STAT 139: STATISTICAL SLEUTHING THROUGH LINEAR MODELS

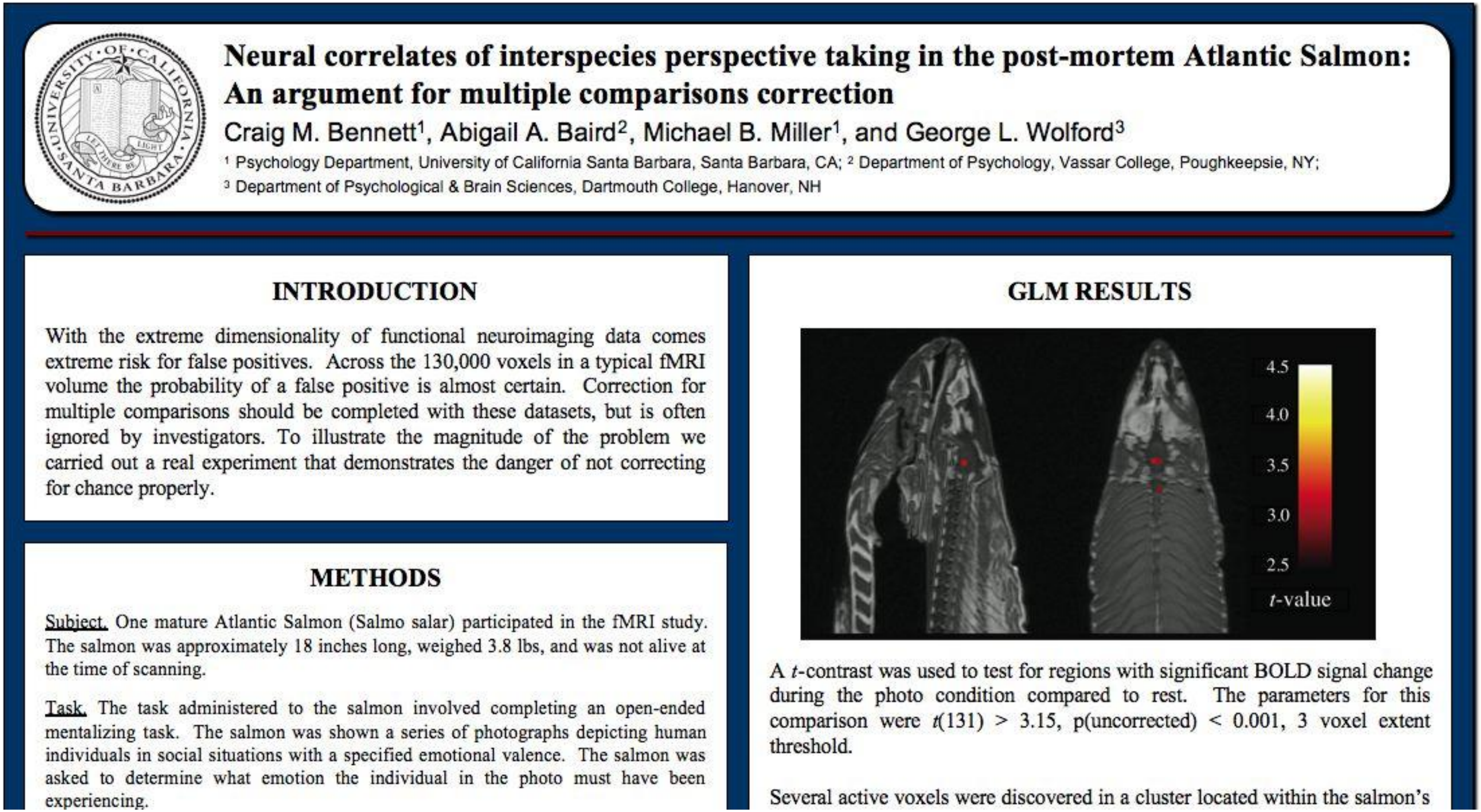
Lecture 14
Oct 21, 2014

Victoria Liublinska

Odds and Ends

- HW 6 Part I is posted and due on Fri, 10/24
  - Part 1(a): ~~reduced~~ model -> full model (thanks to Alister Bent)
  - For part 1(d): $n = n_1 + n_2 + \ldots + n_I$
  - Question 3:
    - Run `install.packages("tcltk")` before using `library(asbio)`;
    - Check and comment on assumptions;
    - Compare the unadjuasted 95% CIs to the modified ones.

- Pick up graded midterms at the end of today's lecture.
  - HW 6 part II is due before class next Tuesday ;
  - No hard copies – upload to the dropbox under "Exams".



Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction

Craig M. Bennett[1], Abigail A. Baird[2], Michael B. Miller[1], and George L. Wolford[3]

[1] Psychology Department, University of California Santa Barbara, Santa Barbara, CA; [2] Department of Psychology, Vassar College, Poughkeepsie, NY;
[3] Department of Psychological & Brain Sciences, Dartmouth College, Hanover, NH

INTRODUCTION

With the extreme dimensionality of functional neuroimaging data comes extreme risk for false positives. Across the 130,000 voxels in a typical fMRI volume the probability of a false positive is almost certain. Correction for multiple comparisons should be completed with these datasets, but is often ignored by investigators. To illustrate the magnitude of the problem we carried out a real experiment that demonstrates the danger of not correcting for chance properly.

METHODS

Subject. One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning.

Task. The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence. The salmon was asked to determine what emotion the individual in the photo must have been experiencing.

GLM RESULTS

A *t*-contrast was used to test for regions with significant BOLD signal change during the photo condition compared to rest. The parameters for this comparison were $t(131) > 3.15$, p(uncorrected) < 0.001, 3 voxel extent threshold.

Several active voxels were discovered in a cluster located within the salmon's

http://prefrontal.org/files/posters/Bennett-Salmon-2009.jpg

fMRI Gets Slap in the Face with a Dead Fish

- Poster presented by Bennett and colleagues at 2009 Human Brain Mapping conference.

- Objective: fMRI scanning of a salmon's brain.

- MRI scan divides the brain up into cubic units called voxels. There are over 40,000 in a typical scan. Most fMRI analysis treats every voxel independently…

- Bennett concludes: *The vast majority of fMRI studies should be utilizing multiple comparisons correction as standard practice.*

- HW6 article (Rothman, 1990, "No Adjustments Are Needed for Multiple Comparisons") makes an opposite argument.

Today's overview

- Course Project
- Midterm: common mistakes
- Simple Linear Regression
  - Motivation
  - Terminology
  - Model
  - Estimation

Reading:

- **Required:** Finish R&S Ch. 7, Ch. 7 R code

Course Project

- Groups of 3 or 4
- Types: Data project or Theory project

- Requirements and deadlines:
  - Group composition and project description (by Nov 3rd).
  - Mid-project update (up to 2 pages, by Nov 24th).
  - Poster (on Dec 9th):
    - Code (if applicable),
    - Reference list,
    - Evaluation of other posters.

- Poster presentation: Tuesday, December 9

Data Project

- Not a “data collection” project!
- Can be related to your outside research or other interests (however, can not be a course project fro another course), and the question must be *new*.
- Analyze data based on principles from this class:
  - Focus on assumptions, methods, validity of results.

Data Project Examples

- *Effect of hometown elevation on marathon performance.*
- *Eat your peas: Study of food waste and dining hall satisfaction at Harvard University.*
- *Empirical analysis of American Time Use Data: Analysis of leisure time and income.*
- *Interstate variation in rates of race-based hate crimes: Evidence for an association with racial diversity.*
- *What makes successful marriage? Factors that explain the duration of marriage.*
- *Explaining Regional Variability in Health Care Spendings.*

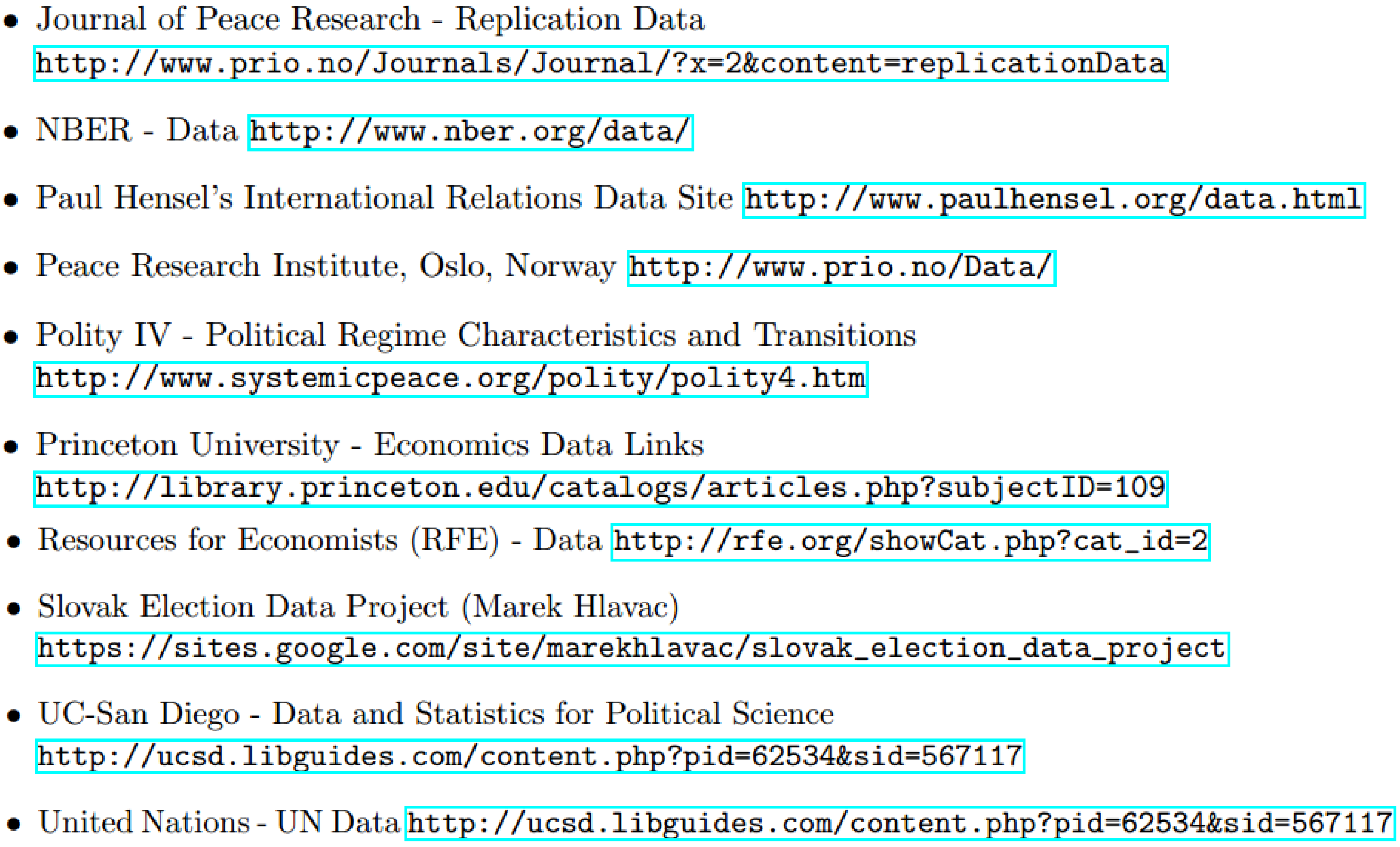
Examples of Data Sources

- A range of data resources for academic community `http://datalib.edina.ac.uk/catalogue/all`
- World Bank Open Data: free and open access to data about development in countries around the globe `http://data.worldbank.org/`.
- Sports Data: `http://it.stlawu.edu/~rlock/sports.html`
- Government data (from more than 70 agencies) `http://www.fedstats.gov/`
- 100+ Interesting Data Sets for Statistics `http://rs.io/100-interesting-data-sets-for-statistics/`
- Real estate sales data in the US (it is free, although registration is required to get a full access) `https://www.redfin.com/`
- Data available though Harvard Library (click "Data" on the left and explore!) `http://library.harvard.edu/`
- Aid Data - Open Data for International Development `http://www.aiddata.org/content/index`
- Center for Economic Policy Research - ceprDATA `http://ceprdata.org/`

Examples of Data Sources

- Correlates of War http://www.correlatesofwar.org/
- European Union - EUROSTAT http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home/
- FBI - Crime Statistics http://www.fbi.gov/stats-services/crimestats
- Federal Reserve Economic Data (FRED) http://research.stlouisfed.org/fred2/
- Gapminder http://www.gapminder.org/
- Harvard University Library: Getting Started in Economic Research http://guides.hcl.harvard.edu/content.php?pid=176637&sid=1487256 (Very comprehensive, many good links)
- IMF Data and Statistics http://www.imf.org/external/data.htm
- Interuniversity Consortium for Political and Social Research (ICPSR) at the University of Michigan http://www.icpsr.umich.edu/icpsrweb/ICPSR/index.jsp
- IQSS Dataverse Network http://dvn.iq.harvard.edu/dvn/
- Journal of Peace Research - Replication Data http://www.prio.no/Journals/Journal/?x=2&content=replicationData

Examples of Data Sources



- Journal of Peace Research - Replication Data
  http://www.prio.no/Journals/Journal/?x=2&content=replicationData
- NBER - Data http://www.nber.org/data/
- Paul Hensel's International Relations Data Site http://www.paulhensel.org/data.html
- Peace Research Institute, Oslo, Norway http://www.prio.no/Data/
- Polity IV - Political Regime Characteristics and Transitions
  http://www.systemicpeace.org/polity/polity4.htm
- Princeton University - Economics Data Links
  http://library.princeton.edu/catalogs/articles.php?subjectID=109
- Resources for Economists (RFE) - Data http://rfe.org/showCat.php?cat_id=2
- Slovak Election Data Project (Marek Hlavac)
  https://sites.google.com/site/marekhlavac/slovak_election_data_project
- UC-San Diego - Data and Statistics for Political Science
  http://ucsd.libguides.com/content.php?pid=62534&sid=567117
- United Nations - UN Data http://ucsd.libguides.com/content.php?pid=62534&sid=567117

Examples of Data Sources

- University of Rochester - Guide to Political Science Data `http://www.lib.rochester.edu/index.cfm?page=3869&this_pageID=53`
- World Bank - Barro-Lee Educational Attainment Data `http://go.worldbank.org/HK0H13Y5D0`
- World Bank - Database of Political Institutions `http://go.worldbank.org/2EAGGLRZ40`
- World Bank - World Development Indicators `http://data.worldbank.org/data-catalog/world-development-indicators`
- World Health Organization (WHO) - Data `http://data.worldbank.org/data-catalog/world-development-indicators`
- World Justice Project - Rule of Law Index `http://worldjusticeproject.org/rule-of-law-index`
- Resources on Duke University web-site `https://stat.duke.edu/resources/datasets`
- Data surfing on the WWW, from Robin Lock (http://it.stlawu.edu/~rlock/datasurf.html)
- *Data Analysis Using Regression and Multilevel/Hierarchical Models* http://stat.columbia.edu/~gelman/arm/
- StatLib at CMU http://lib.stat.cmu.edu, including the Data and Story Library

Theory Project

- Extended simulations to answer a research question you propose.

- Explanation (and application) of
  - Bayesian, non-parametric, or iteratively reweighted version of regression;
  - Partial least squares regression.
  - [Probabilistic] Principal Components Analysis (PCA);
  - Random-effects models (or multilevel models);
  - …

- Critical review of recent papers on a topic related to this class.

…

Theory Project Examples

- *Bonferroni vs. Scheffe: A practical guideline on the choice of correction method in multiple comparison testing.*

- *The Bayesian approach to linear regression.*

- *Identifying Predictors of Childhood Mortality in Sub-Saharan Africa using (p)PCA.*

- *Instrumental Variables: An Econometric Tool for Determining Causality with Non-Randomized Treatment*

Course Project

- More info under “Project”.

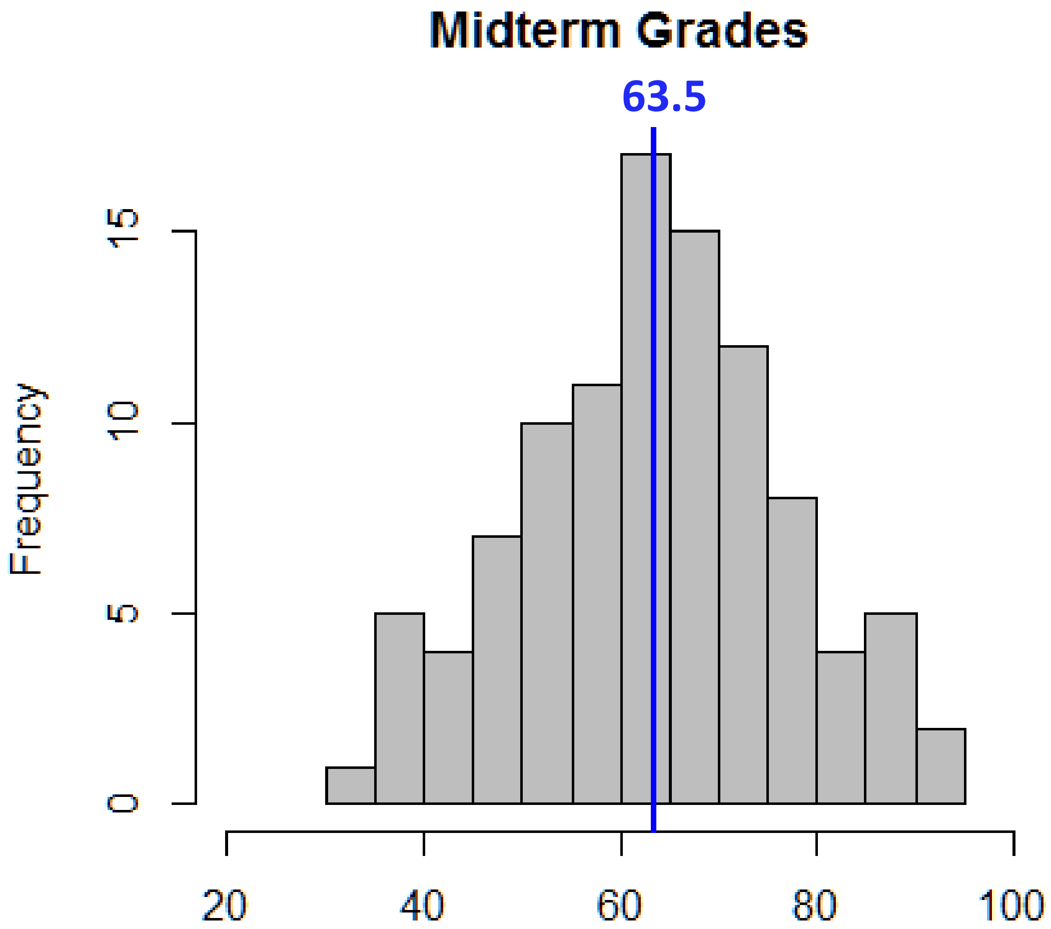
- Use Piazza to find project partners!

First deadline: Upload to the drop-box under “Project” by Monday, 5pm, November 3rd (2 weeks from today):

1. Names of all students in your group;
2. One paragraph describing project proposal;
3. Possible data source (if applicable).

Questions?

https://learningcatalytics.com/class_sessions


Midterm Grades
63.5
Frequency
0
5
10
15
20
40
60
80
100

Midterm Grades Summary

Min	10%	25%	50%	75%	90%	Max
33.5	45.5	53.5	63.5	72.5	80.5	94

Below 50 – alert!

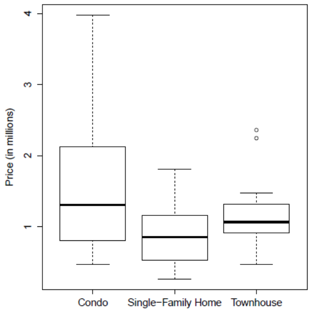
Year in School	Median Score
Sophomore	77
Junior	64
Senior	63
Graduate	63

Component	% of grade
Homework	30
Midterm	15
Final Project	15
Final Exam	30
LC Participation	10

Midterm: Common Mistakes

Problem #1 (four unrelated parts):

- 1a (multiple comparison): Why is it important to use multiple comparison adjustment? *see exercise 9 in Ch. 6
- 1b: Three permutation tests: all three tests were applied to a *randomized experiment*. What hypotheses are being tested?
- 1c: Two-sample *t*-test vs. permutation test vs. rank-sum test: Power!
- 1d: Two-sample *t*-test vs. paired *t*-test: Independent data mistakenly treated as dependent.



Midterm: Common Mistakes

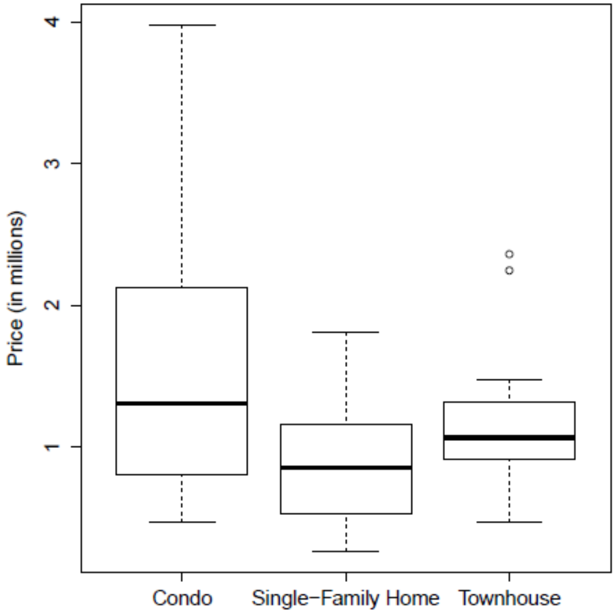
Problem #2 (real estate in greater Boston area)

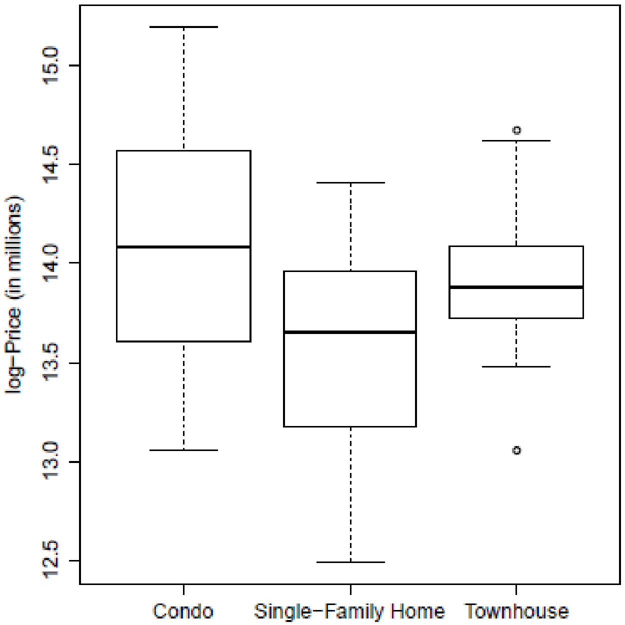
- 2a (study unit): Not just a “home”!
- 2b.i (two-sample *t*-test with three groups: condos, townhouses, single-family homes): can not pool all three groups for $s_{p.}$
- 2b.iii (list and comment on t-test assumptions): independence and homogeneity may also be an issue.
- 2b.iv (state your conclusion for the client): not enough to say “we reject the null hypothesis”.
- 2c.i (hypotheses for ANOVA): Log-transformed data, however, it is incorrect to state

  H0~~: ln(mu1) = ln(mu2) = ln(mu3)~~

Midterm: Common Mistakes

- 2c.iii (conclusion for ANOVA *F*-test *that your colleague can present to her client*): again, not enough to say that “we reject the null”!
- 2c.v (comment on ANOVA assumptions): Is the transformation justified?



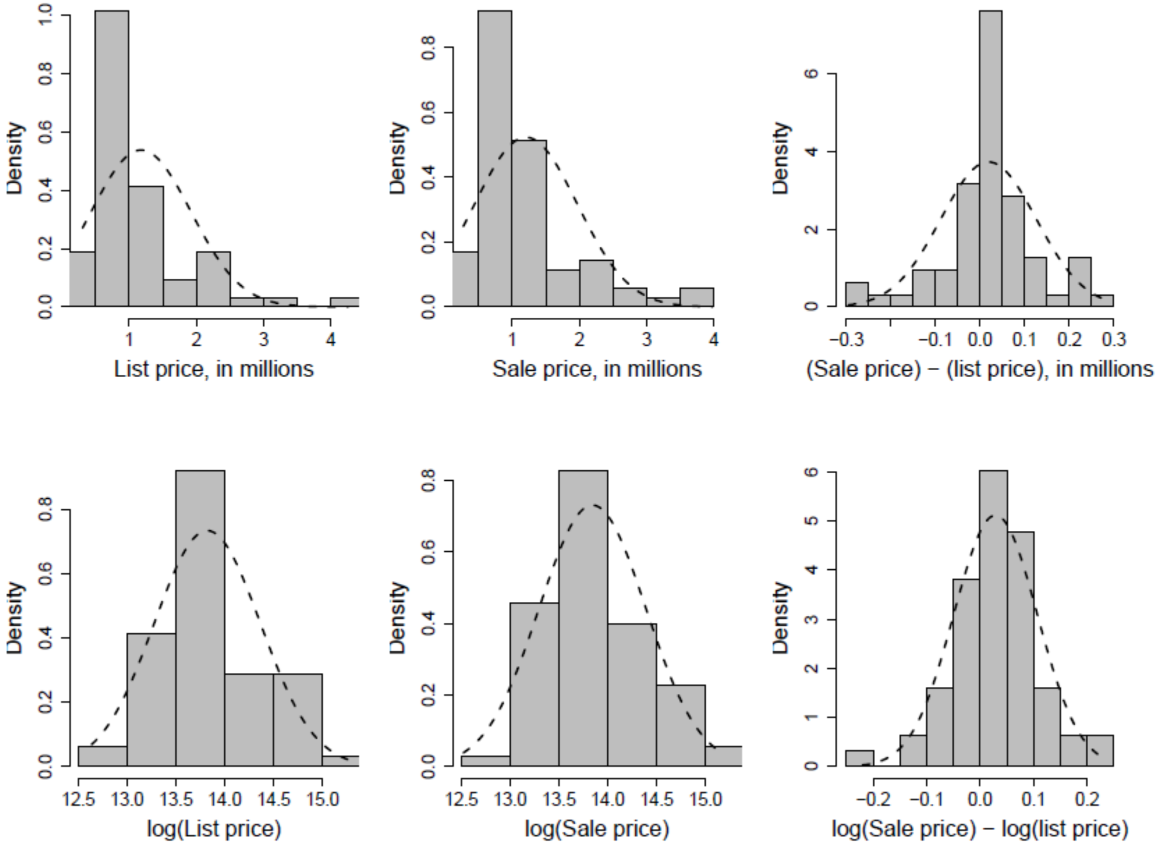


- 2c.vi (nonparametric alternative of ANOVA) - ?

Midterm: Common Mistakes

Part (d) had data on list and sale prices:

- 2d.i (choose an R output): Paired test! Transformation looks best.



Midterm: Common Mistakes

Part (d) had data on list and sale prices:

- 2d.ii (write down hypotheses): which parameter is being estimated?
- 2d.iii (state your conclusion based on the *original question*): is it a buyer's or a seller's market?
- 2d.iv (95% CI): always back-transform intervals
- 2d.v (assumptions): no need in "equal variances between populations" if the test is paired.

Question #3: Sampling distributions of sample variances and SSR.

- 3c: Need to *derive* the distribution.

Questions?

Simple Linear Regression

Chapters 7-10

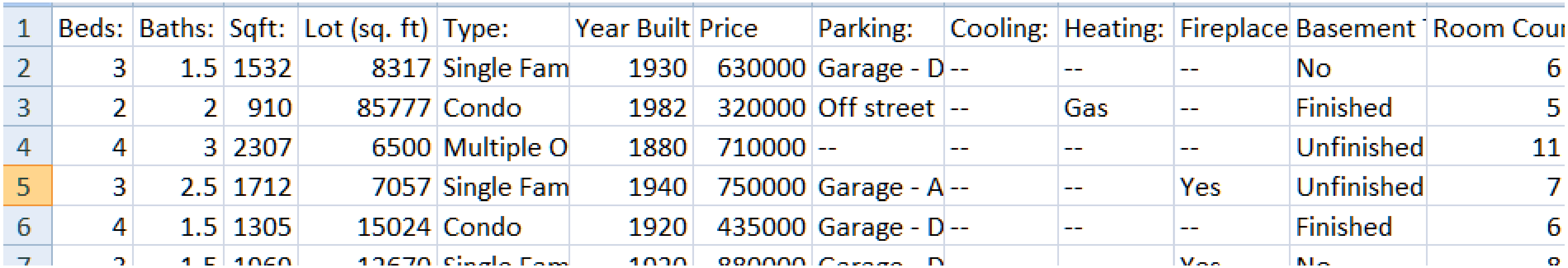
Chapter 7, 8: Simple Linear Regression

Chapter 9, 10: Multiple Regression

Real Estate Appraisal

- Data: 46 recent home sales in Newton, MA.

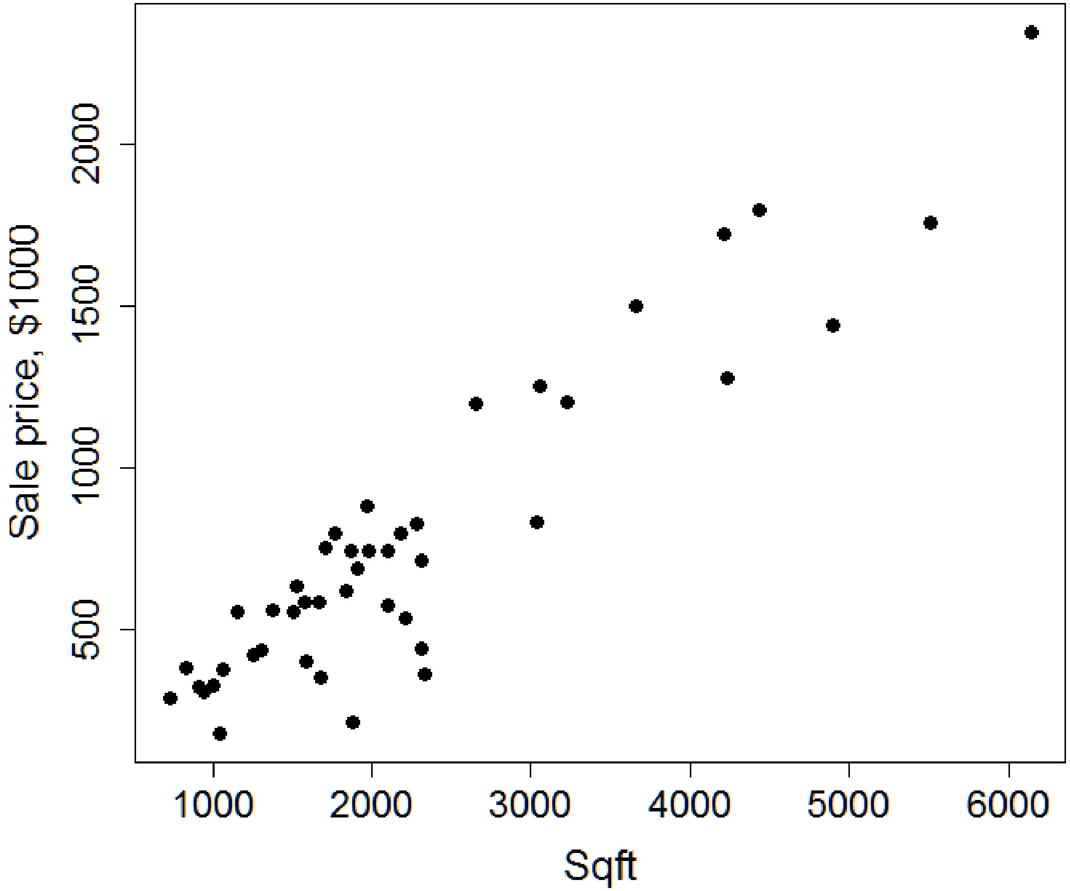
Source: zillow.com



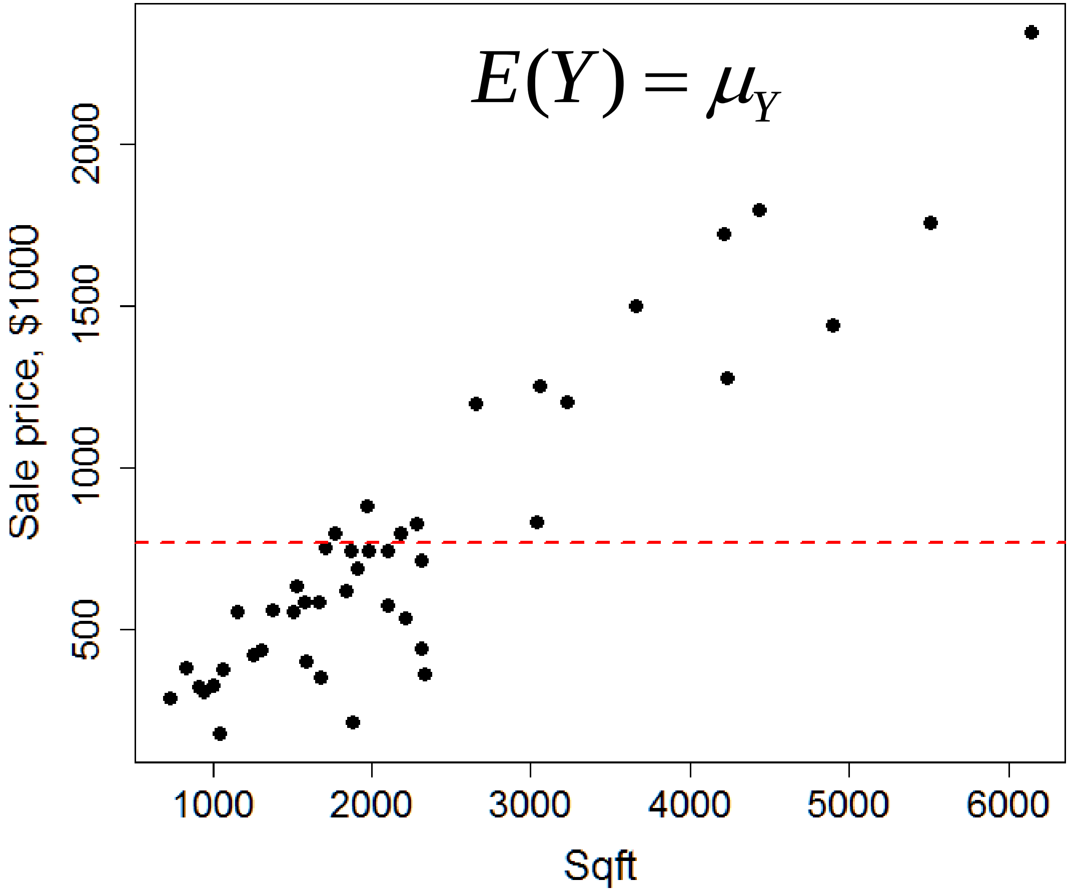
1	Beds:	Baths:	Sqft:	Lot (sq. ft)	Type:	Year Built	Price	Parking:	Cooling:	Heating:	Fireplace	Basement	Room Cou
2	3	1.5	1532	8317	Single Fam	1930	630000	Garage - D	--	--	--	No	6
3	2	2	910	85777	Condo	1982	320000	Off street	--	Gas	--	Finished	5
4	4	3	2307	6500	Multiple O	1880	710000	--	--	--	--	Unfinished	11
5	3	2.5	1712	7057	Single Fam	1940	750000	Garage - A	--	--	Yes	Unfinished	7
6	4	1.5	1305	15024	Condo	1920	435000	Garage - D	--	--	--	Finished	6
7	2	1.5	1060	12670	Single Fam	1920	880000	Garage - D			Yes	No	8

- Can we model and predict sale prices using just the living area size (Sqft)?

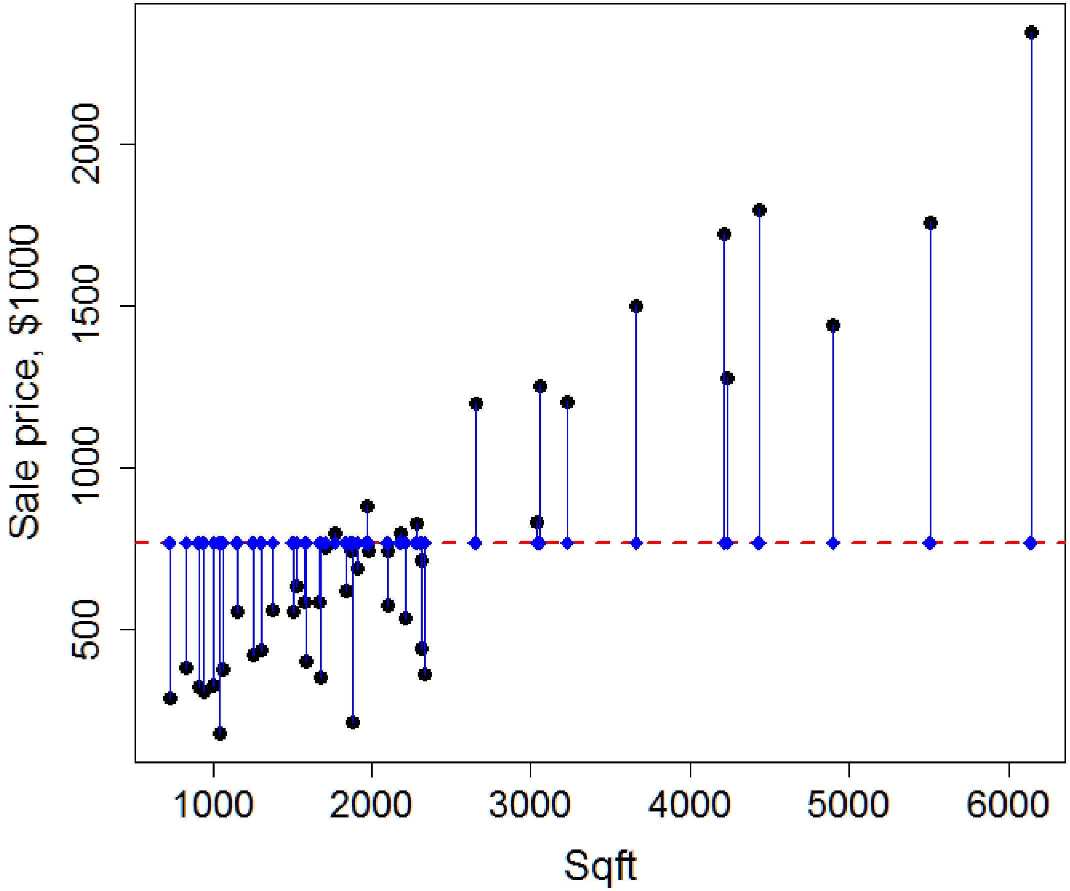
Home Sale Prices vs. Sqft



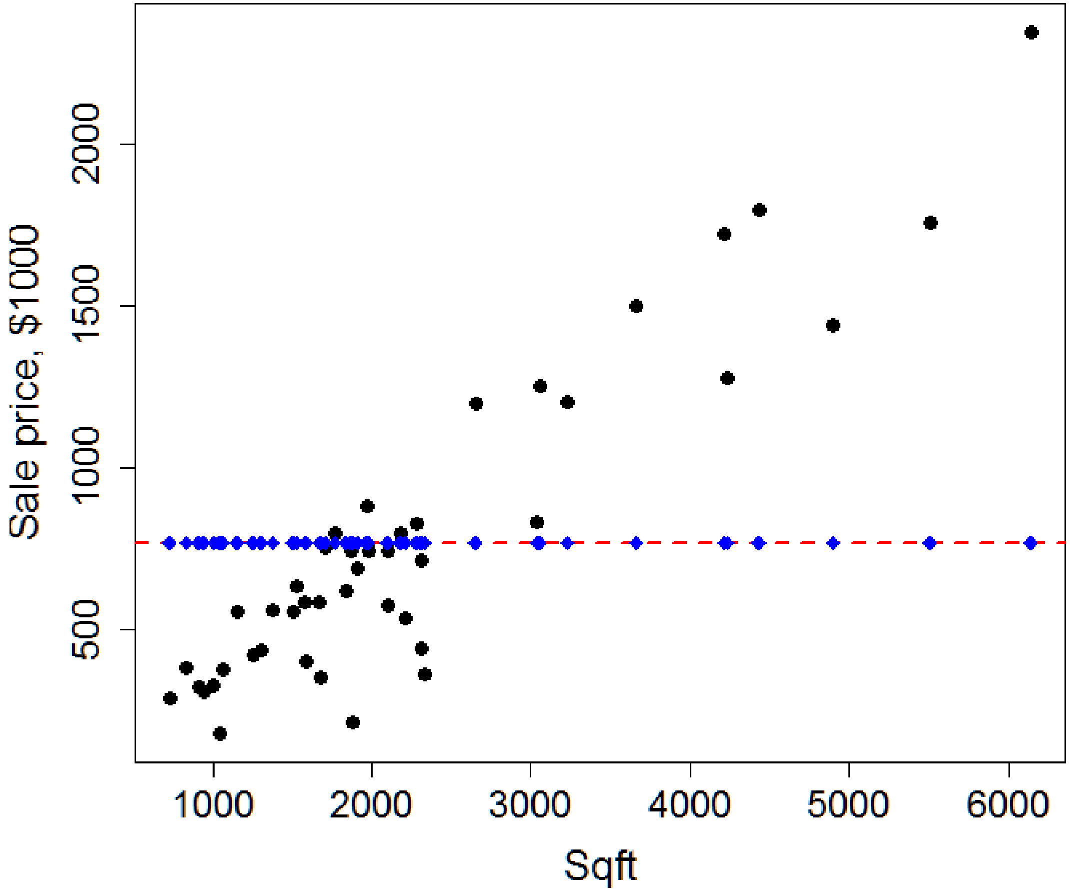
Equal-means model



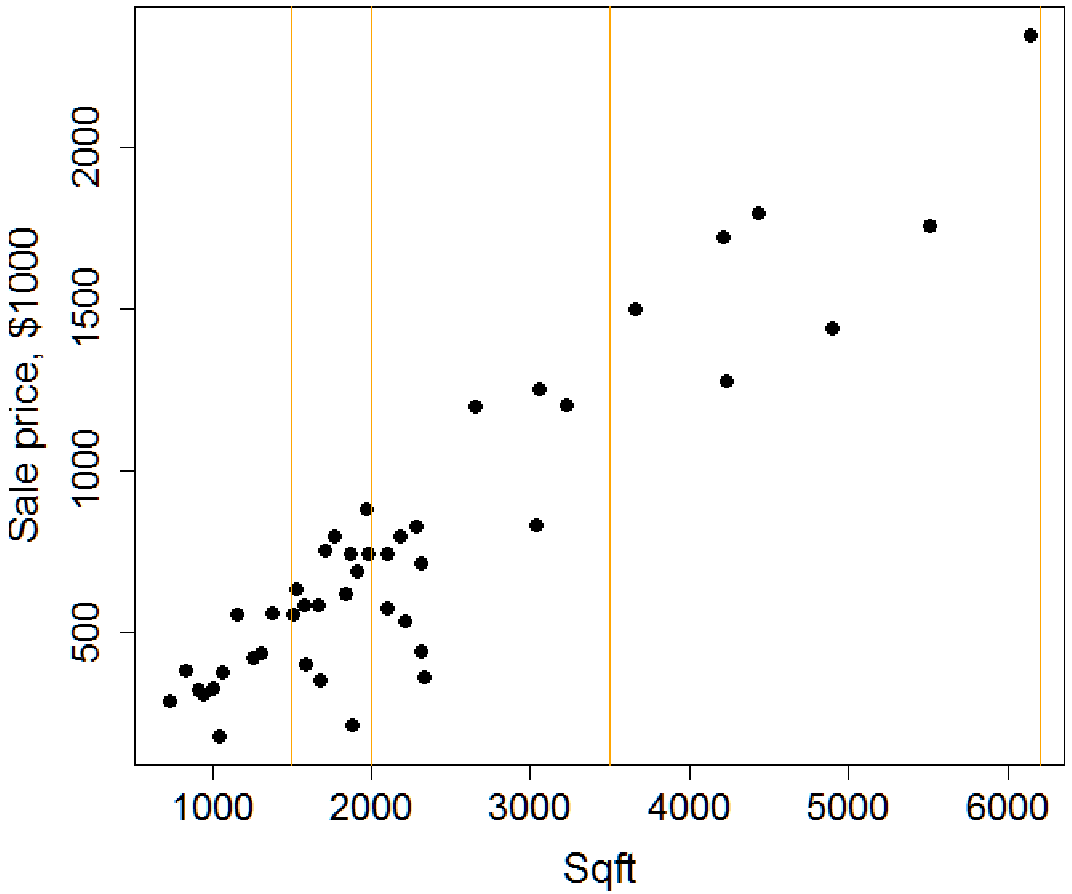
Equal-means model: predicted average prices



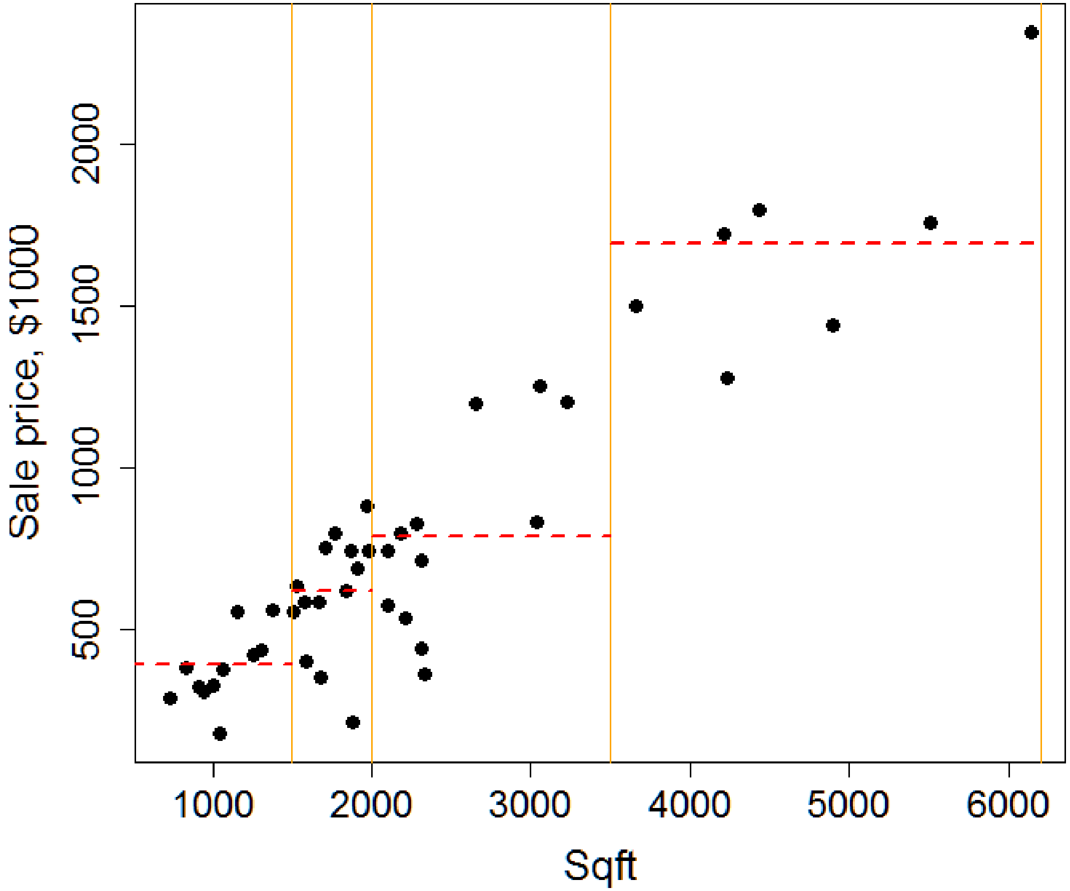
Equal-means model: predicted average prices



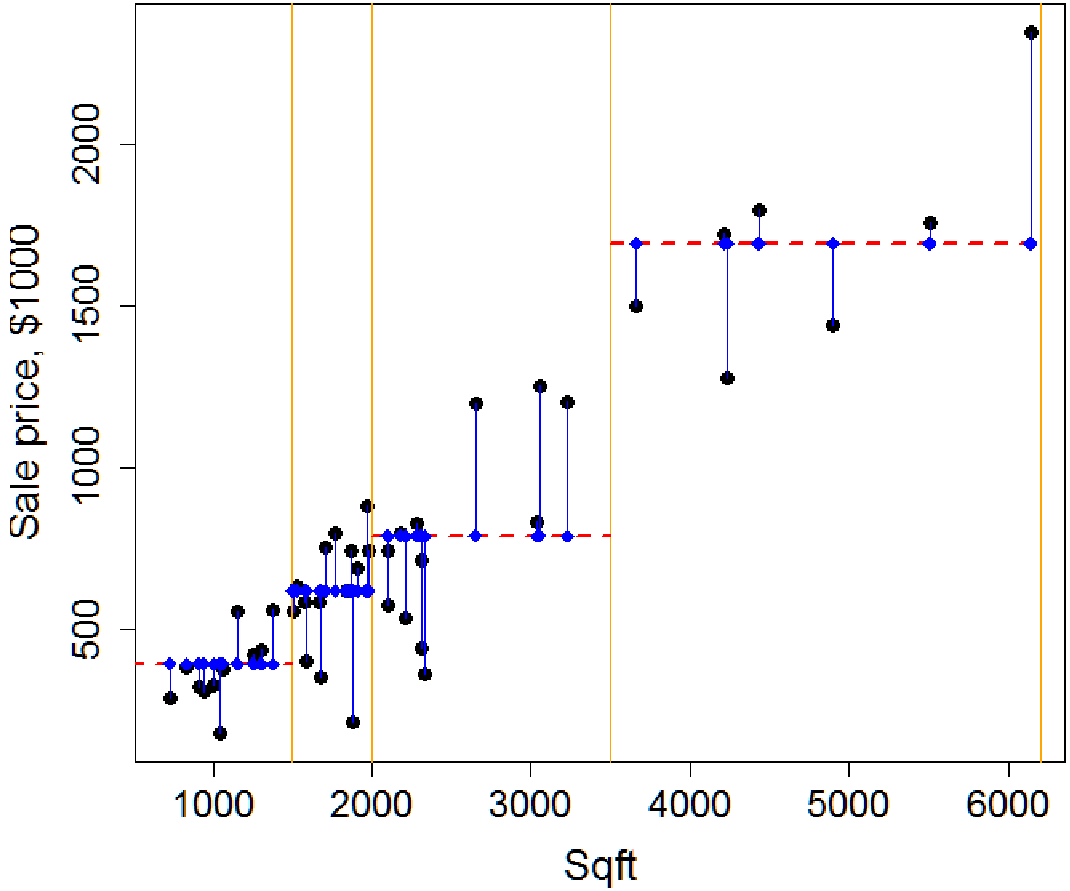
Separate-means model: 4 groups



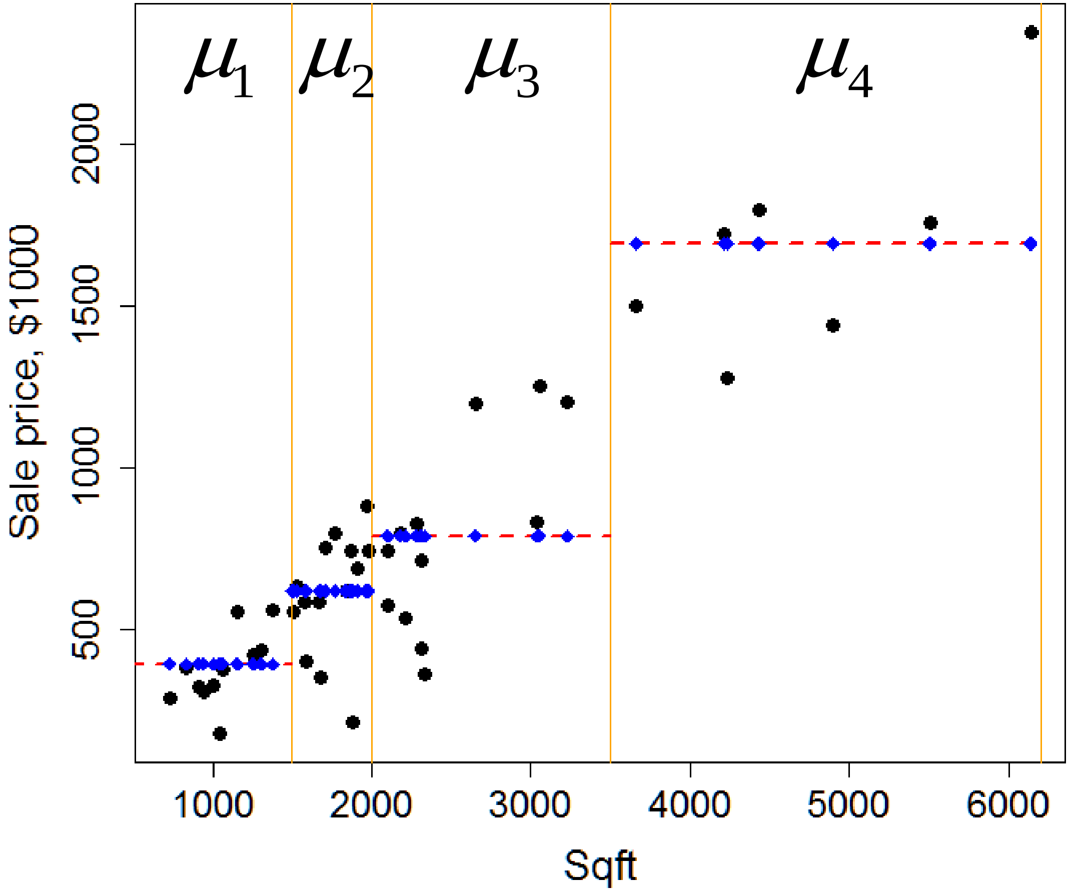
Separate-means model: 4 groups



Separate-means model: 4 groups

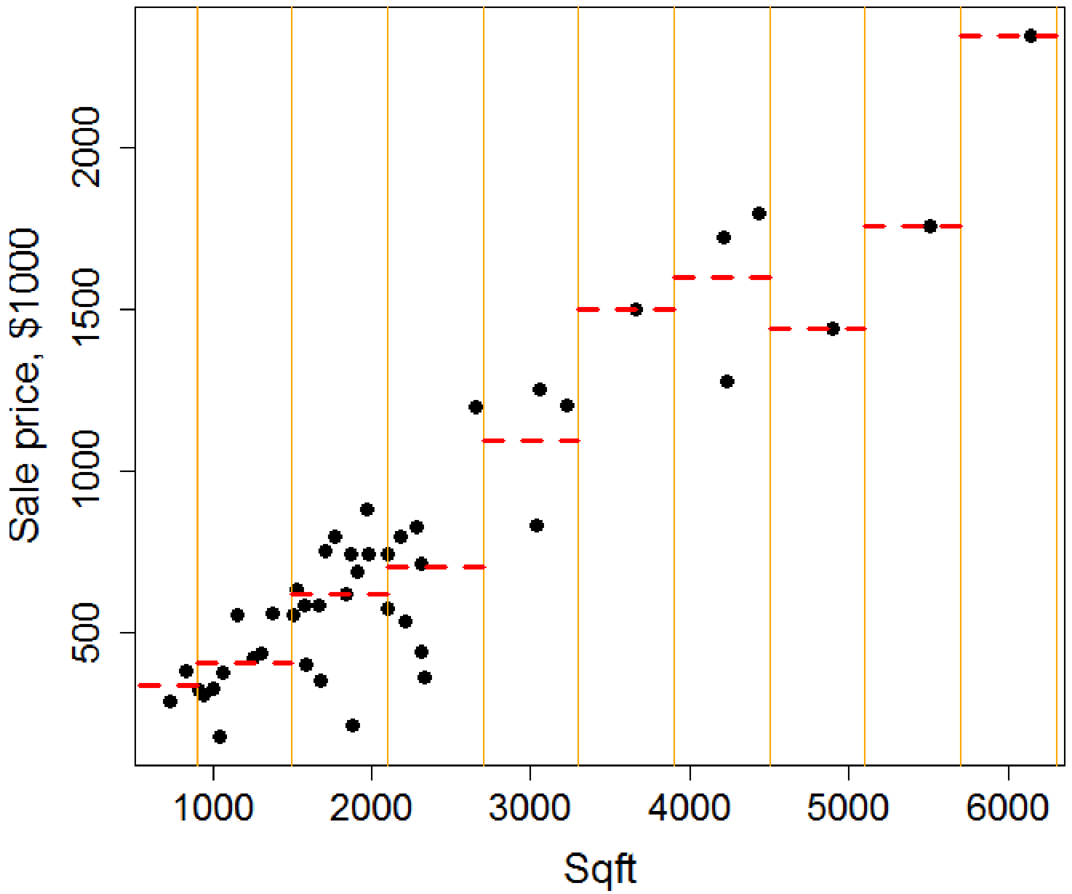


Separate-means model: 4 groups



How many parameters do we have to estimate?

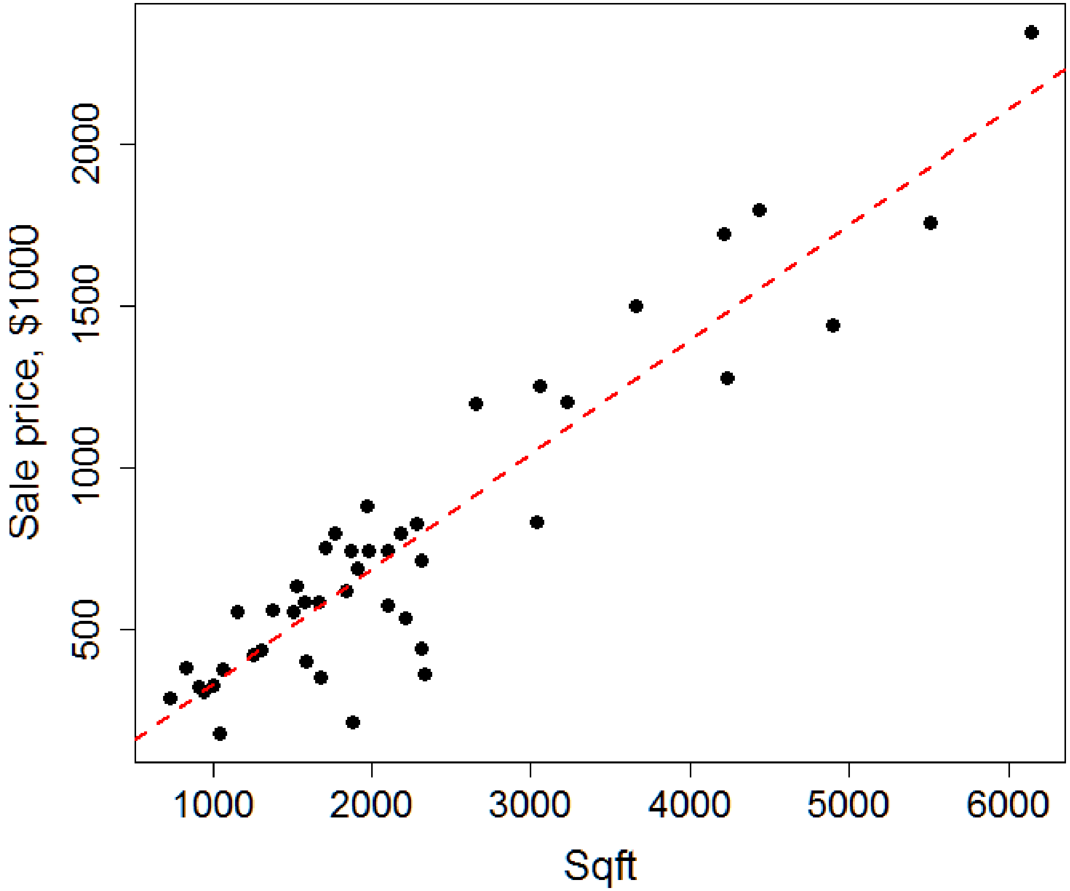
Separate-means model: 10 groups



How many parameters do we have to estimate?

Regression line

$$\mu\{\text{Price} \mid \text{Sqft}\} = \beta_0 + \beta_1 \text{Sqft}$$



- Three parameters: β_0, β_1, and σ^2

When is it appropriate to consider a regression line?

- Different groups correspond to different levels of a quantitative explanatory variable X (e.g., sq.ft.)

- Predicted means of Y (e.g., sale prices) in I consecutive groups fall along a line.

Regression of Response on Explanatory Variable

Data: $(X_i,\ Y_i)$ for $i = 1, 2, \ldots, n.$

Y_i: response or dependent or endogenous random variable.

X_i: explanatory or independent or exogenous variable (not random).

Regression Analysis: describes a probability distribution of Y_i as a function of X_i and a vector of parameters $\boldsymbol{\beta}$, $f(Y_i \mid X_i; \boldsymbol{\beta})$.

Simple Linear Regression

Data: (X_i, Y_i) for $i = 1, 2, \ldots, n$.

Y_i: response or dependent or endogenous random variable.

X_i: explanatory or independent or exogenous variable (not random).

Simple Linear Regression: describes the mean of Y_i as a linear function of X_i,

$$\mu\{Y_i \mid X_i\} = E(Y_i \mid X_i) = \beta_0 + \beta_1 X_i$$

Textbook notation: $\mu\{Y \mid X\} = \beta_0 + \beta_1 X$

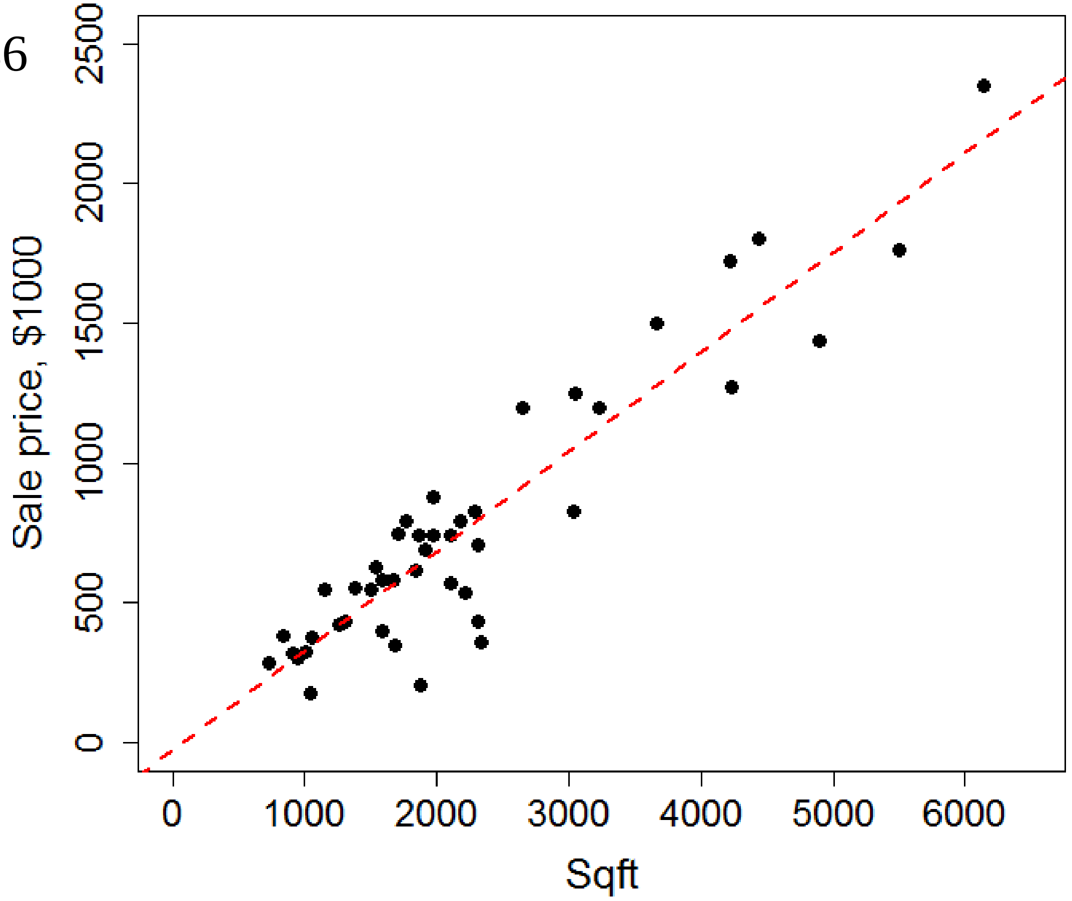
Regression line:
Example of Real Estate Appraisal

True parameters :

$\text{Intercept} : \beta_0 = -26$

$\text{Slope} : \beta_1 = 0.36$

$$\mu\{\text{Price} \mid \text{Sqft}\} = \boxed{-26} + \boxed{0.36} \cdot \text{Sqft}$$



Estimation Using Simple Linear Regression

$$\mu\{\text{Price}\,|\,\text{Sqft}\} = -26 + 0.36 \cdot \text{Sqft}$$

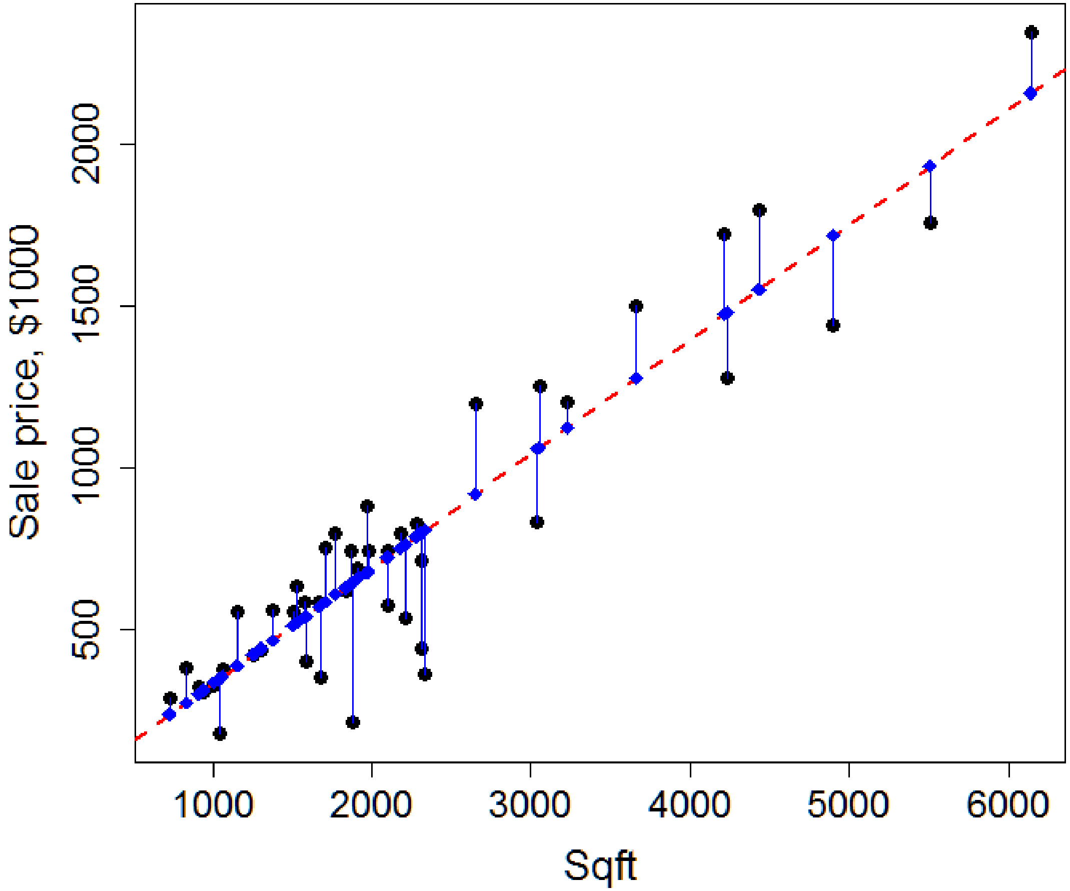
Interpolation: What is the average price of a house with 4000 sq.ft. in Newton, MA?

$$\mu\{\text{Price}\,|\,\text{Sqft} = 4000\} = ?$$

Extrapolation: What is the average price of a "shed" with 72.2 sq.ft.?

$$\mu\{\text{Price}\,|\,\text{Sqft} = 72.2\} = ?$$

Regression line: Estimated average Y_i for each X_i



What about the variability of Y_i?

Simple Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \text{ where } \varepsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$$

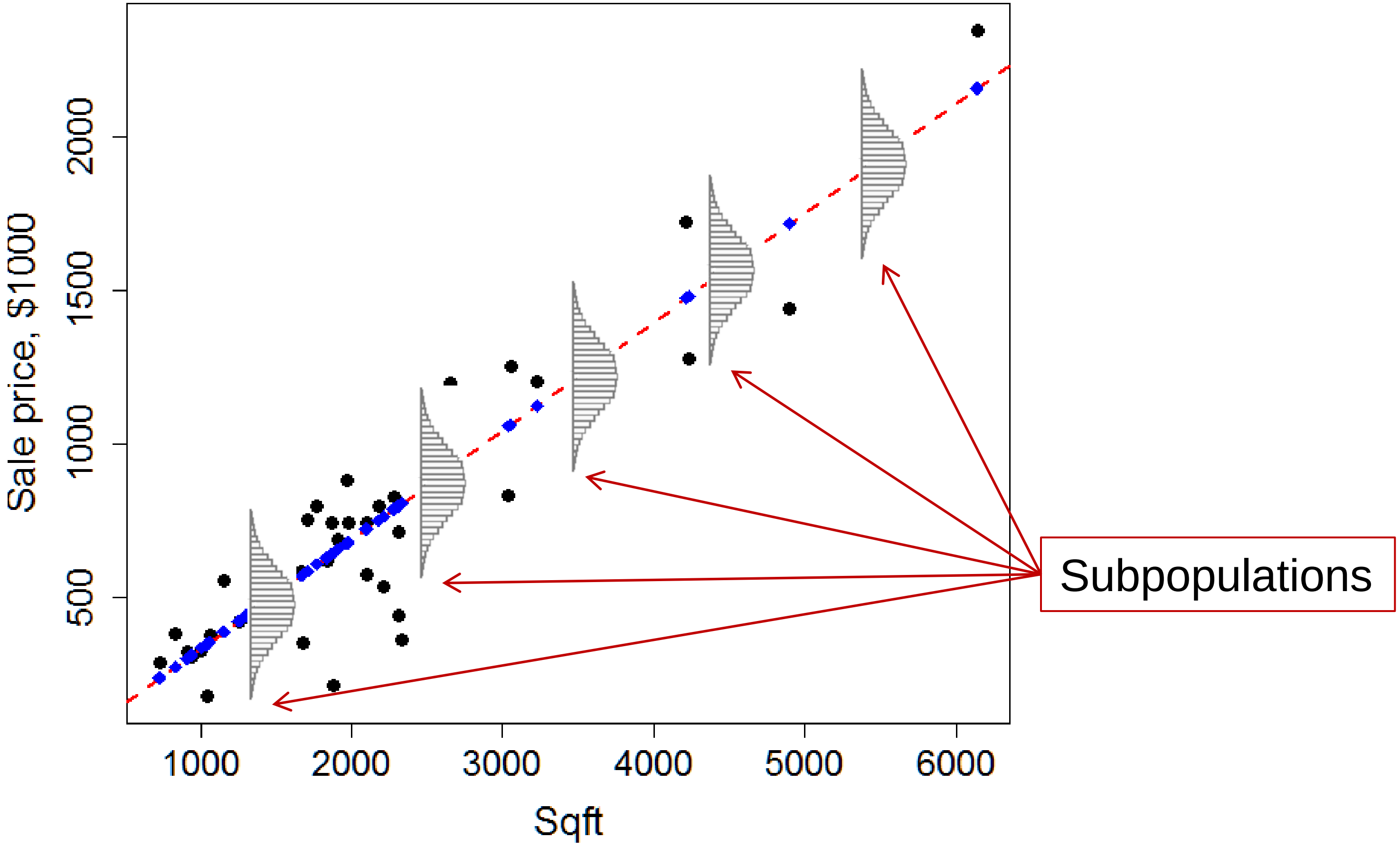
- $E(Y_i \mid X) = \beta_0 + \beta_1 X_i$
- $Var(Y_i \mid X) = \sigma^2 = Var(\varepsilon_i)$

The model assumes that there exists a large (or infinite) real (or hypothetical) population, from which we randomly sample n units and collect a pair of observations $(X_i,\ Y_i)$, $i = 1, 2, \ldots, n$.

Regression line: Assumptions

$$\text{Price}_{\text{i}} = -26 + 0.36 \cdot \text{Sqft}_{\text{i}} + \varepsilon_{\text{i}}, \ \text{ where } \varepsilon_{\text{i}} \overset{i.i.d}{\sim} N(0{,}181^2)$$

- Linearity
- Equal Spread
- Normality
- Independence



Simple Linear Regression: Assumptions

- Linearity: Means of sub-populations of responses for *each value of the explanatory variable* fall along a line.
- Equal Spread (Constant Variance): Variances of sub-populations are all equal.
- Normality: sub-population of responses for each value of explanatory variable are distributed normally around the estimated mean.
- Independence: units' responses are independent.