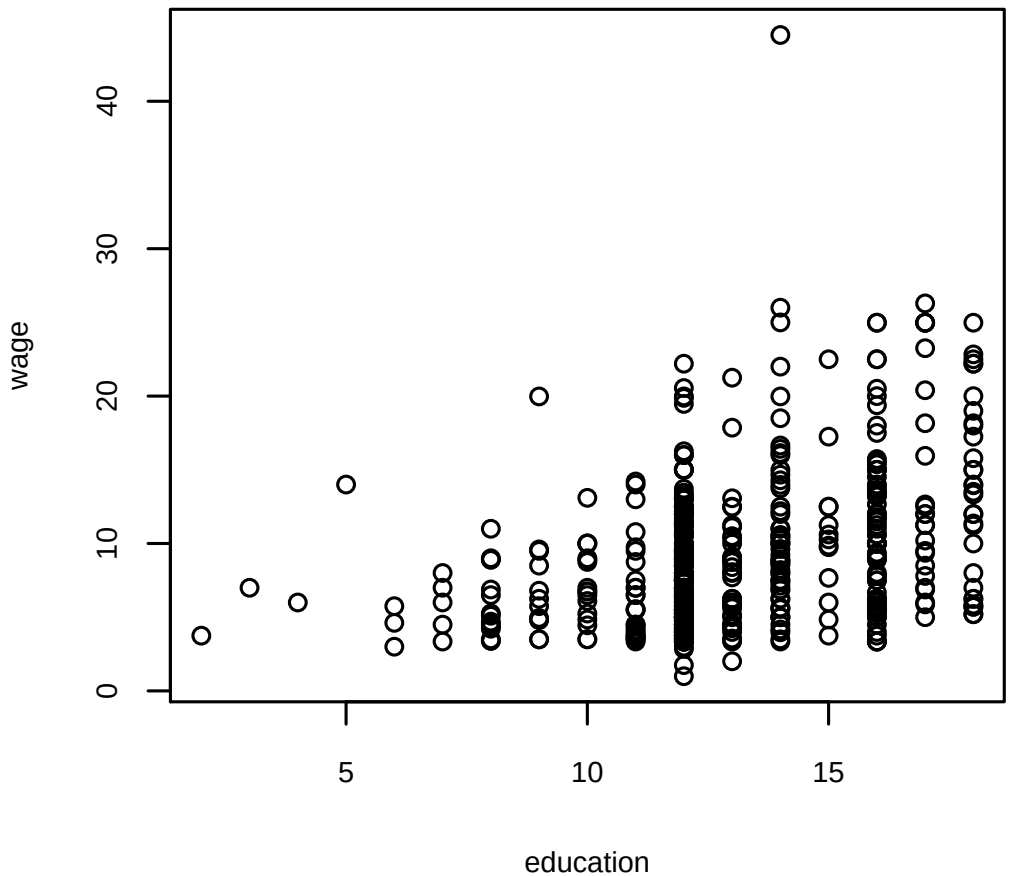
Econometrics I - Heteroskedasticity

Ryan T. Godwin

University of Manitoba

Figure: Possible heteroskedasticity in the CPS data. The variance in `wage` may be increasing as `education` increases. The reasoning is that individuals who have not completed highschool (or university) are precluded from many high-paying jobs (doctors, lawyers, etc.). However, having many years of education does not preclude individuals from low-paying jobs. The spread in wages is higher for highly educated individuals.



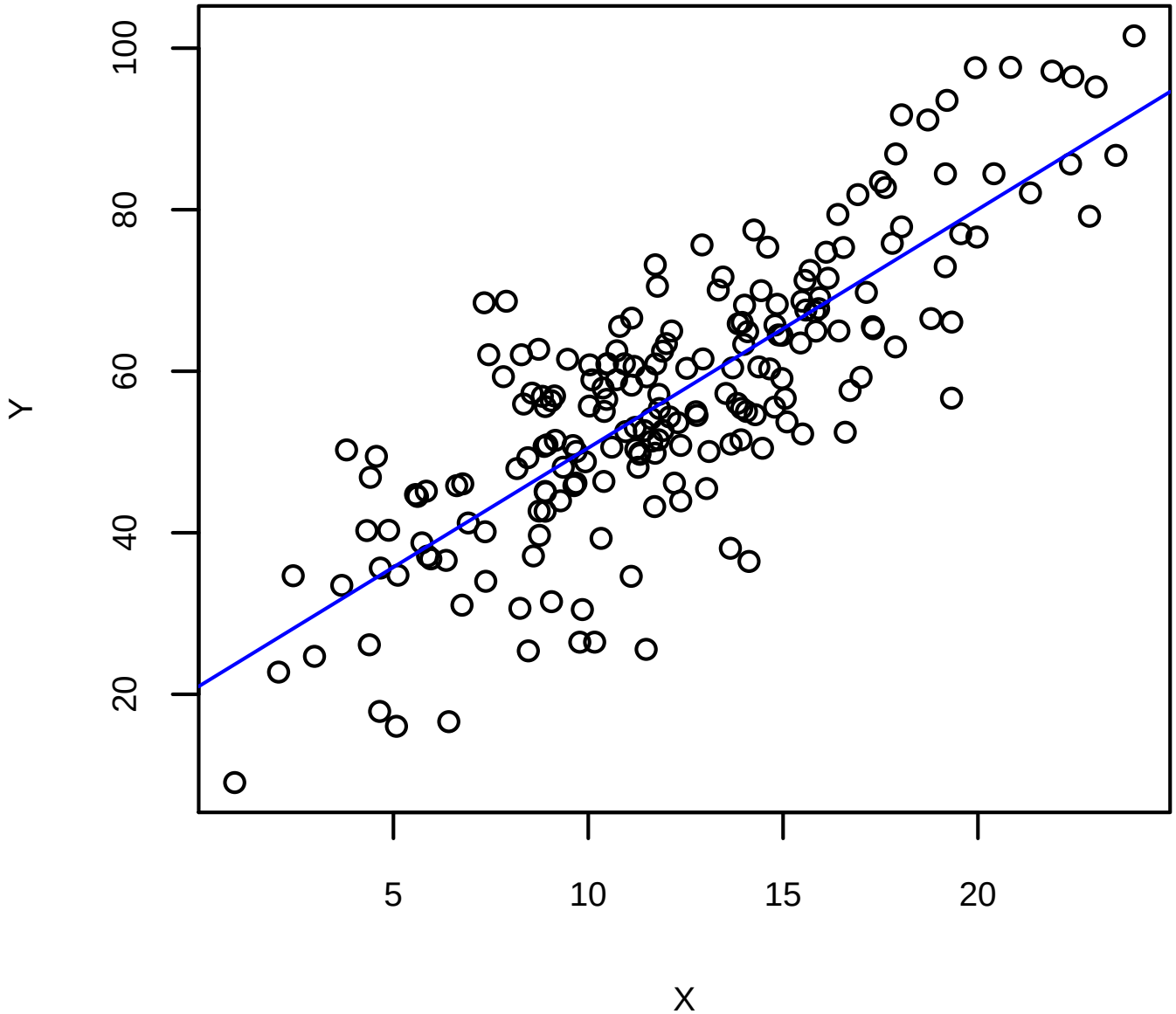
Homoskedasticity

If assumption A4 is satisfied, then ϵ is identically distributed. This means that all of the ϵ_i have the same variance. That is, all of the random effects that determine Y, outside of X, have the same dispersion. The term *homoskedasticity* (same dispersion) refers to this situation of identically distributed error terms.
Stated mathematically, homoskedasticity means:

$$\text{Var}[\epsilon_i|X_i] = \sigma^2 \ , \ \forall i$$

The variance of ϵ is constant, even conditional on knowing the value of X.
Homoskedasticity means that the squared vertical distance of each data point from the (population or estimated) line is, on average, the same. The values of the X variables do not influence this distance (the variance of the random unobservable effects are not determined by any of the values of X). See figure 2.

Figure: Homoskedasticity. The average squared vertical distance from the data points to the OLS estimated line is the same, regardless of the value of X.



Heteroskedasticity

Heteroskedasticity refers to the situation where the variance of the error term ϵ is not equal for all observations. The term heteroskedasticity means *differing dispersion*. Mathematically:

$$\text{Var}[\epsilon_i | X_i] \neq \sigma^2 \ , \ \forall i$$

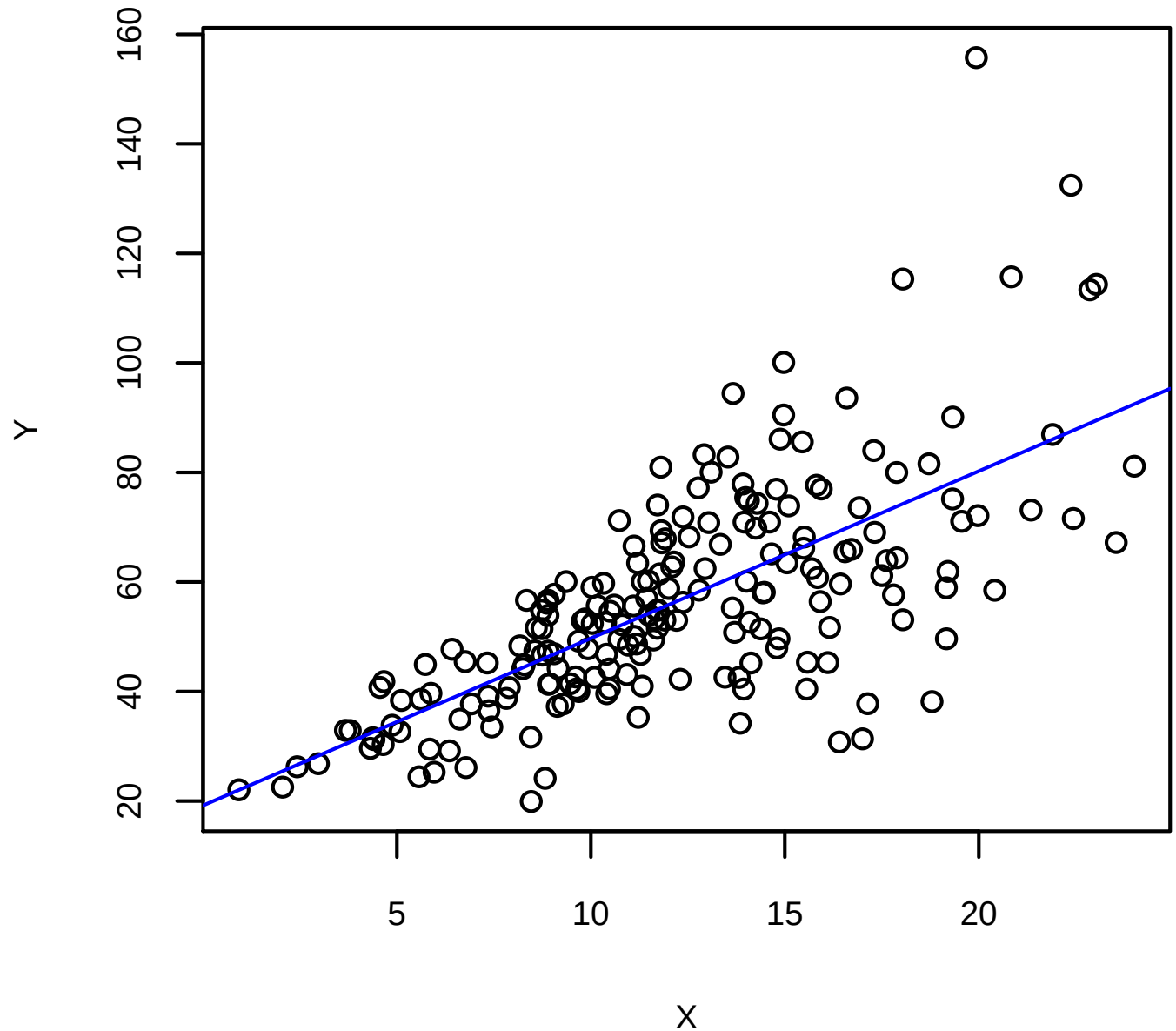
or

$$\text{Var}[\epsilon_i | X_i] = \sigma_i^2$$

Each observation can have its own variance, and the value of X may influence this variance.

Heteroskedasticity means that the squared vertical distance of each data point from the estimated regression line is not the same on average, and may be influenced by one or more of the X variables.

Figure: Heteroskedasticity. The squared vertical distance of a data point from the OLS estimated line is influenced by X.



Heteroskedasticity

In this chapter we revisit assumption A.4, which says:

$$\mathrm{V}\left[\boldsymbol{\epsilon}\right] = \sigma^2 I_n$$

The term "non-spherical disturbances" refers to the situation where $\mathrm{V}\left[\boldsymbol{\epsilon}\right] \neq \sigma^2 I_n$. In this chapter, we instead generalize the specification of the error term in the population model:

$$E[\boldsymbol{\epsilon}] = \mathbf{0} \quad ; \quad V[\boldsymbol{\epsilon}] = \sigma^2 \Omega = \Sigma \tag{1}$$

Equation 1 allows for the possibility of one or both of *heteroskedasticity* and *autocorrelation*. In this chapter we examine the situation of heteroskedasticity, and how this more general situation for the covariance matrix of the error term affects our LS estimator, and hypothesis testing.

The error term is said to be heteroskedastic when $\text{var}[\epsilon_i] = \sigma_i^2$, and there are some $\sigma_i^2 \neq \sigma_j^2$. That is, each observation can have a different variance, and the term "heteroskedasticity" means "differing dispersion." The alternative to heteroskedasticy is *homoskedasticity* (which we have been assuming via A.4), where $\text{var}[\epsilon_i] = \sigma^2$.

In the case of heteroskedasticity, the covariance matrix for the error term takes the form:

$$\sigma^2\Omega = \sigma^2 \begin{bmatrix} \omega_{11} & 0 & \cdots & 0 \\ 0 & \omega_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \omega_{nn} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix} = \text{diag}\left(\sigma_i^2\right)$$

When the error term $\boldsymbol{\epsilon}$ exhibits heteroskedasticity, we will find that:

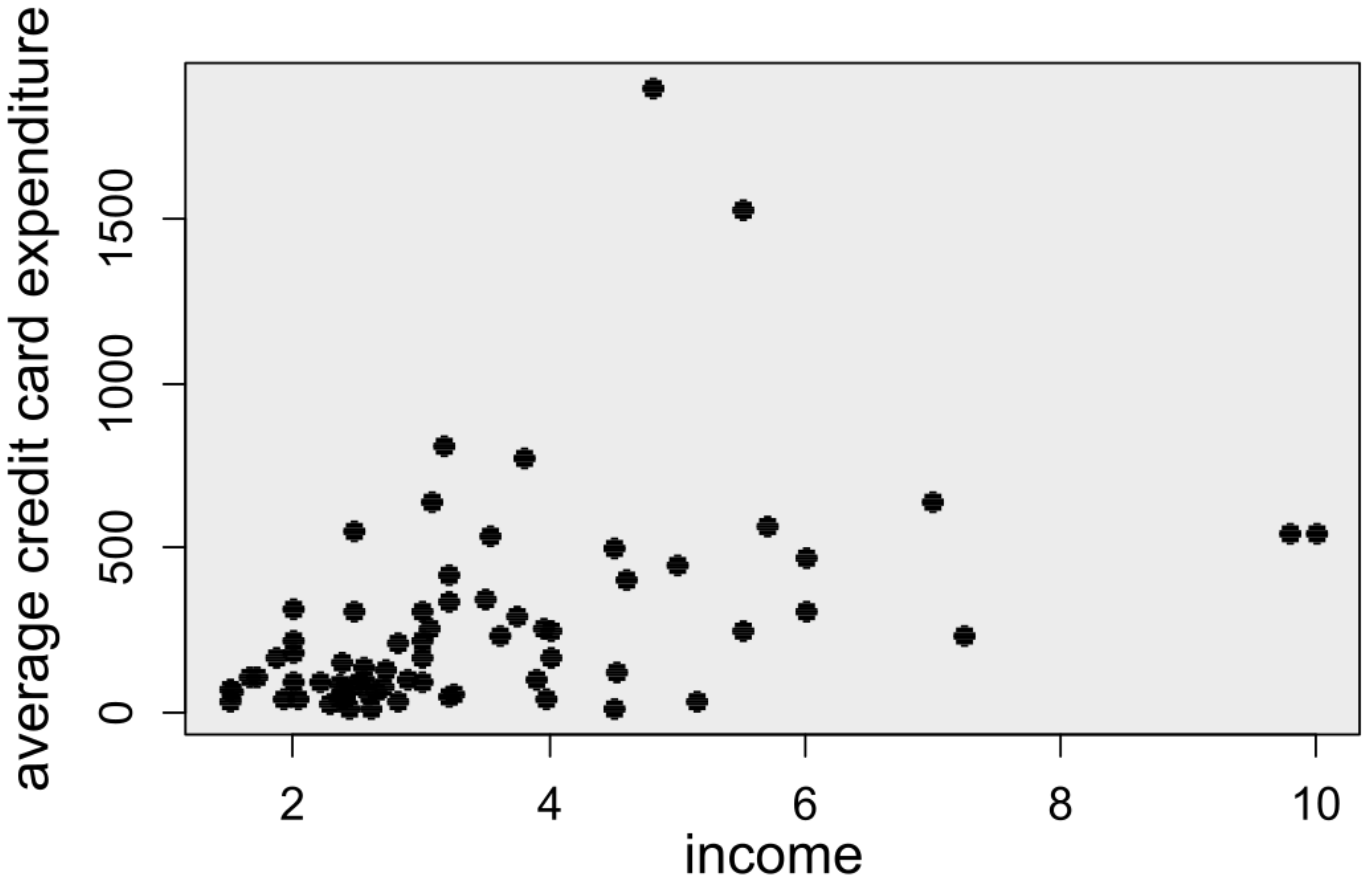
- The LS estimator is still unbiased and consistent.
- The LS estimator is now *inefficient*.
- The usual estimator for $\mathrm{V}[\boldsymbol{b}]$ (which has been $s^2(X'X)^{-1}$ in previous chapters) is now *inconsistent*, which invalidates hypothesis testing.

A solution to the inefficiency of LS is to use the generalized least squares (GLS) or feasible (FGLS) estimator, which also takes care of the inconsistency of the standard errors of $\boldsymbol{b}$. A common practice, however, is to ignore the inefficiency of LS and use a *robust* estimator for $\mathrm{V}[\boldsymbol{b}]$ (such as White's heteroskedastic robust covariance estimator).

Load and plot a dataset that potentially has heteroskedasticity (see Figure 4):

```
ccard <- read.csv("https://rtgodwin.com/data/creditcard.csv")
plot(ccard$income, ccard$avgexp)
```

Figure: Credit card expenditure data possibly exhibits heteroskedasticity.



Statistical properties of LS estimation in the presence of heteroskedasticity

Recall the proof that the LS estimator is unbiased:

$$\begin{aligned} \boldsymbol{b} &= (X'X)^{-1} X'\boldsymbol{y} = (X'X)^{-1} X'(X\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\ &= \boldsymbol{\beta} + (X'X)^{-1} X'\boldsymbol{\epsilon} \\ E(\boldsymbol{b}) &= \boldsymbol{\beta} + (X'X)^{-1} X'E(\boldsymbol{\epsilon}) = \boldsymbol{\beta} \end{aligned}$$

We need to use assumption A.3 and A.5 to establish this result, but we do not need A.4. Hence, heteroskedasticity does not affect the unbiasedness property of LS. Similarly, $\text{plim}[\boldsymbol{b}] = \boldsymbol{\beta}$ whether the error term is heteroskedastic or not. (The IV and NLS estimators will also be consistent in the presence of heteroskedasticity).

Now, let's consider the covariance matrix of our LS estimator under heteroskedasticity:

$$\begin{aligned} V(\boldsymbol{b}) &= V\left[\boldsymbol{\beta} + (X'X)^{-1} X'\boldsymbol{\epsilon}\right] = V\left[(X'X)^{-1} X'\boldsymbol{\epsilon}\right] \\ &= \left[(X'X)^{-1} X' V(\boldsymbol{\epsilon}) X (X'X)^{-1}\right] \\ &= \left[(X'X)^{-1} X' \sigma^2 \Omega X (X'X)^{-1}\right] \\ &\neq \left[\sigma^2 (X'X)^{-1}\right] \end{aligned}$$

where we have used assumption 1 instead of A.4. We can see that if $\Omega = I_n$ then we get the usual expression for $V(\boldsymbol{b})$.

The usual computer output (for example from `summary()`), will be using $s^2 (X'X)^{-1}$, which is the *wrong* formula! The standard errors, t-statistics, confidence intervals, will all be incorrect. The usual estimator for the covariance matrix of $\boldsymbol{b}$, namely $s^2 (X'X)^{-1}$, will be an *inconsistent* estimator of the true covariance matrix of $\boldsymbol{b}$.

The LS estimator will turn out to be *inefficient* under heteroskedasticity, but it is easiest to show this after we develop the generalized least squares (GLS) estimator, and so we postpone this discussion for later. For now, we turn to the most pressing issue - the inconsistency of the estimator for the covariance matrix of $\boldsymbol{b}$.

White's heteroskedastic consistent covariance matrix

If we knew Σ, then the "estimator" of the covariance matrix for $\boldsymbol{b}$ would just be:

$$\hat{\mathrm{V}}[\boldsymbol{b}] = \left[(X'X)^{-1} X'\Sigma X (X'X)^{-1} \right] \tag{2}$$

The covariance matrix in equation 2 is known as a *sandwich covariance matrix*. In practice, $\mathrm{V}[\boldsymbol{\epsilon}] = \Sigma$ will usually be unknown and need to be estimated. But since Σ is $n \times n$ and explodes as $n \to \infty$, it seems hopeless to try to get a consistent estimator for Σ. However, we can find a consistent estimator when we consider the entire *middle* of the sandwich.

For asymptotic theory, what we actually need is an estimator for the covariance matrix of $\sqrt{n}(\boldsymbol{b} - \boldsymbol{\beta})$, not $\boldsymbol{\beta}$. By distributing the factor n we can rewrite equation 2 as:

$$\hat{\mathrm{V}}[\boldsymbol{b}] = \frac{1}{n}\left[\left(\frac{1}{n}X'X\right)^{-1}\left(\frac{1}{n}X'\Sigma X\right)\left(\frac{1}{n}X'X\right)^{-1}\right]$$

where we see that we need to find a consistent estimator of $\frac{1}{n}X'\Sigma X$. While Σ is $n \times n$ and explodes as $n \to \infty$, the matrix $\frac{1}{n}X'\Sigma X$ is a $k \times k$ symmetric matrix, and has k distinct elements in the diagonal (with autocorrelation there would be $\frac{1}{2}(k^2 + k)$ distinct elements).

Let $Q^* = \left(\frac{1}{n} X'\Sigma X\right)$. In the case of just heteroskedasticity (for autocorrelation we would have $\boldsymbol{x}_i \boldsymbol{x}_j'$ terms), Q^* becomes:

$$Q^* = \frac{1}{n} \sum_{i=1}^{n} \sigma_i^2 \boldsymbol{x}_i \boldsymbol{x}_i'$$

White (1980) showed that if we define

$$S_0 = \frac{1}{n} \sum_{i=1}^{n} e_i^2 \boldsymbol{x}_i \boldsymbol{x}_i'$$

then

$$\operatorname{plim}(S_0) = Q^*$$

Since LS is still consistent under heteroskedasticity, the residuals $\boldsymbol{e}$ are still consistent estimators for $\boldsymbol{\epsilon}$. This means that we can estimate the model by LS, get the residuals $\boldsymbol{e}$, and then a consistent estimator of $\mathrm{V}[\boldsymbol{b}]$ will be:

$$\hat{\mathrm{V}}[\boldsymbol{b}] = \frac{1}{n}\left[\left(\frac{1}{n}X'X\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} e_i^2 \boldsymbol{x}_i \boldsymbol{x}_i'\right)\left(\frac{1}{n}X'X\right)^{-1}\right]$$

In practice we ignore the n^{-1} and use:

$$\hat{\mathrm{V}}[\boldsymbol{b}] = (X'X)^{-1}\left(\sum_{i=1}^{n} e_i^2 \boldsymbol{x}_i \boldsymbol{x}_i'\right)(X'X)^{-1} \tag{3}$$

which amounts to replacing every diagonal element of Σ with a squared residual. The sandwich estimator in 3 is called a *heteroskedasticity-consistent covariance matrix estimator*, and is valid regardless of the unknown form of the heteroskedasticity. Taking the square roots of the diagonal elements of 3 gives us the het-consistent, or "robust" standard errors.

There are alternatives to the sandwich estimator in 3. Alternate versions include multiplying the entire matrix by $n/(n-k)$ as a degrees of freedom correction, or using $e_i^2/(1-h_i)$ instead of just e_i^2, where h_i is the i^{th} diagonal element of the P_X matrix. All of the alternatives are consistent estimators, and differ in their *finite* sample properties, which vary depending on the data.

As a result of using a sandwich estimator such as in 3, the t-statistics, F-statistic, standard errors, confidence intervals, etc. will be modified, but only in a manner that is appropriate asymptotically. This means that the usual test statistics will be unreliable in finite samples, and instead of the t-distribution and F-distribution we should use their asymptotic approximations: the standard Normal and Chi-square distributions.

Example - Robust standard errors.

Use the credit card expenditure data to estimate the model:

$$\text{avgexp} = \beta_1 + \beta_2 \text{age} + \beta_3 \text{ownrent} + \beta_4 \text{income} + \beta_5 \text{income}^2 + \epsilon$$

Download the data:

```
ccard<- read.csv("https://rtgodwin.com/data/creditcard.csv")
```

Estimate the model assuming *homoskedasticity*:

```
ccard.mod <- lm(avgexp ~ age + ownrent + income
        + I(income^2), data = ccard)
summary(ccard.mod)
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -237.147    199.352  -1.190  0.23841
age           -3.082      5.515  -0.559  0.57814
ownrent       27.941     82.922   0.337  0.73721
income       234.347     80.366   2.916  0.00482 **
I(income^2)  -14.997      7.469  -2.008  0.04870 *
```

If we have heteroskedasticity, then the standard errors, t-statistics, and associated p-values, are all wrong! Install and load a package capable of "sandwich" covariance matrix estimation:

```
install.packages("sandwich")
library(sandwich)
```

and get White's heteroskedastic consistent covariance matrix estimator from equation 3 (we can change the `type` to use alternate estimators):

```
coeftest(ccard.mod, vcov = vcovHC(ccard.mod, "HC1"))
```

```
               Estimate Std. Error t value Pr(>|t|)
(Intercept) -237.1465   220.7950 -1.0741  0.28665
age           -3.0818     3.4226 -0.9004  0.37112
ownrent       27.9409    95.5657  0.2924  0.77090
income       234.3470    92.1226  2.5439  0.01328 *
I(income^2)  -14.9968     7.1990 -2.0832  0.04105 *
```

The standard errors have either decreased or increased, and some are quite different! The significance of one of the regressors has changed, for example. Ignoring the possibility of heteroskedasticity, and thus using the wrong standard errors, can invalidate hypothesis testing.

Testing for homoskedasticity

Heteroskedasticity reduces the efficiency of the LS estimator of $\boldsymbol{\beta}$ (we still haven't showed this) and has serious implications for the properties of the associated standard errors, confidence intervals, and tests. It would be very useful to have a test of the hypothesis that the errors in our regression model are homoskedastic, against the alternative that they exhibit some sort of heteroskedasticity. Because LS is still a consistent estimator of $\boldsymbol{\beta}$ even if the errors are heteroskedastic, we can use the LS residuals to construct tests that will still be (at least) asymptotically valid.

White's test

Consider the following null and alternative hypotheses under the standard population model:

$$H_0 : \sigma_i^2 = \sigma^2 \quad ; \quad i = 1, 2, \ldots, n \quad \text{vs.} \quad H_A : \text{ Not } H_0$$

The alternative hypothesis is very general, and no specific form of heteroskedasticity has been declared. To implement the test:

1. Estimate the model by LS, and get the residuals, $e_i \quad ; \quad i = 1, 2, \ldots, n$.
2. Using LS again, regress the e_i^2 values on each of the x's in the original model; their squared values; all of the cross-products of the regressors; and an intercept. We are using the information in X to approximate any possible unknown form of heteroskedasticity.
3. The nR^2 from the regression in Step 2 is asymptotically $\chi^2_{(p)}$ (Chi-square distributed) if H_0 is true; where p is the number of parameters that are estimated at Step 2.
4. Reject H_0 in favour of H_A if the p-value for the nR^2 statistic from the chi-square distribution is small.

Note the limitation of this test:

- It is valid only asymptotically.
- The test is "non-constructive", in the sense that if we reject H_0, we don't know what form of heteroskedasticity we may have.
- This means that it won't be clear what form the GLS estimator (in the next section) should take.

Even though White's test is non-constructive, it can provide enough information to alert us to use White's heteroskedasticity-consistent estimator of $\mathrm{V}(\boldsymbol{b})$. In fact, there is little, if anything, to be lost in using this covariance matrix estimator, as long as the sample is large. This is because homoskedasticity is just a *special case* of heteroskedasticity. That is, the heteroskedastic consistent covariance matrix estimators do not rule out the possibility of homoskedasticity.

White's test in R

Use the data and model from Example 21 to test for the presence of heteroskedasticity:

```
ccard <- read.csv("https://rtgodwin.com/data/creditcard.csv")
ccard.mod <- lm(avgexp ~ age + ownrent + income + I(income^2), data = ccard)
```

Install and load a package:

```
install.packages("skedastic")
library(skedastic)
```

and calculate White's test using the `white()` function:

```
white(ccard.mod, interactions = TRUE)
```

```
  statistic p.value parameter method        alternative
      <dbl>   <dbl>     <dbl> <chr>         <chr>
1      14.3   0.426        14 White's Test greater
```

The White test statistic is 14.3, with a Chi-square p-value of 0.426. We fail to reject the null of homoskedasticity after all! Are the degrees of freedom for the Chi-square distribution right?

White's test by hand

Use the data from Example 21 and 26 to test for the presence of heteroskedasticity "by hand". Get the squared residuals from the estimated model:

```
ccard <- read.csv("https://rtgodwin.com/data/creditcard.csv")
ccard.mod <- lm(avgexp ~ age + ownrent + income + I(income^2), data = ccard)
ccard.res.sq <- ccard.mod$residuals ^ 2
```

and regress the squared residuals on all regressors, squared regressors, and cross-products:

```r
summary(lm(ccard.res.sq ~ age + ownrent + income
       + I(income^2) + I(age^2) + age*ownrent
       + age*income + age*I(income^2) + ownrent^2
       + ownrent*income + ownrent*I(income^2) + I(income^2)
       + I(income^3)+ I(income^4), data=ccard))
```

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          1637390.4  1290979.7   1.268   0.2097
age                     5366.2    48893.8   0.110   0.9130
ownrent                812036.8   991630.2   0.819   0.4161
income               -2021697.6  1053559.1  -1.919   0.0598 .
I(income^2)            669055.3   365666.7   1.830   0.0724 .
I(age^2)                 -424.1      627.5  -0.676   0.5018
I(income^3)            -86805.3    51162.6  -1.697   0.0950 .
I(income^4)              3762.7     2277.4   1.652   0.1038
age:ownrent              4661.7    14424.6   0.323   0.7477
age:income              11499.9    15614.3   0.736   0.4643
age:I(income^2)         -1093.3     1568.1  -0.697   0.4884
ownrent:income        -510192.3   469792.6  -1.086   0.2819
ownrent:I(income^2)     51835.1    61799.8   0.839   0.4050
---
Signif.codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 274600 on 59 degrees of freedom
Multiple R-squared:  0.199, Adjusted R-squared:  0.0361
F-statistic: 1.222 on 12 and 59 DF,  p-value: 0.2905
```

We are essentially looking at the "overall fit" of this auxiliary regression, or the "joint significance" of all of the explanatory variables. Typically we would look at the p-value of 0.2905 for the joint significance. But this is an F-test, and we are in an asymptotic setting. So, instead of the F-test and F-distribution we use the Wald test and the Chi-square distribution. The Wald test statistic is $nR^2 = 72 \times 0.199 = 14.3$ (same as from example 26) and the associated p-value is:

```
1 - pchisq(72 * 0.199, 14)
```

```
[1] 0.4255717
```

But the degrees of freedom of 14 is wrong! Two of the cross-products are redundant and have been dropped from the auxiliary regression, leaving us with $p = 12$ and the proper p-value is:

```
1 - pchisq(72 * 0.199, 12)
```

```
[1] 0.280255
```

The `white()` provides the wrong p-value. In any case, we cannot reject the null of homoskedasticity using White's test, even though heteroskedasticity seems apparent from Figure 4. What would be the safe thing to do in this case?