A Mixed Bag of Models

Overview

- Learn about different regression models
- Learn about statistical power and power analysis
- Learn about the effect of measurement error in predictors

Extending Linear Regression

Why Extend Linear Regression?

Our data often does not lend itself well to a linear regression model. It may violate many of the linear regression assumptions:

- Non-normal residuals (Binary or categorical outcomes)
- Correlated observations (time series data or nested data)
- Nonlinear relationships (Spline models)

Modeling Binary Outcomes with Logistic Regression

Logistic regression models essentially use a linear model to predict the odds of an event occurring:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 \boldsymbol{X}_1$$

Estimating a Logistic Regression Model

```
# A tibble: 3 × 3
  selection_test  hire hire_cat
  <chr>          <int> <chr>   
1 pass               1 hire    
2 pass               0 no hire 
3 fail               1 hire    
```

```r
mod_logistic <- glm(hire ~ selection_test, data = data, family = "binomial")
```

Interpreting a Logistic Regression Model

```r
summary(mod_logistic)
```

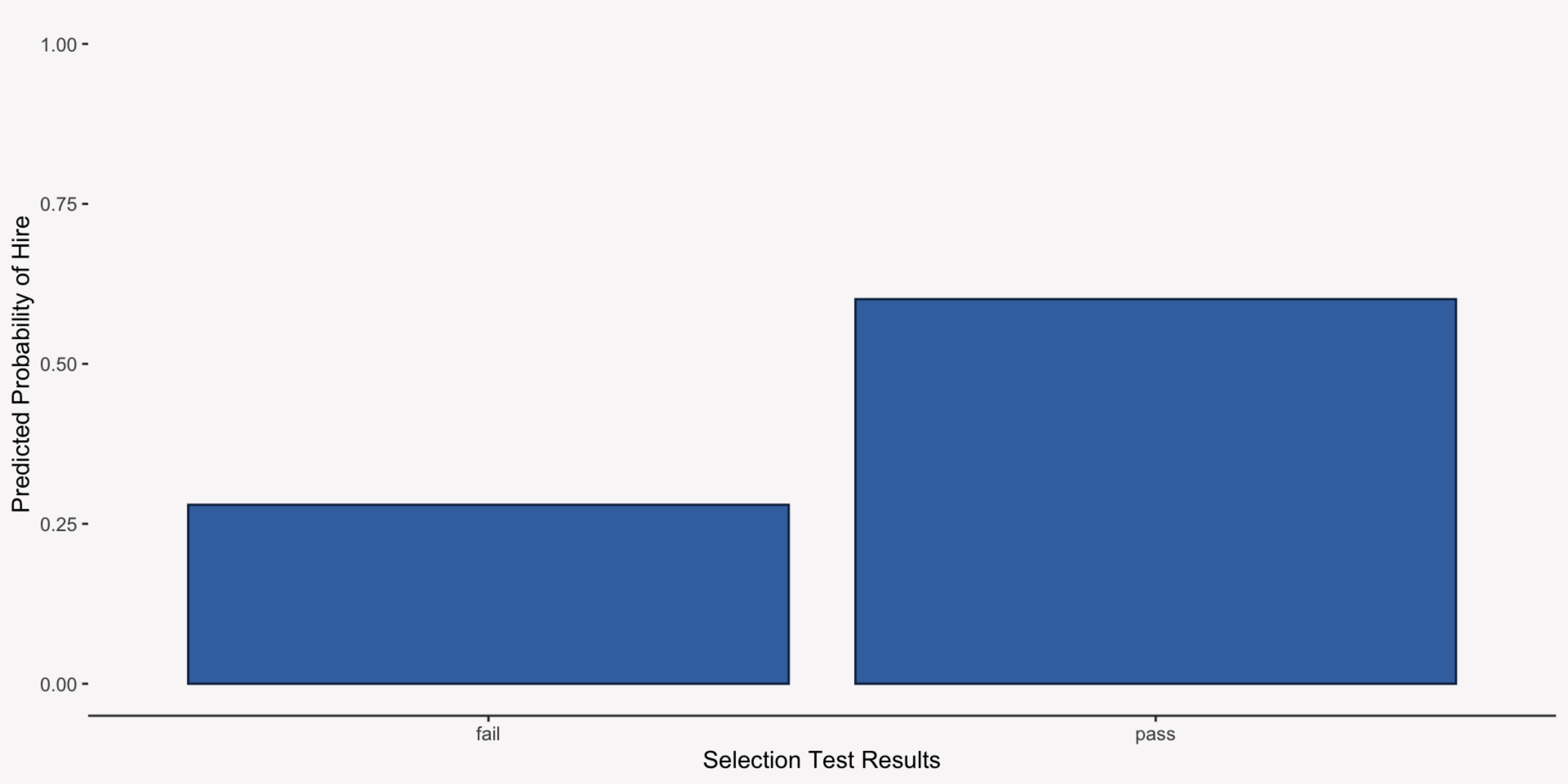
```

Call:
glm(formula = hire ~ selection_test, family = "binomial", data = data)

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)    
(Intercept)        -0.94631    0.04694  -20.16   <2e-16 ***
selection_testpass  1.35624    0.08824   15.37   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3919.4  on 2999  degrees of freedom
Residual deviance: 3675.1  on 2998  degrees of freedom
AIC: 3679.1

Number of Fisher Scoring iterations: 4
```

Interpreting a Logistic Regression Model

You can transform the coefficients from the logistic regression model to get:

- How the odds of an event occurring (e.g. being hired) change as the predictor values change
- How the probability of an event occurring change as the predictor values change

Graphing Logistic Regression Models is Best



Ordinal Logistic Regression: Ordered Multicategorical Outcome

Ordinal logistic regression models allow you to analyze categorical data when there is an inherent order to the categories:

- Survey response data
- Loan default risk (Low, Medium, High)
- Really any variables where you can create low, medium, and high categories

Estimating an Ordinal Logistic Regression Model

```
# A tibble: 5 × 2
  yr_employ default_risk
      <dbl> <fct>
1        12 high
2         8 low
3        11 high
4        14 moderate
5        15 low
```

```r
mod_ordinal <- MASS::polr(default_risk ~ yr_employ, data = data_ord)
```

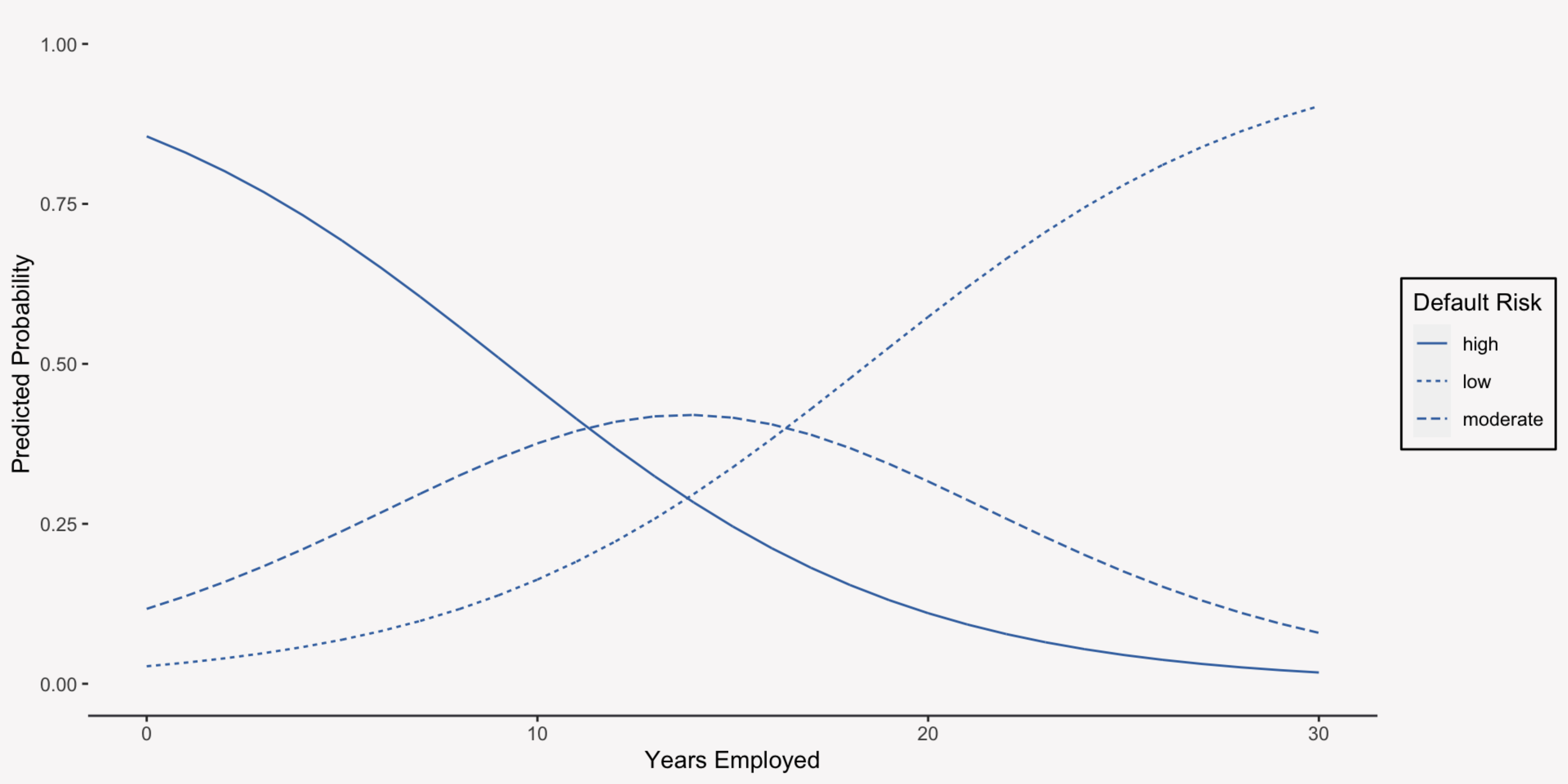
Interpreting an Ordinal Logistic Model

The coefficient tells you how the odds of moving up in a category change for a unit increase in your predictors. A negative value means that higher values on the predictor are associated with lower odds of being in a higher category.

```r
summary(mod_ordinal)
```

```
Call:
MASS::polr(formula = default_risk ~ yr_employ, data = data_ord)

Coefficients:
            Value Std. Error t value
yr_employ -0.1933   0.006338   -30.5

Intercepts:
              Value    Std. Error t value
low|moderate   -3.5709   0.1032   -34.5992
moderate|high  -1.7790   0.0927   -19.1852

Residual Deviance: 9844.334
AIC: 9850.334
```

Plotting the Results of an Ordinal Regression Model



Multilevel Models

Multilevel models allow you to model data that are nested in higher units like employees nested within an organization.

When units are nested in higher-level clusters, your data will likely violate the independence of error assumption–units within the same cluster will be more related to one another than units in a different cluster.

Testing Effects at Different Levels

Multilevel models also allow you to test effects at the unit (level 1) level or the cluster level (level 2):

- Impact of employee engagement on individual sales (level 1) and the impact of office location on individual sales (level 2)
- Impact of parental education on a student's achievement (level 1) and the impact of teacher quality on students' achievement (level 2)

Estimating a Multilevel Model

```
# A tibble: 10 × 4
   ind_perf ind_sat cohesion  team
      <dbl>   <dbl>    <dbl> <int>
 1    3.23   1.01     -1.02      1
 2    2.29  -0.809    -1.02      1
 3    1.99  -0.387    -1.02      1
 4    4.14   1.36     -1.02      1
 5    0.213  0.0215   -1.02      1
 6    3.29  -0.558     0.484     2
 7    4.43   1.03      0.484     2
 8    4.01  -0.394     0.484     2
 9    5.04   0.202     0.484     2
10    5.07   1.34      0.484     2
```

```r
mod_mlm <- lme4::lmer(ind_perf ~ ind_sat + cohesion + (1|team), data = data_mlm)
```

Interpreting a Multilevel Model

```r
summary(mod_mlm)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: ind_perf ~ ind_sat + cohesion + (1 | team)
   Data: data_mlm

REML criterion at convergence: 15033.5

Scaled residuals: 
    Min      1Q  Median      3Q     Max 
-3.4760 -0.6363 -0.0007  0.6450  3.6050 

Random effects:
 Groups   Name        Variance Std.Dev.
 team     (Intercept) 0.2271   0.4766  
 Residual             1.0158   1.0079  
Number of obs: 5000, groups:  team, 1000

Fixed effects:
            Estimate Std. Error t value
```

Time Series Models

Time series models are statistical models that have been developed specifically to analyze data that follow a time ordering:

- Financial data (Daily/Monthly changes in stock prices)
- Personalized medicine (Changes in an individuals blood pressure over time)
- Political data (Changes in congressional approval through time)

Autoregressive Time Series Model

A standard time series model is the AR model, where the outcome variable at Time T is predicted by its value at Time T-1:

$$\boldsymbol{Y}_t = \beta_1 \boldsymbol{Y}_{t-1} + \epsilon_t$$

Statistical Power

What is Power?

The **power** of a statistical test is the probability of obtaining a significant result **given** that the effect actually exists.

Typically, you want the power of your test to be 80% or higher.

Decisions: Power, Type 2 Error, and Type 1 Error

	H_0 is True	**H_0 is True**
Test Rejects H_0	α	1 - β
Test Doesn't Reject H_0	1 - α	β

- **Power** is represented as 1 - β
- **Type 1** error is represented as α
- **Type 2** error is represented as β

Factors Affecting the Power of Any Statistical Test

There are several things that impact the power of all statistical tests:

- **N**: Sample size
- **Effect Size**: Size of the effect of interest
- α**-level**: The significance cut-off (e.g. .05)

Factors Affecting the Power of Regression Slope Test

We can use the formula of a statistical test and the SE of a regression slope to see the factors that directly impact the power of a regression slope test:

$$\text{Statistical Test} = \frac{\beta}{SE_\beta}$$

$$SE_{\beta_j} = \sqrt{\frac{\sigma_e^2}{N \times Var(X_j) \times Tol(X_j)}}$$

What is a Power Analysis?

A power analysis is an analysis undertaken before you begin your study that helps you determine the minimum sample size you need to achieve a certain amount of power–typically 80%.

As we will see, a power analysis is full of a fair amount of educated guessing…

Information Needed to Conduct a Power Analysis

To conduct a power analysis you generally need the following pieces of information:

- The size of the effect of interest (e.g. R^2 or β)
- The significance level being used (e.g. $\alpha = .05$)
- Desired level of power

For multiple regression, you will also need to know the number of predictors being used in the model.

Multiple Regression Power Analysis

There are a few ways to conduct a power analysis for a multiple regression model. You can decide to look at the power to the individual regression coefficients significant or the power to find the R^2 significant.

You will likely need to write a simulation script to determine the sample size needed to find an individual regression coefficient significant, but we can use a package called `pwr` to determine the sample size needed to find the overall R^2 significant.

Example of Power Analysis Using `pwr`

```r
pwr::pwr.f2.test(u = 6, f2 = .3 / (1 - .3), sig.level = .05, power = .80)
```

- f2 = $\frac{R^2}{1 - R^2}$ (Effect Size)
- u = The number of predictors
- v = N - u - 1
- N = v + u + 1

Changing the Significance Level

```r
pwr::pwr.f2.test(u = 6, f2 = .3 / (1 - .3), sig.level = .05, power = .80)
```

```
     Multiple regression power calculation 

              u = 6
              v = 31.57285
             f2 = 0.4285714
      sig.level = 0.05
          power = 0.8
```

```r
pwr::pwr.f2.test(u = 6, f2 = .3 / (1 - .3), sig.level = .10, power = .80)
```

```
     Multiple regression power calculation 

              u = 6
              v = 24.86152
             f2 = 0.4285714
      sig.level = 0.1
          power = 0.8
```

Changing the Effect Size

```r
pwr::pwr.f2.test(u = 6, f2 = .3 / (1 - .3), sig.level = .05, power = .80)
```

```

     Multiple regression power calculation 

              u = 6
              v = 31.57285
             f2 = 0.4285714
      sig.level = 0.05
          power = 0.8
```

```r
pwr::pwr.f2.test(u = 6, f2 = .5 / (1 - .5), sig.level = .05, power = .80)
```

```

     Multiple regression power calculation 

              u = 6
              v = 13.86892
             f2 = 1
      sig.level = 0.05
          power = 0.8
```

Measurement Error in Predictors

What is Measurement Error?

Measurement error is the random noise that affects all of our measurements:

- Mental measurements (intelligence, personality, etc.)
- Health measurements (weight, blood pressure, etc.)
- Economic measurements

Observed Score, True Score, & Measurement Error Variance

Measurement theory defines your **observed score** (X) as your **true score** (T) plus **random measurement error** (E):

$$X_{Obs} = T + E$$

$$Var(X_{obs}) = \sigma^2_T + \sigma^2_E$$

What is the Typical Effect of Measurement Error?

Measurement error in either the outcome or predictor will **weaken** the power of your statistical tests.

Measurement error in the predictor will bias most of your statistical estimates such that the estimate you obtain will be even further than the parameters it is estimating.

How Do We Quantify Measurement Error?

To quantify measurement error, we calculate a ratio of true score variance to observed score variance, which is referred to as the measurement's **reliability**:

$$r_{xx} = \frac{\sigma^2_T}{\sigma^2_{Obs.}}$$

How Do We Estimate Reliability?

There are many different ways to estimate the reliability of a measure, but in the social sciences the most common way is through a coefficient called Coefficient Alpha.

All you need to know right now is that survey measurements should have Coefficient Alpha values equal to or greater than .80.

Measurement Error and Simple Regression

Measurement error in the predictor will bias the estimate of the regression slope by a factor equal to the reliability of the predictor, r_{xx}:

$$b = \beta \times r_{xx}$$

This will also serve to weaken the power of the statistical test as any measurement error shrinks the regression slope.

Measurement Error and Multiple Regression

The effects of measurement error on multiple regression are very similar to its effects in simple regression, but predicting the exact effects are harder if the predictors are correlated with one another.

Can We Correct for Measurement Error?

If we know the amount of measurement error (reliability) in our measures, then we can use some corrections to adjust our statistical estimates (e.g. $\frac{b}{r_{xx}} = \beta$).

However, the best way to adjust for measurement error is to use a more complex family of statistical models called **latent variable models**. Latent variable models explicitly model and correct for measurement error.