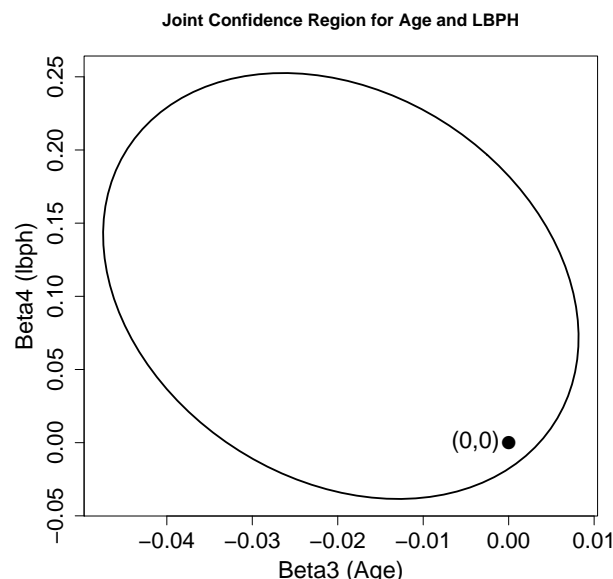


1. Faraway 3.1. For the `prostate` data, fit a model with `lpsa` as the response and the other variables as predictors.

(a) Compute 90 and 95% CIs for the parameter associated with `age`. Using just these intervals, what could we have deduced about the p -value for `age` in the regression summary?

The 90% CI for the age coefficient is $(-0.0382, -0.0011)$, and the 95% CI is $(-0.0418, 0.0026)$. Since 0 is in the first but not the second, we conclude that a two-sided t -test for $H_0 : \beta_{age} = 0$ is between .05 and .1 providing moderate evidence of significance.

(b) Compute and display a 95% joint confidence region for the parameters associated with `age` and `lbph`. Plot the origin on this display. The location of the origin on the display tells us the outcome of a certain hypothesis test. State that test and its outcome.



Whether the origin is within the confidence region tests the joint significance of age and lbph in the model - specifically, $H_0 : \beta_3 = \beta_4 = 0$. Since $(0, 0)$ is within the region, we conclude (at $\alpha = .05$ level) that there is no statistical evidence to suggest that age and lbph are jointly significant in the model.

(c) Suppose a new patient with the following values arrives:

lcpvol	lweight	age	lbph	svi	lcp	gleason	pgg45
1.44692	3.62301	65.00000	0.30010	0.00000	-0.79851	7.00000	15.00000

Predict the `lpsa` for this patient along with an appropriate 95% CI.

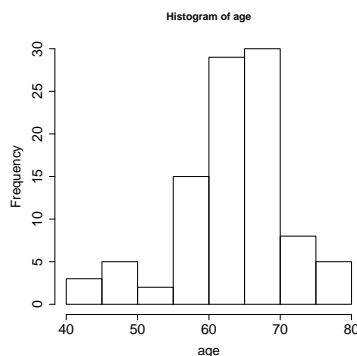
Let \mathbf{x}_0 be the $1 \times p$ vector of values given in the table above (with a one prepended). Then the 95% CI is given by

$$\mathbf{x}_0 \hat{\boldsymbol{\beta}} \pm t_{.975}(n-p) \sqrt{\text{MSE} \cdot \mathbf{x}_0' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_0} \approx 2.389 \pm 0.217 \text{ or } (2.172, 2.606).$$

(d) Repeat the last question for a patient with the same values except that he or she is age 20. Explain why the CI is wider.

In this case, let \mathbf{x}_1 denote the explanatory variables, then

$$\mathbf{x}_1 \hat{\boldsymbol{\beta}} \pm t_{.975}(n-p) \sqrt{\text{MSE} \cdot \mathbf{x}_1' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_1} \approx 3.272 \pm 1.012 \text{ or } (2.260, 4.285).$$



The increase in uncertainty is due to the lack of data points for ages near 20. This is evident in the histogram of ages to the left. As indicated in the book, numerical extrapolation leads to increases in uncertainty.

(e) In the text, we made a permutation test corresponding to the F -test for the significance of all the predictors. Execute the permutation test corresponding to the t -test for age in this model. (Hint: `summay(g)$coef[4,3]` gets you the t -statistic you need if the model is galled `g`.)

We fit the model with 4999 permutations of the age variable and compare the F -statistic given by squaring the t -statistic for the age coefficient and compare it to the unpermuted coefficient's F -statistic. This resulted in a p-value of 0.0838, providing slight evidence for significance.

2. Faraway 3.3. Using the `teengamb` data, fit a model with `gamble` as the response and the other variables as predictors.

(a) Which variables are statistically significant?

In the full model, the Coefficients Table reports:

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	22.55565	17.19680	1.312	0.1968
sex	-22.11833	8.21111	-2.694	0.0101 *
status	0.05223	0.28111	0.186	0.8535
income	4.96198	1.02539	4.839	1.79e-05 ***
verbal	-2.95949	2.17215	-1.362	0.1803

This indicates that sex and income are significant in the model with each predictor.

(b) What interpretation should be given to the coefficient for `sex`?

Since the sex codes are 0=male and 1=female, we conclude that holding all other predictors constant, on average male gambling expenditure is about £22 more than females'.

(c) Predict the amount that a male with average (given these data) status, income, and verbal score would gamble along with an appropriate 95% CI. Repeat the prediction for a male with maximal values (for this data) of status, income and verbal score. Which CI is wider and why is this result expected?

The average male predictor is given by $\bar{\mathbf{x}}_m \approx (1, 0, 45.234, 4.642, 6.660)'$, and the 95% CI for its gambling expenditure is given by

$$\bar{\mathbf{x}}_m' \hat{\boldsymbol{\beta}} \pm t_{.975}(n-p) \sqrt{\text{MSE} \cdot \bar{\mathbf{x}}_m' (\mathbf{X}'\mathbf{X}) \bar{\mathbf{x}}_m} \approx 28.24 \pm 9.46 \text{ or } (18.78, 37.70).$$

(d) Fit a model with just `income` as a predictor and use an F -test to compare it to the full model.

The ANOVA model comparison test gives

$$F = \frac{\text{RSS}_\omega - \text{RSS}_\Omega}{\text{MSE}} \approx 4.134 \quad \text{and} \quad p = 0.012.$$

So, there is statistical evidence that sex, status, and verbal score add explanatory significance to the model.

3. Faraway 3.4. Using the `sat` data:

(a) Fit a model with *total* sat score as the response and *expend*, *ratio* and *salary* as predictors. Test the hypothesis that $\beta_{\text{salary}} = 0$. Test the hypothesis that $\beta_{\text{salary}} = \beta_{\text{ratio}} = \beta_{\text{expend}} = 0$. Do any of these predictors have an effect on the response?

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1069.234	110.925	9.639	1.29e-12 ***
expend	16.469	22.050	0.747	0.4589
ratio	6.330	6.542	0.968	0.3383
salary	-8.823	4.697	-1.878	0.0667 .

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

Residual standard error: 68.65 on 46 degrees of freedom
Multiple R-squared: 0.2096, Adjusted R-squared: 0.1581
F-statistic: 4.066 on 3 and 46 DF, p-value: 0.01209

From the `summary` command on the linear model, we test $H_0 : \beta_{\text{salary}} = 0$ and see that β_{salary} marginally significant ($p = 0.0667$) in the full model. The test for model significance, $H_0 : \beta_{\text{salary}} = \beta_{\text{ratio}} = \beta_{\text{expend}} = 0$, provides evidence contrary to H_0 - that is, there is statistical evidence suggesting significant explanatory power of the model. We note the sample size is 50, and thus the statistical significance may not be practically significant especially in the absence of a scientific explanation.

(b) Now add *takers* to the model. Test the hypothesis that $\beta_{\text{takers}} = 0$. Compare this model to the previous one using an F -test. Demonstrate that the F -test and t -test here are equivalent.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1045.9715	52.8698	19.784	< 2e-16 ***
expend	4.4626	10.5465	0.423	0.674
ratio	-3.6242	3.2154	-1.127	0.266
salary	1.6379	2.3872	0.686	0.496
takers	-2.9045	0.2313	-12.559	2.61e-16 ***

...

> anova mdl2, mdl1

Analysis of Variance Table

Model 1: total ~ expend + ratio + salary + takers

Model 2: total ~ expend + ratio + salary

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	45	48124				
2	46	216812	-1	-168688	157.74	2.607e-16 ***

Note that the percentage of those eligible to take the test is highly significant in the full model with a p-value of 2.61×10^{-16} . The t -test and F -test produce an identical p-value. In fact,

```
> anova mdl2, mdl1 [2,5] - summary mdl2 $ coefficients [5,3]^2
[1] 0
```

4. For each of the following regression models, indicate whether it is a general linear regression model. If it is not, state whether it can be expressed in the general linear model form:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i$$

(a) $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 \log_{10} x_{i2} + \beta_3 x_{i1}^2 + \epsilon_i$

This is a linear regression model.

(b) $y_i = \epsilon_i \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}^2)$

If we log-transform the response, then we have the general linear regression model

$$\log y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}^2 + \log \epsilon_i.$$

(c) $y_i = \beta_0 + \log_{10}(\beta_1 x_{i1}) + \beta_2 x_{i2} + \epsilon_i$

This problem is *not* linear due to the term $\log_{10}(\beta_1 x_{i1})$, and there is no obvious transformation that will make the model linear.

(d) $y_i = \beta_0 \exp(\beta_1 x_{i1}) + \epsilon_i$

This problem is also *not* linear due to the term $\beta_0 \exp(\beta_1 x_{i1})$, and a log transformation will not work as in (b) since the error is additive.

(e) $y_i = [1 + \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}^2 + \epsilon_i)]^{-1}$

Consider the transformation $g(y) = \log(y^{-1} - 1)$. Then, the transformed model is

$$g(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}^2 + \epsilon_i,$$

which is a general linear regression model.

5. The information below relates a response variable y , a second measurement on wood volume, to four explanatory variables defined as: x_1 = a first measurement on wood volume, x_2 = the number of trees, x_3 = the average age of trees, and x_4 = the average volume per tree. Note that $x_4 = x_1/x_2$. Some of the information in the coefficients table and Analysis of Variance table below has not been reported, so that you can figure it out on your own.

					Analysis of Variance Table					
Coefficients Table					Source	df	SS	MS	F	p-val
Predictor	$\hat{\beta}_k$	$SE(\hat{\beta}_k)$	t	p-val	Regression	4	887994			0.000
Intercept	23.45	14.90		0.122	Error					
x_1	0.93209	0.08602		0.000	Total		902773	16718.02		
x_2		0.4721	1.5554	0.126	Sequential Sums of Squares					
x_3	-0.4982	0.1520		0.002	Source	df	Seq. SS			
x_4	3.486	2.274		0.132	x_1	1	883880			
					x_2	1	183			
					x_3	1	3237			
					x_4	1	694			

(a) How many observations are in these data?

$$n = \text{TSS}/\text{MST} + 1 = (902773/16718.02) + 1 = 55.$$

(b) What is R^2 for this model?

R^2 is given by the regression sum of squares (SSReg) over the total sum of squares (TSS). From the partial ANOVA table we calculate

$$\frac{\text{SSReg}}{\text{TSS}} = \frac{887994}{902773} \approx 0.9836.$$

(c) What is the mean squared error?

$$\text{MSE} = \frac{\text{RSS}}{n - p} = \frac{\text{TSS} - \text{SSReg}}{n - p} = \frac{902773 - 887994}{55 - 5} = 295.58$$

(d) Give a 95% confidence interval for β_2 .

The t -statistic in the Coefficients Table tests if the coefficient is 0, hence, is given by

$$t^* = \frac{\hat{\beta}_2}{\text{SE}(\hat{\beta}_2)} \iff \hat{\beta}_2 = \text{SE}(\hat{\beta}_2) \cdot t^* \approx 0.4721 \cdot 1.5554 \approx 0.734.$$

The confidence interval (with Bonferroni correction) is of the form

$$\hat{\beta}_2 \pm t_{1-0.5/(2 \cdot 5)}(50) \cdot \text{SE}(\hat{\beta}_2) \approx 0.734 \pm 1.264. \text{ or } (-0.530, 1.998)$$

(e) Test the null hypothesis $H_0 : \beta_3 = 0$.

The fourth row of the Coefficients Table tests this hypothesis, and the reported p-value is 0.002 which gives fairly strong evidence that the coefficient is 0.

(f) Test the null hypothesis $H_0 : \beta_1 = 1$. Why might this hypothesis be of interest?

We calculate

$$t^* = \frac{\hat{\beta}_1 - 1}{\text{SE}(\hat{\beta}_1)} \approx -0.789$$

and calculate the p-value for two-sided test $p = 2P(t \leq t^*) \approx 0.434$. Hence, we conclude that there is no evidence to suggest that $\beta_1 \neq 1$. This is not surprising since x_1 was a second measurement of the response. In some sense, this test indicates that the methods for measuring the response and second measurement agree.

(g) Give the F -statistic for testing $H_0 : \beta_3 = 0$ relative to the full model.

It can be shown [Renchner 204-205 and HW1.8a] that the F -statistic is given by squaring the associated t -statistic from the Coefficients Table. I.e.

$$F = t^2 = \left(\frac{\hat{\beta}_3}{\text{SE}(\hat{\beta}_3)} \right)^2 = \left(\frac{-0.4982}{0.1520} \right)^2 = 10.74287.$$

(h) Give $R(\beta_3|\beta_1, \beta_2)$ and find $R(\beta_3|\beta_1, \beta_2, \beta_4)$.

From the Sequential Sums of Squares table, we can read directly

$$R(\beta_3|\beta_1, \beta_2) = \text{SSReg}(x_3|x_1, x_2) = 3237.$$

On the other hand, we can recover $R(\beta_3|\beta_1, \beta_2, \beta_4)$ from the F -statistic calculated in (g). That is,

$$F = \frac{R(\beta_3|\beta_1, \beta_2, \beta_4)}{\text{MSE}} \iff R(\beta_3|\beta_1, \beta_2, \beta_4) = \text{MSE} \cdot F \approx 295.58 \cdot 10.74287 \approx 3175.377$$

(i) Test the model with only variables x_1 and x_2 against the model with all of the variables x_1, x_2, x_3, x_4 .

The relevant F -statistic is

$$\begin{aligned} F &= \frac{\text{SSReg}(x_3, x_4|x_1, x_2)/2}{\text{MSE}(x_1, x_2, x_3, x_4)} = \frac{(\text{SSReg}(x_3|x_1, x_2) + \text{SSReg}(x_4|x_1, x_2, x_3))/2}{\text{MSE}(x_1, x_2, x_3, x_4)} \\ &= \frac{(3237 + 694)/2}{295.5} = 6.650. \end{aligned}$$

The p-value for the test is practically 0.0027, hence significant evidence exists for the joint explanatory power x_3 and x_4 .

(j) Test the model with only variables x_1 and x_2 against the model with variables x_1, x_2 , and x_3 .

Note that $\text{RSS}(x_1, x_2, x_3) = \text{RSS}(x_1, x_2, x_3, x_4) + \text{SSReg}(x_4|x_1, x_2, x_3) = 15473$. Hence the relevant F -statistic is

$$F = \frac{\text{SSReg}(x_3|x_1, x_2)}{\text{RSS}(x_1, x_2, x_3)/(n - p + 1)} = \frac{\text{SSReg}(x_3|x_1, x_2)}{\text{MSE}(x_1, x_2, x_3)} \frac{3237/2}{15473/51} = 10.669.$$

The p-value is 0.0001, providing significant evidence for the joint significance of x_3 in the model without x_4 .

(k) Should the test in part (g) be the same as the test in part (j)? Why or why not?

These are distinct tests (note the different F -statistics). The first tests the joint explanatory significance of x_3, x_4 in the context of the full model with each x_1, x_2, x_3, x_4 , while the second tests only how x_3 adds to the model with x_1, x_2 and x_3 . Although, since the coefficient for x_4 is not significant ($p = .123$), the competing explanatory power gained by dropping x_4 was marginal. Hence, both are practically testing how well x_1 and x_2 explain y .

(l) For estimating the point on the regression surface at $(x_1, x_2, x_3, x_4) = (100, 25, 50, 4)$, the standard error of the estimate for the point on the surface is 2.62. Give the estimated point on the surface, a 95% confidence interval for the point on the surface, and a 95% prediction interval for a new point with these x -values.

The prediction is given by

$$\widehat{y|\mathbf{x}_0} = (1, 100, 25, 50, 4)\widehat{\boldsymbol{\beta}} \approx 124.0506.$$

The 95% CI for the mean response given $\mathbf{x}'_0 = (1, 100, 25, 50, 4)$ is given by

$$\widehat{y|\mathbf{x}_0} \pm t_{.975}(n - p) \cdot \text{SE}(\widehat{y|\mathbf{x}_0}) \approx 124.051 \pm 5.262 \text{ or } (118.788, 129.313).$$

Since $\text{SE}(\widehat{y|\mathbf{x}_0}) = \sqrt{\text{MSE} \cdot \mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}$, the prediction interval is given by

$$\begin{aligned} \widehat{y|\mathbf{x}_0} \pm t_{.975}(n - p) \sqrt{\text{MSE} \cdot (1 + \text{SE}(\widehat{y|\mathbf{x}_0})^2/\text{MSE})} &\approx 124.051 \pm 34.93 \\ &\text{or } (89.120, 158.981). \end{aligned}$$

(m) Test the null hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$.

This is the model significance test given by the F -statistic

$$F = \frac{(\text{TSS} - \text{RSS})/(p - 1)}{\text{MSE}} \approx 751.0606,$$

Which yields a p-value that is practically 0. Hence, the model is significant.

6. A random sample of 20 incoming shipments of chemicals in drums arriving at a warehouse was taken where measurements were taken on the number of drums in the shipment (x_1), the total weight of the shipment in hundreds of pounds (x_2), and the number of minutes required to handle the shipment (y). The data are given in three columns in the same order as above in the file `shipment.txt` on the course web page.

(a) Using matrix notation, fit the model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \quad i = 1, \dots, n, \quad \epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2).$$

Report the following vectors and matrices: $\mathbf{X}'\mathbf{X}$, $\mathbf{X}'\mathbf{y}$, $\mathbf{y}'\mathbf{y}$, $(\mathbf{X}'\mathbf{X})^{-1}$, & $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

The model in matrix form is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where \mathbf{y} and $\boldsymbol{\epsilon}$ are $n \times 1$ with entries y_i and ϵ_i respectively, \mathbf{X} is $n \times 3$ with entries x_{ij} with $j = 0, 1, 2$, and $\boldsymbol{\beta}$ is 3×1 with entries β_j . The following table summarizes the desired matrix computations.

$\mathbf{X}'\mathbf{X}$	$\mathbf{X}'\mathbf{y}$	$\mathbf{y}'\mathbf{y}$	$(\mathbf{X}'\mathbf{X})^{-1}$	$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$
$\begin{bmatrix} 20.00 & 238.00 & 182.240 \\ 238.00 & 3420.00 & 2671.060 \\ 182.24 & 2671.06 & 2161.227 \end{bmatrix}$	$\begin{bmatrix} 1889.00 \\ 27246.00 \\ 21648.81 \end{bmatrix}$	219449	$\begin{bmatrix} 0.307 & -0.033 & 0.015 \\ -0.033 & 0.012 & -0.012 \\ 0.015 & -0.012 & 0.014 \end{bmatrix}$	$\begin{bmatrix} 3.324 \\ 3.768 \\ 5.080 \end{bmatrix}$

(b) Estimate the variance-covariance matrix of the vector $\hat{\boldsymbol{\beta}}$.

$$\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1} \cdot \text{MSE} \approx \begin{bmatrix} 9.677 & -1.040 & 0.469 \\ -1.040 & 0.377 & -0.379 \\ 0.469 & -0.379 & 0.443 \end{bmatrix}.$$

(c) Test whether or not $\beta_2 > 50$ and interpret your conclusion clearly in the language of the problem.

We test $H_0 : \beta_2 \leq 50$. The test statistic is

$$t^* = \frac{\hat{\beta}_2 - 50}{\text{SE}(\hat{\beta}_2)} \approx -67.49392.$$

Hence $P(t > t^*) \approx 1$ and we conclude that there is an extreme lack of evidence for $H_a : \beta_2 > 50$. Practically, given that $\hat{\beta}_2 \approx 5.08$ and $\text{SE}(\hat{\beta}_2)^2 = 0.433$, we have absolutely no reason to believe that a shipment weighing 100 more pounds will result in more than 50 minutes of handling time (fixing the amount of drums).

(d) Noting that for random variables U, V and constants a, b , we have:

$$\text{Var}(aU + bV) = a^2\text{Var}(U) + b^2\text{Var}(V) + 2ab\text{Cov}(U, V),$$

compute the standard error of $\hat{\beta}_1 + 2\hat{\beta}_2$ using part (b) above, and test whether or not $\beta_1 = -2\beta_2$.

From above,

$$\text{Var}(\hat{\beta}_1 + 2\hat{\beta}_2) = \text{Var}(\hat{\beta}_1) + 4\text{Var}(\hat{\beta}_2) + 4\text{Cov}(\hat{\beta}_1, \hat{\beta}_2).$$

Hence, an estimate for $\sqrt{\text{Var}(\hat{\beta}_1 + 2\hat{\beta}_2)}$ is

$$\text{SE}(\hat{\beta}_1 + 2\hat{\beta}_2) = \sqrt{\widehat{\text{Var}}(\hat{\beta})_{22} + 4\widehat{\text{Var}}(\hat{\beta})_{33} + 4\widehat{\text{Var}}(\hat{\beta})_{2,3}} \approx 0.7967.$$

Note that $\hat{\beta}_1 + 2\hat{\beta}_2 \sim N(\beta_1 + 2\beta_2, \text{Var}(\hat{\beta}_1 + 2\hat{\beta}_2))$. Hence, we can test $H_0 : \beta_1 + 2\beta_2 = 0$ with the t -statistic

$$t^* = \frac{\hat{\beta}_1 + 2\hat{\beta}_2}{\text{SE}(\hat{\beta}_1 + 2\hat{\beta}_2)} \approx 17.482,$$

with p-value $2P(t > |t^*|) \approx 0$. Hence, there is fairly significant evidence to conclude $\beta_1 \neq -2\beta_2$.

(e) Find a 90% confidence interval for $\beta_1 + 2\beta_2$.

The confidence interval is given by

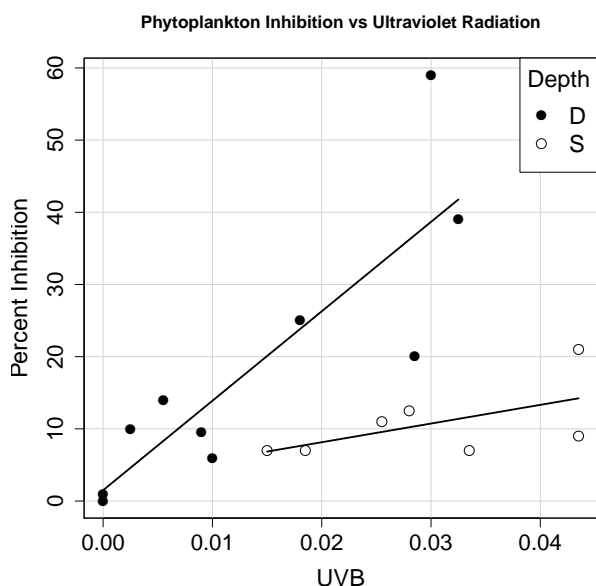
$$\hat{\beta}_1 + 2\hat{\beta}_2 \pm t_{.95}(n - p)\text{SE}(\hat{\beta}_1 + 2\hat{\beta}_2) \approx 13.927 \pm 1.386 \text{ or } (12.541, 15.313).$$

7. Thinning of the protective layer of ozone surrounding the earth may have catastrophic consequences. A team of University of California scientists estimated that increased solar radiation through the hole in the ozone layer over Antarctica altered processes to such an extent that primary production of phytoplankton was reduced 6-12%.

Depletion of the ozone layer allows the most damaging ultraviolet radiation - UVB (280-320 nm) - to reach the earth's surface. An important consequence is the degree to which oceanic phytoplankton production is inhibited by exposure to UVB, both near the ocean surface (where the effect should be slight) and below the surface (where the effect could be considerable).

To measure this relationship, the researchers sampled from the ocean column at various depths at 17 locations around Antarctica during the austral spring of 1990. To account for shifting of the ozone hole's positioning, they considered a measure of UVB exposure integrated over exposure time. The exposure measurements and the percentage of inhibition of normal phytoplankton production were extracted from their graph to produce the data in the file `ozone.txt` on the course webpage. These data contain 4 variables: the location number, percent inhibition, UVB exposure, and depth of measurement (S=surface, D=deep).

Does the effect of UVB exposure on the distribution of percentage inhibition differ at the surface and in the deep? How much difference is there? Analyze these data and write a summary of statistical findings in no more than one page. (Suggestion: Fit the model with different intercepts and different slopes, even if some terms are not significantly different from zero.)



In the figure to the left, we have plotted each percent inhibition against UVB exposure with depth indicated by different symbols. Least squares fitted lines for each depth are plotted on top of the data. We see that at each depth level inhibition increases with ultraviolet radiation, and as expected by the researchers, inhibition is greater at the lower depth. There may be a difference in the effect of UVB exposure on inhibition as indicated by the differing slopes of the least square lines. This suggests a possible interaction between depth and UVB exposure.

So, we fit the linear model with an interaction term,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 (x_{i1} x_{i2}) + \epsilon_i,$$

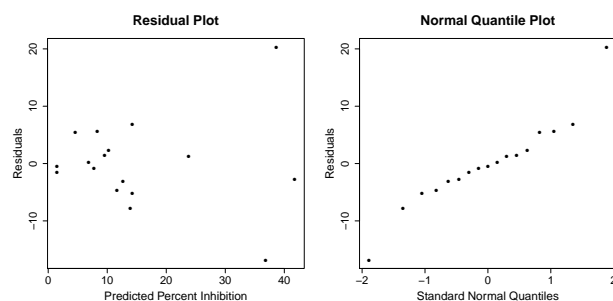
where the response y_i is the percent inhibition, x_{i1} is the UVB exposure, $x_{i2} = 1$ if measured at depth and $x_{i2} = 0$ if measured at the surface, and ϵ_i is assumed to be identically distributed normal error. The resulting estimates are

	β_0	β_1	β_2	β_3
Estimate	2.967	258.936	-1.467	980.039
Std. Error	9.726	309.612	10.538	381.539

To answer the primary question posed by the researchers, we tested the significance of the parameter β_3 in the model, and it was found to be moderately significant ($p = 0.023$). Hence, we have evidence of an interaction between UVB and depth in explaining phytoplankton inhibition. We can quantify the interaction by writing the model in the two depth cases

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} && \text{when at the surface, and} \\ y_i &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_{i1} && \text{at depth.} \end{aligned}$$

That is the slope of the relationship between inhibition and UVB is expected to be $\hat{\beta}_3 \approx 980.039$ higher at depth than at the surface, and the intercept is expected to change by $\hat{\beta}_2 \approx -1.467$.



Regarding the normality assumption, we note that there is no evidence of variance heterogeneity in the residual plot. There may be non-normality as indicated by the outlying values in the qqplot. The Shapiro-Francia test reports a p -value of 0.055, indicating possible non-normality. An arcsin transformation on the response has worse normality problems (SF-test: $p = 0.024$).

However, a permutation test ($n = 5000$) still indicates significance of the interaction at

$p = 0.0216$.