# Probability Distributions in Regression

Andrew Zieffler
January 07, 2021

## Preparation

To illustrate how probability distributions are used in practice, we will will use the data in the file *riverview.csv* (see the data codebook (https://zief0002.github.io/epsy-8252/codebooks/riverview.html) for more information about these data) and fit a regression model that uses education level and seniority to predict variation in employee income. Some (most?) of this content should also be review from EPsy 8251.

```
# Load libraries
library(broom)
library(tidyverse)

# Import and view data
city = read_csv(file = "~/Documents/github/epsy-8252/data/riverview.csv")
head(city)
```

```
# A tibble: 6 x 5
  education income seniority gender     party
      <dbl>  <dbl>     <dbl> <chr>      <chr>
1         8   26.4         9 female     Independent
2         8   37.4         7 Not female Democrat
3        10   34.2        16 female     Independent
4        10   25.5         1 female     Republican
5        10   47.0        14 Not female Democrat
6        12   46.5        11 female     Democrat
```

To begin, we will fit a multiple regression model that uses level of education and seniority to predict variation in employee's incomes. The model is:

$$\begin{aligned} \text{Income}_i &= \beta_0 + \beta_1(\text{Education}_i) + \beta_2(\text{Seniority}_i) + \epsilon_i \\ \text{where} \quad \epsilon_i &\sim \mathcal{N}(0, \sigma^2_\epsilon) \end{aligned}$$

We have four unknowns in this model that need to be estimated: $\beta_0$, $\beta_1$, $\beta_2$, and $\sigma^2_\epsilon$. This last unknown is the error (or residual) variance. As a side note, when you report results from a fitted regression model, you should report the estimated residual variance (or residual standard error) along with the coefficient estimates.

Aside from the estimates for the coefficients and RSE, we are also generally interested in the estimates of uncertainty for the coefficients (i.e., the standard errors). These uncertainty estimates also allow us to carry out hypothesis tests on the effects included in the model.

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

In practice, all of the estimates, SEs, and inferential output are available using functionality in R. For example, the model-level output, including $R^2$, the *F*-statistic, the model and residual *df*, and the residual standard error are all outputted from the `glance()` function from the `{broom}` package. We can also partition the variation using the `anova()` function.

```
# Model-level output
glance(lm.1)
```

```
# A tibble: 1 x 12
  r.squared adj.r.squared sigma statistic p.value    df logLik   AIC   BIC
      <dbl>         <dbl> <dbl>     <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>
1     0.742         0.724  7.65      41.7 2.98e-9     2  -109.  226.  232.
# … with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

```
# Partition the variation
anova(lm.1)
```

```
Analysis of Variance Table

Response: income
           Df Sum Sq Mean Sq F value         Pr(>F)
education   1 4147.3  4147.3  70.944 0.000000002781 ***
seniority   1  722.9   722.9  12.366        0.00146 **
Residuals  29 1695.3    58.5
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Similarly the `tidy()` function from the `{broom}` package outputs coefficient-level output, including the coefficients, standard errors, *t*-values, and associated *p*-values.

```
# Coefficient-level output
tidy(lm.1)
```

```
# A tibble: 3 x 5
  term        estimate std.error statistic     p.value
  <chr>          <dbl>     <dbl>     <dbl>       <dbl>
1 (Intercept)    6.77      5.37       1.26 0.218
2 education      2.25      0.335      6.73 0.000000220
3 seniority      0.739     0.210      3.52 0.00146
```

Our goal here is to understand how the probability distributions play a role in determining some of these values.

# Model-Level Inference: The *F*-Distribution

At the model-level, we are interested in whether or not the model (as a whole) explains variation in the outcome. Our estimate of how much variation the model explains is based on partitioning the total variation of the outcome ($SS_{Total}$) into that which is explained by the model ($SS_{Model}$) and that which is not explained by the model ($SS_{Residual}$). From the `anova()` output:

$$\begin{aligned} SS_{Model} &= 4147.3 + 722.9 = 4870.2 \\ SS_{Residual} &= 1695.3 \\ SS_{Total} &= 4870.2 + 1695.3 = 6565.5 \end{aligned}$$

Then we compute the a statistic called $R^2$ by computing the ratio of the explained variation to the total variation.

$$R^2 = \frac{4870.2}{6565.5} = 0.742$$

The model (differences in education and seniority levels) explains 74.2% of the variation in employee's incomes in the sample. We might also want to test whether this is more variation than we expect because of sampling error. To do this we want to test the hypothesis that:

$$H_0 : \rho^2 = 0$$

To evaluate this we convert the sample $R^2$ value into a test statistic using,

$$F = \frac{R^2}{1 - R^2} \times \frac{df_{\text{Error}}}{df_{\text{Model}}}$$

The degrees of freedom (*df*) is also partitioned in the `anova()` output:

$$\begin{aligned} \text{df}_{\text{Model}} &= 1 + 1 = 2 \\ \text{df}_{\text{Residual}} &= 29 \\ \text{df}_{\text{Total}} &= 2 + 29 = 31 \end{aligned}$$

Converting our $R^2$ value of 0.742 value to an *F*-statistic:

$$\begin{aligned} F &= \frac{0.742}{1 - 0.742} \times \frac{29}{2} \\ &= 41.7 \end{aligned}$$

We write this standardization of $R^2$ as $F(2, 29) = 41.7$.

## Computing F from the ANOVA Partitioning

We can also compute the model-level *F*-statistic directly using the partitioning of variation from the ANOVA table.

```
# Partition the variation
anova(lm.1)
```

```
Analysis of Variance Table

Response: income
          Df Sum Sq Mean Sq F value         Pr(>F)
education  1 4147.3  4147.3  70.944 0.000000002781 ***
seniority  1  722.9   722.9  12.366        0.00146 **
Residuals 29 1695.3    58.5
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The *F*-statistic is a ratio of the mean square for the model and the mean square for the error. To compute a mean square we use the general formula:

$$\text{MS} = \frac{\text{SS}}{\text{df}}$$

The model includes both the education and seniority predictor, so we combine the SS and df. The MS model is:

$$\begin{aligned} \text{MS}_{\text{Model}} &= \frac{\text{SS}_{\text{Model}}}{\text{df}_{\text{Model}}} \\ &= \frac{4147.3 + 722.9}{1 + 1} \\ &= \frac{4870.2}{2} \\ &= 2435.1 \end{aligned}$$

The MS error is:

$$\begin{aligned} \text{MS}_{\text{Error}} &= \frac{\text{SS}_{\text{Error}}}{\text{df}_{\text{Error}}} \\ &= \frac{1695.3}{29} \\ &= 58.5 \end{aligned}$$

Then, we compute the *F*-statistic by computing the ratio of these two mean squares.

$$\begin{aligned} F &= \frac{\text{MS}_{\text{Model}}}{\text{MS}_{\text{Error}}} \\ &= \frac{2435.1}{58.5} \\ &= 41.6 \end{aligned}$$

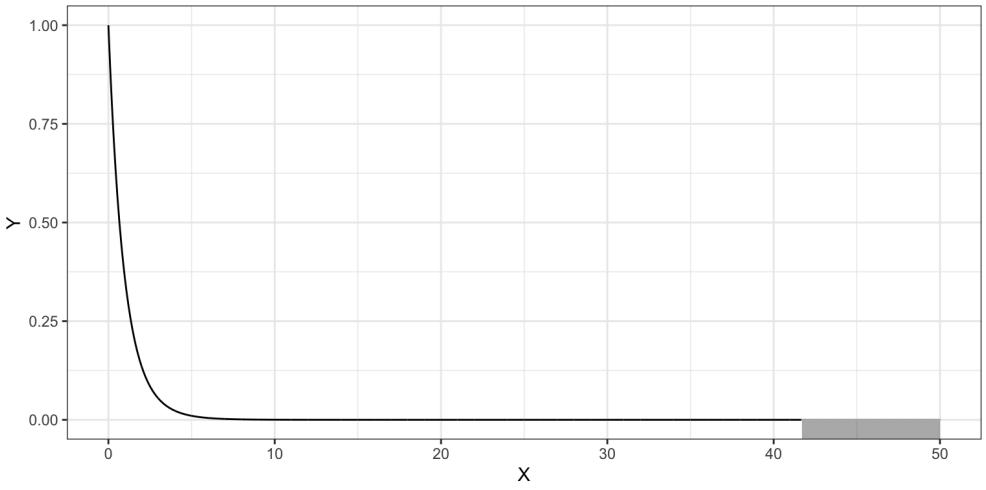
Since a mean square represents the average amount of variation (per degree of freedom), we can see that $F$ is a ratio between the average amount of variation explained by the model and the average amount of variation unexplained by the model. In our example, this ratio is 41.6; on average the model explains 41.6 times the variation that is unexplained.

Note that this is an identical computation (although reframed) as the initial computation for $F$. We can use mathematics to show this equivalence:

$$
\begin{aligned}
F &= \frac{R^2}{1 - R^2} \times \frac{\text{df}_{\text{Error}}}{\text{df}_{\text{Model}}} \\
&= \frac{\frac{\text{SS}_{\text{Model}}}{\text{SS}_{\text{Total}}}}{\frac{\text{SS}_{\text{Error}}}{\text{SS}_{\text{Total}}}} \times \frac{\text{df}_{\text{Error}}}{\text{df}_{\text{Model}}} \\
&= \frac{\text{SS}_{\text{Model}}}{\text{SS}_{\text{Error}}} \times \frac{\text{df}_{\text{Error}}}{\text{df}_{\text{Model}}} \\
&= \frac{\text{SS}_{\text{Model}}}{\text{df}_{\text{Model}}} \times \frac{\text{df}_{\text{Error}}}{\text{SS}_{\text{Error}}} \\
&= \text{MS}_{\text{Model}} \times \frac{1}{\text{MS}_{\text{Error}}} \\
&= \frac{\text{MS}_{\text{Model}}}{\text{MS}_{\text{Error}}}
\end{aligned}
$$

## Testing the Model-Level Null Hypothesis

We evaluate our test statistic ($F$ in this case) in the appropriate test distribution, in this case an $F$-distribution with 2 and 29 degrees of freedom. The figure below, shows the $F(2, 29)$-distribution as a solid, black line. The $p$-value is the area under the curve that is at least as extreme as the observed $F$-value of 41.7.




*Plot of the probability density function (PDF) for the $F(2,\ 29)$-distribution. The cumulative density representing the p-value for a test evaluating whether $\rho^2 = 0$ using an observed F-statistic of 41.7 is also displayed.*

The computation using the cumulative density function, `pf()`, to obtain the $p$-value is:

```
# p-value for F(2,29)=41.7
1 - pf(41.7, df1 = 2, df2 = 29)
```

```
[1] 0.000000002942114
```

## Mean Squares are Variance Estimates

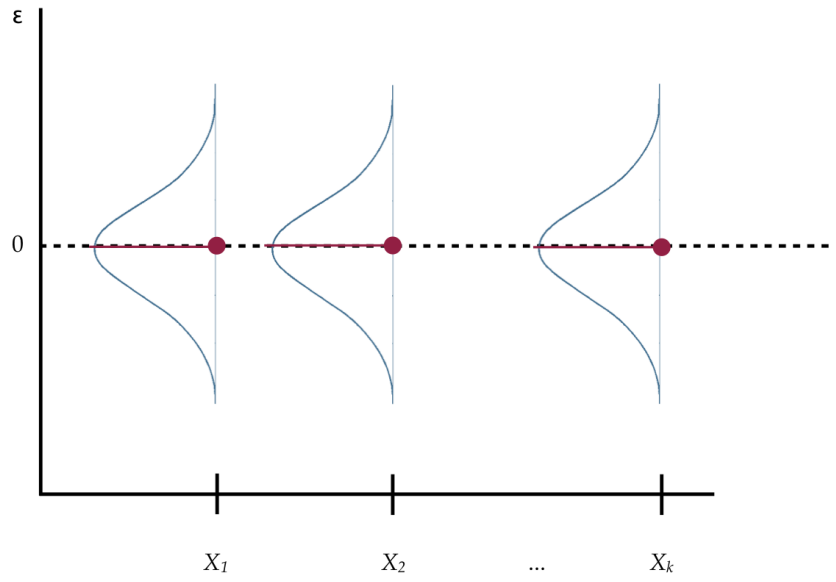
Mean squares are also estimates of the *variance*. Consider the computational formula for the sample variance,

$$\hat{\sigma}^2 = \frac{\sum(Y - \bar{Y})^2}{n - 1}$$

This is the total sum of squares divided by the total *df*. The variance of the outcome variable is interpreted as the average amount of variation in the outcome variable (in the squared metric). Thus, it is also referred to as the mean square total.

When we compute an *F*-statistic, we are finding the ratio of two different variance estimates—one based on the model (explained variance) and one based on the error (unexplained variance). Under the null hypothesis that $\rho^2 = 0$, we are assuming that all the variance is unexplained. In that case, our *F*-statistic would be close to zero. When the model explains a significant amount of variation, the numerator gets larger relative to the denominator and the *F*-value is larger.

The mean squared error (from the `anova()` output) plays a special role in regression analysis. It is the variance estimate for the conditional distributions of the residuals in our visual depiction of the distributional assumptions of the residuals underlying linear regression.




*Visual Depiction of the Distributional Assumptions of the Residuals Underlying Linear Regression*

Recall that we made implicit assumptions about the conditional distributions of the residuals, namely that they were identically and normally distributed with a mean of zero and some variance. Based on the estimate of the mean squared error, the variance of each of these distributions is 58.5.

While the variance is a mathematical convenience, the standard deviation is often a better descriptor of the variation in a distribution since it is measured in the original metric. The standard deviation fro the residuals (error) is 7.6. Because the residuals are statistics (summaries computed from sample data), their standard deviation is referred to as a "standard error." The *residual standard error* (RSE) is sometimes referred to as the *Root Mean Squared Error* (RMSE).

```
# Compute RMSE
sqrt(58.5)
```

```
[1] 7.648529
```

Why is this value important? It gives the expected variation in the conditional residual distributions, which is a measure of the average amount of error. For example, since all of the conditional distributions of the residuals are assumed to be normally distributed, we would expect that 95% of the residuals would fall between $\pm 2$ standard errors from 0; or, in this case, between $-15.3$ and $+15.3$. Observations with residuals that are more extreme may be regression outliers.

More importantly, it is a value that we need to estimate in order to specify the model.

# Coefficent-Level Inference: The *t*-Distribution

Recall that the coefficients and SEs for the coefficients are computed directly from the raw data based on the OLS estimation. These can then be uses to construct a test statistic (e.g., *t*) to carry out a hypothesis test or compute endpoints for a confidence interval. To see how this is done, we will consider the partial effect of education level (after controlling for differences in seniority) in our fitted model.

$$\hat{\beta}_{\text{Education}} = 2.25$$
$$\text{SE}(\hat{\beta}_{\text{Education}}) = 0.335$$

We might want to test whether the partial effect of education level on income, after accounting for differences in seniority level, we observed in the data is more than we would expect because of sampling error. To answer this we need to evaluate the following hypothesis:

$$H_0 : \beta_{\text{Education}} = 0$$

We begin by converting our estimated regression coefficient to a *t*-statistic using:

$$t_k = \frac{\hat{\beta}_k}{\text{SE}(\hat{\beta}_k)}$$

In our example,

$$t_{\text{Education}} = \frac{2.25}{0.335}$$
$$= 6.72$$

Remember this tells us the education coefficient of 2252 is 6.72 standard errors above 0. Since we are estimating the SE using sample data, our test statistic is likely *t*-distributed[1]. Which value should we use for *df*? Well, for that, statistical theory tells us that we should use the error *df* value from the model. In our example, this would be:

$$t(29) = 6.72$$

Using the *t*-distribution with 29 *df*, we can compute the *p*-value associated with the two-tailed test that $\beta_{\text{Education}=0}$:

```
# p-value for the two-tailed test of no effect of education
2 * pt(q = -6.72, df = 29)
```

```
[1] 0.0000002257125
```

The $p$-value is 0.0000002. The data are inconsistent with the hypothesis that there is no partial effect of education level on income (after accounting for differences in seniority level).

We could also carry out these tests for the partial effect of seniority level and, if it is of interest, for the intercept. For both of those tests, we would use the same $t$-distribution, but our test statistic would be computed based on the coefficient estimates and standard errors for those terms, respectively.

# Regression Model and Simulation

We fitted a multiple regression model that uses level of education and seniority to predict variation in employee's incomes. The fitted equation is:

$$\hat{\text{Income}}_i = 6.77 + 2.25(\text{Education}_i) + 0.739(\text{Seniority}_i)$$

We also make a set of assumptions around the residuals, which, using our estimated RSE are:

$$\epsilon_i \sim \mathcal{N}(0, 7.65^2)$$

Note that here the $RSE = 7.65$ and the estimated residual variance is $\hat{\sigma}^2 = 7.65^2 = 58.5$.

The statistical model is a description of the data generating process we believe underlies our observed data. For example the data generating process here is that:

- We have a *fixed* set of $X$ values (education and seniority values) in the observed data.
- These values can be used along with the fitted equation to produce mean income values ($\hat{Y}_i$ values).
- The errors for each of our $n$ cases are then randomly generated from the $\mathcal{N}(0, 7.65^2)$ distribution.
- Since $Y_i = \hat{Y}_i + \hat{\epsilon}_i$, we can add the $\hat{Y}_i$ values to the randomly generated errors to get the $Y_i$ values.
- This gives us a dataset of the $X$ (education and seniority) and $Y$ (income) values.

Remember that the data are assumed to be one possible sample from this data generating process. If we carried out this process an infinite number of times, our observed data would be one of the datasets that we produce.

Here is some R syntax that mimics this data generating process.

```
# Generate Y-values using the same X values that are in the data
new_city = city %>%
  select(education, seniority) %>%
  mutate(
    y_hat = 6.77 + 2.25*education + 0.739*seniority,
    e_i = rnorm(n = 32, mean = 0, sd = 7.65),
    income = y_hat + e_i
  )

# View generated data
new_city
```

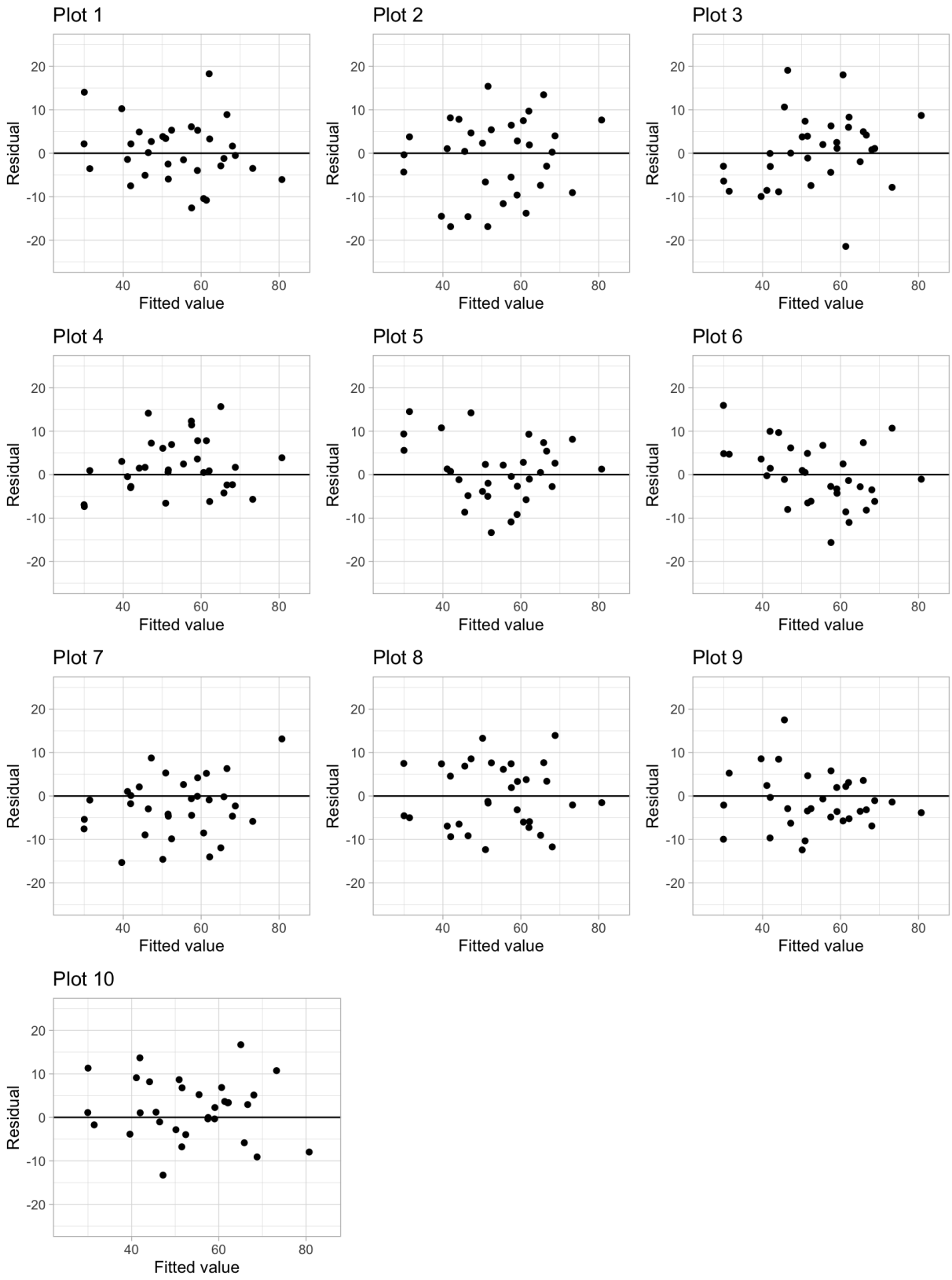
```
# A tibble: 32 x 5
   education seniority y_hat     e_i income
       <dbl>     <dbl> <dbl>   <dbl>  <dbl>
 1         8         9  31.4 -4.40     27.0
 2         8         7  29.9  4.48     34.4
 3        10        16  41.1  6.89     48.0
 4        10         1  30.0 13.6      43.7
 5        10        14  39.6  4.51     44.1
 6        12        11  41.9 -1.48     40.4
 7        12        16  45.6  0.0386   45.6
 8        12        14  44.1  9.22     53.3
 9        12        24  51.5 -1.22     50.3
10        14         5  42.0 -3.30     38.7
# … with 22 more rows
```

This does not exactly reproduce the observed data, but instead are one possible sample that would be produced from the data generating process (i.e., from the model).

It is quite useful to be able to mimic the data generating process. In methodological work, this often forms the basis for examining characteristics of the model or violations of the model's assumptions. For example, if we have questions about whether violations of the normality assumption have an effect on the estimates we obtain, we could generate data using a non-normal distribution and then examine the estimates from a regression fitted to the simulated data. Doing this many times would allow us to evaluate how the non-normality affects the estimates.

This methodology has also been proposed as a way to evaluate the model assumptions. To do this we might create several randomly generated datasets from the model and then create plots of those datasets. We would also create the same plot of the actual observed data and then randomly arrange all the plots. If you can identify the observed data in this series of plots, it is a suggestion that the model was mis-specified; a violation of one or more assumptions.

To illustrate this, consider the 10 plots below. One of them is the observed data and nine of them were randomly generated from the estimated model. Can you tell which is the observed data?




None of the plots look markedly different from the others. This suggests that the plot with the observed data is indistinguishable from plots that used data randomly generated from the model, which implies that the observed data could have been generated from this model (the assumptions are tenable)[2].

# References


1. Whether this is actually *t*-distributed depends on whether the model assumptions are met.↩

2. It turns out the observed data is in Plot 8. Look at the syntax in the RMD file to see how these plots were generated and randomly ordered.↩