NEP 2005

mcdevitt

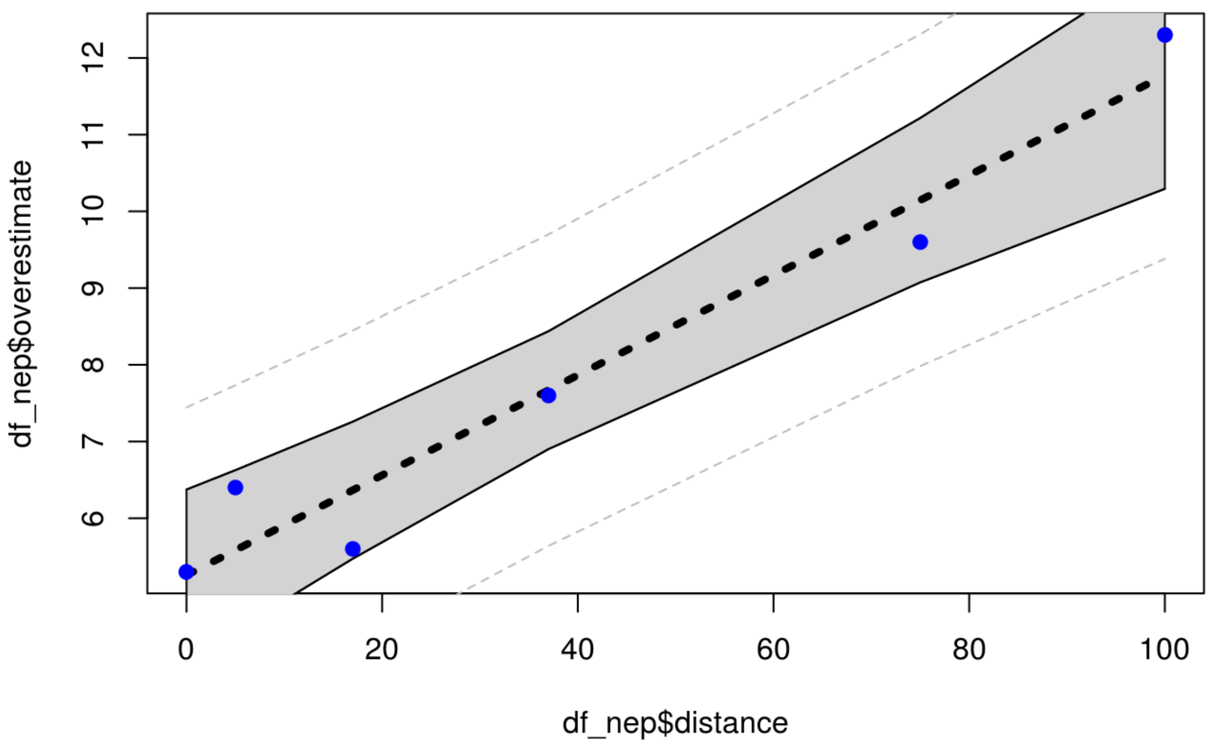
26 mars 2017

Edison - Mitofsky Election Results

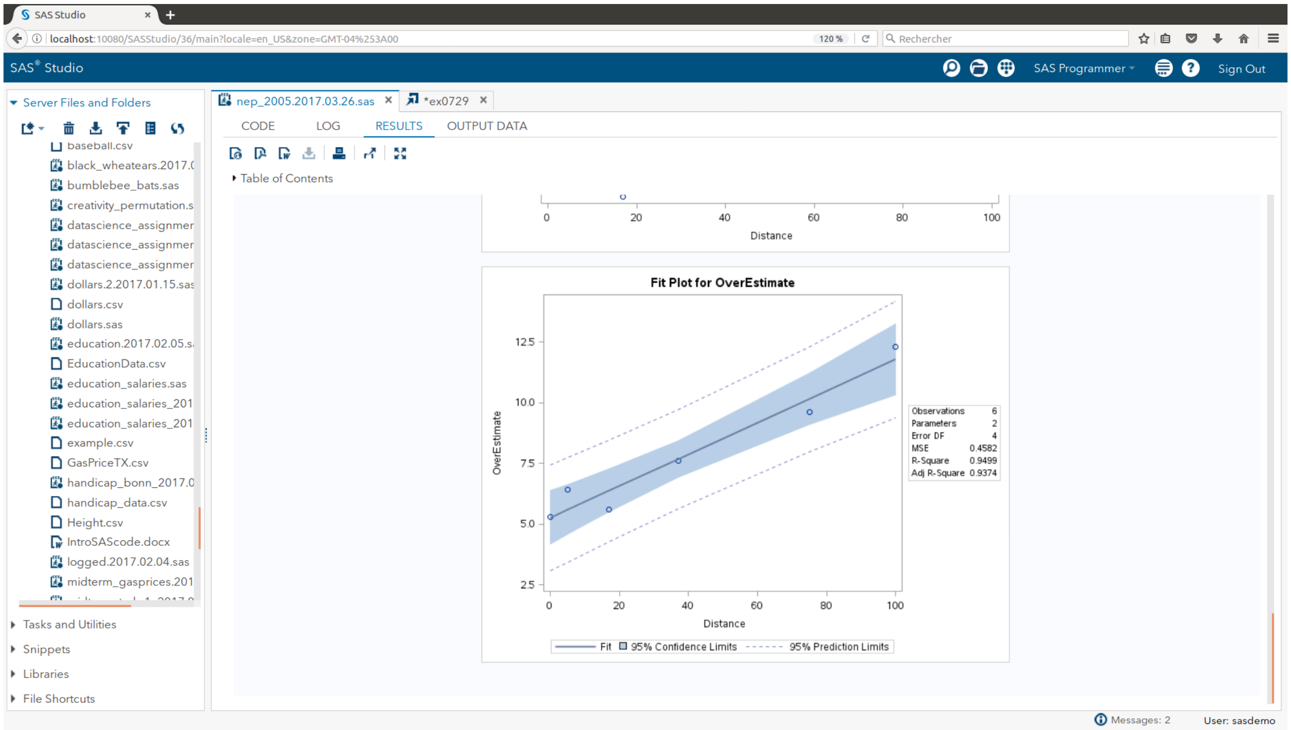
Homework 10 - NEP 2005 Evaluation - Problem 29, Chapter 7

b. Analyze the data providing at least the following:

=-
__ i. A Scatterplot with confidence intervals of the regression line and prediction intervals of the regression line. Please do in SAS and R! __



And the SAS plot …

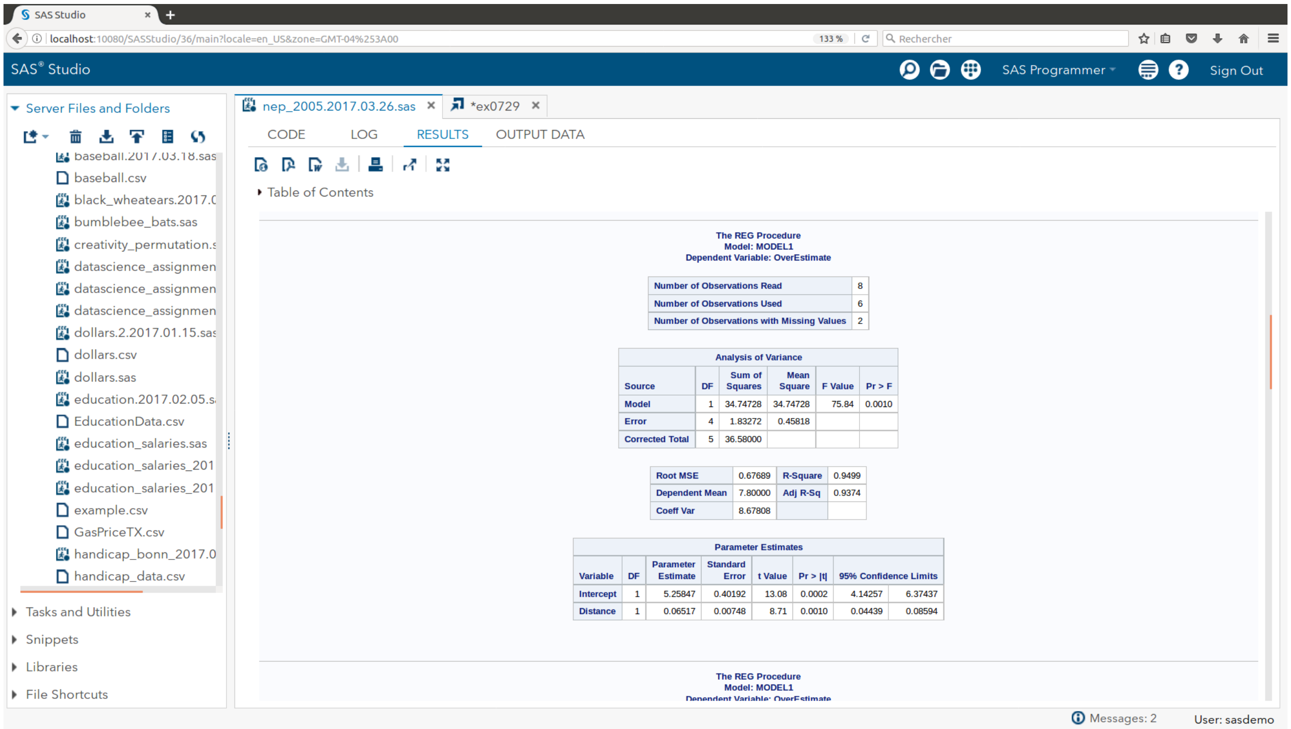


=-

ii. A table showing the t-statistics and pvalues for the significance of the regression parameters: . Please do in SAS and R!

```
##
## Call:
## lm(formula = overestimate ~ distance, data = df_nep)
##
## Residuals:
##        1        2        3        4        5        6
##  0.04153  0.81569 -0.76632 -0.06967 -0.54603  0.52479
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.258470   0.401917  13.083 0.000197 ***
## distance    0.065167   0.007483   8.708 0.000957 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6769 on 4 degrees of freedom
## Multiple R-squared:  0.9499, Adjusted R-squared:  0.9374
## F-statistic: 75.84 on 1 and 4 DF,  p-value: 0.0009575
```

And the SAS table …



Overestimate vs. Poll Distance from SAS PROC REG

=-

iii. Using the data in ii show all 6 steps of each hypothesis test.

=-

Six-Step Hypothesis Test - Slope

1: Ho : b1 = 0; Ha ; b1 != 0

2 : Critical Value : t(0.975, df = **4**) = +/- **2.7764451**

3 : t = 8.7084782

4 : p_value = 9.574970110^{-4} < 0.05 ?

5: Reject Ho

6 : There is sufficient evidence to suggest at the alpha = 0.05 level of significance (p-value = 9.574970110^{-4}) that the slope of the regression line that estimates the OverEstimate for NEP 2005 based on the Polling Distance is equal to zero. A 95% confidence interval for the slope is (0.044, 0.086), which is an interval that does not contain the value zero

Six-Step Hypothesis Test - Intercept

1 : Ho : b0 = 0; Ha ; b0 != 0

2 : Critical Value : t(0.975, df = **4**) = +/- **2.7764451**

3 : t = 13.0834768

4 : p_value = 1.970294310^{-4} < 0.05 ?

5: Reject Ho

6 : There is sufficient evidence to suggest at the alpha = 0.05 level of significance (p-value = 1.970294310^{-4}) that the intercept of the regression line that estimates the OverEstimate for NEP 2005 based on the Polling Distance is not equal to zero. A 95% confidence interval for the slope is (

4.143, 6.374), which is an interval that does not contain the value zero

=-

iv. The regression equation.

The Regression equation : - overestimate = 0.0651674 * distance + 5.2584698

=-

v. Interpretation of the slope and intercept in the model (regression equation.)

The slope represents the rate of change in overestimate estimated for a unit change in poll distance . I.e., for each increase in 1 foot of poll distance there is a corresponding increase in overestimate by 0.0651674 percentage points

The intercept represents the estimated average level of overestimate percentage pointss associated to a polling distance of 0 feet from this sample study. From a practical point of view, since there were some polling distance experience of 0 feet, the intercept provides an estimate of the mean overestimate percentage expected for the population represented from this sample study.

=-

vi. Find and interpret the 95% confidence interval for the mean overestimate response conditional on a poll distance of 37 feet. Please do in SAS and R!

The confidence interval is the upper and lower bound for the expected mean value at the given indpendent value (e.g., distance = 37 feet) for the current regression relationship.

For this particular regression, the confidence interval at poll distance = 37 : overestimate percentage points in the range (6.901, 8.438)

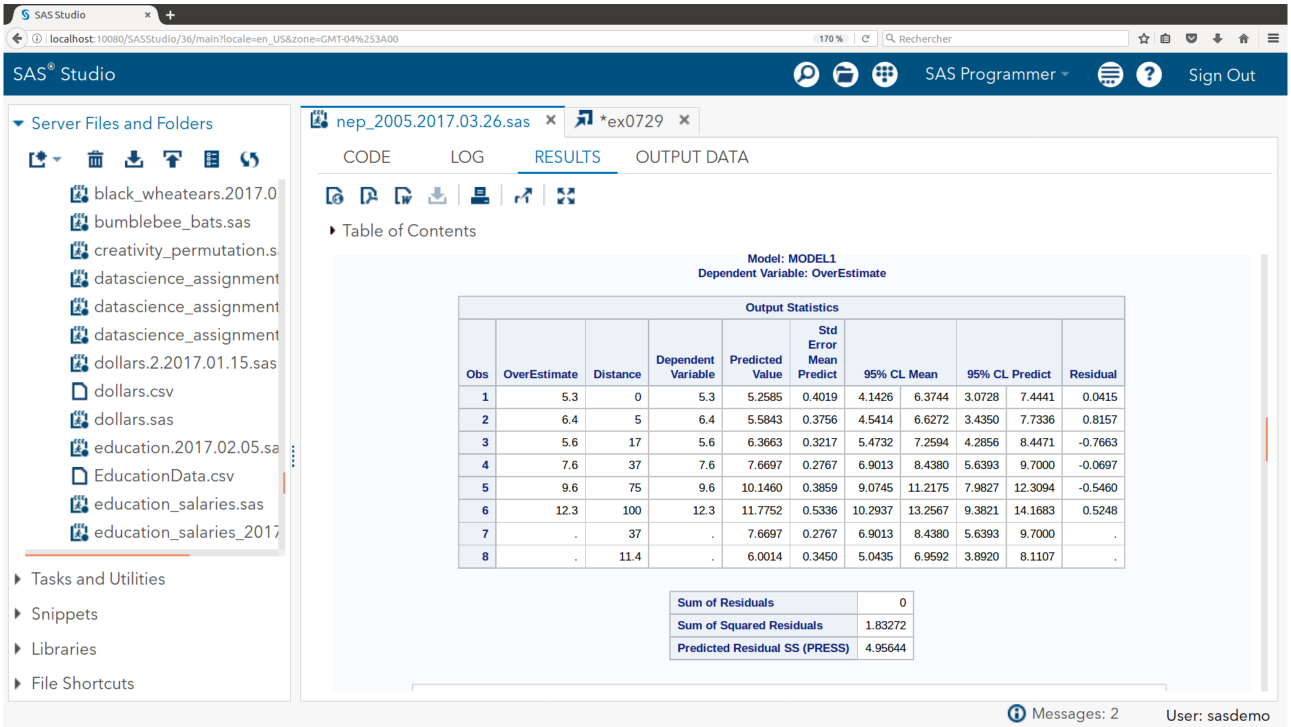
=-

vii. Find and interpret the 95% prediction interval for the predicted overestimate response given a poll distance of 37 feet. Please do in SAS and R!

The prediction interval is the upper and lower bound for an ‘next’ observation of dependent value at the given independent value, based on the current regression relationshhip. That is to say, for an observation of distance and overestimate not included in this analysis, the prediction interval bounds the range of future observations that are expected.

For this particular case, the prediction interval at poll distance = 37 : overestimate percentage points in the range (5.639, 9.700)

And the SAS table …



Model: MODEL1
Dependent Variable: OverEstimate

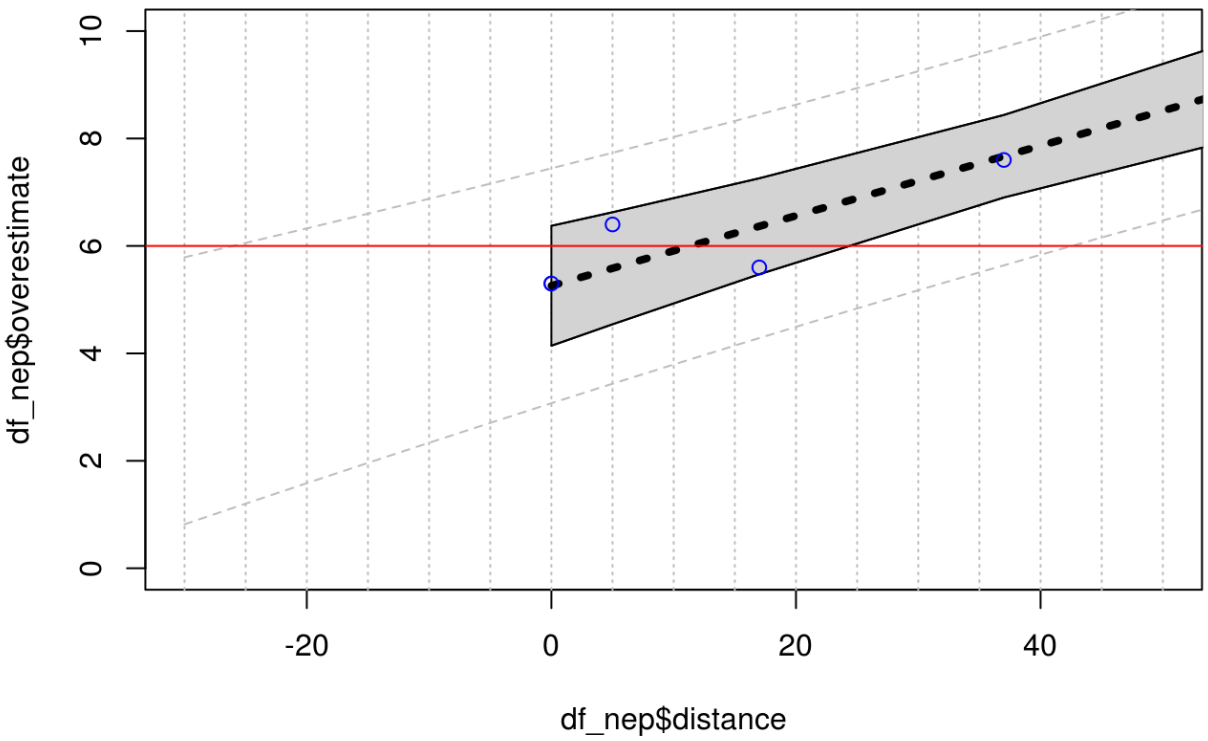
Output Statistics

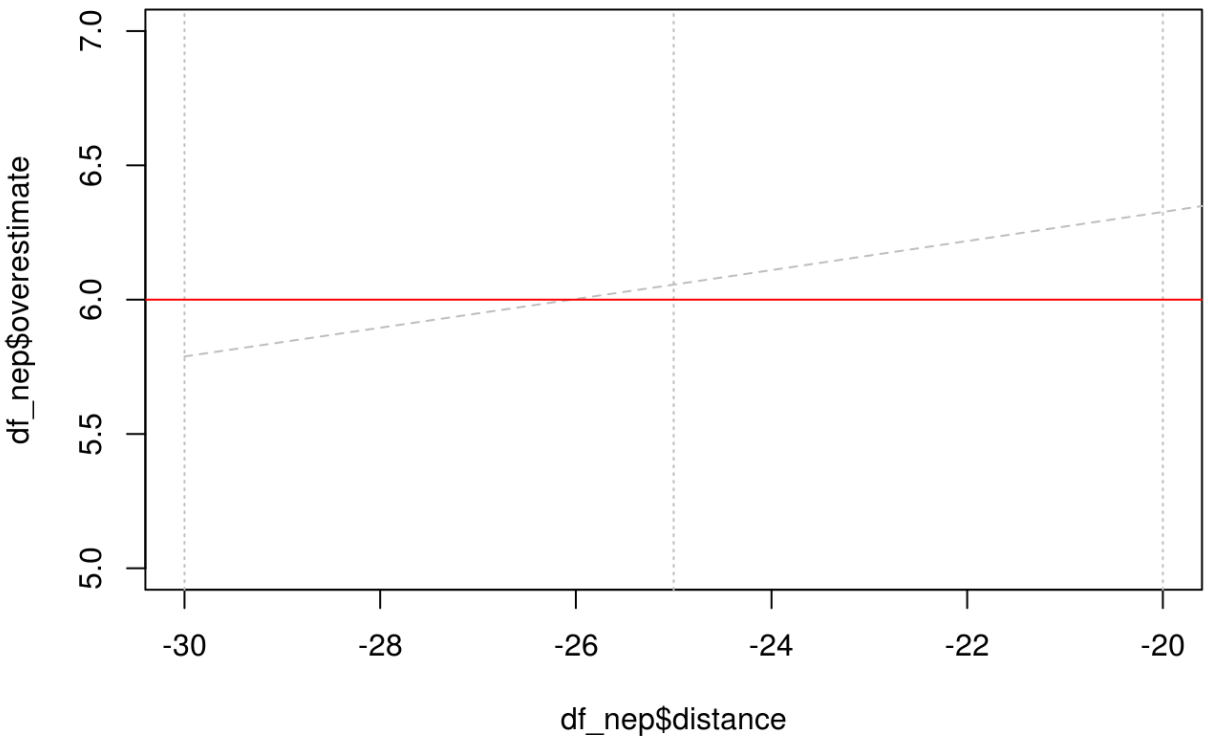
Obs	OverEstimate	Distance	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	5.3	0	5.3	5.2585	0.4019	4.1426	6.3744	3.0728	7.4441	0.0415
2	6.4	5	6.4	5.5843	0.3756	4.5414	6.6272	3.4350	7.7336	0.8157
3	5.6	17	5.6	6.3663	0.3217	5.4732	7.2594	4.2856	8.4471	-0.7663
4	7.6	37	7.6	7.6697	0.2767	6.9013	8.4380	5.6393	9.7000	-0.0697
5	9.6	75	9.6	10.1460	0.3859	9.0745	11.2175	7.9827	12.3094	-0.5460
6	12.3	100	12.3	11.7752	0.5336	10.2937	13.2567	9.3821	14.1683	0.5248
7	.	37	.	7.6697	0.2767	6.9013	8.4380	5.6393	9.7000	.
8	.	11.4	.	6.0014	0.3450	5.0435	6.9592	3.8920	8.1107	.

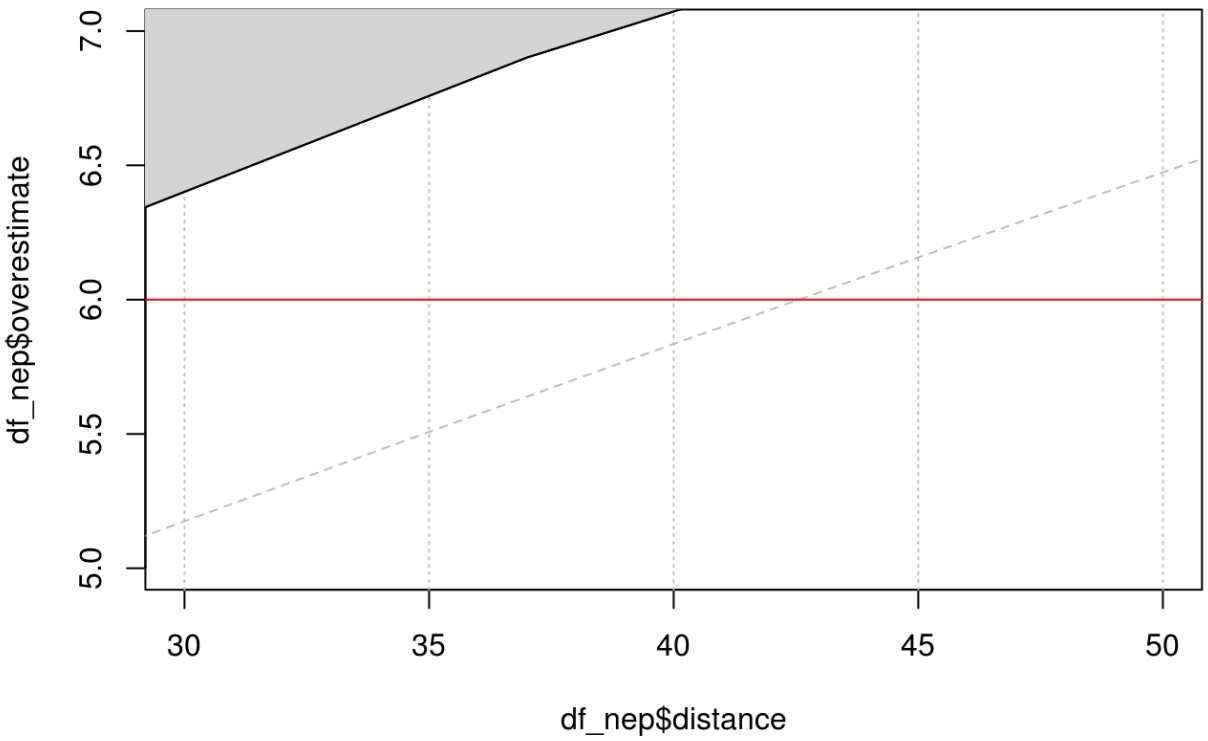
Sum of Residuals	0
Sum of Squared Residuals	1.83272
Predicted Residual SS (PRESS)	4.95644

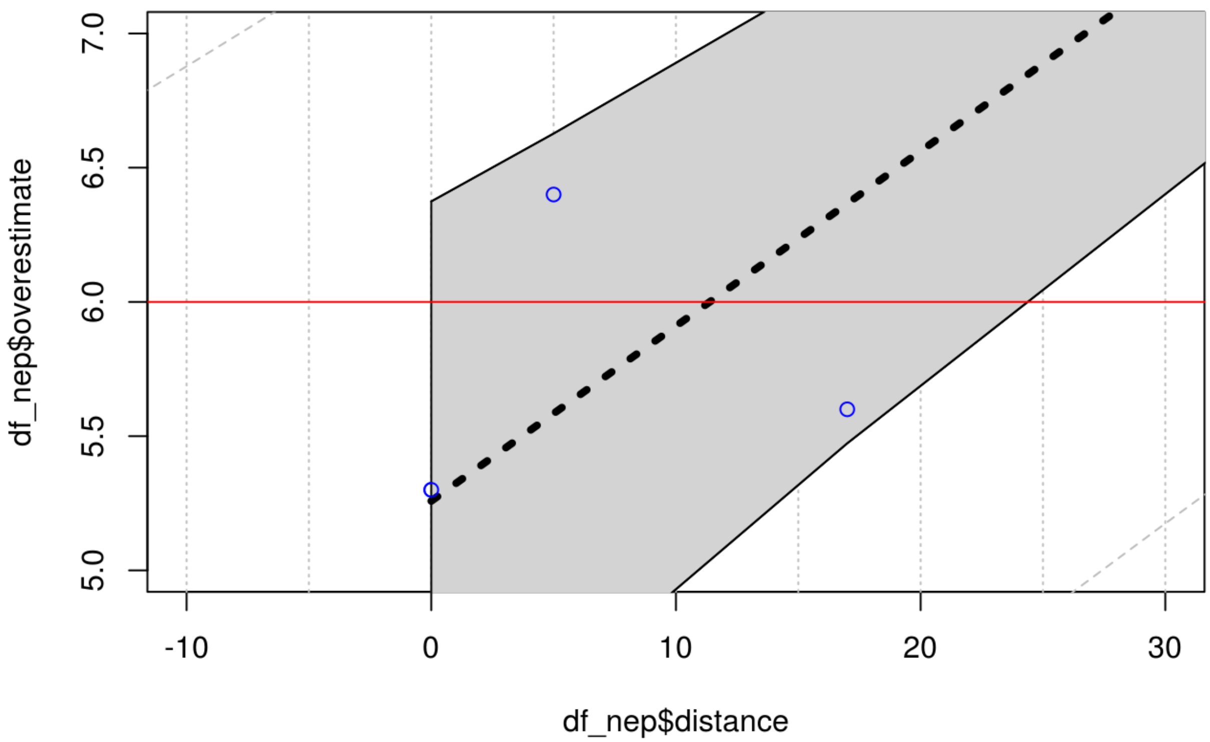
=-

viii. Using the graphical method, find and interpret the calibration intervals for the overestimate response of 6.0 points. (Both for mean overestimate response and for a single overestimate response. Please do in SAS and R! (R: package investr)


df_nep$overestimate
0
2
4
6
8
10
-20
0
20
40
df_nep$distance


df_nep$overestimate
7.0
6.5
6.0
5.5
5.0
-30
-28
-26
-24
-22
-20
df_nep$distance


df_nep$overestimate
7.0
6.5
6.0
5.5
5.0
30
35
40
45
50
df_nep$distance



The calibration estimates based on graphical inspection are as follows :

- calibration interval for the **mean** overestimate response of 6.0 points = (-6, 25)
- calibration interval for **single point** overestimate response of 6.0 points = (-25.5, 42.5)

=-

ix. Find the same calibration intervals analytically using the SE equations given in class and in the book (Version 3 page 194).

```
y_est <- 6.0
x_at_y_est <- (y_est - b0) / b1

x_at_y_est
```

```
## (Intercept)
##    11.37884
```

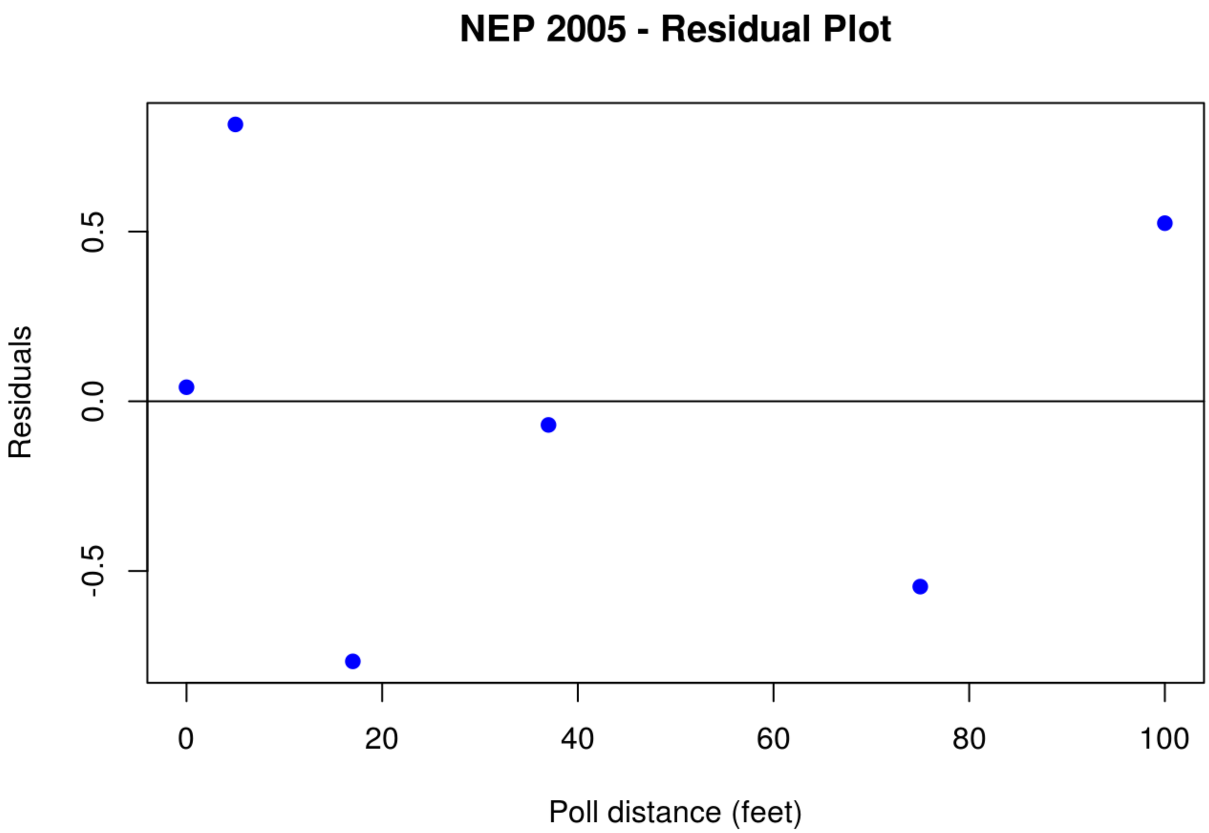
```
se_at_y_est <- predict(fit_all, data.frame(distance = x_at_y_est), se.fit = T
RUE)

se_ci_at_y_est <- se_at_y_est$se.fit / abs(b1)

alpha <- 0.05
dof <- dim(df_nep)[1] - 2
crit_value <- qt(1 - alpha/2, dim(df_nep)[1] - 2)

cal_est_at_y_est_upr <- x_at_y_est + crit_value * se_ci_at_y_est
cal_est_at_y_est_lwr <- x_at_y_est - crit_value * se_ci_at_y_est


# ...   calibration of predicted value

res_std_dev <- sigma(fit_all) * sigma(fit_all)

se_pi_at_y_est <- (sqrt (res_std_dev + se_at_y_est$se.fit * se_at_y_est$se.fi
t) ) / abs(b1)

cal_pi_est_at_y_est_upr <- x_at_y_est + crit_value * se_pi_at_y_est
cal_pi_est_at_y_est_lwr <- x_at_y_est - crit_value * se_pi_at_y_est
```

The calibration estimates are as follows :

- calibration interval for the **mean** overestimate response of 6.0 points = (-3.3235882, 26.0812776)
- calibration interval for **single point** overestimate response of 6.0 points = (-20.9914431, 43.7491325)

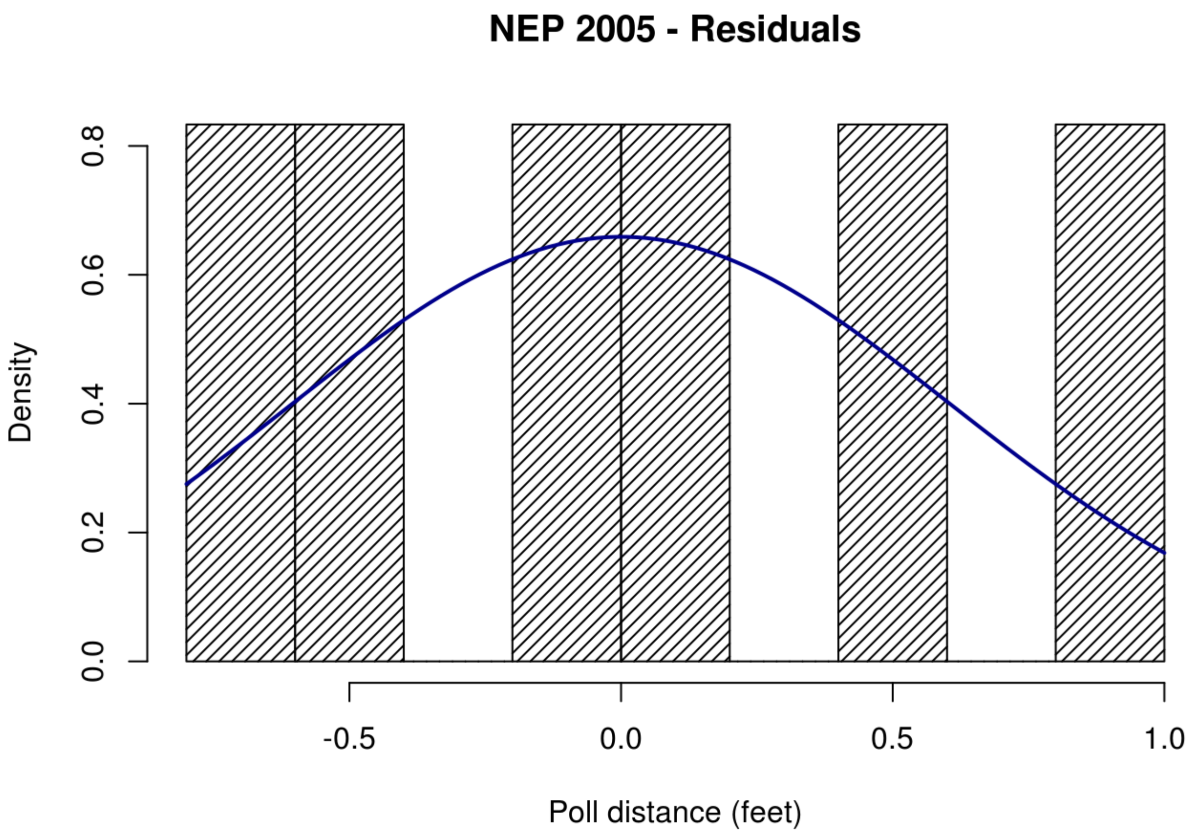
=-

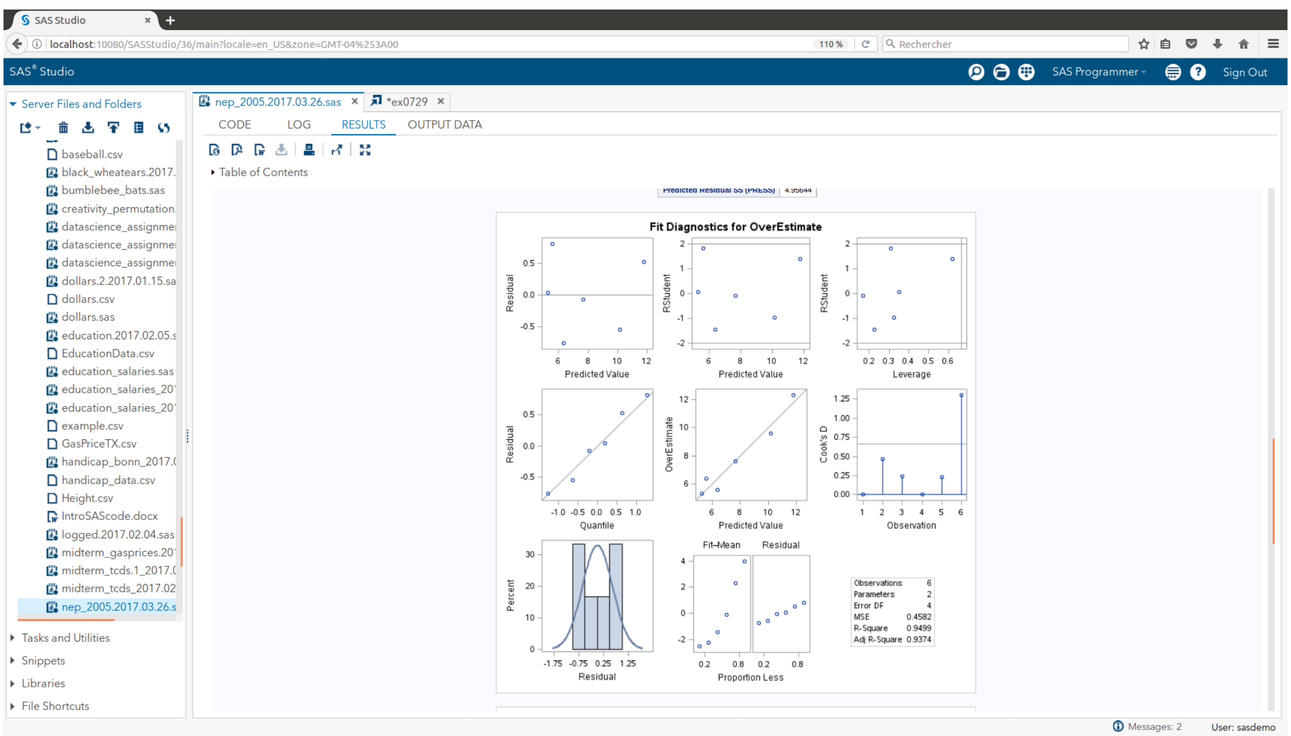
x. A scatterplot of residuals. Please do in SAS and R!



=-

xi. A histogram of residuals with normal distribution superimposed. (from SAS).





Overestimate vs. Poll Distance from SAS PROC REG

=-

xii. Question from the text : How strong is the evidence that the mean Kerry overestimate increases with increasing distance of interviewer from the door ?

The data set for this evaluation is quite small (6 data points !!) in relation to the size of the electorate - there are more than 100,000 polling places in the USA. Despite the smallness of the data set, the correllation is very strong. The r^2 for this study is 94% - suggesting that 94% of the variation in response over-estimation that was observed in this sample can be associated to the variation in the distance from the voter to the exit poll interviewer. This is a remarkable degree of association, considering all of the factors that are potentially contributors to a polling error. One other aspect to limit the inferences from this study is that this is an observational study, so no causal inference can be drawn to state the the overestimation error is due to the distance from the interviewer to the voter.

=-=–=-=-=-

xiii BONUS QUESTION : What is the unit of measurement for distance in this study ?

__Numerous references in the article refer to distances in feet … the distances are related to the distances that each region imposed on the number of feet that exit poll interviewers were required to remain from the doors of the polling stations.

=-=–=-=-=-