Topic Overview
This topic will cover

- Diagnostics and Remedial Measures
- Influential Observations and Outliers

Chapter 9: Regression Diagnostics
We now have more complicated models. The ideas (especially with regard to the residuals) of Chapter 3 still apply, but we will also concern ourselves with the detection of outliers and influential data points. The following are often used for the identification of such points and can be easily obtained from SAS:

- Studentized deleted residuals
- Hat matrix diagonals
- Dffits, Cook’s D, DFBETAS
- Variance inflation factor
- Tolerance

Life Insurance Example
- We will use this as a running example in this topic.
- References: page 364 in NKNW and nknw364.sas.
- \( Y \) = amount of insurance (in $1000)
- \( X_1 \) = Average Annual Income (in $1000)
- \( X_2 \) = Risk Aversion Score (0-10)
- \( n = 18 \) managers were surveyed.

```sas
data insurance;
  infile 'H:\System\Desktop\Ch09ta01.dat';
  input income risk amount;
proc reg data=insurance;
  model amount=income risk/r influence;
```
Just to get oriented...

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2</td>
<td>173919</td>
<td>86960</td>
<td>542.33</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>15</td>
<td>2405.147</td>
<td>160.343</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Correct Total</td>
<td>17</td>
<td>176324</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE = 12.66267, R-Square = 0.9864, Dependent Mean = 134.44444, Adj R-Sq = 0.9845, Coeff Var = 9.41851

Parameter Estimates

| Parameter | DF | Estimate | Standard Error | t Value | Pr > |t| |
|-----------|----|----------|----------------|---------|-------|
| Intercept | 1  | -205.718 | 11.392         | -18.06  | <.0001|
| income    | 1  | 6.28803  | 0.20415        | 30.80   | <.0001|
| risk      | 1  | 4.73760  | 1.37808        | 3.44    | 0.0037|

Model is significant and \( R^2 = 0.9864 \) – quite high – both variables are significant.

The Usual Residual Plots

The plot statement generates the following two residual plots (in the past we have used gplot to create these). These residuals are for the full model. Note the weird syntax
r.*(income risk). It prints the estimated equation and the $R^2$ on it automatically, which is kind of nice. This is an alternative to saving the residuals and using 

```
proc reg data=insurance;
  model amount=income risk/r partial;
  plot r.*(income risk);
```

It looks like there is something quadratic going on with income in the full model. The residuals for risk look okay.
(We should also do a qqplot.)

**Types of Residuals**

**Regular Residuals**

- $e_i = Y_i - \hat{Y}_i$ (the usual).
- These are given in the SAS output under the heading “Residual” when you use the r option in the model statement, and to store them use r = (name) in an output statement.

**Studentized Residuals**

- $e_i^* = \frac{e_i}{\sqrt{MSE \times (1-h_{i,i})}}$
• Studentized means divided by its standard error. (When you ignore the $h_{i,i}$ and just divide by Root MSE they are called semistudentized residuals.)

• Recall that $s^2\{e\} = MSE(I - H)$, so that $s^2\{e_i\} = MSE(1 - h_{i,i})$. These follow a $t_{(n-p)}$ distribution if all assumptions are met.

• Studentized residuals are shown in the SAS output under the heading “Student Residual.” In the output, “Residual” / “Std Error Residual” = “Student Residual”. SAS also prints a little bar graph of the studentized residuals so you can identify large ones quickly.

• In general, values larger than about 3 should be investigated. (The actual cutoff depends on a $t$ distribution and the sample size; see below.) These are computed using the ‘r’ option and can be stored using student=(name).

Studentized Deleted Residuals

• The idea: delete case $i$ and refit the model. Compute the predicted value and residual for case $i$ using this model. Compute the “studentized residual” for case $i$. (Don’t do this literally.)

• We use the notation $(i)$ to indicate that case $i$ has been deleted from the computations.

• $d_i = Y_i - \hat{Y}_{i(i)}$ is the deleted residual. (Also used for PRESS criterion)

• Interestingly, it can be calculated from the following formula without re-doing the regression with case $i$ removed. It turns out that $d_i = \frac{e_i}{(1 - h_{i,i})}$, where $h_{i,i}$ is the $i$th diagonal element of the Hat matrix $H$. Its estimated variance is $s^2\{d_i\} = \frac{MSE(i)}{(1 - h_{i,i})}$.

• The studentized deleted residual is $t_i = \frac{d_i}{\sqrt{s^2\{d_i\}}} = \frac{e_i}{(1 - h_{i,i})\sqrt{MSE(i)}} = \frac{e_i}{\sqrt{MSE(i)(1 - h_{i,i})}}$.

• $MSE(i)$ can be computed by solving this equation: $(n-p)MSE = (n-p-1)MSE(i) + \frac{e_i^2}{1 - h_{i,i}}$.

• The $t_i$ are shown in the SAS output under the heading “Rstudent”, and the $h_{i,i}$ under the heading “Hat Diag H”. To calculate these, use the influence option and to store them use rstudent=(name).

• We can use these to test (using a Bonferroni correction for $n$ tests) whether the case with the largest studentized residual is an outlier (see page 374).

```
proc reg data=insurance;
  model amount=income risk/r influence;
```
### Output Statistics

<table>
<thead>
<tr>
<th>Obs</th>
<th>Dep Var</th>
<th>Std Error</th>
<th>Student Residual</th>
<th>Cook's D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>91.0000</td>
<td>12.216</td>
<td>-1.206</td>
<td>0.036</td>
</tr>
<tr>
<td>2</td>
<td>162.0000</td>
<td>12.009</td>
<td>-0.910</td>
<td>0.031</td>
</tr>
<tr>
<td>3</td>
<td>11.0000</td>
<td>11.403</td>
<td>2.121</td>
<td>0.349</td>
</tr>
<tr>
<td>4</td>
<td>240.0000</td>
<td>11.800</td>
<td>-0.363</td>
<td>0.007</td>
</tr>
<tr>
<td>5</td>
<td>73.0000</td>
<td>12.175</td>
<td>-0.210</td>
<td>0.001</td>
</tr>
<tr>
<td>6</td>
<td>311.0000</td>
<td>10.210</td>
<td>1.013</td>
<td>0.184</td>
</tr>
<tr>
<td>7</td>
<td>316.0000</td>
<td>7.780</td>
<td>2.293</td>
<td>2.889</td>
</tr>
<tr>
<td>8</td>
<td>154.0000</td>
<td>11.798</td>
<td>-0.846</td>
<td>0.036</td>
</tr>
<tr>
<td>9</td>
<td>164.0000</td>
<td>12.239</td>
<td>-0.842</td>
<td>0.017</td>
</tr>
<tr>
<td>10</td>
<td>54.0000</td>
<td>12.009</td>
<td>0.0879</td>
<td>0.000</td>
</tr>
<tr>
<td>11</td>
<td>53.0000</td>
<td>11.878</td>
<td>0.415</td>
<td>0.008</td>
</tr>
<tr>
<td>12</td>
<td>326.0000</td>
<td>10.599</td>
<td>1.177</td>
<td>0.197</td>
</tr>
<tr>
<td>13</td>
<td>55.0000</td>
<td>12.050</td>
<td>0.150</td>
<td>0.001</td>
</tr>
<tr>
<td>14</td>
<td>130.0000</td>
<td>11.258</td>
<td>-1.392</td>
<td>0.171</td>
</tr>
<tr>
<td>15</td>
<td>112.0000</td>
<td>12.042</td>
<td>-0.487</td>
<td>0.008</td>
</tr>
<tr>
<td>16</td>
<td>91.0000</td>
<td>12.162</td>
<td>-1.011</td>
<td>0.029</td>
</tr>
<tr>
<td>17</td>
<td>14.0000</td>
<td>11.454</td>
<td>1.271</td>
<td>0.120</td>
</tr>
<tr>
<td>18</td>
<td>63.0000</td>
<td>12.114</td>
<td>-0.0479</td>
<td>0.000</td>
</tr>
</tbody>
</table>

### Test for Outliers Using Studentized Deleted Residuals

- should use the Bonferroni correction since you are looking at all \( n \) residuals
- studentized deleted residuals follow a \( t_{(n-p-1)} \) distribution since they are based on \( n-1 \) observations
- If a studentized deleted residual is bigger in magnitude than \( t_{n-p-1}(1 - \frac{\alpha}{2n}) \) then we identify the case as a possible outlier based on this test.
- In our example, take \( \alpha = 0.05 \). Since \( n = 18 \) and \( p = 3 \), we use \( t_{14}(0.9986) \approx 3.6214 \).
- None of the observations may be called an outlier based on this test.
- Note that if we neglected to use the Bonferroni correction our cutoff would be 2.1448 which would detect obs. 3 and 7, but this would not be correct.
- Note that “identifying an outlier” does not mean that you then automatically remove the observation. It just means you should take a closer look at that observation and check for reasons why it should possibly be removed. It could also mean that you have problems with normality and/or constant variance in your dataset and should consider a transformation.

### What to Look For

When we examine the residuals we are looking for

- Outliers
• Non-normal error distributions

• Influential observations

Other Measures of Influential Observations

The `influence` option calculates a number of other quantities. We won’t spend a whole lot of time on these, but you might be wondering what they are.

<table>
<thead>
<tr>
<th>Obs</th>
<th>D</th>
<th>H</th>
<th>DFFITS</th>
<th>Intercept</th>
<th>income</th>
<th>risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.036</td>
<td>0.0693</td>
<td>-0.3345</td>
<td>-0.1179</td>
<td>0.1245</td>
<td>-0.1107</td>
</tr>
<tr>
<td>2</td>
<td>0.031</td>
<td>0.1006</td>
<td>-0.3027</td>
<td>-0.0395</td>
<td>-0.1470</td>
<td>0.1723</td>
</tr>
<tr>
<td>3</td>
<td>0.349</td>
<td>0.1890</td>
<td>1.1821</td>
<td>0.9594</td>
<td>-0.9871</td>
<td>0.1436</td>
</tr>
<tr>
<td>4</td>
<td>0.007</td>
<td>0.1316</td>
<td>-0.1369</td>
<td>0.0770</td>
<td>-0.0821</td>
<td>-0.0410</td>
</tr>
<tr>
<td>5</td>
<td>0.001</td>
<td>0.0756</td>
<td>-0.0580</td>
<td>-0.0394</td>
<td>0.0286</td>
<td>0.0011</td>
</tr>
<tr>
<td>6</td>
<td>0.184</td>
<td>0.3499</td>
<td>0.7437</td>
<td>-0.5298</td>
<td>0.3048</td>
<td>0.5125</td>
</tr>
<tr>
<td>7</td>
<td>2.889</td>
<td>0.6225</td>
<td>3.5292</td>
<td>-0.3649</td>
<td>2.6598</td>
<td>-2.6751</td>
</tr>
<tr>
<td>8</td>
<td>0.036</td>
<td>0.1319</td>
<td>-0.3263</td>
<td>0.0816</td>
<td>0.0254</td>
<td>-0.2452</td>
</tr>
<tr>
<td>9</td>
<td>0.017</td>
<td>0.0658</td>
<td>-0.2212</td>
<td>0.0308</td>
<td>-0.0672</td>
<td>-0.0366</td>
</tr>
<tr>
<td>10</td>
<td>0.000</td>
<td>0.1005</td>
<td>0.0284</td>
<td>0.0238</td>
<td>-0.0138</td>
<td>-0.0092</td>
</tr>
<tr>
<td>11</td>
<td>0.008</td>
<td>0.1201</td>
<td>0.1490</td>
<td>0.0863</td>
<td>-0.1057</td>
<td>0.0536</td>
</tr>
<tr>
<td>12</td>
<td>0.197</td>
<td>0.2994</td>
<td>0.7801</td>
<td>-0.5820</td>
<td>0.4495</td>
<td>0.4096</td>
</tr>
<tr>
<td>13</td>
<td>0.001</td>
<td>0.0944</td>
<td>0.0468</td>
<td>0.0348</td>
<td>-0.0294</td>
<td>0.0014</td>
</tr>
<tr>
<td>14</td>
<td>0.171</td>
<td>0.2096</td>
<td>-0.7423</td>
<td>-0.2706</td>
<td>-0.2656</td>
<td>0.6269</td>
</tr>
<tr>
<td>15</td>
<td>0.008</td>
<td>0.0957</td>
<td>-0.1543</td>
<td>-0.0164</td>
<td>0.0532</td>
<td>-0.0953</td>
</tr>
<tr>
<td>16</td>
<td>0.029</td>
<td>0.0775</td>
<td>-0.2934</td>
<td>-0.1810</td>
<td>0.0258</td>
<td>0.1424</td>
</tr>
<tr>
<td>17</td>
<td>0.120</td>
<td>0.1818</td>
<td>0.6129</td>
<td>0.5803</td>
<td>-0.3608</td>
<td>-0.2877</td>
</tr>
<tr>
<td>18</td>
<td>0.000</td>
<td>0.0849</td>
<td>-0.0141</td>
<td>-0.0101</td>
<td>0.0080</td>
<td>-0.0001</td>
</tr>
<tr>
<td>*</td>
<td>0.826</td>
<td>0.3333</td>
<td>0.8165</td>
<td>1 (or 0.4714)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Cook’s Distance

• This measures the influence of case i on all of the \( \hat{Y}_i \)’s. It is a standardized version of the sum of squares of the differences between the predicted values computed with and without case i.

• Large values suggest an observation has a lot of influence. Cook’s D values are obtained via the ‘r’ option in the `model` statement and can be stored with `cookd=(name)`.

• here “large” means larger than the 50th percentile of the \( F_{p,n-p} \) distribution; for our example \( F_{3,15}(0.5) = 0.826 \).

Hat Matrix Diagonals

• \( h_{i,i} \) is a measure of how much \( Y_i \) is contributing to the prediction of \( \hat{Y}_i \). This depends on the distance between the \( X \) values for the \( i \)th case and the means of the \( X \) values. Observations with extreme values for the predictors will have more influence.
- $h_{i,i}$ is sometimes called the leverage of the $i$th observation. It always holds that $0 \leq h_{i,i} \leq 1$ and $\sum h_{i,i} = p$.

- A large value of $h_{i,i}$ suggests that the $i$th case is distant from the center of all $X$’s. The average value is $p/n$. Values far from this average (say, twice as large) point to cases that should be examined carefully because they may have a substantial influence on the regression parameters.

- For our example, $\frac{2p}{n} = \frac{6}{18} = 0.333$ so values larger than 0.333 would be considered large. Observations #6, #7, and maybe #12 seem to have a lot of influence. These can be further examined with the next set of influence statistics.

- The hat matrix diagonals are displayed with the influence option and can be stored with $h=(\text{name})$.

**DEFITS**

- Another measure of the influence of case $i$ on its own fitted value $\hat{Y}_i$. It is a standardized version of the difference between $\hat{Y}_i$ computed with and without case $i$. It is closely related to $h_{i,i}$ (consult the text for formula if you are interested). Values larger than 1 (for small to medium size datasets) or $2 \sqrt{\frac{p}{n}}$ (for large datasets) are considered influential. (In our example, $2 \sqrt{\frac{p}{n}} = 0.816$ but this is a small dataset so we would use 1).

  - these are calculated with the influence option and can be stored with $dffits=(\text{name})$.

**DFBETAS**

- A measure of the influence of case $i$ on each of the regression coefficients.

  - It is a standardized version of the difference between the regression coefficient computed with and without case $i$.

  - Values larger than 1 (for small-to-medium datasets) or $\frac{2}{\sqrt{n}}$ (for large datasets) are considered influential. In this example $\frac{2}{\sqrt{n}} = 0.4714$, but we would use 1 as a cutoff.

  - According to all these measures, observation #7 appears to be influential. This is not surprising because it has the smallest risk (1) and the highest income (79.380) of all the observations.

**Measures of Multicollinearity**

We already know about several identifying factors in dealing with multicollinearity:

- regression coefficients change greatly when predictors are included/excluded from the model

- significant $F$-test but no significant $t$-tests for $\beta$’s (ignoring intercept)
• regression coefficients that don’t “make sense”, i.e. don’t match scatterplot and/or intuition
• Type I and II SS very different
• predictors that have pairwise correlations

There are two other numerical measures that can be used: vif and tolerance

Variance Inflation Factor

• The VIF is related to the variance of the estimated regression coefficients.

\[ VIF_k = \frac{1}{1-R^2_k} \]

where \( R^2_k \) is the squared multiple correlation obtained in a regression where all other explanatory variables are used to predict \( X_k \). We calculate it for each explanatory variable.

• If this \( R^2_k \) is large that means \( X_k \) is well predicted by the other \( X \)’s. One suggested rule is that a value of 10 or more for VIF indicates excessive multicollinearity. This corresponds to an \( R^2_k \) of \( \geq 0.9 \). Use the vif option to the model statement.

Tolerance

• \( TOL = 1 - R^2_k = \frac{1}{VIF} \). A tolerance of \( < 0.1 \) is the same as a VIF > 10, indicating excessive multicollinearity. Use the TOL option to the model statement. Described in comment on p 388.

Typically you would look at either vif or tol, not both.

### Partial Regression Plots

• Also called partial residual plots, added variable plots or adjusted variable plots.

• Related to partial correlations, they help you figure out the net effect of \( X_i \) on \( Y \), given that other variables are in the model.
• One plot for each $X_i$. To get the plot, run two regressions. In the first, use the other $X$’s to predict $Y$. In the second use the other $X$’s to predict $X_i$. Then plot the residuals from the first regression against the residuals from the second regression. The correlation of these residuals was called the partial correlation coefficient.

• A linear pattern in this type of plot indicates that the variable would be useful in the model, and the slope is its regression coefficient. The plots shows the strength of a marginal relationship between $Y$ and $X_i$ in the full model. If the partial residual plot for $X_i$ appears “flat”, $X_i$ may not need to be included in the model. If they appear like a straight line (with non-zero slope), then that suggests $X_i$ should be included as a linear term, etc.

• Nonlinear relationships, heterogeneous variances, and outliers may also be detected in these plots.

• In SAS, the ‘partial’ option in the model statement can be used to get a partial residual plot. This is not a very good plot (useful for first glance, but not something you would want to publish), so it is useful to know how to create a better one.

Coding for the poor resolution plot (they’re kind of ugly):

```
proc reg data=insurance;
  model amount=income risk/r partial;
```

(The number labels on the plot are the first digit of income because we said “id income”.)

The axes are labelled amount and income, but we are actually plotting the residuals for amount (predicted by risk) vs. the residuals for income (when predicted by risk)

(The number labels on the plot are the first digit of income because we said “id income”.)

Obtaining Partial Regression Plots

```
title1 'Partial residual plot';
title2 'for risk';
symbol1 v=circle i=rl;
axis1 label=('Risk Aversion Score');
axis2 label=(angle=90 'Amount of Insurance');
proc reg data=insurance;
  model amount risk = income;
  output out=partialrisk r=resamt resrisk;
proc gplot data=partialrisk;
  plot resamt*resrisk / haxis=axis1 vaxis=axis2 vref = 0;
run;
```

The $y$-axis has the residuals for the model insur = income. The $x$-axis has the residuals for the model risk = income (i.e. treat risk as a $Y$-variable).

The residuals compared to the horizontal line are the residuals for the model that omits risk as a variable. The residuals compared to the “regression” line are the residuals for the
model that includes risk as a variable. Are the points closer to the regression line than to the x-axis? This helps decide if there is much to be gained (i.e. smaller residuals) by including risk in the model. In this case risk clearly should be included.

Similar code for income:

```plaintext
axis3 label=('Income');
title2 'for income';
proc reg data=insurance;
model amount income = risk;
   output out=partialincome r=resamt resinc;
proc gplot data=partialincome;
   plot resamt*resinc / haxis=axis3 vaxis=axis2 vref = 0;
```
The resulting plot has on the y-axis the residuals for the model \( \text{insur} = \text{risk} \), and the x-axis has the residuals for the model \( \text{income} = \text{risk} \). This is the same as the text plot.

This plot shows, first of all, that \( \text{income} \) is clearly needed in the model. Secondly, we can see that the effect of \( \text{income} \) (when \( \text{risk} \) is included) is mostly linear. Third, a close look shows that the residuals curve a bit around the straight line, so that there is a quadratic effect. However, the quadratic effect is small compared to the linear one. A quadratic term will improve the fit of the model, but it may not improve it much. We would have to weigh the improved fit vs. the interpretability and possible multicollinearity problems when deciding on the final model.

Here’s what happens when we include the square of (centered) income:

```r
data quad;
  set insurance;
  sinc = income;
proc standard data=quad out=quad mean=0;
  var sinc;
data quad;
  set quad;
  incomesq = sinc*sinc;
title1 'Residuals for quadratic model';
proc reg data=quad;
  model amount = income risk incomesq / r vif;
  plot r.*(income risk incomesq);
```

### Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3</td>
<td>176249</td>
<td>58750</td>
<td>10958.0</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>14</td>
<td>75.05895</td>
<td>5.36135</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>17</td>
<td>176324</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE: 2.31546
Dependent Mean: 134.44444
Coeff Var: 1.72224

### Parameter Estimates

| Variable  | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| | Inflation |
|-----------|----|--------------------|----------------|---------|-------|---------|------------|
| Intercept | 1  | -200.81134         | 2.09649        | -95.78  | <.0001|         | 0          |
| income    | 1  | 5.88625            | 0.04201        | 140.11  | <.0001|         | 1.35424    |
| risk      | 1  | 5.40039            | 0.25399        | 21.26   | <.0001|         | 1.08627    |
| incomesq  | 1  | 0.05087            | 0.00244        | 20.85   | <.0001|         | 1.26657    |

For the two-variable model, \( R^2 \) was 0.9864, so while this is an improvement, it does not make a big difference. Our assumptions are now more closely met, which is good, but it also appears an outlier now exists where it did not before.
Regression Diagnostics Summary

Check normality of the residuals with a normal quantile plot.
Plot the residuals versus predicted values, versus each of the $X$’s and (when appropriate) versus time
Examine the partial regression plots for each $X$ variable.
Examine

- the studentized deleted residuals (RSTUDENT in the output)
- The hat matrix diagonals
- Dffits, Cook’s D, and the DFBETAS
- Check observations that are extreme on these measures relative to the other observations
- Examine the tolerance or VIF for each $X$

If there are variables with low tolerance / high VIF, or if any of the other indications of multicollinearity problems are present, you may need to do some model building:

- Recode variables
- Variable selection
Remedial Measures (Chapter 10)

- Weighted Regression
- Robust Regression
- Nonparametric Regression
- Bootstrapping

Weighted Regression

Maximum Likelihood

\[ Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad \text{Var}(\epsilon_i) = \sigma_i^2 \]
\[ Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma_i^2) \]
\[ f_i = \frac{1}{\sqrt{2\pi\sigma_i}} e^{-\frac{1}{2} \left( \frac{y_i - \beta_0 - \beta_1 x_i}{\sigma_i} \right)^2} \]
\[ L = f_1 \times f_2 \times \cdots \times f_n - \text{likelihood function} \]

- Variance is no longer constant
- Maximization of \( L \) with respect to \( \beta \)'s.
- Equivalent to minimization of \( \sum \frac{1}{\sigma_i^2} (Y_i - \beta_0 - \beta_1 X_{i,1} - \cdots - \beta_{p-1} X_{i,p-1})^2 \)

Weighted Least Squares

- Used to deal with unequal variances:

\[ \sigma^2 \{\epsilon\} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix} \]

- Least squares minimizes the sum of the squared residuals. For WLS, we minimize instead the sum of the squared residuals each multiplied by an appropriate weight. If the error variances are known, the weights are \( w_i = 1/\sigma_i^2 \).
- Otherwise the variances need to be estimated (see discussion pages 403-405).
- The regression coefficients with weights are: \( b_W = (X'WX)^{-1}(X'WY) \) where \( W \) is a diagonal matrix of weights.
- In SAS, use a 'weight' statement in PROC REG.
Drawbacks to Weighted Least Squares

No clear interpretation for \(MSE\). \(MSE\) will be close to 1 if error variance is modeled well.

Advantages to Weighted Least Squares

Improved parameter estimates, and CI’s. Valid inference in presence of heteroscedasticity.

Determining the Weights

We try to find a relationship between the absolute residual and another variable and use this as a model for the standard deviation; or similarly for the squared residual and the variance. Sometimes it is necessary to use grouped data or approximately grouped data to estimate the variance. With a model for the standard deviation or the variance, we can approximate the optimal weights. Optimal weights are proportional to the inverse of the variance as shown above. If the data have many observations for each value of \(X\) we can get a variance estimate at each value (this happens frequently in ANOVA).

NKNW Example

- NKNW p 406 (nknw406.sas)
- \(Y\) is diastolic blood pressure
- \(X\) is age
- \(n = 54\) healthy adult women aged 20 to 60 years old

```sas
data pressure;
  infile 'H:\System\Desktop\Ch10ta01.dat';
  input age diast;
proc print data=pressure;
title1 'Blood Pressure';
symbol1 v=circle i=sm70;
proc sort data=pressure;
  by age;
proc gplot data=pressure;
  plot diast*age;
```

This clearly has non-constant variance. Run the (unweighted) regression to get residuals.

```sas
proc reg data=pressure;
  model diast=age / clb;
  output out=diag r=resid;
```

<table>
<thead>
<tr>
<th>Analysis of Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Source</td>
</tr>
<tr>
<td>Model</td>
</tr>
</tbody>
</table>

14
Use the output data set to get the absolute and squared residuals. Plot each of them (vs. X) with a smoother.

data diag;
  set diag;
  absr=abs(resid);
  sqrr=resid*resid;

proc gplot data=diag;
  plot (resid absr sqrr)*age;

The absolute value of the residuals appears to have a fairly linear relationship with age (it appears more linear than does the graph of squared residuals vs. age). Thus, we will model standard deviation as a linear function of age. (If the second graph was more linear we would model variance instead.) We will model the absolute residuals as a function of age, and use the predicted values of that regression as weights.

Predict the standard deviation (absolute value of the residual):

proc reg data=diag;
model absr=age;
output out=findweights p=shat;
data findweights;
set findweights;
wt=1/(shat*shat);

We always compute the weights as the reciprocal of the estimated variance. Regression with weights:

proc reg data=findweights;
  model diast=age / clb p;
  weight wt;
  output out = weighted p = predict;

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Squares</th>
<th>Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>83.34082</td>
<td>83.34082</td>
<td>56.64</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>52</td>
<td>76.51351</td>
<td>1.47141</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>53</td>
<td>159.85432</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 1.21302  R-Square 0.5214
Dependent Mean 73.55134  Adj R-Sq 0.5122
Coeff Var 1.64921

16
### Other Methods

#### Robust Regression

- Basic idea is to have a procedure that is not sensitive to outliers.
- Alternatives to least squares, minimize either the sum of absolute values of residuals or the median of the squares of residuals.
- Do weighted regression with weights based on residuals, and iterate.
- See Section 10.3 for details.

#### Nonparametric Regression

- Several versions
- We have used e.g. `i=sm70`
- Interesting theory
- All versions have some smoothing parameter similar to the 70 in `i=sm70`.
- Confidence intervals and significance tests not fully developed.

#### Bootstrap

- Very important theoretical development that has had a major impact on applied statistics
- Based on simulation
- Sample *with* replacement from the data or residuals and get the distribution of the quantity of interest
- CI usually based on quantiles of the sampling distribution

### Model Validation

Three approaches to checking the validity of the model.
- Collect new data: does it fit the model?
- Compare with theory, other data, simulation.
- Use some of the data for the basic analysis ("training set") and some for validity check.
Qualitative Explanatory Variables (Chapter 11)

Example include
- Gender as an explanatory variable
- Placebo versus treatment
- Insurance Co. example from previous notes (Type of company)

Two Categories
Recall from Topic 4 (General Linear Tests):
- Model: \( Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon \)
- When \( X_1 = 0 \), \( \beta_1 \) and \( \beta_3 \) terms disappear: \( Y = \beta_0 + \beta_2 X_2 + \epsilon \). For this group, \( \beta_0 \) is the intercept, and \( \beta_2 \) is the slope.
- When \( X_1 = 1 \), \( \beta_1 \) and \( \beta_3 \) terms are incorporated into the intercept and \( X_2 \) coefficient:
  
  \[
  Y = (\beta_0 + \beta_1) + (\beta_2 + \beta_3)X_2 + \epsilon
  \]

  For this group, \( \beta_0 + \beta_1 \) is the intercept, and \( \beta_2 + \beta_3 \) is the slope.
- \( H_0 : \beta_1 = \beta_3 = 0 \) is the hypothesis that the regression lines are the same.
- \( H_0 : \beta_1 = 0 \) hypothesizes the two intercepts are equal.
- \( H_0 : \beta_3 = 0 \) hypothesizes the two slopes are equal.

More Complicated Models
- If a categorical (qualitative) variable has \( k \) possible values we need \( k - 1 \) indicator variables in order to describe it.
- These can be defined in many different ways; we will do this in Chapter 16 (ANOVA).
- We also can have several categorical explanatory variables, plus interactions, etc.
- Example: Suppose we have a variable \textit{speed} for which 3 levels (high, medium, low) are possible. Then we would need two indicator variables (e.g. \( X_1 = \text{medium} \) and \( X_2 = \text{high} \)) to describe the situation.

<table>
<thead>
<tr>
<th>speed</th>
<th>( X_1 )</th>
<th>( X_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>low</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>medium</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>high</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
**Piecewise Linear Model**

At some (known) point or points, the slope of the relationship changes. We can describe such a model with indicator variables.

Examples:

- tax brackets
- discount prices for bulk quantities
- overtime wages

**Piecewise Linear Model Example**

- NKNW page 476 (nknw476.sas)
- \( Y = \) unit cost, \( X_1 = \) lot size, \( n = 8 \)
- We have reason to believe that a linear model is appropriate, but a slope change should be allowed at \( X_1 = 500 \). (Note the ‘bending’ in the plot.)
- We can do this by including an indicator variable \( X_2 \) that is 1 if \( X_1 \) is bigger than 500 and 0 otherwise and allowing it to interact with \( X_1 \).

```sas
data piecewise;
infile 'H:\System\Desktop\Ch11ta06.dat';
input cost lotsize;
symbol1 v=circle i=sm70 c=black;
proc sort data=piecewise; by lotsize;
proc gplot data=piecewise;
plot cost*lotsize;
```

**Piecewise Model**

Define a new variable \( X_2 \) which is 0 when \( X_1 \leq 500 \) and 1 when \( X_1 > 500 \). Then create an adjusted interaction term \( X_3 = X_2(X_1 - 500) \). This uses \(-500X_2\) to indicate the change in intercept and the product \( X_1X_2 \) to find the change in slope. Note that there is only one parameter since the two lines must join at \( X_1 = 500 \). We will not use \( X_2 \) explicitly in the model, just the interaction term \( X_3 \). Thus the model is

\[
Y = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \epsilon
\]

\[
= \beta_0 + \beta_1 X_1 + \beta_2 X_2(X_1 - 500) + \epsilon
\]

\[
= \beta_0 - 500\beta_2 X_2 + \beta_1 X_1 + \beta_2 X_1X_2 + \epsilon
\]

\[
= \begin{cases} 
\beta_0 + \beta_1 X_1 & X_2 = 0 \quad (X_1 \leq 500) \\
(\beta_0 - 500\beta_2) + (\beta_1 + \beta_2)X_1 & X_2 = 1 \quad (X_1 > 500)
\end{cases}
\]

Our model has
• An intercept ($\beta_0$)
• A coefficient for lot size (the slope $\beta_1$)
• An additional explanatory variable that will add a constant to the slope whenever lot size is greater than 500.

data piecewise; set piecewise;
  if lotsize le 500
    then cslope=0;
  if lotsize gt 500
    then cslope=lotsize-500;
proc print data=piecewise;

<table>
<thead>
<tr>
<th>Obs</th>
<th>cost</th>
<th>lotsize</th>
<th>cslope</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.75</td>
<td>300</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>4.40</td>
<td>340</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>4.52</td>
<td>400</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>3.77</td>
<td>480</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>3.55</td>
<td>570</td>
<td>70</td>
</tr>
<tr>
<td>6</td>
<td>2.57</td>
<td>650</td>
<td>150</td>
</tr>
<tr>
<td>7</td>
<td>2.49</td>
<td>720</td>
<td>220</td>
</tr>
<tr>
<td>8</td>
<td>1.39</td>
<td>800</td>
<td>300</td>
</tr>
</tbody>
</table>

The variable cslope is our $X_3$. Run the regression:

proc reg data=piecewise;
  model cost=lotsize cslope;
  output out=pieceout p=coefficient;

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DF</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

20
Model 2 9.48623 4.74311 79.06 0.0002
Error 5 0.29997 0.05999
Corrected Total 7 9.78620

Root MSE 0.24494  R-Square 0.9693
Dependent Mean 3.43000  Adj R-Sq 0.9571
Coeff Var 7.14106

<table>
<thead>
<tr>
<th>Parameter Estimates</th>
<th>Parameter</th>
<th>Standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable</td>
<td>DF</td>
<td>Estimate</td>
</tr>
<tr>
<td>Intercept</td>
<td>1</td>
<td>5.89545</td>
</tr>
<tr>
<td>lotsize</td>
<td>1</td>
<td>-0.00395</td>
</tr>
<tr>
<td>cslope</td>
<td>1</td>
<td>-0.00389</td>
</tr>
</tbody>
</table>

Plot data with fitted values:

symbol1 v=circle i=none c=black;
symbol2 v=none i=join c=black;
proc sort data=pieceout; by lotsize;
proc gplot data=pieceout;
    plot (cost costhat)*lotsize/overlay;