STAT 139: STATISTICAL SLEUTHING THROUGH LINEAR MODELS

Lecture 17
Oct 30, 2014

Victoria Liublinska

Odds and Ends

- Course Project Deadline #1: Upload the names of group members and a one-paragraph description of the project proposal to a drop-box by **Monday, November 3rd, 5pm**.
  - Should be done by each group member.

The Department of Statistics presents:

a spooky open house for concentrators interested in STATISTICS

Please join us for an information session, lunch & trick-or-treating for those in costume!

Friday, October 31st * 12:00-2:00
Science Center, 7th floor

Previous lecture: Review

- Regression line for standardized variables,

$$\tilde{Y}_i = \frac{Y_i - \overline{Y}}{S_Y}, \tilde{X}_i = \frac{X_i - \overline{X}}{S_X} \Rightarrow \hat{\mu}\left(\tilde{Y}_i \mid \tilde{X}_i\right) = r_{XY}\tilde{X}_i$$

- *Regression toward the mean* and *regression fallacy*.
- Inference for mean response at X=X_0,

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

$$\text{or } \hat{\beta}_0 + \hat{\beta}_1 X_0 \pm \sqrt{2F_{2,n-2,0.95}} SE(\hat{\beta}_0 + \hat{\beta}_1 X_0) \text{ for multiple points,}$$

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

Previous lecture: Review

- Inference for a future response at $X{=}X_0$,

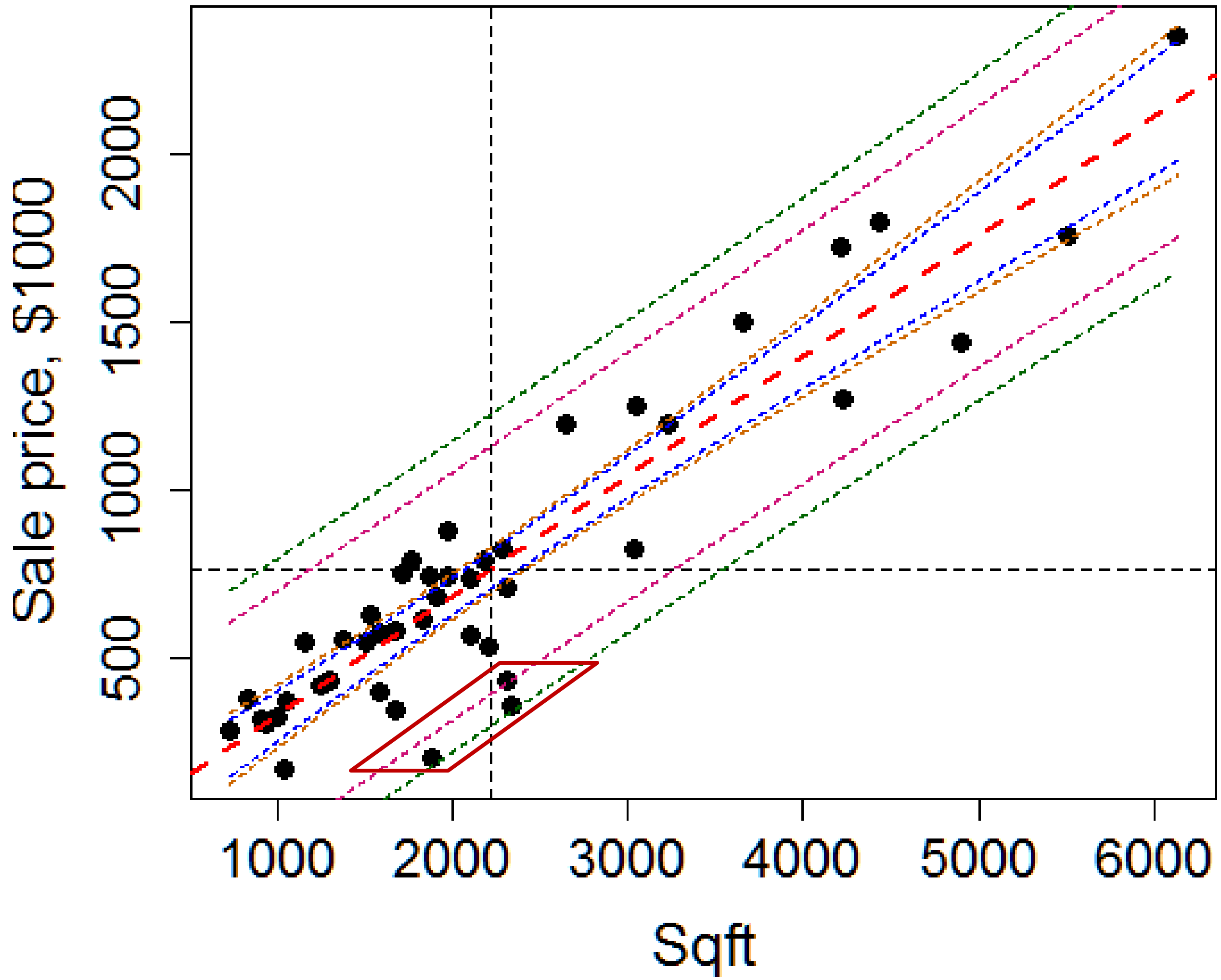
$$\hat{\beta}_0 + \hat{\beta}_1 X_0 \pm t_{n-2,1-\alpha/2} SE\left(\text{Pred}\{Y|X = X_0\}\right) \text{ for one point,}$$

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

- Analogously, for multiple points we can use Scheffe's method,

$$\hat{\beta}_0 + \hat{\beta}_1 X_0 \pm \boxed{\sqrt{2F_{2,n-2,0.95}}}\, SE\left(\text{Pred}\{Y|X = X_0\}\right).$$

Newton Data: Confidence and Prediction Bands



R code for the entire plot (Part III)

```
#Repeated: Prediction band
prdFuture <- predict(regmodel, newdata=data.frame(Sqft.=newx),
                      interval = c("prediction"),
                      type="response")
lines(newx,prdFuture[,2], col="deeppink3", lty=1)
lines(newx,prdFuture[,3], col="deeppink3", lty=1)

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

Alternative Interpretation of the estimator of β_1

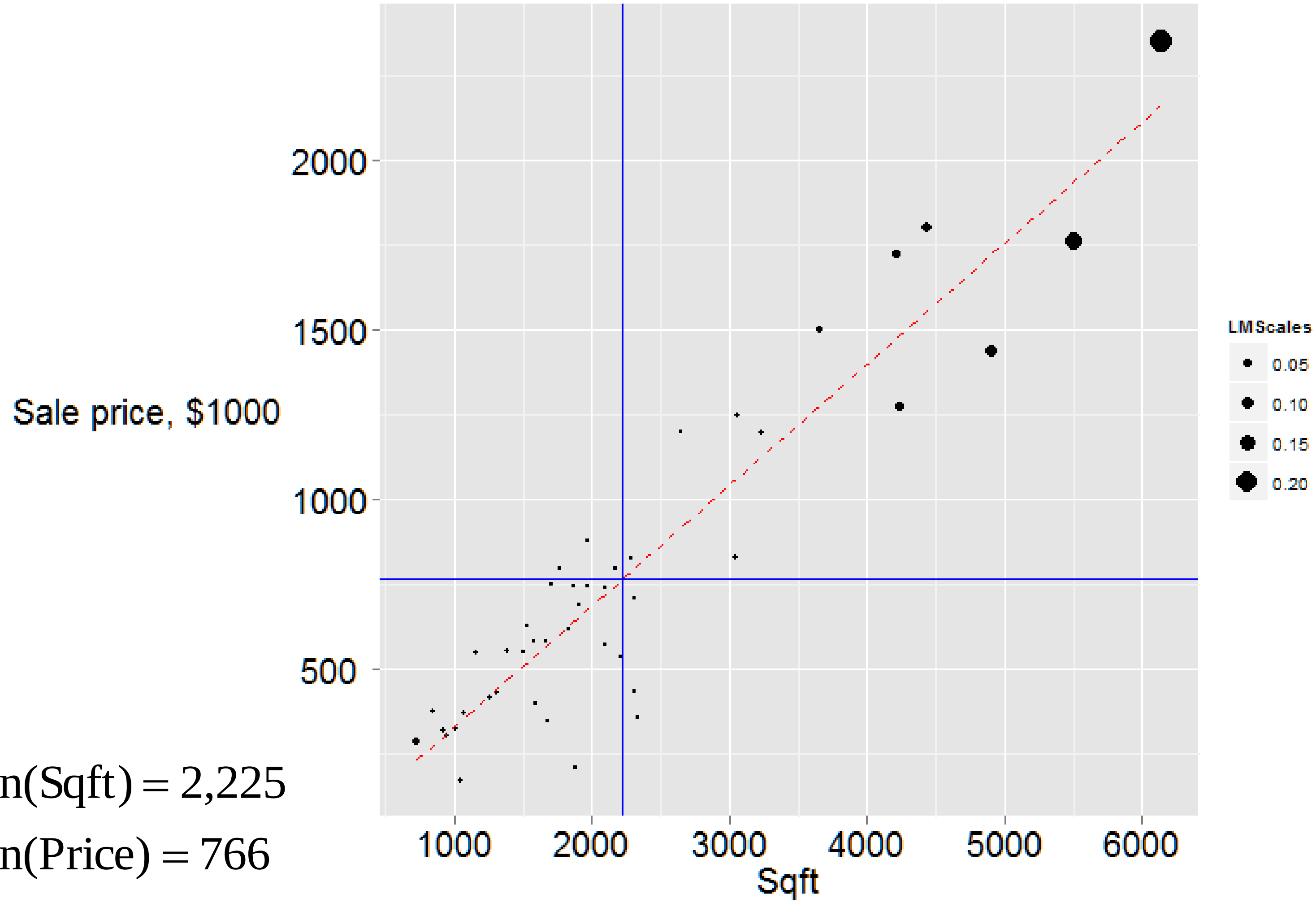
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}$$

$$= \sum_{i=1}^{n}\left[\frac{\left(X_i - \bar{X}\right)^2}{\left(\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right)}\frac{\left(Y_i - \bar{Y}\right)}{\left(X_i - \bar{X}\right)}\right]$$

$$= \sum_{i=1}^{n}\left[\omega_i \frac{\left(Y_i - \bar{Y}\right)}{\left(X_i - \bar{X}\right)}\right], \quad \text{where} \quad \omega_i = \frac{\left(X_i - \bar{X}\right)^2}{\left(\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right)} \text{ and } \sum_{i=1}^{n}\omega_i = 1$$

Weights reflect each observation's leverage. (Ch.11)

Distribution of Weights Among Observations for Newton Data



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

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

Today's overview

- Calibration (or inverse prediction).
- A closer look at assumptions for simple linear regression.
- Interpretation of results after log transformation.

Reading:

- **Required:** R&S Ch. 8, Ch. 8 R code

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 determine 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:

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

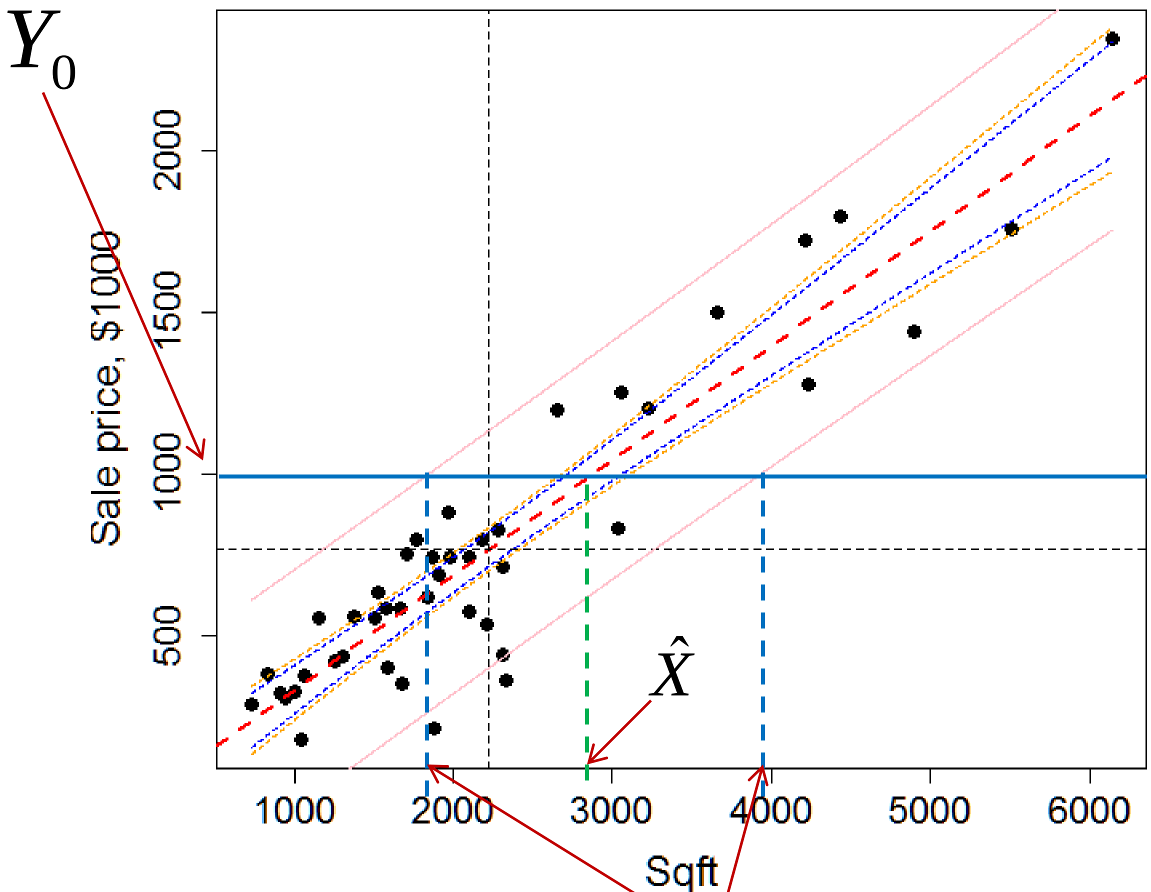
$$SE_{\mu}(\hat{X}) = \frac{SE\left(\hat{\mu}\{Y|X = \hat{X}\}\right)}{|\hat{\beta}_1|} \text{ or } SE_{\text{Pred}}(\hat{X}) = \frac{SE\left(\text{Pred}\{Y|X = \hat{X}\}\right)}{|\hat{\beta}_1|}$$

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

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

Confidence Interval for Mean Response vs. Prediction Interval for The Actual Response

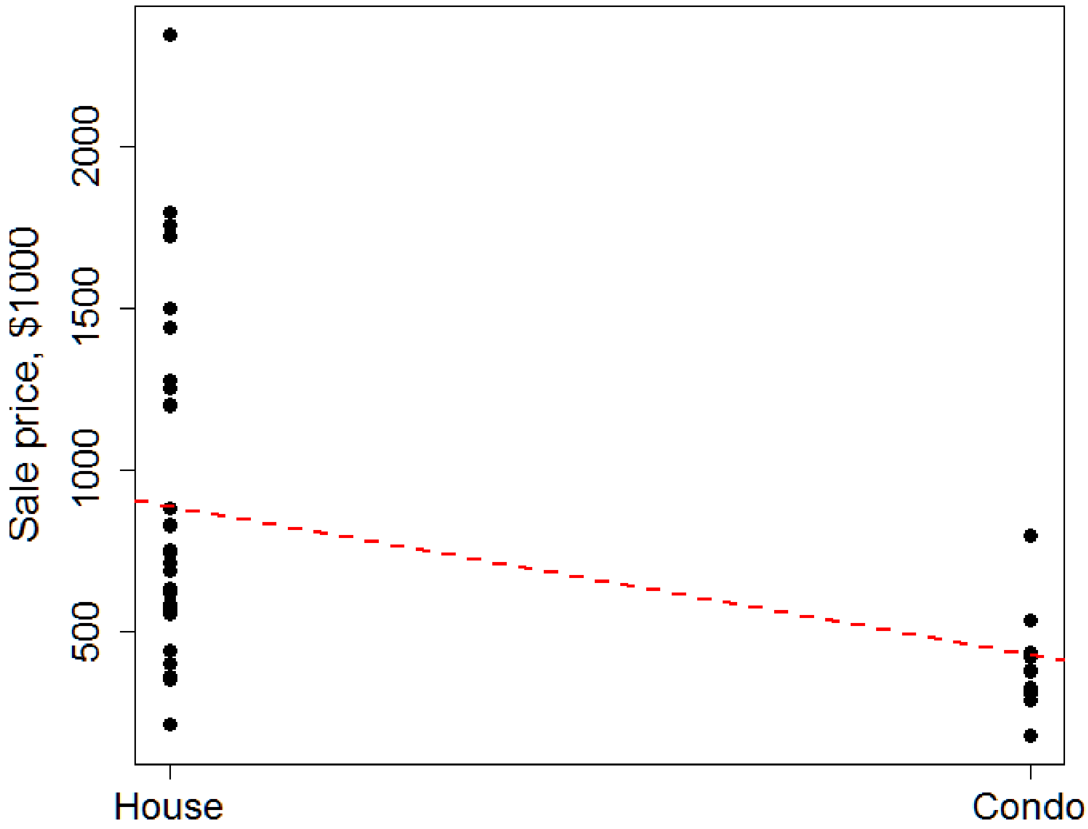
$$SE\left(\hat{\mu}\{\mathrm{Y} \mid \mathrm{X} = \mathrm{X}_0\}\right) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\left(X_0 - \overline{X}\right)^2}{(n-1)S_X^2}}$$

$$SE\left(\mathrm{Pred}\{\mathrm{Y} \mid \mathrm{X} = \mathrm{X}_0\}\right) = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{\left(X_0 - \overline{X}\right)^2}{(n-1)S_X^2}}$$

As the sample size goes to infinity, the width of the confidence interval for $\mu\{Y|X=X0\}$ goes to zero and the width of the prediction interval for $Pred\{Y|X=X0\}$ goes to $2z_{0.975}\sigma$.

Simple Linear Regression vs. Pooled Two-Sample *t*-Test

Simple Linear Regression with Binary X



Simple Linear Regression with Binary X

```
lm(formula = Price ~ Condo, data = SaleData)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   885.25      75.49  11.726 3.94e-15 ***
CondoTRUE    -456.50     147.81  -3.088  0.00348 **
```

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Simple Linear Regression with Binary X vs. a two-sample *t*-test

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   885.25      75.49  11.726 3.94e-15 ***
CondoTRUE    -456.50     147.81  -3.088  0.00348 **
----------------------------------------------------------------
> t.test(Price ~ Condo, data = SaleData, var.equal=TRUE)

  Two Sample t-test

data:  Price by Condo
t = -3.0884, df = 44, p-value = 0.003479
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -158.6094 -754.3905
sample estimates:
mean in group FALSE  mean in group TRUE
             885.25              428.75
```

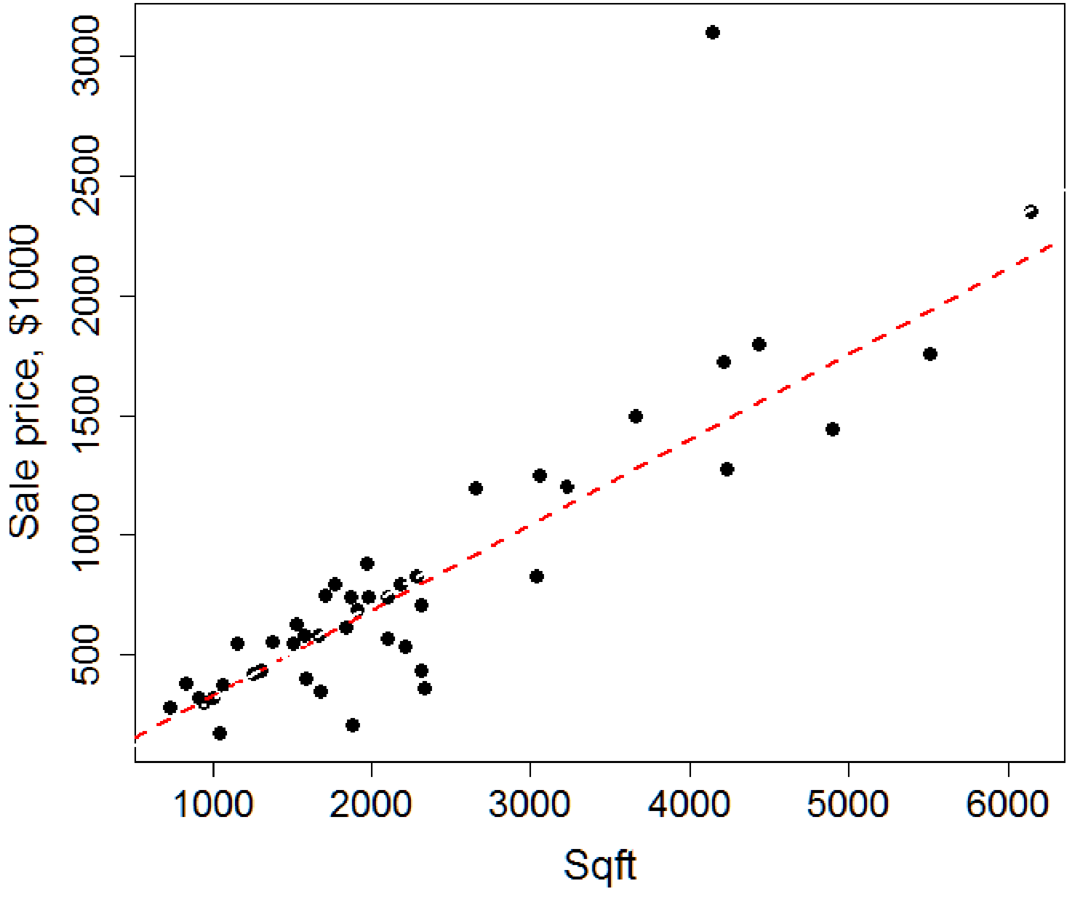
Simple Linear Regression: Assumptions and Diagnostics

Simple Linear Regression: Assumptions and Diagnostics

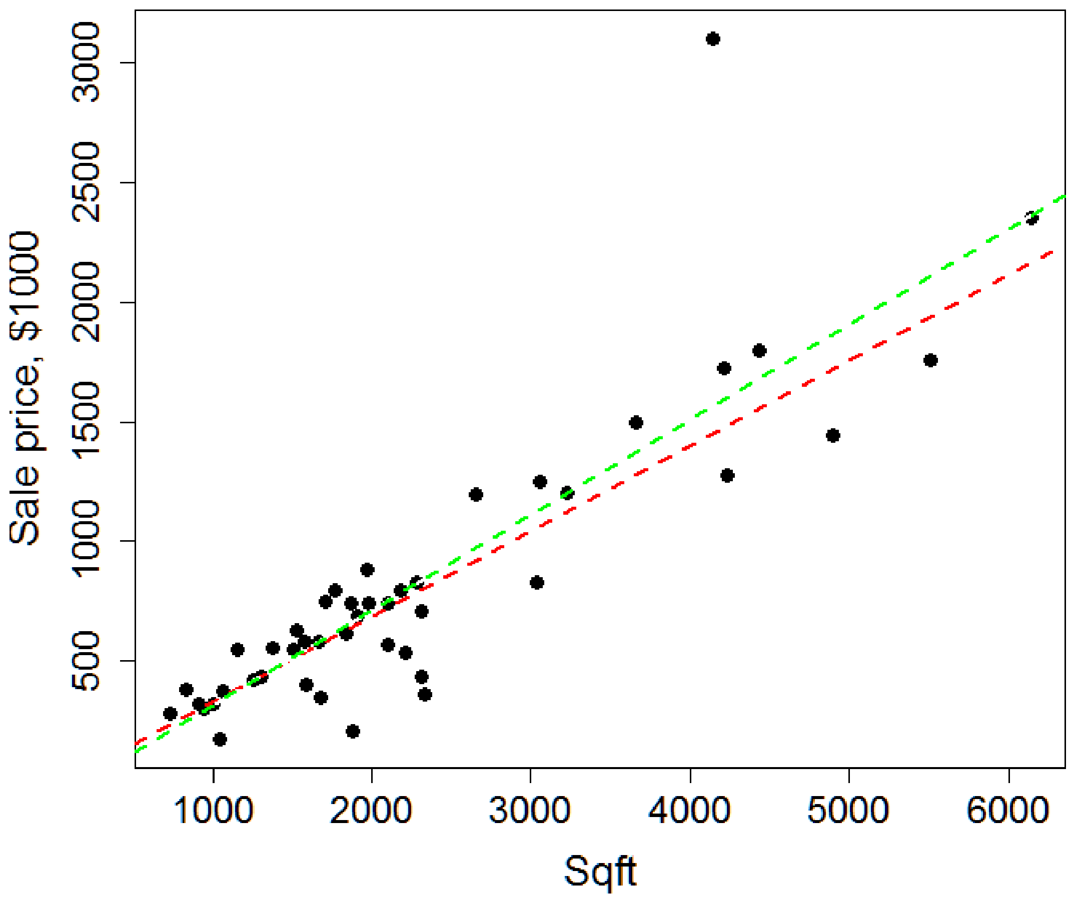
- Linearity, $E(Y \mid X) = \beta_0 + \beta_1 X$
  - Checking: graphically, conceptually (based on the phenomenon of interest and chosen predictors).
    - Look for nonlinearity and/or **outliers**.
  - If violated:
    - Estimators $\hat{\beta}_0, \hat{\beta}_1$, and predictions may be biased.

    Strategies:
    - Consider transformations ($\log(x)$, $1/x$, x^2, $\log(y)$, $1/y$, etc.) or add interactions (Ch 9).
    - Use nonlinear functions of X: **spline (or polynomial) regression** or other **generalized additive models**.

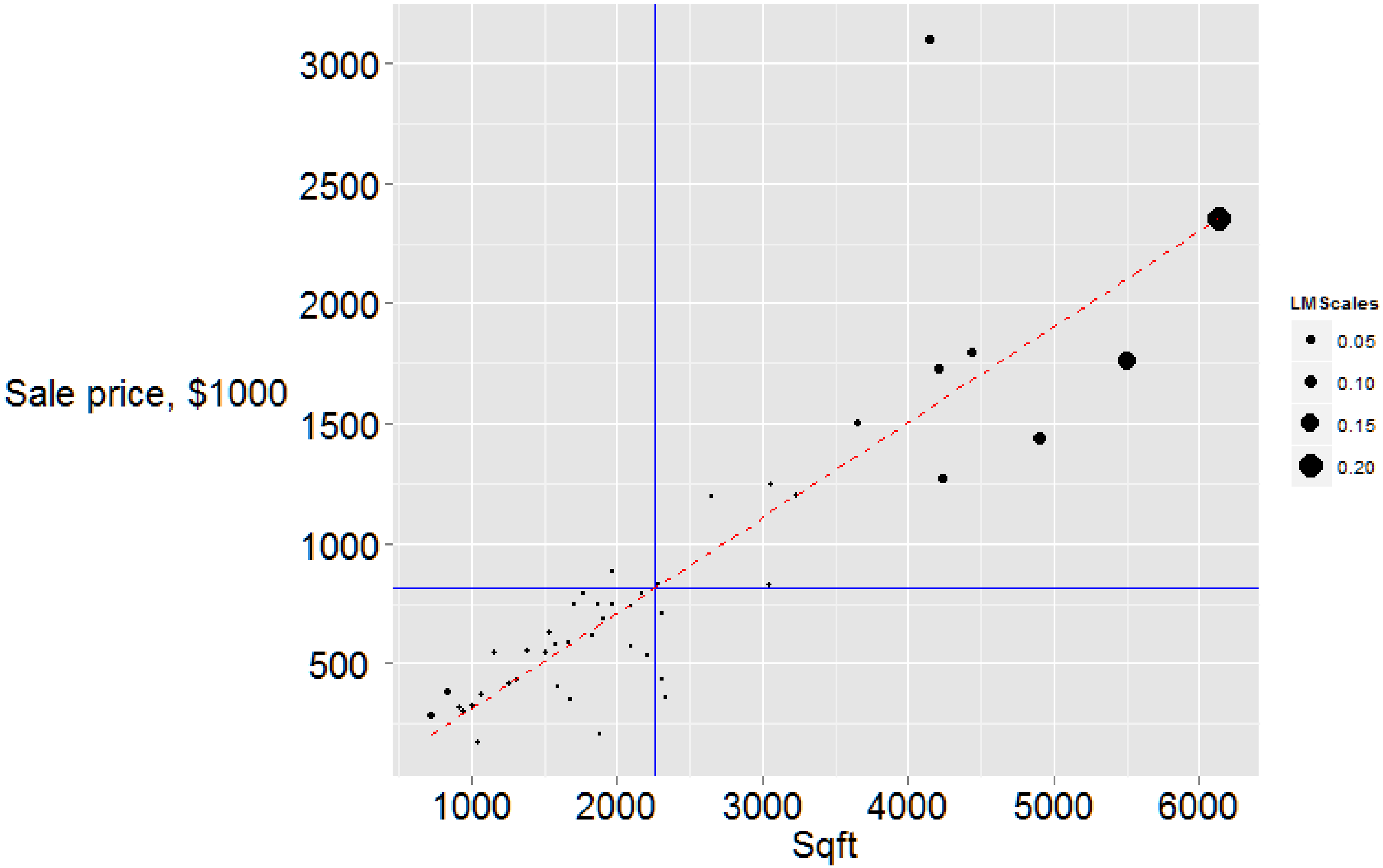
Newton Data with an Outlier



Newton Data with an Outlier



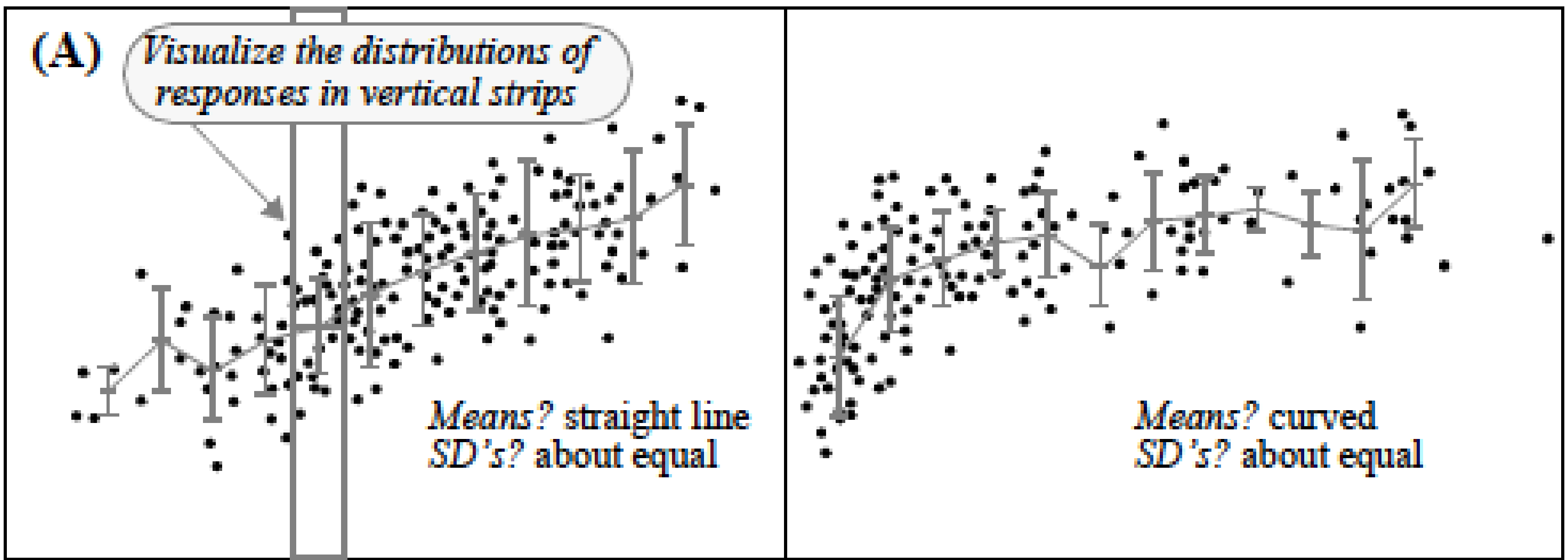
Newton Data: Outlier's leverage



Scatter plot of the Response vs. the Explanatory Variable

Display 8.6 **p. 213**

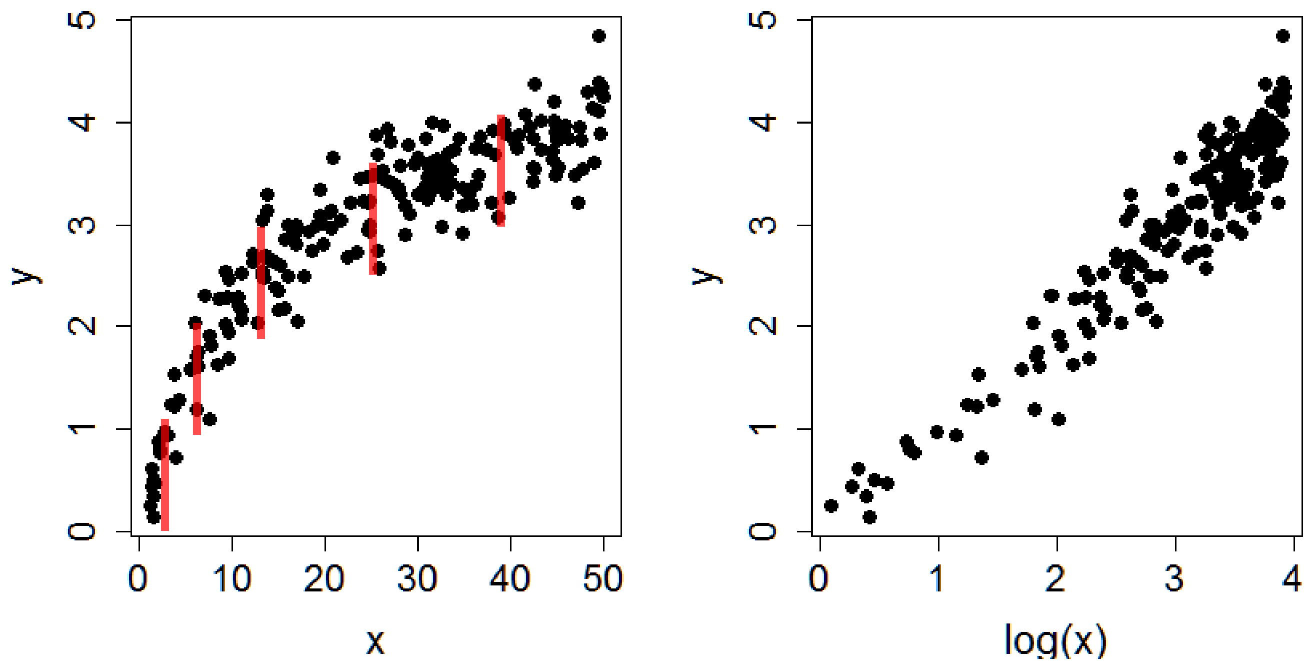
Some hypothetical scatterplots of response versus explanatory variable with suggested courses of action; (A) is ideal



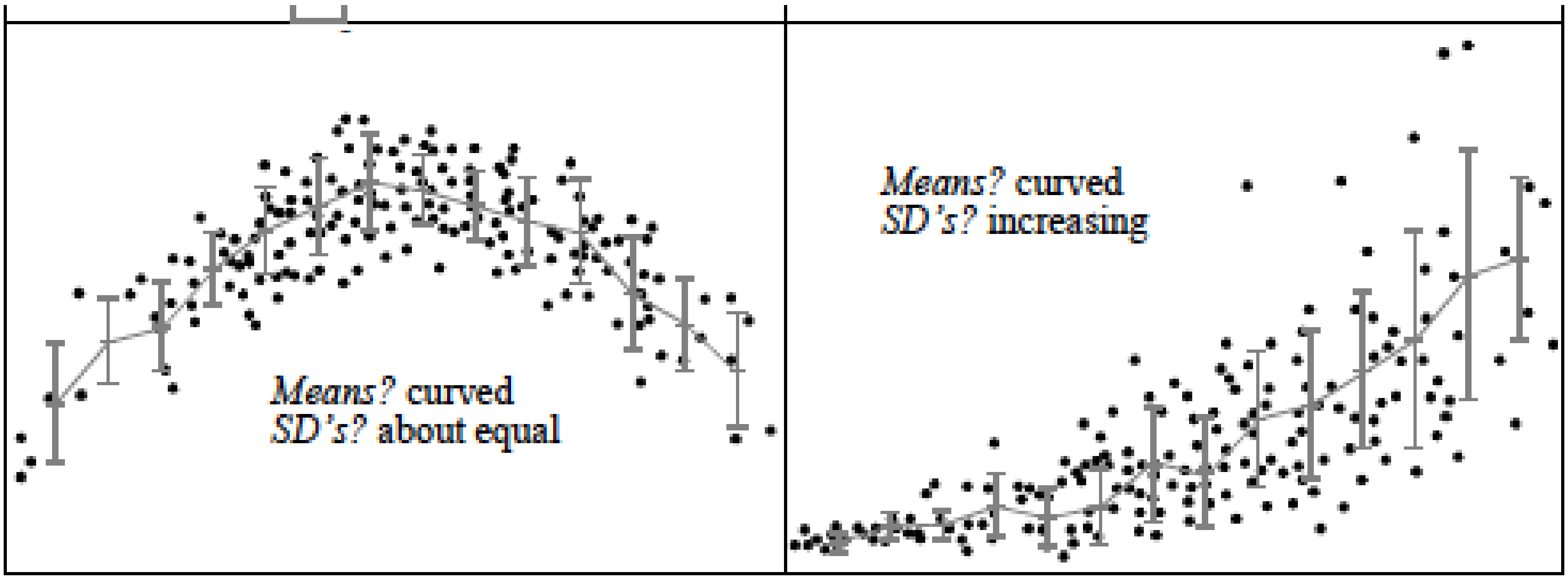
Linear regression can model non-linear relationships between *X* and *Y*, as long as there is a transformation of *X* or *Y* (or both) that makes it linear.

Transforming *X*

```
N=200
x = runif(N,1,50)
y = log(x) + rnorm(N,0,0.3)
lm(y ~ log(x))
```



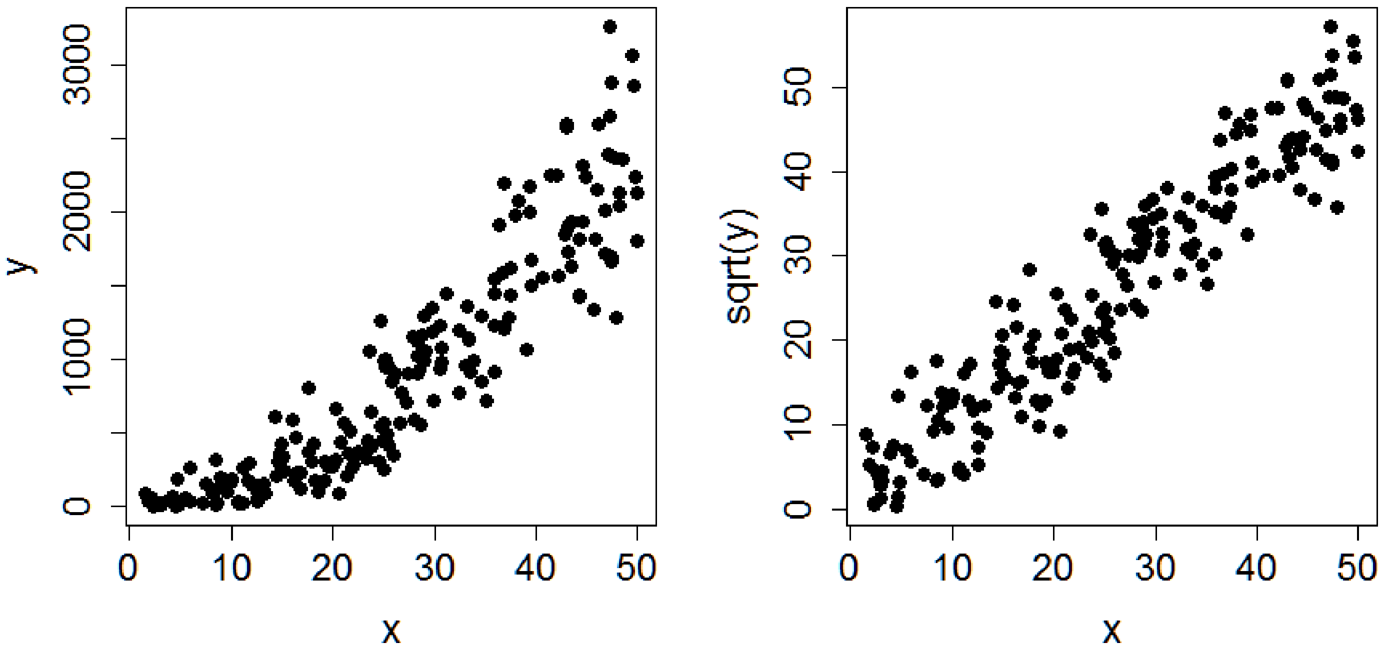
Scatter plot of the Response vs. the Explanatory Variable



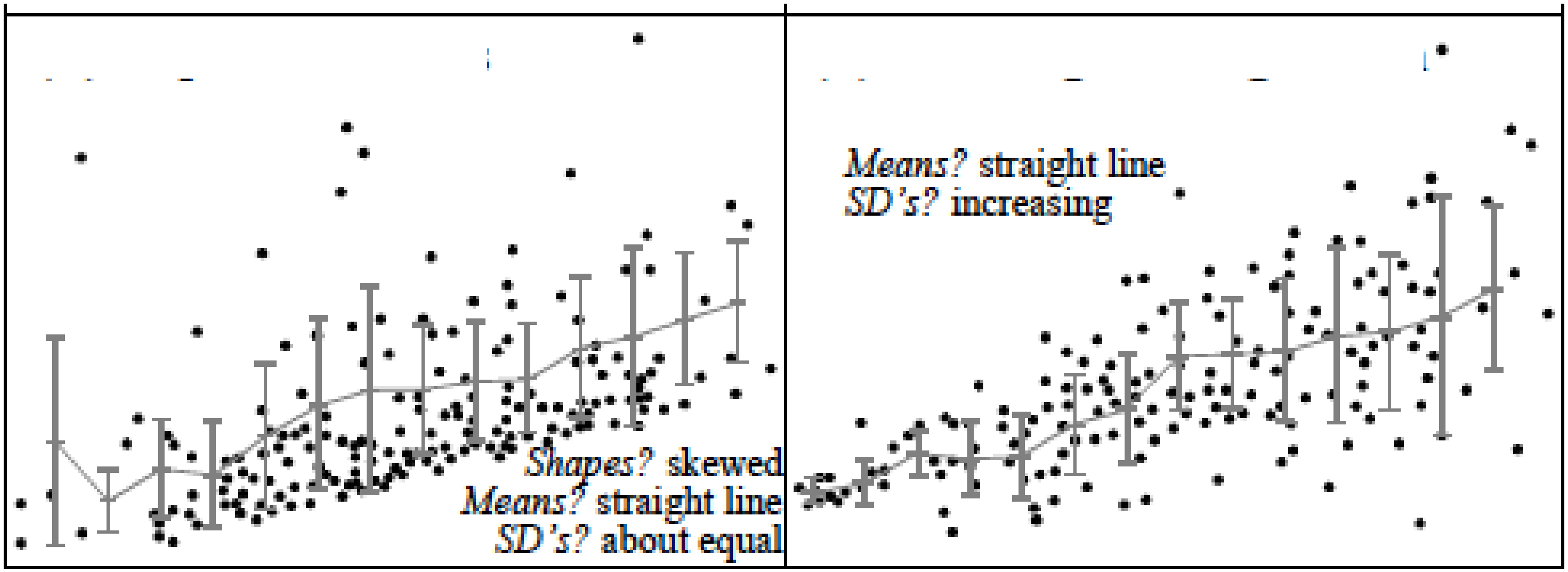
Covered in Ch. 9

Transforming *Y*

```
N=200
x = runif(N,1,50)
y = (x + rnorm(N,0,5))^2
lm(sqrt(y) ~ x)
```



Scatter plot of the Response vs. the Explanatory Variable



Covered in Section 11.6.1

Why is Regression "Linear"?

- .. if we can model *nonlinear* relationships between X and Y and fit models such as

$$\mu\{Y \mid X\} = \beta_0 + \beta_1 \log(X) \text{ or } \mu\{Y^2 \mid X\} = \beta_0 + \beta_1 X.$$

- It's called linear regression because $\mu(Y|X)$ is a linear function of regression coefficients. However, it may be an arbitrary function of the covariates.

Simple Linear Regression: Assumptions and Diagnostics

- Independence of errors ε_i. Residuals for any two observations Y_i and Y_j do not "travel together" after taking into account the corresponding X values.

  - Checking: Were all independent predictors included in the model of $\mu(Y|X)$? Examine the design.
    - Plot residuals vs. time/distance, when applicable.
  - If violated:
    - Doesn't lead to bias in $\hat{\beta}_0, \hat{\beta}_1$ but standard errors are affected (tests and CIs can be misleading).

Simple Linear Regression: Assumptions and Diagnostics

- Independence of errors ε_i. Residuals for any two observations Y_i and Y_j do not "travel together" after taking into account the corresponding X values.

  - Strategies:
    - Add more predictors (Ch. 9), group units in the same cluster.
    - For serial effects see Ch. 15 (models for time series).
    - For cluster effects or repeated observations, consider linear regression with correlated errors, including
      - Multilevel (or random-effect(s)) models (Gelman & Hill, 2007 – on reserve),
      - MANOVA or Repeated Measures ANOVA (Ch. 16).