LR HW 2

2.1. A student working on a summer internship in the economic research department of a large corporation studied the relation between sales of a product (Y, in million dollars) and population (X, in million persons) in the firm's 50 marketing districts. The normal error regression model (2.1) was employed. The student first wished to test whether or not a linear association between Y and X existed. The student accessed a simple linear regression program and obtained the following information on the regression coefficients:

Parameter	Estimated Value	95 Percent Confidence Limits	
Intercept	7.43119	−1.18518	16.0476
Slope	.755048	.452886	1.05721

a. The student concluded from these results that there is a linear association between Y and X. Is the conclusion warranted? What is the implied level of significance?

b. Someone questioned the negative lower confidence limit for the intercept, pointing out that dollar sales cannot be negative even if the population in a district is zero. Discuss.

a) The slope confidence interval of 95% does not include 0

There is strong evidence of a linear association at $\alpha = 0.05$

b) The regression model can run into an issue with extrapolation which is a pitfall of regression models

2.3. A member of a student team playing an interactive marketing game received the following computer output when studying the relation between advertising expenditures (X) and sales (Y) for one of the team's products:

Estimated regression equation: $\hat{Y} = 350.7 - .18X$

Two-sided P-value for estimated slope: .91

The student stated: "The message I get here is that the more we spend on advertising this product, the fewer units we sell!" Comment.

The P value of 0.91 indicates the slope may not be statistically significant. which suggests a lack of evidence that the slope is not just 0. So a strong claim cannot be made

2.10. For each of the following questions, explain whether a confidence interval for a mean response or a prediction interval for a new observation is appropriate.

a. What will be the humidity level in this greenhouse tomorrow when we set the temperature level at 31°C?

b. How much do families whose disposable income is $23,500 spend, on the average, for meals away from home?

c. How many kilowatt-hours of electricity will be consumed next month by commercial and industrial users in the Twin Cities service area, given that the index of business activity for the area remains at its present level?

a) Prediction Interval - we want to make a specific observation

b) Confidence Interval - we are looking at an average and

c) Prediction Interval - we want a specific observation for next month

2.22. Using the normal error regression model (2.1) in an engineering safety experiment, a researcher found for the first 10 cases that R^2 was zero. Is it possible that for the complete set of 30 cases R^2 will not be zero? Could R^2 not be zero for the first 10 cases, yet equal zero for all 30 cases? Explain.

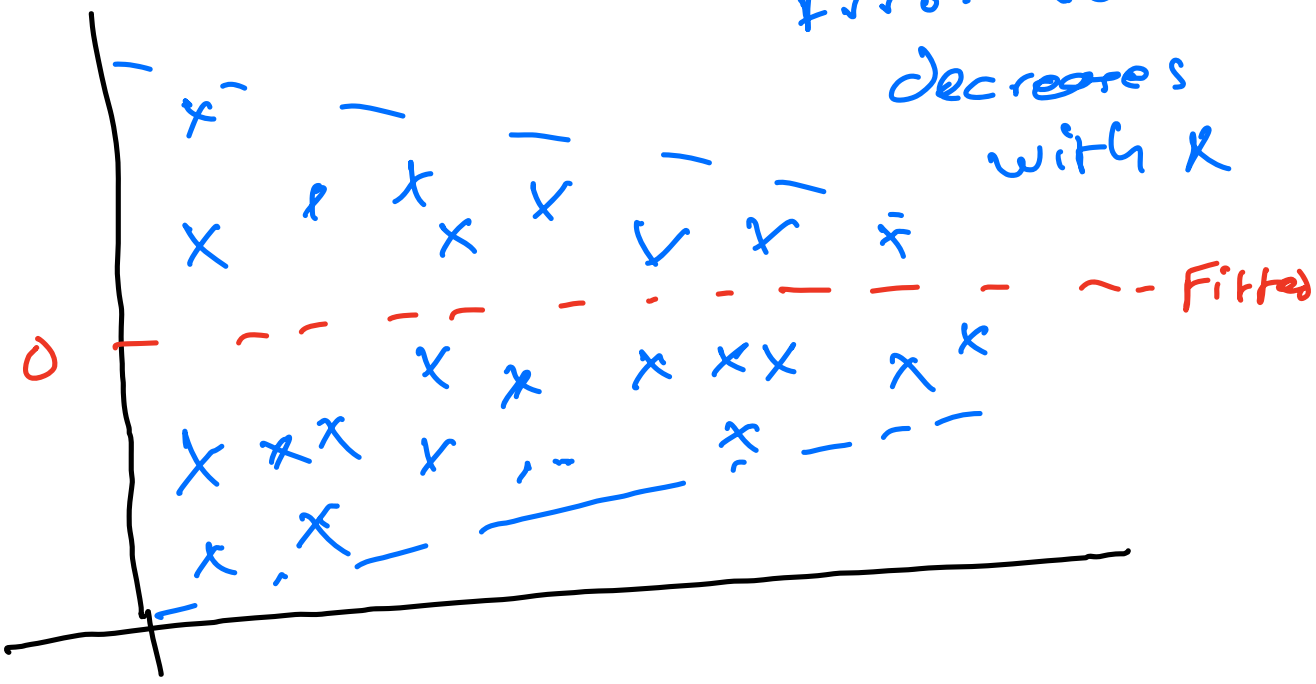
$$R^2 = \frac{SSE}{SST} = \frac{SSE}{SSE + SSR}$$

- It is possible that complete set may not be represented by the first 10 cases and so the remaining 20 may reveal some relationship and $R^2 > 0$

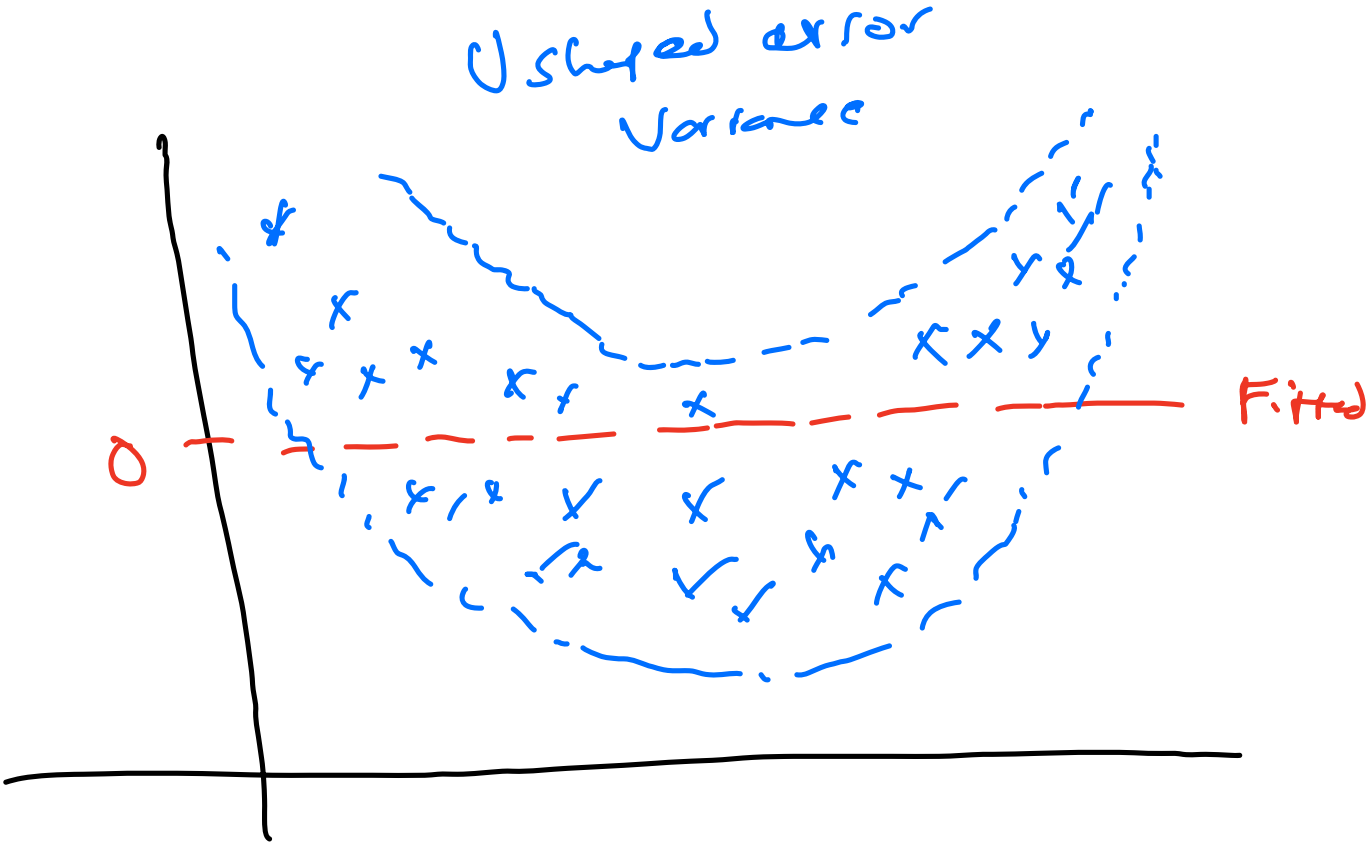
- It is improbable that the 20 remaining cases completely overshadow the relationship found in the first 10 such that it would make R^2 completely 0.

3.2. Prepare a prototype residual plot for each of the following cases: (1) error variance decreases with X; (2) true regression function is $\cup$ shaped, but a linear regression function is fitted.

(1)



(2)



1. Recall the following statement about in-sample predictions: assuming (x_i, y_i) belongs to the training sample we have used to fit the SLR, then the residual e_i has mean 0 and variance $\sigma^2[1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{SSX}]$. Now imagine after fitting a regression line and obtaining the parameter estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, we are predicting an observation that's out of sample, i.e. the observation has independent variable value $X = x_0$, where x_0 does not belong to the data used to fit the model. The true response variable

$$y_0 = \beta_0 + \beta_1 x_0 + \epsilon_0$$

is predicted by

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

What is the mean and variance of the out-of-sample prediction bias $e_0 = y_0 - \hat{y}_0$? Hint: since this is an out-of-sample point, ϵ_0 is uncorrelated with all the in sample random errors ϵ_i, $i = 1, ..., n$.

Unbiased estimators:

$$E(\hat{\beta}_1) = \beta_1 \qquad E(\hat{\beta}_0) = \beta_0$$

$$E(e_0) = E[(\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)x_0 + \epsilon_0]$$

$$= E[0] + x_0 E(0) + 0 \quad \text{by Gauss Markov}$$

$$= 0$$

$$Var(\hat{\beta}_0) = \sigma^2 (1/n + \bar{x}^2/SSx)$$

$$Var(\hat{\beta}_1) = \sigma^2/SSx$$

$$Var(\epsilon_0) = \sigma^2 \quad (\epsilon_0 \sim N(0, \sigma^2))$$

$$Var(e_0) = \sigma^2(1/n + \bar{x}^2/SSx) + x_0^2 \cdot \frac{\sigma^2}{SSx} + \sigma^2$$

$$= \sigma^2 \left[1 + 1/n + \frac{(x_0 - \bar{x})^2}{SSx}\right]$$

2. **DO NOT USE PYTHON OR R.** Calculate manually or just with a simple calculator, to fill out the ANOVA table for SLR using the following data, then perform the F-test by stating the hypothesis (H_0 and H_1) and providing the test statistic and decision of the test:

X	Y
3	10
3.5	11.5
5	12
6	14

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

$\hat{\beta_0} = \bar{Y} - \beta_1 \bar{x}$

$\hat{\beta_1} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$

$\bar{x} = \frac{3 + 3.5 + 5 + 6}{4}$

$= 4.375$

$\bar{y} = \frac{10 + 11.5 + 12 + 14}{4}$

$= 11.875$

$\hat{\beta_1} = \frac{\sum (x_i - 4.375)(y_i - 11.875)}{\sum (x_i - 4.375)^2} \approx 1.13$

$\hat{\beta_0} \approx 6.935$

$\hat{y} = 6.935 + 1.13x$

$$SST = \sum (y_i - \bar{y})^2 \approx 8.1875$$

$$SSR = \sum (\hat{y}_i - \bar{y})^2 \approx 7.26$$

$$SSE = SST - SSR \approx 0.925$$

$k = 2$
$n = 4$

$$MSR = \frac{SSR}{k-1} = \frac{7.26}{1} \approx 7.26$$

$$MSE = \frac{SSE}{n-k} = \frac{0.925}{2} \approx 0.4625$$

$$F = \frac{MSR}{MSE} = \frac{7.26}{0.4625} \approx 15.7$$

Crit val $F_{0.05,1,2} \approx 18.51$ (googled)

Source	SS	d.f.	MS	F
Regression	7.26	1	7.26	15.7
Error	0.925	2	0.4625	
Total	8.1875	3		

Fail to reject H_0, not enough evidence.

Coding questions:

1. The data "hospital_infection.csv" recorded the patients infection rates data from 58 hospitals. Let's take `InfctRisk` as the response and `Stay` (average staying days) as the predictor and perform a regression analysis using Python. You are not restricted to the libraries and functions I have used in class. Upload your python file in the homework submission. In your .pdf, answer the following questions:
   (a) From scatter plot: is there any initial relationship you can observe? i.e. linear or not, positive or negative, etc.
   (b) From simple linear regression: write down the equation of the fitted line.
   (c) From the summary: what is the test statistic and p-value of the coefficient of `Stay`. How does that suggest the impact of `Stay` to `InfctRisk`?
   (d) Calculate the 95% confidence interval for β_1. Does it match with the result from the summary output?
   (e) Run ANOVA: what is the test statistic and p-value of the F test? What does that suggest about the significance of the predictor in your model?
   (f) Verify that squaring the t-statistic associated with testing the null hypothesis $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$ is the same value as F-statistic generated from the ANOVA. Also verify that the p-values are approximately the same for the two tests.
   (g) From the summary: what is the value of R^2? What does it suggest about the goodness-of-fit of your model?
   (h) Predict the infection risk for a hospital with an average stay length of 32 days.

2. Continue with the hospital infection example from above. Use python to draw the residual v.s. fitted value plot and QQ plot based on your model. Based on your plots, evaluate the following assumptions. A brief/roughly observed comment is acceptable (i.e., seems to be false because...; hard to judge becuase...; etc.).
   (a) Infection risk and stay length have a linear relationship.
   (b) The assumption of error terms have constant variance is valid.
   (c) The assumption of error terms are independent is valid.
   (d) The assumption of error terms are normally distributed is valid.

3. Continue with the hospital infection example from above. Use Python to give the 95% confidence interval for the mean infection risk when the average stay is 32 days, and the 95% prediction interval of the infection risk for a new patient who would stay 32 days.

1(a) Seems like positive linear relationship
although with quite some error

b) $\hat{Y} = -1.16 + 0.5689X$

c) $t = 6.041$, $p < 0.001$
Very significant relationship

d) [0.380, 0.7575]

Yes, matches summary

e) $F_{stat} = 36.5$, $p\,val < 0.001$

Strong relationship

f) $t^2 \simeq 36.49$ ✓

g) $R^2 = 0.395$

39.5% of variability in infection Risk is explained by stay

Model fit is decent but stay is not only factor in infection risk

h) Infection Risk Prod = 17.040%

This is very extrapolated

[no data in near that length of stay]

2 a) Doesn't seem to be clear pattern that goes against linearity

b) Variance could be constant doesn't seem to be other patterns

c) Seems hard to assess error independence

d) normality is mostly valid outside the extremes (QQ plot)

3) 95% CE: [12.89, 21.19]

95% PF: [12.41, 21.67]