Types of outliers in linear regression

Types of outliers

How do(es) the outlier(s) influence the least squares line?

To answer this question think of where the regression line would be with and without the outlier(s).

How do(es) the outlier(s) influence the least squares line?
Some terminology

- **Outliers** are points that fall away from the cloud of points.
- Outliers that fall horizontally away from the center of the cloud are called leverage points.
- High leverage points that actually influence the slope of the regression line are called **influential** points.
- In order to determine if a point is influential, visualize the regression line with and without the point. Does the slope of the line change considerably? If so, then the point is influential. If not, then it’s not.

Data are available on the log of the surface temperature and the log of the light intensity of 47 stars in the star cluster CYG OB1.

Clicker question

Which of the below best describes the outlier?

(a) influential  
(b) leverage  
(c) leverage  
(d) none of the above  
(e) there are no outliers

Does this outlier influence the slope of the regression line?
Recap (cont.)

Which of the following is true?

(a) Influential points always change the intercept of the regression line.
(b) Influential points always reduce $R^2$.
(c) It is much more likely for a low leverage point to be influential, than a high leverage point.
(d) When the data set includes an influential point, the relationship between the explanatory variable and the response variable is always nonlinear.
(e) None of the above.

Inference for linear regression

In 1966 Cyril Burt published a paper called “The genetic determination of differences in intelligence: A study of monozygotic twins reared apart?” The data consist of IQ scores for [an assumed random sample of] 27 identical twins, one raised by foster parents, the other by the biological parents.

- Scatter plot showing the relationship between biological IQ and foster IQ with a regression line.
- Coefficient table:
  - (Intercept): 9.20760, Std. Error: 9.29990, t value: 0.990, Pr(>|t|): 0.332
  - bioIQ: 0.90144, Std. Error: 0.09633, t value: 9.358, Pr(>|t|): 1.2e-09

- Summary statistics:
  - Residual standard error: 7.729 on 25 degrees of freedom
  - Multiple R-squared: 0.7779, Adjusted R-squared: 0.769
  - F-statistic: 87.56 on 1 and 25 DF, p-value: 1.204e-09

Which of the following is false?

(a) For each 10 point increase in the biological twin's IQ, we would expect the foster twin's IQ to increase on average by 9 points.
(b) Roughly 78% of the foster twins’ IQs can be accurately predicted by the model.
(c) The linear model is $fosterIQ = 9.2 + 0.9 \times bioIQ$.
(d) Foster twins with IQs higher than average IQs have biological twins with higher than average IQs as well.
Clicker question

Assuming that these 27 twins comprise a representative sample of all twins separated at birth, we would like to test if these data provide convincing evidence that the IQ of the biological twin is a significant predictor of IQ of the foster twin. What are the appropriate hypotheses?

(a) $H_0 : b_0 = 0; H_A : b_0 \neq 0$
(b) $H_0 : \beta_1 = 0; H_A : \beta_1 \neq 0$
(c) $H_0 : b_1 = 0; H_A : b_1 \neq 0$
(d) $H_0 : \beta_0 = 0; H_A : \beta_0 \neq 0$

We always use a t-test in inference for regression.

Remember: Test statistic, $T = \frac{\text{point estimate} - \text{null value}}{\text{SE}}$

Point estimate = $\hat{b}_1$ is the observed slope.

$SE_{\hat{b}_1}$ is the standard error associated with the slope.

Degrees of freedom associated with the slope is $df = n - 2$, where $n$ is the sample size.

Remember: We lose 1 degree of freedom for each parameter we estimate, and in simple linear regression we estimate 2 parameters, $\beta_0$ and $\beta_1$.

Estimate | Std. Error | t value | Pr(>|t|) |
---|---|---|---|
(Intercept) | 9.2076 | 9.2999 | 0.99 | 0.3316 |
bioIQ | 0.9014 | 0.0963 | 9.36 | 0.0000 |

$T = \frac{0.9014 - 0}{0.0963} = 9.36$

$df = 27 - 2 = 25$

$p-value = P(|T| > 9.36) < 0.01$

% College graduate vs. % Hispanic in LA

What can you say about the relationship between % college graduate and % Hispanic in a sample of 100 zip code areas in LA?
Inference for linear regression

HT for the slope

% College educated vs. % Hispanic in LA - another look

What can you say about the relationship between % college graduate and % Hispanic in a sample of 100 zip code areas in LA?

<table>
<thead>
<tr>
<th>% Hispanic</th>
<th>% College graduate</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>25%</td>
<td>25%</td>
</tr>
<tr>
<td>50%</td>
<td>50%</td>
</tr>
<tr>
<td>75%</td>
<td>75%</td>
</tr>
<tr>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>


Inference for linear regression

HT for the slope

% College educated vs. % Hispanic in LA - linear model

Clicker question

Which of the below is the best interpretation of the slope?

| Estimate  | Std. Error | t value | Pr(>|t|) |
|-----------|------------|---------|---------|
| (Intercept) | 0.7290    | 0.0308  | 23.68   | 0.0000  |
| %Hispanic  | -0.7527   | 0.0501  | -15.01  | 0.0000  |

(a) A 1% increase in Hispanic residents in a zip code area in LA is associated with a 75% decrease in % of college grads.
(b) A 1% increase in Hispanic residents in a zip code area in LA is associated with a 0.75% decrease in % of college grads.
(c) An additional 1% of Hispanic residents decreases the % of college graduates in a zip code area in LA by 0.75%.
(d) In zip code areas with no Hispanic residents, % of college graduates is expected to be 75%.


Inference for linear regression

CI for the slope

Confidence interval for the slope

Clicker question

Remember that a confidence interval is calculated as point estimate ± ME and the degrees of freedom associated with the slope in a simple linear regression is n–2. Which of the below is the correct 95% confidence interval for the slope parameter? Note that the model is based on observations from 27 twins.

| Estimate  | Std. Error | t value | Pr(>|t|) |
|-----------|------------|---------|---------|
| (Intercept) | 9.2076    | 9.2999  | 0.99    | 0.3316  |
| bioIQ     | 0.9014    | 0.0963  | 9.36    | 0.0000  |

(a) 9.2076 ± 1.65 × 9.2999
(b) 0.9014 ± 2.06 × 0.0963
(c) 0.9014 ± 1.96 × 0.0963
(d) 9.2076 ± 1.96 × 0.0963


Inference for linear regression

HT for the slope

% College educated vs. % Hispanic in LA - linear model

Do these data provide convincing evidence that there is a statistically significant relationship between % Hispanic and % college graduates in zip code areas in LA?

| Estimate  | Std. Error | t value | Pr(>|t|) |
|-----------|------------|---------|---------|
| (Intercept) | 0.7290    | 0.0308  | 23.68   | 0.0000  |
| hispanic  | -0.7527   | 0.0501  | -15.01  | 0.0000  |


Inference for linear regression

HT for the slope

% College educated vs. % Hispanic in LA - another look

How reliable is this p-value if these zip code areas are not randomly selected?
Inference for linear regression

Recap

- Inference for the slope for a SLR model (only one explanatory variable):
  - Hypothesis test:
    \[
    T = \frac{b_1 - \text{null value}}{SE_{b_1}} \quad df = n - 2
    \]
  - Confidence interval:
    \[
    b_1 \pm t_{\alpha/2; n-2}SE_{b_1}
    \]
- The null value is often 0 since we are usually checking for any relationship between the explanatory and the response variable.
- The regression output gives \( b_1 \), \( SE_{b_1} \), and two-tailed p-value for the t-test for the slope where the null value is 0.
- We rarely do inference on the intercept, so we'll be focusing on the estimates and inference for the slope.

Caution

- Always be aware of the type of data you’re working with: random sample, non-random sample, or population.
- Statistical inference, and the resulting p-values, are meaningless when you already have population data.
- If you have a sample that is non-random (biased), the results will be unreliable.
- The ultimate goal is to have independent observations – and you know how to check for those by now.

Variability partitioning

- We considered the t-test as a way to evaluate the strength of evidence for a hypothesis test for the slope of relationship between \( x \) and \( y \).
- However, we can also consider the variability in \( y \) explained by \( x \), compared to the unexplained variability.
- Partitioning the variability in \( y \) to explained and unexplained variability requires analysis of variance (ANOVA).

ANOVA output - Sum of squares

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>biolQ</td>
<td>1</td>
<td>5231.13</td>
<td>5231.13</td>
<td>87.56</td>
<td>0.0000</td>
</tr>
<tr>
<td>Residuals</td>
<td>25</td>
<td>1493.53</td>
<td>59.74</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>26</td>
<td>6724.66</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
\text{Sum of squares:} \quad SS_{Tot} = \sum (y - \bar{y})^2 = 6724.66 \quad (\text{total variability in } y)
\]
\[
SS_{Err} = \sum (y - \hat{y})^2 = e_i^2 = 1493.53 \quad (\text{unexplained variability in residuals})
\]
\[
SS_{Reg} = \sum (\hat{y} - \bar{y})^2 = SS_{Tot} - SS_{Err} \quad (\text{explained variability in } y)
\]
\[
= 6724.66 - 1493.53 = 5231.13
\]

\[
\text{Degrees of freedom:} \quad df_{Tot} = n - 1 = 27 - 1 = 26
\]
\[
df_{Reg} = 1 \quad (\text{there is only 1 predictor})
\]
\[
df_{Res} = df_{Tot} - df_{Reg} = 26 - 1 = 25
\]
ANOVA output - F-test

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>bioIQ</td>
<td>1</td>
<td>5231.13</td>
<td>5231.13</td>
<td>87.56</td>
</tr>
<tr>
<td>Residuals</td>
<td>25</td>
<td>1493.53</td>
<td>59.74</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>26</td>
<td>6724.66</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Mean squares: $\text{MS}_\text{Reg} = \frac{\text{SS}_\text{Reg}}{\text{df}_\text{Reg}} = \frac{5231.13}{1} = 5231.13$

$\text{MS}_\text{Err} = \frac{\text{SS}_\text{Err}}{\text{df}_\text{Err}} = \frac{1493.53}{25} = 59.74$

F-statistic: $F_{(1,25)} = \frac{\text{MS}_\text{Reg}}{\text{MS}_\text{Err}} = 87.56$ (ratio of explained to unexplained variability)

The null hypothesis is $\beta_1 = 0$ and the alternative is $\beta_1 \neq 0$. With a large F-statistic, and a small p-value, we reject $H_0$ and conclude that the slope is significantly different than 0, i.e. the explanatory variable is a significant predictor of the response variable.

Revisit $R^2$

- Remember, $R^2$ is the proportion of variability in $y$ explained by the model:
  - A large $R^2$ suggests a linear relationship between $x$ and $y$ exists.
  - A small $R^2$ suggests the evidence provided by the data may not be convincing.
- There are actually two ways to calculate $R^2$:
  1. From the definition: proportion of explained to total variability
  2. Using correlation: square of the correlation coefficient

\[ R^2 = \frac{\text{explained variability}}{\text{total variability}} = \frac{\text{SS}_\text{Reg}}{\text{SS}_\text{Tot}} = \frac{5231.13}{6724.66} \approx 0.78 \quad (1) \]

\[ R^2 = \text{square of correlation coefficient} = 0.881^2 \approx 0.78 \quad (2) \]