DSC 152 HW2 Student Answer Key

Question 1

1(a)

```
# Loading in the data as a dataframe
workshop_df <- read.delim("workshop_scores.tsv", sep = "\t")


# Create gain = posttest - pretest
workshop_df$gain <- workshop_df$posttest - workshop_df$pretest

# Create new variables for coding and handwritten gains
coding_gains      <- workshop_df$gain[workshop_df$group == "coding"]
handwritten_gains <- workshop_df$gain[workshop_df$group == "handwritten"]

# Create the requested sample quantities
mean_coding      <- mean(coding_gains)
mean_handwritten <- mean(handwritten_gains)
sd_coding        <- sd(coding_gains)
sd_handwritten   <- sd(handwritten_gains)
obs_diff         <- mean_coding - mean_handwritten
```

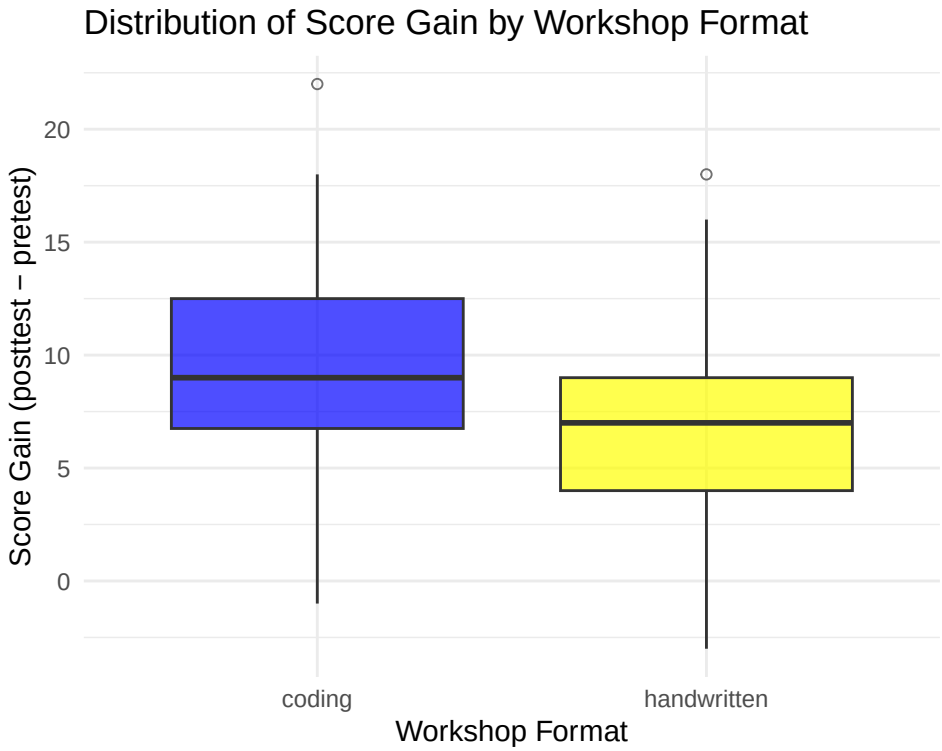
The five required quantities are:

Quantity	Value
Mean gain — coding	9.9063
Mean gain — handwritten	6.5938
SD of gain — coding	5.0248
SD of gain — handwritten	4.6894

Quantity	Value
Observed difference $\bar{x}_{\text{coding}} - \bar{x}_{\text{handwritten}}$	3.3125

1(b)

```
workshop_df %>%
  ggplot( aes(x = group, y = gain, fill = group)) +
  geom_boxplot(alpha = 0.7, outlier.shape = 1) +
  scale_fill_manual(values = c("coding" = "blue",
                               "handwritten" = "yellow")) +
  labs(
    title = "Distribution of Score Gain by Workshop Format",
    x     = "Workshop Format",
    y     = "Score Gain (posttest - pretest)"
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```



The coding group has a higher median gain (approximately 10 points) than the handwritten group (approximately 6 points). The two groups have similar spread, with comparable interquartile ranges, though the handwritten group shows one upper outlier around 21 points. Neither distribution shows severe skewness; both appear roughly symmetric within their respective boxes.

1(c)

The hypotheses are:

$$H_0 : \mu_{\text{coding}} = \mu_{\text{handwritten}} \quad \text{vs.} \quad H_A : \mu_{\text{coding}} \neq \mu_{\text{handwritten}}$$

where μ_{coding} and $\mu_{\text{handwritten}}$ are the true mean score gains in each format. As introduced in Lecture 6, the standard statistical test for a two-group A/B study like this is the **Welch two-sample t-test**, which does not assume equal variances across groups.

```
ttest_result <- t.test(gain ~ group, data = workshop_df)
ttest_result
```

```
    Welch Two Sample t-test

data:  gain by group
t = 2.7263, df = 61.706, p-value = 0.008326
alternative hypothesis: true difference in means between group coding and group handwritten i
95 percent confidence interval:
 0.8835013 5.7414987
sample estimates:
     mean in group coding mean in group handwritten
                  9.90625                   6.59375
```

From the output:

- **Test statistic**: $t = 2.7263$
- **p-value**: 0.008326
- **95% CI for $\mu_{\text{coding}} - \mu_{\text{handwritten}}$**: $(0.884,\ 5.741)$

We reject H_0 ($p = 0.0083$). There is statistically significant evidence at the 0.05 level that the coding and handwritten formats led to different average score gains. Specifically, the coding group gained more on average, with the difference estimated at 3.31 points (95% CI: 0.88 to 5.74 points).

1(d)

Cohen's d using the **pooled standard deviation** is defined as:

$$d = \frac{\bar{x}_{\text{coding}} - \bar{x}_{\text{handwritten}}}{s_p}$$

where the pooled standard deviation is:

$$s_p = \sqrt{\frac{(n_{\text{coding}} - 1)\, s^2_{\text{coding}} + (n_{\text{handwritten}} - 1)\, s^2_{\text{handwritten}}}{n_{\text{coding}} + n_{\text{handwritten}} - 2}}$$

```
n_coding      <- length(coding_gains)
n_handwritten <- length(handwritten_gains)

s_pooled <- sqrt(
  ((n_coding - 1) * sd_coding^2 + (n_handwritten - 1) * sd_handwritten^2) /
  (n_coding + n_handwritten - 2)
)

cohens_d <- obs_diff / s_pooled
cohens_d
```

```
[1] 0.6815792
```

$d = 0.6816$.

The p-value and Cohen's d measure fundamentally different things, and as Lecture 5 emphasized, the p-value is not everything.

- The **p-value** ($p = 0.0083$) answers: "How surprising is this data if H_0 were true?" It depends heavily on sample size: With a large enough n, even a trivially small difference will produce $p < 0.05$.
- **Cohen's** d ($d = 0.68$) answers: "How large is the difference, measured in standard deviation units?" It does not depend on n.

Here, $d \approx 0.68$ is close to the conventional threshold for a medium effect ($|d| = 0.5$) and is approaching large ($|d| = 0.8$). So this is not only statistically significant but also of meaningful practical magnitude: roughly two-thirds of a standard deviation in score gain separates the two formats in this sample.

1(e)

The permutation test proceeds as follows:

1. Write a function that takes a dataframe and returns the test statistic (here, the absolute difference in mean gain between groups).
2. In a loop, shuffle either the group label or the outcome variable and recompute the statistic. Either shuffling approach is valid.
3. Store all shuffled statistics — this is the **null distribution**.
4. The p-value is the proportion of the null distribution that is at least as extreme as the observed statistic.

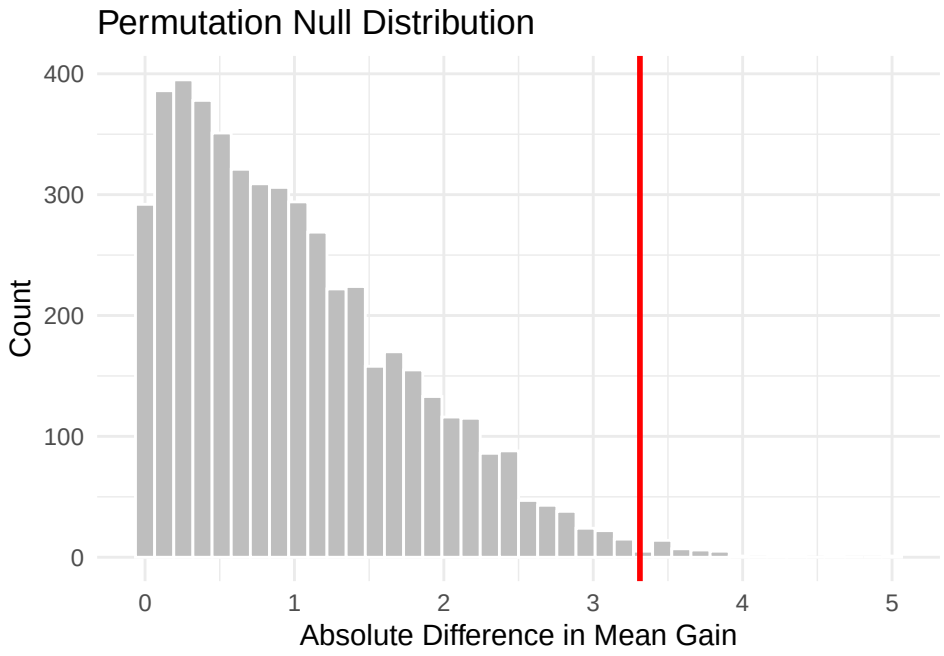
```
diff_in_means <- function(df) {
  means <- tapply(df$gain, df$group, mean)
  abs(means["coding"] - means["handwritten"])
}

obs_stat <- diff_in_means(workshop_df)
obs_stat
```

```
coding
3.3125
```

```
set.seed(152)
n_perms   <- 5000
null_dist <- numeric(n_perms)

for (i in 1:n_perms) {
  shuffled_df       <- workshop_df
  shuffled_df$group <- sample(workshop_df$group)
  null_dist[i]      <- diff_in_means(shuffled_df)
}

perm_pvalue <- mean(null_dist >= obs_stat)
perm_pvalue
```

```
[1] 0.0086
```

```r
null_df <- data.frame(stat = null_dist)

ggplot(null_df, aes(x = stat)) +
  geom_histogram(bins = 40, fill = "grey", color = "white") +
  geom_vline(xintercept = obs_stat, color = "red", linewidth = 1) +
  labs(
    title = "Permutation Null Distribution",
    x     = "Absolute Difference in Mean Gain",
    y     = "Count"
  ) +
  theme_minimal()
```



- **Observed test statistic**: 3.3125
- **Permutation p-value** (seed 152, 5,000 permutations): 0.0086

Comparison with t-test: Both tests address the same null hypothesis and give very similar p-values ($p = 0.0083$ for the t-test, $p = 0.0086$ for the permutation test). Unlike the Welch t-test, the permutation test's validity does not rest on a distributional assumption — it is justified directly by the random assignment of students to groups. The close agreement between the two p-values suggests the t-distribution approximation is performing well for this sample.

1(f)

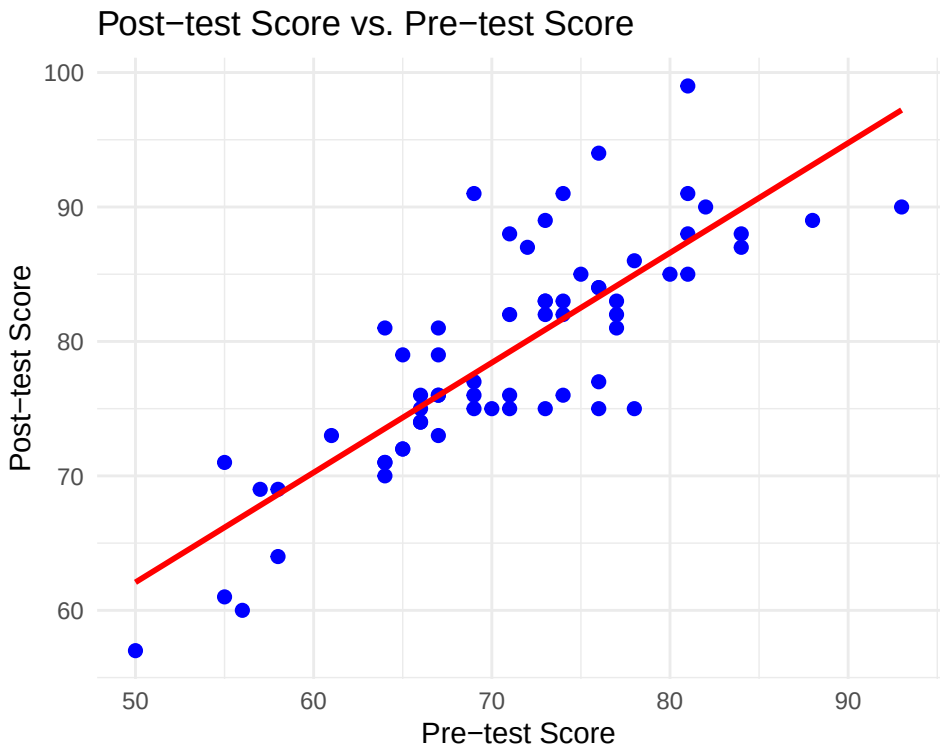
Because the workshop format was randomly assigned to students, causal language is appropriate. Random assignment is the key design feature that allows us to move from association to causation: it ensures the two groups are, on average, comparable on all other variables — both measured and unmeasured — so any observed difference in outcomes can be attributed to the treatment rather than to confounding.

It is therefore reasonable to say: *"The evidence supports that assignment to the coding format caused higher average score gains than assignment to the handwritten format, in this study population."* (Since we rejected H_0, the data are inconsistent with no average treatment effect.)

Question 2

Full solution

```
model_slr <- lm(posttest ~ pretest, data = workshop_df)

ggplot(workshop_df, aes(x = pretest, y = posttest)) +
  geom_point(color = "blue", size = 2) +
  geom_smooth(method = "lm", color = "red", se = FALSE, linewidth = 1) +
  labs(
    title = "Post-test Score vs. Pre-test Score",
    x     = "Pre-test Score",
    y     = "Post-test Score"
  ) +
  theme_minimal()
```



2(b)

```
summary(model_slr)
```

```
Call:
lm(formula = posttest ~ pretest, data = workshop_df)

Residuals:
    Min      1Q  Median      3Q     Max
-9.9570 -3.0132 -0.3793  2.1276 13.3953

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 21.23726    5.26346   4.035 0.000152 ***
pretest      0.81692    0.07369  11.085  2.4e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 4.905 on 62 degrees of freedom
Multiple R-squared:  0.6647,    Adjusted R-squared:  0.6592
F-statistic: 122.9 on 1 and 62 DF,  p-value: 2.396e-16
```

From the `Coefficients` table:

- $\hat{\beta}_1 = 0.8169$
- Standard error of $\hat{\beta}_1 = 0.0737$
- p-value for $H_0 : \beta_1 = 0$: $p < 2.4 \times 10^{-16}$

For each additional point a student scored on the pre-test, their predicted post-test score is 0.82 points higher, on average.

2(c)

```r
confint(model_slr, level = 0.95)
```

```
                2.5 %    97.5 %
(Intercept) 10.7157449 31.758774
pretest      0.6696084  0.964231
```

The 95% CI for β_1 (the `pretest` row) is $(0.670,\ 0.964)$.

Interpretation: We are 95% confident that the true population slope β_1 — the average change in post-test score associated with a one-point increase in pre-test score, for students in the population this sample represents — lies between 0.67 and 0.96 points.

The key is that this is a statement about the **population relationship**, not just the observed sample.

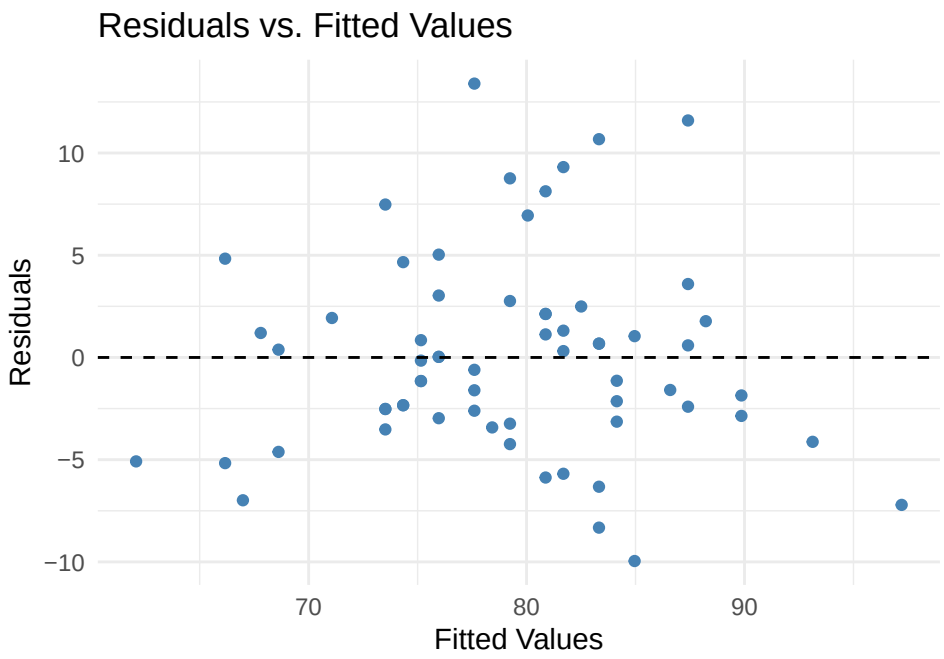
2(d)

As discussed in Lecture 10, residual analysis for simple linear regression checks four conditions:

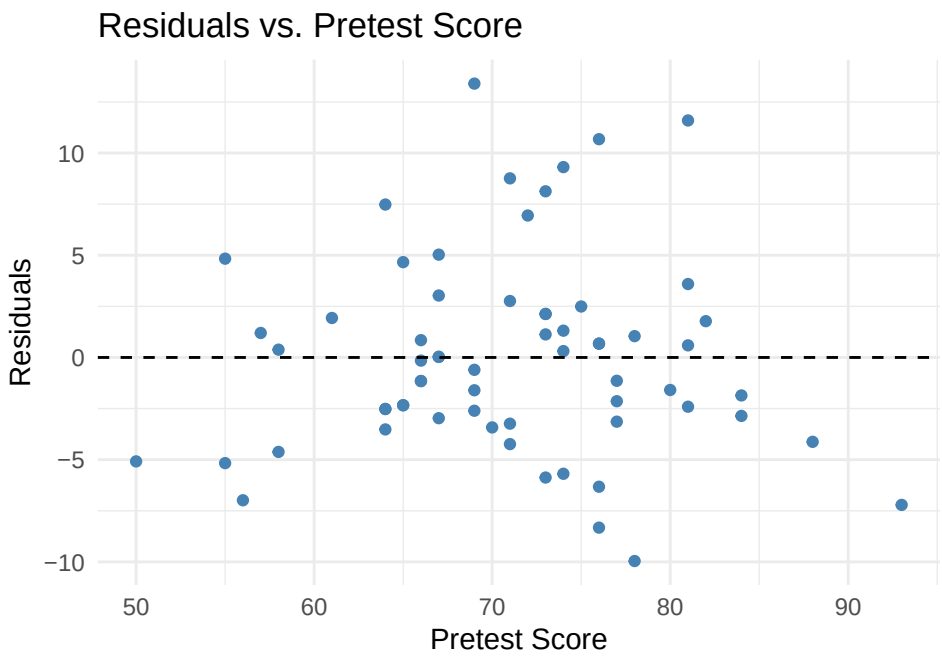
1. **Linearity**: The relationship between predictor and outcome should be linear. Check using a **residuals vs. fitted values** plot — no systematic curved pattern should be present.

2. **Equal variance (homoskedasticity)**: The spread of residuals should be roughly constant across all values of the predictor. Check using a **residuals vs. x values** plot (here, residuals vs. pretest score).
3. **Normality of errors**: Residuals should be approximately normally distributed. Check using all three diagnostics taught in Lecture 10: a **histogram of residuals**, a **Normal Q-Q plot**, and the **Shapiro-Wilk test**.
4. **Independence**: Observations should be independent of each other. The standard diagnostic is a plot of residuals vs. order of entry into the study. Since the order of data collection is not recorded here, this condition must instead be evaluated from the study design: students were separately and randomly assigned individuals, making independence plausible, though it cannot be verified from the data alone.

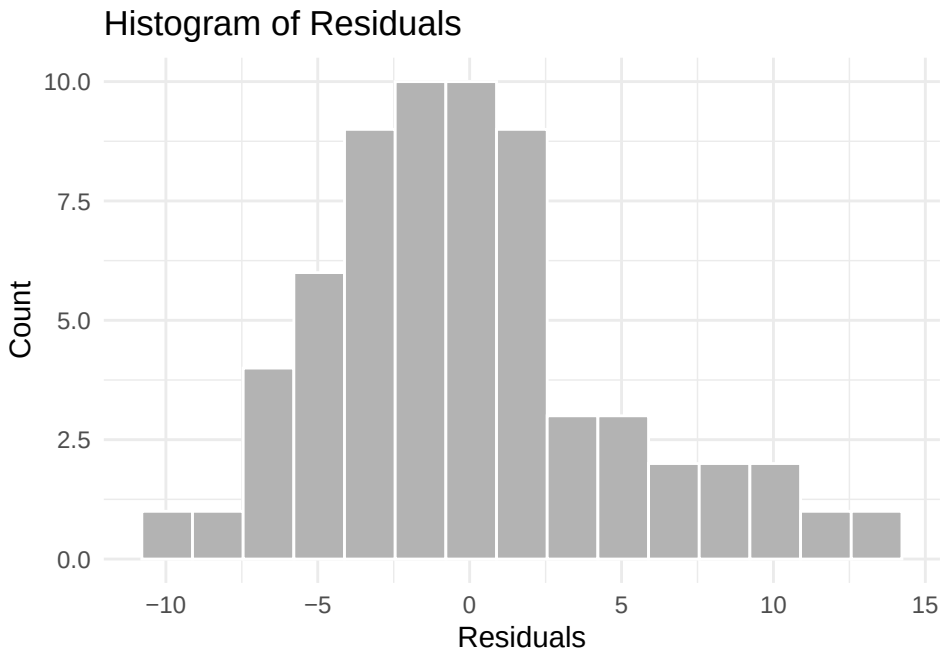
```
resid_slr   <- residuals(model_slr)
fitted_slr  <- fitted(model_slr)
diag_df_slr <- data.frame(
  fitted  = fitted_slr,
  resid   = resid_slr,
  pretest = workshop_df$pretest
)

# Linearity: residuals vs. fitted values
ggplot(diag_df_slr, aes(x = fitted, y = resid)) +
  geom_point(color = "steelblue") +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(title = "Residuals vs. Fitted Values",
       x = "Fitted Values", y = "Residuals") +
  theme_minimal()
```



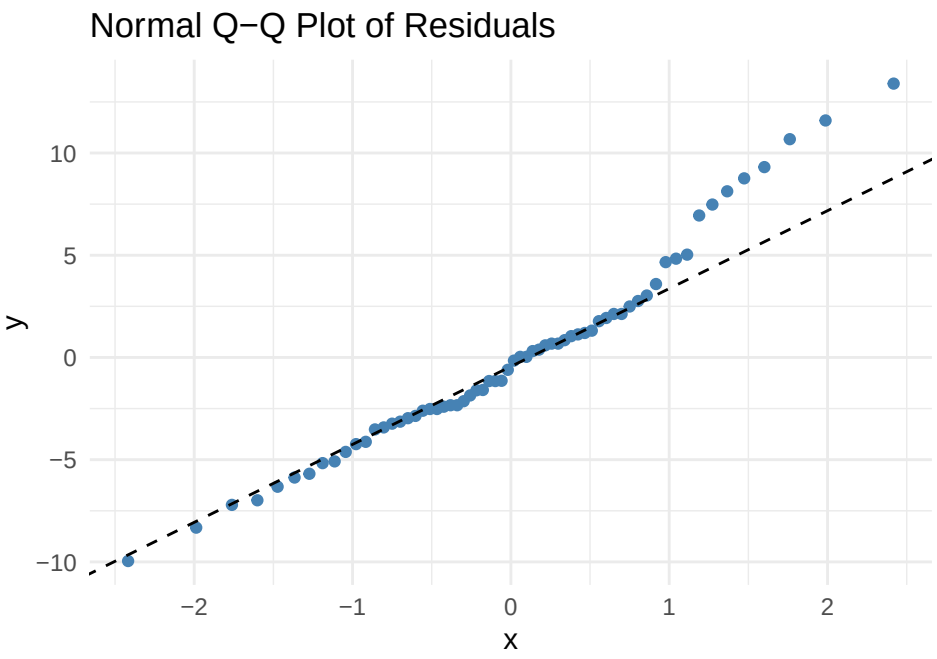

```
# Equal variance: residuals vs. x (pretest)
ggplot(diag_df_slr, aes(x = pretest, y = resid)) +
  geom_point(color = "steelblue") +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(title = "Residuals vs. Pretest Score",
       x = "Pretest Score", y = "Residuals") +
  theme_minimal()
```




```
# Normality: histogram
ggplot(diag_df_slr, aes(x = resid)) +
  geom_histogram(bins = 15, fill = "grey70", color = "white") +
  labs(title = "Histogram of Residuals",
       x = "Residuals", y = "Count") +
  theme_minimal()
```




```
# Normality: QQ plot
ggplot(diag_df_slr, aes(sample = resid)) +
  stat_qq(color = "steelblue") +
  stat_qq_line(linetype = "dashed") +
  labs(title = "Normal Q-Q Plot of Residuals") +
  theme_minimal()
```




```
# Normality: Shapiro-Wilk test
shapiro.test(resid_slr)
```

```
    Shapiro-Wilk normality test

data:  resid_slr
W = 0.96455, p-value = 0.06288
```

- *Residuals vs. Fitted*: No obvious curved pattern; residuals are roughly centered around zero across fitted values. The linearity condition appears reasonably satisfied.
- *Residuals vs. Pretest*: The spread of residuals is roughly consistent across the range of pretest scores, with no clear fan or funnel shape. The equal variance condition appears satisfied.
- *Histogram of Residuals*: The distribution is roughly symmetric and bell-shaped, with no severe skewness. Normality looks plausible.
- *Normal Q-Q*: Points follow the diagonal closely in the middle, with slight departure in the left tail. This is a minor departure unlikely to invalidate inference.
- *Shapiro-Wilk test*: W = 0.96455, p-value 0.06288. This provides some evidence against normality, though the departure appears mild visually; the CLT partially mitigates concerns at this sample size.

- *Independence*: Cannot be assessed from the data since order of entry is not recorded. Plausible by design — students were separately and randomly assigned, so dependence between observations is unlikely.

Overall, no condition appears seriously violated.

Question 3

3(a)

```r
workshop_df$group <- relevel(factor(workshop_df$group), ref = "handwritten")

model_unadj <- lm(posttest ~ group, data = workshop_df)
summary(model_unadj)
```

```
Call:
lm(formula = posttest ~ group, data = workshop_df)

Residuals:
     Min       1Q   Median       3Q      Max
-22.5313  -4.6094  -0.5313   6.1563  19.4687

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  78.8438     1.4960  52.703   <2e-16 ***
groupcoding   0.6875     2.1157   0.325    0.746
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.463 on 62 degrees of freedom
Multiple R-squared:  0.0017,    Adjusted R-squared:  -0.0144
F-statistic: 0.1056 on 1 and 62 DF,  p-value: 0.7463
```

```r
model_adj <- lm(posttest ~ group + pretest, data = workshop_df)
summary(model_adj)
```

```
Call:
```

```
lm(formula = posttest ~ group + pretest, data = workshop_df)

Residuals:
    Min      1Q  Median      3Q     Max
-9.9147 -3.6368 -0.4292  2.2175 11.9965

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.83185    5.25725   3.392  0.00122 **
groupcoding  2.90420    1.19538   2.430  0.01808 *
pretest      0.84446    0.07184  11.754  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.722 on 61 degrees of freedom
Multiple R-squared:  0.6942,    Adjusted R-squared:  0.6842
F-statistic: 69.25 on 2 and 61 DF,  p-value: < 2.2e-16
```

Model	$\widehat{\beta}_{\text{coding}}$	p-value
Unadjusted — handwritten as ref	0.6875	0.746
Adjusted — handwritten as ref	2.9042	0.0181
Unadjusted — coding as ref (default)	-0.6875	0.746
Adjusted — coding as ref (default)	-2.9042	0.0181

3(b)

The **adjusted model** (`posttest ~ group + pretest`) more directly answers the question *"Among students with similar pre-test scores, do the two formats differ in average post-test score?"* The `groupcoding` coefficient in the adjusted model represents the estimated difference in average post-test score between coding and handwritten groups while **holding pre-test score fixed**. The unadjusted model compares raw post-test outcomes between groups without accounting for any pre-test differences, so its `groupcoding` coefficient conflates the format effect with any baseline imbalance.

3(c)

The `groupcoding` coefficient from the adjusted model is $\widehat{\beta}_1 = 2.9042$. This is interpreted as: among students who had the **same pre-test score**, those in the coding group had an average post-test score **2.90 points higher** than those in the handwritten group, on average.

3(d)

Even in a randomized experiment, the **realized sample** may have chance imbalances in pre-test scores between the two groups. If one group starts with higher pre-test scores on average, a raw comparison of post-test means will be misleading: that group will tend to score higher on the post-test simply because of their head start, not because of the treatment.

In this dataset that is exactly what happened: the handwritten group had slightly higher average pre-test scores, which inflated their raw post-test average relative to the coding group and masked the format effect. The unadjusted coefficient (0.69 points, $p = 0.75$) is much smaller and non-significant compared to the adjusted coefficient (2.90 points, $p = 0.018$). The adjusted model controls for this by providing a within-pretest-level comparison — a more precise and less confounded estimate of the format effect.

3(e)

The student's proposed rule — "drop `pretest` if its p-value exceeds 0.05" — is not appropriate when the goal is statistical inference about the group effect. There are two reasons:

First, the purpose of including `pretest` is not to test whether it is significant on its own. It is included because it is a strong predictor of the outcome and is relevant to the research question (comparing groups with similar starting levels). Whether or not `pretest` clears a significance threshold in this particular sample does not change that rationale.

Second, dropping predictors based on their own p-values can distort inference for the remaining coefficients. Specifically:

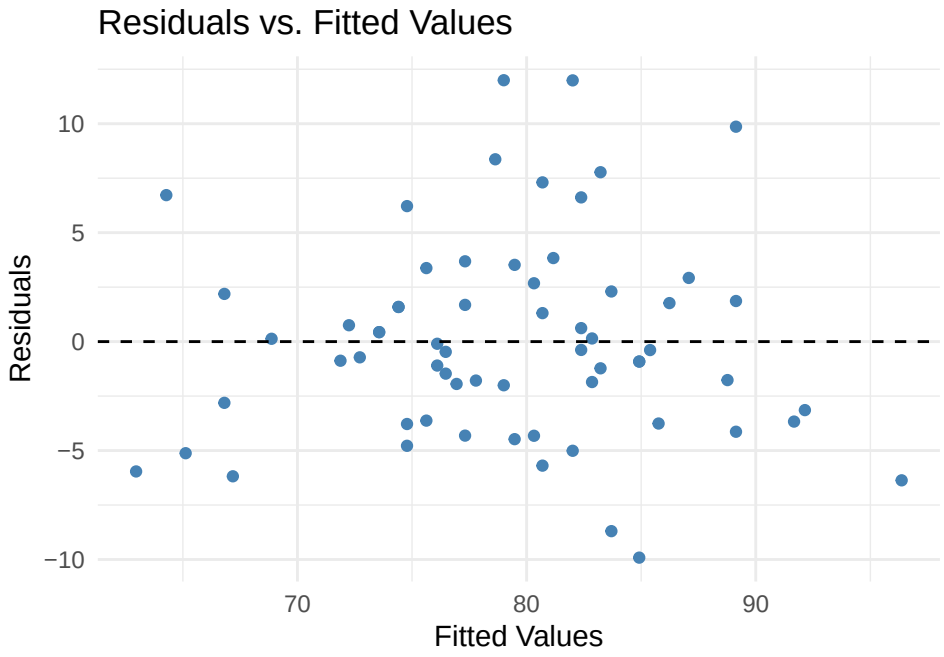
- `pretest` may still meaningfully change the point estimate or standard error of `groupcoding` even if its own p-value exceeds 0.05.
- Performing model selection based on p-values and then doing inference on the selected model inflates the Type I error rate for the tests that remain.

The correct rationale for including or excluding `pretest` is subject-matter / design-based: we know from the study design that a student's pre-test score is a meaningful predictor of their post-test performance. Including it is appropriate regardless of its p-value in this sample. Dropping it would only be defensible if there were a strong domain-specific reason to believe pre-test is irrelevant — not simply because it didn't reach $p < 0.05$ in this dataset.

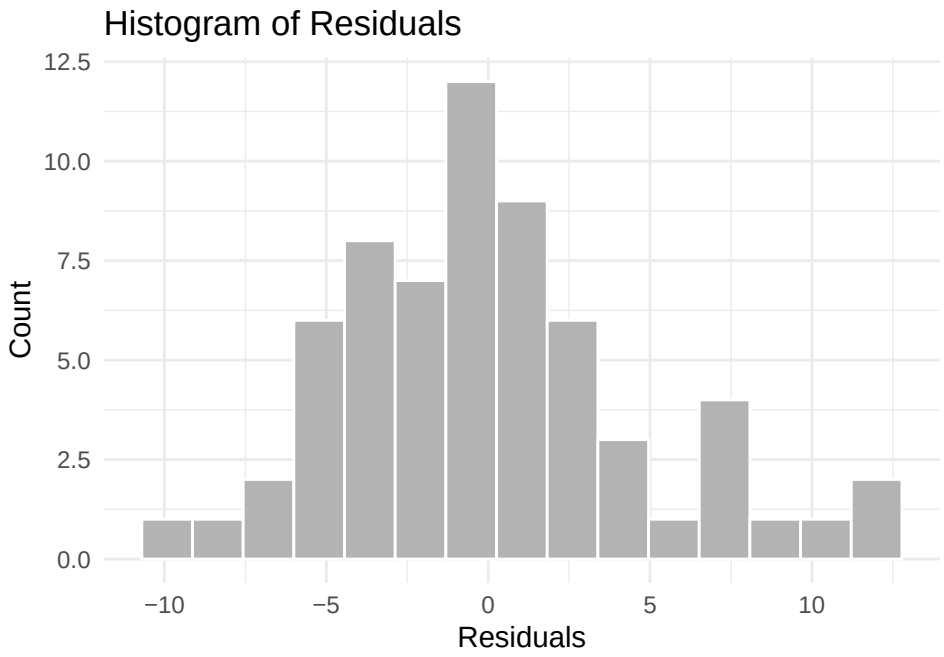
3(f)

For multiple linear regression, the same four conditions apply as in Q2(d), now checked on the **adjusted model** residuals.

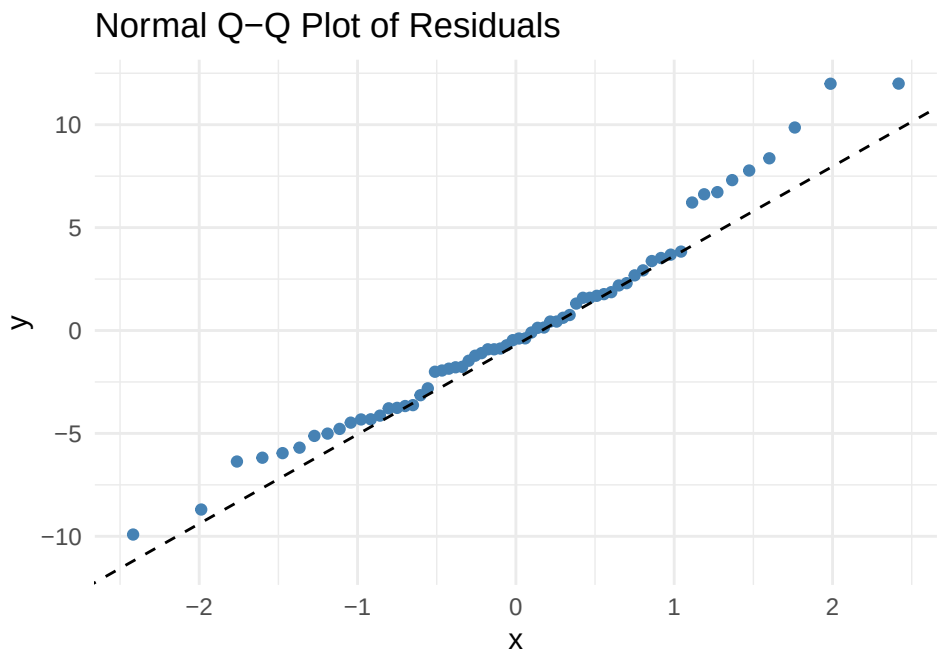
```r
resid_adj   <- residuals(model_adj)
fitted_adj  <- fitted(model_adj)
diag_df_adj <- data.frame(
  fitted  = fitted_adj,
  resid   = resid_adj,
  pretest = workshop_df$pretest
)

ggplot(diag_df_adj, aes(x = fitted, y = resid)) +
  geom_point(color = "steelblue") +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(title = "Residuals vs. Fitted Values",
       x = "Fitted Values", y = "Residuals") +
  theme_minimal()
```




```r
ggplot(diag_df_adj, aes(x = resid)) +
  geom_histogram(bins = 15, fill = "grey70", color = "white") +
  labs(title = "Histogram of Residuals",
       x = "Residuals", y = "Count") +
  theme_minimal()
```




```
ggplot(diag_df_adj, aes(sample = resid)) +
  stat_qq(color = "steelblue") +
  stat_qq_line(linetype = "dashed") +
  labs(title = "Normal Q-Q Plot of Residuals") +
  theme_minimal()
```




```
shapiro.test(resid_adj)
```

```
    Shapiro-Wilk normality test

data:  resid_adj
W = 0.97004, p-value = 0.1216
```

Written commentary:

- *Residuals vs. Fitted*: No systematic curve; residuals are scattered roughly evenly around zero. Linearity and equal variance conditions both appear reasonably satisfied.
- *Histogram of Residuals*: Roughly symmetric and bell-shaped, similar to the SLR model. Normality looks plausible.
- *Normal Q-Q*: Points follow the diagonal with minor departure in the left tail. No severe normality violation.
- *Shapiro-Wilk test*: W = 0.97004, p-value = 0.1216. Interpret as in Q2(d).
- *Independence*: Same argument as Q2(d) — plausible by design; no order-of-entry variable is available to check formally.

The adjusted model satisfies the regression conditions at least as well as the simple linear regression model. No condition appears seriously violated.

3(g)

Based on the adjusted model (`posttest ~ group + pretest`):

Among students with comparable pre-test scores, those in the coding group scored on average **2.90 points higher** on the post-test than those in the handwritten group. This difference was statistically significant at $\alpha = 0.05$ ($p = 0.0181$, 95% CI: roughly 0.51 to 5.30 points). In practical terms, a 2.90-point advantage on a 100-point scale is modest but meaningful in the context of a short workshop intervention; the CI excludes effects smaller than about half a point. Because workshop format was randomly assigned, causal language is appropriate: the evidence supports that assignment to the coding format caused a higher average post-test score, on average, compared to the handwritten format, among students at comparable starting levels.