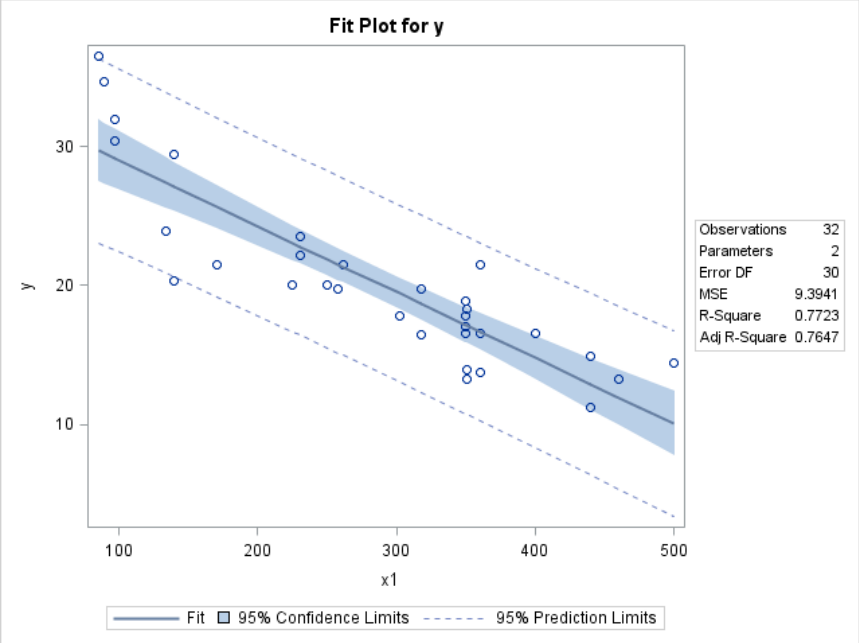
Homework 1

朱强强
17064001
Applied Statistics

2.4

a. From the consequence computed by SAS, we can clearly know that $y = 33.72268 - 0.04736x_1$.



b. The analysis-of-variance table.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	955.71971	955.71971	101.74	<.0001
Error	30	281.82438	9.39415		
Corrected Total	31	1237.54409			

The test for significance of regression.

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	33.72268	1.44390	23.36	<.0001
x1	1	-0.04736	0.00470	-10.09	<.0001

From the table above, we know that the p_values of two parameters are both below 0.05.

c. $R_2 = \frac{SSR}{SST} = 0.7723$

So about 77.23% of the total variability in gasoline mileage is accounted for by the linear relationship with engine displacement.

Root MSE	3.06499	R-Square	0.7723
Dependent Mean	20.22313	Adj R-Sq	0.7647
Coeff Var	15.15585		

d. We know that

$$\begin{aligned} \widehat{E(y|x_0)} = \hat{\mu}_{y|x_0} &= \hat{\beta}_0 + \hat{\beta}_1 x_0 \\ Var &= Var(\hat{\beta}_0 + \hat{\beta}_1 x_0) \\ &= Var[\bar{y} + \hat{\beta}_1(x_0 - \bar{x})] \\ &= Var(\bar{y}) + Var[\hat{\beta}_1(x_0 - \bar{x})] \\ &= \frac{\sigma^2}{n} + \frac{\sigma^2(x_0 - \bar{x})^2}{S_{xx}} \end{aligned}$$

$$\hat{\mu}_{y|x_0} - t_{\alpha/2,n-2}\sqrt{MS_{RES}[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}]} \le E(y|x_0) \le \hat{\mu}_{y|x_0} + t_{\alpha/2,n-2}\sqrt{MS_{RES}[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}]}$$

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		Residual
1	.	20.6988	0.5439	19.5881	21.8095	.

From the output of SAS, we can see that the predicted value of y is 20.6988, and a 95% CI on $E(y|x_0 = 275)$ is $[15.5881, 21.8095]$.

e. Known $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$

Assume $\Psi = y_0 - \hat{y}_0, \therefore E(\Psi) = 0$

$$Var(\Psi) = Var(y_0 - \hat{y}_0) = \sigma^2[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}]$$

$$\hat{y}_0 - t_{\alpha/2,n-2}\sqrt{MS_{RES}[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}]} \le y_0 \le \hat{y}_0 + t_{\alpha/2,n-2}\sqrt{MS_{RES}[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}]}$$

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Predict		Residual
1	.	20.6988	0.5439	14.3415	27.0561	.

From the table above, a 95% prediction interval of y is $[14.3415, 27.0561]$.

f. The difference between the confidence interval and the prediction interval.

A prediction interval is an interval associated with a random variable yet to be observed, with a specified probability of the random variable lying within the interval.

A confidence interval is an interval associated with a parameter and is a frequent concept. The parameter is assumed to be non-random but unknown, and the confidence interval is computed from data. Because the data are random, the interval is random.

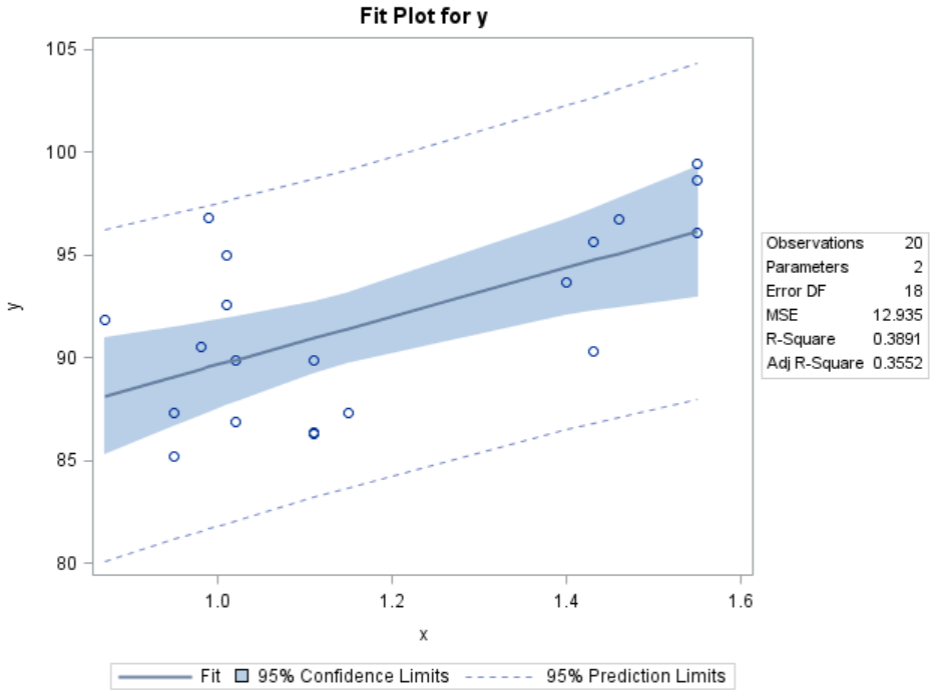
From this problem, we can clearly see that the prediction interval is wider than the confidence interval. The reason is that the prediction interval is the prediction of the variable and the confidence interval is the prediction of the expectation of the variable, so the prediction internal contains the deviation of the variable relative to the the estimated value, compared with the confidence value.

2.7

a. From the consequence computed by SAS, we can clearly know that $y = 77.86328 - 11.80103x$.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	77.86328	4.19889	18.54	<.0001
x	1	11.80103	3.48512	3.39	0.0033



b. For the hypothesis $H_0 : \beta_1 = 0$,
the test statistics $F_0 = 11.47$, and the corresponding $p\_value = 0.0033 < 0.01$.
So we have 99% of confidence to reject the null hypothesis.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	148.31296	148.31296	11.47	0.0033
Error	18	232.83436	12.93524		
Corrected Total	19	381.14732			

c. From the figure below we can see that $R^2 = 0.3891$.

Root MSE	3.59656	R-Square	0.3891
Dependent Mean	91.81800	Adj R-Sq	0.3552
Coeff Var	3.91705		

d. According to the results displayed by SAS, we can find that a 95% CI on the slope is $[4.47907, 19.12299]$.

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	77.86328	4.19889	18.54	<.0001	69.04175	86.68482
x	1	11.80103	3.48512	3.39	0.0033	4.47907	19.12299

e. We find that a 95% CI on y when $x = 1.00$ is $[87.5102, 91.8185]$.

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		Residual
1	.	89.6643	1.0253	87.5102	91.8185	.

2.25

Consider the simple linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$, with $E(\varepsilon) = 0, Var(\varepsilon) = \sigma^2$, and ε uncorrelated.

a. Show that $Cov(\hat{\beta}_0, \hat{\beta}_1) = -\bar{x}\sigma^2 / S_{xx}$.

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \sum_{i=1}^{n} c_i y_i, Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

$$\begin{aligned} Cov(\hat{\beta}_0, \hat{\beta}_1) &= Cov(\bar{y} - \hat{\beta}_1 \bar{x}, \hat{\beta}_1) \\ &= Cov(\bar{y}, \hat{\beta}_1) - \bar{x} Var(\hat{\beta}_1) \\ Cov(\bar{y}, \hat{\beta}_1) &= \frac{1}{S_{xx}} Cov(\bar{y}, \Sigma(x_i - \bar{x}) y_i) \\ &= \frac{\Sigma(x_i - \bar{x})}{S_{xx}} Cov(\bar{y}, y_i) \\ &= \frac{\Sigma(x_i - \bar{x})}{S_{xx}} \cdot \frac{\sigma^2}{n} = 0 \\ \therefore Cov(\hat{\beta}_0, \hat{\beta}_1) &= -\bar{x}\sigma^2 / S_{xx} \end{aligned}$$

b. Show that $Cov(\bar{y}, \beta_1) = 0$.

$Cov(\bar{y}, \beta_1) = E[E(\bar{y}) - \bar{y}][E(\beta_1) - \beta_1] = 0.$

2.28

Consider the maximum-likelihood estimator $\tilde{\sigma}^2$ of σ^2 in the simple linear regression model. We know that $\tilde{\sigma}^2$ is a biased estimator for σ^2.

a. Show the amount of bias in $\tilde{\sigma}^2$.

We know that $\tilde{\sigma}^2$ is a biased estimator for σ^2, and $\hat{\sigma}^2$ is an unbiased estimator for σ^2.

$\therefore E(\hat{\sigma}) = \sigma$

$$\begin{aligned} \tilde{\sigma}^2 &= \frac{\sum_{i=1}^{y} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n} = \frac{SSE}{n} \\ \hat{\sigma}^2 &= \frac{SSE}{n-2} = \frac{n}{n-2} \tilde{\sigma}^2 \\ \therefore E(\hat{\sigma}^2) &= E(\frac{n}{n-2} \tilde{\sigma}^2) = \frac{n}{n-2} E(\tilde{\sigma}^2) = \sigma^2 \\ \therefore E(\tilde{\sigma}^2) &= \frac{n-2}{n} \sigma^2 \end{aligned}$$

b. What happens to the bias as the sample size n becomes larger?

$\tilde{\sigma}^2$ gets closer to σ^2 as n becomes larger.
When n is enough large, $\tilde{\sigma}^2$ can be regarded as an unbiased estimator for σ^2.

2.33

Consider the least-squares residuals $e_i = y_i - \hat{y}_i, i = 1, 2, \ldots,$ from the simple linear regression model. Find the variance of the residuals $Var(e_i)$. Is the variance of the residuals a constant? Discuss.

$$\begin{aligned} Var(e_i) &= E[e_i - E(e_i)]^2 = E(e_i^2) - [E(e_i)]^2 \\ &= E(e_i^2) = SS_{RES} \\ &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \end{aligned}$$

From procedures above we can know that the variance of the residuals is **error sum of squares**, and it will become larger with the increasing number of the sample size n.

So it is not a constant.

SAS Code

```
/*2.6*/
proc import datafile='E:\Applied Regression Analysis\SAS\data-table-B3.csv' out = file_data;
getnames=yes;
run;
proc reg data=file_data;
model y=x1/clm alpha=0.05;
run;
proc reg data=file_data;
model y=x1/cli alpha=0.05;
run;

/*2.7*/
data file_data2;
input y x;
cards;
86.91   1.02
89.85   1.11
90.28   1.43
86.34   1.11
92.58   1.01
87.33   0.95
86.29   1.11
91.86   0.87
95.61   1.43
89.86   1.02
96.73   1.46
99.42   1.55
98.66   1.55
96.07   1.55
```

```
93.65   1.40
87.31   1.15
95.00   1.01
96.85   0.99
85.2    0.95
90.56   0.98
;
run;
proc reg data=file_data2;
model y=x/clb;
run;
data file_data2_new;
input x;
cards;
1.00
;
run;
data file_data2;
set file_data2_new file_data2;
run;
proc reg data=file_data2;
model y=x/clm alpha=0.05;
run;
quit;
```