Regression Diagnostics-
Data Analysis and Visualization

Dr. Farshid Alizadeh-Shabdiz

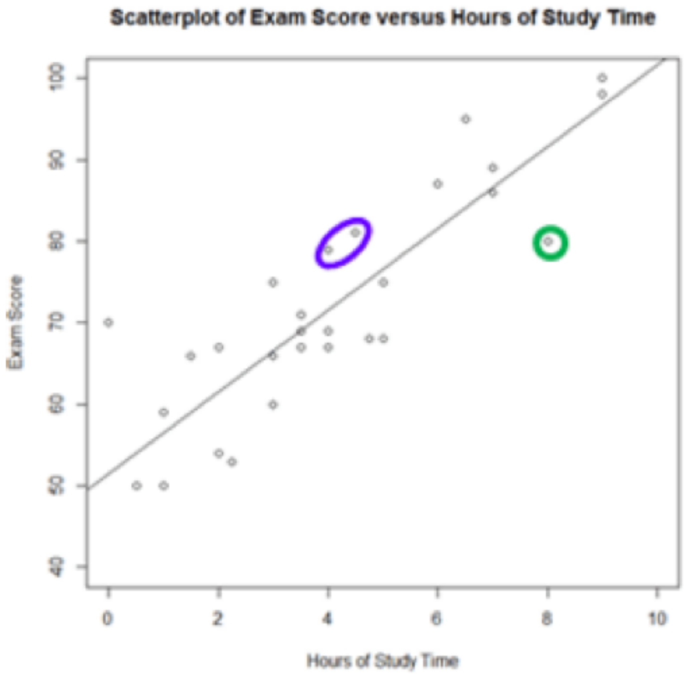
Spring 2021

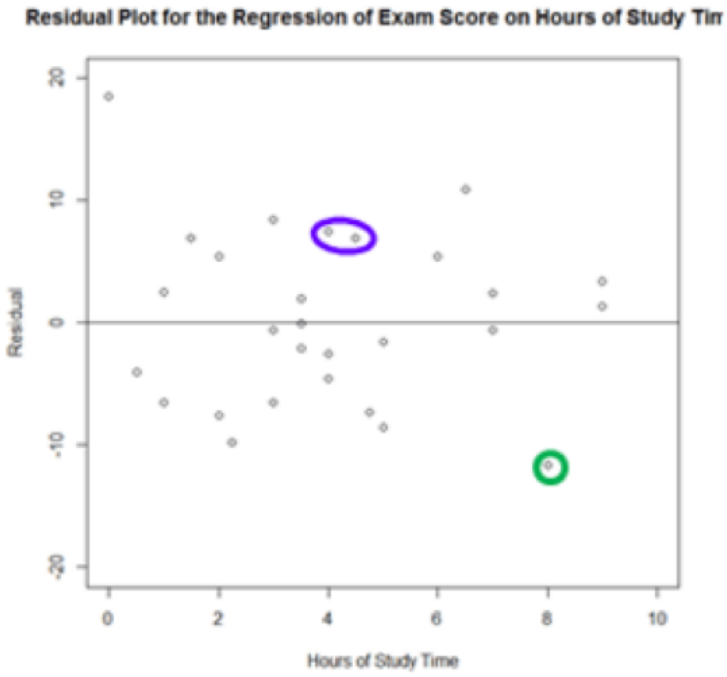
Regression Diagnostics

- ▷ **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.**
- ▷ In Regression Diagnostics we check if the assumptions are met in order to have confidence in the inferences.
- ▷ These techniques involve visualizing.
- ▷ We look for issues with **violations of the assumptions of the regression model.**

Residual Plots

- Residual plots help visualize how well a regression equation fits the sample data.
- Residual plots are scatterplots of the regression **residuals (y-axis)** against the **explanatory variable (x-axis)**.
- The residual plot **turns the regression line on the horizontal line** to see patterns and unusual observations.





Residual Plots

- ▷ Residual plots can also be generated using the **predicted values on the x-axis** as opposed to the explanatory variable **(especially more useful when MLR)**.
- ▷ Residual plots can also be generated by **plotting standardized or studentized residuals**[1] (which involves dividing the residual by an estimate of the variability of the residuals).

[1] Read more `https://en.wikipedia.org/wiki/Studentized_residual` `https://en.wikipedia.org/wiki/William_Sealy_Gosset`

R commands - Residual Plots

```
# Claculates the residual values of a Linear Regression Model
> resid(m)

# Creating 4 plots side by side in 2 rows and 2 columns
> 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))
```

Outliers and Influence Points

Outliers:

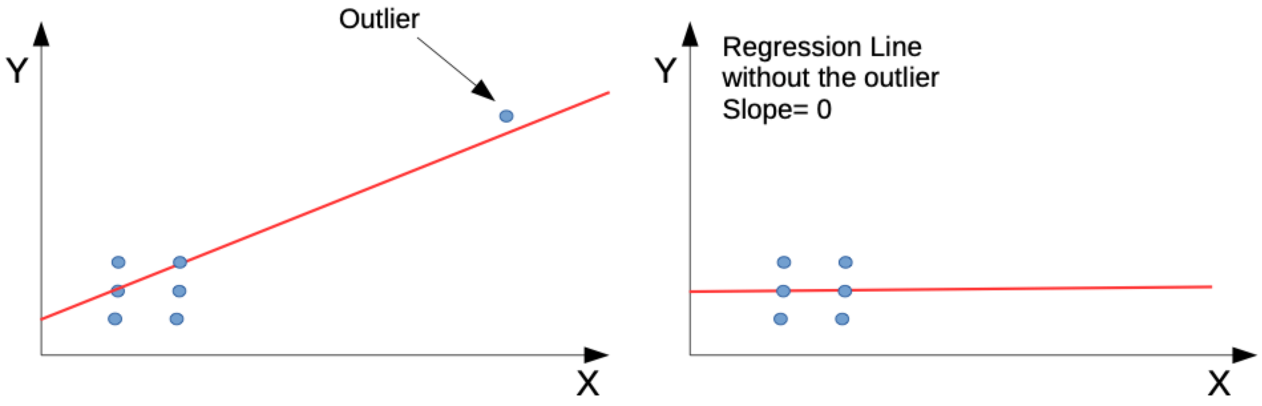
- ▷ Outliers are observations that lie outside the overall pattern of the other observations.
- ▷ Outliers can be identified via review of the scatterplot.
- ▷ Outliers in the y-direction tend to have large residuals.
- ▷ Outliers in the x-direction may or may not have large residuals but have **the potential to be influential**.

Influence Point:

- ▷ An influence point is an observation that **markedly changes the result of the regression if it were to be removed from the calculation.**

Outliers and Influence Points

- Least-squares regression is based on minimizing the squared vertical distances between the observations and the regression line (l2 norm), 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.



Outliers and Influence Points

- When there are outliers, we 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.

4 Principal Assumptions of the least-square regression

The following conditions must all be met before it is appropriate to make inference from a least-squares regression:

- ▷ The true relationship is **linear**.
- ▷ The observations are **independent**.
- ▷ The **variation** of the response variable around the regression line is **constant** (Constance Variance).
- ▷ 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).

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))
```

Linearity and additivity of the relationship

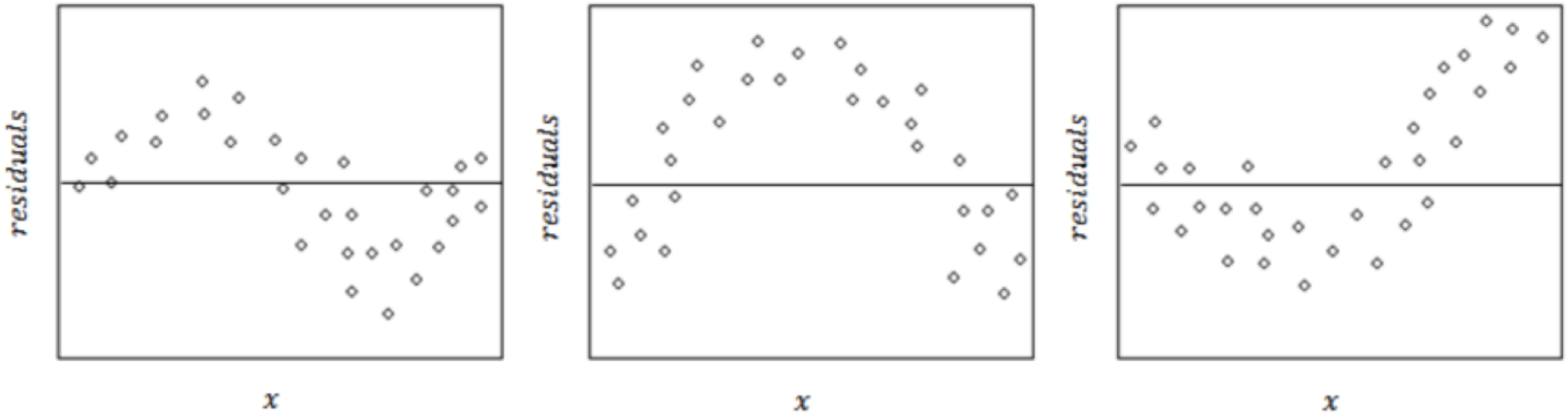
Linearity and additivity of the relationship between dependent and independent variables:

- The expected value of dependent variable is a **straight-line function** of each independent variable, holding the others fixed.
- **Slope 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**.

Linearity

- Generate a scatterplot to visualize a roughly linear trend between factors.
- Be cautious of curved or other non-linear relationships.
- Residual plots can help assess the assumption as they can magnify non-linearity.
- **Violations of linearity or additivity are extremely serious:** your predictions might to be seriously in error when data are non-linearly or non-additively related.

Examples of Violations:



Independence

We make the assumption that the **observations are independent.**

Example:

`Assume 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).`

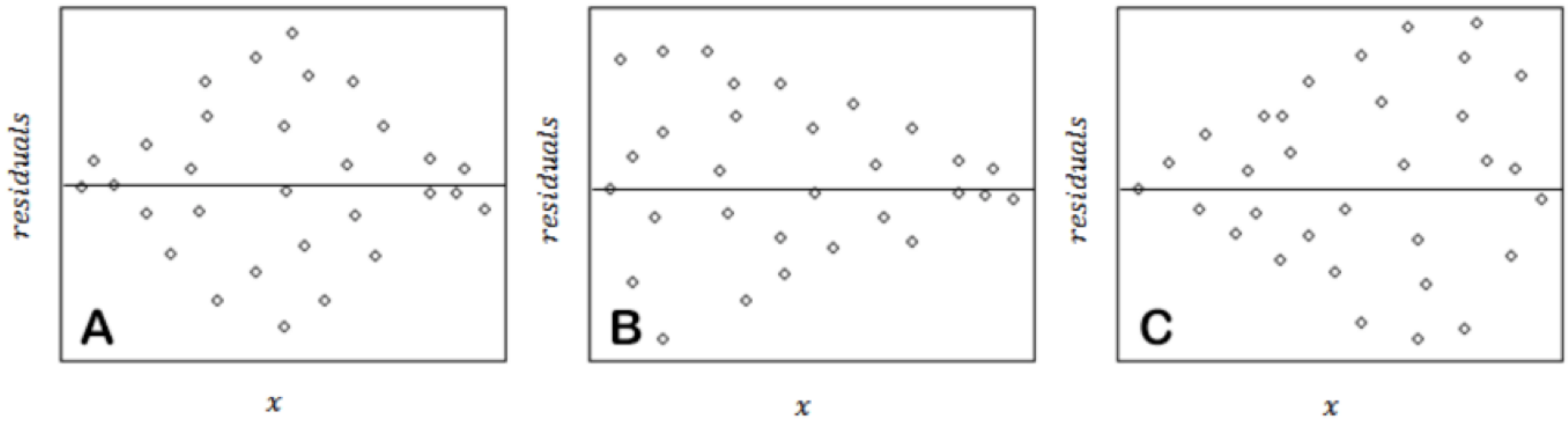
Correlated Data:

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.

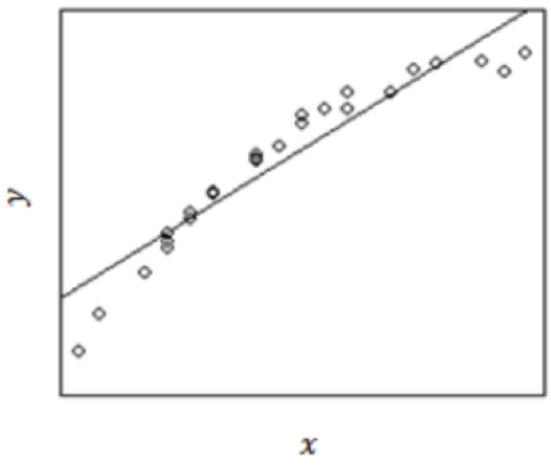
Examples of Violations:

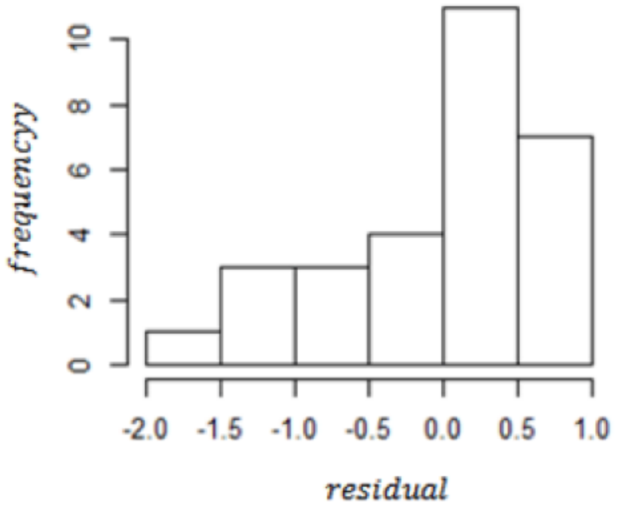


Normality

- 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 use histograms of the residuals.





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**:

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

Where

- ▷ $\boldsymbol{e_i}$ is the residual for observation i and
- ▷ $\boldsymbol{MSE}$ 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.

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.
```

Transformations

Assumptions of regression are violated?

Transformations can often help and be applied. If the variance of the response increases/decreases as the explanatory variable increases/decreases, 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.

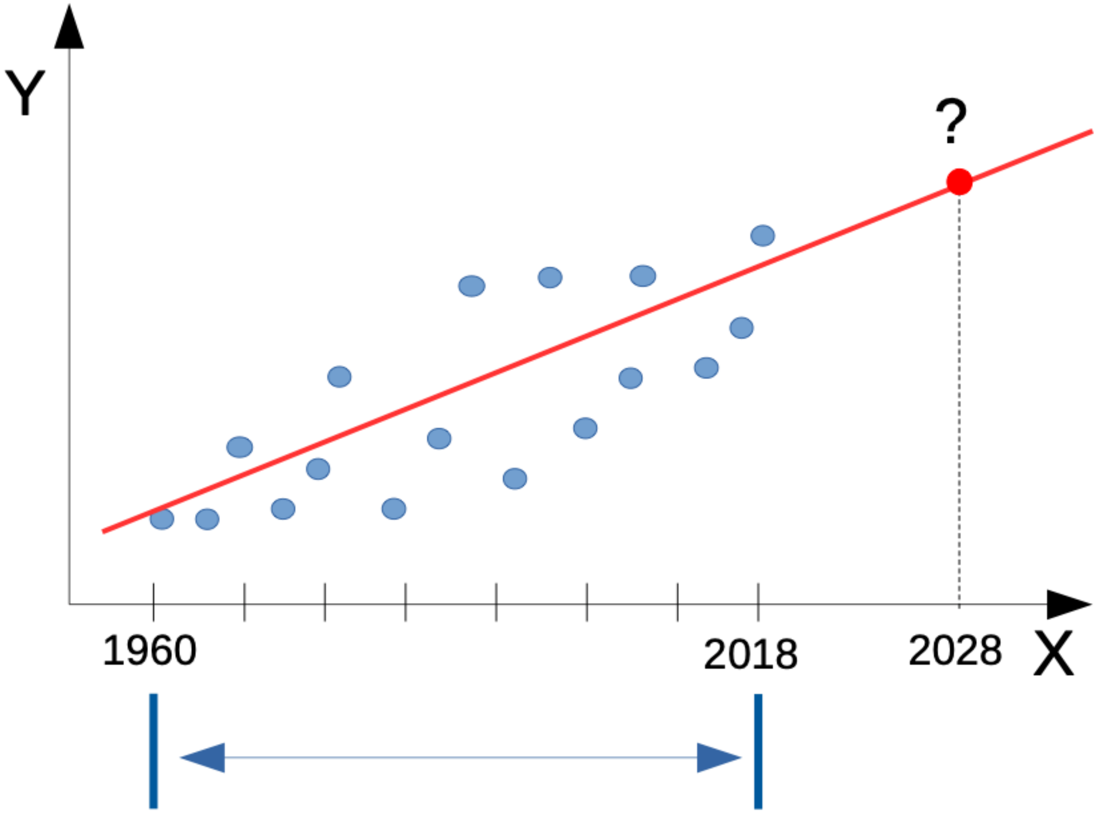
Finding the right transformation often involves some trial and error.[2]

If transformation **only marginally improves**, then **no transformation is preferred** (Due to the increased complexity)

[2] See section 11.10 and 11.1 in R Cookbook on transformations

Extrapolation

- Extrapolation is 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** (aka confounding).
- You should always consider this when interpreting regression results.

Example: Consider the correlation between study time and grade score.

Many other hidden variables might have an impact on final grade score.

Variables like **stress level of a person, amount of sleep per day, motivation score, etc . . .**

Causation and Association

- ▷ Linear regression is useful to understand and quantify **associations** - which is **not the same thing as determining causation**.
- ▷ The regression shows changes in the explanatory variables are associated with changes in the response variable, it **is not the as causing changes in the response variable**.

Example:

Amount of milk consumed in a country might be correlated to number of forest fires.

- But it does not mean that it caused that.

- Sometimes you might find a third factor that connects things, like weather.

It is easy to find correlation but it is hard to proof causation.

Causation and Association

- 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.
- For an experiment, we **actually try to provoke a response by doing some actions**.
- Experiments are often expensive and not always ethical to conduct (Example medical research)
- **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).

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 SE means some **regression coefficients are close to 0**. *Multicollinearity makes some variables statistically insignificant when they should be significant.*

Multicollinearity

What to do?

- 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 ($correlation > 0.8$) 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:

1. **A regression coefficient is not significan**t even though, theoretically, that variable should be highly correlated with Y.
2. When you **add or delete an X variable**, the **regression coefficients change dramatically.**
3. You see a **negative regression coefficient** when your **response should increase** along with X.
4. You see a **positive regression coefficient** when the **response should decrease** as X increases.
5. Your X variables have **high pairwise correlations**.

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 coefficients 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

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")
```