Applied and our with of under outlier calculate it v Lev but b to not Influence w data the Diagnostics: is loss assumptionsW the suc c the YX\text{distribution} problems in Regr SV\text{vss} T ass b resp 11 righ w an often there matrix this sectionW we again look for characteristics in the data that can cause problems with our fitted modelsY

In multiple regression, we often investigate with a scatterplot matrix first:

collinearity
outliers
marginal pairwise linearity

These scatterplots can be useful, but it only shows pairwise relationships.

After fitting the model, we check for:
nonlinearity
non-constant variance
non-normality
multicollinearity (VIFs)
outliers and influential observations

Outliers, Leverage, and Influence

- In the general sense, an outlier is any unusual data point.

- It helps to be more specific:

  ![Scatterplot](image)

  - In the above plot there is a strong outlier in the bottom right.

This is an outlier with respect to the X-distribution (independent variable). It is not an outlier with respect to the Y-distribution (dependent variable).
• Here is the fitted line after removing the high leverage point:

![Fitted line after removing high leverage point]

• This one point had a a lot of influence on the fitted line.

• The second picture shows a much better fit to the bulk of the data. But one must be careful about removing outliers. In this case it was a value that was reportedly incorrectly fitted the in picture but did fitted the in pline the influence on ivalues. A point with high leverage doesn’t HAVE to have a large impact on the fitted line.

• A point with high leverage doesn’t HAVE to have a large impact on the fitted line.

• Recall the cigarette data:

![Cigarette data]

• Above is the fitted line with the inclusion of the point of high leverage (top right data point is far outside the distribution of x-values).

• The blue point below is an outlier, but not with respect to the independent variable (not high leverage).

• The before and after in this case are quite similar. This leverage point did NOT have a large influence on the fitted model.

• Here is the fitted line after removal of the outlier, it did not have a big influence on the fitted line.

• A point with high leverage has the potential to greatly affect the fitted model.
• What about higher dimensions (more predictors)?

• A point with high leverage will be outside of the ‘cloud’ of observed x-values.

• With 2 predictors $X_1$ and $X_2$:

The point at the top left has high leverage because it is outside the cloud of $(X_1, X_2)$ independent values.

• The overall potential impact from $Y_i$ on the fitted model is found by summing all the squared $h_{ij}$ values (a sum across all $j$’s).

• The hat value for obs $i$: $h_i = \sum_{j=1}^{n} h_{ij}^2$

• How do we determine if $h_i$ is large?
  * $1/n \leq h_i \leq 1$
  * $\sum_{i=1}^{n} h_i = (k + 1)$ (where $k$ is the number of predictors)
  * $\bar{h} = (k + 1)/n$

• We can compare $h_i$ to the average. If it is much larger, then the point has high leverage.

• Often use $2(k+1)/n$ or $3(k+1)/n$ as a guide for determining high leverage.

Assessing Leverage: Hat Values

• Beyond graphics, we have a quantity called the hat value which is a measure of leverage.

• Leverage is a measure of “potential” influence.

• The hat value $h_{ij}$ captures the contribution of observation $Y_i$ to the fitted value of the $j$th observation, or $\hat{Y}_j$.

• If $h_{ij}$ is large, $Y_i$ CAN have a substantial impact on the $j$th fitted value.

In fact, we have the relationship:

$$\hat{Y}_j = h_{1j}Y_1 + h_{2j}Y_2 + \cdots + h_{nj}Y_n$$

$$= \sum_{i=1}^{n} h_{ij}Y_i$$

• In SLR, $h_i$ just measures the distance from the mean $\bar{X}$.

$$h_i = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{j=1}^{n}(X_j - \bar{X})^2}$$

• In multiple regression, $h_i$ measures distance from the centroid (center of the cloud of data points in the X-space).

• The dependent-variable values are not involved in determining leverage.
Example: Duncan data

Response: Prestige
Prestige measure of occupation.

Predictors: Income
Percent of males in occupation earning $3500 or more in 1950.

Education
Percent of males in occupation in 1950 who were high school graduates.

> attach(Duncan)
> head(Duncan)

<table>
<thead>
<tr>
<th>type</th>
<th>income</th>
<th>education</th>
<th>prestige</th>
</tr>
</thead>
<tbody>
<tr>
<td>accountant</td>
<td>62</td>
<td>86</td>
<td>82</td>
</tr>
<tr>
<td>pilot</td>
<td>72</td>
<td>76</td>
<td>83</td>
</tr>
<tr>
<td>architect</td>
<td>75</td>
<td>92</td>
<td>90</td>
</tr>
<tr>
<td>author</td>
<td>55</td>
<td>90</td>
<td>76</td>
</tr>
<tr>
<td>chemist</td>
<td>64</td>
<td>86</td>
<td>90</td>
</tr>
<tr>
<td>minister</td>
<td>21</td>
<td>84</td>
<td>87</td>
</tr>
</tbody>
</table>

> nrow(Duncan)
[1] 45

Fit the model, get the hat values:

> lm.out=lm(prestige ~ education + income)
> hatvalues(lm.out)

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>0.05092832 0.05732001 0.06963699 0.06489441</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Plot the hat values against the indices 1 to 45, and include thresholds for $2 \times (k + 1)n$ and $3 \times (k + 1)n$:

> plot(hatvalues(lm.out),pch=16,cex=2)
> abline(h=2*3/45,lty=2)
> abline(h=3*3/45,lty=2)
> identify(1:45,hatvalues(lm.out),row.names(Duncan))

Two Predictors:
> plot(education,income,pch=16,cex=2)
## The next line produces an interactive option for the
## user to identify points on a plot with the mouse.
> identify(education,income,row.names(Duncan))

From this bivariate plot of the two predictors, it looks like the ‘RR.engineer’, ‘conductor’, and ‘minister’ may have high leverage.

These three points have high leverage (potential to greatly influence the fitted model) using the 2 times and 3 times the average hat value criterion.

To measure influence, we’ll look at another statistic, but first... consider studentized residuals.
Studentized Residuals

• In this section, we’re looking for outliers in the Y-direction for a given combination of independent variables (x-vector).

• Such an observations would have an unusually large residual, and we call it a regression outlier.

• If you have two predictors, the mean structure is a plane, and you’d be looking for observations that fall exceptionally far from the plane compared to the other observations.

• First, we start with standardized residuals.

• Though we assume the errors in our model have constant variance \( [\epsilon_i \sim N(0, \sigma^2)] \), the estimated errors, or the sample residuals \( e_i \)’s DO NOT have equal variance...

We can get away from the problem by estimating \( \sigma^2 \) with a sum of squares that does not include the \( i \)th residual.

- Delete the \( i \)th observation, re-fit the model based on \( n - 1 \) observations, and get

\[
\hat{\sigma}^2(-i) = \frac{RSS}{n-1-k-1}
\]

and

\[
SE(-i) = \sqrt{\frac{RSS}{n-1-k-1}}
\]

• This gives us the studentized residual

\[
e_i^* = \frac{e_i}{SE(-i)\sqrt{1 - h_i}}
\]

• Now, the \( \hat{\sigma} \) in the denominator is not correlated with the numerator and

\[
e_i^* \sim t_{n-1-k-1}
\]

• Observations with high leverage tend to have smaller residuals. This is an artifact of our model fitting. These points can pull the fitted model close to them, giving them a tendency toward smaller residuals.

• \( \text{Var}(e_i) = \sigma^2(1 - h_i) \)

• We know...

\[
1/n \leq h_i \leq 1.
\]

• With this variance, we can form a standardized residual \( e_i' \) which all have equal variance as

\[
e_i' = \frac{e_i}{\sigma\sqrt{1 - h_i}} = \frac{e_i}{SE\sqrt{1 - h_i}}
\]

but the distribution of \( e_i' \) isn’t a \( t \) because the numerator and denominator are not independent (part of theory of \( t \)).

\[
e_i^* \sim t_{n-1-k-1}
\]

• We can now use the \( t \)-distribution to judge what is a large studentized residual, or how likely we are to get a studentized residual as far away from 0 as the ones we get.

• As a note, if the \( i \)th observation has a large residual and we left it in the computation of \( SE \), it may greatly inflate \( SE \), deflating the standardized residual \( e_i' \) and making it hard to notice that it was large.

• The standardized residuals is also referred to as the internally studentized residual.

• The studentized residual listed here is also referred to as the externally studentized residual.
• Test for outliers by comparing $e_i^*$ to a $t$ distribution with $n - k - 2$ df and applying a Bonferroni correction (multiply the p-values by the number of residuals).

• Really, we only have to be concerned about the large studentized residuals. Perhaps take a closer look at those with $e_i^* > 2$

**Example:** Returning to the Duncan model:

```r
> nrow(Duncan)
[1] 45
> lm.out=lm(prestige ~ education + income)
> sort(rstudent(lm.out))
-2.3970223990  -1.9309187757  -1.7604905300  ...
```

Can look at adjusted p-values from $t_{41}$, but there is a built-in function to help with this...

• If an observation has a big impact on the fitted model and it’s not justified to remove it, one option to report both models (with and without) commenting on the differences.

• An analysis called Robust Regression will allow you to leave the observation in, while reducing it’s impact on the fitted model.

This method essentially ‘weights’ the observations differently when computing the least squares estimates.

Get the adjusted p-value for the largest $|e_i^*|$:  

```r
> outlier.test(lm.out, row.names(Duncan))
max|rstudent| = 3.134519, degrees of freedom = 41, unadjusted p = 0.003177202, Bonferroni p = 0.1429741
```

Observation: minister

Since $p=0.1429741$ is larger than 0.05, we would conclude that this model doesn’t have any extreme residuals.

• An outlier may indicate a sample peculiarity or may indicate a data entry error. It may suggest an observation belongs to another ‘population’.

• Fitting a model with and without an observation gives you a feel for the sensitivity of the fitted model to the observation.

**Measuring influence**

• We’ve described a point with high leverage as having the potential to greatly influence the fitted model.

• To greatly influence the fitted model, a point has high leverage and it’s $Y$-value is not ‘inline’ with the general trend of the data (has high leverage and looks ‘odd’).

• High leverage + large studentized residual = High influence

• To check influence, we can delete an observation, and see how much the fitted regression coefficients change. A large change suggests high influence.

$$\text{Difference} = \hat{\beta}_j - \hat{\beta}_j(-i)$$
• Fortunately this can be done analytically.

We will use subscript (−i) to indicate quantities calculated without case i...

• DFBETAS - effect of $Y_i$ on a single estimated coefficient

$$DFBETAS_{j,i} = \frac{\hat{\beta}_j - \hat{\beta}_{j(-i)}}{SE(\hat{\beta}_{j(-i)})}$$

$|DFBETAS| > 1$ is considered large in a small or medium sized sample

$|DFBETAS| > 2n^{-1/2}$ is considered large in a big sample

• COOKSD - effect on all fitted values

Based on:

how far x-values are from the mean of the x’s

how far $Y_i$ is from the regression line

$$COOKSD = \sum_j (\hat{Y}_j - \hat{Y}_{j(-i)})^2 \frac{e_i^2h_i}{(k + 1)S_E^2}$$

$$= \frac{S_E^2}{S_E^2(k + 1)(1 - h_i)^2}$$

$$COOKSD > \frac{4}{n-k-1}$$ Unusually influential case.

Look at COOKSD relative to each other.

• DFFITS - effect of $i$th case on fitted value for $Y_i$

$$DFFITS = \frac{\hat{Y}_i - \hat{Y}_{i(-i)}}{S_E(-i)\sqrt{h_i}} = e_i^* \sqrt{\frac{h_i}{1 - h_i}}$$

$|DFFITS| > 1$ is considered large in a small or medium sized sample

$|DFFITS| > 2\sqrt{\frac{k+1}{n}}$ is considered large in a big sample

Look at DFFITS relative to each other.

• Example: Returning to the Duncan model:

```
> influence.measures(lm.out)
Influence measures of
  lm(formula = prestige ~ education + income) :
      dfb.1  dfb.edct dfb.incm  dfst cov.r cook.d hat.info
  1 -2.25e-02  0.035944  6.64e-04  0.070398  1.125  1.69e-03  0.0509
  2 -2.54e-02  -0.008118  5.09e-02  0.084067  1.311  2.41e-03  0.0573
  3 -9.19e-03  0.005619  6.48e-03  0.019768  1.155  1.33e-04  0.0696
  4 -4.72e-03  0.000140  -6.02e-05  0.000187  1.150  1.20e-08  0.0649
  5 -6.58e-02  0.086777  1.70e-02  0.192261  1.078  1.24e-02  0.0513
```

You get a DFBETA for each regression coefficient and each observation.

We’re not really interested in the intercept DFBETA though. That said, a bivariate plot of the other DFBETAS may be useful...
The ‘minister’ observation may have a large impact on both regression coefficients.

We can bring together leverage, studentized residuals and cooks distance in the following “Bubble plot”:

```
> plot(dfbetas(lm.out)[,c(2,3)],pch=16)
> identify(dfbetas(lm.out)[,2],dfbetas(lm.out)[,3],
  row.names(Duncan))
```

The Cook’s distance is perhaps most commonly looked at (effect on all fitted values).

```
> plot(cookd(lm.out),pch=16,cex=1.5)
> abline(h=4/(45-2-1),lty=2)
> identify(1:45,cookd(lm.out),row.names(Duncan))
```

• ‘conductor’, ‘minister’, and ‘RR engineer’ have high leverage
• ‘minister’ and ‘reporter’ have large regression residuals
• ‘minister’ has the highest influence on the fitted model