PSY 3307

Regressions

Jonathan A. Pedroza PhD

Cal Poly Pomona

2021-11-30

We're at the Finish Line

Agenda (1/2)

- The difference between correlation and regression
- The linear model
- Basic linear regression
- Assessing fit
- Assessing individual predictors
- Bias in regression
- Assumptions in linear regression
- Sample size in regression
- Simple regression in SPSS
- Interpretation of the overall model
    - Model parameters

Agenda (2/2)

- Multiple linear regression
  - hierarchical regression (not hierarchical modeling)
  - stepwise methods/forced entry
- Comparing models
- Multicollinearity
- Multiple regression in SPSS
- Robust regression (Bootstrapping)
- Interpretation of multiple regression
- How to report multiple regression

There are a variety of different regression techniques

- linear regression
- logistic regression
- negative-binomial regression
- multinomial regression
- ridge regression
- lasso regression
- elastic net regression
- spatial regression
- quantile regression
- poisson regression
- structural equation modeling

- mixed effect model/multi-level model

- these also all have different estimation types (which we are not getting into)

For this class

- we are only focusing on (multiple) linear regression
- if we can get to it (interactions in linear regression)

Holy Smokes!

- the book only covers linear regression in 3 pages!

Difference between Correlation and Regression

- Both focus on the strength of a relationship
    - regression focuses more on the direction of the relationship
- correlation only states if the two variables are positively or negatively related
    - there is a relationship present
- regression is scale dependent in that coefficients are the expected change on average in y given a one-point/unit increase in X
    - for a one point increase in X, there is a _ increase/decrease in Y
- That being said, a standardized regression coefficient in a simple linear regression is the same thing as a correlation coefficient

```
##
##      Pearson's product-moment correlation
##
## data:  jp$tv and jp$smartphone
## t = 5.4083, df = 370, p-value = 0.0000001144
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1737730 0.3623755
## sample estimates:
##       cor
## 0.2706695
```

```
## lm(formula = tv ~ smartphone, data = jp)
##             coef.est coef.se t value Pr(>|t|)
## (Intercept) 1.88     0.61    3.09    0.00
## smartphone  0.46     0.09    5.41    0.00
## ---
## n = 372, k = 2
## residual sd = 2.43, R-Squared = 0.07
```

```
##
##     Pearson's product-moment correlation
##
## data:  jp$tv and jp$smartphone
## t = 5.4083, df = 370, p-value = 0.0000001144
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1737730 0.3623755
## sample estimates:
##       cor
## 0.2706695
```

```
##
## Call:
## lm(formula = tv ~ smartphone, data = jp)
##
## Standardized Coefficients::
## (Intercept)  smartphone
##   0.0000000   0.2706695
```

Simple Linear Model

- while these models are similar in that we are looking at one IV and one DV, there are some additional components to a linear regression
    - we must note that the regression includes the *unstandardized* measure of the relationship
- we are looking at the relationship between X and Y with a parameter (`b1`) that quantifies the relationship between X and Y
- additionally, we also have b0 (**the intercept**), which is the value of the outcome when your IV is at zero

Simple Linear Model

$$Y_i = mx + b$$

$$Y_i = a + bX_i + \epsilon$$

- the equation is the equation of a straight line
- the straight line can be defined as two things
    - the slope of the line (b1)
    - the point at which the line crosses the vertical (y) axis of the graph (b0 or intercept)

Simple Linear Regression

$$Y_i = b_0 + b_1 X_i + \epsilon_i$$

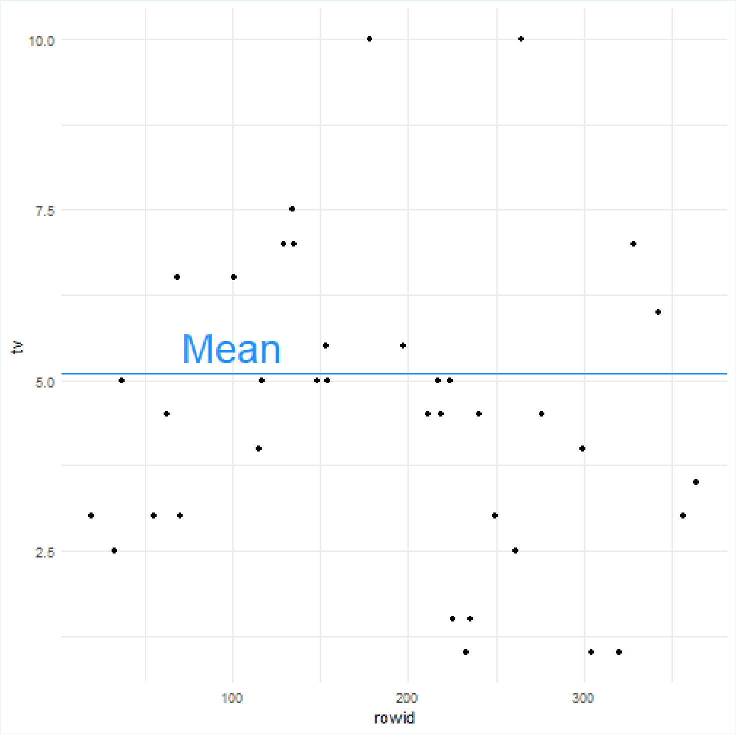
- IV = predictor variable/indicator variable
  - `X` in the equation
- DV = outcome variable/criterion variable
  - `Y` in the equation
- `b0` is the y-intercept (we'll get to this shortly)
- `b1` is the slope of the association between X and Y
- `e` is the error/residual of what is unexplained in our model

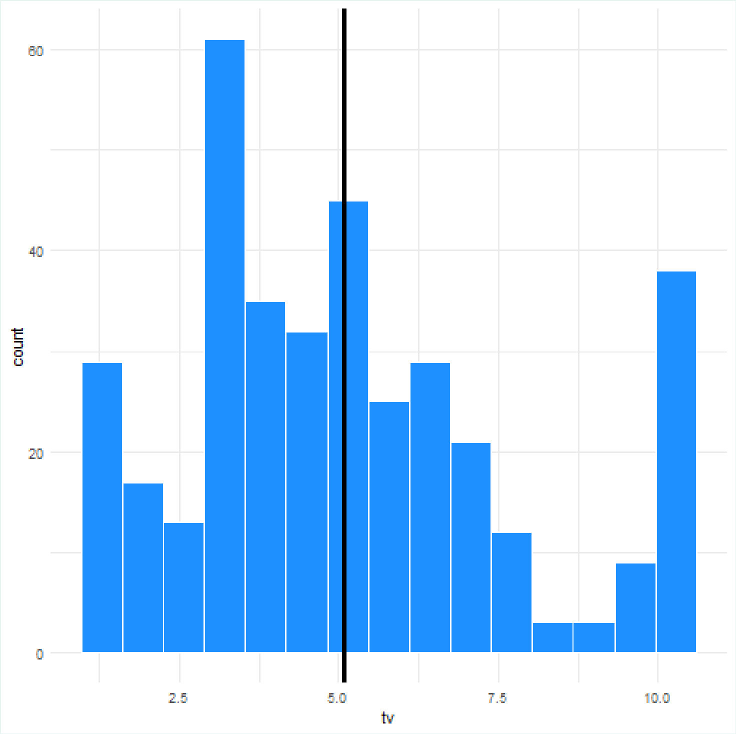
Regression Scores

- in simple and multiple linear regressions, you will have actual scores and predicted scores
    - **actual scores** are values that participants answered in your study/survey/experiment
    - **predicted values** are values that are predicted on the regression line
    - values that fall on the regression line

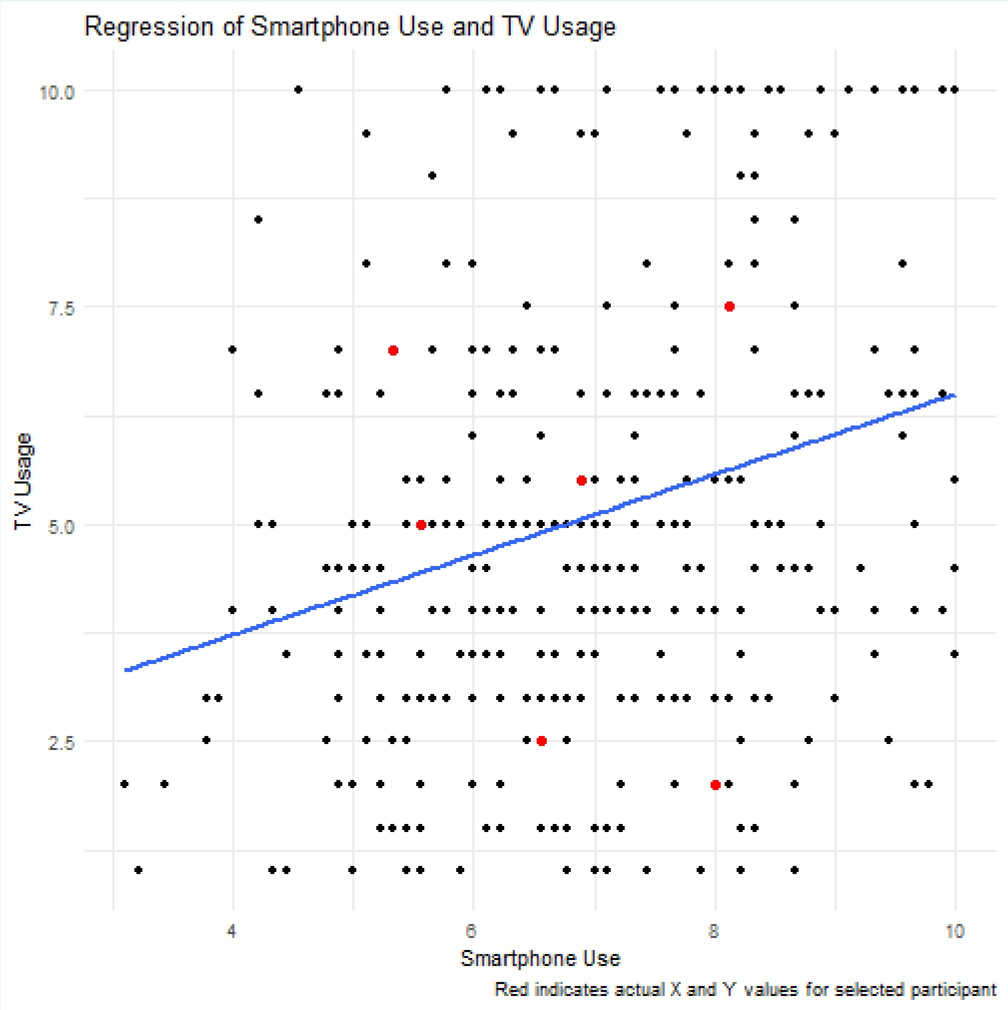
Actual Values

$$Y = actual\ scores$$



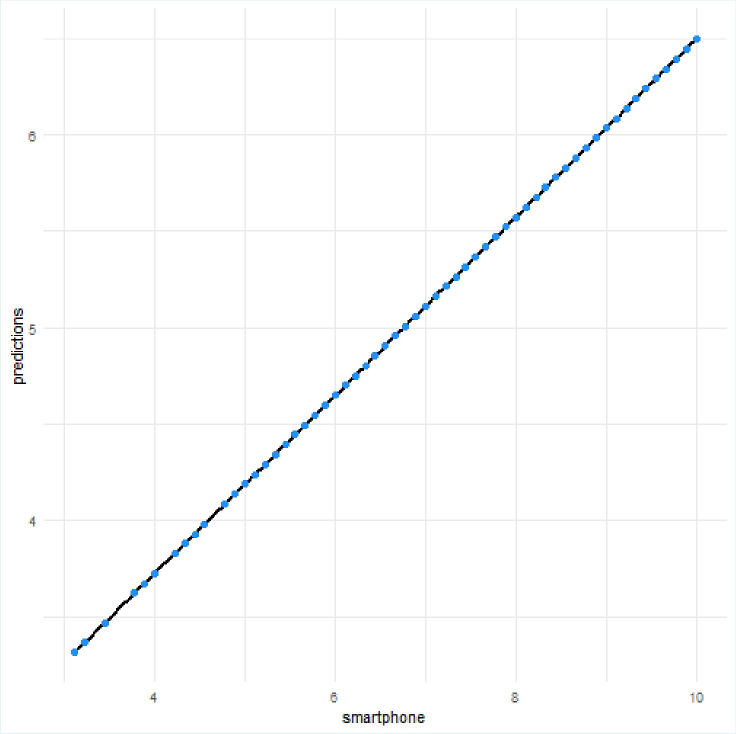


More Actual Values



Predicted Values

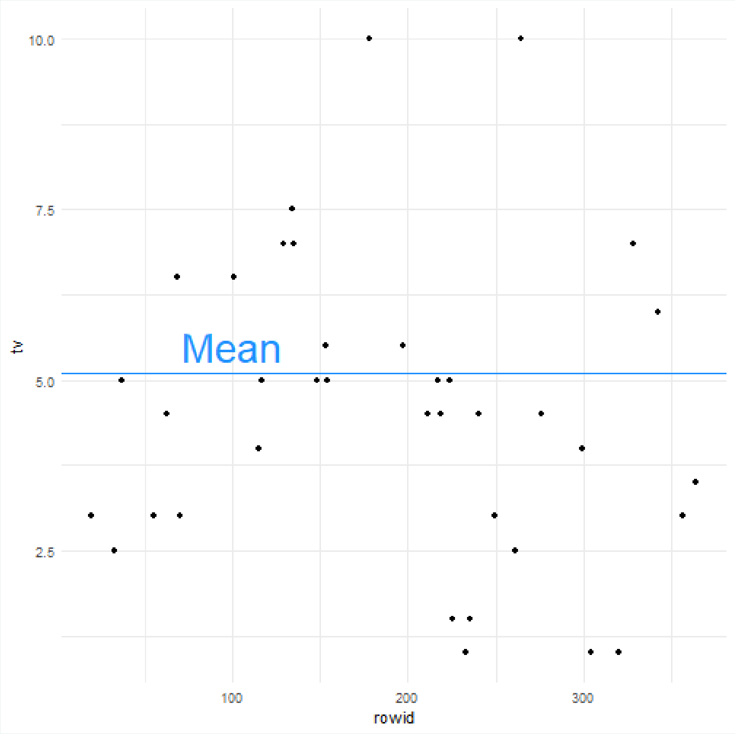
$$\hat{Y} = predicted\ scores$$

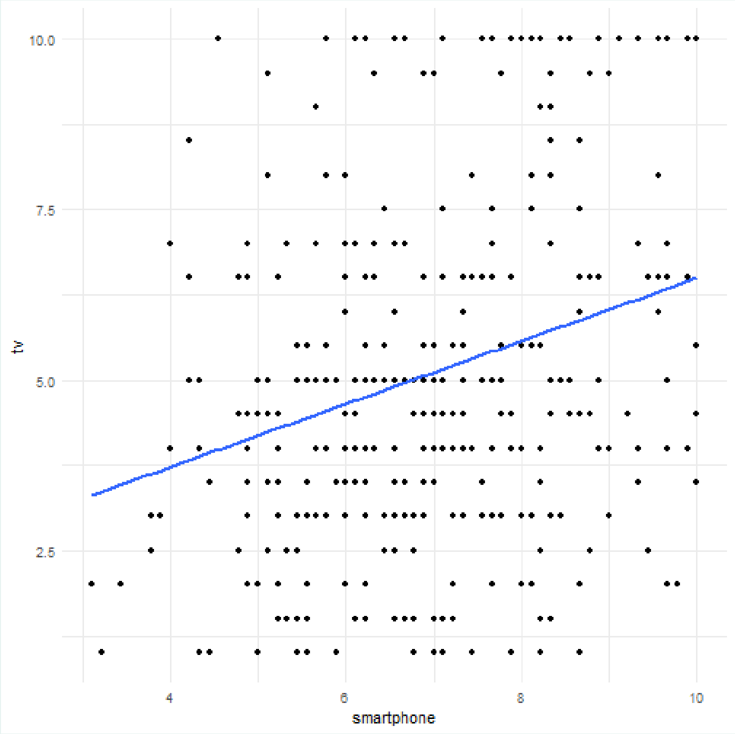


Deviation & Residuals

- when estimating the model, there is always some error left
  - in regressions, the difference between what the model predicts and what the actual scores are is referred to as our **residual**
- this is similar when we looked at the mean and we saw that participants deviated from the mean

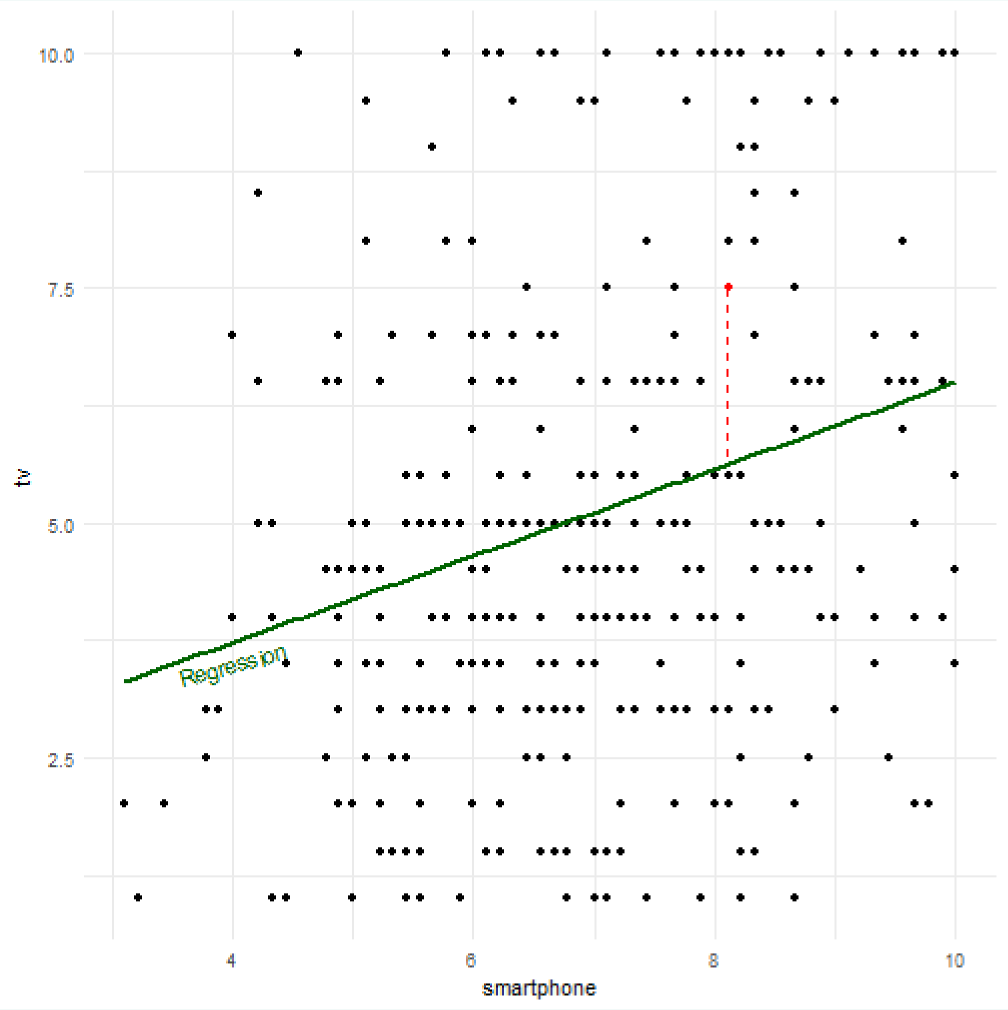
Examples of Deviation & Residuals





What is Left in Our Model

$$Total\ Error = \Sigma(observed_i - model_i)^2$$



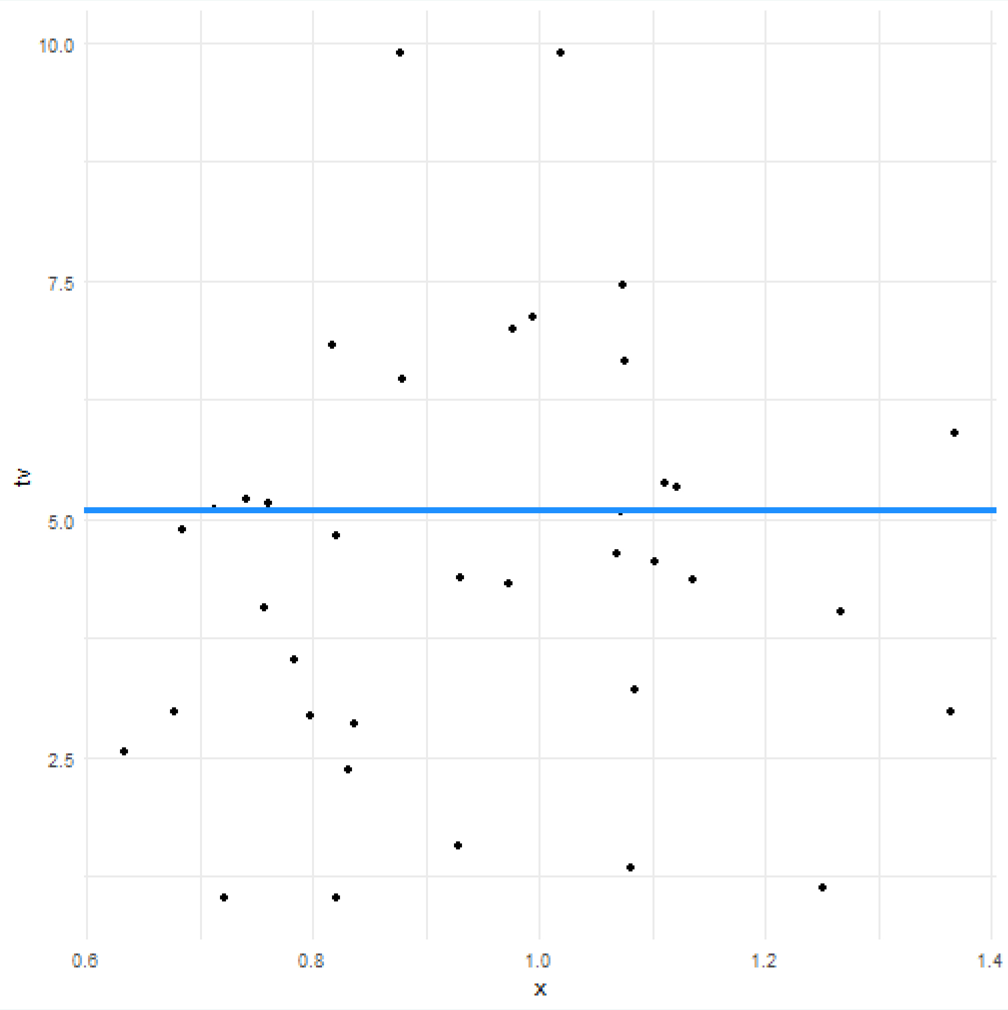
10.0
7.5
5.0
2.5
tv
Regression
4
6
8
10
smartphone

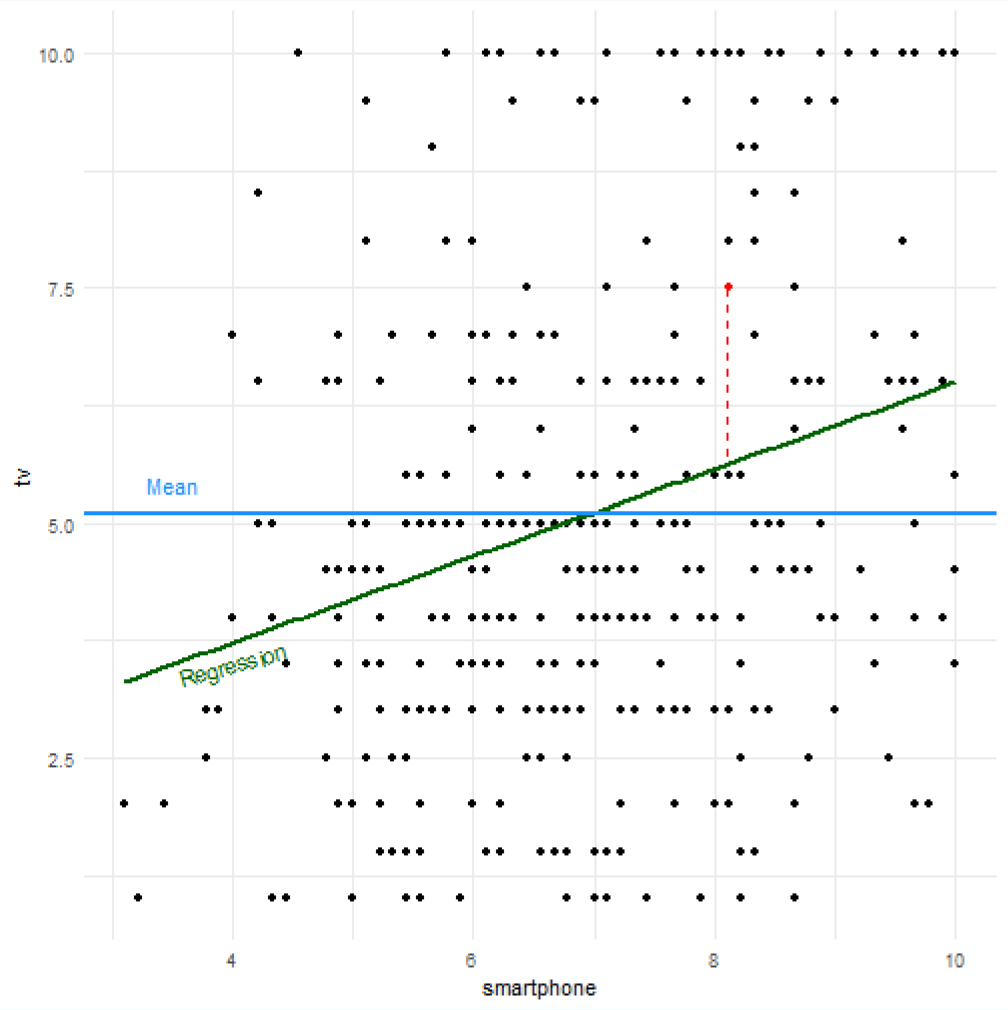
Residuals in Regression

- to assess how much error/residual/unknown in our regression model, we use a sum of squared errors
  - in the regression framework, we refer to this as **sum of squared residuals** or **residual sum of squares**
- this tells us how well our regression line fits the data
  - larger sum of squared residuals = regression not representative of the data
  - smaller sum of squared residuals = more representative regression
- the method of least squares to estimate the parameters where the sum of squared residuals is lowest is known as **ordinary least squares (OLS)** regression

Goodness of Fit & R2

- **total sum of squares** is the sum of squared differences
  - it uses the differences between the observed/actual data and the mean value of your outcome
- **sum of squared residuals** shows the differences between the observed/actual data and the regression line
- **model sum of squares** shows the differences between the mean value of your outcome and the regression line


10.0
7.5
5.0
2.5
tv
0.6
0.8
1.0
1.2
1.4
x


10.0
7.5
5.0
2.5
tv
Mean
Regression
4
6
8
10
smartphone

Goodness of Fit & R2

- similar to ANOVA, we can then get the amount of variation accounted for by our model using the model sum of squares

$$R^2 = \frac{model\ sum\ of\ squares}{total\ sum\ of\ squares}$$

$$R^2 = \frac{SS_M}{SS_T}$$

Goodness of Fit & R2

- our F statistic is based on the improvement in our model and the difference between the model and the observed/actual data
  - because our sum of squares values depend on the number of differences we have added up, we rely on our meqan squares values
- the **F-ratio** is the measure of how much the model has improved in the relationship/association/prediction of the outcome compared to how inaccurate your model was

$$F = \frac{MS_M}{MS_R}$$

Goodness of Fit & R2

- if you wanted to know if your R2 value was statistically significant, then you use the following formula, where:
  - `N - k - 1` is your degrees of freedom
  - `k` is the number of predictors/IVs

$$F = \frac{(N - k - 1)R^2}{k(1 - R^2)}$$

Assessing Individual Predictors

- every variable has its own slope (b)
- hypothesis testing is in the form of a t-statistic

Individual Predictors

- t-statistic tests whether the value of b is different from zero
  - H0: b is zero
  - H1: b is significantly different from zero

$$t = \frac{b_{observed} - b_{expected}}{SE_b}$$

- since our null of our expected b is zero the formula then becomes

$$t = \frac{b_{observed}}{SE_b}$$

- df is `N - k - 1` for multiple regression, simple regression is `N - 2`

Individual Predictors

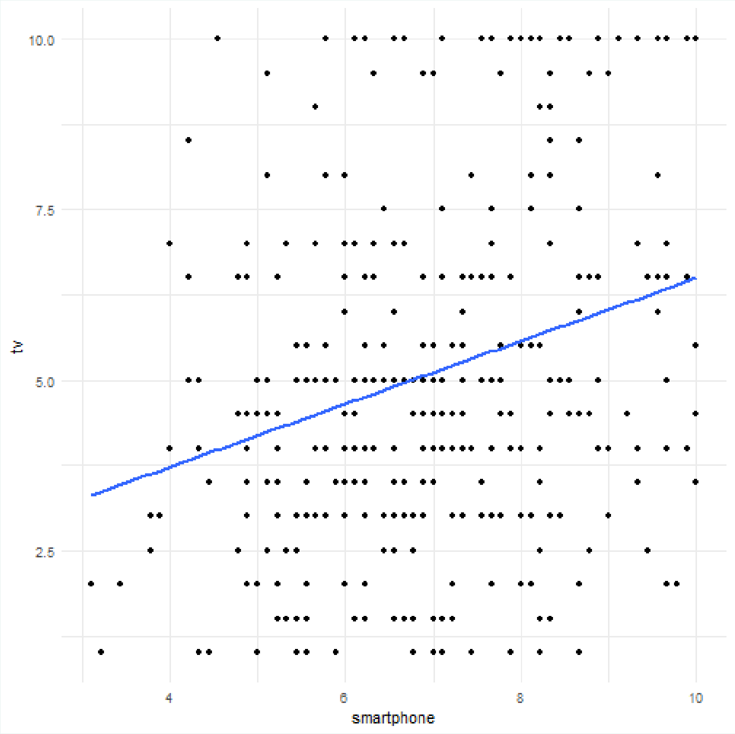
- while we use unstandardized regression coefficients to explain the association/relationship between IV and DV
    - **standardized regression coefficients** are useful for seeing the strength of the association
    - they are not however true values for effect size
    - they are z-transformed so their values should range from 0-1 but if you have IVs that are severely correlated your standardized regression coefficients can be over 1
- additionally, remember that R2 is the amount of variance accounted for in your outcome by your IV(s)

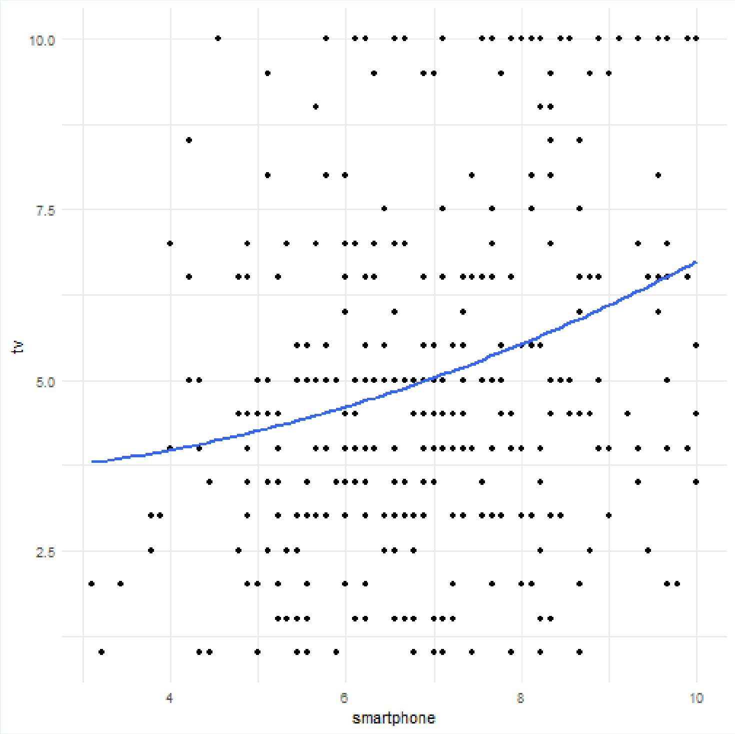
Bias in Regression Models

- we want to make sure our data can be generalizable to other samples
    - we'll do this by making sure our data is not biased by unusual cases and by diagnosing our model

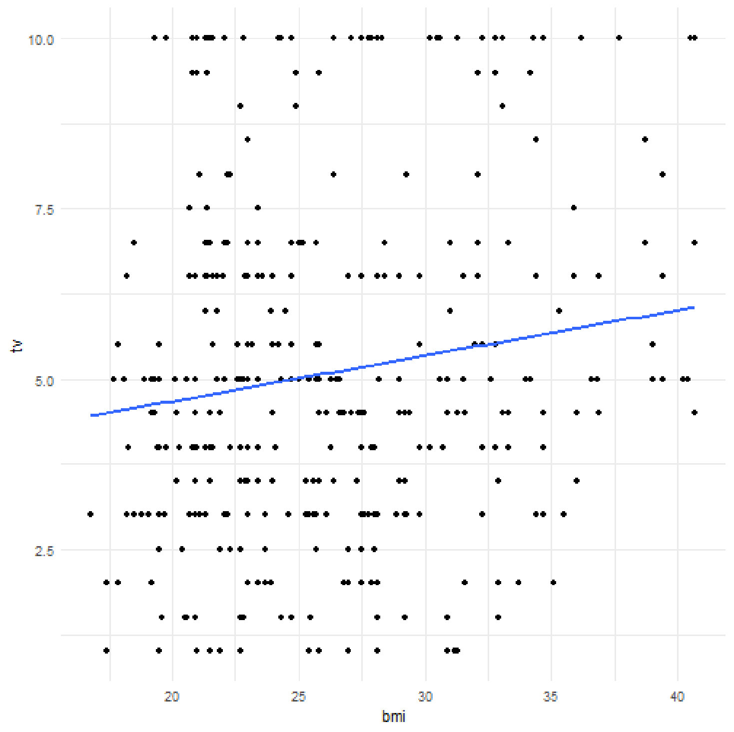
Linearity

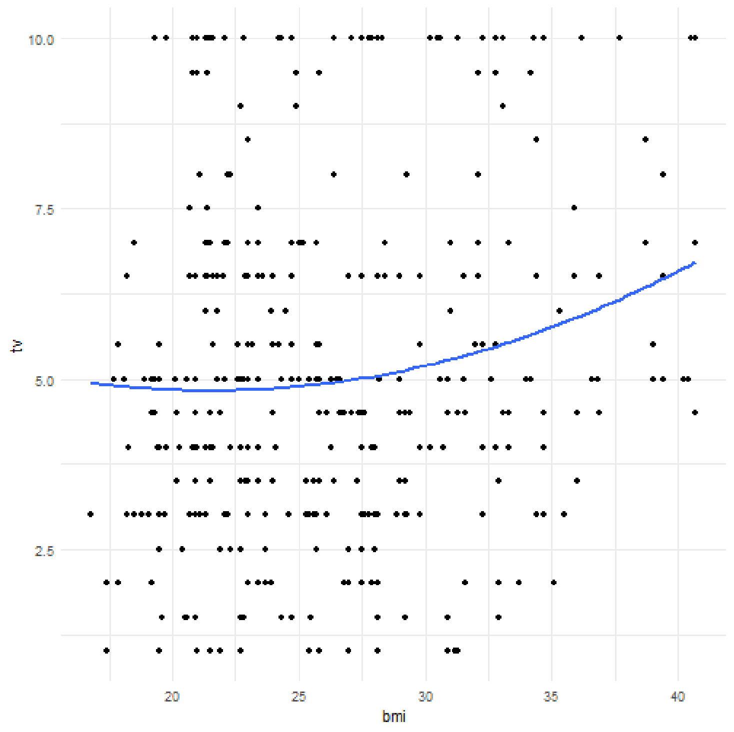
- you should have a linear relationship between every IV and your DV





Linearity pt 2





Outliers

- can detect univariate (one variable at a time) outliers using histograms/boxplots
- can detect bivariate (two variables at a time) outliers with scatterplots
- if there are severe outliers think about either deleting them
- using Cook's distance is influence of cases on the model
    - some state over |1| could be influential
- Leverage is the influence of observed value on the outcome across the predicted values
    - influential is a value 2-3x grater than the average value
- Mahalanobis distance
    - distance from the mean (highest = bad)

Outliers

- could delete extreme 5% of tails of the scores
- could delete values +-3 SD from mean

- "Winsorizing" replace the outlier with +-3 SD value

- JP Note: don't touch it if it could be a valid case
    - or run the model with the outliers and without the outliers to see if they are influential

Homoscedasticity

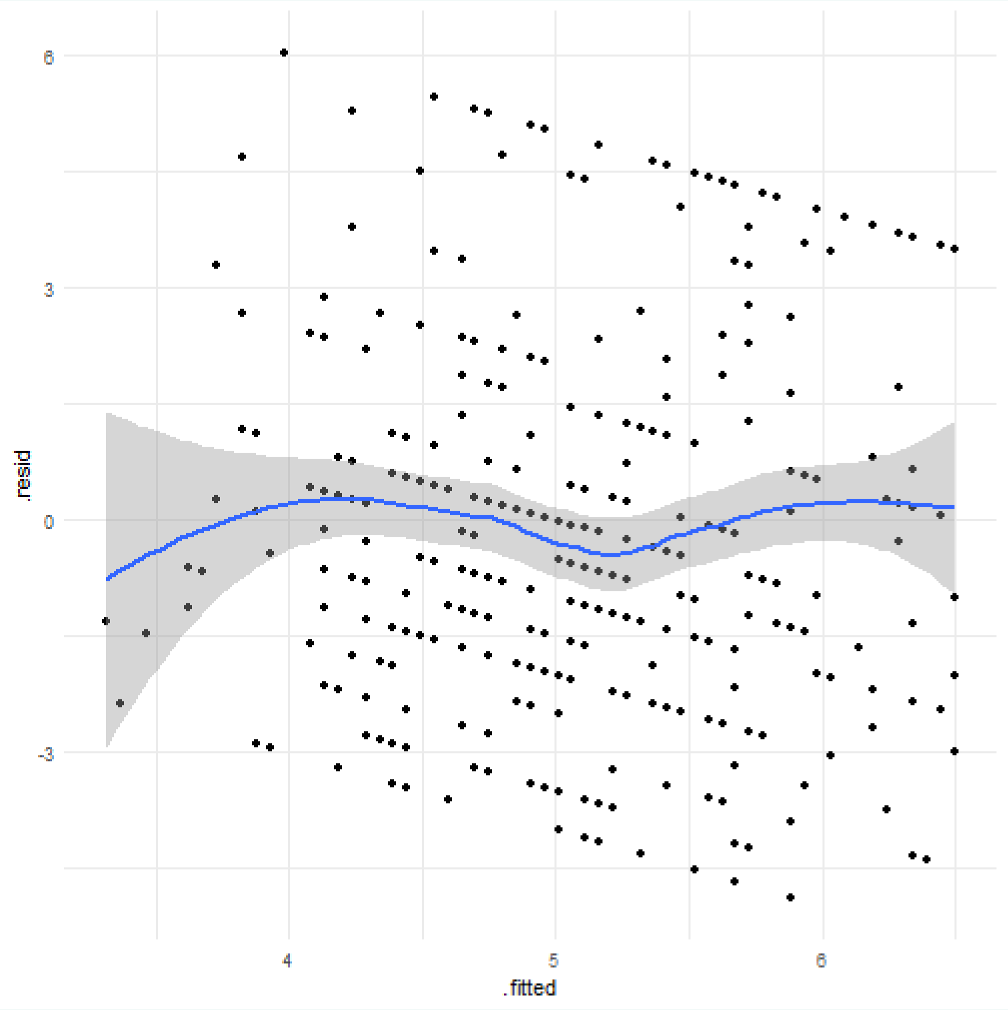
- **homoscedasticity** is when the error has constant variance across the values of the IVs
- **heteroscedasticity** is when variance changes across values of the IVs
- you can put some trust in the Levene's test, which we want to be nonsignificant
    - states that the variance is equal across the values of the IVs

Independence

- residual terms should not be related to one another
- can also be tested through the **Durbin-Watson test**
  - examines if adjacent residuals are correlated
- if you violate independence, which could be easily violated with the same variables collected over multiple years
  - multi-level modeling or include a control variable, such as year may remove the violation

Independence of Residual Errors

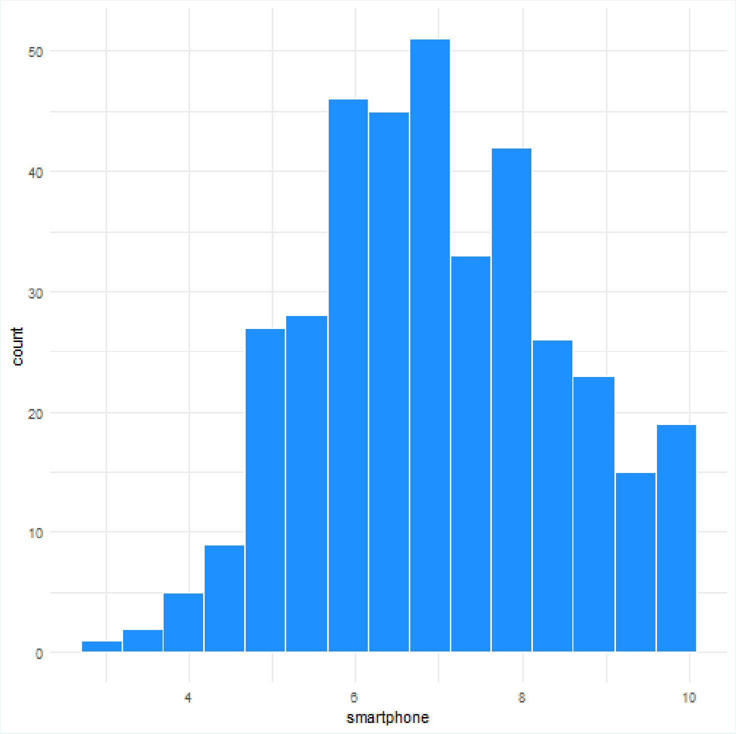
- expect to see no relationship between our fitted values (predictions) and our residuals (distance away from the regression line)

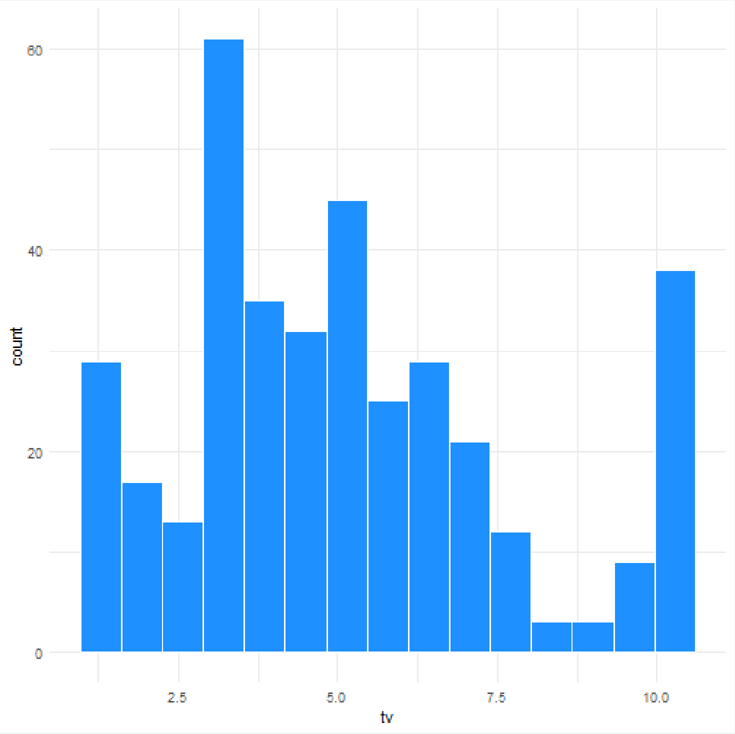

6
3
0
-3
.resid
4
5
6
.fitted

Normality

- histograms of residuals
- p-p plots (probability - probability)
- q-q plots (quantile - quantile)
- normality test statistics
  - Kolmogorov-Smirnov
  - Shapiro-Wilk
- if non-normal --> transform
  - makes it more difficult to interpret

Noramlity (Univariate)





Residuals

- **unstandardized** are in the original measurement units of the outcome
  - difficult to use when comparing across models
- **standardized** are unit free residuals because they are z-scores
  - can compare across models (>3 is problematic)
  - in standard deviation units
  - assumes equal variance across values of IVs
- **studentized** are unit free residuals which are unstandardized residuals divided by an estimate of its SD that varies from point-to-point
  - doesn't assume equal variances across values of IVs
  - often the best option

Sample Size in Regression

- bigger sample will always be better
- be aware of how many IVs you include in your model
    - you should have at least 20 (preferably more) participants per IV included in your model
    - realistically you should have much more for your sample size, this is the bare minimum

Multiple Linear Regression

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \epsilon_i$$

- in addition to everything from our simple linear regression
- `b2` is the slope of our second IV
- `X2` is the second variable in our model
- useful for controlling for other variables and examining the unique association/relationship between your IV of interest and DV

More Complex Linear Regression

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + b_n X_{ni} + \epsilon_i$$

Different Types of Multiple Regression

- hierarchical (sequential)
  - IVs are entered in steps/blocks based on theoretical knowledge
  - could also be in the form of block 1: control variables, block 2: IVs of interest, block 3: possible interactions
- simultaneous (standard)
  - all IVs are entered together
- automated regression
  - let the computers do everything for you in choosing IVs
  - **do not use** because there is no theory with this method
  - simplistic way of predictive modeling/machine learning/simply put…

Skynet

Model Comparisons

- we may be interested in comparing two multiple regression models
  - these models must be nested
- to put it simply **nested** models are when models contain all the same variables, with the second model containing additional variables
- good way to see if adding additional variables made your model better/account for more variation in your outcome
- compares model by using ANOVA

Model Comparisons

```
## Analysis of Variance Table
##
## Model 1: tv ~ smartphone
## Model 2: tv ~ smartphone + bmi
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    370 2189.5
## 2    369 2147.9  1    41.555 7.1389 0.007877 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## [1] 0.017589
```

- the finding tells us that the model with bmi and smartphone use is a significantly better fitting model than the model with only smartphone use
- similarly, we can see the corresponding change in R2 values
  - the addition of BMI helped account for .02 or 2% more variability in TV viewing

Akaike Information Criteria

```
## [1] 1721.071
```

```
## [1] 1715.943
```

```
## [1] -5.128224
```

- complicated fit criteria but to keep it simple, lower AIC = better fitting model
  - penalizes model for having more variables
- comparing these AIC values is interpretable
  - Recommendations by Burnham and Anderson (2002)

Multicollinearity, VIF, & Tolerance

- **multicollinearity** is when one IV correlates strongly with another IV ($r > .7$)
- **variance inflation factor (VIF)** is when an IV has a strong linear relationship with one or more IV(s)
  - VIF > 10 = concern in the model, diagnose it for multicollinearity
  - average VIF > 1 there may be bias in model
- **tolerance** is similar to VIF in that tolerance = 1/VIF
  - tolerance below .1 is a serious problem
  - tolerance below .2 may indicate bias in model

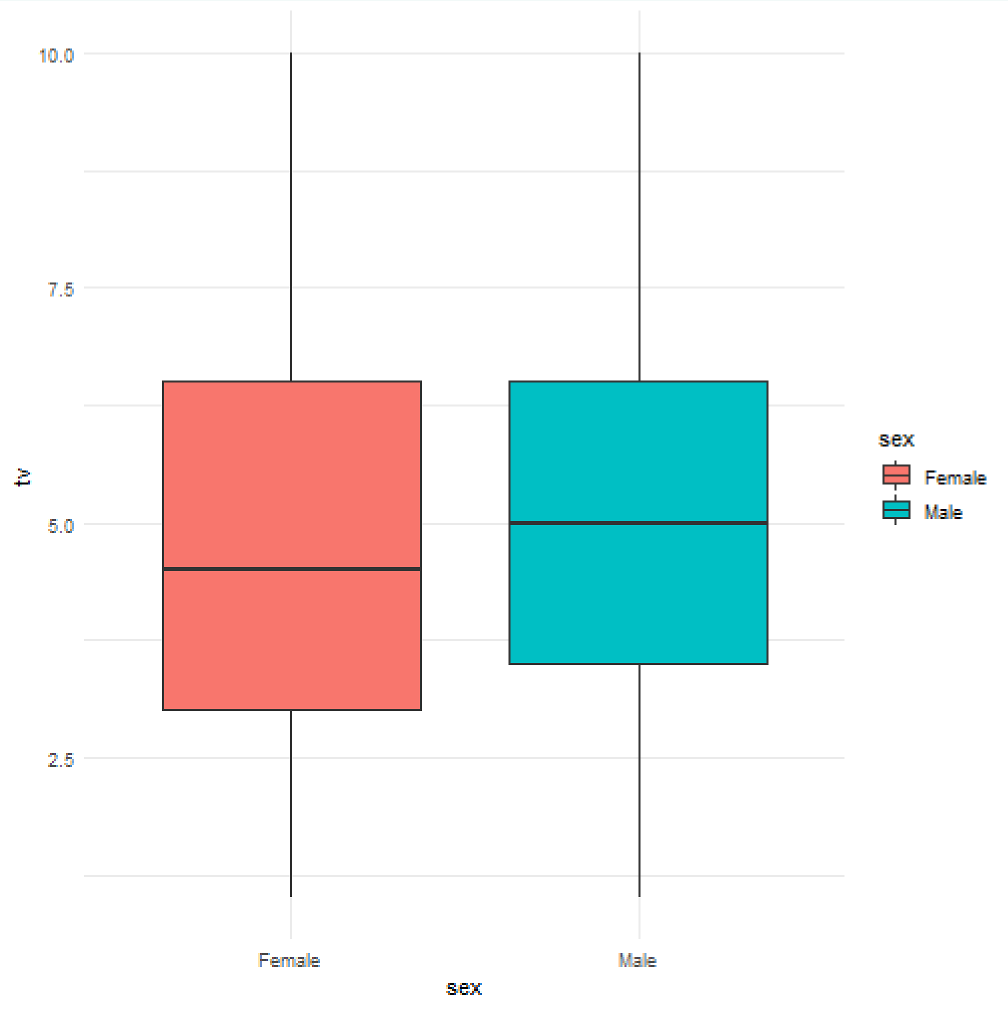
Simple Linear Regression w/ Categorical IV

$$Y_i = b_0 + b_1 X_i + \epsilon_i$$

$$tv = b_0 + b_1(male) + \epsilon_i$$

- comparing males and females in their tv scores is the equivalent of a one-way ANOVA
  - IV = sex
  - levels (male/female)
  - DV = tv scores

Example



```
## lm(formula = tv ~ male, data = jp)
##             coef.est coef.se t value Pr(>|t|)
## (Intercept)  5.03     0.16   31.85    0.00
## male         0.20     0.28    0.70    0.49
## ---
## n = 372, k = 2
## residual sd = 2.53, R-Squared = 0.00
```

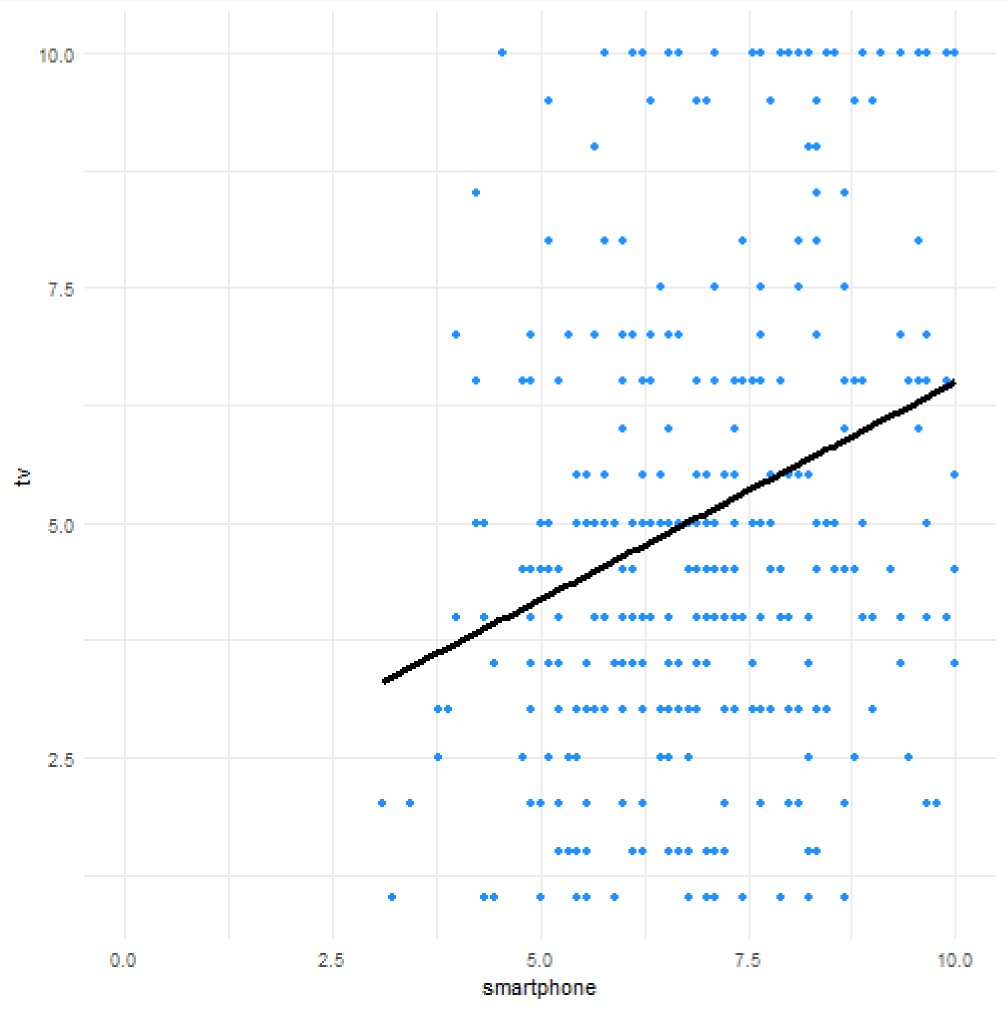
- the intercept is the average tv value for female participants
  - the average tv viewing score was 5.03 for female participants
- the slope is the difference in tv values comparing males to females
  - the average tv viewing score for males was 0.20 higher compared to females

Simple Linear Regression with Continuous IV

$$Y_i = a + bX_i + \epsilon$$

$$tv = b_0 + b_1(smartphone) + \epsilon_i$$

Example



```
## lm(formula = tv ~ smartphone, data = jp)
##             coef.est coef.se t value Pr(>|t|)
## (Intercept) 1.88     0.61    3.09    0.00
## smartphone  0.46     0.09    5.41    0.00
## ---
## n = 372, k = 2
## residual sd = 2.43, R-Squared = 0.07
```

- the intercept (`b0`) is 1.88 or the point at which the regression line hits the y axis
  - the average TV value when smartphone use is at zero
- the slope of the association between smartphone use and tv (`b1`) is 0.46

Predictions - by hand

$$tv = 1.88 + 0.46 * smartphone$$

- this is the prediction/association/relationship for any one participant
- if a participant used their smartphone all the time (`10 on the MTUAS scale`), what would we expect for their TV usage

$$tv = 1.88 + 0.46 * 10$$

```
.46*10
```

```
## [1] 4.6
```

$$tv = 1.88 + 4.6$$

```
4.6 + 1.88
```

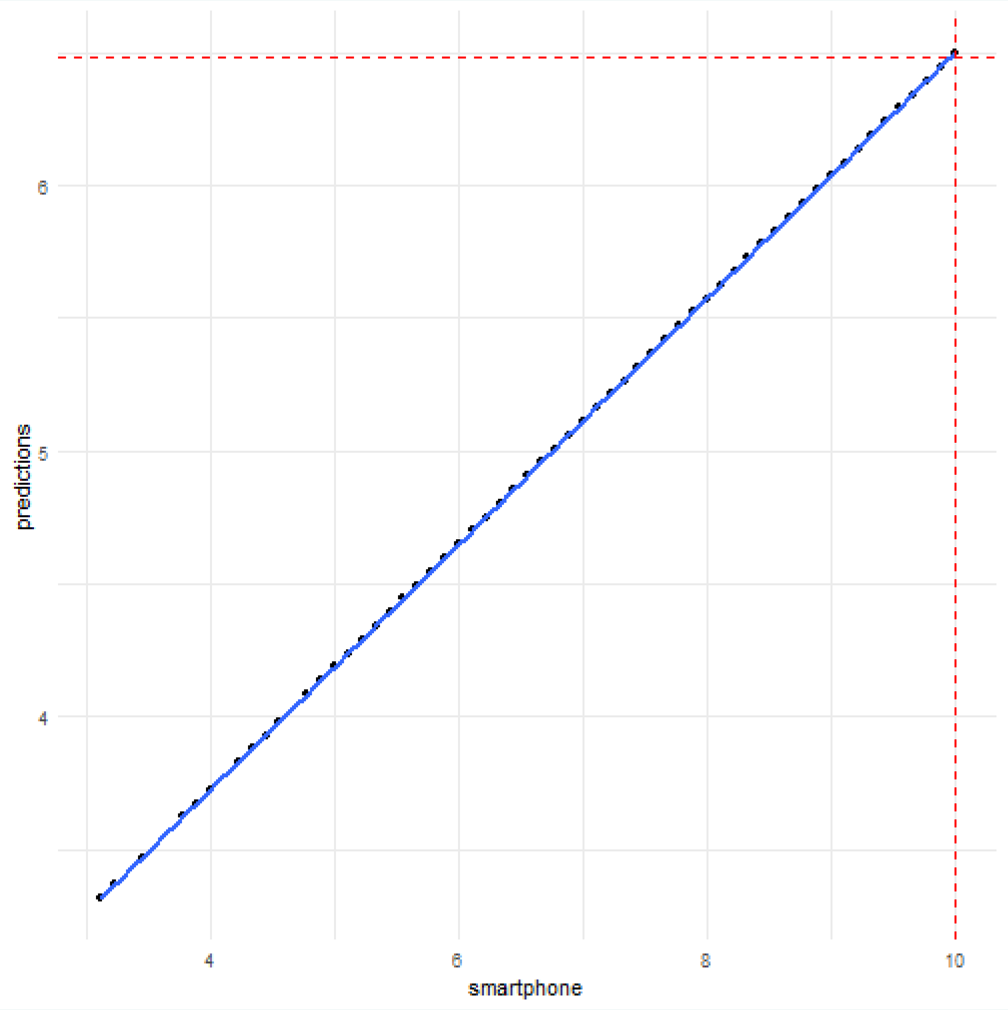
```
## [1] 6.48
```

$$tv = 6.48$$

Predictions - through R or SPSS

- From this, we can get the predictions from our model

```
## [1] 5.265329 5.624638 6.035276 4.187401 5.983947 4.957349
```



predictions
6
5
4
4
6
8
10
smartphone

Several combinations of Multiple Regressions

1. all continuous IVs
2. all categorical IVs
3. continuous and categorical IVs

Multiple continuous IVs

```
## lm(formula = tv ~ smartphone + bmi, data = jp)
##             coef.est coef.se t value Pr(>|t|)
## (Intercept) 0.31     0.84    0.37    0.71
## smartphone  0.45     0.08    5.34    0.00
## bmi         0.06     0.02    2.67    0.01
## ---
## n = 372, k = 3
## residual sd = 2.41, R-Squared = 0.09
```

$$Y_i = a + b_1 X_{1i} + b_2 X_{2i} + \epsilon$$

$$tv = b_0 + b_1(smartphone) + b_2(BMI) + \epsilon_i$$

- the intercept (`b0`) is 0.31 or the point at which the regression line hits the y axis
    - the average TV value when smartphone use and BMI are both at zero
- the slope of the association between smartphone use and tv (`b1`) is 0.45 when BMI is held constant
    - a one unit increase in smartphone use is associated with a 0.45 average increase in tv viewing when BMI is held constant
- the slope of the association between BMI and tv (`b2`) is 0.06 when smartphone use is held constant
    - a one unit increase in BMI is associated with a 0.06 increase in tv viewing when smartphone use is held constant

Multiple categorical IVs

```
## lm(formula = tv ~ male + latino, data = jp)
##             coef.est coef.se t value Pr(>|t|)
## (Intercept)  5.32     0.22   23.99    0.00
## male         0.22     0.28    0.79    0.43
## latino      -0.50     0.27   -1.88    0.06
## ---
## n = 372, k = 3
## residual sd = 2.52, R-Squared = 0.01
```

$$Y_i = a + b_1 X_{1i} + b_2 X_{2i} + \epsilon$$

$$tv = b_0 + b_1(male) + b_2(Latino) + \epsilon_i$$

- the intercept is 5.32 or the point at which the regression line hits the y axis
    - the average tv viewing value when sex is female and race is non-Latino
- the slope is the difference in tv values comparing males to females when holding race constant
    - males watch 0.22 more tv than females when holding race constant
- the slope is the difference in tv values comparing Latinos and non-Latinos when holding sex constant
    - Latinos watch 0.50 less tv than non-Latinos when holding sex constant

Continuous and Categorical IVs

```
## lm(formula = tv ~ smartphone + male, data = jp)
##             coef.est coef.se t value Pr(>|t|)
## (Intercept) 1.53     0.64    2.41    0.02
## smartphone  0.49     0.09    5.65    0.00
## male        0.48     0.28    1.75    0.08
## ---
## n = 372, k = 3
## residual sd = 2.43, R-Squared = 0.08
```

$$Y_i = a + b_1 X_{1i} + b_2 X_{2i} + \epsilon$$

$$tv = b_0 + b_1(smartphone) + b_2(male) + \epsilon_i$$

- the intercept is 1.53 or the point at which the regression line hits the y axis
    - the average tv value when smartphone use is zero and sex is female
- the slope of the association between smartphone use and tv (`b1`) is 0.49 when sex is male/held constant
    - a one unit increase in smartphone use is associated with a 0.49 average increase in tv viewing when sex is male/held constant
- the slope is the difference in tv values comparing males to females when holding smartphone use constant
    - males watch 0.48 more tv than females when smartphone use was held constant

Calculating Predictions Practice

```
## lm(formula = snacks ~ depression, data = jp)
##              coef.est coef.se t value Pr(>|t|)
## (Intercept)  2.16     0.12    17.93   0.00
## depression   0.15     0.09     1.75   0.08
## ---
## n = 372, k = 2
## residual sd = 0.72, R-Squared = 0.01
```

For the model findings above, what would you predict is the level of snack eating in a participant with a depression score of 3?

$$snacks = b_0 + b_1(depression)$$

$$snacks = 2.16 + 0.15(depression)$$

```
## lm(formula = tv ~ smartphone + bmi, data = jp)
##             coef.est coef.se t value Pr(>|t|)
## (Intercept) 0.31     0.84    0.37    0.71
## smartphone  0.45     0.08    5.34    0.00
## bmi         0.06     0.02    2.67    0.01
## ---
## n = 372, k = 3
## residual sd = 2.41, R-Squared = 0.09
```

For the model findings above, what would you predict is the TV viewing score for a participant with a smartphone use score of 5 and a BMI of 30?

$$tv = b_0 + b_1(smartphone) + b_2(BMI)$$

$$tv = 0.31 + 0.45(smartphone) + 0.06(BMI)$$

```
## lm(formula = video_game ~ bmi + depression + snacks, data = jp)
##             coef.est coef.se t value Pr(>|t|)
## (Intercept) 1.53     0.84    1.83    0.07
## bmi         0.02     0.02    1.04    0.30
## depression  0.26     0.29    0.90    0.37
## snacks      0.47     0.17    2.74    0.01
## ---
## n = 372, k = 4
## residual sd = 2.36, R-Squared = 0.02
```

For the model above, what would you predict is the video gaming score for a participant with a BMI of 27, a depression score of 1.5, and a snacking score of 3?

$$video\ games = b_0 + b_1(BMI) + b_2(depression) + b_3(snacks)$$

$$video\ games = 1.53 + 0.02(BMI) + 0.26(depression) + 0.47(snacks)$$