Regression Diagnostics Notes

Lt Col Ken Horton Professor Bradley Warner

30 July, 2020

Objectives

1) Obtain and interpret R-squared and the F-statistic.
2) Use `R` to evaluate the assumptions of a linear model.
3) Identify and explain outliers and leverage points.

Introduction

Over the last two lessons, we have detailed simple linear regression. First, we described the model and its underlying assumptions. Next, we obtained parameter estimates using the method of least squares. Finally, we obtained the distributions of parameter estimates and used that information to conduct inference on parameters and predictions. Implementation was relatively straightforward; once we obtained the expressions of interest, we used `R` to find parameters estimates, interval estimates, etc.

We have been using the `lm()` function. It is simple and intuitive. The first argument is the formula. A formula is given by the response variable, followed by a tilde (`~`) and followed by the predictor variables. If there are more than one predictors, they are separated by `+`. The output is an object of class "lm" which is a list containing several components. For more information, consult the documentation (`?lm`). In this lesson we will explore more tools to assess the quality of our linear regression model. Some these will generalize when we move to multiple predictors.

Assessing our model

There is more that we can do with the output from the `lm()` function. Let's load our data from the Starbucks example and build the linear regression model.

```
library(openintro)
```

```
star_mod <- lm(calories~carb,data=starbucks)
```

```
summary(star_mod)
```

```
##
## Call:
## lm(formula = calories ~ carb, data = starbucks)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -151.962  -70.556   -0.636   54.908  179.444
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 146.0204    25.9186   5.634 2.93e-07 ***
## carb          4.2971     0.5424   7.923 1.67e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 78.26 on 75 degrees of freedom
## Multiple R-squared:  0.4556, Adjusted R-squared:  0.4484
## F-statistic: 62.77 on 1 and 75 DF,  p-value: 1.673e-11
```

You may have noticed some other information that appeared in the summary of our model. Some of these quantities are familiar and others are new.

Residual Standard Error

The "residual standard error" is the estimate of σ. In our example, this turned out to be 78.26. If we would like more precision, we first recognize that `summary(my.model)` is also a list with several components:

```
names(summary(star_mod))
```

```
##  [1] "call"          "terms"         "residuals"     "coefficients"
##  [5] "aliased"       "sigma"         "df"            "r.squared"
##  [9] "adj.r.squared" "fstatistic"    "cov.unscaled"
```

As expected, the `sigma` component shows the estimated value of σ.

```
summary(star_mod)$sigma
```

```
## [1] 78.25956
```

Obviously, if this value is smaller the closer the points will be to the regression fit. It is a measure of unexplained variance in the data.

R-squared

Another quantity that appears is R-squared. You may have heard of this value before. R-squared is one measure of goodness of fit. Essentially, R-squared is a ratio of variance (in the response) explained by the model to overall variance of the response. It helps to describe the decomposition of variance:

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{SS_{\text{Total}}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{SS_{\text{Regression}}} + \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{SS_{\text{Error}}}$$

In other words, the overall variation in y can be separated into two parts: variation due to the linear relationship between y and the predictor variable(s) and residual variation (due to random scatter or perhaps a poorly chosen model).

R-squared simply measures the ratio between $SS_{\text{Regression}}$ and SS_{Total}. A common definition of R-squared is the proportion of overall variation in the response that is explained by the linear model. R-squared can be between 0 and 1. Values of R-squared close to 1 indicate a tight fit (little scatter) around the estimated regression line. Value close to 0 indicate the opposite (large remaining scatter).

We can obtain R-squared "by hand" or by using the output of the `lm()` function:

```
summary(star_mod)$r.squared
```

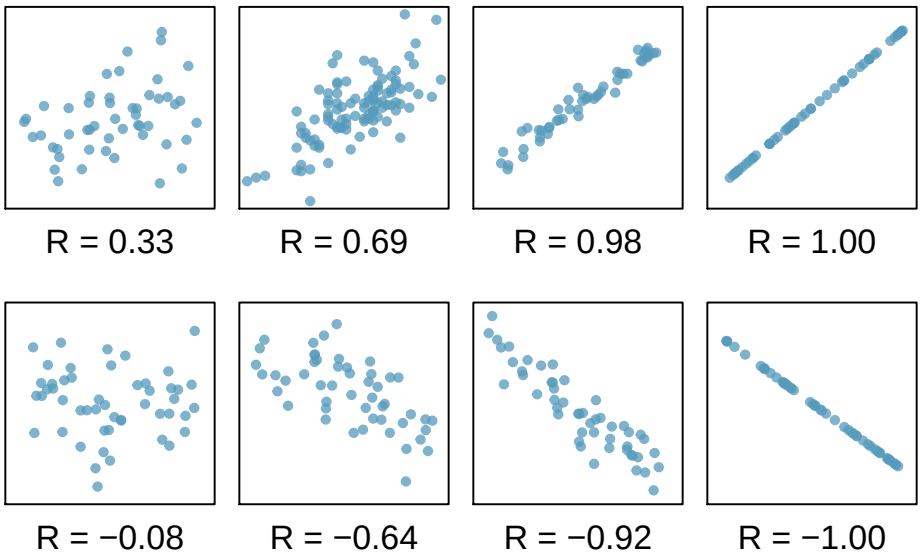
```
## [1] 0.4556237
```

For simple linear regression, R-squared is related to **correlation**. We can compute the correlation using a formula, just as we did with the sample mean and standard deviation. However, this formula is rather complex,[1] so we let `R` do the heavy lifting for us.

```
starbucks %>%
  summarize(correlation=cor(carb,calories),correlation_squared=correlation^2)
```

```
## # A tibble: 1 x 2
##   correlation correlation_squared
##         <dbl>               <dbl>
## 1       0.675               0.456
```

The figure below shows eight plots and their corresponding correlations. Only when the relationship is perfectly linear is the correlation either -1 or 1. If the relationship is strong and positive, the correlation will be near +1. If it is strong and negative, it will be near -1. If there is no apparent linear relationship between the variables, then the correlation will be near zero.



[1] Formally, we can compute the correlation for observations (x_1, y_1), (x_2, y_2), $\ldots$, (x_n, y_n) using the formula

$$R = \frac{1}{n-1} \sum_{i=1}^{n} \frac{x_i - \bar{x}}{s_x} \frac{y_i - \bar{y}}{s_y}$$

where $\bar{x}$, $\bar{y}$, s_x, and s_y are the sample means and standard deviations for each variable.

Exercise
If a linear model has a very strong negative relationship with a correlation of -0.97, how much of the variation in the response is explained by the explanatory variable?[2]

Note that one of the components of `summary(lm())` function is `adj.r.squared`. This is a value of R-squared adjusted for number of predictors. We'll cover this concept more closely in Math 378.

F-Statistic

Another quantity that appears in the summary of the model is the F-statistic. This value evaluates the null hypothesis that all of the non-intercept coefficients are equal to 0. Rejecting this hypothesis implies that the model is useful in the sense that at least one of the predictors shares a significant linear relationship with the response.

H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$ H_a: At least one coefficient not equal to 0.

where p is the number of predictors in the model. Just like in ANOVA, this is a simultaneous test of all coefficients and does not inform us which one(s) are different from 0.

The F-statistic is given by

$$\frac{n-p-1}{p}\frac{\sum(\hat{y}_i - \bar{y})^2}{\sum e_i^2}$$

Under the null hypothesis, the F-statistic follows the F distribution with parameters p and $n-p-1$.

In our example, the F-statistic is redundant since there is only one predictor. In fact, the p-value associated with the F-statistic is equal to the p-value associated with the estimate of β_1. However, when we move to cases with more predictor variables, we may be interested in the F-statistic.

```
summary(star_mod)$fstatistic
```

```
##    value    numdf    dendf
## 62.77234  1.00000 75.00000
```

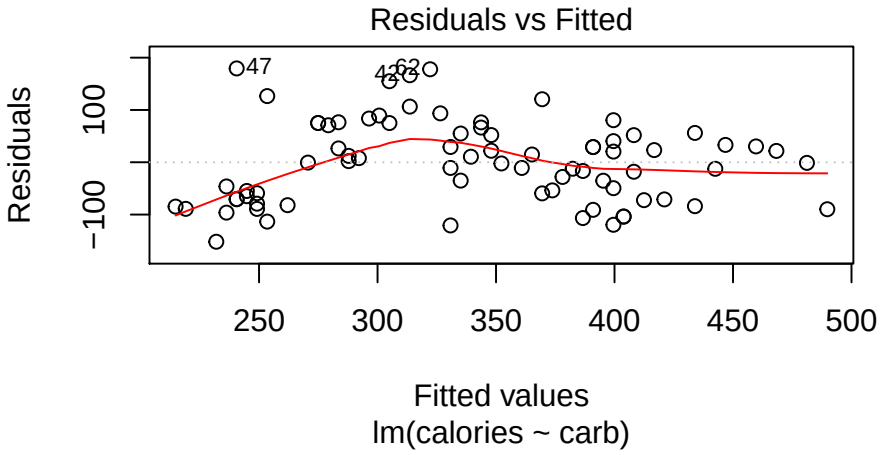
Checking Assumptions

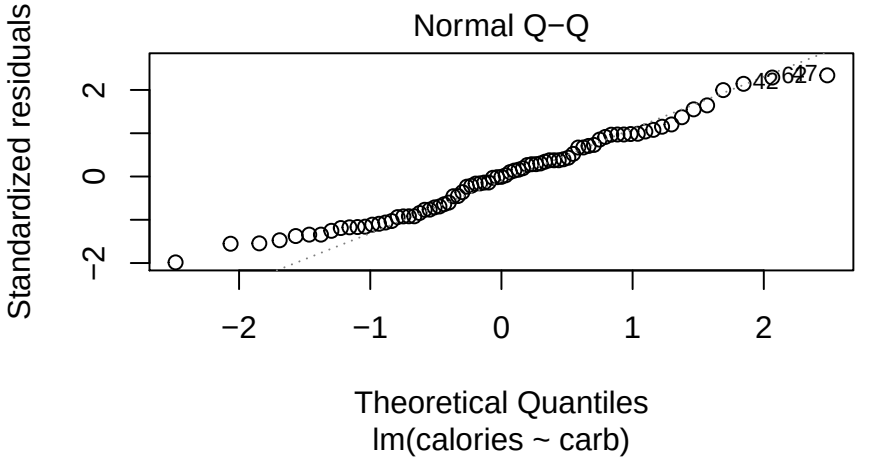
Finally, we can use the "lm" object to check the assumptions of the model. We have discussed the assumptions before but in this lesson we will use `R` to generate visual checks. We will also introduce the ideas of outliers and leverage points.

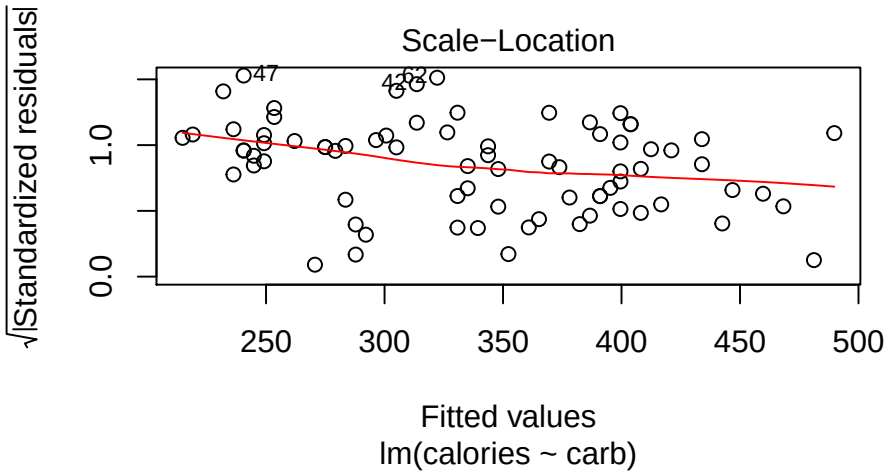
Applying the `plot()` function to an "lm" object provides several graphs that allow us visually evaluate a linear model's assumptions. Also, applying `plot()` to an "lm" object returns four different plots by default. To obtain all four at once, simply use `plot(my.model)`. However, it's best to walk through each of these four plots in our Starbucks example.

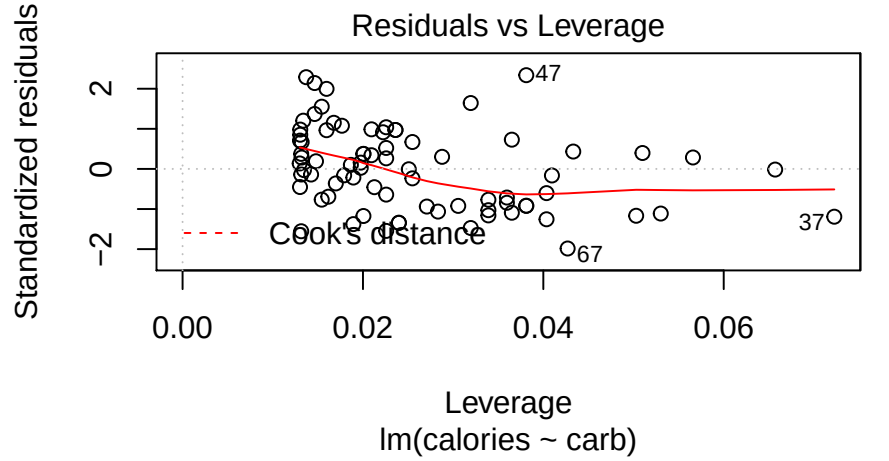
```
plot(star_mod)
```

[2] About $R^2 = (-0.97)^2 = 0.94$ or 94% of the variation is explained by the linear model.


Residuals vs Fitted
Residuals
100
−100
47
62
250
300
350
400
450
500
Fitted values
lm(calories ~ carb)

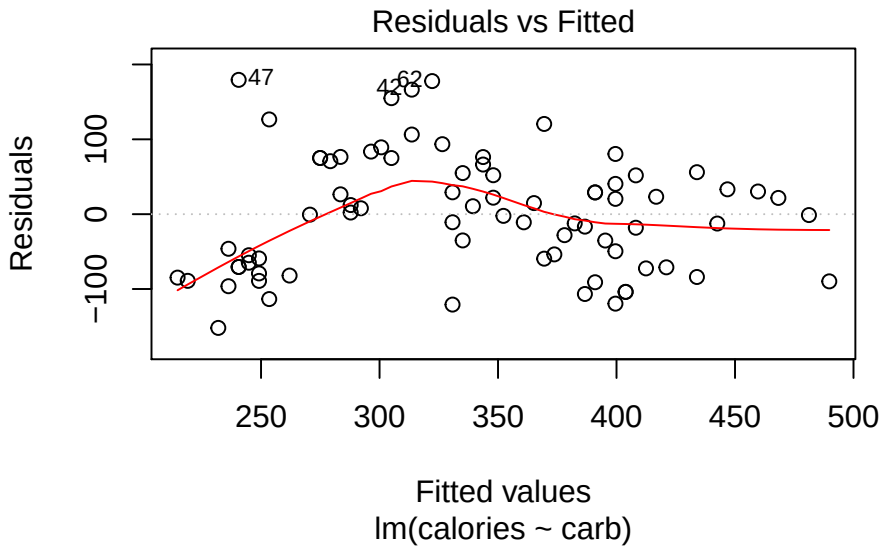

Normal Q−Q
Standardized residuals
2
0
−2
−2
−1
0
1
2
Theoretical Quantiles
lm(calories ~ carb)





Residuals vs Fitted

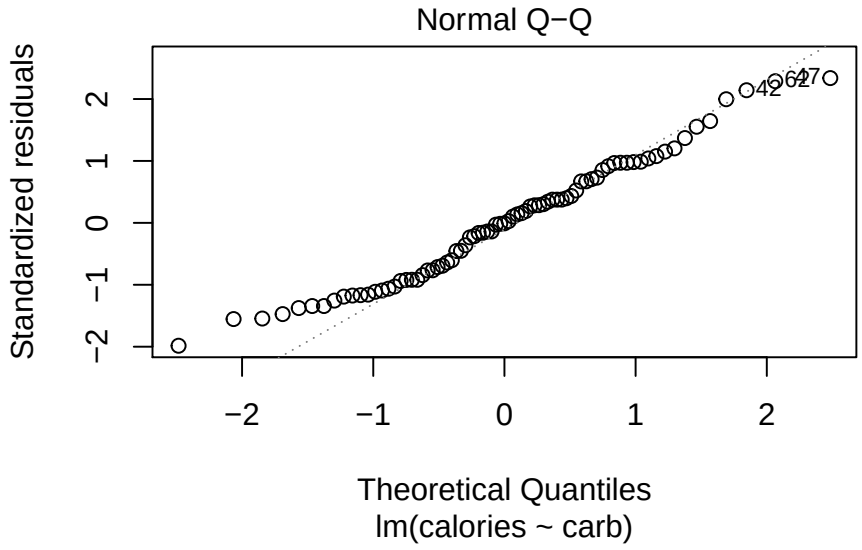
```
plot(star_mod,1)
```



This plot assesses linearity of the model and homoscedasticity (constant variance). Ideally, the red line should be centered around the dashed horizontal line. Furthermore, the scatter around the dashed line should be relatively constant across the plot. In this case, it looks like there is some minor concern over linearity and non-constant error variance. We noted this earlier with the cluster of points in the lower left hand corner of the scatterplot. The points that are labeled are points with a high residual value. They are extreme. We will discuss outliers shortly.

Normal Q-Q Plot

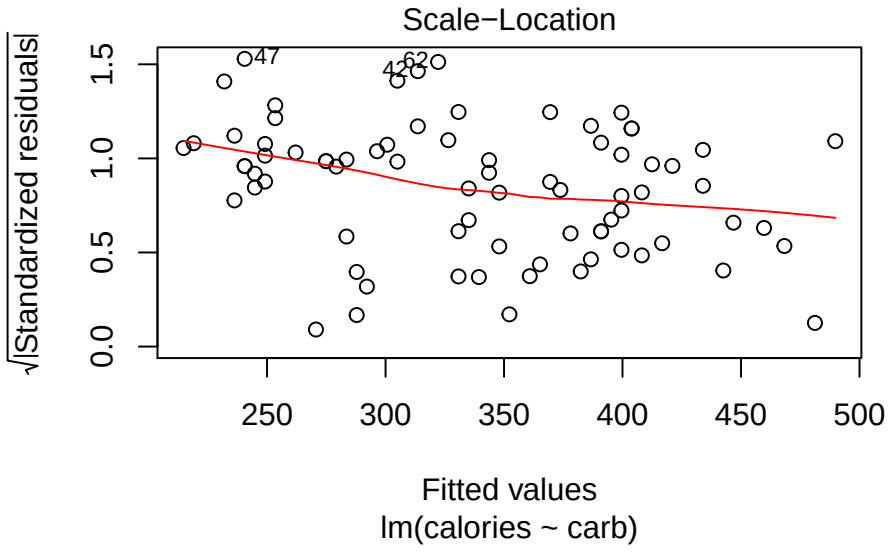
```
plot(star_mod,2)
```



As it's name suggests, this plot evaluates the normality of the residuals. Along the y-axis is the actual standardized residuals. Along the x-axis is where those should be if the residuals were actually normally distributed. Ideally, the dots should fall along the diagonal dashed line. In this case, it appears there is some positive skew. This is concerning.

Scale-Location Plot

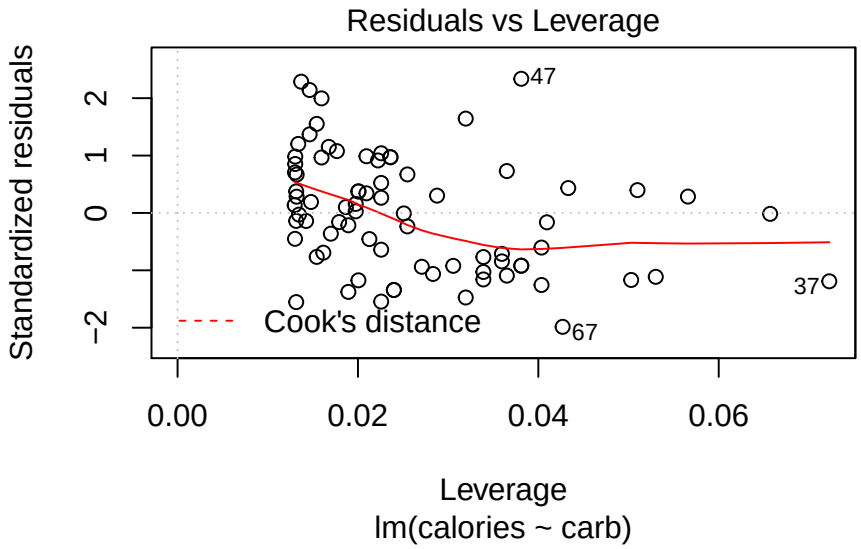
```
plot(star_mod,3)
```



The scale-location plot is a better indicator of non-constant error variance. A straight horizontal red line indicates constant error variance. In this case, there is some indication error variance is higher for lower carb counts.

Residuals vs Leverage Plot

```
plot(star_mod,5)
```



The residuals vs leverage plot is a good way to identify influential observations. Sometimes, influential observations are representative of the population, but they could also indicate an error in recording data, or an otherwise unrepresentative outlier. It could be worth looking into these cases. In this example, there are three points that may be overly influential.

Outliers and leverage

Outliers in regression are observations that fall far from the "cloud" of points. These points are especially important because they can have a strong influence on the least squares line.

> **Exercise**:
> There are six plots shown in figure below along with the least squares line and residual plots. For each scatterplot and residual plot pair, identify any obvious outliers and note how they influence the least squares line. Recall that an outlier is any point that doesn't appear to belong with the vast majority of the other points.

(1) There is one outlier far from the other points, though it only appears to slightly influence the line.

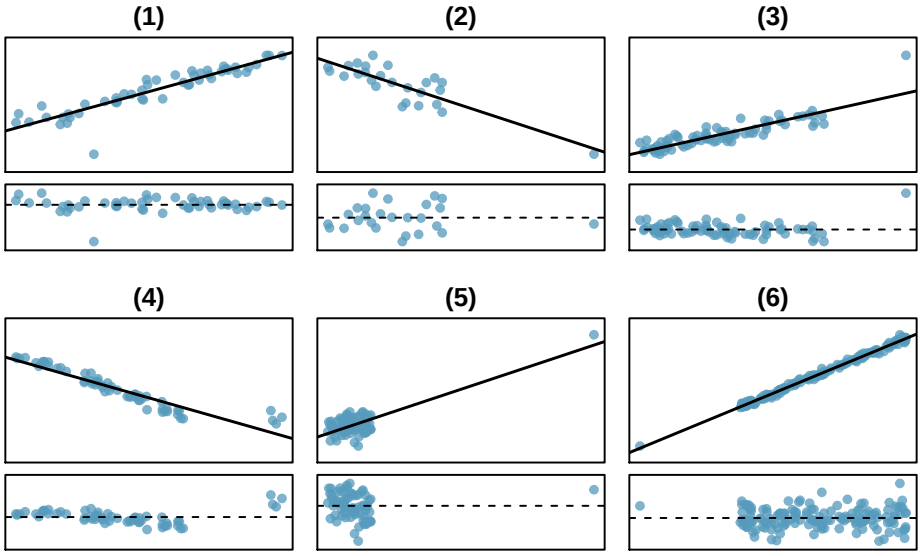
(2) There is one outlier on the right, though it is quite close to the least squares line, which suggests it wasn't very influential.

(3) There is one point far away from the cloud, and this outlier appears to pull the least squares line up on the right; examine how the line around the primary cloud doesn't appear to fit very well.

(4) There is a primary cloud and then a small secondary cloud of four outliers. The secondary cloud appears to be influencing the line somewhat strongly, making the least square line fit poorly almost everywhere. There might be an interesting explanation for the dual clouds, which is something that could be investigated.

(5) There is no obvious trend in the main cloud of points and the outlier on the right appears to largely control the slope of the least squares line.

(6) There is one outlier far from the cloud, however, it falls quite close to the least squares line and does not appear to be very influential.



Examining the residual plots in the figure, you will probably find that there is some trend in the main clouds of (3) and (4). In these cases, the outliers influenced the slope of the least squares lines. In (5), data with no clear trend were assigned a line with a large trend simply due to one outlier (!).

> Leverage Points that fall horizontally away from the center of the cloud tend to pull harder on the line, so we call them points with **high leverage**.

Points that fall horizontally far from the line are points of high leverage; these points can strongly influence the slope of the least squares line. If one of these high leverage points does appear to actually invoke its influence on the slope of the line – as in cases (3), (4), and (5) – then we call it an **influential point**. Usually we can say a point is influential if, had we fitted the line without it, the influential point would have been unusually far from the least squares line.

It is tempting to remove outliers. Don't do this without a very good reason. Models that ignore exceptional (and interesting) cases often perform poorly. For instance, if a financial firm ignored the largest market swings – the "outliers" – they would soon go bankrupt by making poorly thought-out investments.

What If Our Assumptions Are Violated

We will leave it to the reader to research what to do if there appear to be violations of assumptions. Sometimes, it is appropriate to transform the data (either response or predictor). Other times, it is appropriate

to explore other models. When confronted with clear violated assumptions, There are entire courses on regression where blocks of material are devoted to diagnostics and transformations. We will use resampling in the next lesson and it does not assume linearity.

File Creation Information

- File creation date: 2020-07-30
- Windows version: Windows 10 x64 (build 18362)
- R version 3.6.3 (2020-02-29)
- `mosaic` package version: 1.7.0
- `tidyverse` package version: 1.3.0