# Violations of the Constant Variances Assumption

James H. Steiger

Department of Psychology and Human Development
Vanderbilt University

# Violations of the Constant Variances Assumption

1

- A Diagnostic for Nonconstant Variance

# Introduction

- The assumption of constant conditional variance is a staple of the standard linear regression model, both in the case of a single predictor-regressor (bivariate regression) or in the case of several predictors (multiple regression).
- Violation of this assumption occurs quite frequently in practice, for a number of reasons.
- In this module, we'll explore a diagnostic significance test sometimes used to assess departures from the equal variances assumption.

# A Diagnostic for Nonconstant Variance

- Breusch and Pagan (1979) gave a test for nonconstant variance. This was also developed independently by Cook and Weisberg(1983) and discussed in section 7.2.2 of the ALR4 text.
- The test assumes that the conditional variance of $Y$ given $X$ is an exponential function of an unknown parameter vector and some set of regressors $Z$. The assumption is that

$$\mathrm{Var}(Y|X, Z = \mathbf{z}) = \sigma^2 \exp(\boldsymbol{\lambda}'\mathbf{z}) \tag{1}$$

- If $\boldsymbol{\lambda} = \mathbf{0}$, then the right side of the equation evaluates to $\sigma^2$, and we have constant variance.
- Under that assumption, a score test that $\boldsymbol{\lambda} = \mathbf{0}$ can be computed using regression software.

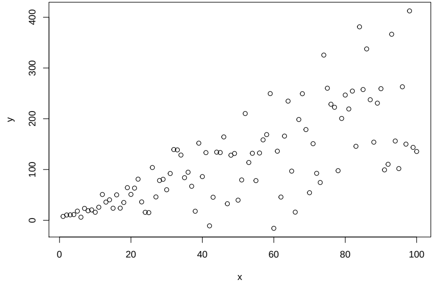
# A Diagnostic for Nonconstant Variance

- Compute the standard OLS fit. Save the residuals $\hat{e}_i$.
- Compute scaled residuals $u_i = \hat{e}_i^2/\tilde{\sigma}^2$. The maximum likelihood estimator $\tilde{\sigma}^2$ is simply $\sum \hat{e}_i^2/n$, i.e., it uses $n$ instead of $n-p-1$ as a denominator. The variable $U$ is simply composed of the $u_i$.
- Compute the regression for the mean function $E(U|Z=\mathbf{z}) = \lambda_0 + \boldsymbol{\lambda}'\mathbf{z}$. Obtain $SSreg$ for this regression with degrees of freedom equal to $q$, the number of components in $Z$. If variance is thought to be a function of the responses (i.e., the dependent variable $Y$), then in this regression replace $Z$ by the fitted values of the regression in step 1, in which case the test will have 1 degree of freedom.
- The score test statistic is $S = SSreg/2$. The reference distribution is $\chi^2_q$.

# A Diagnostic for Nonconstant Variance

- If you have more than one predictor, you can perform the B-P test on different combinations of regressors based on those predictors, in order to develop a model for the variance function.
- Let's start with a simple bivariate regression.
- Suppose we generate some artificial data in which the residual variance is a function of $X$.

```
> set.seed(12345)
> x <- 1:100
> y <- 2*x + 5 + x * rnorm(100)
> plot(x,y)
```



# A Diagnostic for Nonconstant Variance

Here are the manual calculations:

```
> m0 <- lm(y ~ x)
> sig2 <- sum(residuals(m0)^2)/length(x)
> U <- residuals(m0)^2/sig2
> m1 <- lm(U~x)
> anova(m1)

Analysis of Variance Table

Response: U
          Df  Sum Sq Mean Sq F value    Pr(>F)
x          1  65.233  65.233  34.272 6.383e-08 ***
Residuals 98 186.533   1.903
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> S <- anova(m1)$'Sum Sq'[1]/2
> p.value <- 1-pchisq(S,1)
> S

[1] 32.61656

> p.value

[1] 1.122545e-08
```

# A Diagnostic for Nonconstant Variance

## The Snow Geese Data

- You can also use the car library and its `ncv.test` function to get the same result.

```
> library(car)
> ncvTest(m0,~x)
```

```
Non-constant Variance Score Test
Variance formula: ~ x
Chisquare = 32.61656    Df = 1     p = 1.122545e-08
```

# A Diagnostic for Nonconstant Variance

## The Snow Geese Data

- Aerial surveys sometimes rely on visual methods to estimate the number of animals in an area. For example, to study snow geese in their summer range areas west of Hudson Bay in Canada, small aircraft were used to fly over the range, and when a flock of geese was spotted, an experienced person estimated the number of geese in the flock.
- To investigate the reliability of this method of counting, an experiment was conducted in which an airplane carrying two observers flew over $n = 45$ flocks, and each observer made an independent estimate of the number of birds in each flock.
- Also, a photograph of the flock was taken so that a more or less exact count of the number of birds in the flock could be made.
- The resulting data are given in the data file `snowgeese.txt` (Cook and Jacobson, 1978). The three variables in the data set are $Photo$ = photo count, $Obs1$ = aerial count by observer one and $Obs2$ = aerial count by observer 2.

# A Diagnostic for Nonconstant Variance

## The Snow Geese Data

- Here we demonstrate calculation of the test statistic. This demonstration uses the `snowgeese` data.

```
> data(snowgeese)
> attach(snowgeese)
> library(xtable)
> m1 <- lm(photo~obs1,snowgeese)
> sig2 <- sum(residuals(m1)^2)/length(snowgeese$obs1)
> U <- residuals(m1)^2/sig2
> m2 <- lm(U~snowgeese$obs1)
> anova(m2)

Analysis of Variance Table

Response: U
               Df Sum Sq Mean Sq F value    Pr(>F)
snowgeese$obs1  1 162.83 162.826  50.779 8.459e-09 ***
Residuals      43 137.88   3.207
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> S <- anova(m2)$'Sum Sq'[1]/2
> p.value <- 1-pchisq(S,1)
> S

[1] 81.41318

> p.value

[1] 0
```

# A Diagnostic for Nonconstant Variance

## The Snow Geese Data

- However, a much easier way to do it is to use the library `lmtest`, then employ the `bptest` function,
- To get the same output as ALR, you have to set the option `studentize=FALSE`.

```
> library(lmtest)
> bptest(photo~obs1,studentize=F)

        Breusch-Pagan test

data:  photo ~ obs1
BP = 81.4132, df = 1, p-value < 2.2e-16
```

- You can also use the `car` library and its `ncv.test` function

```
> library(car)
> ncvTest(m1,~obs1)

Non-constant Variance Score Test
Variance formula: ~ obs1
Chisquare = 81.41318    Df = 1     p = 1.831324e-19
```

# A Diagnostic for Nonconstant Variance

## The Sniffer Data

- The sniffer data example on page 166–167 of ALR4 implement the Breusch-Pagan statistic in diagnosing and compensating for nonconstant variance.
- When gasoline is pumped into a tank, hydrocarbon vapors are forced out of the tank and into the atmosphere.
- To reduce this significant source of air pollution, devices are installed to capture the vapor.
- In testing these vapor recovery systems, a "sniffer" measures the amount recovered.
- To estimate the efficiency of the system, some method of estimating the total amount given off must be used.

# A Diagnostic for Nonconstant Variance

## The Sniffer Data

- In a controlled experiment, 4 predictors of the response $Y$ (amount given off) were measured:
  - *TankTemp*, the initial tank temperature in F$^\circ$
  - *GasTemp*, temperature of the dispensed gasoline in F$^\circ$
  - *TankPres*, initial vapor pressure in the tank in psi.
  - *GasPres* vapor pressure of the dispensed gasoline in psi.
- The reponse $Y$ is the hydrocarbons emitted, in grams.

# A Diagnostic for Nonconstant Variance

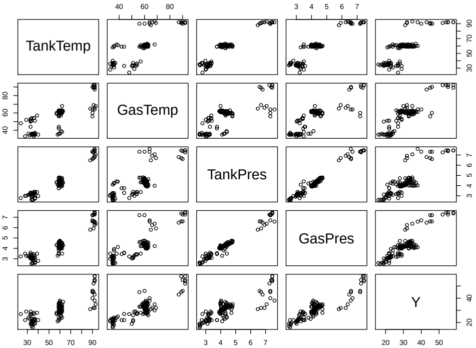
## The Sniffer Data

- We can start by looking at a scatterplot matrix for all the variables.
- Three notable trends are evident:
    - First, there several of the plots show concentration in some regions, indicating, selection of specific values, probably for substantive reasons.
    - Second, there is substantial linearity, indicating that transformations are not necessary.
    - There is substantial linear redundancy between the pressure predictors

# A Diagnostic for Nonconstant Variance

## The Sniffer Data

```
> data(sniffer)
> attach(sniffer)
> pairs(sniffer)
```



# A Diagnostic for Nonconstant Variance

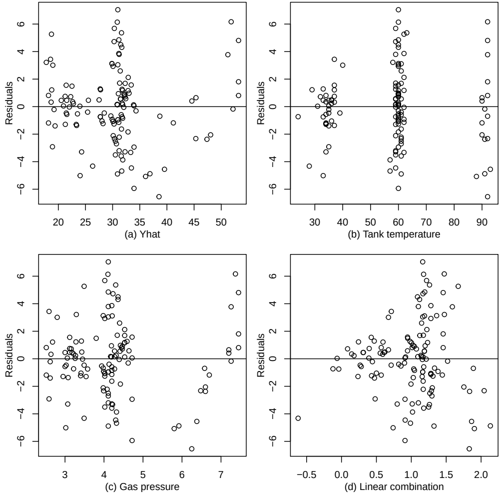
## The Sniffer Data

- Examining some residual plots using the code below, we see that variance does seem to increase from left to right in plots of *TankTemp* and *GasPres*.

```
> pdf("ALR_FIG0810.pdf", onefile=T)
> m1 <- lm(Y~TankTemp+GasTemp+TankPres+GasPres,sniffer)
> op<-par(mfrow=c(2,2),mar=c(4,3,0,.5)+.1,mgp=c(2,1,0))
> plot(predict(m1),residuals(m1),xlab="(a) Yhat", ylab="Residuals")
> abline(h=0)
> plot(TankTemp,residuals(m1),xlab="(b) Tank temperature",
+       ylab="Residuals")
> abline(h=0)
> plot(GasPres,residuals(m1),xlab="(c) Gas pressure",
+       ylab="Residuals")
> abline(h=0)
> U <- residuals(m1)^2*125/(sum(residuals(m1)^2))
> m3 <- update(m1,U~.)
> plot(predict(m3),residuals(m1),xlab="(d) Linear combination",
+       ylab="Residuals")
> abline(h=0)
```

# A Diagnostic for Nonconstant Variance

## The Sniffer Data



# A Diagnostic for Nonconstant Variance

## The Sniffer Data

- Weisberg goes on to conduct a sequence of tests for nonconstant variance under choice of various predictors. Results are displayed in Table 7.4, page 167:

**Table 7.4 Score Tests for the Sniffer Data**

| Choice for *Z* | *df* | Test stat. | *p*-Value |
|---|---|---|---|
| `GasPres` | 1 | 5.50 | .019 |
| `TankTemp` | 1 | 9.71 | .002 |
| `TankTemp, GasPres` | 2 | 11.78 | .003 |
| `TankTemp, GasTempTankPres, GasPres` | 4 | 13.76 | .008 |
| Fitted values | 1 | 4.80 | .028 |

- By subtraction, we can compare nested models, with a $\chi^2$ *difference test*. The difference between two nested model $\chi^2$ statistics, $\chi^2_a - \chi^2_b$, has an approximate $\chi^2$ distribution with $df_a - df_b$ degrees of freedom.
- In this case, if we first compare the statistic for `TankTemp, GasPres` with the statistic for `TankTemp`, we find that the difference test has a $\chi^2 = 11.78 - 9.71 = 2.07$ with 1 degree of freedom, which is not significant, indicating that `GasPres` does not improve the prediction of variance significantly better than `TankTemp`.
- We also see that adding three additional predictors does not improve significantly over the use of `TankTemp`, as the difference statistic is $\chi^2_3 = 13.76 - 9.71 = 4.05$, which is also nonsignificant.

# A Diagnostic for Nonconstant Variance

## The Sniffer Data

- We arrive at the decision to model the variance as

$$\text{Var}(Y|X,Z) = \sigma^2 \times \texttt{TankTemp} \tag{2}$$

  thereby using $1/\texttt{TankTemp}$ values as weights in weighted least squares.
- Note. If you compare the above with the discussion on page 166 of ALR4, you will discover that the textbook has an error. This traces back to ALR3, the previous edition.
- In ALR3, the corresponding table (Table 8.4 in ALR3) had the test statistics for `TankTemp` and `GasPres` reversed.
- The discussion in ALR3 was based on those reversed values.
- The table was corrected in ALR4, but unfortunately the discussion was not.