# Inference for Linear Regression

## ANOVA for SLR

Grinnell College

December 6, 2024

# Review

- Hypothesis testing
  - test-statistics
  - p-values
  - need to be careful what $H_0$ and $H_A$ actually are
- ANOVA
  - testing equality of group means
  - $H_0$: $\mu_1 = \mu_2 = \cdots = \mu_k$
  - F $= \frac{MSG}{MSE} = \frac{SSG/(k-1)}{SSE/(n-k)}$
  - MSG measures how far (on average) group means are from overall mean
  - MSE measures how far (on average) observations are from their group means

# ANOVA and Regression

ANOVA Null hypothesis:

$$H_0 : \mu_1 = \mu_2 = \dots \mu_k$$

- comparing mean values of a continuous variable for $k$ different groups
- $H_0$ true $\implies$ each group has same *overall* mean $\mu$

We are going to see how this ANOVA stuff can be applied to linear regression

# ANOVA and Regression

We might ask if it is better to predict an outcome ($\hat{y}$) using an overall mean or if we are better off predicting with a group mean:

$$H_0 : \hat{y}_j = \mu, \qquad H_A : \hat{y}_j = \mu_j$$

In this case by *better*, we mean that we minimize the residual sum of squares, or the squared difference between our prediction and the true value

$$\begin{aligned} \text{Sums of Squared Residuals} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} e_i^2 \end{aligned}$$

# Regression

Recall that regression formulas are of the form:

$$y_i = \beta_0 + X_i \beta_1 + \epsilon_i$$

- $\beta_0$ represents an intercept
- $\beta_1$ indicates a slope associated with $X_i$

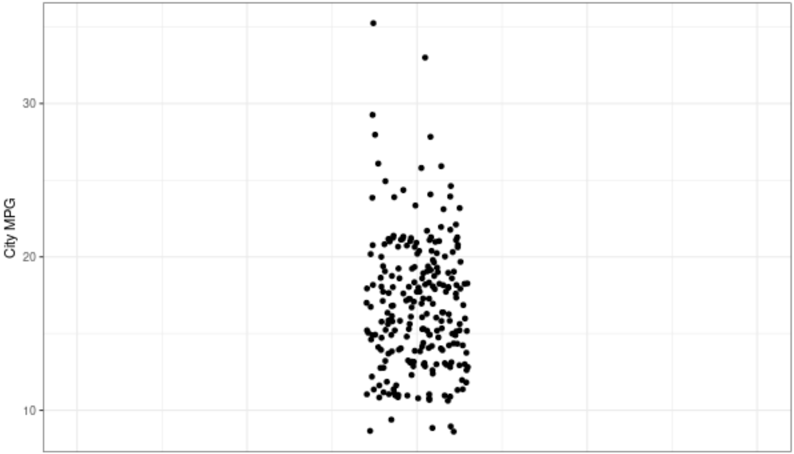
Once we fit to line to the data, we have an estimated line of

$$\hat{y}_i = \hat{\beta}_0 + X_i \hat{\beta}_1 \ \ (= b_0 + b_1 X_i)$$

- residual $e_i = y_i - \hat{y}_i$ is an estimate of the error $\epsilon_i$

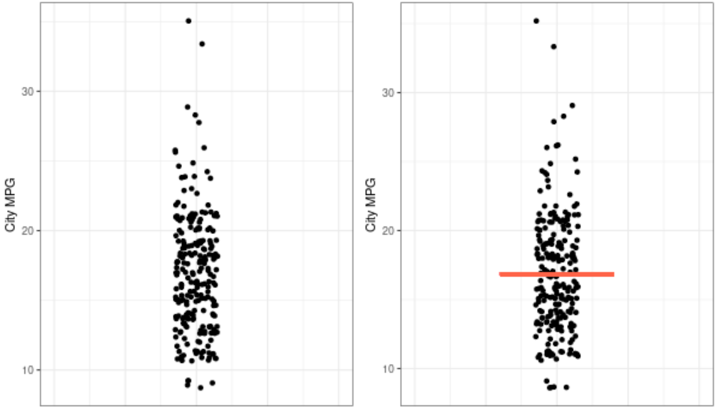
# mpg Example

Consider the mpg dataset, where we might be interested in estimating the city miles per gallon of various vehicles



# mpg Example

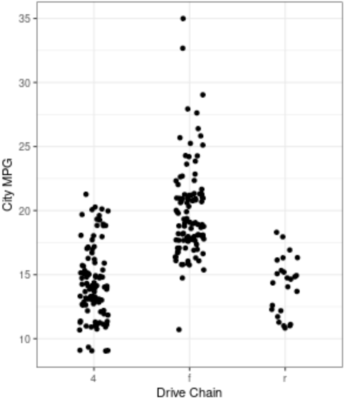
Using simply the overall mean, we would have total squared error of 4220

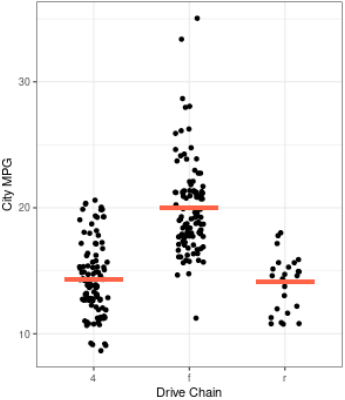



| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| Residuals | 233 | 4220.35 | 18.11 | | |

# mpg Example

Consider the alternative, where we predict city mileage based on drive train




| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| drv | 2 | 1878.81 | 939.41 | 92.68 | <0.0001 |
| Residuals | 231 | 2341.53 | 10.14 | | |

- SSR has gone down (good!) and is sequestered into SSG (drv)

# mpg Example

In terms of a regression model, we could frame this as

$$\hat{y} = \mathbb{1}_{\text{4wd}}\hat{\beta}_1 + \mathbb{1}_{\text{Fwd}}\hat{\beta}_2 + \mathbb{1}_{\text{Rwd}}\hat{\beta}_3$$

where $\mathbb{1}$ represents our *indicator variable* and, in the case of categorical variable regression, $\hat{\beta}$ represents the mean value for each group. This is exactly what we saw towards the beginning of the semester

```
> lm(cty ~ -1 + drv, mpg)

Coefficients:
 drv4   drvf   drvr
14.33  19.97  14.08
```

$$\hat{y} = (14.33 \times \mathbb{1}_{\text{4wd}}) + (19.97 \times \mathbb{1}_{\text{Fwd}}) + (14.08 \times \mathbb{1}_{\text{Rwd}})$$

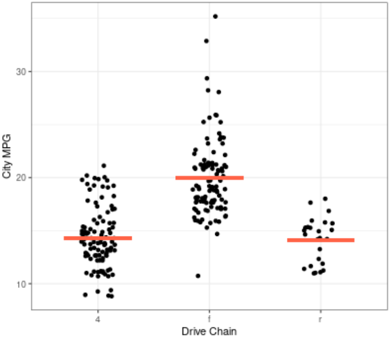
# Baseline Category

By default, R will choose one category as the "reference" variable

- usually based on 1st alphabetic category or lowest numeric

```
> lm(cty ~ drv, mpg)
(Intercept)          drvf          drvr
    14.3301        5.6416       -0.2501
```

$$\hat{y} = \hat{\beta}_0 + \mathbb{1}_{\text{Fwd}}\hat{\beta}_1 + \mathbb{1}_{\text{Rwd}}\hat{\beta}_2 = 14.33 + 5.64 \times \mathbb{1}_{\text{Fwd}} - 0.25 \times \mathbb{1}_{\text{Rwd}}$$



So, what we have just seen tells us:

- SLR with one categorical variable as a predictor is actually a special case of ANOVA
- both attempted to minimize SSE (=SSR) by partioning that variance into something else (SSG)

However, instead of simply assessing whether or not there is *any* difference between groups, we may be interested specifically in estimating values of $\beta$ in the expression

$$y = \beta_0 + X\beta_1 + \epsilon$$

where X is a *quantitative* variable

# Inference and Regression

$$y = \beta_0 + \beta_1 X + \epsilon$$

When considering a regression line, we are actually trying to find out if there is a linear relationship between the variables.

We could test this by structuring a null hypothesis like so:

$$\begin{aligned} & H_0 : \text{there is no linear relationship} \\ \text{(equivalently)} \quad & H_0 : \beta_1 = 0 \end{aligned}$$

Given our estimate of $\hat{\beta}$, we can make the test statistic,

$$t = \frac{\hat{\beta}_1}{SE_{\beta_1}}$$

# mpg Example

Comparing residuals and F statistic for ANOVA and regression

```
> aov(cty ~ drv, mpg) %>% summary()
             Df Sum Sq Mean Sq F value Pr(>F)
drv           2   1879   939.4   92.68 <2e-16 ***
Residuals   231   2342    10.1
```

```
> lm(cty ~ drv, mpg) %>% summary()

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.3301     0.3137  45.680   <2e-16 ***
drvf          5.6416     0.4405  12.807   <2e-16 ***
drvr         -0.2501     0.7098  -0.352    0.725


Residual standard error: 3.184 on 231 degrees of freedom
Multiple R-squared:  0.4452,  Adjusted R-squared:  0.4404
F-statistic: 92.68 on 2 and 231 DF,  p-value: < 2.2e-16
```

# mpg Example

Comparing pairwise differences for TukeyHSD and regression (reference/intercept variable is 4WD)

```
> aov(cty ~ drv, mpg) %>% TukeyHSD()
  Tukey multiple comparisons of means
    95% family-wise confidence level

          diff       lwr       upr     p adj
f-4  5.6416010  4.602497  6.680705 0.0000000
r-4 -0.2500971 -1.924554  1.424359 0.9338857
r-f -5.8916981 -7.561520 -4.221876 0.0000000
```

```
> lm(cty ~ drv, mpg) %>% summary()

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.3301     0.3137  45.680   <2e-16 ***
drvf          5.6416     0.4405  12.807   <2e-16 ***
drvr         -0.2501     0.7098  -0.352    0.725
```

# ANOVA and Regression

ANOVA is a generalization of the t-test for multiple groups

- regression is a generalization of ANOVA for any combination of variables
- only tells us that a difference exists, not *what* the difference actually is

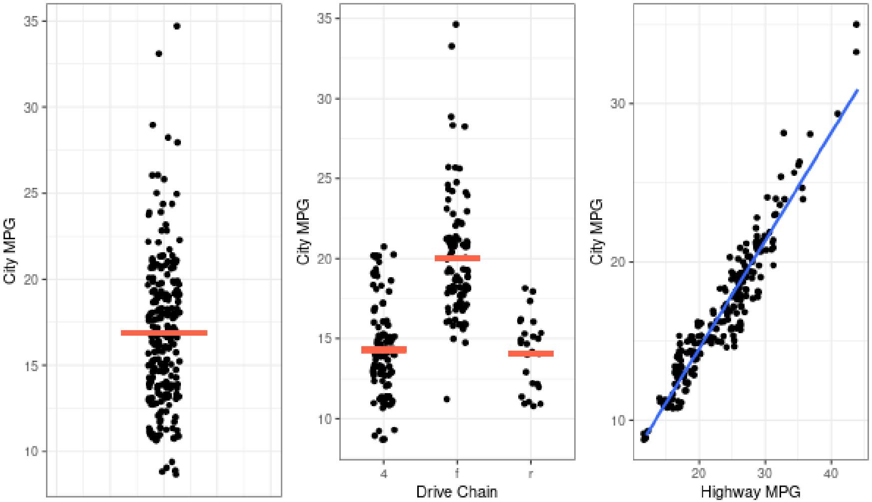
Benefits of Regression:

- requires fewer assumptions about data
  - ANOVA has a hidden assumption of Normal groups
- provides statistical tests for each of the group categories
- allows us to predict quantitative outcome using a quantitative predictor

# Regression Example

Which of these do you suspect will have the smallest residual error?

- think about how far observations are from predictions



# mpg Example

```
> lm(cty ~ hwy, mpg) %>% summary()

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.84420    0.33319   2.534   0.0119 *
hwy          0.68322    0.01378  49.585   <2e-16 ***

Residual standard error: 1.252 on 232 degrees of freedom
Multiple R-squared:  0.9138,  Adjusted R-squared:  0.9134
F-statistic:  2459 on 1 and 232 DF,  p-value: < 2.2e-16
```

$$\hat{y} = b_0 + b_1 \times (hwy) = 0.84 + 0.68 \times (hwy)$$

- F is testing whether both intercept and slope are zero
- t is testing for specifically slope/intercept one at a time
- it is possible that the F-test shows a linear model works well, but that the intercept is not significant

# Interpretations

Interpretations of coefficients is exactly the same as before:

**Slope ($\mathbf{b}_1$):** how much the prediction for y ($\hat{y}$) changes when we change the X variable

**Intercept ($\mathbf{b}_0$):** our prediction for y ($\hat{y}$) when X $= 0$

MPG example: $\widehat{city} = b_0 + b_1 \times (hwy) = 0.84 + 0.68 \times (hwy)$

**Slope:**

- when we change the hwy mpg of a vehicle by 1, the predicted city mpg changes by 0.68

**Intercept:**

- when the highway mpg of a vehicle is 0, the predicted city mpg is 0.84

# Key Takeaways

- Regression is a generalization of ANOVA
- The $\beta$ coefficients indicate how much a change in $X$ impacts a change in $Y$
- Under the null, $H_0 : \beta = 0$
- $R^2$ gives an estimate of explained variance that, in the case of regression with a categorical variable, is identical to the sum of between-group variability
- Likewise, the residuals correspond to the total within-group variability