STAT 139: STATISTICAL SLEUTHING THROUGH LINEAR MODELS

Lecture 16
Oct 28, 2014

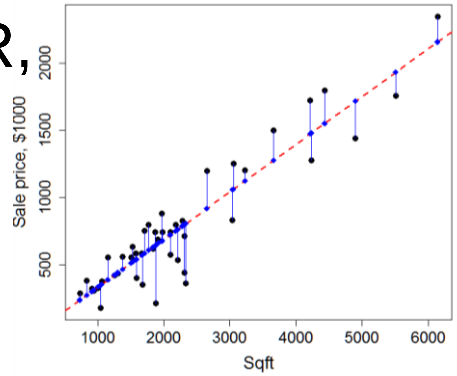
Victoria Liublinska

Odds and Ends

- HW 7 is due on Fri, 10/31
  - Problem 2: 2nd ed. of the textbook has a typo on Display 7.17 (corrected in the 3rd ed.): the SE for Netherlands is 0.028.
- HW6 solution has been posted.
- Midterm solution will be posted by the end of the day.
- Check-out Piazza if you are still looking for a project partner.

Previous lecture: Review

- The line of best fit is found by minimizing SSR,

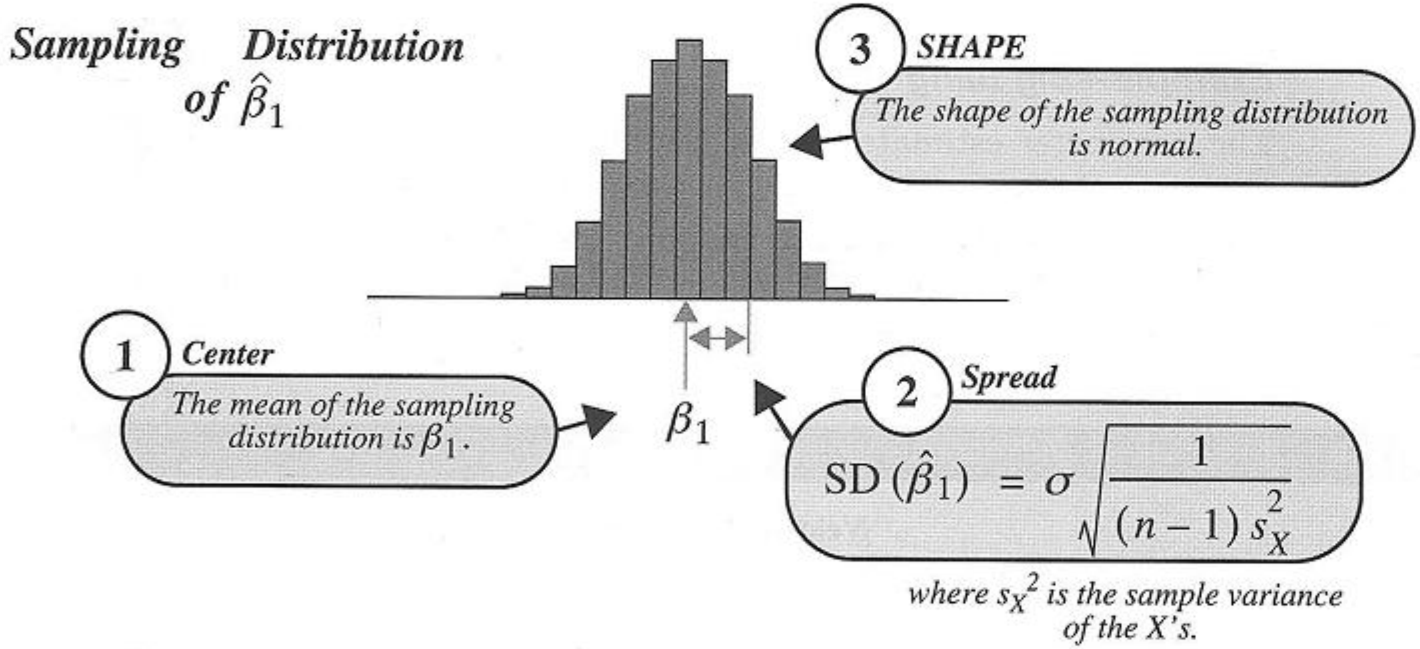


$$\sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} \left(Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right)\right)^2.$$

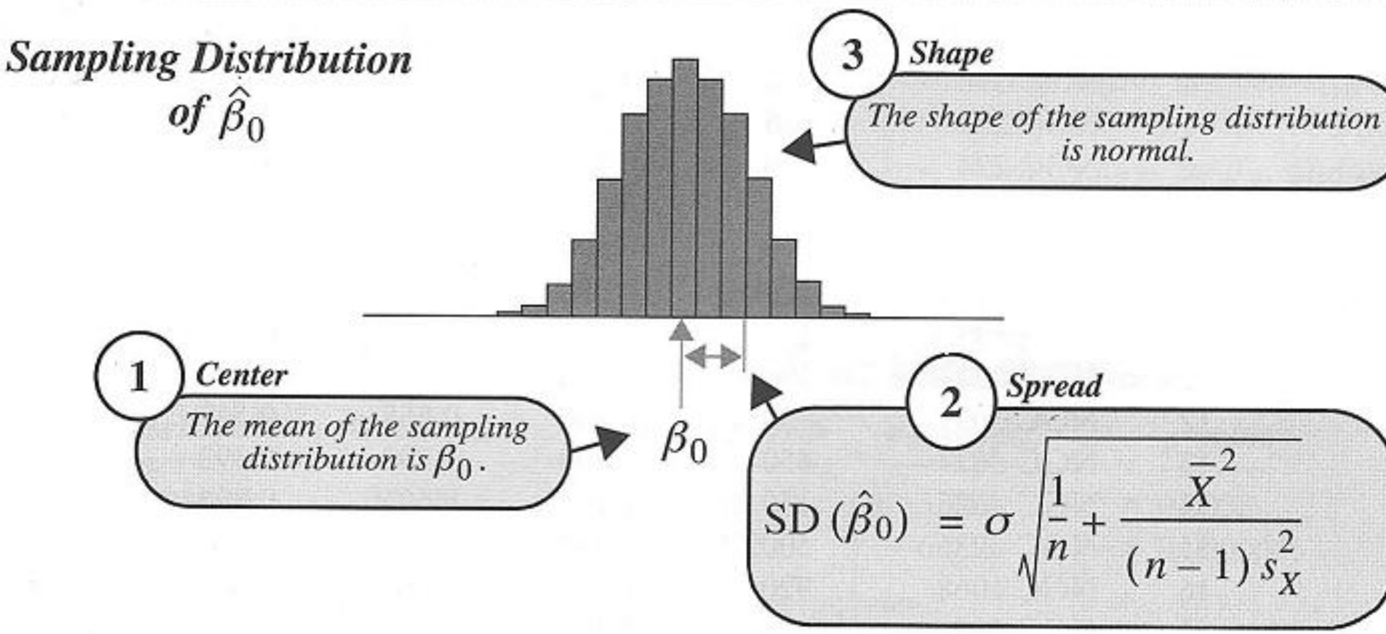
- **Slope**: $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2} = \frac{r_{XY} S_Y}{S_X}$$

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

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



Sampling Distribution of $\hat{\beta}_1$
3 SHAPE
The shape of the sampling distribution is normal.
1 Center
The mean of the sampling distribution is β_1.
β_1
2 Spread
$\mathrm{SD}(\hat{\beta}_1) = \sigma \sqrt{\frac{1}{(n-1)\, s_X^2}}$
where s_X^2 is the sample variance of the X's.



Sampling Distribution of $\hat{\beta}_0$
3 Shape
The shape of the sampling distribution is normal.
1 Center
The mean of the sampling distribution is β_0.
β_0
2 Spread
$\mathrm{SD}(\hat{\beta}_0) = \sigma \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n-1)\, s_X^2}}$

Previous lecture: Review

- Residual variance and its sampling distribution:

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

- Interpretation of regression coefficients;
- *t*-tests and confidence intervals for slope and intercept;
- R function `lm()`;
- Properties of the least-squares line.

Today's overview

- Regression equation for standardized variables.
- Regression to the mean and regression fallacy.
- Prediction of mean response and future response at $X=X_0$.

Reading (same is in lect. 15):

- **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).

Regression Line for Standardized Variables: Interpretation

$$\tilde{Y}_i = \frac{Y_i - \bar{Y}}{S_Y};\ \tilde{X}_i = \frac{X_i - \bar{X}}{S_X} \Rightarrow \bar{\tilde{X}} = \bar{\tilde{Y}} = 0 \text{ and } S_{\tilde{X}} = S_{\tilde{Y}} = 1.$$

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

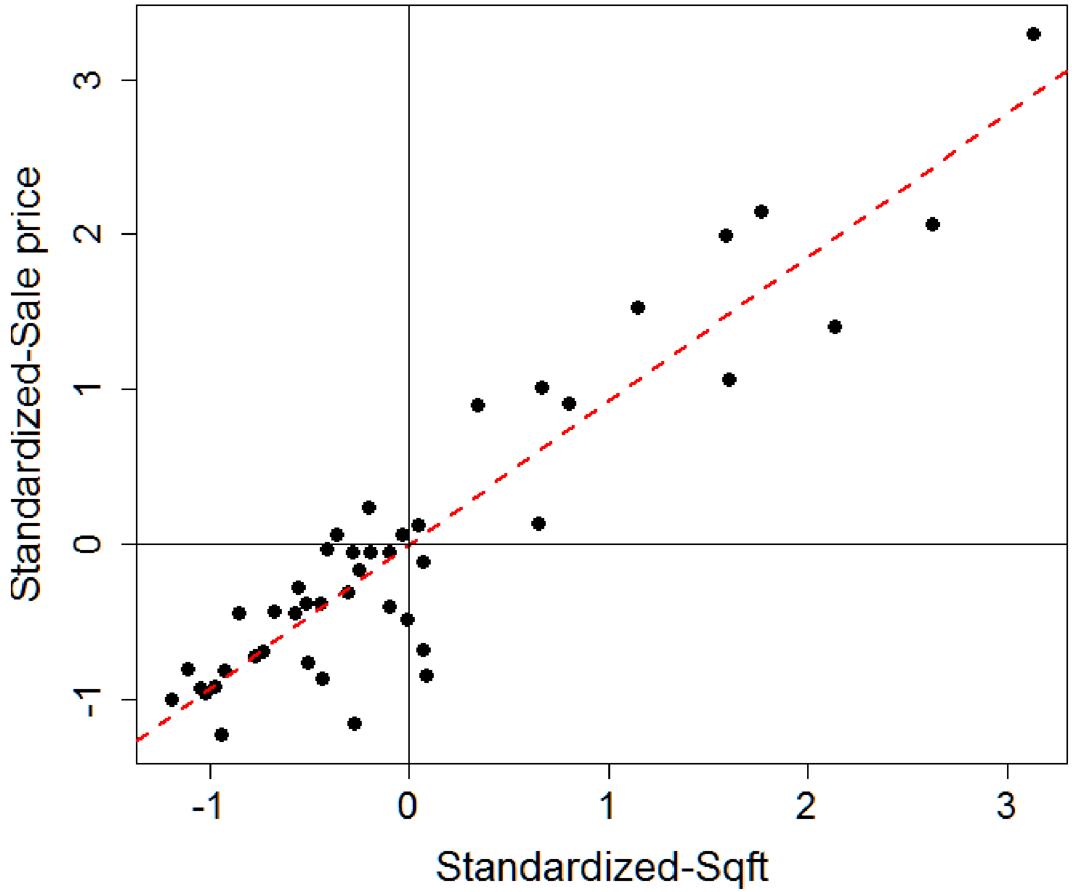
$$r_{\tilde{X}\tilde{Y}} = \frac{\sum_{i=1}^{n}\left(\tilde{X}_i - \bar{\tilde{X}}\right)\left(\tilde{Y}_i - \bar{\tilde{Y}}\right)\Big/(n-1)}{S_{\tilde{X}} S_{\tilde{Y}}} = \sum_{i=1}^{n}\tilde{X}_i\tilde{Y}_i\Big/(n-1) = r_{XY}$$

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

$$\hat{\mu}\{\widetilde{\text{Price}} \mid \widetilde{\text{Sqft}}\} = 0.93 \cdot \widetilde{\text{Sqft}}, \text{ and } \hat{\sigma} = \sqrt{\sum_{i=1}^{n}\left(\tilde{Y}_i - r_{XY}\cdot\tilde{X}_i\right)^2 \Big/ (n-2)} = 0.376$$

Regression Line for Standardized Variables: Interpretation

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

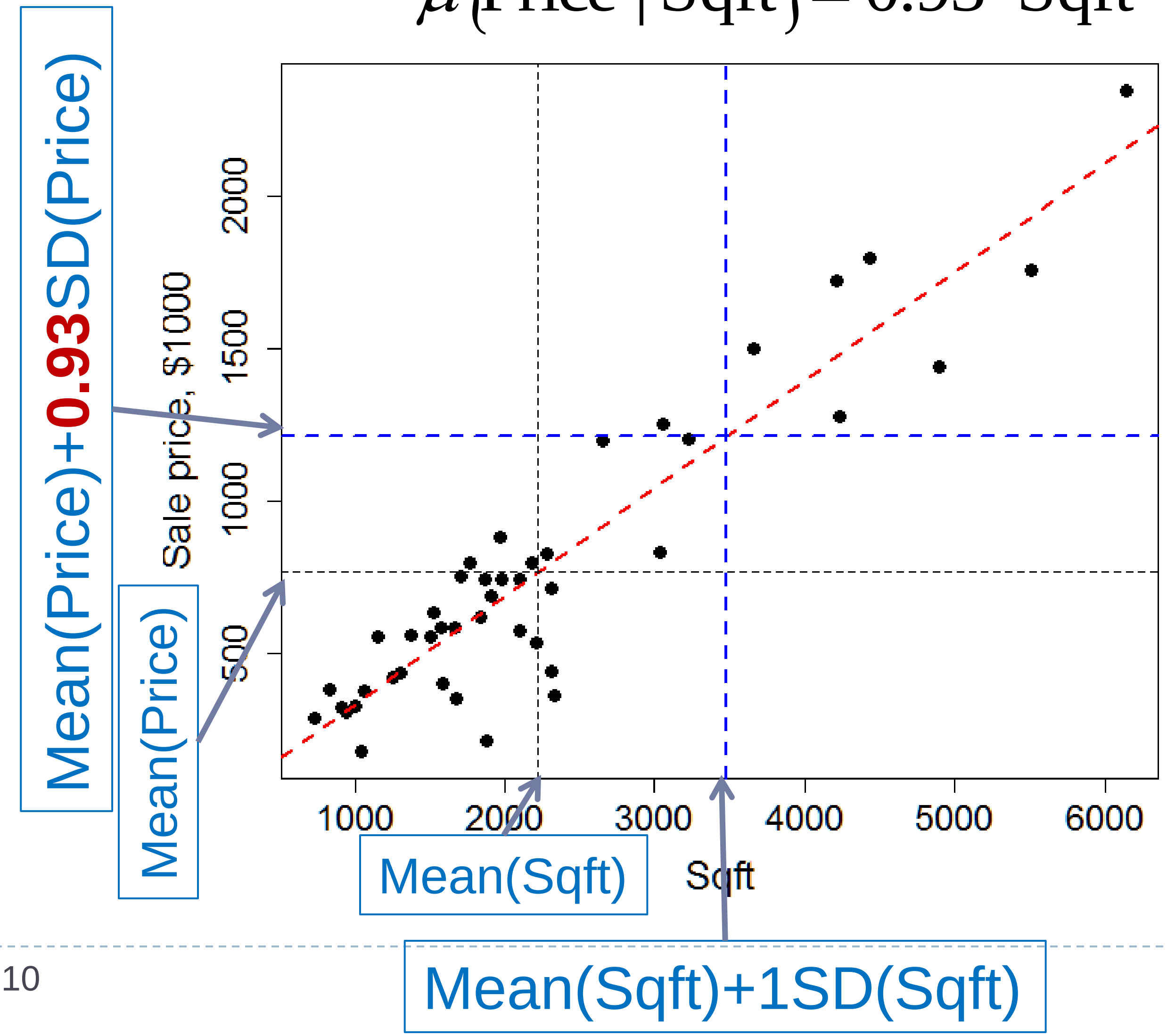
One-unit change in $\tilde{X}_i$ is equivalent to a one-SD change in X_i .

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

One-unit change in $\tilde{Y}_i$ is equivalent to a one-SD change in Y_i .

Regression Line for Standardized Variables: Interpretation

$$\hat{\mu}\{\widetilde{\text{Price}} \mid \widetilde{\text{Sqft}}\} = 0.93 \cdot \widetilde{\text{Sqft}}$$



$\text{Mean(Sqft)} = 2{,}225$

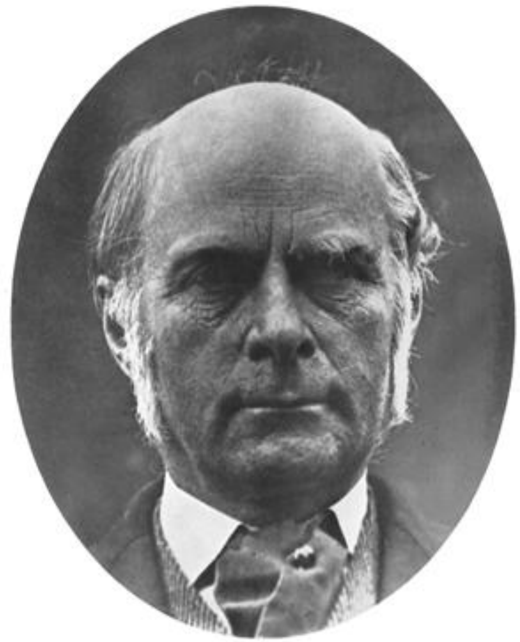
$\text{Mean(Price)} = 766$

$\text{SD(Sqft)} = 1{,}252$

$\text{SD(Price)} = 480$

Regression to the Mean

Sir Francis Galton
England, (1822 - 1911)



Geographer, meteorologist, inventor of fingerprint identification, pioneer of statistical correlation and regression, convinced hereditarian, eugenicist.

Sir Francis Galton: Facts and achievements

- Half-cousin of Charles Darwin.
- Fellow of Royal Society of London, was knighted in 1909.
- Inventor of the silent dog whistle.
- Coiner of the term “anticyclone” (meteorology).
- Inventor of **quincunx** (“Galton box”) (1873) (video).
- Author of 3 books on fingerprints in forensic science (at the age of 80!).

Regression to the Mean

- Sir Francis collected data on 1,078 of adult sons and their fathers and estimated

$$\mu(H_S \mid H_F),$$

where H_S is son's height and H_F is father's height.

- For standardized heights, Sir Francis estimated that

$$\hat{\mu}(\widetilde{H}_S \mid \widetilde{H}_F) = 0.44\widetilde{H}_F$$

"*It is a universal rule that the unknown kinsman in any degree of any specified man, is probably more mediocre than he*." (Francis Galton, 1886)

Regression to the Mean

$$\hat{\mu}\left(\tilde{H}_S \mid \tilde{H}_F\right) = 0.44\tilde{H}_F$$

- As compared to father's height, H_F, son's height, H_S, *regresses to the mean*: *Tall fathers will have, on average, shorter sons, and short fathers will have, on average, taller sons.*

Then why don't we all have the same height by now?

- Note that, if we regress father's heights on son's heights, we will get the same equation!

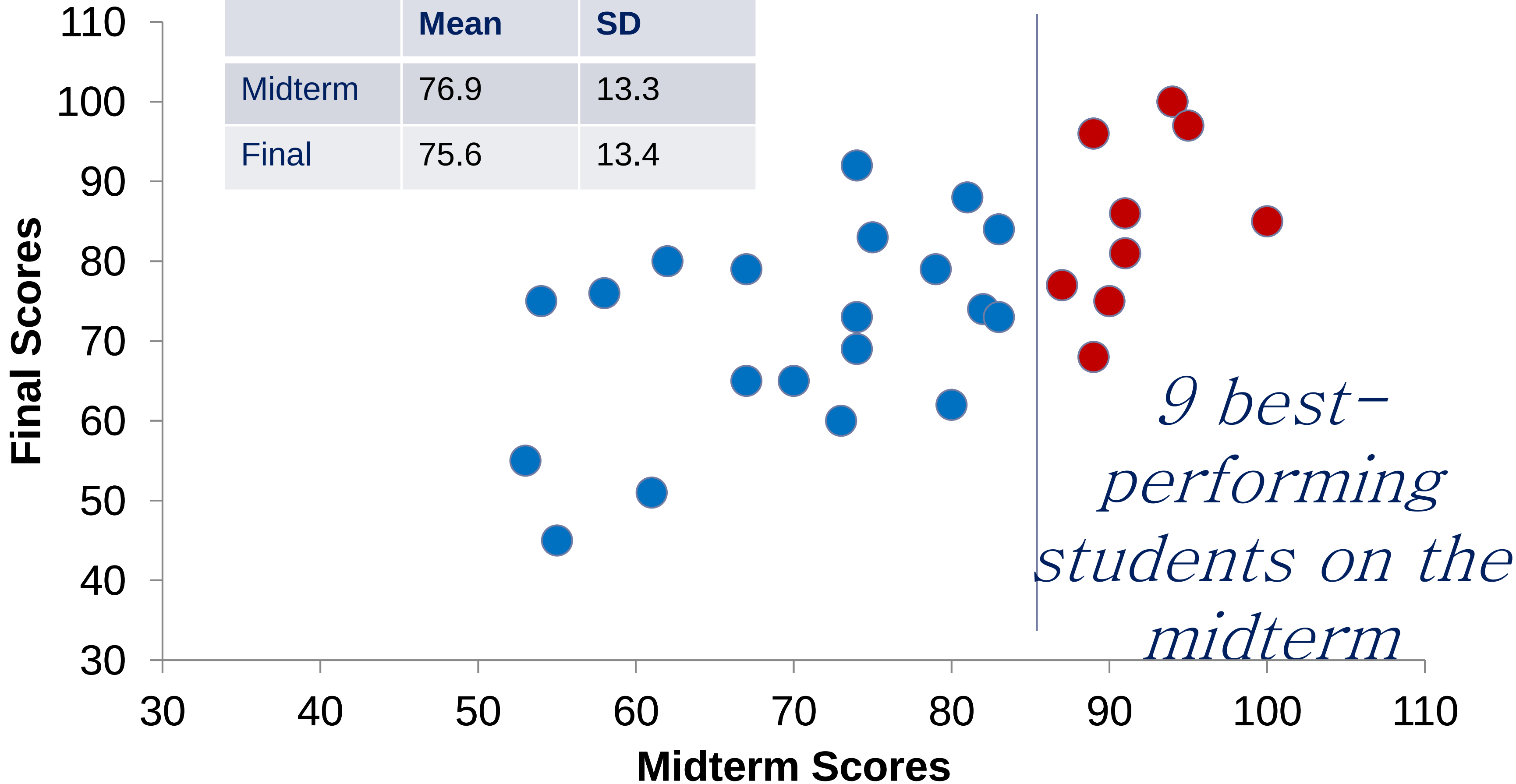
$$\hat{\mu}\left(\tilde{H}_F \mid \tilde{H}_S\right) = 0.44\tilde{H}_S$$

Regression Effect

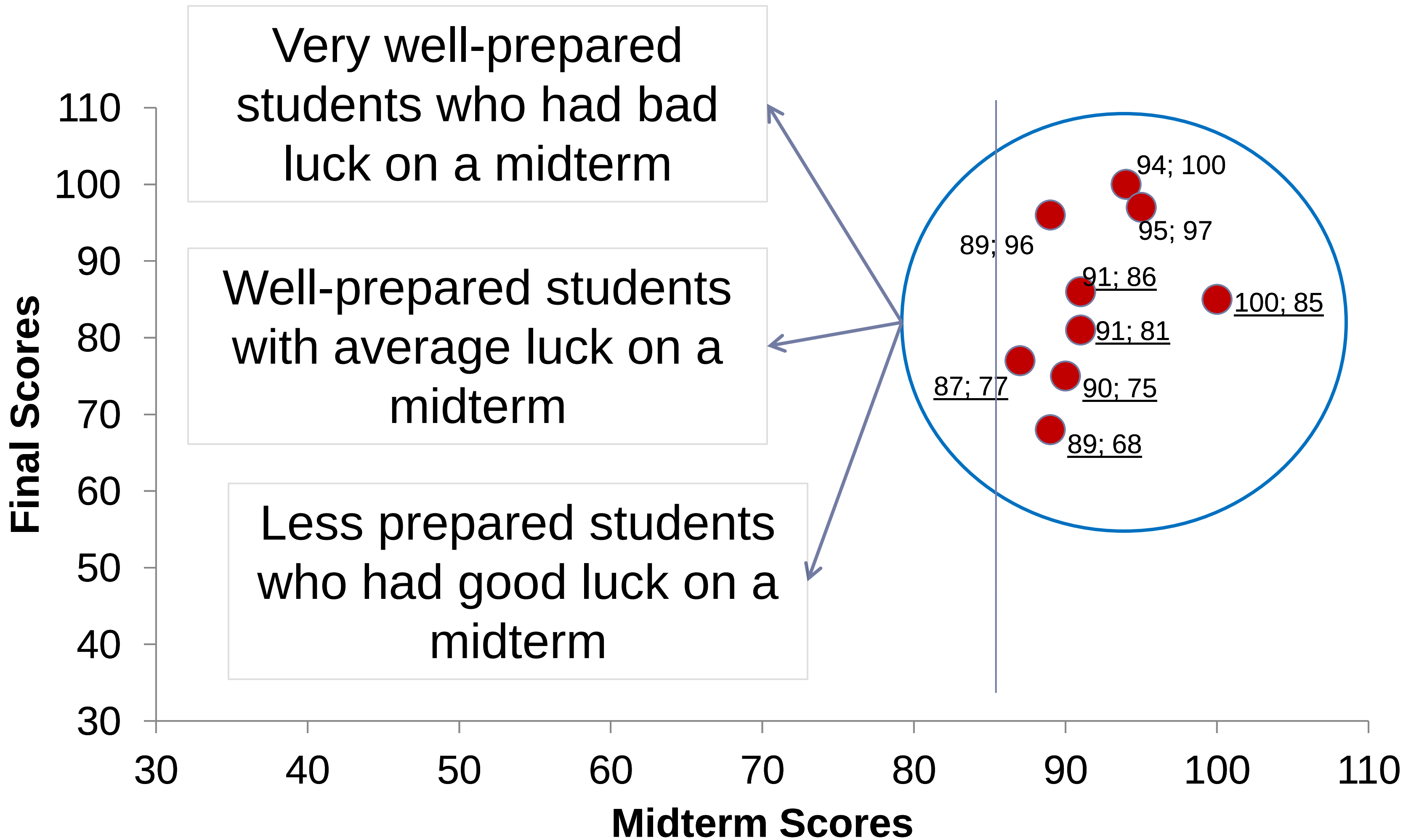
Regression effect (*regression toward the mean*) implies that if you take *new measurements* of something that can vary, the mean of the group that measured lower than population mean will go up, and the mean of the group that measured higher than population mean will go down.

Midterm and Final Exam Scores

	Mean	SD
Midterm	76.9	13.3
Final	75.6	13.4

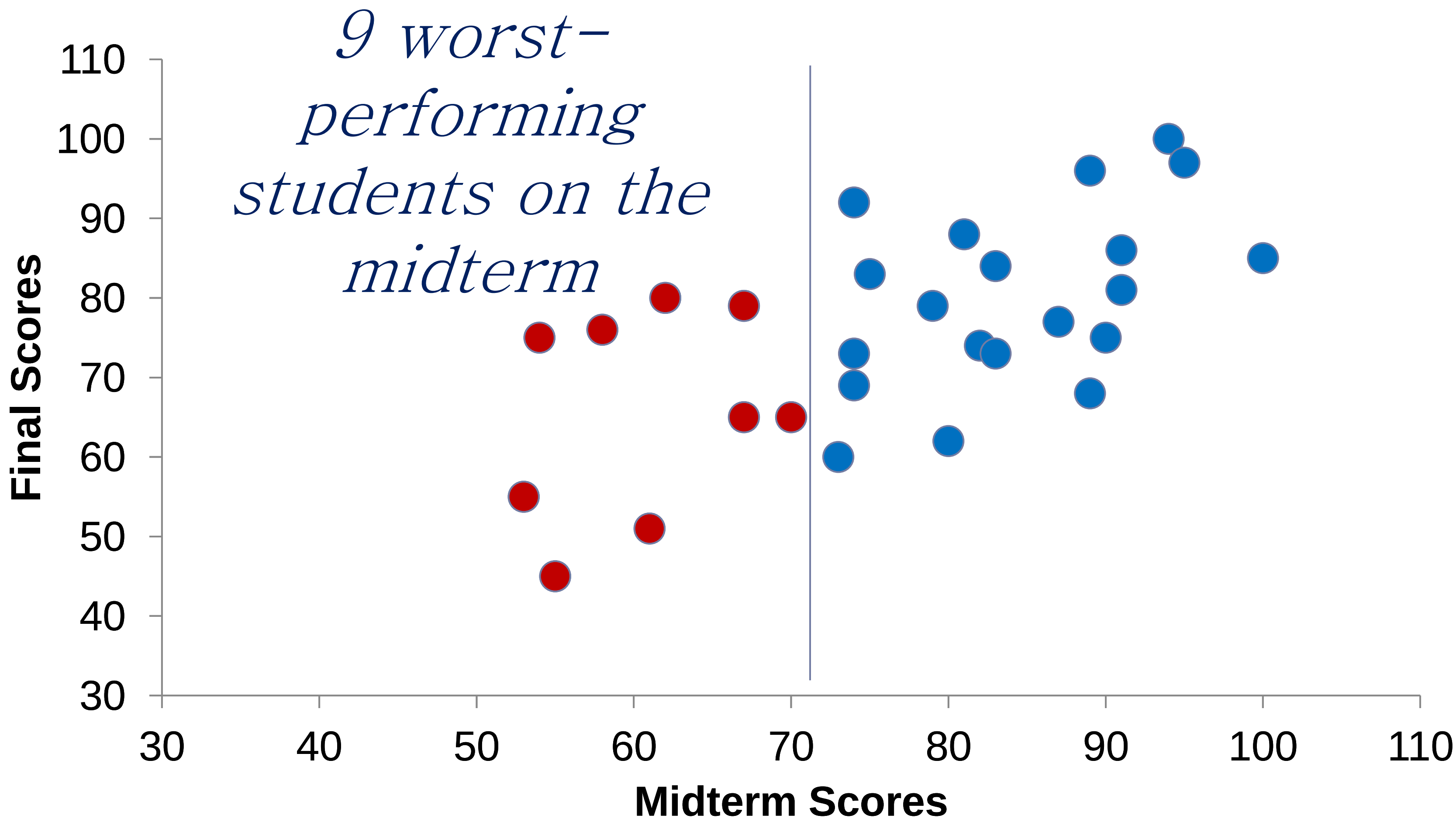


Three types of students with very high midterm scores

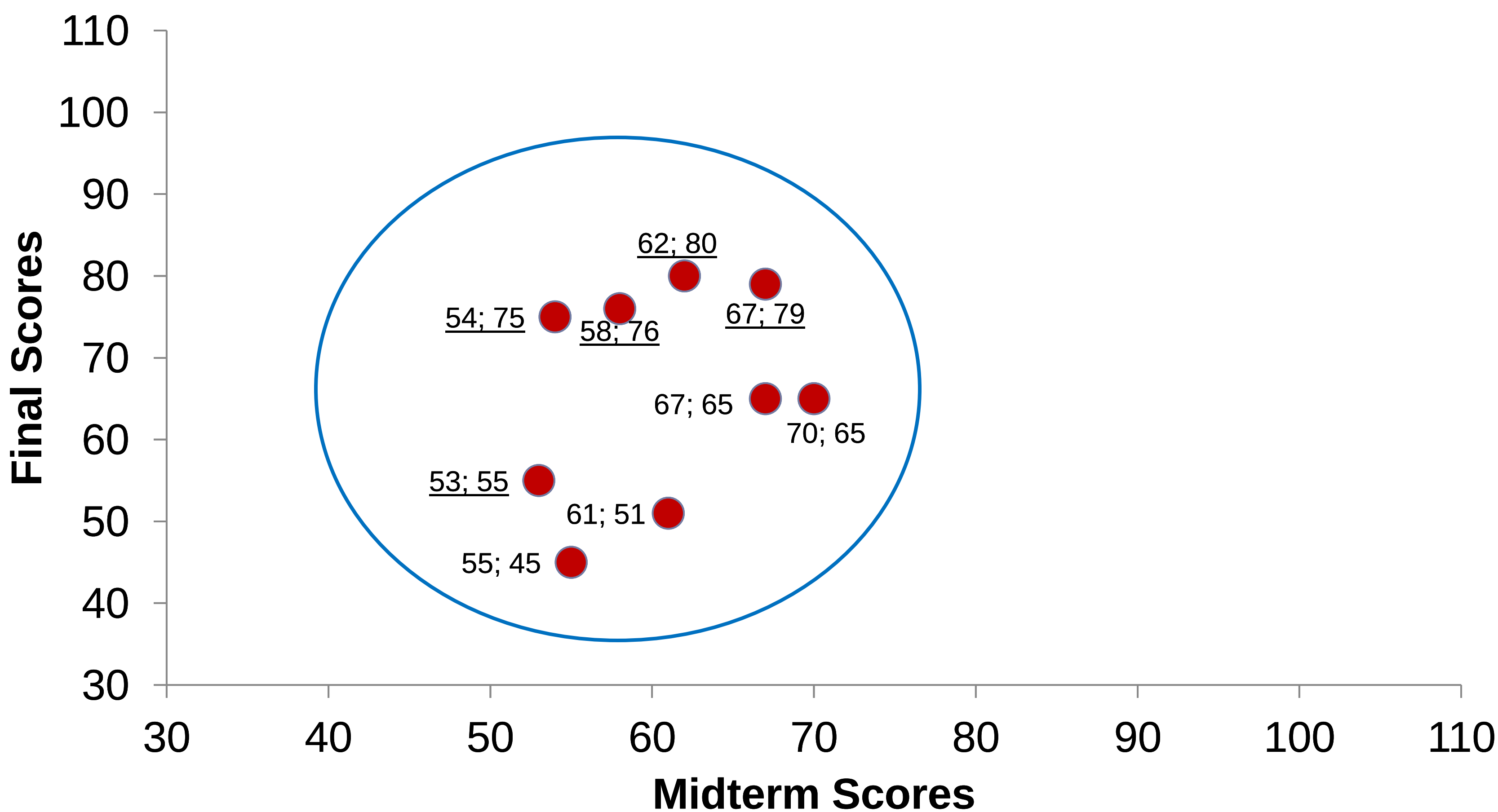


Midterm average is 91.8 → Final average is 85

Midterm and Final Exam Scores



Midterm and Final Exam Scores



Midterm average is 60.8 → Final average is 65.7

Regression Fallacy in Observational Studies

Regression fallacy is attributing the change in mean from the regression effect to some *cause*.

Subjects enrolled into a study on the basis of an extreme value of some measurement and a treatment is declared effective because subsequent measurements are not as extreme:

- Education programs applied to poorly performing schools;
- Diet pills provided to a group of overweight individuals, etc.

- Control group will help but only if the effect is *additive*.

Bonus question

Suppose a treatment is expected to:

- lower the post-test measurements of those with high pre-test measurements and
- raise the post-test measurements of those with low pre-test measurements.

For example, a broad-based health care program might be expected to raise mean birthweight in villages where birthweight was too low and lower mean birthweight in villages where birthweight was too high.

How can we detect the treatment effect and distinguish it from regression to the mean?

Estimation of Mean Response

Estimating Mean Response at a Particular Value of $X=X_0$

Last time we found μ(*Price*|*Sqft*=4000) = \$1,398K, i.e., the estimated average price of a 4,000 sq.ft. house is \$1,398K.

What is the error of this estimate?

More generally, what is the sampling distribution of the estimate of the mean response?

$$\hat{\mu}\{Y \mid X = X_0\} = \hat{\beta}_0 + \hat{\beta}_1 X_0$$

Estimating Mean Response at a Particular Value of $X=X_0$

$$\hat{\mu}\{Y \mid X = X_0\} = \hat{\beta}_0 + \hat{\beta}_1 X_0$$

Using the fact that

$$\hat{\beta}_1 \sim N\left(\beta_1, \sigma^2 \frac{1}{(n-1)S_X^2}\right) \text{ and } \hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left[\frac{1}{n} + \frac{\bar{X}^2}{(n-1)S_X^2}\right]\right),$$

$$E\left(\hat{\mu}\{Y \mid X = X_0\}\right) = \beta_0 + \beta_1 X_0,$$

$$Var\left(\hat{\mu}\{Y \mid X = X_0\}\right) = \sigma^2\left[\frac{1}{n} + \frac{\left(X_0 - \bar{X}\right)^2}{(n-1)S_X^2}\right].$$

Sampling Distribution of Mean Response at a Particular Value of $X=X_0$

$$\hat{\mu}\{Y \mid X = X_0\} = \hat{\beta}_0 + \hat{\beta}_1 X_0$$

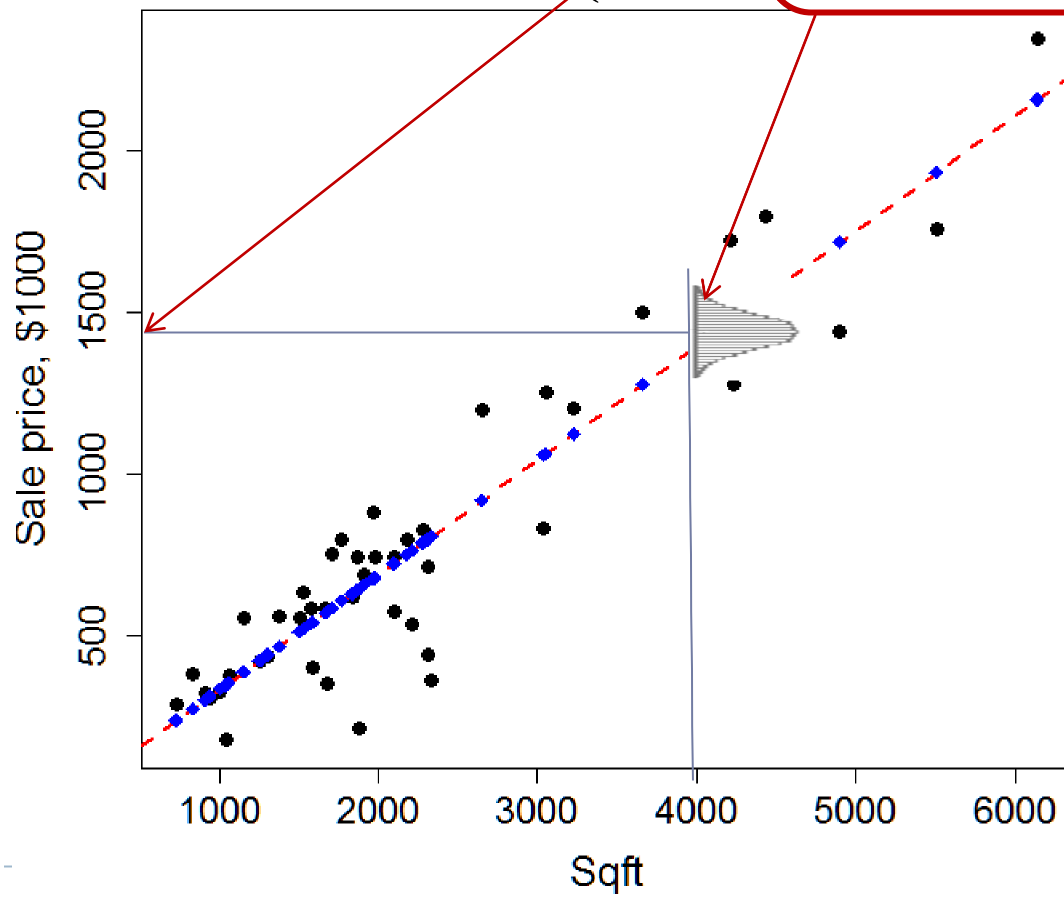
Therefore,

$$\hat{\mu}\{Y \mid X = X_0\} \sim N\left(\beta_0 + \beta_1 X_0, \sigma^2\left[\frac{1}{n} + \frac{\left(X_0 - \overline{X}\right)^2}{(n-1)S_X^2}\right]\right).$$

Normality comes from the fact that both summands are (jointly) normally distributed.

Sampling Distribution of Mean Response at a Particular Value of X=X_0

$$\hat{\mu}\{\text{Price} \mid \text{Sqft} = 4000\} \sim N\left(1{,}398,\ \boxed{\sigma^2\left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{(n-1)S_X^2}\right]}\right)$$



Inference for Mean Response at a Particular Value of X=X$_0$: *t*-Test

$$H_0 : \mu\{Y \mid X = X_0\} = \mu_0$$

$$H_a : \mu\{Y \mid X = X_0\} \neq \mu_0$$

Corresponding t-statistic:

$$\frac{\hat{\beta}_0 + \hat{\beta}_1 X_0 - \mu_0}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\left(X_0 - \overline{X}\right)^2}{(n-1)S_X^2}}} \sim t_{n-2}$$

Inference for Mean Response at a Particular Value of $X=X_0$: CI

(1-α)100% CI for the estimate of the mean response at X_0:

$$\hat{\beta}_0 + \hat{\beta}_1 X_0 \pm t_{n-2,1-\alpha/2} SE(\hat{\beta}_0 + \hat{\beta}_1 X_0),$$

$$\text{where } SE(\hat{\beta}_0 + \hat{\beta}_1 X_0) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\left(X_0 - \overline{X}\right)^2}{(n-1)S_X^2}}$$

For Newton data, show that 95% CI for the mean price at Sqft = 4000 is (1304, 1492) and interpret it.

$$\hat{\beta}_0 + \hat{\beta}_1 X_0 = 1{,}398, n = 46,\ \bar{x} = 2{,}225,\ s_X = 1{,}252\ ,\ \hat{\sigma} = 181$$

Newton Data: Estimation Precision

$$\hat{\beta}_0 + \hat{\beta}_1 X_0 \pm t_{n-2,1-\alpha/2} SE(\hat{\beta}_0 + \hat{\beta}_1 X_0),$$

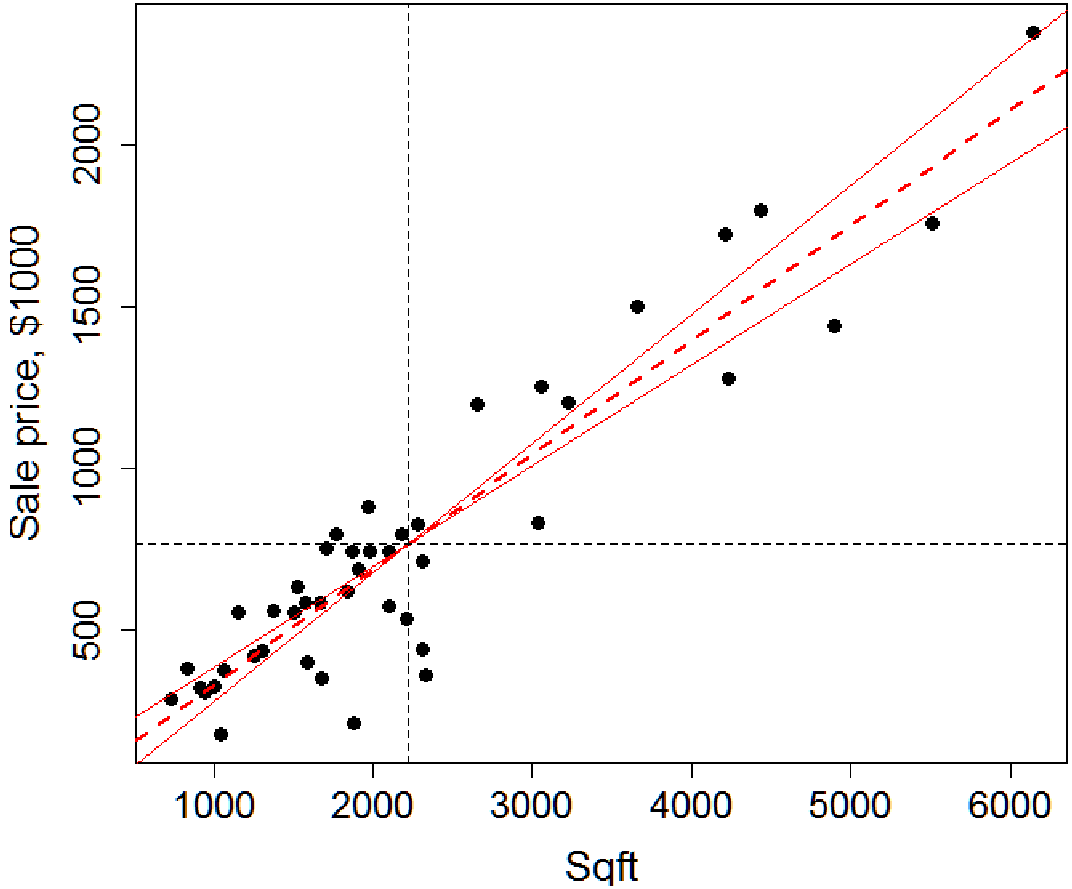
$$\text{where } SE(\hat{\beta}_0 + \hat{\beta}_1 X_0) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\left(X_0 - \overline{X}\right)^2}{(n-1)S_X^2}}$$

Why is the estimation error not the same along the regression line?

Newton Data: Range of Regression Lines

95% CI for the slope is (0.32, 0.4)

95% CI for the intercept it is (-136, 84)

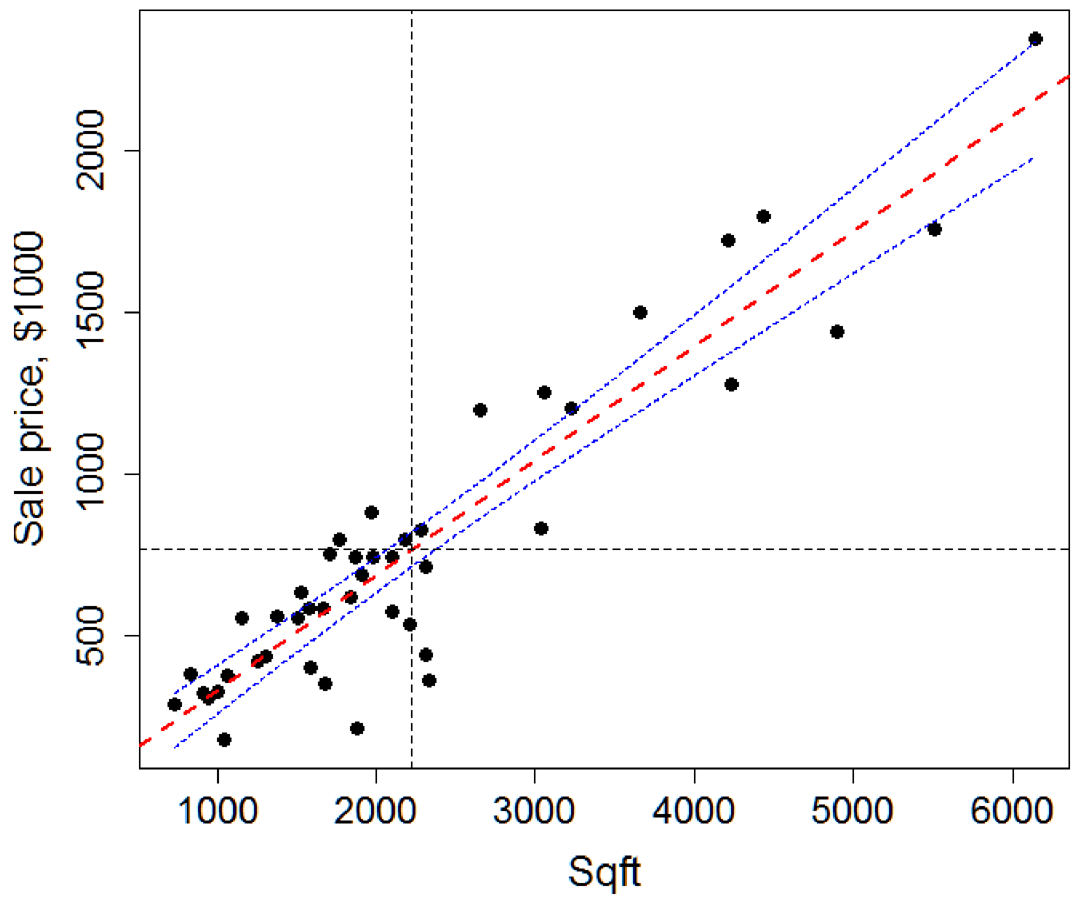


Inference for Mean Response: Centering "Trick" in R

```
> SaleData$ShiftedSqft. = SaleData$Sqft.-4000
> regmodel <- lm(Price/1000 ~ ShiftedSqft., data = SaleData)
> summary(regmodel)

Call:
lm(formula = Price/1000 ~ ShiftedSqft., 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)  1.398e+03  4.657e+01   30.03   <2e-16 ***
ShiftedSqft. 3.561e-01  2.152e-02   16.55   <2e-16 ***
```


Basis of the trick: $SE\left(\hat{\beta}_0\right) = SE\left(\hat{\mu}\{Y \mid X = 0\}\right)$

Newton Data:
95% CI for Mean Response Estimates




```
predict(RegModel, newdata=data.frame(VarX=newx),
interval = c("confidence"), type="response")
```

Mean Response for Many Values of X: Compound Uncertainty

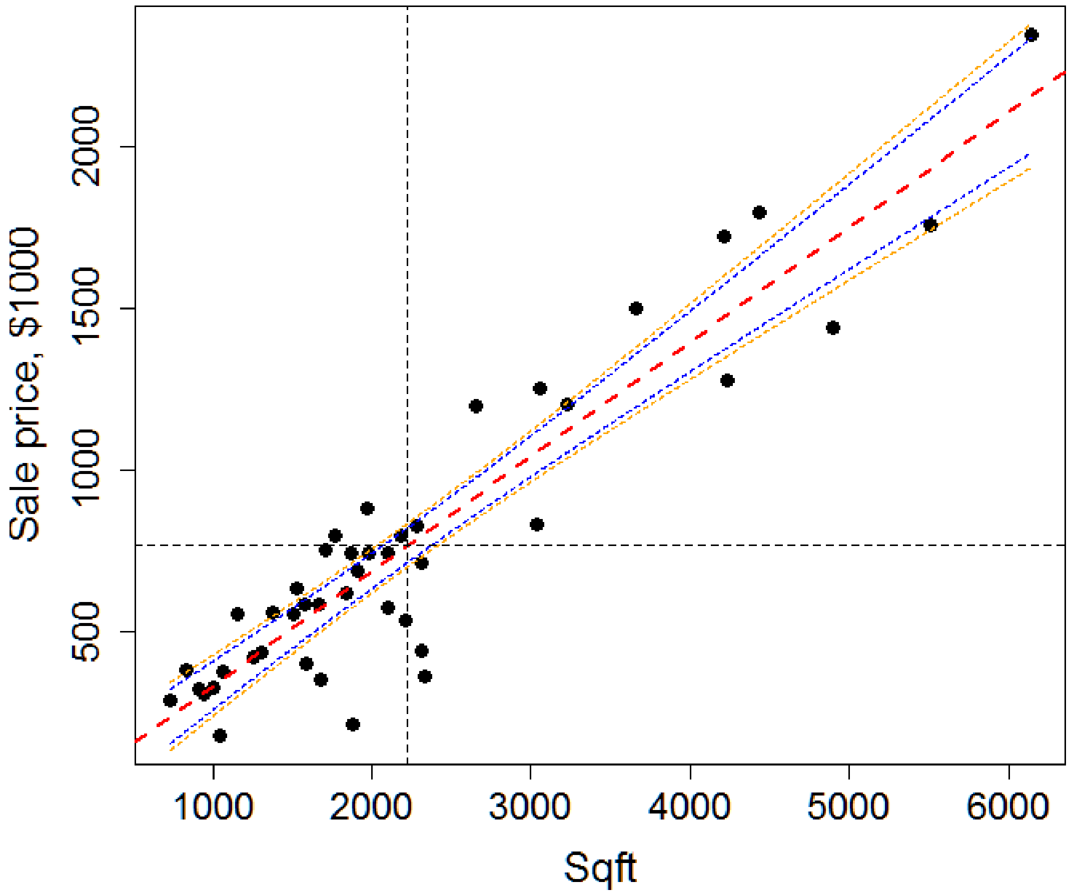
- If we are looking for an estimate in a couple of points, X_0 and X_1, then Bonferroni procedure can be used to adjust CIs.

- In order to build a confidence band (i.e., mean responses for all values of X), we can use Scheffe's method:

$$\hat{\beta}_0 + \hat{\beta}_1 X_0 \pm \boxed{\sqrt{2F_{2,n-2,0.95}}} SE(\hat{\beta}_0 + \hat{\beta}_1 X_0),$$

$$\text{where } SE(\hat{\beta}_0 + \hat{\beta}_1 X_0) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\left(X_0 - \overline{X}\right)^2}{(n-1)S_X^2}}$$

Newton Data: 95% confidence band for Mean Response Estimates (with Scheffe's adjustment)



R code will be shown later.

Prediction of a Future Response

Predicting Future Response at X=X_0

What will be the price of a *particular* house among all 4,000-sq.ft. houses?

Point-estimate is the same:

$$\text{Pred}\{Y \mid X = X_0\} = \hat{\mu}\{Y \mid X = X_0\} = \hat{\beta}_0 + \hat{\beta}_1 X_0$$

Variance is <u>larger!</u>

Variance of the Predicted Future Response at $X=X_0$

$$Y - \text{Pred}\{Y \mid X = X_0\} = Y - \hat{\mu}\{Y \mid X = X_0\}$$
$$= \left[Y - \mu\{Y \mid X = X_0\}\right] + \left[\mu\{Y \mid X = X_0\} - \hat{\mu}\{Y \mid X = X_0\}\right]$$

Prediction error = Random Sampling error + Estimation error

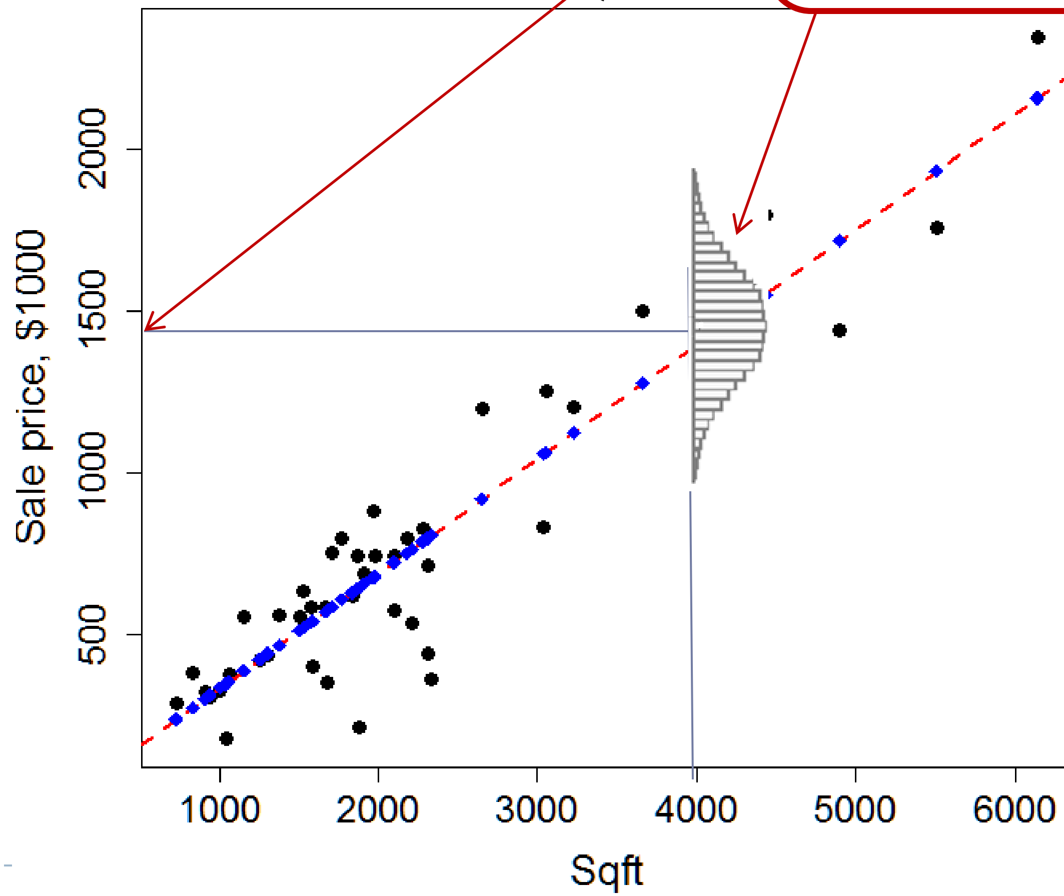
$$\text{Var}\left(\text{Pred}\{Y \mid X = X_0\}\right) = \sigma^2 + Var\left(\hat{\mu}\{Y \mid X = X_0\}\right)$$

Therefore,

$$\text{Pred}\{Y \mid X = X_0\} \sim N\left(\beta_0 + \beta_1 X_0, \sigma^2\left[1 + \frac{1}{n} + \frac{\left(X_0 - \overline{X}\right)^2}{(n-1)S_X^2}\right]\right)$$

Predicting Future Response at $X=X_0$

$$\text{Pred}\{\text{Price}\mid\text{Sqft}=4000\}\sim N\left(1398,\ \boxed{\sigma^2\left[1+\frac{1}{n}+\frac{(X_0-\bar{X})^2}{(n-1)S_X^2}\right]}\right)$$



Inference for the Predicted Future Response at $X=X_0$

(1-α)100% CI for the prediction of future response at X_0:

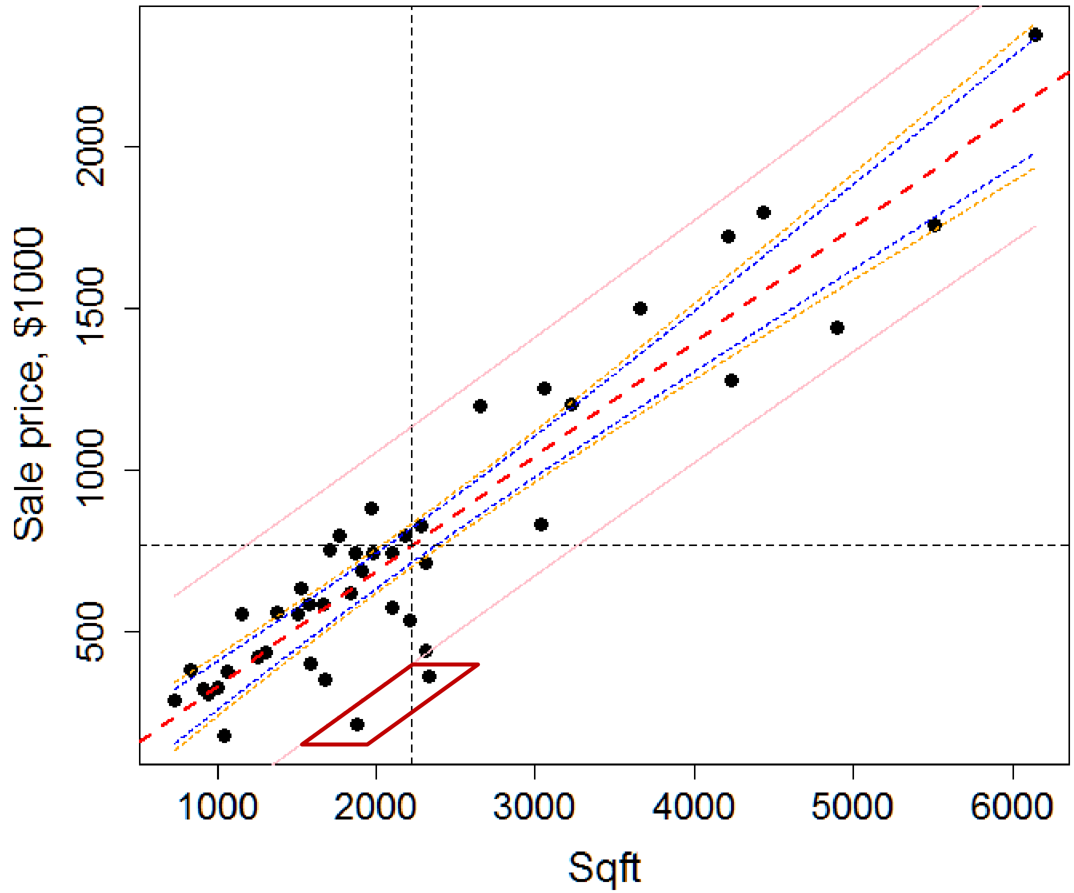
$$\hat{\beta}_0 + \hat{\beta}_1 X_0 \pm t_{n-2,1-\alpha/2} SE\left(\text{Pred}\{Y|X = X_0\}\right),$$

$$\text{where } SE\left(\text{Pred}\{Y|X = X_0\}\right) = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{\left(X_0 - \bar{X}\right)^2}{(n-1)S_X^2}}$$

For Newton data, show that 95% CI for the prediction at Sqft=4000 is (1021, 1775) and interpret it.

$$\hat{\beta}_0 + \hat{\beta}_1 X_0 = 1{,}398, n = 46, \ \bar{X} = 2{,}225, \ S_X^2 = 1{,}252^2, \ \hat{\sigma} = 181$$

Newton Data: 95% Prediction Band for Future Response Estimates




```
predict(RegModel, newdata=data.frame(VarX=newx),
interval = c("prediction"), type="response")
```

R code for the entire plot (Part I)

```
plot(SaleData$Sqft., SaleData$Price/1000, xlab="Sqft", ylab="Sale
  price, $1000", pch=19, cex.axis=1.5, cex.lab=1.6, cex=1.3)
regmodel <- lm(PriceScaled ~ Sqft.,
               data = SaleData)
abline(reg=regmodel, col="red", lwd=2, lty=2)
abline(h=mean(SaleData$Price/1000), v=mean(SaleData$Sqft.), lty=2)

# Confidence band
Newx <- seq(min(SaleData$Sqft.), max(SaleData$Sqft.))
Prd <- predict(regmodel,
               newdata=data.frame(Sqft.=newx),
               interval = c("confidence"),
               type="response")
lines(newx,prd[,2], col="blue", lty=2)
lines(newx,prd[,3], col="blue", lty=2)
```

R code for the entire plot (Part II)

```
# Confidence band with Scheffe adjustment
prdScheffe <- predict(regmodel, newdata=data.frame(Sqft.=newx),
                      interval = c("confidence"),
                      type="response")
prdScheffe[,2] <- prdScheffe[,1] - (prd[,2]-
               prd[,1])/qt(0.975,44)*sqrt(2*qf(0.95,2,44))
prdScheffe[,3] <- prdScheffe[,1] + (prd[,2]-
               prd[,1])/qt(0.975,44)*sqrt(2*qf(0.95,2,44))
lines(newx,prdScheffe[,2], col="orange", lty=2)
lines(newx,prdScheffe[,3], col="orange", lty=2)

# Prediction band
prdFuture <- predict(regmodel, newdata=data.frame(Sqft.=newx),
                      interval = c("prediction"),
                      type="response")
lines(newx,prdFuture[,2], col="pink", lty=1)
lines(newx,prdFuture[,3], col="pink", lty=1)
```

Calibration (or Inverse Prediction): Estimating X That Results in $Y=Y_0$

Suppose you have a certain budget ($1,000K) for a new home and you are trying to see how big of a house you can buy in Newton.

Ideally: Regress X on Y (if makes sense).

An <u>approximate</u> analytical method (that works on values closer to the middle) is:

$$\hat{X} = (Y_0 - \hat{\beta}_0)/\hat{\beta}_1$$

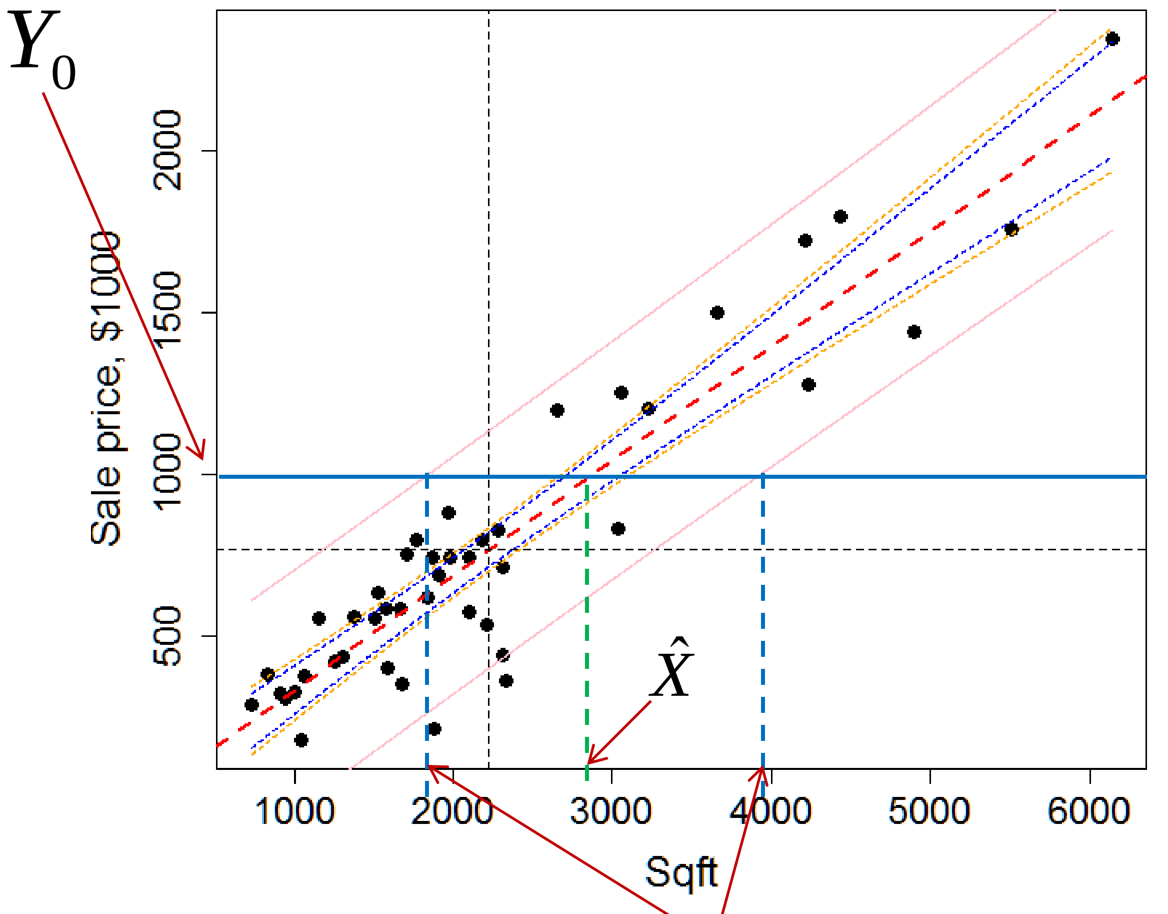
$$SE_\mu(\hat{X}) = \frac{SE\left(\hat{\mu}\{Y|X = \hat{X}\}\right)}{|\hat{\beta}_1|} \text{ or } SE_{\text{Pred}}(\hat{X}) = \frac{SE\left(\text{Pred}\{Y|X = \hat{X}\}\right)}{|\hat{\beta}_1|}$$

For CI use t-multiplier with d.f. = n-2.

Newton data: $\hat{X} = 2882, \quad SE_\mu(\hat{X}) = 131 \text{ or } SE_{\text{Pred}}(\hat{X}) = 525.3$

Calibration (or Inverse Prediction): Estimating X That Results in $Y=Y_0$

Graphical method:



Calibration intervals may by asymmetric!

Limits of the calibration interval