Stat 849: Linear regression diagnostics

Sündüz Keleş

Department of Statistics
Department of Biostatistics and Medical Informatics
University of Wisconsin, Madison

Regression Diagnostics

- Checking the Gauss-Markov assumptions on the error vector using graphical and numerical methods.
- Checking the normality assumption.
- Weighted least squares: Heteroscedascity.
- Outlier and influential point detection.

Assumptions on the error vector

$$\begin{aligned} E[\epsilon] &= 0 \\ var(\epsilon) &= \sigma^2 I_n \\ \epsilon &\sim \mathcal{N}(0, \sigma^2 I_n) \end{aligned}$$

First two are the G-M assumptions under which $\hat{\beta}$ is BLUE.

Residuals

$$\hat{\epsilon}_i = y_i - \hat{y}_i, \quad i = 1, \cdots, n.$$

Residuals are useful for answering questions such as

- The regression function is not linear.
- The error terms do not have a constant variance.
- The model fits pretty well but there are a few outliers.
- The error terms are not normally distributed.
- One or several important predictors have been omitted from the model.

Informal diagnostic plots of residuals

- residuals vs predictor variable,
- absolute or squared residuals vs predictor variable,
- residuals vs fitted values,
- residuals vs time or other sequence,
- residuals vs omitted predictor variables,
- box plot of residuals,
- normal probability plot (quantile plot) of residuals.

Nonlinearity of the regression function

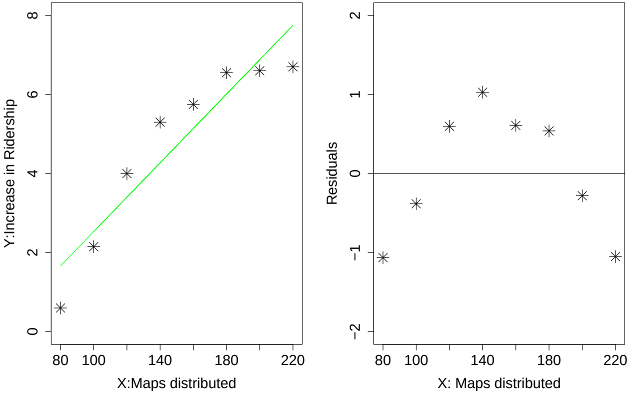


Figure: *Relation between maps distributed and bus ridership in eight test cities.* Whether a linear regression function is appropriate for the data being analyzed can be studied from a *residual plot against the predictor variable* or equivalently from a *residual plot against the fitted values*.

Checking $var(\epsilon) = \sigma^2 I_n$

- Plot residuals $\hat{\epsilon}$ versus fitted values $\hat{y}$ *or* residuals $\hat{\epsilon}$ versus predictor variable X.

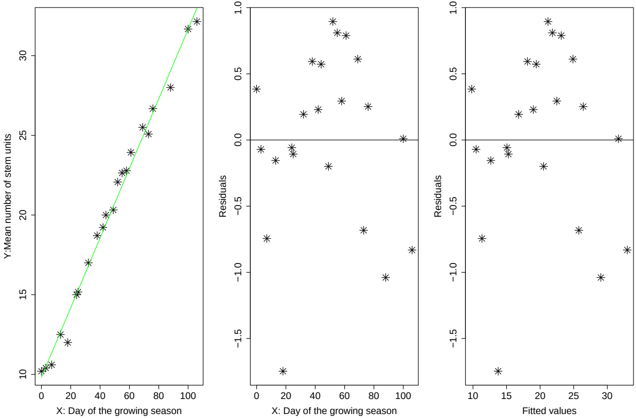


Figure: Heteroscasdicity?.

Checking $var(\epsilon) = \sigma^2 I_n$

- Plotting the absolute values of the residuals or of the squared residuals against the predictor variable X or against the fitted values $\hat{y}$ are also useful for diagnosing nonconstancy of the error variance. Especially useful when there are not many observations in the data set.

How can we deal with these?

- Weighted least squares (when the linearity assumption is ok, but the only concern is nonconstant variance).
- Transformation of the variables, i.e., response or the predictors (nonlinearity and nonconstant variance might go together).
- Changing the linear model.

Weighted Least Squares

- Although the mean of the dependent variable might be a linear function of the regressors, the variance of the error terms might also depend on those same regressors.
- Heteroscedasticity is often seen with *aggregate data*.
- Assume we are collecting data from n groups, e.g., different days, regions etc...
- For individual i in group j, we have the following linear model

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \epsilon_{ij}, \quad \mathrm{var}(\epsilon_{ij}) = \sigma^2.$$

Now, assume that we only observe *group averages* (Can you think of examples??):

$$\bar{Y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} Y_{ij}, \quad \bar{X}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{ij}.$$

Then,

$$\bar{Y}_j = \beta_0 + \beta_1 \bar{X}_j + \bar{\epsilon}_j, \quad \mathrm{var}(\bar{\epsilon}_j) = \sigma^2 \left(\frac{1}{n_j} \right) = \sigma^2 h_j.$$

Apple trees

[Weisberg, 1985] Apple trees produce "long shoots" (which may grow as much as 15 to 20 cm over a growing season) as well as "short shoots". Samples of both of these types of shoots were taken from trees in an orchard every few days during a growing season (106 days), and they were then analyzed in a laboratory. We will look at only long shoots. Among the measurements taken was a count of the number of "stem units" on each shoot.

- day of the growing season ($\bar{X}_j$),
- # shoots sampled that day (n_j),
- mean number of stem units among the sampled shoots ($\bar{Y}_j = 1/n_j \sum_{i=1}^{n_j} Y_{ij}$),
- standard deviation of the stem unit counts ($\hat{\sigma}\sqrt{n_j}$).

Interested in modelling

mean # stem units ∼ day of the growing season.

General linear model:

$$Y = X\beta + \epsilon, \quad E[\epsilon] = 0, \quad cov(\epsilon) = \sigma^2 V = \Sigma.$$

This no longer satisfies the G-M conditions.

- What is the least squares estimator for β?

$$(X^T X)^{-1} X^T Y$$

- What happens if we blindly use $\hat{\beta}_{OLS} = (X'X)^{-1}X'Y$ as an estimator of β?
- Statistical properties to check: *unbiasedness*, *efficiency*!
- Can we find the unique BLUE for β (when X is full rank)?

Consider transforming this linear model into a form that we know about!

$$\begin{aligned} V &= KK^T \quad \text{(Cholesky decomposition)} \\ Y^* &= K^{-1}Y, \quad X^* = K^{-1}X \end{aligned}$$

Now multiply $Y = X\beta + \epsilon$, $E[\epsilon = 0]$, $\text{var}(\epsilon) = \sigma^2 V$.

$$\begin{aligned} \underbrace{K^{-1}Y}_{Y^*} &= \underbrace{K^{-1}X}_{X^*}\beta + K^{-1} + K^{-1}\epsilon, \quad \text{var}(K^{-1}\epsilon) = \sigma^2 I_n \\ Y^* &= X^*\beta + \epsilon^* \quad \text{G-M satisfied!} \\ 0 &= \frac{\partial}{\partial\beta}(Y^* - X^*\beta)^T(Y^* - X^*\beta) \\ &\Longrightarrow \hat{\beta} = ((X^*)^T X^*)^{-1}(X^*)^T Y^* = (X^T V^{-1} X)^{-1} X^T V^{-1} Y \\ &\Longrightarrow var(\hat{\beta}) = \sigma^2 (X^t V^{-1} X)^{-1}. \end{aligned}$$

$\hat{\beta}$ is called the *generalized least squares estimator*, and it is called *weighted least squares estimator* when V is diagonal.

How about the statistical properties of $\hat{\beta}$?

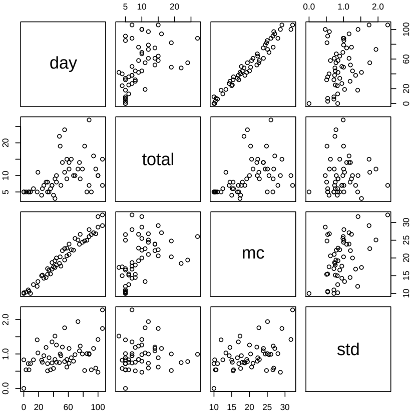
Properties of the GLE $\hat{\beta}$?

GLE $\hat{\beta}$ is simply the ordinary least squares estimator for the transformed model, we would expect $\hat{\beta}$ to have the same statistical properties of the ordinary least squares estimator in this transformed model.

- $E[\hat{\beta}] = \beta$ (OLS for the transformed model) and $E[\hat{\beta}_{OLS}] = \beta$ (OLS for the original model).
- $var(a'\hat{\beta}) \leq var(a'\hat{\beta}_{OLS})$, $\forall a_{p\times 1}$ and $var(a'\hat{\beta}) \leq var(a'\tilde{\beta})$ where $\tilde{\beta}$ is any unbiased estimator in the general linear model.

So, $\hat{\beta}_{OLS}$ is inefficient but unbiased. Can we use it for inference?

Apple trees data



R script for the apple trees example

```
all = read.table("apple_shoots.txt", header = F)
names(all) = c("day", "total", "mc", "std")
attach(all)
lm.wls = lm(mc ~ day, data = all, weights = total)
lm.ols = lm(mc ~ day, data = all)
? lm
...
 weights: an optional vector of weights to be used
         in the fitting process.  Should be ’NULL’ or a
         numeric vector. If non-NULL, weighted least squares
         is used with weights ’weights’ (that is, minimizing
         ’sum(w*e^2)’); otherwise ordinary least
         squares is used.
 > summary(lm.ols)$coef
              Estimate  Std. Error  t value     Pr(>|t|)
 (Intercept) 9.7688811 0.337028691 28.98531 6.310058e-33
  day        0.1963376 0.005648492 34.75930 1.094398e-36

#Utilize total number of counts as weights.
 > summary(lm.wls)$coef
               Estimate  Std. Error  t value     Pr(>|t|)
 (Intercept) 10.0354917 0.414189300 24.22924 2.736109e-29
  day         0.1909468 0.006364845 30.00023 1.237558e-33

#Utilize 1/(standard errors^2) as weights
 lm.wls1 = lm(mc ~ day, data = all, weights = std^-2)
 > summary(lm.wls1)$coef
              Estimate  Std. Error  t value     Pr(>|t|)
 (Intercept) 9.8150263 0.271764428 36.11593 5.655701e-37
  day        0.1894794 0.004640411 40.83246 1.671095e-39
```

Diagnostic plots for the apple trees example

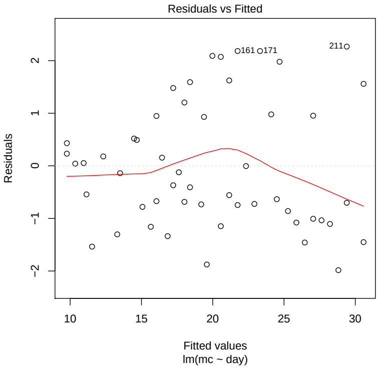


Figure: Ordinary least squares fit.

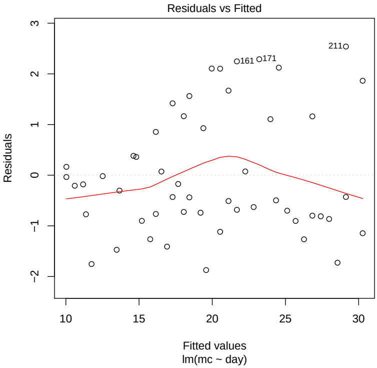


Figure: Weighted least squares fit using n as weights.

No substantial differences between the two fits!

Diagnostic plots for the apple trees example

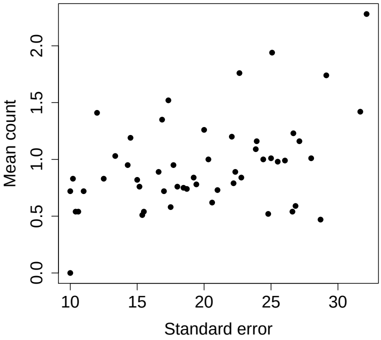


Figure: Mean counts vs standard errors.

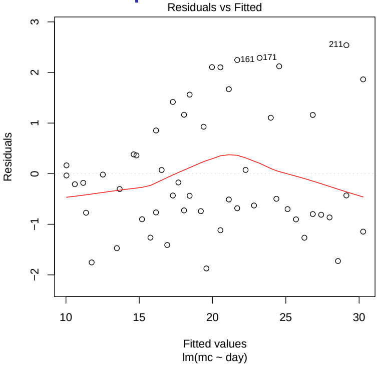


Figure: Weighted least squares fit using $\frac{1}{\hat{\sigma}_i^2}$ as weights.

Which of three models is preferable?

```
> c(summary(lm.ols)$r.squared, summary(lm.wls)$r.squared,
    summary(lm.wls1)$r.squared)
[1] 0.9602610 0.9473692 0.9714500
```

Checking normality of the error terms: $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$

- We observe estimates $\hat{\epsilon}$ of the error terms.
- We can use these estimates to compare observed values to the theoretical values.
- This can be done via a normal probability plot or a quantile-quantile (Q-Q) plot.
- Each residual is plotted against its expected value under normality (comparing empirical cumulative distribution with the theoretical one). A plot that departs substantially from linearity provides evidence against normality of the error distribution.

Q-Q plots

Quantile point q_p for random variable X is the point such that

$$F_X(q_p) = P(X \leq q_p) = p, \quad q_p = F^{-1}(p).$$

If q_p^X and q_p^Y are quantile functions of random variables X and Y, Q-Q plot of X and Y is the plots of (q_p^X, q_p^Y) for all p.

1. Sort the residuals $\hat{\epsilon}_{(1)} \leq \hat{\epsilon}_{(2)} \leq \cdots, \leq \hat{\epsilon}_{(n)}$.
2. Compute $u_i = \Psi^{-1}\left(\frac{i}{n}\right)$, where Ψ is the probability distribution function of the standard normal distribution.
3. Plot $\hat{\epsilon}_{(i)}$ against u_i.

Example

```
set.seed(1)
n = 300
x = rnorm(n, 1, 3)
y = 0.5 + 4*x + rnorm(n, 0, 1)
lm1 = lm(y ~ x)
> summary(lm1)
Call: lm(formula = y ~ x)
Residuals:
     Min       1Q   Median       3Q      Max
-2.97733 -0.67295 -0.02005  0.70834  3.85267

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.47995    0.06459   7.431 1.15e-12 ***
             4.00851    0.02091 191.708  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.045 on 298 degrees of freedom Multiple
R-Squared: 0.992,      Adjusted R-squared: 0.9919 F-statistic:
3.675e+04 on 1 and 298 DF,  p-value: < 2.2e-16
```

Example

```
sigma = sqrt(sum(lm1$resid^2)/(n-2))
> sigma
[1] 1.045298

sresid = sort(lm1$resid)
tsresid1 = qnorm(c(1 : n) / n)
postscript("qqplot.eps")
qqnorm(lm1$resid)
points(tsresid1, sresid, pch = 3, col = "red")
qqline(lm1$resid)
dev.off()
```

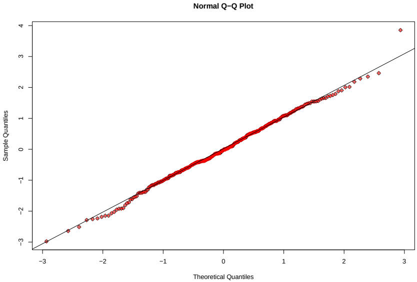


Figure: $\hat{\epsilon}_{(i)}$ versus $u_i = \Psi^{-1}\left(\frac{i}{n}\right)$.

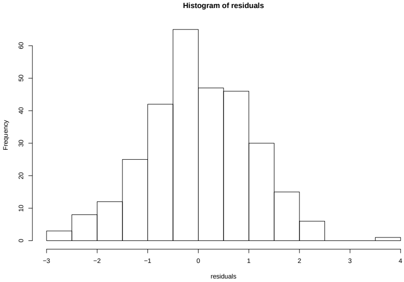


Figure: Histogram of residuals is also useful for checking normality.

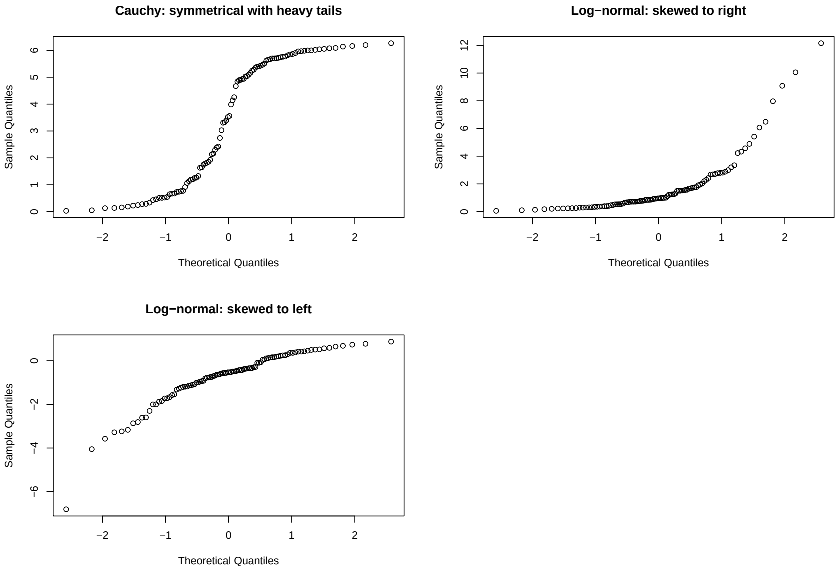
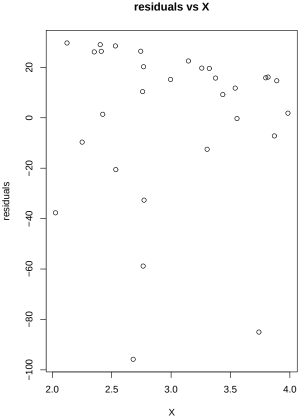


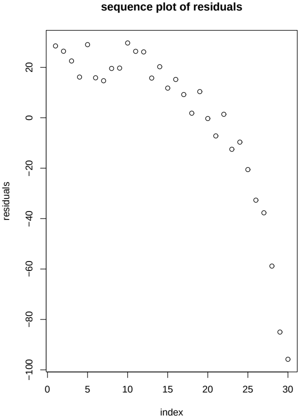
Figure: Examples of deviation from normality.

Are the errors independent?

In general no way to check this assumption from the plots alone, need to know how the data were collected.

Whenever the data is obtained in a time sequence or some other type of sequence, such as for adjacent geographic areas, it is useful to prepare a sequence plot of the residuals.





Omission of important predictor variables

Residuals should also be plotted against variables omitted from the model that might have important effects on the response.

E.g.

Consider three explanatory variables X_1, X_2, X_3 where X_2 and X_3 are binary variables.

True model $Y \sim X_1 + X_2$.
Fitted model $Y \sim X_1$.

```
X1 = runif(n, 20, 45)

X2 = rbinom(n, c(0, 1), 0.7)

X3 = rbinom(n, c(0, 1), 0.5)

Y = 0.4 + 0.1 * X1 + 2 * X2 + rnorm(n, 0, 0.5)
```

Plot residuals from the fitted model $Y \sim X_1$ versus X_2 and X_3 to see whether either of these should be included in the model.
Any ideas how these plots might look like?

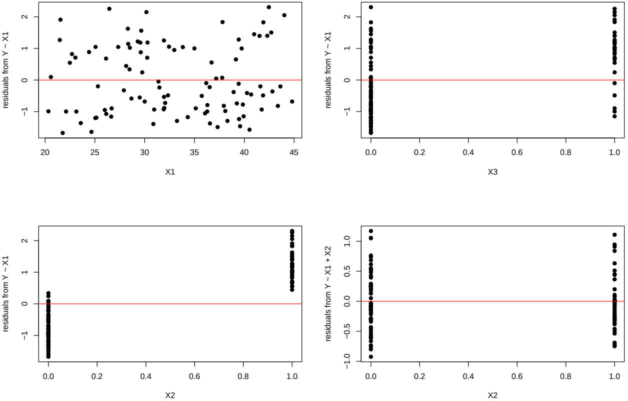


Figure: *Omission of important predictor variables*. Residuals from the fitted model $Y \sim X_1$ versus X_1, X_2, and X_3.

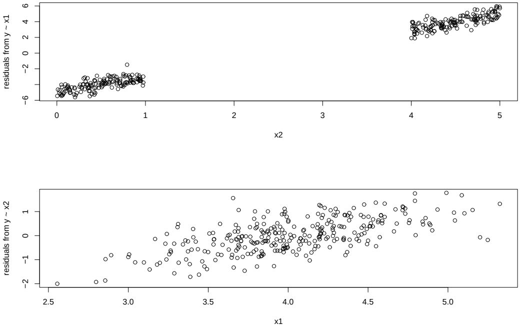


Figure: *Omission of important predictor variables*. Y: blood pressure in the morning. X_1: Calories at dinner. X_2: Age range.

Omission of important predictor variables

Does not mean the original model is *wrong*.
Implies that it can be improved by adding another predictor variable.

Added-variable plots

- Residual plots may not properly show the nature of the marginal effect of a predictor variable, given the other predictor variables in a model.
- *Added-variable plots*, also called *partial regression plots* or *adjusted variable plots*, are refined residual plots that provide graphic information about the marginal importance of a predictor variable X_j, given the other predictor variables already in the model.
- In addition, these plot might be useful for identifying the nature of the marginal relation for a predictor variable in the regression model.

How to obtain added-variable plots?

- Both the response variable Y and the predictor variable X_j under consideration are regressed against the other predicted variables in the model and the residuals are obtained for each.
- These residuals reflect the part of each variable that is *not linearly associated* with the other predicted variables already in the regression model.
- The plot of these residuals against each other (1) shows the marginal importance of this variable in reducing the residual variability and (2) may provide information about the nature of the marginal regression relation for the predictor variable X_j under consideration for possible inclusion in the regression model.

Consider Y, X_1, X_2.

$$\begin{aligned} \hat{Y}(X_1) &= \hat{\beta}_0 + \hat{\beta}_1 X_1 \\ \hat{\epsilon}(Y \mid X_1) &= Y - \hat{Y}(X_1) \\ \hat{X}_2(X_1) &= \hat{\beta}_0^* + \hat{\beta}_1^* X_1 \\ \hat{\epsilon}(X_2 \mid X_1) &= X_2 - \hat{X}_2(X_1) \end{aligned}$$

The added-variable plot for predictor variable X_2 consists of a plot of the Y residuals $\hat{\epsilon}(Y \mid X_1)$ against the X_2 residuals $\hat{\epsilon}(X_2 \mid X_1)$.

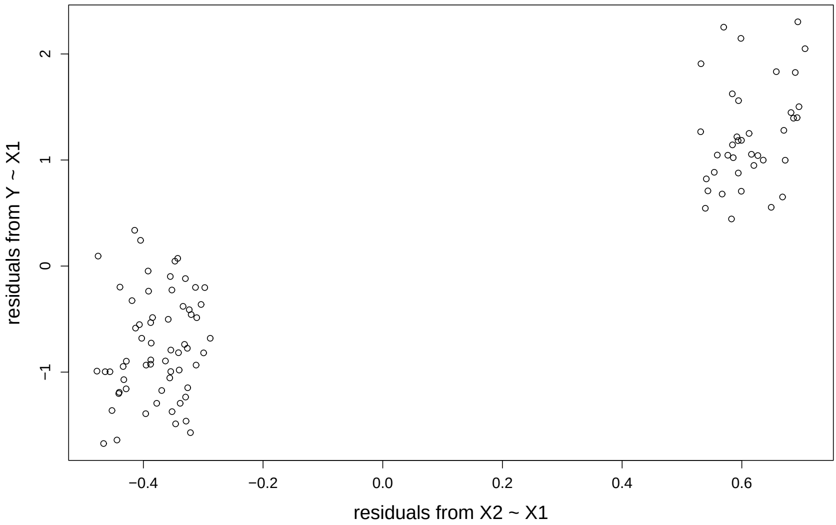


Figure: *Added-variable plot*. True model: $Y \sim X_1 + X_2$.

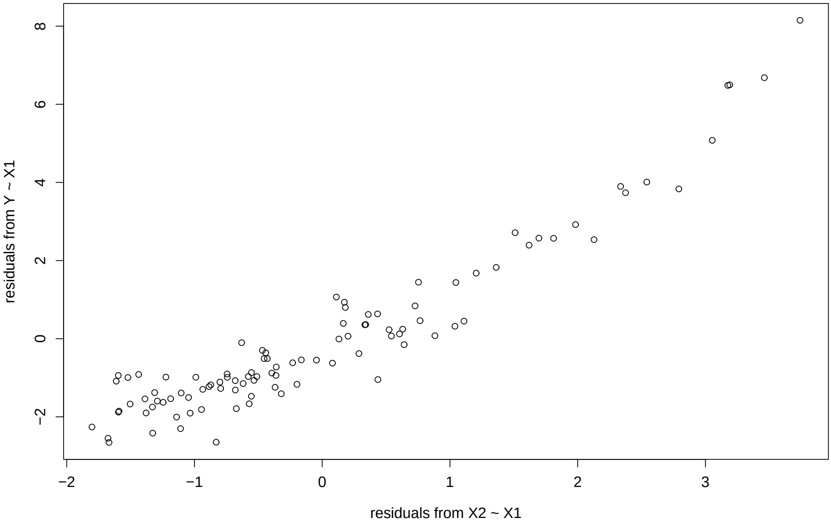


Figure: *Added-variable plot*. True model: $Y \sim X_1 + X_2^2$.

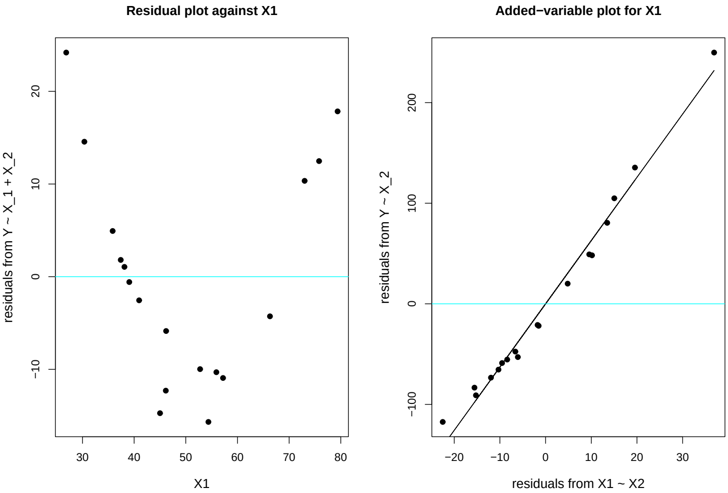


Figure: *Added-variable plot*. Question: How does $\hat{\beta}_1$ from $Y \sim X_1 + X_2$ relate to the slope of the regression line on the right panel?

```
lm0 = lm(Y ~ X1 + X2)

lm1 = lm(Y ~ X2)

lm2 = lm(X1 ~ X2)
```

```
 > lm0
 Call: lm(formula = Y ~ X1 + X2)
 Coefficients:
 (Intercept)           X1       X2
    -205.719        6.288    4.738

 > lm(lm1$resid ~ lm2$resid)
 Call: lm(formula = lm1$resid ~ lm2$resid)
  Coefficients: (Intercept)    lm2$resid
                -2.458e-15    6.288e+00
```

Exercise: Can you argue this analytically?

Notes on added-variable plots

- An added variable plot only suggests the nature of the functional relation in which a predictor variable should be added to the regression model but does not provide an analytical expression.
- The relation shown is for predictor X_j adjusted for the other predictor variables in the regression model, not for X_j directly.
- These plots may not show the proper form of the marginal effect of a predictor variable if the functional relations for some or all of the variables already in the regression model are misspecified.

Outliers and Influential observations

- Regression outliers.
- High leverage points.
- Influential points.

What are these and how can we measure them?

One should be cautious about unusual data points in linear models since they can influence the results of the analysis, and their presence might may be a signal that the model fails to capture important characteristics of the data.

Outliers

- Outliers are extreme observations.
- It is common practice to distinguish between two types of outliers.
- Outliers in the response variable are called residual outliers.
- Outliers with respect to the predictors are called leverage points. They can affect the regression model, too. Their response variables need not be outliers. However, they may almost uniquely determine regression coefficients. They may also cause the standard errors of regression coefficients to be much smaller than they would be if the observation were excluded.

Outlier detection

- Residual outliers can be identified from (1) residual plots against X or $\hat{Y}$; (2) box plots, stem-and-leaf plots, and dot plots of the residuals.
- Plotting standardized residuals might be helpful.
- Standardized residuals: $\hat{\epsilon}_i / \sqrt{\text{var}(\hat{\epsilon}_i)}$.

$$\text{var}(\hat{\epsilon}) = ?$$

Standardized (internally studentized) residuals

$$\begin{aligned}\mathrm{var}(\hat{\epsilon}) &= \mathrm{var}(Y - \hat{Y}) \\ &= \mathrm{var}(Y - PY) = \sigma^2(I - P),\end{aligned}$$

where $P = X(X^T X)^{-1} X^T$ is the projection matrix. It is also called the *hat matrix H*.
Then, $\mathrm{var}(\hat{\epsilon}_i) = \sigma^2(1 - h_{ii})$ where h_{ii} is the i-th element on the main diagonal of the hat matrix, and the covariance between residuals $\hat{\epsilon}_i$ and $\hat{\epsilon}_j$ ($i \neq j$) is $\sigma^2(0 - h_{ij}) = -h_{ij}\sigma^2$, where h_{ij} is the element in the i-th row and j-th column of the hat matrix.
Note: This can only correct for the natural non-constant variance in residuals when errors ϵ_i have constant variance.

Leverage

- h_{ii} is called the leverage (in terms of the X values) of the i-th case.
- Properties of h_{ii}: $0 \leq h_{ii} \leq 1$, $\sum_{i=1}^{n} h_{ii} = p$.
- The h_{ii} values theoretically range from $1/n$ to 1. Those that exceed $2p/n$ are said to be large.
- Interpretation: It is a measure of distance between the X values for the i-th case and the means of the X values for all of the observations. It measures the degree of conformity of a single observation to the linear pattern established by the other $n-1$ observations.
- For a linear regression with model with one predictor, the leverage associated with the specific observation (x_i, y_i) is

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

```
#Computing standardized  residuals
xx = cbind(1, x)
H = xx %*% solve((t(xx)%*%xx)) %*% t(xx) #Hat Matrix
dhat = (1-diag(H))
#Could also use lm.influence(lm1)$hat
lm1.stdres0 = lm1$resid/sqrt((sum(lm1$resid^2)/(n-2 ))*dhat)
#Two R functions that will give us standardized residuals
lm1.stdres2 = rstandard(lm1)
library(MASS)
lm1.stdres1 = stdres(lm1)
```

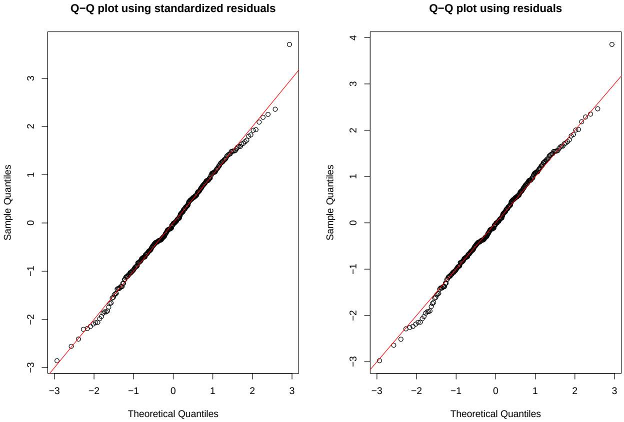


Figure: *Outlier detection.*

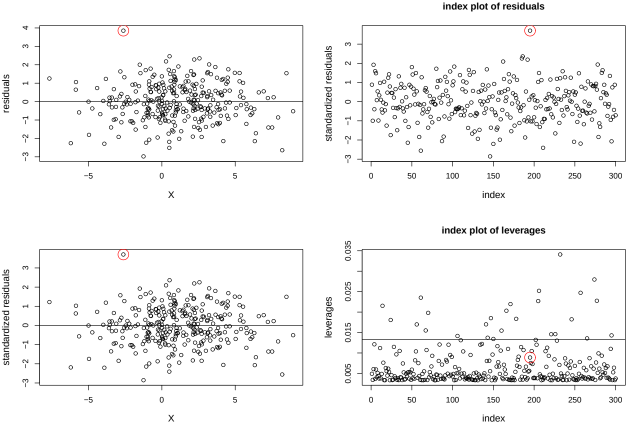


Figure: *Outlier detection*. What to do with outliers? Should only discard these if there is a direct evidence that it represents an error in recoding, a miscalculation, a malfunctioning of equipment, or a similar type of circumstance. Sometimes detection of an outlier itself might be of interest.

```
#Model fit with all observations:
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.47995    0.06459   7.431 1.15e-12 *** x
             4.00851    0.02091 191.708  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.045 on 298 degrees of freedom Multiple
R-Squared: 0.992,      Adjusted R-squared: 0.9919 F-statistic:
3.675e+04 on 1 and 298 DF,  p-value: < 2.2e-16

#Model fit omitting the circled outlier point:
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.46061    0.06340   7.265  3.3e-12 *** x[-195]
             4.01431    0.02051 195.681  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.023 on 297 degrees of freedom Multiple
R-Squared: 0.9923,     Adjusted R-squared: 0.9923 F-statistic:
3.829e+04 on 1 and 297 DF,  p-value: < 2.2e-16
```