

For each of the first three problems, there may be more than one correct response. Think of these problems as multiple true/false rather than multiple choice.

Problem 1 (5 points)

A scientist acquires thirty laboratory rats from a supplier. The scientist uses randomization to assign the rats to two separate treatment groups of fifteen rats each. One group receives a calcium dietary supplement and the other group (the control group) does not. The mean concentration of zinc in the blood of the rats receiving the calcium dietary supplement is higher than the mean of the control group, and there is strong, significant statistical evidence (from both a permutation test and a t test) that the difference is not easily explained by chance variation.

CIRCLE ALL TRUE RESPONSES.

- (a) It is justified to conclude that in these rats, the calcium dietary supplement caused an increase in zinc concentration in the blood.
- (b) The rats in the experiment were randomly sampled from a population.
- (c) It is justified, based on this statistical analysis alone, to conclude that, on average, a calcium dietary supplement would increase zinc concentration in blood in a large population of similar rats.
- (d) It is justified, based on this statistical analysis alone, to conclude that, on average, a calcium dietary supplement would increase zinc concentration in blood in humans.

Solution:

- (a) TRUE. The groups are decided by randomization, so it is justified to assume causality.
- (b) FALSE. The rats come from a specific supplier, and were not randomly sampled in any formal sense.
- (c) FALSE. Any justification for generalization of results to other groups of rats depends on outside information or subjective judgment about how representative these rats are of the population of interest. The rats were not randomly sampled, and so generalization to other rats is speculative. (In this setting, I would expect the results to generalize, but this expectation is based on subjective judgment, not *the statistical analysis alone*.)
- (d) FALSE. Generalization to a completely different species is even more speculative than to other groups of rats. Many results do carry over to humans from animal studies, but some do not. My subjective judgment is less certain that the results would generalize to humans, but this cannot be justified by the statistical analysis alone.

Problem 2 (5 points)

In a simple linear regression problem, examination of residual plots indicates that a model $\mu\{\log Y|X\} = \beta_0 + \beta_1 X$ fits well for the observed range of X data between 10 and 60. **CIRCLE ALL TRUE RESPONSES.**

- (a) A 95% prediction interval for $\log Y$ for a new observation with $X = 40$ will be wider than a 95% confidence interval for $\mu\{\log Y|X = 40\}$.
- (b) The lower and upper endpoints of a confidence interval for $\mu\{\log Y|X = 40\}$ will be an equal distance below and above the estimated mean $\hat{\beta}_0 + (\hat{\beta}_1 \times 40)$.
- (c) The lower and upper endpoints of a 95% prediction interval for a new observation Y with $X = 40$ will be an equal distance below and above the predicted value for Y .
- (d) The formula for the standard error in the estimate for $\mu\{\log Y|X = 140\}$ is wider than the formula for the standard error in the estimate for $\mu\{\log Y|X = 40\}$ (see page 187) because it accounts for the possibility that the linear relationship observed in the range of the data may not extend beyond this range.

Solution:

- (a) TRUE. Prediction intervals are predictions for the next observation and include uncertainty in the actual mean as well as an estimate of the inherent individual variability. Confidence intervals only estimate uncertainty in the mean.
- (b) TRUE. The confidence interval for the mean of the transformed response variable is centered at the estimate.
- (c) FALSE. After transforming back to the original scale from the log scale, the estimate is no longer centered between the endpoints.
- (d) FALSE. The formula is wider because uncertainty in the slope has a larger effect far from the mean than close to the mean, but the formula assumes that a linear relationship extends.

Problem 3 (5 points)

In a study that predicts IQ test score for 8-year-olds who were born prematurely using several explanatory variables including *mother's education*, the estimated coefficient for an indicator variable that the child drank breast milk as an infant was 8.3 and was significant with $p < 0.0001$. **CIRCLE ALL TRUE RESPONSES.**

- (a) If we changed the model by removing or adding some explanatory variables, we would still expect the estimated coefficient for breast milk to be close to 8.3.
- (b) This study provides strong evidence that giving an infant born prematurely breast milk causes an increase in IQ at age 8.
- (c) If an infant born prematurely received half of its food as breast milk and half as formula as an infant, we would expect the increase in IQ at age 8 to be only about 4 points.
- (d) If the study infants who received breast milk were randomly determined instead of selected by their parents' choices, the results of the study would be stronger.
- (e) It is proper to include the variable *mother's education*, which is categorized to five levels from a low of 1 to a high of 5, as a single quantitative variable with these values.

Solution:

- (a) FALSE. Regression coefficients only have meaning with respect to the other variables included in the model.
- (b) FALSE. Mothers make their own decisions about whether or not to breast feed. We are not justified in inferring causality from observational studies.
- (c) FALSE. This conclusion is based on the assumptions of causality as well as a linear effect of IQ based on the proportion of breast milk in the diet. Neither of these assumptions is justified.
- (d) TRUE. There would be more evidence of a causal relationship if the treatment condition were randomly assigned.
- (e) FALSE. Categorical variables should not be treated as quantitative variables. While $5 - 4 = 1$ and $2 - 1 = 1$, it makes less sense to say that the difference between a postgraduate education and a college education is the same as the difference between a high school education and an eight grade education, for example.

For these five problems, it suffices to circle **True** or **False**. If you wish to add *one sentence* of explanation, this might result in some partial credit if your true/false response is incorrect.

Problem 4 (5 points)**True or False:**

In a one-way analysis of variance with a categorical variable with three levels, three t tests to separately test null hypotheses $\mu_1 = \mu_2$, $\mu_1 = \mu_3$, $\mu_2 = \mu_3$ are equivalent to a single F test with null hypothesis $\mu_1 = \mu_2 = \mu_3$.

Solution: **False.** An F test is not equivalent to a set of t tests.

Problem 5 (5 points)**True or False:**

A log transformation of the response variable can be a remedy when a plot of residuals versus fitted values shows a wedge shaped pattern with larger spread for larger fitted values.

Solution: **True.** This is not guaranteed to work, but log transformations lessen the differences between large values and can result in data that comes closer to the model assumptions.

Problem 6 (5 points)**True or False:**

When comparing different multiple regression models, one should always prefer the model with the smallest residual sum of squares.

Solution: **False.** The residual sum of squares will always decrease when additional variables are added, but a model with so many variables that it fits the data perfectly is usually not a good model.

Problem 7 (5 points)**True or False:**

When a categorical variable with five levels is added to a multiple regression model, there are five dummy variables added to the model matrix (design matrix).

Solution: **False.** Only four dummy variables would be added (assuming an intercept is already in the model.).

Problem 8 (5 points)**True or False:**

In a regression model $\mu\{Y|X\} = \beta_0 + \beta_1 X + \beta_2 X^2$, the coefficient $\hat{\beta}_1$ is an estimate of the change in Y given a unit change in X keeping all other explanatory variables fixed.

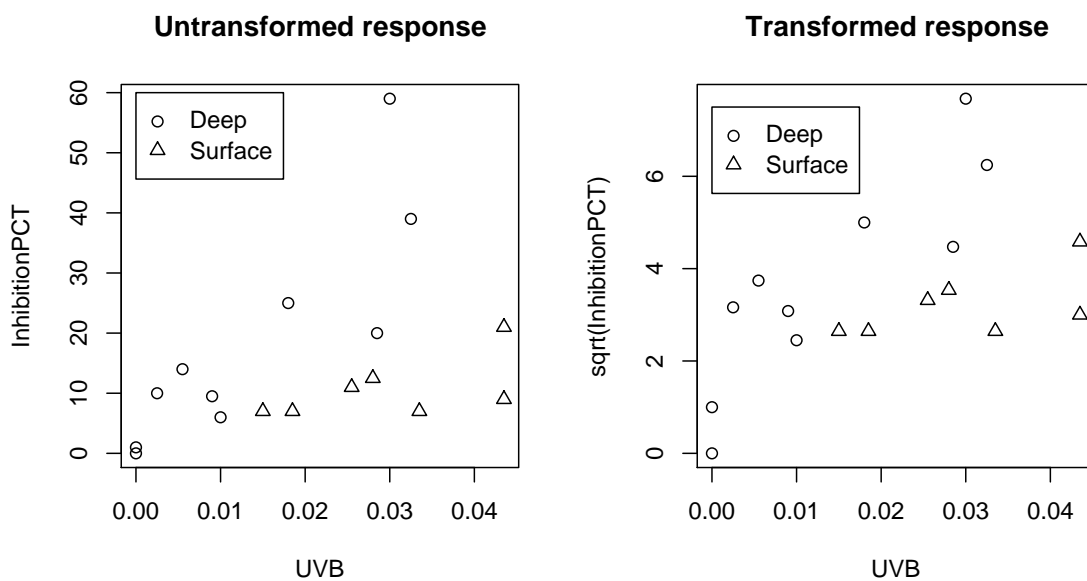
Solution: You cannot keep X^2 fixed while changing X .

On page 299 of your textbook, Exercise 26 describes a data set in which a measure of ultraviolet radiation that reaches the earth's surface is an explanatory variable for the percentage inhibition of phytoplankton production. There are ten sites where measurements are reported at deep water and seven where the measurements are at the surface. The data is on page 300 of your textbook and repeated here.

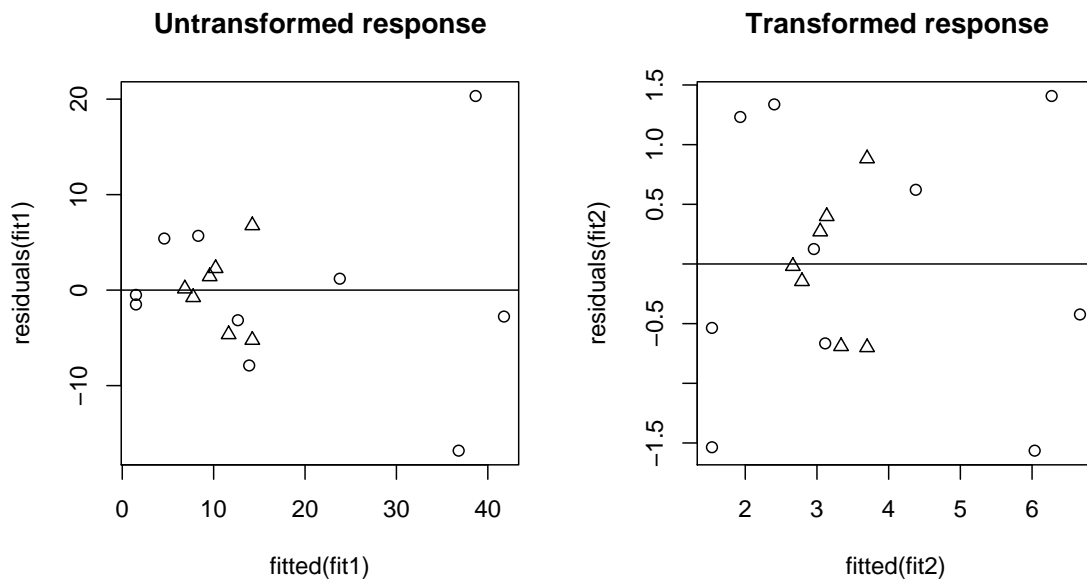
```
> ex1026
```

	InhibitionPCT	UVB	Depth
1	0.0	0.0000	DEEP
2	1.0	0.0000	DEEP
3	6.0	0.0100	DEEP
4	7.0	0.0150	SURFACE
5	7.0	0.0185	SURFACE
6	7.0	0.0335	SURFACE
7	9.0	0.0435	SURFACE
8	9.5	0.0090	DEEP
9	10.0	0.0025	DEEP
10	11.0	0.0255	SURFACE
11	12.5	0.0280	SURFACE
12	14.0	0.0055	DEEP
13	20.0	0.0285	DEEP
14	21.0	0.0435	SURFACE
15	25.0	0.0180	DEEP
16	39.0	0.0325	DEEP
17	59.0	0.0300	DEEP

Here are side-by-side coded scatter plots of the data. The plot on the left is the original data. The plot on the right uses a square root transformation of the response variable.



The next two plots are residual plots for the untransformed and transformed data after fitting a model in each case with UVB , $Depth$, and their interaction as explanatory variables.



Problem 9 (10 points)

Briefly explain why a statistician may prefer using the square-root-transformed response variable instead of the response variable measured on the original scale. Briefly explain why a square root transformation is preferable to a log transformation for this data set.

Solution: The residual plot shows more variability among residuals with large fitted values than small fitted values. There is more constant variance of residuals for the square root transformed data. The log transformation is not a good choice in this example because one of the response values is 0, for which the log is not defined. (Sometimes people use $\log(1 + Y)$ in this situation.)

Problem 10 (10 points)

In the residual plot for the transformed data, the triangles (surface points) are less spread out than the circles (deep points) in the horizontal direction. (a) Provide a brief explanation based the scatter plot on the previous page. (b) The triangles are also less spread out than the circles in the vertical direction. Does this observation shed some doubt on a model assumption? If so, which one? Briefly explain.

Solution:

- The horizontal axis in the residual plot shows the fitted values. The triangles are more closely bunched horizontally because the range of their response values (and hence the fitted values) is much smaller than those of the circles.
- The model fits two lines with independent slopes and intercepts. The lesser spread in the vertical direction indicates that the triangle points are more tightly clustered around their line than the circles are around theirs. This sheds doubt on the assumption of equal variance for all observations. A model that allowed different variances for surface and deep measurements may be more appropriate. (This would be equivalent to fitting two separate lines. The estimated coefficients would be identical, but the estimates of sigma would not pool information, so the standard errors and p-values would be different from their values in the combined analysis.)

We will continue with the analysis of the transformed data only. Here is a summary of the estimated model coefficients.

```
> summary(fit2)
```

Call:

```
lm(formula = sqrt(InhibitionPCT) ~ UVB * Depth)
```

Residuals:

Min	1Q	Median	3Q	Max
-1.56431	-0.66538	-0.01843	0.62175	1.40781

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.5356	0.4834	3.177	0.00729 **
UVB	157.9232	26.5715	5.943	4.88e-05 ***
DepthSURFACE	0.5838	1.2559	0.465	0.64976
UVB:DepthSURFACE	-121.6040	45.4698	-2.674	0.01910 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.015 on 13 degrees of freedom

Multiple R-Squared: 0.7412, Adjusted R-squared: 0.6815

F-statistic: 12.41 on 3 and 13 DF, p-value: 0.0004086

Problem 11 (10 points)

Write a single equation for the response variable $\sqrt{\text{inhibitionPCT}}$ in terms of the explanatory variables using the symbols β_0, β_1 , and so on in the order in which these variables appear in the summary. (It may be useful to let *surface* be an indicator variable measurements at the surface depth.) Then, write two separate equations using numerical values for $\sqrt{\text{inhibitionPCT}}$ as a function of *UVB*, one each for deep and surface measurements.

Solution:

$$\sqrt{\text{inhibitionPCT}} = \beta_0 + \beta_1(\text{UVB}) + \beta_2(\text{surface}) + \beta_3(\text{UVB} \times \text{surface}) + (\text{error})$$

For deep measurements,

$$\sqrt{\text{inhibitionPCT}} = 1.5356 + 157.9232(\text{UVB})$$

while for surface measurements,

$$\sqrt{\text{inhibitionPCT}} = 2.1194 + 36.3192(\text{UVB}).$$

Problem 12 (15 points)

Conduct a formal hypothesis test for the null hypothesis that the slope of the regression lines of inhibition percentage versus UVB for deep and surface measurements are the same versus the alternative that the slope for deep measurements is larger. (a) State the null and alternative hypotheses in terms of the regression coefficients β_i you defined in the previous problem. (b) Report a p-value for this test. (c) Briefly summarize the results in the

context of the problem. (Say something about the relationship between UVB, depth, and the percentage inhibition of phytoplankton production.)

Solution:

- (a) $H_0: \beta_3 = 0$ versus $H_a: \beta_3 < 0$. $t = -2.674$.
- (b) The one-sided p-value is half the computed two-sided p-value. $p = 0.0190/2 = 0.00955$.
- (c) There is strong statistical evidence that UVB has a greater inhibiting effect on phytoplankton production in deep water than at the surface.

Problem 13 (5 points)

One of the estimated regression coefficients has a nonsignificant p-value. This means that there is little evidence that the intercepts for the two lines are different. If the intercepts were the same, the predicted values of the response variable would be the same when UVB had the value 0 for both deep and surface measurements. In the context of the problem, briefly explain why might we expect this to be the case.

Solution: Here is a possible explanation. If there were no UVB, then we would not expect any inhibition due to UVB for either surface or deep measurements.

Problem 14 (10 points)

Below is the variance-covariance matrix of the estimated regression coefficients. In Problem 11, you found a numerical estimate of the slope of the regression line for the surface measurements. Find the standard error for this estimate. (Before starting, recognize that the numerical estimate of this slope is a sum of two estimated regression coefficients. Equations on page 288 or my related notes may be useful.)

```
> vcov(fit2)
```

	(Intercept)	UVB	DepthSURFACE	UVB:DepthSURFACE
(Intercept)	0.2337016	-9.60224	-0.2337016	9.60224
UVB	-9.6022397	706.04704	9.6022397	-706.04704
DepthSURFACE	-0.2337016	9.60224	1.5773162	-49.95979
UVB:DepthSURFACE	9.6022397	-706.04704	-49.9597899	2067.50656

Solution:

$$SE(\hat{\beta}_1 + \hat{\beta}_3) = \sqrt{706.04704 + 2067.50656 + 2(-706.04704)} = 36.9.$$