7. Scatterplots, association, and correlation

- Scatterplots, $y =$ response variable, $x =$ explanatory variable
- Association: direction, form, strength.
- Outliers, clusters.
- Correlation: measures direction and strength of linear association.
- Lurking variables: correlation does not imply causation.

- Old cars data.
- Introduce correlation coefficient using simulated data.

$$r = \frac{1}{n-1} \sum z_x z_y$$

- Cars again.
- Atlantic hurricanes: Max wind speed (mph) and Central pressure (mb).
- Belgravia 2003 real estate prices and property taxes.
- Cigarettes: nicotine, tar, and carbon monoxide.

- Imagine what a scatterplot might look like for:
  - drug dosage and degree of pain relief.
  - calories consumed and weight loss.
  - hours of sleep and score on a test.
  - shoe size and grade-point average.
  - age of car and cost of repairs.

- Is the correlation coefficient resistant to outliers? How much can the correlation vary by moving one point?

- Correlation should only be reported if:
  - both variables are quantitative;
  - association (if any) is linear;
  - there are no outliers.

Big melons mean big trouble on ice Study links aggression with facial dimensions
Joanne Laucius, Canwest News Service, Thursday, August 21, 2008.

Fatheaded hockey players are more aggressive than their slimmer-faced counterparts, a St. Catharines, Ont., study has found.

Results of the study, published Wednesday in the prestigious Proceedings of the Royal Society, concluded of the six Canadian-based NHL teams, the faces of the Ottawa Senators are dead giveaways when it comes to predicting how much time players spend in the penalty box.

"We're not saying that Ottawa is more aggressive than any other team. But each individual player's face predicts how much time he had in the box," said Brock University neuroscience researcher Justin Carre.

Carre, who studies fluctuations in hormone levels, wanted to test recent theories that link male facial width-to-height ratio to behaviour such as aggression. Changes in male facial shape start at puberty, when boys are exposed to the influences of testosterone, a hormone that also sparks aggressive behaviour.

The researcher devised a lab experiment comparing facial ratios of a group of student volunteers with their aggressiveness while playing a video game.

The measurements, performed with the help of a digital ruler, compare the width of the face at the cheekbones with the height between the bottom of the eyebrows and the upper lip. An unusually wide male face has a ratio of about 2.3, while a relatively narrow face has a ratio of about 1.6.

Among the male students, those with wide faces were more likely to play the video game aggressively, even downright vengefully.

Carre, who had played American college-level hockey and is currently assistant coach of the Brock Badgers, decided to take the theory into the real world.

"We wanted to come up with the idea of readily available statistics – penalty minutes," said Cheryl McCormick, co-author of the paper and the Canada research chair in behavioural neuroscience at Brock.

He shifted his gaze to the NHL, and calculated the facial ratio for the players on Canada's NHL teams using 2007-08 roster photos and compared the results with the average number of penalty minutes per game the player racked up for aggressive behaviour such

as slashing, cross-checking, high-sticking, boarding, elbowing, checking from behind and fighting. Goalies were not analyzed.

Of the 18 Senators, Carre looked at defenceman Mike Commodore, who has since left the Senators, with a facial ratio of about 1.6 and only about a minute per game in the penalty box, was at the low end of the scale.

Right-winger Chris Neil, with a facial ratio of almost 2.4 and about three minutes per game in the box, was at the opposite end.

"(Chris) Neil was off the chart in the face ratio," said Carre.

Of the Canadian teams, the Ottawa Senators had the strongest "correlation" between facial width and aggression – although overall, the Senators were relatively gentlemanly players with relatively few penalties.

The teams had an average correlation ranking of .30. The Senators scored the highest with .51, with the Montreal Canadiens next at .39, closely followed by the Toronto Maple Leafs at .37. The Vancouver Canucks rated .24, followed by the Edmonton Oilers at .20 and the Calgary Flames at .17.

Brian Morris, a spokesman for the Senators, was at a loss for a comment on Carre's findings. It would be hard to draw conclusions based on facial measurements working from photographs alone, he suggested.

"Seemingly, it's more of a theory than a scientific fact," he said.

Carre spent Wednesday juggling requests from the international media to perform his calculations on the faces of other sports figures, mostly British footballers.

And Carre believes there's a fascinating follow-up study to be done on how facial ratios affect hockey referees. Perhaps wide-faced players are more likely to be penalized than players whose faces are less threatening.

"It might have implications for the type of officiating they get," he said.

8. Linear regression

- $y =$ response variable, $x =$ predictor or explanatory variable.
- Linear regression model. Least squares regression of y on x. Slope and intercept, interpretation, standardized variables, correlation.
- Residuals and their standard deviation, plots of residuals.
- $R^2 =$ fraction of variability of y accounted for by the least squares linear regression on x.
- Examples: Belgravia real estate, Hurricane winds, Income and housing.

Example: Belgravia real estate prices.

- Property taxes in Edmonton are based on market assessments of property value. We would thus expect property tax to be a useful predictor of sale price.

- The following data set shows $x =$ property tax and $y =$ sale price for 27 homes sold in Belgravia in 2003. Both variables are reported in thousands of dollars.

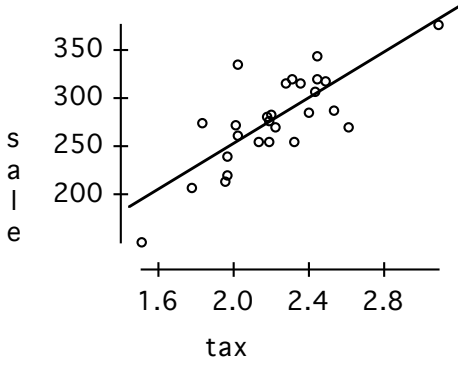
- How can we describe the relationship between tax and sale price?

- How do we use this description to make predictions?

- How good are these predictions?

tax	sale
1.507	151
1.774	207
1.960	214
1.966	220
1.965	238.5
2.320	253.9
2.018	260
2.192	255
2.128	255
2.221	270
2.616	270
2.007	272
1.832	273
2.181	280
2.402	285
2.198	283
2.189	275.5
2.534	287
2.488	318
2.308	319.5
2.436	307.5
2.279	316
2.355	315
2.442	320
2.440	344
2.024	335
3.084	377
1.962x	683
5.476x	720

Belgravia 2003 real estate: property tax and sale price



Multivariate Summaries
No Selector
29 total cases of which 2 are missing

Variable	Mean	StdDev
sale	277.848	47.5080
tax	2.21726	0.308982

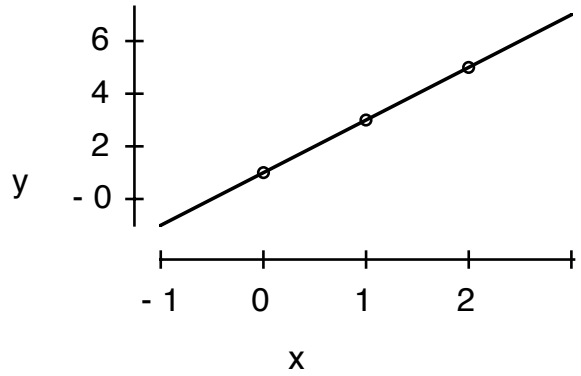
Correlation r = 0.780

Dependent variable is: sale
No Selector
29 total cases of which 2 are missing
R squared = 60.9% R squared (adjusted) = 59.3%
s = 30.29 with 27 - 2 = 25 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	35741.7	1	35741.7	39.0
Residual	22940.5	25	917.620	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	11.7853	43.03	0.274	0.7864
tax	119.996	19.23	6.24	≤ 0.0001

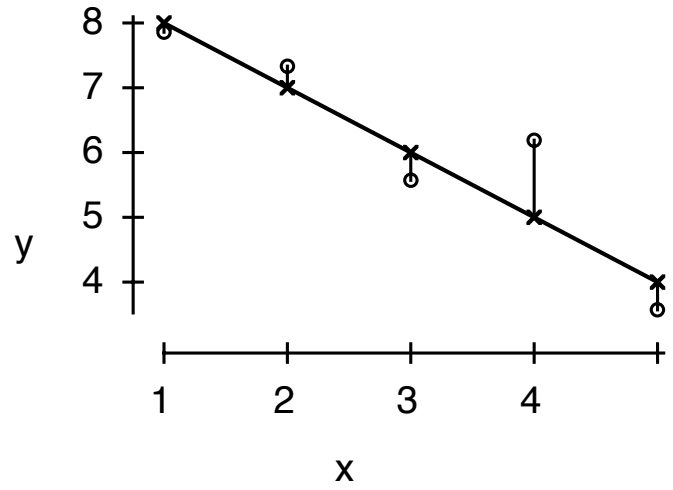
- Formula for a straight line: $y = b_0 + b_1 x$
- $b_0 =$ intercept
- $b_1 =$ slope.
- Identified points (x, y) are $(0, 1), (1, 3), (2, 5)$. Equation?



- Given two points $(5, 20), (8, 10)$, find the equation.

Best fit to data? Least squares regression line.

- For any given line $\hat{y} = b_0 + b_1 x$, the values $y - \hat{y}$ represent vertical deviations of the data points from the line.



- Choose (b_0, b_1) to minimize the sum of squared deviations:

$$\sum (y - b_0 - b_1 x)^2$$

- Solution: $$b_1 = \frac{s_y}{s_x} r \quad \text{and} \quad b_0 = \bar{y} - b_1 \bar{x}.$$

- Least squares regression line: $\hat{y} = b_0 + b_1 x$.
  We call the values $\hat{y}$ the *fitted* or *predicted* values.
  We use $\hat{y}$ to predict a new value of y given the value x.

- The regression line passes through the point $(\bar{x}, \bar{y})$. So, if $x = \bar{x}$, then we predict $\hat{y} = \bar{y}$.

- Standard deviation as a ruler: if x is increased by the value s_x, then we increase or decrease our prediction $\hat{y}$ by the value $r\, s_y$.
  Francis Galton: "regression to the mean" phenomenon.

- Least squares regression for standardized data: $\hat{z}_y = r\, z_x$.
  If z_x is increased by 1, then we increase or decrease our prediction $\hat{z}_y$ by the value r.

Belgravia example

- Calculate b_0 and b_1 for regression line $\widehat{\texttt{sale}} = b_0 + b_1\,\texttt{tax}$.
  Units? b_0 is in y units, b_1 is in (y units) per (x unit).

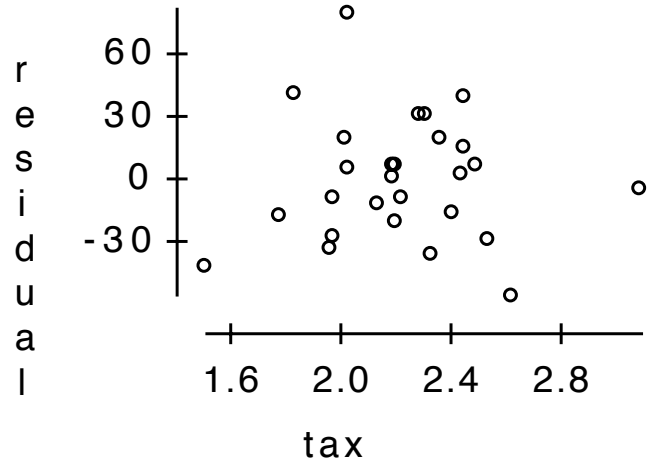
- Predict the sale price when tax $=$ \$2400.

- Predict the difference in prices for properties with taxes \$2500 and \$2000.

- Predict the sale price when tax is \$0. Is this meaninful?

How well does the line fit the data?

- Residuals: $e = y - \hat{y} = y - b_0 - b_1 x$
- What to check for in a plot of the residuals:
  - Is there an association between residuals and x variable?
  - Does the residual variation remain constant as x varies (or does the plot thicken)?



- If the residual variation is roughly constant, then it is useful to calculate the residual standard deviation:

$$s_e = \sqrt{\frac{\sum e^2}{n-2}} = s_y \sqrt{\frac{n-1}{n-2}(1-r^2)}$$

- Warning: Do not confuse s_e with s_y or s_x.

- Why divide by $n-2$?

- Belgravia example. Examine residual distribution using histogram and normal probability plot. Interpretation if residual distribution is near normal? Simple description of conditional distribution of y given x using the 68-95-99.7 rule.

- What fraction of the variability of y is accounted for by the regression of y on x? Two answers, but we will use only the first.

$$\begin{aligned} R^2 &= \frac{\sum(y-\bar{y})^2 - \sum e^2}{\sum(y-\bar{y})^2} \\ R^2(\text{adjusted}) &= \frac{s_y^2 - s_e^2}{s_y^2} \end{aligned}$$

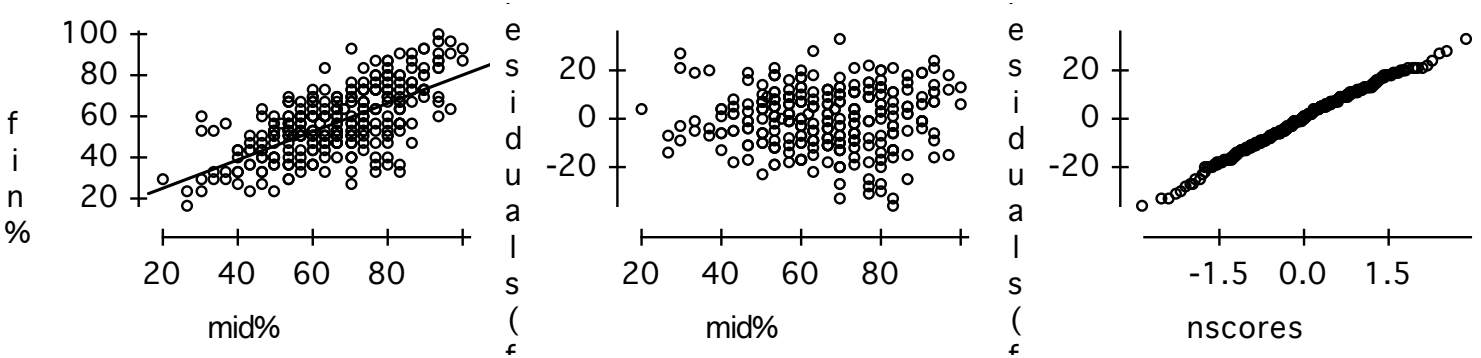
- R^2 is slightly larger than $R^2(\text{adjusted})$. Both measures are applicable for general linear regression models (with one or more predictors). When there is just one predictor, we have $R^2 = r^2$.

- Belgravia: differences in property tax account for 60.9% of the variability in sales price. Is that R^2 value large?

- How well do "facial ratios" predict penalty minutes in NHL hockey?

- We looked at the regression of `sale` on `tax`. We could also regress `tax` on `sale`.
  - Do the two regressions produce the same line?
  - If not, why not? How are the two lines related and where do they cross?
  - Hint: consider the standardized variables.

Midterm and Final marks (percent) combined from three STAT 141 classes.



Dependent variable is: **fin%**
No Selector
309 total cases of which 1 is missing
R squared = 44.0% R squared (adjusted) = 43.8%
s = 12.14 with 308 - 2 = 306 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	35444.7	1	35444.7	241
Residual	45094.5	306	147.368	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	11.8562	3.019	3.93	0.0001
mid%	0.688883	0.0444	15.5	≤ 0.0001

Summaries
No Selector

Variable	Count	Mean	StdDev
fin%	308	57.4350	16.1970
mid%	308	66.1632	15.5977

Examples from the text:

- Atlantic hurricanes: regress Max wind speed (mph) on Central pressure (mb).

- Income and housing: regress Housing Cost Index on Median family income (per state). See page 218, Ex 35.

9. Regression wisdom

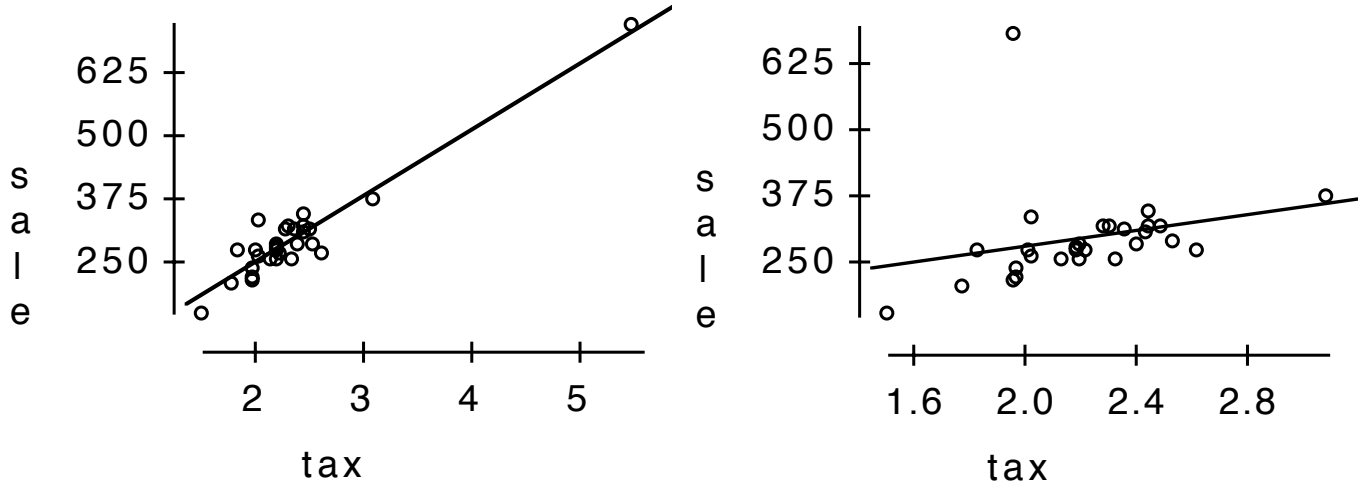
- Getting the "bends": when the residuals aren't straight.
  Fetal weight, birth weight, and gestational age.

  - Does it make sense to regress `BirthWt` on `GestAge`?
  - Does it make sense to regress `FetalWt` on `GestAge`?
  - Does it help to transform one or both variables?
  - Compare rates of growth for `GestAge` > 37.

- Extrapolation: Reaching beyond the data.
  Investment caveat: "Past performance is no guarantee of future results."
  Fetal growth data.

- Working with summary values.
  Birth weight and gestational age.

- Outliers, leverage, and influence.
  Two properties were omitted from the Belgravia data: (5.476, 720) and (1.962, 683). What effect would their inclusion have on the least squares regression line?



- Sifting residuals for groups. Determined by other variables?
  Regression lines for different subsets.
  Cigarettes data: regress tar on carbon monoxide.

- Lurking variables and causation, fallacies.
  Post hoc ergo propter hoc (after this, therefore because of this).
  Golden Palace Restaurant.

- Bird songs: 587 MPEG audio files from *Smithsonian Field Guide to the Birds of North America*. Regress file size (MB) on time length (minutes).