LINEAR REGRESSION

Applied Analytics: Frameworks and Methods 1

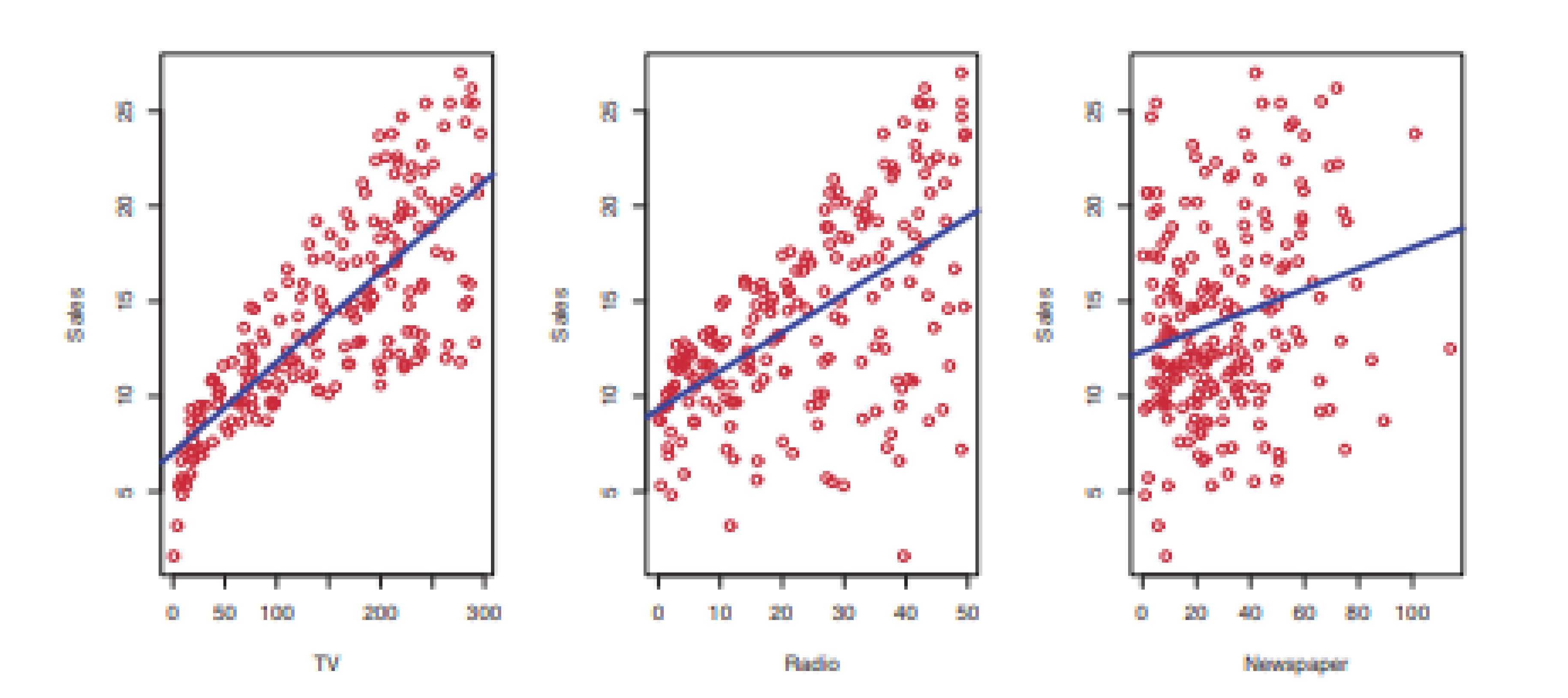
Outline

- About Regression
- Mechanics of Estimation
- Prediction and Inference
- Regression Models using Wages Data
- Regression Assumptions

Linear Regression

- Oldest, most basic predictive modeling (or supervised learning) technique
- Yet, it remains a useful tool for predicting a numerical outcome and continues to be widely used
- Many modern machine learning approaches are generalizations or extensions of linear regression

Consider this Advertising Data



Source: James et al (2017), Introduction to Statistical Learning with Applications in R

Questions Regression May Answer

- Is there a relationship between advertising budget and sales?
- How strong is the relationship between advertising budget and sales?
- Which media contribute to sales?
- How accurately can we predict future sales?
- Is the relationship linear?
- Is there synergy among the advertising media?

Regression

1. Estimate Regression Equation
2. Prediction
3. Inference

Let's begin by examining the estimation process

MECHANICS OF ESTIMATION

Estimate Regression Equation

- Estimate parameters of the population regression equation
- $Y = \beta_0 + \beta_1 X + \varepsilon$
  - *where X is the predictor,*
  - *Y is the outcome,*
  - β_0 *and* β_1 *are regression coefficients*
  - ε *is random error*
- Coefficients estimated using an optimization procedure like Ordinary Least Squares (OLS)
  - *Construct a linear combination of predictors such that* $\Sigma e_i = 0$ *and* Σe_i^2 *is minimum*
- Next few slides will illustrate this optimization process using an example.

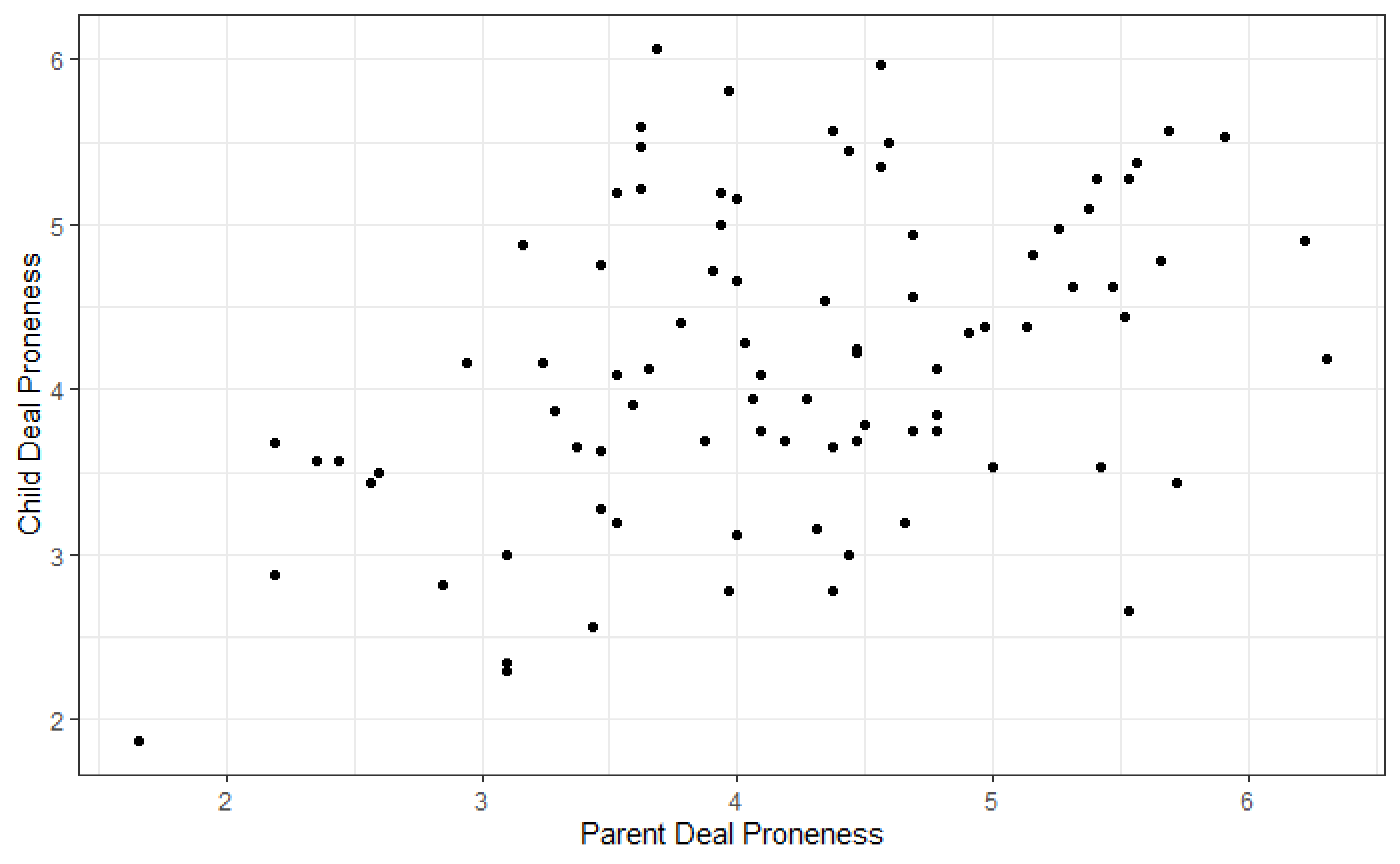
Example

- Deal proneness is the tendency of shoppers to buy products that that offer a good deal such as coupon discounts, sales and buy-one get-one free offers.
- Does deal proneness of parents affect deal proneness of children?
- Schindler, Lala, and Grussenmeyer (2014) gathered data on deal proneness of parents and their children using a 32-item scale for deal proneness. The scores were averaged to construct an index.

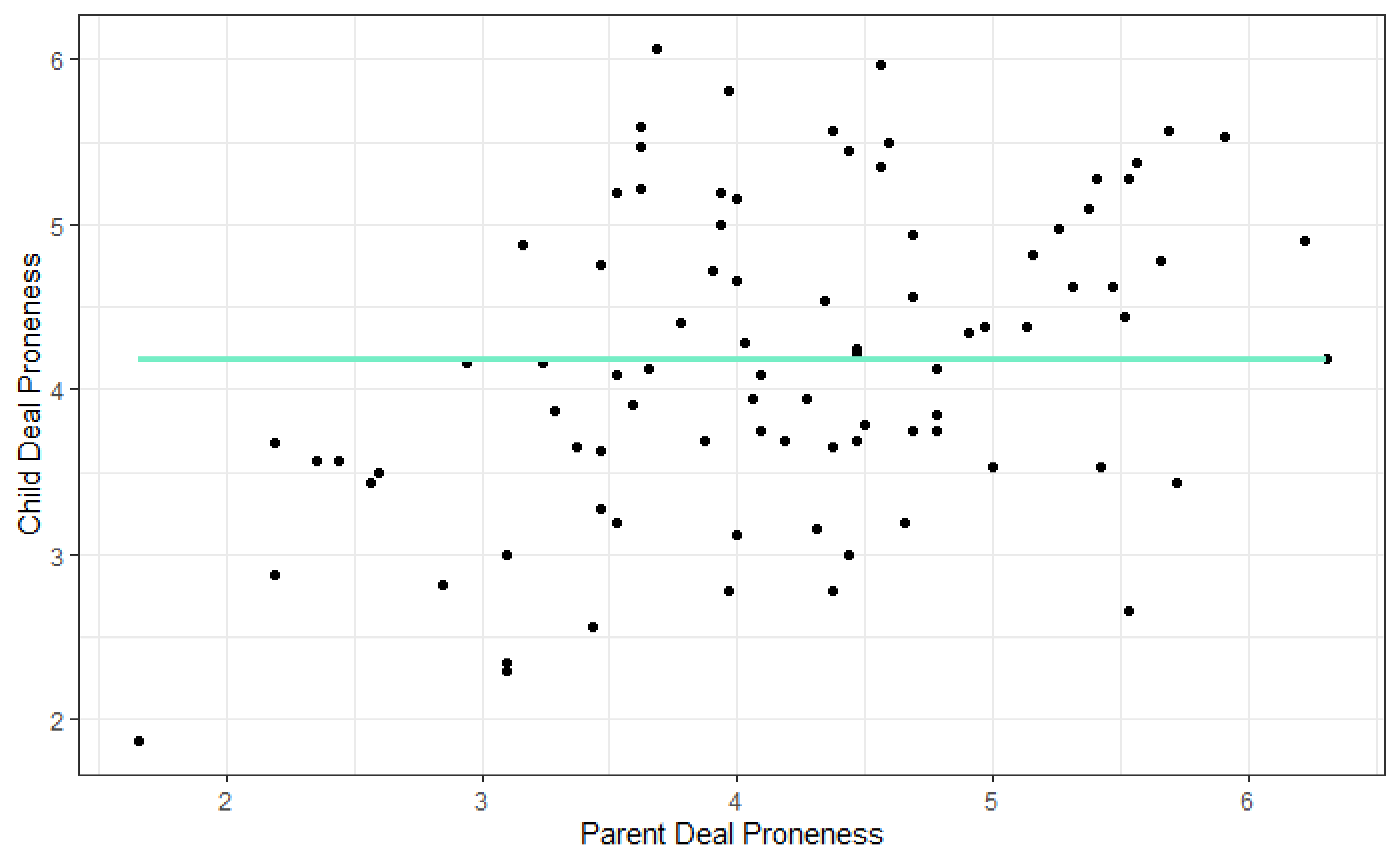
Source: Schindler, Robert. M., Vishal Lala, and Colleen Corcoran (2014). "Intergenerational Influence in Consumer Deal Proneness," Psychology & Marketing, 31 (5), 307-320

id	Parent (X)	Child (Y)
1	5.0	3.5
2	3.9	5.0
3	5.5	4.6
4	3.4	2.6
5	3.6	5.6
6	5.9	5.5
7	2.6	3.5
8	5.7	3.4
9	4.4	2.8
10	4.1	3.9
..	..	..
..	..	..

Scatter Plot

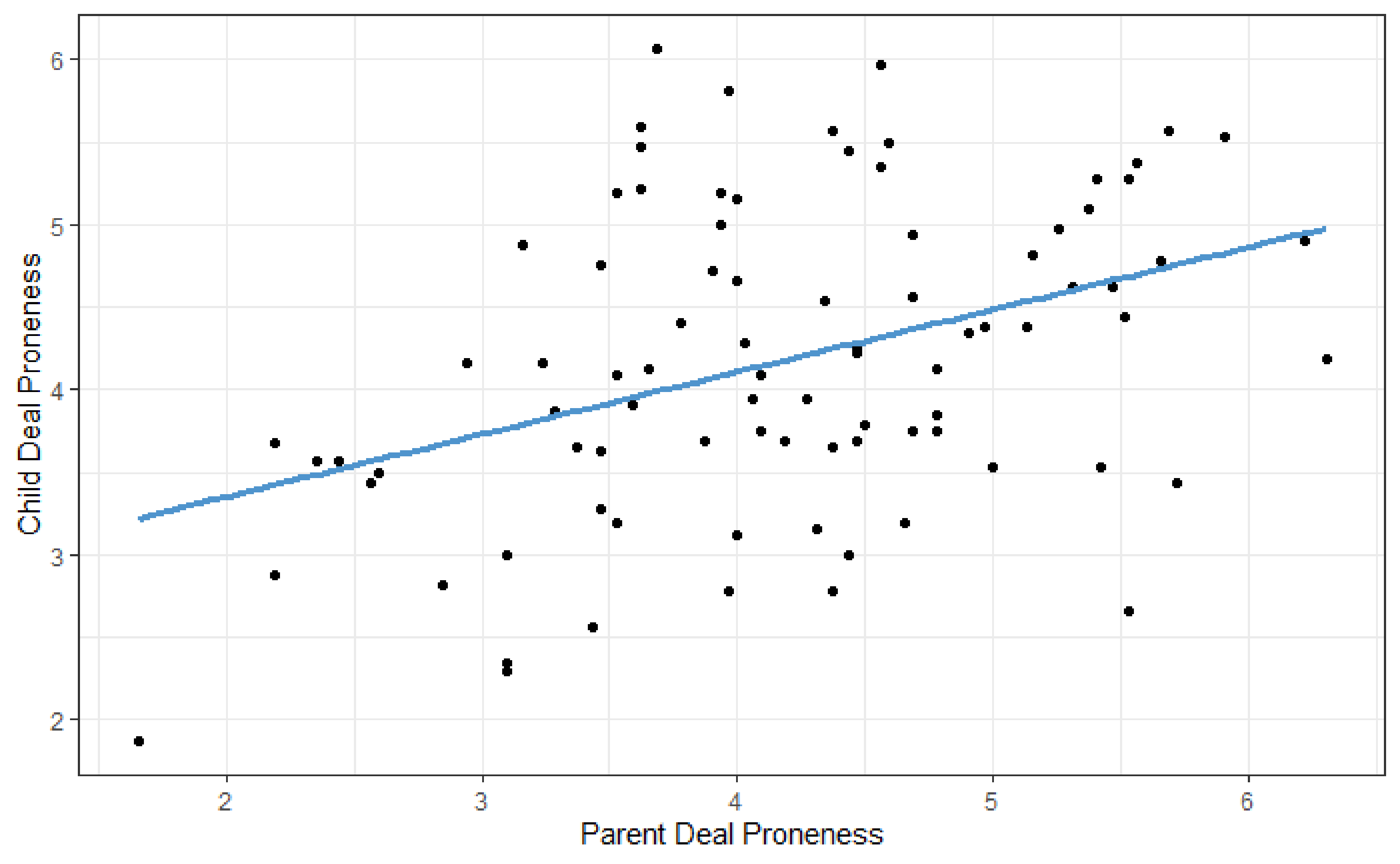


Baseline Model



Credit: Colors selected by Rohan Lala

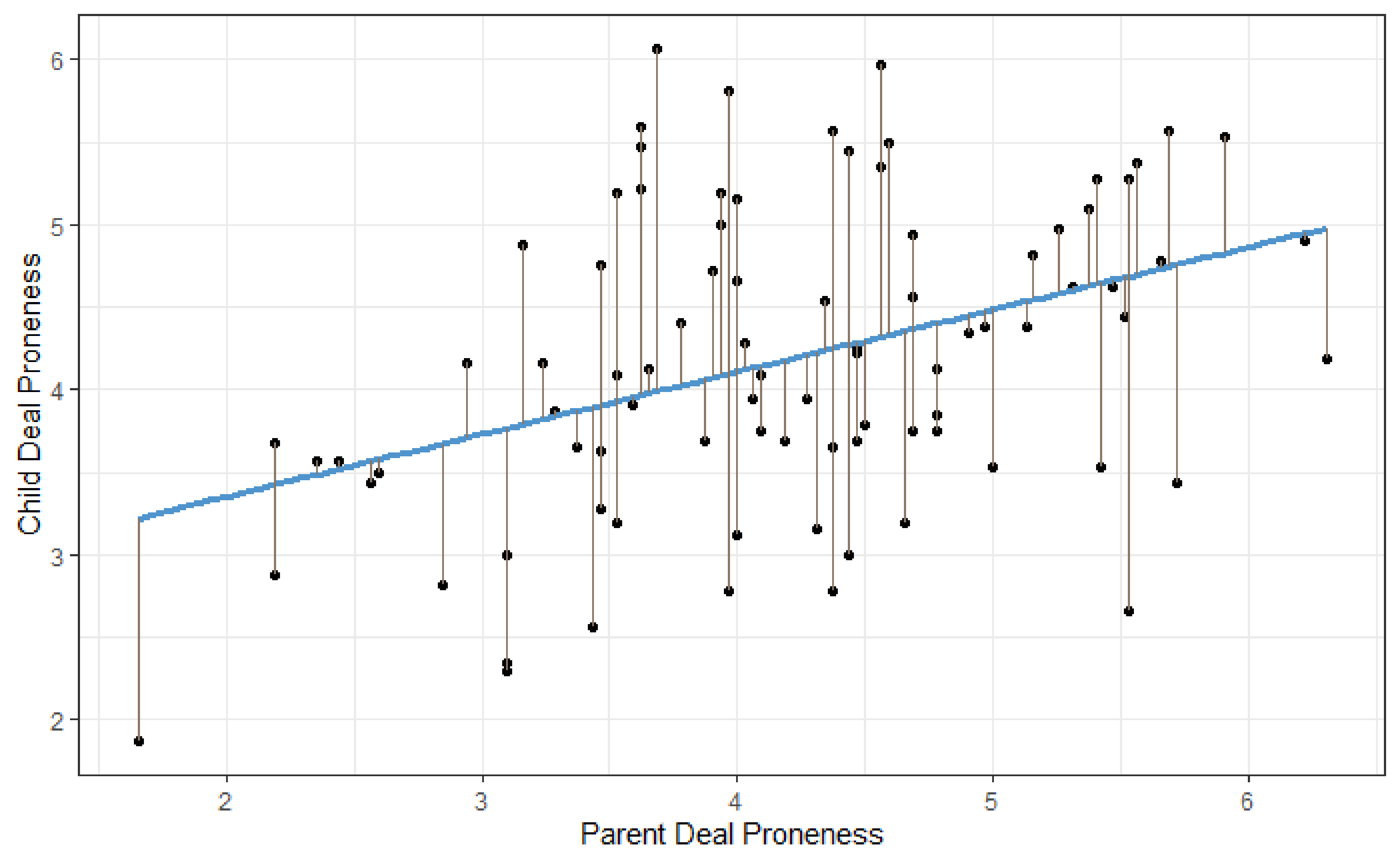
Regression Model



Credit: Colors selected by Nikhil Lala

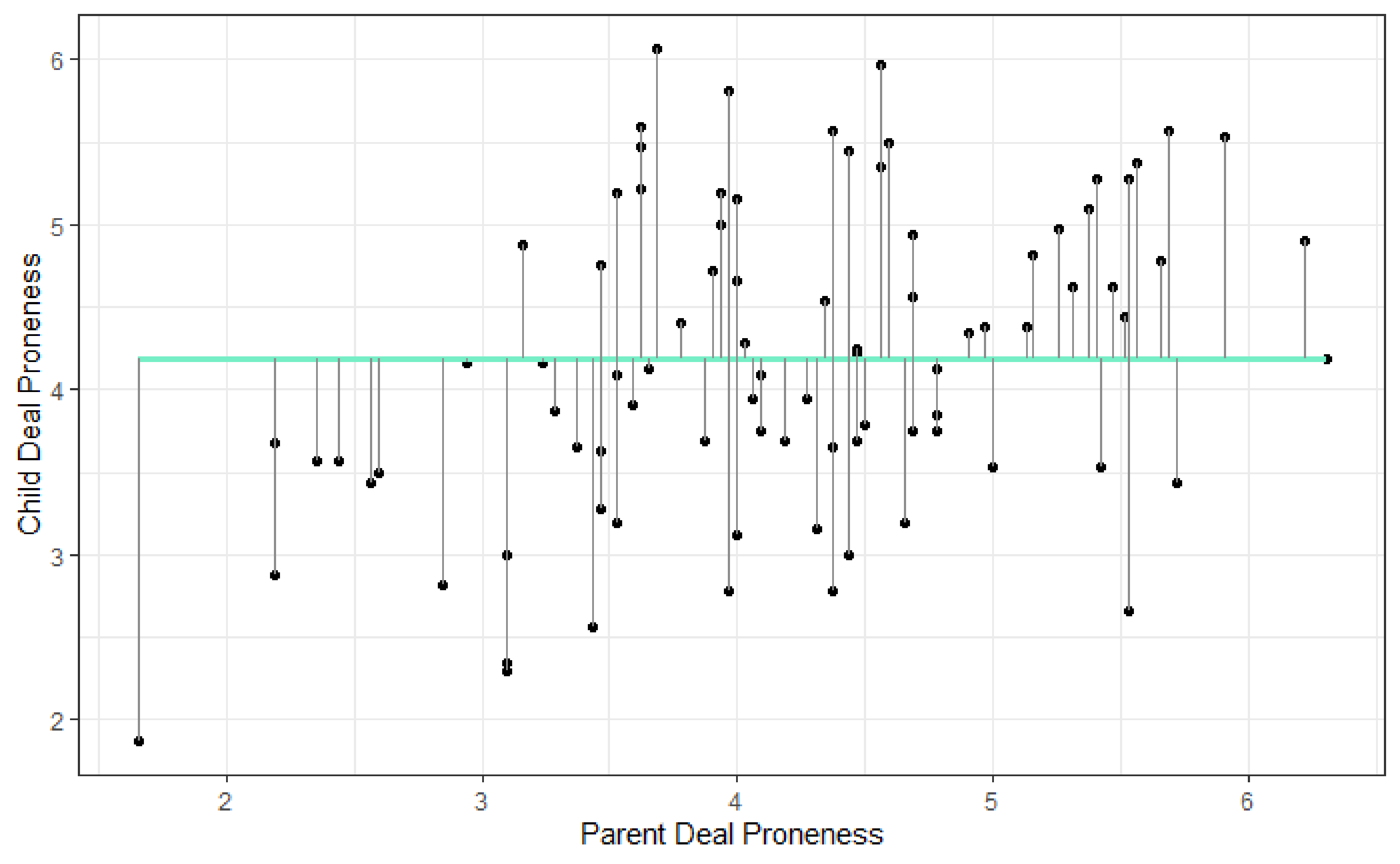
Regression Model (with errors)

sse = $\min(\Sigma e_i^2)$ = sum of squared errors



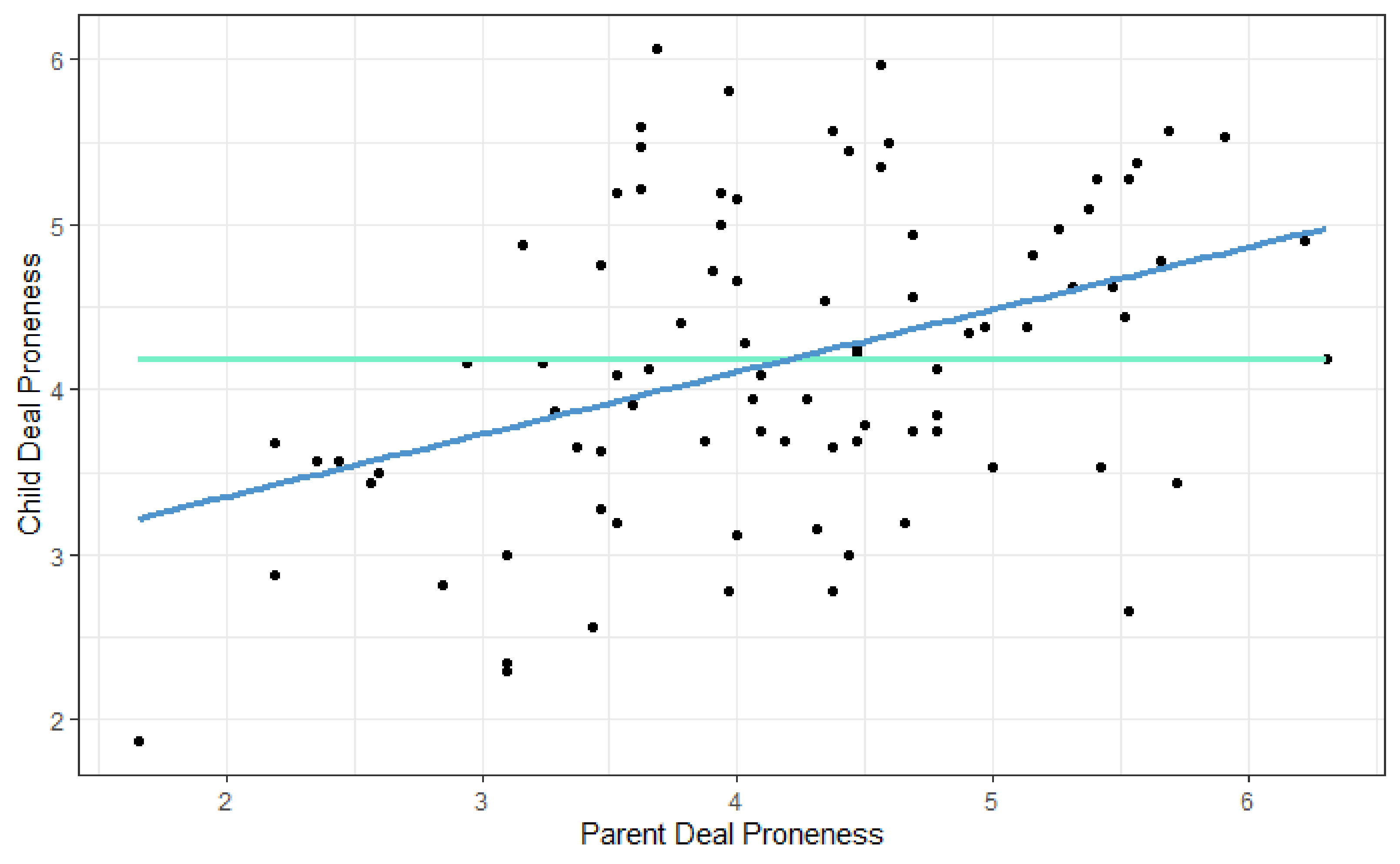
Baseline Model (with errors)

sst = sum of squared total errors



Regression vs. Baseline

$R^2 = 1 – sse/sst$



PREDICTION AND INFERENCE

Prediction

- Is there a relationship between outcome and predictors?
  - *Statistical test to see if at least one of the coefficients is non-zero*
  - *F = ((sst – sse)/p) / (sse/(n-p-1))*
  - *Statistical significance indicates a relationship*
- How strong is the relationship?
  - $R^2 = 1 - sse/sst$
  - $0 < R^2 < 1$
  - *Heuristics: Weak:* $R^2 < 0.1$*, Moderate:* $0.1 <= R^2 < 0.5$*; Strong:* $R^2 >= 0.5$
- How accurate are the predictions?
  - *Various indices that incorporate residuals/errors*
  - *Residual error, Sum of squared errors (sse), Mean squared error (mse), Root mean squared error (rmse)*
  - *Cannot be used for comparisons across samples.*

Inference

- Which predictors influence the outcome?
  - *Statistical test to examine individual coefficients*
  - *$t = b_1/se(b_1)$; where b_1 is estimate of coefficient for first predictor*
  - *Statistical significance indicates an effect*
- Interpretation of coefficients
  - *A unit change in X_1 will result in a change of b_1 units in Y while holding all other predictor variables constant.*
- Nature of the relationship (e.g., linear, quadratic, exponential)
  - *Examine scatterplot between predictor and outcome; Statistical significance of non-linear term will reflect nature of relationship.*
- Relative strength of variables
  - *Standardized regression coefficients; Can only be used for predictors in the same model.*
  - *Standardized_b1 = b1*sd(X)/sd(Y)*

Regression Types

- Regression generates an optimal linear combination of predictor variables to come up with best prediction of the outcome variable.
- In the slides that follow, we will examine each of the following using an example dataset
  - *Simple regression: When there is one predictor*
  - *Multiple Regression: When there are multiple predictors*
  - *How to model categorical predictors*
  - *How to test variable interactions*
  - *Non-linear effects*
  - *Estimate out of sample error*
- Multicollinearity (will be discussed in the module on feature selection)

REGRESSION MODELS

Using Wages Data

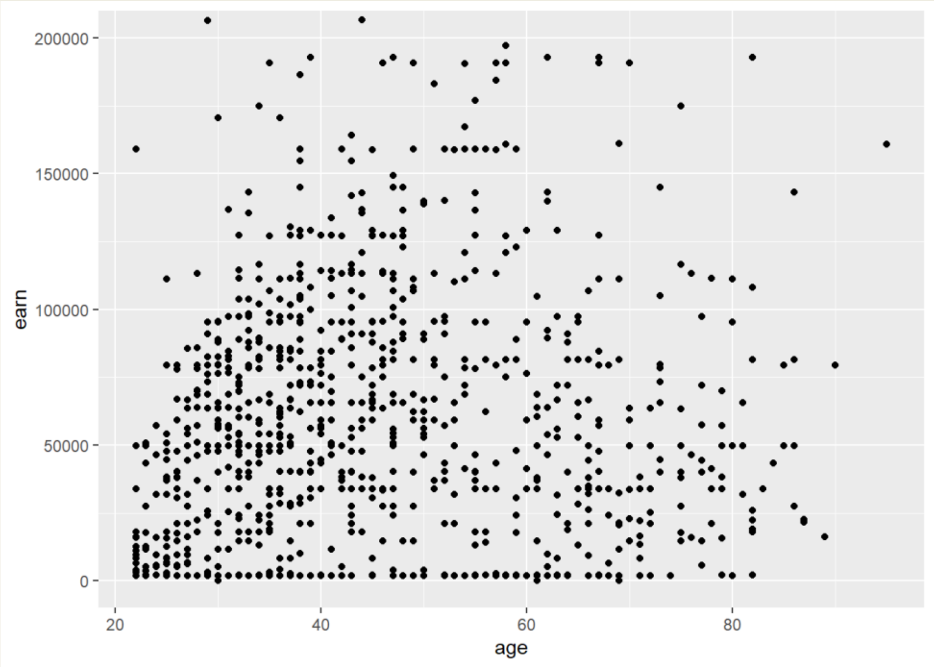
Wages Data

```
'data.frame':   1368 obs. of  6 variables:
 $ earn  : int  159142 192794 97422 160956 164178 30626 94208 101920 6426 85994 ...
 $ height: num  73.9 66.2 63.8 63.2 63.1 ...
 $ sex   : Factor w/ 2 levels "female","male": 2 1 1 1 1 1 1 2 2 2 ...
 $ race  : Factor w/ 4 levels "african-american",..: 4 4 4 2 4 4 4 4 3 4 ...
 $ ed    : int  16 16 16 16 17 15 12 17 15 12 ...
 $ age   : int  49 62 33 95 43 30 53 50 25 30 ...
```

Simulated dataset based on a real dataset in Data Analysis using Regression and Multilevel/Hierarchical Models by Andrew Gelman and Jennifer Hill

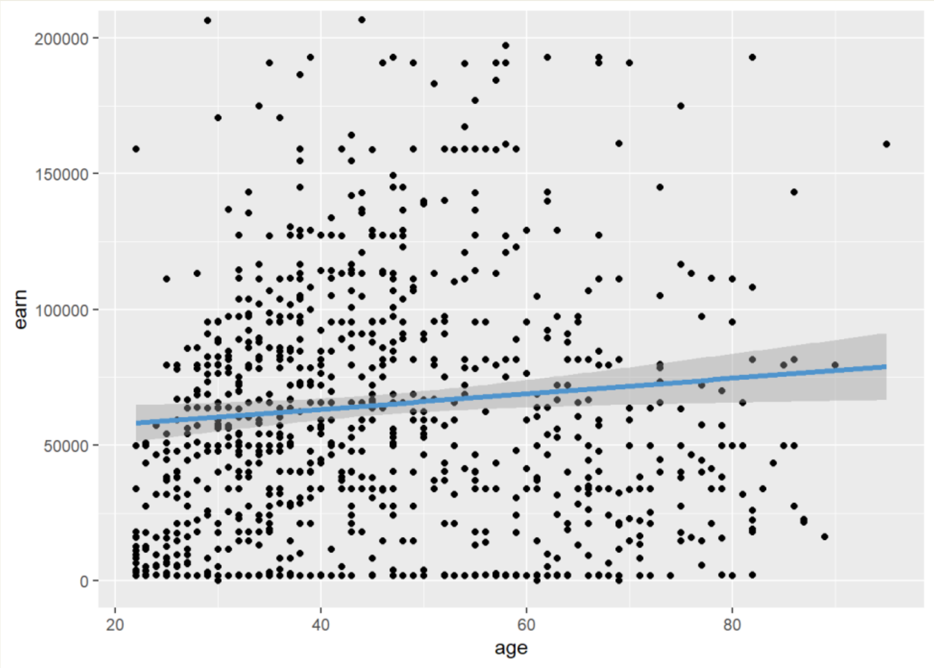
Model 1: Simple Regression
earn = f(age)

- Does age influence how much a person earns?
- Scatterplot is a handy way to visualize bivariate relationships



Model 1: Simple Regression
earn = f(age)

- Linear regression fits a straight line through the data so as to minimize sum of squared errors.
- Gray area indicates confidence bands as the line represents the sample regression function.



Model 1: Simple Regression
earn = f(age)

- Estimate Regression equation
  - *earn = 51806 + 286age*
- Prediction
- Inference

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  51805.6     5832.5   8.882   <2e-16 ***
age            286.0      121.2   2.360   0.0185 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 59270 on 966 degrees of freedom
Multiple R-squared:  0.005731,   Adjusted R-squared:  0.004702
F-statistic: 5.568 on 1 and 966 DF,  p-value: 0.01849
```

Model 1: Simple Regression
earn = f(age)

- Estimate Regression equation
- Prediction
  - *Raw predictions for ten observations*
- Inference

	earn <int>	prediction <dbl>
143	75146	64961.40
144	2010	64103.41
146	47758	61815.44
147	14690	73255.28
149	97418	62101.44
150	89422	61529.45
152	59222	68965.34
153	136514	65533.39
154	103778	65533.39
155	1990	67535.36

1-10 of 10 rows

Model 1: Simple Regression

earn = f(age)

- Estimate Regression equation
- Prediction
  - $F = 5.568$, $p < 0.05$
  - $R^2 = 0.005731$
  - *rmse (computed)=59212.82*
- Inference

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  51805.6     5832.5   8.882   <2e-16 ***
age            286.0      121.2   2.360   0.0185 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 59270 on 966 degrees of freedom
Multiple R-squared:  0.005731,	Adjusted R-squared:  0.004702
F-statistic: 5.568 on 1 and 966 DF,  p-value: 0.01849
```

Model 1: Simple Regression
earn = f(age)

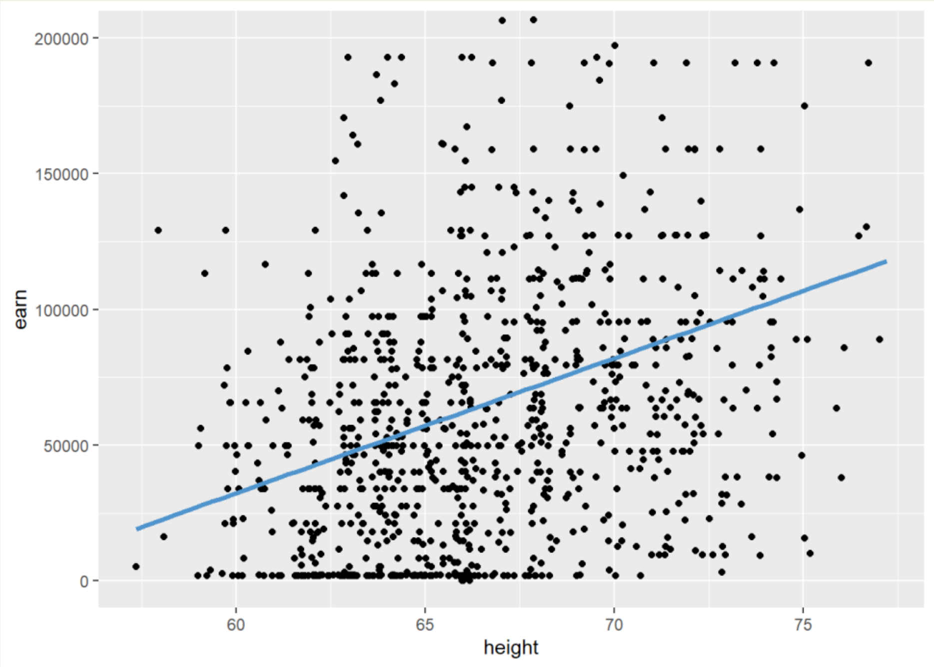
- Estimate Regression equation
- Prediction
- Inference
  - *Age: $t = 2.36$, $p < 0.05$*
  - *Age influences earn*
  - *The model predicts the earn for a 35 year old to be*
    - 51805.6 + 286*35
  - *A person ten years older will make on average 10*286 = $2860 more.*

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  51805.6     5832.5   8.882   <2e-16 ***
age            286.0      121.2   2.360   0.0185 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 59270 on 966 degrees of freedom
Multiple R-squared:  0.005731,	Adjusted R-squared:  0.004702
F-statistic: 5.568 on 1 and 966 DF,  p-value: 0.01849
```

Model 2: Simple Regression
earn = f(height)

- Does height influence how much a person earns?



Model 2: Simple Regression
earn = f(height)

- Estimate Regression equation
  - *earn = -265589.6 +4966 height*
- Prediction
  - *F = 103.5, p < 0.05*
  - R^2 *= 0.0968*
  - *rmse (computed) = 56435.93*
  - *Is height a better predictor than age?*
- Inference

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -265589.6    32522.7  -8.166 9.88e-16 ***
height         4966.0      488.1  10.175  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 56490 on 966 degrees of freedom
Multiple R-squared:  0.0968, Adjusted R-squared:  0.09587
F-statistic: 103.5 on 1 and 966 DF,  p-value: < 2.2e-16
```

Model 2: Simple Regression
earn = f(height)

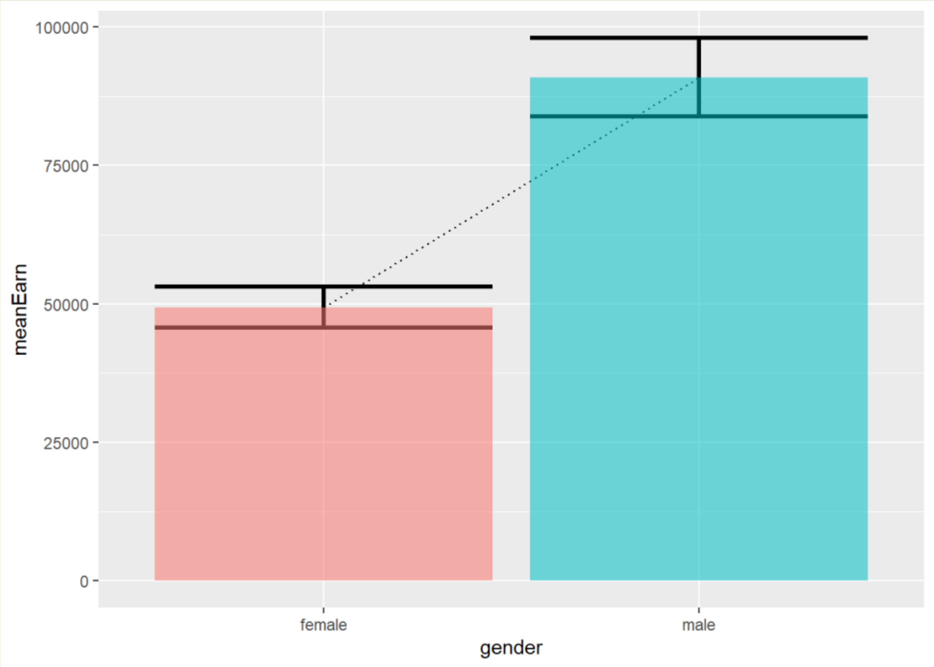
- Estimate Regression equation
- Prediction
- Inference
  - *Height: $t = 10.175$, $p < 0.05$*
  - *Height influences earn*
  - *What is the impact of a 2 inch increase in height on earn?*
  - *How much will a six foot person earn (all else being equal)?*

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -265589.6    32522.7  -8.166 9.88e-16 ***
height         4966.0      488.1  10.175  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 56490 on 966 degrees of freedom
Multiple R-squared:  0.0968, Adjusted R-squared:  0.09587
F-statistic: 103.5 on 1 and 966 DF,  p-value: < 2.2e-16
```

Model 3: Simple Regression (categorical predictor)
earn = f(gender)

- Does a person's gender have an effect on their earning?
- Since gender is a categorical variable, a bar chart is a more meaningful than a scatterplot.
- Error bars represent the 95% confidence intervals



Model 3: Simple Regression (categorical predictor)
earn = f(gender)

- gender is a categorical variable with two levels.
- This variable has a class factor. The factor is unordered, therefore the levels are listed in alphabetical order
- When faced with a predictor that is a factor with two levels, R will treat it as a dummy variable, coding the first level as 0 and the second one as 1.

```
> class(wages$gender)
[1] "factor"
> levels(wages$gender)
[1] "female" "male"
```

Model 3: Simple Regression (categorical predictor)
earn = f(gender)

- Estimate Regression equation
  - *earn = 49367 + 41536*gendermale*
- Prediction
  - *$F = 124.6$, $p < 0.05$*
  - *$R^2 = 0.1143$*
  - *rmse (computed) = 55887.16*
  - *Is gender a better predictor than age/height?*
- Inference

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    49367       2269   21.76   <2e-16 ***
gendermale     41536       3720   11.16   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 55940 on 966 degrees of freedom
Multiple R-squared:  0.1143, Adjusted R-squared:  0.1134
F-statistic: 124.6 on 1 and 966 DF,  p-value: < 2.2e-16
```

Model 3: Simple Regression (categorical predictor)
earn = f(gender)

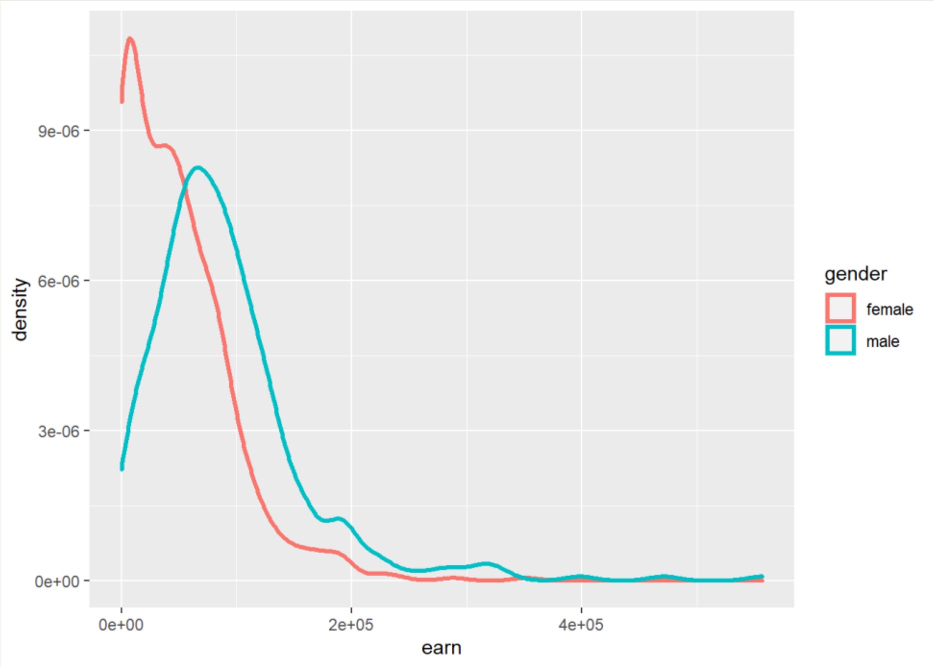
- Estimate Regression equation
- Prediction
- Inference
  - *gender influences earn ($p<0.05$)*
  - *Predicted earn of a female = 49367 + 41536 * 0*
  - *Predicted earn of a male = 49367 + 41536 * 1*
  - *A male makes $41536 more than a female.*

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)    49367       2269   21.76   <2e-16 ***
gendermale     41536       3720   11.16   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 55940 on 966 degrees of freedom
Multiple R-squared:  0.1143, Adjusted R-squared:  0.1134 
F-statistic: 124.6 on 1 and 966 DF,  p-value: < 2.2e-16
```

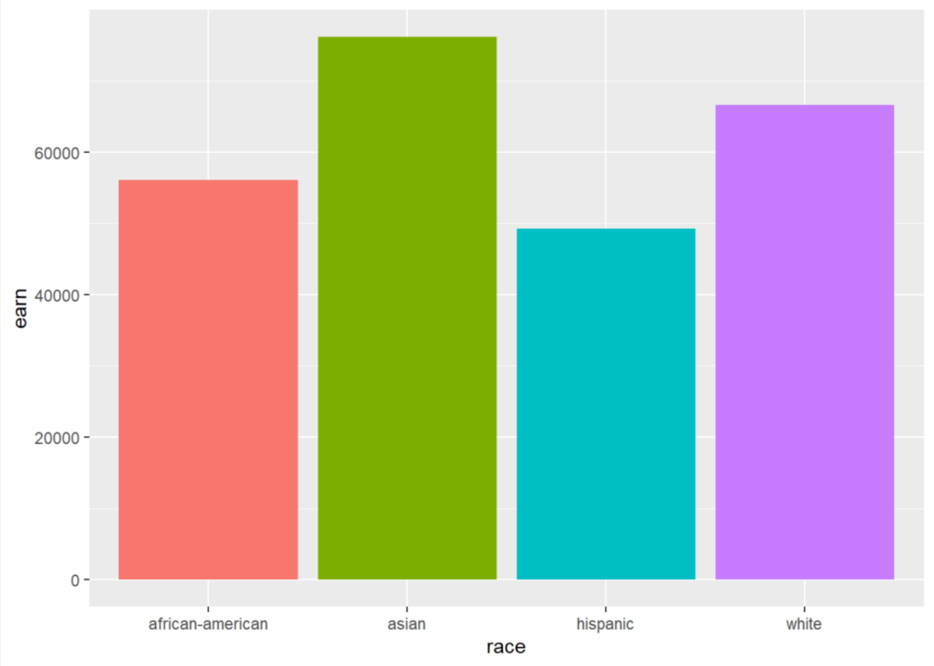
Model 3: Simple Regression (categorical predictor)
earn = f(gender)

- Review density curves to see if you can find a reason for the discrepancy in earn?



# Model 4: Simple Regression (categorical predictor)
earn = f(race)

- Does a person's race have an effect on their earning?



Model 4: Simple Regression (categorical predictor) earn = f(race)

- Race is a factor with four levels.
- This variable has to be dummy coded.
- k levels implies k-1 dummy variables.
- By default, first level becomes the reference level and does not get a dummy variable.
- Remember, for an unordered factor, R will organize levels in alphabetical order.

```
> class(wages$race)
[1] "factor"
> levels(wages$race)
[1] "african-american" "asian"            "hispanic"         "white"
```

Model 4: Simple Regression (categorical predictor) earn = f(race)

- Estimate Regression equation
  - *earn = 56079 + 20040raceAsian – 6865raceHispanic + 10480raceWhite*
- Prediction
  - *F = 2.306, p > 0.05*
  - *Race does not influence earn.*
- Inference

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     56079       6285   8.922   <2e-16 ***
raceasian       20040      15694   1.277    0.202
racehispanic    -6865      10288  -0.667    0.505
racewhite       10480       6622   1.583    0.114
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 59290 on 964 degrees of freedom
Multiple R-squared:  0.007126,	Adjusted R-squared:  0.004037
F-statistic: 2.306 on 3 and 964 DF,  p-value: 0.07522
```

Model 4: Simple Regression (categorical predictor)
earn = f(race)

- Estimate Regression equation
- Prediction
- Inference
  - *Race does not influence earn. Why?*
  - *BUT, IF race influenced earn, who would you say earns more, those who are “white” or “asian” and what is the difference?*

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)      56079       6285   8.922   <2e-16 ***
raceasian        20040      15694   1.277    0.202
racehispanic     -6865      10288  -0.667    0.505
racewhite        10480       6622   1.583    0.114
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 59290 on 964 degrees of freedom
Multiple R-squared:  0.007126,	Adjusted R-squared:  0.004037
F-statistic: 2.306 on 3 and 964 DF,  p-value: 0.07522
```

Model 5: Multiple Regression
earn = f(height, gender)

- A multiple regression will consider the effects of multiple predictors on the outcome
- Do height and gender influence how much a person earns?

Model 5: Multiple Regression
earn = f(height, gender)

- Estimate Regression equation
  - *earn = -101841.9 + 2343 height + 28961.2 genderMale*
- Prediction
  - *F = 69.15, p < 0.05*
  - *R^2 = 0.1254*
  - *rmse (computed) = 55536.7*
  - *Do height and gender jointly predict earn better than either one alone?*
- Inference

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -101841.9    43318.5  -2.351 0.018923 *
height         2343.0      670.3   3.495 0.000495 ***
gendermale    28961.2     5159.9   5.613  2.6e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 55620 on 965 degrees of freedom
Multiple R-squared:  0.1254, Adjusted R-squared:  0.1235
F-statistic: 69.15 on 2 and 965 DF,  p-value: < 2.2e-16
```

Model 5: Multiple Regression
earn = f(height, gender)

- Estimate Regression equation
- Prediction
- Inference
  - *Both height ($p<0.05$) and gender ($p<0.05$) influence earn*
  - *A 4 inch difference in height will correspond to a 4*2343 increase in earn, while holding gender constant*
  - *Of the two, gender is a stronger predictor of earn*

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -101841.9    43318.5  -2.351 0.018923 *
height         2343.0      670.3   3.495 0.000495 ***
gendermale    28961.2     5159.9   5.613  2.6e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 55620 on 965 degrees of freedom
Multiple R-squared:  0.1254, Adjusted R-squared:  0.1235
F-statistic: 69.15 on 2 and 965 DF,  p-value: < 2.2e-16
```

```
Standardized Coefficients::
(Intercept)      height  gendermale
  0.0000000   0.1467905   0.2357114
```

Model 6: Multiple Regression
earn = f(height, gender, race, ed, age)

- A multiple regression will consider the effects of multiple predictors on the outcome
- Generally speaking, more predictors are likely to
  - *Reduce specification bias and presenting a complete picture*
  - *improve predictions*
  - *lead to overfitting*
  - *reduce interpretability*

Model 6: Multiple Regression
earn = f(height, gender, race, ed, age)

- Estimate Regression equation
- Prediction
  - *$F = 47.44$, $p < 0.05$*
  - *$R^2 = 0.257$*
  - *rmse (computed) = 51185.93*
- Inference
  - *Based on the model, how much will a 22 year old, 64 inch tall, White Female with 16 yrs of ed earn?*
  - *Which is the strongest predictor of earn?*

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -204809.1    42188.0  -4.855 1.41e-06 ***
height          1777.8      632.9   2.809  0.00507 **
gendermale     30313.4     4790.2   6.328 3.80e-10 ***
raceasian      17176.3    13660.9   1.257  0.20894
racehispanic   -4240.7     8962.7  -0.473  0.63621
racewhite       5900.6     5760.6   1.024  0.30595
ed              8173.6      674.7  12.114  < 2e-16 ***
age              562.6      107.2   5.248 1.90e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 51400 on 960 degrees of freedom
Multiple R-squared:  0.257,  Adjusted R-squared:  0.2516
F-statistic: 47.44 on 7 and 960 DF,  p-value: < 2.2e-16
```

```
Standardized Coefficients::
 (Intercept)      height  gendermale   raceasian racehispanic   racewhite
  0.00000000  0.11138357  0.24671668  0.03799303  -0.01624610  0.03681513
          ed         age
  0.34367652  0.14893277
```

Model 7: Multiple Regression (with interaction)
earn = f(age, gender, age*gender)

- Previously, we examined effects of predictors acting independently
- Often, variables may interact such that a particular combination of predictors may maximize the outcome.
- Or one variable may be said to modify the relationship of another with the outcome.
- In the scatterplot here, we can see that gender has modified the regression of age on earn.

Model 7: Multiple Regression (with interaction) earn = f(age, gender, age*gender)

- Estimate Regression equation
  - *earn = 40329.8 + 195.3 age + 21569.8 gendermale + 461.4age*gendermale*
- Prediction
  - $F = 46.89, p < 0.05$
  - $R^2 = 0.1273$
  - *rmse (computed) = 55473.6*
- Inference

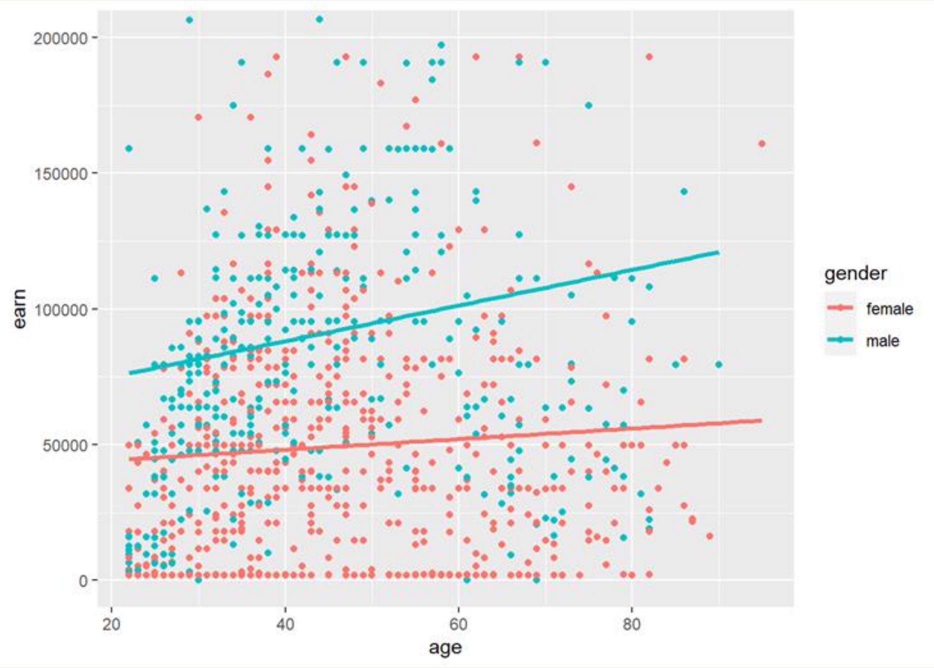
```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)     40329.8     7057.3   5.715 1.47e-08 ***
age               195.3      144.5   1.351   0.1769    
gendermale      21569.8    11186.7   1.928   0.0541 .  
age:gendermale    461.4      234.8   1.965   0.0497 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 55590 on 964 degrees of freedom
Multiple R-squared:  0.1273, Adjusted R-squared:  0.1246 
F-statistic: 46.89 on 3 and 964 DF,  p-value: < 2.2e-16
```

Model 7: Multiple Regression (with interaction)
earn = f(age, gender, age*gender)

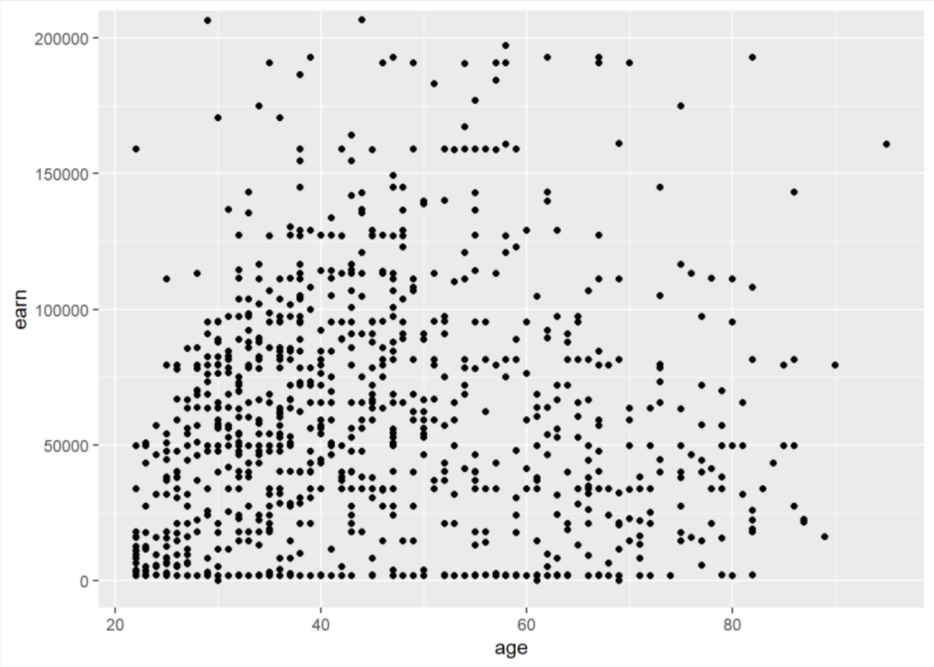
- Estimate Regression equation
- Prediction
- Inference
  - *Age and gender interact ($p<0.05$)*
  - *Statisticians recommend not interpreting the main effects (i..e, effects of age or gender) if the interaction is significant*
  - *Age is positively related to earn BUT only for males*

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     40329.8     7057.3   5.715 1.47e-08 ***
age               195.3      144.5   1.351   0.1769
gendermale      21569.8    11186.7   1.928   0.0541 .
age:gendermale    461.4      234.8   1.965   0.0497 *
```



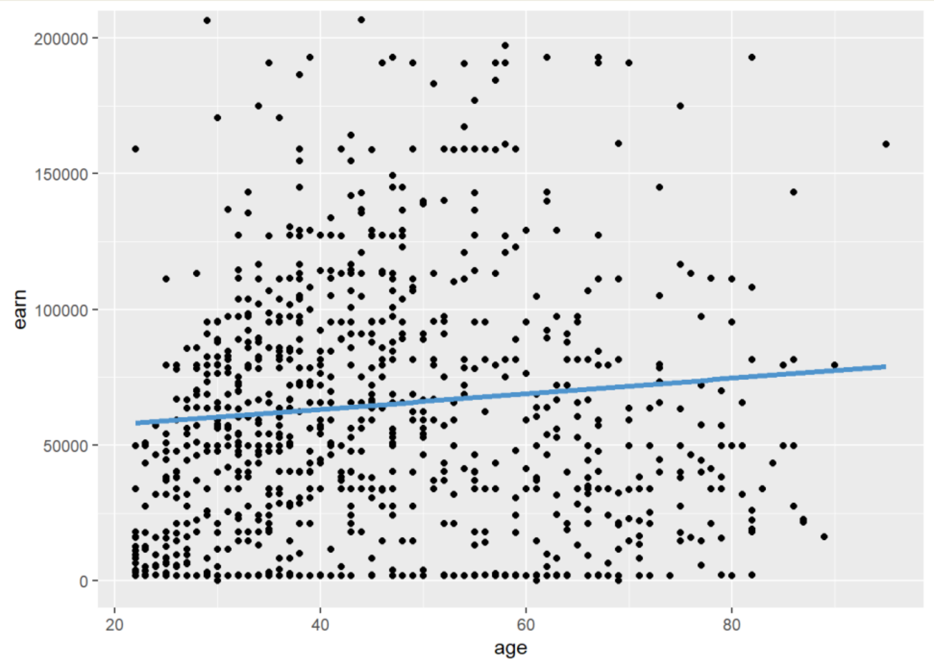
Model 8: Linear Regression to Examine a Non-Linear Relationship
earn = f(age, age^2)

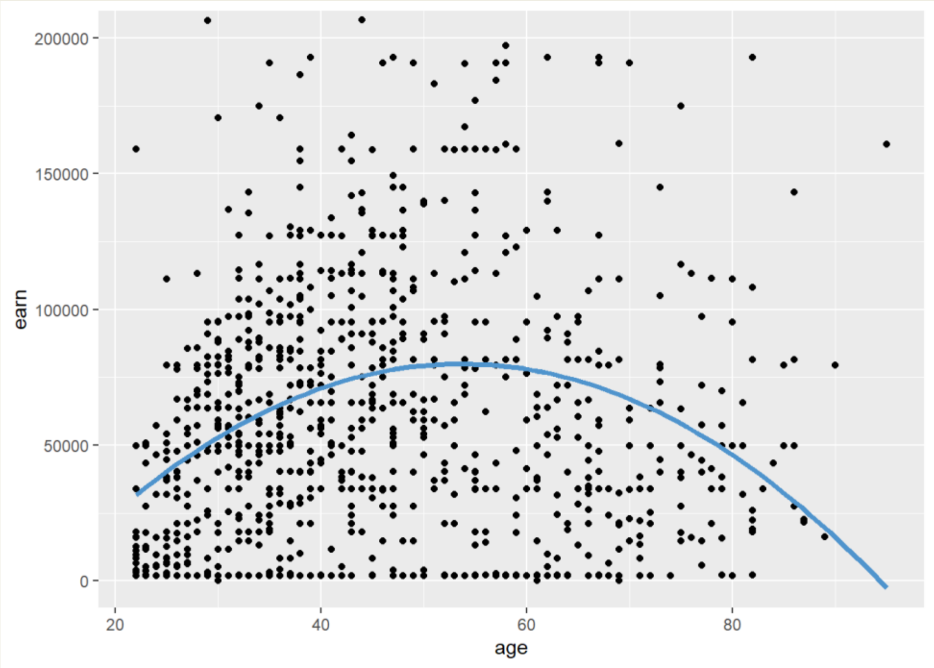
- What is the functional form of the relationship of age on earn?
  - *Linear?*
  - *Quadratic*
  - *Cubic?*
  - *Exponential?*
- A scatterplot may offer a hint, however it is best to consult theory or domain knowledge first.
- Model is linear regression since the parameters for the non-linear predictors (e.g., age^2) are linear.



Model 8: Linear Regression to Examine a Non-Linear Relationship

earn = f(age, age^2)





Model 8: Linear Regression to Examine a Non-Linear Relationship
earn = f(age, age^2)

- Estimate Regression equation
  - *earn = 64814 + 139869 age - 405551 age^2*
- Prediction
  - $F = 27.5, p < 0.05$
  - $R^2 = 0.05391$
  - *rmse (computed) = 57760.27*
  - *Is a nonlinear model better than a linear model?*
- Inference

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)       64814       1859  34.858  < 2e-16 ***
poly(age, 2)1    139869      57850   2.418   0.0158 *
poly(age, 2)2   -405551      57850  -7.010 4.46e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 57850 on 965 degrees of freedom
Multiple R-squared:  0.05391,    Adjusted R-squared:  0.05195
F-statistic:  27.5 on 2 and 965 DF,  p-value: 2.436e-12
```

Model 8: Linear Regression to Examine a Non-Linear Relationship
earn = f(age, age^2)

- Estimate Regression equation
- Prediction
- Inference
  - *The coefficient of Age^2 is significant ($p<0.05$)*
  - *Therefore age has a quadratic relationship with earn*
  - *In your opinion, does this model better represent reality?*

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)       64814       1859  34.858  < 2e-16 ***
poly(age, 2)1    139869      57850   2.418   0.0158 *
poly(age, 2)2   -405551      57850  -7.010 4.46e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 57850 on 965 degrees of freedom
Multiple R-squared:  0.05391,    Adjusted R-squared:  0.05195
F-statistic:  27.5 on 2 and 965 DF,  p-value: 2.436e-12
```

Model 9: Multiple Regression
(evaluate on test sample)
earn = f(height, gender, race, ed, age)

- So far, we have examined prediction performance of the models on the same data used to build them
- Model performance is generally,
  - *better on the sample used to train the model*
  - *but worse on data not used to train the model*
- This problem is exacerbated as the model becomes more complex by say adding more variables, or introducing nonlinear terms.

Model 9: Multiple Regression
(evaluate on test sample)
earn = f(height, gender, race, ed, age)

- Since, getting new data is often too costly, difficulty or not possible, one solution is to split the sample into two parts: train and test
- Estimate the model on train set and evaluate using the test set.
- Performance of model on test set can be used as an indication of out-of-sample performance.

Model 9: Multiple Regression

(evaluate on test sample)
earn = f(height, gender, race, ed, age)

- Estimate Regression equation
- Prediction
- Inference

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -204809.1    42188.0  -4.855 1.41e-06 ***
height           1777.8      632.9   2.809  0.00507 ** 
gendermale      30313.4     4790.2   6.328 3.80e-10 ***
raceasian       17176.3    13660.9   1.257  0.20894    
racehispanic    -4240.7     8962.7  -0.473  0.63621    
racewhite        5900.6     5760.6   1.024  0.30595    
ed               8173.6      674.7  12.114  < 2e-16 ***
age               562.6      107.2   5.248 1.90e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 51400 on 960 degrees of freedom
Multiple R-squared:  0.257,  Adjusted R-squared:  0.2516 
F-statistic: 47.44 on 7 and 960 DF,  p-value: < 2.2e-16
```

Model 9: Multiple Regression

(evaluate on test sample)
earn = f(height, gender, race, ed, age)

- Estimate Regression equation
- Prediction
  - *Train sample*
    - F = 47.44, $p < 0.05$
    - $R^2 = 0.257$
    - rmse (computed) = 51185.93
  - *Test sample*
    - $R^2 = 0.232$
    - rmse (computed) = 60810.34
- Inference

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -204809.1    42188.0  -4.855 1.41e-06 ***
height         1777.8      632.9   2.809  0.00507 **
gendermale    30313.4     4790.2   6.328 3.80e-10 ***
raceasian     17176.3    13660.9   1.257  0.20894
racehispanic  -4240.7     8962.7  -0.473  0.63621
racewhite      5900.6     5760.6   1.024  0.30595
ed             8173.6      674.7  12.114  < 2e-16 ***
age             562.6      107.2   5.248 1.90e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 51400 on 960 degrees of freedom
Multiple R-squared:  0.257,  Adjusted R-squared:  0.2516
F-statistic: 47.44 on 7 and 960 DF,  p-value: < 2.2e-16
```

But wait: Regression Assumptions

- Regression makes a number of assumptions.
- Generally speaking, regression is robust against *small* violations of assumptions.
- It is best to check for these assumptions before conducting analysis.
- A discussion of ways to remedy violations of assumptions is beyond the scope of this course.

- Linear in parameters
- Mean of residuals is zero
- Homoscedasticity
- No autocorrelation
- IVs and residuals are not correlated
- n > number of parameters
- Variance of IVs > 0
- No perfect multicollinearity
- No specification bias
- Errors are normally distributed

[How to test using R]

Conclusion

- In this module, we reviewed
  - *what regression is*
  - *mechanics of estimation*
  - *use of regression for prediction and inference*
  - *estimation of various regression models*
  - *regression assumptions*