Regression Model Assumptions

2017-08-17

Introduction and Research Question

In this set of notes, we will explore the assumptions underlying the regression model. We will do this using the *riverside.csv* data from C. Lewis-Beck & Lewis-Beck (2016).

Preparation

```r
# Load libraries
library(broom)
library(dplyr)
library(ggplot2)
library(readr)
library(sm)

# Read in data
city = read_csv(file = "~/Dropbox/epsy-8251/data/riverside.csv")

# Fit the simple regression model
lm.1 = lm(income ~ 1 + education, data = city)
summary(lm.1)
```

```
Call:
lm(formula = income ~ 1 + education, data = city)

Residuals:
   Min     1Q Median     3Q    Max
-15808  -5783   2088   5127  18379

Coefficients:
            Estimate Std. Error t value     Pr(>|t|)
(Intercept)  11321.4     6123.2   1.849       0.0743 .
education     2651.3      369.6   7.173 0.0000000556 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8978 on 30 degrees of freedom
Multiple R-squared:  0.6317,    Adjusted R-squared:  0.6194
F-statistic: 51.45 on 1 and 30 DF,  p-value: 0.00000005562
```

Model Assumptions

There are five primary assumptions for the regression model.

- Linearity
- Independence
- Normality
- Homogeneity of variance (homoskedasticity)

We will be more specific about each of these, but it can be instructive to examine a visual representation of some of these assumptions. Imagine that we had the population of (x, y) values and we plotted them and fitted a regression to them. That picture would look like this,

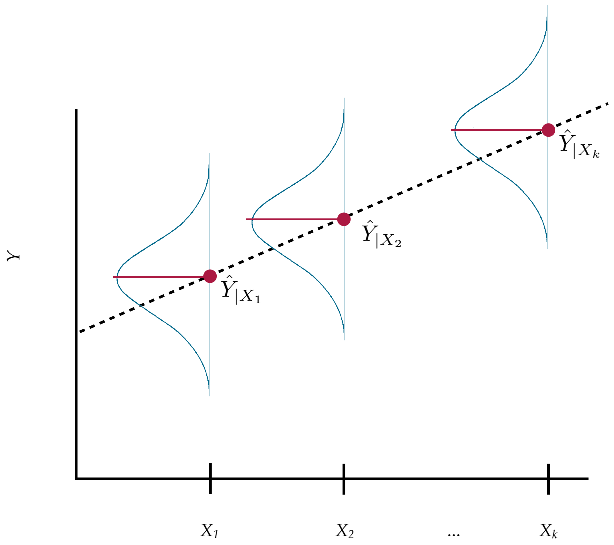

Figure 1. A visual depiction of the simple regression model's assumptions.

In Figure 1, we the normal distibutions being shown are the distribution of Y-values a particular X-value; the distribution of Y is conditioned on X, and is thus called a conditional distribution. Although only three of these distributions are shown in the figure, there is a conditional distribution for EVERY value of X. Now that we understand this picture, we can expand upon the regression assumptions listed earlier.

- **Linearity:** The linearity assumption implies that the MEAN values of Y from all the conditional distributions all fall on the same line. If this is the case, we would say that the conditional mean Y-values are linear.
- **Independence:** This is not shown in the figure. The assumption is that each Y-value in a particular conditional distribution is independent from every other Y-value in that same distribution.
- **Normality:** This assumption indicates that every one of the conditional distributions of Y-values is normally distributed.
- **Homoskedasticity:** This is the homogeneity of variance assumption. It says that the variance (or standard deviation) of all of the conditional distributions is exactly the same.

]

Assumptions are about Residuals

Note that so far we have stated these assumptions in terms of the Y-values and the conditional distributions of Y. Technically, all model assumptions (for regrression, ANOVA, t-test, etc.) all refer to the residuals. Think about how we compute the residuals

$$\epsilon_i = Y_i - \hat{Y}_i$$

In Figure 1, the $\hat{Y}_i$ value is the Y-value that corresponds to the point on the line. If we transform every Y-value in the population, from Figure 1, to an ϵ value, and re-plot them, the visual depiction now looks like this.

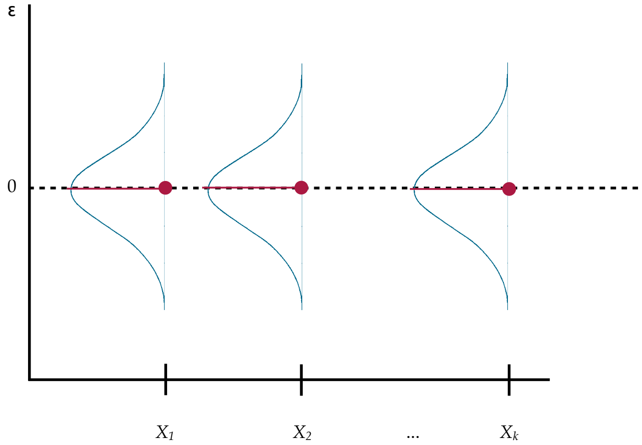

Figure 2. A visual depiction of the simple regression model's assumptions about the residuals.

So if we restate the assumptions in terms of the residuals and the conditional distributions of the residuals,

- **Linearity:** The MEAN values of each of the conditional distributions of the residuals is zero.
- **Independence:** Again, this is not shown in the figure. The assumption is that each residual value in a particular conditional distribution is independent from every other residual value in that same distribution.
- **Normality:** This assumption indicates that each of the conditional distributions of residuals is normally distributed.
- **Homoskedasticity:** The variance (or standard deviation) of all of the conditional distributions of residuals is exactly the same.

Thes assumptions can be expressed mathematically as,

$$\epsilon|X \sim \text{i.i.d } \mathcal{N}\left(0, \sigma^2\right)$$

The "i.i.d" is read, *independent and identically distributed.* The mathematical expression says the residuals conditioned on X are independent and identically normally distributed with a mean of 0 and some variance, represented as σ^2.

Evaluating the Regression Model's Assumptions

Before beginning to evaluate the assumptions using our data, it is important to point out that the assumptions are about the population's residuals. Because in most analyses, we only have a sample of data, we can never really evaluate these assumptions. We can only offer a guess as to whether they are tenable given the data we see.

This can be somewhat difficult since we expect that a sample of residuals won't actually meet these assumptions (remember sampling error). Examining the sample residuals, is however, a reasonable way to evaluate the tenability of assumptions in practice. We just have to keep in mind that the sample residuals may deviate a bit from these assumptions.

Linearity

The linearity assumption is critical in specifying the structural part of the model. This assumption can be initially evaluated theoretically (literature supporting a linear relationship between X and Y) or empirically, by examining scatterplots of the outcome vs. predictor.

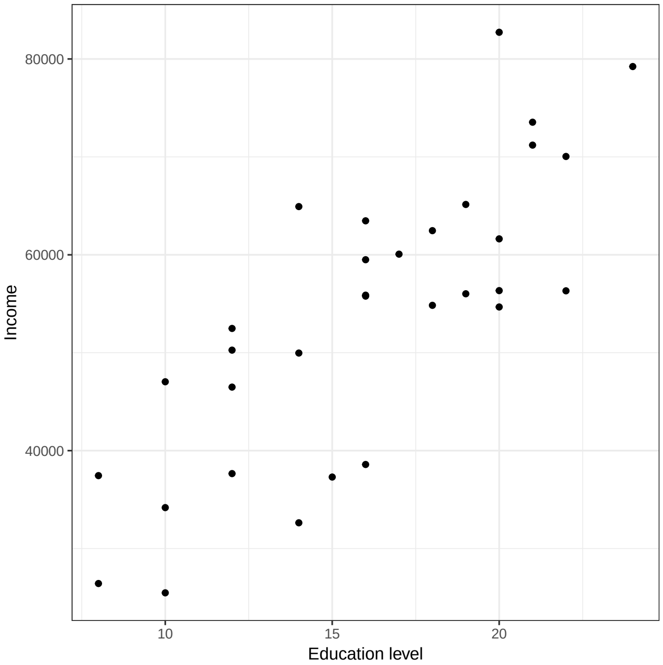

Figure 3. Scatterplot of $n = 32$ employee incomes versus their education level. The plot indicates that linearity may be a tenable based on the pattern observed in these data.

The linearity assumption seems reasonable for these data. However, we will double-check that this holds with the residuals.

Fitting a linear model when the TRUE relationship between X and Y is non-linear may, or may not be problematic. Coefficients may be wrong. Predictions may also be wrong, especially at the extreme values for X More importantly, mis-specified models lead to misinformed understandings of the world.

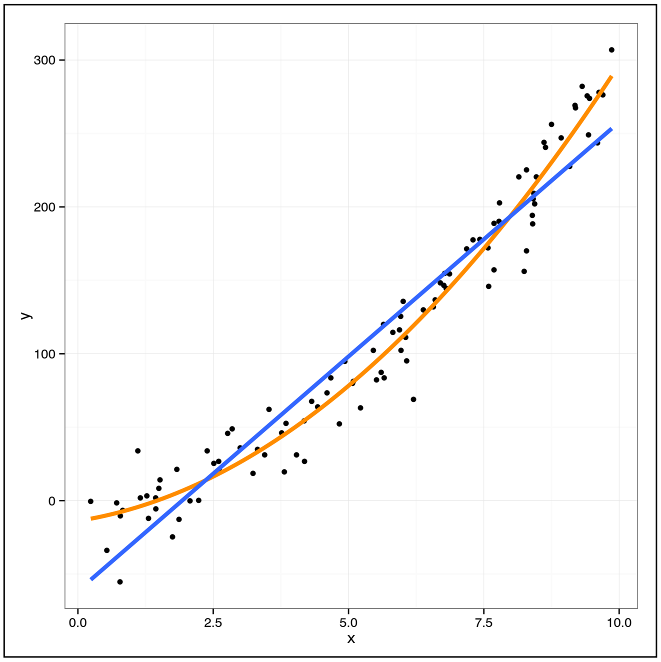

Figure 4. The plot show the differences between the true non-linear model (orange) and a mis-specified linear fitted model (blue). Using the linear fitted model to make predictions would be quite misleading, especially at extreme values of X.

Notice that when a linear model is fitted to data generated from a non-linear function (as in Figure 4) that the data are consistently above, or below the line, dependng on the X-value. This type of pattern would be evidence that the linearity assumption is not tenable. When evaluating this assumption, we want to see data in the scatterplot that is "equally" above and below the line at each value of X. Going back to the scateerplot of incomes versus education level, the fitted regression line seems to meet this criteria that at most values of X roughly half of the points are above the line, and half are below.

Independence

The definition of independence relies on formal mathematics. Loosely speaking a set of observations is independent if knowing that one observation is above or below its mean value conveys no information about whether any other observation is above or below its mean value. If observations are not independent, we say they are dependent or correlated.

Using a random chance in the design of the study, to either select observations (random sampling) or assign them to levels of the predictor (random assignment) will guarantee independence of the observations. Outside of this, independence is often difficult to guarantee, and often has to be a logical argument.

There are a few times that we can ascertain that the independnce assumption would be violated. These instances often result from aspects of the data collection process. One such instance common to social science research is when the observations (i.e., cases, subjects) are collected within a physical or spatial proximity to one another. For example, this is typically the case when a researcher gathers a convenience sample based on location, such as sampling students from the same school. Another violation of independence occurs when observations are collected over time (longitudinally), especially when the observations are repeated measures from the same subjects.

One last violation of independence occurs when the observation level used to assign cases to the different predictor values (e.g., treatment or control) does not correspond to the observation level used in the analysis.

For example, in educational studies whole classrooms are often assigned to treatment or control. That means that the cases used in the analysis, in order to satisfy the independnece assumption, would need to be at the classroom level (e.g., the cases would need to be classroom means), not individual students. This can be deleterious for sample size.

If the independence assumption is violated, almost every value you get from the `summary()` output—the standard errors, t-values, p-values, F-statistics, residual standard errors—are wrong. If you suspect that you have violated the independence assumption, then you will need to use a method (not OLS regression) that accommodates non-independence. (We cover some of these methods in EPsy 8252.)

Normality and Homoskedasticity

The assumptions about normality and homoskedasticity are about the distribution of errors at each level of X. Both of these assumptions are less critical than the assumptions of linearity and independence. It is only problematic for the OLS regression results if there are egeregious violations of the distributional assumptions. In general, if these assumptions are only minorly violated, the results of the OLS regression are still valid; we would say the results from an OLS regression are *robust* to violations of normality and homoskedasticity. Even if the violations are bad, there are many transformations of the data that can alleviate those problems. We will cover some of those transformations later in the course.

Empirically Evaluating the Assumptions

We can use the data to empirically evaluate the assumptions of linearity, normality, and homoskedasticity. (The assumption of independence is difficult to evaluate using the data, and is better left to a logical argument.) Recall that the assumptions are about the residuals, and that they also are about the population. Thus, we need to compute the residuals and recognize that as a sample, the assumptions may not be perfectly met.

To compute the residuals, we will use the `augment()` function from the **broom** package. We will also write those results into an object so we can compute on it later.

```
# Fortify the model to get residuals
out_1 = augment(lm.1)

# View fortified data
head(out_1)
```

```
  income education  .fitted  .se.fit      .resid        .hat   .sigma
1  37449         8 32531.75 3355.997    4917.248 0.13972458 9078.376
2  26430         8 32531.75 3355.997   -6101.752 0.13972458 9049.516
3  47034        10 37834.35 2727.145    9199.655 0.09226695 8953.829
4  34182        10 37834.35 2727.145   -3652.345 0.09226695 9103.810
5  25479        10 37834.35 2727.145 -12355.345 0.09226695 8808.354
6  46488        12 43136.94 2169.077    3351.061 0.05836864 9109.054
      .cooksd .std.resid
1 0.028316630  0.5904976
2 0.043602007 -0.7327412
3 0.058785810  1.0754922
4 0.009265591 -0.4269800
5 0.106032557 -1.4444104
6 0.004585452  0.3846420
```

The normality and homogeneity of variance (homoskedasticity) assumption are about the distributions of the residuals at each value of X. We evaluate these assumptions by examining a scatterplot of the residuals (plotted on the y-axis) versus the model's predictor (plotted on the x-axis). This is often referred to as a residual plot. We will also add a horizontal line at $Y = 0$.

```r
ggplot(data = out_1, aes(x = education, y = .resid)) +
  geom_point() +
  theme_bw() +
  geom_hline(yintercept = 0) +
  xlab("Education") +
  ylab("Residuals")
```

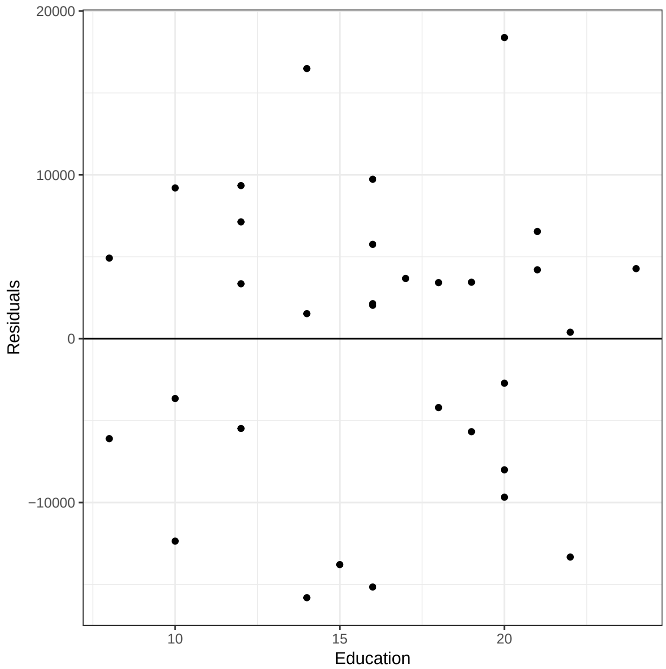

Figure 5. Scatterplot of the residuals versus education level. A horizontal line at $y = 0$ has been added to the plot to indicate the expected residual value of 0.

To evaluate homoskedasticity, we want to look to see if the variation in residuals seems consistent at different values of X (education). In other words, does the range of the residuals look about the same across the different education values?

Adding the horizontal line at $Y = 0$ allows us to evaluate the linearity assumption. If the linearity assumption is tenable, we would expect to see about half the residuals above the line and half below at each value of X. There should also be no visible pattern in the scatterplot, since the correlation of the predictor with the residuals should be 0.

What about normality of the residuals at each X value? This can be difficult to evaluate in a scatterplot. Perhaps we could look at a density plot of the residuals at each X value, but this has its own set of problems. For example, there are only three residuals at $X = 8$. It would be difficult to evaluate whether they come from a population that was normally distributed or not. This data scarcity issue is a problem at many of the X values.

In practice, researchers combine all the residuals together and examine the *marginal distribution* rather than the separate conditional distributions. This is easier to do, but is not satisfactory. Even if the marginal distribution seems normal it does not guarantee that the conditional distributions are all normally distributed. Luckily, recall that the normality assumption is not quite as critical, so in general, examining the marginal distrbution is a reasonable alternative to not being able to actually examine the conditional distributions.

We will use the `sm.density()` function from the **sm** package to plot the density curve of the residuals. We also include the `model="normal"` argument to add a confidence envelope of where we would expect to see the density IF the population were normally distributed.

```r
sm.density(out_1$.resid, xlab = "Residuals", model = "normal")
```

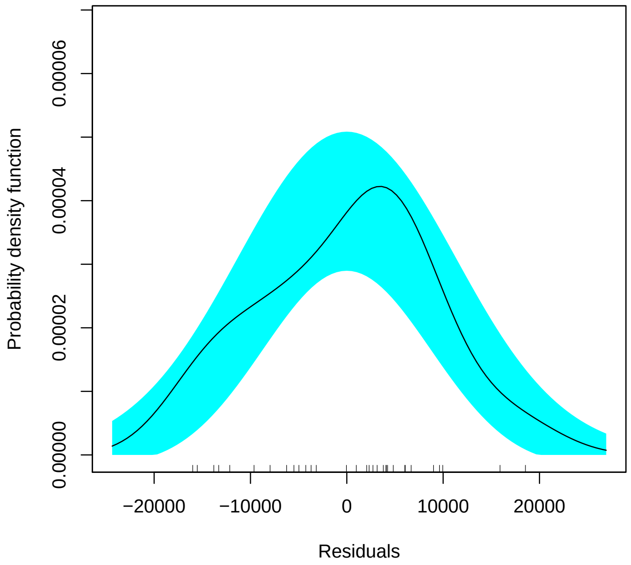

Figure 6. A density plot of the residuals from the fitted regression model. The confidence envelope (in blue) shows the sampling variation in density expected under the normality assumption.

Since the actual density curve of the residuals lies within the confidence envelope, we conclude that the normality assumption seems reasonably met (despite not looking at the conditional distributions).

Studentized Residuals

Often researchers standardize the residuals before performing the assumption checking. This does not change any of the findings from previously, in fact whether you use the raw residuals or the standardized residuals the scatterplot and density plot look identical, albeit with different scales. (Recall from introductory statistics that standardizing is a linear function which does not change the shape of the distribution.)

$$z_{\epsilon} = \frac{\epsilon - 0}{\sigma_{\epsilon}}$$

(Remember the mean value of the residual at each X is 0, so that is why we subtract 0.) Unfortunately, we do not know what the value for σ_{ϵ} (the standard error ofthe residuals) is, so we need to estimate it from the data. This adds error to the normal distribution and makes the distribution t-distributed.

$$t_{\epsilon} = \frac{\epsilon - 0}{\hat{\sigma}_{\epsilon}}$$

Since the t-distribution is also referred to as *Student's distribution*, this transformation of the residuals is called *studentizing*. What studentizing does for us is to put the residuals on a scale that uses the standard error.

This allows us to judge whether particular residuals that look extreme (either highly positive or negative) are actually extreme or not. The studentized residuals are given in the fotified output as `.std.resid`.

```r
# Plot the studentized residuals versus education level
ggplot(data = out_1, aes(x = education, y = .std.resid)) +
  geom_point() +
  theme_bw() +
  geom_hline(yintercept = 0) +
  geom_hline(yintercept = -2, linetype = "dotted") +
  geom_hline(yintercept = 2, linetype = "dotted") +
  xlab("Education") +
  ylab("Studentized residuals")

# Density plot of the studentized residuals
sm.density(out_1$.std.resid, xlab = "Studentized residuals", model = "normal")
```

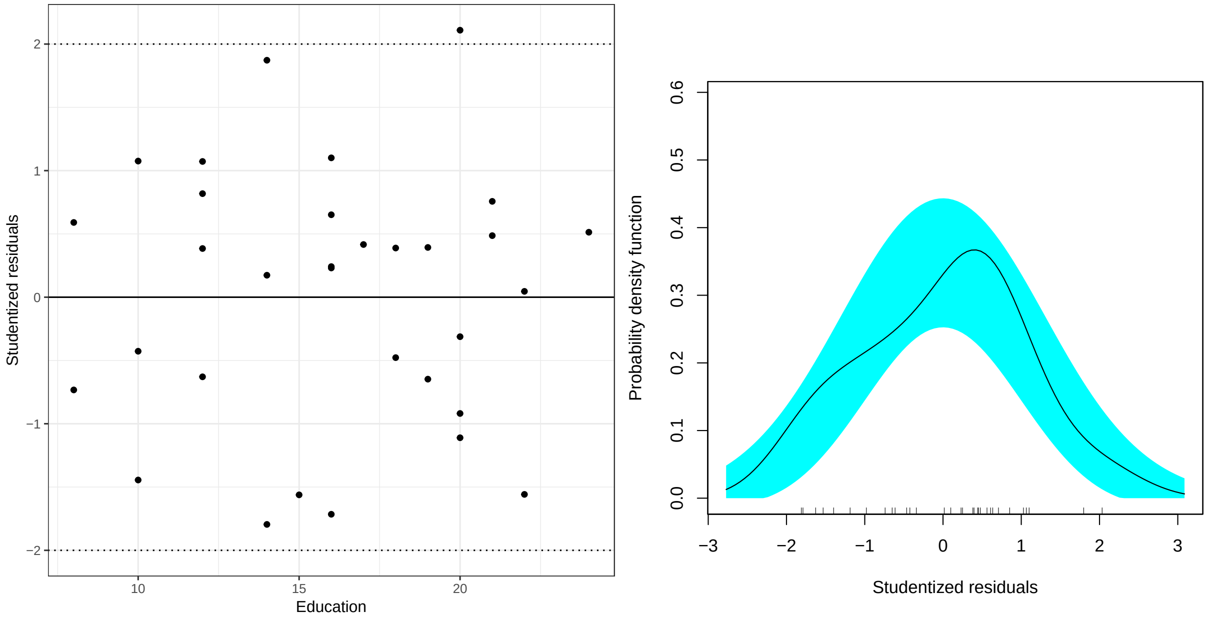

Figure 7. Scatterplot of the studentized residuals versus education level (left-hand side). Horizontal lines at $y = -2$ and $y = 2$ have been added to the plot to help identify observations that have a much higher or lower income than would be expected given their education level. A density plot of the studentized residuals is also shown (right-hand side). The confidence envelope (in blue) shows the sampling variation in density expected under the normality assumption.

The only thing that has changed between these plots and the previous plots of the raw residuals is the scale. However, now we can identify observations with extreme residuals, because we can make use of the fact that most of the residuals (~95%) should fall within two standard errors from the mean of 0. The employee with an education level of 20 years and a studentized residual of over two is slightly extreme. Given that person's education level, we would expect her/him to have an income that is much lower than it actually is. These observations are often worth a second look because they are interesting and may point to something going on in the data.

We can also `filter()` the fortified data to find these observations and to determine the exact value of their studentized residual. Recall that the vertical pipe (`|`) means "OR".

```r
out_1 %>% filter(.std.resid <= -2 | .std.resid >= 2)
```

```
  income education  .fitted  .se.fit   .resid       .hat   .sigma
1  82726        20 64347.31 2169.077 18378.69 0.05836864 8427.138
    .cooksd .std.resid
1 0.1379261   2.109545
```

Multiple Regression Assumptions

Recall that the model for a multiple regression (with two predictors) is a fitted plane that is composed of (x_1, x_2, y) ordered triples. Figure 8 visually shows the multiple regression model's assumptions.

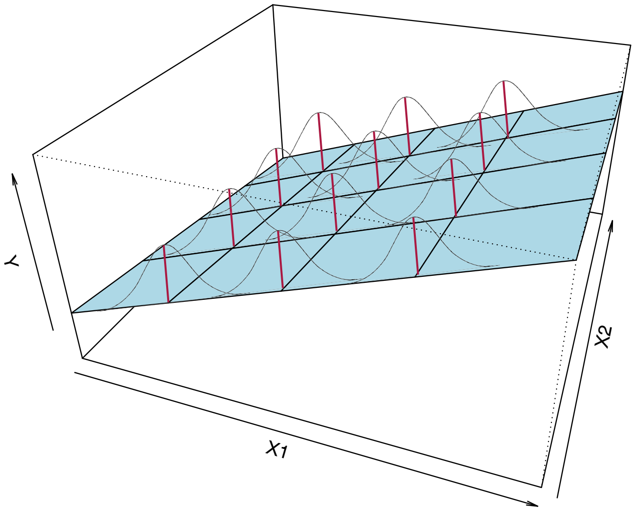

Figure 8. A visual depiction of the multiple regression model's assumptions.

Now the Y-values, and thus the residuals, at each combination $(x_1,\ x_2)$ are the conditional distributions we need to be thinking about. The assumptions for the multiple regression model are similar to those for the simple model, namely,

- **Linearity:** The MEAN values of each combination $(x_1,\ x_2)$ are linear in both the X_1 and the X_2 directions. The mean of each of the conditional distributions of the residuals is zero.
- **Independence:** Again, this is not shown in the figure. The assumption is that each residual value in a particular conditional distribution is independent from every other residual value in that same distribution.
- **Normality:** This assumption indicates that each of the conditional distributions of residuals is normally distributed.
- **Homoskedasticity:** The variance (or standard deviation) of all of the conditional distributions of residuals is exactly the same.

To evaluate these assumptions, we will create the exact same plots we created to evaluate the assumptions in the simple regression model, with one twist. Rather than creating the scatterplot by plotting the studentized residuals versus the predictor value, we will plot them against the FITTED values (i.e., the $\hat{Y}_i$ values). The fitted values from a multiple regression represent the weighted combination of both predictors, and thus give us the appropriate conditioning when we examine the distributions. (Remember, we want to consider the distribution of residuals at each $(x_1,\ x_2)$ combination.)

```r
# Fit the multiple regression model
lm.2 = lm(income ~ 1 + education + seniority, data = city)
```

```r
# Fortify the model to obtain the fitted values and residuals
out_2 = augment(lm.2)
head(out_2)
```

```
  income education seniority  .fitted  .se.fit    .resid       .hat
1  37449         8         7 29955.51 2950.407  7493.488 0.14890588
2  26430         8         9 31433.11 2875.027 -5003.105 0.14139431
3  47034        10        14 39630.78 2377.987  7403.221 0.09673126
4  34182        10        16 41108.37 2502.139 -6926.372 0.10709541
5  25479        10         1 30026.42 3213.087 -4547.425 0.17660089
6  46488        12        11 41918.08 1879.446  4569.919 0.06042377
    .sigma     .cooksd .std.resid
1 7628.276 0.065819020  1.0623539
2 7713.995 0.027374771 -0.7061822
3 7640.672 0.037051096  1.0187937
4 7656.894 0.036745076 -0.9586880
5 7723.339 0.030713542 -0.6554421
6 7730.011 0.008150562  0.6166185
```

```r
# Plot the studentized residuals versus education level
ggplot(data = out_1, aes(x = education, y = .std.resid)) +
  geom_point() +
  theme_bw() +
  geom_hline(yintercept = 0) +
  geom_hline(yintercept = -2, linetype = "dotted") +
  geom_hline(yintercept = 2, linetype = "dotted") +
  xlab("Education") +
  ylab("Studentized residuals")

# Density plot of the studentized residuals
sm.density(out_1$.std.resid, xlab = "Studentized residuals", model = "normal")
```

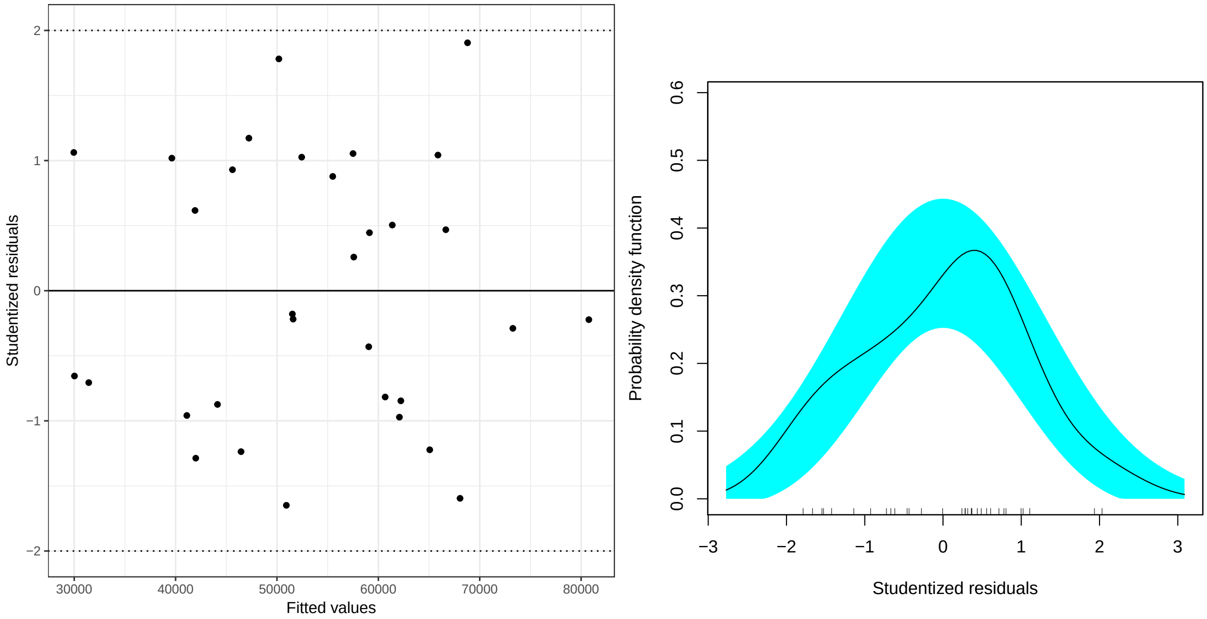

Figure 9. Scatterplot of the studentized residuals versus the fitted values from a regression model using education level and seniority level to predict income (left-hand side). Horizontal lines at $y = -2$ and $y = 2$ have been added to the plot to help identify observations that have a much higher or lower income than would be expected given their education and seniority level. A density plot of the studentized residuals is also shown (right-hand side). The confidence envelope (in blue) shows the sampling variation in density expected under the normality assumption.

The scatterplot shows random scatter around the $Y = 0$ line which indicates that the linearity assumption seems satisfied. The range of the studentized residuals at each fitted value also seem roughly the same indicating that the homoskedasticity assumption seems satsified. The density plot of the studentized residuals lies within the confidence band showing the random variation expected under the normal distribution, which suggests that normality is probably not a problem. Lastly, since the observations were randomly sampled we have satisfied the independence assumption.

If any of the assumptions (aside from the independence assumption) do not seem reasonably satisfied, you can re-plot the residual plots based on the different simple regression models. (In this case we would look at the residuals versus education and then the residuals versus seniority). This might help you identify if one, or both. of the predictors is the cause of the problem.

Advanced Plotting: Loess Smooth to Help Evaluate Linearity

In the scatterplot of the residuals (or studentized residuals) versus the fitted values, we would expect that the average value of the residual at a given fitted value would be 0. The loess smoother helps us visualize the mean pattern in the actual data. We can then compare this to what would be expected (a constant mean pattern of 0) to evaluate the linearity assumption.

To add a loess smoother, we use the `geom_smooth()` function with the argument `method="loess"`. This will plot the loess line and also the confidence envelope around that loess line. This gives us an indication of the mean pattern in the data and its uncertainty. We would hope to see the line $Y = 0$ (our expected pattern under linearity) encompassed in the uncertainty.

```
ggplot(data = out_2, aes(x = .fitted, y = .std.resid)) +
  geom_point() +
  geom_smooth(method = "loess") +
  theme_bw() +
  geom_hline(yintercept = 0) +
  xlab("Fitted values") +
  ylab("Studentized residuals")
```

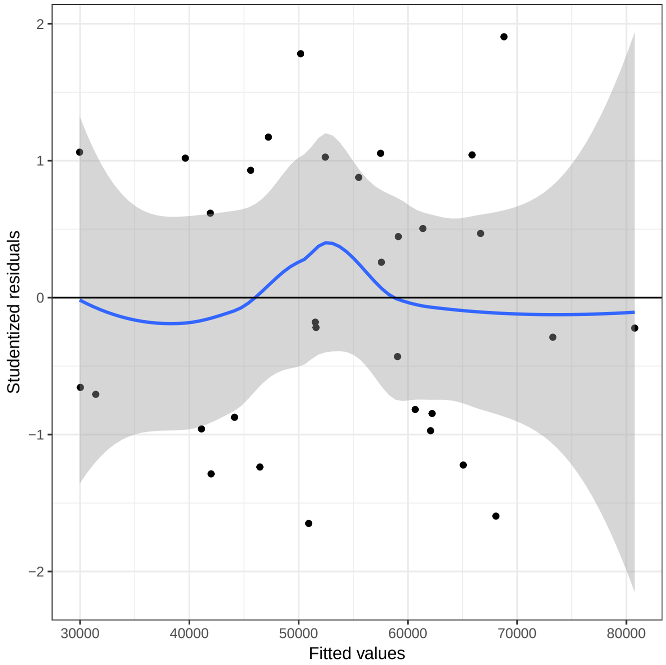

Figure 10. Scatterplot of the studentized residuals versus the fitted values from a regression model using education level and seniority level to predict income. A horizontal lines at $Y = 0$ shows the expected mean residual under the linearity assumption. The loess line (blue) and uncertainty bands (grey shaded area) indicate that the average conditional residuals for each of the fitted values in the data are not statistically different from 0 (the linearity assumption seems tenable).

Advanced Plotting: Identify Observations with Extreme Residuals

It can be useful to identify particular observations in the residual plots directly. This can be useful as you explore the plots, and also to create plots for publications in which you wish to highlight particular cases. In the plot below, we will use the fortified data from the initial simple regression model we fitted to identify the cases in the scatterplot.

Rather than plotting points (`geom_point()`) for each observation, we can plot text for each observation using `geom_text()`. For example, you might imagine writing the name of each employee in place of their point on the scatterplot. To do this, we first need to create an ID variable in the fortified data, then use `geom_text()` rather than `geom_point()` in the ggplot syntax. In the `geom_text()` function we will set `label=` to the newly created ID variable, and since it is a variable in the data set, we will put that in an `aes()` function.

Since the original data set does not include an ID variable (e.g., names), we will use the employee's row number as their ID. In other words the employee in the first row will have an ID of 1, etc.

```r
# Create ID variable in the fotified data
out_1 = out_1 %>% mutate( id = 1:nrow(out_1) )
head(out_1)
```

```
  income education  .fitted  .se.fit     .resid        .hat   .sigma
1  37449         8 32531.75 3355.997   4917.248 0.13972458 9078.376
2  26430         8 32531.75 3355.997  -6101.752 0.13972458 9049.516
3  47034        10 37834.35 2727.145   9199.655 0.09226695 8953.829
4  34182        10 37834.35 2727.145  -3652.345 0.09226695 9103.810
5  25479        10 37834.35 2727.145 -12355.345 0.09226695 8808.354
6  46488        12 43136.94 2169.077   3351.061 0.05836864 9109.054
      .cooksd .std.resid id
1 0.028316630  0.5904976  1
2 0.043602007 -0.7327412  2
3 0.058785810  1.0754922  3
4 0.009265591 -0.4269800  4
5 0.106032557 -1.4444104  5
6 0.004585452  0.3846420  6
```

```r
# Plot the id variable as text rather than points in the scatterplot
ggplot(data = out_1, aes(x = .fitted, y = .std.resid)) +
  geom_text(aes(label = id), size = 4) +
  theme_bw() +
  geom_hline(yintercept = 0) +
  geom_hline(yintercept = -2, linetype = "dotted") +
  geom_hline(yintercept = 2, linetype = "dotted") +
  xlab("Fitted values") +
  ylab("Studentized residuals")
```

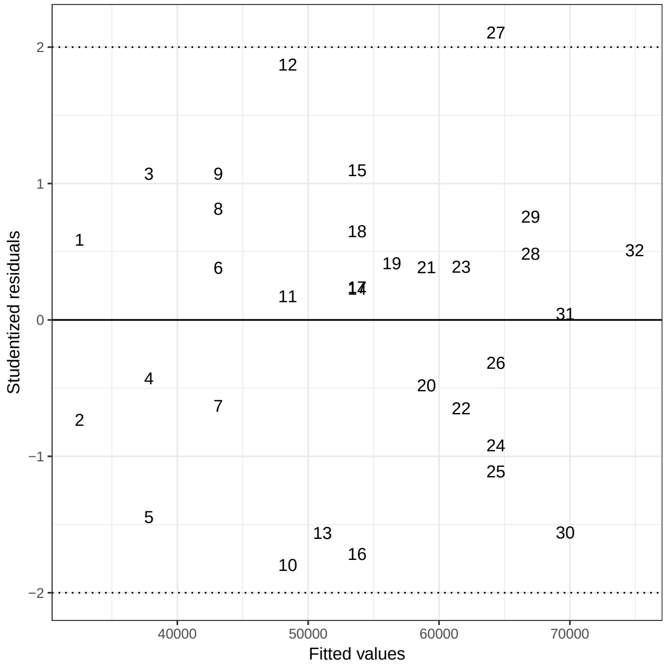

Figure 11. Scatterplot of the studentized residuals versus the fitted values from a regression model using education level to predict income. A horizontal lines at $Y = 0$ shows the expected mean residual under the linearity assumption. The values plotted indicate the employee row numbers in the data.

We can also plot points for some employees and their ID label for others. For example, suppose we wanted to give the ID number for only those employees with a studentized residual that was less than −2 or greater than 2, and plot a point otherwise. To do this, we would create the ID variable in the fortified data (which we have already done), then split the dataset into two datasets: one for those employees with extreme residuals and one for those that have a non-extreme residual. Then we will call `geom_point()` for those in the non-extreme data set, and `geom_text()` for those in the extreme set. We do this by including a `data=` argument in one of those functions to call a different dataset.

```
# Create different data sets for the extreme and non-extreme observations
extreme = out_1 %>% filter(.std.resid <= -2 | .std.resid >= 2)
nonextreme = out_1 %>% filter(.std.resid > -2 & .std.resid < 2)

# Plot using text for the extreme observations and points for the non-extreme
ggplot(data = extreme, aes(x = .fitted, y = .std.resid)) +
  geom_text(aes(label = id), size = 4, color = "red") +
  geom_point(data = nonextreme) +
  theme_bw() +
  geom_hline(yintercept = 0) +
  geom_hline(yintercept = -2, linetype = "dotted") +
  geom_hline(yintercept = 2, linetype = "dotted") +
  xlab("Fitted values") +
  ylab("Studentized residuals")
```

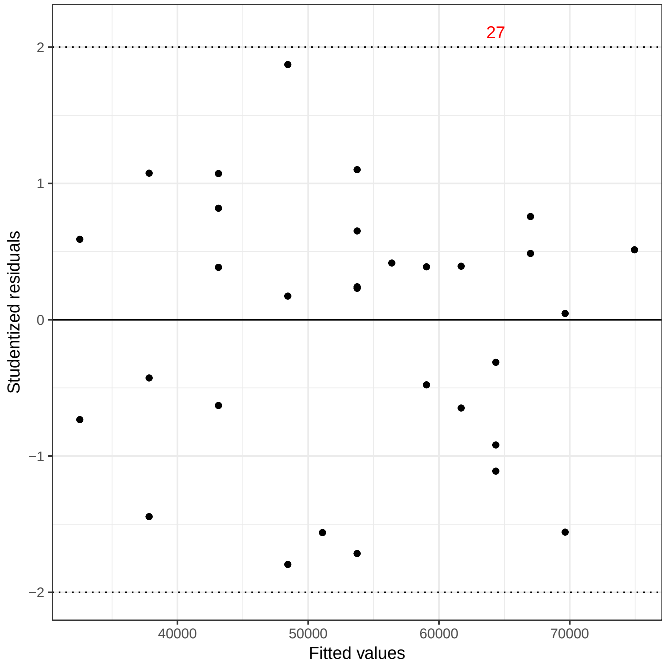

Figure 12. Scatterplot of the studentized residuals versus the fitted values from a regression model using education level to predict income. A horizontal lines at $Y = 0$ shows the expected mean residual under the linearity assumption. The values plotted indicate the employee row numbers in the data. These values are only shown for employees with residuals larger than two standard errors from 0.

References

Lewis-Beck, C., & Lewis-Beck, M. (2016). *Applied regression: An introduction* (2nd ed.). Thousand Oaks, CA: Sage.