Introduction

Every statistical method is developed based on assumptions. The validity of results derived from a given method depends on how well the model assumptions are met. Many statistical procedures are “robust”, which means that only extreme violations from the assumptions impair the ability to draw valid conclusions. Linear regression falls in the category of robust statistical methods. However, this does not relieve the investigator from the burden of verifying that the model assumptions are met, or at least, not grossly violated. In addition, it is always important to demonstrate how well the model fits the observed data, and this is assessed in part based on the techniques we’ll learn in this lecture.
Different types of residuals

Recall that the residuals in regression are defined as $y_i - \hat{y}_i$, where $y_i$ is the observed response for the $i$th observation, and $\hat{y}_i$ is the fitted response at $x_i$.

There are other types of residuals that will be useful in our discussion of regression diagnostics. We define them on the following slide.

---

Different types of residuals (cont.)

**Raw residuals:** $r_i = y_i - \hat{y}_i$

**Standardized residuals:** $z_i = \frac{r_i}{s}$ where $s$ is the estimated error standard deviation (i.e. $s = \hat{\sigma} = \sqrt{MSE}$).

**Studentized residuals:** $r^*_i = \frac{z_i}{\sqrt{1-h_i}}$ where $h_i$ is called the *leverage*. (More later about the interpretation of $h_i$.)

**Jackknife residuals:** $r_{(-i)} = r^*_i \frac{s}{s_{(-i)}}$ where $s_{(-i)}$ is the estimated error standard deviation computed with the $i$th observation deleted.
Which residual to use?

The standardized, studentized and jackknife residuals are all scale independent and are therefore preferred to raw residuals. Of these, jackknife residuals are most sensitive to outlier detection and are superior in terms of revealing other problems with the data. For that reason, most diagnostics rely upon the use of jackknife residuals. Whenever we have a choice in the residual analysis, we will select jackknife residuals.

Analysis of residuals - Normality

Recall that an assumption of linear regression is that the error terms are normally distributed. That is $\varepsilon \sim \text{Normal}(0, \sigma^2)$. To assess this assumption, we will use the residuals to look at:

- histograms
- normal quantile-quantile (qq) plots
- Wilk-Shapiro test
Histogram and QQ plot of residuals in SAS

ods rtf style = Analysis;
ods graphics on;

ods select ResidualHistogram;
ods select QQPlot;

proc reg data = one plots(unpack);
   model infrisk = los cult beds;
run;
quit;

ods graphics off;
ods rtf close;

---

Histogram of residuals in SAS

---

Regression diagnostics – p. 7/48

Regression diagnostics – p. 8/48
What is a normal QQ plot?

- Let \( q \) be a number between 0 and 1. The \( q \)th quantile of a distribution is that point, \( x \), at which \( q \times 100 \) percent of the data lie below \( x \) and \((1 - q) \times 100 \) percent of the data lie above \( x \). Specially named quantiles include quartiles, deciles, etc.

- The quantiles of the standard normal distribution are well known. Here are a few with which you should be familiar.

<table>
<thead>
<tr>
<th>( q )</th>
<th>Quantile</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.025</td>
<td>-1.96</td>
</tr>
<tr>
<td>0.05</td>
<td>-1.645</td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
</tr>
<tr>
<td>0.95</td>
<td>1.645</td>
</tr>
<tr>
<td>0.975</td>
<td>1.96</td>
</tr>
</tbody>
</table>

What is a normal QQ plot? (cont.)

- If data come from a normal distribution, then the quantiles of their standardized values should be approximately equivalent to the known quantiles of the standard normal distribution.

- A normal QQ plot graphs the quantiles of the data against the known quantiles of the standard normal distribution. Since we expect the quantiles to be roughly equivalent, then the QQ plot should follow the 45° reference line.
Normal QQ plot of residuals in SAS

What a normal QQ plot shouldn’t look like ...
The Wilk-Shapiro test

\( H_0 \): The data are normally distributed  
\( H_A \): The data are not normally distributed

```
proc reg data = one noprint;
   model infrisk = los cult beds;
   output out = fitdata rstudent = jackknife;
run;
quit;

proc univariate data = fitdata normal;
   var jackknife;
run;
```

The Wilk-Shapiro test (cont.)

Tests for Normality

<table>
<thead>
<tr>
<th>Test</th>
<th>-Statistic---</th>
<th>-----p Value------</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shapiro-Wilk</td>
<td>0.994445</td>
<td>Pr &lt; W 0.9347</td>
</tr>
</tbody>
</table>

We fail to reject the null hypothesis and conclude that there is insufficient evidence to conclude that the model errors are not normally distributed.
Identifying departures from homoscedasticity

We identify departures from homoscedasticity by plotting the residuals versus the predicted values.

``` SAS
ods html style = Journal;
ods graphics on;
ods select ResidualByPredicted;

proc reg data = one plots(unpack);
   model infrisk = los cult beds;
run;
quit;

ods graphics off;
ods html close;
```

Departures from homoscedasticity (cont.)

![Residual by Predicted for INFRISK](image)
Outliers

Outliers are observations that are *extreme* in the sense that they are noticeably different than the other data points. What causes outliers?

1. Data entry errors (the biggest culprit!)
2. Data do not represent a homogeneous set to which a single model applies. Rather, the data are a heterogeneous set of two or more types, of which one is more frequent than the others.
3. Error distributions that have “thick tails” in which extreme observations occur with greater frequency. (What does that mean?)
Outliers (cont.)

$t$ dist has 'thicker tails' than Normal

Regression diagnostics – p. 19/48

Outliers (cont.)

Sample from Normal (0,1) dist

Sample from $t_2$ df dist

Regression diagnostics – p. 20/48
Outliers (cont.)

Although linear regression is robust to departures from normality, this is not the case when the error distribution has thick tails. Ironically, sampling distributions that look quite different from a normal distribution cause little trouble, while these thick tail distributions flaw the inference based on $F$ tests.

Outlier detection - visual means

1. Simple scatterplots of the data (useful primarily for SLR)
2. Plots of the residuals versus the fitted values
3. Plot of the residuals versus each predictor

```sas
ods html style = analysis;
ods graphics on;
ods select ResidualByPredicted;
ods select ResidualPanel1;

proc reg data = one plots(unpack);
   model infrisk = los cult beds;
run; quit;

ods graphics off;
ods html close;
```
Outlier detection - Resid's vs. predicted values

Outlier detection - Resid's vs. predictors

Regression diagnostics – p. 23/48

Regression diagnostics – p. 24/48
Outlier detection - numerical means

Some rules of thumb about jackknife residuals

- Jackknife residuals with a magnitude less than 2 (i.e. between -2 and +2) are not unusual.
- Jackknife residuals with a magnitude greater than 2 deserve a look.
- Jackknife residuals with a magnitude greater than 4 are highly suspect.

Outlier detection in SAS

```sas
proc reg data = one;
   model infrisk = los cult beds;
   output out = fitdata rstudent = jackknife;
run;
quit;
```

```sas
proc print data = fitdata;
   where abs(jackknife) > 2;
run;
```

<table>
<thead>
<tr>
<th>Obs</th>
<th>INFRISK</th>
<th>LOS</th>
<th>CULT</th>
<th>BEDS</th>
<th>jackknife</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>5.4</td>
<td>11.18</td>
<td>60.5</td>
<td>640</td>
<td>-2.66653</td>
</tr>
<tr>
<td>35</td>
<td>6.3</td>
<td>9.74</td>
<td>11.4</td>
<td>221</td>
<td>2.29488</td>
</tr>
<tr>
<td>53</td>
<td>7.6</td>
<td>11.41</td>
<td>16.6</td>
<td>535</td>
<td>2.60976</td>
</tr>
<tr>
<td>63</td>
<td>5.4</td>
<td>7.93</td>
<td>7.5</td>
<td>68</td>
<td>2.20954</td>
</tr>
<tr>
<td>96</td>
<td>2.5</td>
<td>8.54</td>
<td>27.0</td>
<td>98</td>
<td>-2.15224</td>
</tr>
</tbody>
</table>
Leverage and influence

- The *leverage* of a data point refers to how extreme it is relative to $\bar{x}$. Observations far away from $\bar{x}$ are said to have “high leverage”.
- The *influence* of a data point refers to its impact on the fitted regression line. If an observation “pulls” the regression line away from the fitted line that *would* have resulted if that point had not been present, then that observation is deemed “influential”.

---

Dotted = with point; Solid = without point

---

Low lev, low inf

Low lev, high inf

High lev, low inf

High lev, high inf
Measuring leverage

For each data point, there is a corresponding quantity known as the hat diagonal that measures the standardized distance of the \(i\)th observation to the center of the predictors. In symbols, the hat diagonal is written \(h_i\). (We saw this term in the definition of studentized residuals.)

As a general rule of thumb, any value of \(h_i\) greater than \(2(k + 1)/n\) is cause for concern, where \(k\) is the number of predictors in the model and \(n\) is the number of data points.

Leverage and residual plots in SAS

```sas
ods html style = analysis;
ods graphics on;
ods select RStudentByLeverage;

proc reg data = one plots(unpack);
   model infrisk = los cult beds;
run;
quit;

ods graphics off;
odsl html close;
```
Measuring influence

For each data point, there is a corresponding quantity known as *Cook’s D* _i_ (‘D’ for ‘distance’) that is a standardized distance measuring how far \( \hat{\beta} \) moves when the _i_ th observation is removed. As a general rule of thumb, any observation with a value of Cook’s *D* _i_ greater than \( 4/n \), where _n_ is the number of observations, deserves a closer look.
Graphing of Cooks D in SAS

ods html style = analysis;
ods graphics on;
ods select CooksD;

proc reg data = one plots(unpack);
    model infrisk = los cult beds;
run;
quit;

ods graphics off;
ods html close;

Graphing of Cooks D in SAS (cont.)
Leverage/influence - numerical assessment

proc reg data = one;
   model infrisk = los cult beds;
   output out = fitdata cookd = cooksd h = hat;
run;
quit;

*(2*4)/113 = 0.071;
proc print data = fitdata;
   where hat ge (2*4)/113;
run;

*4/113 = 0.035;
proc print data = fitdata;
   where cooksd ge 4/113;
run;

Regression diagnostics – p. 35/48

Leverage/influence - numerical assessment

<table>
<thead>
<tr>
<th></th>
<th>INFRISK</th>
<th>LOS</th>
<th>CULT</th>
<th>BEDS</th>
<th>cooksd</th>
<th>hat</th>
<th>jackknife</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>5.4</td>
<td>11.18</td>
<td>60.5</td>
<td>640</td>
<td>0.45427</td>
<td>0.21252</td>
<td>-2.66653</td>
</tr>
<tr>
<td>11</td>
<td>4.9</td>
<td>11.07</td>
<td>28.5</td>
<td>768</td>
<td>0.03058</td>
<td>0.08368</td>
<td>-1.15910</td>
</tr>
<tr>
<td>13</td>
<td>7.7</td>
<td>12.78</td>
<td>46.0</td>
<td>322</td>
<td>0.02080</td>
<td>0.09230</td>
<td>0.90371</td>
</tr>
<tr>
<td>20</td>
<td>4.1</td>
<td>9.35</td>
<td>15.9</td>
<td>833</td>
<td>0.02697</td>
<td>0.11055</td>
<td>-0.93115</td>
</tr>
<tr>
<td>46</td>
<td>4.6</td>
<td>10.16</td>
<td>8.4</td>
<td>831</td>
<td>0.00055</td>
<td>0.10609</td>
<td>-0.13510</td>
</tr>
<tr>
<td>47</td>
<td>6.5</td>
<td>19.56</td>
<td>17.2</td>
<td>306</td>
<td>0.01081</td>
<td>0.30916</td>
<td>-0.30957</td>
</tr>
<tr>
<td>54</td>
<td>7.8</td>
<td>12.07</td>
<td>52.4</td>
<td>157</td>
<td>0.04106</td>
<td>0.13462</td>
<td>1.02773</td>
</tr>
<tr>
<td>78</td>
<td>4.9</td>
<td>10.23</td>
<td>9.9</td>
<td>752</td>
<td>0.00066</td>
<td>0.07977</td>
<td>0.17330</td>
</tr>
<tr>
<td>110</td>
<td>5.8</td>
<td>9.50</td>
<td>42.0</td>
<td>98</td>
<td>0.00083</td>
<td>0.08236</td>
<td>0.19197</td>
</tr>
<tr>
<td>112</td>
<td>5.9</td>
<td>17.94</td>
<td>26.4</td>
<td>835</td>
<td>0.20239</td>
<td>0.19506</td>
<td>-1.84793</td>
</tr>
</tbody>
</table>

Regression diagnostics – p. 36/48
ods html style = analysis;
ods graphics on;

proc reg data = one plots;
   model infrisk = los cult beds;
run;
quit;

ods graphics off;
ods html close;
Collinearity

Collinearity is a problem that exists when some (or all) of the independent variables are strongly linearly associated with one another. If collinearity exists in your data, then the following problems result.

- The estimated regression coefficients can be highly inaccurate.
- The standard errors of the coefficients can be highly inflated.
- The p-values, and all subsequent inference, can be wrong.

Symptoms of collinearity

- Large changes in coefficient estimates and/or in their standard errors when independent variables are added/deleted.
- Large standard errors.
- Non-significant results for independent variables that should be significant.
- Wrong signs on slope estimates.
- Overall test significant, but partial tests insignificant.
- Strong correlations between independent variables.
- Large variance inflation factors (VIFs).
Variance inflation factor

For each independent variable $X_j$ in a model, the variance inflation factor is calculated as

$$\text{VIF}_j = \frac{1}{1 - R^2_{X_j \sim \text{all other } X\text{s}}}$$

where $R^2_{X_j \sim \text{all other } X\text{s}}$ is the usual $R^2$ obtained by regressing $X_j$ on all the other $X$s, and represents the proportion of variation in $X_j$ that is explained by the remaining independent variables. Therefore, $1 - R^2_{X_j \sim \text{all other } X\text{s}}$ is a measure of the variability in $X_j$ that isn’t explained by the other independent variables.

Variance inflation factor (cont.)

When there is strong collinearity,

- $R^2_{X_j \sim \text{all other } X\text{s}}$ will be large
- $1 - R^2_{X_j \sim \text{all other } X\text{s}}$ will be small, and so
- VIF$_j$ will be large.

As a general rule of thumb, strong collinearity is present when VIF$_j > 10$. 

Regression diagnostics – p. 41/48

Regression diagnostics – p. 42/48
Collinearity example

Consider the MLR of INFRISK on LOS, CENSUS and BEDS.

ods html;
ods graphics on;

proc corr data = one plots=matrix;
   var los census beds;
run;

ods graphics off;
ods html close;

proc reg data = one;
   model infrisk = los/vif;
   model infrisk = los census/vif;
   model infrisk = los census beds/vif;
run;

Collinearity: PROC CORR output
Collinearity: PROC CORR output (cont.)

Pearson Correlation Coefficients, N = 113
Prob > |r| under H0: Rho=0

<table>
<thead>
<tr>
<th></th>
<th>LOS</th>
<th>CENSUS</th>
<th>BEDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOS</td>
<td>1.00000</td>
<td>0.47389</td>
<td>0.40927</td>
</tr>
<tr>
<td>LENGTH OF STAY</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td>CENSUS</td>
<td>0.47389</td>
<td>1.00000</td>
<td>0.98100</td>
</tr>
<tr>
<td>AVG DAILY CENSUS</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td>BEDS</td>
<td>0.40927</td>
<td>0.98100</td>
<td>1.00000</td>
</tr>
<tr>
<td>NUMBER OF BEDS</td>
<td>&lt;.0001</td>
<td>&lt;.0001</td>
<td></td>
</tr>
</tbody>
</table>

Collinearity: PROC REG output

INFRISK REGRESSED ON LOS

Root MSE 1.13929 R-Square 0.2846
Dependent Mean 4.35487 Adj R-Sq 0.2781
Coeff Var 26.16119

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| | Variance Inflation |
|----------|----|--------------------|----------------|---------|-------|-----|-------------------|
| Intercept | 1  | 0.74430            | 0.55386        | 1.34    | 0.1817|     | 0                 |
| LOS      | 1  | 0.37422            | 0.05632        | 6.64    | <.0001|     | 1.00000           |

Regression diagnostics – p. 45/48

Regression diagnostics – p. 46/48
Collinearity: PROC REG output (cont.)

INFRISK REGRESSED ON LOS AND CENSUS

<table>
<thead>
<tr>
<th>Parameter Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable</td>
</tr>
<tr>
<td>Intercept 1</td>
</tr>
<tr>
<td>LOS 1</td>
</tr>
<tr>
<td>CENSUS 1</td>
</tr>
</tbody>
</table>

Collinearity: PROC REG output (cont.)

INFRISK REGRESSED ON LOS, CENSUS AND BEDS

<table>
<thead>
<tr>
<th>Parameter Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable</td>
</tr>
<tr>
<td>Intercept 1</td>
</tr>
<tr>
<td>LOS 1</td>
</tr>
<tr>
<td>CENSUS 1</td>
</tr>
<tr>
<td>BEDS 1</td>
</tr>
</tbody>
</table>