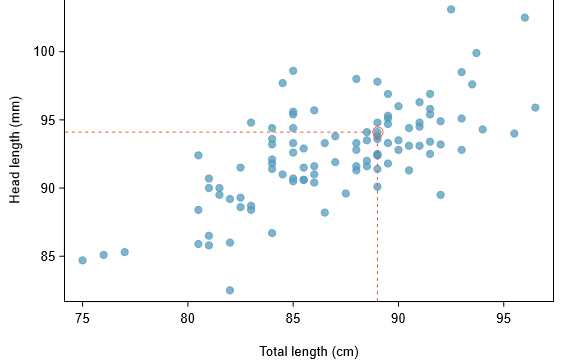
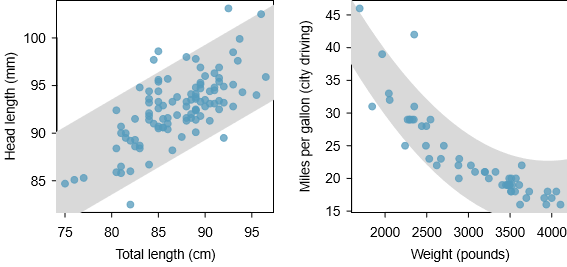
# Chapter 5: Introduction to linear regression

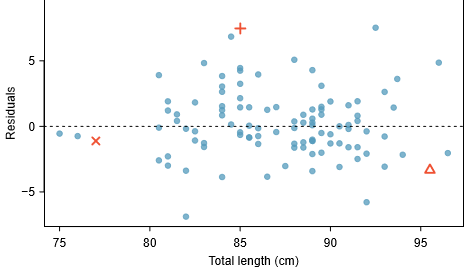
* Linear regression = very powerful statistical technique where we try to model relationships w/ a straight line
* perfect linear relationship = know exact value of y just by knowing value of x (unrealistic)
* Family income X, would provide useful info about how much ﬁnancial support Y a college may oﬀer a prospective student but there’d still be variability in ﬁnancial support, even when comparing students w/ families w/ similar ﬁnancial backgrounds.
* **Linear regression *assumes* relationship between variables can be modeled by straight line: y = β0 + β1x** where β0 + β1 = model parameters estimated using data (written as **point estimates** = b0 + b1)

1. 

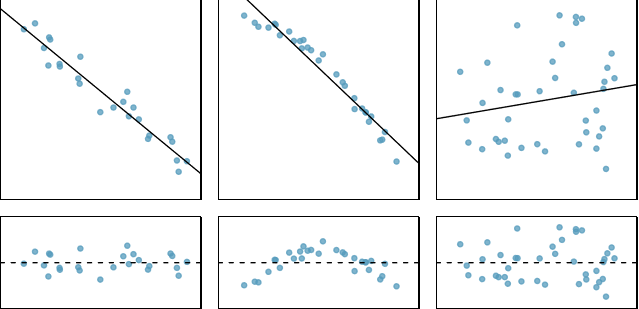
* head length + total length of 104 brushtail possums from Australia are associated.
* above average total length = tend to have above average head lengths.
* relationship is not perfectly linear, but could be helpful to partially explain connection between these variables w/ a straight line.
* **Straight lines should only be used when data appear to have a linear relationship** vs. a curved line

1. 

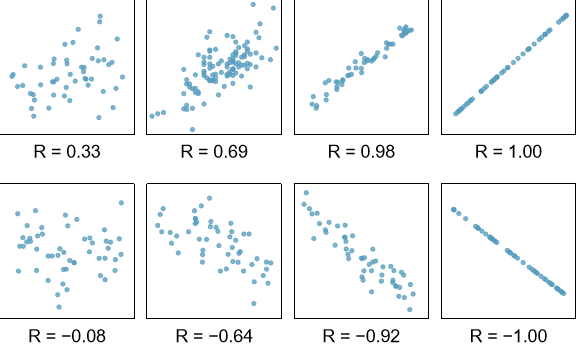
* If data show a nonlinear trend, use more advanced/complex, non-linear techniques
* We want to describe relationship between head length + total length using a line w/ total length as predictor, x, to predict head length, y 🡺 **y^ = 41 + 0.59x**
* equation predicts possum w/ total length = 80 cm will have a head length = **y^ = 41 + 0.59\*80 = 88.2**
* **y^ estimate** may be viewed as an **average** 🡺 predict possums w/ total length = 80 cm will have **an average head length = 88.2 mm**.
* **\*\*\*Absent further info about an 80 cm possum, prediction for head length that uses the average is a reasonable estimate\*\*\***
* **Residuals** = leftover variation in data after accounting for model ﬁt: **Data = Fit + Residual**
* **=** vertical distance from observation to the line, is positive (Each observation has a residual)
* If observation above regression line, residual = positive, if below the line, residual = negative
* **\*\*\*Goal in picking right linear model = have residuals to be as small as possible, in terms of its absolute value\*\*\***
* Residual = diﬀerence between observed + expected
* residual of the ith observation (xi, yi) = diﬀerence of observed response (yi) + response predicted based on model ﬁt (yi^ = identify by plugging xi into model) 🡺 
* Guided Practice 5.4: If a model underestimates an observation, will the residual be positive or negative? What about if it overestimates the observation?
* If underestimated true value, residual = positive, if overestimated, residual = negative
* Guided Practice 5.5 Compute residuals for observations (85.0, 98.6) + (95.5,94.0) using a linear relationship y^ = 41 + 0.59x
* 98.6 - 91.15 = 7.45 94 - 97.345 = -3.345
* Residuals = helpful in evaluating how well linear model ﬁts data set, often displayed in **residual plot**

1. 

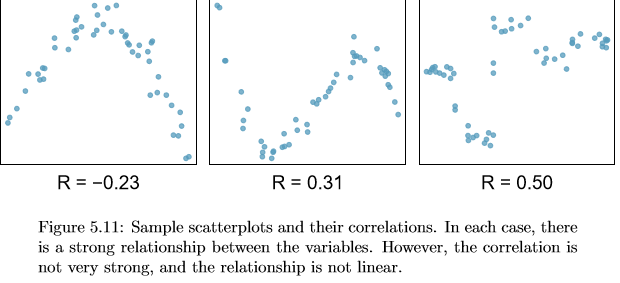
* Residuals are plotted @ original horizontal locations but w/ vertical coordinate = residual.
* residual plot = sort of like tipping scatterplot over so regression line is horizontal
* **\*\*\*1 purpose of residual plots** = **ID** characteristics/**patterns** **still apparent in data after ﬁtting a model**\*\*\*

1. 
   * **Middle model’s residual plot has upside-down U pattern = non-linear structure of true *f***
   * 1st + 3rd datasets’ residuals = no obvious patterns + appear to be scattered randomly around dashed line that = 0
   * But, 3rd plot = very little upwards trend, so reasonable to try to ﬁt a linear model to the data.
   * However, unclear if there’s statistically signiﬁcant evidence the slope parameter is diﬀerent from 0.
   * **\*\*\*Point estimate of the slope parameter, b1,** is NOT = 0, but can wonder if this could just be due to chance\*\*\*

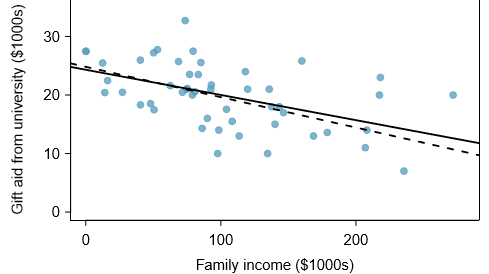
* **Correlation (r)** = strength of linear relationship (always between -1 and 1) between 2 variables.
* Can compute correlation via formula, but it’s complex () so generally perform calculations on a CPU

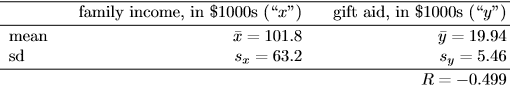
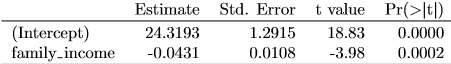
1. 

* **correlation** = intended to **quantify the strength of a linear trend**.
* **\*\*\*Nonlinear trends, even when strong, sometimes produce correlations that do not reﬂect the strength of the relationship\*\*\***

1. 

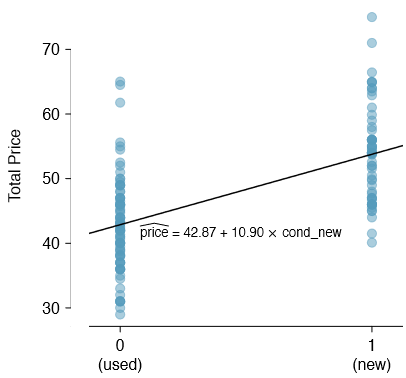
* Guided Practice: It appears no straight line would ﬁt any of the datasets represented above. Instead, try drawing nonlinear curves on each plot. Once you create a curve for each, describe what is important in the ﬁt.
  + Curve should be close to the DP’s such that residuals are minimized
* **Least squares regression** = more rigorous approach.

1. 

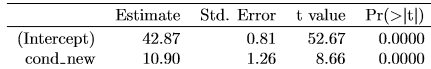
* family income + gift aid data from a random sample of 50 students in ‘11 freshman class of Elmhurst College in Illinois
* Negative trend/correlation🡺 students w/ higher family incomes tend to have lower gift aid
* Begin by thinking about “**best**” 🡺 Mathematically want a line w/ small residuals, so criterion could *minimize the sum of the residual magnitudes:*  (dashed line in above plot)
* More common = **choose line that minimizes SUM OF SQUARED RESIDUALS**: (solid line)
* most commonly used method.
* Computing this line = much easier by hand + in most statistical software.
* **\*\*\*In many applications, residual 2x as large as another residual == *more* than 2x as bad + squaring the residuals accounts for this discrepancy\*\*\***
* \*\*\*There ARE applications where *sum of the residual magnitudes* may be more useful, + there are plenty of other criteria we might consider\*\*\*
* Can write equation of least squares regression line as 
* Other than estimating parameters using observed data, can apply \*\*\*2 properties of the least squares line:\*\*\*
* Slope of least squares line can be estimated by  where sx + sy = sample SD’s
* **If x­- = mean of the predictor + y¯ = mean of outcome, point (x¯, y¯) is ON the least squares line**
* Use b­­0 + b1 as **point estimates** of parameters β0 + β1.. However, this book only applies the least squares criterion.
* Guided Practice:  🡺 How to plot point (101.8,19.94) on to verify it falls on the least squares line
  + **draw straight line up from x-value = ~101 + draw a horizontal line @ y = ~20 + these lines should intersect on the least squares line**
* Guided Practice: Using summary statistics, compute slope for regression line of gift aid against family income
  + **-.499\*(5.46/63.2) == -0.04310981012**
* Given slope + point on a line, (x0,y0), equation for line can be written as
* Common exercise to become more familiar w/ foundations of least squares regression = use basic summary stats + point-slope form to produce least squares line.
* TIP: IDing least squares line from summary statistics:
  + **\*\*\*Estimate slope parameter, β1, by calculating b1 using \*\*\***
  + \*\*\*Noting point (x¯, y¯) is ON least squares line, use x0 = x¯ and y0 = y¯ along w/ slope b1 in the point-slope equation: \*\*\*
  + Simplify the equation.
* \*\*Ex\*\*: Using point (101.8,19.94) from the **sample means** + slope estimate b1 = −0.0431, ﬁnd least-squares line for predicting aid based on family income.
* Apply **point-slope equation** using the DP + slope b1:  = 
* Expanding RHS + adding 19.94 to each side to simplify: 
* CPU output: 
  + 1st row = β0, 2nd row = β1, \*\*\***1st col = point estimate for β1, 2nd col, 2nd row = standard error for a point estimate** (0.0108), **3rd col = t-test statistic for the null H0: β1 = 0** (−3.98), **last col = p-value for the t-test statistic for the null H0: β1 = 0 + the 2-sided H1 (0.0002)**
* Suppose a high school senior is considering Elmhurst College. Can she simply use the linear equation that we have estimated to calculate her ﬁnancial aid from the university?
  + \*\*\*May use it as an **ESTIMATE**, though some qualiﬁers are important 🡺 data all come from 1 freshman class, the way aid is determined by the university may change year to year\*\*\*
  + Also, \*\*\***equation will provide an imperfect estimate**\*\*\* = linear equation is good at capturing the trend in the data, but no individual aid will be perfectly predicted
* **\*\*\*Interpreting parameters** in a regression model = often one of the most important steps in the analysis\*\*\*
* Example 5.18: Slope + intercept estimates for Elmhurst data = -0.0431 and 24.3. What do these numbers really mean?
* Family income = $0 predicts **AVERAGE** financial aid = $24.3K, + for each $1K increase in family income, financial aid would **ON AVERAGE** decreases in -.0431 financial aid units ($1K) = −$43.10
* higher family income corresponds to less aid b/c coeﬃcient of family income < 0 in model.
* \*\*\*Must be cautious in this interpretation: while there is a *real* association, **we cannot interpret a causal connection between the variables because these data are observational\*\*\***
  + **Increasing a student’s family income may not cause the student’s aid to drop.** (reasonable to contact + ask if relationship is causal/if aid decisions are partially based on family income)
* **\*\*\*Meaning of the intercept is relevant to this application since family income for some students at Elmhurst is $0\*\*\***
* **\*\*\*In other applications, intercept may have little/no practical value if there are no observations where x is near 0\*\*\***
* Interpreting parameters estimated by least squares: **Slope** describes estimated diﬀerence in outcome if predictor for a case happened to be 1 unit larger. **Intercept** describes average outcome if predictor = 0 + **that the linear model is valid all the way to x = 0** (in many applications = not case)
* **\*\*\*Extrapolation is treacherous\*\*\***
* Linear models have real limitations 🡺 it’s simply a **modeling framework** + the **truth = almost always much more complex** than simple line (don’t know how data outside of training will behave)
* Use aid model income to estimate aid of student w/ family income = $1 million (Recall units of family income = $1K, so calculate aid as 
* Model predicts student will have -$18,800 in aid, i.e. pay extra on top of tuition to attend
* **Extrapolation** = applying a model estimate to values outside the realm of the original data
* Generally, a **linear model is only an *approximation*** of **real relationship** **between** **variables**.
* \*\*\*Extrapolating = making an \*\*\***unreliable**\*\*\* bet that the *approximate* linear relationship will be valid in places where it has not been explored\*\*\*
* Compared to R, more common to explain strength of a linear ﬁt using, R-squared, R2
* If provided a linear model, might like to describe how closely data cluster around linear ﬁt + **R2 describes amount of variation in the response explained by the least squares line**.
* Ex: Aid data 🡺 variance of response = 
* Applying our least squares line, model reduces uncertainty in predicting aid using a student’s family income.
* **\*\*\*Variability in *residuals* describes how much variation remains *after* using model\*\*\***  == In short, there was a reduction of:

1. 

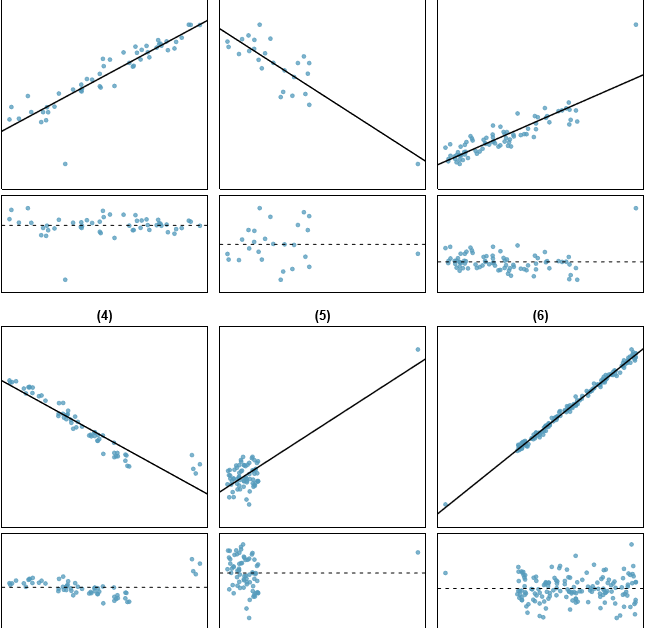
* i.e. we get ~25% reduction in data’s variation by using info about family income for predicting aid *using a linear model.*
* This corresponds exactly to R2 = 0.25, compared to R = −0.499
* Guided Practice: If a linear model has a very strong negative relationship w/ correlation r = -0.97, how much variation in the response is explained by the explanatory variable?
* Compute R2 🡪 -.972 = .9409 🡺 ~94% of variation in response is explained by model
* **Categorical variables** are also useful in predicting outcomes.
* Ex: categorical predictor w/ 2 levels 🡺 Ebay auctions for a video game, Mario Kart for Nintendo Wii, where both total price of auction + condition of the game were recorded
* Want to predict total price based on **game condition = used and new**.

1. 

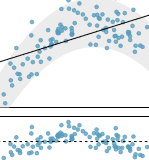
* To incorporate game condition into a regression, **convert the categories into a numerical form** using an **indicator (dummy) variable** called condnew, which takes value = 1 when game is new + 0 when game is used.
* Using this indicator variable, linear model is 

1. 

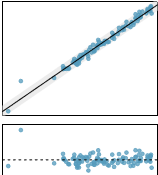
* Interpret the 2 parameters estimated in the model for the price of Mario Kart in eBay auctions.
  + Intercept value of 42.87 estimates that, on average, used games selling price = $42.87, and if new, on average, new games selling price = $42.87 + $10.9 = $53.77
  + Coefficient for indicator says on average, new games sell for ~$10.90 more than used games
* **\*\*\*TIP: Interpreting model estimates for categorical predictors. Estimated intercept = value of response for 1st category (i.e. category corresponding to indicator value = 0) + estimated slope = avg. change in response between the 2 categories\*\*\***
* **Outliers** in regression = observations that fall **far** from the “cloud” of points
* **\*\*\*Outliers** = especially important b/c they **can have a strong inﬂuence on least squares line\*\*\***
* 6 plots, each w/ fit + residuals (each dataset has *at least 1* outlier)

1. 

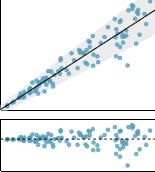
* (1) 1 outlier, only appears to *slightly* inﬂuence the line
* (2) 1 outlier, though quite close to least squares line = suggests not very inﬂuential
* (3) 1 outlier, appears to pull least squares line up on the right 🡺 see how line around the primary cloud doesn’t appear to ﬁt very well
* (4) 4 outliers, appears to be inﬂuencing line somewhat strongly (pulling upwards), making least square line ﬁt poorly almost everywhere (**might be interesting explanation 🡪 investigate**)
* (5) No obvious trend in main cloud + **outlier appears to largely control the slope** 
  + *data w/ no clear trend were assigned a line w/ a large trend simply due to one outlier*
* (6) 1 outlier far from the cloud, but falls quite close to line + doesn’t appear to be very inﬂuential
* **\*\*\*Leverage Points** = fall *horizontally* away from the center of the cloud + **tend to pull harder on the line 🡺 call them *high leverage points\*\*\****
* **High leverage points can strongly inﬂuence slope of a least squares line.**
* **Inﬂuential Point 🡺** ahigh leverage DP *DOES* appear to inﬂuence slope of line (cases (3), (4), (5))
* \*\*\*Usually can say a point is **inﬂuential** if, *had we ﬁtted the line without it*, the inﬂuential point would have been unusually far from the least squares line\*\*\*.
* \*\*\*Tempting to remove outliers 🡺 **Don’t do this without a very good reason**\*\*\*
* Models that ignore exceptional (+ interesting) cases often perform poorly.
* Ex: If financial ﬁrm ignored largest market swings (“outliers”), they’d soon go bankrupt by making poorly thought-out investments.
* Caution: **Don’t ignore outliers when ﬁtting a ﬁnal model**
* Outliers should NOT be removed or ignored **without a good reason**.
* Whatever ﬁnal model is ﬁt to the data would NOT be very helpful if it ignores the most exceptional cases
* **Caution: Outliers for a categorical predictor w/ 2 levels Be cautious about using categorical predictor when 1 level = very few observations as the few observations become inﬂuential points**
* Inference for linear regression
* **There is uncertainty** in estimates of slope + y-intercept for a regression line.
* **\*\*\*When performing inference on a least squares line, we generally require the following:\*\*\***
* **Linearity**: Data should show linear trend + if nonlinear, apply more advanced regressions

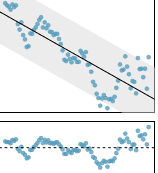
1. 

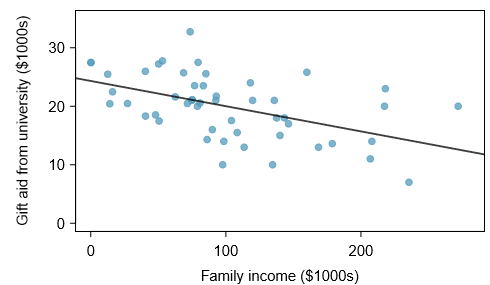
* **Nearly normal residuals**: When this = unreasonable, it is ***usually b/c of outliers or concerns about inﬂuential points***.

1.  = non-normal residuals due to 2 outliers

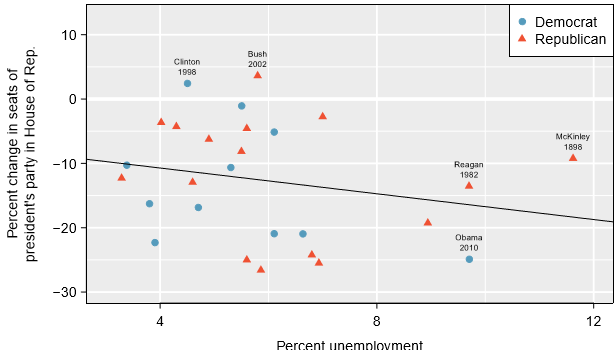
* **Constant variability** = *variability of points around least squares line remains roughly constant.*

1. 

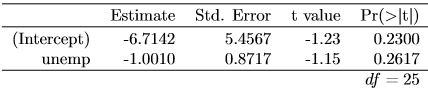
* **Independent observations**. Be cautious about applying regression to data collected *sequentially* in a **time series**
  + Such data may have an underlying structure that should be considered in a model/analysis
  + Ex: time series where independence is violated 🡺 
* Should we have concerns about applying inference to the Elmhurst data?

1. 
   * trend appears to linear, data fall around the line (no obvious outliers), variance = roughly constant
   * also NOT time series observations 🡺 would be reasonable to analyze model w/ **inference**

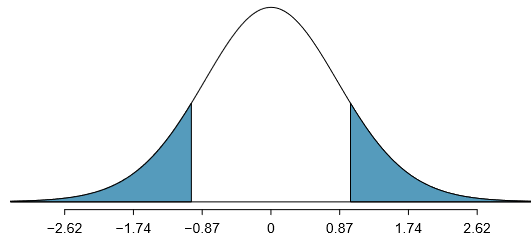
* House of Reps Elections occur every 2 years, coinciding every 4 years w/ Presidential election + House elections in middle of a Presidential term = **midterm elections**.
* In America’s 2-party system, 1 political theory suggests higher unemployment rate = worse President’s party will do in the midterm elections.
* To assess validity of this claim, compile historical data + look for a connection, considering every midterm election from 1898-2010, w/ exception of elections during Great Depression.

1. 
2.  

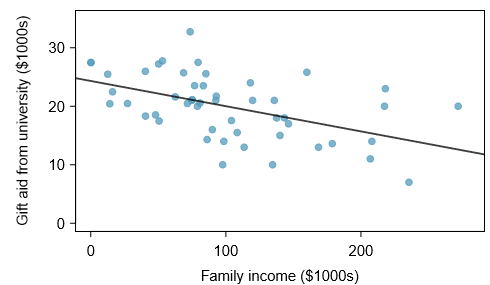
* Consider % change in # of seats of President’s party (e.g. % change # of seats for Democrats in 2010) against unemployment rate.
* Examining data 🡺 no clear deviations from linearity, constant variance condition, or normality of residuals (don’t examine a normal probability plot here) + while data are collected sequentially, a separate analysis made sure of no any apparent correlation between successive observations
* Guided Practice: Data for Great Depression (1934 + 1938) were removed b/c unemployment rate was 21% + 18%, respectively. Do you agree they should be removed for this investigation?
  + Yes, b/c very unusual numbers that could have large influence on model
  + Each of these DP’s would have **very high leverage** on ANY least-squares regression line, + **years w/ such high unemployment may not help us understand what would happen in other years where unemployment is only modestly high**
  + On the other hand, these are exceptional cases, + we’d be discarding important info if excluded from a ﬁnal analysis
* Negative slope but \*\*\*it (+ the y-intercept) are only **estimates** of the parameter values\*\*\*.
* Is this convincing evidence that the “true” linear model ***f*** has a negative slope? 🡺 Do the data provide strong evidence the political theory is accurate?
* **\*\*\*Frame investigation into a 2-sided statistical hypothesis test**\*\*\*
  + *2-sided test since a statistically signiﬁcant result in either direction would be interesting*
* **H0: β1 = 0 🡺 true linear model has slope = 0.**
* **HA: β1 <> 0 🡺 true linear model has a slope diﬀerent than 0 + higher unemployment = greater loss for President’s party in the House, or vice-versa**
* **\*\*\*Reject H0 in favor of HA if data provide strong evidence the true slope parameter < 0\*\*\***
* To assess the hypotheses, we ID a **standard error for the estimate**, **compute** an appropriate **test statistic,** + **ID the p-value**
* Generally label test statistic w/ ***t***, since it **follows the t-distribution**.

1. 

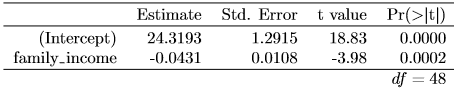
* Ex: What do the 1st + 2nd columns represent?
* **1st col = least squares *estimates*, b0 + b1, of the parameters β0 + β1**
* **2nd col = standard errors of *each estimate***
* In the hypotheses, null value for slope = 0, so compute the test statistic using the T (or Z) score formula: 
* Look for the 2-tailed p-value using the probability table for the t-distribution

1. 
   * **See sampling distribution for b1, \*\*\**if the null hypothesis was true*\*\*\*.**
   * Shaded tail = p-value for the hypothesis test evaluating for convincing evidence higher unemployment corresponds to greater loss of House seats for President’s party in midterms

* dF = 25 gives absolute value of the test statistic (1.15) is smaller than any value listed 🡺 means the tail area + therefore also the p-value *is larger* than 0.200 (2 tails).
* B/c p-value is so large, we **fail to reject the null**.
* That is, the data do NOT provide convincing evidence that unemployment = good predictor of how well a president’s party will do in midterms for the House
* Can also ID the t-test statistic from the software output 🡺 2nd row (unemp) + 3rd column (t-value).
* 2nd row + last column = p-value for the 2-sided hypothesis test where the null value = 0
* **Inference for regression**: Usually rely on statistical software to ID **point estimates**+ **standard errors** for parameters of a regression line.
* After verifying conditions hold for ﬁtting a line, can use the methods for the t-distribution to create CI’s for regression parameters or to evaluate hypothesis tests.
* \*\*\***Caution: Don’t carelessly use the p-value from regression output**\*\*\*
* last column in regression output often lists p-values for *one particular hypothesis* = 2-sided test where null value = 0.
* If hypothesis test should be ONE-sided or a **comparison** is being made to *a value other than zero*, be cautious about using the software output to obtain the p-value
* Ex: How sure are we the slope is statistically signiﬁcantly diﬀerent from 0 for Elmhurst data? That is, do we think a formal hypothesis test would reject the claim that true slope of the line should be = 0?

1. 

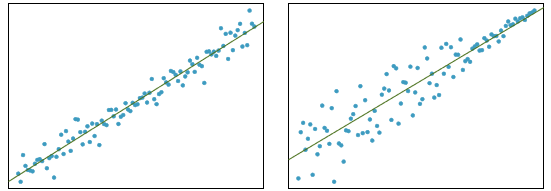
* While relationship is not *perfect*, there IS an evident decreasing trend in the data 🡺 suggests the hypothesis test will reject the null claim that slope = 0
* Guided Practice: Use below to formally evaluate the following hypotheses.
* H0: True coeﬃcient for family income = 0 HA: True coeﬃcient for family income is NOT zero

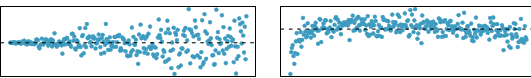
1. 

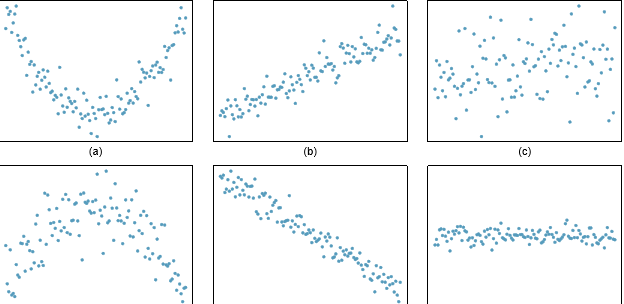
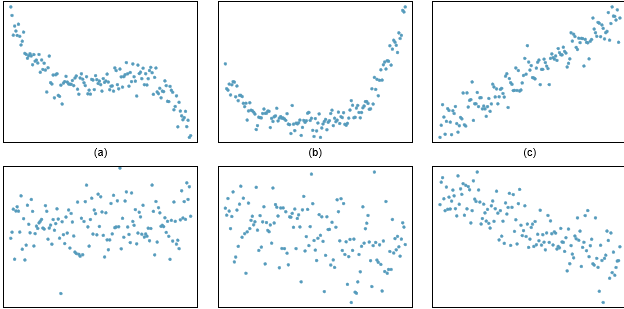
* Point estimate of slope of line = -0.0431, standard error of this estimate = 0.0108, + t-test statistic = -3.98
* **p-value corresponds exactly to the 2-sided test we’re interested in 🡺 0.0002 = so small that we reject the null + conclude family income + ﬁnancial aid at Elmhurst College for freshman entering in the year 2011 are negatively correlated + true slope parameter is indeed < 0,**
* **TIP: Always check assumptions** 🡺 If conditions for a regression line do NOT hold, methods presented above should NOT be applied.
* **Standard error/distribution assumption of the point estimate (assumed to be normal when applying the t-test statistic) may not be valid.**
* We considered t-statistic as a way to evaluate strength of evidence for a hypothesis test
* Could focus on R2 (proportion of variability in response explained by predictor)
* If this R2 proportion is large 🡺 suggests linear relationship exists between the variables.
* If small 🡺 evidence provided by the data may not be convincing.
* Concept of **considering amount of variability in response explained by predictor** is a key component in some statistical techniques.
  + **Analysis of variance (ANOVA)** uses this general principle.
* The method states that **if enough variability is explained away by the categories, we’d conclude the mean varied between the categories.**
* **On the other hand, we might NOT be convinced if only a little variability is explained.**
* ANOVA can be further employed in advanced regression modeling to evaluate the of explanatory variables

# 5.5 Exercises

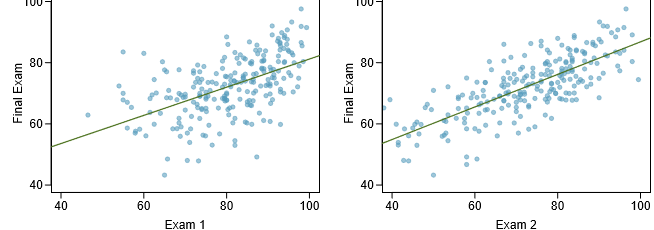
* Visualize the residuals: Scatterplots below each have a superimposed regression line. If we were to construct a residual plot (residuals vs. x) for each, describe what those plots would look like.

1. 
   * Plot 1) **Residuals** = cloud of points huddled around horizontal line = 0 w/ constant variance, as this regression line is quite close to majority of DP’s
   * Plot 2) Residuals = more scattered around horizontal line = 0 for lower values of X, huddling close together as X increases (tail-end of plot) == fan-shape pattern == bad model fit

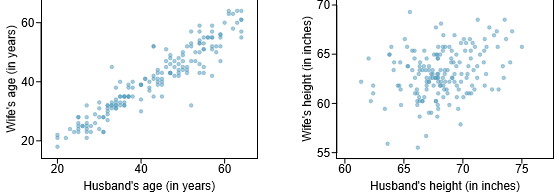
* Trends in residuals. Shown below = 2 plots of residuals remaining after ﬁtting a linear model to 2 diﬀerent sets of data. Describe important features + determine if a linear model would be appropriate for these data. Explain reasoning.
  + 
  + Plot 1) Fan-shape = a pattern, so linear model = not a good choice == **Strong relationship, but it’s non-linear**
  + Plot 2) DP’s have mostly constant variance after initial underestimation in lower values of X
* ID relationships:. For each plot, ID the relationship strength + whether ﬁtting a linear model would be reasonable

1. 
   * 1 = strong, non-linear 2 = moderately strong, linear 3 = weak*, could use* linear
   * 4 = weak-moderate, non-linear 5 = strong, linear 6 = no/weak relationship, linear
2. 
   * 1 = strong, non-linear 2 = strong, non-linear 3 = strong, linear
   * 4 = moderate, linear could work 5 = weak, linear could work 6 = moderate, linear

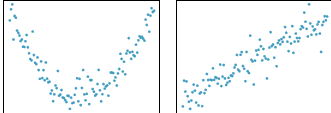
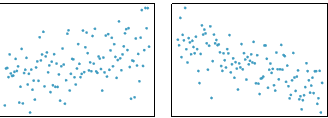
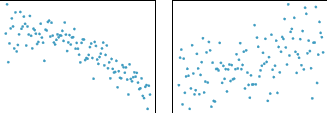
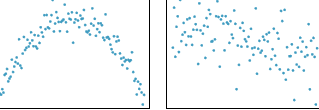
* 2 scatterplots show relationship between ﬁnal + mid-semester exam grades recorded during several years for a course at a university.

1. 
   * Based on these graphs, which of the 2 exams has the strongest correlation w/ the ﬁnal?
     + **Exam 2, as the DP’s are more closely aligned to the regression line**
   * Can you think of a reason why correlation between exam chosen above + the ﬁnal is higher?
     + **Further on in course = more knowledge of what final exam entailed**

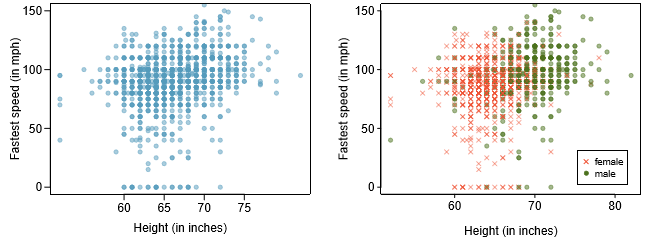
* The Great Britain Oﬃce of Population Census + Surveys once collected data on a random sample of 170 married couples in Britain, recording **age** (in years) + **heights** (converted here to inches) of the husbands + wives. Left scatterplot = wife’s age plotted against husband’s age, Right = wife’s height plotted against husband’s height.

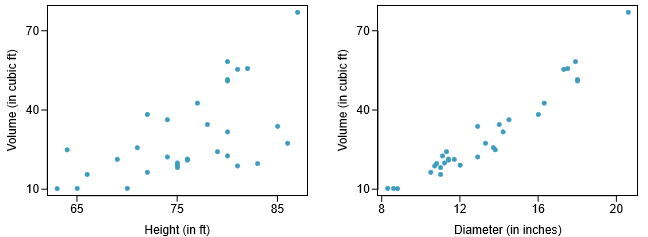
1. 
   * (a) Describe the relationship between husbands’ and wives’ ages.
     + Strong, positive relationship (as husband ages, wife ages linearly)
   * (b) Describe the relationship between husbands’ and wives’ heights.
     + No real relationship, possibly weak linear
     + As husband’s grow taller, it’s not necessarily true that wives do as well
   * (c) Which plot shows a stronger correlation? Explain your reasoning.
     + Plot 1 = more closely aligns with a line, plot 2 = more of a random scatter
   * (d) Data on heights were originally collected in cm, + then converted to inches. Does this conversion aﬀect the correlation between husbands’ and wives’ heights?
     + **\*\*\*No, Change of units will NOT impact mathematical relationship\*\*\***

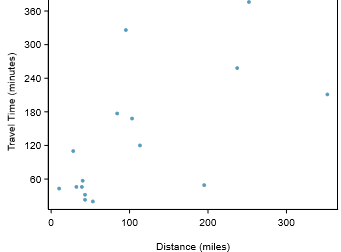
* Match calculated correlations to corresponding scatterplot.

1.  
   * (a) R = −0.7 (b) R = 0.45 (c) R = 0.06 (d) R = 0.92
   * **1 = c 2 = 4 3 = b 4 = a**
2.  
   * . (a) R = 0.49 (b) R = −0.48 (c) R = −0.03 (d) R = −0.85
   * **1 = d 2 = a 3 = c 4 = b**

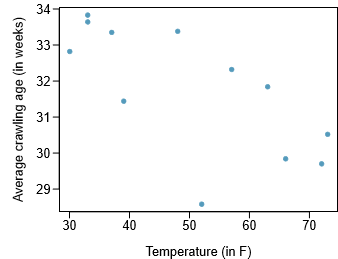
* 1302 students ﬁlled out a survey, asked about height, fastest speed ever driven, + gender.
* Left scatterplot = relationship between height + fastest speed, Right scatterplot = breakdown by gender in this relationship.

1. 
   * Describe the relationship between height and fastest speed
     + Seems relationship is quite **weak**, but **positive**, but w/ *mayyybe* some influence:
       - taller = more likely to drive faster
     + Many students have not driven a car (DP’s = 0 for a value of Y (speed))
     + Have some leverage DP’s = low height, all the same (error?) + some very tall
   * Why do you think these variables are positively associated?
     + No real clear reason in 1st plot
   * What role does gender play in the relationship between height and fastest driving speed?
     + **\*\*\*2nd plot = reveals CONFOUNDER of gender 🡪 taller = more male, males = more likely to drive fast**
     + **Seems a binary classification line can bisect the cloud into males vs. females w/ males = higher speeds\*\*\***
     + **Males > females on avg. for height, also for speed (anecdotal)** 
       - **\*\*\*sociological studies conﬁrm this anecdote\*\*\***

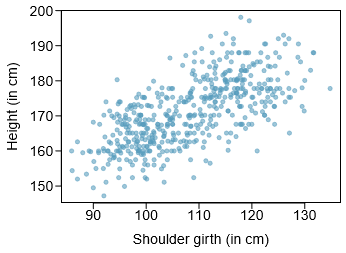
* Scatterplots show relationship between height, diameter, + volume of timber in 31 felled black cherry trees. The diameter of the tree is measured 4.5 feet above the ground
  + - * 
  + (a) Describe the relationship between volume + height or diameter of these trees.
    - Volume + height = weak/moderate, linear, positive
    - Volume + diameter = linear, strong, positive
  + (b) Suppose you have height + diameter measurements for another black cherry tree. Which variable would be preferable to use to predict volume of timber in this tree using a simple linear regression model? Explain your reasoning.
    - Diameter 🡺 relationship has larger correlation (more closely resembles a line)
* The Coast Starlight Amtrak train runs from Seattle to LA. Scatterplot displays distance between each stop (in miles) + amount of time it takes to travel from 1 stop to another (in minutes).

1. 
   * (a) Describe the relationship between distance and travel time.
     + Very weak, could use linear 🡺 **\*\*\*Note: small cluster in lower-left\*\*\***
     + **Did not expect such a weak relationship TBH**
   * (b) How would the relationship change if travel time was instead measured in hours, + distance was instead measured in KM?
     + Would decrease range of y-axis, increase x-axis
     + **\*\*\*Would NOT change form, direction, or strength of relationship\*\*\***
       - *If longer distances in mi. = associated w/ longer travel time in min., longer distances in KM would be associated w/ longer travel time in hours.*
   * (c) Correlation between travel time (in mi.) + distance (in min.) is R = 0.636. What is correlation between travel time (in km) + distance (in hours)?
     + **\*\*\*Same (changing units does NOT affect correlation)\*\*\***

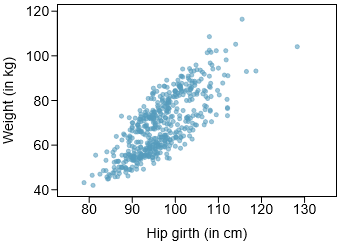
* University of Denver study investigated whether babies take longer to learn to crawl in cold months, when often bundled in clothes that restrict movement, then in warmer months. Infants born during study year were split into 12 groups, 1/month. Consider average crawling age of babies in each group against average temp when babies = 6 months old (often when begin trying to crawl). Temp = measured in ◦F + age in weeks

1. 
   * (a) Describe the relationship between temperature and crawling age.
     + Appears to be a moderate, negative correlation/relationship
     + *Suggests* as temps increase, **on average,** average crawling age in weeks would decrease, supporting above hypothesis
     + *1 outlier of very low average crawling age in weeks @ around low 50’s F*
   * (b) How would the relationship change if temp was measured in ◦C + age in months?
     + **changing units does NOT affect correlation**
   * (c) Correlation between temp in ◦F + age in weeks was R = −0.70. If we converted temp to ◦C + age to months, what would the correlation be?
     + **Same (changing units does NOT affect correlation)**

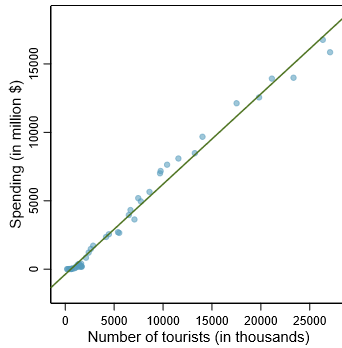
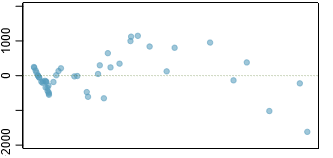
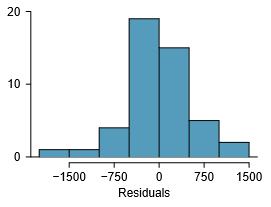
* Researchers studying anthropometry collected body girth + skeletal diameter measurements, as well as age, weight, height, + gender for 507 physically active individuals. Scatterplot shows relationship between height + shoulder girth (over deltoid muscles), both measured in cm.

1. 
   * (a) Describe the relationship between shoulder girth and height.
     + Quite strong positive relationship = as girth increases, can ASSUME a higher height, (both in cm) *on average*
   * (b) How would relationship change if shoulder girth was measured in in. while units of height *remained* in cm?
     + **\*\*\*Same (changing units, \*\*\*even if just for 1 of variable\*\*\* does NOT affect correlation)\*\*\***

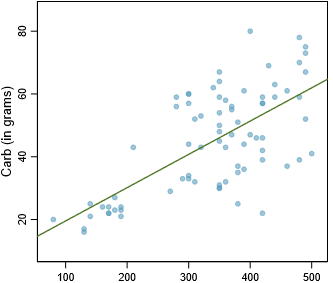
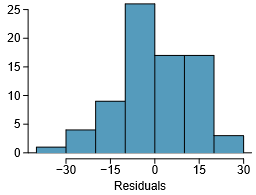
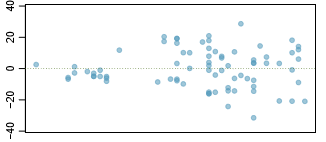
* See relationship between weight measured in kg + hip girth measured in cm from the data described

1. 
   * (a) Describe the relationship between hip girth and weight.
     + Quite a strong, positive relationship but some outliers appear to minorly influence relationship
   * (b) How would relationship change if weight was measured in lbs. while units for hip girth remained in cm?
     + **\*\*\*Same (changing units, \*\*\*even if just for 1 of variable\*\*\* does NOT affect correlation)\*\*\***

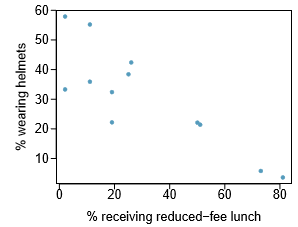
* What would the correlation be between ages of husbands + wives if men *always married woman who were*
  + (a) 3 years younger? (b) 2 years older? (c) half as old as themselves?
    - Model as: ageH = ageW + 3; (b) ageH = ageW −2; and (c) ageH = 2×ageW
    - **correlation will be exactly = 1 in all 3 parts**
    - \*\*\*as husband ages 1 year, so will wife, no matter starting age\*\*\*
* What would be the correlation between annual salaries of males + females at a company if for a certain type of position men always made
  + (a) $5k more than women? (b) 25% more than women? (c) 15% less than women?
    - Models: salaryM = salaryW + 5000; salaryM = 1.25\*salaryW; salaryM = .75\*salaryW
    - **Same absolute value for all (1), then for a) and b) = positive, for c) = negative**
* Association of Turkish Travel Agencies reports # of foreign tourists visiting Turkey + tourist spending by year. Scatterplot shows relationship between these 2 variables along w/ least squares ﬁt.

1. 
   * (a) Describe the relationship between number of tourists and spending.
     + Very strong positive relationship (more tourists = more tourist spending)
   * (b) What are the explanatory and response variables?
     + Explanatory = # of tourists in thousands, response = tourist spending in $1M’s
   * (c) Why might we want to ﬁt a regression line to these data?
     + To predict how much income country gets from tourist spending based on how many tourists we expect
     + Set aside money to advertise more tourism to get more tourists to come to get more income from tourism spending
   * (d) Do the data meet the conditions required for ﬁtting a least squares line? In addition to the scatterplot, use the residual plot + histogram to answer this question.
2. 
   * + The residuals **histogram’ looks normal**, which is required
     + The **residuals plot shows not visible pattern**, which suggests our residuals are independent and identically distributed (**IID**) BUT **\*\*\*residual plot *actually* shows a nonlinear relationship\*\*\***
       - \*\*\*This is NOT a contradiction 🡺 **residual plots can show divergences from linearity that can be diﬃcult to see in a scatterplot**\*\*\*
     + **Simple linear model = inadequate for modeling these data**.
     + **Must consider that *these data are observed sequentially* = means there may be a hidden structure that’s not evident in current data but is important to consider**

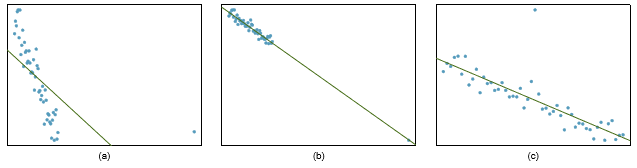
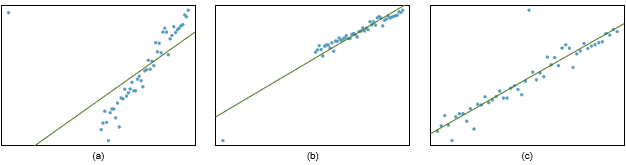
* See relationship between # of calories + amount of carbs (g) that Starbucks food menu items contain. Since Starbucks only lists # of calories on display items, we’re interested in predicting amount of carbs a menu item has based on calorie content.
  + (a) Describe relationship between # of cals + amount of carbs (g) in Starbucks food items

1. 
   * + Moderate to poor positive relationship
   * (b) In this scenario, what are the explanatory and response variables?
     + Explanatory = # of cals, response = amount of carbs in g
   * (c) Why might we want to ﬁt a regression line to these data?
     + To predict carbs we’d be consuming based on presented calorie counts
   * (d) Do these data meet the conditions required for ﬁtting a least squares line?
     +  
     + No pattern in **histogram**, appears mostly **normal**
     + **Residual** plot shows a **fan shape**, w/ higher variability for higher x, indicating non-linearity, so **linear models may not be best** for this data

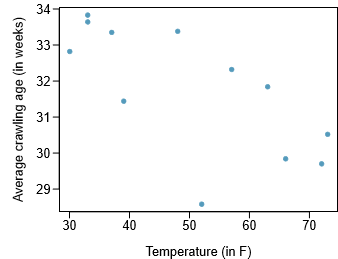
* Mean travel time on Coast Starlight Amtrak train from Seattle to LA from 1 stop to the next = 129 mins, w/ a SD = 113 mins. Mean distance traveled from 1 stop to the next = 107 mi. w/ SD = 99 mi. Correlation between travel time + distance = 0.636.
  + (a) Write the equation of the regression line for predicting travel time.
    - Calculate slope 🡺  🡺 (113/99)\*.636 = **.72593939394**
    - Noting that point (x¯, y¯) is ON least squares line, use x0 = x¯ and y0 = y¯ along w/ slope b1 in point-slope equation: **y-- = b0 + x­­--\*b1 🡺 b0 = 129 - .726\*107** = **51.318**
    - So, y = 51.318 +.726\*x 🡺 **travelTime = 51.318 + .726\*milesBetweenStops**
  + (b) Interpret the slope and the intercept in this context.
    - The mean average time between stops w/ a distance of 0 miles = 51.318 == NO REAL INTERPRETATION 🡺 just serves to set up “default” height on y-axis
    - For each mile between stops, travel time, on average, will increase by .726 minutes
  + (c) Calculate R2 of the regression line for predicting travel time from distance traveled for the Coast Starlight + interpret it in the context of the application.
    - R2 = given correlation squared = .6362 = **.404496**
    - This means ~40% of variability in mean travel time is explained by the distance between stops (accounted for by the model)
  + (d) Distance between Santa Barbara + Los Angeles is 103 miles. Use the model to estimate the time it takes for the Starlight to travel between these two cities.
    - 51.318 +.726\*(103) = **126.096**
  + (e) It actually takes Coast Starlight ~168 mins to travel from Santa Barbara to Los Angeles. Calculate the residual + explain the meaning of this residual value.
    - Residual = y – y^ = 168 - 126.096 = **41.904 = ~42** 🡺 positive residual = model underestimates travel time by 42 minutes
  + (f) Suppose Amtrak is considering adding a stop to that’s 500 miles away from LA. Would it be appropriate to use this linear model to predict travel time from LA to this point?
    - No, as this requires extrapolation 🡺 see original max x-value (distance) == ~350
* Mean shoulder girth = 108.20 cm w/ SD = 10.37 cm, + mean height = 171.14 cm w/ SD = 9.41 cm, + correlation between height + shoulder girth = 0.67.
  + (a) Write the equation of the regression line for predicting height.
    - Calculate slope 🡺  🡺 (9.41/10.37)\*.67 = **.60797492765988**
    - Noting point (x¯, y¯) is ON least squares line, use x0 = x¯ and y0 = y¯ along w/ slope b1 in point-slope equation: y­­-- = b0 + x--\*b1 🡺 b0 = 171.14 - .608\*108.2 **= 105.3544**
    - So, y = 105.3544 +.608\*x 🡺 **shoulderGirth = 105.3544 + .608\*height**
  + (b) Interpret the slope and the intercept in this context.
    - Height = 0 gives shoulder girth of 105 cm, which is meaningless in context 🡺 just serves to set up “default” height on y-axis
  + (c) Calculate R2 of the regression line for predicting height from shoulder girth, + interpret it in the context of the application.
    - R2 = given correlation squared = .6362 = **.4489**
    - This means about 45% of variability in mean shoulder girth is explained by height (is accounted for by the model)
  + (d) A randomly selected student from your class has a shoulder girth of 100 cm. Predict the height of this student using the model.
    - 105.3544 + .608\*(100) = **166.1544**
  + (e) The student from part (d) is 160 cm tall. Calculate the residual + explain what it means.
    - ei = y – yi = 160 – 166 = -6 🡺 negative residual = model overestimates height
      * Height is usually about 6 cm lower than predicted by model
  + (f) A 1 year old has a shoulder girth = 56 cm. Would it be appropriate to use this linear model to predict the height of this child?
* 21) Scatterplot shows relationship between socioeconomic status measured as % of children in a neighborhood receiving reduced-fee lunches at school (**lunch**) + % of bike riders in the neighborhood wearing helmets (**helmet**). Average % of children w/ reduced-fee lunches = 30.8% w/ a SD = 26.7% + average % of bike riders wearing helmets = 38.8% w/ SD = 16.9%

1. 
   * (a) If R2 for least-squares regression line for these data = 72%, what is the correlation between lunch and helmet? 🡺 **sqrt(.72) = r = 0.84852813742, which is negative b/c trend in plot is negative**
   * (b) Calculate slope + intercept for the least squares regression line for these data
     + Calculate slope 🡺  🡺 (16.9/26.7)\*-.849 = -**.53738202247**
     + Noting point (x¯, y¯) is ON least squares line, use x0 = x¯ and y0 = y¯ along w/ slope b1 in point-slope equation: y\_ = b0 + x\_\*b1 🡺 **b0 = 38.8 +.537\*30.8 = 55.3396**
     + So, y = **55.34** -.**537**\*x 🡺 **helmet = 55.34** -.**537**\***reducedFeeLunch**
   * (c) Interpret the intercept of the least-squares regression line in context of the application
     + Neighborhood having no students w/ reduced-fee lunches means ~55% of children in neighborhood wear helmets, which is meaningless in context 🡺 serves to set up “default” % on y-axis
   * (d) Interpret the slope of the least-squares regression line in the context of the application.
     + As the % of children receiving reduced-fee lunches increases, % of children wearing helmets decreases by ~1/2%
   * (e) What would the value of the residual be for a neighborhood where 40% of children receive reduced-fee lunches + 40% of bike riders wear helmets? Interpret the meaning of this residual in the context of the application
     + ei = y – yi = 40 – (**55.34** -.**537**\***40**) = 6.14 🡺 positive residual = model underestimates % of children who wear helmets
       - There are ~6% more children who wear helmets than predicted by model

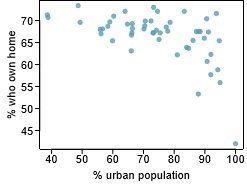
* 5.22 ID the outliers in the scatterplots + determine what type they are. Explain your reasoning.

1. 
2. 
   * A) + B) have **high leverage points** in both rows = horizontally far from center of the data
   * W/out outliers in both A)’s our regresisno line slope would be v. different = **influential pt.**
   * W/out outlier in 1st B), regression slope wouldn’t change, but would in 2nd B) = another influential pt.
   * For C)’s both points are not high leverage and also do not seriously affect our regression lines

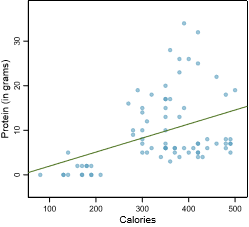
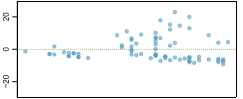
* Scatterplot of average monthly temp. during the month babies ﬁrst try to crawl (~6 months) + the avg. ﬁrst crawling age for babies born in a given month reveals a potential outlying month when average temp = ~53F◦ + average crawling age is ~28.5 weeks. Does this point have high leverage? Is it an inﬂuential point?

1. 
   * This point is not far from center of data = not high leverage
   * Does not appear that it would significantly change our regression line, but it could, but I’d say it’s not an influential point

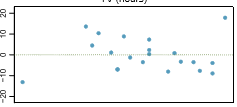
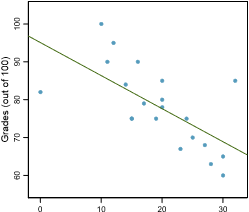
* See the % of families who own their home vs. % of population living in urban areas in 2010. There are 52 observations, each corresponding to a state in the US., Puerto Rico, + DC

1. 

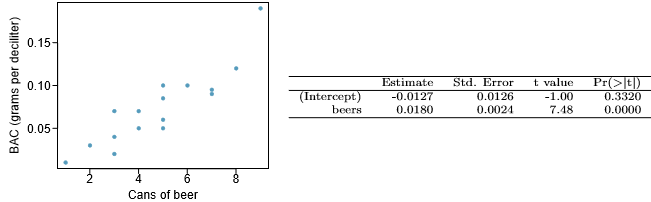
* (a) Describe the relationship between % of families who own their home + % of the population living in urban areas in 2010.
  + Appears moderately negative, possibly strong, w/ 1 outlier of 100% urban population
* (b) Outlier in bottom right = DC, where 100% of the population = considered urban. What type of outlier is this observation?
  + This is a high leverage point, and an influential point
* Back to nutrition info on Starbucks food menu items. Based on the scatterplot + residual plot provided, describe the relationship between protein content + calories of these menu items, + determine if a simple linear model is appropriate to predict amount of protein from # of calories.

1.  
   * Appears to have numerous outliers, and a weak positive relationship
   * Linear model would be bad, as seen from both plots (pattern in residual)

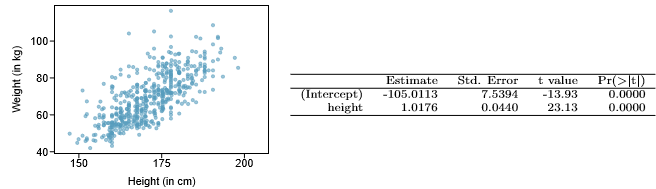
* 27) Data were collected on the # of hours/week students watch TV + grade they earned in a biology class on a 100-point scale. Based on the scatterplot + residual plot provided, describe the relationship between the 2 variables, + determine if a simple linear model is appropriate to predict a student’s grade from # of hours/week a student watches TV

1. 
   * Moderately strong negative relationship, no real pattern in residuals
   * Linear model may work

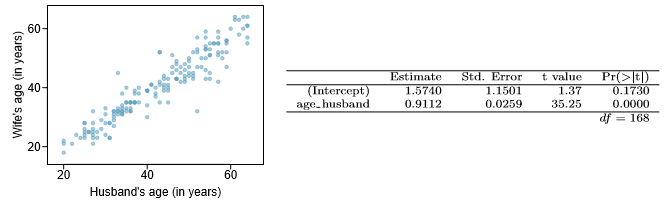
* 28) Many people believe gender, weight, drinking habits, + many other factors = much more important in predicting BAC than simply considering # of drinks consumed. Here we examine data from 16 student volunteers at OSU who drank a randomly assigned # of cans of beer. Students were evenly divided between men + women, + diﬀered in weight + drinking habits. 30 minutes later, BAC was measured in **g of alcohol per dL of blood**. The scatterplot + regression table summarize ﬁndings

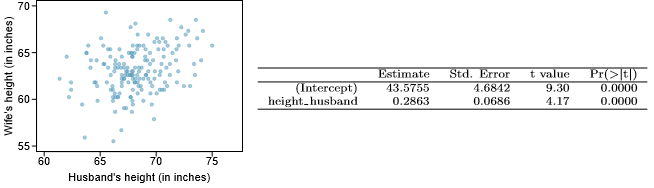
1. 

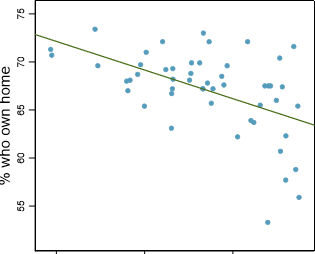
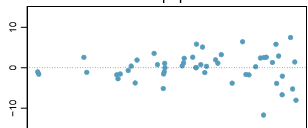
* (a) Describe the relationship between # of cans of beer + BAC.
  + As # of cans drank increases, as does BAC, with 1 potential outliers at 9 cans and a BAC = ~ .2
* (b) Write the equation of the regression line. Interpret the slope and intercept in context.
  + **y = -.0127 + .018\*cans\_of\_beer**
  + This model assumes we have a slight negative BAC before drinking any alcohol, on average, which is meaningless, and just serves to set the y-intercept
  + Then, the model assumes that for every 1 can of beer drank, our BAC increases by .180, all else held constant
* (c) Do the data provide strong evidence that drinking more cans of beer = associated w/ an increase in blood alcohol? State the null + alternative hypotheses, report the p-value, + state your conclusion.
  + H0 = There is no significant association between BAC and # of beers drank
    - true slope coeﬃcient of # of cans of beer is zero (β1 = 0).
  + H1 = There is a significant association between BAC and # of beers drank
    - true slope coeﬃcient of # of cans of beer > zero (β1 > 0)
      * A 2-sided test would also be acceptable for this application (β1 != 0)
  + Our reported p-value = .0000, which is indeed < .05, so we state that we can reject H0 and our data present evidence of a significant association between BAC and # of beers drank
* (d) correlation coeﬃcient for # of cans of beer + BAC = 0.89. Calculate R2 + interpret it in context
  + R2 = .892 = 0.7921 = ~79% of the variability in BAC = attributable to/explained by # of beers drank
* (e) Suppose we visit a bar, ask people how many drinks they have had + also take their BAC. Do you think the relationship between # of drinks + BAC would be as strong as the relationship found in the Ohio State study?
  + No, because this model is only fit to these specific data (training).
  + Our model could perform well on these new observations, but not as well as to this data
* 29) Scatterplot + least squares summary below show the relationship between weight measured in kg + height measured in cm of 507 physically active individuals

1. 

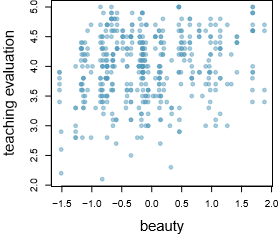
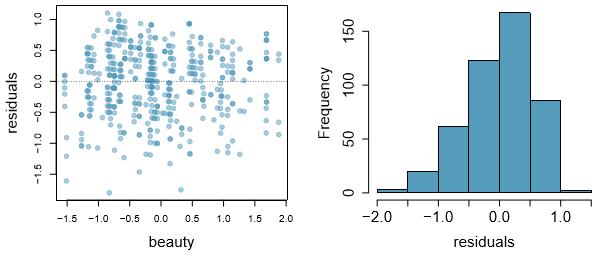
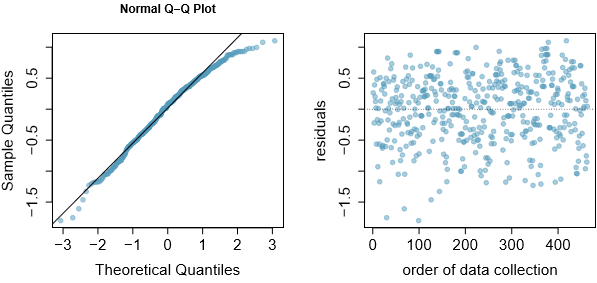
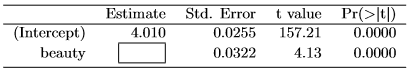
* (a) Describe the relationship between height + weight.
  + There is a moderate-to-strong correlation, with some *possible* outliers
* (b) Write the equation of the regression line. Interpret the slope and intercept in context.
  + **y = -105.0133 + 1.0176\*height**
  + This model assumes a weight of -105kg for a height of 0 cm, on average, which is meaningless, + just serves to set the y-intercept
  + Then, the model assumes that for every 1 cm grown, height increases by 1.0176 kg, all else held constant
* (c) Do the data provide strong evidence an increase in height is associated w/ an increase in weight? State the null + alternative hypotheses, report the p-value, and state your conclusion.
  + H0 = There is no significant association between height + weight of physically active individuals
    - true slope coeﬃcient of height is NOT zero (β1 != 0)
  + H1 = There is a significant association between height + weight of physically active individuals
    - true slope coeﬃcient of height is > zero (β1 > 0)
      * A 2-sided test would also be acceptable for this application (β1 != 0)
  + Our reported p-value = .0000, which is indeed < .05, so we state that we can reject H0 and our data present evidence of a significant association between height + weight of physically active individuals
* (d) The correlation coeﬃcient for height and weight is 0.72. Calculate R2 and interpret it in context.
  + R2 = .722 = 0.5184 = ~52% of the variability in weight = attributable to/explained by height
* 30. See the relationship between husbands’ + wives’ ages in a random sample of 170 married couples in Britain, where both partners’ ages < 65 years. Given below is summary output of the least squares ﬁt for predicting wife’s age from husband’s age.

1. 

* (a) We might wonder, is the age diﬀerence between husbands + wives consistent across ages? If this were the case, the slope parameter would be β1 = 1. Use the info above to evaluate if there is strong evidence the diﬀerence in husband and wife ages diﬀers for diﬀerent ages.
  + Our p-value for the association of husbands’ + wives’ ages is .0000, is indeed < .05, and our β1 is very near to 1, so we state that our data present evidence of a significant association between husbands’ + wives’ ages and that this age diﬀerence is quite consistent across ages
* (b) Write the equation of the regression line for predicting wife’s age from husband’s age.
  + **y = 1.5740 + .9112\*husband\_age**
* (c) Interpret the slope and intercept in context
  + This model assumes a wife age of 1.5740 for a husband age of 0, on average, which is meaningless, + just serves to set the y-intercept, but it could mean that the model assumes wives are ~1.5 years older than husbands on average, and the increase is consistent (β1 is close to 1)
  + Then, the model assumes that for every 1 year of age for husbands, wives age by .9112
* (d) Given that R2 = 0.88, what is the correlation of ages in this data set?
  + **R = .9381**, which is very strong
* (e) You meet a married man from Britain who is 55 years old. What would you predict his wife’s age to be? How reliable is this prediction?
  + **y = 1.5740 + .9112\*55 = 51.69,** and given our model evidence above, we can assume this estimate is reliable
* (f) You meet another married man from Britain who is 85 years old. Would it be wise to use the same linear model to predict his wife’s age? Explain.
  + **y = 1.5740 + .9112\*85 = 79.026**
  + ~~Yes, because the fact that β1 is close to 1 indicates that the age diﬀerence between husbands + wives is consistent across ages~~
  + Not really, as this data was only collected on a sample of ages < 65 🡺 \*\*\***avoid extrapolating\*\*\***
* 31) Scatterplot summarizes husbands’ + wives’ heights in a random sample of 170 married couples in Britain, where both partners’ ages < 65 years. Summary output of the least squares ﬁt for predicting wife’s height from husband’s height is also provided in the table.
  + 
* (a) Is there strong evidence that taller men marry taller women? State the hypotheses and include any information used to conduct the test.
  + H0 = There is no significant association between husband height and wife height
    - true slope coeﬃcient of husband height is zero (β1 = 0).
  + H1 = There is a significant association between husband height and wife height
    - true slope coeﬃcient of husband height > zero (β1 > 0)
      * A 2-sided test would also be acceptable for this application (β1 != 0)
  + Our reported p-value = .0000, which is indeed < .05, so we state that we can reject H0 and our data present evidence of a significant positive association between husband height and wife height
* (b) Write the equation of the regression line for predicting wife’s height from husband’s height.
  + **y = 43.5755 + .2863\*husband\_height**
* (c) Interpret the slope and intercept in the context of the application.
  + This model assumes a wife height of 43.5755 in for a husband height of 0 in, on average, which is meaningless, + just serves to set the y-intercept
  + Then, the model assumes that for every 1 in in husband height, wives grow by .2863 in
* (d) Given that R2 = 0.09, what is the correlation of heights in this data set?
  + R = .3, which is quite weak
* (e) You meet a married man from Britain who is 5’9” (69 inches). What would you predict his wife’s height to be? How reliable is this prediction?
  + The model is quite confident if only looking at the p-value, but the R2 value is too low to be a reliable estimate
  + **y = 43.5755 + .2863\*69 = 63.3302**
* (f) You meet another married man from Britain who is 6’7” (79 inches). Would it be wise to use the same linear model to predict his wife’s height? Why or why not?
  + No, as this data was only collected on a sample of heights < 75 in 🡺 \*\*\***avoid extrapolating\*\*\***
* 32) See a scatterplot displaying relationship between % of families that own their home + % of the population living in urban areas, excluding DC + including Puerto Rico, for the US as well as the residuals plot for the 51 cases.

1.  

* (a) For these data, R2 = 0.28. What is the correlation? How can you tell if it is positive or negative?
  + R = -.52915, which is moderately negative (judging by the regression line in the scatterplot
* (b) Examine the residual plot. What do you observe? Is a simple least squares ﬁt appropriate for these data?
  + Appears to have a slight fan-shape = not good for linear models
* 33. Is **gestational age** (time between conception + birth) of a low birth-weight baby useful in predicting head circumference at birth? 25 low birth-weight babies were studied at a Harvard teaching hospital + investigators calculated the regression of head circumference (measured in cm) against gestational age (measured in weeks). The estimated regression line: **head circumference = 3.91 + 0.78×gestational age**
* (a) What is the predicted head circumference for a baby whose gestational age is 28 weeks?
  + **head circumference = 3.91 + 0.78×28 = 25.75 cm**
* (b) Std. error for the coeﬃcient of gestational age = 0.35, which is associated w/ dF = 23. Does the model provide strong evidence gestational age = signiﬁcantly associated w/ head circumference?
  + t = **2\*(estimate / std. error) = 2\*(3.91/.35) = 22.3428**
  + p-value = Looking @ t-table for dF = 23, we see t = **22.3428** givesa p-value between .05 and .025, so we can say the model provides strong evidence that gestational age is signiﬁcantly positively associated w/ head circumference
* See the relationship between teaching evaluation score (higher score = better) + standardized beauty score (0 = average, negative = below average, positive = above average) for a sample of 463 professors on RateMyProfessors.com. Given below are associated diagnostic plots. Also given is a regression output for predicting teaching evaluation score from beauty score.

1.    

* (a) Given the average standardized beauty score = -0.0883 + average teaching evaluation score = 3.9983, calculate the slope (or using just the info provided in the model summary table)
  + Calculate slope 🡺 **t\*std.error = .132986**
* (b) Do these data provide convincing evidence that the slope of the relationship between teaching evaluation and beauty is positive? Explain your reasoning.
  + Yes, p < .05
* (c) List the conditions required for linear regression and check if each one is satisﬁed for this model.
  + Residuals are normally distributed = almost, from histogram (left skew) + QQ-plot, but slightly high fitted quantiles for higher theoretical quantiles + slightly low fitted quantiles for lower theoretical quantiles
  + Residuals = IID 🡺 yes, form the residual plots showing no patterns