We’ve seen that there are certain assumptions made when we model the data.

These assumptions allow us to calculate a test statistic, and know the distribution of the test statistic under a null hypothesis.

Besides the specified distributional assumptions, there are other things to check for in regression.

In this section, we look for characteristics in the data that can cause problems with our fitted models.
In multiple regression, we often investigate with a scatterplot matrix first:

- collinearity (highly correlated $x$’s)
- outliers
- marginal pairwise linearity

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

After fitting the model, we check for:
- nonlinearity (with residual plot)
- non-constant variance
- non-normality
- multicollinearity (VIFs, yet to come in Ch. 13)
- *outliers and influential observations*
Outliers, Leverage, and Influence

- In general, an outlier is any unusual data point. It helps to be more specific...

- 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).
• A strong outlier with respect to the independent variable(s), is said to have high leverage.

• Such a point could have a big impact on the fitted model.

• Here is the fitted line for the full data set:
• Here is the fitted line after removing the high leverage point:

![Graph showing the fitted line with removed high leverage point.]

• This one point had a lot of influence on the fitted line. The $\hat{\beta}_0$ and $\hat{\beta}_1$ changed greatly in the re-fit.

• 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 reported incorrectly, and removal was justified.
• A point with high leverage doesn’t HAVE to have a large impact on the fitted line.

• Recall the 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).
• Here is the fitted line after removing the outlier:

![Graph showing the fitted line after removing the outlier.]

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

• A point with high leverage has the potential to greatly affect the fitted model.
• The blue point below is an outlier, but not with respect to the independent variable (not high leverage).

• Here is the fitted line after removal of the outlier, it did not have a big influence on the fitted line.
• 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 $X_2$ (no response shown)

![Graph showing points with high leverage]

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

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

• Leverage measures “potential” influence.

• In regression, the hat value $h_i$ (or $h_{ii}$) is a common measure of leverage.

• A high $h_i$ hat value equates to high leverage.

• What is $h_i$? How is it calculated?
  – First, why is it called a hat value?
  – In matrix notation for OLS, we have:
    $$
    \hat{Y} = X\hat{\beta} = X \left( X'X \right)^{-1} X'Y
    $$
    an $n \times n$ matrix called the Hat matrix
    $$
    \hat{Y} = HY
    $$
Thus, \( \hat{Y}_j \) is the \( j^{th} \) row of \( H \) times \( Y \)
But \( H \) is symmetrical, so the book shows it as...

\[
\hat{Y}_j = h_{1j}Y_1 + h_{2j}Y_2 + \cdots + h_{nj}Y_n
= \sum_{i=1}^{n} h_{ij}Y_i
\]

The value \( h_{ij} \) captures the contribution of observation \( Y_i \) to the fitted value of the \( j^{th} \) observation, or \( \hat{Y}_j \) (consider \( h_{ij} \) a weight)

If \( h_{ij} \) is large, \( Y_i \) CAN have a substantial impact on the \( j^{th} \) fitted value.

It can be shown that...

\[
h_i = h_{ii} = \sum_{j=1}^{n} h_{ij}^2
\]

and \( h_i \) summarizes the potential influence of \( Y_i \) on the fit of ALL other observations.
• The hat value for obs \(i\) is the \(i^{th}\) element in the diagonal of \(H\) (thus, \(h_{ii}\)).

• How do we determine if \(h_i\) is large?
  * \(\frac{1}{n} \leq h_i \leq 1\)
  * \(\sum_{i=1}^{n} h_i = (k + 1)\)
    (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.
• 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. Leverage is a statement about the X-space.
• **Example:** Duncan data

**Response:** **Prestige**

Percent of raters in an opinion study rating occupation as excellent or good in prestige.

**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
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.
Fit the model, get the hat values:

```r
> lm.out=lm(prestige ~ education + income)
> hatvalues(lm.out)
  1     2     3     4    ... 
  0.05092832 0.05732001 0.06963699 0.06489441 ... 
```

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

```r
> 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))
```
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 observation 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 \( [\varepsilon_i \sim N(0, \sigma^2)] \), the estimated errors, or the sample residuals \( e_i \)’s DO NOT have equal variance...
• 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

$$e'_i = \frac{e_i}{\hat{\sigma} \sqrt{1 - h_i}} = \frac{e_i}{S_E \sqrt{1 - h_i}}$$

but the distribution of $e'_i$ isn’t a $t$ because the numerator and denominator are not independent (part of theory of a $t$ statistic).
We can get away from the problem by estimating $\sigma^2$ with a sum of squares that does not include the $i$th residual.

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

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

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

or

$$S_E(-i) = \sqrt{\frac{RSS}{n-1-k-1}}$$

- This gives us the studentized residual

$$e_i^* = \frac{e_i}{S_E(-i)\sqrt{1 - h_i}}$$

- Now, the $\hat{\sigma}$ in the denominator, or $S_E(-i)$, is not correlated with the numerator and

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

- 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 \( S_E \), it may greatly inflate \( S_E \), 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 largest studentized residuals. Perhaps take a closer look at those with $|e_i^*| > 2$. 
• **Example**: Returning to the Duncan model:

```r
> nrow(Duncan)
[1] 45
```

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

```r
> 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 in the *car* library to help with this...

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

```r
> outlierTest(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.

• If an observation has a big impact on the fitted model and it’s not justified to remove it, one option is 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.
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 ‘in-line’ 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 (and we don’t have to re-fit the model \( n \) times).

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
• 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, in other words, look for large values.
**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

\[
COOKSD = \sum_j (\hat{Y}_j - \hat{Y}_j(-i))^2 \frac{1}{(k + 1)S_E^2}
\]

\[
= \frac{e_i^2 h_i}{S_E^2 (k + 1)(1 - h_i)^2}
\]

\[
= \left( \frac{e_i^{*2}}{k + 1} \right) \left( \frac{h_i}{1 - h_i} \right)
\]

So, **COOKSD** can be high if $h_i$ is very large (close to 1) and $e_i^{*2}$ is moderate, or if $e_i^{*2}$ is very large and $h_i$ is moderate, or if they’re both extreme.
\[ \text{COOKSD} > \frac{4}{n-k-1} \] Unusually influential case.

Look at COOKSD relative to each other.

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

```r
> influence.measures(lm.out)
Influence measures of
  lm(formula = prestige ~ education + income) :

   dfb.1_  dfb.edct  dfb.incm   dffit   cov.r  cook.d   hat   inf
     1  -2.25e-02  0.035944  6.66e-04  0.070398 1.125  1.69e-03  0.0509
     2  -2.54e-02 -0.008118  5.09e-02  0.084067 1.131  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-05  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 (using the > 1 threshold).
The Cook’s distance is perhaps most commonly looked at (effect on all fitted values).

```
> plot(cooks.distance(lm.out),pch=16,cex=1.5)
> abline(h=4/(45-2-1),lty=2)
> identify(1:45,cooks.distance(lm.out),row.names(Duncan))
```
We can bring together leverage, studentized residuals and cooks distance in the following “Bubble plot”:

```r
## Plot Residuals vs. Leverage:
> plot(hatvalues(lm.out), rstudent(lm.out), type="n")
> cook = sqrt(cooks.distance(lm.out))
> points(hatvalues(lm.out), rstudent(lm.out),
       cex = 10 * cook / max(cook))

## Include diagnostic thresholds:
> abline(h = c(-2, 0, 2), lty = 2)
> abline(v = c(2, 3) * 3/45, lty = 2)

## Overlay names:
> identify(hatvalues(lm.out), rstudent(lm.out),
          row.names(Duncan))
```
• High Leverage: conductor, minister, RR engineer
• Large regression residuals: minister, reporter
• ‘minister’ has the highest influence on the fitted model
It turns out, **R** will do some work for us in our diagnostic plotting...

```r
> par(mfrow=c(2,2))
> plot(lm.out)
```

![Residuals vs Fitted](image1.png)
![Normal Q-Q](image2.png)
![Scale-Location](image3.png)
![Residuals vs Leverage](image4.png)
1. Residuals vs. fitted  
   (checking constant variance)

2. Normal QQ plot  
   (checking normality)

3. Scale-Location plot  
   (similar to 1., but uses $\sqrt{|e_i^*|}$ vs. $\hat{Y}_i$)

4. Residuals vs. Leverage  
   (checking influence, plus it gives you a sweet contour plot of $COOKSD$)
Jointly influential data points

- The previous section considered the impact that individual data points can have on the fitted regression coefficients.

- Their impact can be seen with leave-one-out deletion diagnostics. See...

  ```
  > influence.measures(lm.out)
  ```

- But data points can have a joint influence on the fitted model.

- Graphical methods can help identify such sub-sets.

- For this purpose, we will use the Partial Regression Plot that we saw earlier (in the multiple regression introduction). Also known as Added Variable Plot.
Recall an Added Variable Plot for $X_i$

For $X_i$,

1. get the residuals from the regression of $Y$ on all predictors except $X_i$, and plot on the Y axis (what’s left for $X_i$ to explain).

2. get the residuals from the regression of $X_i$ on all other predictors, and plot on the X axis (the non-redundant part of $X_i$ in the model).

3. The plot provides information on adding the $X_i$ variable to the model.

The ‘special’ simple linear regression related to the plot above (i.e. of Residuals($Y \sim X_{(-X_i)}$) on Residuals($X_i \sim X_{(-X_i)}$)) gives the multiple regression coefficient for $X_i$.

After we fit the regression to this plot, the residuals are the same as the full model multiple regression residuals.
As this ‘special’ plot is used to fit the multiple regression coefficient, the relationship we see should be linear.

• In the Duncan data set, the data point ‘minister’ had the most influence when considering leave-one-out diagnostics.

MINISTER DATA POINT:
This data point had a large positive studentized residual... the prestige for the ‘minister’ was higher than expected for the given income and education (seems reasonable for the occupation though).
This data point also had an odd combination of income/education with a low income for their given amount of education (high leverage, but the combination again seems reasonable).

This lead to minister being highly influential.

Let’s fit the model with and without this data point...

And let’s see if it is jointly influential by considering added variable plots...
• The fitted regression coefficients **with**
  the ‘minister’ data point:
  
  > summary(lm.out)

  Coefficients:

  |            | Estimate | Std. Error | t value | Pr(>|t|) |
  |------------|----------|------------|---------|---------|
  | (Intercept)| -6.06466 | 4.27194    | -1.420  | 0.163   |
  | education  | 0.54583  | 0.09825    | 5.555   | 1.73e-06*** |
  | income     | 0.59873  | 0.11967    | 5.003   | 1.05e-05*** |

• The fitted regression coefficients **without**
  the ‘minister’ data point:

  > summary(lm(prestige[-6]~education[-6]+income[-6]))

  Coefficients:

  |            | Estimate | Std. Error | t value | Pr(>|t|) |
  |------------|----------|------------|---------|---------|
  | (Intercept)| -6.62751 | 3.88753    | -1.705  | 0.0958  |
  | education[-6]| 0.43303 | 0.09629    | 4.497   | 5.56e-05*** |
  | income[-6]| 0.73155  | 0.11674    | 6.266   | 1.81e-07*** |

The education coefficient is lower without ‘minister’.
The income coefficient is higher without ‘minister’.
(These changes are understandable when we look at the added variable plots.)
**Example**: Returning to Duncan model:

```r
> avPlots(lm.out,"education",labels=row.names(Duncan),
  id.method=cooks.distance(lm.out),id.n=2)
```

This is the plot used to fit the *education* multiple regression coefficient.

Here we see how the multiple regression coefficient for education is higher when ‘minister’ is included (influential point).
The ‘conductor’ data point influences the fitted line in the same direction (removal of both at once may show a large change in the *education* regression coefficient).

```r
> avPlots(lm.out,"income",labels=row.names(Duncan),
          id.method=cooks.distance(lm.out),id.n=2)
```

Here we see that ‘minister’ and ‘conductor’ work together to *flatten* the income multiple regression coefficient. Removal of both at once may show a large change in the *income* regression coefficient.
Removing outliers (careful)

- Removal of an outlier should only be done after careful investigation, and perhaps discussion.

- It shouldn’t be removed if it’s not an error, or can’t be shown that the data point belongs to a different ‘population’.

- Duncan Data set:
  When reporting the analysis, one could report both analyses (with and without ‘minister’). Otherwise, if all involved agree that ‘minister’ is not representative of the population at hand, it could be removed.