Chapter 13:

Assumptions of Linear Regression

In the previous chapter, we examined how to do linear regression. We know that calculating the line of best fit is rather straight-forward, an application of either calculus or matrix algebra. As we have defined "best fit," this line is unique. We also interpreted that line of best fit, given that the assumptions of ordinary least squares regression are not violated by the model.

While the previous chapter did state the assumptions, it did not cover how to test those assumptions. This chapter does just that. Here, we offer several methods for testing the model to look for violations.

❧❧❧

Explaining a person's income is an interesting exercise as it allows us to determine what contributes to wealth. To determine some factors affecting personal income, we surveyed 243 people, asking a variety of questions, ranging from their parent's average IQ to the animals the respondent owned. Which of those variables most affected personal income?

The previous chapter focused on the mechanics of ordinary least squares regression and of interpreting and graphing the results. While the assumptions were mentioned, the chapter did not cover how to test them. We will here.

Recall that a line of best fit is a line that minimizes some function of the errors. Frequently, this function is the sum of the squared errors, with the error measured as vertical distance between the observed value and the line. This line of best fit is unique *if and only if* there is variation in the independent variable. Without that variation, how can the non-varying independent variable explain the varying dependent variable?

regression line

change

That the independent variable varies is the only requirement for being able to fit the data with a line of best fit. That the independent variables are independent of each other is the requirement for being able to partition the effect of the independent variables on the dependent variable. To see why this is a logical requirement, consider the following.

partition

Example 13.1: A former student of mine once told me that she should not study for my test because she always does worse when she studies. Trying to get to the bottom of this counter-intuitive result, I asked her several questions. Apparently, she only studied for tests when the course was difficult.

selection bias

Notice that the dependent variable here is "performance on the test," and the independent variables are "effort studying" and "difficulty of the course." By her own admission, the two independent variables are not independent of each other—she would only study if the course was difficult. Thus, the question remains: What explains her test performance? Is it the difficulty of the course or her studying? From the data she supplied, there is no way to tell because the two variables are highly correlated.

dependent

❧ ❧ ❧

Those are the only two requirements for linear regression. To draw valid conclusions using ordinary least squares estimation, however, we need to make one assumption:

statistics

$$\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2) \tag{13.1}$$

Here, ε_i are the residuals (errors), the vertical distance between the predicted (expected) value and the observed value. Also, as usual, $\overset{\text{iid}}{\sim}$ indicates that the variable is independent and identically distributed.

i.i.d.

While this is a single assumption, there are several aspects to it. First, the residuals need to be Normally distributed. Second, they need to be identically distributed (constant expected value and constant variance). Third, they need to be independent.

4 assumptions

Finally, as with all inference, the sample must be from a single population. If this is not so, then to what population are you generalizing?

generalization

The next sections cover how to test these assumptions and requirements.

13.1: Multicollinearity

Recall that in order to partition the effect of the independent variables on the dependent variable, the independent variables needed to be independent of each other. If they are not, then they are termed **multicollinear**. The effect of multicollinearity is that the standard errors are greater than they should be—they are inflated. This means that the test statistics are smaller than they should be. This means that the p-values are greater than they should be. Thus, multicollinearity makes more difficult to detect dependence between the dependent variable and the independent variables because the standard errors are inflated. Multicollinearity reduces the power of the test.

dependent

power loss

Testing for multicollinearity is done using the **variance inflation factor** (VIF).[1] Note that a variable is multicollinear with others if it can be predicted using those other variables. From last chapter, we can test this using regression—regressing the several independent variables on the one. If the R^2 value is high, then there is evidence of multicollinearity. The value of the VIF for that independent variable is

tolerance

$$\text{VIF} = \frac{1}{1 - R^2} \tag{13.2}$$

Note that if R^2 is close to 1 (high level of multicollinearity), the VIF is close to infinity. Conversely, if R^2 is close to 0 (low level of multicollinearity), the VIF is close to 1.

[1] Some researcher prefer to calculate the tolerance, the reciprocal of the VIF, $\text{TOL} = \text{VIF}^{-1}$.

Example 13.2: Let us return to the `crime` datafile and model the violent crime rate in 2000 using the violent crime rate in 1990 and the property crime rates in 1990 and 2000 using an additive model. There are three independent variables. Let us determine the variance inflation factor associated with the violent crime rate in 1990.

Solution: To perform the regression, we can use these lines:

```
crime=read.csv("http://rfs.kvasaheim.com/data/crime.csv")
mod=lm(vcrime00~vcrime90+pcrime90+pcrime00, data=crime)
summary(mod)
```

Note that I did not attach the data here. I used the `data` parameter in the `lm` function to specify the data source.

To calculate the variance inflation factor for the violent crime rate in 1990 in this model, we can perform this regression

independent variable

```
summary( lm(vcrime90~pcrime90+pcrime00) )
```

and note that the R^2 value is 0.525. With this, and Equation 13.2, we calculate the variance inflation factor for this variable in this model

$$\begin{aligned} \text{VIF} &= \frac{1}{1-R^2} \\ &= \frac{1}{1-0.525} \end{aligned}$$

Thus, the VIF for `vcrime90` is 2.11.

For `pcrime90`,

```
summary( lm(pcrime90~vcrime90+pcrime00) )
```

and the R^2 is 0.751. Using the formula,

$$\begin{aligned} \text{VIF} &= \frac{1}{1-R^2} \\ &= \frac{1}{1-0.751} \end{aligned}$$

Thus, the VIF for `pcrime90` is 4.02.

I leave it as an exercise for you to calculate the VIF for the third independent variable, `pcrime00`. ◇

VIF=5 $\Leftrightarrow$ R^2=0.8

How high of a VIF is too high? There are several rules of thumb for the cutoff: five and eight and 10 are all popular. Let us use five as the cutoff. If the VIF for any of the independent variables is greater than 5, then there is evidence of severe multicollinearity and one (or more) of the independent variables should be dropped from consideration.

What does it take for the VIF to be 5?

$$\begin{aligned} \text{VIF} &= \frac{1}{1-R^2} \\ R^2 &= 1 - \frac{1}{\text{VIF}} \\ &= 1 - \frac{1}{5} \\ &= 0.80 \end{aligned}$$

That is, if the other independent variables explain 80% of the variation in the independent variable, then VIF=5. In other words, if the other independent variables explain 80% of the variation in *this* independent variable, we are saying it adds so little information to the model that we will consider remove it from consideration.

adjustment

inflation

If your theory requires all of your independent variables, then you can adjust the standard errors and recalculate the p-values. This is sub-optimal as the adjustment is approximate. The VIF is the factor by which the variance is inflated, thus the square root of the VIF is the factor by which the standard error is inflated. Thus, the adjusted standard error is the calculated standard error multiplied by the square root of the VIF. Divide the parameter estimate by this adjusted standard error to get the adjusted test statistic. Comparing the test statistic to the appropriate t-distribution provides the adjusted p-value.

While you can calculate the variance inflation factor for each independent variable as in Example 13.2, it is easier to let `R` do the work. The `vif` function requires loading an additional package. Both `car` and `HH` packages provide identical `vif` functions. From the previous example, typing `vif(mod)` gives

```
vcrime90 pcrime90 pcrime00
2.105168 4.022934 2.800058
```

Thus, the variance inflation factor for the violent crime rate in 1990 is 2.1 (as above); for the property crime rate in 1990, 4.0; and for the property crime rate in 2000, 2.8. Note that none of these is greater than 5, thus there is no evidence of significant multicollinearity, and model `mod` passes *this* requirement.

VIF<5 ⇒ pass

Example 13.3: In the previous example, we estimated the violent crime rate in 2000 using the violent crime rate in 1990 and the property crime rates in 1990 and 2000. For this example, let us predict the violent crime rate in 2000 using the violent crime rate in 1990 and the gross state product (GSP) in 1990 and the population in 1990. Is there an issue with multicollinearity in this model?

Solution: Reading in the data and fitting the model is as usual. To use the `vif` function, we need to load either the `car` or the `HH` package. For no reason, I choose the second. With this, the script is

```
library(HH)
crime=read.csv("http://rfs.kvasaheim.com/data/crime.csv")
mod=lm(vcrime00~vcrime90+gsp90+pop90, data=crime)
vif(mod)
```

From this, the variance inflation factors are

VIF

```
vcrime90     gsp90     pop90
1.067801 43.746135 43.945327
```

As two of the VIF values are greater than five, we should conclude that either the GSP in 1990 or the population in 1990 (or both) should be removed from the analysis.

If, however, theory tells us that we need to keep both variables, we should adjust the standard errors (and the p-values), which we do here.

theory

13.1.1 Adjusting for Multicollinearity

Adjusting for the effect of multicollinearity is the lesser of two evils—the greater is ignoring it. Recall that multicollinearity inflates the standard error estimates; while the parameter estimates are still unbiased, the standard error estimates are not. The correction is to scale the standard error estimates by some factor—the square root of the variance inflation factor.

bias

Thus, let us first take note of the unadjusted regression table provided by R, using `summary(mod)`:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.095e+02  2.329e+01   4.703 2.28e-05 ***
vcrime90     5.873e-01  3.207e-02  18.314  < 2e-16 ***
gsp90       -5.987e-04  5.708e-04  -1.049    0.300
pop90        1.296e-05  1.473e-05   0.880    0.383
```

According to the output from `vif(mod)`, the variance inflation factor for the GSP in 1990 is 43.746135. Thus, the standard error inflation factor is $\sqrt{43.746135} = 6.614$. For the 1990 population, the VIF is 43.945327 and the standard error inflation factor is 6.629. Thus, the adjusted standard errors will be $5.708 \times 10^{-4}/6.614 = 8.630 \times 10^{-5}$ and 2.222×10^{-6}, respectively. Dividing the estimates by the standard errors give the adjusted test statistic values of $(-5.987 \times 10^{-4})/(8.630 \times 10^{-5}) = -6.937$ and 5.833, respectively. The corresponding adjusted p-values are 1.017193×10^{-8} and 4.820074×10^{-7}.

With the above information, the adjusted regression table is

```
               Estimate  Std. Error  t value    Pr(>|t|)
(Intercept)  1.09522e+02 2.32863e+01  4.70331 2.27639e-05
vcrime90     5.87313e-01 3.10345e-02 18.92450 0.00000e+00
gsp90       -5.98722e-04 8.63041e-05 -6.93735 1.01728e-08
pop90        1.29574e-05 2.22183e-06  5.83188 4.83135e-07
```

Thus, while the original (improper) analysis indicated that the effect of neither the GSP in 1990 nor the population in 1990 was statistically significant, the corrected analysis indicates that both are highly so.

There is no R package with a command providing the VIF-adjusted regression table (yet). There is, however, a function on the book's website called `summaryVIFA`. As usual, you need to source the file to use it. Thus, the following will provide the VIF-adjusted regression table:

source

```
source("http://rfs.kvasaheim.com/Rfctns/summaryVIFA.R")
summaryVIFA(mod)
```

While the table produced by this function is utilitarian, it does provide the needed information. Variance inflation only affects the standard error estimates (and the test statistic value and the p-value). Thus, you should also use the standard regression output, `summary(mod)`, for additional model information. ◇

13.2: Normality

This, and the next couple sections, deal with assumptions about the residuals. Technically, the assumptions are about the conditional distribution of the dependent variable given the independent variables. This point will become important when discussing generalized linear models.

The assumption on the residuals, ε_i, is

$$\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2) \tag{13.3}$$

with $\overset{\text{iid}}{\sim}$ indicating that the variable is *independent* and *identically* distributed. In this section, we tackle testing the Normality assumption. **Normality**

❧ ❧ ❧

Testing the Normality of the residuals requires calculating the residuals, then using one of the available Normality tests. Residuals are defined as the observed value less the predicted value. In R, there are two methods for calculating the residuals. The first method follows the definition of residual and the regression equation. **errors**

Example 13.4: Let us continue Example 13.3 in which we modeled the violent crime rate in 2000 using the violent crime rate in 1990, the GSP in 1990, and the population in 1990. Calculate the residuals for this model.

First, let us use the definition of residual.

```
res=crime$vcrime00-predict(mod)
```

Second, let us use the R function.

```
res=residuals(mod)
```

These two methods are equivalent.

Now that we have the residuals, testing for Normality is straight forward. Graphically, one can use a Normal quantile-quantile plot or a histogram. If the plotted points fall near the diagonal line in the quantile-quantile plot, then there is sufficient evidence that the residuals are Normally distributed. Likewise if the histogram is bell-shaped. **Q-Q plot**

Figure 13.1 shows the two graphics for the residuals. Note that both graphics indicate non-Normality and leptokurtosis in the residuals (excess

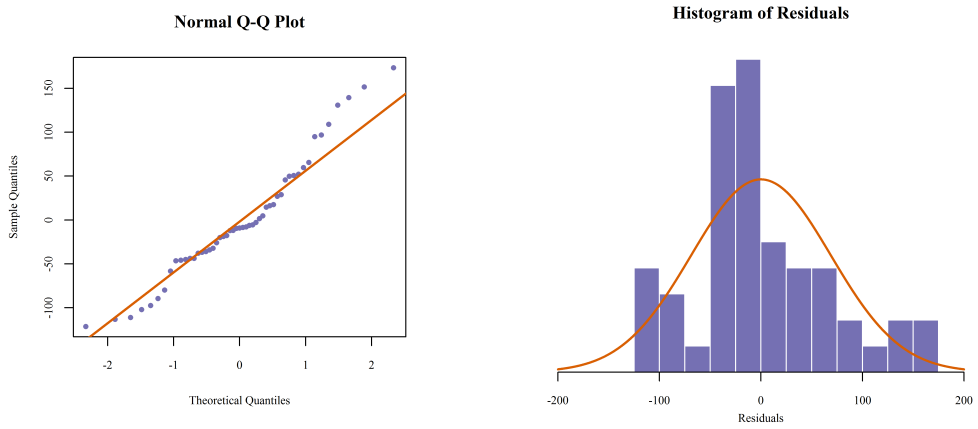

Figure 13.1: *Normal quantile-quantile plot (left) and histogram (right) of the residuals from Example 13.4. Note that both graphics indicate non-Normality.*

kurtosis = 1.28619). Thus, we should reject the assumption that the residuals are Normally distributed.

Shapiro-Wilk test

In addition to the graphical methods, we can use numeric methods to test the assumption (null hypothesis) of Normality. As in earlier chapters, you can use the Shapiro-Wilk test or any of a slew of Normality tests. The `nortest` package offers many options, as does the `fBasics` package. The plethora of Normality tests allows you some options. First, you can select one test and always use that test. Second, you can run multiple tests and let majority decide. Finally, you can become an expert in the several tests and use the one that best fits the type of data.

When discussing t-tests (Chapters 5 and 6) and analysis of variance (Chapter 7) we only used the Shapiro-Wilk test to test the assumption of Normality. Here, let us show many of the Normality tests available in R.

Running the following code

```
shapiro.test(res)          # Base package

ad.test(res)               # Package: nortest
cvm.test(res)
lillie.test(res)
sf.test(res)

dagoTest(res)              # Package: fBasics
ksnormTest(res)
```

Test Name	R Function	Package	Statistic	P-Value
Anderson-Darling	`adTest` `ad.test`	`fBasics` `nortest`	1.9126	5.977×10^{-5}
Cramér-von Mises	`cvmTest` `cvm.test`	`fBasics` `nortest`	0.3235	1.656×10^{-4}
D'Agostino	`dagoTest`	`fBasics`	6.5923	0.03702
Jarque-Bera	`jarqueberaTest`	`fBasics`	6.6493	0.03599
Kolmogorov-Smirnov	`ksnormTest`	`fBasics`	0.6274	6.661×10^{-16}
Lillifors	`lillieTest` `lillie.test`	`fBasics` `nortest`	0.146	0.008391
Shapiro-Francia	`sfTest` `sf.test`	`fBasics` `nortest`	0.8937	5.389×10^{-4}
Shapiro-Wilk	`shapiro.test` `shapiroTest`	— `fBasics`	0.9022	4.994×10^{-4}

Table 13.1: *The results from several tests of Normality on the residuals. The test name, R function, and needed package are listed for reference.*

```
jarqueberaTest(res)
adTest(res)
cvmTest(res)
lillieTest(res)
sfTest(res)
shapiroTest(res)
```

produces the results given in Table 13.1, assuming the needed libraries are installed. Note that all tests indicate significant deviation for Normality. Thus, we can safely conclude that the residuals are *not* Normally distributed. $p < \alpha \Rightarrow$ **fail**

With this said, do not forget the Central Limit Theorem (Appendix Section C.3). If the distribution is bell-shaped, then the sample size needed to forgo the Normality assumption is usually fifty or less.[2] Here, the histogram indicates the residuals have a vaguely bell-shaped distribution. The sample size is 51. Is this large enough to ignore the non-Normality? **CLT**

It depends on the cost of you being wrong. **costs**

[2] This really depends on the kurtosis. When the excess kurtosis is close to zero, the required sample size is smaller.

13.3: Identically Distributed

The second assumption is that the residuals are identically distributed. In light of the previous section, this assumption actually reduces to two simpler assumptions about the residuals: constant expected value and constant variance.

constant

13.3.1 Constant Expected Value

Testing for constant expected value is beneficial in that violations of this assumption indicate that you are not using the correct transformation (Chapter 14). If your model violates this assumption, then your predictions will be biased. While bias is not necessarily a bad thing, it should be avoided unless you gain something from it (see Chapter 11).

bias

Because of the wide variety of possible violations (graph shapes), this test is graphical: Plot the residuals against the dependent variable. If a functional pattern appears, then you have a violation.

pattern

Example 13.5: To demonstrate a common violation of this assumption, let us create some data.

```
set.seed(370)
x = runif(1000, min=0, max=5)
e = rnorm(1000, m=0, s=1)
y = x^2 + e
```

By construction (line 4), we know that the relationship between `x` and `y` is quadratic; that is, we know $y = x^2$. However, for the sake of illustration, let us fit it with both a linear and a quadratic curve and compare the two residual plots.

Figure 13.2 provides plots of the residuals against the dependent variable for both the linear model and the quadratic model. Note the pattern in the residual plot of the linear model. This pattern suggests the independent variable should be squared.

Simple linear regression of `y` against `x` gives the following table

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.29223    0.13669   -31.4   <2e-16 ***
x            5.05380    0.04662   108.4   <2e-16 ***
```

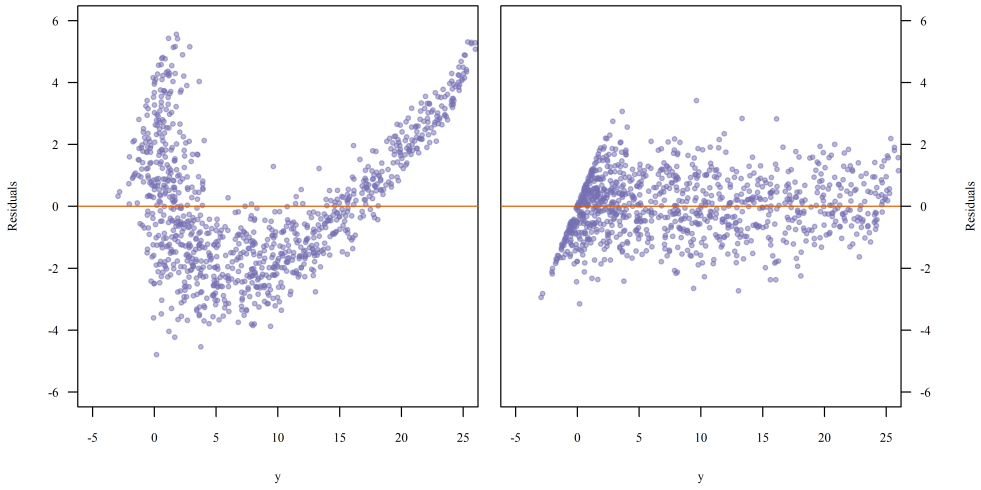

Figure 13.2: *Two residual plots. The left plot corresponds to the linear model of Example 13.5; the right, the quadratic model.*

Notice that the variable `x` is highly significant in terms of explaining the variable `y`. Without checking the assumptions, one would be tempted to stop here and report the results. Note, however, that the residual plot indicates non-constant expected values. If you trace out the 'middle' of the residual values for each value of the dependent variable, you will trace out a U-shaped curve—the expected value (middle) of the residuals is not constant (with respect to the dependent variable). In fact, the shape of the residual plot suggests that the independent variable should be squared.

p-value

more information

The second model, `lm(y~I(x^2))`, produces the following regression table

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.031769   0.047452   -0.67    0.503
I(x^2)       0.999637   0.004131  241.96   <2e-16 ***
```

Again, the independent variable is highly significant. However, this quadratic model is preferred as the regression plots suggests a constant expected value. If you trace out the 'middle' of the residual values for each value of the dependent variable, you will trace out a horizontal line—the expected value

constant

(middle) of the residuals is constant (with respect to the dependent variable).[3]

By construction, we know that the prediction equation is $y = 0 + x^2 + \varepsilon$ (line 4 of the data-generating process in Example 13.5). It is interesting to see how close the estimates are to reality. Using `confint(mod2)`, we see that the true values of parameters β_0 and β_1 (0 and 1, respectively) are within the 95% confidence intervals estimated from the data:

confidence interval

```
                2.5 %     97.5 %
(Intercept) -0.1248869 0.0613483
I(x^2)       0.9915303 1.0077446
```

Warning: *If there are multiple independent variables, it is usually difficult to determine which independent variable needs to be squared—there may be more than one. This is why statisticians need to be patient and test all models. Remember, you are trying to move towards the truth.*

Warning: *With the above said, you need to be careful here. If you are doing exploratory analysis, you need to test your models on different data. If you are doing confirmatory analysis, you need to* not *do model selection.*

This helps avoid false discoveries. If you look hard enough you can find a lot of interesting relationships in the data. The question is which are real relationships in the population? Separating your exploratory and confirmatory analyses and using different data for each will make your discoveries more genuine.

13.3.2 Constant Variance

The assumption that gave rise to the need for a constant expected value likewise requires a constant variance. Constant variance is called **homoskedasticity**; non-constant variance, **heteroskedasticity**.[4] The problem with heteroskedasticity is that the standard errors are a function of the dependent variable, which means the p-values are, too. As such, one cannot make a blanket statement about the statistical significance

bias

[3] Also note that the adjusted R^2 values indicate the second model is preferred (0.9216 versus 0.9832).

[4] Heteroskedasticity is from the Greek prefix *hetero-* (ἑτερο-), meaning "different," and the Greek root *skedannumi* (σκεδάννυμι), meaning "to scatter." Frequently, the word is spelled "heteroscedasticity." However, as McCulloch (1985) pointed out, this is not correct.

of a variable. While the parameter estimates are not biased, the standard error estimates are.

As with most assumptions, graphical and numeric methods exist to test homoskedasticity. One numeric test is the Breusch-Pagan test (Breusch and Pagan, 1979; Cook and Weisberg, 1983). A graphical test for heteroskedasticity is a residuals plot—the residuals plotted against the dependent variable.

Breusch-Pagan test

Given a residuals plot, one would expect the vertical 'spread' of the residuals to be approximately constant across the values of the dependent variable in the presence of homoskedasticity. Look at the 'thickness' of the residual values at each dependent variable value. The thickness should remain approximately constant.

concentration

To illustrate this, let us return to Figure 13.2. The left panel shows heteroskedasticity—the variance (thickness) for smaller values appears larger than the variance for larger values. The right panel, model `mod2`, does not show this. The concentration of the residuals seems approximately constant across all values of the dependent variable. The exception occurs at the far-left, which is at the edge of the values.

edge effects

`R` implements the Breusch-Pagan test through the `car` package function `ncvTest` and the `lmtest` package function `bptest`. The two functions are slightly different in their output, but are identical in their input—both take the model as the input. The `ncvTest` function is a score test, whereas the `bptest` function is not. This means `bptest` offers a slight improvement in power over `ncvTest` (Koenker 1981). With that said, the two tests rarely give different *substantive* conclusions.

Performing the two tests, `ncvTest(mod2)` and `bptest(mod2)`, produces the following output

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 1.442045    Df = 1     p = 0.2298087
```

and

```
        studentized Breusch-Pagan test

data:  mod2
BP = 1.4748, df = 1, p-value = 0.2246
```

Note that in both cases, the conclusion is that there is no significant evidence of heteroskedasticity.

sandwich estimator

ADJUSTING FOR HETEROSKEDASTICITY: It may happen that you cannot rid your model of heteroskedasticity. In such a case, you will want to adjust your standard error estimates. The usual method to produce heteroskedasticity-consistent standard errors is the Huber-White method (White, 1980).

White's 1980 paper tackled the issue of estimating the standard errors. While methods existed, they were computationally expensive and had poor properties. White needed a method that produced good estimates *and* that could be done with the available computing power. In his paper, White (1980: 817) stated

> It is well known that the presence of heteroskedasticity in the disturbances of an otherwise properly specified linear model leads to consistent but inefficient parameter estimates and inconsistent covariance matrix estimates. As a result, faulty inferences will be drawn when testing statistical hypotheses in the presence of heteroskedasticity.

By the end of the paper, White not only calculated the corrected variance estimates, he created a new test for heteroskedasticity. White's adjusted standard errors, s_W had a different multiplier than that presented in Equation 13.4. Work subsequent to White's determined that the $\frac{n}{n-k}$ produced better estimates than his original, which was 1.

$$s_W^2 = \left(\frac{n}{n-k}\right)(\mathbf{X}'\mathbf{X})^{-1}\left(\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'e_i^2\right)(\mathbf{x}'\mathbf{x})^{-1} \tag{13.4}$$

Here, n is the sample size, k is the number of independent variables, $\mathbf{X}$ is the observed design matrix, and e_i is the i^{th} residual.[5]

This function is available on the book's website as `summaryWASE`. As usual, you need to source the file to use it. Thus, the following will provide the regression table with heteroskedasticity-consistent standard errors for the second model:

```
source("http://rfs.kvasaheim.com/Rfctns/summaryWASE.R")
summaryWASE(mod2)
```

[5] Recall from the previous chapter that the design matrix is the data matrix prepended with a column of 1s.

The heteroskedastic-adjusted regression table is

```
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) -0.0317693 0.048156549  -0.6597089 0.5094407
I(x^2)       0.9996374 0.004044144 247.1814665 0.0000000
```

That there is little difference between this table and the unadjusted regression table (page 335) is due to the lack of heteroskedasticity in the model.

> ***Note***: The parameter estimates remain unbiased in the presence of heteroskedasticity. The standard error estimates do not. This means the statistical significance of your variables is in doubt. If you cannot fix heteroskedasticity, you should at least adjust for it.
>
> Measurement error correlated with an independent variable is one cause of heteroskedasticity. Missing independent variables is another. Thus, the presence of heteroskedasticity tells you that your model is lacking and there is more information in the data than you are accounting for. Thus, heteroskedasticity is a hint that there is more than meets the eye. **more information**

13.4: Independence

The first 'i' of $\overset{\text{iid}}{\sim}$ stands for **independence**, namely that each residual is independent of all others. Whereas violations of the previous two assumptions still produced unbiased estimates of the parameter estimates, violations of independence assumption do *not*. Whereas violations of the constant expected value assumption produced unbiased estimates of the standard errors, violations of independence do not. In the presence of dependence, there is little you can infer. At the very least, you need to know the nature of that dependence before being able to make inferences. **garbage**

Dependence can arise in many ways, but most are due either to the data collection technique or to the structure of the data itself.

Cross-sectional data is collected in such a manner that neither time nor position matters. It is data collected across the sampled population at a single moment in time. Usually, survey data is collected in this manner. If cross-sectional data is collected using random sampling, then it is independent. **survey**

13.4.1 Temporal Data

time series

When data is collected in such a manner that time is an issue, the data is termed **temporal** (or time-series) **data**. Repeated measurements in time is typical time-series data. Modeling stock prices over time, modeling height over time, modeling performance over time are all examples of time-series data.

By its very nature, time-series data is dependent data. The value at one moment in time is (most-likely) similar to that at a close time. This type of data is so important, and this type of assumption violation is so prevalent, that time-series analysis composes an entire subfield within statistics. To fully understand the data, the process, and the violation, you will need to take one or more courses in time-series analysis.

While the data may be time-series data, the assumption of independence may not be *strictly* violated. It is possible. In such cases, the time-series data behaves like cross-sectional data. One possible test to determine if the residuals are "independent enough" is the Durbin-Watson test (Durbin and Watson 1950, 1951). This test is performed by the function `dwtest` in the `lmtest` package and by the function `durbinWatsonTest` in the `car` package. In both cases, the null hypothesis is independence. Thus, if the p-value is less than α, there is significant evidence of serial *dependence* in the data. The only difference between the two functions is the default behavior in terms of the alternative hypothesis; `dwtest` defaults to testing if the correlation is positive, `durbinWatsonTest` defaults to testing if the correlation is not zero.

$p > \alpha \Rightarrow$ pass

A poor fix exists for temporal data. One can use the value of the dependent variable in the previous time unit (one lag) as an independent variable in the model. Such a model is called a **lagged dependent variable model**. This is better than ignoring time-dependence of the data. The drawback is that parameter estimates are biased (although less so than just ignoring the issue). Note that the residuals of a lagged dependent variable model may still be correlated. If so, you may wish to try multiple lags and functions of lags.

LDV model

bias

A *better* fix is to model the time aspect. However, this requires time-series analysis and is beyond the scope of this book. Introductory time-series books include Wei (2005) and Shumway and Stoffer (2013).

model

13.4.2 Spatial Data

When data is collected such that location is a factor, it is termed **spatial** (or geographical) **data**. People within neighborhoods tend to be more similar than people in two neighborhoods. Neighboring US states are more similar than states on opposite coasts. Countries in a continent tend to be more similar than countries on different continents. Each of these cases suggests geographical correlation. As with time-series data, geographical dependence constitutes an entire subfield within statistics (and within geography). Also as with time-series data, one should take a course dedicated to the topic.

geography

Again, while the data may be geographical data, the assumption of independence may not be *strictly* violated. In such cases, geographical data behaves like cross-sectional data. One test of many to determine if the residuals are "independent enough" is Moran's I test. In `R`, this test is implemented in the function `Moran.I` in the `ape` package and in the function `moran.test` in the `spdep` package. In both cases, the the function requires the residuals as well as a neighbor matrix—an $n \times n$ matrix specifying which units are neighbors and how neighborly they are. This is the most important part about doing spatial analysis, as the computer program needs to know the structure of space.

contiguity matrix

Again, the null hypothesis of Moran's I test is independence. Thus, if the p-value is less than α, there is significant evidence of spatial dependence.

As with temporal data, there is a poor fix for spatially-correlated data. Since the correlation originates from the neighboring units, including the neighbor-averaged dependent variable as an independent variable may control the spatial correlation. This is another type of **lagged dependent variable model**. Again, the parameter estimates are biased, but less so than in the original model.

LDV model

Again, this fix is suboptimal as you are ignoring relevant data. It is much better to model the spatial relationships. This, too, is beyond the scope of this text. Introductory time-series books include Bivand et al. (2008), Fotheringham et al. (2002), and Forsberg (forthcoming).

13.4.3 Dependence Combinations

Combinations and extensions of these types of data exist and create their own correlation issues. Cross-sectional time-series data is repeated measurements taken over time on the same units. This may also be called panel data. Temporal-spatial data is geographic data taken over time.

panel data

In every case, the key to modeling the data correctly is to determine the correlation between (and among) the records. While this is also beyond the scope of this book, you may wish to consult Hsaio (2003) and Wooldridge (2001) for treatment of cross-sectional time-series data; Cressie and Wikle (2011) for temporal-spatial data; Hardin and Hilbe (2012) for general correlation among the experimental units (records).

13.5: Single Population

target population

The assumption undergirding all inference is that your sample represents the target population. In Chapter 3, we learned about the target populations, sampled populations, and generalization. You wish to draw conclusions about the target population. Your sample is drawn from the sampled population. Unless the latter is similar to the former, you cannot generalize the conclusions of your analysis to the former.

random sampling

Random sampling, which ensures independence, also ensures that your sample is representative of your sampled population on average. However, the final check you must make is that your sample also comes from your target population.

error

Frequently, miskeying your data into the computer is the cause of this violation. A subtler way to violate this assumption is to have data come from outside the population of interest. For instance, when studying the US states, researchers usually include Washington, DC. However, is DC really a part of the target population?

theory

The answer is between you and your theory. Statistics does not—*should* not—enter in to this discussion. The only place for statistics here is to highlight questionable units and to determine if they really matter. The first is a question about outlying points; the second, influential points.

13.5.1 Outlying Points

distance

An outlier test should be used to determine if any data values are questionable. It can also be used to highlight units that do not "follow the model." For the former case, you will want to create box-and-whiskers plots of all variables and examine those outlying units.

model fit

For the latter case, in addition to box-and-whiskers plots of the data, you will want to check the residuals for outliers. Outliers in the residuals are those units that do not fit the model well. While a box-and-whiskers plot of

the residuals will work here, the Bonferroni outlier test can speed things up a bit.

In R, the function outlierTest from the car package performs this test. The output is a list of the unit farthest from the line of best fit in addition to all units whose distance is beyond the t-distribution's critical value for the specified α (where $\nu = n - 1$). Those units with outlying residuals should be examined to determine *why* the model fit that unit so poorly. **why**

Performing this test on model mod2 produces this output

```
No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
    rstudent unadjusted p-value Bonferonni p
241 3.453026         0.00057763      0.57763
```

From this, we can conclude that there is no unit whose residual is of note. All units follow the model well. Unit 241 is the worst, but the residual of that unit is still well within the t-distribution critical bounds. $p > \alpha \Rightarrow$ **pass**

Performing this test on model mod1 produces this output

```
No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
    rstudent unadjusted p-value Bonferonni p
482 2.844429          0.0045401           NA
```

Recall that the Bonferroni adjustment multiplies the calculated p-value by the number of tests performed, which is the sample size n. When that product is greater than 1, this test returns the value NA. Thus, no unit is extremely far from the prediction for model mod1. $p > 1$

While the outlier test makes things easy, nothing replaces exploring and knowing your data. **know your data!**

13.5.2 Influential Points

An **influential data point** is one that markedly alters the estimated model. Outliers may *or may not* be influential. If the outlier is located towards the middle of the range of the independent variable, it will most likely *not* be an influential point. If the outlier is located towards the extremes of the range of the independent variable, it *will* most likely be an influential point. **influence**

The test for influential points is based on its definition. The model is fit both with and without the unit. The amount of change in the predictions is a measure of influence. Two functions are of interest here. First,

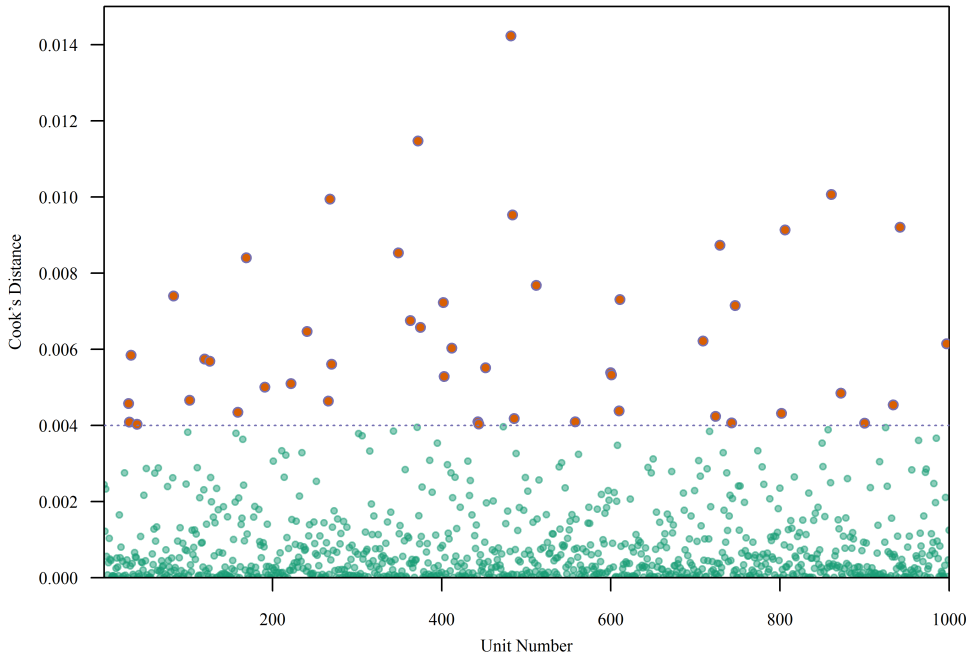

Figure 13.3: *An index plot of Cook's distance for each unit. Using the* $4/n$ *cutoff , 48 units are marked as influential.*

the `influence.measures` function provides several measures of influence for each data point. The final column provides asterisks to indicate the units that the function determines are influential.

Running `influence.measures(mod2)` provides an extensive list of influence measures for all 1000 units. Of the 1000, this function identifies 70 which are influential. As you can see, this may not be helpful for large datasets. That is, unless you wish to assert that these 70 units are from a separate population.

In such cases, you may wish to produce index plots the various influence statistics. I suggest plotting Cook's distance (1977) for each unit (e.g., Figure 13.3). Such a plot will help you see if there are any unusual units. Determining which values are influential using Cook's distance is as easy as selecting the correct cutoff value (and as difficult). Cook and Weisberg (1982) suggest a cutoff of 1. Bolen and Jackman (1990) suggest a cutoff of $4/n$. I suggest just looking at the plot to see if there are a few units whose Cook's distance is a lot different from that of the others.

The following plots the Cook's distances for model `mod2` and adds the $4/n$ cutoff.

```
plot(cooks.distance(mod2))
n <- length(y)
abline(h=4/n)
```

The resulting graphic is provided (gussied up a bit) as Figure 13.3. While there are 48 influential units according to the graphic, I would not find the plot alarming. From experience, I am only concerned by influential points when there are just a few of them that do not appear to be a part of the general variation.

Once you have your list of influential points, you will want to double check that the values entered for those points are correct. You may also wish to determine why those few points are influential. Is there something interesting about those units? If analyzing the US states, the influential states may be the source of your most important exploration.

How interesting!

13.6: A Full Example

To illustrate the concepts, tests, and procedures in this chapter, let us perform a full analysis. The datafile we will use is the `animal` datafile. The goal will be to predict a person's annual income using the data.

There are 223 records and 10 variables. This is a lot of data. To quickly focus on variables of interest, one could plot all pairwise combinations using `plot(animal)`. Alternatively, one could calculate all pairwise correlations between the 10 variables.

The pairwise plots suggest that the best predictor of the respondent's income (`rinc`) is the parent's income (`pinc`). None of the other variables seem to be markedly correlated with the respondent's income. The pairwise correlations concur. The only variable whose correlation with the respondent's income is above 0.50 is parent's income ($\hat{\rho} = 0.9808$).

From this, it appears as though our model should be a simple linear model. Performing the regression of respondent's income on parent's income, we see that the respondent, on average, has an income 94% that of the parents ($p \ll 0.0001$).

effect size

To make sure that this model is appropriate, let us test the assumptions. There is no need to test for multicollinearity as there is only one in-

multicollinearity

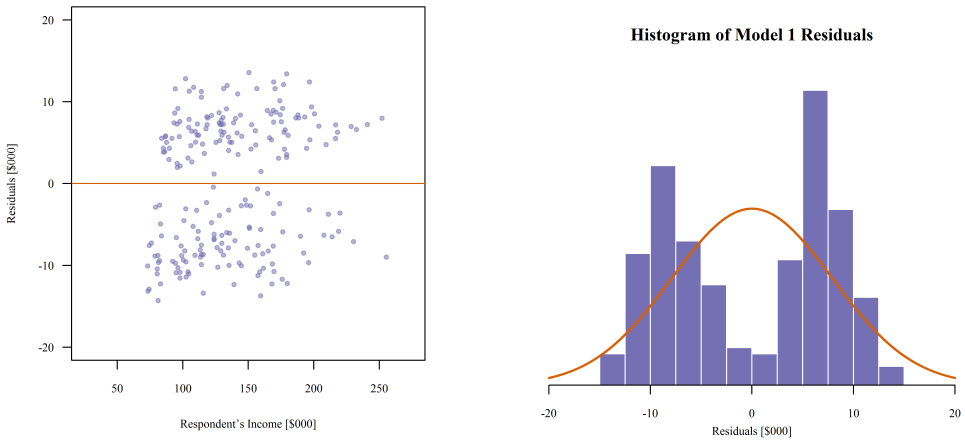

Figure 13.4: *A plot of the residuals against the dependent variable (left) and a histogram of the residuals (right) for model* `mod1`.

dependent variable. The Shapiro-Wilk test indicates severe departure from Normality ($p \ll 0.0001$). Figure 13.4 (left) is a plot of the residuals against the observed respondent's incomes (the dependent variable). The expected value of the residual appears constant. The plot also suggests homoskedasticity, as the "thickness" of the residuals appears to be constant with respect to the dependent variable; the Durbin-Watson score test concurs ($p = 0.7910$). There are no records with outlying residuals. Finally, since the data is cross-sectional data (and presumably a random sample), there is no dependence to worry about. In fact, the only problem is the lack of Normality.

Normality and the CLT

Do we *really* need to worry about the lack of Normality? Probably not. As the sample size is $n = 223$, the Central Limit Theorem suggests that the inferences we make will still be close.

more information!

insight

However, we are missing a chance to better understand what influences personal income. The question is: What is causing the lack of Normality? The answer will give us more insight, and we would be foolish to skip this opportunity to learn.

two populations

Figure 13.4 (right) is a histogram of the residuals with a Normal curve superimposed. This graphic shows the problem. There appears to be two different groups represented in this data. What could those two groups be?

a new variable

The only dichotomous variable in the dataset is the `male` variable. From an economic standpoint, including this variable makes sense as it is quite likely males make more than females, all things held equal. Thus, we

continue our exploration by including the respondent's gender in our analysis.

As stated last chapter, we first need to test the interaction model. If the interaction term is not statistically significant, we remove it from consideration and use the additive model. **interaction model**

The analysis of variance summary for the interaction model is[6]

```
             Df    Sum Sq   Mean Sq   F value Pr(>F)
pinc          1 3.514e+11 3.514e+11 40873.775 <2e-16 ***
male          1 1.199e+10 1.199e+10  1394.927 <2e-16 ***
pinc:male     1 1.564e+06 1.564e+06     0.182   0.67
Residuals   219 1.883e+09 8.596e+06
```

The p-value of the interaction term ($p = 0.67$) is greater than our usual $\alpha = 0.05$. Thus, this term is not needed in the final model. Removing it gives us the additive model.

The regression table for the additive model is **additive model**

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.106e+04  4.862e+02  105.02   <2e-16 ***
pinc        9.511e-01  4.659e-03  204.13   <2e-16 ***
male        1.470e+04  3.929e+02   37.42   <2e-16 ***
```

Both independent variables are highly significant. According to this model, on average as parental income increased by $1000, the respondent's income increased by only $951, with a 95% confidence interval of $942 to $960. This does not mean people tend to make less than their parents. The intercept indicates people start out earning $5106 more than their parents. **constant term**

Also, according to this model, men tend to earn $14,700 more than women. In fact a 95% confidence interval for this income differential is from $13,927 to $15,475.

These values came from `confint(mod2a)`:

```
                  2.5 %       97.5 %
(Intercept) 5.009970e+04 5.201598e+04
pinc        9.418888e-01 9.602535e-01
male        1.392681e+04 1.547541e+04
```

[6]When *any* of the independent variables are categorical, you will need to run `summary.aov` instead of `summary`. Here is the reason: The former determines the statistical significance of the variable; the latter, of the level.

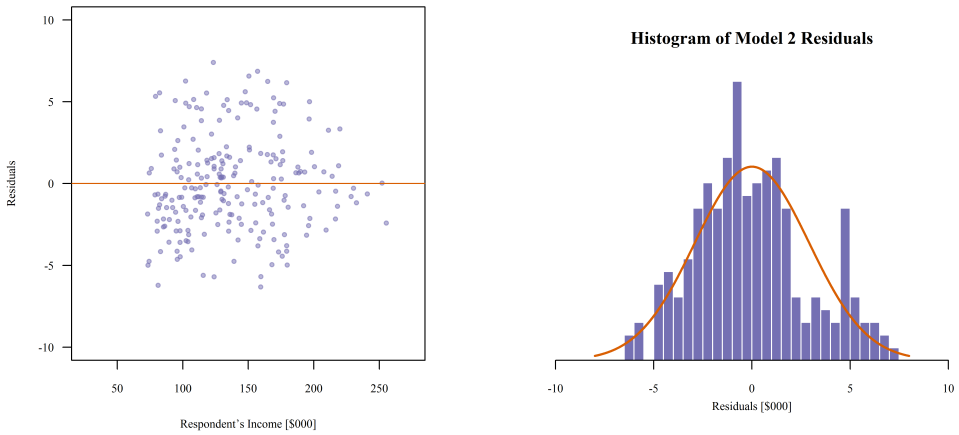

Figure 13.5: *A plot of the residuals against the dependent variable (left) and a histogram of the residuals (right) for model* `mod2a`.

All that remains with our analysis is determining if this model violates any of the assumptions of ordinary linear regression (and graphing the results).

The Shapiro-Wilk test indicates the residuals are not Normally distributed ($p = 0.0034$). A plot of the residuals against the dependent variable suggests the expected values are constant (Figure 13.5, left). The same figure suggests no heteroskedasticity. The Breusch-Pagan score test concurs ($p = 0.410$). There are no records with outlying residuals.

Again, the only violation is that of Normality. The histogram of the residuals (Figure 13.5, right) does not seem to indicate two populations. We could explore the variables in greater detail, creating model after model after model.

The problem with blindly trying multiple models is that you may discover relationships that exist only in the sample and not in the population. This is called a **false discovery**. There is extensive literature on it.

false discovery

The bottom line is that the models should have some (non-statistical) explanation for variable inclusion. In this model, we wanted to model personal income (`rinc`). There is much literature in the economics and sociological fields suggesting that the parent's income should offer information. The original pairwise correlations performed merely showed importance of parental income.

theory

Gender was added for two reasons. First, the original model was lacking. The histogram of the residuals suggested the existence of two popula-

rationale

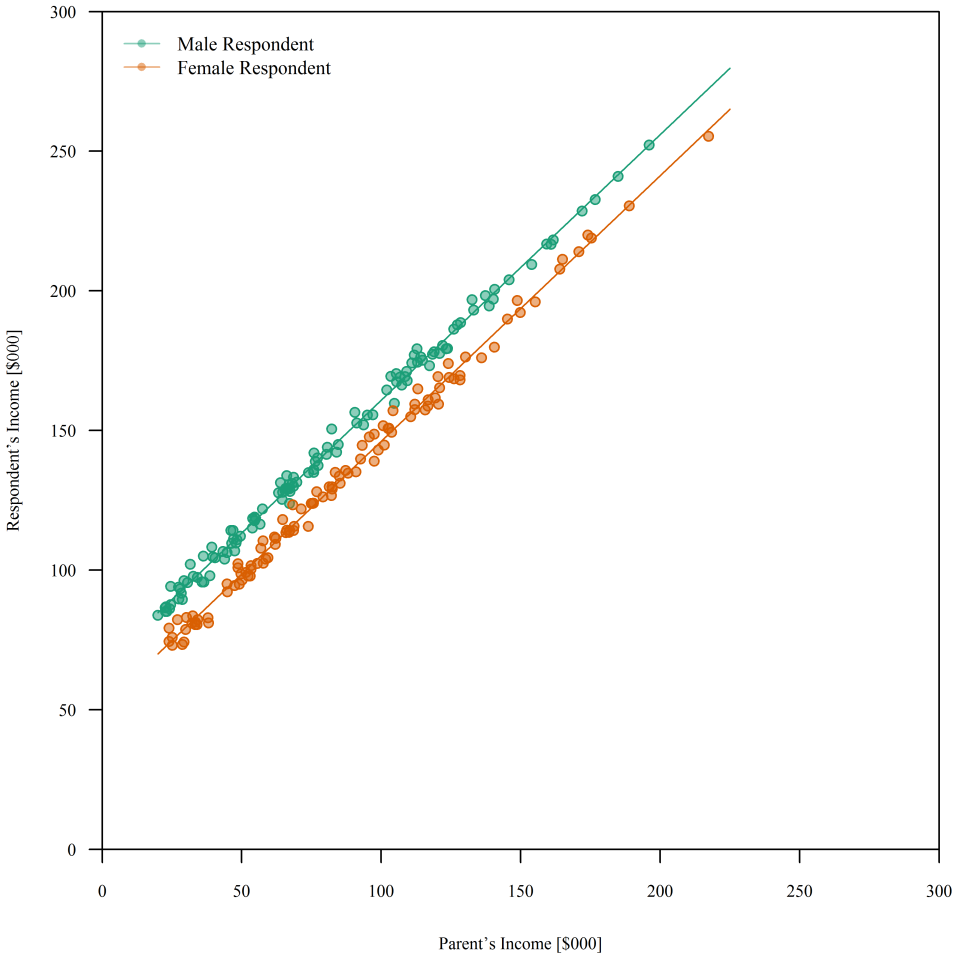

Figure 13.6: *A prediction plot for the additive model discussed in the text. Note that males earn, on average, $14,700 more than females, ceteris paribus.*

tions in the sample. Second, economics and sociological literature supported its inclusion.

None of the other variables were included, as the residual histogram of the additive model did not indicate multiple populations, and none of the other variables seem to be connected to personal income.

As the additive model violated only the Normality assumption, and as the sample size was large ($n = 223$), we invoked the Central Limit Theorem and decided this was a sufficient model. CLT

Warning: *For statisticians, the variables are just numbers. For social scientists, the variables have meaning. That meaning must be respected when modeling. If you do not have a theoretically (or logically) sound reason for adding a variable, then do not add it.*

predictions

The last thing we need to do is create the prediction plot. I leave it as an exercise to have you create something akin to Figure 13.6.

13.7: Conclusion

This chapter built upon the previous chapter. In that chapter, we introduced linear models and saw how to interpret and graph them. In this chapter, we covered the assumptions of the ordinary least squares method. There is still a requirement that the independent variables vary. There is a requirement that the independent variables are independent.

In addition, there are several assumptions that need to be checked. All are based on the residuals:

$$\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

This one statement tells us that we assume the expected values of the residuals are constant, the variances of the residuals are constant, the distribution of the residuals is Normal, the residuals are independent, and the data comes from the target population.

The assumption of constant expected value was only briefly discussed here. In the next chapter, we cover some methods to help take care of non-constant expected values.

Warning: *Remember: The purpose of analysis is two-fold: to get closer to truth and to learn about the data-generating process. In some ways, coming up with a perfect model is disappointing; there is nothing else to explore.*

13.8: End of Chapter Materials

13.8.1 R Functions

In this chapter, we were introduced to many, many R functions that will be useful in the future. These are listed here.

Packages: These are the packages used in this chapter. The information provided for each is copied from the official package description.

ape Ape provides functions for reading, writing, plotting, and manipulating phylogenetic trees, and for similar types of analyses.

car This package accompanies J. Fox and S. Weisberg, *An R Companion to Applied Regression*, Second Edition, Sage, 2011.

fBasics Environment for teaching a course on "Financial Engineering and Computational Finance."

HH Support software for *Statistical Analysis and Data Display* (Springer, ISBN 0-387-40270-5). This contemporary presentation of statistical methods features extensive use of graphical displays for exploring data and for displaying the analysis. The authors demonstrate how to analyze data—showing code, graphics, and accompanying computer listings—for all the methods they cover. They emphasize how to construct and interpret graphs, discuss principles of graphical design, and show how accompanying traditional tabular results are used to confirm the visual impressions derived directly from the graphs. Many of the graphical formats are novel and appear here for the first time in print. All chapters have exercises.

lmtest A collection of tests, data sets, and examples for diagnostic checking in linear regression models. Furthermore, some generic tools for inference in parametric models are provided.

nortest Five omnibus tests for the composite hypothesis of normality.

spdep A collection of functions to create spatial weights matrix objects from polygon contiguities, from point patterns by distance and tessellations, for summarizing these objects, and for permitting their use in spatial data analysis, including regional aggregation by minimum spanning tree; a collection of tests for spatial autocorrelation, including

global Moran's I, APLE, Geary's C, Hubert/Mantel general cross product statistic, Empirical Bayes estimates and Assuno/Reis Index, Getis/Ord G and multicoloured join count statistics, local Moran's I and Getis/Ord G, saddlepoint approximations and exact tests for global and local Moran's I; and functions for estimating spatial simultaneous autoregressive (SAR) lag and error models, impact measures for lag models, weighted and unweighted SAR and CAR spatial regression models, semi-parametric and Moran eigenvector spatial filtering, GM SAR error models, and generalized spatial two stage least squares models.

STATISTICS: These functions perform statistical tests. Make sure you install the appropriate library before using the function. The Normality tests are also covered on page 332.

ad.test(x) This performs the Anderson-Darling test for the composite hypothesis of Normality. It is in the `nortest` package.

adTest(x) This performs the Anderson-Darling test for the composite hypothesis of Normality. It is in the `fBasics` package.

bptest(model) One of the more important numerical tests in detecting heteroskedasticity is the Breusch-Pagan test. This is in the `lmtest` package.

cvm.test(x) The Cramér-Von Mises test of Normality. It is in the `nortest` package.

cvmTest(x) The Cramér-Von Mises test of Normality. It is in the `fBasics` package.

dagoTest(x) The D'Agostino test of Normality. It is in the `fBasics` package.

durbinWatsonTest(model) The Durbin-Watson test detects serial correlation in the residuals. It is in the `car` package.

dwtest(model) The Durbin-Watson test detects correlation in the residuals. It is in the `lmtest` package.

influence.measures(mod) This function computes some of the regression diagnostics to check for influential points.

jarqueberaTest(x) The Jarque-Bera test of Normality. It is in the `fBasics` package.

ksnormTest(x) This is the Kolmogorov-Smirnov Normality test. It is in the `fBasics` package.

lillie.test(x) The Lilliefors Normality test. It is in the `nortest` package.

lillieTest(x) The Lilliefors Normality test. It is in the `fBasics` package.

lm(formula) This function performs linear regression on the data, with the supplied formula. As there is much information contained in this function, you will want to save the results in a variable.

Moran.I(x,N) This function computes Moran's I autocorrelation coefficient of x giving a matrix of weights using the method described by Gittleman and Kot (1990). It is in the `ape` package.

moran.test(x,N) This is Moran's test for spatial autocorrelation using a spatial weights matrix in weights list form. It is in the `spdep` package.

ncvTest(model) This is also the Breusch-Pagan test, which tests for heteroskedasticity. It computes a score test of the hypothesis of constant error variance against the alternative that the error variance changes with the level of the response (fitted values), or with a linear combination of predictors. This is in the `car` package.

outlierTest(model) This function reports the Bonferroni p-values for Studentized residuals in linear and generalized linear models, based on a t-test for linear models (and a Normal-distribution test for generalized linear models). It is in the `car` package.

predict(model, newdata) As with almost all statistical packages, `R` has a predict function. It takes two parameters, the model, and a dataframe of the independent values from which you want to predict. If you omit `newdata`, then it will predict based on the independent variables of the data itself, which can be used to calculate residuals. The dataframe must list all independent variables with their associate new values. You can specify multiple new values for a single independent variable.

residuals(model) `R` also has a function that calculates the residuals in the data; that is, it calculates the difference between the actual value and the predicted value. This function is extremely useful when doing analyses on the residuals.

set.seed(n) This function specifies the random number seed. It allows for replication of work.

sf.test(x) The Shapiro-Francia test of Normality. It is in the `nortest` package.

sfTest(x) The Shapiro-Francia test of Normality. It is in the `fBasics` package.

shapiro.test(x) The Shapiro-Wilk test is used to quantify the degree of Normality in a group of data. The null hypothesis is that the data is Normally distributed. Thus, a p-value greater than $\alpha = 0.05$ signifies that there is not enough evidence to conclude the data is *not* Normally distributed.

shapiroTest(x) Also the Shapiro-Wilk test of Normality. It is in the `fBasics` package.

vif(model) This function calculates the variance-inflation factor. It is in the `car` package.

summaryVIFA(model) This function provides a regression table with standard errors adjusted for multicollinearity. This function needs to be sourced from the book's site on the Internet.

summaryWASE(model) This function provides a regression table with standard errors adjusted for heteroskedasticity. This function needs to be sourced from the book's site on the Internet.

Graphing:

abline() This function adds one or more straight lines through the current plot. If `h` is specified, a horizontal line is drawn. if `v`, a vertical line is drawn. If `a` and `b` are specified, a line with intercept `a` and slope `b` is drawn.

lines() This is s generic function that takes coordinates given in various ways and joins the corresponding points with line segments. The coordinates can be specified either as `x,y` or as `y~x`.

plot() This draws a scatter plot with decorations such as axes and titles in the active graphics window. The values of x and y can be specified as `x,y` or as `y~x`.

points() This is s generic function that takes coordinates given in various ways and plots the corresponding points. The coordinates can be specified either as `x,y` or as `y~x`.

Programming:

I(·) This allows one to use arithmetical functions in `R` formulas. It inhibits the formula interpretation of operators such as "+", "−", "*", and "^" so that they are used as arithmetical operators.

13.8.2 Exercises and Extensions

This section offers suggestions on things you can practice from this chapter. Save the scripts in your Chapter 13 folder.For each of the following problems, please save the associated `R` script in the chapter folder as `ext0x.R`, where `x` is the problem number.

Summary:

1. Why does there need to be variation in the independent variable?
2. What is the one assumption of ordinary least squares regression? What are the components of that one assumption?
3. Define multicollinearity. What is the effect of multicollinearity on inference?
4. How does the coefficient of determination (R^2) affect the variance inflation factor?
5. Define heteroskedasticity. What is an appropriate test for it? What is its affect on inference?
6. In general, what is a lagged dependent variable model?
7. What is the difference between the target population and the sampled population?
8. Define false discovery and explain why it is a bad thing.

Data:

For the next three questions, refer to the data and the additive model of Section 13.6.

9. In the sample, what is the average difference between male and female incomes? Is this within the 95% confidence interval predicted?
10. Given that my parent's average income was $40,000 and I am male, predict my income.
11. Create Figure 13.6.

13.8.3 References and Additional Readings

This section provides a list of statistical works. Those works cited in the chapter are here. Also here are works that complement the chapter's topics.

- Roger S. Bivand, Edzer J. Pebesma, and Virgilio Gómez-Rubio. (2008) *Applied Spatial Data Analysis with R*. New York: Springer.
- Kenneth A. Bollen and Robert W. Jackman. (1990) "Regression Diagnostics: An expository treatment of outliers and influential cases," in John Fox and J. Scott Long *Modern Methods of Data Analysis*. Los Angeles: Sage Publications.
- Trevor S. Breusch and Adrian R. Pagan. (1979) "Simple Test for Heteroscedasticity and Random Coefficient Variation." *Econometrica* **47**(5): 1287–1294.
- R. Dennis Cook and Sanford Weisberg. (1982) *Residuals and Influence in Regression*. New York: Chapman & Hall.
- R. Dennis Cook and Sanford Weisberg. (1983) "Diagnostics for heteroscedasticity in regression." *Biometrika* **70**(1), 1–10.
- Noel A. Cressie and Christopher K. Wikle. (2011) *Statistics for Spatio-Temporal Data*. Hoboken, NJ: John Wiley & Sons.
- James Durbin and Geoffrey S. Watson. (1950). "Testing for Serial Correlation in Least Squares Regression, I". *Biometrika* **37**(3-4): 409–428.
- James Durbin and Geoffrey S. Watson. (1951). "Testing for Serial Correlation in Least Squares Regression, II". *Biometrika* **38**(1-2): 159–179.
- A. Stewart Fotheringham, Chris Brunsdon, and Martin G. Charlton. (2002) *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Hoboken, NJ: John Wiley & Sons.
- James W. Hardin and Joseph M. Hilbe. (2012) *Generalized Estimating Equations*, Second Edition. New York: CRC Press.
- Cheng Hsaio. (2003) *Analysis of Panel Data*. New York: Cambridge University Press.

- Peter J. Huber. (1967) "The Behavior of Maximum Likelihood Estimates under Nonstandard Conditions." 1967 Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66), Vol. I: Statistics: 221–233.
- Roger Koenker. (1981) "A Note on Studentizing a Test for Heteroscedasticity." *Journal of Econometrics* **17**(1): 107–112.
- Frank J. Massey, Jr. (1951) "The Kolmogorov-Smirnov Test for Goodness of Fit." *Journal of the American Statistical Association*. **46**(March): 68–78.
- J. Huston McCulloch. (1985) "Miscellanea: On Heteros*edasticity." *Econometrika*. **53**(2): 483.
- Samuel S. Shapiro and Martin B. Wilk. (1965) "An Analysis of Variance Test for Normality (Complete Samples)." *Biometrika* **52**(3-4): 591–611.
- Robert H. Shumway and David S. Stoffer. (2013) *Time Series Analysis and Its Applications: With R Examples*, Third edition. New York: Springer.
- William W. S. Wei. (2005) *Time Series Analysis: Univariate and Multivariate Methods*, Second edition. Upper Saddle River, NJ: Pearson.
- Halbert L. White, Jr. (1980) "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica* **48**(4): 817–838.
- Jeffrey M. Wooldridge. (2001) *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: The MIT Press.