Review of Statistical Concepts

Welcome Back Everyone!

Hope you all had a refreshing summer!

What Are We Doing this Semester?

Extend the regression model in two ways:

1. Relax the normality assumption: Logistic Regression (GLMs)
2. Relax the independent residuals assumption: Mixed-effects regression models

Semester Assignments

- Homework (~5-6 over the course)
- In-Class Projects (For immersion days)
- Project

Overview for Today

- Probability & Statistics Review
- R/RStudio Review

What is Probability?

Probability is the language of uncertainty.

Anytime we are dealing with random events such as the outcome of a coin toss or the response to a survey question, we rely on probability to talk about these events.

Axioms (Rules) of Probability

Probability theory is built on three rules:

1. $P(\text{Event}) \geq 0$
2. $P(\text{Any Event} = 1$
3. $P(\text{A or B}) = P(\text{A}) + P(\text{B})$ for Mutually Exclusive events

Joint & Conditional Probabilities

When dealing with two or more random variables, we can describe the probability of multiple events happening using joint probabilities and conditional probabilities:

1. *Joint Probability*: Probability of rolling a 1 and a 2
2. *Conditional Probability*: Probability of rolling a 1 given (conditional on) your first roll was a 1

Simulating a Roll of Two Dice

```
set.seed(435)
roll_1 <- sample(1:6, size = 20000, replace = TRUE)
roll_2 <- sample(1:6, size = 20000, replace = TRUE)
xtabs(~roll_1 + roll_2) |> prop.table() |> round(2)
```

Simulating a Roll of Two Dice

```
      roll_2
roll_1    1    2    3    4    5    6
     1 0.03 0.03 0.03 0.03 0.03 0.03
     2 0.03 0.03 0.03 0.03 0.03 0.03
     3 0.03 0.03 0.03 0.03 0.03 0.03
     4 0.03 0.03 0.03 0.03 0.03 0.03
     5 0.03 0.03 0.03 0.03 0.03 0.03
     6 0.03 0.03 0.03 0.03 0.03 0.03
```

Independent Events

Two or more events are independent when the occurrence of one event has no impact on the occurrence of the other events:

$$P(A|B) = P(A)$$

Are Die Rolls Independent?

If you roll a pair of dice, is the first roll independent of the second?

Calculating Conditional Independence

```
xtabs(~roll_1 + roll_2) |> prop.table(1) |> round(2)
```

Calculating Conditional Independence

```
      roll_2
roll_1    1    2    3    4    5    6
     1 0.17 0.17 0.16 0.16 0.17 0.17
     2 0.17 0.18 0.18 0.16 0.17 0.16
     3 0.17 0.17 0.17 0.16 0.16 0.18
     4 0.15 0.18 0.17 0.16 0.17 0.17
     5 0.16 0.17 0.16 0.18 0.16 0.17
     6 0.16 0.17 0.16 0.17 0.17 0.18
```

Probability Mass/Distribution Function

Probability Mass and Density Functions (PMF & PDF, respectively) are functions that take the value of a random variable as an input and output the probability of that value occurring. Every statistical model we will use will assume a certain PMF or PDF.

- PMF is a probability distribution function for **discrete random variables**
- PDF is a probability distribution function for **continuous random variables**

Bernouli Distribution

The Bernoulli Distribution is a PMF used for a random variable that takes on two different values:

- Coin toss: Heads or Tails
- Football game: Win or Loss
- Clicked on an ad: Yes or No

PMF for UGA Winning the College Football National Championship

$$p(\text{Win}) = \pi^Y (1 - \pi)^{1-Y}$$

$$\pi = \text{Probability UGA Wins}$$

$$Y = 1 \text{ if they win, } 0 \text{ if they lose}$$

Using PMFs in R

$$p(\text{Win}) = .25^{Y}(1 - .25)^{1-Y}$$

```r
dbinom(1, 1, prob = .25)
```

```
[1] 0.25
```

```r
dbinom(0, 1, prob = .25)
```

```
[1] 0.75
```

Binomial Distribution

The binomial distribution is a PMF used for a random variable that is the count of successes of n independent experiments/trials (multiple, independent Bernoulli variables):

- Probability of 10 heads out of 15 tosses (head = success)
- Probability a college football team wins 10 of its 12 games
- Probability a user clicks on 3 of the 5 ads presented to them

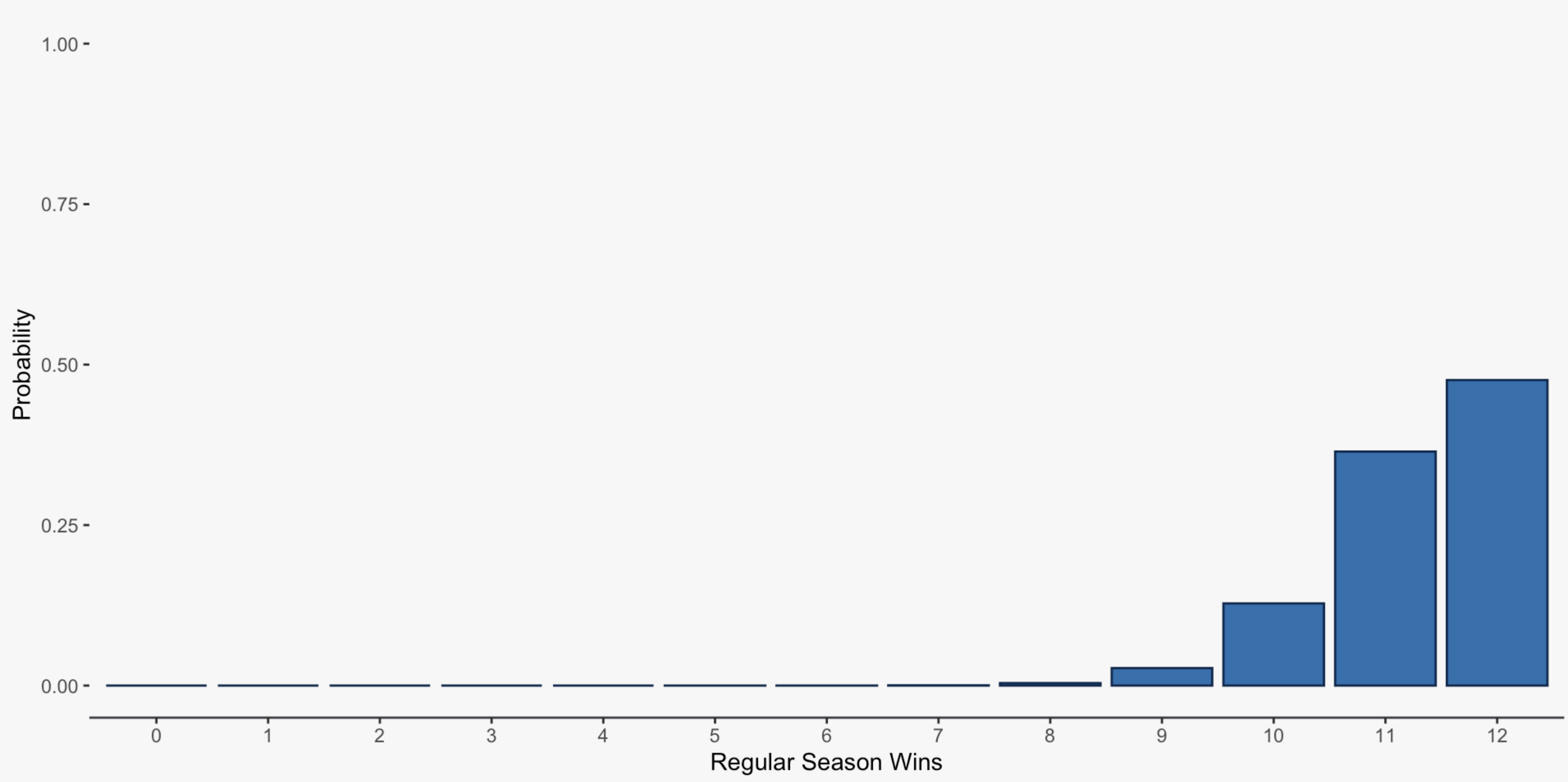
The Probability Distribution of UGA's Regular Season Record

UGA's record under their current head coach: 94-16 (94%). So let's say they have a 94% chance of winning each game – what does the probability distribution of their 12 game season win-loss record look like?

```r
data_record <-
  tibble::tibble(
    record = 0:12,
    prob = dbinom(record, 12, .94)
  )

ggplot2::ggplot(
  data = data_record,
  ggplot2::aes(x = as.factor(record), y = prob)
) +
  ggplot2::geom_bar(stat = "identity") +
  ggplot2::ylim(c(0, 1))
```

The Probability Distribution of UGA's Regular Season Record

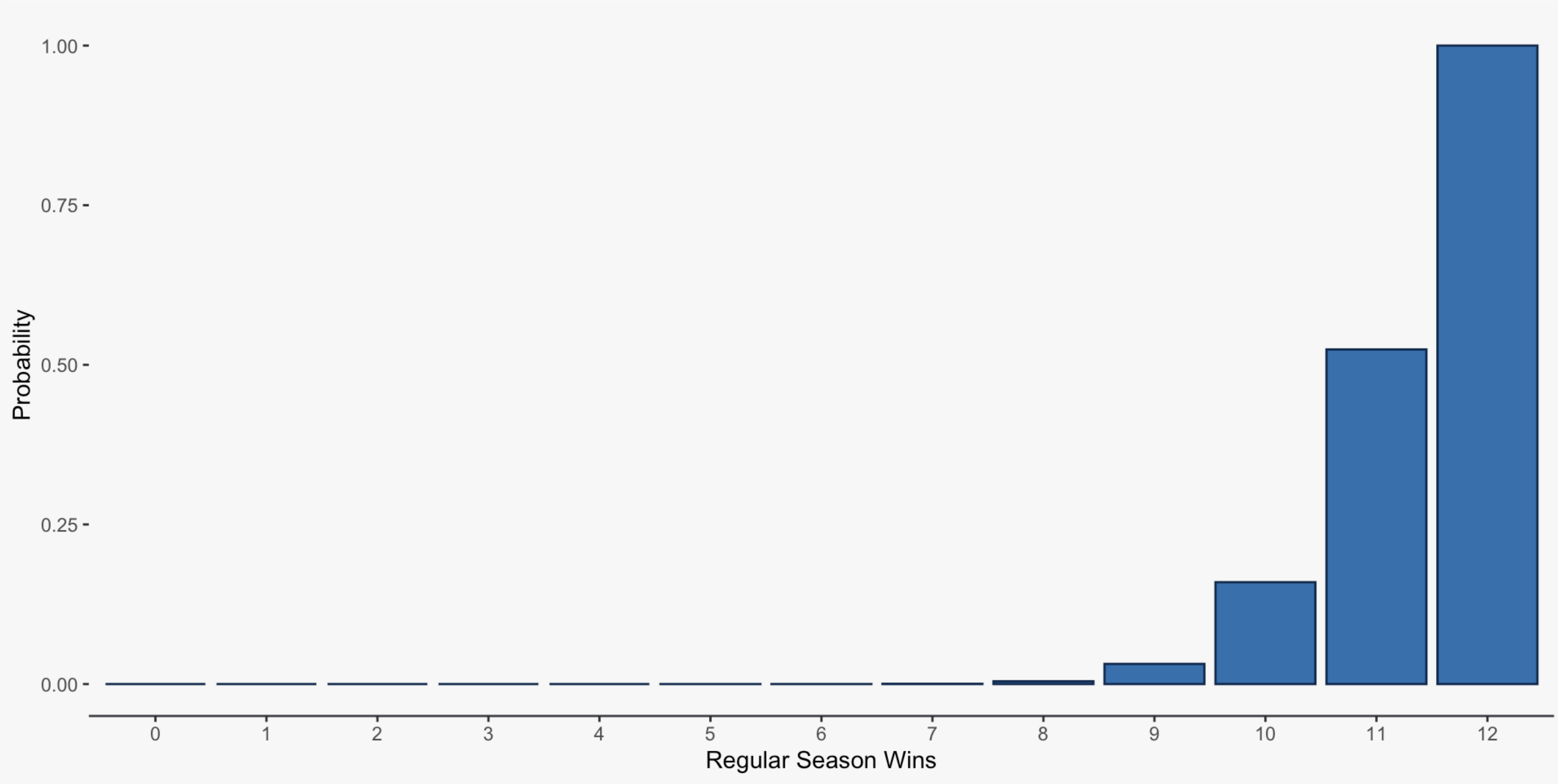


Cumulative Distribution Function

The Cumulative Distribution Function (CDF) specifies the probability that a random variable takes a value, Y, or any value less than Y (think of percentiles).

Probability UGA Wins 10 or Less Games

$$F(\text{UGA Record} = 10) = P(\text{UGA Record} \leq 10)$$



How Does Regression Connect to Probability?

The simple linear regression model we've seen before:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i$$

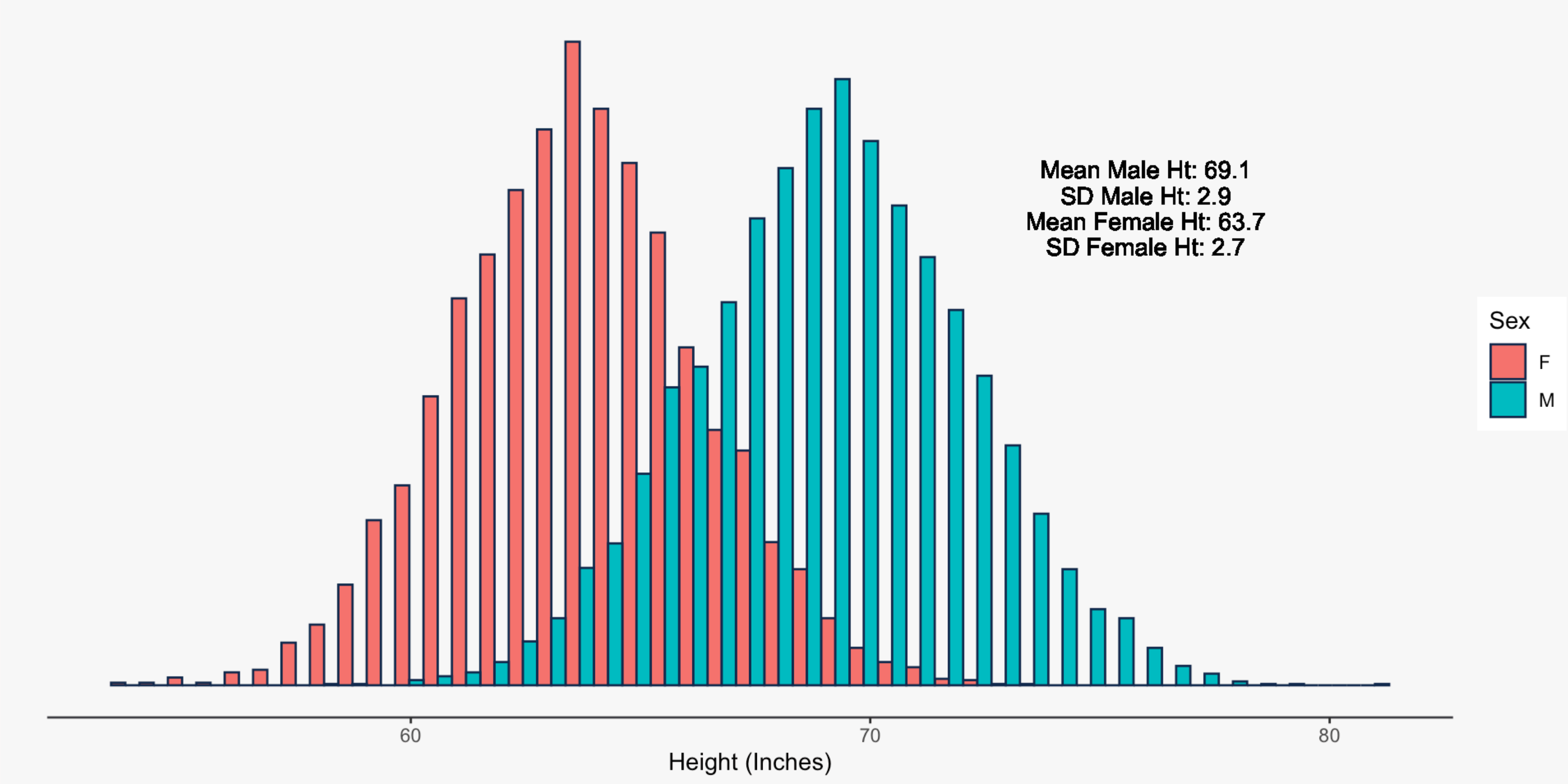
$$\epsilon_i \sim N(0, \sigma)$$

Regression as a Probability Model

Rewriting linear regression as a probability model:

$$P(Y_i | X_{i1}) = N(\beta_0 + \beta_1 X_{i1}, \sigma)$$

US Heights by Sex



Using Linear Regression to Describe Heights

```r
mod_ht <- lm(ht ~ sex, data = data_ht)
```

```
# A tibble: 10,000 × 2
      ht sex
   <dbl> <chr>
 1  70.7 M
 2  67.9 M
 3  73.6 M
 4  68.4 M
 5  67.0 M
 6  71.2 M
 7  67.5 M
 8  68.8 M
 9  63.9 M
10  69.9 M
# ℹ 9,990 more rows
```

What Does the Model Tell Us?

How do we translate our model results into a probability model?

```
Call:
lm(formula = ht ~ sex, data = data_ht)

Residuals:
     Min       1Q   Median       3Q      Max 
-11.1388  -1.8635   0.0065   1.8557  11.6430 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 63.68069    0.04038 1576.85   <2e-16 ***
sexM         5.39268    0.05610   96.13   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.803 on 9998 degrees of freedom
Multiple R-squared:  0.4803,    Adjusted R-squared:  0.4803 
F-statistic:  9242 on 1 and 9998 DF,  p-value: < 2.2e-16
```

Why It's Important to Think of Regression as a Probability Model

Conceptualizing linear regression as a probability model allows us to generalize the ideas of linear regression to a larger number of probability distributions than just the normal distribution.

It opens up the world of **Generalized Linear Models**, which we will become more familiar with throughout the semester.

Regression Question

You want to understand the impact that an employee's job demands and resources have on their work engagement.

A Look at Our Simulated Data

```
# A tibble: 6 × 4
  job_demand job_res part_time   eng
       <dbl>   <dbl> <chr>     <dbl>
1      0.341 -1.14   no         1.30
2     -0.703 -1.02   no         3.39
3     -0.380 -0.575  no         1.38
4     -0.746 -0.0909 yes        6.44
5     -0.898 -0.0192 no         4.52
6     -0.335 -1.51   no         3.20
```

Estimating a Regression Model with R

```
mod_engage <- lm(eng ~ job_demand + job_res, data = data_jdr)
```

Interpreting the Model Output

What does the output below tell us about the relationships between engagement and job demands and job resources?

```r
summary(mod_engage)
```

```

Call:
lm(formula = eng ~ job_demand + job_res, data = data_jdr)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.5697  -1.7671   0.0077   1.6561  10.1450 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.80444    0.05883   64.67   <2e-16 ***
job_demand  -0.98796    0.05832  -16.94   <2e-16 ***
job_res      0.91971    0.06021   15.28   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.631 on 1997 degrees of freedom
Multiple R-squared:  0.2053,	Adjusted R-squared:  0.2045 
```

Communicating the Model Results

- While adjusting for a worker's level of job resources, for every one unit increase in job demands, worker engagement should decrease by .99 units, on average.
- While adjusting for a worker's level of job demands, for every one unit increase in job resources, worker engagement should increase by .92 units, on average.
- Overall, our model accounts (or explains) 20% of the variance in worker engagement.

Statistical Significance and Regression

Statistical significance asks the question: “If I believe the null hypothesis is true (usually no effect), what is the probability that my estimate would be this large or larger?”

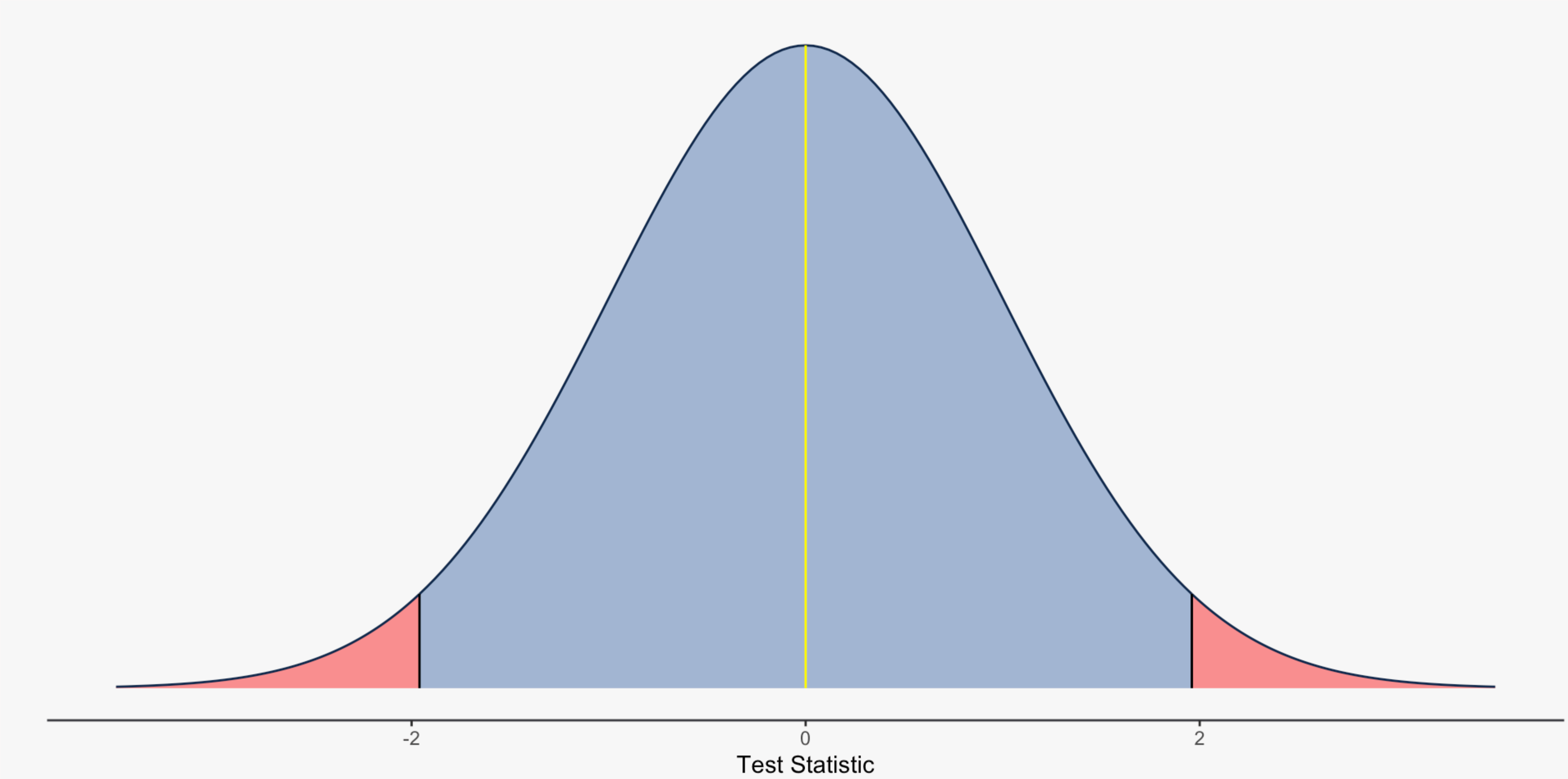
The p-value (probability value) tells us this probability and it is up to us to decide if the probability is small enough for us to reject the null hypothesis (usually if the probability is less than .05).

Standard Errors, Test Statistics, and Null Distributions

Significance testing relies heavily on the concepts of standard errors, test statistics, and null distributions:

- **Standard Errors**: Amount of uncertainty in our estimate.
- **Test Statistics**: The number of standard deviations the estimate is away from the null value.
- **Null Distributions**: The probability distribution specified by the null hypothesis.

Visualizing the Significance Test



Understanding Model Predictions and Errors (Residuals)

- **Model Prediction**:
  $3.80 + -.99 * .341 + .92 * -1.14 = 2.41$
- **Model Error**: Observed - Predicted

Calculating Model Predictions and Errors

```r
data_jdr |>
  dplyr::select(job_demand, job_res, eng) |>
  dplyr::mutate(
    prediction = predict(mod_engage),
    error = mod_engage$residuals
  )
```

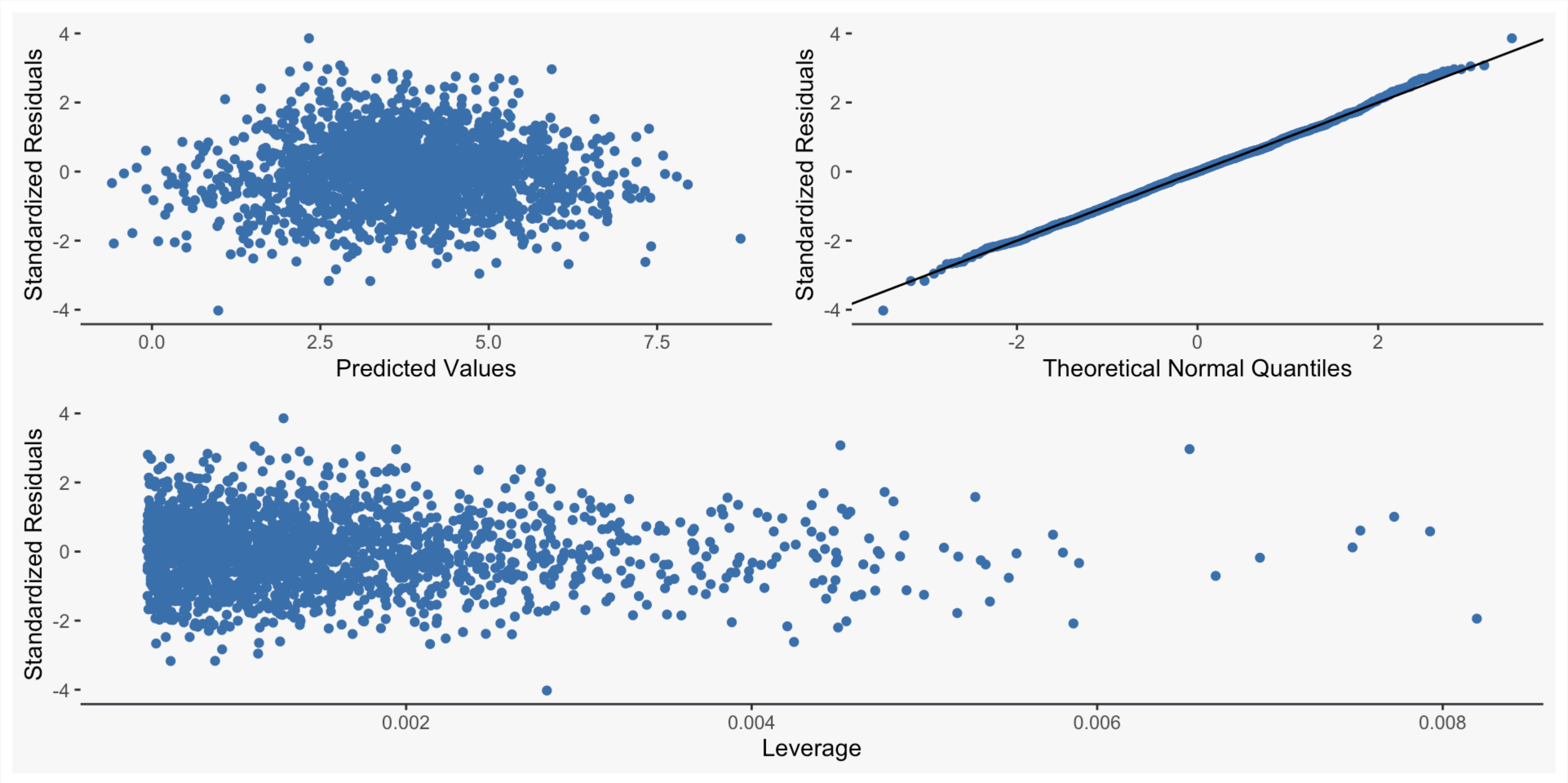
```
# A tibble: 2,000 × 5
   job_demand job_res   eng prediction  error
        <dbl>   <dbl> <dbl>      <dbl>  <dbl>
 1      0.341 -1.14    1.30       2.42 -1.12
 2     -0.703 -1.02    3.39       3.56 -0.172
 3     -0.380 -0.575   1.38       3.65 -2.28
 4     -0.746 -0.0909  6.44       4.46  1.98
 5     -0.898 -0.0192  4.52       4.67 -0.156
 6     -0.335 -1.51    3.20       2.75  0.452
 7     -0.501 -0.585   7.61       3.76  3.85
 8     -0.175 -1.76    1.13       2.36 -1.23
 9      1.81    1.39   4.99       3.30  1.70
10     -0.230  0.545   7.03       4.53  2.49
# ℹ 1,990 more rows
```

Assessing Model Fit with R-Squared

The R^2 can be calculated by squaring the correlation between our model predictions of the outcome variable and the actual values of the outcome variable.

Although it was developed for normal linear models, the R^2 can still be a helpful measure of fit for generalized linear models.

Assessing Model Diagnostics Using Residuals



Categorical Predictors and Indicator Coding

To use a categorical predictor with K groups in a regression model, you have to transform the variable into K - 1 indicator variables (variables that only take on 0 and 1 values), where the group coded as 0 is referred to as the **reference group**:

```
x3
```

```
# A tibble: 3 × 3
  Group         `Did Not Start` Incomplete
  <chr>         <chr>           <chr>
1 Completed     0               0
2 Incomplete    0               1
3 Did Not Start 1               0
```

Interpreting the Effects of Indicator Variables

For a model where the only predictor is the indicator variable:

- Intercept is the mean of the outcome variable for the reference group
- The remaining K - 1 coefficients compare the outcome variable mean for the K - 1 groups to the outcome variable mean for the reference group

Impact Part-Time Status has on Engagement

```r
mod_engage_cat <- lm(eng ~ part_time, data = data_jdr)
summary(mod_engage_cat)
```

```
Call:
lm(formula = eng ~ part_time, data = data_jdr)

Residuals:
     Min       1Q   Median       3Q      Max 
-12.6198  -1.8585   0.0857   1.9740   9.7262 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)   3.99385    0.07345  54.372  < 2e-16 ***
part_timeyes -0.95968    0.16125  -5.951 3.13e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.924 on 1998 degrees of freedom
Multiple R-squared:  0.01742,	Adjusted R-squared:  0.01693 
F-statistic: 35.42 on 1 and 1998 DF,  p-value: 3.13e-09
```

Interaction (Moderation) Effects

An interaction effect allows us to test if the impact of a predictor variable on an outcome variable changes at different levels of another predictor variable:

- The relationship between job demands and engagement is strong and negative when job resources are low, but weak, and likely non-significant, when job resources are high.
- Too Much of a Good Thing Effect (Vitamins are good for you unless you take a lot at once!)

Estimating & Interpreting Interaction Effects

```r
mod_engage_int <- lm(eng ~ job_demand * job_res, data = data_jdr)
summary(mod_engage_int)
```

```
Call:
lm(formula = eng ~ job_demand * job_res, data = data_jdr)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.6683 -1.7158  0.0427  1.6745 10.3648 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)         3.80063    0.05794  65.591  < 2e-16 ***
job_demand         -0.96718    0.05750 -16.821  < 2e-16 ***
job_res             0.92433    0.05930  15.588  < 2e-16 ***
job_demand:job_res  0.47113    0.05949   7.919 3.94e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.591 on 1996 degrees of freedom
```

Always Plot Interaction Effects