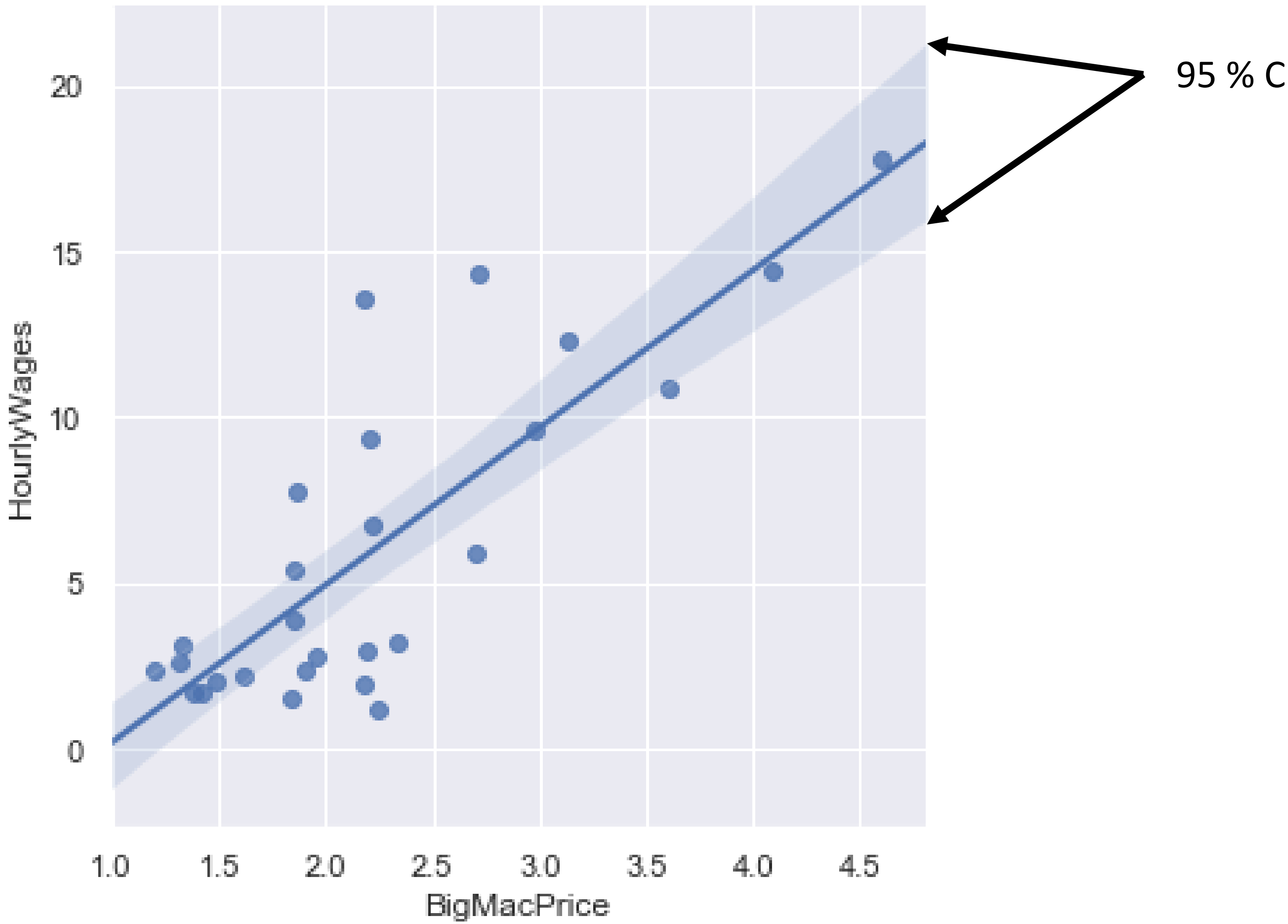
Estimation – Confidence Interval

Estimation – Confidence Intervals

A regression line provides a point estimate from a sample. A different sample may yield a different point estimate. A Confidence Interval for estimating a average values of y for a give x is more useful.

$$E(y_x) = \hat{y} \pm t_{n-2,\frac{\alpha}{2}} * SE * \sqrt{\frac{1}{n} + \frac{(x_o - \bar{x})^2}{SS_{xx}}}$$

Where x_o = a particular value of x


95 % CI
HourlyWages
BigMacPrice
20
15
10
5
0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5

Simple Linear Regression - Steps

•

Get familiar with data
- Plots
- Descriptive stats

Formulate a linear model and fit to data
- Do Regression

Inadequate fit

Check model and assumptions
- Look at residual plots
- Look at unusual observations
- Look at R-Squared
- Look at p-values

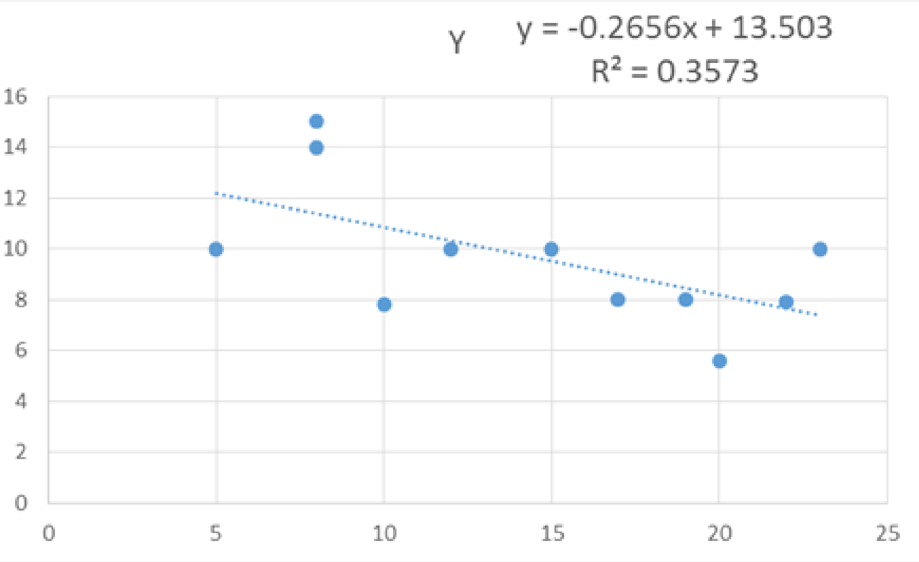
Good fit

Report results and equation
- Make predictions for values of interest

R^2

as metric for quality of fit- some caveats

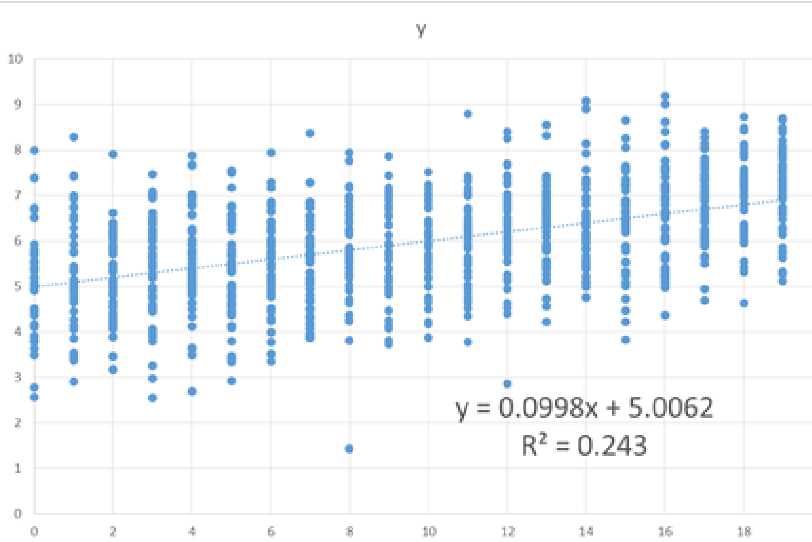
R-Squared and Significance - Caution



23	Regression Analysis						
24							
25	OVERALL FIT						
26	Multiple R	0.597718					
27	R Square	0.357267					
28	Adjusted R Square	0.285852					
29	Standard Error	2.335299					
30	Observations	11					
31							
32	ANOVA				Alpha	0.05	
33		*df*	*SS*	*MS*	*F*	*p-value*	*sig*
34	Regression	1	27.282857	27.28285699	5.002704	0.052125754	no
35	Residual	9	49.082598	5.453621951			
36	Total	10	76.365455				
37							
38		*coeff*	*std err*	*t stat*	*p-value*	*lower*	*upper*
39	Intercept	13.50289	1.8553076	7.277980002	4.67E-05	9.305894086	17.699888
40	X	-0.26561	0.1187518	-2.23667255	0.052126	-0.53424402	0.0030263

- R-Sq suggests that 35% of variation in y can be explained by variation in x.
- t and F tests show that coefficient is <u>not significant</u> and null hypothesis cannot be rejected.
- The 95% confidence interval of the slope, $b_1 \pm t_{crit} * s_b$, is (-0.534,0.003).

R-Squared and Significance - Caution



Regression Statistics	
Multiple R	0.492914799
R Square	0.242964999
Adjusted R Square	0.242206447
Standard Error	1.016805138
Observations	1000

ANOVA

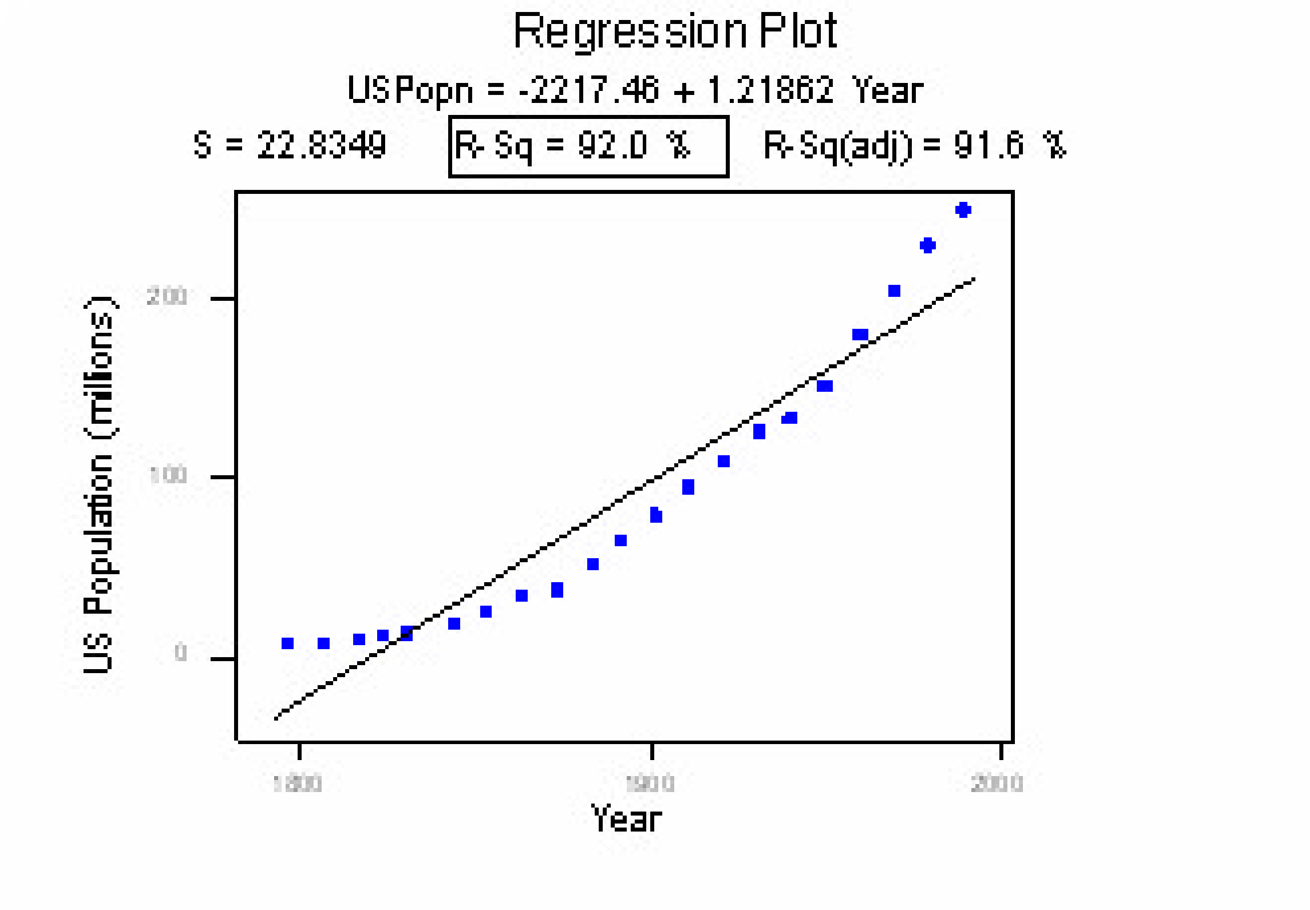
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	331.1568641	331.1568641	320.3010019	2.43789E-62
Residual	998	1031.824904	1.033892689		
Total	999	1362.981768			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	5.006235417	0.061969128	80.78595852	0	4.88463068	5.127840154
x	0.09979782	0.005576246	17.8969551	2.43789E-62	0.088855309	0.110740332

- R-Sq suggests that 24% of variation in y can be explained by variation in x.
- t and F tests show that coefficient is <u>significant</u> and null hypothesis should be rejected.
- The 95% confidence interval of the slope, $b_1 \pm t_{crit} * s_b$, is (0.089,0.111).
- *Statistical significance* doesn't necessarily mean *practical significance*.

Caution: High R^2 doesn't imply a good fit !

- US population from 1790 to 1900 (decade wise data)

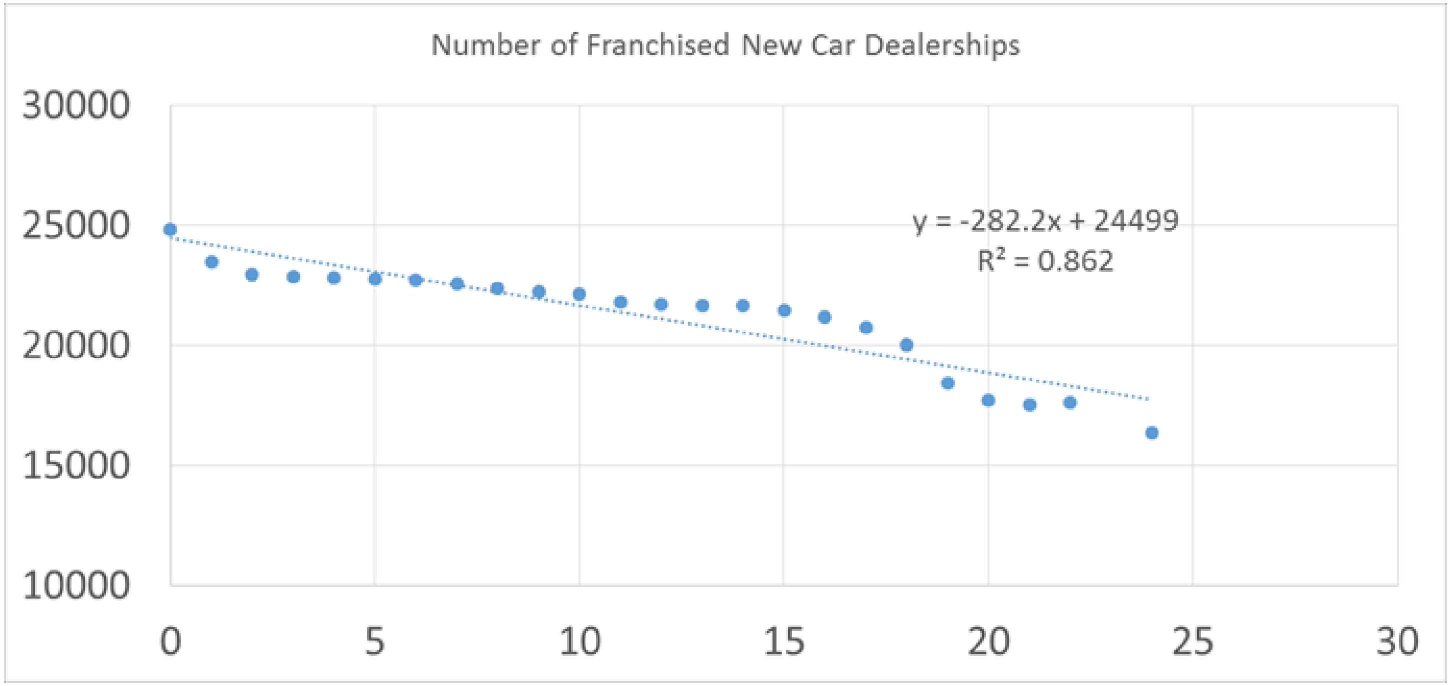


New Car Dealerships data

National Automotive Dealers Association (NADA) of US publishes state-of-the-industry report each year.

You want to know if there is any linear relationship between the time since 1990 and the number of franchised new car dealerships.

R-Squared, Significance and Residuals - Caution



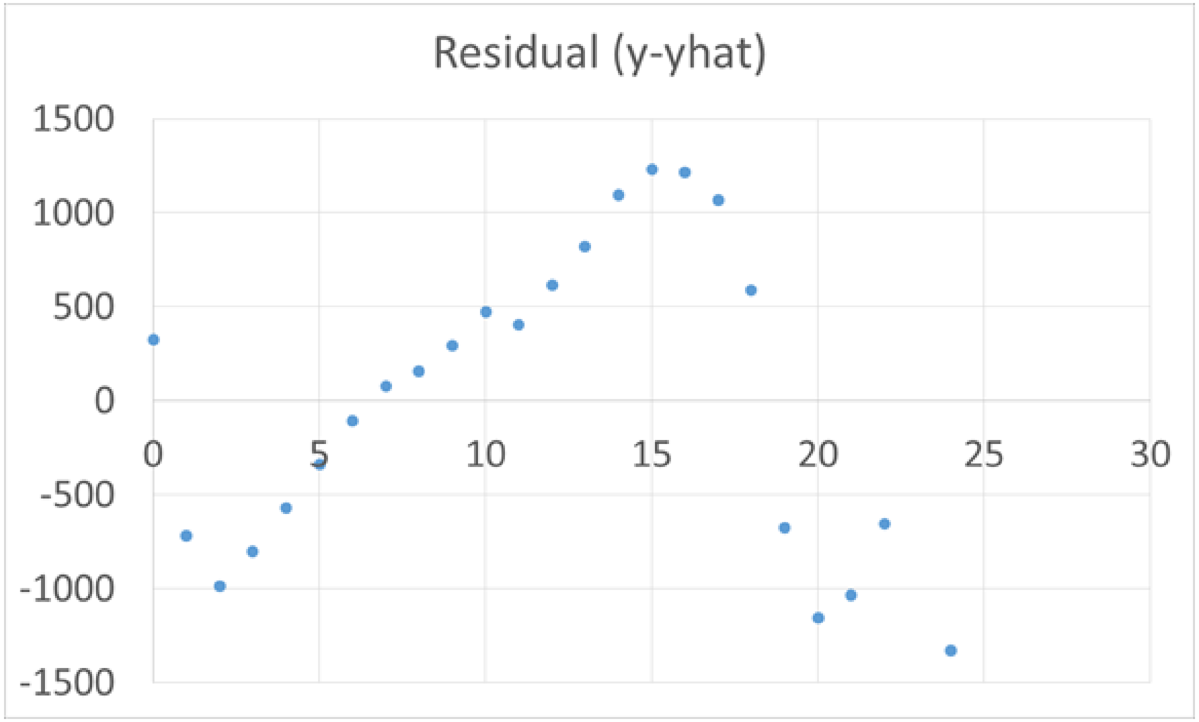
- Based on the shape of the scatter plot, do you think a linear fit looks good?
- Does R^2 imply a good fit?
- What can you infer from the intercept and the slope?

R-Squared, Significance and Residuals - Caution

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.928448566							
R Square	0.862016739							
Adjusted R Square	0.855744773							
Standard Error	824.748263							
Observations	24							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	93487768.66	93487768.66	137.4396293	6.21261E-11			
Residual	22	14964613.34	680209.6973					
Total	23	108452382						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 99.0%*	*Upper 99.0%*
Intercept	24498.51368	324.8477406	75.41537349	4.68438E-28	23824.8207	25172.20666	23582.84714	25414.18022
Time Since 1990 (in years)	-282.1961313	24.07105183	-11.7234649	6.21261E-11	-332.1164374	-232.2758252	-350.0465546	-214.3457081

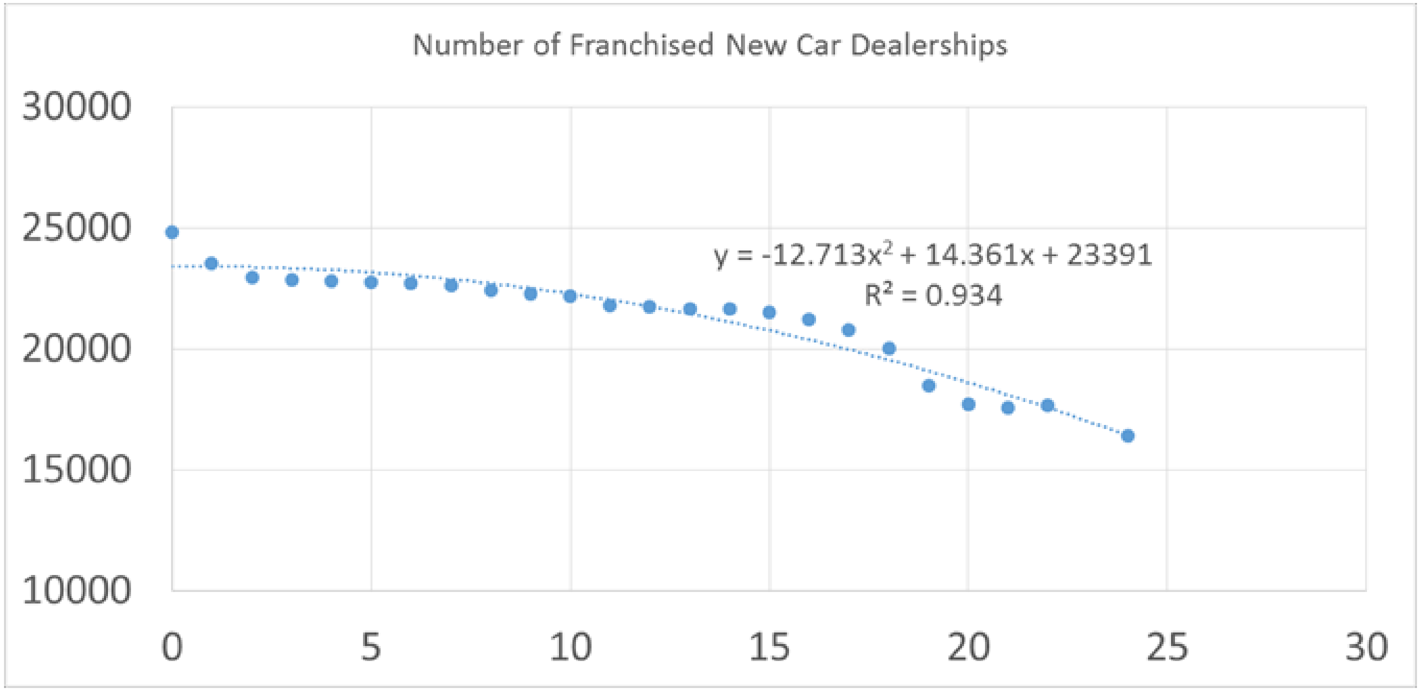
- Is the slope significant?
- Is the model significant?

R-Squared, Significance and Residuals - Caution

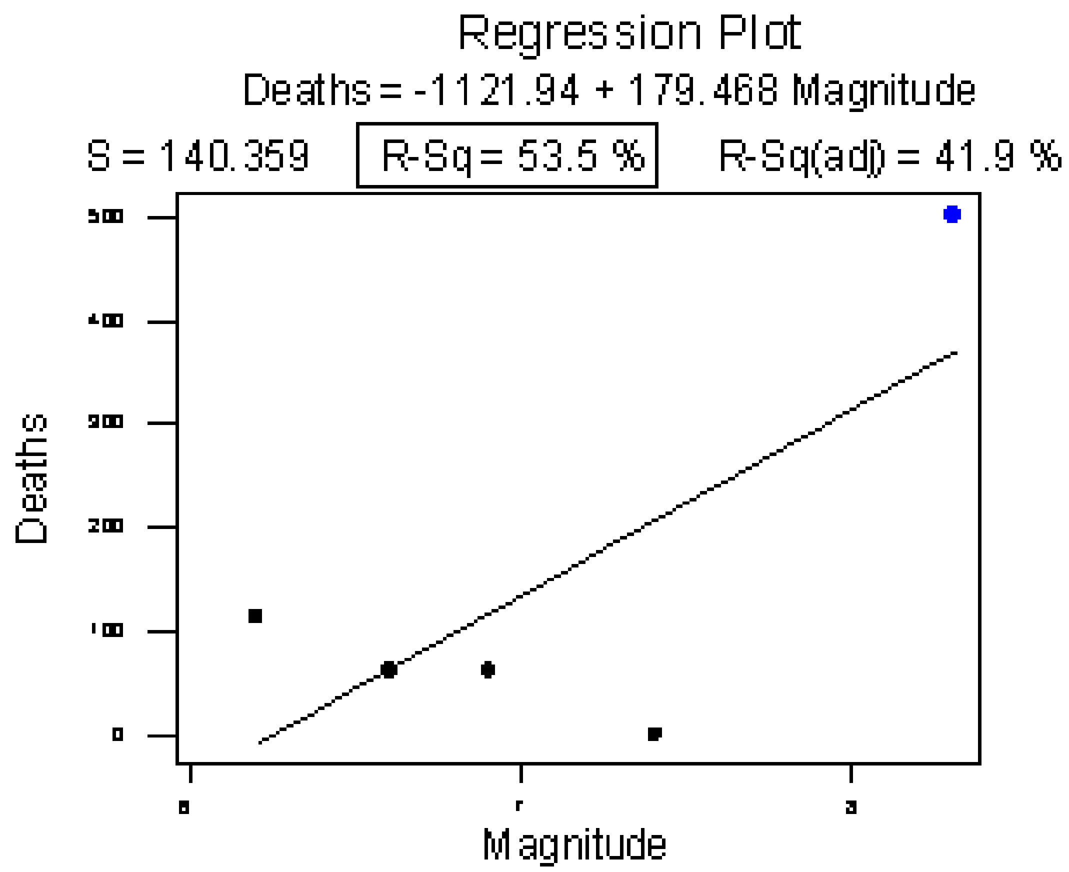


- Based on the residual plot, do you think a linear model is a good fit?

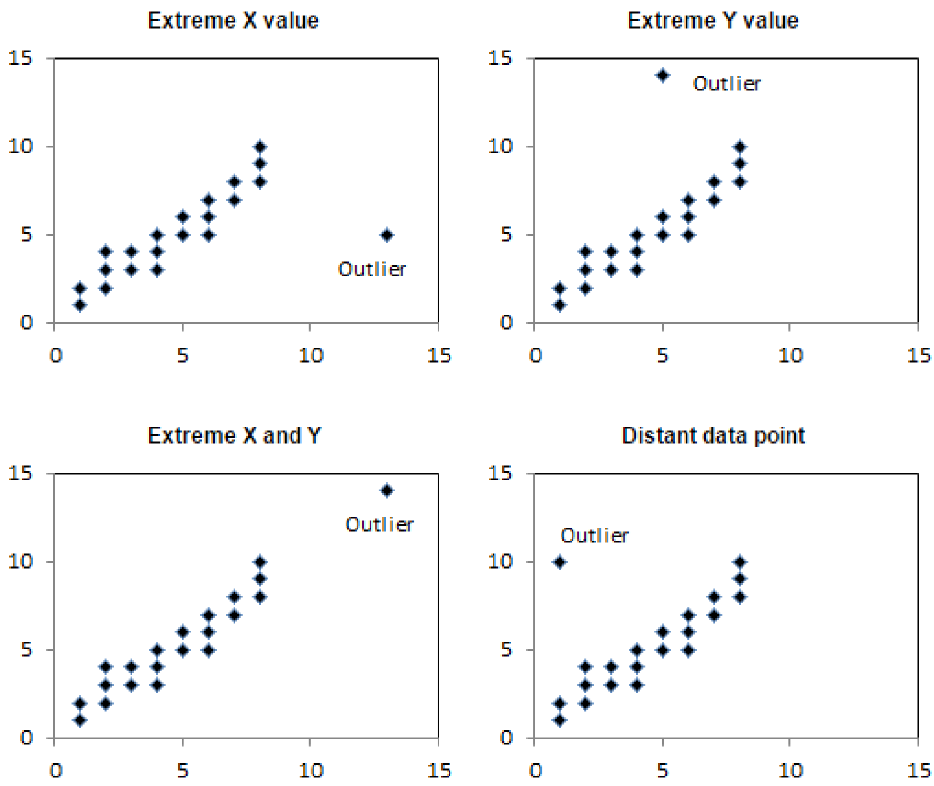
R-Squared, Significance and Residuals - Caution



Caution: Single point can change the result



Outliers



- Outliers do not follow the general trend of the rest of the data
- Outlier typically have a large residual

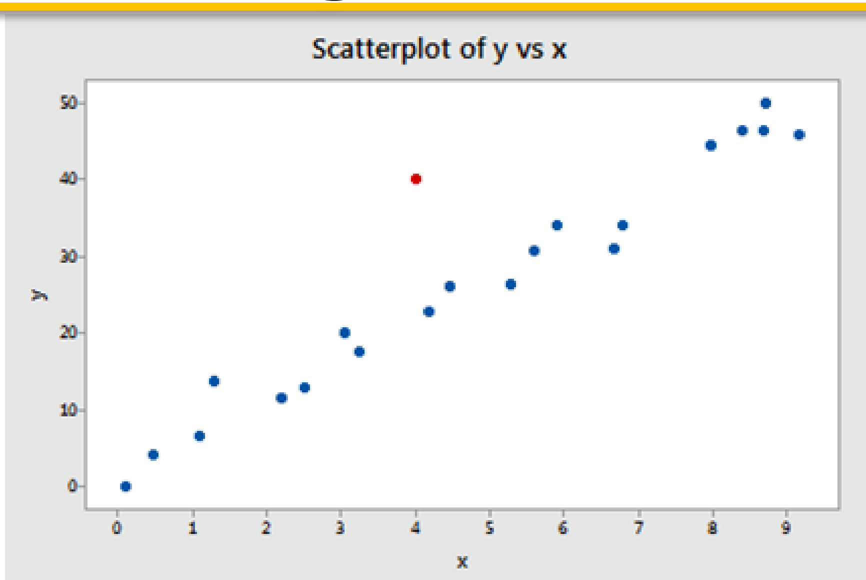
Influential Observations - Leverage

How much the observation's value on the predictor variable differs from the mean of the predictor variable. That is it tells us about extreme x values, which have the potential to highly influence the regression in certain conditions.

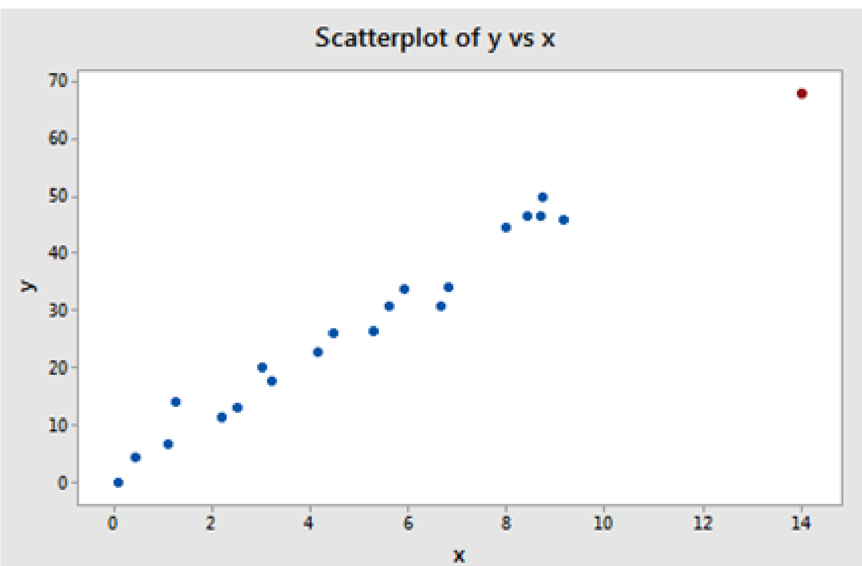
$$Leverage, h = \frac{(Standardized\ predictor\ value)^2+1}{n}$$

The sum of leverages = # of parameters, *p* (regression coefficient including intercept).

Influential Observations - Leverage



Low leverage



High leverage

Flat observation

Whose h >3* avg(h) or h > 2* avg(h)

$$\text{Avg(h)} = \frac{sum(h)}{n} = \frac{p}{n}$$

Influential Observations

An observation which, when not included, greatly alters the predicted scores of other observations.

Cook's D is a measure of the influence and is proportional to the sum of the squared differences between predictions made with all observations in the analysis and predictions made leaving out the observation in question.

If Cook's D > 1, the observation can be considered as having too much influence.

Points with Cook's D> 0.5 should be investigated

Influence is a function of leverage and residual.

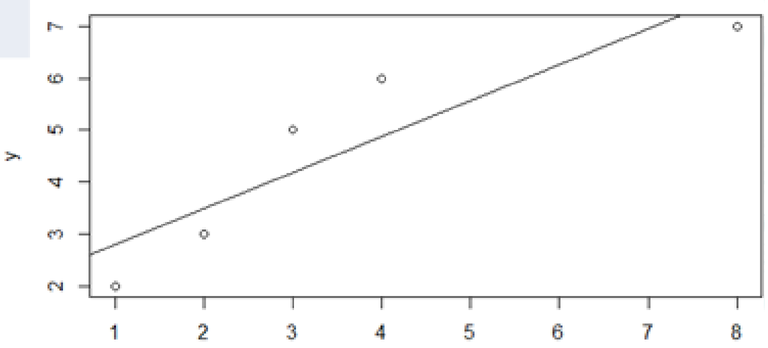
Influential Observations - Distance

Based on error of prediction and is measured by Studentized Residual, which is related to error of prediction of that observation divided by the standard deviation of the errors of prediction.

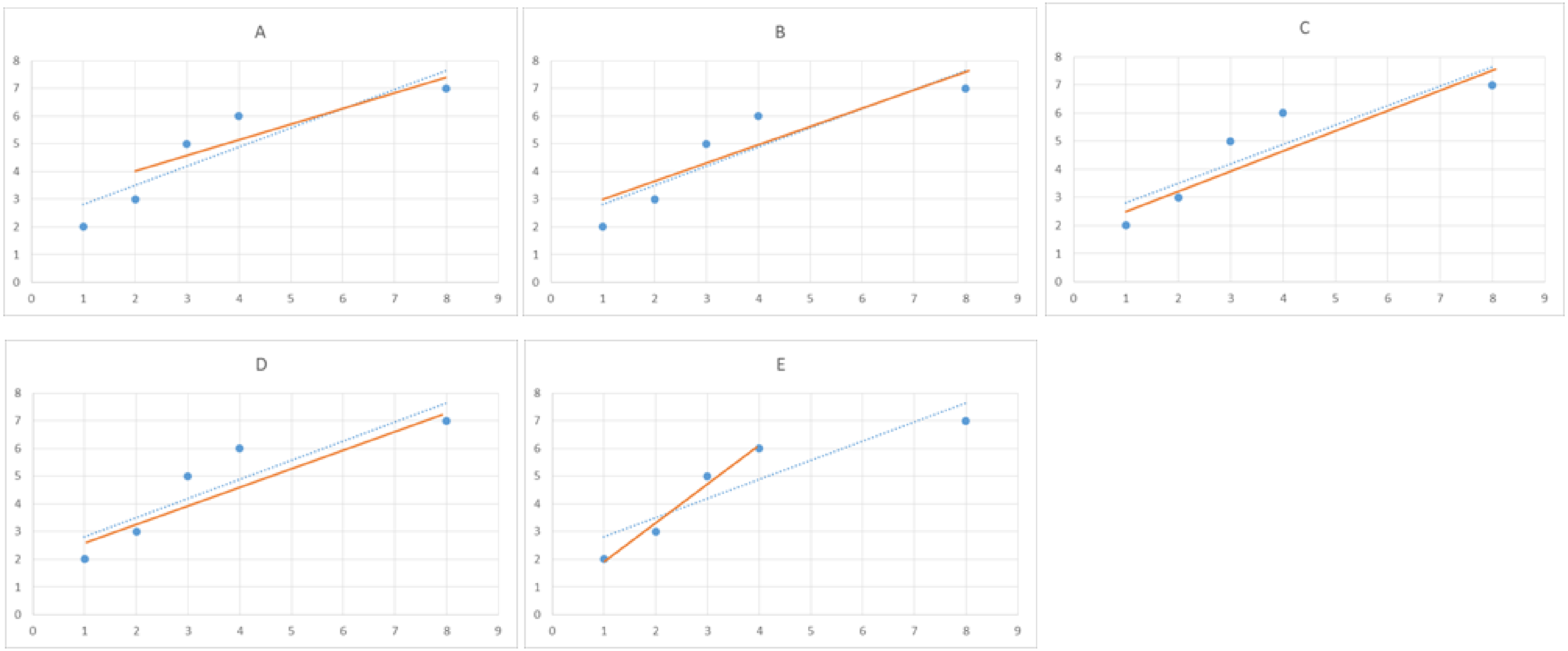
ID	X	Y	h	R	D
A	1	2	0.39	-1.02	0.4
B	2	3	0.27	-0.56	0.06
C	3	5	0.21	0.89	0.11
D	4	6	0.2	1.22	0.19
E	8	7	0.73	-1.68	8.86

h is the leverage, R is the studentized residual, and D is Cook's measure of influence.

D> 0.5 : Investigate
D>1 : Influential point



Influential Observations



Influential Observations

So what does one do when you find influential observations in your dataset?

- Check if its bad data or there was a procedural error in data collection – delete/correct it
- If data not representative of intended study population– delete it
- Use business intelligence to figure out if different physics or processes involved for the region near the influential point. Maybe a different model applies there.
- Are there other relevant variables that you are ignoring? Redo model with those.
- If unsure – report results with both including the data point and excluding it.