Lecture 10: Assumption Diagnostics

Reading Assignment:

- Muller and Fetterman, Chapter 7: "GLM Assumption Diagnostics" (Required)

Wait, what are those assumptions agagin?

Homogeneity, Independence, Linearity, Existence, and Gaussian distribution.

Checking model assumptions focuses on analysis of residuals and outliers. Ethically, one must avoid allowing exploratory analysis (data snooping) to bias confirmatory analysis (inflate α). It is, however, acceptable to examine outliers or to transform variables to get the best test of the Gaussian distribution of residuals.

The First Step: Get to Know Your Data

Checking the Basics

- Determine the sampling unit (child, family, mouse, basketball team)
- Investigate the data collection procedure (counting number of vegetable servings consumed daily, measuring tumor size, measuring body mass index (BMI))
- Obtain units of measurement for all variables (number, cubic centimeters, kilograms per square meter)
- Obtain plausible range of values (0-10 servings, 0-1000 cubic centimeters, 17-40 kg per square meter)
- Obtain typical values for all variables (2 servings, 25 cubic cm, 25 kg per square meter)

Examining Individual Values

- Print the first 50 or so observations (PROC PRINT DATA=OZONE (OBS=50);)
- Calculate the minimum and maximum values for all numeric variables (PROC MEANS MIN MAX;)
- Generate frequency tables for character variables (PROC FREQ)

Summarizing the Data

- Report descriptive statistics (mean, median, quartiles, largest and smallest few observations - PROC UNIVARIATE)
- Generate appropriate graphical displays (histograms, stem and leaf plots, etc.using PROC UNIVARIATE)
- Look for patterns and unusual values

Check correlations, covariances, plot X_j versus Y, X_j versus $X_{j'}$

Example: Basic Diagnostics for the Ozone Data

The following SAS code is used to provide basic diagnostics.

```
proc print data=ozone (OBS=50);
run;

proc means min max data=ozone;
run;

proc univariate plot normal data=ozone;
var personal;
run;

proc capability data=ozone;
qqplot personal outdoor home time_out;
histogram time_out;
run;

proc corr noprob data=ozone;
var personal outdoor home time_out;
run;
```

The art in reporting these diagnostics lies in doing enough to provide necessary information with sufficiently few pages to be read (if you are

bored with your report, certainly everyone else is!). See Example 7.1 in the textbook for a nice illustration.

Selected diagnostics from the ozone data are presented below.

1. Basics: the sampling units in the ozone data are young children. Personal ozone exposures were measured using a portable monitor attached to the child's clothing, outdoor ozone exposures were measured at EPA monitoring stations, indoor (home) exposures were measured using monitors in each child's home, and the proportion of time spent outdoors was reported by a parent. Ozone exposures are measured in parts per billion. The plausible range of time spent outdoors would be between 0 and 1.
2. Individual values: in the list of data, one subject spent 90% of time outdoors, leading the biostatistician to make a note to ask the investigator if this variable is "proportion of play time spent outdoors" because 90% of one day would be 21.6 hours.
3. Data Summary: personal ozone exposures range from 0.46 to

67.18 ppb, with one outlier on the boxplot (67.18 ppb, which is not disturbing in a sample of 64 observations) and a mean and median in the range of 22-24 ppb. This variable does not look perfectly Gaussian, which is not disturbing as the Gaussian assumption applies not to Y unconditionally but to Y conditional on the covariates. (We should describe the covariates in a similar manner, noting that we need not assume they follow Gaussian distributions.)

4. Further details: personal ozone exposure appears to be most highly correlated with home exposure, and we do not see evidence of collinearity of covariates from the correlations.

Violations of Independence

The independence assumption applies to all statistical tests we have discussed thus far. The consequences of violating the independence assumption range from finding spurious relationships to missing significant ones.

Common violations of the independence assumption include

- repeated measures
  - weight loss of individuals is monitored over time
  - CD-4 counts of AIDS patients are recorded over the course of the disease
- nested data
  - schools are randomly assigned different physical education programs with student fitness as the outcome

- family data
  - researchers at NIEHS are studying sisters and daughters of women with breast cancer
  - pregnant rats are exposed to toxins and litters are studied for malformations

In general, independence must be evaluated by careful consideration of the study design. Warning signs include "too many" data points (often lab scientists run multiple assays on tissues from one specimen, for example) or "strange" results.

Residual Analysis

Recall that the raw residuals are estimates of errors for the model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

The HILE Gauss assumptions refer to errors ($\boldsymbol{\varepsilon}$), which correspond to $Y|X$ and not Y unconditionally.

Thus to evaluate the assumptions we study the residuals $\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$, with $\widehat{\varepsilon}_i = y_i - \widehat{y}_i$.

Example: Residuals

Suppose we plan to sample undergraduate males at UNC and measure their lung function.

What might the density (histogram) of their lung function look like?

(What fraction smoke?) Y may be bimodal (if so, certainly not normal!). However, errors, if we include the amount smoked in our model, may be Gaussian.

Properties of Residuals

Recall that the raw residuals, which are differences between the data points and fitted values, follow a singular normal distribution

$$\widehat{\boldsymbol{\varepsilon}} \sim \mathcal{SN}_n\{\mathbf{0}, \sigma^2[\boldsymbol{I} - \boldsymbol{H}]\}$$

and that we can estimate σ^2 using the residuals as follows:

$$\widehat{\sigma}^2 = \sum_{i=1}^{n} \widehat{\varepsilon}_i^2/(n-r).$$

The covariance among residuals, $\sigma^2[\boldsymbol{I} - \boldsymbol{H}]$, implies non-zero correlation and non-independence among residuals, though such correlations are modest in many settings.

If the model spans an intercept, then $\overline{\widehat{\varepsilon}} = \sum_{i=1}^{n} \widehat{\varepsilon}_i/n = 0$. In this case, it is clear that residuals are *not* independent!

Standardized Residuals

Because $\text{Var}(\widehat{\varepsilon}_i) = \sigma^2(1 - h_i)$, the residuals may have different variances because the variance of the residuals depends on the pattern of covariate values.

It would be easier to use the residuals to evaluate model assumptions if they all had the same variance. For this reason, we construct *standardized residuals*

$$r_i = \frac{\widehat{\varepsilon}_i}{\sqrt{\widehat{\sigma}^2(1 - h_i)}}.$$

The standardized residuals (also called internally studentized residuals) will follow a Gaussian distribution with mean 0 and variance 1 in large samples. Because of this property, one rule of thumb is to examine any observations with standardized residuals > 2 in absolute value in further detail (the justification is that 95% of the mass of the Gaussian distribution lies between -1.96 and 1.96, so that studentized residuals with absolute values greater than 2 represent observations in the

extreme tails of the distribution). Standardizing residuals thus helps us to assess their magnitude relative to the precision of the estimated regression analysis. Although these residuals are standardized (and thus have mean zero and variance one), they do depend on a variance estimate that may be affected by outliers. Because of this, we may miss those outliers in a residual analysis.

The jack-knifed or studentized residuals provide a solution to this problem based on more robust variance estimates. Recall that the studentized residuals (our preference) are defined as

$$r_{(-i)} = \frac{\widehat{\varepsilon}_i}{\sqrt{\widehat{\sigma}^2_{(-i)}(1-h_i)}} \sim T(n-r-1),$$

where the jack-knifed estimate of σ^2 is more robust to outliers than the usual estimate. These residuals are also called externally studentized residuals.

Note that a sample set of standardized or studentized residuals have approximate mean zero and approximate variance one. As $(n-r)$ becomes large, $T \rightarrow$Gaussian. Studentized residuals can still be thought of in "standard deviation" units so that the usual cutoff of 2 is approximately valid.

Evaluating Assumptions with (or Without) Residuals

- Homogeneity: violations seen in the pattern of residuals.
- Independence: assessed through logic of sampling scheme.
- Linearity: examine pattern of residuals.
- Existence: (finite sample...).
- Gaussian distribution: distributional assessment involves box plot of residuals, histogram of residuals, and test of Gaussian distribution of residuals. (The discrepancy between T and Gaussian random variables somewhat inflates the probability of rejecting the null...why?)

Sometimes, Gaussian errors, homogeneity, and linearity come and go as a package.

Evaluating Heterogeneity and Linearity

Plotting $\{r_{(-i)}\}$ vs. $\{\widehat{y}_i\}$ in an *R/P plot* provides the most useful diagnostic because predicted values capture all the information in predictors that is available as a linear combination of them. The R/P plot allows us to assess both heterogeneity and linearity. The R/P plot should resemble a rectangular cloud with no obvious trends or pattern. When examining an R/P plot, look for the following indicators of problems:

- any trend, such as a tendency for negative residuals for small predicted values and positive residuals for large predicted values, which may indicate non-linearity
- non-constant spread of the residuals, such as a tendency for tightly clustered residuals for small predicted values and widely dispersed residuals for large predicted values, which may indicate heteroscedasticity.

One issue in multiple regression is the detection of *multivariate* outliers. The R/P plot replaces a vector of predictors with one univariate predicted value. This is just one way to move from the multivariate predictor space to a univariate space.

A *probability plot* or *Q-Q* plot graphs the ordered residuals versus the expected order statistics of the standard normal distribution. To create a Q-Q plot, we order the residuals from smallest to largest, letting $r_{(1)}$ be the smallest, $r_{(i)}$ the i^{th} smallest, and $r_{(n)}$ be the largest. As you may recall from BIOS 550/660, $r_{(i)}$ is called an *order statistic*.

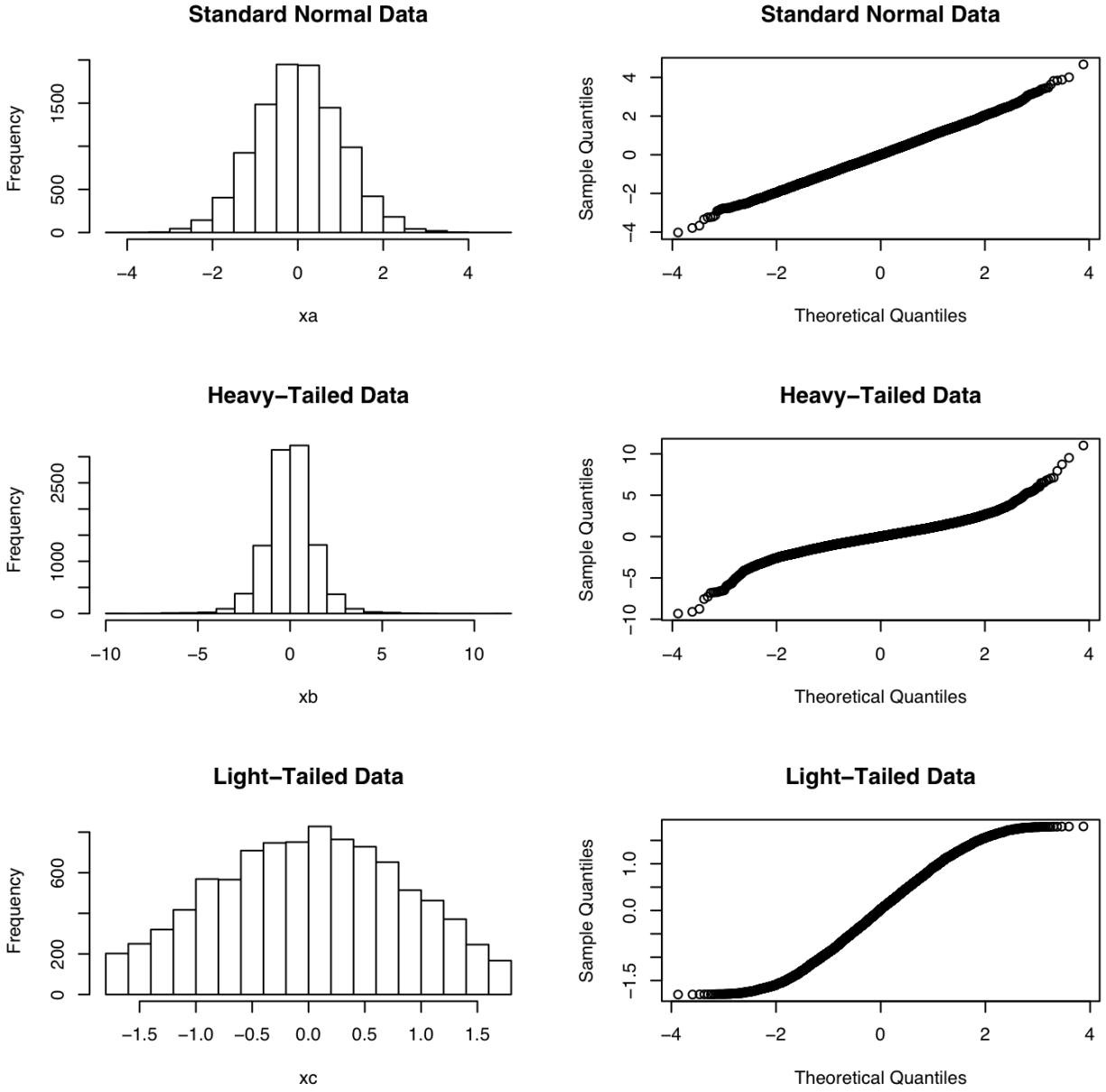
Next, imagine taking a sample of size n from a standard normal distribution and ordering it to obtain the order statistics $z_{(i)}$, $i = 1, \ldots, n$. Using results from probability theory, we can obtain the expected values of these order statistics. (For example, we expect the median to be 0, the 5^{th} percentile to be -1.96, and the 95^{th} percentile to be 1.96.)

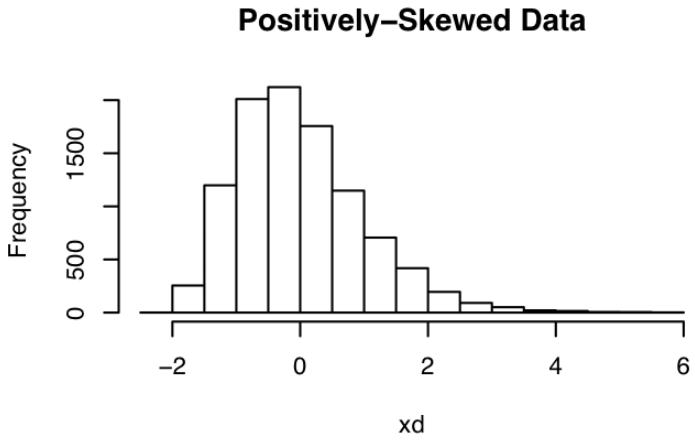
If our observations come from a Gaussian distribution, we expect our observed order statistics to be pretty close to the expected order statistics from the standard normal distribution. Specifically, if we plot the observed order statistics against the expected ones, we expect to

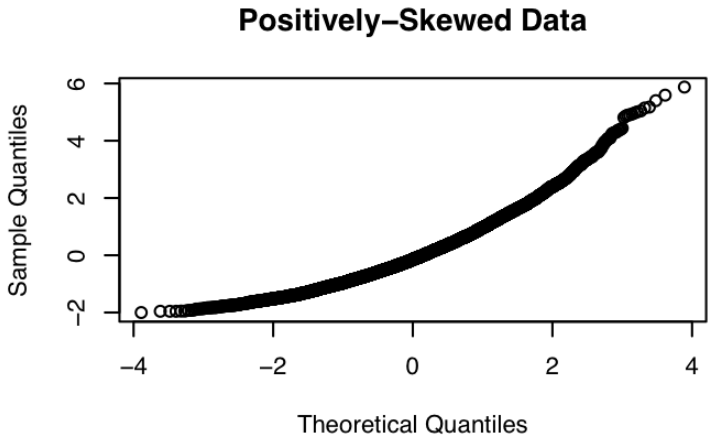
see a straight line through the origin with unit slope. When this is not the case, the observed order statistics deviate from what we would expect from a standard normal distribution, and the Gaussian errors assumption may be violated.

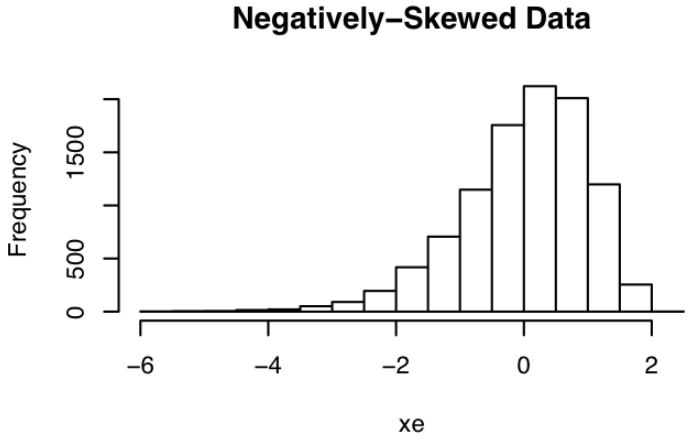
For example, if the residuals are from a distribution with heavy tails, we expect to see values consistently above a straight line through the origin with unit slope in the upper tail and consistently below the line in the lower tail. If the residuals are from a distribution with light tails, then we expect to see values consistently below the line in the upper tail and above the line in the lower tail.

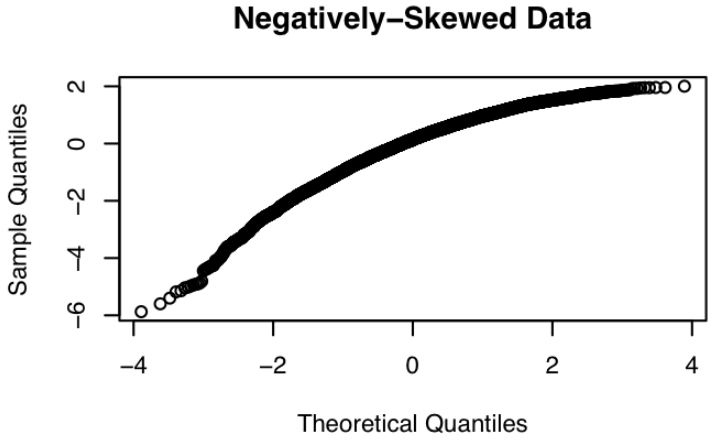
A curve that is concave upwards indicates positively-skewed data, and a curve that is concave downwards indicates negatively-skewed data.


Standard Normal Data
Frequency
0
500
1500
−4
−2
0
2
4
xa
Standard Normal Data
Sample Quantiles
−4
−2
0
2
4
−4
−2
0
2
4
Theoretical Quantiles
Heavy−Tailed Data
Frequency
0
1000
2500
−10
−5
0
5
10
xb
Heavy−Tailed Data
Sample Quantiles
−10
−5
0
5
10
−4
−2
0
2
4
Theoretical Quantiles
Light−Tailed Data
Frequency
0
200
600
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
xc
Light−Tailed Data
Sample Quantiles
−1.5
0.0
1.0
−4
−2
0
2
4
Theoretical Quantiles


Positively–Skewed Data
Frequency
0
500
1500
−2
0
2
4
6
xd


Positively–Skewed Data
Sample Quantiles
−2
0
2
4
6
−4
−2
0
2
4
Theoretical Quantiles


Negatively–Skewed Data
Frequency
0
500
1500
−6
−4
−2
0
2
xe


Negatively–Skewed Data
Sample Quantiles
−6
−4
−2
0
2
−4
−2
0
2
4
Theoretical Quantiles

The following code creates an R/P plot as well as a Q-Q plot of the studentized residuals.

```
proc reg data=ozone;
model personal=outdoor home time_out;
output out=out rstudent=studresid predicted=predicted h=leverage;
run;

proc plot data=out;
plot studresid*predicted/vref=0;
run;

proc capability data=out;
qqplot studresid;
run;
```

Now we consider a variety of example R/P plots in order to see what types of plots are acceptable and what types of plots are cause for concern.

Figure 1: Acceptable residuals: simulated $x_i \sim \mathcal{N}(0,1)$ independent of $\varepsilon_i \sim \mathcal{N}(0,1)$ with true model $y_i = x_i + \varepsilon_i$ and fit valid model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

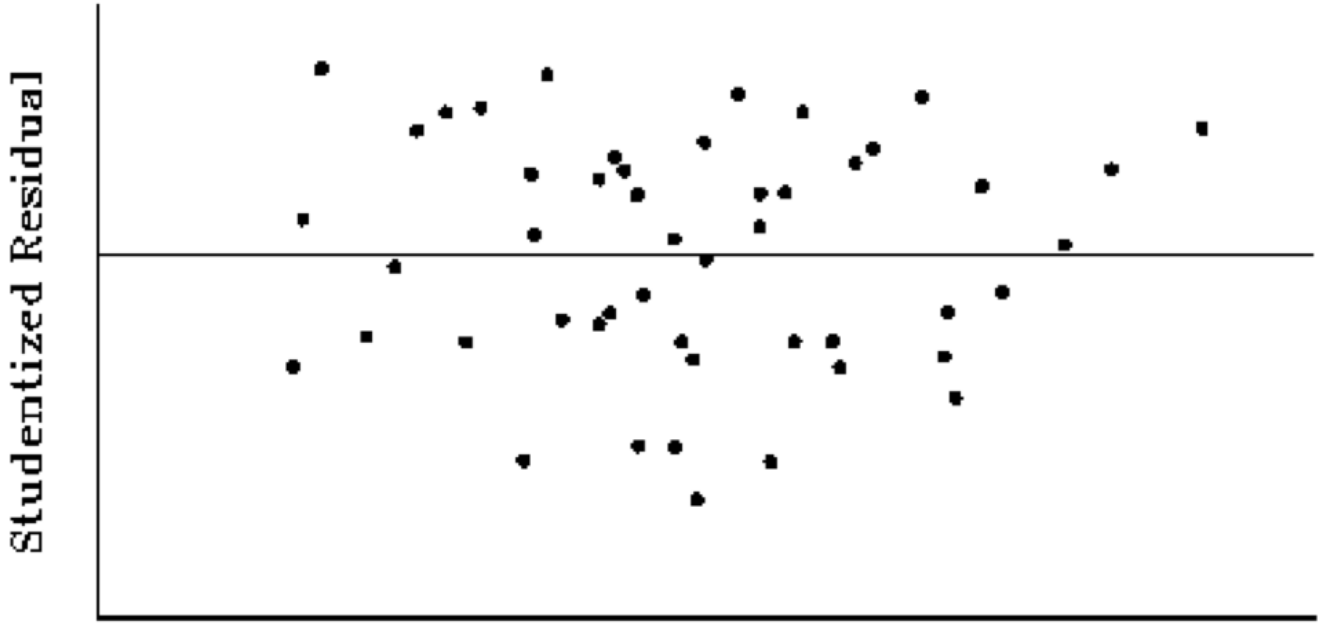


Figure 2: Unacceptable residuals (violation of homogeneity): simulated $x_i \sim U(0,1)$ independent of $\varepsilon_i \sim \mathcal{N}(0,1)$ with true model $y_i = x_i + x_i\varepsilon_i$ and fit invalid model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

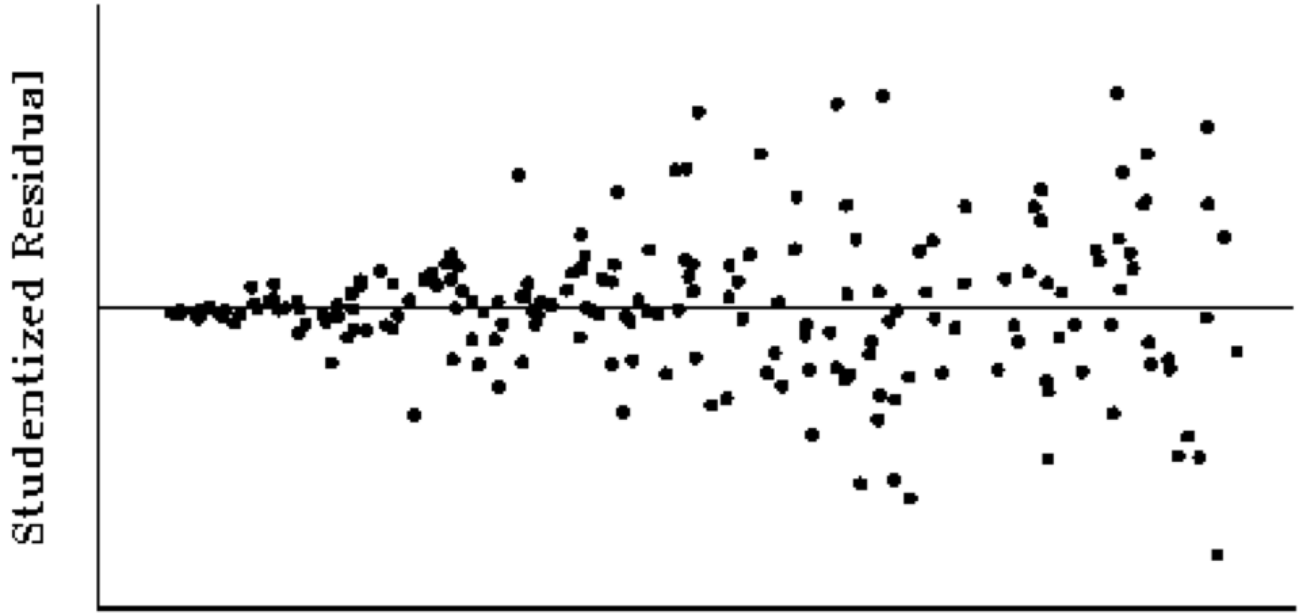


Figure 3: Unacceptable residuals (violation of homogeneity): simulated $x_i \sim U(0,1)$ independent of $\varepsilon_i \sim \mathcal{N}(0,1)$ with true model $y_i = x_i + (1 - x_i)\varepsilon_i$ and fit invalid model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

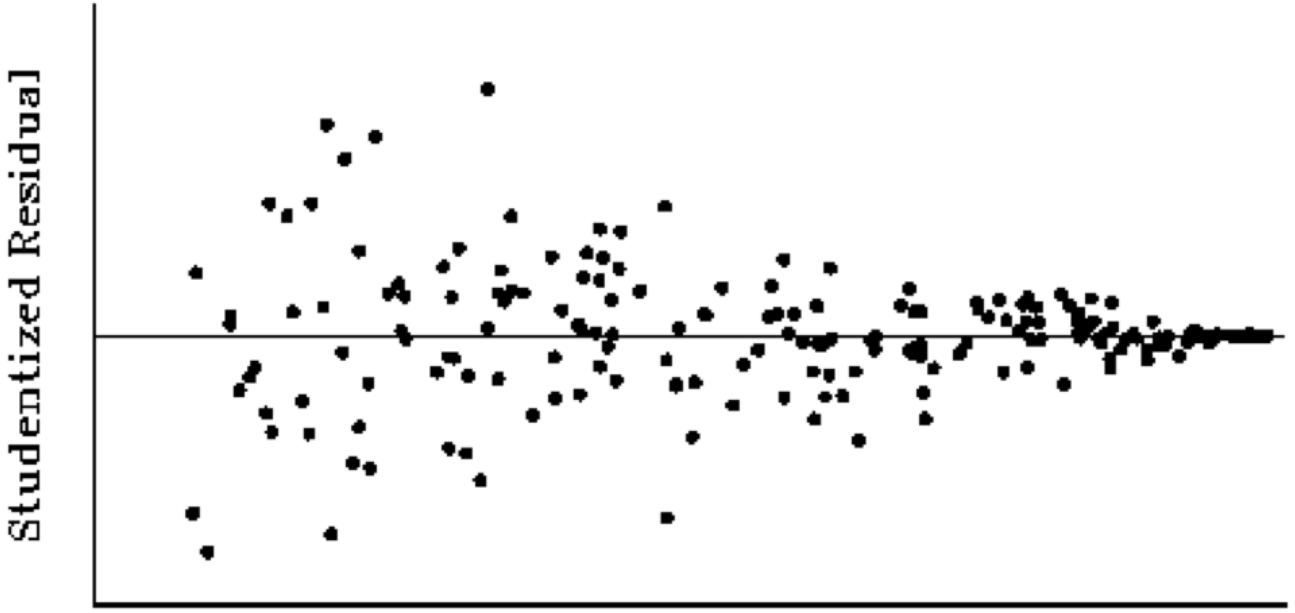


Figure 4: Unacceptable residuals (violation of homogeneity): simulated $x_i \sim U(0,1)$ independent of $\varepsilon_i \sim \mathcal{N}(0,1)$ with true model $y_i = x_i + 2 \mid 0.5 - x_i \mid \varepsilon_i$ and fit invalid model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

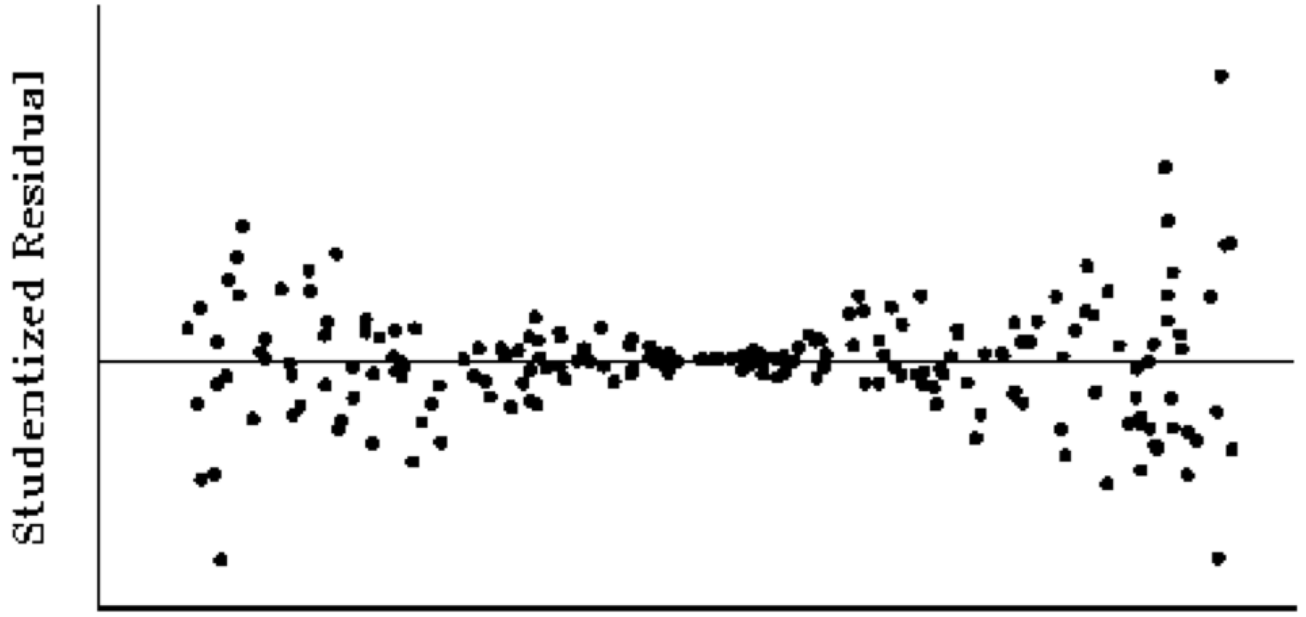


Figure 5: Unacceptable residuals (violation of homogeneity): simulated $x_i \sim U(0,1)$ independent of $\varepsilon_i \sim \mathcal{N}(0,1)$ with true model $y_i = x_i + (1 - \mid 0.5 - x_i \mid)\varepsilon_i$ and fit invalid model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

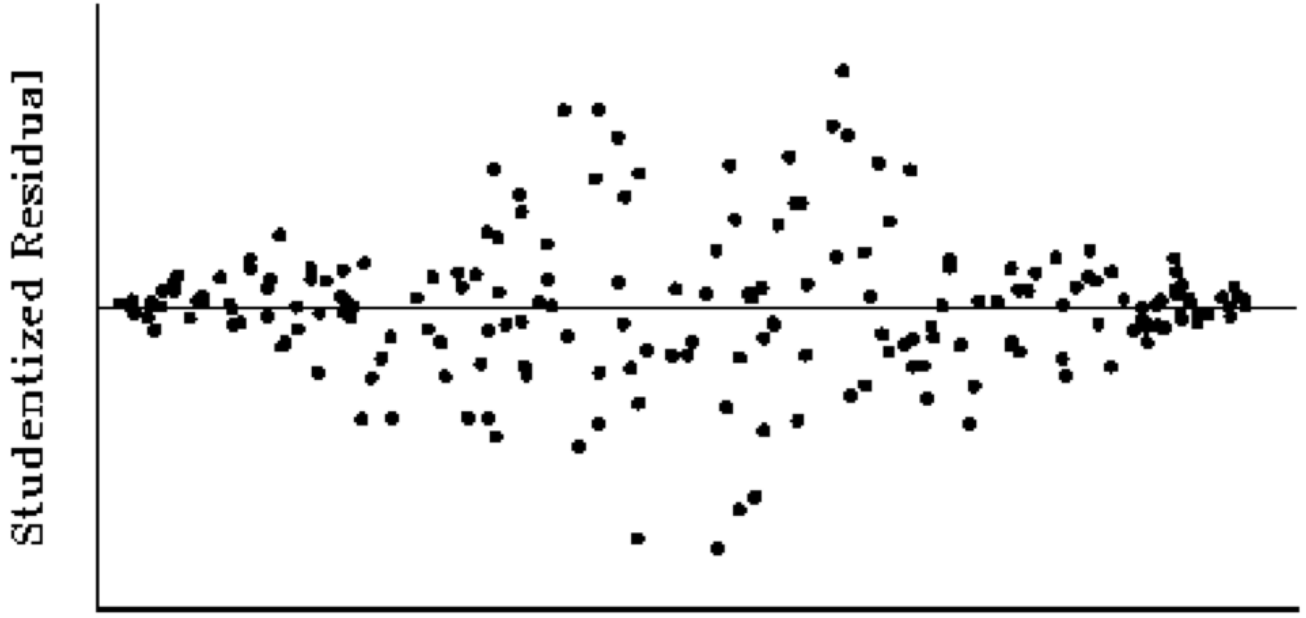
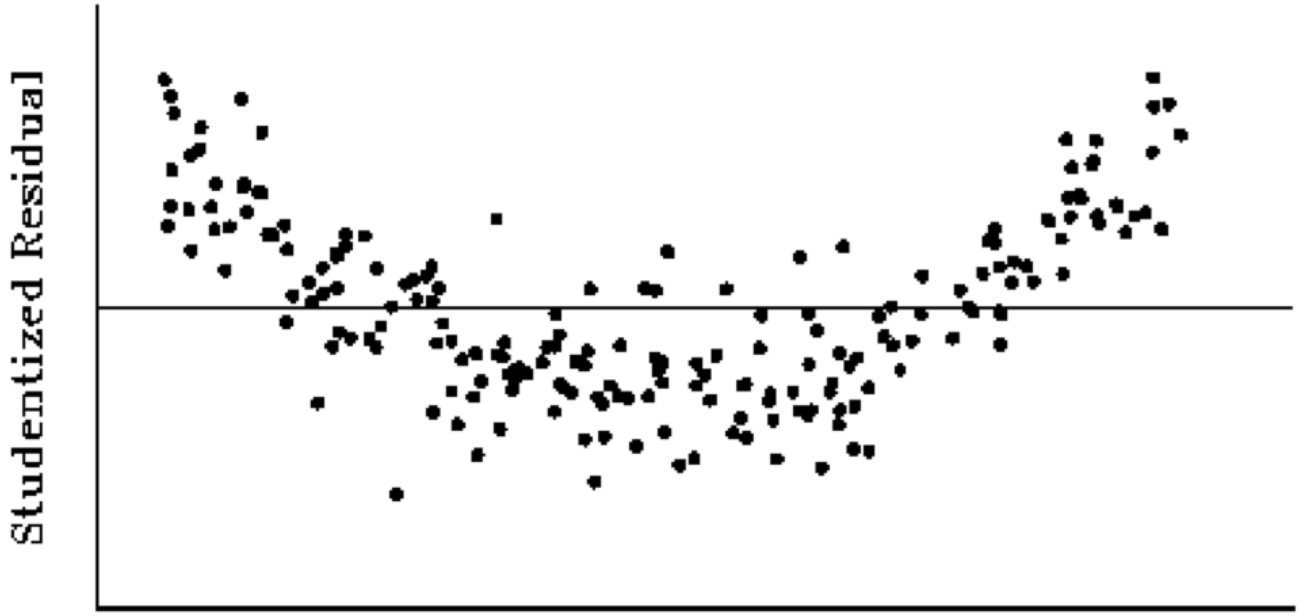


Figure 6: Unacceptable residuals (violation of linearity): simulated $x_i \sim U(0,1)$ independent of $\varepsilon_i \sim \mathcal{N}(0,1)$ with true model $y_i = x_i^2 + 0.05\varepsilon_i$ and fit invalid model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.



Some analysts plot the response versus each individual predictor, a tactic that ignores the impact of other predictors. A better idea is to create the *partial plot* of $\{\mathbf{y}\}$ versus $\{\mathbf{x}_j | \{\mathbf{x}_1, \ldots \mathbf{x}_{j-1}, \mathbf{x}_{j+1} \ldots, \mathbf{x}_p\}\}$, the residuals from a regression with $\mathbf{x}_j$ as response and the other covariates as the predictors. The univariate correlation in the plot is the semi-partial correlation

$$r(\mathbf{y}, \mathbf{x}_j \mid \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, \ldots, \mathbf{x}_p\}).$$

These plots are most useful for identifying culprits in known problems and not for detecting any new problems.

Evaluating Gaussian Distribution

SAS PROC UNIVARIATE provides the following information:

- a Q-Q ("normal") plot (use PROC CAPABILITY to get a nicer one),
- a box and whisker plot,
- a stem and leaf plot (a very basic histogram),
- moments (mean, variance, skewness, kurtosis),
- the largest and smallest 5 values with ID's, and
- quartiles.

Example: Ozone Data Residuals

The following code is used to evaluate the residuals for the ozone data from the model

$$O_{PERSONAL} = \beta_0 + \beta_1 O_{OUTDOOR} + \beta_2 O_{HOME} + \beta_3 TIME_{OUT} + \varepsilon.$$

```
proc reg data=ozone;
```

```
model personal=outdoor home time_out;
output out=out rstudent=studresid predicted=predicted;
run;

proc univariate plot normal data=out;
id subject;
var studresid;

proc capability data=out;
qqplot studresid;
run;

proc capability data=out;
histogram studresid/kernel;
run;
```

```
************************************************************************
                      The UNIVARIATE Procedure
   Variable:  studresid  (Studentized Residual without Current Obs)

                                Moments

   N                          64    Sum Weights               64
   Mean                0.00618379    Sum Observations    0.39576231
   Std Deviation        1.0298702    Variance            1.06063263
   Skewness            0.90708362    Kurtosis              0.893602
```

Uncorrected SS	66.8223027	Corrected SS	66.8198554
Coeff Variation	16654.3631	Std Error Mean	0.12873377

Basic Statistical Measures

Location		Variability	
Mean	0.00618	Std Deviation	1.02987
Median	-0.33120	Variance	1.06063
Mode	.	Range	5.12982
		Interquartile Range	1.02182

Tests for Location: Mu0=0

Test		-Statistic-		-----p Value------	
Student's t	t	0.048035	Pr > |t|		0.9618
Sign	M	-4	Pr >= |M|		0.3817
Signed Rank	S	-133	Pr >= |S|		0.3779

Tests for Normality

Test		--Statistic---		-----p Value------
Shapiro-Wilk	W	0.935601	Pr < W	0.0024
Kolmogorov-Smirnov	D	0.148292	Pr > D	<0.0100

```
Cramer-von Mises      W-Sq  0.279391    Pr > W-Sq  <0.0050
Anderson-Darling      A-Sq  1.565583    Pr > A-Sq  <0.0050
```

We see that all the normality tests agree that the residuals do not appear to be normal. (P-value shopping is not advised – it is best to pick one test before doing the analysis.) The differences among the tests are described below. For all tests, the null hypothesis is H_0 : the data are normally distributed.

- Shapiro-Wilks: roughly, a measure of the straightness of a Q-Q plot (appropriate for small sample sizes)
- Kolmogorov-Smirnov: largest discrepancy between the empirical cdf and the estimated hypothesized one (based on observed mean and variance)
- Cramer-von Mises: considers squared difference between empirical and estimated cdf
- Anderson-Darling: considers a weighted squared difference between empirical and estimated cdf

We can look at the quantiles, boxplot, and histogram to check whether residuals are skewed.

```
Quantiles (Definition 5)

Quantile       Estimate

100% Max       3.054968
99%            3.054968
95%            2.155071
90%            1.483757
75% Q3         0.388099
50% Median    -0.331204
25% Q1        -0.633717
10%           -0.907937
5%            -1.151776
1%            -2.074857
0% Min        -2.074857
```

We see some evidence of skewness in the quantiles.

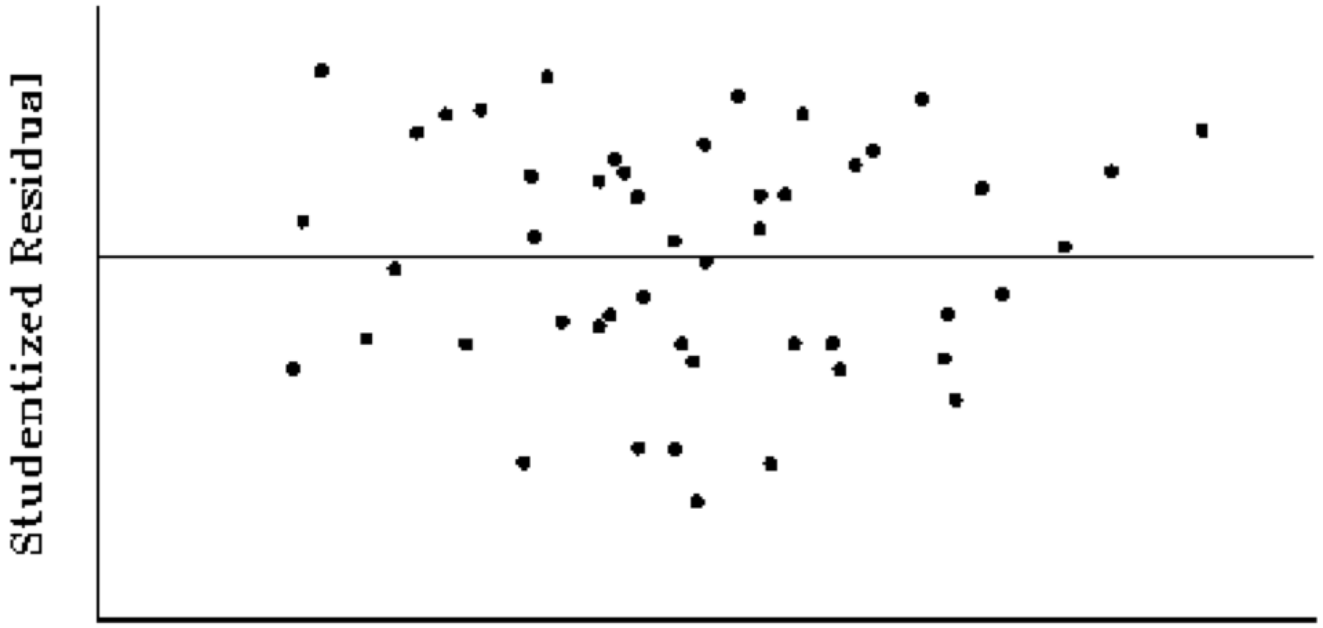
In addition, we can look at the most extreme residuals to gauge whether any outliers deserve further attention.

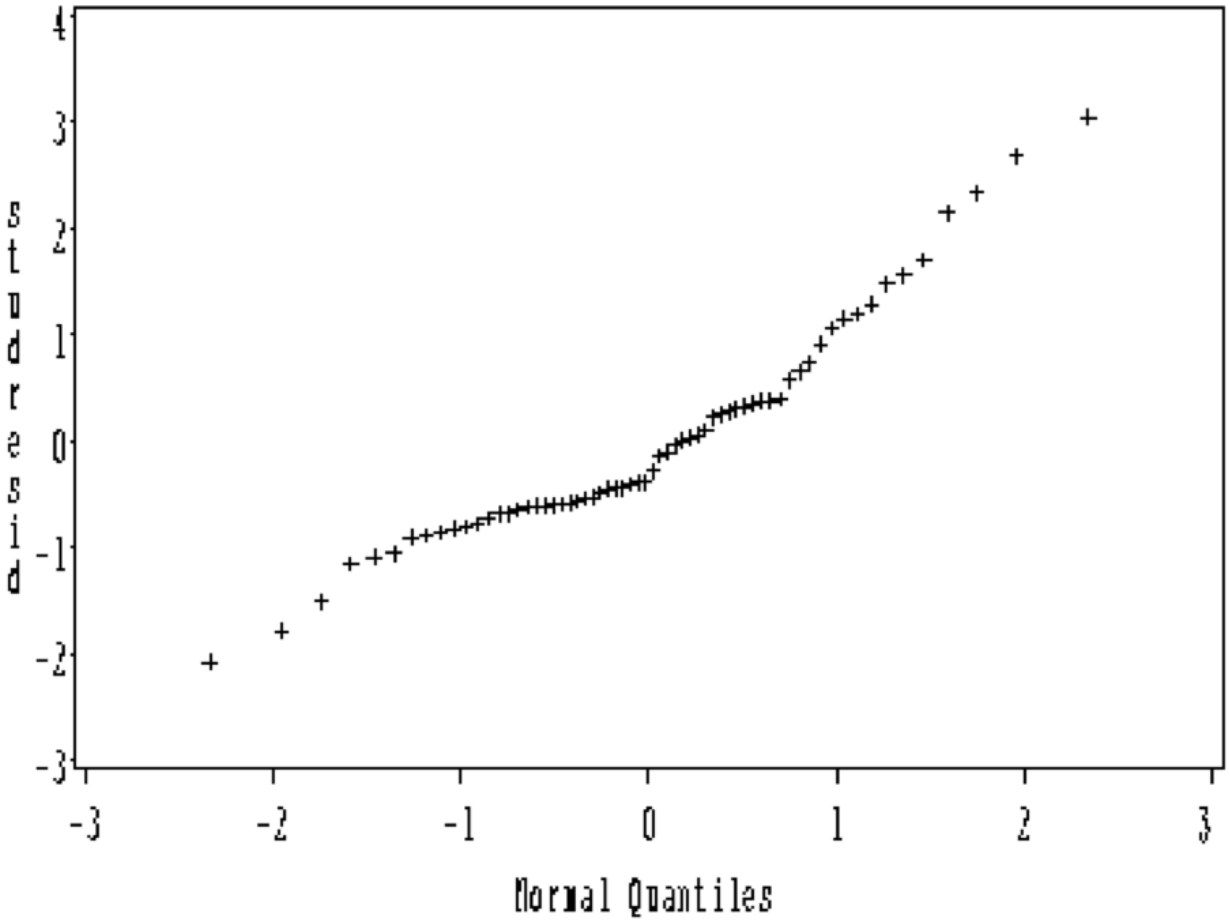
```
Extreme Observations
```

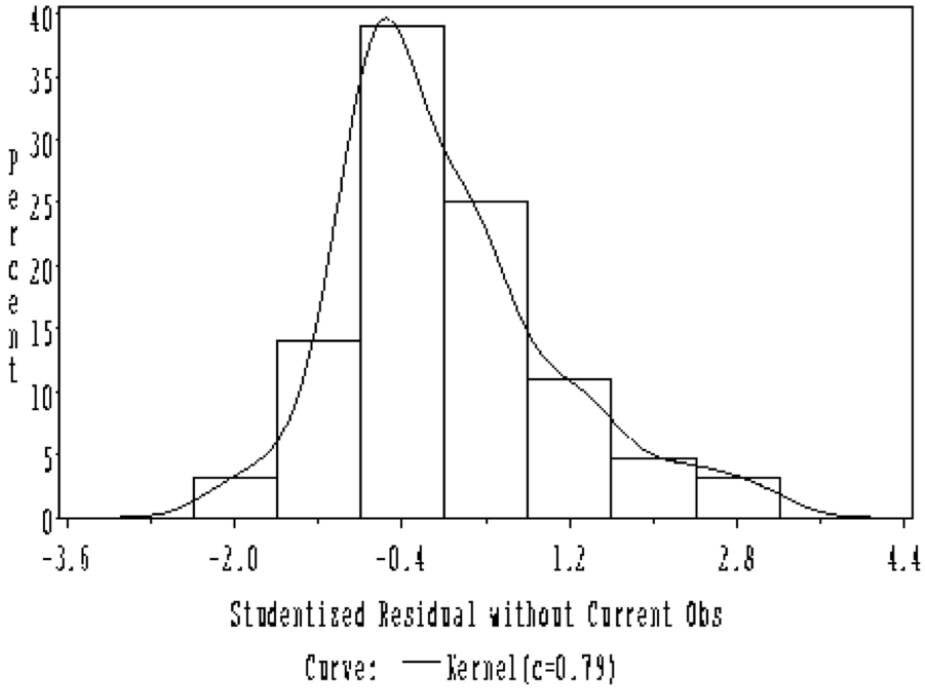
```
-----------Lowest----------        ----------Highest----------

   Value    subject      Obs           Value    subject      Obs

-2.07486         25       25         1.70690         49       49
-1.78561         36       36         2.15507         64       64
-1.51369          4        4         2.33176         17       17
-1.15178          2        2         2.67945         39       39
-1.08553         20       20         3.05497         26       26
```

There do seem to be some outliers on the right.

```
Stem Leaf                     #             Boxplot
  18
  16 1                        1                |
  14 87                       2                |
  12 09                       2                |
  10 65                       2                |
   8 1                        1                |
   6 53                       2                |
   4 8                        1                |
   2 358135889                9             +-----+
   0 0360                     4             |  +  |
  -0 415                      3             |     |
  -2 998                      3             *-----*
  -4 9863385441              10             |     |
  -6 729852110                9             +-----+
  -8 19520                    5                |
 -10 596                      3                |
 -12                                           |
 -14 1                        1                |
 -16 9                        1                |
 -18                                           |
 -20 7                        1                |
     ----+----+----+----+
 Multiply Stem.Leaf by 10**-1
```


Studentized Residual


4
3
2
1
0
-1
-2
-3
stdresid
-3
-2
-1
0
1
2
3
Normal Quantiles



From this output, we see that the data appear to be skewed to the right a little bit, and there is some peakedness in the distribution. Again, we see several outliers in the right tail of the distribution. A *kernel density estimate* is superimposed on the histogram in PROC CAPABILITY in order to facilitate interpretation.

Evaluating Extreme Residuals

One residual must be the most extreme, but too many extreme residuals (or too extreme residuals) are evidence of poor model fit.

Testing residuals (the null hypothesis is H_0: $E[r_{(-i)}] = 0$) leads to a multiple comparisons issue: we do n tests (one for each residual), so we should use a Bonferroni correction (use α/n instead of α and the $\alpha/(2n)$ critical value for T).

Any positive correlation among residuals makes the correction conservative.

If $(n - r) \to \infty$ then $\sigma^2(\boldsymbol{I} - \boldsymbol{H}) \to \sigma^2\boldsymbol{I}$, implying that correlations among residuals converge to zero for large samples.

Example: Extreme Residuals in Ozone Data

Based on previous analyses, do any residuals appear to be extreme? If so, which ones? What do you conclude based on a test of the hypothesis that any extreme residuals come from a population with mean 0?

Outliers

An *outlier* is a value (of a predictor or a response) much larger in absolute value than next nearest value. On a box plot, outliers are often marked using 0 or $*$.

Least squares estimation is rather sensitive to outliers.

An outlier may be an anomaly or merely a chance event (e.g.,a child reports 27 vegetable servings per day, or height is reported as 60 feet rather than 60 inches).

Automatically discarding extreme observations is WRONG! Exceptions: verifiable instrument malfunction, recording error.

Consider whether an outlier is plausible, implausible, or impossible.

- A woman weighs 250 pounds, 400 pounds, or 1000 pounds...
- Body temperature of 36.8 (centigrade or Farenheit? average

person, heart surgery patient, or crocodile?)

Extreme residuals may indicate an anomaly, while extreme values of Y or X_j need not be important.

Leverage

A *leverage* value depends only on $\mathbf{X}$ and measures how extreme the ith observation is in terms of the predictor space. An observation with high leverage has the potential to have great influence on the model fit.

The hat matrix, $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$, has ith diagonal element h_i, the *leverage* for the ith observation.

$h_i = \boldsymbol{X}_i(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_i'$,with $\boldsymbol{X}_i$ $1 \times p$, the ith row of $\boldsymbol{X}$

For $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, with $s_x^2 = \sum_{i=1}^n (x_i - \overline{x})^2/n$,

$$h_i = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{n s_x^2} = \left[1 + \frac{(x_i - \overline{x})^2}{s_x^2}\right]/n\,.$$

With an intercept and all $p-1$ predictors mean zero and uncorrelated,

$$h_i = \frac{1}{n} + \sum_{j=1}^{p-1} \frac{x_{ij}^2}{n s_{x_j}^2} = \left[1 + \sum_{j=1}^{p-1} \frac{x_{ij}^2}{s_{x_j}^2}\right]/n\,.$$

We can prove $\sum_{i=1}^n h_i = r$, the rank of $\boldsymbol{X}$, so full rank $\boldsymbol{X}$ implies $\overline{h} = \sum_{i=1}^n h_i/n = p/n$.

A handy rule of thumb is to examine further any subjects with $h_i > \frac{2p}{n}$, which is leverage greater than twice the average value.

With an intercept and $(p-1)$ multivariate Gaussian predictors (NOT part of HILE Gauss!) with each row i.i.d., we see that a function of

the leverage values follow an F distribution,

$$F_i = \frac{(h_i - 1/n)/(p-1)}{(1-h_i)/(n-p)} = F(p-1, n-p)\,.$$

Multiple testing leads to using a Bonferroni correction and using α/n.

This F statistic tests the hypothesis that the observation in question comes from a multivariate Gaussian distribution with the same mean as the other observations.

If the model does not span an intercept, then

$$F_i = \frac{h_i/p}{(1-h_i)/(n-p)} \sim F(p, n-p)\,.$$

Example: Leverage Values for Ozone Data

The following code is used to print and test leverage values.

```
proc reg data=ozone;
model personal=outdoor home time_out;
output out=out rstudent=studresid predicted=predicted h=leverage;
run;
proc sort data=out;
by descending leverage;
run;
data out_l;
set out (obs=10);
p=4; n=64;
F=((leverage-(1/n)/(p-1)))/((1-leverage)/(n-p));
pvalue=1-probf(F,p-1,n-p);
if pvalue <=0.05/n then BONF="*";
                   else BONF=" "; *Bonferroni correction;
label Bonf="Signif at 0.05/n?";
run;
proc print data=out_l uniform label noobs;
var subject personal outdoor home time_out leverage F pvalue Bonf;
run;
```

```
*************************************************************************
           Personal                          Home Indoor     Proportion
```

subject	Ozone Exposure (ppb)	Outdoor Ozone Concentration (ppb)	Ozone Concentration (ppb)	of Time Spent Outdoors
9	28.55	92.563	7.14	0.26
13	38.28	30.435	35.38	0.69
2	14.63	43.792	13.97	0.90
12	37.58	104.100	45.10	0.42
35	11.81	20.649	5.71	0.78
7	31.97	88.863	44.12	0.05
42	61.19	54.318	46.04	0.45
17	53.18	70.003	11.13	0.65
8	32.45	81.916	45.74	0.33
30	13.94	71.878	9.51	0.38

Leverage	F	pvalue	Signif at 0.05/n?
0.21178	15.7246	.000000117	*
0.15601	10.7210	.000009731	*
0.15214	10.3980	.000013240	*
0.14355	9.6918	.000026235	*
0.13310	8.8519	.000060328	*
0.12067	7.8783	.000162823	*

```
0.11869     7.7259    .000190717       *
0.11679     7.5798    .000222079       *
0.09578     6.0096    .001193472
0.09544     5.9852    .001225971
```

How do you interpret the results?

Mahalanobis Distance

Assume the model spans an intercept and let $\mathbf{X}^*$ indicate the $p-1$ predictors other than the intercept (removing the intercept column from the $\mathbf{X} = \begin{bmatrix} \mathbf{J}_n & \mathbf{X}^* \end{bmatrix}$ matrix).

Define $\mathbf{C}^*$ as the $(p-1) \times (p-1)$ sample covariance matrix for the predictors in $\mathbf{X}^*$:

$$\mathbf{C}^* = \frac{\mathbf{X}^{*\prime}\left(\mathbf{I}_n - \frac{\mathbf{J}_n\mathbf{J}_n^{\prime}}{n}\right)\mathbf{X}^*}{n}.$$

(Note that the sample variances of the predictors lie on the diagonal of $\mathbf{C}^*$.)

Compute the unbiased estimate of population covariances as

$$\widehat{\mathbf{\Sigma}^*} = \left(\frac{n}{n-1}\right)\mathbf{C}^*.$$

The *Malhalanobis distance*

$$m_i = (\boldsymbol{X^*}'_i - \overline{\boldsymbol{x^*}})' \widehat{\boldsymbol{\Sigma}^*}^{-1} (\boldsymbol{X^*}'_i - \overline{\boldsymbol{x^*}})$$

is the deviation of one observation's predictors from the center of the predictor space.

If all predictors are uncorrelated (in the sample at hand), then

$$m_i = \sum_{j=1}^{p-1} \frac{(x^*_{ij} - \overline{x^*}_j)^2}{\widehat{\sigma}^2_{x^*_j}} \, .$$

The Essential Equivalence of Leverage and Malhalanobis Distance

Both leverage and Malhalanobis distance measure the "extremeness" of an observation in $\mathbf{X}$ space.

For any model that spans an intercept,

$$h_i = n^{-1} + (n-1)^{-1} m_i.$$

Thus only leverage or Malhalanobis distance need be examined (leverage is more commonly used).

Influence: Cook's Distance

Cook proposed an influence measure based on the extent to which parameter estimates would change if we had deleted the i^{th} observation from the sample. *Cook's distance* assesses the impact of deleting one observation from the sample.

Why is this a sensible diagnostic?

Cook's statistic measures the standardized shift in predicted values and the shift in $\widehat{\boldsymbol{\beta}}$ due to deleting the ith observation:

$$\begin{aligned} D_i &= \frac{(\widehat{\boldsymbol{y}}_{(-i)} - \widehat{\boldsymbol{y}})'(\widehat{\boldsymbol{y}}_{(-i)} - \widehat{\boldsymbol{y}})}{p\widehat{\sigma}^2} \\ &= \frac{(\widehat{\boldsymbol{\beta}}_{(-i)} - \widehat{\boldsymbol{\beta}})'(\boldsymbol{X}'\boldsymbol{X})(\widehat{\boldsymbol{\beta}}_{(-i)} - \widehat{\boldsymbol{\beta}})}{p\widehat{\sigma}^2} \\ &= r_i^2 \cdot \frac{h_i}{p(1-h_i)} \end{aligned}$$

where r_i is standardized residual. As a quadratic form, $D_i \geq 0$. Distributional results are not straightforward, and there is no perfect rule of thumb for evaluating any particular D_i. One proposed rule of thumb is that values of D_i close to 1 are indicative of excessive influence.

DFBETAS and DFFITS

Two other common measures of influence are DFBETAS and DFFITS.

The *DFFITS* statistic is a scaled measure of the change in the predicted value of the i^{th} observation if that observation were omitted from the analysis. That is,

$$DFFITS_i = \frac{\widehat{y}_i - \widehat{y}_{(-i)}}{\sqrt{\widehat{\sigma}^2_{(-i)} h_i}}.$$

A general rule of thumb is to examine further any observations with $DFFITS_i > 2\sqrt{\frac{p}{n}}$.

The *DFBETAS* statistic is a normalized measure of the effect of observations on the estimated regression coefficients. There are multiple DFBETAS statistics for each subject (one for each regression

coefficient). They are computed as

$$DFBETAS_{ij} = \frac{\widehat{\beta}_j - \widehat{\beta}_{j(-i)}}{\sqrt{\widehat{\sigma}^2_{(-i)}(\mathbf{X}'\mathbf{X})^{-1}_j}}.$$

One typically examines observations with $DFBETAS_{ij} > \frac{2}{\sqrt{n}}$.

Concluding Comments

"One should be cautioned that deleting the most deviant observations will in all cases slightly improve, and sometimes substantially improve, the fit of the model. One must be careful not to data snoop simply in order to polish the fit of the model by discarding troublesome data points." (KKM, 1988, p201)

KKM: Kleinbaum, Kupper, Muller, and Nizam, *Applied Regression Analysis and Other Multivariable Methods*

One *must* report any deletions, and observations should be deleted only in the most extreme circumstances.

Leverage evaluates design and extremeness in predictor space, while residuals evaluate model adequacy. Cook's D_i values evaluate influence (both of above).

It may be argued that a measure of adequacy of a sample is that no

observation is influential, even though it may be high in leverage and have an extreme residual.

Use $\{h_i\}$, $\{m_i\}$, $\{r^2_{(-i)}\}$, $\{\widehat{\boldsymbol{\beta}}_{(-i)}\}$, and $\{\widehat{\sigma}^2_{(-i)}\}$ to define "bad data," not $\boldsymbol{y}$ and $\boldsymbol{X}$.

Next: Computation Diagnostics

Reading Assignment:

- Muller and Fetterman, Chapter 8: "GLM Computation Diagnostics"