STP 530
Lecture 3: Basic Diagnostics and Remedial Measures

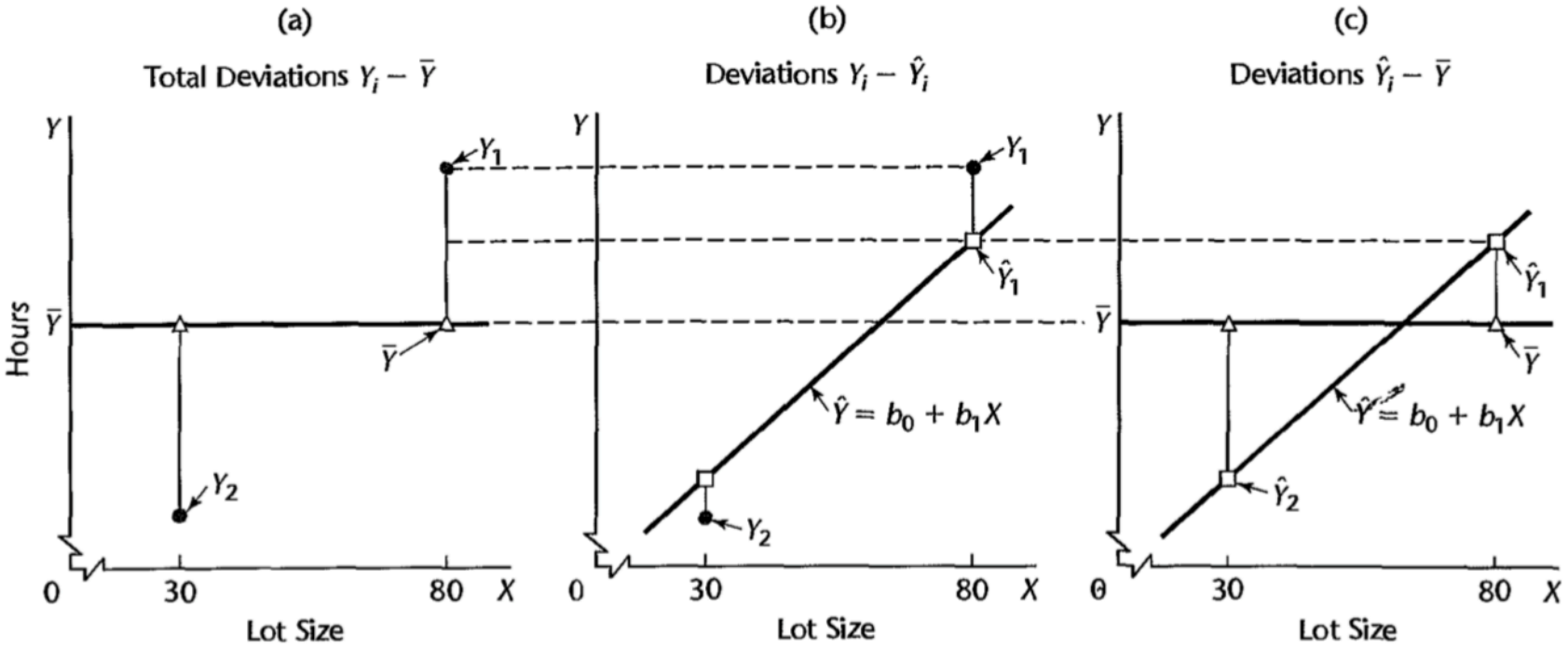
Yi Zheng, Ph.D.
yi.isabel.zheng@asu.edu

ARIZONA STATE UNIVERSITY
SCHOOL OF MATHEMATICAL AND STATISTICAL SCIENCES

September 12, 2023

Outline

Partitioning of Deviations



$$\underbrace{Y_i - \bar{Y}}_{\text{Total deviation}} = \underbrace{Y_i - \hat{Y}_i}_{\substack{\text{Deviation around} \\ \text{fitted regression line}}} + \underbrace{\hat{Y}_i - \bar{Y}}_{\substack{\text{Deviation of fitted} \\ \text{regression value} \\ \text{around the mean}}}$$

Partitioning of Sums of Squares

Total Sum of Squares: $SSTO = \sum(Y_i - \bar{Y})^2$

Error Sum of Squares: $SSE = \sum(Y_i - \hat{Y}_i)^2$

Regression Sum of Squares: $SSR = \sum(\hat{Y}_i - \bar{Y})^2$

When the regression model is estimated by the least square method,

$$SSTO = SSE + SSR$$

The corresponding partitioning of degrees of freedom is:

$$n - 1 = (n - p) + (p - 1)$$

where n is the sample size and p is the total number of regression parameters being estimated.

Proof of $SSTO = SSE + SSR$

$$\begin{aligned}
\sum(Y_i - \bar{Y})^2 &= \sum(Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2 \\
&= \sum[(Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})]^2 \\
&= \sum[(Y_i - \hat{Y}_i)^2 + (\hat{Y}_i - \bar{Y})^2 + 2(\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i)] \\
&= \sum(Y_i - \hat{Y}_i)^2 + \sum(\hat{Y}_i - \bar{Y})^2 + 2\sum(\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i)
\end{aligned}$$

The last term equals zero per least square estimation properties #4 and #7 given in Lecture 1: simple linear regression (recalled next slide).

$$\begin{aligned}
\sum(\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i) &= \sum(\hat{Y}_i - \bar{Y})e_i \\
&= \sum \hat{Y}_i e_i - \sum \bar{Y} e_i \\
&= 0(\text{per property \#7}) - \bar{Y}\sum e_i \\
&= 0 - 0(\text{per property \#4}) \\
&= 0
\end{aligned}$$

Recall: Least square estimation properties

From the partial derivative equations

$$\frac{\partial Q}{\partial \beta_0} = -2\sum(Y_i - \beta_0 - \beta_1 X_i) = 0; \quad \frac{\partial Q}{\partial \beta_1} = -2\sum X_i(Y_i - \beta_0 - \beta_1 X_i) = 0$$

We know that the solutions b_0 and b_1 satisfy the following **normal equations**:

$$\sum(Y_i - b_0 - b_1 X_i) = 0; \quad \sum X_i(Y_i - b_0 - b_1 X_i) = 0$$

So the least square solution has the following properties:

4. $\sum e_i = 0$, because $\sum e_i = \sum(Y_i - b_0 - b_1 X_i) = 0$
5. $\sum Y_i = \sum \hat{Y}_i$, because $\sum Y_i = \sum(\hat{Y}_i + e_i) = \sum \hat{Y}_i + \sum e_i = \sum \hat{Y}_i$
6. $\sum X_i e_i = 0$ (a.k.a. $\boldsymbol{X} \perp \boldsymbol{e}$), because $\sum X_i e_i = \sum X_i(Y_i - b_0 - b_1 X_i) = 0$
7. $\sum \hat{Y}_i e_i = 0$ (a.k.a. $\hat{\boldsymbol{Y}} \perp \boldsymbol{e}$), because
   $\sum \hat{Y}_i e_i = \sum(b_0 + b_1 X_i)e_i = b_0 \sum e_i + b_1 \sum X_i e_i = 0$

Mean Squares

A MS is a Sum of Squares (SS) divided by its associated degrees of freedom.

Total Mean Square:

$$\text{“}MSTO\text{”} = \frac{SSTO}{n-1} = s_Y^2$$

Error Mean Square:

$$MSE = \frac{SSE}{n-p}$$

Regression Mean Square:

$$MSR = \frac{SSR}{p-1}$$

Note: Mean Squares are not additive:

$$\text{“}MSTO\text{”} \neq MSE + MSR$$

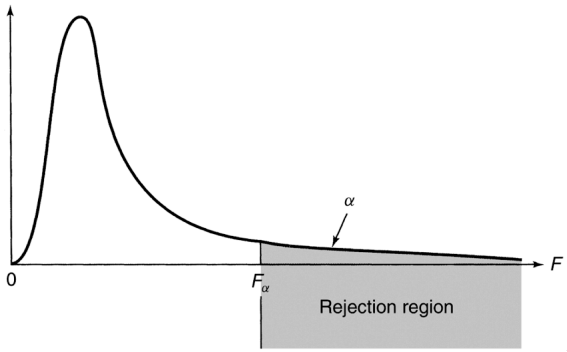
The ANOVA table for simple linear regression

Source	SS (Sum of Squares)	d.f.	MS (Mean Squares)
Regression	$SSR = \sum(\hat{Y}_i - \bar{Y})^2$	$p-1$	$MSR = SSR/(p-1)$
Error	$SSE = \sum(Y_i - \hat{Y}_i)^2$	$n-p$	$MSE = SSE/(n-p)$
Total	$SSTO = \sum(Y_i - \bar{Y})^2$	$n-1$	$MSTO = SSTO/(n-1)$

The F Test of Global Model Utility

Five steps:

1. **Assumptions:** ε i.i.d. $\sim \mathrm{N}(0, \sigma^2)$
2. **Hypotheses:** $H_0 : \beta_1 = 0$ $H_1 : \beta_1 \neq 0$
3. **Test-statistic:**
$$F_{obs} = \frac{MSR}{MSE}$$
4. **P-value:**
The right-tail probability in the reference distribution: $F(p-1, n-p)$
(See plot of F-distribution with various df's here:
https://www.medcalc.org/manual/f-distribution_functions.php)



5. **Conclusion:** Reject H_0 if p-value is less than a pre-determined significance level (typically 0.05) and conclude that the true population slope is not 0, in other words, there is linear association between X and Y.

The F test and the t test in a simple linear regression

In a simple linear regression, for hypotheses
$H_0 : \beta_1 = 0 \quad vs. \quad H_1 : \beta_1 \neq 0,$
the F-test of global model utility is mathematically equivalent with the two-sided t test. The test-statistic $F_{obs} = t^2_{obs}$, the p-values are the same.

```
Call:
lm(formula = sales_y ~ adv_x)

Residuals:
         1          2          3          4          5
 4.000e-01 -3.000e-01 -3.886e-16 -7.000e-01  6.000e-01

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.1000     0.6351  -0.157   0.8849
adv_x         0.7000     0.1915   3.656   0.0354 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6055 on 3 degrees of freedom
Multiple R-squared:  0.8167, Adjusted R-squared:  0.7556
F-statistic: 13.36 on 1 and 3 DF,  p-value: 0.03535
```

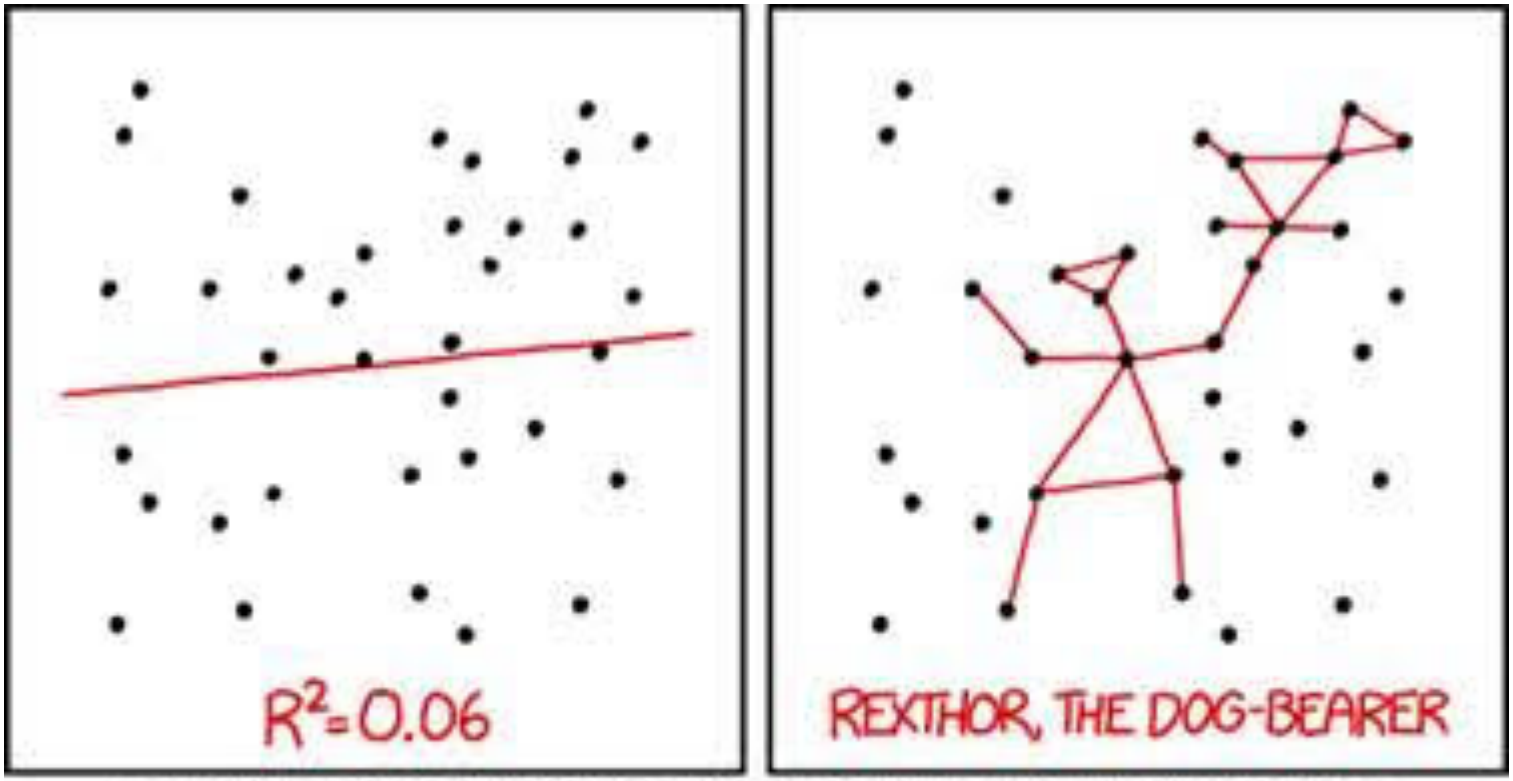
Outline

The coefficient of determination R^2

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2}$$

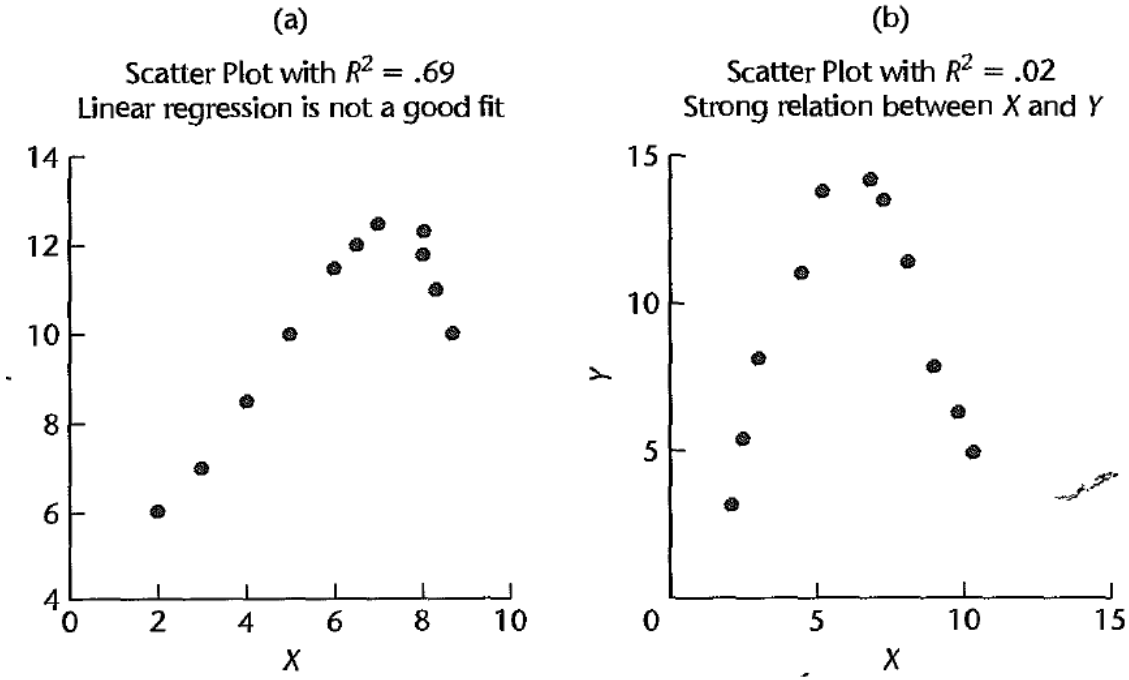
- **Interpretation:** About $R^2 100\%$ of variation in Y can be explained by the regression model.
- $0 \leqslant R^2 \leqslant 1$.
- $R^2 = 1$ when X accounts for all variation in Y.
- $R^2 = 0$ when there is no linear association between X and Y in the sample data.

The coefficient of determination R^2



Cautions using R^2

- A high R^2 does not necessarily indicate that the estimated regression line is a good fit. (Figure a)
- R^2 near zero does not indicate that X and Y are not related; it indicates, however, that X and Y are not linearly associated. (Figure b)



Outline

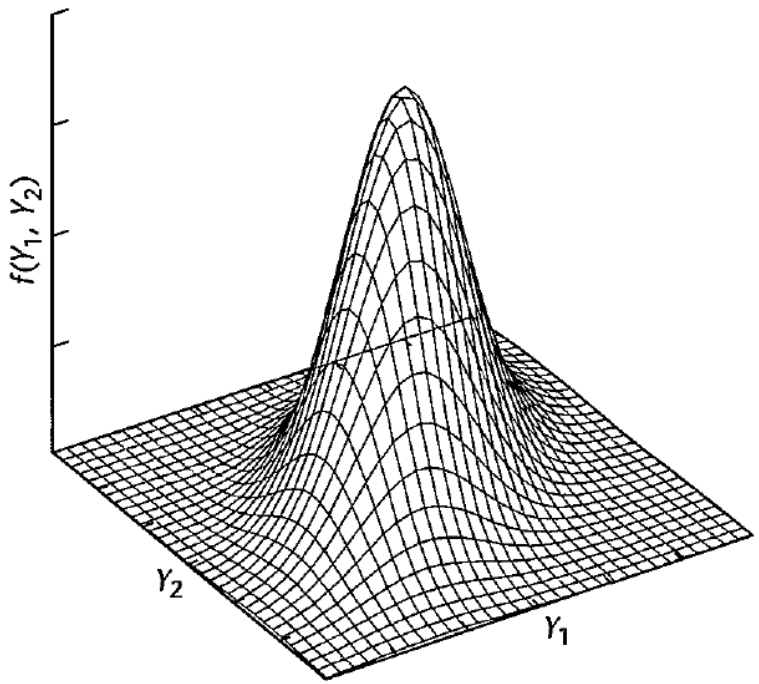
The Bivariate Normal distribution

$$f(Y_1, Y_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho_{12}^2}} \exp\left\{ -\frac{1}{2(1-\rho_{12}^2)}\left[\left(\frac{Y_1-\mu_1}{\sigma_1}\right)^2 - 2\rho_{12}\left(\frac{Y_1-\mu_1}{\sigma_1}\right)\left(\frac{Y_2-\mu_2}{\sigma_2}\right) + \left(\frac{Y_2-\mu_2}{\sigma_2}\right)^2\right]\right\}$$

- The distribution has five parameters: $\mu_1, \sigma_1, \mu_2, \sigma_2, \rho_{12}$.
- ρ_{12} is the coefficient of correlation between Y_1 and Y_2.
- Point estimator of ρ_{12} is Pearson product-moment correlation coefficient

$$r_{12} = \frac{\sum(Y_{i1}-\bar{Y}_1)(Y_{i2}-\bar{Y}_2)}{\sqrt{\sum(Y_{i1}-\bar{Y}_1)^2\sum(Y_{i2}-\bar{Y}_2)^2}}$$

.



Confidence interval of ρ_{12}

Because the sampling distribution of r_{12} is complicated when $\rho_{12} \neq 0$, confidence interval of ρ_{12} is usually carried out by approximation based on the *Fisher z transformation*:

$$z' = \frac{1}{2} \ln \left(\frac{1 + r_{12}}{1 - r_{12}} \right)$$

When n is large (25 or more as rule of thumb), the distribution of z' is approximately normal with approximate mean and variance:

$$E\{z'\} = \zeta = \frac{1}{2} \ln \left(\frac{1 + \rho_{12}}{1 - \rho_{12}} \right)$$

$$\sigma^2\{z'\} = \frac{1}{n - 3}$$

Confidence interval of ρ_{12} and ρ_{12}^2

Assume that Y_1 and Y_2 are distributed as bivariate normal.

When the sample size is large ($n \geq 25$), the approximate $1-\alpha$ confidence limits for ζ are:

$$z' \pm z(1-\alpha/2) \cdot \sigma\{z'\}$$

where $z(1-\alpha/2)$ is the $(1-\alpha/2)100$ percentile of the standard normal distribution.

Then the approximate $1-\alpha$ confidence limits for ρ_{12} are obtained by transforming the limits on ζ back to ρ based on the equation given on the next slide.

Then we can obtain the approximate $1-\alpha$ confidence limits for ρ_{12}^2 by squaring the two confidence limits, provided that the two confidence limits of ρ_{12} do not differ in signs. Otherwise, set the lower limit of the confidence interval of ρ_{12}^2 to 0.

Deriving the transformation from z' back to r:

$$z' = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$$

$$2z' = \ln\left(\frac{1+r}{1-r}\right)$$

$$e^{2z'} = \frac{1+r}{1-r}$$

$$(1-r)e^{2z'} = 1+r$$

$$e^{2z'} - e^{2z'}r = 1+r$$

$$e^{2z'} - 1 = (e^{2z'}+1)r$$

$$r = \frac{e^{2z'}-1}{e^{2z'}+1}$$

Relation between R^2 and Pearson correlation coefficient r

For simple linear regression,

$$R^2 = r_{XY}^2$$

For multiple regression,

$$R^2 = r_{Y\hat{Y}}^2$$

Relation between R^2 and r in simple linear regression

$$
\begin{aligned}
R^2 &= \frac{SSR}{SSTO} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} && \text{(by definition of } R^2\text{)} \\
&= \frac{\sum(\hat{Y}_i - \bar{Y})^2/(n-1)}{\sum(Y_i - \bar{Y})^2/(n-1)} = \frac{\text{Var}\{\hat{Y}\}}{\text{Var}\{Y\}} && \text{(by definition of sample variance)} \\
&= \frac{\text{Var}\{b_0 + b_1 X)}{\text{Var}\{Y\}} && \text{(by linear regression model)} \\
&= \frac{b_1^2 \text{Var}\{X\}}{\text{Var}\{Y\}} && \text{(by property of variance)} \\
&= \frac{\left(\frac{\text{Cov}\{X,Y\}}{\text{Var}\{X\}}\right)^2 \text{Var}\{X\}}{\text{Var}\{Y\}} && \text{(by least square formula of } b_1\text{)} \\
&= \frac{(\text{Cov}\{X,Y\})^2}{\text{Var}\{X\}\text{Var}\{Y\}} = r^2 && \text{(by definition of Pearson's } r\text{)}
\end{aligned}
$$

Outline

Univariate plots and scatter plot matrix

Always plot your data before you run any numerical analysis!

For each variable (including Xs and Y): examine histogram, min, max, missing data patterns. Check for odd observations.

For the entire dataset, examine the scatterplot matrix.
R code: `pairs(my.data)`

Residuals

Each observation i has a residual e_i:

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

"Semistudentized" residuals:

$$e_i^* = \frac{e_i - \bar{e}}{s} = \frac{e_i}{\sqrt{MSE}} = \frac{e_i}{\sqrt{\frac{SSE}{n-p}}}$$

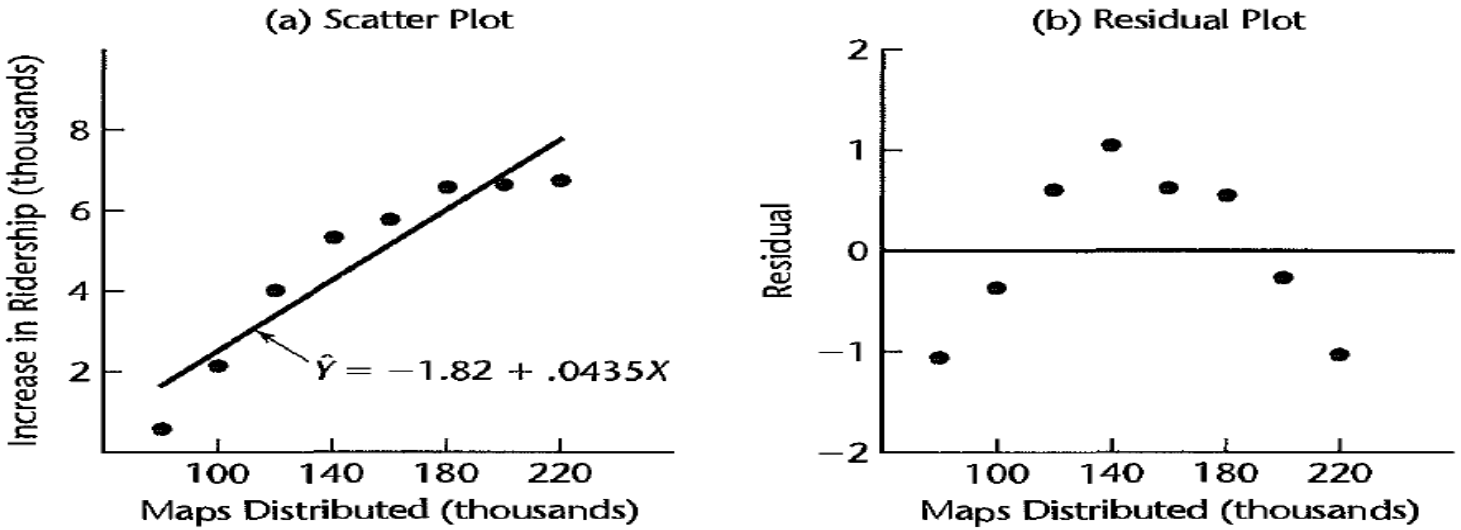
Note: s is not the precise standard deviation of the residuals, so it's called "semistudentized". The more precise way, the studentized residual will be introduced later.

Residual analysis

1. The regression function is linear.
2. No excessive influences by outliers.
3. The error terms have constant variance (homoskedasticity).
4. The error terms are normally distributed.
5. The error terms are independent among each other.

(1) Checking for lack of linear fit

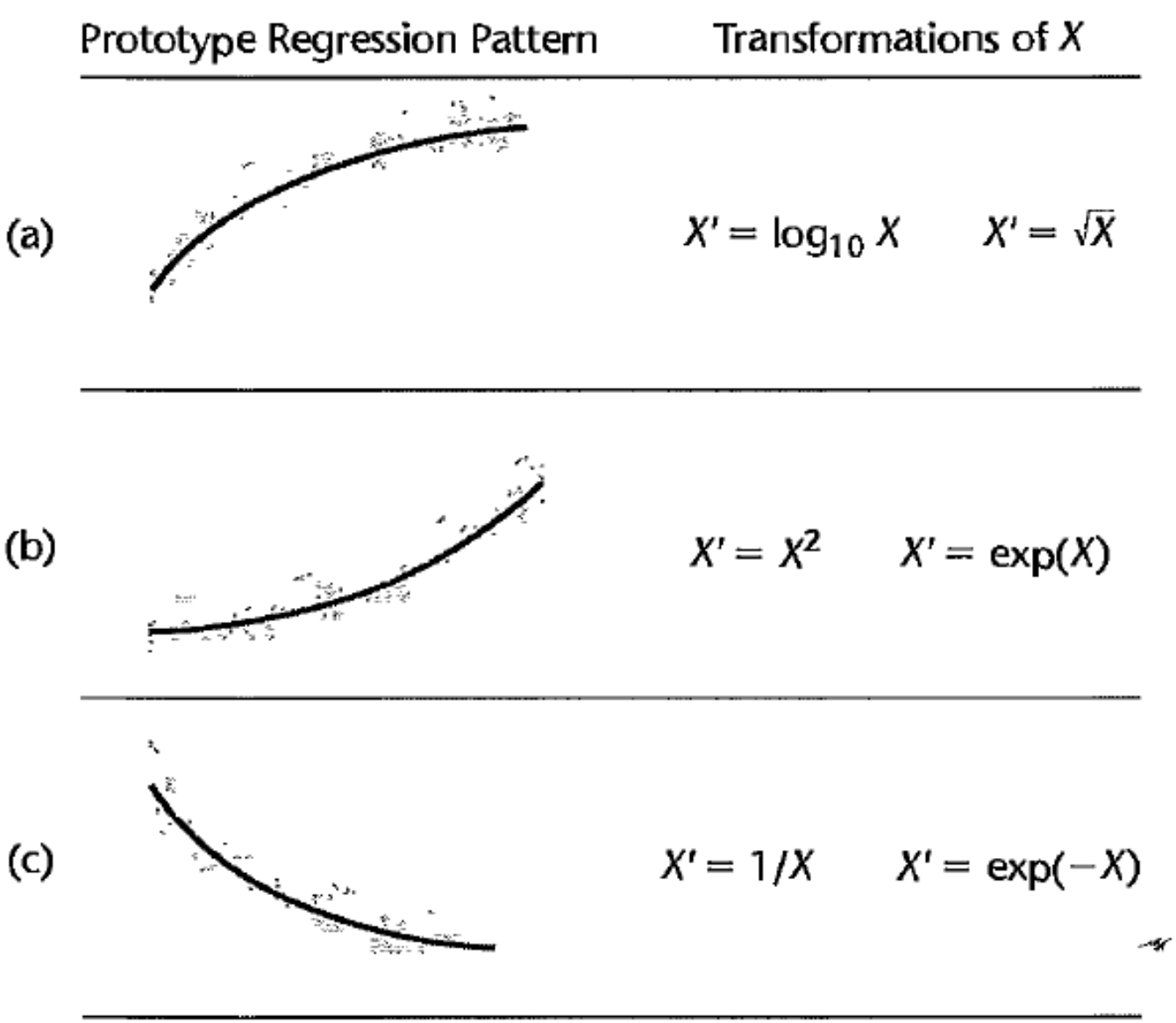
- Plot residuals against each predictor variable — Have insights on the functional form of Y on specific predictors.
- Plot residuals against fitted values — Have an overall picture of the entire model's linear fit.



Remedial measure: fit nonlinear regression models (polynomial, exponential, etc.)

Transformation of X for nonlinear fit

FIGURE 3.13
Prototype Nonlinear Regression Patterns with Constant Error Variance and Simple Transformations of X.

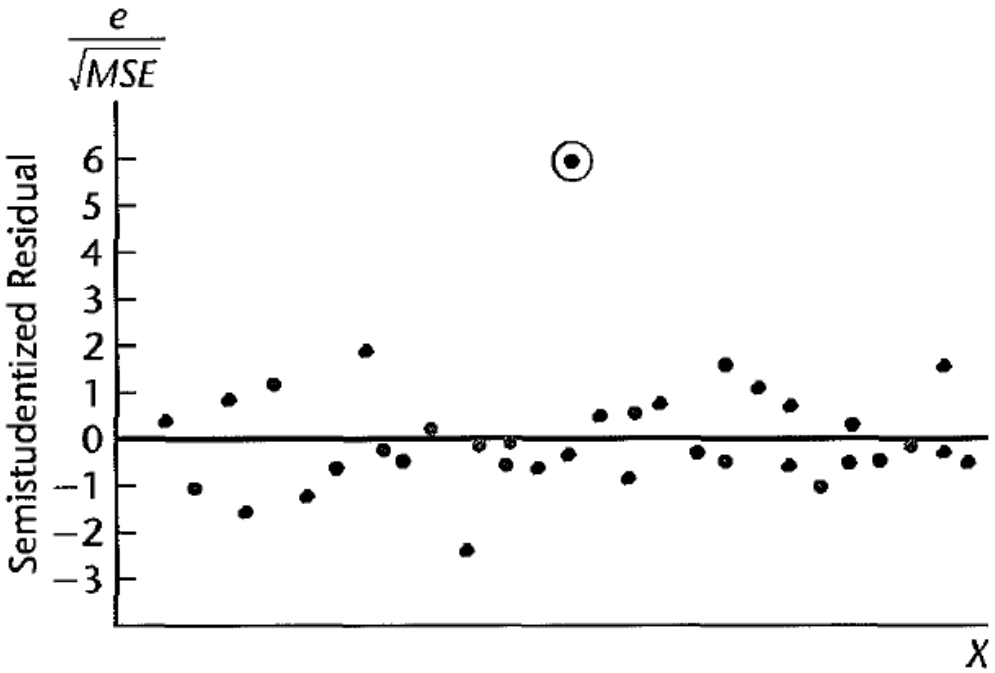


Residual analysis

1. The regression function is linear. ✔
2. No excessive influences by outliers.
3. The error terms have constant variance (homoskedasticity).
4. The error terms are normally distributed.
5. The error terms are independent among each other.

(2) Checking for outliers

Plot the "standarized" (semistudentized) residuals against $\hat{Y}$.



Look for absolute values of the "standarized" (semistudentized) residuals beyond 2.5 or 3.

(2) Checking for outliers

Remedial measures:

- Check for data entry errors.
- Investigate possible reasons for large departure from the group.
- Consider removing the outliers or separating data — if you do this, make sure to justify your action and remember to discuss the removed points separately.

Residual analysis

1. The regression function is linear. ✔
2. No excessive influences by outliers. ✔
3. The error terms have constant variance (homoskedasticity).
4. The error terms are normally distributed.
5. The error terms are independent among each other.

(3-5) Checking for constant variance, normality, and independence of error terms

Why do we need to check for (1) constant variance, (2) normality, and (3) independence of error terms?

Because of the assumption of the normal error model:

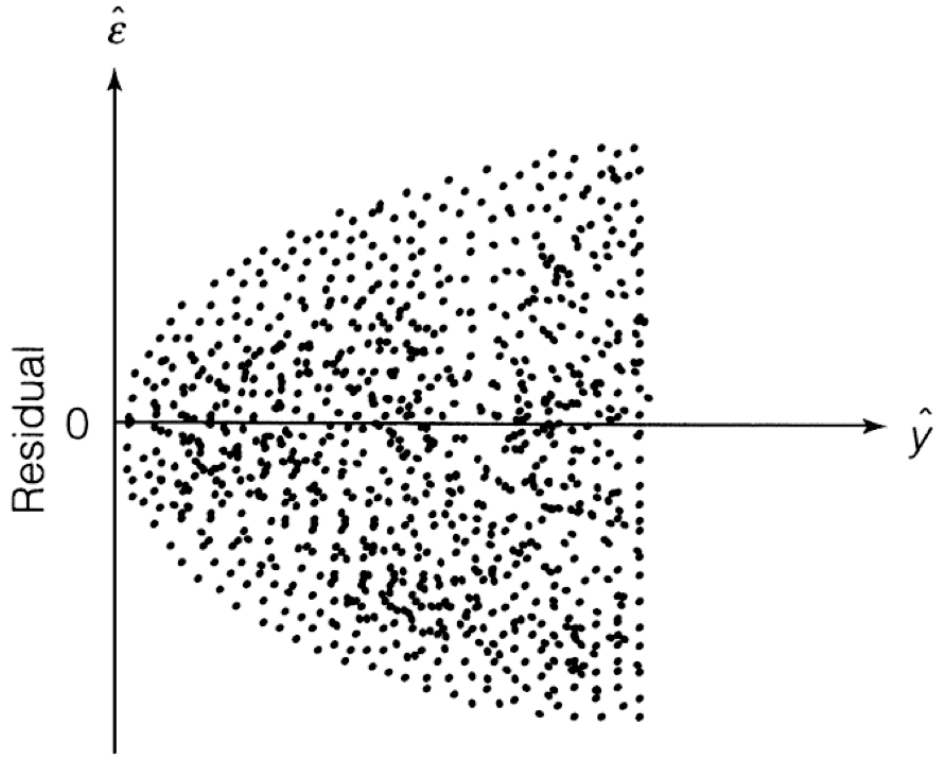
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad \text{where } \varepsilon_i \text{ iid} \sim N(0, \sigma^2)$$

"iid" — independently, identically (constant variance) distributed

Violation to the model assumption will impair the validity of the inferences (hypothesis tests, confidence intervals).

(3) Checking for constant variance of the residuals

Plot residuals against the fitted values.
R code: `plot(predict(m), residuals(m))`



If the response variable is a count that has a Poisson distribution, $\text{Var}(y) = E(y)$.

(3) Checking for constant variance of residuals

Remedial measures:

- Transform Y.
  - Most commonly log or sqrt.
  - Box-Cox transformation (p.134): a family of power transformation in the form of

    $$Y' = Y^{\lambda}$$

    where λ is estimated using MLE based on the sample data.

    Regression results based on the BC-transformed Y are statistically more efficient and associated with more powerful hypothesis tests and have tighter CIs.
  - **Beware:**"Transformations may obscure the fundamental interconnections between the variables, though at other times they may illuminate them"(p.128).

(3) Checking for constant variance of residuals

Other remedial measures:

- Weighted least square (WLS) estimators of the regression coefficients and their covariance matrix.
- Heteroskedastic corrected covariance matrix (HCCM; e.g., the sandwich estimators of the regression coefficients' covariance matrix).
- Bootstrapping to construct empirical confidence intervals of the regression coefficients.

(3) Checking for constant variance of residuals

Below illustrates how a sqrt transformation stabilizes the Poisson type residual variance.

Suppose h is some smooth function. Then by *Taylor's Theorem*:

$$\begin{aligned} h(y) &= h(E\{y\}) + \frac{h'(E\{y\})(y - E\{y\})}{1!} + \frac{h''(E\{y\})(y - E\{y\})^2}{2!} + \cdots \\ &\approx h(E\{y\}) + h'(E\{y\})(y - E\{y\}) \end{aligned}$$

The first term is a constant, so the variance of $h(y)$ is

$$\text{Var}\{h(y)\} \approx [h'(E(y))]^2 \text{Var}\{y\}$$

For count data that follow a Poisson distribution, $\text{Var}(y) = E(y)$, and with a sqrt transformation, $h(y) = \sqrt{y}$, so $h'(y) = \frac{1}{2}y^{-1/2}$, and

$$\text{Var}\{\sqrt{y}\} \approx [\frac{1}{2}E\{y\}^{-1/2}]^2 E\{y\} = 1/4$$

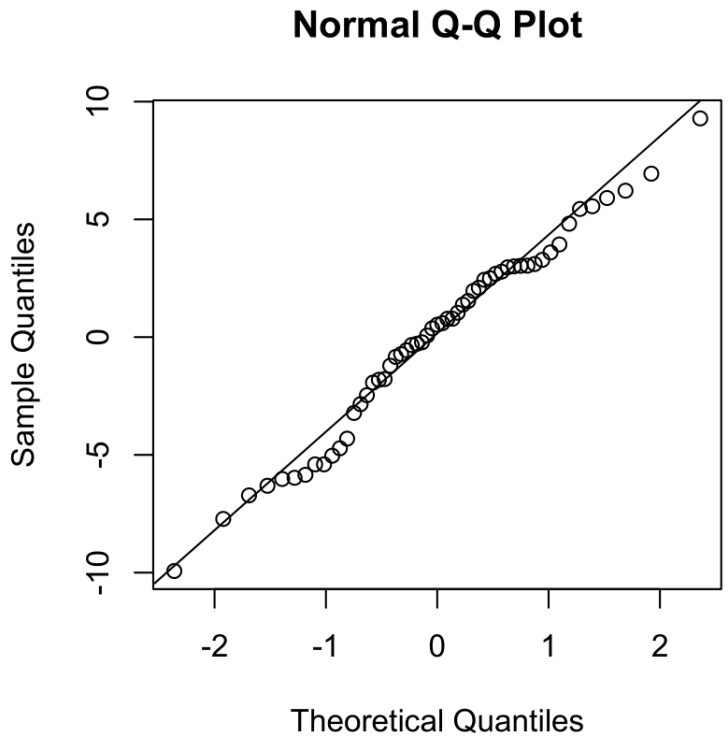
The above is a constant regardless of the value of $E\{y\}$).

Residual analysis

1. The regression function is linear. ✔
2. No excessive influences by outliers. ✔
3. The error terms have constant variance (homoskedasticity). ✔
4. The error terms are normally distributed.
5. The error terms are independent among each other.

(4) Checking for normal distribution of residuals

Generate a Q-Q plot of the residuals.
R code: `qqnorm(residuals(m)); qqline(residuals(m))`



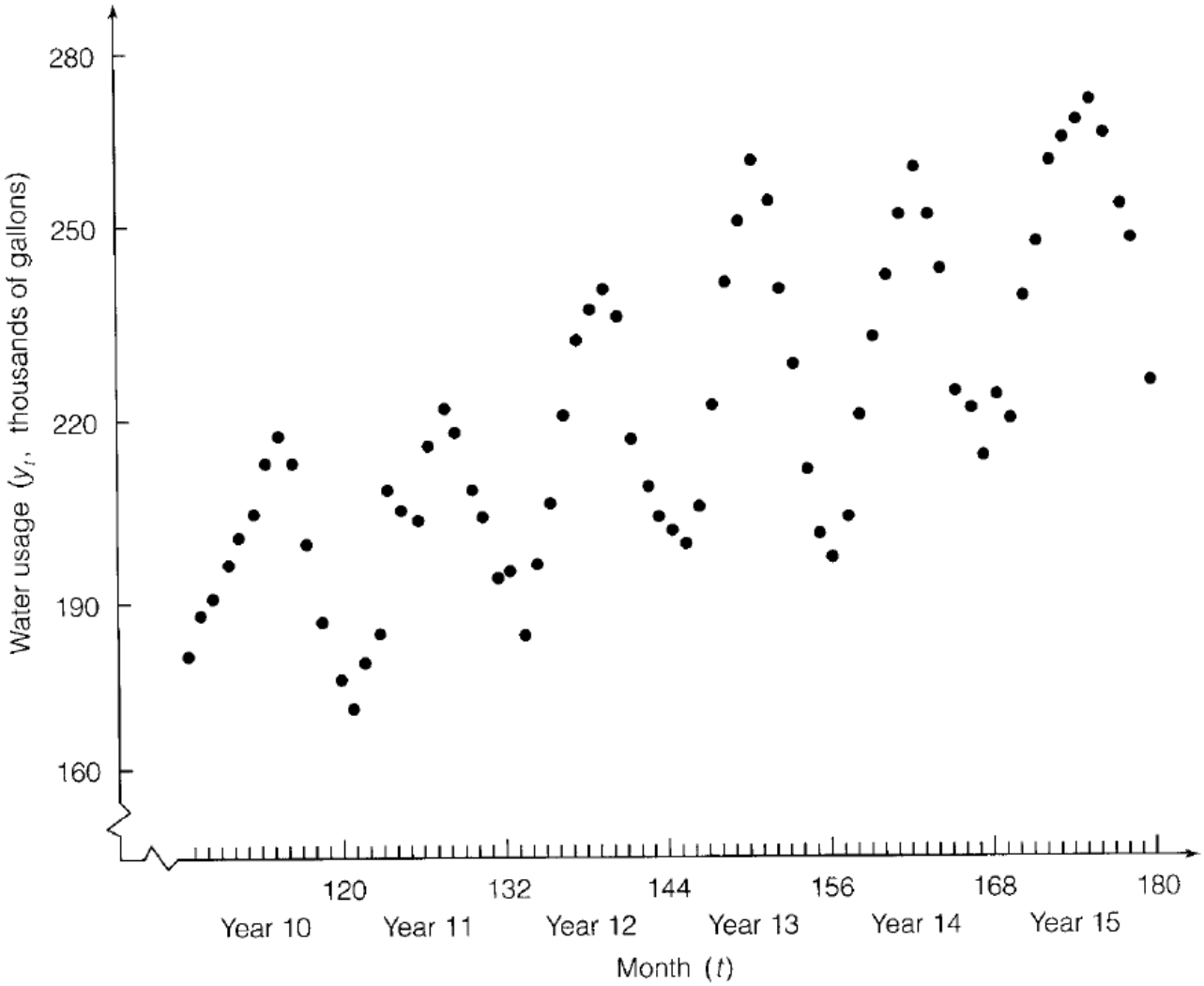
Remedial measure:

- Inferences based on t-distribution are robust to violation of normality, especially with large samples.
- Transform Y (see slide 34).
- Bootstrapping to obtain empirical CIs

Residual analysis

1. The regression function is linear. ✔
2. No excessive influences by outliers. ✔
3. The error terms have constant variance (homoskedasticity). ✔
4. The error terms are normally distributed. ✔
5. The error terms are independent among each other.

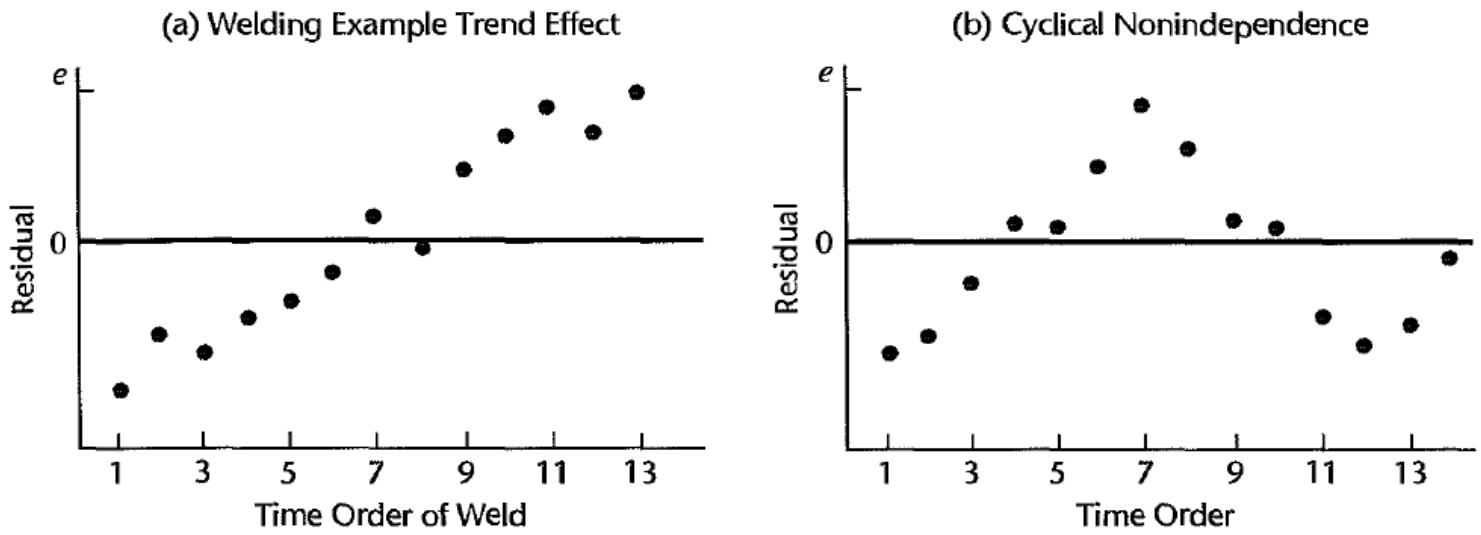
(5) Checking for dependent residuals



(5) Checking for dependent residuals

Examine the sequence plot of the residuals:

FIGURE 3.8 Residual Time Sequence Plots Illustrating Nonindependence of Error Terms.



When the error terms are independent, we expect the residuals in a sequence plot to fluctuate in a more or less random pattern around the baseline 0. Lack of randomness can take the form of too much or too little alternation of points around the zero line.

(5) Checking for dependent residuals

Another type of residual dependency: **Nested data**

Example: Students nested within classes, classes nested within Schools.

The unit of analysis is students. But students are not simple random samples. Students are clustered based on classes. Students from the same class share some common characteristics unique from the other classes. The regression residuals tend to correlate among students in the same class (intra-class correlation, ICC).

Remedial:

- Multilevel Modeling
- a.k.a., Mixed Effects Modeling
- a.k.a., Hierarchical Linear Modeling

Residual analysis

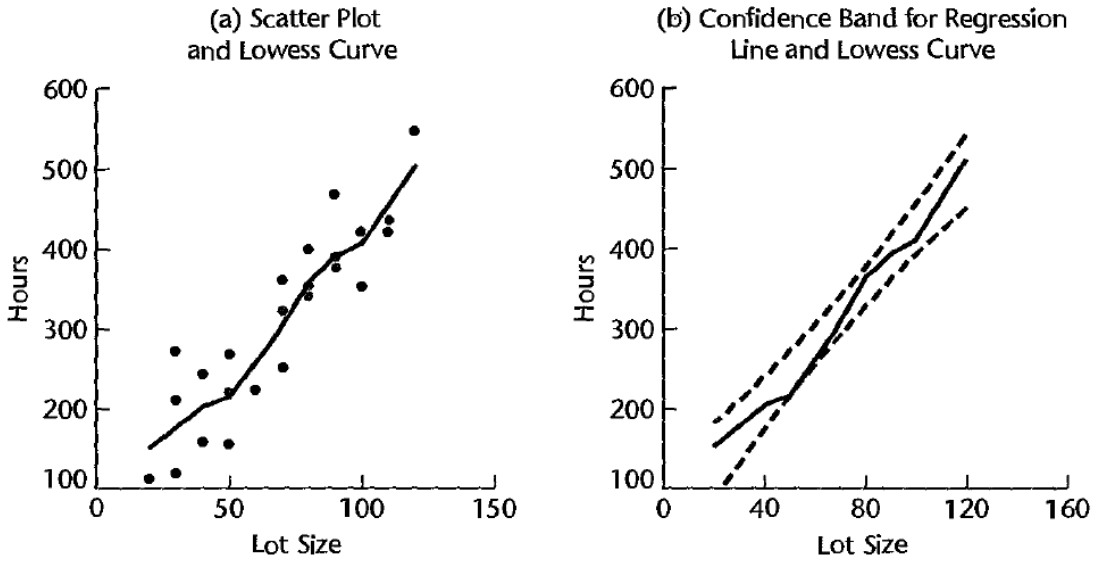
1. The regression function is linear. ✔
2. No excessive influences by outliers. ✔
3. The error terms have constant variance (homoskedasticity). ✔
4. The error terms are normally distributed. ✔
5. The error terms are independent among each other. ✔

Outline

Exploring the shape of the regression function

Nonparametric smoothing methods can be used to explore the shape of the regression function:

- Moving averages
- LOWESS (locally weighted regression scatter plot smoothing): fits a weighted linear regression to the data in the neighborhood of the X value and then using the fitted value at X as the smoothed value.



- LOESS (locally estimated scatterplot smoothing): works with multiple predictors and generates a surface.