Heteroskedasticity

Note on required packages:

The following code requires the packages `sandwich` and `lmtest` to estimate regression error variance that may change with the explanatory variables. If you have not done so on your machine, download the package `lmtest`. This should only need to be done once on your computer.

```
install.packages("lmtest")
```

You can call the libraries `sandwich` and `lmtest` with the following calls to `library`:

```
library("sandwich")
library("lmtest")
```

```
## Loading required package: zoo
##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

---

1. Introduction

Homoskedasticity is the property that the variance of the error term of a regression (estimated by the variance of the residual in the sample) is the same across different values for the explanatory variables, or the same across time for time series models.

Heteroskedasticity is the property when the variance of the error term changes predictably with one or more of the explanatory variables.

Heteroskedasticity is common with financial variables related to income or spending. Suppose you are interested in predicting spending on housing and one of your explanatory variables is income. When income is low and the predicted value for spending on housing is low, the errors you make in the regression are also small. For people in low income groups, the amount each spends on housing varies little from other people in the low income group. For people in high income groups, the amount each spending on housing can vary greatly from other people in the high income group. The variance of your error term increases as the predicted value for y increases. This is heteroskedasticity.

Problems and non-problems with heteroskedasticity:

1. T-statistics and p-values from ordinary least squares (OLS) results assumes homoskedasticity.

2. Even with heteroskedasticity, the OLS estimates for your coefficients and predicted values for y are still *unbiased*. OLS is still useful.

3. OLS estimates for the *variances* of your coefficients are biased with heteroskedasticity.

4. There is a straightforward correction to the OLS variances to allow for heteroskedasticity, holding on to your OLS estimates for the coefficients and predicted values.

2. Example: Factors affecting monthly earnings

Let us examine a data set that explores the relationship between total monthly earnings (`MonthlyEarnings`) and a number of factors that may influence monthly earnings including including each person's IQ (`IQ`), a measure of knowledge workplace environment (`Knowledge`), years of education (`YearsEdu`), years experience (`YearsExperience`), and years at current job (`Tenure`).

The code below downloads a CSV file that includes data on the above variables from 1980 for 935 individuals and assigns it to a data set that we name `wages`.

```
wages <- read.csv("http://murraylax.org/datasets/wage2.csv");
```

The following call to `lm()` estimates a multiple regression predicting monthly earnings based on the five explanatory variables given above. The call to `summary()` displays some summary statistics from the regression.

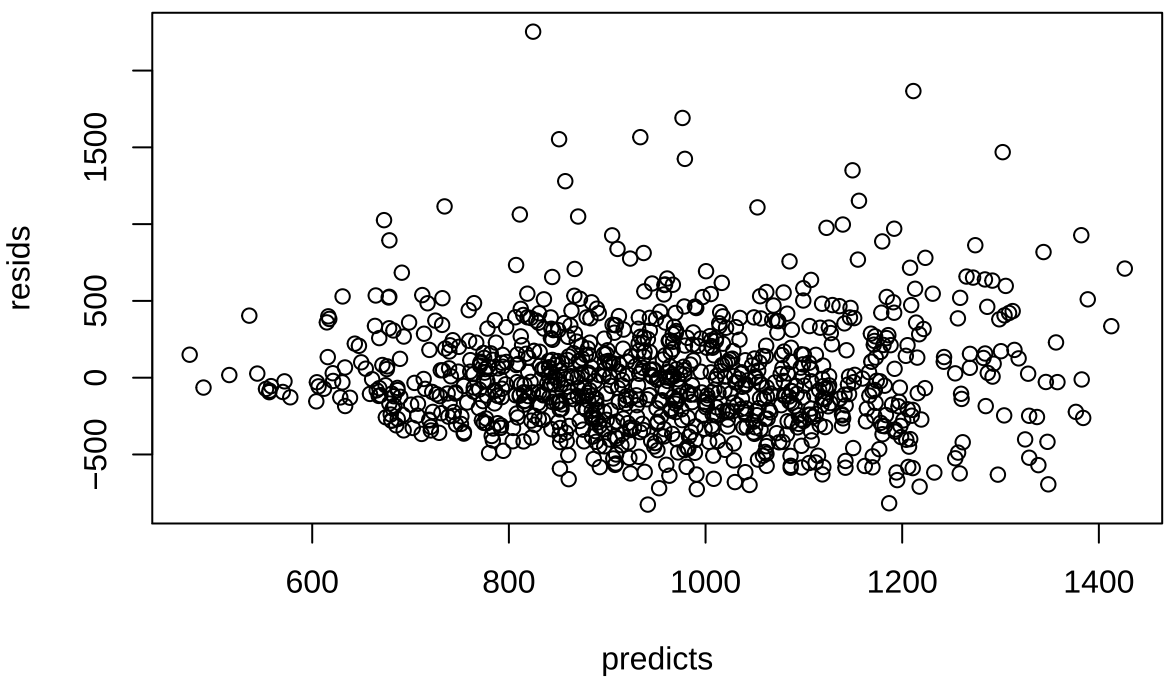
```
wages <- read.csv("http://murraylax.org/datasets/wage2.csv");
lmwages <- lm(MonthlyEarnings
              ~ IQ + Knowledge + YearsEdu + YearsExperience + Tenure,
              data = wages)
```

Let's save the residuals and predicted values from the regression:

```
resids <- lmwages$residuals
predicts <- lmwages$fitted.values
```

Let's plot the residuals to see if they are more spread out for larger values of the predicted values:

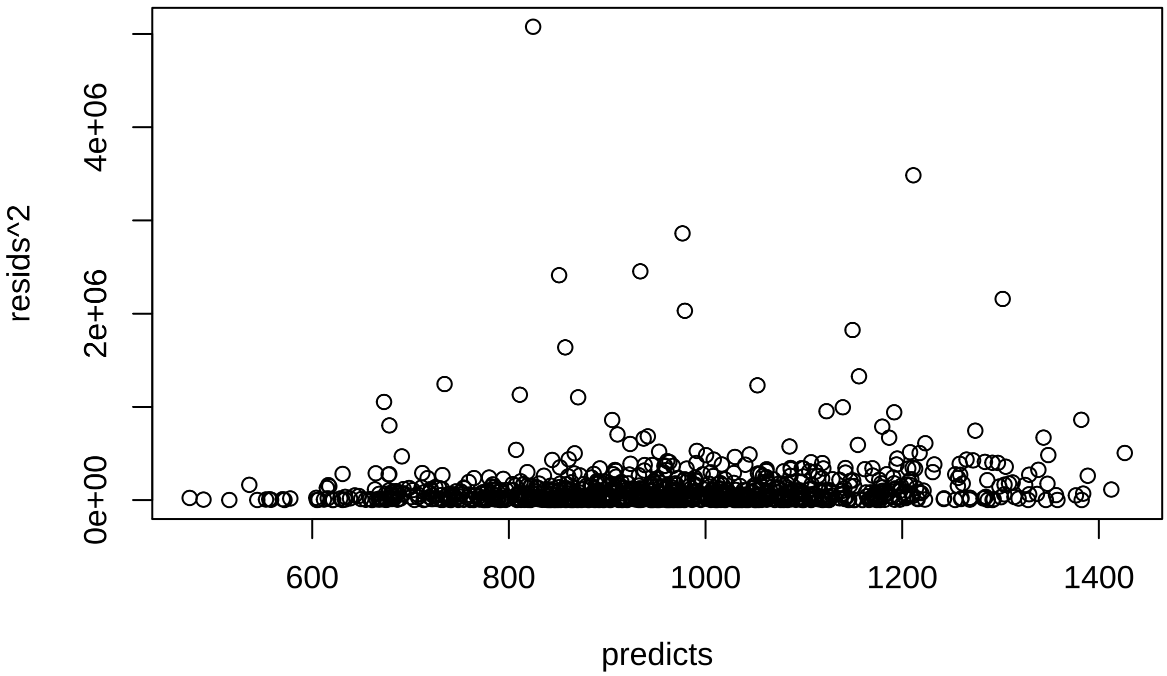
```
plot(y=resids, x=predicts)
```



It does appear that the distribution is less spread out at the left end. That is, at the lowest levels for predicted monthly earnings, the residuals are smaller.

Let's plot the *squared* residual against the predicted values. The squared residual is an estimate of the variance of the error term:

```
plot(y=resids^2, x=predicts)
```



A visual inspection suggests there may be a pattern. As the predicted values increase, the variance of the error term increases.

Ordinary least squares assumes homoskedasticity which assumes the variance for the error term is the same for all predicted values. The estimate for the variance is simply the average of the squared residuals (corrected for degrees of freedom = n-1). When we have heteroskedasticity, the variance for the residual on each observation is different, and given by each observation's squared residual shown in the graph above.

In Section 3 below, we learn how to correct the OLS estimates for the variances for the coefficients and predicted values making use of these squared residuals.

In Section 4 below, we learn how to more formally test for the presence for heteroskedasticity. It is based on the relationships we see in these graphs between the squared residuals and the predicted values.

3. Heteroskedasticity robust standard errors

White's correction for heteroskedasticity is a method for using OLS estimates for coefficients and predicted values (which are unbiased), but fixing the estimates for the variances of the coefficients (which are biased). It uses as an estimate for the *possibly changing* variance the squared residuals estimated from OLS which we computed and graphed above.

In R, we can compute the White heteroskedastic variance/covariance matrix for the coefficients with the call to `vcovHC` (which stands for Variance / Covariance Heteroskedastic Consistent):

```
vv <- vcovHC(lmwages, type="HC1")
```

The first parameter in the call above is our original output from our call to `lm()` above. The second parameter `type="HC1"` tells the function to use the White estimate for the variance covariance matrix which uses as an estimate for the changing variance the squared residuals from the OLS call.

Then we can use this estimate for the variance / covariance to properly compute our standard errors, t-statistics, and p-values for the coefficients:

```r
coeftest(lmwages, vcov = vv)
```

```
##
## t test of coefficients:
##
##                    Estimate Std. Error t value  Pr(>|t|)
## (Intercept)     -531.03920  112.19185 -4.7333 2.553e-06 ***
## IQ                 3.69656    0.93439  3.9561 8.198e-05 ***
## Knowledge          8.27030    2.01590  4.1025 4.445e-05 ***
## YearsEdu          47.26976    7.46751  6.3301 3.809e-10 ***
## YearsExperience   11.85891    3.22845  3.6733 0.0002532 ***
## Tenure             6.24651    2.51596  2.4828 0.0132128 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The first parameter to `coeftest` is the result of our original call to `lm()` and the second parameter is the updated variance / covariance matrix to use.

Let's compare the result to the OLS regression output:

```r
summary(lmwages)
```

```
##
## Call:
## lm(formula = MonthlyEarnings ~ IQ + Knowledge + YearsEdu + YearsExperience +
##     Tenure, data = wages)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -826.33 -243.85  -44.83  180.83 2253.35
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -531.0392   115.0513  -4.616 4.47e-06 ***
## IQ                 3.6966     0.9651   3.830 0.000137 ***
## Knowledge          8.2703     1.8273   4.526 6.79e-06 ***
## YearsEdu          47.2698     7.2980   6.477 1.51e-10 ***
## YearsExperience   11.8589     3.2494   3.650 0.000277 ***
## Tenure             6.2465     2.4565   2.543 0.011156 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 365.4 on 929 degrees of freedom
## Multiple R-squared:  0.1878, Adjusted R-squared:  0.1834
## F-statistic: 42.97 on 5 and 929 DF,  p-value: < 2.2e-16
```

The coefficients are exactly the same, but the estimates for the standard errors, the t-statistics, and p-values are slightly different.

One reason the standard errors and p-values are only slightly different is that there may actually be very little heteroskedasticity. The differences in variance of the error term may be small even for large differences in the predicted value for x.

The White heteroskedastic robust standard errors are valid for **either homoskedasticity or heteroskedasticity**, so it is always safe to use these estimates if you are not sure if the homoskedasticity assumption holds.

4. Testing for heteroskedasticity

In section 2 we examined visual evidence that the magnitudes of the residuals or squared residuals were on average larger for larger predicted values for monthly earnings. In this section, we will test this relationship more formally.

The goal is to determine whether the squared residuals can be explained by the predicted values. Both, the squared residuals and predicted values from the OLS regression are unbiased, so we will use these values.

We run the following regression of the squared residuals on the predicted values and squared predicted values:

```r
resids2 <- resids^2;  ## Save the squared residuals
predicts2 <- predicts^2; ## Save the suared predicted values
lmhet <- lm(resids2 ~ predicts + predicts2)
summary(lmhet)
```

```
##
## Call:
## lm(formula = resids2 ~ predicts + predicts2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -244651 -108628  -76561   14043 4973137
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.309e+05  2.427e+05   0.540    0.590
## predicts    -2.013e+02  5.092e+02  -0.395    0.693
## predicts2    2.052e-01  2.623e-01   0.782    0.434
##
## Residual standard error: 313300 on 932 degrees of freedom
## Multiple R-squared:  0.01235,    Adjusted R-squared:  0.01023
## F-statistic: 5.829 on 2 and 932 DF,  p-value: 0.003051
```

Do the predicted values from the regression explain the squared residuals? We answer this with the joint F-test.

$$H_0 : \beta_{predicts} = 0;\ \beta_{predicts^2} = 0 \quad \text{(i.e. homoskedasticity)}$$
$$H_A : \text{Either } \beta_{predicts} \neq 0 \text{ or } \beta_{predicts^2} \neq 0 \quad \text{(i.e. heteroskedasticity)}$$

If the null hypothesis is true and both coefficients are equal to zero, then the predicted values of the regression do not explain the squared residuals. In this case we say we have homoskedasticity: the variance of the residuals do not change predictably with the predicted values.

If the alternative hypothesis is true then we have statistical evidence that either the predicted values or the squared predicted values helps explain the variance of the error term. In this case we have heteroskedasticity: As the predicted values change, so does the variance of the error term.

The p-value is 0.003. We find sufficient statistical evidence that the model does exhibit heteroskedasticity, so the White heteroskedastic robust test above is the most appropriate.

This procedure for testing for heteroskedasticity is called the *White test for heteroskedasticity.*

5. Functional Form

Heteroskedasticity on its own is a small problem. OLS estimates for coefficients are still unbiased, and it is easy to correct standard errors and p-values to allow for the possibility of heteroskedasticity.

However, sometimes heteroskedasticity is the result of a *mis-specified model*. That is, you estimated a functional form for the model that is not appropriate, that is not true in the data.

It is common for financial variables and variables related to income, that larger numbers grow exponentially. When we take the natural log of these variables, the transformed values grow linearly, and linear regression becomes more applicable.

Let us revisit the model above, but instead use the *natural log of monthly earnings* as the outcome variable:

```
lmlogwages <- lm(log(MonthlyEarnings)
              ~ IQ + Knowledge + YearsEdu + YearsExperience + Tenure,
              data = wages)
```

Let us compute the White test for heteroskedasticity:

```
resids <- lmlogwages$residuals
resids2 <- resids^2
predicts <- lmlogwages$fitted.values
predicts2 <- predicts^2
lmhet <- lm(resids2 ~ predicts + predicts2)
summary(lmhet)
```

```
##
## Call:
## lm(formula = resids2 ~ predicts + predicts2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.1550 -0.1308 -0.0812  0.0337  3.7723
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -12.5404     8.7662  -1.431    0.153
## predicts      3.7979     2.5853   1.469    0.142
## predicts2    -0.2840     0.1905  -1.491    0.136
##
## Residual standard error: 0.2535 on 932 degrees of freedom
## Multiple R-squared:  0.004038,   Adjusted R-squared:  0.001901
## F-statistic:  1.89 on 2 and 932 DF,  p-value: 0.1517
```

The p-value for the joint F-test for regression significance is 0.1517. Now we fail to reject. We fail to find statistical evidence that the variance of the residual changes with the predicted or squared predicted values. We failed to find statistical evidence for heteroskedasticity.

It is likely that the log-model is the appropriate model to use. The White test for heteroskedasticity can be a useful check on the appropriateness of a model.