Statistical Method
Potential Problems in Regression

I-Chen Lee, STAT, NCKU

Section 3.3.3
Gareth et al. (2021). An Introduction to Statistical Learning with Applications in R.

Nov 28, 2023

Potential problems in regression

1. Non-linearity of the response-predictor relationships
2. Correlation of error terms (residual v.s order)
3. Non-constant variance of error terms
4. Outliers
5. High-leverage points
6. Collinearity

1. Non-linearity relationships

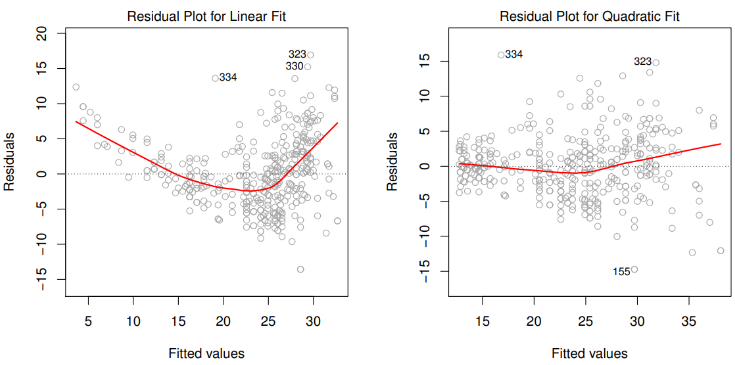


FIGURE 3.9. *Plots of residuals versus predicted (or fitted) values for the* `Auto` *data set. In each plot, the red line is a smooth fit to the residuals, intended to make it easier to identify a trend.* Left: *A linear regression of* `mpg` *on* `horsepower`. *A strong pattern in the residuals indicates non-linearity in the data.* Right: *A linear regression of* `mpg` *on* `horsepower` *and* `horsepower`^{2}. *There is little pattern in the residuals.*

3. Non-constant variance

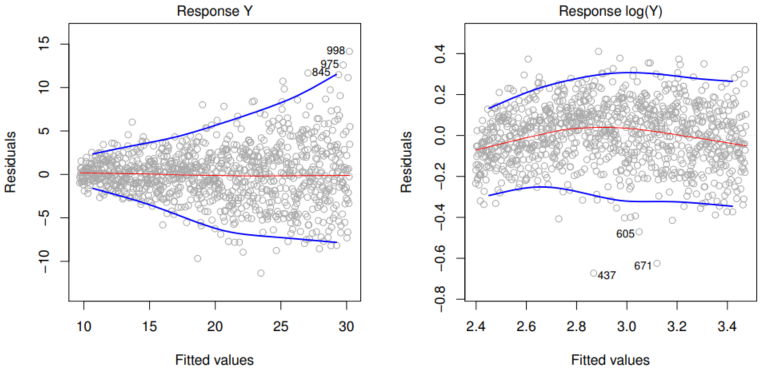


FIGURE 3.11. *Residual plots. In each plot, the red line is a smooth fit to the residuals, intended to make it easier to identify a trend. The blue lines track the outer quantiles of the residuals, and emphasize patterns.* Left: *The funnel shape indicates heteroscedasticity.* Right: *The response has been log transformed, and there is now no evidence of heteroscedasticity.*

3. Constant variance is violated? (Variance-stabilizing transformation)

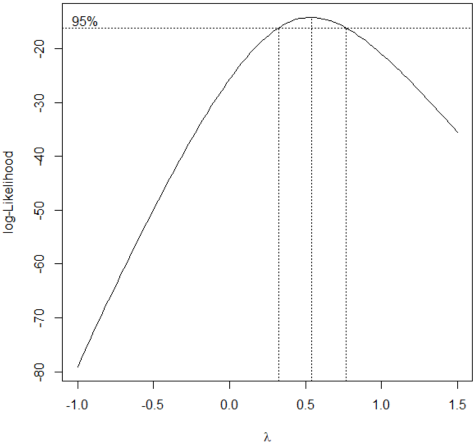
Assume $\sigma_y \propto \mu^\alpha$.

- Make the power transformation to yield a constant variance:

$$y^* \propto y^\lambda.$$

- $y^* \propto \begin{cases} y^\lambda & \text{if } \lambda \neq 0 \\ \log y & \text{if } \lambda = 0 \end{cases}.$
- It is also called the Box-Cox transformation.

3. Constant variance is violated?
Box-Cox transformation



4. Outliers
Unknown reasons on residuals

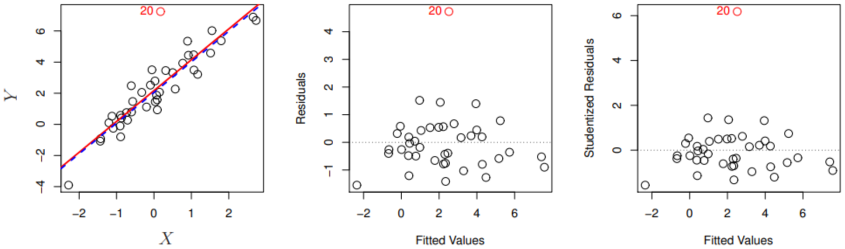


FIGURE 3.12. Left: *The least squares regression line is shown in red, and the regression line after removing the outlier is shown in blue.* Center: *The residual plot clearly identifies the outlier.* Right: *The outlier has a studentized residual of 6; typically we expect values between* −3 *and* 3.

5. High-leverage points
(Distance of X between observation to the central dataset)

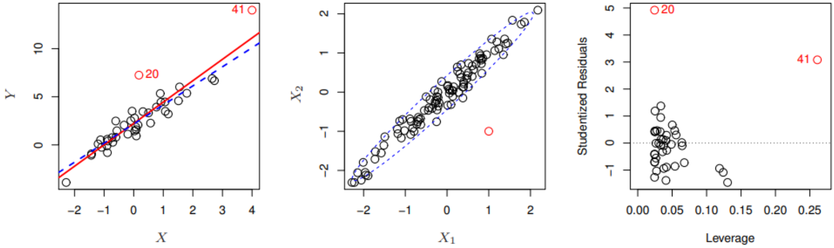


FIGURE 3.13. Left: *Observation 41 is a high leverage point, while 20 is not. The red line is the fit to all the data, and the blue line is the fit with observation 41 removed.* Center: *The red observation is not unusual in terms of its X_1 value or its X_2 value, but still falls outside the bulk of the data, and hence has high leverage.* Right: *Observation 41 has a high leverage and a high residual.*

6. Collinearity (Credit dataset)

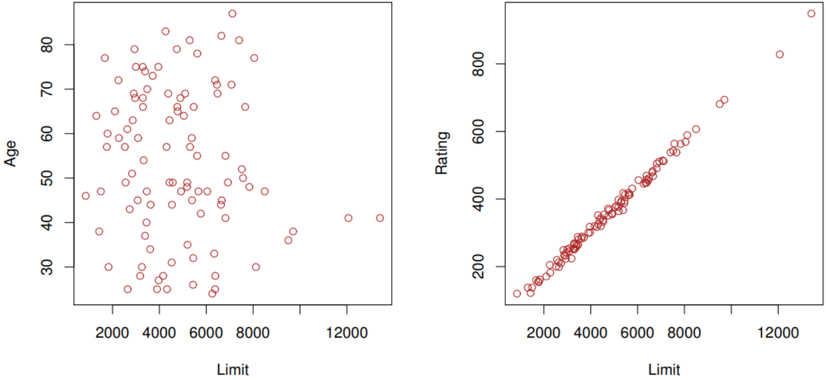


FIGURE 3.14. *Scatterplots of the observations from the* Credit *data set.* Left: *A plot of* age *versus* limit*. These two variables are not collinear.* Right: *A plot of* rating *versus* limit*. There is high collinearity.*

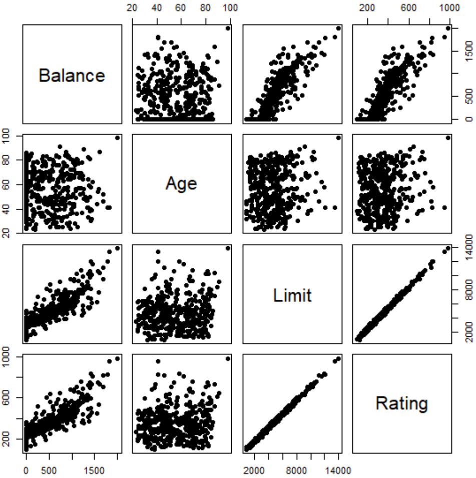
6. Collinearity

- The collinearity refers to the situation in which two or more predictor variables are closely related to one another. (like limit and rating)
- It could be difficult to separate out the individual effects of collinear variables on the response. Assume $x_1 = a + bx_2 + e$, which indicates $x_1 \sim a + bx_2$.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i = \beta_0 + \beta_1 (a + bx_{i2} + e_i) + \beta_2 x_{i2} + \varepsilon_i$$

$$= (\beta_0 + \beta_1 a) + (\beta_1 * b + \beta_2) x_{i2} + (\beta_1 e_i + \varepsilon_i) = \gamma_0 + \gamma_1 x_{i2} + \nu_i.$$

6. Collinearity (Credit dataset)



6. Collinearity (Credit dataset)

fit1:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -259.51752   55.88219  -4.644 4.66e-06 ***
Age           -2.34575    0.66861  -3.508 0.000503 ***
Limit          0.01901    0.06296   0.302 0.762830
Rating         2.31046    0.93953   2.459 0.014352 *
Residual standard error: 229.1 on 396 degrees of freedom
Multiple R-squared:  0.7536,    Adjusted R-squared:  0.7517
```

fit2:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.734e+02  4.383e+01  -3.957 9.01e-05 ***
Age         -2.291e+00  6.725e-01  -3.407 0.000723 ***
Limit        1.734e-01  5.026e-03  34.496  < 2e-16 ***
---
Residual standard error: 230.5 on 397 degrees of freedom
Multiple R-squared:  0.7498,    Adjusted R-squared:  0.7486
```

6. Collinearity (Credit dataset)

fit3:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -269.58110   44.80616  -6.017 4.05e-09 ***
Age           -2.35078    0.66764  -3.521  0.00048 ***
Rating         2.59328    0.07443  34.840  < 2e-16 ***
---
Residual standard error: 228.8 on 397 degrees of freedom
Multiple R-squared:  0.7535,    Adjusted R-squared:  0.7523
```

fit4:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -377.53680   45.25418  -8.343 1.21e-15 ***
Limit          0.02451    0.06383   0.384   0.7012    
Rating         2.20167    0.95229   2.312   0.0213 *  
---
Residual standard error: 232.3 on 397 degrees of freedom
Multiple R-squared:  0.7459,    Adjusted R-squared:  0.7447
```

Detecting multicollinearity

Use variance inflation factors (VIFs) to examine the possible multicollinearity.

$$\mathrm{VIF}(\hat{\beta}_k) = \frac{1}{1 - R_k^2},$$

where R_k^2 is the R^2 from the regression model

$$x_k = \alpha_0 + \alpha_1 x_1 + \cdots + \alpha_{k-1} x_{k-1} + \alpha_{k+1} x_{k+1} + \cdots + \epsilon.$$

General rule of thumb:

- VIF > 4: further investigation.
- VIF > 10: serious multicollinearity requiring correction.
- In the Credit data, a regression of balance on age, rating, and limit indicates that the predictors have VIF values of 1.01, 160.67, and 160.59. Collinearity in the data!

6. Collinearity
(VIF in Credit dataset)

```
> vif(fit1)
       Age      Limit     Rating
  1.011385 160.592880 160.668301
> vif(fit2)
     Age    Limit
1.010283 1.010283
> vif(fit3)
     Age   Rating
1.010758 1.010758
> vif(fit4)
   Limit   Rating
160.4933 160.4933
```

Possible solution to collinearity

1. The first is to drop one of the problematic variables from the regression.
2. The second solution is to combine the collinear variables together into a (new) single predictor. (The methods could be the principle component analysis (PCA) or Partial least squares (PLS) regression).

Summary

Potential problems in regression:

1. Non-linearity of the response-predictor relationships
   (Transform on X or Y.)
2. Correlation of error terms (residual v.s order)
   (Fit time series models or detrend first.)
3. Non-constant variance of error terms
   (Transform on Y.)
4. Outliers
   (Can not be explained by X. We can remove it.)
5. High-leverage points
   (Report it but keep it in the set.)
6. Collinearity
   (Use VIF to detect collinearity or other methods.)