# 36-401 Class Notes and Exercise for 10/12/00

**Howard Seltman**

## Testing and Breaking the Regression Assumptions

### Checking the assumptions

We check assumptions by thinking about the experimental design, viewing plots, and sometimes by performing special tests.
1) Covariates are measured with no (little) error. Is there a clear, precise, accurate method for measuring each covariate? Could we "theoretically obtain many random samples, each with the same X but with different Y's due to different $\epsilon$'s.?"

2) Linearity. Do we think that it is reasonably likely that within the region of the covariate space within which our observations lie (the "design space"), the relationship between each covariate and the outcome may be monotone, and further, nearly linear?
2b) Additivity. We need to add interaction terms if the effects of each separate covariate is not identical at all levels of the other covariates.

3) Normality of errors (as checked by looking at normality of residuals. From the problem description, is it reasonable to say $Y \sim N(\mu, \sigma^2)$? Use a histogram with overlaid normal curve and/or a Quantile-Normal plot to detect deviations from normality. Boxplots and symmetry plots will indicate skew or lack of skew, but lack of skew alone does not prove normality. Non-normal symmetric distributions include those with heavy or light tails. On a box-plot short length whiskers suggest lighter that Normal tails, and too many outliers suggests heavier than normal tails. Formal tests like Kolmogorov-Smirnov exist.

4) Independent errors. Does knowing whether a certain unit has a positive residual tell us something about the residual of another specific unit (e.g. when we have 20 observations on 5 subjects)? Is there a clear relationship in a plot of $e_i$ vs. $e_{i-1}$? Formal tests for serial correlation like Durbin-Watson exist. Adding appropriate covariates may restore

independence.

5) Equal variances. Is it clear that at different parts of the covariate space the variability of $Y$ will vary more than at others on an absolute scale? Check the fitted vs. residual plot for different spread in different regions. Don't be fooled by the effects of different numbers of cases in different regions.

**"Error in variables"**
Assume the usual simple regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \;\; \epsilon_i \sim N(0, \sigma^2), \;\; \text{i.i.d}$$

But we cannot see the X's. Instead we see a noisy version, $\tilde{X}$.

$$\tilde{X} = X + u, \;\; u \sim N(0, \tau^2), \;\; \text{i.i.d}$$

Then the relationship between the usual estimate $b_1$ and the true value, $\beta_1$ is:

$$\mathbf{b} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\mathbf{Y} \quad \Rightarrow \quad b_1 = \frac{\sigma^2}{\sigma^2 + \tau^2}\beta_1$$

What conclusions can we make about simple regression when X is measured with error?

**Unequal variance** Consider the following code, that simulates simple regression with heteroscedasticity. The second version uses an alternate technique called "weighted" regression.

```
b3_matrix(NA,500,4)
for (i in 1:500) {
  x_seq(1,3,len=15)
  y_3-5*x+rnorm(15,0,sqrt(x))   # 1<variance<3
  rslt_lm(y~x)
  b3[i,]_as.vector(summary(rslt)$coef[,1:2])
}
# as.vector unwraps matrix so b3 has b0, b1, SEb0, SEb1

print(apply(b3,2,mean))  # 3.0191229 -5.0031621  1.2060848  0.5762269
```

```
print(apply(b3[,1:2],2,sd))  # 1.0801233 0.5701735

tval_qt(0.975,13)
cat(100*round(sum(b3[,1]-tval*b3[,3]<3 & b3[,1]+tval*b3[,3]>3)/500,3),
  "% beta0 coverage\n")
# 97.4% beta0 coverage
cat(100*round(sum(b3[,2]-tval*b3[,4]<(-5) & b3[,2]+tval*b3[,4]>(-5))/500,3),
  "% beta1 coverage\n")
# 96.4% beta1 coverage

b3w_matrix(NA,500,4)
for (i in 1:500) {
  x_seq(1,3,len=15)
  y_3-5*x+rnorm(15,0,sqrt(x))  # 1<variance<3
  rslt_lm(y~x,weights=1/x)
  b3w[i,]_as.vector(summary(rslt)$coef[,1:2])
}
print(apply(b3w,2,mean))  # 2.9549608 -4.9696424  1.0418717  0.5502395
print(apply(b3w[,1:2],2,sd))  # 1.0436101 0.5536613

cat(100*round(sum(b3w[,1]-tval*b3w[,3]<3 & b3w[,1]+tval*b3w[,3]>3)/500,3),
  "% beta0 coverage\n")
# 94.8% beta0 coverage
cat(100*round(sum(b3w[,2]-tval*b3w[,4]<(-5) & b3w[,2]+tval*b3w[,4]>(-5))/500,3),
  "% beta1 coverage\n")
# 96.2% beta1 coverage
```

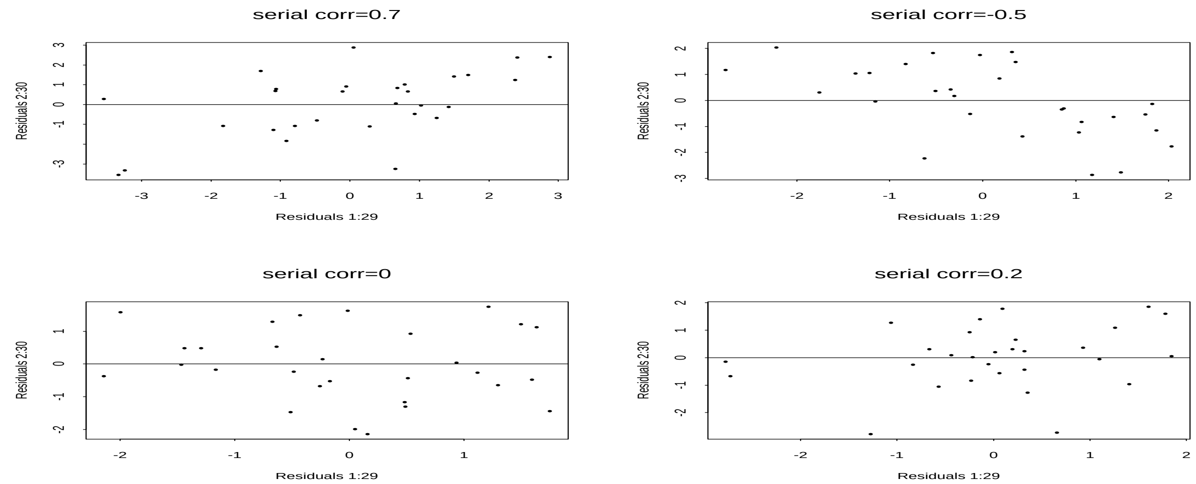
**Correlated errors**

Correlated errors cause regression results to have unbiased coefficient estimates, but the estimated standard errors of the coefficients are biased. Positive correlation causes low estimates of the standard errors of the coefficients.

**Non-normality of errors** Non-normality causes loss of the property of "most efficient unbiased coefficient estimates" due to highly technical reasons. In this case p-values and CI's are incorrect also.

| **Problem** | **Biased b** | **Biased SE** | **Invalid t & F tests** | **High Var(b) (Inefficient)** |
|---|---|---|---|---|
| Nonlinear relationship | yes | yes | yes | – |
| Omit "relevant" X | yes | yes | yes | – |
| Include irrelevant X | no | no | no | yes |
| X measured with error | yes | yes | yes | – |
| Heteroscedasticity | no | yes | yes | yes |
| Autocorrelation | no | yes | yes | yes |
| X correlated with $\epsilon$ | yes | yes | yes | – |
| Non-normal $\epsilon$ | no | no | yes | yes |
| Multicollinearity | no | no | no | yes |

## Serially Correlated Errors




**Durbin-Watson test**

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{(i-1)})^2}{\sum_{i=1}^{n} e_i^2}$$

Above graphs have true serial residual correlations of 0.7, -0.5, 0, 0.2 and d=0.89, 2.92, 2.23, 1.38 with $d_L$=1.35 and $d_U$=1.49 (from table in back of RwG book). $H_0$ is that the errors are uncorrelated. This is rejected for "A" only. Values between $d_L$ and $d_U$ are considered inconclusive. Values of $d$ greater than $d_U$ allow acceptance of $H_0$.
For negative correlation, $d > 2$ and we need to look up 4-d instead. For "B" we use d=4-2.23=1.77 which is significant.

## Generalized Least Squares

More general model:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},\ \ \boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}_n, \boldsymbol{\Sigma})$$

New estimator:

$$\mathbf{b} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}\boldsymbol{\Sigma}^{\mathbf{-1}}\mathbf{Y}$$