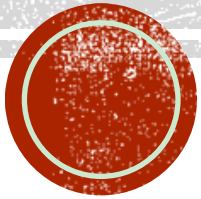
Lecture 10.6: model diagnosis



CONTENTS

- Assessing residual
- Multicollinearity
- Influential points
- Other pitfalls

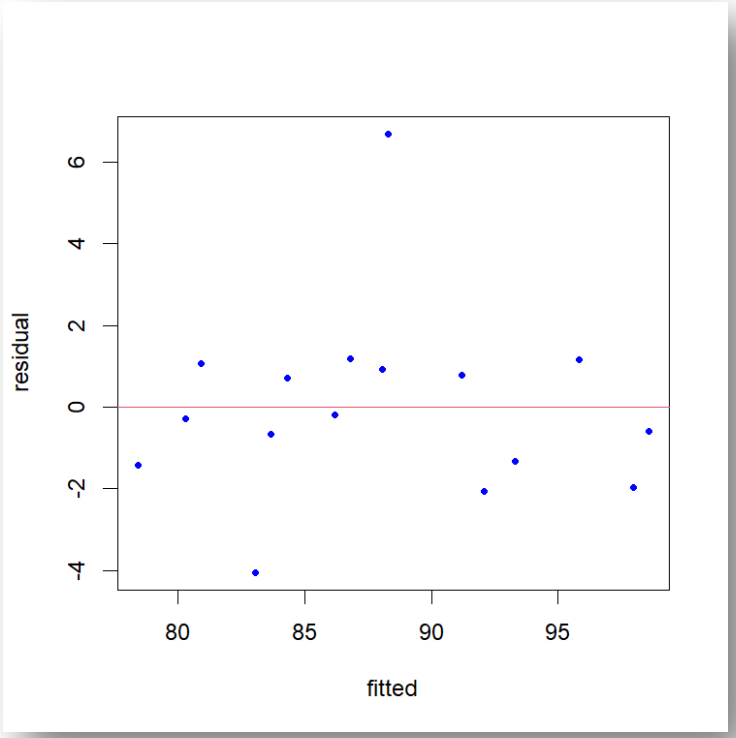
FOUR CONDITIONS ("LINE") $\epsilon i \sim N(0,\sigma^2)$

- **Linear Function:** The mean of the response, E(Yi) , at each set of values of the predictors, (x1i, x2i, …….), is a Linear function of the predictors.
- **Independent:** The errors, εi, are Independent.
- **Normally Distributed:** The errors, εi, at each set of values of the predictors, (x1i, x2i, …….), are Normally distributed.
- **Equal variances:** The errors, εi , at each set of values of the predictors, (x1i, x2i, …….), have equal variances (denoted).

ASSESSING RESIDUAL

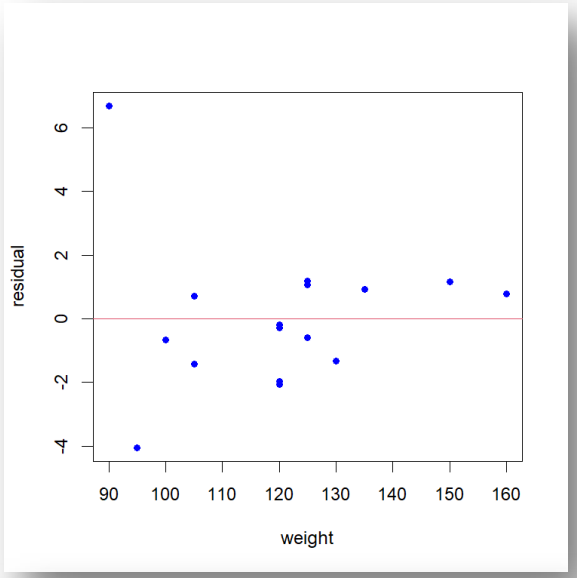
Scatterplot (residuals vs. fitted)

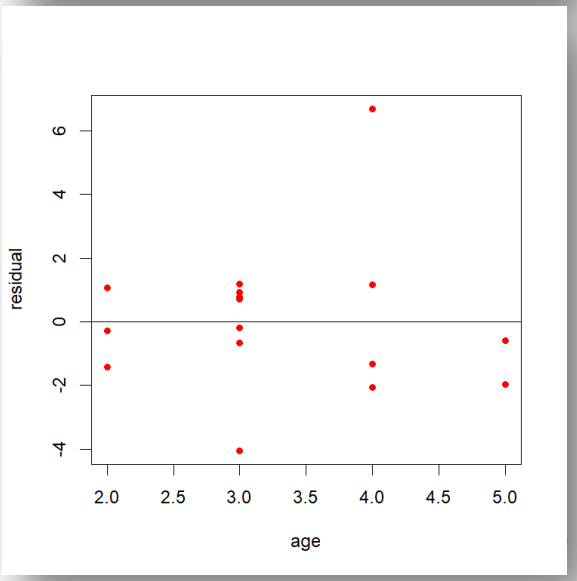
- the (vertical) average of the residuals remains close to 0 as we scan the plot from left to right (this affirms the "L" condition);
- the (vertical) spread of the residuals remains approximately constant as we scan the plot from left to right (this affirms the "E" condition);
- there are no excessively outlying points (we'll explore this in more detail in "influential points" section).
- violation of any of these three may necessitate remedial action (such as transforming one or more predictors and/or the response variable), depending on the severity of the violation



Scatterplot (residuals vs. predictor)

- the (vertical) average of the residuals remains close to 0 as we scan the plot from left to right (this affirms the "L" condition);
- the (vertical) spread of the residuals remains approximately constant as we scan the plot from left to right (this affirms the "E" condition);
- violation of either of these for at least one residual plot may suggest the need for transformations of one or more predictors and/or the response variable





NORMALITY OF RESIDUAL

- Histogram
- Box plot
- PP plot
- QQ plot

MULTICOLLINEARITY

Multicollinearity

- **Multicollinearity** in regression analysis occurs when two or more predictor variables are highly correlated to each other, such that they do not provide unique or independent information in the regression model.

Assumption: no explanatory variable is a perfect linear function of any other explanatory variables

Multicollinearity

- **1.1 Perfect Multicollinearity**: occurs if at least two explanatory variables are linearly dependent.

$$\lambda_1 x_1 + \lambda_2 x_2 + ... + \lambda_k x_k = 0$$

- **Example**: including the same information twice (weight in pounds and weight in kilograms), not using dummy variables correctly (falling into the dummy variable trap), etc..

$$Yi = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

$$x1i = a_0 + a_1 x_{2i}$$

- **Consequence**: OLS cannot generate estimates of regression coefficients (mathematically, they do not exit)!
- **Solution**: Drop one of the variables!

Examples of Perfect Multicollinearity

- **One Predictor Variable is a Multiple of Another**

"height in centimeters" and "height in meters"

- **One Predictor Variable is a Transformed Version of Another**

$$BMIZ = \frac{BMI - \mu}{\sigma}$$

- **The Dummy Variable Trap**

Income	Age	Marital Status
$45,000	23	Single
$48,000	25	Single
$54,000	24	Single
$57,000	29	Single
$65,000	38	Married
$69,000	36	Single
$78,000	40	Married
$83,000	59	Divorced
$98,000	56	Divorced
$104,000	64	Married
$107,000	53	Married

→

Income	Age	Single	Married	Divorced
$45,000	23	1	0	0
$48,000	25	1	0	0
$54,000	24	1	0	0
$57,000	29	1	0	0
$65,000	38	0	1	0
$69,000	36	1	0	0
$78,000	40	0	1	0
$83,000	59	1	0	1
$98,000	56	1	0	1
$104,000	64	0	1	0
$107,000	53	0	1	0

Multicollinearity

- **Imperfect (near) Multicollinearity**: offen arises in the time series regression model, especially in data involving economic time series, while variables over time tend to move in the same direction..

$$\lambda_1 x_1 + \lambda_2 x_2 + ... + \lambda_k x_k + v_i = 0$$

- **Example**:

$$Yi = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

$$x1i = a_0 + a_1 x_{2i} + u_i$$

- **Consequence**: does not violate assumption. The regression will still yield unbiased estimates of all of the coefficients, But!!! (see next section)

Multicollinearity

- **2.1 Structural multicollinearity**: is a mathematical artifact caused by creating new predictors from other predictors — such as, creating the predictor x^2 from the predictor x.
- **2.2 Data multicollinearity**: This type of multicollinearity is present in the data itself rather than being an artifact of our model. Observational experiments are more likely to exhibit this kind of multicollinearity.

Why do we care about Multicollinearity

- The variances and the standard errors of the regression coefficient estimates will increase. This means lower t-statistics.
- The overall fit of the regression equation will be largely unaffected by multicollinearity. This also means that forecasting and prediction will be largely unaffected.
- Regression coefficients will be sensitive to specifications. Regression coefficients can change substantially when variables are added or dropped.

The detection of Multicollinearity

- **High Correlation Coefficients**
- Pairwise correlations among independent variables might be high (in absolute value). Rule of thumb: If the correlation > 0.8 then severe multicollinearity may be present.
- **High R2 with low t-Statistic Values**
- Possible for individual regression coefficients to be insignificant but for the overall fit of the equation to be high.
- **High Variance Inflation Factors (VIFs)**
- A VIF measures the extent to which multicollinearity has increased the variance of an estimated coefficient. It looks at the extent to which an explanatory variable can be explained by all the other explanatory variables in the equation.

VIF

$$\text{var}(bj) = \sigma^2 (X'X)^{-1}_{jj} = \frac{\sigma^2}{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j^2)(1 - R_j^2)}$$

Where R_j^2 is the R^2-value obtained by regressing the jth predictor on the remaining predictors, variables xj is the dependent, and the remaining x's are explanatory variables.

$$VIF_j = \frac{1}{1 - R_j^2}, TOL_j = \frac{1}{VIFj} = 1 - R_j^2$$

The multicollinearity problem is serious if $R_j^2 > 0.8$, VIFj > 5 or TOLj < 0.2

Remedies for Multicollinearity

- **Drop a Redundant Variable**
- **Transform the Multicollinear Variables**
- **Increase the sample size**
- **Do nothing**

INFLUENTIAL POINTS

Outlier

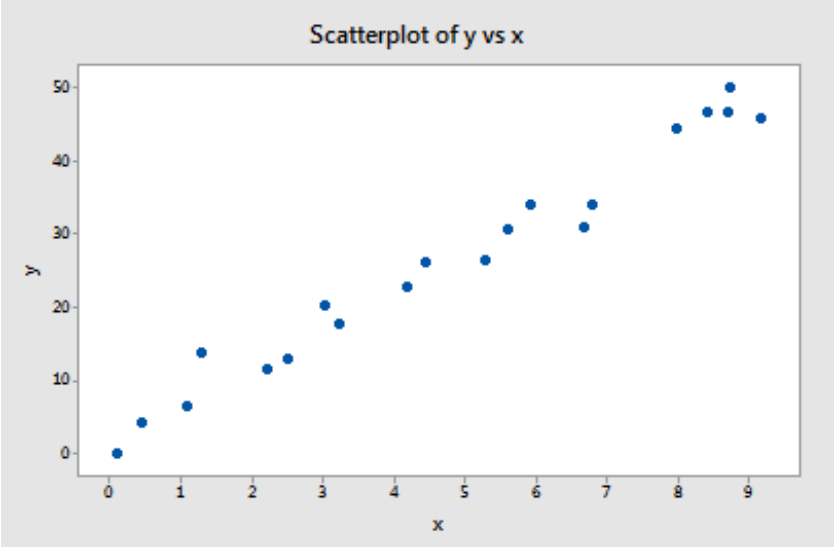
leverage

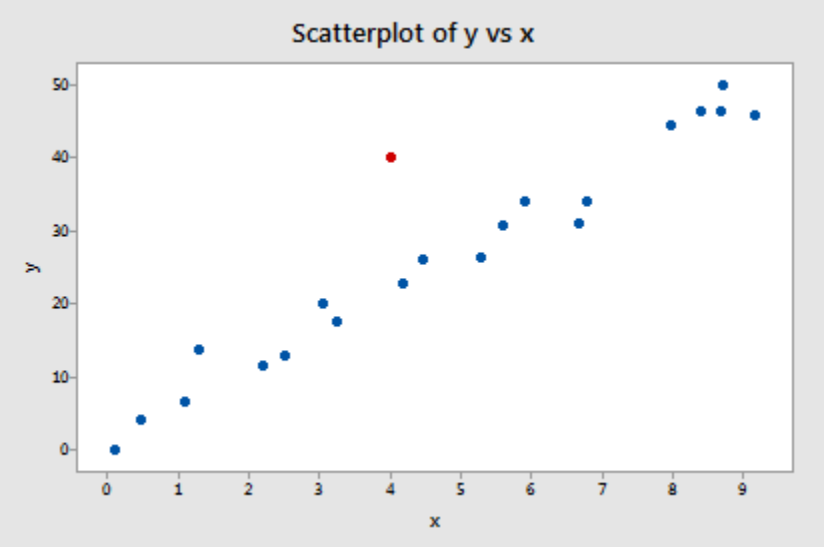
Influential points

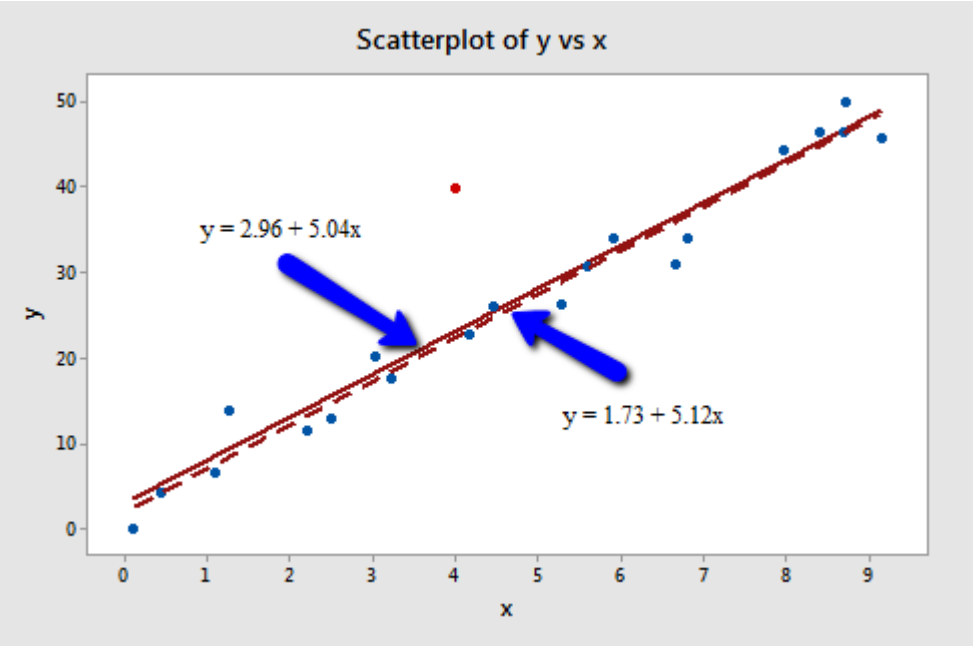
- If an observation has a response value that is very different from the predicted value based on a model, then that observation is called an **outlier**. On the other hand, if an observation has a particularly unusual combination of predictor values (e.g., one predictor has a very different value for that observation compared with all the other data observations), then that observation is said to have high **leverage**.

Outlier

- An **outlier** is a data point whose response *y* does not follow the general trend of the rest of the data.







Do the two samples yield different results when testing H_0: $\beta_1 = 0$? Well, we obtain the following output when :

the **red data point is excluded**:

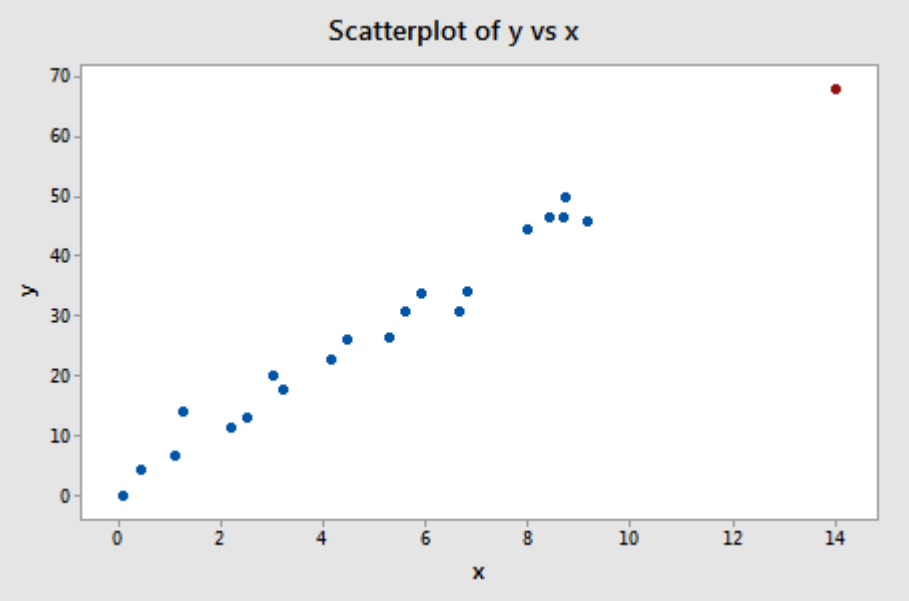
```
Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
4.71075  91.01%     90.53%      89.61%


Coefficients

Term       Coef  SE Coef  T-Value  P-Value   VIF
Constant   2.96     2.01     1.47    0.157
x         5.037    0.363    13.86    0.000  1.00


Regression Equation

y = 2.96 + 5.037 x
```

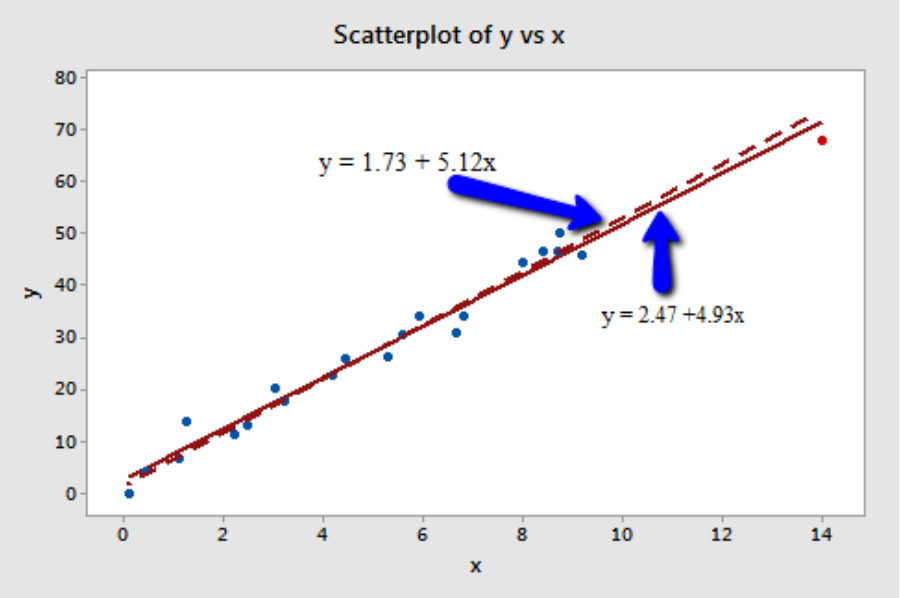
the **red data point is excluded**:

```
Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
2.59199  97.32%     97.17%      96.63%


Coefficients

Term       Coef  SE Coef  T-Value  P-Value   VIF
Constant   1.73     1.12     1.55    0.140
x         5.117    0.200    25.55    0.000  1.00


Regression Equation

y = 1.73 + 5.117 x
```

Leverage

- A data point has high **leverage** if it has "extreme" predictor *x* values.



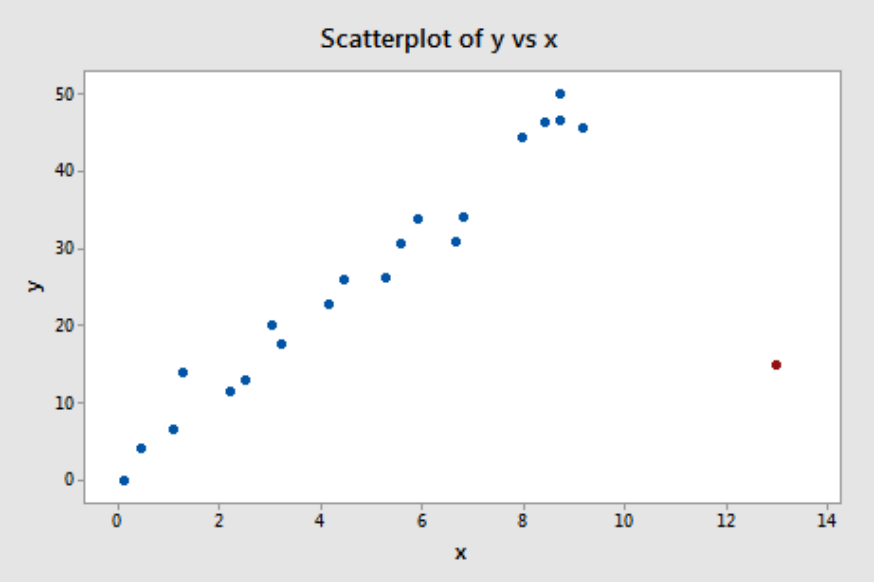


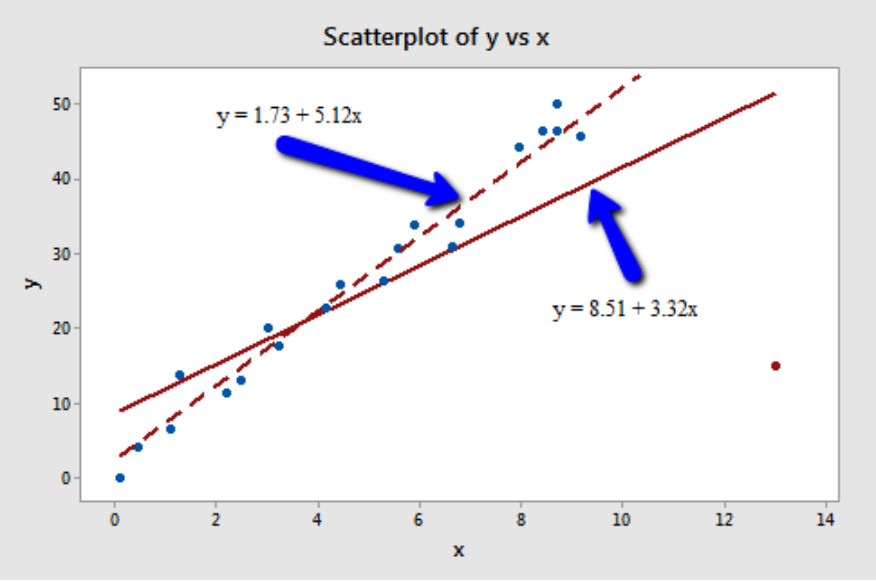
the **red data point is included**

```
Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
2.70911  97.74%     97.62%      97.04%


Coefficients

Term       Coef  SE Coef  T-Value  P-Value   VIF
Constant   2.47     1.08     2.29    0.033
x         4.927    0.172    28.66    0.000  1.00


Regression Equation

y = 2.47 + 4.927 x
```

the **red data point is excluded**

```
Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
2.59199  97.32%     97.17%      96.63%


Coefficients

Term       Coef  SE Coef  T-Value  P-Value   VIF
Constant   1.73     1.12     1.55    0.140
x         5.117    0.200    25.55    0.000  1.00


Regression Equation

y = 1.73 + 5.117 x
```





the **red data point is included**:

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
10.4459	55.19%	52.84%	19.11%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	8.50	4.22	2.01	0.058	
x	3.320	0.686	4.84	0.000	1.00

Regression Equation

y = 8.50 + 3.320 x

the **red data point is excluded**:

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
2.59199	97.32%	97.17%	96.63%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1.73	1.12	1.55	0.140	
x	5.117	0.200	25.55	0.000	1.00

Regression Equation

y = 1.73 + 5.117 x

Definition and properties of leverages

- $\mathbf{Y=X\beta+\epsilon}$ $\qquad\qquad$ $\mathbf{b=(X'X)^{-1}X'y}$

Therefore, the predicted responses can be represented in matrix notation as:

$\mathbf{\hat{y}=Xb}$ ⟶ $\mathbf{\hat{y}=X\,(X'X)^{-1}X'y}$

That is, the predicted responses can be obtained by pre-multiplying the $n \times 1$ column vector, $\boldsymbol{y}$, containing the observed responses by the $n \times n$ matrix $\boldsymbol{H}$:

$\mathbf{H=X\,(X'X)^{-1}X'}$ $\qquad\qquad$ $\mathbf{\hat{y}=Hy}$

$$y_i=h_{i1}y_1+h_{i2}y_2+\ldots+h_{ii}y_i+\ldots+h_{in}y_n \text{ for } i=1,\ldots,n$$

Identifying data points whose x values are extreme

- $y_i = h_{i1}y_1 + h_{i2}y_2 + \ldots + h_{ii}y_i + \ldots + h_{in}y_n$ for i=1,…,n
- H_{ii}, the **leverage**, quantifies the influence that the observed response y_i has on its predicted value $\hat{y}_i$.

$$\bar{h} = \frac{\sum_{i=1}^{n} h_{ii}}{n} = \frac{k+1}{n}$$

$$h_{ii} > 3\left(\frac{k+1}{n}\right)$$

Large leverage

$$h_{ii} > 2\left(\frac{k+1}{n}\right)$$

Identifying Outlier (unusual Y values)

- Residuals
- ei=yi−$\widehat{yi}$
- Standardized Residuals

$$r_i = \frac{e_i}{s(e_i)} = \frac{e_i}{\sqrt{MSE(1-h_{ii})}}$$

The good thing about standardized residuals is that they quantify how large the residuals are in standard deviation units, and therefore can be easily used to identify outliers:

An observation with a standardized residual that is larger than **3** (in absolute value) is deemed by some to be an **outlier**. Some statistical software flags any observation with a standardized residual that is larger than 2 (in absolute value)

Identifying Influential Data Points

- (1) Difference in Fits (DFFITS)

The difference in fits for observation i, denoted $DFFITS_i$, is defined as:

$$DFFUTSi = \frac{\hat{y}i - \hat{y}(i)}{\sqrt{MSE_{(i)}h_{ii}}}$$

An observation is deemed influential if the absolute value of its *DFFITS* value is greater than:

$$2\sqrt{\frac{k+2}{n-k-2}}$$

where, n = the number of observations and k = the number of predictor terms (i.e., the number of regression parameters excluding the intercept).

Identifying Influential Data Points

- (2) Cook's Distance

Cook's distance measure, denoted D_i, is defined as:

$$D_i = \frac{(y_i - \hat{y}_i)^2}{(k+1) \times MSE} \left[\frac{h_{ii}}{(1-h_{ii})^2} \right]$$

- D_i directly summarizes how much *all* of the fitted values change when the i^{th} observation is deleted.
- A data point having a large D_i indicates that the data point strongly influences the fitted values.

- If D_i is greater than 0.5, then the i^{th} data point is worthy of further investigation as it **may be influential**.
- If D_i is greater than 1, then the i^{th} data point is **quite likely to be influential**.
- Or, if D_i sticks out like a sore thumb from the other D_i values, it is **almost certainly influential**.

Dealing with Problematic Data Points

First, check for obvious data errors:

- If the error is just a data entry or data collection error, correct it.
- If the data point is not representative of the intended study population, delete it.
- If the data point is a procedural error and invalidates the measurement, delete it.

Consider the possibility that you might have just misformulated your regression model:

- Did you leave out any important predictors?
- Should you consider adding some interaction terms?
- Is there any nonlinearity that needs to be modeled?

Decide whether or not deleting data points is warranted:

- Do not delete data points just because they do not fit your preconceived regression model.
- You must have a good, **objective** reason for deleting data points.
- If you delete any data after you've collected it, justify and describe it in your reports.
- If you are not sure what to do about a data point, analyze the data twice — once with and once without the data point — and report the results of both analyses.

OTHER PITFALLS

- Nonconstant Variance and Weighted Least Squares
- Autocorrelation and Time Series Methods

- Nonconstant Variance and Weighted Least Squares
- Autocorrelation and Time Series Methods