# BACS - HW 11

**Question 1)** Model fit is often determined by $R^2$ so let's dig into what this perspective of model fit is all about. Download `demo_simple_regression_rsq.R` from Canvas – it has a function that runs a regression simulation. This week, the simulation also reports $R^2$ along with the other metrics from last week.

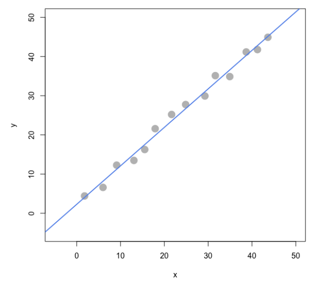
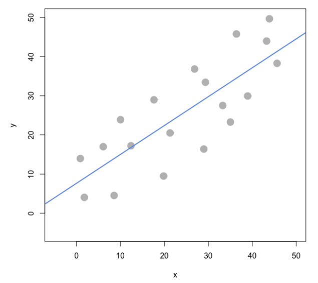
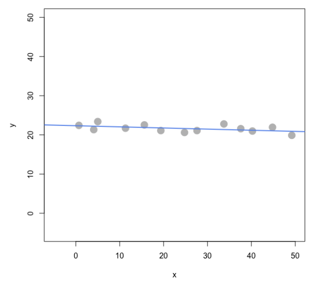
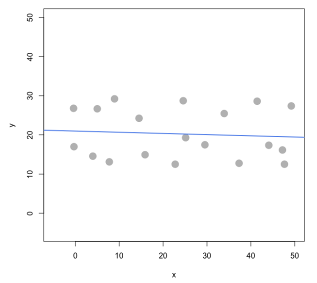
To answer the questions below, understand each of these four scenarios by simulating them:

Scenario 1: Consider a very <u>narrowly dispersed</u> set of points that have a negative or positive <u>steep</u> slope

Scenario 2: Consider a <u>widely dispersed</u> set of points that have a negative or positive <u>steep</u> slope

Scenario 3: Consider a very <u>narrowly dispersed</u> set of points that have a negative or positive <u>shallow</u> slope

Scenario 4: Consider a <u>widely dispersed</u> set of points that have a negative or positive <u>shallow</u> slope

| *Scenario 1* | *Scenario 2* | *Scenario 3* | *Scenario 4* |
|---|---|---|---|
|  |  |  |  |

a. Let's dig into what regression is doing to compute model fit:
   i. Plot Scenario 2, storing the returned points: `pts <- interactive_regression_rsq()`
   ii. Run a linear model of x and y points to confirm the $R^2$ value reported by the simulation:
   `regr <- lm(y ~ x, data=pts)`
   `summary(regr)`
   iii. Add line segments to the plot to show the regression residuals (errors) as follows:
   - Get values of $\hat{y}$( regression line's estimates of y, given x): `y_hat <- regr$fitted.values`
   - Add segments: `segments(pts$x, pts$y, pts$x, y_hat, col="red", lty="dotted")`
   iv. Use only `pts$x`, `pts$y`, `y_hat` and `mean(pts$y)` to compute SSE, SSR and SST, and verify $R^2$

b. Comparing scenarios 1 and 2, which do we expect to have a stronger $R^2$ ?

c. Comparing scenarios 3 and 4, which do we expect to have a stronger $R^2$ ?

d. Comparing scenarios 1 and 2, which do we expect has bigger/smaller SSE, SSR, and SST?
   *(do not compute SSE/SSR/SST here – just provide your intuition)*

e. Comparing scenarios 3 and 4, which do we expect has bigger/smaller SSE, SSR, and SST?
   *(do not compute SSE/SSR/SST here – just provide your intuition)*

*(Question 2 on next page)*

**Question 2)** We're going to take a look back at the early heady days of global car manufacturing, when American, Japanese, and European cars competed to rule the world. Take a look at the data set in file `auto-data.txt`. We are interested in explaining what kind of cars have higher fuel efficiency (`mpg`).

```
1. mpg:           miles-per-gallon (dependent variable)
2. cylinders:     cylinders in engine
3. displacement:  size of engine
4. horsepower:    power of engine
5. weight:        weight of car
6. acceleration:  acceleration ability of car
7. model_year:    year model was released
8. origin:        place car was designed (1: USA, 2: Europe, 3: Japan)
9. car_name:      make and model names
```

Note that the data has missing values ('?' in data set), and lacks a header row with variable names:

```
auto <- read.table("auto-data.txt", header=FALSE, na.strings = "?")
names(auto) <- c("mpg", "cylinders", "displacement", "horsepower", "weight",
                 "acceleration", "model_year", "origin", "car_name")
```

a. Let's first try exploring this data and problem:
   i. Visualize the data in any way you feel relevant (report only relevant/interesting ones)
   ii. Report a correlation table of all variables, rounding to two decimal places (in the `cor()` function, set `use="pairwise.complete.obs"` to handle missing values)
   iii. From the visualizations and correlations, which variables seem to relate to `mpg`?
   iv. Which relationships might not be linear? *(don't worry about linearity for rest of this HW)*
   v. Are there any pairs of independent variables that are highly correlated (*r > 0.7*)?
b. Let's create a linear regression model where `mpg` is dependent upon all other suitable variables *(Note: `origin` is categorical with three levels, so use `factor(origin)` in `lm(...)` to split it into two dummy variables)*
   i. Which independent variables have a 'significant' relationship with mpg at 1% significance?
   ii. Looking at the coefficients, is it possible to determine which independent variables are the *most effective* at increasing mpg? If so, which ones, and if not, why not? (hint: units!)
c. Let's try to resolve some of the issues with our regression model above.
   i. Create fully standardized regression results: are these slopes easier to compare? (note: consider if you should standardize `origin`)
   ii. Regress mpg over each *nonsignificant* independent variable, individually. Which ones become significant when we regress `mpg` over them individually?
   iii. Plot the density of the *residuals*: are they normally distributed and centered around zero? (get the residuals of a fitted linear model, e.g. `regr <- lm(...)`, using `regr$residuals`