Lecture 4

Diagnostics and transformations

Aim: to understand when the SLR model is a valid model for the data

1. Anscombe's four data sets
2. Regression diagnostics
3. Transformations

Last week we learned

Confidence intervals on the slope and intercept

Estimating mean response

Prediction of a new observations

Testing significance of regression

Coefficient of determination R^2

The slope and the intercept

- A test statistic to test the null hypothesis $H_0 : \beta_i = \beta_i^*$ is

$$T = \frac{\hat{\beta}_i - \beta_i^*}{\text{se}(\hat{\beta}_i)} \sim t_{n-2}$$

  We reject H_0 if $t_{cal} = |t| > t_{\alpha/2,n-2} = t_{crit}$.

- A $100(1-\alpha)\%$ CI on β_i is

$$\text{CI}(\beta_i) = \left[\hat{\beta}_i - t_{\alpha/2,\, n-2} \cdot \text{se}(\hat{\beta}_i),\ \hat{\beta}_i + t_{\alpha/2,\, n-2} \cdot \text{se}(\hat{\beta}_i)\right]$$

- The estimated standard errors of β_1 and β_0 are

$$\text{se}(\hat{\beta}_1) = \sqrt{\hat{\sigma}^2 / s_{xx}} \qquad \text{se}(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2\left(1/n + \bar{x}^2/s_{xx}\right)}$$

  where

$$\hat{\sigma}^2 \equiv MS_E = \frac{SS_E}{n-2} = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

  is an estimate of σ^2.

Mean Responses and New Observations

- The LSE of the mean response μ_0 at $X = x_0$

$$\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 \sim N\left(\mu_0, \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}}\right)\sigma^2\right)$$

- A $100(1-\alpha)\%$ CI on μ_0 is

$$\text{CI}(\mu_0) = \left[\hat{\mu}_0 - t_{\alpha/2,\, n-2} \cdot \text{se}(\hat{\mu}_0),\ \hat{\mu}_0 + t_{\alpha/2,\, n-2} \cdot \text{se}(\hat{\mu}_0)\right]$$

- The LSE of a new observation y_0 at $X = x_0$

$$\hat{Y}_0 = \hat{\mu}_0 + \varepsilon_0 \sim N\left(\mu_0, \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{xx}} + 1\right)\sigma^2\right)$$

- A $100(1-\alpha)\%$ PI on y_0 is

$$\text{PI}(y_0) = \left[\hat{y}_0 - t_{\alpha/2,\, n-2} \cdot \text{se}(\hat{y}_0),\ \hat{y}_0 + t_{\alpha/2,\, n-2} \cdot \text{se}(\hat{y}_0)\right]$$

Significance of Regression

- Analysis of Variance Identity

  $$SS_T = SS_R + SS_E$$

  where

  $$SS_T = \sum_{i=1}^{n}(y_i - \bar{y})^2 \quad SS_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \quad SS_E = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

  – SS_T measures the total variation in y around its mean $\bar{y}$.

  – SS_R measures the total variation in $\hat{y}$ around the mean $\bar{y}$.

  – SS_E measures how closely model fits the data.

- Statistic for testing significance of regression, $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$,

  $$F = \frac{MS_R}{MS_E} = \frac{SS_R/\nu_R}{SS_E/\nu_E} \sim F_{1,n-2}$$

- Coefficient of determination

  $$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T} \in [0,1]$$

Analysis of Variance Table

- Analysis of variance table

Source of variation	d.o.f.	SS	MS	F
Regression	$\nu_R = 1$	SS_R	$MS_R = \frac{SS_R}{\nu_R}$	$F = \frac{MS_R}{MS_E}$
Residual (Error)	$\nu_E = n-2$	SS_E	$MS_E = \frac{SS_E}{\nu_E}$	
Total	$\nu_T = n-1$	SS_T		

- We reject the null hypothesis $H_0 : \beta_1 = 0$ if $F_{cal} > F_{crit} = F_{\alpha,1,n-2}$.

Lecture 4

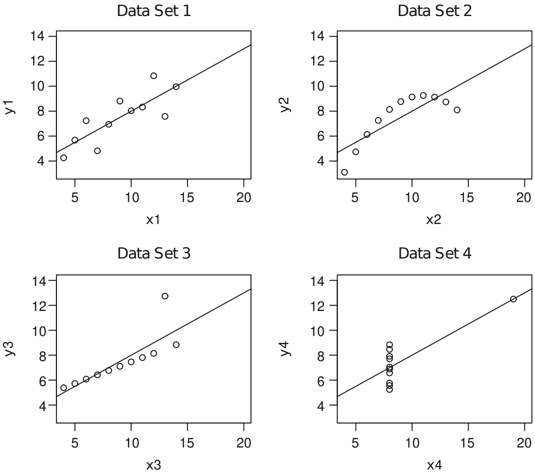
Diagnostics and transformations

Aim: to understand when the SLR model is a valid model for the data

1. Anscombe's four data sets
2. Regression diagnostics
3. Transformations

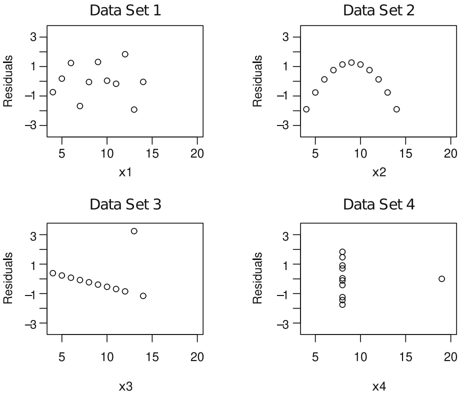
Anscombe's four data sets

- Statistician Francis Anscombe in 1973 created four data sets to illustrate the point that looking only at the numerical regression output may lead to very misleading conclusions about the data, thus lead to adopting a wrong model.



Anscombe's four data sets

- Residual plots can reveal if the regression model is valid or invalid:
  - Set 1: the SLR model is a valid model.
  - Set 2: a quadratic pattern indicates that a quadratic regression model, $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$, might suit the data better.
  - Set 3: residuals indicate an outlier.
  - Set 4: residuals indicate a leverage point.



Lecture 4

Diagnostics and transformations

1. Anscombe's four data sets
2. Regression diagnostics
3. Transformations

Regression diagnostics

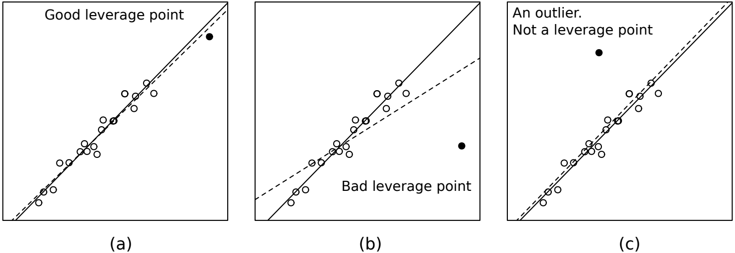
- Examine “crude” residuals and standardised residuals to determine if the proposed regression model is a valid model.
- Find leverage cases – data points with extreme X-values.
- Find outliers – data points that do not follow the general pattern.
- Examine whether the assumption of constant variance of the errors is reasonable.
- Adjust the model and repeat the diagnostics.

Leverage points

- A leverage point is a point whose X-value is distant from the other X-values.
- We will describe leverage points in a simplistic manner as “good” or “bad”:

  – A leverage point is a bad leverage point if its Y-value does not follow the pattern set by the other data points, i.e. a bad leverage point is a leverage point which is also an outlier.

  – A leverage point is a good leverage point if it is not also an outlier.



Leverage points

- We would like to have a numerical rule that will identify leverage points. Recall that

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

where

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \qquad \hat{\beta}_1 = \sum_{j=1}^{n} c_j y_j, \qquad c_j = \frac{x_j - \bar{x}}{s_{xx}}$$

- We rewrite $\hat{y}_i$ as

$$\begin{aligned} \hat{y}_i = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i = \bar{y} + \hat{\beta}_1 (x_i - \bar{x}) \\ &= \frac{1}{n} \sum_{j=1}^{n} y_j + \sum_{j=1}^{n} \frac{x_j - \bar{x}}{s_{xx}} \cdot y_j \cdot (x_i - \bar{x}) \\ &= \sum_{j=1}^{n} \left(\frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{s_{xx}} \right) y_j = \sum_{j=1}^{n} h_{ij}\, y_j \end{aligned}$$

Note that

$$\sum_{j=1}^{n} h_{ij} = \sum_{j=1}^{n} \left(\frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{s_{xx}} \right) = 1$$

Leverage points

- We can predict the value $\hat{y}_i$ as

$$\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j = h_{ii} y_i + \sum_{j \neq i} h_{ij} y_j \qquad h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{s_{xx}}$$

  where h_{ii} is commonly called the leverage or the hat-value of the ith case.

- If $h_{ii} \cong 1$, then, because of $\sum_{j=1}^{n} h_{ij} = 1$,

$$\hat{y}_i \cong 1 \times y_i + \text{other terms} \cong y_i$$

  Thus $\hat{y}_i$ is very close to y_i no matter what the rest of the data are, i.e. the ith case is a point of high leverage, i.e. a leverage point.

- A popular rule, which we shall adopt, to classify x_i as a leverage point in a simple linear regression model is if

$$h_{ii} > 2 \times \text{average}(h_{ii}) = 2 \times \frac{2}{n} = \frac{4}{n}$$

- It remains to determine if a leverage point is “good” or “bad”. This can done by inspecting the value of its standardised residual, which we shall discuss a bit later.

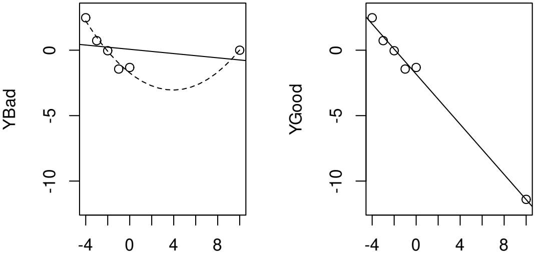
Leverage points

- Statistician Peter Huber in 1981 constructed two data sets to illustrate “good” and “bad” leverage points. Both data sets have the same X-values, while the Y's, $YBad$ and $YGood$, only differ when $X = 10$

Case	X	$YBad$	$YGood$	Leverage
1	-4	2.48	2.48	0.2897
2	-3	0.73	0.73	0.2359
3	-2	-0.04	-0.04	0.1974
4	-1	-1.44	-1.44	0.1744
5	0	-1.32	-1.32	0.1667
6	10	0	-11.4	0.9359

- The $X = 10$ point is a leverage point:

$$h_{66} = 0.9359 > 2 \times \text{average}(h_{ii}) = 4/n = 4/6 = 0.67$$



Standardised residuals

- The residuals e_i are often used to detect any problems with the proposed model.
- It can be shown that the least squares estimators E_i of e_i satisfy

$$E_i \sim N(0, \sigma^2(1-h_{ii})) \qquad \mathrm{Cov}(E_i, E_j) = -\sigma^2 h_{ij}$$

where

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{s_{xx}}$$

Thus the E_i's do not quite mimic the properties of ε_i, since $\varepsilon_i \overset{ind}{\sim} N(0, \sigma^2)$, but can be thought of as their proxies.

- The problem of E_i's having different variances can be overcome by a standardisation

$$\frac{E_i}{\sqrt{\sigma^2(1-h_{ii})}} \sim N(0,1)$$

- Consequently, the standardised residuals, r_i, defined by

$$r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2(1-h_{ii})}}$$

are generally more informative that the "crude" residuals e_i.

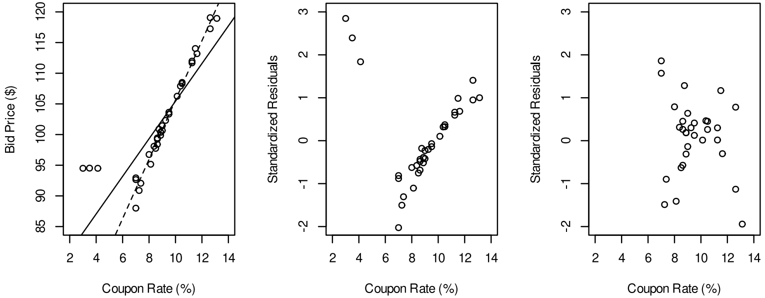
Standardised residuals

- Standardised residuals are more informative when cases of high leverage are present. In such a scenario the plots of crude residuals have nonconstant variance even if the errors have constant variance.

- Standardised residuals can be used to identify outliers. The ith case is labelled as an outlier if its standardised residual satisfies

  – $|r_i| > 2$ for small and medium data sets,

  – $|r_i| > 4$ for large data sets.



(a) (b) (c)

Influential cases

- Sometimes one or more data points can strongly control or influence the estimated regression model.

- For instance, in the previous example, the three “flower bond“ cases dramatically influenced the model.

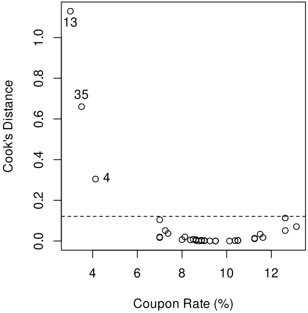
- Statistician Ralph Dennis Cook in 1977 proposed a widely used measure of the influence of individual cases which in the case of the simple linear regression is given by

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

  Here $\hat{y}_{j(i)}$ denotes the jth fitted value based on the fit obtained when the ith case has been deleted from the fit.

Influential cases

- It can be shown that
$$D_i = \frac{r_i^2}{2} \cdot \frac{h_{ii}}{1-h_{ii}}$$
- A large value of D_i may be due to a large value of r_i, a large value of h_{ii}, or both.
- If the largest $D_i \ll 1$, deletion of that case will not change the estimate by much.
- A recommended rough cut-off for D_i for the simple linear regression is $4/(n-2)$.
- In practice, it is important to look for gaps in the values of Cook's distance and not just whether values exceed the suggested cut-off.



Normality of errors

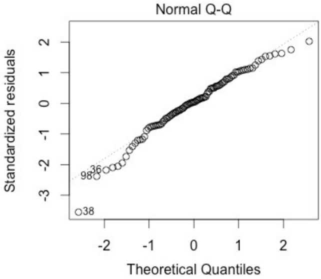
- The assumption of normality of the errors is needed in small samples for the validity of t-distribution based hypothesis tests and confidence intervals and for all sample sizes for prediction intervals.
- This assumption is generally checked by looking at the distribution of the residuals or standardised residuals.
- It can be shown that

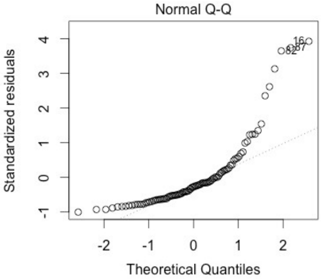
$$e_i = y_i - \hat{y}_i = y_i - \sum_{j=1}^{n} h_{ij} y_j = \ldots = \varepsilon_i - \sum_{j=1}^{n} h_{ij} \varepsilon_j$$

- In small to moderate samples, the second term in the last equation can dominate the first and the residuals can look like they come from a normal distribution even if the errors do not.
- As n increases, the second term in the last equation has a much smaller variance than that of the first term and as such the first term dominates the last equation. This implies that for large samples the residuals can be used to assess normality of the errors.

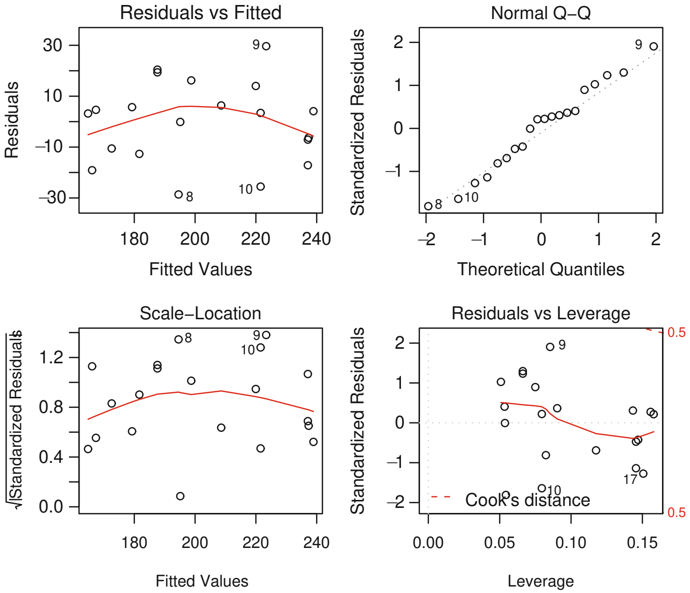
Normality of errors

- In spite of what we have just discovered, a common way to assess normality of the errors is to look at what is commonly referred to as a *normal probability plot* or a **normal Q-Q plot** of the standardised residuals.
- A normal Q-Q plot is obtained by plotting the empirical standardised residuals against the theoretical ones coming from a normal distribution.
- If the resulting plot produces points "close" to a straight line, then the data are said to be consistent with that from a normal distribution. On the other hand, departures from linearity provide evidence of non-normality.





Example: residual analysis of a regression model for the production data



Constant variance

- We have assumed that errors have a **constant variance**

$$\varepsilon_i \overset{ind}{\sim} N(0, \sigma^2) \qquad \text{where} \qquad \sigma^2 := \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_n^2$$

  This is necessary for all the inferential tools (t-tests, CIs, PIs, etc.) to be valid.

- When the variance is found to be nonconstant, there are two main methods for overcoming this:

  – **transformations** and

  – **weighted least squares**.

  We will discuss the former method only.

Constant variance

Example. A building maintenance company is bidding for a room cleaning contract and wants to estimate the number of cleaning crews it will need.

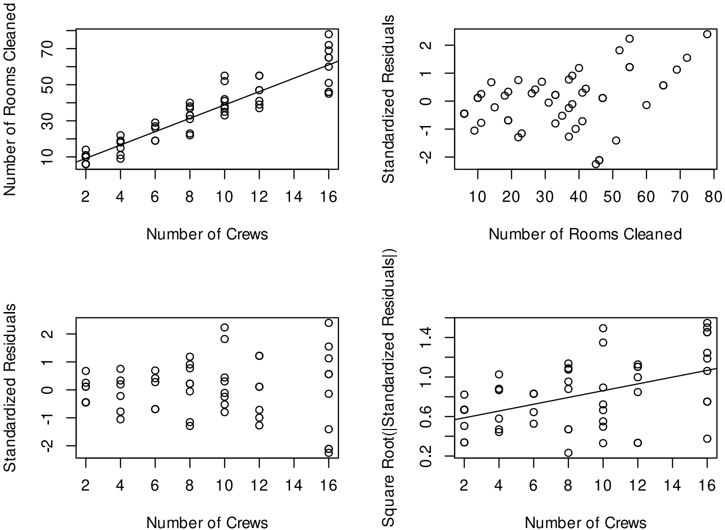
The company has a 53-day record of its cleaning service

Number of Crews	2	4	6	...
Number of Rooms Cleaned	10	15	25	...

We need to build a regression model for this data.

Constant variance

- Regression diagnostics of the room cleaning data.



Lecture 4

Diagnostics and transformations

1. Anscombe's four data sets
2. Regression diagnostics
3. Transformations

Transformations

- When nonconstant variance exists, it is often possible to transform one or both of the regression variables to produce a model in which the error variance is constant.

- Typical transformations are:
  - Square root transformation, i.e. $y_i \to \sqrt{y_i}$ or $x_i \to \sqrt{x_i}$
  - Log transformation, i.e. $y_i \to \log y_i$ or $x_i \to \log x_i$
  - Power transformations, i.e. $y_i \to y_i^\lambda$ or $x_i \to x_i^\lambda$

Square root transformation

- Consider again the room cleaning data. Count data are often modelled using the Poisson distribution. Suppose that Y follows a Poisson distribution with mean and variance λ, so that

  $$\mathrm{P}(Y = y) = \lambda^y e^{-\lambda}/y! \qquad \mathbb{E}(Y) = \mathrm{Var}(Y) = \lambda$$

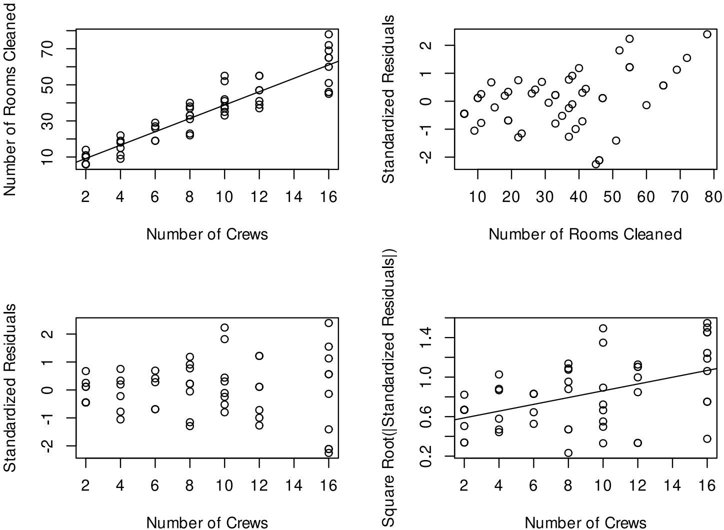
  In such a case, the appropriate transformation of Y for stabilizing variance is the square root.

- We shall try the square root transformation for both the predictor and response variables. (*When both Y and X are measured in the same units then it is often natural to consider the same transformation for both X and Y.*) We thus fit the model

  $$\sqrt{Y} = \beta_0 + \beta_1 \sqrt{x} + \varepsilon.$$

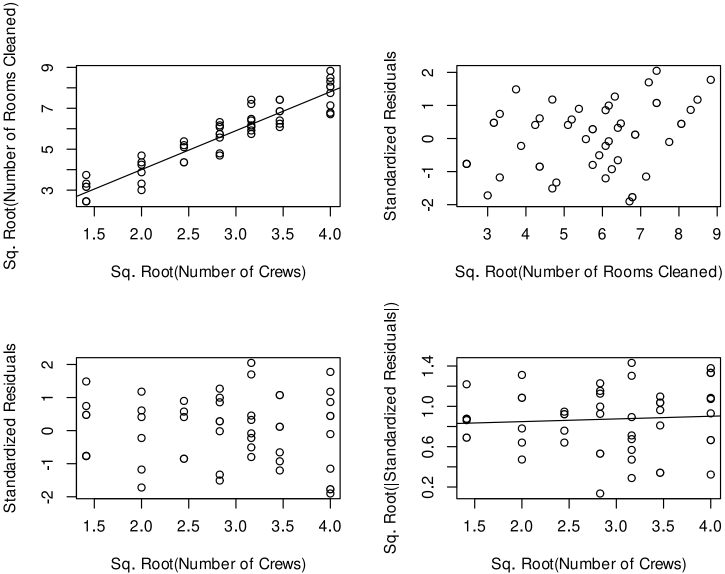
Square root transformation

- Regression diagnostics of the “raw“ or “original“ data.



Square root transformation

- Regression diagnostics of the square root transformed data.



Log transformation

- We consider data of maximum salary for 495 nonunionized job classes in a Midwestern Government unit in US in 1986.

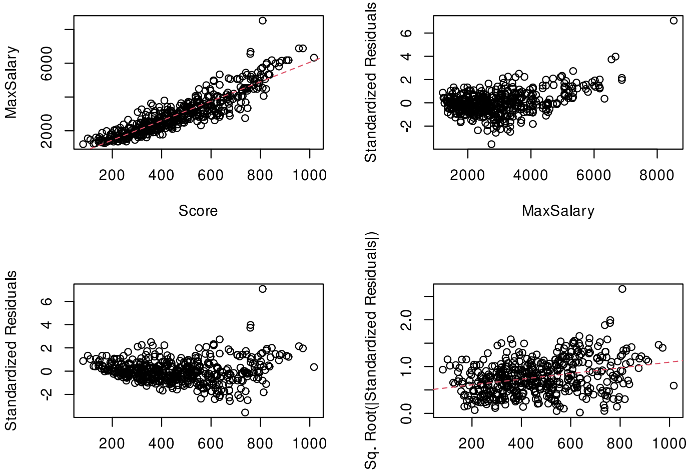
- We shall build a regression model
  - to predict the **maximum salary** (in \$) for employees in this job class
  - using single predictor, the **score** for job class, based on difficulty, skill level, training requirements and level of responsibility.

- We begin by considering a simple linear regression model for the salary data:

$$\text{MaxSalary} = \beta_0 + \beta_1 \text{Score} + \varepsilon$$

Log transformation

- Regression diagnostics of the salary data:

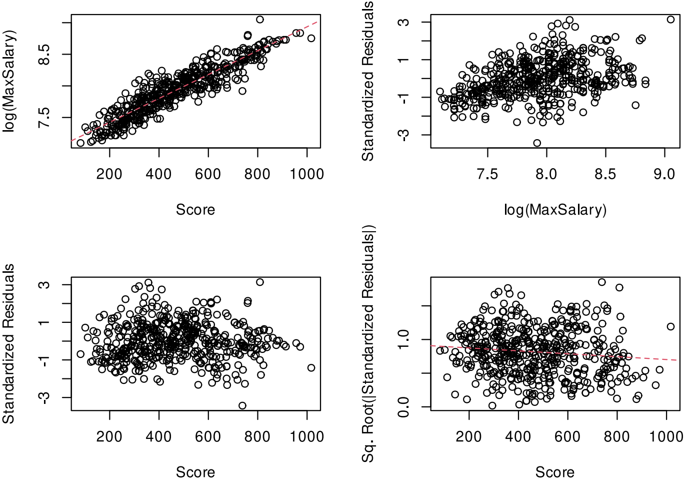


- There is a clear evidence of nonlinearity and nonconstant variance in these plots

Log transformation

- We log-transform the response variable and fit the model

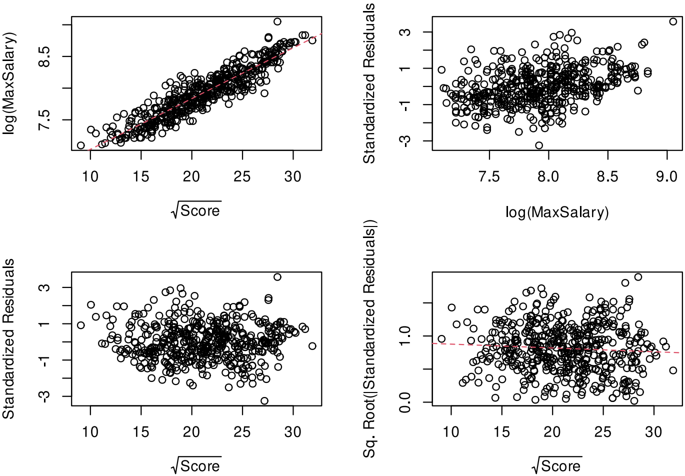
$$\log(\text{MaxSalary}) = \beta_0 + \beta_1 \text{Score} + \varepsilon$$



Log transformation

- We further square-root-transform the predictor variable and fit the model

$$\log(\text{MaxSalary}) = \beta_0 + \beta_1 \sqrt{\text{Score}} + \varepsilon$$



Power transformation

- Statisticians Box and Cox in 1964 considered a family of transformations

$$Y^{(\lambda)} = \begin{cases} \mathrm{gm}(Y)^{1-\lambda}(Y^{\lambda}-1)/\lambda & \text{if } \lambda \neq 0 \\ \mathrm{gm}(Y)\log(Y) & \text{if } \lambda = 0 \end{cases}$$

  where

$$\mathrm{gm}(Y) = \prod_{i=1}^{n} Y_i^{1/n} = \exp\left(\frac{1}{n}\sum_{i=1}^{n}\log(Y_i)\right)$$

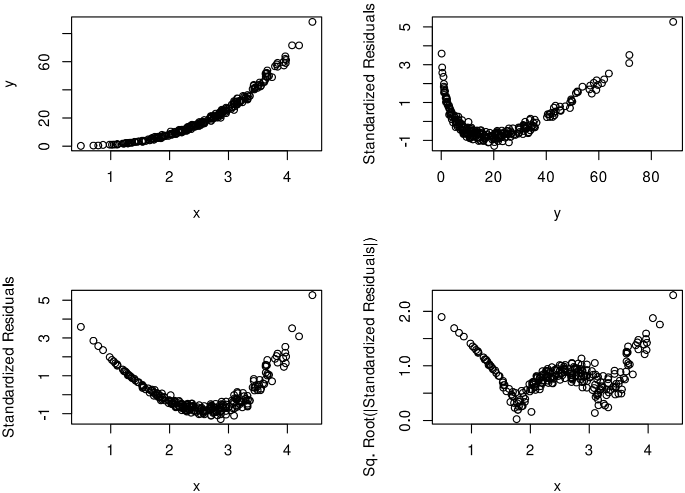
  is the geometric mean of Y.

- The Box-Cox method is based on the notion that for some value of λ the transformed version of Y, namely, $Y^{(\lambda)}$, is normally distributed. Likelihood methods can then be used to find the wanted value of λ, and the right model is then

$$\begin{aligned} Y^{\lambda} &= \beta_0 + \beta_1 x + \varepsilon \quad \text{if } \lambda \neq 0 \\ \log(Y) &= \beta_0 + \beta_1 x + \varepsilon \quad \text{if } \lambda = 0 \end{aligned}$$

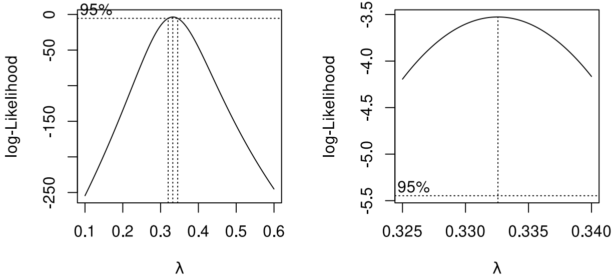
Power transformation

Example. Regression diagnostics for a data set of 250 data points shown below. It is evident that there is a power dependence between Y and x.



Power transformation

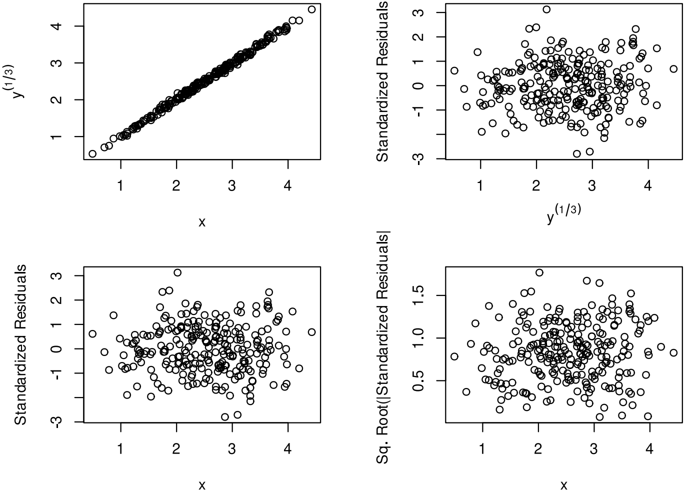
- A better model for this data is $Y^{\lambda} = \beta_0 + \beta_1 x + \varepsilon$ for certain λ.
- Log-likelihood for the Box-Cox transformation is shown below



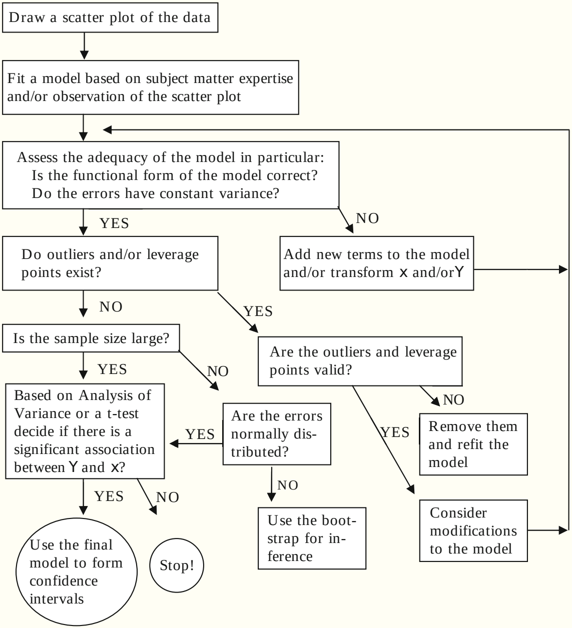
The value of λ that maximizes the log-likelihood and 95% confidence limits for λ are marked on each plot. This shows that $\lambda = 0.333 = 1/3$.

Power transformation

- Regression diagnostics for the model $Y^{1/3} = \beta_0 + \beta_1 x + \varepsilon$:



Flow chart for simple linear regression



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

Add new terms to the model and/or transform x and/or Y

Do outliers and/or leverage points exist?

NO

YES

Are the outliers and leverage points valid?

NO

Remove them and refit the model

YES

Consider modifications to the model

Is the sample size large?

YES

NO

Are the errors normally distributed?

YES

NO

Use the bootstrap for inference

Based on Analysis of Variance or a t-test decide if there is a significant association between Y and x?

YES

NO

Use the final model to form confidence intervals

Stop!

Next week

Matrix approach to linear regression