Review of Quiz 2

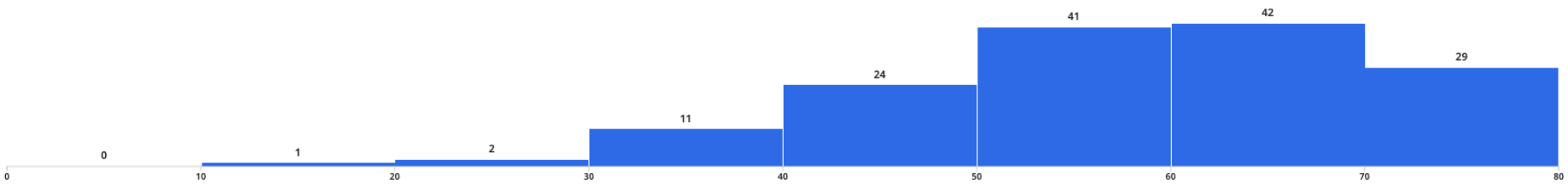
DSC 152 – Week 8 Discussion

May 20th, 2026

Reminders

- Lab #7 due Monday
- Homework #3 to be released tomorrow, due next Thurs.
- Regrade requests open for Quiz #2, open until next Tuesday

Some Statistics for Quiz #2



Minimum	Median	Maximum	Mean	Std Dev
16.0	58.75	79.0	57.31	12.81

Lower mean/median than last time, and slightly higher standard deviation than Quiz 1 (12.81 as compared ~ 9.5).

Reminder: your lowest quiz score will be dropped!

Question #1

Researchers want to see how adherence rates relate to recovery times, fitting a simple linear model between the two variables. The abridged output is:

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      38.1950     2.5255  15.123  < 2e-16 ***
adherence_rate  -0.1892      0.0312  -6.063 6.65e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.42 on 198 degrees of freedom
Multiple R-squared:  0.1566,        Adjusted R-squared:  0.1523
F-statistic: 36.76 on 1 and 198 DF,  p-value: 6.646e-09
```

(a) Write a sentence w/ the correct interpretation of the slope coefficient.

For every 1 percentage point increase in adherence rate, the expected recovery time decreases by 0.1892 weeks, on average.

Question #1

(b) What does the intercept (`Intercept`) represent in this model?

- The expected recovery time for a patient with 100% adherence.
  FALSE: The intercept term corresponds to when the adherence rate is 0%, not 100%.
- The expected recovery time for a patient with 0% adherence.
  TRUE: The intercept term is the predicted outcome of the model when the predictor variable takes the value 0. In this case, this would mean 0% adherence.
- The average recovery time in the dataset.
  FALSE: The intercept term depends on the mean of the outcome (recovery time), the coefficient, and the mean of the predictor. This would only be the case when the slope is 0 or if the mean of the predictor is 0, which is a special case of the second choice.
- The change in recovery time per 1% increase in adherence.
  FALSE: This answer choice describes the interpretation of the slope coefficient β_1, not the intercept.

Question #1

(c) Residual Sum of Squares (RSS) is $\text{RSS} = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$. Which are true?

- [x] The RSS is used to find the least squares regression line, by finding the coefficient values that minimize it.
  TRUE: This choice describes the procedure correctly given that the estimated coefficients are those that minimize the RSS over all possible coefficient values, making RSS the objective function for OLS.

- [ ] This quantity will always be larger when the relationship between x and y is non-linear compared to when it is linear.
  FALSE: RSS measures model fit, not whether the true relationships is linear or nonlinear, and there is no theoretical guarantee of produced RSS values solely based on the underlying relationship being linear or nonlinear.

- [ ] This quantity only works for simple linear regression, and not multiple linear regression.
  FALSE: RSS remains the same residual-based objective function minimized by OLS given any number of predictors.

- [x] In simple linear regression, the definition of $\widehat{y}_i$ is the predicted value of y at the point x_i.
  TRUE: In class, we defined $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$ to be the predicted values at each i.

- [x] If we add `HIV_status` to `model1` as a covariate and run a linear model, the RSS from this new model must be less than or equal to the RSS from the original `model1`.
  TRUE: Adding predictors expands the parameter space over which RSS gets minimized. The original model is contained within a larger model, so optimizing for the expanded model cannot yield a larger minimum RSS. Thus, the RSS must decrease or remain unchanged.

Question #1

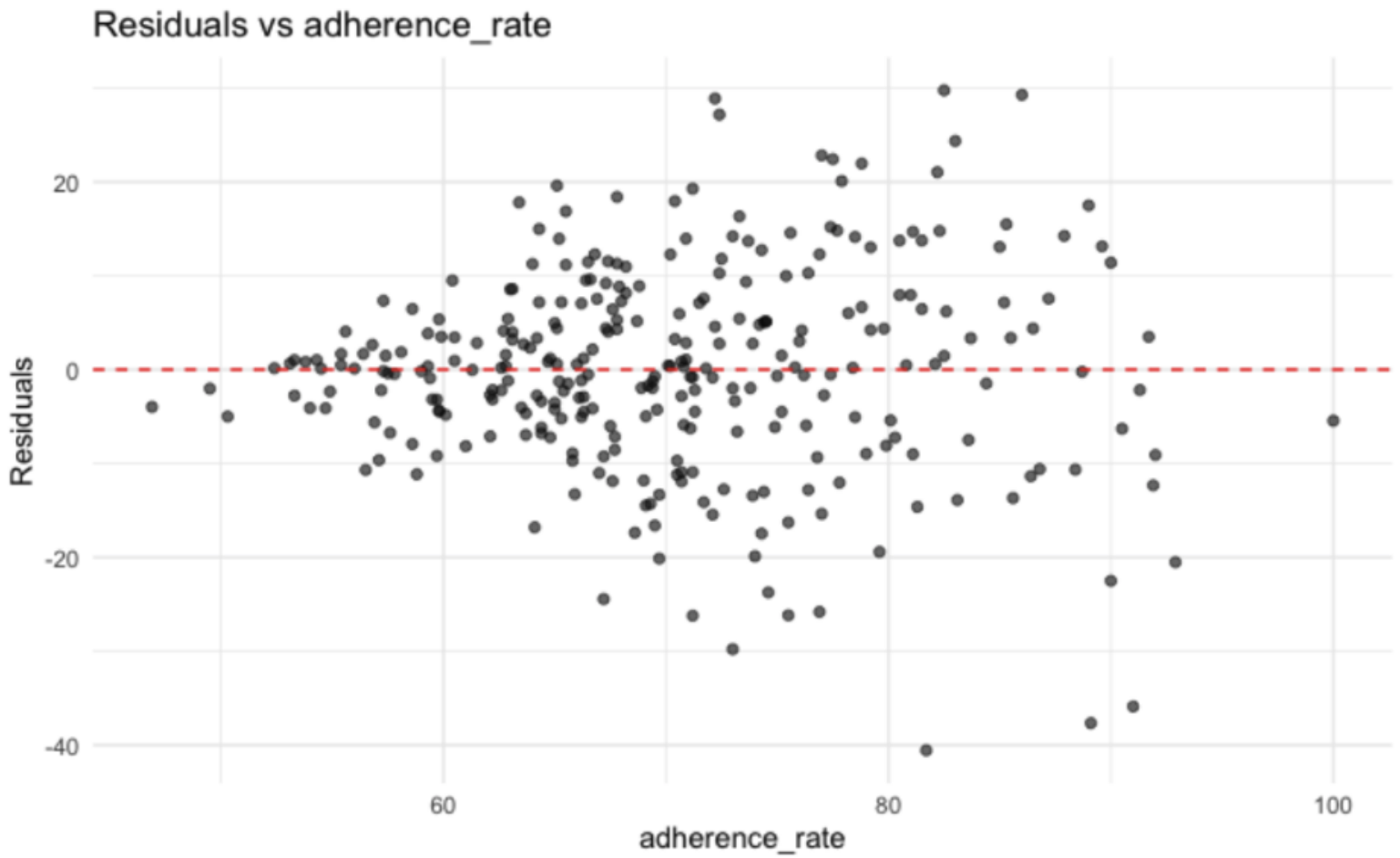
(d) Write code to produce a residual plot most suitable for checking **linearity**.

```
tb_data$resid <- model1$residuals          # Get residuals
tb_data$fitted <- model1$fitted.values     # Get fitted vals
# Plot fitted values vs. residuals to check for linearity
ggplot(data=tb_data, aes(x=fitted,y=resid))+geom_point()
```

We use the fitted values for linearity.

Question #1

Plot in question for Question 1(e). Notice the plot is between the actual predictor values and the residuals.



Question #1

(e) Which condition for linear regression is/are violated based on the previous plot?

- ☐ Linearity.
  FALSE: To check linearity, we would need to check for systematic curvature in a plot of residuals versus fitted values, to check if the conditional mean of the response is a linear function of the predictor. This plot is between residuals and predictor values (not fitted values), which we tend to check for linearity with respect to any curvature away from 0.

- ☐ Independence of observations.
  FALSE: Independence refers to whether residuals are correlated across observations, often due some time ordering or clustering. This plot does not help determine whether this condition is violated.

- ☐ Normality of ε_i.
  FALSE: Normality is associated with the distributional shape of the errors, seen using a histogram or QQ-plot. This plot does not necessarily indicate a deviation from normality.

- ■ Equal variances.
  TRUE: The assumption for homoskedasticity (equal variance) requires the residuals to be mostly constant across all predictor values. This funnel-shaped plot, where residuals start spreading away from 0 as the adherence rate increases, is evidence of the variance changing with the predictor.

- ☐ None of the above.
  FALSE: Since choice 4 is correct, this cannot hold.

Question #2

Primary new question is whether adherence to prescribed medication is associated with faster recovery, while adjusting for undernourishment. Thus, we fit

```
mod_adj <- lm(time_to_recovery ~ adherence_rate +
    undernourishment, data=tb_data)
```

(a) What must be true about `undernourishment` for it to be a confounder?

Undernourishment must have some relationship to both adherence rate (the primary predictor) and the time to recovery (the outcome).

Question #2

(b) Suppose the estimated `adherence_rate` coefficient is $\widehat{\beta}_1 = -0.08$. What does this indicate?

- Each 1% increase in adherence causes recovery time to decrease by 0.08 weeks for all patients.
  FALSE: This model shows an association, not necessarily a causal effect, and it is incorrect to say "for all patients" since the coefficient is conditional on patients with the same undernourishment status.
- Patients tend to recover 0.08 weeks slower per 1% increase in adherence, among patients with the same undernourishment status.
  FALSE: The coefficient is negative, meaning an increase in adherence decreases recovery time, or that patients recover quicker.
- Patients tend to recover 0.8 weeks faster per 10% increase in adherence, among patients with the same undernourishment status.
  TRUE: A 1% increase in adherence is associated with an 0.08 decrease in recovery time, adjusting for undernourishment. Multiplying by 10 yields this statement's interpretation.
- Undernourished patients recover 0.08 weeks faster on average than well-nourished patients.
  FALSE: This statement incorrectly attributes the coefficient to the confounder (undernourishment) instead of the primary predictor (adherence rate).
- Adherence has no association with recovery time after adjusting for undernourishment.
  FALSE: With an estimated non-zero coefficient $\widehat{\beta}_1$, we understand there is an association between adherence and recovery time, after adjustment.

Question #2

(c) Suppose estimated `undernourishmentyes` coefficient were $\widehat{\beta}_2 = 1.5$ and we flip reference categories. What would then be true?

- [x] $\widehat{\beta}_2$ would equal -1.5.
  TRUE: Changing the reference level reverses the coding of the indicator variable, meaning the coefficient now represents the difference (no) − (yes). Thus, the coefficient changes sign and becomes -1.5.
- [ ] $\widehat{\beta}_2$ would equal $1/1.5$ $(= 2/3)$.
  FALSE: "Releveling" a categorical variable does not cause the coefficient to be inverted; it simply changes sign since the difference is now reverted.
- [ ] $\widehat{\beta}_1$ could possible change and be different from -0.08.
  FALSE: "Releveling" a categorical variable would not alter the fitted values, it simply reparameterizes the model in a slightly different way. Thus, the slope estimate for adherence would be unchanged.
- [ ] P-value for $H_0 : \beta_2 = 0$ would be different in new model versus original.
  FALSE: Flipping the reference category only changes the sign of the coefficient, not its magnitude. Since the model has not changed, its corresponding p-value will be the same.
- [x] RSS for this new model would be the same as the original.
  TRUE: Since flipping the reference only changes how the coefficients are expressed, the predicted values remain unchanged. Thus, the RSS would be identical across both parameterizations.

Question #2

(d) Which are true about overall global $\mathcal{F}$-test for original model?

- ☐ It tests $H_0 : \beta_0 = \beta_1 = \beta_2 = 0$.
  FALSE: The intercept term (β_0) should not be included in the null hypothesis for the global $\mathcal{F}$-test since the null should test whether the predictor coefficients are zero (intercept-only model).
- ■ It tests $H_0 : \beta_1 = \beta_2 = 0$.
  TRUE: The global $\mathcal{F}$-tests whether all non-intercept coefficients are equal to zero, testing the full model versus an intercept only model. In this case, those coefficients are β_1, β_2.
- ☐ It tests $H_0 : \beta_1 = \beta_2$.
  FALSE: The global $\mathcal{F}$-test does not compare non-intercept coefficients to each other. The provided null is a test for equal coefficients, not for the presence of a predictor.
- ■ It is an inappropriate test to do fro the specific question of interest.
  TRUE: The original model looked at whether adherence is associated with recovery time after adjusting for undernourishment, or $H_0 : \beta_1 = 0$. This null is different to the one in the global $\mathcal{F}$-test, which addresses overall model significance instead of targeted adherence effect.
- ☐ None of the above.
  FALSE: Choices 2 and 4 are correct, so this statement cannot hold.

Question #2

(e) Suppose the p-value for `adherence_rate` in the original model was 0.003. State the null hypothesis and the conclusion at $\alpha = 0.05$.

We test $H_0 : \beta_1 = 0$ (adherence rate has no association with recovery time, adjusting for undernourishment), and since $p = 0.003 < 0.05 = \alpha$, we would reject H_0. That is, there is statistically significant evidence to suggest that the adherence rate is associated with recovery time among patients with the same undernourishment status.

Question #3

We want to test whether adherence rates differ between patients enrolled in adherence support programs or not, running the test:

```
model2 <- lm(adherence_rate ~ intervention, data=tb_data)
```

Assume `intervention = 'standard'` is the reference category.

(a) State null/alternative hypotheses, in math and in words.

We test

$$H_0 : \beta_1 = 0 \quad \text{vs.} \quad H_A : \beta_1 \neq 0$$

In words, the null assumes the mean adherence rate is the same for patients in both intervention groups, while the alternative suggests the mean adherence rates differ between the two groups.

Question #3

(b) Write R code to perform partial $\mathcal{F}$-test for hypotheses listed in (a), using the null model `null_model`.

```
1 # SOLUTION
2 null_model <- lm(adherence_rate ~ 1, data=tb_data)
3 anova(null_model, model2)
```

Since `intervention` is a categorical predictor with reference category ``standard'', we can write the model as

$$y = \beta_0 + \beta_1 \cdot \mathbb{1}\{\texttt{intervention = ``program''}\}$$

testing $H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$. Under H_0, intervention has no explanatory power for adherence rate, and so our null model only contains an intercept term, precisely what `~ 1` is doing in line 1.

Then, `anova` allows us to perform our partial $\mathcal{F}$-test, comparing the reduced, intercept-only (null) model with the full model (`model2`).

Question #3

(c) Would an equivalent p-value to that of the partial test from (b) appear in the `summary(model2)` output?

YES: Because `intervention` is the model's only predictor, the partial $\mathcal{F}$-test tests $H_0 : \beta_1 = 0$, which is the same as the null tested by a t-test for the intervention coefficient (tests if coefficient is non-zero). Moreover, this null is the same as that for the global $\mathcal{F}$-test (tests whether all non-intercept coefficients are zero, of which there is only one). Thus, the same p-value for this test would appear in:

(i) The `interventionprogram` row in the coefficients table, and

(ii) The global $\mathcal{F}$-statistic p-value at the bottom of the summary output.

```
## Call:
## lm(formula = adherence_rate ~ intervention, data = tb_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -29.6909  -5.4971  -0.1409   7.2904  22.4091
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          77.5909     0.9741  79.652  < 2e-16 ***
## interventionprogram   5.4249     1.3708   3.958 0.000105 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.692 on 198 degrees of freedom
## Multiple R-squared:  0.0733, Adjusted R-squared:  0.06862
## F-statistic: 15.66 on 1 and 198 DF,  p-value: 0.0001055
```

Question #3

(d) Since the approach in the previous parts used a linear model, it required the equal variance condition to be valid. Write **one line** of R code to perform a test for the same hypotheses as in (a) that **DOES NOT** require the equal variance condition.

```r
t.test(adherence_rate ~ intervention, data = tb_data)
```

Recall that we discussed Welch's two-sample t-test, which is a test to see if group means are equal without assuming equal variances. In this case, our two groups correspond to the patients in each intervention group, and this t-test allows us to do correctly perform the same hypotheses as in part (a) without having to check that the variances across the groups are equal.

(Permutation tests assume exchangeability under the null hypothesis, an assumption stronger than those used for a t-test.)

Question #4

Researchers want to now understand tuberculoses recovery outcomes different according to patients based on HIV status ("positive" or "negative) and undernourishment ("yes" or "no"), fitting

```
model3 <- lm(time_to_recovery ~ HIV_status*
    undernourishment, data=tb_data)
```

(a) Why is an interaction term included in this model?

- ☐ To account for baseline differences in recovery times between HIV status groups.
  FALSE: Baseline differences between groups are already captured by the main effect terms in the model; the interaction term is not needed to model this baseline level and captures how the effects combine.
- ■ To test whether the effect of HIV status on recovery time differs by undernourishment status.
  TRUE: This defines why models might have an interaction term, namely to allow the slope (difference in means due to HIV status) to change depending on undernourishment status.
- ☐ To control for confounding between HIV status and undernourishment.
  FALSE: Interaction terms do not "control for confounding" since confounding is addressed by including both variables as main effects in the model.
- ☐ To test whether both predictors have additive effects on recovery time.
  FALSE: Having additive effects would mean there is no interaction; the interaction term allows for deviations from additivity, not to test for additivity. (The test for additivity would have been $H_0 : \beta_{\text{interaction}} = 0$.)
- ☐ None of the above.
  FALSE: Choice 2 is correct.

Question #4

(b) Fill in the code to calculate mean recovery time for each combination:

```
tb_data %>% group_by(HIV_status, undernourishment) %>%
  summarize(mean_recovery = mean(time_to_recovery))
```

We just use the group_by function on each combination.

(c) The global $\mathcal{F}$-test p-value reported by `summary(model3)` corresponds to which null hypothesis?

- $H_0 : \beta_3 = 0$
  FALSE: This corresponds to the null of a partial test for the interaction only.
- $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$
  TRUE: Global $\mathcal{F}$-test checks whether all non-intercept coefficients are zero, which is exactly what is described here.
- $H_0 : \beta_0 = \beta_1 = \beta_2 = \beta_3 = 0$
  FALSE: The intercept is never included in the global $\mathcal{F}$-test, it can be non-zero.
- $H_0 : \mu_{\text{yes}} = \mu_{\text{no}}$
  FALSE: This represents a marginalized comparison of undernourishment (HIV status and interaction are ignored). The global $\mathcal{F}$-test has to consider all predictors jointly.

Question #4

(d) Suppose researchers want to know whether HIV status has **any** relationship with recovery time, accounting for both the interaction term and undernourishment. If the appropriate p-value is in the `summary(model3)` output, identify where it is. If not, write R code to obtain the appropriate p-value.

```
1 null_model <- lm(time_to_recovery ~ undernourishment,
      data = tb_data)
2 anova(null_model, model3)
```

Given the interaction term, no one coefficient in `summary(model3)` captures relationship between HIV status and recovery time among all possible undernourishment states. To determine if HIV status has any relationship at all, we test $H_0 : \beta_1 = \beta_3 = 0$, and so the null model in line 1 removes those terms. Then, `anova` allows us to compare this reduced, null model against the full one.

```
## Call:
## lm(formula = time_to_recovery ~ HIV_status + undernourishment +
##     HIV_status * undernourishment, data = tb_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.6543 -2.5779 -0.4516  2.2539 19.2511
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         20.7489     0.4794  43.284  < 2e-16 ***
## HIV_statuspositive                   5.8727     0.8027   7.316 6.39e-12 ***
## undernourishmentyes                  1.9054     0.9131   2.087   0.0382 *
## HIV_statuspositive:undernourishmentyes  3.3594     1.4864   2.260   0.0249 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Question #4

(e) When would it be appropriate to utilize the p-value for the `HIV_statuspositive` coefficient of `model3`?

- ☐ Anytime we want to know something about the relationship between HIV status, time to recovery.
  FALSE: Given an interaction, the effect of HIV status cannot be constant under undernourishment groups. Its coefficient p-value would then pertain to a specific undernourishment status, not the overall relationship.

- ☐ When we want to test whether HIV status is associated with recovery time among those that are undernourished.
  FALSE: Among those that are undernourished, the HIV effect is $\beta_1 + \beta_3$. To test whether this effect is associated with recovery time, we would need a linear combination of the coefficients, not the p-value for `HIV_statuspositive`.

- ■ When we want to test whether HIV status is associated with recovery time among those that are not undernourished.
  TRUE: Since "no" is the reference category for undernourishment, β_1 represents the effect of HIV status among those who are not undernourished. Thus the p-value for `HIV_statuspositive` is appropriate for this test.

- ☐ When we want to test whether undernourishment status is associated with recovery time, among those that are HIV positive.
  FALSE: This tests for effect of undernourishment, not effect of HIV status. The HIV coefficient p-value would be unrelated to this test.

- ☐ When we want to test whether undernourishment status is associated with recovery time, among those that are HIV negative.
  FALSE: This tests the effect of undernourishment on HIV-negative patients, which corresponds to the coefficient β_2 (not β_1).

- ☐ There is no situation in which it would be correct to utilize this p-value.
  FALSE: The `HIV_statuspositive` p-value is meaningful for the subgroup corresponding to the reference level of the interacting variable; here, this is patients that are not undernourished.