Business Analytics

Exploratory Data Analysis (EDA) & Linear Model: Simple Linear Regression

Dr. Jiun-Yu Yu
BA, NTU
18 Sep 2018

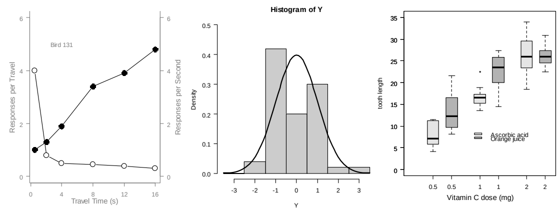
Outlines

- Exploratory Data Analysis (EDA)
  - Summarizing data
  - Representing data
  - Examples with R
- Simple Linear Regression (SLR) Model
  - Least squares estimators
  - Goodness of fit – coefficient of determination
  - Assumptions about SLR model
  - Statistical inference of least squares estimators
  - Model diagnosis – checking model assumptions
  - Prediction

Exploratory Data Analysis (EDA)

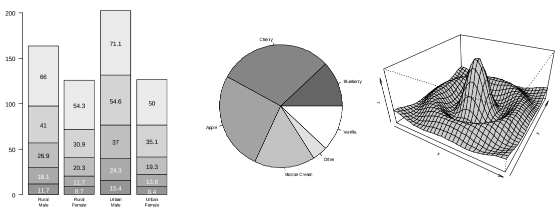
R Graphics

- R provides the usual range of standard statistical plots, including scatterplots, histograms, boxplots, barplots, piecharts, and basic 3D plots.



R Graphics

- In the first four cases, the basic plot type has been augmented by adding additional labels, lines, and axes.



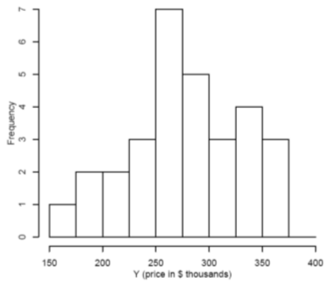
- Murrell, P. (2006), *R Graphics*, Chapman & Hall / CRC.

EDA: Summarizing Data

- Overall task: analyze data to inform a (business) decision.
- Assume data relevant to the problem has collected.
- Intermediate task: identify and summarize the data.
- Example:
  we've moved to a new city and wish to buy a home.
- Data:
  Y = selling price (in $ thousands) for n = 30 randomly sampled single-family homes
  (HOMES1):

155.5	195.0	197.0	207.0	214.9	230.0
239.5	242.0	252.5	255.0	259.9	259.9
269.9	270.0	274.9	283.0	285.0	285.0
299.0	299.9	319.0	319.9	324.5	330.0
336.0	339.0	340.0	355.0	359.9	359.9

Histogram



R on Examples

- Data input
  - `HOMES1 <- read.table("d1802_HOMES1.txt",header=TRUE)`
  - `attach(HOMES1)`
  - `Y`
- Histogram
  - `histY <- hist(Y, freq = FALSE, breaks = c(150,175,200,225,250,275,300,325,350,375,400), ylab = "Frequency", xlab = "Y (price in $ thousands)")`
- Box plot
  - `boxplot(Y)`

Summarizing the Data

- Measures of Location
  - `mean(Y);  median(Y)`
- Measures of Spread / Dispersion
  - `sd(Y);  var(Y)`
  - `min(Y);  max(Y);  range(Y)`
  - `quantile(Y, c(0.25,0.5,0.75));  summary(Y)`
  - `length(Y)`
- Measures of Shape
  - `skewness(Y)  # works with package "fBasics"`
  - `kurtosis(Y)  # works with package "fBasics"`
- End of Analysis
  - `detach`

EDA: Representing Multivariate Data

Example – Air Pollution (*Everitt, Chap 2*)

- 60 regions in the United States, variables include
  - *Rainfall*: mean annual precipitation in inches
  - *Education*: median school years completed for those over 25 in 1960
  - *Popden*; population/mile2 in urbanized area in 1960
  - *Nonwhite*: percentage of urban area population that is non-white
  - *NOX*: relative pollution potential of oxides of nitrogen
  - *SO2*: relative pollution potential of sulfur dioxide
  - *Mortality*: total age-adjusted mortality rate, deaths per 100,000
- R
  - `airpoll <- source("d1802_airpoll.dat")$value`
  - `attach(airpoll)`

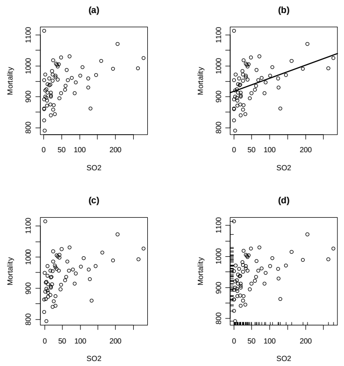
Scatter Plot – Air Pollution

- Setting plot area
  - `par(mfrow=c(2,2))`
  - `par(pty="s")`
    - "s" means square plot
- First scatter plot
  - `plot(SO2, Mortality, pch=1, lwd=1)`
    - pch means point type, lwd means line width
  - `title("(a)", lwd=2)`
- Second scatter plot with regression line
  - `plot(SO2, Mortality, pch=1, lwd=1)`
  - `abline(lm(Mortality~SO2), lwd=2)`
  - `title("(b)", lwd=2)`

Scatter Plot – Air Pollution

- Jittered plot
  - Sometimes some points are overplotting because their values are too close, so some small random amounts are added
  - `table(SO2)`
  - `subset(airpoll, SO2==1)$Mortality`
  - `Airpoll1 <- jitter(cbind(SO2,Mortality), amount=3)`
  - `plot(airpoll1[,1], airpoll1[,2], xlab="SO2", ylab="Mortality", pch=1, lwd=1)`
  - `title("(c)", lwd=2)`
- Rugged plot
  - Display marginal distributions of the two variables
  - `plot(SO2, Mortality, pch=1, lwd=1)`
  - `rug(jitter(SO2), side=1)`
  - `rug(jitter(Mortality), side=2)`
  - `title("(d)", lwd=2)`

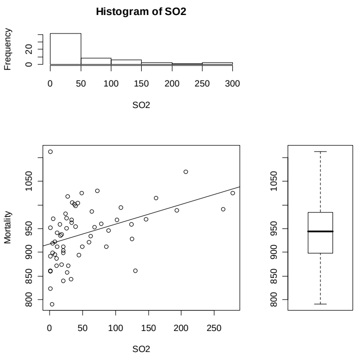
Scatter Plot – Air Pollution



Integrated Plots – Air Pollution

- Scatter plot on the left-bottom corner
  - `par(fig=c(0,0.7,0,0.7))`
    - fig = c(x1, x2, y1, y2), the partial area for plot
  - `plot(SO2, Mortality, lwd=1)`
  - `abline(lm(Mortality~SO2), lwd=1)`
- Add SO2 histogram on the top
  - `par(fig=c(0,0.7,0.65,1), new=TRUE)`
  - `hist(SO2, lwd=1)`
- Add Mortality boxplot on the right
  - `par(fig=c(0.65,1,0,0.7), new=TRUE)`
  - `boxplot(Mortality, lwd=1)`

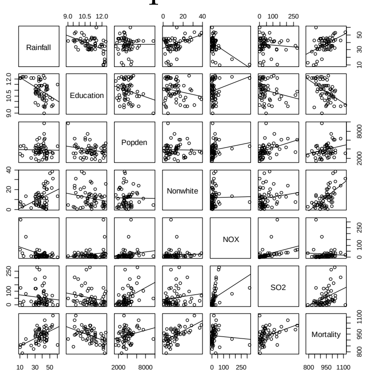
Integrated Plots – Air Pollution



Scatterplot Matrix

- Pair-wise comparisons
  - `pairs(airpoll)`
- Add regression lines in each plot
  - `pairs(airpoll, panel=function(x,y)`
    `{abline(lsfit(x,y)$coef, lwd=1); points(x,y)})`

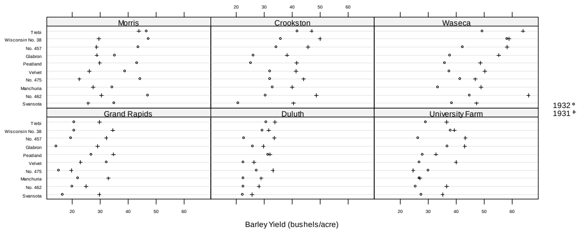
Scatterplot Matrix



R Graphics: Trellis Plots

- In addition to the traditional statistical plots, R provides an implementation of Trellis plots via the package `lattice`.
- Trellis plots provide a feature known as "multi-panel conditioning," which creates multiple plots by splitting the data being plotted according to the levels of other variables.
- Figure below shows an example of a Trellis plot.
  - The data are yields of several different varieties of barley at six sites, over two years.
  - The plot consists of six "panels," one for each site. Each panel consists of a dotplot showing yield for each site with different symbols used to distinguish different years, and a "strip" showing the name of the site.

R Graphics: Trellis Plots



Simple Linear Regression (SLR)

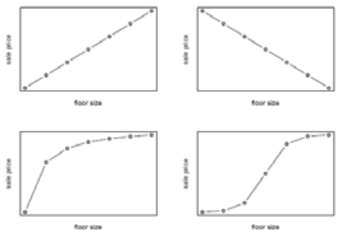
Simple Linear Regression Model: X & Y

- Y is a quantitative *response* variable
  (a.k.a. dependent, outcome, or output variable).
- X is a quantitative *explanatory* variable
  (a.k.a. predictor, independent or input variable, or covariate).
- Two variables play different roles, so important to identify which is which and define carefully, e.g.:
  - Y is sale price, in $ thousands;
  - X is floor size, in thousands of square feet.
- How much do we expect Y to change by when we change the value of X?
- What do we expect the value of Y to be when we set the value of X at 2?
- Note: association (observational data) not causation (experimental data).

Possible Relationships Between X and Y

Which factors might lead to the different relationships?

→ To conceptualize possible relationships in a *scatterplot* with Y plotted on the vertical axis and X plotted on horizontal axis.

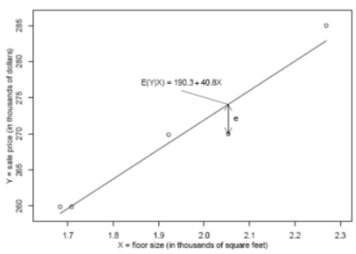


Straight-Line Model

- Simple linear regression models straight-line relationships (like upper two plots on last slide).
- Suppose sale price is (on average) $190,300 plus 40.8 times floor size:
  - $E(Y|X) = \mu(Y|X) = 190.3 + 40.8\,X$.
- Individual sale prices can deviate from this expected value by an amount ε_i (called a "random error").
  - $Y_i = 190.3 + 40.8\,X_i + \varepsilon_i \quad (i = 1, \ldots, n)$.
  - Y_i = deterministic part + random error.
- Error, ε_i, represents variation in Y due to factors other than X which we haven't measured,
  e.g., lot size, # beds/baths, age, garage, schools…

HOMES2 Data

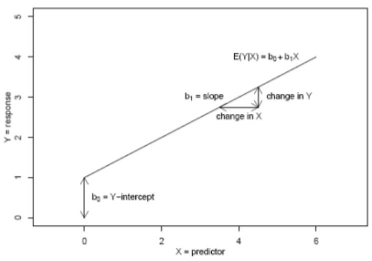
Y	259.9	259.9	269.9	270.0	285.0
X	1.683	1.708	1.922	2.053	2.269



Simple Linear Regression Model

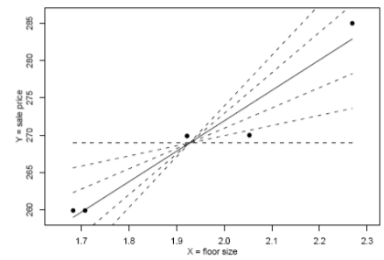
Population: $E(Y|X) = b_0 + b_1 X$

b_0 and b_1 are regression coefficients (*a.k.a.* regression parameters).



Least Squares

Which line fits the data best?



Least Squares Estimators

- Model: $Y_i = b_0 + b_1 X_i + \varepsilon_i$
- Sample: $\hat{Y} = \hat{\mu}(Y \mid X) = \hat{b}_0 + \hat{b}_1 X$ (estimated model).
- Obtain $\hat{b}_0$ and $\hat{b}_1$ by finding best fit line (least squares line).
- Mathematically, *minimize* the sum of squared errors (*SSE*):

$$SSE = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{b}_0 - \hat{b}_1 X_i)^2$$

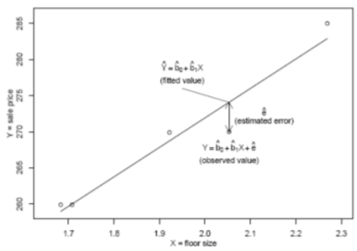
- Thus

$$\hat{b}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n} (X_i - \bar{X}) Y_i}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n} X_i (Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

$$\hat{b}_0 = \bar{Y} - \hat{b}_1 \bar{X}$$

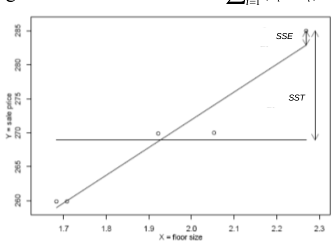
Least Squares Estimators

- Model: $Y_i = b_0 + b_1 X_i + \varepsilon_i$
  - $\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_i$: estimated expected value of response variable
    $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$: residual from least squares
  - Some properties:
    (1) $\sum_{i=1}^{n} \hat{\varepsilon}_i = 0$
    (2) $\sum_{i=1}^{n} X_i \hat{\varepsilon}_i = 0$
    (3) $\sum_{i=1}^{n} \hat{Y}_i \hat{\varepsilon}_i = 0$



Goodness of Fit

- Without model, estimate Y with sample mean $\bar{Y}$.
- With model, estimate Y using fitted $\hat{Y}$ - value .
- Total error without model: $SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$
- Remaining error with model: $SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i^{\,2}$



Coefficient of Determination R^2

- Proportional reduction in error: $R^2 = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$
- R^2 measures the proportion of variation in Y (about its mean) that can be explained by a straight-line relationship between Y and X. Thus, $0 \le R^2 \le 1$, and value closer to 1 refers to better fit.
- Using R^2 to compare the goodness of fits among different models, these models must have the same response variables.
- Correlation coefficient, r, measures the strength and direction of linear association between Y and X, and $r = \sqrt{R^2}$. (SLR only)

Residual Standard Error, $\hat{\sigma}$

- **Note: No assumption on ε_i is needed so far.**
- From now on, some assumptions about ε_i are made:

$$E(\varepsilon_i) = 0,\ \ Var(\varepsilon_i) = \sigma^2,\ \text{and}\ \ Cov(\varepsilon_i, \varepsilon_j) = 0\ \ \forall i \neq j.$$

- Thus, : $E(Y_i) = b_0 + b_1 X_i$, $Var(Y_i) = \sigma^2$.
- *Residual standard error*, $\hat{\sigma}$, estimates σ:

$$\hat{\sigma} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}{n-2}}$$

Least Squares Estimators for Regression Coefficients

$$E(\hat{b}_1) = b_1$$

$$Var(\hat{b}_1) = \sigma^2 \left(\frac{1}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right)$$

$$E(\hat{b}_0) = b_0$$

$$Var(\hat{b}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right)$$

$$Cov(\hat{b}_0, \hat{b}_1) = -\sigma^2 \left(\frac{\bar{X}}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right)$$

Further Assumptions on ε_i

- Now we further assume that ε_i are independent identical distributed (i.i.d.) normal random variable:

  $\varepsilon_i \sim N(0, \sigma^2)$
- Thus,

$$\hat{b}_1 \sim N\left(b_1,\ \ \sigma^2 \left(\frac{1}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right) \right)$$

$$\hat{b}_0 \sim N\left(b_0,\ \ \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right) \right)$$

$$\hat{\varepsilon}_i \sim N\left(0,\ \ Var(\hat{\varepsilon}_i)\right)$$

Something too complicated to consider in this course

Inference in Regression Analysis

- Recall from last week that σ is difficult to know so that it is usually replaced by its estimator $\hat{\sigma}$.
  By doing so the *t* distribution should be applied.
- Thus, let $SE_{\hat{b}_1} = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}} = \sqrt{\frac{SSE/(n-2)}{\sum_{i=1}^{n} (X_i - \bar{X})^2}}$, we have

$T_{\hat{b}_1} = \frac{\hat{b}_1 - b_1}{SE_{\hat{b}_1}} \sim t_{n-2}$. Similarly,

$T_{\hat{b}_0} = \frac{\hat{b}_0 - b_0}{SE_{\hat{b}_0}} \sim t_{n-2}$, where $SE_{\hat{b}_0} = \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right)}$.

Example: HOMES2

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  190.318     11.023  17.266 0.000423 ***
X             40.800      5.684   7.179 0.005569 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Estimated equation: $\hat{Y} = \hat{b}_0 + \hat{b}_1 X = 190.3 + 40.8X$.
- We expect *Y* to change by $\hat{b}_1$ when *X* increases by one unit, i.e., we expect sale price to increase by \$40,800 when floor size increases by 1000 sq. feet.
- For this example, more meaningful to say we expect sale price to increase by \$4,080 when floor size increases by 100 sq. feet.

Hypothesis Test for b_1

- We know *t*-ratio = $T_{\hat{b}_1} = \frac{\hat{b}_1 - b_1}{SE_{\hat{b}_1}} \sim t_{n-2}$
- H_0: $b_1 = 0$ vs. H_A: $b_1 \neq 0$
- *t*-ratio = $\frac{\hat{b}_1 - b_1}{SE_{\hat{b}_1}} = \frac{40.8 - 0}{5.684} = 7.18$
- Significance level = 5%, and *p*-value is 0.0056.
- Since *p*-value < significance level, reject H_0 in favor of H_A.
- In other words, the sample data favor a nonzero slope (at a significance level of 5%).

Slope Confidence Interval

- Calculate a 95% confidence interval for b_1 .
- 97.5^{th} percentile of t_3 is 3.182.
- $\hat{b}_1 \pm 97.5^{th}$ percentile $(SE_{\hat{b}_1})$
  $= 40.8 \pm 3.182 \times 5.684 = 40.8 \pm 18.1 = (22.7, 58.9)$.
- Loosely speaking: based on this dataset, we are 95% confident that the population slope, b_1 , is between 22.7 and 58.9.
- More precisely: if we were to take a large number of random samples of size 5 from our population of homes and calculate a 95% confidence interval for each, then 95% of those confidence intervals would contain the (unknown) population slope.

Interpreting R^2

```
Residual standard error: 2.786 on 3 degrees of freedom
Multiple R-squared: 0.945,      Adjusted R-squared: 0.9266
```

- Home prices example: $R^2 = \dfrac{423.4 - 23.3}{423.4} = 0.945$
- 94.5% of the variation in sale price (about its mean) can be explained by a straight-line relationship between sale price and floor size.

HOMES2 Example R Results

```
> fit <- lm(Y~X)
> summary(fit)

Call:
lm(formula = Y ~ X)

Residuals:
      1       2       3       4       5
 0.9152 -0.1048  1.1640 -4.0808  2.1064

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  190.318     11.023  17.266 0.000423 ***
X             40.800      5.684   7.179 0.005569 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.786 on 3 degrees of freedom
Multiple R-squared: 0.945,      Adjusted R-squared: 0.9266
F-statistic: 51.53 on 1 and 3 DF,  p-value: 0.00557

> confint(fit)
                2.5 %    97.5 %
(Intercept) 155.23828 225.39804
X            22.71243  58.88782
```


Regression Model Assumptions Revisited

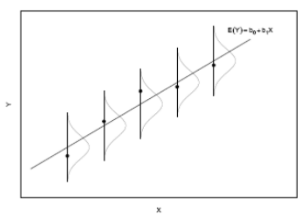
Four assumptions about random errors,

$$\varepsilon = Y - E(Y) = Y - b_0 - b_1 X$$

- Probability distribution of ε at each value of X has a **mean of zero**;
- Probability distribution of ε at each value of X has **constant variance**;
- Value of ε for one observation is **independent** of the value of ε for any other observation;
- Probability distribution of ε at each value of X is **normal**.

Viewing Assumptions on Scatterplot

Random error probability distributions.

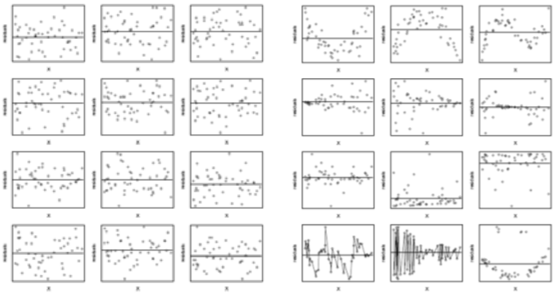


Checking Model Assumptions

- Calculate residuals, $\hat{\varepsilon} = Y - \hat{Y} = Y - \hat{b}_0 - \hat{b}_1 X$
- Draw a residual plot with $\hat{\varepsilon}$ along the vertical axis and X along the horizontal axis.
  - Assess **zero mean** assumption—do the residuals average out to zero as we move across the plot from left to right?
  - Assess **constant variance** assumption—is the (vertical) variation of the residuals similar as we move across the plot from left to right?
  - Assess **independence** assumption—do residuals look "random" with no systematic patterns?
- Draw a histogram and QQ-plot of the residuals.
  - Assess **normality** assumption—does histogram look approximately bell-shaped and symmetric and do QQ-plot points lie close to line?

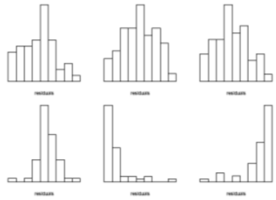
Residual Plots which Pass / Fail

- Left parts pass, right ones fail

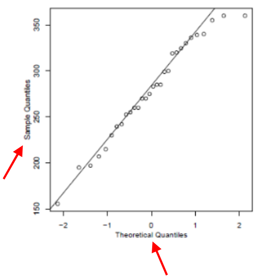


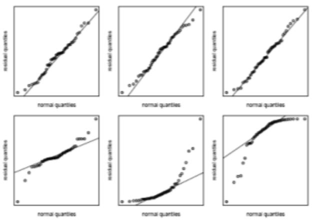
Histograms of Residuals

Upper three pass, lower three fail



QQ-plots of Residuals





Upper three pass, lower three fail

Assessing Assumptions in Practice

- Assessing assumptions in practice can be difficult and time-consuming.
- Taking the time to check the assumptions is worthwhile and can provide additional support for any modeling conclusions.
- Clear violation of one or more assumptions could mean results are questionable and should probably not be used (possible remedies to come in the following lectures).
- Regression results tend to be quite robust to mild violations of assumptions.
- Checking assumptions when n is very small (or very large) can be particularly challenging.

Estimated & Predicted Response

- Recall the confidence interval (CI) for a univariate population mean, μ : $\bar{Y} \pm t\text{-percentile}(SE_Y \sqrt{1/n})$
- Also, a prediction interval (PI) for an individual univariate Y-value: $\bar{Y} \pm t\text{-percentile}(SE_Y \sqrt{1+1/n})$
- Similar distinction between CI and PI for SLR.

- CI for $E(Y)$ at a particular X-value : $\hat{Y} \pm t\text{-percentile } (SE_{\hat{Y}})$
- PI for a Y-value given a new X-value is : $\hat{Y}_{new} \pm t\text{-percentile } (SE_{\hat{Y}_{new}})$

- Which should be wider? Is it harder to estimate a mean or predict an individual value?

Confidence Interval for $E(Y)$

- Formula: $\hat{Y} \pm t\text{-percentile } (SE_{\hat{Y}})$

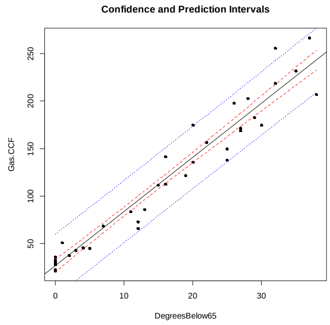
  where $\hat{Y} = b_0 + b_1 X_k, \quad SE_{\hat{Y}} = \hat{\sigma}\sqrt{\dfrac{1}{n} + \dfrac{(X_k - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$
- CI is narrower when:
  - n is large;
  - X_k is close to its sample mean, $\bar{X}$;
  - the residual standard error, $\hat{\sigma}$, is small;
  - the level of confidence is lower.
- Example: for home prices-floor size dataset, the 95% CI for $E(Y)$ when $X_k = 2$ is (267.7, 276.1).
- Interpretation: we're 95% confident that the average sale price for 2000 square-foot homes is between \$267,700 and \$276,100.

Prediction Interval for a Y-value

- Formula: $\hat{Y}_{new} \pm t$-percentile $(SE_{\hat{Y}_{new}})$
  where $\hat{Y}_{new} = b_0 + b_1 X_{new}$, $\quad SE_{\hat{Y}_{new}} = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(X_{new} - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$
- PI is narrower:
  - n is large;
  - X_{new} is close to its sample mean, $\bar{X}$;
  - the residual standard error, $\hat{\sigma}$, is small;
  - the level of confidence is lower.
- Since $SE_{\hat{Y}_{new}} > SE_{\hat{Y}}$, PI is wider than CI.
- Example: home prices-floor size dataset, the 95% PI for $\hat{Y}_{new}$ given $X_{new} = 2$ is (262.1, 281.7).
- Interpretation: we're 95% confident that the sale price for a 2000 square-foot home is between \$262,100 and \$281,700.

Confidence and Prediction Intervals

Compare widths of confidence and prediction intervals.

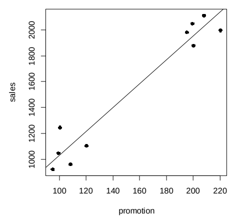


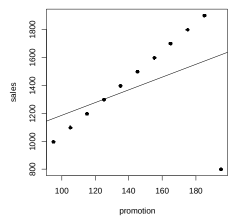
Simple Linear Regression Analysis

- Construct EDA on Y and X, individually and jointly.
  - Always look at the scatterplot.
- Formulate model.
  - Know the substantive context of the model.
- Estimate model parameters using least squares.
- Analyze model:
  - Residual standard error, $\hat{\sigma}$;
  - Coefficient of determination, R^2 ;
  - Coefficient for slope, b_1 .
- Diagnose model.
- Interpret model.
- Estimate $E(Y)$ and predict Y.
  - Limit predictions to the range of observed conditions.

Simple Linear Regression Pitfalls

- Do not assume that changing x causes changes in y.
  - Correlation is NOT causation.
- Do not forget lurking variables.
- Do not trust summaries like R^2 without looking at plots.





Summary

- EDA
- Simple linear regression (SLR) model:
  $Y_i = b_0 + b_1 X_i + \varepsilon_i$
- Least squares estimators
- Goodness of fit – coefficient of determination
- Assumptions about SLR model
- Statistical inference of least squares estimators
- Model diagnosis – checking model assumptions
- Prediction

Reading & Assignment

- Paradis (2005), *R for Beginners*
- Marques (2007), Chapter 2
  - Both available on the course website
- Ramsey & Schafer (2002):
  - Chapter 7
  - 8.1, 8.2, 8.7
- Assignment 0
  - Install R onto your personal computer and practice script '`s01_Intro.R`', downloadable from course website.