Chapter 6 Multiple Regression

Timothy Hanson

Department of Statistics, University of South Carolina

Stat 704: Data Analysis I

6.7 CI for mean response and PI for new response

Let's construct a CI for the mean response corresponding to a set of values

$$\mathbf{x}_h = \begin{bmatrix} 1 \\ x_{h1} \\ x_{h2} \\ \vdots \\ x_{hk} \end{bmatrix}.$$

We want to make inferences about

$$E(Y_h) = \mathbf{x}_h' \boldsymbol{\beta} = \beta_0 + \beta_1 x_{h1} + \cdots + \beta_k x_{hk}.$$

Some math...

- A point estimate is $\hat{Y}_h = \widehat{E(Y_h)} = \mathbf{x}_h'\mathbf{b}$.
- Then $E(\hat{Y}_h) = E(\mathbf{x}_h'\mathbf{b}) = \mathbf{x}_h' E(\mathbf{b}) = \mathbf{x}_h'\boldsymbol{\beta}$.
- Also $\text{var}(\hat{Y}_h) = \text{cov}(\mathbf{x}_h'\mathbf{b}) = \mathbf{x}_h'\text{cov}(\mathbf{b})\mathbf{x}_h = \sigma^2 \mathbf{x}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_h$.

So...

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

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

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

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

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.

```
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
     Dependent  Predicted    Std Error
Obs   Variable      Value  Mean Predict         95% CL Mean           95% CL Predict        Residual
  1   174.4000   187.1841        3.8409   179.1146   195.2536   162.6910   211.6772   -12.7841
 21   166.5000   157.0644        4.0792   148.4944   165.6344   132.4018   181.7270     9.4356
                                            ...et cetera...
 22          .   191.1039        2.7668   185.2911   196.9168   167.2589   214.9490          .
```

6.8 Checking model assumptions

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.

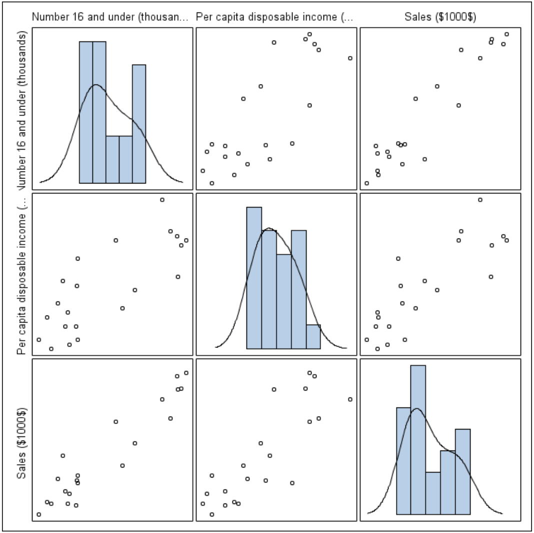
We estimate the unknown $\epsilon_1, \ldots, \epsilon_n$ with the residuals $e_1, \ldots, e_n$. Assumptions can be checked informally using plots and formally using tests.
Note: We can't check $E(\epsilon_i) = 0$ because $e_1 + \cdots + e_n = 0$, i.e. $\bar{e} = 0$, by construction.

Assumption 1: Linear mean

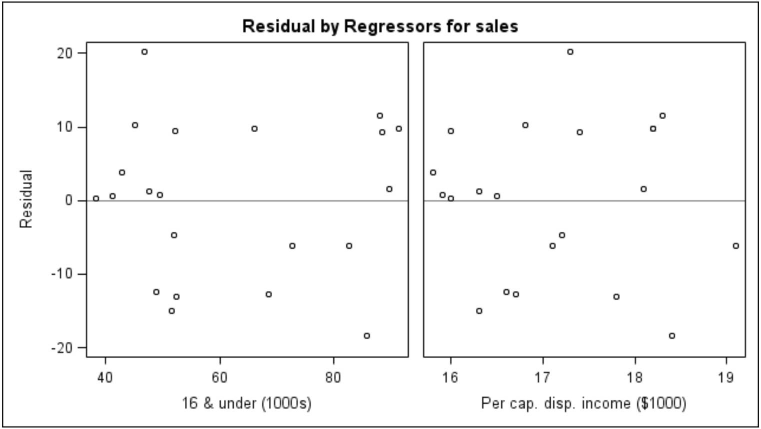
- Scatterplots of $\{(x_{ij}, Y_i)\}_{i=1}^n$ for each predictor $j = 1, \ldots, k$. Look for "nonlinear" patterns. These are *marginal* relationships, and do not get at the simultaneous relationship among variables.
- Look at residuals versus each predictor $\{(x_{ij}, e_i)\}_{i=1}^n$, *and* (or?) residuals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

Scatterplot matrix

```
proc sgscatter; matrix people16 income sales / diagonal=(histogram kernel); run;
```

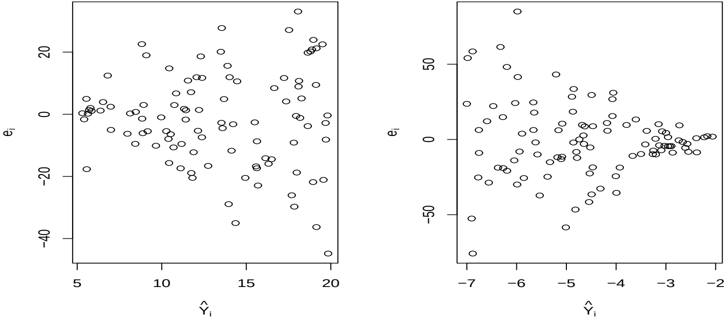


Standard diagnostics from ODS GRAPHICS



Assumption 2: Constant variance

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



- **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 $Y_i \sim N(\mathbf{x}_i'\boldsymbol{\beta}, \sigma_i^2)$, set $\log \sigma_i^2 = \alpha_0 + \alpha_1 x_{i1} + \cdots \alpha_k x_{ik}$ and test $H_0 : \alpha_1 = \cdots = \alpha_k = 0$, i.e. $\log \sigma_i^2 = \alpha_0$. 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.

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

```
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

Equation	DF Model	DF Error	SSE	MSE	Root MSE	R-Square	Adj R-Sq	Label
sales	3	18	2180.9	121.2	11.0074	0.9167	0.9075	Sales (1000)

Nonlinear OLS Parameter Estimates

Parameter	Estimate	Approx Std Err	t Value	Approx Pr > |t|
beta0	-68.8571	60.0170	-1.15	0.2663
beta1	1.45456	0.2118	6.87	<.0001
beta2	9.3655	4.0640	2.30	0.0333

Heteroscedasticity Test

Equation	Test	Statistic	DF	Pr > ChiSq	Variables
sales	Breusch-Pagan	2.10	2	0.3503	1, income, sales

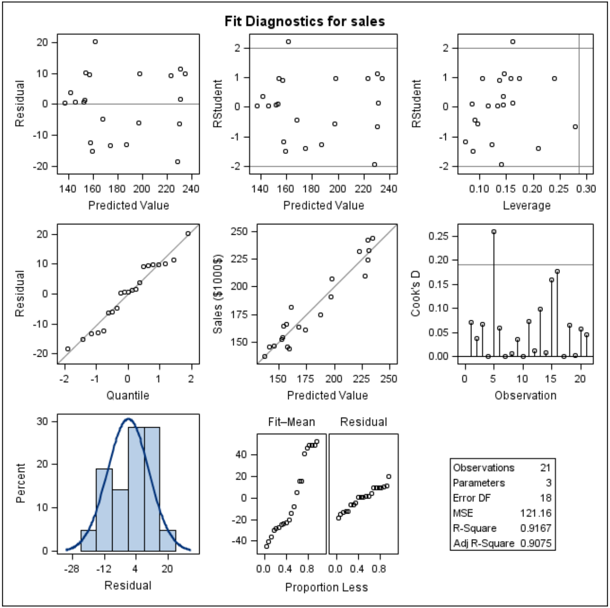
With $p = 0.35$ we do not reject $H_0 : \sigma_i = \sigma$ at $\alpha = 0.05$, no evidence of non-constant variance.

Assumption 3: errors are normally distributed

Caution: *your estimate of ϵ, given by $\mathbf{e} = \mathbf{Y} - \mathbf{X}\mathbf{b}$, is only as good as the model for your mean!* Changing the mean can *drastically* change the residuals $\mathbf{e}$ 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



Test for normal residuals in Portrait data

```
proc reg data=studio;
 model sales=people16 income;
 output out=temp r=residual;
proc univariate data=temp normal; var residual; run;

                   Tests for Normality

Test                  --Statistic---    -----p Value------
Shapiro-Wilk          W     0.954073    Pr < W      0.4056
Kolmogorov-Smirnov    D     0.147126    Pr > D     >0.1500
Cramer-von Mises      W-Sq  0.066901    Pr > W-Sq  >0.2500
Anderson-Darling      A-Sq  0.432299    Pr > A-Sq  >0.2500
```

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.

Comments

- 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).

Assumption 4: Independence

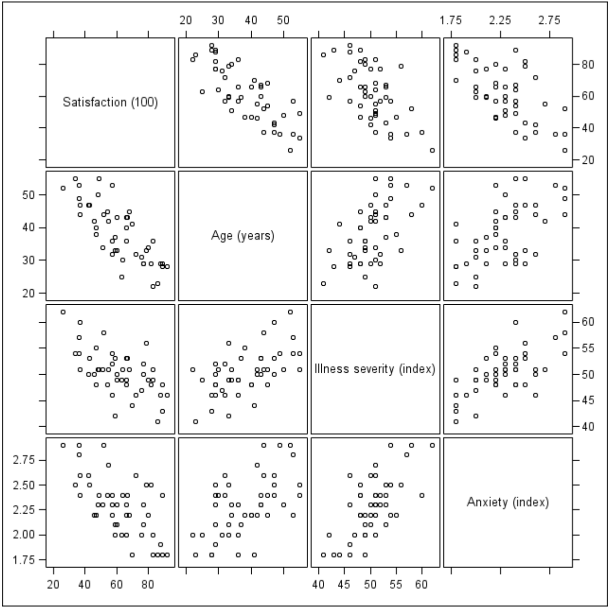
- Chapter 12 discusses time-series methods. Handles correlated errors over time (or space). Can also include time as a predictor.
- If willing to assume some *structure* on the errors, e.g. AR(1), then can do a formal test (Chapter 12, e.g. Durbin-Watson test pp. 484–488).
- Christensen, R. and Bedrick, E. (1997). Testing the independence assumption in linear models. *JASA*, 92, 1006–1016. Uses "near-replicates" instead of replicates. (Replicates needed for standard LOF test).
- In general, need to test $H_0 : \text{cov}(\epsilon) = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$ (diagonal), or even stronger $H_0 : \text{cov}(\epsilon) = \sigma^2 \mathbf{I}_n$ (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...


Satisfaction (100)
Age (years)
Illness severity (index)
Anxiety (index)
20 30 40 50
1.75 2.25 2.75
80 60 40 20
50 40 30 20
60 55 50 45 40
2.75 2.50 2.25 2.00 1.75
20 40 60 80
40 45 50 55 60

SAS code

```
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;
```

Regression output

Analysis of Variance

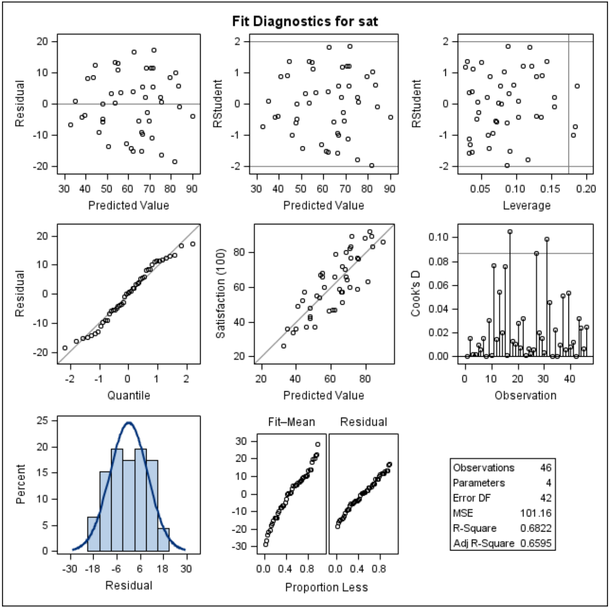
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	9120.46367	3040.15456	30.05	<.0001
Error	42	4248.84068	101.16287		
Corrected Total	45	13369			

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

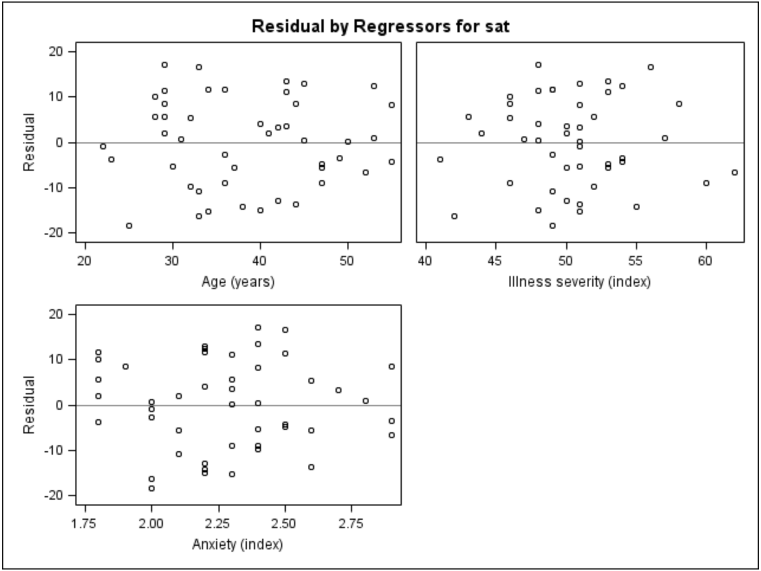
Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard 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

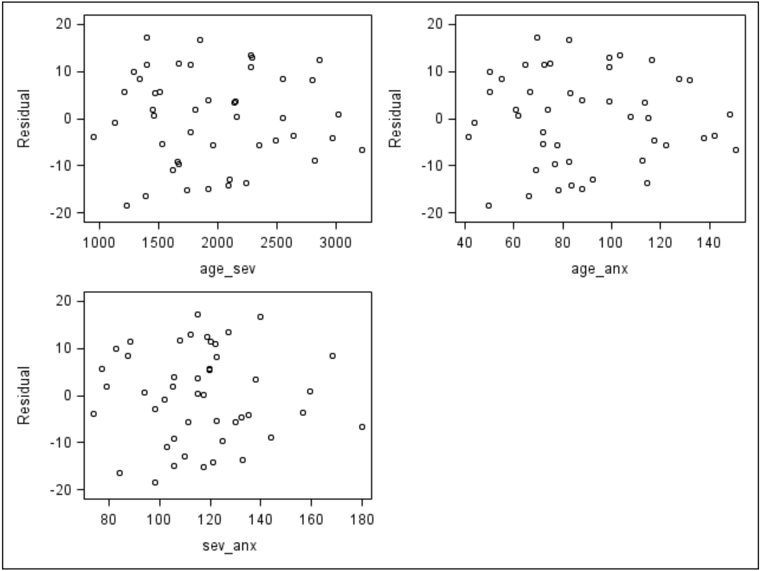
Diagnostics



Residuals vs. predictors



Residuals vs. interactions



Some answers to textbook problems...

- 6.15(c) $\widehat{\text{sat}} = 158.5 - 1.14\text{age} - 0.442\text{sev} - 13.5\text{anx}$.
  $b_2 = -0.442$; 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.

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

Heteroscedasticity Test

Equation	Test	Statistic	DF	Pr > ChiSq	Variables
sat	Breusch-Pagan	2.56	3	0.4648	1, age, sev, anx

Some answers to textbook problems...

- 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

```
    Dependent Predicted    Std Error
Obs  Variable     Value Mean Predict    90% CL Mean        90% CL Predict
  1         .   69.0103        2.6646  64.5285  73.4920  51.5097   86.5109
```

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.