STAT 139: STATISTICAL SLEUTHING THROUGH LINEAR MODELS

Lecture 15
Oct 23, 2014

Victoria Liublinska

Odds and Ends

- If you have issues installing `asbio` package (usually, in MACs) and using `pairw.anova()`
  - Install XQuartz from http://xquartz.macosforge.org/landing/
  - Restart

Previous lecture: Review

- Simple Linear Regression:

X_i are fixed for all i.

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

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

$$Var(Y_i \mid X_i) = \sigma^2 = Var(\varepsilon_i)$$

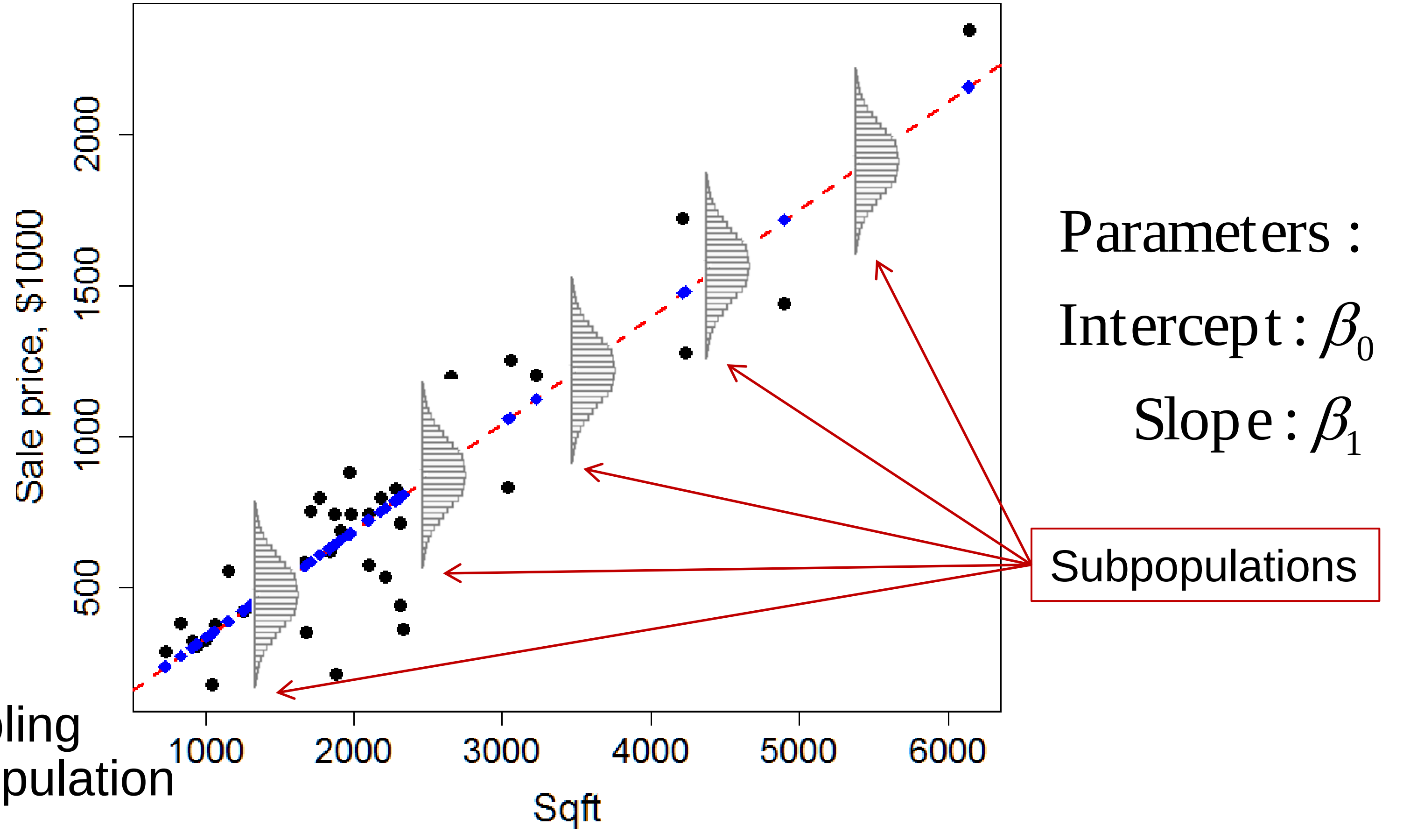
- Estimation: interpolation / extrapolation;
- Data: 46 recent home sales in Newton, MA.

Regression line: Model and Assumptions

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

Assumptions:

- Linearity
- Equal Spread
- Normality
- Independence
- Random sampling from a large population



Parameters:

Intercept: β_0

Slope: β_1

Today's overview

- Simple Linear Regression, cont.
  - Motivation
  - Model
  - Terminology
  - Estimation & testing
  - Computational Tricks
  - Prediction

Reading:

- **Required:** Finish R&S Ch. 7, Ch. 7 R code
- **Supplementary Theory**: A. Sen and M. Srivastava. "Regression Analysis: Theory, Methods, and Applications", Chapter 1: **Introduction** (you may skip Sec. 1.7 for now).

Simple Linear Regression: Estimation

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

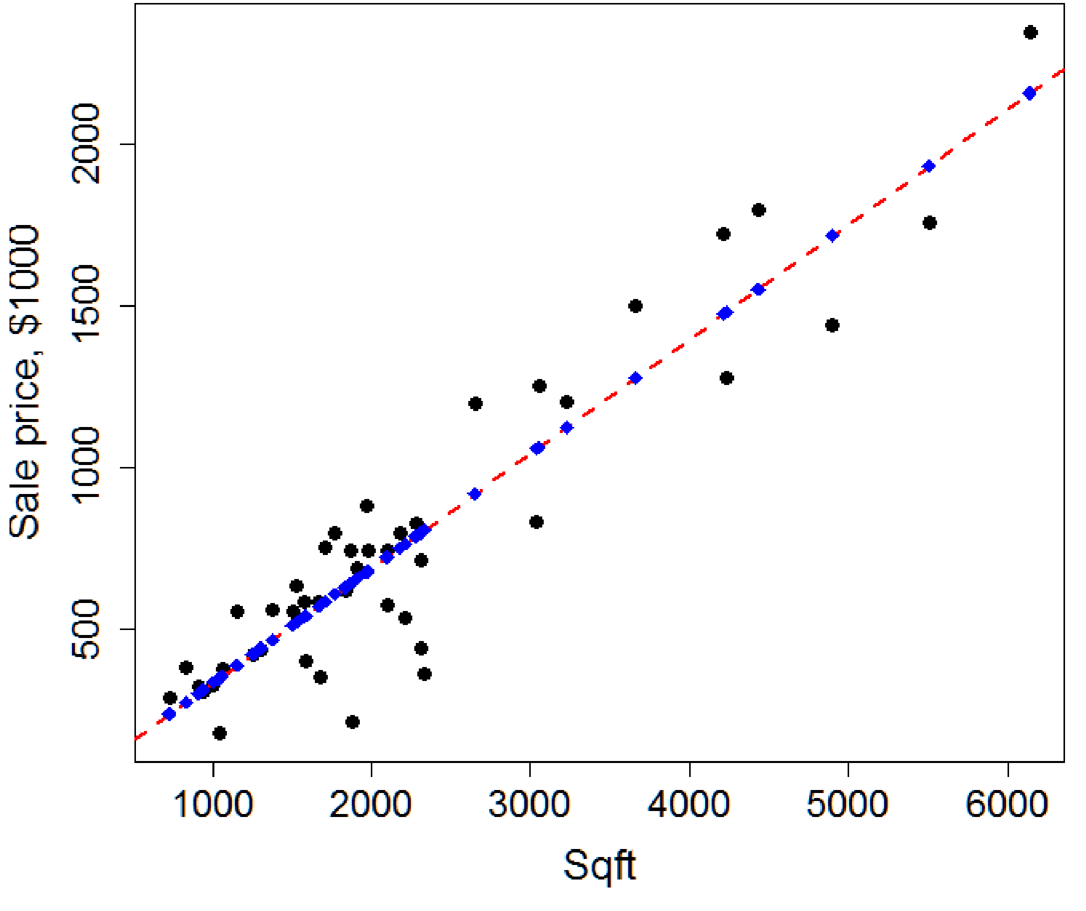
Suppose $\hat{\beta}_0, \hat{\beta}_1, \text{and } \hat{\sigma}^2$ are functions of the data, X_1 , … , X_n and Y_1 , … , Y_n, that estimate $\beta_0, \beta_1, \text{and } \sigma^2$, respectively.

Fitted Values: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

Residuals: $r_i = Y_i - \hat{Y}_i = Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right)$

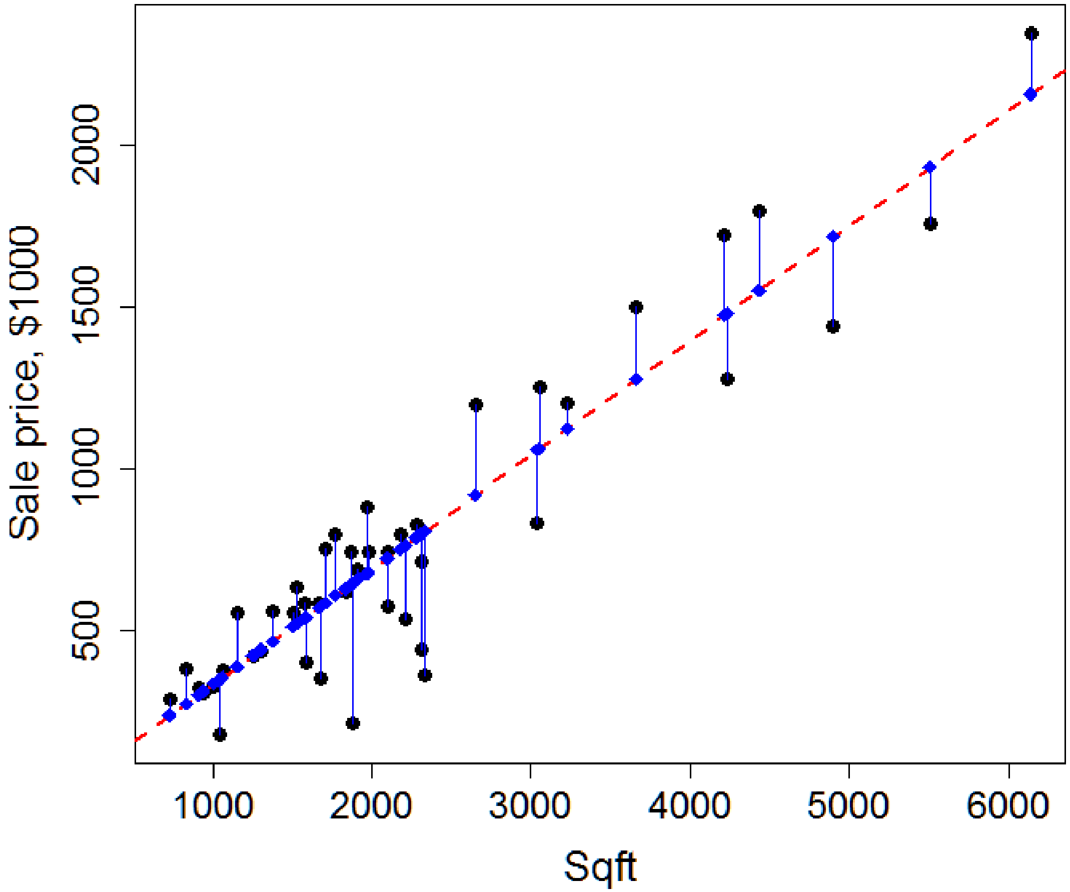
Regression line: Fitted Values

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$



Regression line: Residuals

$$r_i = Y_i - \hat{Y}_i$$



Simple Linear Regression: Line of Best Fit

Idea: Find $\hat{\beta}_0$ and $\hat{\beta}_1$ such that they minimize a certain function of the magnitudes of all residuals, $|r_i|$.

Historically, the default was to minimize the SSR,

$$\sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n}\left(Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right)\right)^2.$$

However, we can also minimize $\sum_{i=1}^{n} |r_i| = \sum_{i=1}^{n}\left|Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right)\right|$ (a form of robust regression)

or any other *distance* between Y_i and $\hat{\beta}_0 + \hat{\beta}_1 X_i$,

$$d(Y_i, \hat{\beta}_0 + \hat{\beta}_1 X_i)$$

http://sambaker.com/courses/J716/demos/LeastSquares/LeastSquaresDemo.html

Simple Linear Regression: Minimizing $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ is the same as using MLE's of $\hat{\beta}_0$ and $\hat{\beta}_1$

Maximum Likelihood Estimation: finds parameters that maximize $f(Y_1, Y_2, \ldots, Y_n \mid X_1, X_2, \ldots, X_n; \boldsymbol{\theta})$.

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

$$f(Y_1, Y_2, \ldots, Y_n \mid X_1, X_2, \ldots, X_n; \beta_0, \beta_1, \sigma^2) \overset{(i.i.d.)}{=} \prod_{i=1}^{n} f(Y_i \mid X_i; \beta_0, \beta_1, \sigma^2)$$

$$= \prod_{i=1}^{n} \left(\frac{1}{\sigma\sqrt{2\pi}} e^{-(Y_i - (\beta_0 + \beta_1 X_i))^2 / 2\sigma^2} \right)$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}} \right)^{\mathrm{n}} \prod_{i=1}^{n} \left(e^{-(Y_i - (\beta_0 + \beta_1 X_i))^2 / 2\sigma^2} \right) \propto \frac{1}{\sigma^n} e^{-\sum_{i=1}^{n}(Y_i - (\beta_0 + \beta_1 X_i))^2 / 2\sigma^2}$$

Simple Linear Regression: Minimizing $\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$ is the same as using MLE's of $\hat{\beta}_0$ and $\hat{\beta}_1$

Maximum Likelihood Estimation: finds parameters that maximize $f(Y_1, Y_2, \ldots, Y_n \mid X_1, X_2, \ldots, X_n; \boldsymbol{\theta})$.

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

$$\arg\max_{\beta_0,\beta_1}\left[f(Y_1, Y_2, \ldots, Y_n \mid X_1, X_2, \ldots, X_n; \beta_0, \beta_1, \sigma^2)\right]$$

$$= \arg\max_{\beta_0,\beta_1}\left[\frac{1}{\sigma^n} e^{-\sum_{i=1}^{n}\left(Y_i - (\beta_0 + \beta_1 X_i)\right)^2 / 2\sigma^2}\right] = \arg\max_{\beta_0,\beta_1}\left[-\sum_{i=1}^{n}\left(Y_i - (\beta_0 + \beta_1 X_i)\right)^2 / 2\sigma^2\right]$$

$$= \arg\min_{\beta_0,\beta_1}\left[\sum_{i=1}^{n}\left(Y_i - (\beta_0 + \beta_1 X_i)\right)^2\right]$$

Least Squares Estimation

$$\left(\hat{\beta}_0, \hat{\beta}_1\right) = \arg\min_{\beta_0, \beta_1} \left[\sum_{i=1}^{n} \left(Y_i - \left(\beta_0 + \beta_1 X_i\right)\right)^2 \right]$$

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

Intercept: $$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

What is the sampling distribution of these estimators?

Exact Sampling Distributions of the Least Squares Estimators

X_i are fixed for all i

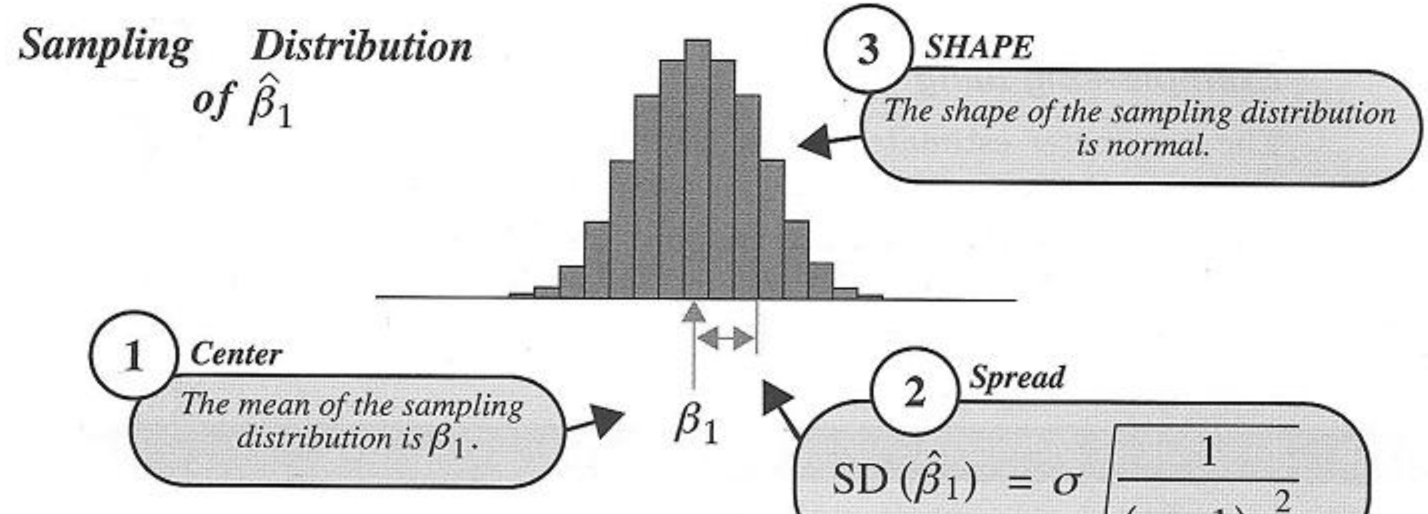
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \text{ where } \varepsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$$

Slope: $\hat{\beta}_1 \sim N\left(\beta_1, \sigma^2 \frac{1}{\sum_{i=1}^n \left(X_i - \bar{X}\right)^2}\right)$

Unbiased!

Intercept: $\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n \left(X_i - \bar{X}\right)^2}\right]\right)$

$$\sum_{i=1}^n \left(X_i - \bar{X}\right)^2 = (n-1)S_X^2$$



$$\hat{\beta}_1 \sim N\left(\beta_1, \sigma^2 \frac{1}{(n-1)S_X^2}\right)$$

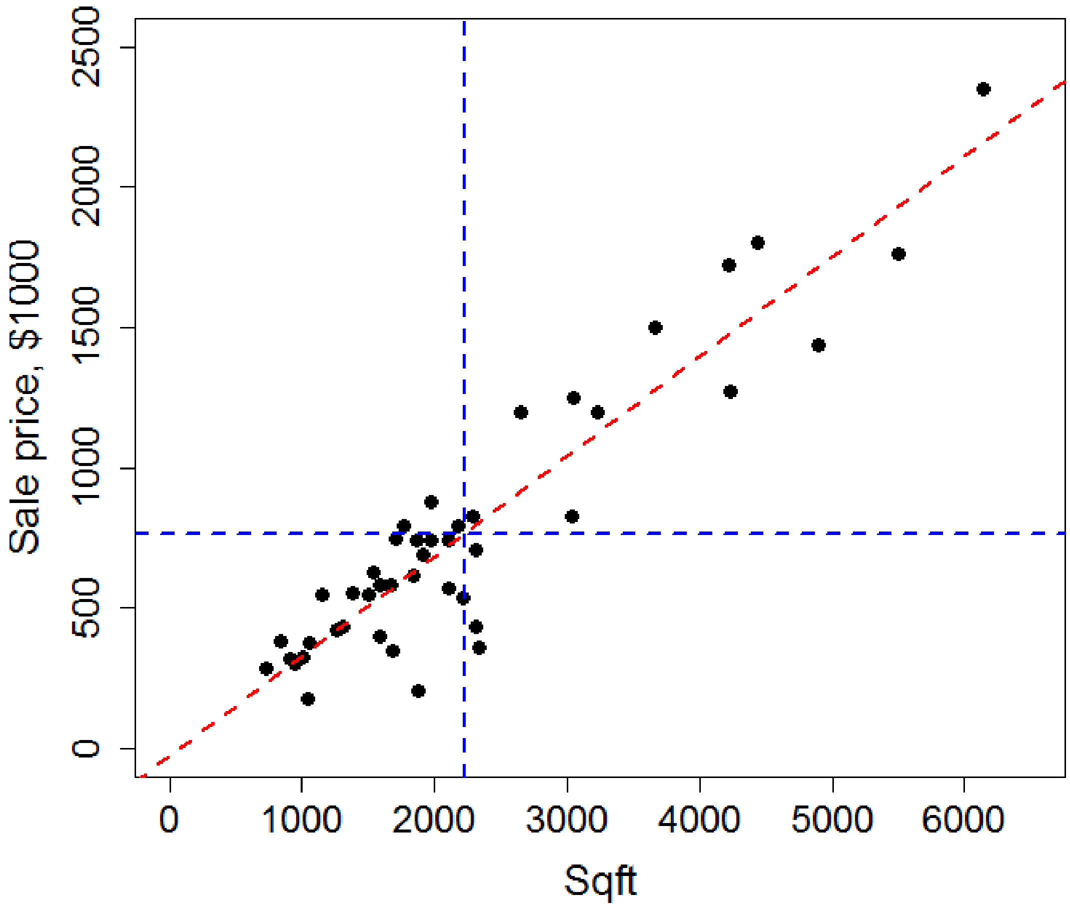
Regression line: Properties

The regression line passes through points $(0, \hat{\beta}_0)$ and $(\bar{X}, \bar{Y})$.

$\hat{\beta}_0 = -25.97$

$\bar{Y} = 766$

$\bar{X} = 2{,}225$



Estimator of Residual Variance and its Sampling Distribution

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

σ^2 is unknown. It is estimated as follows:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2}{n-2} = \frac{\sum_{i=1}^{n} \left(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)\right)^2}{n-2} = \frac{\text{SSR}}{\text{d.f.}}$$

Sampling distribution of the sample variance:

$$\hat{\sigma}^2 \sim \frac{\sigma^2 \chi^2_{n-2}}{n-2}$$

t-tests for Least Squares Estimates

Slope:

$$H_0 : \beta_1 = \beta_1^0$$

$$H_A : \beta_1 \neq \beta_1^0 \text{ or } \beta_1 > \beta_1^0 \text{ or } \beta_1 < \beta_1^0$$

$$\frac{\hat{\beta}_1 - \beta_1^0}{\hat{\sigma}\sqrt{\frac{1}{(n-1)S_X^2}}} \sim t_{\boxed{n-2}}$$

Intercept:

$$H_0 : \beta_0 = \beta_0^0$$

$$H_A : \beta_0 \neq \beta_0^0 \text{ or } \beta_0 > \beta_0^0 \text{ or } \beta_0 < \beta_0^0$$

$$\frac{\hat{\beta}_0 - \beta_0^0}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{(n-1)S_X^2}}} \sim t_{\boxed{n-2}}$$

$$\text{where } S_X^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1}$$

CIs for Least Squares Estimates

$$\hat{\beta}_1 \pm t_{n-2,1-\alpha/2} SE(\hat{\beta}_1), \text{ where } SE(\hat{\beta}_1) = \hat{\sigma}\sqrt{\frac{1}{(n-1)S_X^2}},$$

$$\hat{\beta}_0 \pm t_{n-2,1-\alpha/2} SE(\hat{\beta}_0), \text{ where } SE(\hat{\beta}_0) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{(n-1)S_X^2}}.$$

Newton houses: find 95% CIs for the slope and the intercept.

$$n = 46,\ \overline{X} = 2{,}225,\ S_X = 1{,}251,\ \hat{\beta}_0 = -25.97,\ \hat{\beta}_1 = 0.356,\ \hat{\sigma} = 181$$

```
qt(0.975,44) = 2.015
```

For $\hat{\beta}_1$ it is $(0.32, 0.40)$, for $\hat{\beta}_0$ it is $(-136, 84)$.

CIs for Least Squares Estimates

$$\hat{\beta}_1 \pm t_{n-2,1-\alpha/2} SE(\hat{\beta}_1), \text{ where } SE(\hat{\beta}_1) = \hat{\sigma}\sqrt{\frac{1}{(n-1)S_X^2}},$$

$$\hat{\beta}_0 \pm t_{n-2,1-\alpha/2} SE(\hat{\beta}_0), \text{ where } SE(\hat{\beta}_0) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{(n-1)S_X^2}}.$$

- These CIs are *individual,* they do not preserve the familywise confidence level at (1-α)100%.
- A joint confidence region can be constructed and tested using an *F*-test (covered in Ch. 8).

Linear Regression in R

```
> regmodel <- lm(Price/1000 ~ Sqft., data = SaleData)
> summary(regmodel)

Call:
lm(formula = Price/1000 ~ Sqft., data = SaleData)

Residuals:
    Min      1Q  Median      3Q     Max 
-445.09 -125.97   36.45  107.27  281.39 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -25.96758   54.77713  -0.474    0.638    
Sqft.         0.35607    0.02152  16.549   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 180.7 on 44 degrees of freedom
Multiple R-squared: 0.8616, Adjusted R-squared: 0.8584 
F-statistic: 273.9 on 1 and 44 DF,  p-value: < 2.2e-16
```

$$\hat{\sigma}\sqrt{\frac{1}{n}+\frac{\overline{X}^2}{(n-1)S_X^2}}$$

$$\hat{\sigma}\sqrt{\frac{1}{(n-1)S_X^2}}$$

$$\hat{\sigma}$$

$H_0 : \beta_0 = 0$

$H_A : \beta_0 \neq 0$

and

$H_0 : \beta_1 = 0$

$H_A : \beta_1 \neq 0$

Interpretation of Least Squares Estimates

- Intercept estimates $\mu\{Y|X=0\}$
- Slope estimates the change from $\mu\{Y|X=x\}$ to $\mu\{Y|X=x+1\}$

- If levels of X were randomized among study units (e.g., drug dose) then *causal interpretation is allowed.*
- Otherwise, only *association,* i.e., it is estimated that 1-unit increase in X *is associated with* $\hat{\beta}_1$ change in $\mu\{Y\}$ (or in "average outcome", $E(Y)$).

Pearson Correlation and its Connection to Simple Linear Regression

Pearson Correlation

- Correlation $\rho \in (-1, 1)$ is a measure of a degree of association between two random variables.

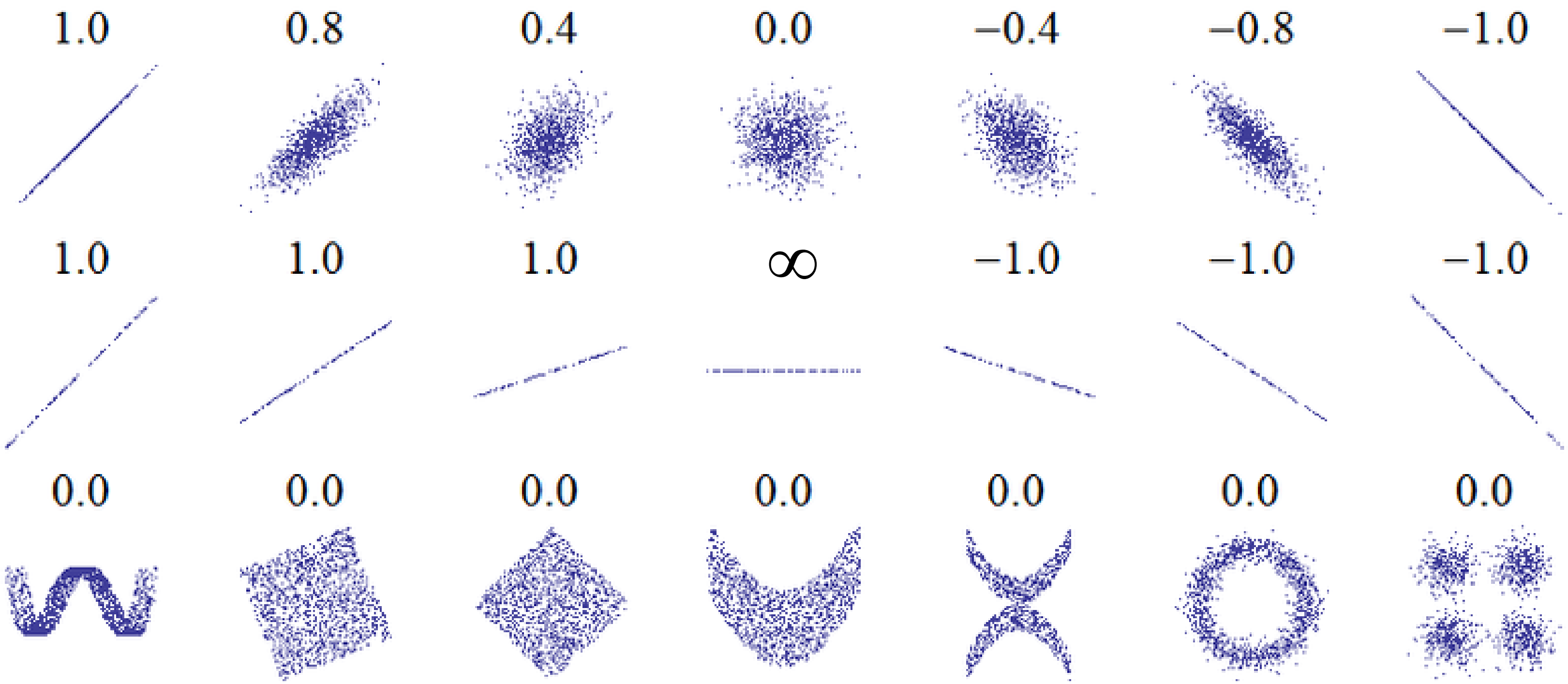
- Pearson product-moment correlation coefficient,

$$\rho_{XY} = \frac{E\big((X - \mu_X)(Y - \mu_Y)\big)}{\sigma_X \sigma_Y},$$

is a measure of linear association between two random variables.

Pearson Correlation

$$\rho_{XY} = \frac{E\left(\left(X - \mu_X\right)\left(Y - \mu_Y\right)\right)}{\sigma_X \sigma_Y}$$



Sample Estimator of Pearson Correlation

$$\rho_{XY} = \frac{E\left(\left(X - \mu_X\right)\left(Y - \mu_Y\right)\right)}{\sigma_X \sigma_Y}$$

Pearson correlation is estimated from the observed data as follows:

$$\hat{\rho}_{XY} = r_{XY} = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right) \Big/ (n-1)}{S_X S_Y}$$

Note that correlation does not have units.

Connection between Sample Correlation and Least Squares Estimates

Slope: $$\hat{\beta}_1 = \frac{r_{XY} S_Y}{S_X}$$

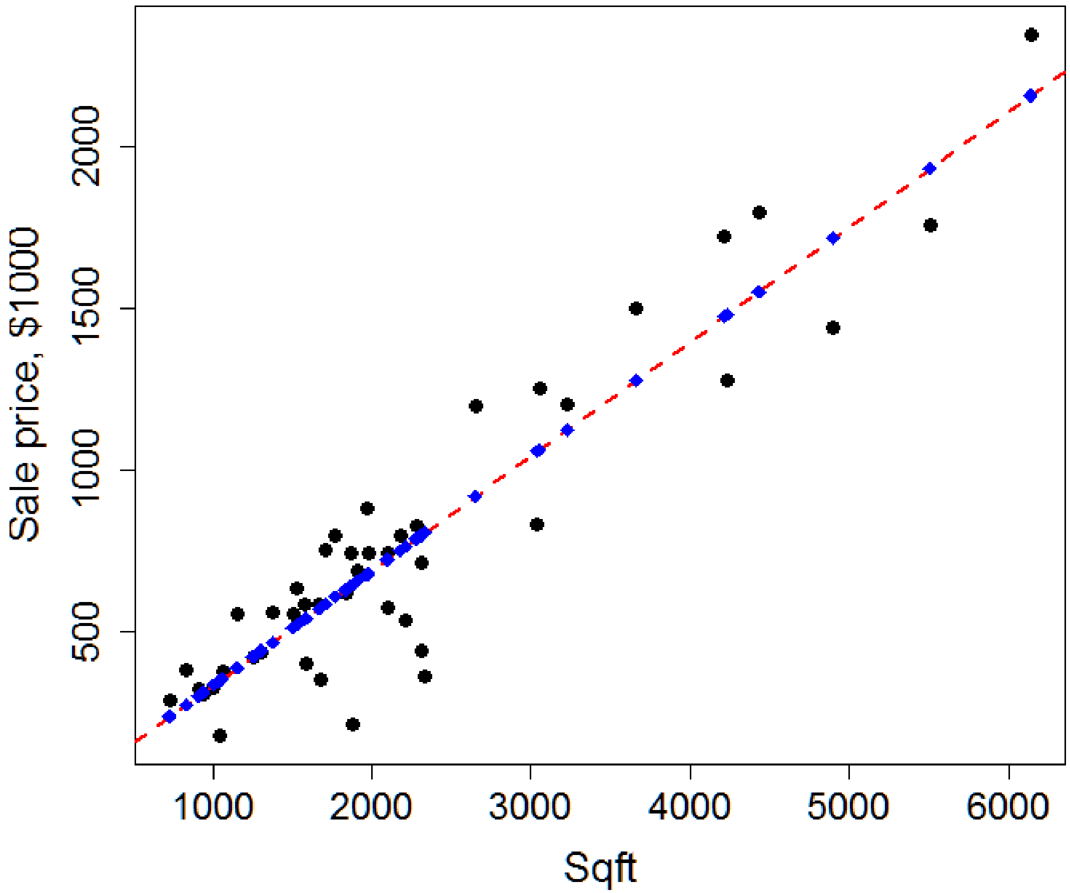
Intercept: $$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Newton houses: show that the slope ≈ 0.356, and the intercept ≈ -26:

$$S_X = 1{,}252;\quad S_Y = 480;\quad r_{XY} = 0.93;\quad \bar{X} = 2{,}225;\quad \bar{Y} = 766$$

Newton Homes: Estimated Regression Line

$$\hat{\mu}\{\text{Price} \mid \text{Sqft}\} = -26 + 0.356 \cdot \text{Sqft}, \text{ and } \hat{\sigma} = 181$$



$$r_{XY} = 0.93$$

$$\hat{\beta}_1 = \frac{r_{XY} S_Y}{S_X}$$

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

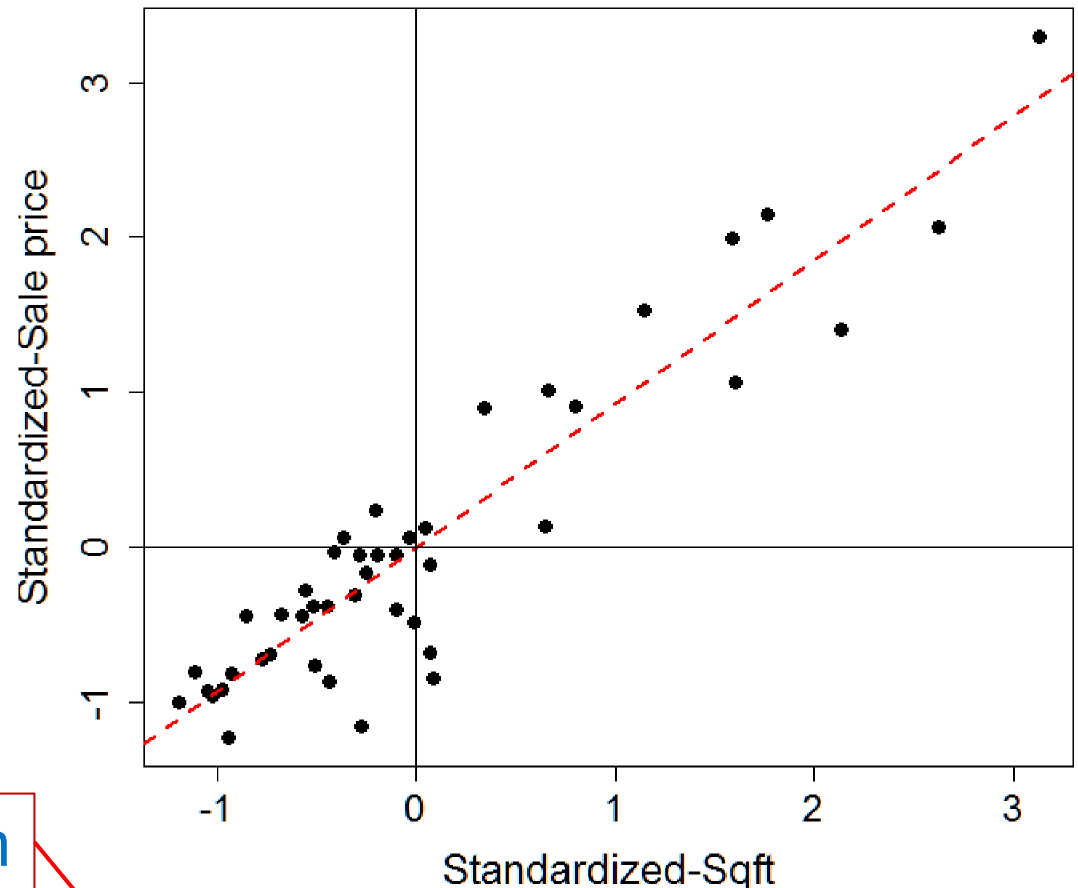
Interpret parameters and name units for each of them.

Regression Line for Standardized Variables

$$\tilde{Y}_i = \frac{Y_i - \bar{Y}}{S_Y}$$

$$\tilde{X}_i = \frac{X_i - \bar{X}}{S_X}$$

$$\hat{\mu}\left(\tilde{Y}_i \mid \tilde{X}_i\right) = r_{XY}\,\tilde{X}_i$$



Sample correlation

$$\hat{\mu}\{\widetilde{\text{Price}} \mid \widetilde{\text{Sqft}}\} = 0.93 \cdot \widetilde{\text{Sqft}}, \text{ and } \hat{\sigma} = 0.376$$