Chapter 6 Multiple Regression

Timothy Hanson

Department of Statistics, University of South Carolina

Stat 704: Data Analysis I
Let’s construct a CI for the mean response corresponding to a set of values

\[ x_h = \begin{bmatrix} 1 \\ x_{h1} \\ x_{h2} \\ \vdots \\ x_{hk} \end{bmatrix}. \]

We want to make inferences about

\[ E(Y_h) = x_h' \beta = \beta_0 + \beta_1 x_{h1} + \cdots + \beta_k x_{hk}. \]
A point estimate is $\hat{Y}_h = E(Y_h) = x'_hb$.

Then $E(\hat{Y}_h) = E(x'_hb) = x'_hE(b) = x'_h\beta$.

Also $\text{var}(\hat{Y}_h) = \text{cov}(x'_hb) = x'_h\text{cov}(b)x_h = \sigma^2x'_h(X'X)^{-1}x_h$.

So...

A $100(1 - \alpha)\%$ CI for $E(Y_h)$ is

$$\hat{Y}_h \pm t_{n-p}(1 - \alpha/2)\sqrt{\text{MSE} \ x'_h(X'X)^{-1}x_h},$$

A $100(1 - \alpha)\%$ prediction interval for a new response $Y_h = x'_h\beta + \epsilon_h$ is

$$\hat{Y}_h \pm t_{n-p}(1 - \alpha/2)\sqrt{\text{MSE}[1 + x'_h(X'X)^{-1}x_h]},$$
Say we want to estimate mean sales in cities with $x_1 = 65.4$ thousand people 16 or younger and per capita disposable income of $x_2 = 17.6$ thousand dollars. Now say we want a prediction interval for a new city with these covariates. We can add these covariates to the data step, with a missing value “.” for sales, and ask SAS for the CI and PI.

```plaintext
data studio;
    input people16 income sales @@;
    label people16='16 & under (1000s)' income ='Per cap. disp. income ($1000)'
        sales = 'Sales ($1000$)';
datalines;
    68.5 16.7 174.4 45.2 16.8 164.4 91.3 18.2 244.2 47.8 16.3 154.6
    46.9 17.3 181.6 66.1 18.2 207.5 49.5 15.9 152.8 52.0 17.2 163.2
    48.9 16.6 145.4 38.4 16.0 137.2 87.9 18.3 241.9 72.8 17.1 191.1
    88.4 17.4 232.0 42.9 15.8 145.3 52.5 17.8 161.1 85.7 18.4 209.7
    41.3 16.5 146.4 51.7 16.3 144.0 89.6 18.1 232.6 82.7 19.1 224.1
    52.3 16.0 166.5 65.4 17.6 .
;
proc reg data=studio;
    model sales=people16 income / clm cli alpha=0.05;
```

Output Statistics

<table>
<thead>
<tr>
<th>Obs</th>
<th>Dependent Variable</th>
<th>Predicted Value</th>
<th>Std Error</th>
<th>95% CL Mean</th>
<th>95% CL Predict</th>
<th>Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>174.4000</td>
<td>187.1841</td>
<td>3.8409</td>
<td>179.1146</td>
<td>195.2536</td>
<td>162.6910</td>
</tr>
</tbody>
</table>

...et cetera...

4 / 25
The general linear model assumes the following:

1. A linear relationship between $E(Y)$ and associated predictors $x_1, \ldots, x_k$.
2. The errors have constant variance.
3. The errors are normally distributed.
4. The errors are independent.

We estimate the unknown $\epsilon_1, \ldots, \epsilon_n$ with the residuals $e_1, \ldots, e_n$. Assumptions can be checked informally using plots and formally using tests.

**Note:** We can’t check $E(\epsilon_i) = 0$ because $e_1 + \cdots + e_n = 0$, i.e. $\bar{e} = 0$, by construction.
Assumption 1: Linear mean

- Scatterplots of \( \{(x_{ij}, Y_i)\}_{i=1}^{n} \) for each predictor \( j = 1, \ldots, k \). Look for "nonlinear" patterns. These are marginal relationships, and do not get at the simultaneous relationship among variables.

- Look at residuals versus each predictor \( \{(x_{ij}, e_i)\}_{i=1}^{n}, and \) (or?) residuals versus fitted values \( \{\hat{Y}_i, e_i\}_{i=1}^{n} \).

- Book suggests looking at residuals versus pairwise interactions, e.g. \( e_i \) versus \( x_{i1}x_{i2} \).

- Look for non-random (especially curved) pattern in the residual plots, indicating violation of linear mean.

**Remedies:** (i) choose different functional form of model, (ii) transformation of one or more predictor variables.

- Formal "lack of fit" test is available (Section 3.7, also p. 235), but requires replicate observations at each distinct predictor value.
proc sgscatter; matrix people16 income sales / diagonal=(histogram kernel); run;
Assumption 2: Constant variance

- Often the most worrisome assumption.
- Violation indicated by “megaphone shape” in residual plot:

![Residual plots](image)

- **Easy remedy**: transform the response, e.g. $Y^* = \log(Y)$ or $Y^* = \sqrt{Y}$.
- **Advanced method**: weighted least squares (Chapter 11).
Non-constant variance

- **Breusch-Pagan test** (pp. 118–119): tests whether the log error variance increases or decreases linearly with the predictor(s). Where $Y_i \sim N(x_i' \beta, \sigma^2_i)$, set
  \[
  \log \sigma^2_i = \alpha_0 + \alpha_1 x_{i1} + \cdots + \alpha_k x_{ik}
  \]
  and test
  \[
  H_0 : \alpha_1 = \cdots = \alpha_k = 0, \text{ i.e. } \log \sigma^2_i = \alpha_0.
  \]
  Requires large samples & assumes normal errors.

- **Brown-Forsythe test** (pp. 116–117): Robust to non-normal errors. Requires user to break data into groups and test for constancy error variance across groups (not natural for continuous data).

- Graphical methods have advantage of checking for general violations, not just violation of a specific type.
Breusch Pagan test in SAS

PROC MODEL carries out a modified version of the test where $\sigma_i = \sigma + \alpha_1 x_{i1} + \cdots + \alpha_k x_{ik}$ and $H_0 : \alpha_1 = \cdots = \alpha_k = 0$. If $H_0$ is true then $\sigma_i = \sigma$ for $i = 1, \ldots, n$.

```sas
proc model data=studio;
  parms beta0 beta1 beta2;
  sales=beta0+people16*beta1+income*beta2;
  fit sales / breusch=(1 income sales);
```

### Nonlinear OLS Summary of Residual Errors

<table>
<thead>
<tr>
<th>Equation</th>
<th>DF Model</th>
<th>DF Error</th>
<th>SSE</th>
<th>MSE</th>
<th>Root MSE</th>
<th>R-Square</th>
<th>Adj R-Sq</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>sales</td>
<td>3</td>
<td>18</td>
<td>2180.9</td>
<td>121.2</td>
<td>11.0074</td>
<td>0.9167</td>
<td>0.9075</td>
<td>Sales ($1000$)</td>
</tr>
</tbody>
</table>

### Nonlinear OLS Parameter Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Std Err</th>
<th>t Value</th>
<th>Pr &gt;</th>
<th>Approx</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta0</td>
<td>-68.8571</td>
<td>60.0170</td>
<td>-1.15</td>
<td>0.2663</td>
<td></td>
</tr>
<tr>
<td>beta1</td>
<td>1.45456</td>
<td>0.2118</td>
<td>6.87</td>
<td>&lt;.0001</td>
<td></td>
</tr>
<tr>
<td>beta2</td>
<td>9.3655</td>
<td>4.0640</td>
<td>2.30</td>
<td>0.0333</td>
<td></td>
</tr>
</tbody>
</table>

### Heteroscedasticity Test

<table>
<thead>
<tr>
<th>Equation</th>
<th>Test</th>
<th>Statistic</th>
<th>DF</th>
<th>Pr &gt; ChiSq</th>
<th>Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>sales</td>
<td>Breusch-Pagan</td>
<td>2.10</td>
<td>2</td>
<td>0.3503</td>
<td>1, income, sales</td>
</tr>
</tbody>
</table>

With $p = 0.35$ we do not reject $H_0 : \sigma_i = \sigma$ at $\alpha = 0.05$, no evidence of non-constant variance.
Caution: your estimate of $\epsilon$, given by $e = Y - Xb$, is only as good as the model for your mean! Changing the mean can drastically change the residuals $e$ and any residual plots or formal tests based on them. Diagnostics include...

- Q-Q plot of $e_1, \ldots, e_n$.
- Formal test for normality: Shapiro-Wilk (Section 3.5), essentially based on the correlation coefficient $r$ for expected versus observed in normal Q-Q plot.
- Remedy: transformation of $Y$ and or any of $x_1, \ldots, x_k$, nonparametric methods (e.g. additive models), robust regression (least sum of absolute distances), median regression.
Standard diagnostics from ODS GRAPHICS

**Fit Diagnostics for sales**

- Residual vs Predicted Value
- RStudent vs Predicted Value
- RStudent vs Leverage
- Residual vs Quantile
- Sales ($1000$) vs Predicted Value
- Cook's D vs Observation
- Percent vs Residual
- Fit-Mean vs Residual
- Proportion Less vs Residual

Observations: 21
Parameters: 3
Error DF: 18
MSE: 121.16
R-Square: 0.9167
Adj R-Square: 0.9075
proc reg data=studio;
   model sales=people16 income;
   output out=temp r=residual;
proc univariate data=temp normal; var residual; run;

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>W 0.954073</td>
<td>Pr &lt; W 0.4056</td>
</tr>
<tr>
<td>Kolmogorov-Smirnov</td>
<td>D 0.147126</td>
<td>Pr &gt; D &gt;0.1500</td>
</tr>
<tr>
<td>Cramer-von Mises</td>
<td>W-Sq 0.066901</td>
<td>Pr &gt; W-Sq &gt;0.2500</td>
</tr>
<tr>
<td>Anderson-Darling</td>
<td>A-Sq 0.432299</td>
<td>Pr &gt; A-Sq &gt;0.2500</td>
</tr>
</tbody>
</table>

We accept (or “do not reject” if you are a purist) $H_0 : e_1, \ldots, e_n$ are normal.

The Anderson-Darling tests looks primarily for evidence of non-normal data in the tails of a distribution; the Shapiro-Wilk emphasizes lack of symmetry in the distribution; i.e. less emphasis placed on the tails.
With large sample sizes, the normality assumption is not critical *unless you are predicting new observations*.

The formal test will not tell you the *type* of departure from normality (e.g. bimodal, skew, heavy or light tails, et cetera).

Q-Q plots help answer these questions (*if* the mean is specified correctly).
Chapter 12 discusses time-series methods. Handles correlated errors over time (or space). Can also include time as a predictor.

If willing to assume some *structure* on the errors, e.g. AR(1), then can do a formal test (Chapter 12, e.g. Durbin-Watson test pp. 484–488).


In general, need to test $H_0 : \text{cov}(\varepsilon) = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$ (diagonal), or even stronger $H_0 : \text{cov}(\varepsilon) = \sigma^2 I_n$ (spherical – constant variance).
Another example

Problems 6.15, 6.16, 6.17

- $Y =$ patient satisfaction (100 point scale)
- $x_1 =$ age in years
- $x_2 =$ illness severity (an index)
- $x_3 =$ anxiety level (an index)

Let’s analyze these data with SAS...
6.15(b) scatterplot matrix
data sat;
    input sat age sev anx;
    age_sev=age*sev; age_anx=age*anx; sev_anx=sev*anx; * interactions;
    label sat='Satisfaction (100)'
        age='Age (years)'
        sev='Illness severity (index)'
        anx='Anxiety (index)';
datalines;
    48  50  51  2.3
    57  36  46  2.3
    66  40  48  2.2
    ...et cetera...
    68  45  51  2.2
    59  37  53  2.1
    92  28  46  1.8
;
proc sgscatter data=sat; matrix sat age sev anx; run;

options nocenter;
proc reg data=sat;
    model sat=age sev anx;
    output out=resid r=residual; run;

proc sgscatter data=resid;
    plot residual*age_sev residual*age_anx residual*sev_anx; run;
### 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>9120.46367</td>
<td>3040.15456</td>
<td>30.05</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>42</td>
<td>4248.84068</td>
<td>101.16287</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>45</td>
<td>13369</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Root MSE: 10.05798
- R-Square: 0.6822
- Dependent Mean: 61.56522
- Adj R-Sq: 0.6595
- Coeff Var: 16.33711

### Parameter Estimates

| Variable                  | Parameter Label                  | DF | Estimate | Error   | t Value | Pr > |t| |
|---------------------------|----------------------------------|----|----------|---------|---------|------|---|
| Intercept                 | Intercept                        | 1  | 158.49125| 18.12589| 8.74    | <.0001|
| age                       | Age (years)                      | 1  | -1.14161 | 0.21480 | -5.31   | <.0001|
| sev                       | Illness severity (index)         | 1  | -0.44200 | 0.49197 | -0.90   | 0.3741|
| anx                       | Anxiety (index)                  | 1  | -13.47016| 7.09966 | -1.90   | 0.0647|
Residuals vs. predictors

![Residual by Regressors for sat graphs](image_url)
Residuals vs. interactions

- Residual vs. age_sev
- Residual vs. age_anx
- Residual vs. sev_anx
6.15(c) $\hat{\text{sat}} = 158.5 - 1.14\text{age} - 0.442\text{sev} - 13.5\text{anx}$.  
$b_2 = -0.442$; for every unit increase in the illness severity index, mean satisfaction is reduced by 0.442 units.

6.15(e) There is a slight increase in variability for $e_i$ vs. $\hat{Y}_i$, but overall it looks okay. The normal probability plot of the residuals looks fine (relatively straight). The plots of the residuals vs. each predictor and each two-way interaction all look appropriately “random.”

6.15(g) The Breusch-Pagan test gives $p = 0.46$, no evidence of non-constant variance.

```plaintext
proc model data=sat;
    parms beta0 beta1 beta2 beta3;
    sat=beta0+age*beta1+sev*beta2+anx*beta3;
    fit sat / breusch=(1 age sev anx);
```

<table>
<thead>
<tr>
<th>Equation</th>
<th>Test</th>
<th>Statistic</th>
<th>DF</th>
<th>Pr &gt; ChiSq</th>
<th>Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>sat</td>
<td>Breusch-Pagan</td>
<td>2.56</td>
<td>3</td>
<td>0.4648</td>
<td>1, age, sev, anx</td>
</tr>
</tbody>
</table>
6.16(a) We reject $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ at the 1% level ($p < 0.0001$ from the ANOVA table). One or more regressors are important in the model.

6.16(c) $R^2 = 0.68$ so 68% of the variability is explained by the regression surface.

6.17(a,b) I added " . 35 45 2.2" to the data and changed the model statement to 

```
model sat=age sev anx / clm cli alpha=0.1;
```

obtaining

<table>
<thead>
<tr>
<th>Obs</th>
<th>Variable</th>
<th>Value</th>
<th>Mean</th>
<th>Predict</th>
<th>90% CL Mean</th>
<th>90% CL Predict</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.</td>
<td>69.0103</td>
<td>2.6646</td>
<td>64.5285</td>
<td>73.4920</td>
<td>51.5097</td>
</tr>
</tbody>
</table>

We are 90% confident that the true mean satisfaction is between 64.5 and 73.5 units for 35 year-olds with severity 45 and anxiety 2.2. We would predict a new patient from this population to have a satisfaction in the range 51.5 to 86.5.