **regression residuals [plotted on the y-axis]** against the **explanatory variable [plotted on the x-axis])**.

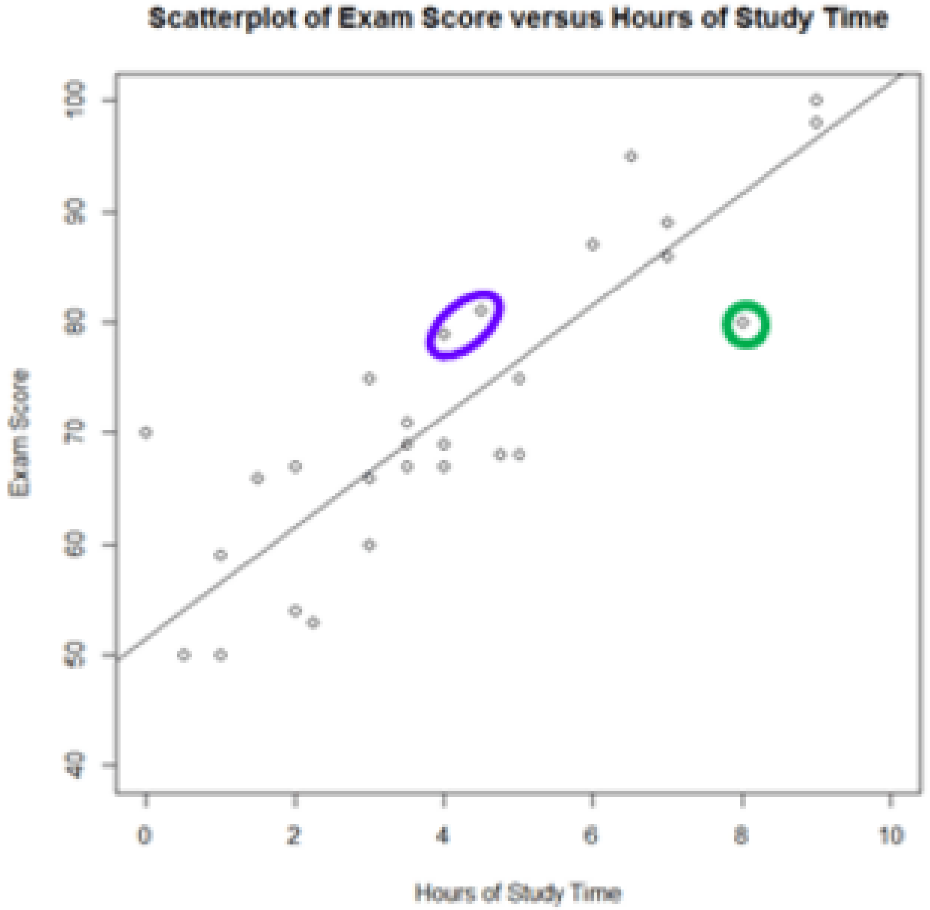
- The residual plot turns the regression line on the horizontal so it is easier to see patterns and pick out unusual observations.

- Residual plots can also **be generated using the predicted values on the x-axis as** opposed to the explanatory variable (which is especially helpful in **multiple regression analysis** when there are more than one explanatory variables).

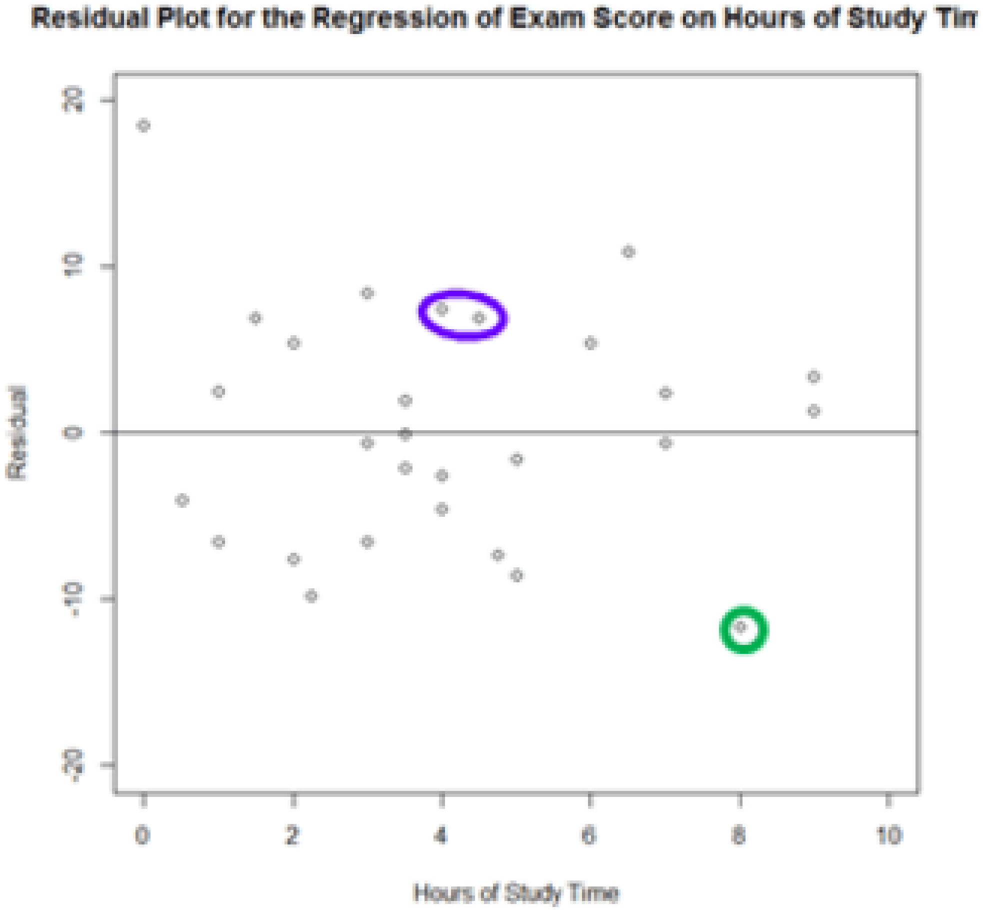
- Residual plots can also be generated **by plotting standardized or studentized residuals** (which involves dividing the residual by an estimate of the variability of the residuals).

Residual Plots

- The residual plot turns the regression line on the horizontal so it is easier to see patterns and pick out unusual observations.



Scatterplot of Exam Score versus Hours of Study Time



Residual Plot for the Regression of Exam Score on Hours of Study Tim

Residual Plots

```
> resid(m)
> par(mfrow=c(2,2))
> plot(fitted(m), resid(m), axes=TRUE, frame.plot=TRUE, xlab='fitted values',
ylab='residue')
> plot(age, resid(m), axes=TRUE, frame.plot=TRUE, xlab='age', ylab='residue')
> plot(height, resid(m), axes=TRUE, frame.plot=TRUE, xlab='height', ylab='residue')
> hist(resid(m))
```

fitted() is a generic function which extracts fitted values from objects returned by modeling functions.

Standardized and Studentized Residuals

The residuals of a model (e_i) can tell us a lot about the model fit.

We generally do not study the residuals themselves, we study a standardized form of the residuals:

$$e_i^* = \frac{e_i - \bar{e}}{\sqrt{MSE}}$$

Where e_i is the residual for observation *i* and *MSE* is the mean squared error of residuals.

If $\sqrt{MSE}$ were an estimate of the standard deviation of the residual , we call e_i^* a studentized residual.

Assumptions of the linear regression

There are **4 principal assumption**s which justify the use of linear regression models for purposes of inference or prediction:

1) Linearity and additivity of the relationship between dependent and independent variables:

- The expected value of dependent variable is a straight-line function of each independent variable, holding the others fixed.
- The slope of that line does not depend on the values of the other variables.
- The effects of different independent variables on the expected value of the dependent variable are additive.

2) The observations are independent.

3) The **variation of the response variable** around the regression line is **constant**.

4) Normality of the error distribution: the residuals are normally distributed.

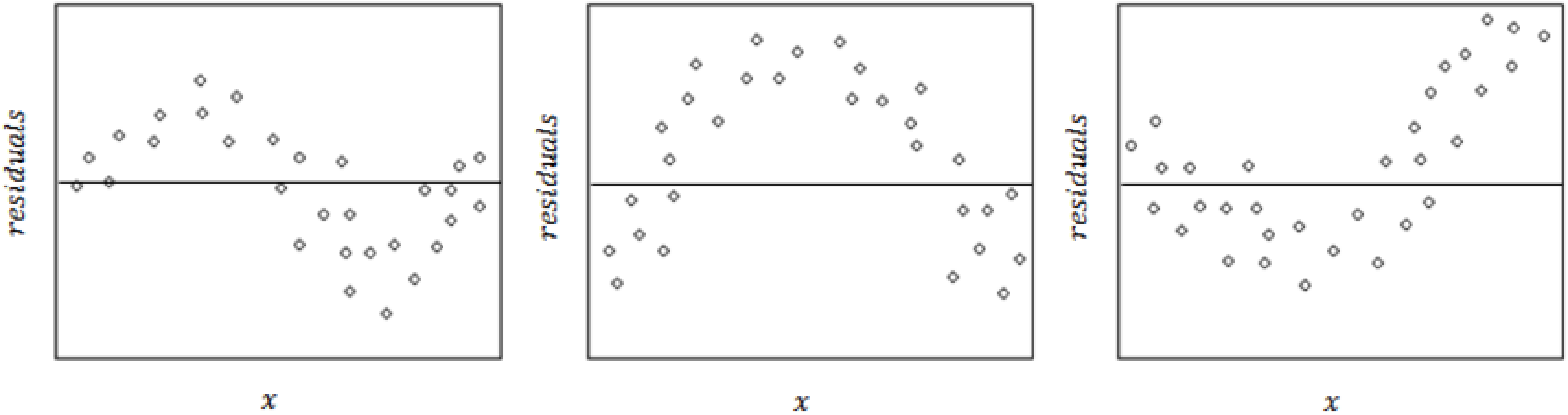
If any of these are violated, then the inference, prediction and interpretation of the regression equation (or correlation) are inefficient (at best) or misleading/biased/incorrect (at worst).

Linearity

We can use the scatterplot to show a roughly linear trend between factors. We should be cautious of curved or other non-linear relationships.

Residual plots can help us assess this assumption as they can magnify non-linearity.

Violations of linearity or additivity are extremely serious: if you fit a linear model to data which are non-linearly or non-additively related, your predictions are likely to be seriously in error.



Independence

In regression, we make the assumption that the observations are independent.

That is, we assume that if we are summarizing data on heights and weights in children, for example, that we only take one observation per child (as opposed to multiple observations of the same child over time or observations on various sets of identical twins).

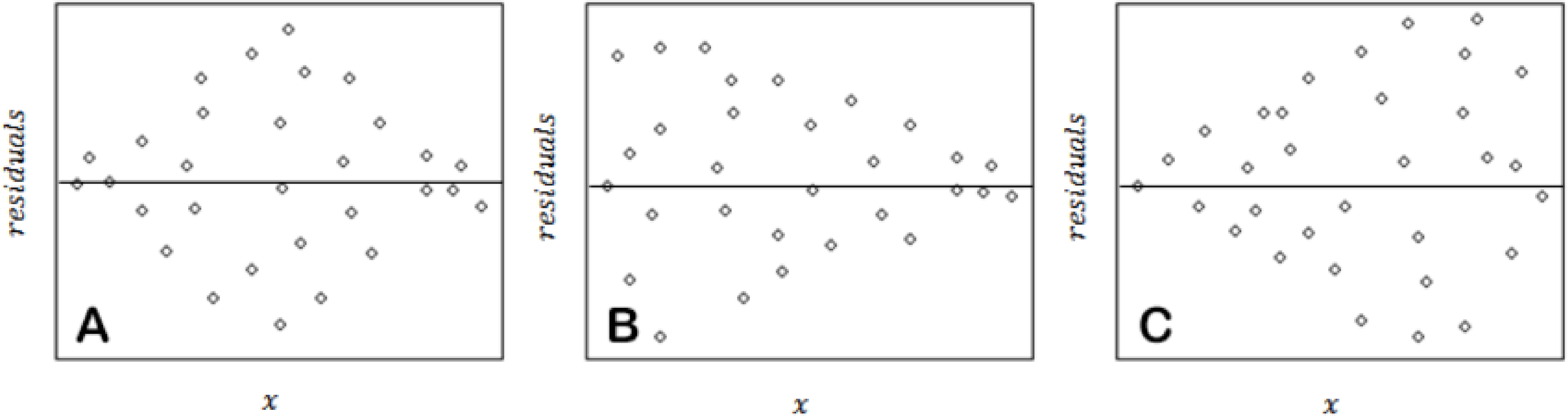
Observations on correlated data require more sophisticated analysis to account for the correlation between observations.

Constant Variance

We assume that the variability of the response is constant across the regression line.

This particular assumption can be checked via a residual plot. The residual plot should show approximately the same amount of scatter from left to right.

The figure below shows generic residual plots for regressions on associations that **violate the constant variance assumption**. In all of the plots, the **variability around of the residuals varies for different values of x.**

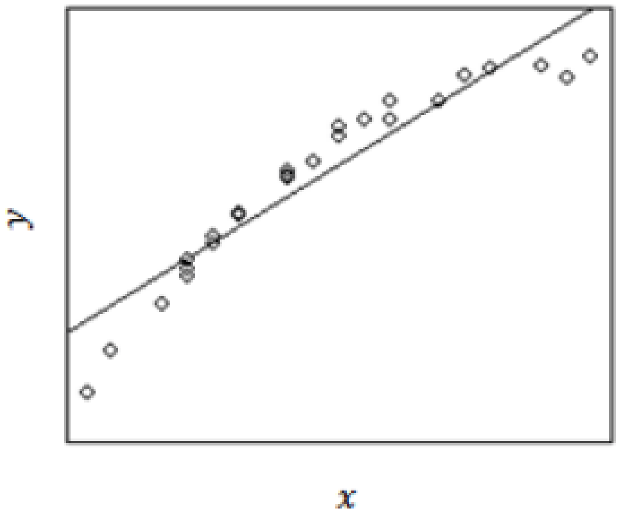


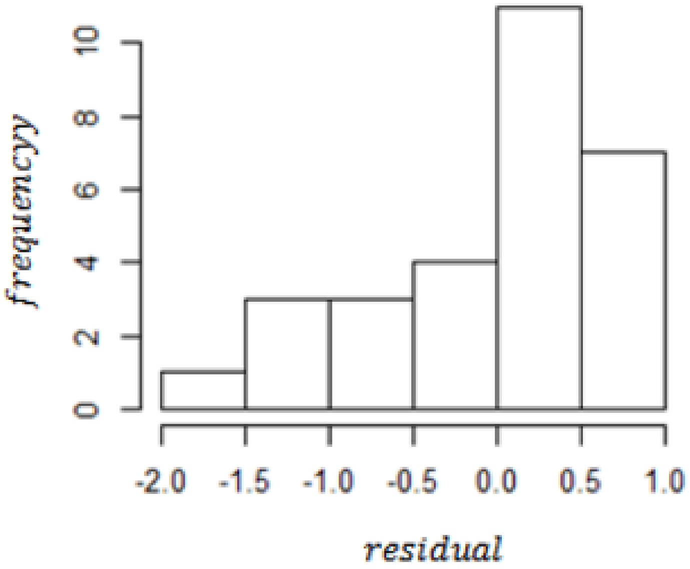
Normality

The residuals should **follow a normal distribution**. Severe deviations from this assumption could be due to outliers or non-normality of the explanatory or response variables.

Residuals may not be normally distributed if the linearity assumption has been violated.

Fortunately, inference is not as sensitive to departures from this assumption, especially when the number of observations is large.
To check this assumption, **histograms of the residuals can be used** to display the distribution of the residuals and ensure that they are approximately normal.





Regression Diagnostics

Check the assumptions:

- **Linearity**
- **Independence**
- **Constance variance**
- **Normally distributed residuals**

Residual Plots (to assess linearity and constant variance

> **plot**([variable for x-axis], **resid**(m))

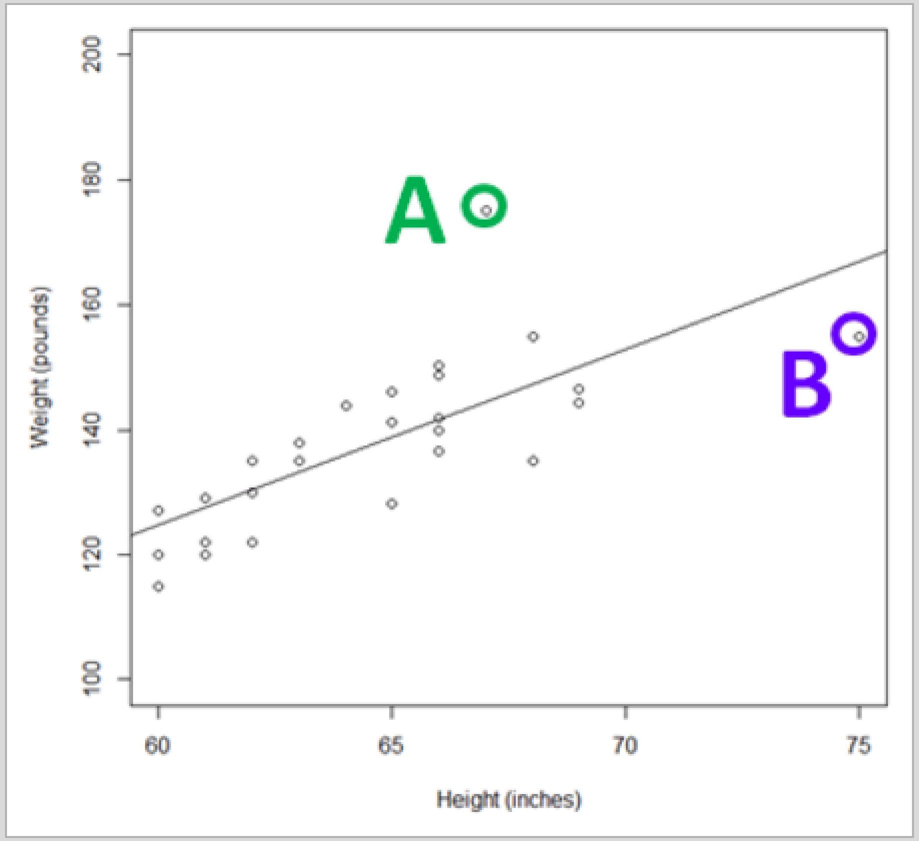
Check each explanatory variable, and the fitted values (**fitted**(m)) the linearity or variance.

Histograms (to check the distribution of the residuals)

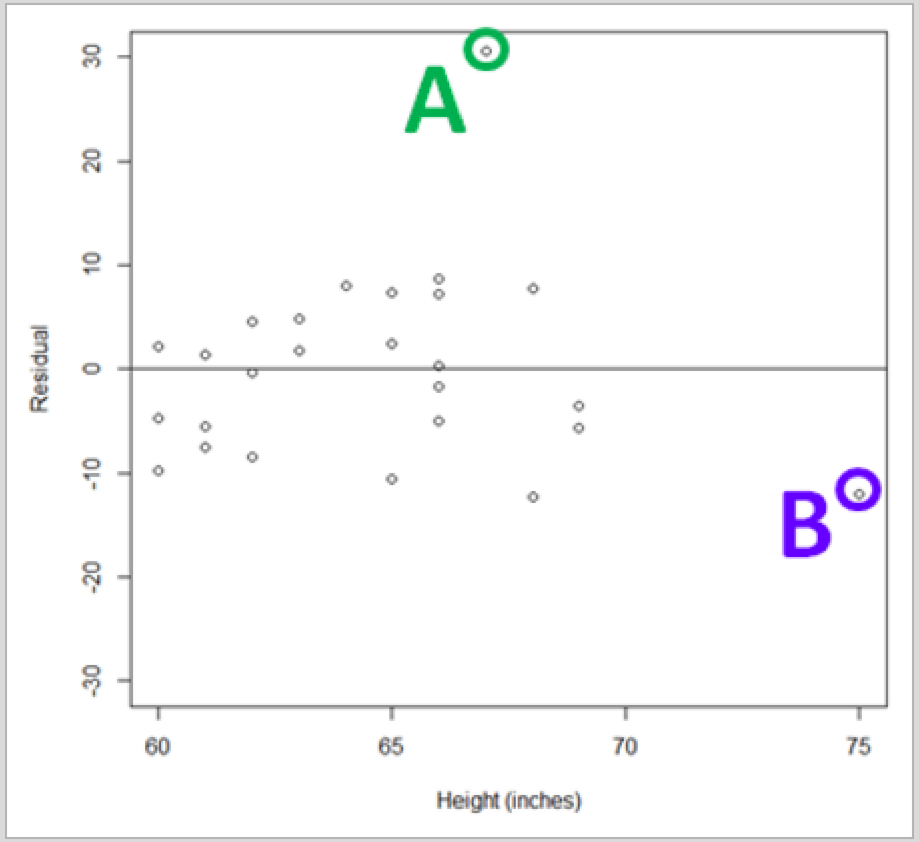
> **hist(resid(m))**

Outliers and Influence Points

- Outliers in the y-direction tend to have large residuals.
- Outliers in the x-direction may or may not have large residuals but have the potential to be influential.
- An influence point is an observation that markedly changes the result of the regression if it were to be removed from the calculation.



Scatterplot on Height and Weight



Residual Plot for the Regression of Weight on Height

Outliers and Influence Points

- Given that least-squares regression is based on **minimizing the squared vertical distances between the observations and the regression line**, extreme points in the **x-direction tend to pull the regression line close** to itself.
- In these cases, the regression line equation may be quite different with or without the points and thus the points influences the regression equation.
- The influence of a particular point should be examined by removing it from the regression calculation and checking how the equations, inference, and conclusions change with its removal.
- When there appears to be observations out of range, one should always check to ensure that there was not an issue with data entry/recording.
- If an outlier in the x-direction, for example, is kept, it may be desirable to collect additional data within the same range to better characterize the relationship and so that the regression doesn’t depend so heavily on the data from a single observation.

Transformations

- If any of the assumptions of regression are violated, **transformations** can often be applied which **may improve normality, linearity, and/or stabilize the variance**.

- If the variance of the response increases or decreases as the explanatory variable increases or decreases, respectively, then either the **natural log (ln) or the square root function can be applied to the response variable to help "stabilize" the variance**.

- If there is a non-linear relationship between factors, then sometimes **squaring the explanatory variable (or adding a squared term to a multiple linear regression model) can help**.

- **Finding the right transformation often involves some trial and error.**

- Due to the increased complexity of the interpretation when a transformation is applied, **no transformation is often preferred if the transformation only marginally improves the linearity or variance**.

Cautions about Regression

- Regression is a powerful tool for describing the relationships between variables. With software, it is very easy to run a regression analysis even when it may not be appropriate to do so. Earlier, we discussed the assumptions underlying the regression theory. Though there is **some room for slight deviations from these assumptions, especially when the number of observations is high**, we need to **check** each of them before making inferences from the regression.

- It is important to **always plot your data** so that the assumptions can be confirmed and so that you can identify if there are any outliers or (more specifically) influence points that may require additional investigation and evaluation.

Extrapolation

- **Extrapolation** is a term given to the incorrect application of a regression equation **outside of the range of data** studied.

- **Interpolation** produces estimates **between known observations**.

- **Extrapolation is subject to greater uncertainty and a higher risk of producing meaningless results.**

Lurking Variables

- Often there are variables that **are not measured** but have **strong influence on a dependent or an independent variable**. Such variables are **called lurking variables**.

- You should always consider this when interpreting regression results.

Causation and Association

- Linear regression is a tool for understanding and quantifying **associations** - which is **not the same thing as determining causation**.

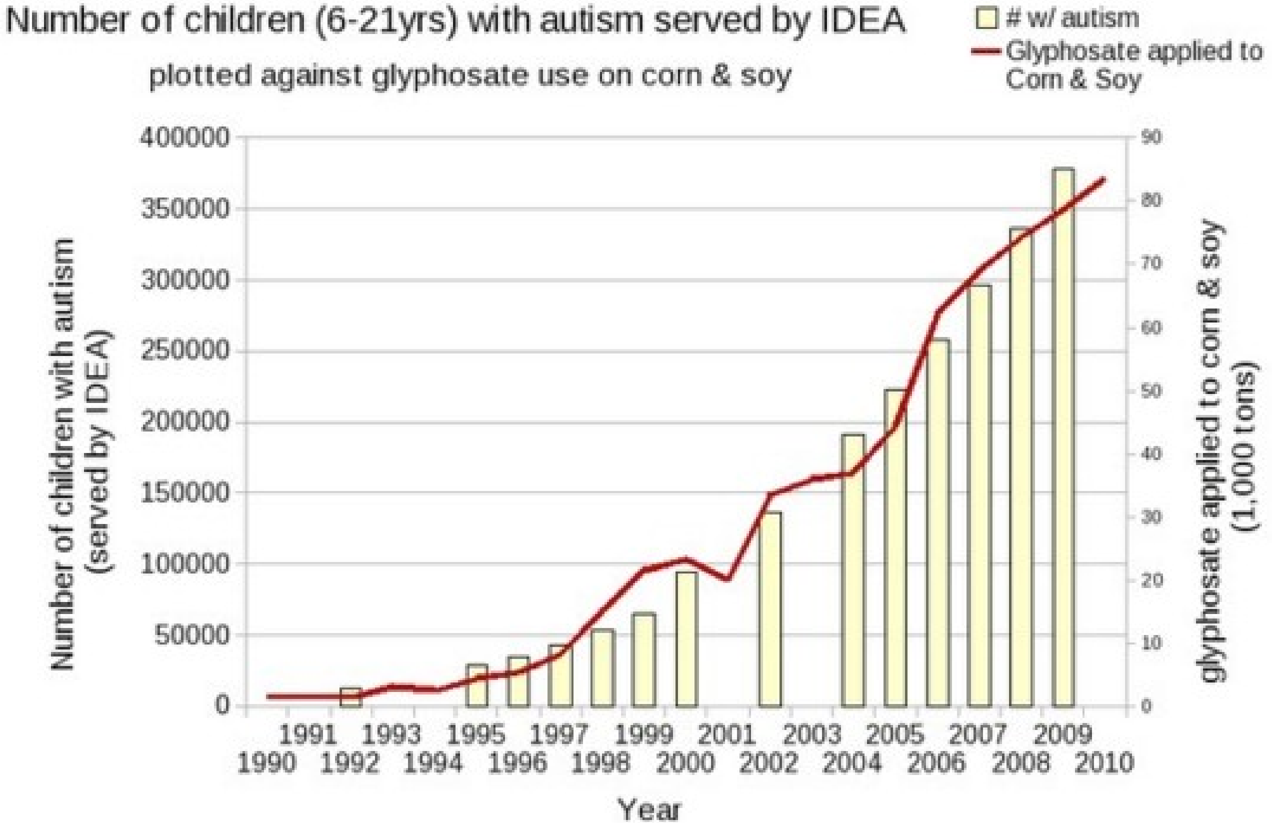
- Though the regression analysis hopes to show that changes in the explanatory variables are **associated with changes in the response variable,** it should **not be interpreted as causing changes in the response variable.**

- If one aims to show causation, an experiment where you change the independent variable in specific ways and observe the resulting effect on the response is the most convincing way to evaluate it. However, such experiments are often expensive and not always ethical to conduct.

- If an experiment cannot be performed, then in order to begin to suggest that **a causal relationship may apply, we'd want to see that the association was strong**, the association is consistent across many different studies, the cause precedes the effect temporally, and the cause is plausible (scientifically and practically).

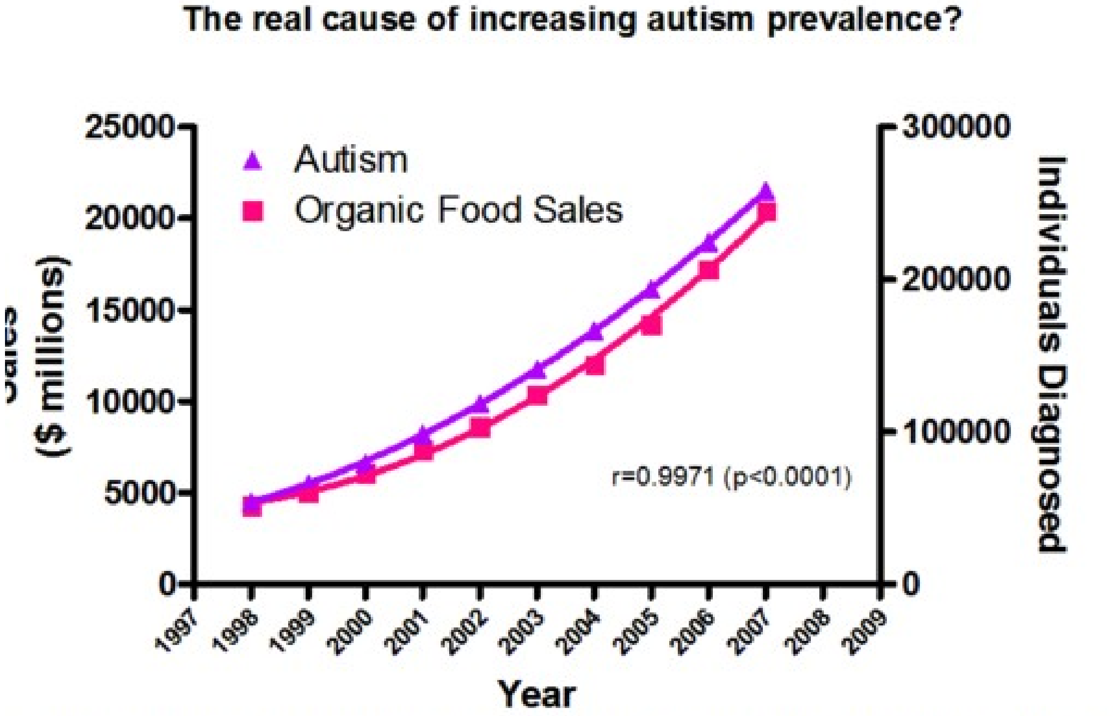
At today’s rate, by 2025, 1 in 2 children will be autistic?

- Stephanie Seneff noted that the side effects of autism closely mimic those of glyphosate toxicity, and presented data showing a consistent correlation between the use of Roundup on crops with rising rates of autism.
- The increasing prevalence of autism is largely due to increasing rates of diagnosis
- “Yes, she just **extrapolates** from current trends, assuming they’ll continue indefinitely! “
- It was one of the most hilarious examples of confusing correlation with causation I’ve ever seen.



Number of children (6-21yrs) with autism served by IDEA

plotted against glyphosate use on corn & soy



The real cause of increasing autism prevalence?

Sources: Organic Trade Association, 2011 Organic Industry Survey; U.S. Department of Education, Office of Special Education Programs, Data Analysis System (DANS), OMB# 1820-0043: "Children with Disabilities Receiving Special Education Under Part B of the Individuals
with Disabilities Education Act

http://bigthink.com/neurobonkers/no-half-of-all-children-wont-be-autistic-by-2025-whatever-facebook-tells-you

Multicollinearity

- When **two or more independent variables are highly correlated**, entering both of them into **a multiple linear regression model** may be problematic **as the effect of each may cancel the other out**.

- **This phenomenon is called collinearity or multicollinearity**. Multicollinearity increases the standard errors of the coefficients.

- Increased standard errors in turn means that coefficients for some **independent variables may be found not to be significantly different from 0**. In other words, by overinflating the standard errors, multicollinearity makes some variables statistically insignificant when they should be significant.

- It is important to look at each independent variable's association with the dependent variable **separately** before looking at them **together** as well as examining the **relationship between independent variables.**

- If **two independent variables are highly correlated (correlation >.8)** then it may be advisable to only select one for inclusion in the regression.

Warning Signs of Multicollinearity

Severe multicollinearity is a major problem, because it increases the variance of the regression coefficients, making them unstable.

Here are some things to watch for:

- **A regression coefficient is not significant** even though, theoretically, that variable should be highly correlated with Y.

- When you add or delete **an X variable, the regression coefficients change dramatically**.

- You see **a negative regression coefficient when your response should increase along with X.**

- You see **a positive regression coefficient when the response should decrease as X increases**.

- **Your X variables have high pairwise correlations.**

3D plots

```
> install.packages("rgl")
```

rgl Provides medium to high level functions for 3D interactive graphics, including functions modelled on base graphics (plot3d(), etc.) as well as functions for constructing representations of geometric objects (cube3d(), etc.). Output may be on screen using OpenGL, or to various standard 3D file formats including WebGL, PLY, OBJ, STL as well as 2D image formats, including PNG, Postscript, SVG, PGF.

```
> library(rgl)
> plot3d(age, height, salary, type = "s", size = .75, xlab="Age", ylab="Height",
    zlab="Annual Salary")
```

Check the linear model you built

Is the model statistically significant?

- Check the F statistic (at the bottom of the summary)

Are the coefficients significant?

- Check the coefficient’s t statistics and p-values in the summary, or check their confidence intervals

Is the model useful?

- Check the R^2 near the bottom of the summary

Does the model fit the data well?

- Plot the residuals and check the regression diagnostics

Does the data satisfy the assumptions behind linear regression?

- Check if the diagnostics confirm that a linear model is reasonable for your data

Page 267-8, **R cookbook**