Stat 608 Chapter 3

Summary

- Chapter 2:
  - Set up model:
    - Inferences about model parameters & regression line
    - Dummy (categorical) variable regression
    - Assumptions: **L**inear, **I**ndependent, **N**ormal, **E**rrors have constant variance

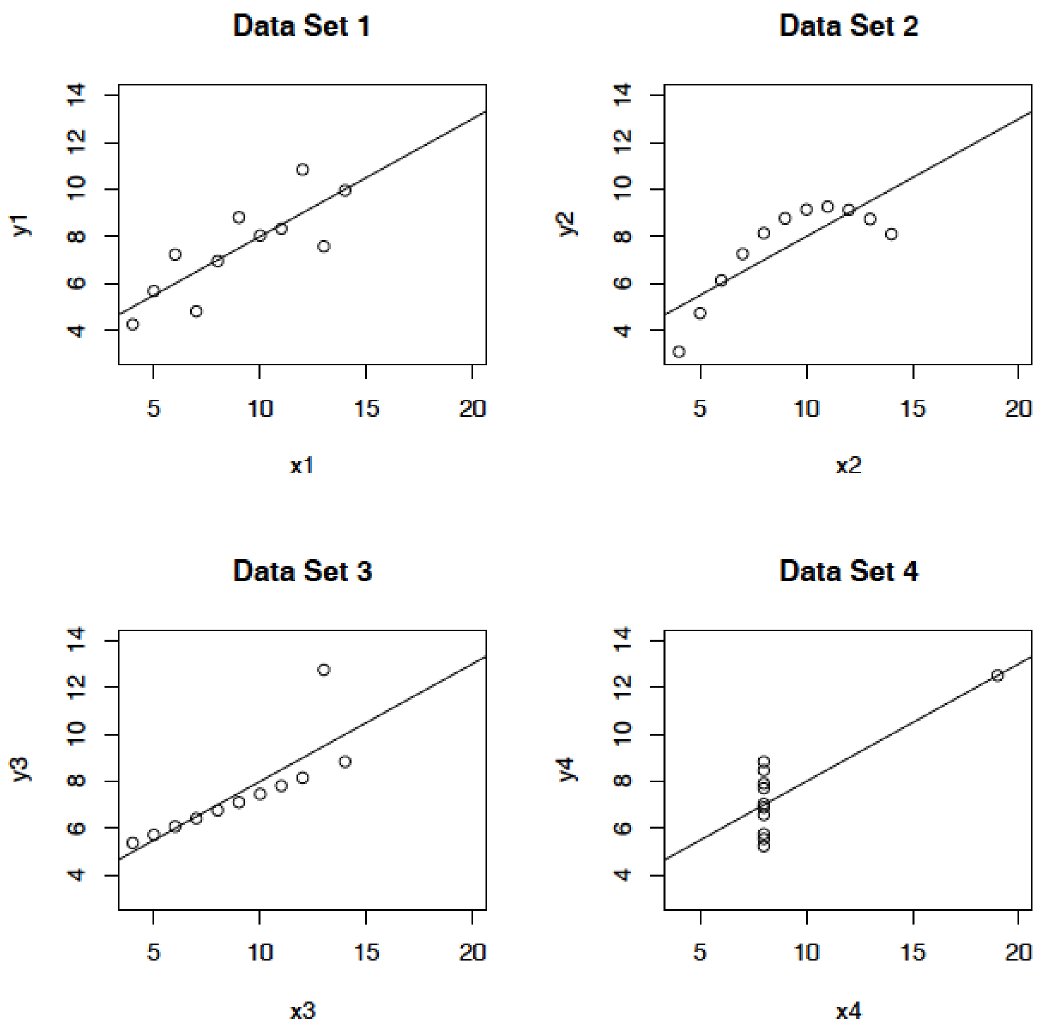
- Chapter 3:
  - Check the model assumptions:
    - Residual plots
    - Leverage & Influence
    - Transformations

+ Anscombe's Data Sets

- Valid Model
  - Is the mean structure correct?
    - For valid models, $E[Y \mid X = x] = \beta_0 + \beta_1 x$
  - Is the variance structure correct?
    - The most common assumption is $Var(Y \mid X = x) = \sigma^2$.

Anscombe's Data Sets

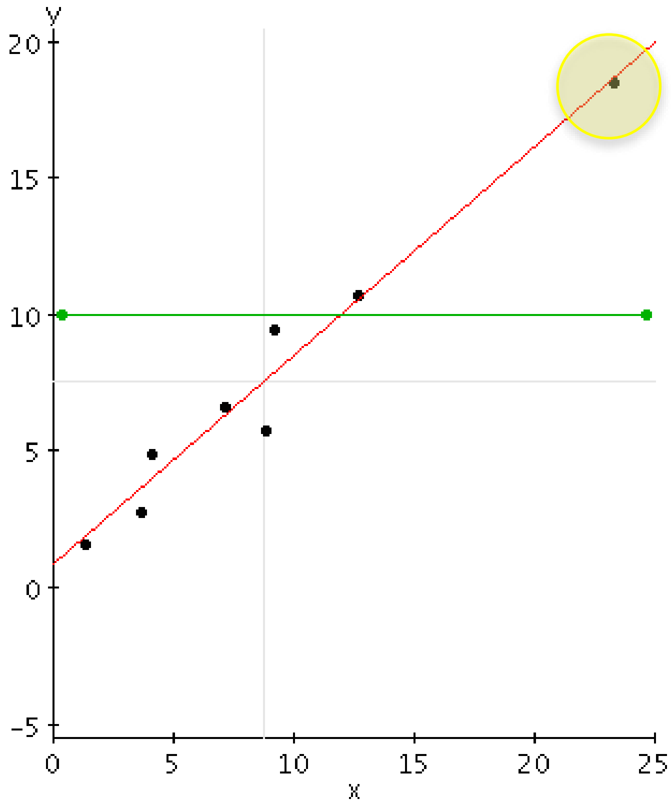
- Is SLR a vaild model for any of these datasets?
  - Same slope
  - Same intercept
  - Same R^2
  - Same s.e.
  - Same p-values



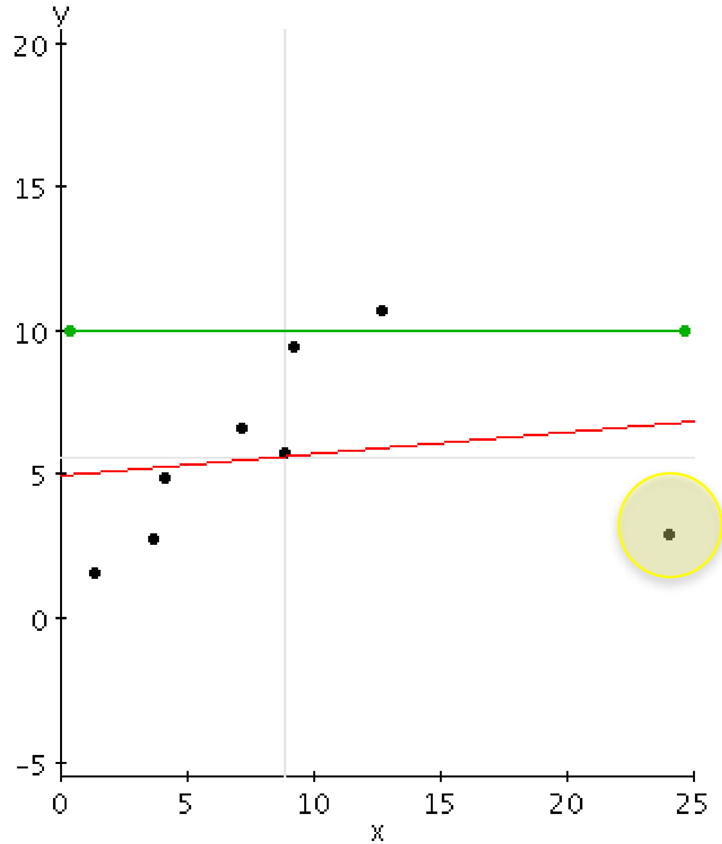
Using Residuals

- Plot residuals:
  - No pattern => model could be valid
  - Pattern => residual plot gives information on what to do
- Look at individual points:
  - **Outliers** don't follow the pattern of the rest of the data, after taking into account the model.
  - **Leverage** points have an unusually large effect on the estimated regression model. (Does the model change much when a single point is removed?)

Leverage Points



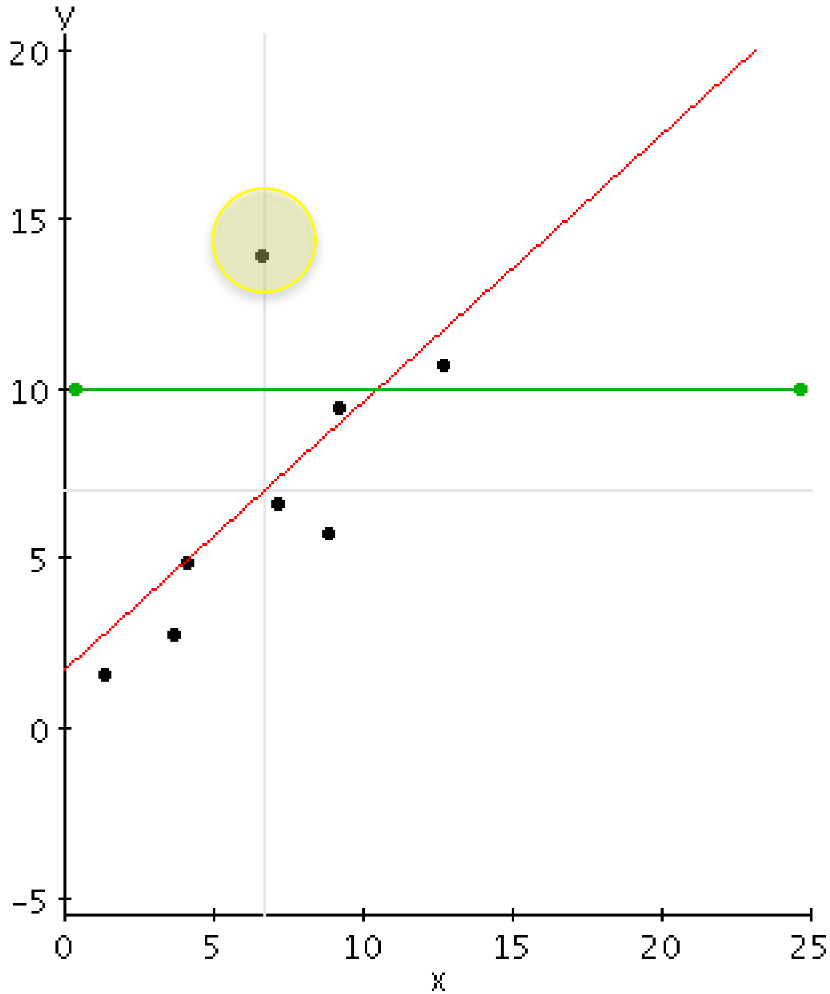
Good leverage point



Bad leverage point

http://www.stat.tamu.edu/~west/ph/regeye.html

Not a Leverage Point



+
Leverage Points

- A **leverage point** is a point whose x-value is far from the other x-values in the data set.
- A leverage point is a **bad leverage point** if its y-value does not follow the pattern set by the other points. That is, a bad leverage point is a leverage point which is also a regression outlier.

Leverage

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$

$$h_{ii} = i^{th} \text{ diagonal of } \mathbf{H}$$

$$\begin{aligned}\sum_i h_{ii} &= tr(\mathbf{H}) \\ &= tr(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \\ &= tr((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}) \\ &= tr(\mathbf{I}_{\mathrm{p}+1}) = p + 1\end{aligned}$$

+ Leverage Rule

- Leverage of i^{th} point = h_{ii}
- A popular rule is to classify the i^{th} point as a leverage point in a multiple linear regression model with p predictors (and one intercept, so that the design matrix has p+1 columns) if:

$$h_{ii} > 2 \times \text{average}(h_{ii}) = 2 \times \frac{p+1}{n}$$

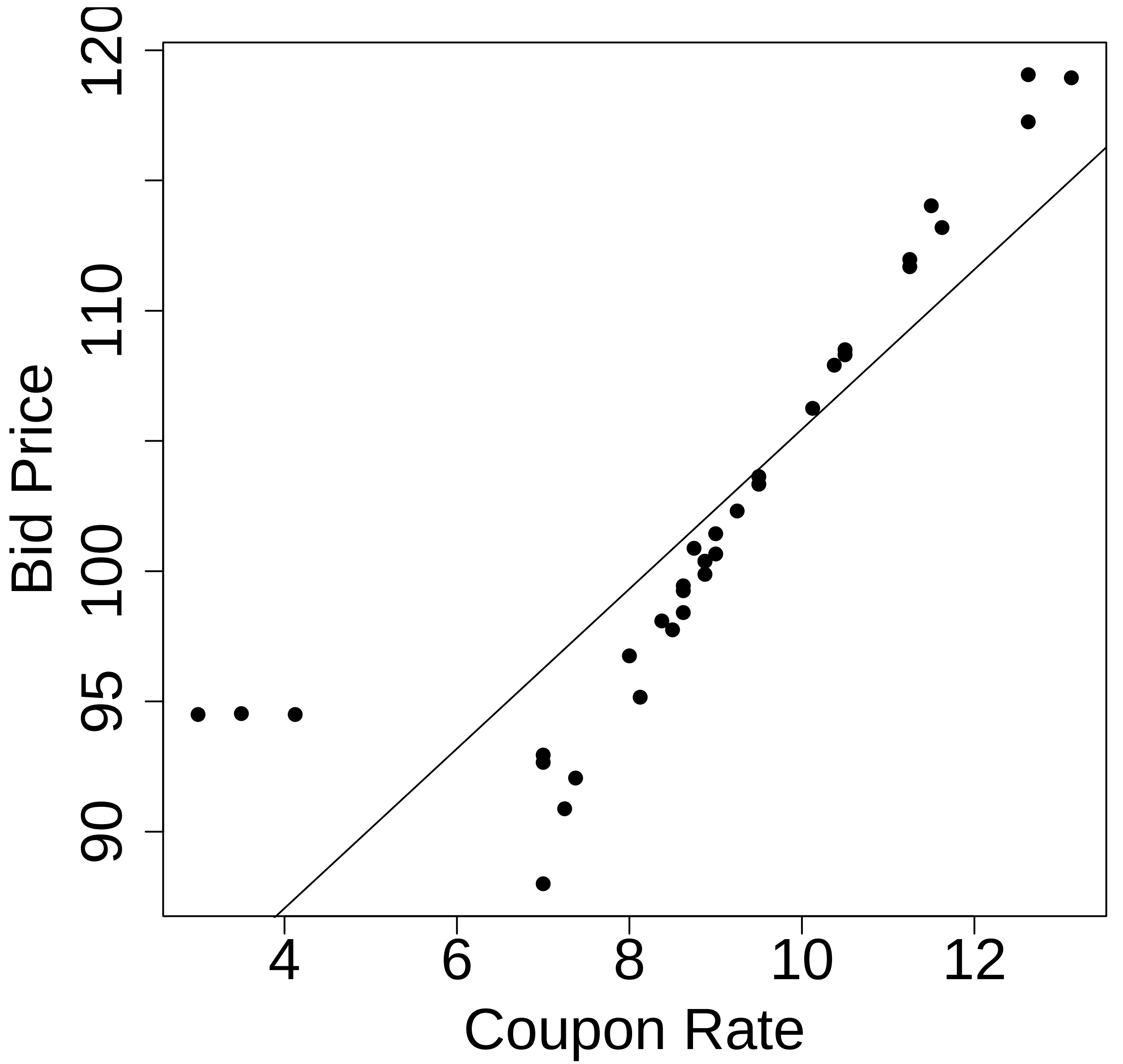
- In a simple linear regression, the i^{th} point is a leverage point if $h_{ii} > 4/n$.

Treasury Bonds

U.S. Treasury bonds maturing between 1994 and 1998.

Half the coupon rate is paid every six months (e.g. a 7% bond pays $3.50 every six months) until maturity, at which time it pays $100.

Treasury Bonds: Regression Line



+ Treasury Bonds: Leverage Points

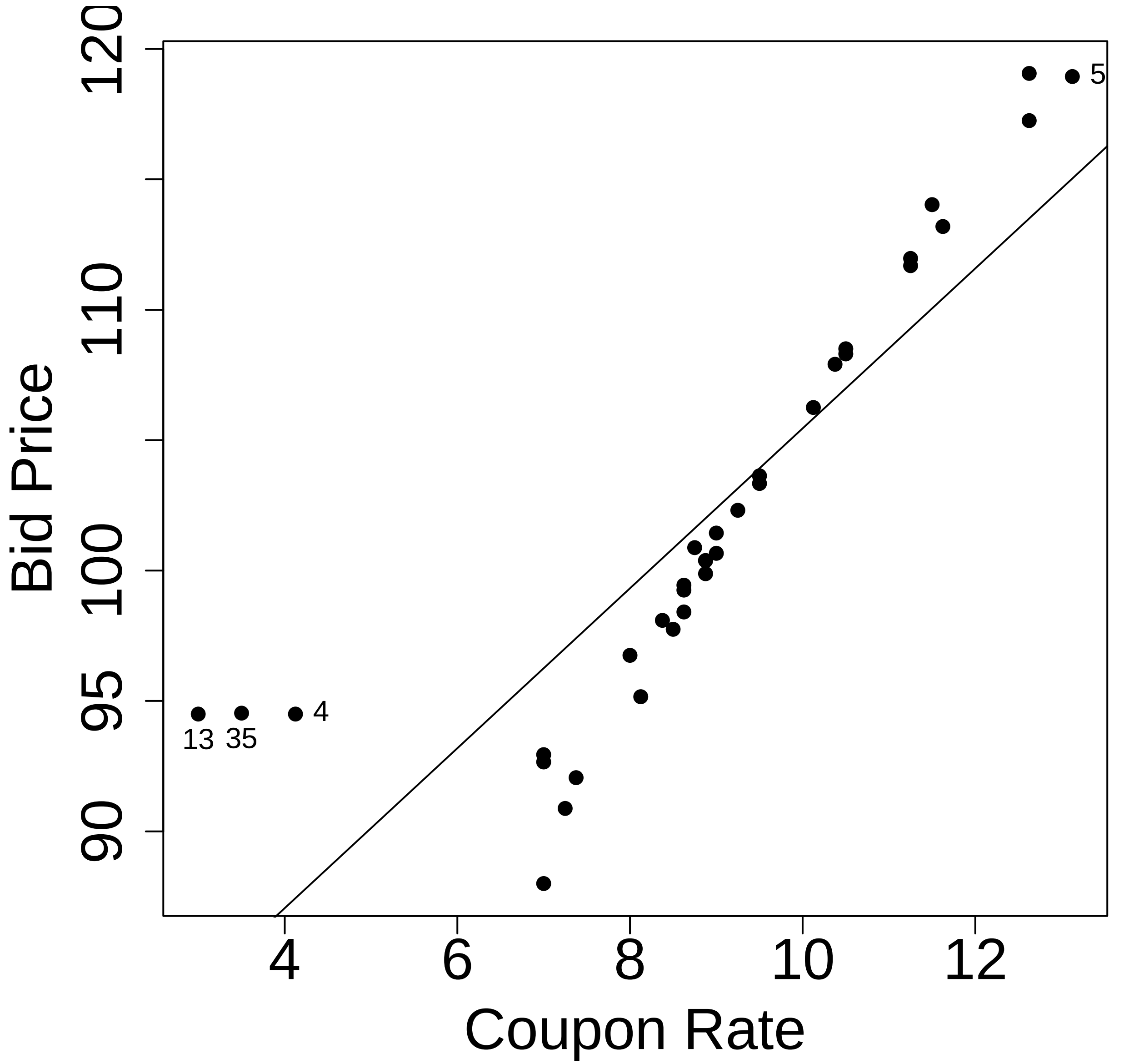
$$n = 35$$

$$4/35 = 0.11$$

- Cases 4, 5, 13, and 35 all have leverage values greater than 0.11.

```
> hatvalues(my.lm)
         1          2          3          4          5          6          7
0.04850372 0.02860476 0.04850372 0.15277805 0.12397083 0.03315530 0.02873009
         8          9         10         11         12         13         14
0.10262572 0.03037870 0.03639225 0.06803392 0.02904583 0.21787629 0.04202499
        15         16         17         18         19         20         21
0.05784582 0.03018347 0.03997869 0.05784582 0.10262572 0.02858307 0.04202499
        22         23         24         25         26         27         28
0.02904583 0.03037870 0.06446917 0.02858307 0.04148268 0.04365431 0.02904583
        29         30         31         32         33         34         35
0.02953029 0.02858307 0.03199597 0.02860476 0.02915429 0.04850372 0.18725657
```

Treasury Bonds: High Leverage



+ Strategies for bad leverage points

1. Remove invalid data points

Are these data unusual or different in some way from the rest of the model? If so, should we use a different model for these data? Our three worst leverage points correspond to flower bonds, which have tax advantages over other bonds.

2. Fit a different regression model

Has an incorrect model been fit to the data? Consider a different model:

- Add predictor variables
- Transform Y and / or x.

+

Good leverage points

Caution:

While “good” leverage points may not affect the estimated values for the regression coefficients, they do affect the estimated standard errors, p-values, correlation, and R^2.

+ Leverage Properties

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{SXX}$$

- What happens to leverage if the point is farther from the rest of the data in the x-direction?
- Does whether a point has high leverage depend on the value of y for that point?

Standardized Residuals

$$Var(\hat{\mathbf{y}}) = \sigma^2 \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \sigma^2 \mathbf{H}$$

$$Var(\mathbf{y}) = \sigma^2 \mathbf{I}$$

$$Var(\hat{\mathbf{e}}) = Var(\mathbf{y} - \hat{\mathbf{y}}) = \sigma^2 (\mathbf{I} - \mathbf{H})$$

$$r_i = \frac{\hat{e}_i}{s\sqrt{1 - h_{ii}}}$$

+ Standardized Residuals

$$r_i = \frac{\hat{e}_i}{s\sqrt{1 - h_{ii}}}$$

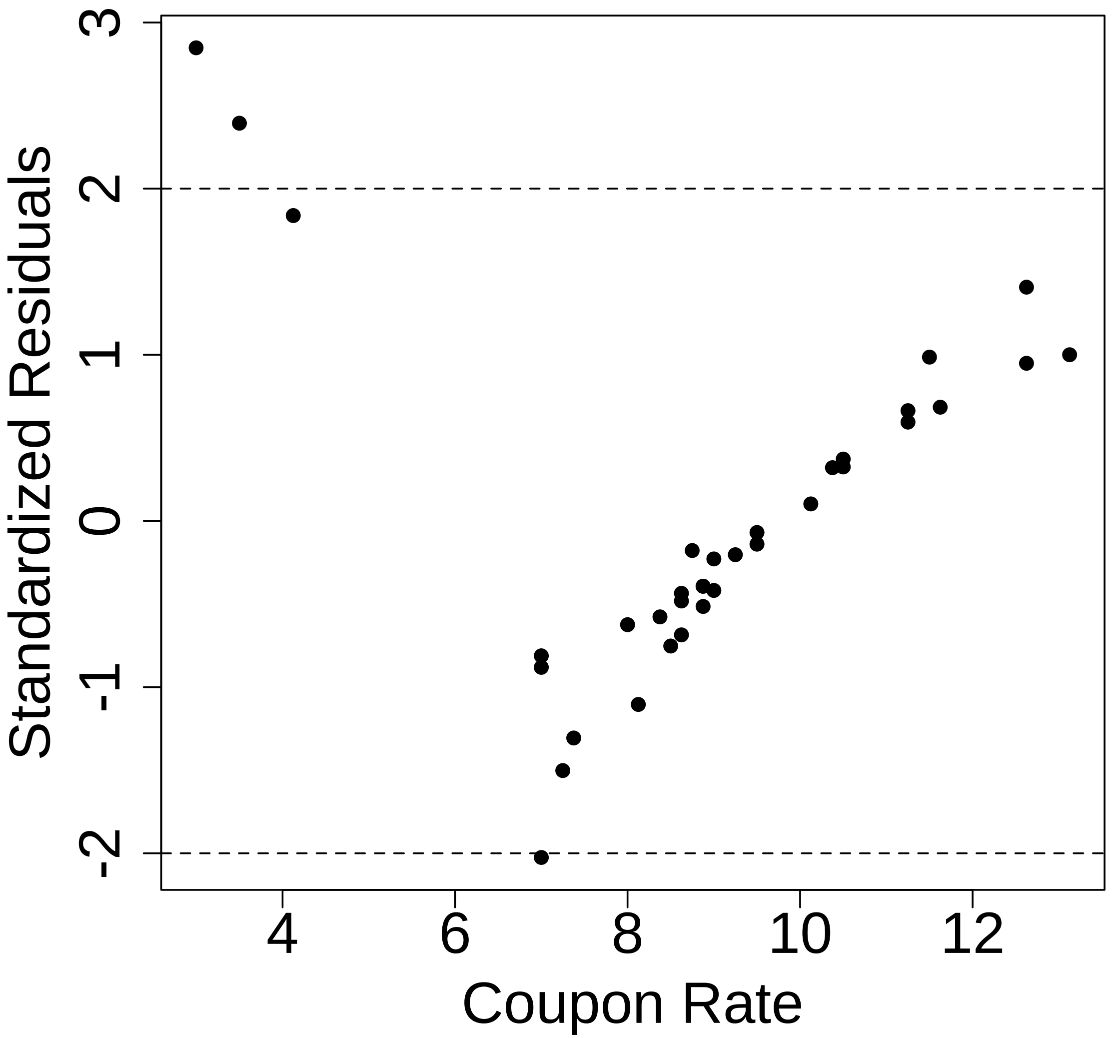
- Two advantages to standardizing (taking z-scores of) residuals:
  1. Residuals (sample) will have nonconstant variance even if the errors (population) have constant variance in the presence of high leverage.
  2. The standardization immediately tells us how many standard deviations any point is away from the fitted regression model. If the errors are normally distributed, 95% of them should be within 2 standard deviations of 0; for small data sets, this is our rule of thumb for outliers. For very large data sets, we expand this rule to +/- 4 standard deviations.

Rule

Recall that a bad leverage point is a leverage point which is also an outlier. Thus, a bad leverage point is a leverage point whose standardized residual falls outside the interval from –2 to 2.

On the other hand, a "good" leverage point is a leverage point whose standardized residual falls inside the interval from –2 to 2 .

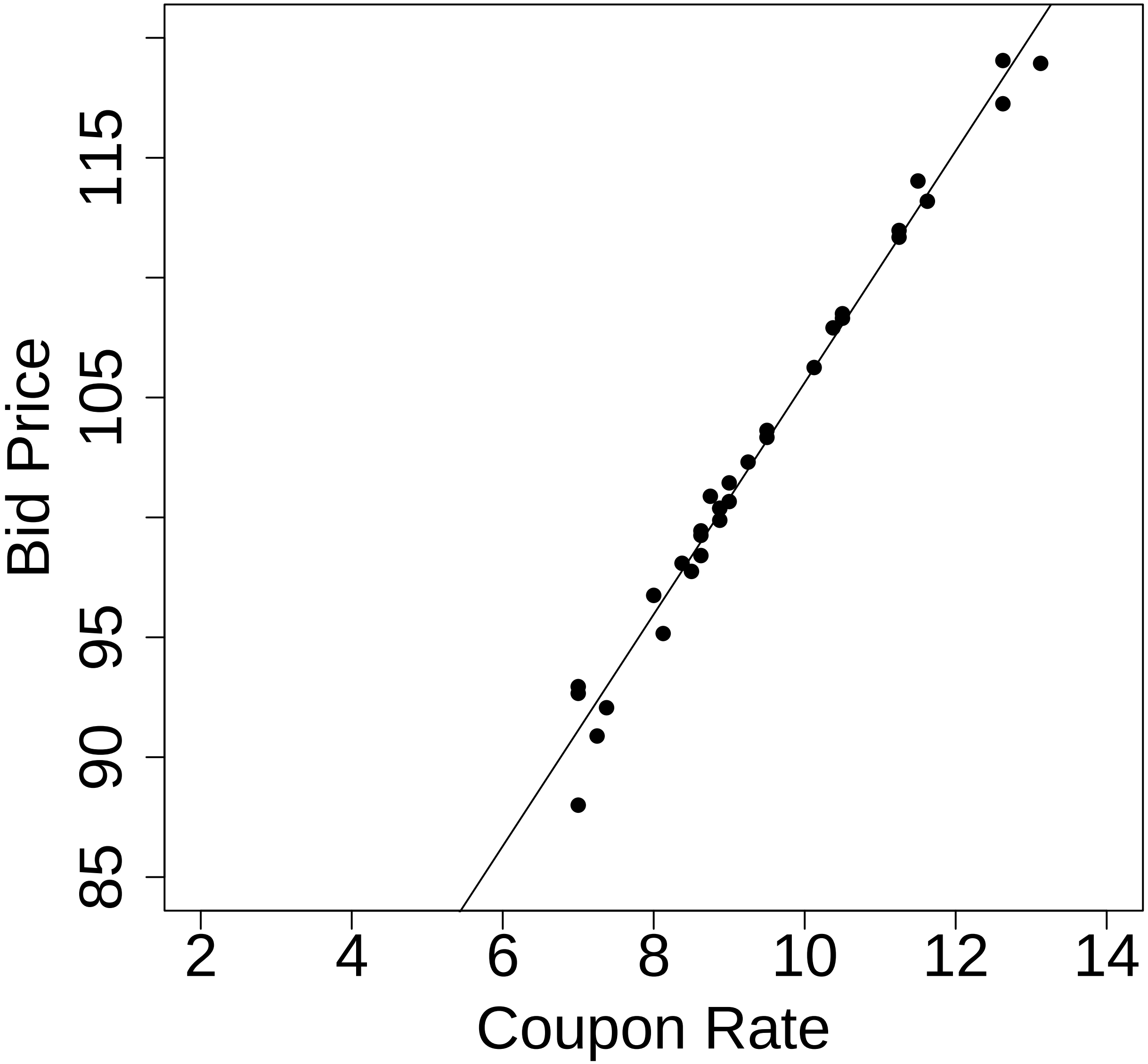
Treasury Bonds: Residuals



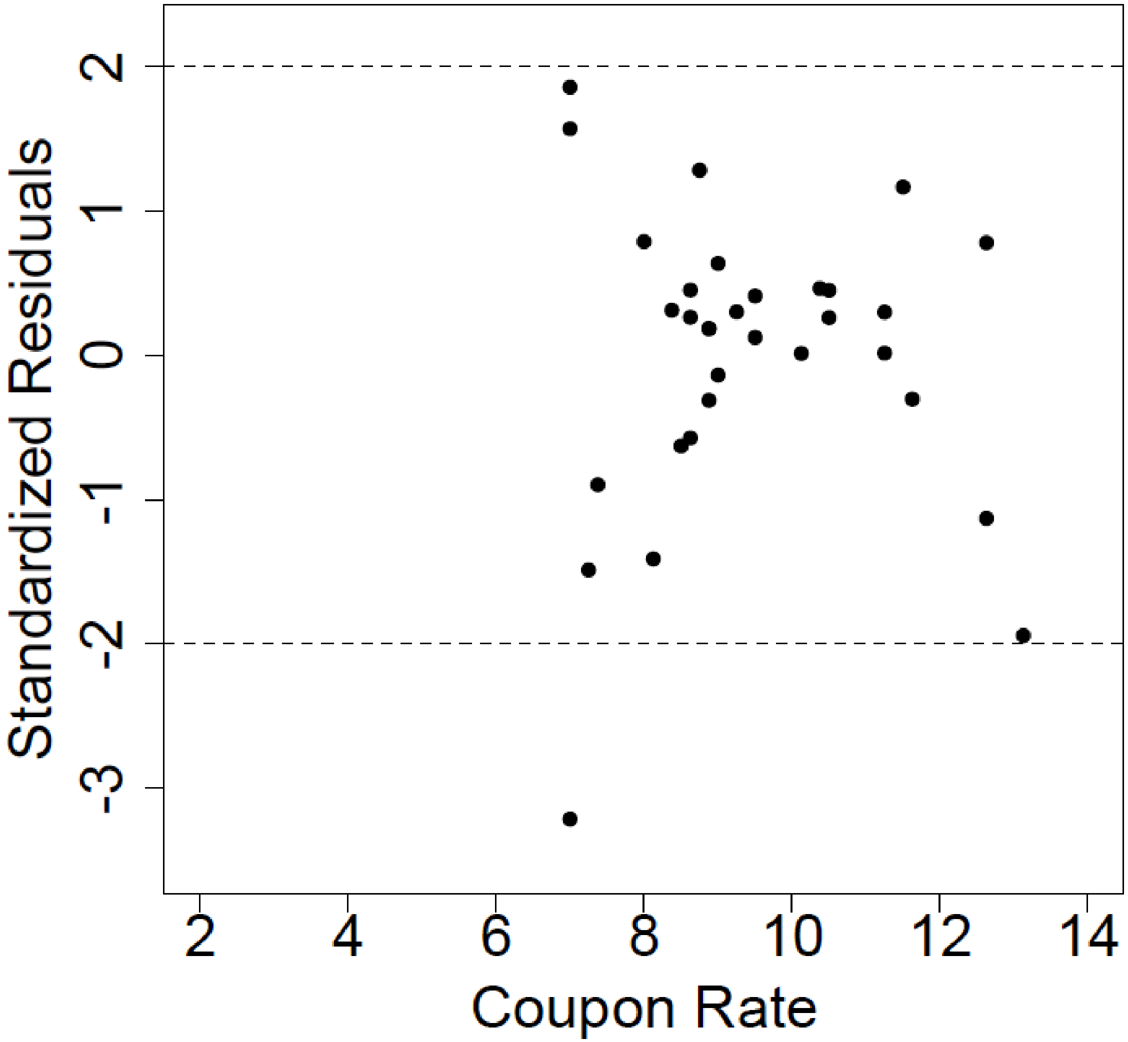
Treasury Bonds: Residuals

- First, the pattern of residuals is not a random scatter, so our model is not valid.
- We saw earlier that cases 4, 5, 13, and 35 could be classified as leverage points.
- Cases 13, 35, and 34 have standardized residual values greater than -2, and case 4 has standardized residual equal to -1.8.
- Cases 13 and 35 (and to a lesser extent, case 4) are thus of high leverage that are also outliers; thus, they are **bad leverage points**.

Treasury Bonds: New Model



+ Treasury Bonds: New Model



Recommendations for Handling Outliers and Leverage Points

- Points shouldn't be routinely deleted just because they do not fit the model!
- Outliers often point to an alternative model. Including one or more dummy variables is one way of coping with outliers that point to an important feature.

+

Correlation between estimated errors

There is a small amount of correlation present in standardized residuals, even if the errors are independent. In fact it can be shown that

$$\mathrm{Cov}(\hat{e}_i, \hat{e}_j) = -h_{ij}\sigma^2 \, (i \neq j)$$

$$\mathrm{Corr}(\hat{e}_i, \hat{e}_j) = \frac{-h_{ij}}{\sqrt{(1-h_{ii})(1-h_{jj})}} (i \neq j)$$

However, the size of the **correlations inherent in the least squares residuals** are generally so small in situations in which correlated errors is an issue (e.g., data collected over time) that they can be effectively ignored in practice.

Assessing Influence

Cook's D:

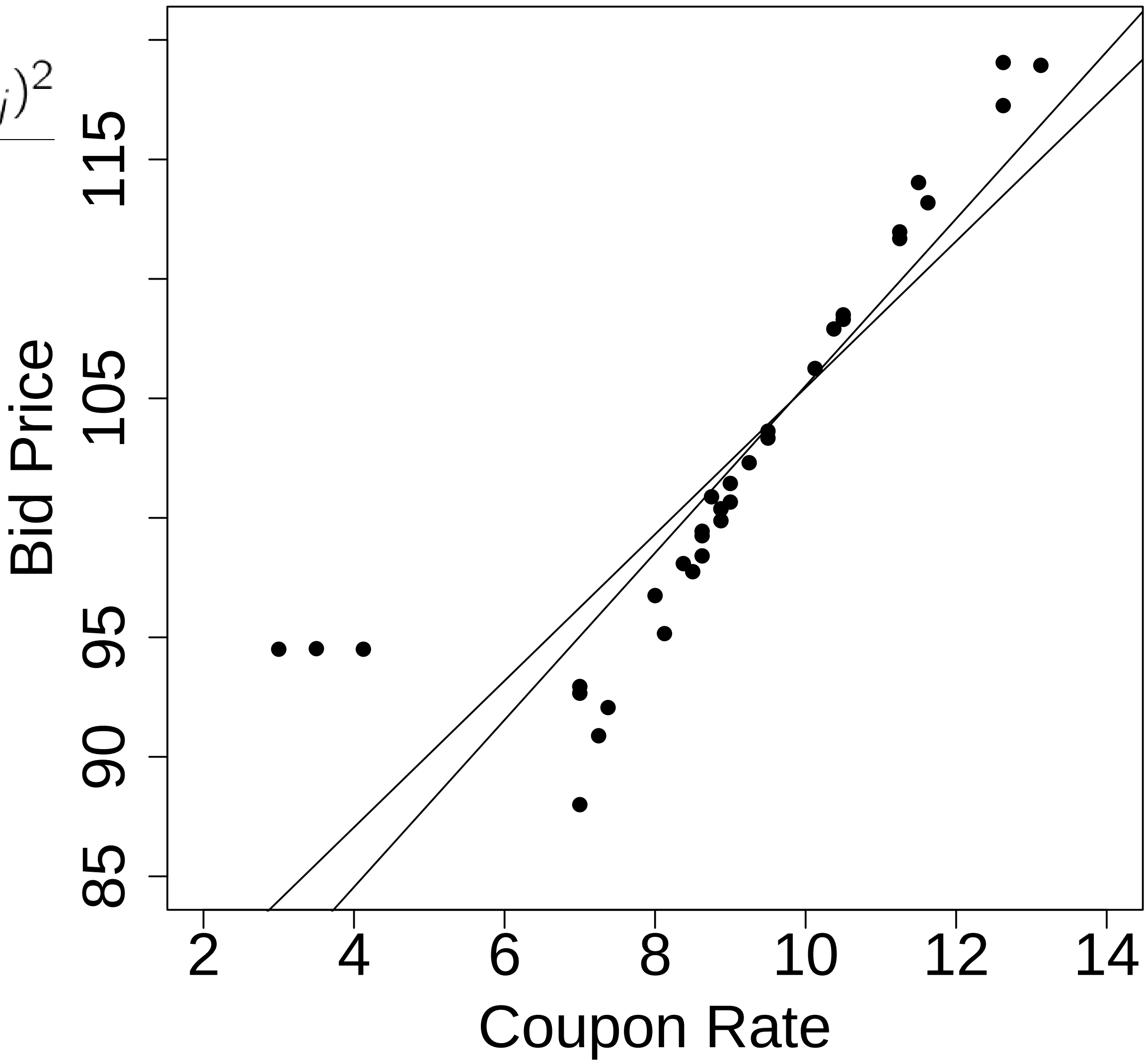
$$D_i = \frac{\sum_{j=1}^{n}(\hat{y}_{j(i)} - \hat{y}_i)^2}{(p+1)S^2} = \frac{r_i^2}{(p+1)}\frac{h_{ii}}{(1-h_{ii})}$$

- The notation j(i) means the value of y-hat with the i^{th} observation deleted.
- When is D_i large?
- One rule of thumb is to classify an observation as noteworthy if D_i is greater than 4/(n-2).
- In practice, look at gaps in the D_i values.

Assessing Influence

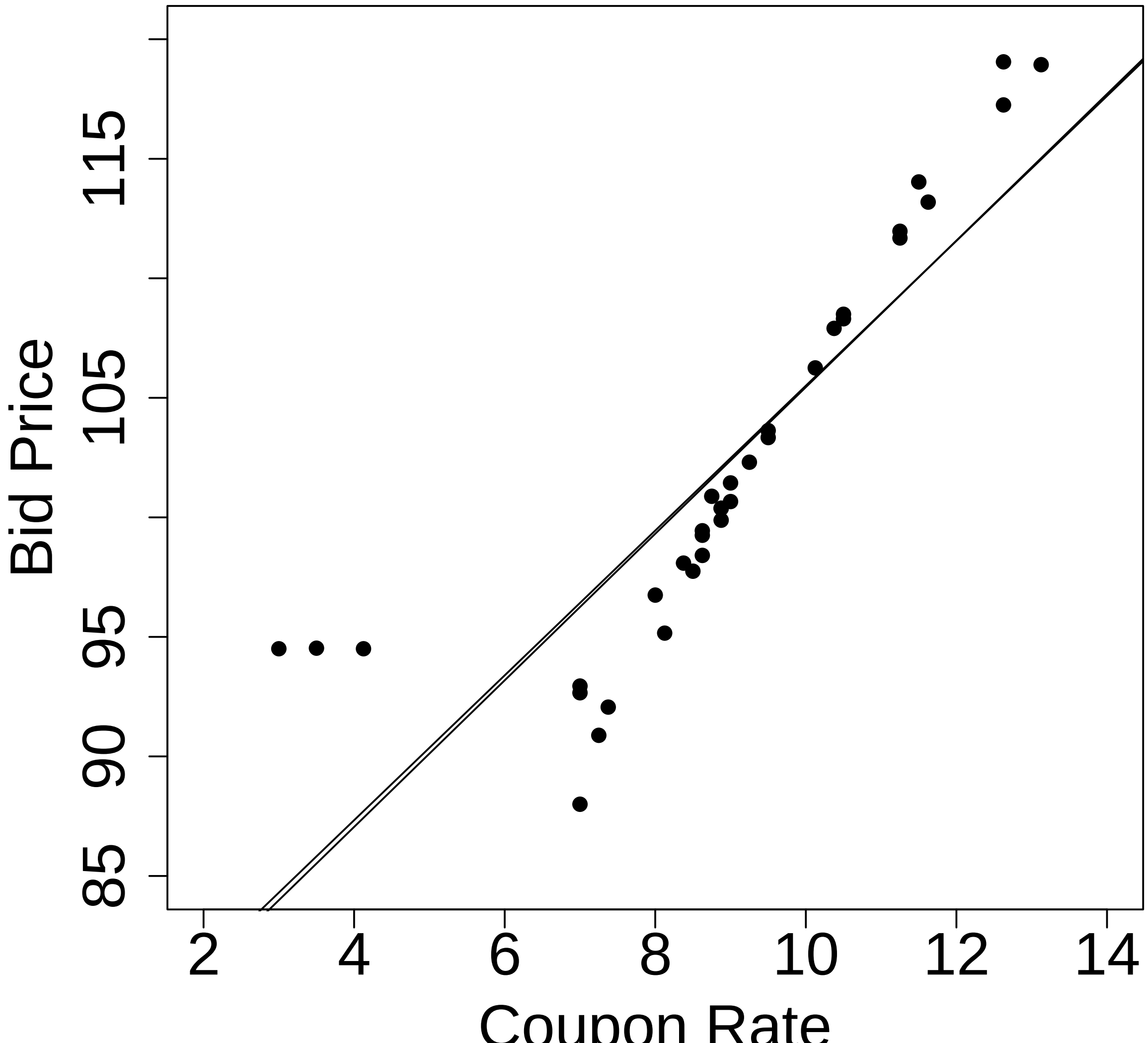
$$D_i = \frac{\sum_{j=1}^{n} (\hat{y}_{j(i)} - \hat{y}_j)^2}{(p+1)S^2}$$

Remove point #13 and compare

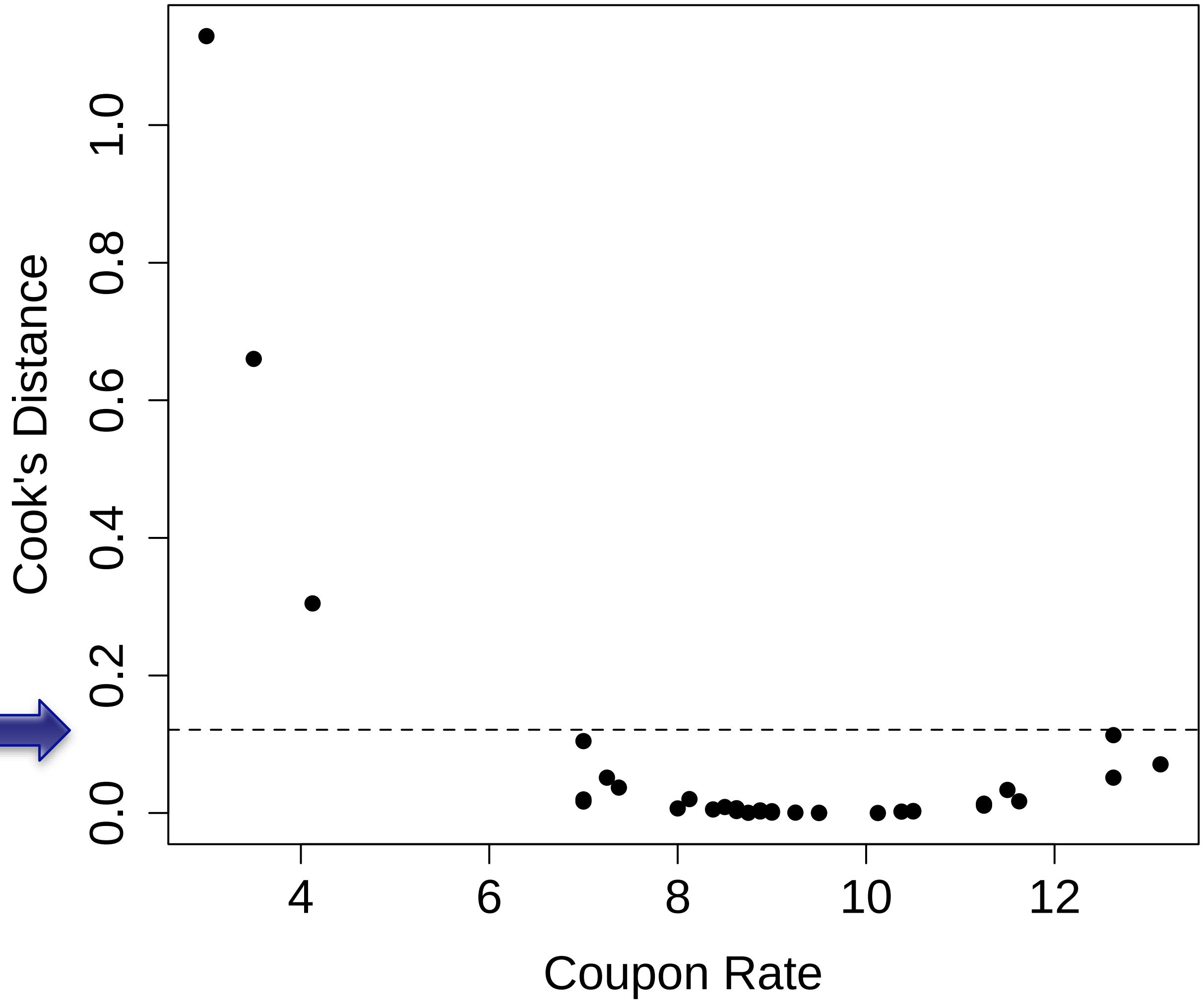


Assessing Influence

Remove point #1:

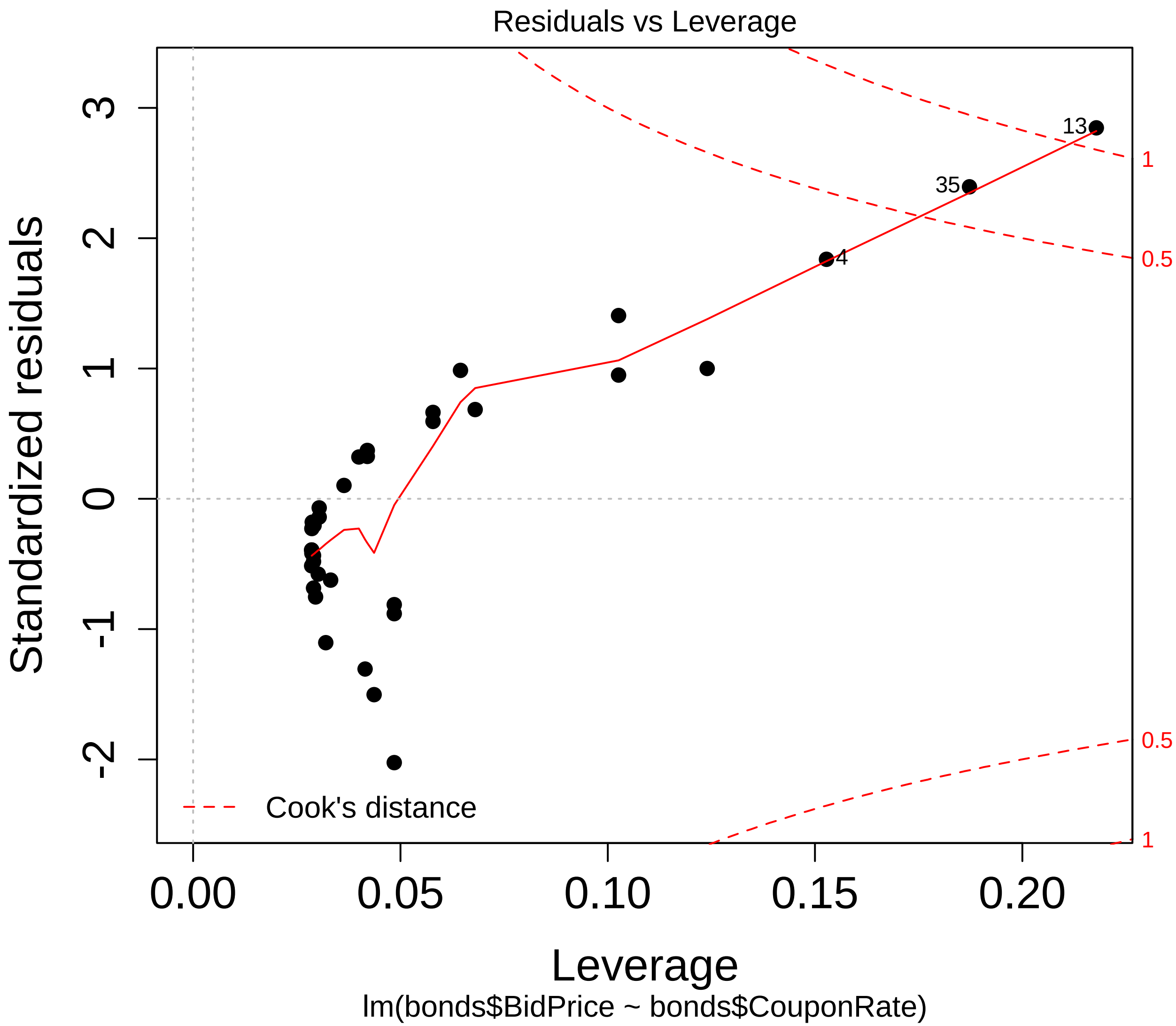


Cook's D: Bond data



33/2 = 0.12

Cook's D: Bond data (default plot)



Normality of the Errors

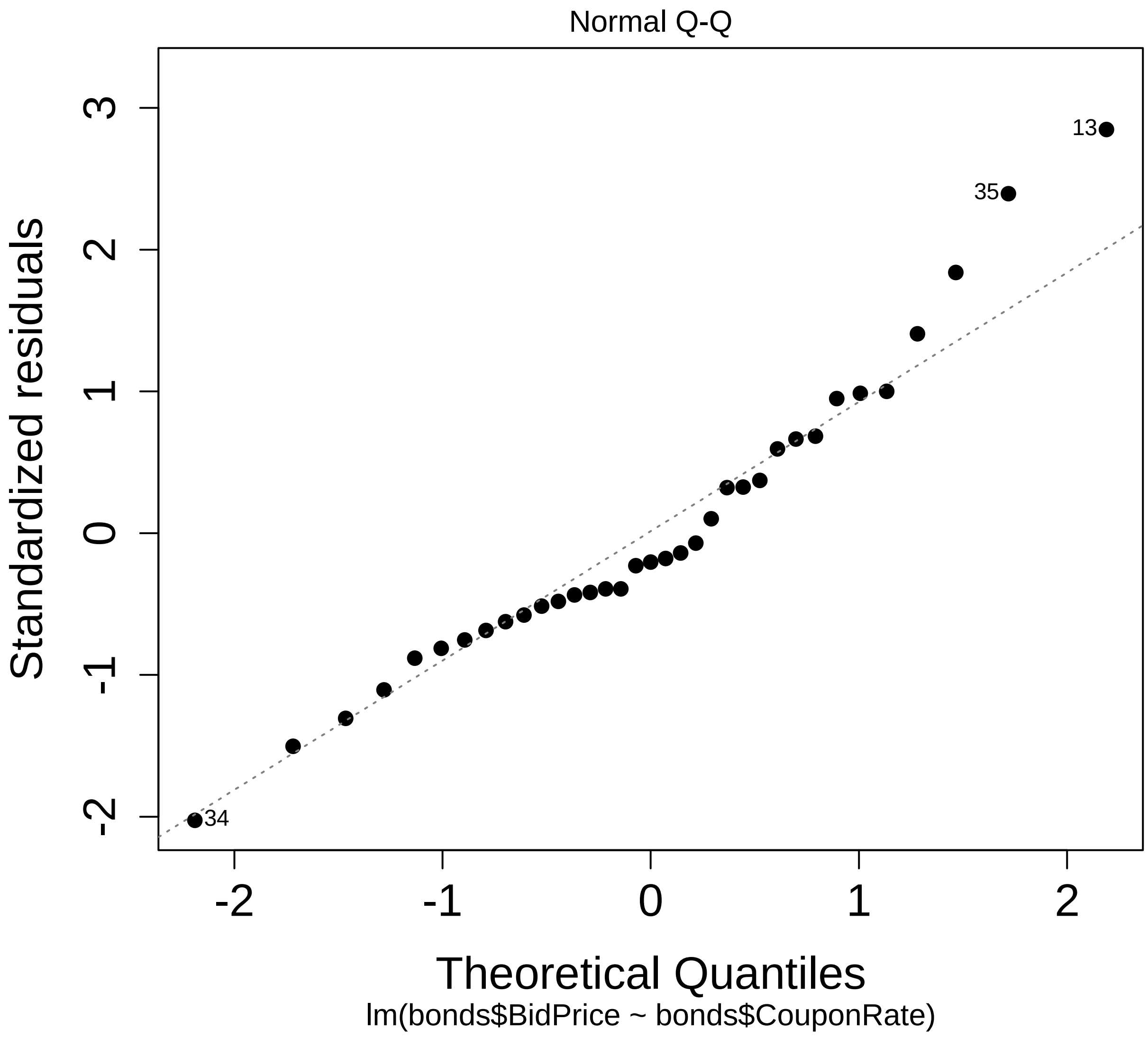
Necessary for:

- Small sample sizes: Hypothesis tests and confidence intervals for the regression parameters and regression line
- All sample sizes: Prediction intervals

To check:

- Q-Q plots of residuals (Warning: residuals appear normal though errors are not normal)
- Formal tests of normality (Anderson-Darling, but remember significance is not the same as practical importance)

Q-Q Plot: Bonds Data (default)



Constant Variance

Necessary for: Inference

Tools to assess constant variance (homoskedasticity):

- Plots
- Formal tests (e.g. White, 1980)

Tools to deal with non-constant variance (heteroskedasticity):

- Transformations
- Weighted least squares (Chapter 4)

Constant variance: Plots

An effective plot to diagnose nonconstant error variance is a plot of

$|\text{Residuals}|^{0.5}$ against x

or

$|\text{Standardized Residuals}|^{0.5}$ against x

(The square root is taken to reduce skewness in the absolute values.)

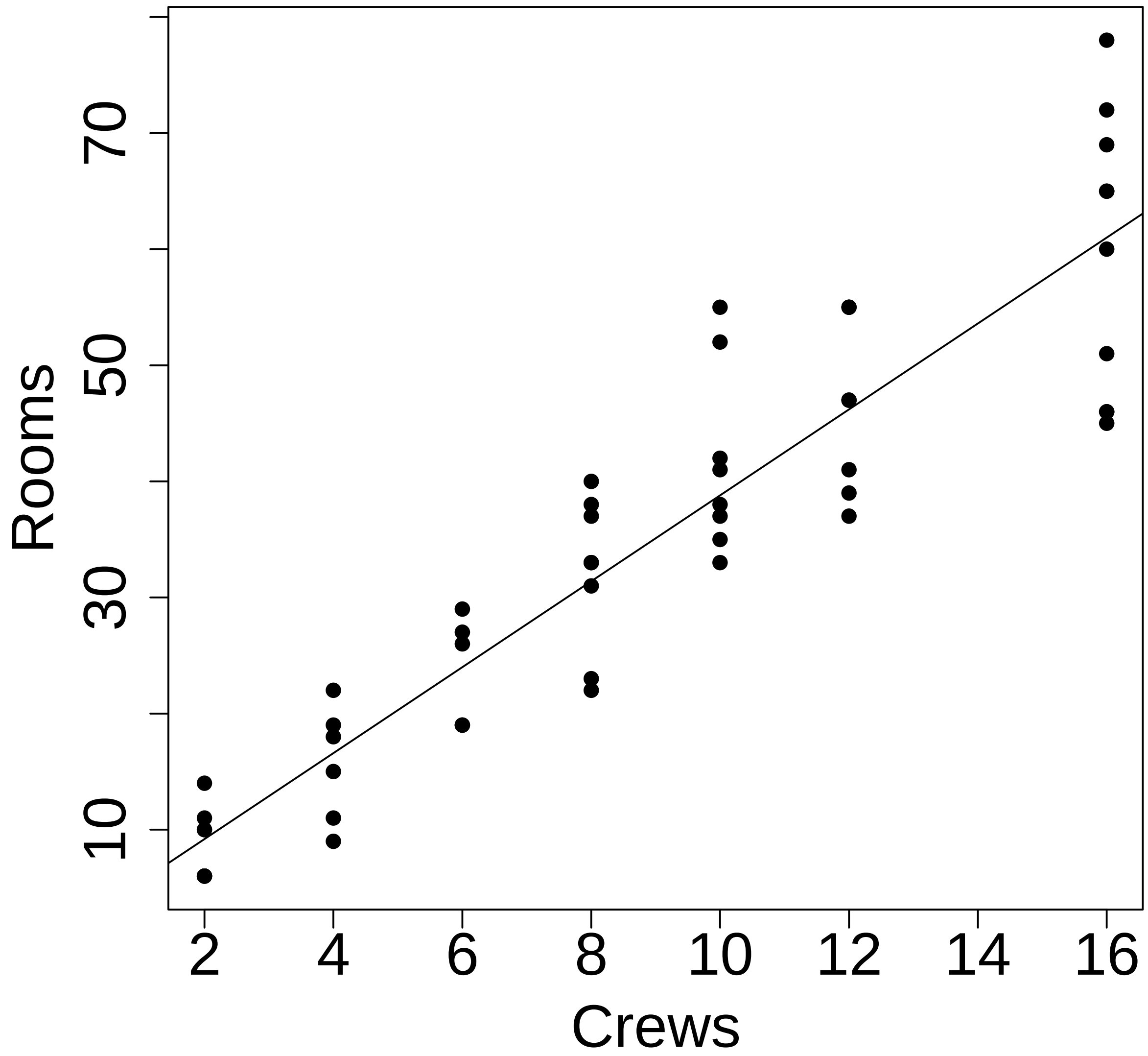
+

Constant Variance: Example

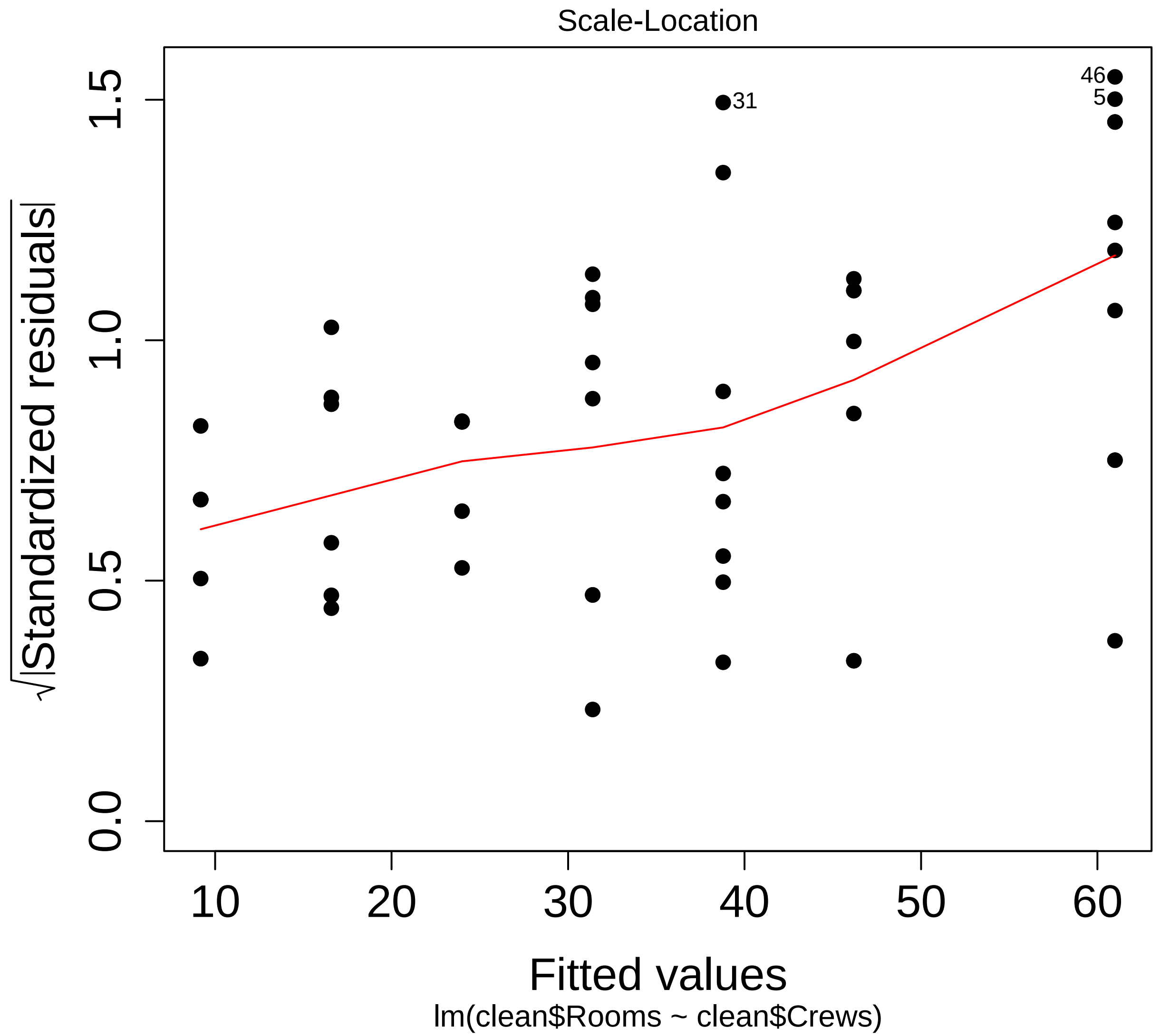
Developing a bid on contract cleaning

- Goal: Bid on a contract to clean offices.
- Costs proportional to # cleaning crews needed.
- 53 days of records: # crews & # rooms cleaned

Constant Variance: Example



Constant Variance: Example (default)

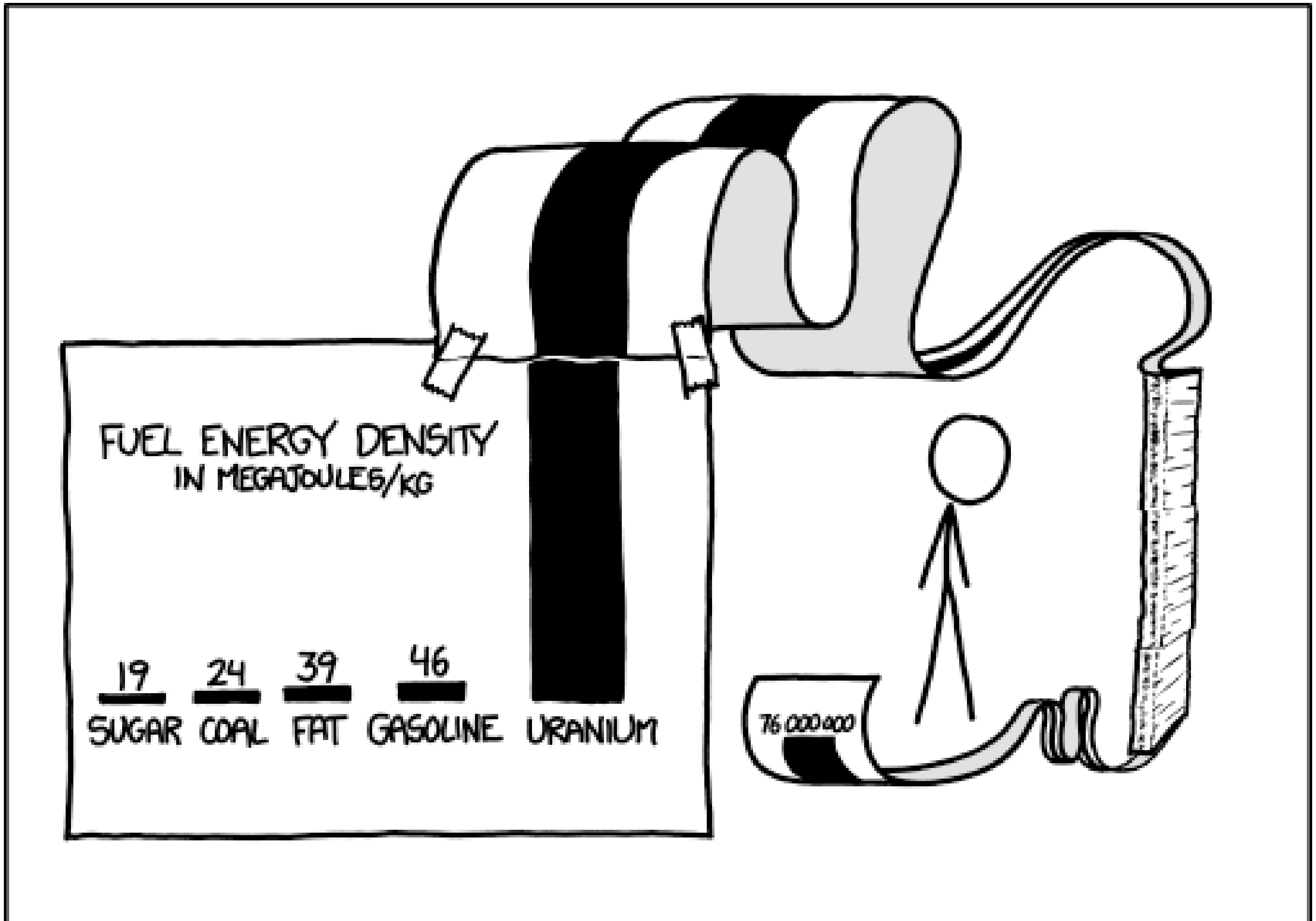


Transformations

Transformations can be used to:

- Overcome problems due to nonconstant variance
- Estimate percentage effects
- Overcome problems due to nonlinearity

Transformations



SCIENCE TIP: LOG SCALES ARE FOR QUITTERS WHO CAN'T FIND ENOUGH PAPER TO MAKE THEIR POINT *PROPERLY*.

Transformations: Count Data

Count data are often modeled using the Poisson distribution. Suppose that *Y* follows a Poisson distribution with mean λ , then it is well-known that the variance of *Y* is also equal to λ . In this case, the appropriate transformation of Y for stabilizing variance is square root.

Justification: Consider the following Taylor series expansion

$$f(Y) = f(E(Y)) + f'(E(Y))(Y - E(Y)) + \ldots$$

The well-known delta rule for calculating first-order variance terms is obtained by considering just the terms given in this last expansion. In particular, taking the variance of each side of this equation gives

$$\mathrm{Var}(f(Y)) \simeq \left[f'(\mathrm{E}(Y))\right]^2 \mathrm{Var}(Y).$$

Transformations: Count Data

$$\text{Suppose } f(Y) = \sqrt{Y} \text{ and } Var(Y) = \lambda = E[Y].$$

$$\text{Then : } Var(\sqrt{Y}) \approx \left\{ \frac{1}{2} (\lambda)^{-1/2} \right\}^2 \lambda = 1/4$$

- In our crew cleaning example, both # crews and # rooms cleaned are count data, so try the square root transformation on both variables.
- It’s natural to use the same transformation on both variables when they are measured in the same units.

Cleaning Data: Intervals

Table 3.7 Predictions and 95% prediction intervals for the number of rooms

x, Crews	Prediction	Lower limit	Upper limit
4 (transformed data)	16 =(4.003^2)	8 =(2.790^2)	27=(5.217^2)
4 (raw data)	17	2	32
16 (transformed data)	61 =(7.806^2)	43 =(6.582^2)	82 =(9.031^2)
16 (raw data)	61	46	76

- Note: When using confidence or prediction intervals after transforming, back-transform the endpoints to get an interval in the original units.

Using Logarithms to Estimate Percentage Effects

$$log(Y) = \beta_0 + \beta_1\, log(x) + e$$

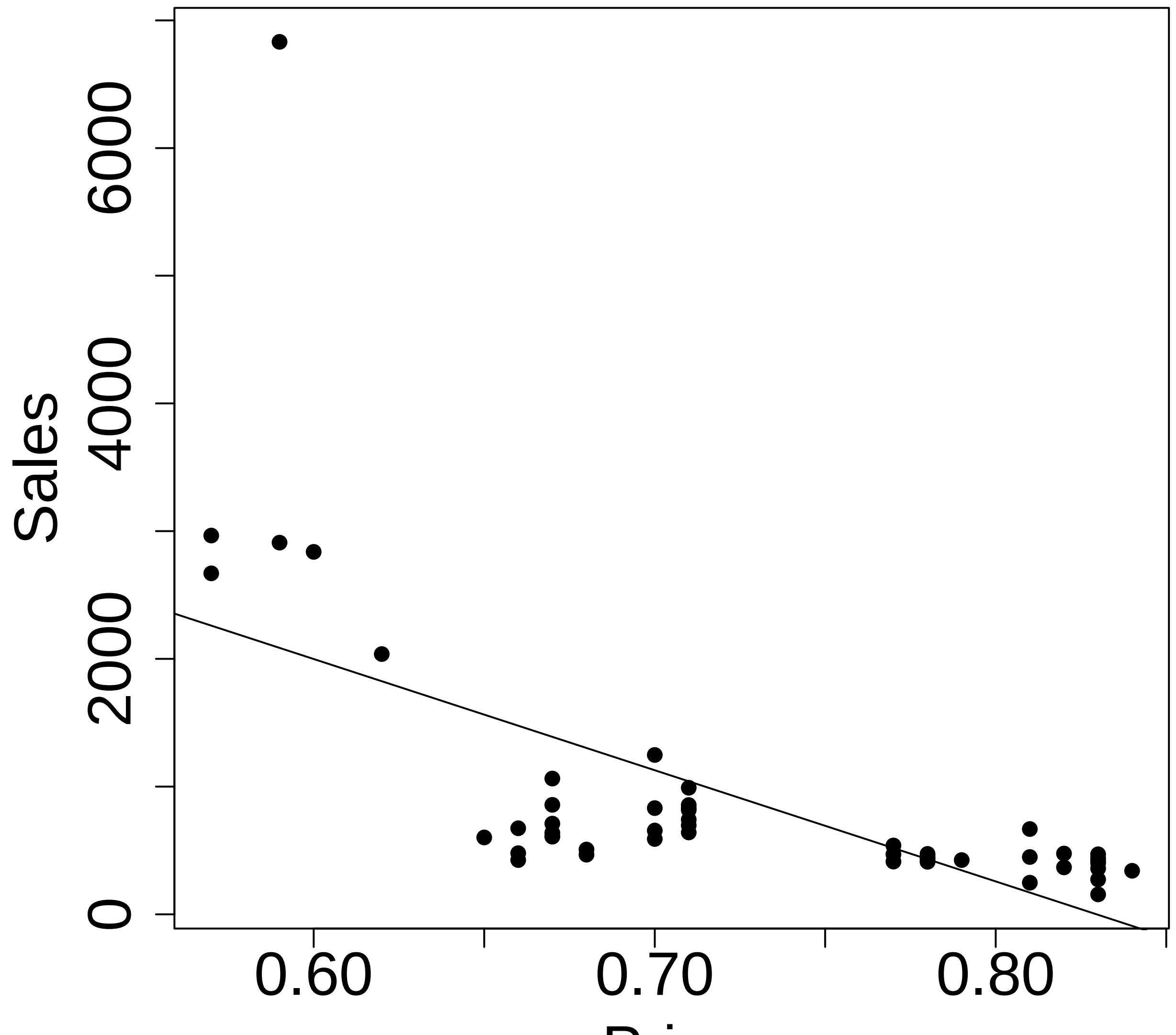
$$\begin{aligned}
\beta_1 &= \frac{\Delta\, log(Y)}{\Delta\, log(x)} \\
&= \frac{log(Y_2) - log(Y_1)}{log(x_2) - log(x_1)} \\
&= \frac{log(Y_2/Y_1)}{log(x_2/x_1)} \\
&\approx \frac{Y_2/Y_1 - 1}{x_2/x_1 - 1} (\text{using } log(1+z) \approx z \text{ and assuming } \beta_1 \text{ is small}) \\
&= \frac{100(Y_2/Y_1 - 1)}{100(x_2/x_1 - 1)} \\
&= \frac{\%\Delta Y}{\%\Delta x}
\end{aligned}$$

Business & Marketing

- Elasticity: If we increase supply by 1%, what happens to demand (in percentages)?
- A manager of Consolidated Foods, Inc. wants to know the relationship between price (P) and the quantity sold (Q).
- First we develop the model

$$Q = \beta_0 + \beta_1 P + e$$

Log Transformation Example



+ Business & Marketing

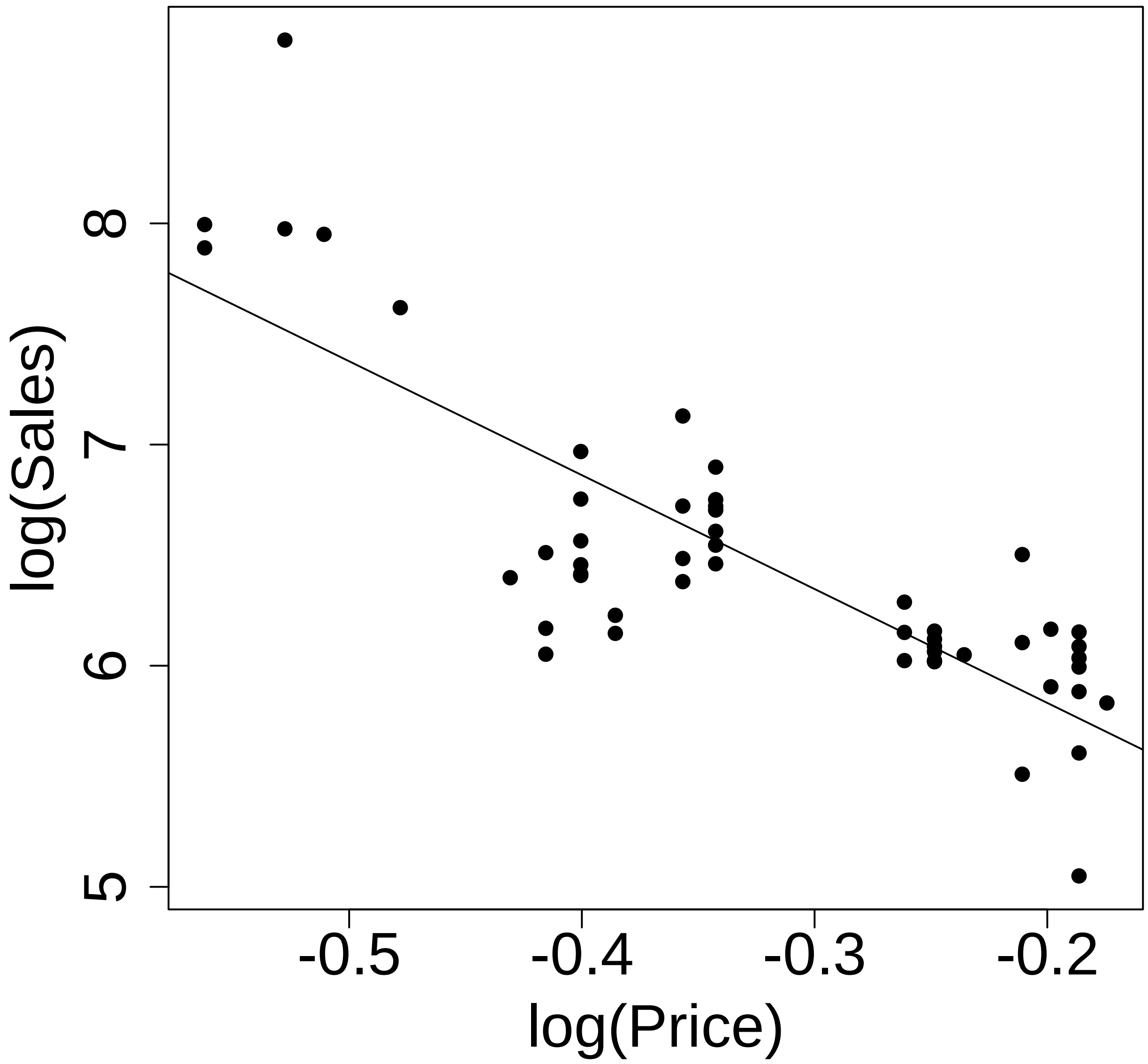
- To study the relationship between a 1% increase in price and the percent change on quantity sold, we take the logarithms of both variables.

- So we try out the model

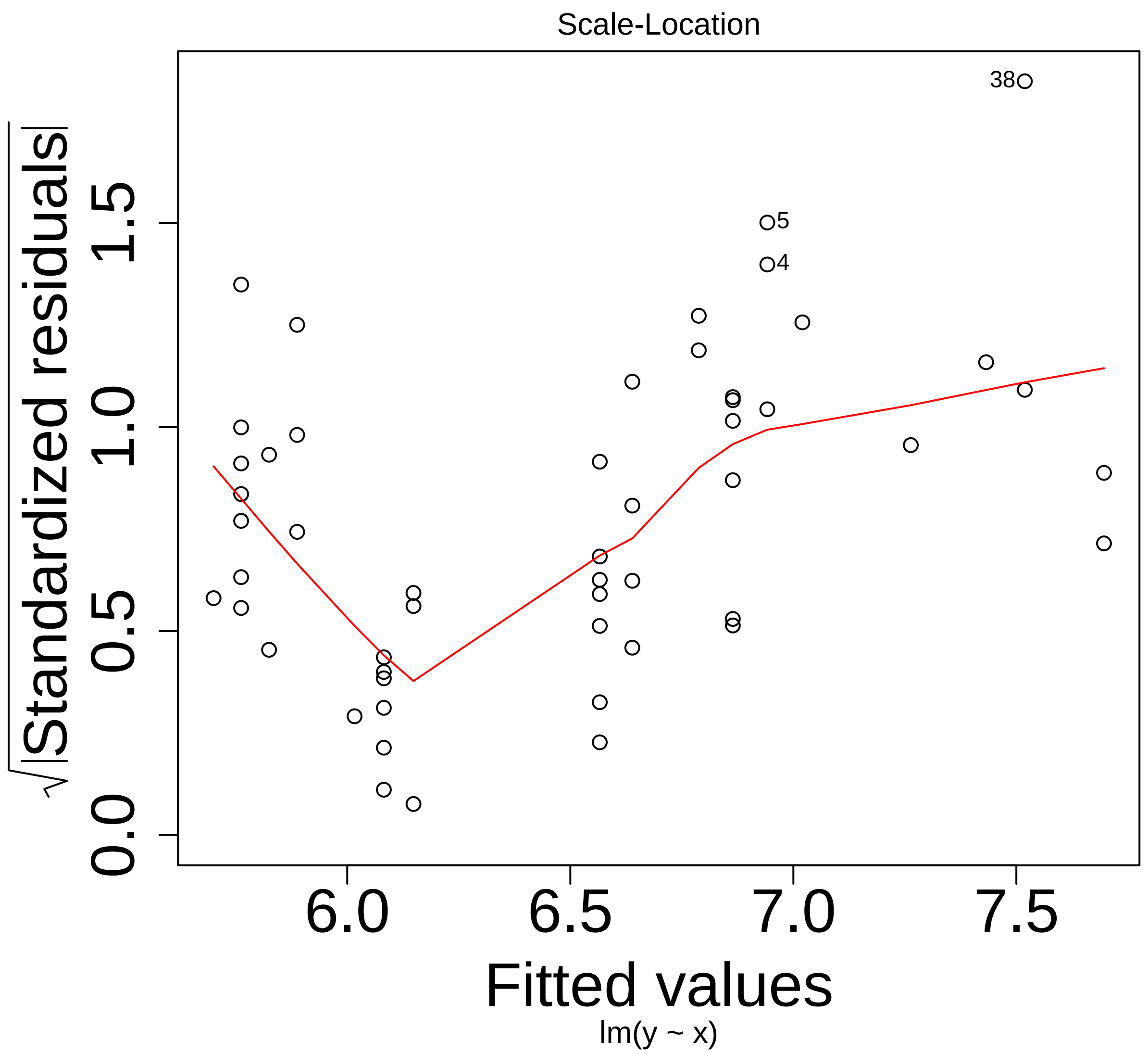
$$log(Q) = \beta_0 + \beta_1 \, log(P) + e$$

- The estimated slope is -5.14. Interpret in context:

Log Transformation Example



+ Log Transformation Example



Using Transformations to Overcome Problems due to Nonlinearity

We've discussed transformations to stabilize variance and estimate percentage effects. Next we attempt to overcome **nonlinearity in the relationship between x and y**. Two methods to do this are:

- Inverse response plots
- Box-Cox procedure

There are three main situations to be considered:

- Only the response variable needs to be transformed
- Only the predictor variable needs to be transformed
- Both need to be transformed.

Transforming only the response variable Y using inverse regression

Suppose that the true regression model between Y and X is:

$$Y = g(\beta_0 + \beta_1 x + e)$$

where *g* is some unknown function. We can turn this model into a simple linear regression model by transforming Y by g^{-1}, since:

$$g^{-1}(Y) = \beta_0 + \beta_1 x + e$$

For example, if:

$$Y = (\beta_0 + \beta_1 x + e)^3$$

Then the inverse transformation would be:

$$g(Y) = Y^3, \text{ and so } g^{-1}(Y) = Y^{1/3}$$

Transforming Y

Randomly generated example:

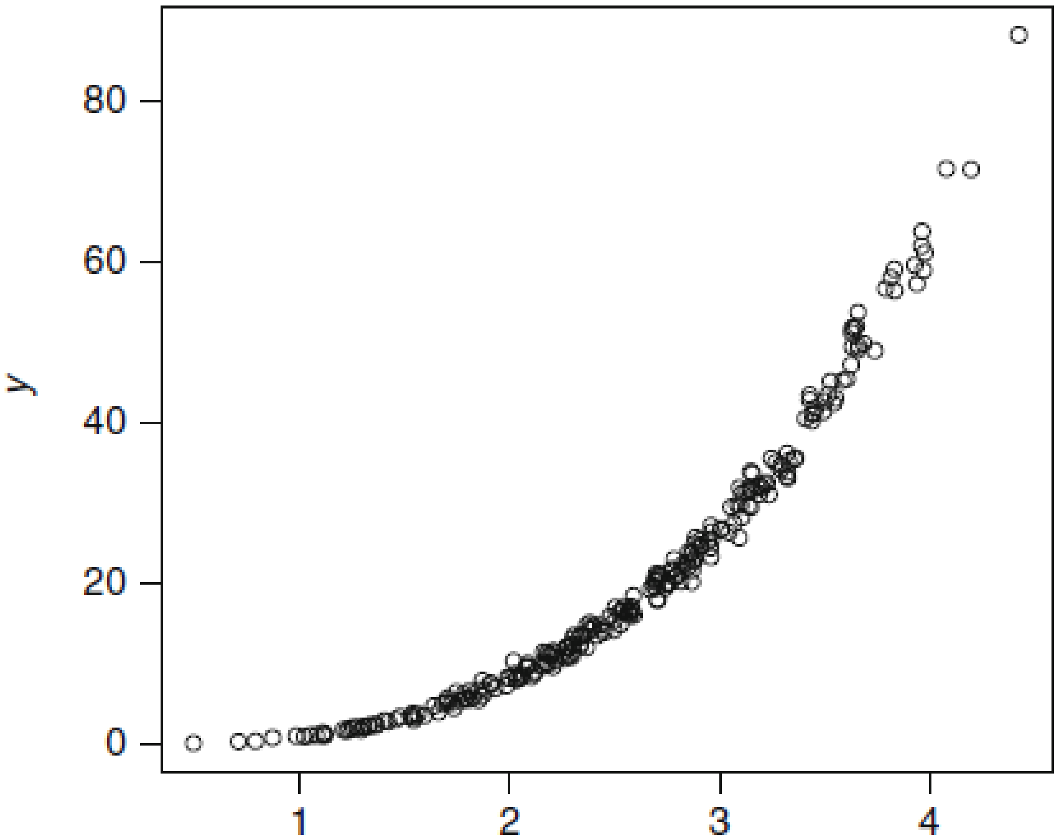
$$Y = (\beta_0 + \beta_1 x + e)^3$$

x and e both normally distributed.

Goal: figure out g(Y). Let's start with fitting a simple linear regression model and see how terrible it is.

$$Y = \beta_0 + \beta_1 x + e$$

Transforming Y



+ Transforming Y

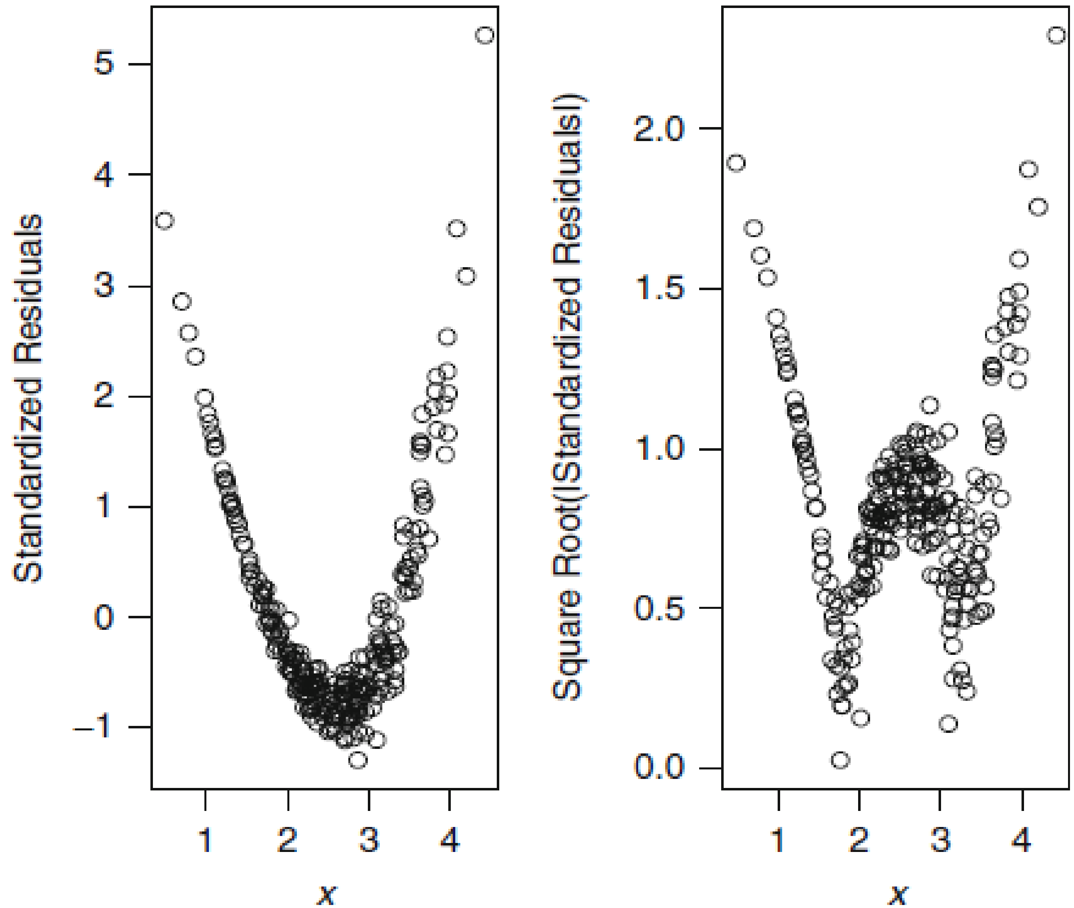


Figure 3.26 Diagnostic plots for model (3.2)

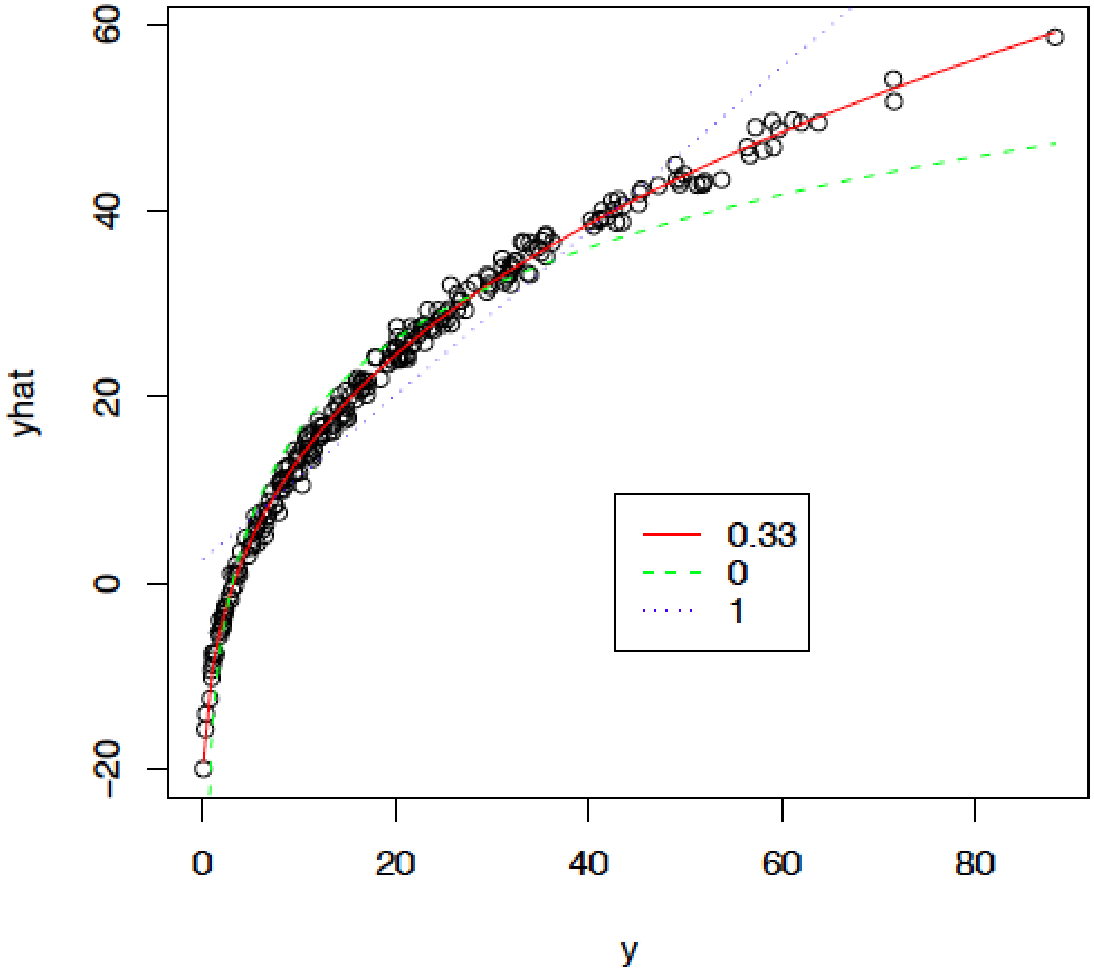
Transforming Y: Inverse Response Plots

Recall that:

$$Y = g(\beta_0 + \beta_1 x + e)$$
$$g^{-1}(Y) = \beta_0 + \beta_1 x + e$$

So if we plot Y on the x-axis and $\beta_0 + \beta_1 x$ on the Y- axis, we should discover the shape of g^{-1}.

Inverse Response Plot



+

Transforming Y: Box Cox

To estimate g^{-1} in general we consider the following family of *scaled power transformations*, defined for strictly positive Y by

$$\Psi_S(Y,\lambda) = \begin{cases} (Y^{\lambda} - 1)/\lambda \text{ if } \lambda \neq 0 \\ \log(Y) \quad \text{if } \lambda = 0 \end{cases}$$

Scaled power transformations have the following three properties:

1. $\Psi_S(Y,\lambda)$ is a continuous function of λ
2. The logarithmic transformation is a member of this family, since

$$\lim_{\lambda \to 0} \Psi_S(Y,\lambda) = \lim_{\lambda \to 0} \frac{\left(Y^{\lambda} - 1\right)}{\lambda} = \log(Y)$$

3. Scaled transformations preserve the direction of the association between Y and X in the sense that if Y and X are positively (negatively) related than $\Psi_S(Y,\lambda)$ and X are positively (negatively) related for all values of λ

Thus, to estimate g^{-1}, we consider fitting models of the form

$$\mathrm{E}(\hat{y} \mid Y = y) = \alpha_0 + \alpha_1 \Psi_S(y,\lambda) \qquad (3.3)$$

Transforming Y: Box Cox

$$E[\hat{y}|Y = y] = \alpha_0 + \alpha_1 \psi(y, \lambda)$$

The explanatory variable in the model above is the transformed y, and the response variable is $\hat{y}$ from the simple linear regression model.

We fit the model above for a range of values λ and choose an estimated optimal value of λ that minimizes the residual sum of squares RSS(λ).

Usually choosing a λ from {-1, -1/2, -1/3, -1/4 0, 1/4, 1/3, 1/2, 1} is adequate.

Transforming Y: Box-Cox

- Box-Cox assumes there exists a power transformation g(Y) such that the transformed version of Y is **normally distributed**.
- If X and Y are normally distributed, the relationship between them will be ______________.
- We're not as concerned with the distribution of the residuals as we are with the relationship between x and y.

Transforming Y: Box-Cox

Box and Cox multiply the function $\psi_S(x, \lambda)$ by a factor to get $\psi_M(x, \lambda)$, the goal being to keep all units the same for all values of λ. The method is based on the maximum likelihood estimates for Y. Maximizing the likelihood with respect to λ is equivalent to minimizing RSS(λ) with respect to λ.

$$RSS(\lambda) = \sum_{i=1}^{n} (\psi_M(y_i, \lambda) - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

This time, the response variable is the transformed y and the explanatory variable is x.

Transformations: Caution

- Transformations don't perform well in every situation.
- X may not explain Y very well, no matter what transformation is used.
  - Ex. Important predictors not included.
- Box-Cox might result in a transformed variable that is not very close to normally distributed.
  - Ex. heavy tails on both sides.

Transforming X

- Try scaled power transformations, as for Y.

$$\psi_S(X,\lambda) = \begin{cases} (X^\lambda - 1)/\lambda & \text{if } \lambda \neq 0 \\ log(X) & \text{if } \lambda = 0 \end{cases}$$

$$E[Y|X = x] = \alpha_0 + \alpha_1 \psi_S(x,\lambda)$$

Choose the value of λ that minimizes the residual sum of squares from fitting the above model.

- The Box-Cox method can also be directly applied to X.

+

Transforming both the response and the predictor variables

Approach 1:

1. Transform X to $\psi(x, \lambda)$ so that the distribution of the transformed version of X is as normal as possible. Univariate Box-Cox is one way to do this.
2. Consider a simple linear regression model of the form $Y = g(\beta_0 + \beta_1 \psi(x, \lambda) + e)$. Use an inverse response plot to decide on the transformation g^{-1} for Y.

Approach 2:

Transform X and Y simultaneously to joint normality using multivariate Box-Cox.

Many times, the two approaches lead to the same answer. Approach 1 is more robust, and may give reasonable answers when approach 2 fails.

Transformations: Confidence and Prediction Intervals

Suppose we are interested in the middle 95% of a distribution (as in a confidence or prediction interval.

What happens if we shift the interval to the left or right?

Transformations: Confidence and Prediction Intervals

If the relationship between Y and some transformation g(Y) were linear, then it would be true that E[g(Y)] = g(E[Y]). But if the relationship between X and g(Y) were linear, and the relationship between g(Y) and Y were linear, then the relationship between X and Y would be linear.

$$
\begin{aligned}
E[X] &= \int g(x) f(x) dx = \int (ax+b) f(x) dx \\
&= \int ax\, f(x) dx + \int b\, f(x) dx \\
&= a \int x\, f(x) dx + b \int f(x) dx \\
&= a\, E[X] + b
\end{aligned}
$$

Moral: g(Y) is probably non-linear. Linear transformations are mostly useless.

Review

$$
\begin{aligned}
Var(X) &= E[(X-\mu)^2] \\
&= E[X^2 - 2X\mu + \mu^2] \\
&= E[X^2] - 2\mu\, E[X] + \mu^2 \\
&= E[X^2] - 2\mu^2 + \mu^2 \\
&= E[X^2] - \mu^2
\end{aligned}
$$

$$E[X^2] = \mu^2 + Var(X)$$

Transformations: Confidence and Prediction Intervals

We sometimes want to be able to find confidence intervals and prediction intervals for the least squares regression line in models with transformed X and Y.

Remember, we're modeling the mean of Y using x:

$$\mu_Y = \beta_0 + \beta_1 x$$

But in order to get a linear relationship between X and Y, we had to transform Y, so we had:

$$g(Y) = W, \mu_W = \beta_0 + \beta_1 x$$

If we also found that g(Y) was normally distributed, then to go back and find the mean of Y at each value of x*, we would need:

$$E[g^{-1}(W)]$$

+ Example: Square Root Transformation

Square root transformation to stabilize variance (e.g. Cleaning. Everything is conditional on X):

$$g(Y) = \sqrt{Y} = W$$

$$\begin{aligned} E[g^{-1}(Y)] &= E[W^2] \\ &= E[W]^2 + Var(W) \\ &= \mu_W^2 + \sigma_W^2 \\ &= (\beta_0 + \beta_1 x)^2 + \sigma_W^2 \end{aligned}$$

The variance of W, given X, is the variance of the errors. So we would estimate the confidence or prediction interval (if W were normal) by transforming the endpoints of the usual confidence or prediction interval using W plus MSE.

+ Example: Square Root Transformation

Square root transformation to stabilize variance (Everything is conditional on X):

$$E[g^{-1}(Y)] = (\beta_0 + \beta_1 x)^2 + \sigma_W^2$$

- The correction to the confidence interval endpoints, then, is $(\text{endpoint})^2$ + RMSE.
- Adding the correction factor, the variance of the errors, is less important when it is small.
- In general, if we don't add the correction factor, simply transforming the endpoints leads to terribly biased intervals that do not contain the population mean the advertised percentage of the time.

Example: Log Transformation

Suppose now that log(Y) is normally distributed. Then Y has the log normal distributed.

The mean of the lognormal distribution is:

$$E[Y] = e^{\mu+\sigma^2/2}$$

So the correction is exp(endpoint + RMSE/2).

+ Example: Inverse Transformation

Suppose we take W = g(Y) = 1/Y, which is normally distributed. Then the inverse transformation is also $g^{-1}(W) = 1/W$. If W is normally distributed, finding the distribution of 1/W becomes more complicated; let's try a Taylor series expansion:

$$g^{-1}(W) \approx \frac{1}{\mu} + (W-\mu)\frac{-1}{\mu^2} + \frac{1}{2}(W-\mu)^2\frac{2}{\mu^3}$$

$$\begin{aligned} E[g^{-1}(W)] &\approx \frac{1}{\mu_W} + 0 + \sigma_W^2 \frac{1}{\mu_W^3} \\ &= \frac{1}{\mu_W}\left(1 + \frac{\sigma_W^2}{\mu_W^2}\right) \end{aligned}$$

Back Transformation Adjustments

Transformation	Transformed Model	Back Transformation with Adjustment
Logarithmic	$log(Y) = \beta_0 + \beta_1 X + e$	$\hat{E}[Y] = exp(\hat{\beta}_0 + \hat{\beta}_1 X + \frac{\hat{\sigma}^2}{2})$
Square Root	$\sqrt{Y} = \beta_0 + \beta_1 X + e$	$\hat{E}[Y] = (\hat{\beta}_0 + \hat{\beta}_1 X)^2 + \hat{\sigma}^2$
Inverse	$\frac{1}{Y} = \beta_0 + \beta_1 X + e$	$\hat{E}[Y] = \frac{1}{\hat{\beta}_0 + \hat{\beta}_1 X}\left(1 + \frac{\hat{\sigma}^2}{(\hat{\beta}_0 + \hat{\beta}_1 X)^2}\right)$
Inverse Square Root	$\frac{1}{\sqrt{Y}} = \beta_0 + \beta_1 X + e$	$\hat{E}[Y] = \frac{1}{(\hat{\beta}_0 + \hat{\beta}_1 X)^2 + \hat{\sigma}^2}\left(1 + \frac{2\hat{\sigma}^4 + 4(\hat{\beta}_0 + \hat{\beta}_1 X)^2\hat{\sigma}^2}{\left[(\hat{\beta}_0 + \hat{\beta}_1 X)^2 + \hat{\sigma}^2\right]^2}\right)$

Back Transformation Adjustments

What if we want to use some other transformation?

Bootstrap.

High R^2 does not imply valid model

- R^2 (the square of the correlation) is often defined as the proportion of the variability in the random variable Y explained by the regression model.

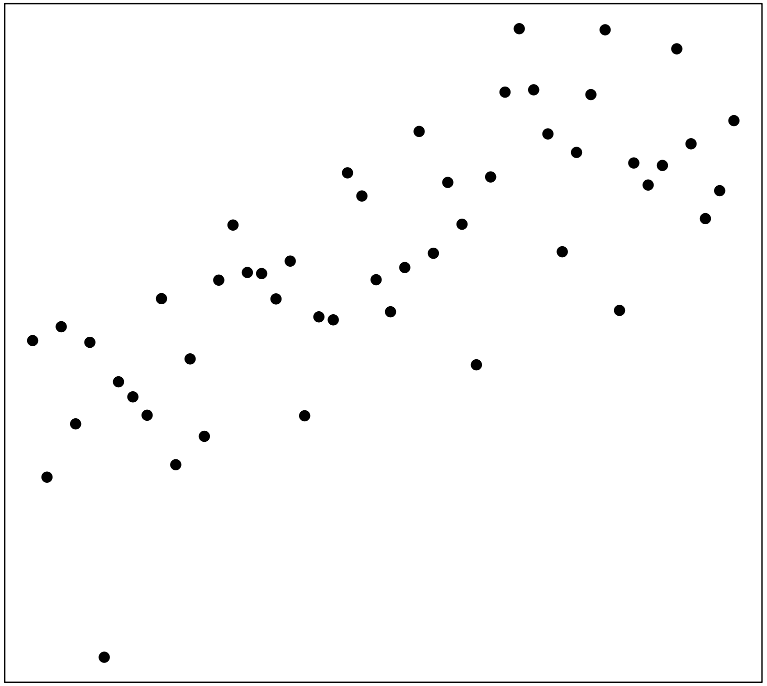
$$SSreg + RSS = SST$$

$$R^2 = \frac{SSreg}{SST} = 1 - \frac{RSS}{SST}$$

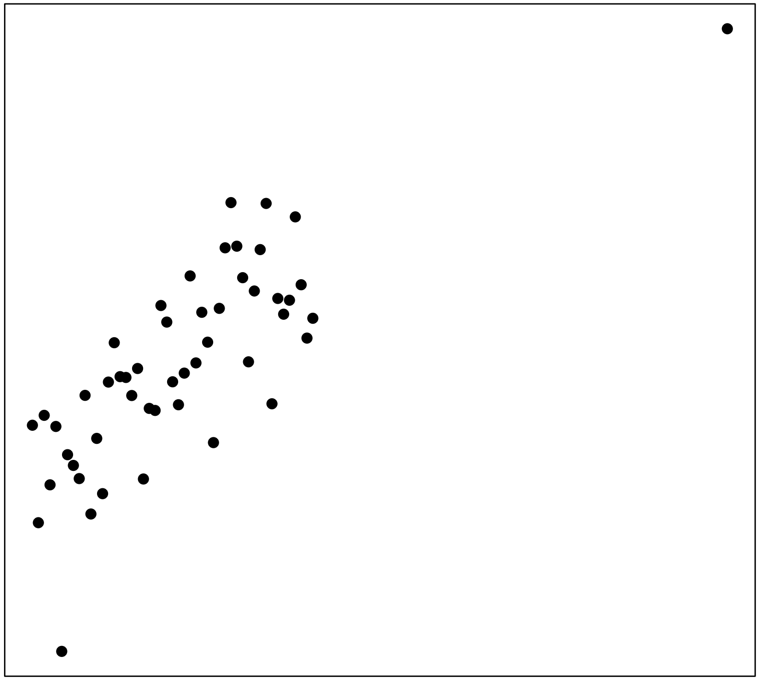
R^2: Outliers

- R^2 is not robust to outliers.

r = 0.75



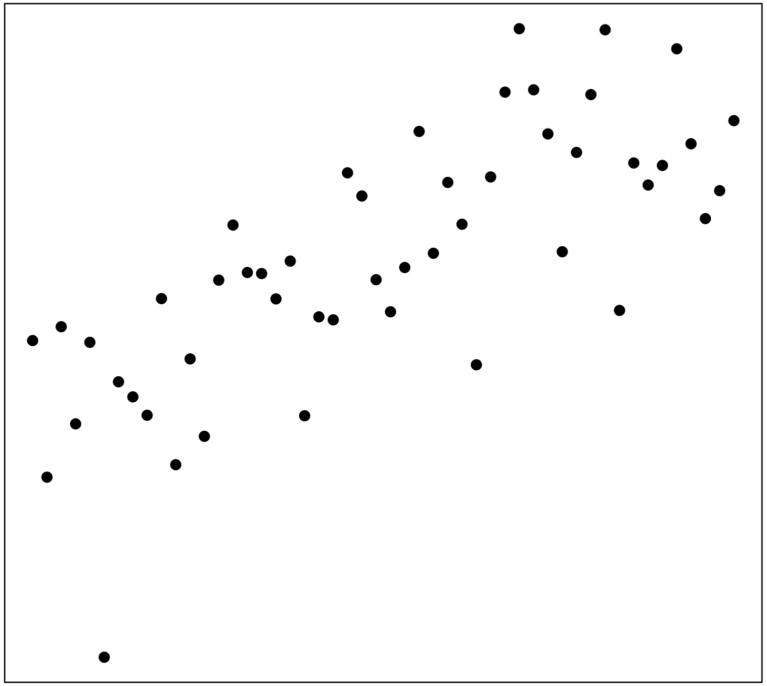
r = 0.80



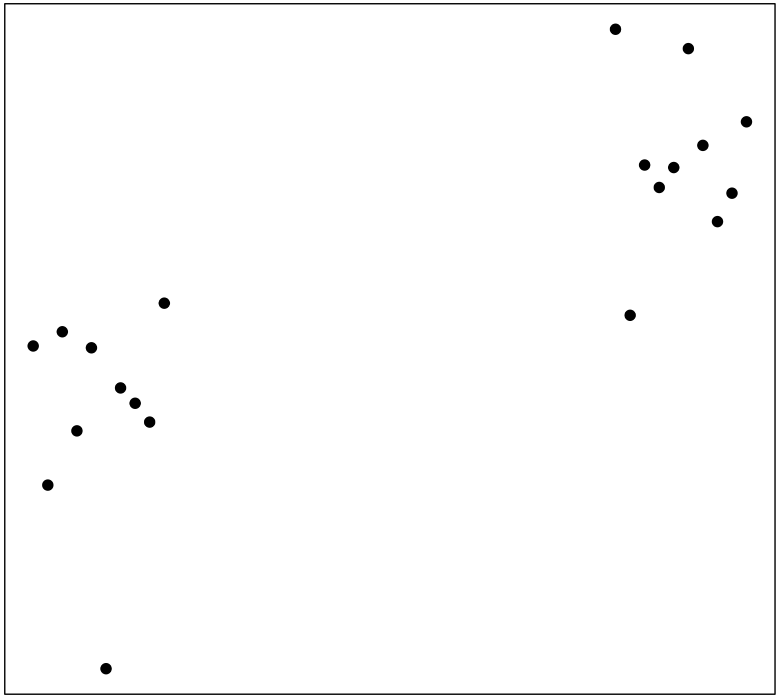
R^2: Stratification

- R^2 takes different values when different values of X are observed.

r = 0.75



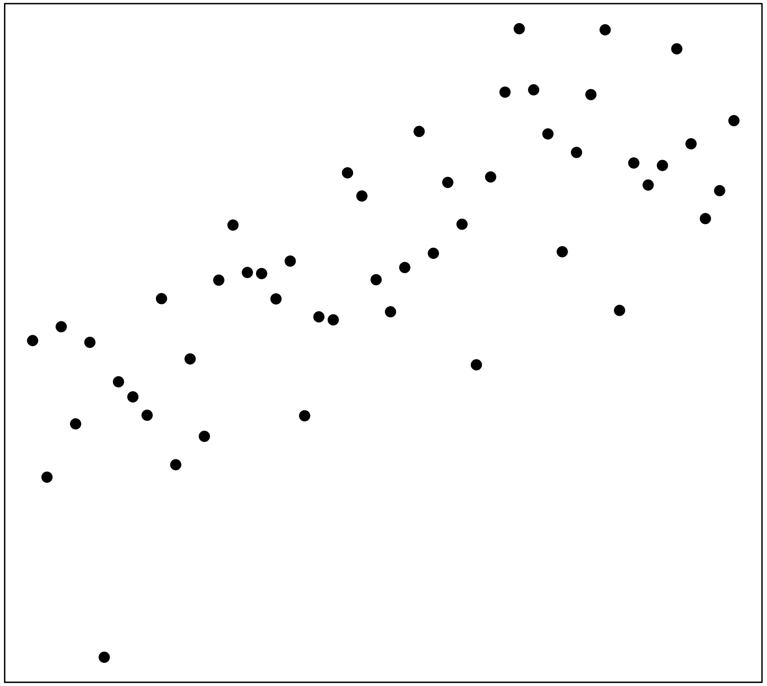
r = 0.81



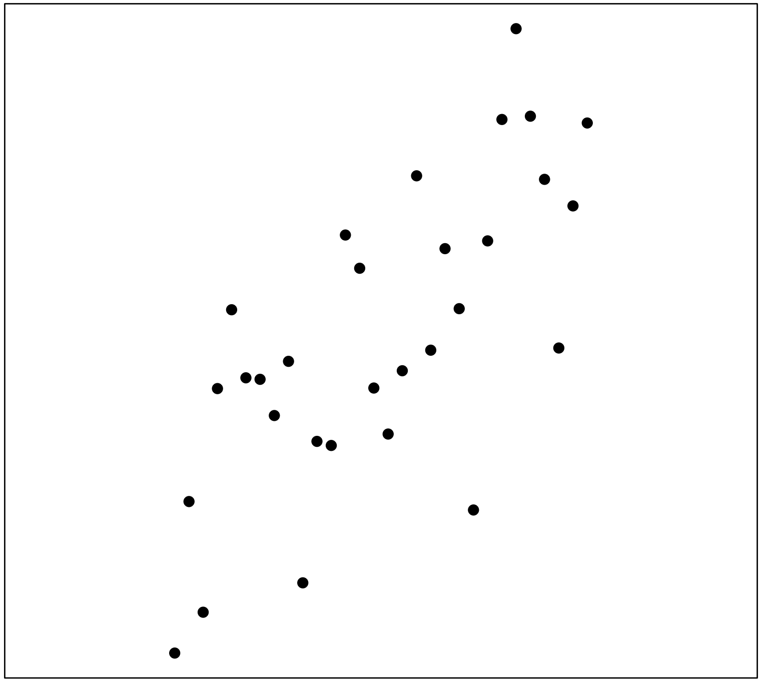
R^2: Stratification

- R^2 takes different values when different values of X are observed.

r = 0.75



r = 0.72



+

R^2: Regression through the origin

- R^2 is not invariant under location changes like Fahrenheit to Celsius when we fit a line through the origin (no intercept).

$$R^2 = 1 - \frac{SSreg}{\sum y_i^2}$$

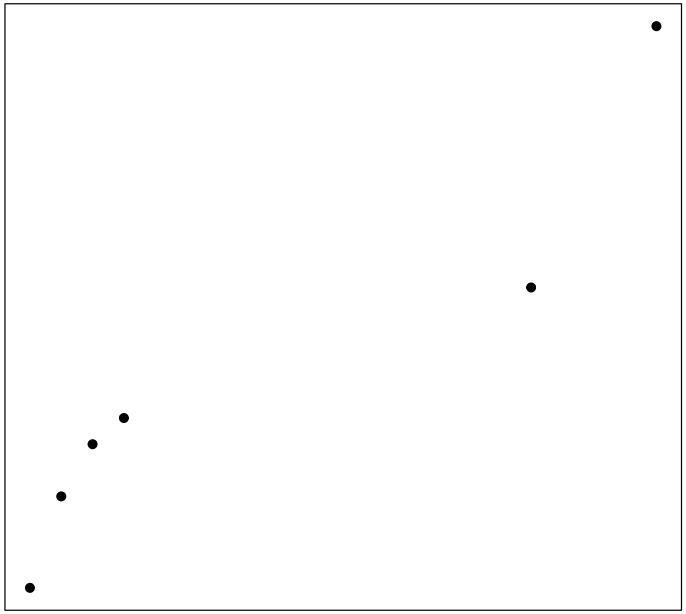
Uncorrected SS

R^2: Transformations

- R^2 is not invariant under transformations
- Example 1:

  (0, 0.5), (1, 4), (2, 6), (3, 7), (16, 12), (20, 22)

  SLR has $R^2 = 0.88$; using log(Y), $R^2 = 0.56$, but the fit is approximately equivalent



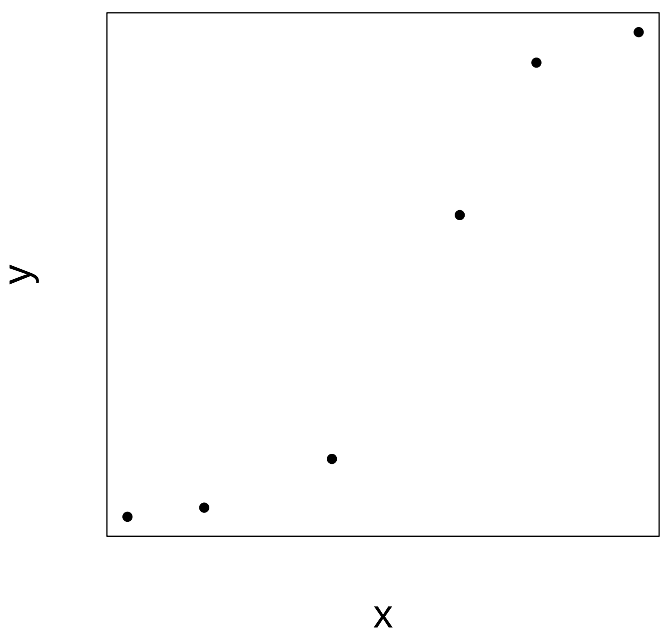
Scott & Wild, 1991. "Transformations and R^2."

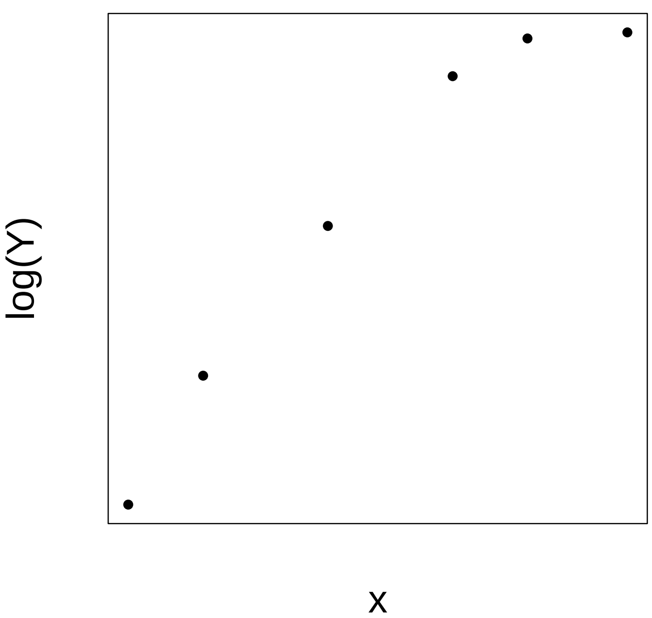
R^2: Transformations

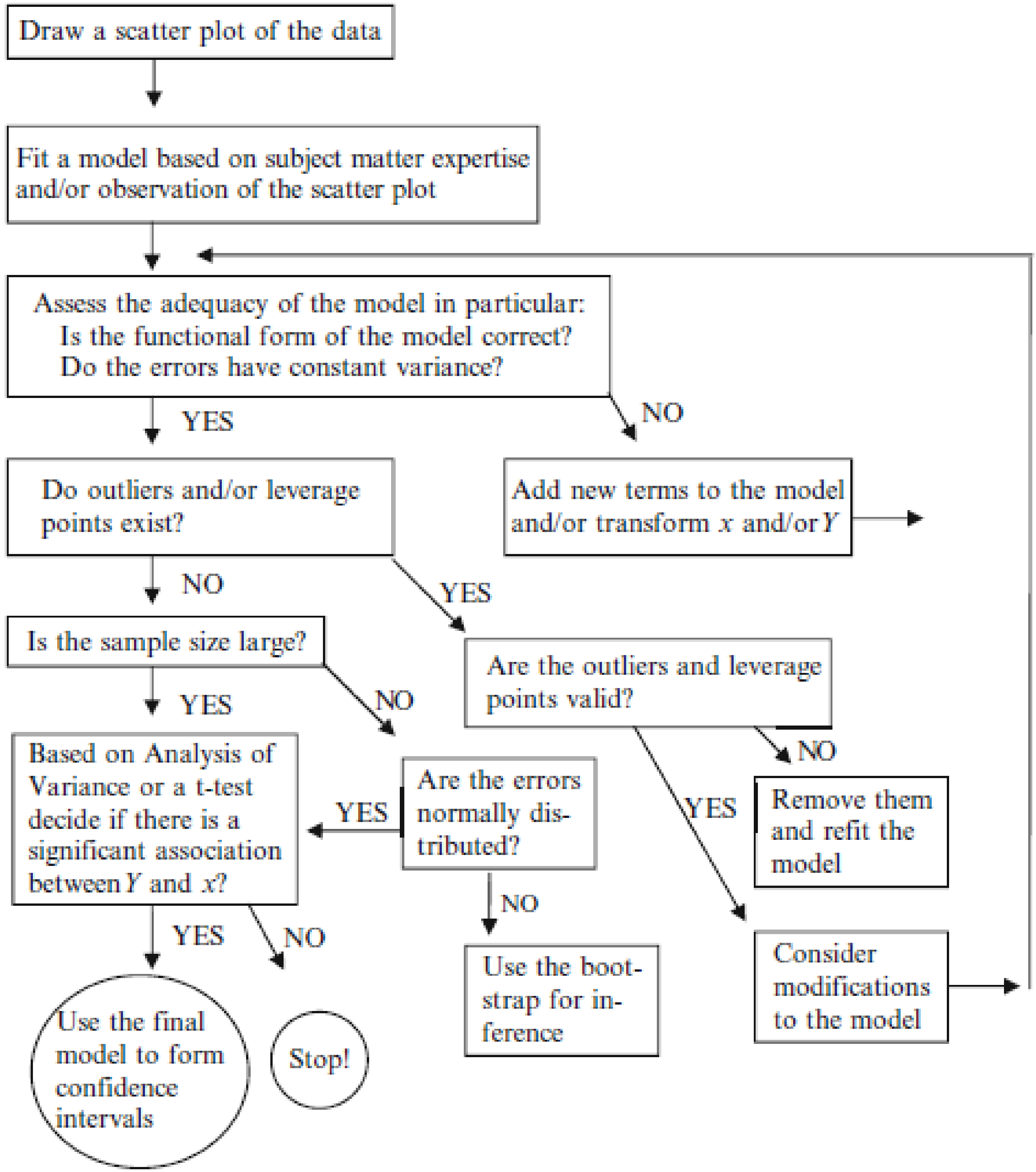
- R^2 is not invariant under transformations
- Example 2:

(0, 0.1), (3, 0.4), (8, 2), (13, 10), (16, 15), (20, 16)

SLR has $R^2 = 0.92$; using log(Y), $R^2 = 0.94$, but the second model is arguably poorer.






Draw a scatter plot of the data
Fit a model based on subject matter expertise and/or observation of the scatter plot
Assess the adequacy of the model in particular: Is the functional form of the model correct? Do the errors have constant variance?
YES
NO
Do outliers and/or leverage points exist?
Add new terms to the model and/or transform x and/or Y
NO
YES
Is the sample size large?
Are the outliers and leverage points valid?
YES
NO
NO
Based on Analysis of Variance or a t-test decide if there is a significant association between Y and x?
YES
Are the errors normally distributed?
YES
Remove them and refit the model
NO
YES
NO
Use the bootstrap for inference
Consider modifications to the model
Use the final model to form confidence intervals
Stop!